Moleculaire modellering van de synthese van zeolieten
en gerelateerde nanoporeuze materialen

Molecular Modeling of the Synthesis of Zeolites
and Related Nanoporous Materials

Toon Verstraelen

Promotoren: prof. dr. ir. V. Van Speybroeck, prof. dr. M. Waroquier
Proefschrift ingediend tot het behalen van de graad van
Doctor in de Ingenieurswetenschappen: Toegepaste Natuurkunde

Vakgroep Subatomaire en Stralingsfysica
Voorzitter: prof. dr. D. Ryckbosch
Faculteit Wetenschappen
Academiejaar 2008 - 2009

UNIVERSITEIT
GENT

ISBN 978-90-8578-282-7
NUR 952, 928
Wettelijk depot: D/2009/10.500/40

Dit onderzoekswerk werd uitgevoerd binnen het Centrum voor Moleculaire Modellering en werd financieel ondersteund door het Instituut voor de Aanmoediging van Innovatie door Wetenschap en Technologie in Vlaanderen (IWT) in het kader van het SBO-BIPOM project en door het Federaal Wetenschapsbeleid (BELSPO) in het kader van het netwerk IAP 6/27.

Samenstelling van de examencommissie

Prof. Dr. ir. Dirk Aeyels
Prof. Dr. Paul W. Ayers (leescommissie)
Prof. Dr. Christine Kirschhock (leescommissie)
Prof. Dr. ir. Johan Martens
Dr. Ewald Pauwels
Prof. Dr. ir. Joris Thybaut (leescommissie)
Prof. Dr. Rutger van Santen
Prof. Dr. ir. Veronique Van Speybroeck (promotor, leescommissie)
Prof. Dr. Michel Waroquier (promotor)

Voorzitter: Prof. Dr. ir. Rik Van de Walle
Secretaris: Prof. Dr. Dimitri Van Neck

Dankwoord

Ik besef dat er binnenkort een belangrijk hoofdstuk in mijn leven zal afgesloten worden. De voorbije 2005 dagen ben ik langzaam veranderd van een frivole student in een gepassioneerd onderzoeker. Dit gaat uiteraard niet zomaar en gelukkig heb ik tijdens mijn doctoraat op veel mensen kunnen rekenen om dit waar te maken. Het hele proces heeft niet alleen geleid tot dit proefschrift, maar het was ook een enorme persoonlijke en professionele verrijking. Mijn dank gaat uit naar iedereen die me rechtstreeks of onrechtstreeks ondersteund en geholpen heeft. Enkele mensen verdienen toch een speciale vermelding.

In eerste instantie zou ik mijn promotoren, Veronique Van Speybroeck en Michel Waroquier, willen bedanken. Ze hebben de *job van mijn leven* mogelijk gemaakt en geven me ook de vrijheid om mijn creativiteit bot te vieren en mijn onderzoekstalent ongebreideld te ontplooien. Van hen heb ik ook de stiel geleerd, evenals de eerste wet van de academische jungle: *Publish or Perish*. Ze hebben me alle kansen gegeven voor de start van een mooie carrière. Dat zal ik nooit vergeten.

Michel, de geestelijke vader van het Centrum voor Moleculaire Modellering (CMM), waakt steeds over de balans tussen onze internationale competitiviteit en de aangename werksfeer. Op zijn memorabele culinaire uitspattingen zijn menig experimentele chemici stikjaloers. Veronique moet op dit vlak zeker niet de duimen leggen, maar ze fungeerde in eerste instantie toch als mijn constante kritische reality-check. Dankzij haar is het doel van al de ontwikkelde modellen en software nooit uit het oog verloren. Ook Dimitri Van Neck mag ik niet vergeten. Hij is een ongelooflijke theoretische *Brut Nature* die, samen met Paul Ayers, bij het schrijven van de GCM-paper de rode draad tussen de formules teruggevonden heeft.

Door de diverse samenwerkingsverbanden waarbij het CMM betrokken is, ben ik in contact gekomen met tal van nationale en internationale collega's. Mede door de vele discussies en kritische bedenkingen, telt dit werk mee in de internationale onderzoeksgemeenschap. Pierre Jacobs, Johan Martens en Christine Kirschhock van het Centrum voor Oppervlaktechemie en Katalyse (COK, KULeuven) waren de drijvende krachten achter het SBO-BIPOM project waarbinnen een groot deel van dit onderzoek uitgevoerd werd. Ook na dit doctoraat zal de samenwerking met het COK een blijvende bron van wetenschappelijke output vormen. Op dezelfde manier schept de samenwerking met het team van Rutger van Santen mooie toekomstperspectieven.

Paul, you deserve a separate paragraph. I do not dare to imagine where I would have been now without your help and inspiration. You are like an older brother, showing the way through the maze of theoretical chemistry. I hope to return the favor, but if it is at all possible, it will take at least a long time before this balance is restored. With your altruistic style, you are one of my few true

examples. Despite all this, you still act as a normal person. I hope our friendship remains so that we can battle at the frontiers of knowledge like brothers in arms.

Elke doctoraatsstudent gaat minstens door enkele morele dieptepunten voor hij de eindstreep bereikt. Gelukkig waren er genoeg mensen om deze dipjes te reduceren tot kleine lokale minima. Mijn bureaugenoten, Karen en Bart, zijn twee bijoukes. Ze klagen niet wanneer mijn hoofd uit frustratie weer in het toetsenbord valt of wanneer ik de ganse dag kwaad naar mijn computerscherm zit te kijken. Als het allemaal te veel wordt, kan er eens goed gezeverd worden en zijn we drie surrealistische wereldverbeteraars. Het hele CMM is hoe dan ook een zeer genietbare bende met wie het altijd tof is om pinten te pakken of te werken aan de meest diverse modelleringsproblemen.

Ook op het thuisfront wordt er steevast hevig gesupporterd, bijvoorbeeld wanneer er een artikel bijna aanvaard is. Mijn ouders, Myriam en Aloïs, hebben altijd het beste met me voorgehad. Ondanks mijn fleurige studentenleven hebben ze altijd geloofd dat het wel in orde zou komen. En voila, hier staan we dan. Pieter, Joke en Jan (broers en zus) en hun lieven, tante Julia, het hele Oosteneind, in het bijzonder Freddy en Sylvie met hun twee kapoenen, de burgies van weleer, Danny Pompier en Anne, ... ze moesten allemaal wel eens weten hoe het ging met mijn doctoraat. Nen dikke merci.

Ik heb dikwijls het uithoudingsvermogen van mijn liefste schat, Hera, op de proef gesteld. Ze heeft bijzonder goed stand gehouden. Met haar pa, Johan, heb ik nooit honger of dorst moeten lijden, integendeel. Isis, de tweelingzus, bleef verbazen met haar talent om steeds met beide voeten op de grond te landen. Ze heeft nu ook haar levensdoel, Wim, eindelijk gevonden. We hebben samen veel lol getrapt en dat zullen we hopelijk nog lang blijven doen. Een minder gekende en stille supporter is ons konijn, Candy, dat als een trouwe hond aan mijn zijde zit tijdens de slapeloze nachten. Ze is altijd vrolijk, net als geluk met twee oren, vier poten en een pomponneke.

Hera, ik weet niet hoe ik je moet bedanken voor de tijd die we samen doorgebracht hebben. De voorbije zes jaren waren zalig, en toch is dit alleen nog maar het begin. Er is niemand die me beter begrijpt dan jij, en er is niemand bij wie ik me meer thuis voel. Het is niet te verklaren, gelukkig maar.

Table of Contents

Nederlandstalige samenvatting

Zeolieten zijn microporeuze keramische kristallen waarvan het rooster hoofdzakelijk bestaat uit silicium, zuurstof en aluminium. De chemische samenstelling van zeolieten is vergelijkbaar met die van kwarts, maar de kristalstructuur is fundamenteel verschillend. Zeolieten bevatten regelmatige holtes (kanalen en kooien) met een maximum diameter van 1.2 nanometer, net groot genoeg om kleine moleculen te bevatten. De wanden van deze caviteiten vormen een groot intern oppervlak (orde 1000 m^2/g). Door de goed gedefinieerde vorm van de kanalen en kooien zijn ze enkel toegankelijk voor moleculen met de juiste vorm. Dit principe noemt men vormselectiviteit en maakt het mogelijk om chemisch gelijkaardige maar structureel verschillende moleculen van elkaar te scheiden. Bovendien is een zeoliet katalytisch actief door de aanwezigheid van Si/Al substituties. Al deze bijzondere eigenschappen zijn de drijfveer achter 60 jaar ontwikkeling van nieuwe types zeolieten en de uitwerking van nieuwe industriële toepassingen op basis van zeolieten. De jaarlijkse wereldwijde productie van zeolieten bedraagt 4.2 miljoen ton, met toepassingen in zeer diverse markten zoals bijvoorbeeld petrochemie, landbouw, veeteelt, farmacie, cosmetica, etc. De belangrijkste toepassingen in de westerse wereld zijn het petrochemisch kraken van ruwe olie, ionenwisselaars (ontharden en zuiveren van water) en de separatie en extractie van gassen en solventen.

Een recente ontwikkeling in dit veld is de zoektocht naar geordende micro- en mesoporeuze materialen. De IUPAC definieert een microkanaal als een porie met een diameter van 0.25 nm tot 2 nm, terwijl een mesokanaal een diameter heeft van 2 nm tot 50 nm. De beperkte diffusiesnelheid van gastmoleculen in de microporiën van een conventionele zeoliet is een gekende belemmering voor de aanvoer van reagentia en de afvoer van reactieproducten. De levensduur van een zeoliet wordt ook beperkt door de vorming van bijproducten die de toegang tot katalytisch actieve sites beperken. De regeneratie van de katalysator waarbij bijproducten verbrand worden, is een energieverslindende en onpraktische tussenstap. Men kan deze negatieve effecten vermijden of op zijn minst reduceren door het voorzien van mesoporiën. Deze mesokanalen fungeren als snelwegen voor het transport van gastmoleculen. Aan het Centrum voor Oppervlaktechemie en Katalyse (COK) van de KULeuven is een nieuwe route ontwikkeld voor de synthese van dergelijke micro- en mesoporeuze materialen met goed gecontroleerde afmetingen van zowel de micro- als de mesoporiën. Deze revolutionaire ontwikkeling heeft aanleiding gegeven tot het strategisch basis onderzoek (SBO) project naar biporeuze materialen (BIPOM's). Dit doctoraatsonderzoek is hoofdzakelijk uitgevoerd in het kader van het SBO-BIPOM project.

Een gedetailleerde beschrijving van de moleculaire mechanismen die aanleiding geven tot de nucleatie en de groei van zeolietkristallen is nog niet voorhanden. Men zou van zulke kennis zeer dankbaar gebruik kunnen maken

om nieuwe zeolieten te synthetiseren op maat van een beoogde industriële toepassing. Vooral de exacte rol van organische templaatmoleculen is nog niet volledig begrepen. Templaten bevorderen zeer selectief de synthese van welbepaalde zeolietstructuren waarvan de kanalen en kooien complementair zijn met de vorm van het gebruikte templaat. In de literatuur worden er ook verschillende modellen gepostuleerd voor de structuur van de precursoren die voorafgaan aan de vorming van zeolietkristallen. Er is wel reeds een indrukwekkende hoeveelheid experimenteel onderzoek verricht waardoor een aantal fragmenten van het verhaal gekend zijn, maar de vele resterende onbekenden vormen een onuitputtelijke bron van discussie. Moleculaire modellering is een onderzoeksgebied dat in deze context een complementaire rol speelt. Aan de hand van computersimulaties worden de interacties tussen moleculen bestudeerd en kan men net die eigenschappen bestuderen die (door de kleine schaal) moeilijk toegankelijk zijn voor het experiment.

Om de moleculaire interacties theoretisch te beschrijven zijn er twee grote families van modellen. In eerste instantie zijn er de kwantummechanische methodes die (bij benadering) het elektronische veeldeeltjesprobleem oplossen om de potentiële energie van een moleculair systeem te bepalen. Een belangrijke subklasse hiervan zijn de zogenaamde ab initio methodes die geen gebruik maken van empirische parameters en enkel gebaseerd zijn op de postulaten van de kwantummechanica. De tweede grote familie bestaat uit de moleculaire mechanica methoden. Dit zijn zuiver empirische modellen die met een zeer beperkte computationele kost moleculaire interacties benaderen. De keerzijde van de medaille is dat deze empirische modellen minder nauwkeurig zijn en enkele belangrijke processen zoals het breken en vormen van een chemische binding niet of slechts zeer benaderend kunnen beschrijven. Beide families worden aangewend om microscopische eigenschappen van moleculaire systemen te bepalen zoals optimale geometrieën, vibrationele modes, reactiemechanismes, enzovoort. Wanneer men echter chemische reacties betrouwbaar wil beschrijven zijn enkel kwantummechanische methodes toereikend. Aan de hand van statistische fysica worden deze microscopische data vertaald in relevante macroscopische grootheden. Moleculaire dynamica is in deze context een zeer veelzijdige techniek. Door het integreren van de bewegingsvergelijkingen van een moleculair systeem, doorloopt men alle relevante microscopische toestanden bij een gegeven druk en temperatuur. Dit laat ook toe om rekening te houden met de moleculaire omgeving zoals een solvent of een zeolietrooster. Met behulp van statistische technieken worden uit het ensemble van microscopische toestanden macroscopische parameters afgeleid.

Deze thesis behandelt de verschillende aspecten van moleculaire modellering die aangewend zijn om nieuwe inzichten te verschaffen in de synthese van zeolieten. Om het gevoerde onderzoek mogelijk te maken, is er nieuwe software ontwikkeld. Het meest zichtbare voorbeeld is ZEOBUILDER,

een programma om atomaire modellen van biporeuze zeolieten te construeren. De geometrische modellen gemaakt met ZEOBUILDER dienen als input voor moleculaire simulaties. Een ander belangrijk onderdeel van dit werk is de ontwikkeling van nieuwe moleculaire mechanica modellen die efficiënte simulaties toelaten op modellen van zeolietprecursoren. In het bijzonder werden er nieuwe methodes voorgesteld om betrouwbare krachtveld parameters te schatten. Deze technologische en methodologische hulpmiddelen hebben we ingezet in theoretische studies van de verschillende tussenstappen in de synthese van zeolieten.

Aan de hand van moleculaire dynamica simulaties zijn infrarood spectra van zeolietprecursoren en -nanoblokjes afgeleid. Dit onderzoek heeft de experimentele observatie bevestigd dat de verschuiving van de zogenaamde MFI-fingerprint in het infrarood spectrum kan geassocieerd worden met de vorming van nanoscopische zeolietkristallen.

Het tetrapropylammonium (TPA) templaat is gekend voor de synthese van MFI-gestructureerde zeolieten zoals ZSM-5 en Silicaliet-I. Met diverse modelleringstechnieken hebben we de interacties van TPA bestudeerd met enkele zeolietprecursoren voorgesteld door het COK. De belangrijkste conclusie is dat TPA bij voorkeur een positie inneemt ten opzichte van de zeolietprecursoren waar in een latere fase de kruising tussen twee types kanalen ontstaat. Dit is in overeenkomst met NMR metingen op MFI-gestructureerde zeolieten waaruit de TPA moleculen nog niet verwijderd zijn. Zonder de aanwezigheid van TPA zijn de onderzochte precursoren instabiel en vallen de voorlopers van kanalen en kooien in elkaar. Dit bevestigt de functie van templaatmoleculen, namelijk dat ze nodig zijn voor de stabilisatie van de initiële caviteiten in zeolietprecursoren.

De ontwikkelde software en theoretische modellen in deze thesis vormen een eerste belangrijke bouwsteen van de modellering van zeolietsyntheseprocessen. De verdere ontwikkeling van polarizeerbare krachtveldmodellen en de implementatie van deze modellen in moleculaire dynamica software zullen in de toekomst moeten verder gezet worden om bijkomende inzichten te verwerven in de moleculaire mechanismen die aanleiding geven tot de synthese van zeolieten.

English Summary

Zeolites are microporous inorganic crystals with a framework that mainly contains silicon, oxygen and aluminum. The chemical composition is comparable to quartz, but the crystal structure is fundamentally different. Zeolites contain regular cavities (channels and cages) with a maximum diameter of 1.2 nanometer, large enough to contain small organic molecules. The walls of the cavities represent a huge internal surface (of the order of 1000 m^2/g). Due to the well-defined shape of these channels and cages, only the molecules with a compatible structure can enter the internal cavities of a zeolite. This principle is called shape selectivity and it enables the separation of chemically similar but structurally different molecules. A Si/Al substitution in the framework introduces a catalytically active site on the walls of the zeolite cavities. During the past 60 years, all these specific properties have driven the development of new zeolites and the implementation of zeolites in industrial applications. The world-wide production of zeolites amounts to 4.2 million tons per annum, with applications in diverse markets such as petrochemistry, agriculture, animal husbandry, pharmacy, and so on. The most important applications in the western world include catalytic cracking in the petrochemical industry, ion exchange (softening and purification of water), and the separation and extraction of gases and solvents.

A recent development in this field is the search for micro- and mesoporous materials. The IUPAC defines a microchannel as a pore with a diameter ranging from 0.25 nm to 2 nm, while a mesopore has a diameter between 2 nm and 50 nm. The limited diffusion rate of guest molecules in the micropores of a conventional zeolite is a well-known obstacle for the transport of reagents and reaction products. Another issue is the formation of side-products that reduce the diffusion rates and block the access to catalytically active sites. The regeneration of a zeolite catalyst, in which side products are burned, is an expensive and impractical procedure. One can avoid or at least reduce these limitations by introducing mesopores in the zeolite that have a diameter of the order of 1 to 50 nanometer. These mesopores act as highways for the transport of guest molecules. The Center for Surface Chemistry and Catalysis (COK) of the KULeuven has developed a new route for the synthesis of such micro- and mesoporous zeolites with well-controlled sizes of both types of pores. This revolutionary concept was the onset for the strategic basic research (SBO) project on biporous materials (BIPOM's). This PhD is mainly carried out in the context of the SBO-BIPOM project.

A detailed description of the molecular mechanisms that give rise to the nucleation and growth of zeolite crystals is not yet available. Such insights are however of great interest for the tailor-made synthesis of zeolite catalysts in industrial applications. One of the major questions is the exact role of organic template molecules. Templates direct the synthesis process in a selective way to the formation of zeolites whose channels and cavities are

structurally complementary to the shape of the template molecule. Another controversy is the exact structure of the precursors that precede the formation of zeolite crystals. A few (incompatible) models are postulated in the literature. Although several fragments of the synthesis mechanism have been unraveled through an impressive history of experimental studies, the remaining unknowns are not easily disentangled. Molecular modeling is a research field that plays a complementary role in this context. With the aid of computer simulations, one gains insights into molecular interactions that are not easily accessible to the experiment due to the small scale.

There are two large categories of models to describe molecular interactions on a theoretical basis. In first instance there are the quantum-mechanical methods that solve the electronic many-body problem (approximately) to obtain the potential energy of a molecular system. An important subcategory are the ab initio methods, which rely only on the elementary quantum-mechanical postulates and do not depend on any empirical input. The second large category consists of the molecular mechanics methods that approximate molecular interactions with mainly empirical models. These models are computationally very efficient, but the downside is that molecular mechanics methods have only a limited accuracy and can not (or at least not correctly) describe important chemical processes such as the formation and the breaking of chemical bonds. Methods from both categories are applied to obtain microscopic properties of molecular systems such as optimal geometries, vibrational modes, reaction mechanisms, and so on. Only the quantum mechanical methods are reliable for the description of chemical reactions. Statistical mechanics provides the theoretical foundations to translate these microscopic data into relevant macroscopic observables. Molecular dynamics is a very versatile technique to apply the laws of statistical physics. By integrating the equations of motion of the molecular system, one runs through all the relevant microscopic states at a given temperature and pressure. Such simulations take into account the complete molecular environment such as a solvent or a zeolite framework. By using the proper statistical techniques, one can derive macroscopic parameters as averages over the ensemble of microscopic states.

This thesis covers all the aspects of molecular modeling that we have addressed to gain more insights into the synthesis of zeolites. The larger part of the research deliverables in this PhD were only possible through the development of new software. The most visible example is ZEOBUILDER, a computer program to construct atomic models of biporous zeolites. The geometric models made with ZEOBUILDER are an essential starting point for molecular simulations. Another important aspect of this work is the development of molecular mechanics models that allow efficient simulations on zeolite precursors. In particular, we proposed new methods to obtain reliable force-field parameters. These technological and methodological tools

are applied in theoretical studies on distinct intermediate steps of the synthesis of zeolites

Infrared spectra of zeolite precursors and zeolite nanocrystals have been derived from molecular dynamics simulations. This investigation confirms the experimental observation that a shift of the so-called MFI-fingerprint in the infrared spectrum can be associated with the formation of nanoscopic zeolite crystals.

The tetrapropylammonium (TPA) template is known for the synthesis of MFI-structured zeolites such as ZSM-5 and Silicalite-I. We have studied the interactions between TPA and zeolite precursors proposed by the COK, using a broad scala of modeling techniques. The principal conclusion is that the preferential position of TPA with respect to the precursors corresponds to the place where the crossing of two channels will form in a later phase of the synthesis. This is in agreement with NMR measurements on macroscopic MFI-structured zeolite crystals from which the TPA has not been removed yet. Without the presence of the TPA, the investigated precursors would collapse, which is a confirmation of the function of the template molecule: providing a structural support for the initial zeolitic species.

The new software and theoretical models in this thesis are the cornerstone for the modeling of zeolite synthesis processes. The development of polarizable force fields must be continued and their implementation in molecular dynamics software is indispensable to gain additional insights in the molecular mechanisms that lead to the formation of zeolites.

General Introduction

The Swedish mineralogist Axel Frederik Cronstedt originally discovered zeolites in the 18th century.[1] Upon heating, he observed that water evaporated from the natural mineral, which lead to the name "zeolite" based on the ancient Greek for "I boil" (zeo) and "stone" (lithos). This curious property is a direct consequence of the microporous structure of the zeolite crystal, which is a periodic arrangement of pores and cages with a radius from varying 0.4 to 1.2 nanometer, separated by a rigid framework with the following chemical composition: $M_{x/n}Si_{1-x}Al_xO_2$. The cavities in the crystal structure are large enough to accommodate small guest molecules such as water, small organics and cations. The framework is built from tetrahedral SiO_4 or $Al0_4$ units that are connected at the corners by a shared oxygen atom. The Al atoms are negative charge defects that must be compensated by cations M such as H^+, Na^+, Mg^{2+} and so on. A model representation is given in figure 1. Today, there are in total 186 known zeolite crystal structures of which 40 occur naturally. The larger part of the zeolite structures is only present in synthetic minerals that have been developed over the past 50 years for commercial and research purposes. Each crystal structure is identified with a three-letter code, e.g. MFI, FAU, and so on. The Structure Commission of the International Zeolite Association (IZA) assigns such a three-letter code to each unique and confirmed structure, and regularly publishes the complete database in The Atlas of Zeolite Framework Types.[2] For each crystal structure, there exist multiple chemical compositions of the framework, e.g. Silicalite-1 and ZSM-5 both have the MFI structure but the latter has a lower Si/Al ratio.

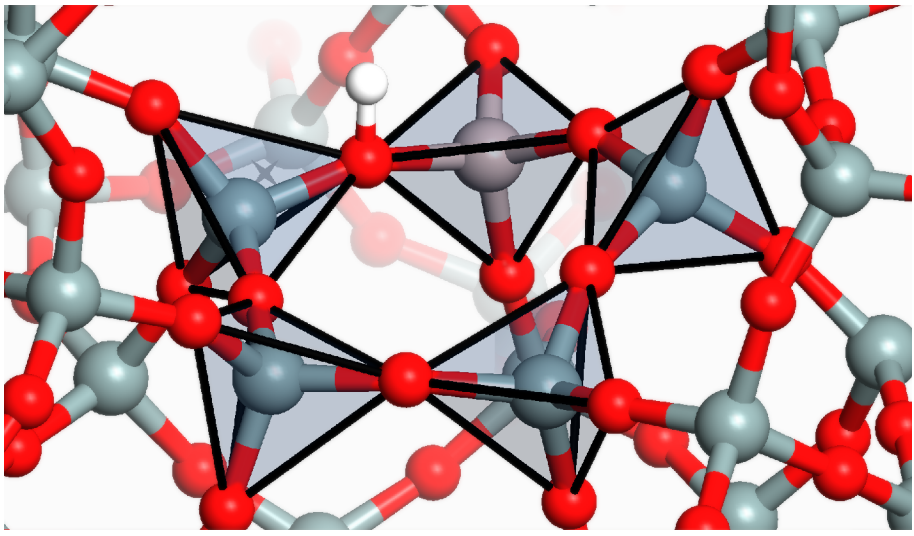


Figure 1: Model representation of a zeolite. The T-atoms (at the center of each tetrahedron) are silicon (blue gray) or aluminum (purple gray). The corners of the tetrahedrons or oxygen atoms (red). The negative charge defect due to the aluminum is compensated by a proton (white).

From all ceramic materials, zeolites are encountered in the widest variety of industrial and domestic applications. They act as shape-selective molecular sieves that can separate similar but structurally distinct molecules thanks to their well-defined micropores. The walls of the internal cavities represent a huge internal surface (of the order of 1000 m^2/g) that can host catalytically active sites. Since only properly shaped reagent molecules can enter the zeolite and likewise only certain types of products can leave, zeolites are used as efficient shape-selective and thus environmentally friendly heterogeneous catalysts. Zeolites are also used as ion-exchangers since certain charge compensating cations bind preferentially over others. These three principal properties find their applications in diverse fields such as catalytic cracking of olefins, oil refining, softening and purification of water, separation and extraction of solvents and gases, controlled drug delivery in medicine, and so on. The adoption of zeolites is driven by environmental regulations and their superior properties in various applications.[3,4]

During the past sixty years, many new zeolites and related zeotype materials were synthesized in an attempt to design new catalysts and sorbents or to device cheaper production processes for known zeolites. A detailed overview of the history of zeolite synthesis is given by Cundy and Cox.[5] We will only highlight the events that are of importance for this thesis. From 1940 to 1960, the experimental foundations of the hydrothermal synthesis of zeolites were established. Until 1961, the synthesis mixtures only contained inorganic ingredients. Two groups proposed simultaneously new recipes that involved quaternary ammonium cations and synthesized more silica-rich zeolites.[6,7] In the following years new and even more silica-rich zeolites were synthesized with similar procedures. Already in these early days, different opinions on the actual synthesis mechanisms were circulating. The two competing proposals were gel rearrangement into a zeolite crystal versus gel dissolution into the solvent and consequent aggregation of solute precursor species into a separate crystal phase. At present both mechanisms are still supported, but the picture is extended with the role of colloidal particles (amorphous silica or nanocrystals surrounded by a surfactant) in the solvent. Another milestone was reached in 1972 with the synthesis of ZSM-5[8], which initiated a decade of new developments. In 1977 a different synthesis procedure for ZSM-5 was proposed without organic ingredients.[9] Most noticeable was the first all-silica zeolite, Silicalite-I, in 1978[1] which is isostructural with ZSM-5. This course of events shows that the mechanism for the formation of the MFI structure is probably not unique, and that evidence for a certain mechanism does not exclude other possibilities. In 1982 the AlPO (Aluminum Phosphate) class of molecular sieves was discovered, containing no silicon at all.[5] In 1988 the first AlPO was generated that broke the maximum 12T ring size of previous molecular sieves.[12] The number of tetrahedrons making up the size-limiting ring determines the maximum size of guest molecules, and consequently also

affects the potential applications. In 1983 the first zeolite with Silicon/Titanium substitutions was discovered and proved not only to be a very useful catalyst, but also demonstrated the feasibility of a wider range of elements acting as T-atoms in zeolite crystals.[13] A next step was the synthesis of mesoporous silica in 1990, which contain regularly ordered and relatively large pores (2 to 50 nm diameter) that are separated by amorphous silica walls.[14,15,16] Compared to conventional zeolites, mesoporous silica have much higher transport coefficients but lack shape selectivity and are not stable under high temperature or pressure.

The invention of mesoporous silica has inspired many researchers to combine the benefits of conventional zeolites (shape selectivity and thermal stability) and mesoporous silica (less diffusion limitations and bypassing of local cokes formation in catalytic applications). Several approaches have been proposed in the literature and we refer to a review of Tao et al.[17] for a detailed discussion. Although there have been attempts to recrystallize the amorphous walls of mesoporous silica, many methods follow the opposite procedure and impose mesopores in the zeolite crystal. One can explicitly cast mesopores into conventional zeolites, either by leaching, steaming or other chemical treatments. Other techniques include the synthesis of nanosized zeolite particles, leaving mesopores between stacked particles. One can also include mesoporous channels and cages during a conventional synthesis process with mesoscopic carbon templates (nanotubes, wires, ...), carbon aerogels or polymer aerogels. Both regular micro and mesopores with a narrow size distribution are of great importance to control the shape selective properties and to guarantee a homogeneous accessibility of micropores.

The Center for Surface Chemistry and Catalysis (COK) of the KULeuven is prominent in this field, and hence, involved in many national and international networking programs. Zeotiles and zeogrids are biporous materials discovered by COK.[18-29] These new zeotype materials arise from a unique approach to control accurately both micro- and mesopore sizes, based on the clear-solution synthesis of zeolite precursors and nanocrystals with a well-defined structure and shape. Nanoslabs, nanotablets, and other nanocrystals depicted in figure 2 have the intrinsic properties of the MFI topology such as the 3D channel structure. These nanosized particles are arranged in a specific order imposed by block copolymers that act as secondary templates for the mesostructure. The well-defined size of the nanoparticles and the choice of the secondary template determine the size of the mesopores. The existence of well-defined colloidal nanoslabs is the subject of serious debate in the literature[30], which is no exception compared to previously proposed zeolite synthesis mechanisms.[5] However, recently different models on the aggregation of nanoscopic zeolite precursors started to merge.[31-33] All aspects that were necessary to bring the nanocrystal based biporous zeolites to the application level, were investigated in the strategic

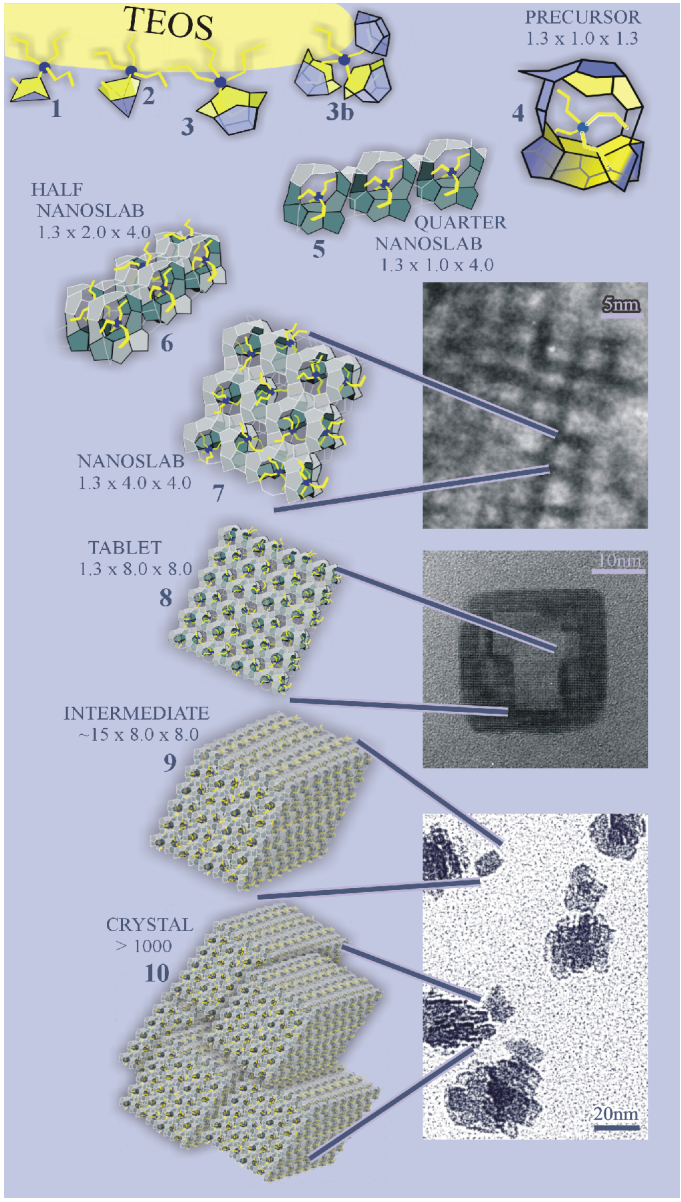


Figure 2: Overview of the MFI-precursor based synthesis hypothesis for Silicalite-I. In the initial stages (1-3) surfactant TPA shields the hydrophobic parts of the initial silica oligomers from the solvent. The colloidal 33T MFI precursor (4) bears already the topological features of the Silicalite-I crystal. Aggregation of precursor units leads to nanocrystals (5-9) and finally to macroscopic particles (10). The slabs, obtained during extraction, are not necessarily present during the synthesis, and are not yet of perfect crystallinity. Dimensions are given in nanometer and TEM images are given for the final stages.
(Image kindly provided by Christine Kirschhock of the COK, KULeuven.)

basic research program (SBO) on biporous materials (BIPOM's). This includes the assessment of catalytic, sorption and separation properties, the exploration of new zeotile and zeogrid systems, the development of nanocrystal-based membranes, a search for industrial processes whose sustainability can be seriously improved with the advent of biporous materials and the molecular modeling of the synthesis mechanisms that lead to the formation of nanocrystals. The larger part of this PhD is carried out in the context of the SBO-BIPOM project. The molecular modeling task is the subject of this thesis and we will outline the work, the choices made, and the results in the remainder of this general introduction.

Molecular modeling is the theoretical study of interactions and processes at the molecular scale. With the aid of statistical mechanics these data are translated into meaningful macroscopic observables. In general, one assumes that the Born-Oppenheimer approximation is valid and that atomic nuclei can be treated as classical particles. In practice, this means that the potential energy of a molecular system is fully determined by the position of the atomic nuclei and that the electronic degrees of freedom are always in their ground state. There are two approaches to compute the molecular energy. The first one is often called the quantum mechanical (QM) approach. The Schrödinger equation of the many-body electron problem is solved (approximatively) to obtain the molecular energy and other properties of interest. The second approach is called Molecular Mechanics (MM) and uses a sum of empirical interatomic potentials to obtain the molecular energy.

QM methods are typically more accurate by construction and there is a wide range of levels of theory, which allows an appropriate trade-off between computational cost and accuracy. The poor description of Van Der Waals interactions is one of the well-known drawbacks, which is only cured in the most computationally expensive QM methods. MM methods are far more efficient in terms of computer resources, but the assessment of the reliability of the empirical parameters in an MM model is difficult. Most MM methods cannot describe chemical reactions because it is assumed that the molecular bond graph does not alter during a simulation. The term force field (FF) is a common synonym for a molecular mechanics model.

Given an appropriate level of theory to compute molecular interactions, there are several techniques to derive macroscopic properties, which can be divided roughly into two classes: static and dynamic computations. The accessible time and length scales, shown in figure 3, depend on the computational cost of the level of theory. When the intrinsic computational cost becomes too high, one is limited to static calculations. First the ground state configuration of the nuclei at zero Kelvin is determined and consequently all thermodynamic properties are computed with a second order extrapolation of the potential energy

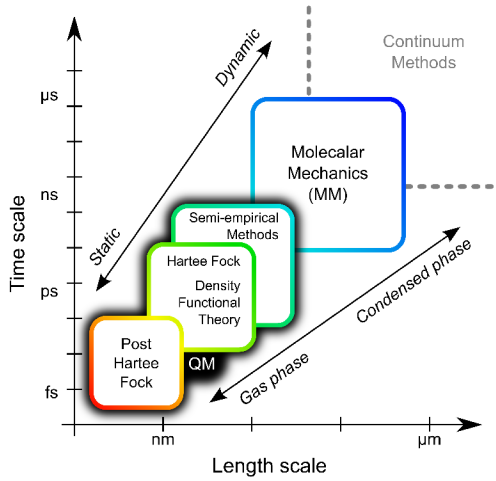


Figure 3: The relevant time and length scales in molecular modeling. Each region in this plot is associated with specific methods to compute the potential energy of the molecular system.

surface. A static calculation is a good approximation for systems in the gas-phase that only have a few important conformers. When one is interested in solvent effects and contributions of a vast number of conformers, the only reliable techniques are molecular dynamics (MD) and Monte Carlo (MC) simulations. Those techniques sample a representative set of states at a finite temperature to compute thermodynamic averages, taking into account multiple local minima on the potential energy surface. Proper averages require a large number of samples and the computational cost of one sample must be small, otherwise dynamic techniques are not feasible. Therefore one often uses force fields to compute the energy during MD or MC simulations. Due to a steadily increasing computer power and algorithmic developments that allow more efficient quantum mechanical simulations, the popularity of QM/MD and QM/MC is systematically increasing.

From the theoretical viewpoint the final (and for now still unattainable) goal is a complete understanding of the zeolite synthesis mechanisms from molecular simulations. This would allow a true de novo design of zeolite catalysts on a computer, by "simply" benchmarking the effectiveness of a broad scala of hypothetical synthesis conditions with computer simulations. An exact and complete molecular mechanism starting from the initial ingredients towards the final crystal is not readily available in the literature. There is still a long way to go, but this thesis is a modest step forward. In this work, we investigated some of the crucial intermediate steps of the zeolite synthesis scheme based on nanocrystals, both using QM and MM techniques.

In addition to the purely scientific results from our modeling studies, it also became clear that, for the study of zeolite synthesis mechanisms, the quality of the currently available molecular mechanics potentials is inadequate. The development of such an empirical model is a laborious and long-term struggle. To make the situation even worse, there are no clearly defined procedures in the literature for deriving force-field parameters successfully. We used this obstacle as an inspiration to search for well-defined protocols that eventually lead to a complete and reliable molecular mechanics model. The development of such a model is still in progress but the fundamentals of the new approach are definitely settled, and preliminary numerical applications are very promising.

A large part of this PhD thesis is devoted to model and software development. New software tools have been developed that are indispensable to tackle research domains that would otherwise remain practically inaccessible. They are discussed in the first section. The theoretical models and methodological developments are outlined in the second section. The third section gives an overview of the modeling applications concerning zeolite synthesis, which extensively relied on the efforts discussed in the first two sections. Finally some conclusions and prospects on future research are given. The thesis has already resulted in 11 papers in ISI journals. They are included in the remainder of this thesis.

Software Development

Software development is crucial in the field of molecular modeling, since the scientific output is mainly based on computer simulations. An overview of the technical work flow of a typical molecular modeling study is given in figure 4. One recognizes three stages in this scheme. In the first place one must construct an input file with all the atomic coordinates of the molecular system, which is often an initial guess of the geometry. One also needs to decide on the level of theory (QM or MM), which can be readily available in the literature, or which must be developed for the application in mind. All these data are the input for molecular simulation software that use theoretical models to predict static or dynamic microscopic properties of the molecular system in the input file. The final step is the interpretation of the simulation output, which often involves the transformation of microscopic data into macroscopic observables based on statistical mechanics.

One uses computer programs for each step in this work flow. The central part is a computationally intensive task that is typically performed with well-known software packages used in most molecular modeling research groups. Such simulation software is non-interactive and computationally efficient. Computer programs in the preparation and post-processing steps deal with highly specific tasks that are not computationally intensive, but involve a

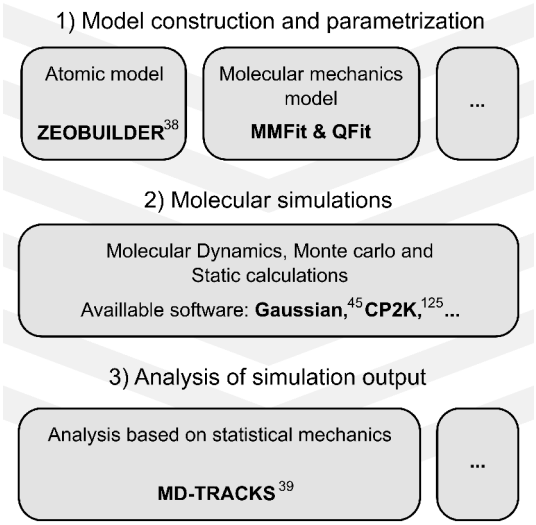


Figure 4: The technical work flow followed during a molecular modeling study.

higher level of complexity, such as the interpretation of results, development of a representative model, or the outline of successful research strategy. This remarkable difference has an important consequence on the software development process. For molecular simulations the computing time is the bottleneck, while for preparation and post-processing tasks, the time to implement and manage the algorithms is the limiting factor. For the former, one typically uses compiled languages such as Fortran, C or C++.[27,28] For the latter, high-level interpreted languages and functional programming (spreadsheets) are more suitable. A high-level language does not require knowledge of computer technical details and leaves more time to transform ideas and algorithms into a computer implementation.

The software discussed in this section belongs to the category of preparation or post-processing. 99% of the code is written in Python[34,35] and only a few time-consuming routines are written in C or Fortran. There are several reasons why Python was chosen over other high-level interpreted programming languages such as Ruby, Perl, BASIC, Matlab, R, Tcl, and so on. Although most of these motivations are discussed extensively in a dedicated issue of Computing in Science and Engineering,[35] we summarize the important factors below.

- The Python language is easily adopted by newcomers. The syntax is minimalistic and designed for readability.
- A selected number of concepts from object oriented programming are present, just enough to write and refactor complex programs quickly.

This is especially useful for model development work. One can benchmark a large variety of ideas (e.g. energy terms in a force field) in little time without writing spaghetti code.

- Python supports a large number of high-level language concepts that are efficient and easy to use: lists, associative arrays (dictionaries), unordered sets, iterators, generator expressions, duck typing, garbage collection, exception handling, and so on. This allows a quick implementation of complex algorithms.
- There is a huge amount of specialized libraries for Python. The most important for this work are NumPy (array operations based on BLAS and LAPACK), MatPlotLib (scientific plots), SciPy (a broad range of scientific and numerical tools), PyOpenGL (3D visualization) and PyGTK (graphical user interfaces).
- It is easy to call Fortran and C routines from Python without a noticeable overhead. We can isolate the computationally intensive routines and write them in a low-level language, while the large remainder of our programming efforts benefit from the advantages that are inherent to Python.
- Python, all the software libraries used for this work, and nearly all other Python libraries (thousands!) are free and open source software. There are no limitations on their usage. This is not only profitable, but we can also build on a solid foundation that will prevail. Python is maintained by a very diffuse group of developers.

Python has one major drawback: it is not useful for pure computationally intensive programs. When the computational bottlenecks in a program can not be isolated in small routines, there is no simple solution to improve the runtime performance. One could argue that C++ offers the advantages of both worlds, being a high-level language that can be compiled into efficient machine code. Unfortunately C++ is relatively difficult to learn and to apply, inhibiting a rapid application development. A new successor of C/C++, namely D, recently emerged and is an attempt to combine the advantages of C++ and recent high-level interpreted languages such as Python or Ruby. Many of the advantages of Python listed above are also present in D. The language is not commonplace yet, but it should be on the radar of those who are interested in computational software development.

Zeobuilder

The first milestone in this work is the construction of atomic models of the various silica oligomers, zeolite precursors and zeolite nanocrystals that are experimentally observed during the synthesis of biporous zeolites. Atomic models are the prerequisite for further theoretical studies and are the input

for molecular simulations. Starting from the well-known periodic MFI structure, several manipulations are required to obtain properly terminated nanocrystals. Suitable software was simply not available at the beginning of the SBO-BIPOM project, probably because complex operations such as joining inorganic nanocrystals into larger units were never encountered before. Kriber, DLS-76[36] and GULP[37] are the (non-graphical) legacy software tools that can be used to build models of conventional zeolites, but are only useful in conventional crystallographic problems where the building units are individual atoms. We have considered the extension of existing graphical software with extra capabilities to build nanocrystals, but both technical and/or legal limitations do not allow this. Therefore, we have written a new

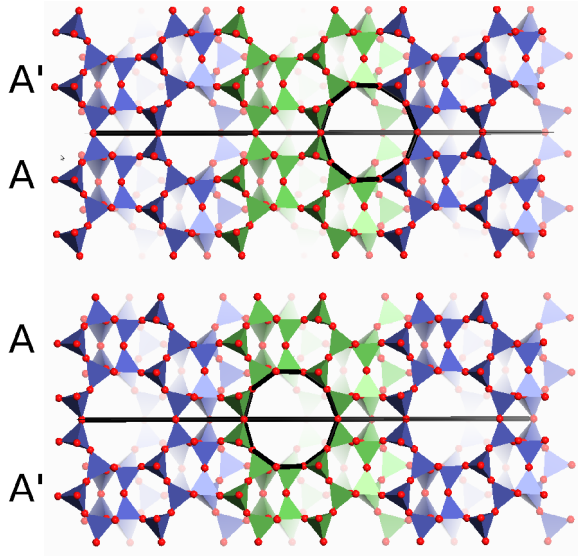


Figure 5: Two different models for a half MFI nanoslab. The black line corresponds to the reflection plane discussed in the text. Alternating colors are used to distinguish the three precursors.

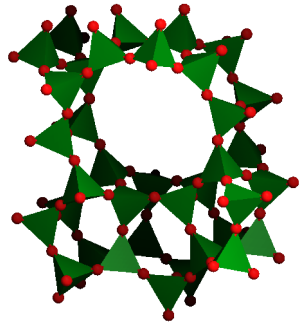


Figure 6: Model representation of the 33T MFI-precursor.[20]

program from scratch, ZEOBUILDER[38], with a software design that allows complex geometric manipulations on molecular systems. A survey of similar software reveals that most programs are not easily extensible with new capabilities. For this reason, extensibility is the second big feature of ZEOBUILDER. The larger part of the program consists of plug-ins that implement the actual features of interest. A small core program implements all the technical aspects that are not of interest from the scientific point of view such as communication with graphics hardware, the data structure to represent the model, management of plug-ins, and so on. Adding a new feature to ZEOBUILDER is only a matter of writing a new plug-in file. At present, the scope of ZEOBUILDER goes beyond the realm of zeolites with applications in biochemistry and inorganic chemistry in general. We refer to paper 1 for a detailed description of ZEOBUILDER.

The first applications of ZEOBUILDER was the construction of a model for Zeotile-I.[25] The instructions to build the model are explained in detail in paper 1. Several variations of these instructions lead to genuinely different atomic models that also match the experimental picture of Zeotile-I. Figure 5 shows for example two models of a half MFI nanoslab.[19] This nanocrystal is an arrangement of six 33T MFI precursors,[20] shown in figure 6. The MFI precursor unit is repeated three times along the a axis of the MFI crystal (with a pure translation). Consequently a copy of the row of three precursors is reflected with respect to a mirror plane orthogonal to the [010] direction. There are two possible positions for this reflection plane, each resulting in a nanocrystal of the same size, with the same type of micropores, but with a different atomic structure. Note that both outcomes are not each other's mirror images. With these two variants, one can again make different types of double units that correspond to the tiles in the Zeotile-I. It is plausible that the different variations of this tile are present in one Zeotile-I sample, which could be an additional reason for the absence of visible microstructure in XRD and HREM measurements.[25]

The atomic models used in paper 6 and paper 7 are constructed with ZEOBUILDER. The program is also used in ongoing research to model silica nanotubes and to prepare Oniom simulations in Gaussian03. At present, new plug-ins for ZEOBUILDER are under development at the Center for Molecular Modeling and at the Catalysis department of the TU Eindhoven. Upcoming features include a convenient tool to build small molecules, an improved user interface and a communication layer between ZEOBUILDER and computational applications.

The MolMod software library is a component of ZEOBUILDER, which is also exploited as a separate tool to process large molecular databases in a non-interactive way. It contains elementary data structures for molecular geometries, molecular graphs, results from molecular simulations, ... and

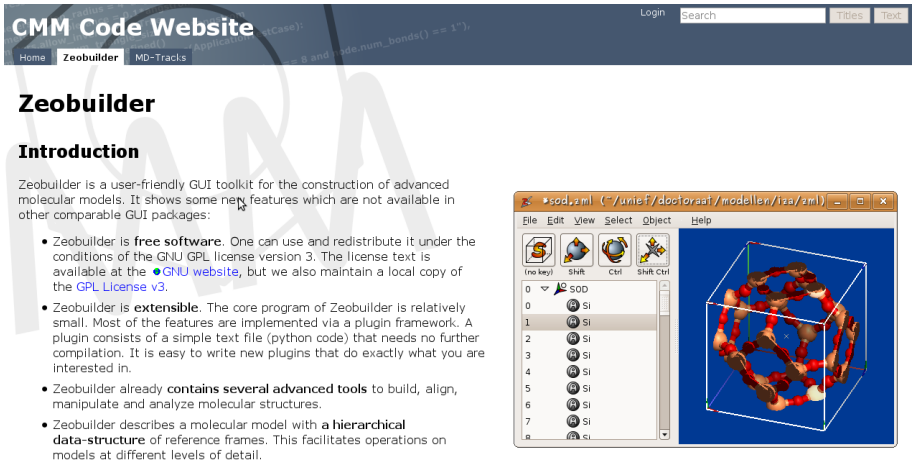


Figure 7: The CMM code Website: http://molmod.ugent.be/code/

algorithms to construct, analyze and manipulate these data. Both methodological papers 3 and 4 are based on large datasets with thousands or millions of molecules. These papers would have been intractable without the MolMod library.

Practical information to use (and extend) ZEOBUILDER is available on the CMM Code website (http://molmod.ugent.be/code). A screenshot of the website is given in figure 7. This website contains installation instructions and a tutorial that covers all the features in the most recent version of ZEOBUILDER. The latest developments in the code can be followed with the git repository viewer. One can subscribe to mailing lists and there is an on-line bug tracker. In the near future we will provide more technical documentation concerning ZEOBUILDER plug-ins and batch scripts based on the MolMod library.

MD-Tracks

An essential aspect of Molecular Dynamics and Monte Carlo simulations is the statistical analysis of the output. MD and MC Simulations perform millions of iterations to integrate equations of motion or to sample statistical distributions respectively. At each iteration the state of the molecular system changes, and the collection of all these states is called the "trajectory data". Because of the ergodic principle, the trajectory data are a representative subset of a thermodynamic ensemble for the molecular system under scrutiny. Thermodynamic quantities can be derived, for example by taking the proper averages. In practice, there is a wide variety of statistical techniques to analyze the trajectory data, which can be implemented on demand in high-level languages such as R and Python. This approach of ad-hoc implementations is

not optimal, mainly due to a lack of software reusability. Before the advent of MD-Tracks[39], we often had to alter previously written scripts and programs to make them compatible with new trajectory file formats or to apply them in slightly different circumstances. MD-Tracks introduces a compact, efficient and cross-platform trajectory database format to solve both issues. Before the actual analysis, the trajectory data is converted to the database format. Then one employs a collection of MD-Tracks scripts that operate in this database to analyze the simulation output. Most MD-Tracks scripts can be applied to any type of trajectory data (atomic positions, internal coordinates, velocities, all kinds of energies, and so on), and when possible the output is also stored in the same database so that it can be further processed with other tools. An exposition on MD-Tracks is given in paper 2, including a detailed description of the software design and step-by-step examples.

Prototypes of MD-tracks and ad-hoc scripts have been used for the analysis of MD simulations in paper 11, 12 and 13. A beta version of MD-Tracks was tested in the work on the MFI-fingerprint (paper 6). The current version of MD-Tracks was extensively used in the master thesis of Marc Van Houteghem[40], and also by other members of the CMM.

A recent CECAM workshop[41] illustrates the importance of a standard database format for molecular dynamics and Monte Carlo trajectory data. It is obviously advantageous that different molecular dynamics programs would write their output in a common format for post-processing purposes. This avoids the file conversions in MD-Tracks and also facilitates the usage of multiple simulation codes in one study. The database format in MD-Tracks must be upgraded before it can be used for this purpose. There are technical limitations that prohibit the analysis of simulations with millions of atoms, which is a common system size in some molecular modeling applications, e.g. research on crack propagation in metals. A redesign of the database format is therefore the next challenge for MD-Tracks. We can take advantage of existing formats for large amounts of numerical data such as netCDF or HDF5, and define an additional set of conventions that guarantee interoperability of trajectory data. This new format should be a good prototype for the adoption in various MD or MC simulation codes.

To be released: HiPart, TAMKin, MMFit and QFit

During my PhD, many software projects were initiated. This subsection is dedicated to the projects that have not been announced yet in scientific publications, nor are they available on our website. Some of these programs are already mature, while others are still evolving. A short description of each project will be given below, including some future prospects. The discussion starts with HiPart and TAMKin, which are already in use by other CMM members. Two other projects, MMFit and QFit, were only used for this thesis.

For the work on the empirical molecular electrostatic models (vide infra) one needs reference values for atomic partial charges for a large set of molecules to estimate empirical parameters. Atomic partial charges can be obtained from quantum mechanical computations, although a rigorous definition of the atomic partial charge does not exist. Hence there are various schemes to compute partial charges, which are all different algorithms that divide the electronic charge distribution into atomic contributions. One of the more recent algorithms is the Iterative Hirshfeld scheme.[42] It has several interesting properties for the derivation of empirical models, such as its robustness with respect to the basis set used in quantum mechanical calculations,[43] and a good reproduction of the molecular electrostatic potential.[44] Since this method is only very recent, it is not readily available in the conventional molecular simulation software. Therefore, we implemented HiPart, which is a post-processing tool for Gaussian03[45] that can derive normal Hirshfeld charges[46], Iterative Hirshfeld charges[42] and ISA charges[47]. The software design of HiPart facilitates the derivation of related properties such as a complete multipole expansion of each atom. This will be used in the future to augment our empirical electrostatic models with inducible dipoles and quadrupoles. HiPart is currently also used for the partitioning of density matrices by other members of the Center for Molecular Modeling.

The Center for Molecular Modeling has a long history on the molecular modeling of thermodynamic quantities and kinetic parameters for gas phase systems. We used and/or developed extensions to the conventional harmonic oscillator approximation such as the free rotor, the hindered rotor and the extended hindered rotor.[48-55] (See paper 8.) More recently also a series of coarse graining techniques were implemented to reduce complexity of the vibrational modes in large systems such as proteins.[56,57] (See paper 9 and 10.) In this context, TAMKin was recently developed as a framework to implement all these techniques in an consistent program. In its current form, TAMKin is a proof-of-concept that can easily combine the output of various quantum chemical calculations related to the reagents, transition states and/or reaction products. It immediately derives the thermodynamic and kinetic parameters of chemical reactions without any manual intervention. The program extensively uses the MolMod library, which is also used by ZEOBUILDER. In the future, concepts like free rotors, hindered rotors, and mobile blocks will be added to TAMKin.

MMFit and QFit are two programs of the same kind. They are both used in the development of empirical models in this thesis. MMFit estimates valence parameters for molecular mechanics models based on ab initio data. QFit is similar, but operates on empirical electrostatic models such as EEM[58] and SQE[59] instead of molecular mechanics models. QFit and MMFit are unique since they can systematically benchmark two important aspects of empirical

models: (i) the accuracy of different variations of a model and (ii) the robustness of different protocols to estimate the parameters.

Model Development

Zeolite synthesis is a very complex process: zeolite crystallization is currently performed by heating a solution or a hydrogel containing an organic structure-directing agent (template). Today, scientific insight into the molecular aspects of the formation of zeolite crystals is still rather limited. There is an ongoing discussion between research groups proposing crystal growth models based on monomer addition and others who explain the crystal formation processes based on aggregation mechanisms.[26,30] Molecular modeling could help in unraveling the molecular aspects of the zeolite synthesis processes. However, new theoretical models and algorithms are required to accomplish this goal. Their development is one of the main efforts in this thesis. We illustrate the necessity of model development with an important example, i.e. understanding the crucial role of the template molecule in different stages of the synthesis of zeolites.

The tetrapropylammonium (TPA) cation is a well-known template molecule that has different functions in several stages of the synthesis of MFI-structured materials such as ZSM-5 and Silicalite-I. TPA acts as a surfactant between the solvent (water, continuous) and the silica source (TEOS, discontinuous) when mixing synthesis ingredients.[26,61] During the initial synthesis stages, TPA shields the hydrophobic parts of silica precursors from the solvent and acts as a molecular scaffold to stabilize nanoscopic cage-like structures.[26,62,63] (steps 1-4 in figure 2) The exact role of TPA in this stage is the onset of a long debate in the literature. Martens et al. claim that nanometer sized zeolite particles are well shaped nanoslabs and already have the MFI porosity due to the presence of TPA.[23] Tsapatsis et al. consider these precursors as amorphous silica that happen to contain TPA cations.[30] TPA imposes the well-known MFI pore structure during the crystallization of larger particles.[26,62] In the case of the nanoslab-based synthesis scheme, the propyl chains of the TPA occluded in the nanoslab prevent direct contact between neighboring slabs.[23] (steps 5-10 in figure 2) This promotes an ordered liquid phase that facilitates long range order of nanoslabs before they aggregate. It is clear that a detailed knowledge of the behavior of TPA during the synthesis is a key to the understanding of zeolite synthesis mechanisms.

Simulations with molecular mechanics methods can already contribute substantially to our insight in the behavior of TPA. Although the formation of silica species is a process of condensation (and hydrolysis) reactions, the directing forces are mainly non-bonding electrostatic interactions that can be studied with force fields. In the process of continuous formation and breaking of oxygen bridges, certain silica intermediates and crystal structures are more

prevalent than others. If condensation is the rate limiting step, the outcome is controlled by the condensation kinetics. These can only be modeled with quantum mechanical techniques. When other processes are much slower (diffusion and reorganization of template molecules, diffusion of silica nutrients, collision and aggregation of colloidal particles, ...) the silica species are allowed to evolve towards their free energy minimum and the final structure is not determined by reaction kinetics. Although both regimes are possible, it is highly interesting to study complexation free energies of various silica particles and template molecules to identify preferred combinations. Molecular mechanics simulations are suitable for this type of work.

Another application of force fields is the computation of IR spectra to assist the identification of silica intermediates during in-situ measurements. Molecular mechanics methods are also attractive from the technical perspective because they allow both long simulations required to compute reliable thermodynamic averages and large molecular systems that take into account a complete explicit model of the environment. The accessible time and length scales are depicted in figure 3.

The inevitable drawback of force-field models is their empirical nature. The accuracy of the interactions predicted by the empirical model will determine whether simulations lead to reliable conclusions. Today only three sets of empirical parameters are published that attempt to describe all the interactions between the substances present during the hydrothermal synthesis of zeolites. One of them is the Universal Force Field (UFF),[60] which has one well-known strength: it can describe nearly all possible molecules with a generic set of empirical parameters that cover a large part of the periodic table of elements. This marvelous property is overshadowed by the poor accuracy of the UFF and the absence of a proper treatment of the electrostatic interactions. The two other models, CVFF[64] and CFF91_CZEO, have been tested by Lewis and coworkers to study the prenucleation stage of zeolite synthesis.[65] The latter model is a variant of CVFF tuned to reproduce gas phase quantum mechanical calculations of small silica fragments. These authors conclude that both empirical potentials are not sufficient for a quantitative description of silica-template complexes. The development of a new force-field model is unavoidable for future studies on the synthesis of zeolites.

The first efforts to parameterize force fields go back to the twenties of the previous century, when force field models were used to explain IR spectra of small molecules. Morse proposed the well-known potential for a diatomic molecule in 1929.[66] Urey and Bradley proposed a model for regular tetrahedral molecules in 1931 (XY_4 where X=C,Si,Sn and Y=Cl,Br).[67] These foundations are still present in current molecular mechanics models. The energy term proposed by Urey and Bradley is also used in this work to obtain a proper description of the vibrations in tetrahedral units present in zeolites. In the

seventies, considerably more complex force field models for large polyatomic systems were developed. Warshel et al proposed the first all-atom consistent force field model to describe proteins.[68] Allinger published the MM2 model for an accurate description of hydrocarbons.[69] During the following decades many improved variations of these models were published, although there has always been a strong focus on models with organic and biochemical applications. Models that are relevant today include MM3[70], MM4[71], MMFF94[72], GROMOS[73], CHARMM[74], AMBER[75] and OPLS[76]. Recent developments systematically rely more and more on post Hartree-Fock ab initio training data instead of experimental measurements to derive the empirical force field parameters. In this work we continue along this line and exclusively use ab initio training data. This allows a systematic generation of training data.

Although a large amount of the force-field development research is committed to biochemical applications, a considerable number of studies is devoted to the derivation of empirical silica potentials. One of the earliest models for silica is reported in 1976[77], which can not describe crystalline silica. Catlow[78] and Van Santen[79] developed the first useful force fields for silica that were able to describe quartz phase diagrams, quartz spectra and stable zeolite structures. Hill and Sauer have performed extensive parameterizations of a consistent force field model and an ion-shell model potential for silica based on ab initio training data.[80-83] From the mid-nineties until today, a vast amount of silica potentials is published, often with very specific purposes, such as the description of Na-Sodalite[84], or the parameterization of water in zeolites.[85] The reactive force fields that attempt to describe the dissociation of Si-O bonds properly are a remarkable achievement. They allow that the bond graph alters during a simulation, which makes the analytical form of the empirical model extremely complex.[86] For the adsorption of alkanes in zeolites, united atom models have been parameterized on the basis of experimental data.[87,88] The number of parameters in these models is limited to keep the parameterization feasible.

At present there are many diverse approaches in the literature to model silica and related materials with empirical potentials. There is no clear consensus on the optimal choice of functional form or parameterization strategy. The introduction of a next generation model that can describe silica with a broad set of organic guest molecules will only be possible with the aid of well-defined protocols. Otherwise, empirical model development is more an art than a science. In the remainder of this section, we outline our efforts to establish new methods and fixed protocols. This methodological aspect of our work has a much broader scope than the synthesis of zeolites. These techniques are transferable to a wide range of application domains including biochemistry, drug design, metal organic frameworks and other ceramic materials than zeolites.

To guarantee the quality of the empirical models developed in this work, we adhered to the principles outlined below. They are generic in the sense that they apply to the development of both molecular mechanics models and empirical electrostatics models.

- All training data are obtained with post Hartree-Fock ab initio quantum-mechanical calculations. The term "training data" refers to the data used to calibrate the empirical parameters in the model. Parameters are considered optimal when the model reproduces the training data as accurately as possible.
- Due to the first requirement, training data are only available for relatively small molecules. One must assure that models based on these small molecules are transferable to the large systems of interest. This means that the model must show the proper behavior over the entire range of length scales. One should also avoid parameters that can only describe correctly the molecules in the training set due to a cancellation of errors.
- The algorithm that determines the empirical parameters must be autonomous and objective. No manual tuning of parameters is allowed and the end result should not depend on auxiliary parameters in the parametrization algorithm. These requirements are a prerequisite to obtain reproducible results. Manual tuning also becomes unfeasible and suboptimal when the number of parameters increases.

The general form of the molecular mechanics energy (for polyatomic molecules) is as follows:

$$U_{\mathrm{FF}} = U_{\mathrm{Valence}} + U_{\mathrm{Non\ bonding}}.$$

The first term is the valence force field and describes the interactions due to the presence of chemical bonds. The latter term describes the non-bonding contributions such as electrostatic , Van Der Waals and hydrogen bonding interactions. This model is a mathematical representation of combined insights in the physical properties of molecular systems. It can not be strictly derived from quantum-mechanical considerations. Hence there are also different analytical forms for these three terms, and it is beyond the scope of this introduction to discuss all possibilities proposed in the literature. An overview of the most elementary ingredients is given in figure 8. A common form for the valence force field is as follows:

$$U_{\mathrm{Valence}} = U_{\mathrm{Bond}} + U_{\mathrm{Bending\ angle}} + U_{\mathrm{Urey\ Bradley}} + U_{\mathrm{Torsion}}$$

Each term corresponds to a specific type of geometrical internal coordinates. For example, a chemical bond is often modeled as a harmonic spring,

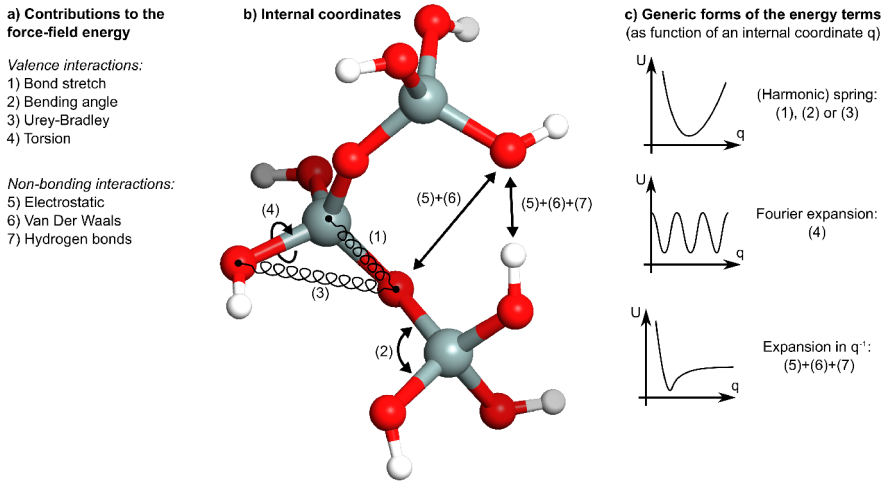


Figure 8: The elementary terms in a force-field model. Each type of interaction (a) has a corresponding energy term, which is an empirical function of a specific internal coordinate (b) and which has a typical functional form (c).

$$U_{\mathrm{Bond}} = \sum_{b=1}^{N_{\mathrm{bonds}}} K_b \left(r_b - r_{\mathrm{rest},b}\right)^2.$$

The sum runs over all chemical bonds. The reference energy corresponds to the equilibrium bond length, $r_{\mathrm{rest},b}$. The force constant, K_b, affects the molecular vibrational spectrum. These are two empirical parameters. The harmonic spring model does clearly not describe bond-dissociation, which is a general limitation of force-field models. Similar simple analytical models are used for the remaining energy terms in the valence force field. For a more elaborate example we refer to the work of Ewig et al.[90] The non-bonding part at least comprises three contributions,

$$U_{\text{Non-bonding}} = U_{\text{Electrostatic}} + U_{\text{Van Der Waals}} + U_{\text{Hydrogen bonds}},$$

which are also depicted in figure 8. It is a common practice to omit the last term and adapt the parameters in the electrostatic and Van Der Waals terms to describe hydrogen bonds implicitly. The most basic form of the electrostatic term is written as

$$U_{\mathrm{Electrostatic}} = \sum_{i=1}^{N_{\mathrm{atoms}}} \sum_{j=1}^{i-1} \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}}.$$

The sums run over all atom pairs. Atoms are treated as interacting point charges with a fixed net atomic charge, q_i. More elaborate models include interatomic charge transfer and atomic polarization (vide infra). In this case the reference energy corresponds to the situation where all atoms are at infinite separation, which is different compared to the valence terms. Therefore neither the total force field has a physical reference energy, nor can one compare absolute values of valence energies with absolute values for the electrostatic term. The term "force field" refers to the fact that it is only useful to consider derivatives of the total energy (i.e. forces).

Our global scheme for the development of a complete empirical force-field model is as follows.

1. A large training set of representative small molecules is constructed. Their gas-phase equilibrium geometries are determined with ab initio calculations.
2. The electrostatic part of the model, including polarization effects, is derived with QFit, based on partial charges from ab initio ground-state calculations and from similar calculations on states that are perturbed by an external electrostatic field.
3. Ab initio energies are computed for a large set of dimers of randomly selected molecules from the training set. The electrostatic interaction, as predicted by the model from the previous step, is subtracted from each dimer calculation. This yields training data that only include Van Der Waals and hydrogen bonding effects. The corresponding parameters are obtained with MMFit.
4. Ab initio calculations on perturbed geometries of single molecules are computed and all the non-bonding interactions, as predicted by the models from the two previous steps, are subtracted. The residual interactions are parameterized with MMFit to obtain the valence force field.

MMFit originally implemented the gradient curves method presented in paper 3, but both the parameterization technique and the implementation have sensitively evolved after this first publication. Only the basic ideas from the gradient curves method are still present. The algorithms implemented in QFit are explained in detail in paper 4. An application of an empirical electrostatic model (obtained with QFit) is discussed in paper 5.

Parameterization of Valence interactions

All force field parameterization procedures are essentially least squares methods. Experimental measurements or theoretical ab initio calculations for a set of representative (small) molecules are used as training data. The

parameters in the force field model are tuned as to minimize the overall error in the prediction of the training data by the force-field model. The overall error is measured by a least-squares cost function (also known as objective function), which is a weighted sum of squared errors between the training data and the predictions of the model:

$$Cost_{\mathrm{LS}}(p_1, \ldots, p_J) = \sum_{i=1}^{N_{\mathrm{observations}}} w_i \left(X_{i,\mathrm{train}} - X_{i,\mathrm{model}}(p_1, \ldots, p_J)\right)^2,$$

where p_j are the empirical parameters, w_i is the weight associated with data point $X_{i,\mathrm{train}}$ and $X_{i,\mathrm{model}}$ is the corresponding prediction by the model. All kinds of training data have been used in history[91], e.g. IR spectra, XRD crystal structures, ab initio atomic forces, ab initio second order derivatives of the energy, and so on. For the derivation of valence force fields, the ab initio forces are a common choice. Except for small molecules, this can only work when the non-bonding contributions are first subtracted from the theoretically computed atomic forces.

Although the method of force matching is appealing and simple at first sight, it easily leads to statistically ill defined parameters.[92] This means that the covariance matrix of the parameters is ill conditioned. We refer to chapter 15.4 of the book "Numerical recipes"[27] for a detailed discussion of the statistical error on parameters estimated with a least squares technique. There are in principle two reasons for this difficulty. Even a simple valence force field model for alkanes contains at least 22 parameters. This number quickly increases for the more accurate models, or when one includes organic functional groups. With such a number of parameters, an ill-conditioned covariance matrix is unavoidable. The second reason is inherent to the nature of valence force fields. The atomic forces in a force field model are the sum of many contributions, each associated with an internal coordinate such as a bond length, bending angle, dihedral angle, and so on. All these internal coordinates are not independent, but are a redundant description of the molecular geometry. Hence the Jacobian that transforms forces along the redundant internal coordinates (given by the force field model) into Cartesian atomic forces, has no unique inverse. During a parameterization based on force matching, it is exactly this ill-defined inverse transformation that is solved simultaneously for many molecular geometries. This eventually causes ill-defined parameters. The mathematical derivation of this principle is given in detail in paper 3.

It is important to realize in how ill-defined parameters can have a negative impact on the transferability of a force field. Ill-defined parameters lead in practice to a cancellation of the errors in the different empirical energy terms. Consider for example the methane molecule. A nonphysical high repulsion term between the geminal hydrogen atoms can be compensated by an

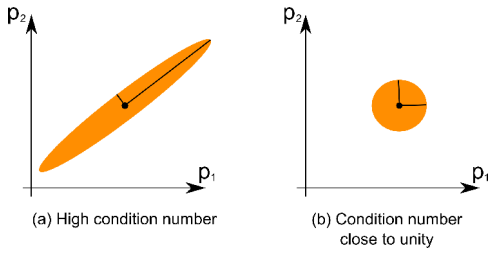


Figure 9: Analysis of the Hessian of a two-dimensional least squares problem. The black dot corresponds to the optimal parameters, while the orange ellipses are the regions where the least-squares cost is low. The lengths of the black lines correspond to the inverse of the square roots of the eigenvalues of the Hessian.

unphysically strong attraction term between the carbon and hydrogen atoms. The resulting net forces on the atoms can still be reasonable, but the model is clearly wrong. When these energy terms are used to describe a methyl moiety in linear alkanes, the compensation of errors will no longer work because the ratio of geminal hydrogen atoms over carbon-hydrogen bonds will change. The same type of error occurs when one mixes parameters from two different force-field models. In principle one can partially cure this problem by checking that each energy term has a sound physical interpretation, but subtle errors of this type are not easily recognized.

In conventional linear least squares applications, one can detect correlated parameters as low eigenvalues of the covariance matrix of the estimated parameters, but this concept is not simply applicable when fitting force-field parameters. The definition of the covariance matrix assumes that the measurement errors on the training data are uncorrelated and normally distributed.[27] When fitting force field parameters to reproduce ab initio training data, these criteria are not fulfilled: ab initio data can be computed up to an arbitrary numerical precision and the errors with respect to experimental data are systematic. For example Hartree-Fock calculations overestimate vibrational frequencies up to 10%.[89] Also the residual error between the ab initio training data and the force field predictions is purely systematic since there is no statistical noise on the ab initio data. Therefore the conventional definition of the covariance matrix is not applicable. As an alternative, we can analyze the Hessian of the cost function, i.e. that is the matrix with second order derivatives of the cost with respect to the parameters. This matrix corresponds to the covariance matrix in the conventional least squares approach when the weight w_i is equal to inverse of the squared measurement error on $X_{i,\mathrm{train}}$. In our case, absolute eigenvalues of the Hessian have no interpretation, but the condition number of this matrix, i.e. the ratio of the highest over the lowest eigenvalue, is a good measure for correlated parameters. This is illustrated in figure 9 for a model with two

parameters. In figure 9a the condition number is high and the sum of p_1 and p_2 is much more sensitive to irrelevant details in the training data than the difference. In other words, only the difference between the two parameters is reliable. When this is the case, one observes exaggerated values of the parameters. When estimating force-field parameters, the condition number quickly amounts to several orders of magnitude. In figure 9b the condition number of the Hessian is close to unity and the quality of both parameters is comparable.

A pragmatic approach to reduce this type of error is given in the work of Ewig et al.[90] The least squares cost function is extended with quadratic terms that penalize high values of the parameters:

$$Cost(p_1, \ldots, p_J) = Cost_{\text{LS}}(p_1, \ldots, p_J) + \sum_{j=1}^{J} v_j p_j^2.$$

In this expression, the sum runs over all the parameters, p_j, and the strength of each penalty term is given by the constants v_j. Unfortunately, this technique also introduces a new difficulty: what are reasonable values for the auxiliary parameters v_j? If they are too high, the parameters are restrained to zero and the force field is useless. When they are too low, the parameters are ill-defined.

The gradient curves method[93] (GCM, paper 3) is our first effort to rationalize the choice of the penalty function. Additionally, we consider the possibility to make no prior assumptions about the analytical form used in the force field model. The method can be summarized as follows:

1. A training set of data is available for the molecules of interest, containing high-level ab initio forces on the atoms in Cartesian coordinates for different distorted molecular geometries. It is assumed that the contributions from the non-bonding interactions are removed from these data. The second type of input are the internal coordinates with which we want to associate energy terms, e.g. bond lengths, bending angles, and so on.

2. For each molecule, the forces acting on the atoms are transformed into forces along the internal coordinates. Without further specifications this transformation would be very ill-defined. We also demand that the transformed data belonging to a certain type of internal coordinate (e.g. C-H bond lengths) should have a minimal deviation from a polynomial of a high degree, e.g. up to power 10. At this point, the transformation is still ill-defined and a second criterion must be introduced: the penalty for unreasonably high forces along the internal coordinates. This penalty contribution is simply proportional to the norm of the vector with all the transformed

forces. Even when the strength of the penalty is negligible compared to the goodness of fit of the transformed data to the auxiliary polynomials, the outcome is insensitive to small perturbations in the training data.

3. The result from step 2 is a set of data points for each energy term, that can be used to fit the derivative of the corresponding energy term. In practice these data points almost coincide with a nice curve, except at the boundaries of the interval for which we have data. Therefore each data set is called a gradient curve (see paper 3), which lies at the origin of the name of the method.

An illustrative example is given in figure 10. The training data consist of 200 random geometries of a single water molecule. For this small molecule, the effect of non-bonding interactions can be included in the valence force field without further difficulties. We considered six force-field models. The most basic model has three types of energy terms: a bond stretch term (function of the OH bond length), a bending angle term (function of the cosine of the HOH angle) and a Urey-Bradley term (function of the HH distance). The subsequent models contain an increasing amount of additional energy terms that are functions of products of the three basic internal coordinates. Each energy term is modeled with a fifth order polynomial. For each parameterization, the molecules were divided in two random subsets of 100 molecules. In principle both sets contain the same information and should therefore lead to the same parameters. We also expect that an additional energy term should not cause major changes in the terms that were already present. Finally, we do not expect ridiculously large forces along the internal coordinates. Model V1 in both part a and b of figure 10 is not suffering from ill-defined parameters. The same result is obtained for both training sets, with or without penalty terms in the cost function. This is not the case for model V2 to V6 obtained with the least squares method. Additional energy terms lead to large and non-systematic fluctuations in the original energy terms. The parameters for the models V2 to V6 obtained with the gradient curves method show completely the opposite behavior: the results are the same for the two training sets, there are no extremely large energy terms and the results only change slightly when a new energy term is added. This example clearly demonstrates that a completely automatic algorithm to derive force field parameters is feasible.

The gradient curves method as presented in paper 3 is only a starting point. Additional improvements and simplifications are still mandatory before it can be used for the parameterization of more generic force fields with a large variety of atom and bond types. An extension of the original GCM is currently under development as part of the MMFit program. This new method,

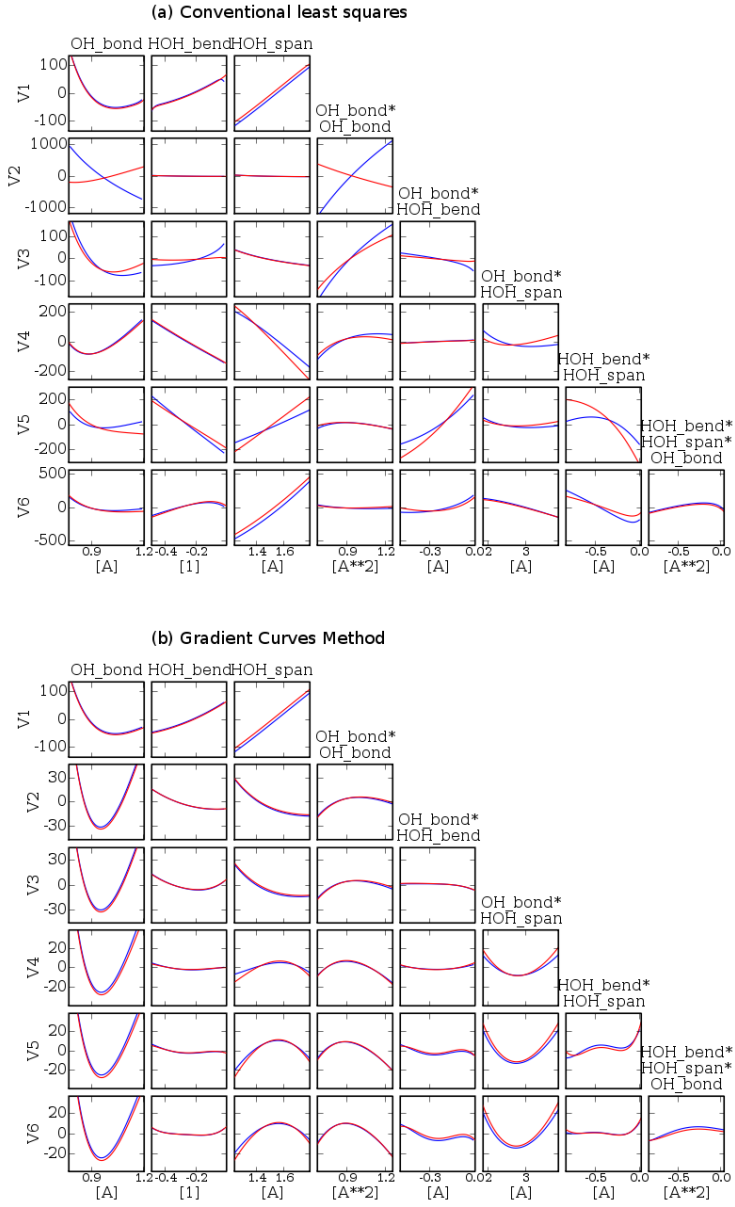


Figure 10: Comparison of a conventional least squares method with the gradient curves method. The red and the blue curves represent results for two different data sets based on 100 randomly distorted geometries of the water molecule. Each row corresponds to a force field model, while each column corresponds to an energy term. The first three columns correspond to the bond stretch term, the bending angle term (as function of the cosine of the angle), and the Urey Bradley term. The following columns are energy terms that depend on products of these three internal coordinates. Each energy term is a polynomial up to power 5.

GCM2, is still preliminary and is therefore not discussed in detail in this thesis. The following improvements are already implemented in GCM2.

1. A different penalty term is associated with each energy term and the optimal value of the corresponding strength is determined as to maximize the robustness of the optimal parameters. This means that there is not a single manually tuned auxiliary parameters in the parameterization procedure. For a given set of training data and an analytical description of the model, the parameters are computed without any manual intervention.
2. Van Der Waals terms can be obtained from post Hartree-Fock ab initio calculations on dimers. It is mandatory to correct for basis set superposition errors.
3. The transformation of the Cartesian energy gradients is omitted and one must specify a complete analytical model a priori. This is computationally more efficient and makes it easier to obtain models that can be implemented in existing molecular dynamics software. All parameters must be linear coefficients.
4. The training data can also contain energy differences and second order derivatives of the ab initio molecular energy.

Preliminary benchmarks on a series of linear alkanes show that nearly all final parameters do not depend on the length of the alkane used for the parameterization. This is a sufficient condition for transferability. Some noticeable variations are still present in the parameters of the torsion energy terms for different linear alkanes. Further research is required to correct for this discrepancy.

There is a second important complication that requires our attention: the preparation of training data from which the non-bonding contributions have been subtracted. This is only possible if we have a good model for the non-bonding interactions. Acceptable Van Der Waals parameters can be derived with GCM2, but the electrostatic interactions don't seem to be a simple function of interatomic distances. Benchmarks on a series of linear ethers and tetrahydrofuran showed that there is no possibility to define transferable rules for fixed atomic charges to model the electrostatic interactions. We could find ad hoc atomic partial charges for each individual case, but this is of little use. For this reason, we started working on more accurate empirical models for the electrostatic interactions, which will be discussed in the next subsection. They are a prerequisite for future work on GCM2.

Empirical electrostatic models

The most basic empirical electrostatic model treats atoms as point charges with a fixed net charge, which cannot account for all types of electrostatic interactions. Electronic polarization effects and higher atomic multipoles are completely neglected. This approximation is still mainstream in current force-field applications,[94] but recently biochemical force fields such as CHARMM[74], AMBER[75] and OPLS[76] are extended with energy terms that account for explicit polarization and atomic dipoles.[95-98]

A few inherently deficient workarounds are commonly used for the absence of electronic polarization in empirical models. One can for example postulate a homogeneous background dielectric constant. This type of approximation is only valid for atom pairs with a large separation in a chemically homogeneous system. At short interatomic distances the macroscopic picture of a uniform dielectric constant is no longer valid. This simple technique also fails for inhomogeneous systems that contain interfaces between dissimilar environments, such as cell membranes.[96] The intermolecular attraction due to mutually induced dipoles can be mimicked with disproportional atomic charges and dispersion parameters. This approach neglects the many-body nature of mutual polarization. The interaction between dimers in the gas-phase has a lower contribution from such many-body interactions compared to a dimer in the liquid phase. The empirical parameters derived for non-polarizable models are therefore not transferable from ab initio gas-phase dimer calculations to condensed phase systems.[97,100,101]

The major applications of force-field models are host-guest, substrate-ligand and solvent-solute simulations, which critically depend on the accuracy of the non-bonding interactions. In the case of zeolite synthesis, the oxygen atom is known to be highly polarizable[102] and this will certainly have a relevant contribution to the interaction between small silica oligomers and organic cations. Even more important is the charge distribution in partially deprotonated silica oligomers. A proper force field on the basis of ab initio training data, which can be used to model the role of organic cations in the synthesis of zeolites, must therefore include an explicit description of polarization effects.

There are two types of electronic polarization: one due to charge rearrangement within an atom and the other due to interatomic charge transfer. The first empirical model for the former effect is about fifty years old. The latter is only described properly with empirical models in the last few years. Dick and Overhauser introduced the charge-on-a-spring model in 1958.[103] The model describes an atom as a positive point charge (nucleus and core electrons) and a massless negative point charge (valence electrons) connected by a harmonic spring. Repulsive Pauli interactions are only present between negative shells to account for the influence of steric hindrance on the

polarizability. Applequist et al. followed a different approach and modeled atoms in a molecule as a set of interacting point dipoles.[104] The first empirical model to describe charge transfer is presented by Mortier et al.[58] He developed the electronegativity equalization method (EEM), which is a second order expansion of the molecular energy in terms of atomic partial charges based on density functional theory. This energy is minimized under a total charge constraint to obtain the partial charges. The original purpose of EEM was to predict partial charges of a molecule in the absence of an external electrostatic field, but the formalism is easily extended with an extra term that takes into account such a perturbation. In this way, one can use the EEM for the prediction of the polarizability due to charge transfer. These three essential papers are the foundation for the work on empirical electrostatic models in the past decades.[97,105,106] A large amount of subsequent papers in the literature deals with the calibration of suitable parameters that can predict charge distributions and linear response properties for a set of molecules.[95-99,105-119]

Only recently it is recognized that the EEM has a fundamental issue with the dependence of the polarizability on the size of a molecule.[120] EEM predicts that a the polarizability of a linear organic chain molecule is a cubic function of the chain length[121], while organic systems and ceramic materials should show a linear dependence on the system size in the macroscopic limit. This means that the EEM parameters that optimally describe linear response properties for small molecules are not transferable to large condensed phase systems. In practical applications one can introduce ad hoc charge constraints on molecules or molecular fragments to reduce erroneous effects.[121,122] This discrepancy can be cured in a more physical way when the EEM is extended with an additional quadratic energy term associated with the charge transfered through a chemical bond. Nistor et al. implemented this concept in the Split Charge Equilibration (SQE) model.[59] A more intuitive approach such as the damping of the long-range coulomb interaction in the EEM is known to fail.[123]

Our contribution in this field is the parameterization of the SQE and the EEM model for a broad range of organic molecules. The most important conclusion is that the parameterization of the SQE model is feasible for a very diverse group of 500 randomly selected organic molecules, when using the proper parameterization techniques. The correct trends for the polarizability for chain molecules are reproduced with the SQE model, while the EEM fails to do so. This is not only the case for ordinary linear alkanes, but also for more complex structures like an alanine alpha helix. The polarizability per monomer as function of the number of monomers in the alpha helix is plotted in figure 11. Also the prediction of atomic partial charges improves drastically compared to the original EEM: the RMS error is reduced by a factor two. An extensive description of the parameterization procedure and the performance of the original EEM versus the SQE model are discussed in paper 4.

In Paper 5 we investigate the local polarizability and the electrostatic potential of the active sites in five proteins to explain ligand docking. The energy due to a perturbation with a Gaussian charge distribution is decomposed into a component due to the equilibrium charge distribution (electrostatic) and due to the induced charge distribution (polarization). The gradient in the electrostatic potential predicted by the SQE model correlates well with the position of the active site. The fluctuations in the electrostatic potential are much less pronounced and uncorrelated when the protein is modeled with the EEM model. Similar conclusions can be drawn from the local polarization energy. In the case of the EEM model it is uniform over the entire protein, while the SQE model shows a nice correlation with the position of the active site. The relative trends in the local polarizability of different active sites can be correlated with the surrounding amino acids.

The positive assessment of our variant of the SQE model after an extensive benchmark study clearly demonstrates that this model is a sound starting point and even a cornerstone for further developments of polarizable force field models. The SQE model can still only describe polarization along chemical bonds, which is the major remaining limitation. This shortcoming can be removed by extending the SQE model with atomic inducible dipoles.[97,105,106]

The next step is the derivation of SQE parameters for silica and other substances present during the hydrothermal synthesis of zeolites. In principle we could limit the training set to small silica oligomers, to water and ethanol as solvent and to tetrapropylammonium (TPA) as template molecule. However, to guarantee transferability and reliability of the parameterization an extension of the training set is recommendable. The work of Catlow and Lewis[102] indicates that the bridging oxygen in silica is highly polarizable and therefore we must consider an extension of the SQE model with inducible

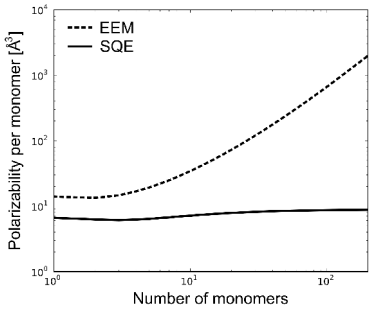


Figure 11: The polarizability per monomer of an alanine alpha helix as function of the number of monomers. The value predicted by the EEM model diverges in the limit of long chains. The SQE model does not show this discrepancy.

dipoles. A proper description of hydrogen bonds between silanol groups, alcohol groups and water even implies an extension up to atomic quadrupoles for the relevant oxygen atoms. An additional complication is the inter-molecular charge transfer through hydrogen bonds, which has a relevant contribution to the electronic polarizability of water.[124]

After the construction of suitable models for the electrostatic interactions in silica is completed, one can proceed with the development of a global parameterization scheme to derive the residual Van Der Waals interactions and consequently also the valence terms.

The final purpose of this work is the development of a complete force field model that is applicable to all forms of silica (both amorphous silica, small oligomers and crystalline zeolites) in contact with polar guest molecules, templates and solvent molecules. This will be an important milestone with a large variety of potential modeling applications. In order to enhance the visibility and the impact of these new developments in the large community of force fields, we will not only publish our results in high-impact scientific journals, but we also plan to implement the entire code in the open source software package CP2K[125]. The program is developed for high performance computing and scales effectively to a large number of processors. It is therefore an excellent tool for simulations on micro- and mesoporous materials.

Modeling the Synthesis of Zeolites

The MFI-Fingerprint

The discovery of the clear solution synthesis of Silicalite-I[126], an all-silica crystal with the MFI-topology, enables in-situ measurements on the initial stages of the synthesis of zeolites. The low concentration of the synthesis ingredients results in an optically transparent mixture. This allows a characterization of the zeolite precursors while they grow, using techniques such as dynamic light scattering (DLS) and Fourier transform infrared (FTIR) spectroscopy. The clear solution method is highly reproducible and one systematically observes the presence of colloidal nanoscopic silica particles.[126,127] There are very different hypotheses in the literature about the exact shape and structure of these silica nanoparticles, but it is generally accepted that these species are an intermediate in the formation of zeolites.[18-30]

Several groups have performed IR spectroscopy measurements on clear solution mixtures.[30,63,128-132] Kirschhock et al recorded detailed spectra, both in-situ as well as on extracted nanoparticles.[19] An essential observation is the evolution of an adsorption band from 650 cm^{-1} in the initial synthesis mixture to 550 cm^{-1} in the final Silicalite-I crystal. The latter peak is associated with the

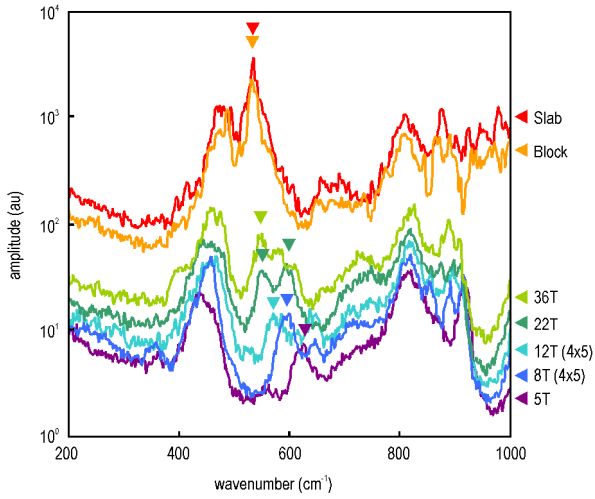


Figure 12: Theoretical IR spectra of an entire range of silica oligomers and nanocrystaline intermediates for the synthesis of MFI-structured zeolites. The spectra are sorted by the size of the corresponding particle, starting with the smallest species at the bottom.

presence of five-membered rings in MFI-structured zeolites.[63,128,129] This peak is absent in amorphous silica or in the individual synthesis ingredients.[19] In paper 6, we investigate theoretical IR spectra of several zeolite precursors suggested in the literature.[133] Several hypothetical structures were also examined to test the systematic dependence of the IR spectrum on structural features.

The theoretical IR spectra are derived from molecular dynamics simulations[134], using a force-field parameterization, GCM-SiOH-0.2, obtained with the gradient curves method.[93] This model should be considered as a prototype, rather than a completely fine-tuned force-field model. The GCM is mainly suitable for the derivation of valence interactions and therefore the quality of the non-bonding interactions is sub-optimal. GCM-SiOH-0.2 is similar to a consistent valence force field, including all possible bond, valence angle and Urey-Bradley energy terms. This facilitates the reproduction of the vibrational frequencies of the molecules in the training set.[67,93]

The main purpose of the theoretical spectra is an underpinning of the correlation between the size or structure of pentacyclic silica species and the position of the absorption peaks in the IR spectrum. The theoretical spectra are shown in figure 12. A single 5T-ring and loosely connected 5T-rings, the smallest relevant species, have an adsorption peak at 650 cm^{-1}. Only more condensed combinations of 5T-rings, with a higher degree of interconnectivity, yield an additional peak around 620 cm^{-1}. When considering

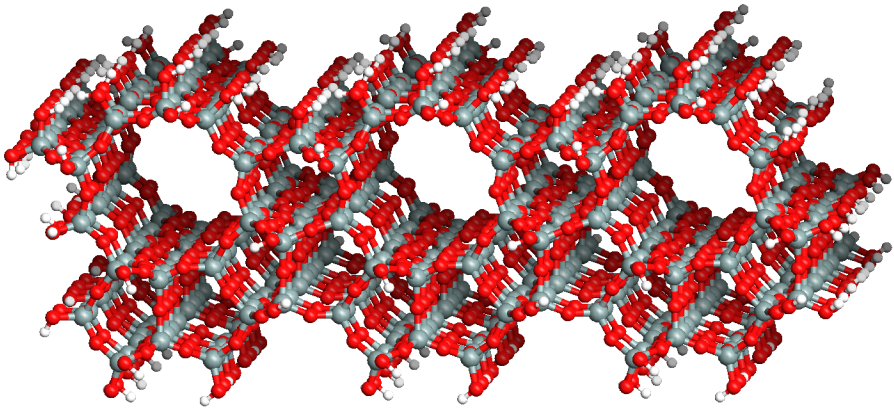


Figure 13: Model representation of the MFI nanoslab. Both the geometry and the topology of this model match exactly the MFI crystal structure.

larger fragments of the MFI structure, the latter peak moves gradually towards lower wavenumbers. The lowest value, 550 cm^{-1}, is observed for the MFI-nanoslab depicted in figure 13.

Future molecular dynamics simulations for the derivation of IR spectra of silica species should take into account two shortcomings in the current work. For the computation of IR spectra, ad-hoc values for the atomic charges were selected to compute the time dependent dipole moment. In a more complete model, the computation of electrostatic properties (such as the dipole moment) will form an integral part of the force field. This is one of our principal motivations to develop and benchmark advanced empirical electrostatic models. (vide supra) A second issue is the absence of a solvent medium during the simulations. The motion of the solvent molecules will couple with vibrational modes in silica species, which will mainly affect the spectra of the smallest oligomers such as the 5T-ring. The coupling of these modes is mainly determined by electrostatic interactions and Pauli repulsion, which is again an incentive for the development of more reliable empirical non-bonding energy terms. It is possibly also interesting to compute spectra on more realistic models of nanocrystals that have fractions of Q^n silicon atoms that correspond to NMR experiments.[19,30]

Multi-level modeling

Despite our efforts to develop more reliable empirical models with a stronger emphasis on theoretically supported models, it remains appropriate to study several aspects of the synthesis of zeolites with well-established techniques. Thermodynamic and kinetic properties of the initial condensation reactions that lead to small silica oligomers[135-137] can only be studied with quantum

mechanical calculations. Such studies can also reveal the limitations of the currently available force fields[65] and quantum mechanical methods.

The elementary behavior of silica-TPA pairs are studied in paper 7, where the silica species are the 22T and 33T MFI precursors as suggested by Kirschhock et al.[20] The results from static quantum mechanical computations are compared with molecular dynamics simulations based on empirical force fields. The main conclusion is that the template molecule can not reside in the 10T-ring channel of the MFI precursor as soon as this ring is completely formed. The template quickly moves to a "half in - half out" position visualized in figure 14, with two propyl chains still interacting with the preliminary channel in the MFI-precursor. A geometric concatenation of multiple precursors into a nanoslab with ZEOBUILDER[38] reveals that the optimal position of the TPA template corresponds to the intersection of the straight and sinusoidal channels of the MFI-topology. The template keeps its optimal position without imposing steric hindrance when the nanoslab is constructed, which is a prerequisite for a straightforward aggregation of MFI precursors.

The quantum mechanical simulations also point out that a 33T precursor without a template collapses due to the formation of internal hydrogen bonds. This supports the idea that a template is essential to impose the channel structure during the formation of the MFI-precursor.[65] A similar collapse is not present in the molecular dynamics simulation, but we should not expect an accurate description either. The UFF[60] model used in the molecular dynamics simulations poorly describes both the force constants that determine the stiffness of the 10T-ring and the hydrogen bonds that are the driving force for the collapse.

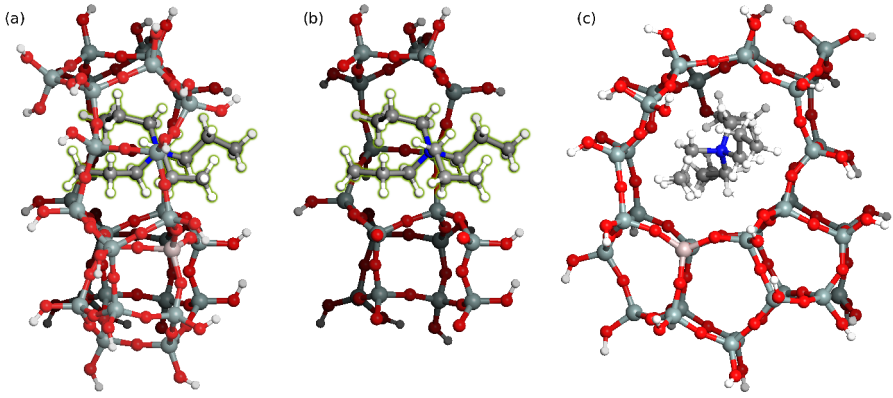


Figure 14: One of the possible complexes of the 33T MFI-precursor[20] with the tetrapropylammonium cation (TPA). Part (a) shows the side view of the complex and part (b) is a similar projection where half the precursor is not visualized to clarify the structure of TPA. Part b is a projection along the preliminary 10T channel.

The conclusions in paper 7 draw a rough picture of the interplay between TPA and the MFI precursors. The large energy difference between the situation with the TPA completely embedded inside the precursor and the TPA in a position "half in - half out" leaves little room for speculation. When one would like to investigate more subtle questions, statistical inference becomes infeasible with quantum mechanical calculations and unreliable with the current force field models. It is for example interesting to know the mean residence time of a TPA cation in contact with a precursor in a water solvent. If this time is long compared to the collision time of two precursors, one has an extra argument that TPA will also play a role in the supramolecular organization of multiple precursors. This question can be answered with the Fast-Marching method[138-140] or transition path sampling techniques.[141,142] The Fast-Marching method is an efficient technique to sample the (free) energy landscape, which can be used to compute kinetic properties. Transition path sampling is computationally more demanding, but it surmounts the limitations of transition-state theory and it is less sensitive to the choice of the reaction coordinate. Both techniques require a large amount of molecular dynamics simulations, which becomes unfeasible with accurate quantum mechanical methods. The same type of molecular dynamics simulations can be performed with an empirical force field model, but then at least the quality of the electrostatic interactions between the precursor, TPA and the solvent must improve.[65] A similar dilemma shows up in other potentially interesting investigations. It would for example be more realistic to perform simulations with an entire shell of template molecules surrounding the precursor particle. The self-diffusion of these TPA molecules is relatively slow and the simulation time required to take proper statistical averages easily reaches beyond the nanosecond. Quantum mechanical simulations become unfeasible, while empirical potentials are not reliable enough to describe the interactions. This type of frustration is the principal motivation for our continuous effort to develop an empirical force field model that will be a good compromise for future simulations.

Part 1: Software Development

Paper 1: "ZEOBUILDER: A GUI Toolkit for the Construction of Complex Molecules on the Nanoscale with Building Blocks"

Toon Verstraelen, Veronique Van Speybroeck, Michel Waroquier

Journal of Chemical Information and Modeling , **2008** , 48, 1530 - 1541

1530 *J. Chem. Inf. Model.* **2008,** *48,* 1530–1541

ZEOBUILDER: A GUI Toolkit for the Construction of Complex Molecular Structures on the Nanoscale with Building Blocks

T. Verstraelen, V. Van Speybroeck, and M. Waroquier*

Center for Molecular Modeling, Gent University, Proeftuinstraat 86, B-9000 Gent, Belgium

Received February 29, 2008

In this paper, a new graphical toolkit, ZEOBUILDER, is presented for the construction of the most complex zeolite structures based on building blocks. Molecular simulations starting from these model structures give novel insights in the synthesis mechanisms of micro- and mesoporous materials. ZEOBUILDER is presented as an open-source code with easy plug-in facilities. This architecture offers an ideal platform for further development of new features. Another specific aspect in the architecture of ZEOBUILDER is the data structure with multiple reference frames in which molecules and molecular building blocks are placed and which are hierarchically ordered. The main properties of ZEOBUILDER are the feasibility for constructing complex structures, extensibility, and transferability. The application field of ZEOBUILDER is not limited to zeolite science but easily extended to the construction of other complex (bio)molecular systems. ZEOBUILDER is a unique user-friendly GUI toolkit with advanced plug-ins allowing the construction of the most complex molecular structures, which can be used as input for all ab initio and molecular mechanics program packages.

1. INTRODUCTION

Specialized graphical computer programs are indispensable for the construction of atomic models in computational chemistry research. The advances in the development of nanoscale materials and the modeling of increasingly complex biological systems imply new demands on the software toolkits that are used to construct and manipulate the molecular models in these challenging research fields. Computational chemistry is playing an increasingly important role in all aspects of zeolite science.

There is large demand for graphical tools that build large frameworks with different compositions and topologies. All existing commercial (Cerius2,[1] Materials Studio,[2] CrystalMaker,[3] Diamond,[4] etc.) software packages are of high quality and are indispensable in most molecular modeling applications using well-validated techniques. A survey of all available open-source tools (Gamgi,[5] PyMol,[6] JMol,[7] Avogadro,[8] AGM,[9] ect.) reveals that these GUIs have rather limited features and are not widely used due to a lack of applicability, feasibility, or transferability. We could consider the extension of existing graphical toolkits, but even the most adequate package poses serious problems when implementing new features such as the growth of a zeolite framework starting from elementary building blocks. An essential ingredient for a molecular building toolkit is a suitable data structure with multiple reference frames in which molecules and molecular building blocks are placed and are hierarchically ordered. A powerful plug-in framework for the extension of a GUI toolkit with new features is equally important and unfortunately not present in many existing codes.

Inorganic crystals such as zeolites are conventionally classified on the basis of subunits in their framework structure.[10] The basic building unit for a zeolite structure is the tetrahedron formed by a central T atom and four oxygen atoms at the corners, where the T atom is typically silicon or aluminum. All zeolite frameworks can be constructed from a small set of larger secondary building units (SBUs), although the list of SBUs has been extended over the years when new zeolite frameworks were developed.[10] It is common practice to describe the structure of a zeolite in terms of even larger composite building units (CBUs) such as the sodalite cage, the double six ring, and so on.[11] At an even larger scale, one recognizes periodic building units (PerBU's), for example, chains or channels formed by an infinite series of SBUs and CBUs. An overview of the standard secondary, composite, and periodic building units for zeolites is given in the Atlas of Zeolite Framework types.[12] The concept of building units is a recurring theme in many fields: for example, proteins are built from peptides; metal−organic frameworks consist of metal centers and ligands in between them, and so on. From this point of view, a graphical molecular editor must fully embrace the concept of molecular building blocks.

Next to the graphical toolkits, a distinct category of nongraphical software tools is used for the manipulation and study of inorganic structures. Zebedde[13,14] is a powerful computational tool focusing on the design of suitable template molecules for a given pore structure but does not address the construction of the framework itself. There are also many popular structure refinement and predicition tools, for example, DLS-76,[15] GRINSP,[16] SHELX,[17] and so on. We are convinced that a graphical toolkit should not duplicate all of the features in the existing nongraphical software. Instead, a GUI toolkit must be able to interoperate with these programs through open standards for file formats and communication protocols.[18]

Taking into account all of these considerations, the authors preferred to build a new GUI toolkit—called "ZEOBUILDER"—from scratch, which involves all of the ingredients

* To whom all correspondence should be addressed. Tel.: 32 (0)9 264 65 59. E-mail: michel.waroquier@UGent.be.

10.1021/ci8000748 CCC: $40.75
Published on Web 06/11/2008

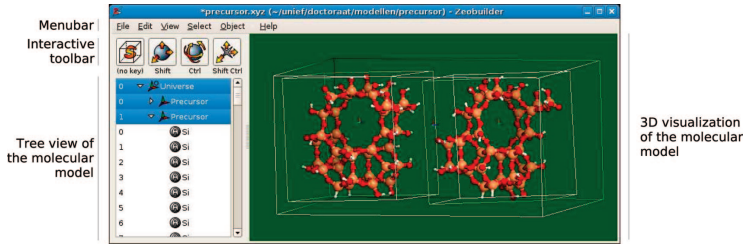


Figure 1. The graphical user interface of ZEOBUILDER consists of four major parts: a menu bar, an interactive toolbar, a tree view of the model, and a 3D visualization of the molecular system.

required for the construction of complex hierarchical zeolite models and which offers an ideal platform for the further development of new features in an open-source facility.

The two key-advantages of ZEOBUILDER can be summarized as follows:

(i) ZEOBUILDER has a user-friendly graphical interface that presents all its features in a toolkit fashion. With an appropriate choice of well-selected features and creativity, the user should be able to manage the most complex tasks.

(ii) Additionally, the program is also developer-friendly. The toolkit approach guarantees that a minimum of coding effort results in maximal functionality. Further, it will be distributed as an open-source cross-platform tool, and new features are easily implemented as plug-ins.

The plug-in architecture is the main technical advantage of ZEOBUILDER, when compared to the open-source alternatives. To encourage new developers, a ZEOBUILDER plug-in is a simple self-contained text file: it is cross-platform by definition, and no compilation is required. The program is written in the popular and highly portable Python programming language. ZEOBUILDER as presented in this paper is not restricted to building large zeolite structures but is easily extended to the construction of other complex (bio)molecular systems. ZEOBUILDER has all of the features needed to act as a generic molecular editor that is also capable of manipulating organic species, for example, ligand-substrate coordination, internal rotations, pseudorotations, and so forth. Transferability is a key property of the new toolkit.

The structure of the paper is as follows: Section 2 outlines the main features implemented in ZEOBUILDER that make it a unique tool from the viewpoint of the user. It offers the reader a first impression of the wealth of possibilities embedded in the code when building and manipulating the most complex molecular systems. Section 3 is devoted to the methodology of some key operations like condensation reactions of zeolite blocks. The next section shows how user-friendly ZEOBUILDER is. With the help of three examples the reader gets started with a list of step-by-step instructions. Finally, in section 5, some practical instructions are given on how to obtain the latest version of ZEOBUILDER and how to get support.

2. MAIN FEATURES

A screenshot of the user interface of ZEOBUILDER is displayed in Figure 1. The large display window in the bottom right gives a 3D visualization of the molecular system. At the left, a tree structure is displayed with a list of all objects constituting the molecular model. The menu bar on top of the display gives access to all noninteractive tools to modify the displayed molecular model. The interactive tools that operate directly on the 3D view are configured with the help of the four toolbar buttons in the top left of the window. Each toolbar button associates an interactive tool with a combination of modifier keys (shift and ctrl) on the keyboard. For example, when the user drags with the mouse button while holding the shift key pressed, (a part of) the model is translated. The default settings are as follows: when no keys are pressed, the selection tool is active. When the shift key is pressed, one performs translations in the 3D view. Similarly, the ctrl key is associated with interactive rotations. When both the shift and the ctrl key are pressed, one translates the global rotation center, or one changes the scale of the 3D display by zooming in or out. These interactive functions are not fixed. Other interactive tools, such as a measurement tool and a sketch tool, are configured by clicking on one of the four toolbar buttons.

ZEOBUILDER can be started up without any preprogrammed input file. Molecules can be built on a free basis. If desirable, any XYZ file can be used as input facilitating the construction of complex molecular structures. However, the default file format is ZML, which is an internal format based on the XML standard. It is flexible toward future extensions, and it is capable of storing all aspects of a ZEOBUILDER model. Other formats such as PDB and XYZ are also supported, but these formats cannot include all aspects of a molecular structure constructed with ZEOBUILDER. In addition to loading and saving models with different file formats, one can also import a file into the current model or one can export a part of the current model into a file. For the user's convenience, a library is added containing all preprogrammed ZML files of unit cells for all commonly used zeolite frameworks (IZA database[12]). At any time, molecules or parts of molecular systems can be saved in the desired format (XYZ, PDB, or ZML).

The molecular structure is visualized by balls with their size and color adapted to commonly used conventions. An algorithm is built in to optionally display bonds between atoms. The procedure on how to carry out this operation is outlined in section 4. Sometimes, it can be useful to use a less-cluttered representation omitting the oxygens, and straight lines are drawn connecting the tetrahedral atoms. An illustration is given in Figure 2, where a sodalite cage unit is displayed in an FAU framework.

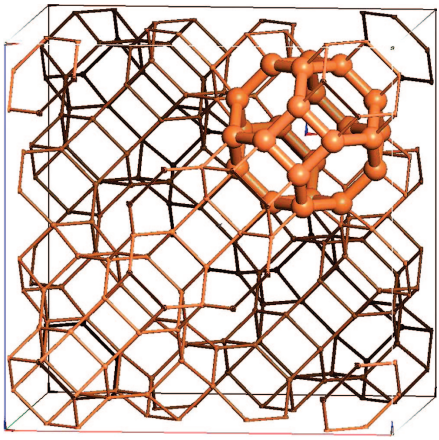


Figure 2. The sodalite cage unit in the FAU framework. Oxygen atoms are omitted, and straight lines are drawn connecting the tetrahedral (T) atoms.

One of the essential concepts of ZEOBUILDER is the novel data structure that represents the molecular model. The user can define a hierarchical structure of reference frames in which molecules or molecular building blocks are placed. This feature is especially useful for constructing models based on building blocks. On the top of the hierarchy stands a space-fixed reference frame. Molecules and other body-fixed reference frames can be inserted freely. They have subhierarchical character. The number of hierarchical levels of the reference frames is unlimited. They all can again serve as containers of molecular structures. When one body-fixed reference frame is submitted to a geometrical transformation (rotation, translation, and inversion), all of the objects contained in the frame will undergo the same transformation.

The hierarchical structure of the reference frames is also practical for other complex models like ligand−receptor systems and coordination complexes. This tranferability is one of the main advantages associated with ZEOBUILDER.

For the convenience of the user, a series of features that are common to usual editor tools has been implemented (be it a word processor or drawing software). Each operation in ZEOBUILDER can be undone, redone, and repeated without limitations. One can cut, copy, and paste parts of molecules or any other group of objects. Additionally, all of the attributes of each object or group of objects can be inspected and modified.

ZEOBUILDER also incorporates some special basic features allowing complex transformations not present in current graphical toolkits. Unique is the ability to insert arbitrary points, vectors, planes, and so forth at any place in space and to associate a specific function to it. A point in space can be regarded as a rotation center or an inverse center, a vector as a rotation axis, a plane for a reflection operation, and so forth. One selects the object (it can be part of the molecular system), and all geometrical operations can be executed. To illustrate with an example, we demonstrate an internal rotation of part of a molecule about an axis. First, that part of the molecule is selected and stored in a separate frame. Next, the rotation axis is determined by a vector defined by two selected atoms. By means of the right interactive tools, the rotation is easily performed.

The most extended and most innovative feature of ZEOBUILDER, which makes it unique and different from other available graphical GUIs, is the multiple builder tools built into the code. Two zeolite building blocks can merge to one larger system with the help of condensation reactions. The mechanism is illustrated in Figure 3. The building blocks are assumed to be rigid structures, and a connection between two building blocks is obtained by rotating and translating one of the two blocks in an attempt to search for a possible set of oxygen pairs that fulfill requirements for a successful condensation process. Terminating hydrogens are hidden in order to make the figure more transparent. The selected oxygen pairs are then connected with a spring, as shown in Figure 3. Introducing springs in a structure always precedes some optimization procedure. In this condensation reaction, the rigid body optimization brings together the two merging blocks until a situation is obtained with partially overlapping oxygens (those connected with a spring). The algorithm applied to perform this condensation will be described in the next section. Finally, each pair of partially overlapping oxygens is replaced by one bridging oxygen atom that connects the two building blocks. In this way, one of the multiple conformations is constructed by merging the two

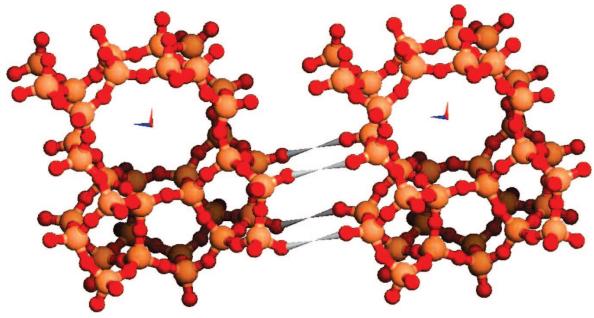


Figure 3. Illustration of how the condensation takes place after choosing manually some pairs of merging oxygens in the two building blocks.

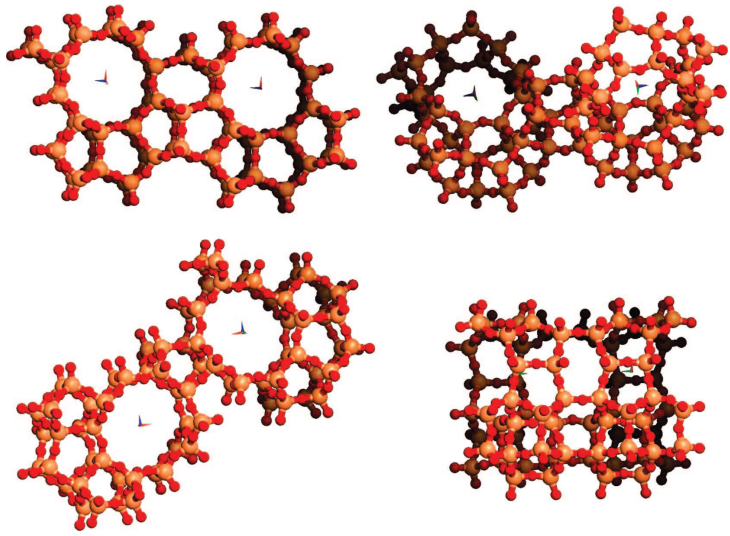


Figure 4. Four plausible conformations after a condensation scan of two zeolite building blocks (MFI precursors[19–21]).

zeolite blocks. The manual search for possible pairs of oxygens appropriate for condensation can be very time-consuming. So, an automatic scanner of all plausible condensation reactions is built in and leads to a series of conformations ordered following a quality factor, which depends on a series of parameters mostly related to steric hindrance and other geometry effects resulting from the merging process. This quality factor reflects the plausibility of formation of the particular conformation.

As an example, in Figure 4, we show four plausible compositions (conformations) resulting from a scan of condensation reactions between two MFI precursors.[19–21] This procedure can be repeated without any limitation and opens perspectives to construct new zeotype materials with different microporosities and topologies of the framework. Obviously, the applicability of this building ability can be easily extended to all types of inorganic nanosized building blocks.

The concept of springs, as introduced in the building process of molecular blocks, can be extended to other applications. A spring can be added to keep a particular (organic) molecule or cations localized in a channel or cage of a zeolite framework, after which an optimization can be performed. There are also other applications outside the realm of zeolite science: for example, in building complexes of molecules where some geometrical constraints are imposed pushing the complex in a well-directed local minimum of the PES, the concept of springs can be very useful. In addition, these tools are very helpful in creating input files for molecular modeling program packages such as Gaussian 03,[22] ADF,[23,24] CPMD,[26] CP2K,[27,28] GULP,[29] CHARMM,[30] GROMACS,[31] GROMOS,[32] LAMMPS,[33] NAMD,[34] and so on. During the past few decades, many force-field models have been developed that are applicable to zeolite systems. In particular, the Catlow library[35] and the parameters derived by Sauer et al.[36,37] are widely used.

In addition to these advanced model manipulations, ZEOBUILDER also contains a series of suitable tools to incorporate elements such as B, Ge, Zn, P, and transition elements into the framework creating different molecular sieves. Additionally, ZEOBUILDER also supports one-, two-, and three-dimensional periodic systems and the tools to transform a periodic system into a cluster, to create super cells, or to wrap a cluster model into a periodic box.

All of the features above focus on model manipulation; that is, given one or more input structures, one can modify and compose these structures into any reasonable model that serves as input for computational chemistry packages. ZEOBUILDER also contains tools to extract (nontrivial) geometrical or topology-related information from a molecular model. With one of the menu options, one can also make some statistical analysis on bond lengths, bending, dihedral angles, and so forth in a molecule or parts of a molecule. One can take into account all internal coordinates in such an analysis, or one can impose a specific subset of internal coordinates. Another interesting plug-in module is the counter tool. To illustrate with an example: in a complex zeolite structure, an interesting piece of information could be the number of TO_4 tetrahedra in the model. Due to the finite size of zeolite models, the T-centra are not all connected with four bridging oxygens, and different silicon environments are present. It is not obvious to extract the distribution of the various Q1, Q2, Q3, and Q4 centers.

A plug-in is also implemented in ZEOBUILDER to count the number of rings in the framework, which is not a trivial task especially in large structures. Pore sizes are also easily measured.

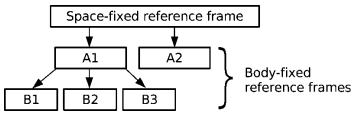


Figure 5. A schematic example of a hierarchical structure of reference frames.

3. CONCEPTS AND ALGORITHMS

3.1. Hierarchical Structure—Multiple Reference Frames. The internal data structure, employed by ZEOBUILDER to represent the molecular model, is based on a hierarchical concept and enables the development of diverse builder algorithms in the program.

In most builder programs, the standard approach is to define all atom coordinates with respect to one space-fixed reference frame. ZEOBUILDER abandons this concept by introducing multiple body-fixed reference frames to be defined in a hierarchical structure. Each frame contains a building block or an aggregation of building blocks. The builder algorithms in ZEOBUILDER will manipulate these reference frames instead of the individual atoms. A schematic example is given in Figure 5. In this example, the space-fixed frame contains two body-fixed frames (A1 and A2), and A1 is again composed of three other body-fixed reference frames (B1, B2, and B3).

Each frame contains a list of objects such as points, subframes, and so on. Within the terminology of ZEOBUILDER, these objects are called children as elements of the frame (the parent). Each object has a coordinate, t, with respect to the parent frame. Body-fixed reference frames also store a rotation matrix, **R**, that describes the orientation relative to the parent frame. The absolute position of an atom C1 in reference frame B2 is then given by $t_{\mathrm{C1,\,abs}} = R_{\mathrm{A1}}(R_{\mathrm{B2}}t_{\mathrm{C1}} + t_{\mathrm{B2}}) + t_{\mathrm{A1}}$. The notation and the implementation of this concept is largely facilitated by merging the rotation matrix and translation vector of each reference frame into a single transformation matrix, **U**, with the following structure:

$$U_{\mathrm{A1}} = \begin{bmatrix} R_{\mathrm{A1}} & t_{\mathrm{A1}} \\ 0 & 1 \end{bmatrix} \tag{1}$$

This reduces the expression for the absolute position of an atom C1 in reference frame B2 to

$$\begin{bmatrix} t_{\mathrm{C1,abs}} \\ 1 \end{bmatrix} = U_{\mathrm{A1}}U_{\mathrm{B2}} \begin{bmatrix} t_{\mathrm{C1}} \\ 1 \end{bmatrix} \tag{2}$$

Similarly, one can express the position of atom C1 relative to the reference frame A2:

$$\begin{bmatrix} t_{\mathrm{C1,A2}} \\ 1 \end{bmatrix} = U_{\mathrm{A2}}^{-1}U_{\mathrm{A1}}U_{\mathrm{B2}} \begin{bmatrix} t_{\mathrm{C1}} \\ 1 \end{bmatrix} \tag{3}$$

This data structure also has a major technical advantage; that is, the accelerated graphics hardware that is responsible for the 3D visualization of the model can efficiently process a ZEOBUILDER data structure because the hardware implementation is also based on algebra of four-dimensional vectors and a hierarchical structure of reference frames.

3.2. Condensation Algorithm. One of the main operations in building nanosized zeolite structures is the polycondensation reaction of two zeolite blocks. The mechanism is shown in Figure 6. Two or more terminating oxygen atoms are selected in both blocks A and B. By rotating and translating the rigid body B, one can find a situation where sets of oxygen pairs fulfill all requirements for a successful condensation process. The two blocks are connected by the newly formed oxygen bridges (see Figure 6a). The algorithm consists of fine-tuning the position of the bridging oxygen atoms so that they are exactly coinciding. Condensation takes place with a release of water molecules. Hydrogens are hidden in order to make the visualization of the process more transparent. In the case of a condensation reaction with only two oxygens, a degree of freedom is left (rotation of block B about the O−O axis). For three or more coinciding oxygens, the position of block B is definitely determined. Due to the fact that the building blocks are handled as rigid structures, the overlap will only be approximate. An algorithm is developed that minimizes the cost function J defined as the sum of the squared O−O distances that are expected to coincide. An example of two building units with two oxygen pairs logically connected is illustrated in Figure 6b. The algorithm as implemented in ZEOBUILDER joins the two blocks by minimizing the cost function J. At the end of the operation, a minimum value of J remains. This minimum value is displayed, and the user is free to decide whether the condensation reaction is accepted. Next, a fine-tuning mechanism is put into operation in order to induce exact overlap of the siloxane bridges. This condensation mechanism creates a new conformation consisting of the two originally separate building blocks.

We have illustrated the condensation algorithm in a specific polycondensation reaction of zeolite blocks, but it should be stressed that the built-in algorithm can be applied to any kind of elimination reaction.

3.3. Multiple Scan of Various Conformations. The manual selection of possible condensation reactions can be of some utility but is not of practical use when multiple condensation reactions belong to the possibilities. The search for all possible conformations formed by merging two blocks A and B is automated in ZEOBUILDER. A conformation scanner has been built in that finds out all possible sets of oxygen pairs c_{ik} ($k = 1, K$) that fulfill requirements for a successful condensation process (label i defines the specific conformation under study). It is not the intention of the authors to go into detail how the algorithm is set up, but the main strategy is based on the search for equivalent or nearly equivalent triangles formed by three oxygen atoms in both blocks A and B. The triangles should overlap with each other within a fixed threshold value (cost function J represents a highly suitable measurement instrument). Other constraints are also built in to avoid unphysical constructions. For example, the merging of triangles formed by oxygens at large distances from each other generates spurious situations with tetrahedra in an unrealistic spatial ordering in most of the cases. All of these spurious constructions should be removed, and ZEOBUILDER only retains N suitable conformations, completely determined by the set of three or more oxygen pairs c_{ik} ($k = 1$, K). Each conformation is stored by means of the transformation matrix belonging to the rotation and translation of block B to form the actual conformation (see subsection 3.1). A search is performed to remove all duplicates leading to similar conformations. A course quality

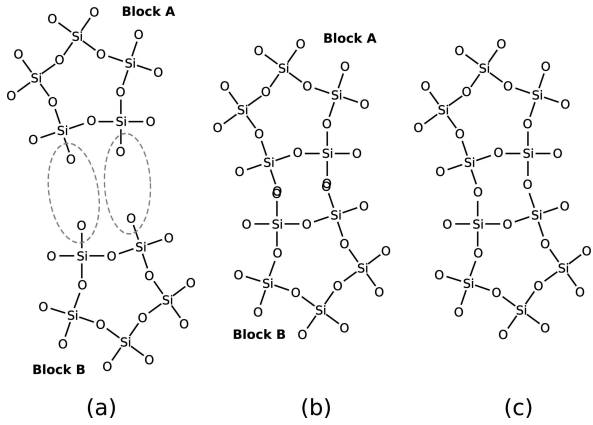


Figure 6. A schematic illustration of the condensation reaction.

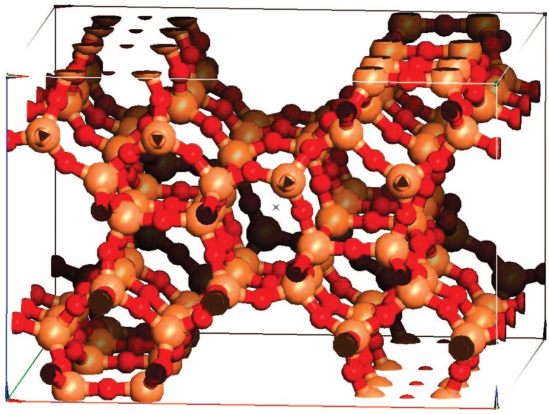


Figure 7. The unit cell of silicalite-1 (MFI).

factor is calculated for each conformation, and the results are sorted by this quality factor.

This procedure can be repeated with a third block, C, merging with one of the retained conformations $[AB]_i$. In this way, ZEOBUILDER is able to build nanoscale zeolite structures.

4. HOW TO WORK WITH ZEOBUILDER

4.1. Example 1: A Model for Zeotiles. In the first example, we focus on zeotiles, which are crystalline silica materials with two levels of porosity and structural order, and which are built from nanoslabs of uniform size with the Silicalite-1 zeolite framework generated by the TPA template.[38] Different tiling patterns have been imaged by high-resolution electron microscopy,[39] and they form a suitable and advanced application of ZEOBUILDER in an attempt to model the observed tiling patterns. The initial building block of an MFI nanoslab is the MFI precursor described in the work of Kirschhock et al.[19–21] In these papers, the mechanism of formation of the nanoslabs from the precursor is discussed.

First, we demonstrate how the MFI precursor as a building block is derived from the silicalite-1 unit cell. Next, nanoslabs are built from these building blocks. Finally, nanoslabs are tiled into the characteristic hexagonal pattern of Zeotile-1.[39]

Construction of the MFI Precursor. With the menu function "File" > "Open", one loads the Silicalite-1 unit cell from the IZA library[12] (see Figure 7). The space-fixed reference frame is represented by a right-hand coordinate system in the lower-left corner of the unit cell. By convention, the *x* axis is colored red, the *y* axis green, and the *z* axis blue. As this unit cell is rectangular, the *x*, *y*, and *z* axes coincide with the crystal directions *a*, *b*, and *c*. In the initial 3D view on the structure, the *z* axis lies perpendicular to the screen. In the tree display, all atoms of the unit cell are labeled. The next step is the transformation of the unit cell

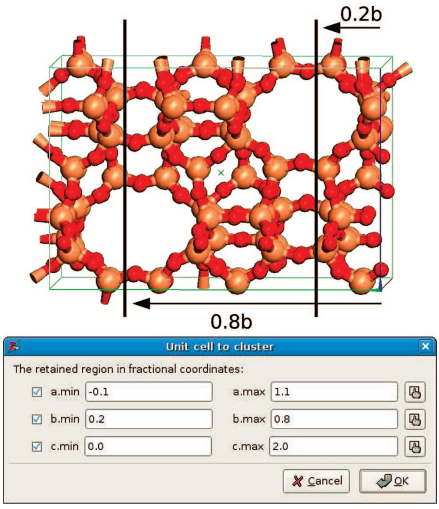


Figure 8. The Sodalite-1 unit cell viewed along the 100 direction. The selected interval along the 010 direction is indicated by vertical lines.

into a cluster that is large enough to contain at least the desired MFI precursor, as suggested by Kirschhock et al.[21,40] From this cluster model, one can remove the redundant atoms with the interactive selection tool. For the transformation of a periodic model into a cluster, one must specify the fractional coordinates of the region in the unit cell that should be retained. In order to estimate this region, it is instructive to rotate the unit cell in all directions with the interactive rotation tool and to examine in particular the crystal views along the three crystal directions. One proceeds as follows:

1. Select the global reference frame, that is, the top item in the tree structure.

2. Click "Object" > "Unit cell" > "To cluster".

3. A dialogue box appears, prompting the user to specify the cutoff values in fractional coordinates in the three crystal directions (see Figure 8). Click OK.

4. A transformation of the space-fixed frame to the center of mass is desirable for further operations, and this is carried out by selecting the frame and clicking "Object" > "Transform" > "Center of mass frame". This menu option keeps the crystal axes in the original directions.

5. The redundant atoms can now be removed with the interactive selection tool. Use the left mouse button to select a part of the model. All objects in the drawn rectangle become selected. Alternatively, one can also point and click at objects. Additionally, one can also use the middle mouse button to extend the current selection. Similarly, the right mouse button can be used to remove parts from the selected region. The function "Edit" > "Delete" will remove the selected objects from the model. Once the MFI precursor is constructed, it can be saved for further applications.

Construction of an MFI Nanoslab. Following the lines given in refs 19–21 for the construction of a half-nanoslab, we start with joining three building blocks (MFI precursors) along the *c* direction of the MFI structure. This is achieved with the following steps:

1. It is desirable to arrange the MFI precursor in a separate body-fixed reference frame. This is extremely useful when introducing multiple precursors in a building process. They can each individually be submitted to all geometrical manipulations. To carry out this operation, first select the global reference frame. Then, choose the menu option "Select" > "Children" followed by "Object" > "Arrange" > "Frame".

2. Then, this frame (and its contents) is duplicated. Select the body-fixed reference frame, and click "Edit" > "Duplicate".

3. Select the duplicated frame, and click "Object" > "Transform" > "Translate". A dialogue box prompts the user to specify the desired translation. Enter the 001 cell vector ($t.x = 0$ Å, $t.y = 0$ Å, and $t.z = 13.42$ Å), and click OK.

4. A third duplicate of the precursor is created in a similar way, but now the duplicate is translated in the opposite direction. With a translation over the exact periodicity in the *c* direction, a situation is created with physically distinct but completely overlapping oxygens. It is obvious that one oxygen atom should be removed from each duo, but we will postpone this operation until the end of this example.

If necessary, the scale of the 3D view can be modified with the interactive zoom function, that is, keep the Ctrl and the Shift keys pressed while dragging with the left mouse button in the 3D view.

In a second step, this row of three precursors is again arranged in a separate reference frame, and this new frame is duplicated. The duplicate is reflected with respect to a plane orthogonal to the *b* direction. Carry out following operations:

1. Select the three precursor frames, and activate the menu function "Object" > "Arrange" > "Frame".

2. Place the view of the three-precursor ensemble orthogonal to the *b* axis. In this position, the atoms at both the left and right edges are aligned.

3. Select all of the atoms in one of the two edges, and click "Object" > "Add" > "Plane". The reflection plane is created and is visualized on the display.

4. Select the frame that contains the three MFI precursors.

5. Duplicate this frame with "Edit" > "Duplicate".

6. Select the duplicated frame, and add one of the two reflection planes to the selection.

7. Perform the reflection by choosing "Object" > "Transform" > "Reflection". A dialogue window with the reflection parameters pops up. The numbers are filled in automatically, on the basis of the selected reflection plane. Click OK.

8. Select the reflection planes, and delete them with "Edit" > "Delete".

9. Select both the original and the reflected reference frame, and click "Object" > "Arrange" > "Frame". As indicated above and shown in Figure 9, there are two possible reflection planes, and both options result in different (pure MFI) geometries for the half-nanoslab.

Construction of Zeotile-1. Following the description as given in the work of Kremer et al.,[39] two half-nanoslabs are now joined into face-sharing double units, measuring 2.6 nm × 2.0 nm × 4.0 nm. As we succeeded in building two types of half-nanoslabs, the construction of a double unit is not unique, even when this double unit must exhibit a proper MFI topology. In

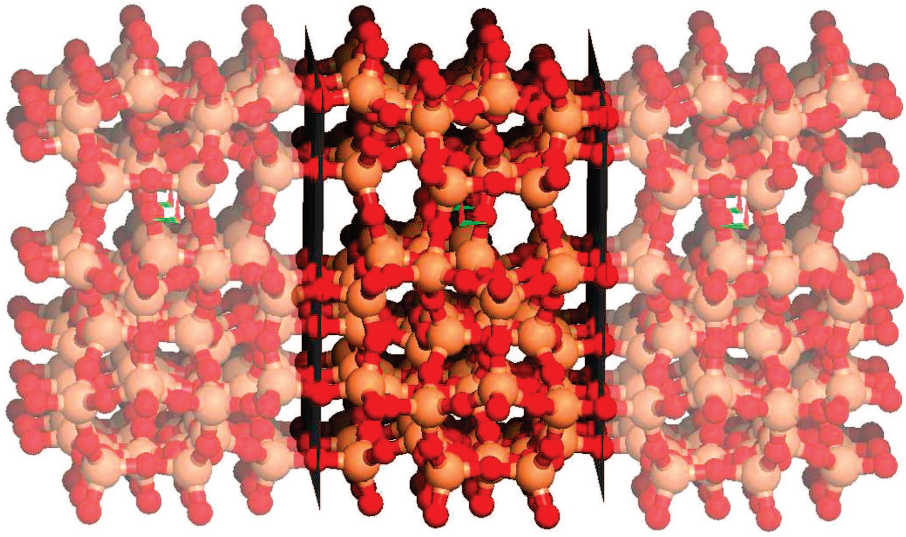


Figure 9. The two possible mirror planes for duplicating the frame with three precursors.

this example, we restrict ourselves to a double unit constructed by identical half-nanoslabs. In this configuration, the two dense edges of both half-nanoslabs—edges orthogonal to the *a* direction of the MFI structure—are shared. This can be performed by the following steps:

1. One chooses and creates a reflection plane containing all terminating oxygens at the dense side of the half-nanoslab.
2. Duplicate the whole construction. Carry out a reflection of the duplicated frame with respect to the reflection plane.
3. Eliminate the reflection planes.
4. The final structure is stored in a separate reference frame. Finally, in order to create a hexagonal pattern based on the double units, some more advanced techniques are required. First, two duplicates of the double unit are rotated +120° and −120°, respectively, along the long direction (*c* axis) of the double unit. The rotation of one double unit is carried out with the following steps:

1. One first defines the rotation axis parallel to the *c* direction of the MFI structure by connecting two atoms with a vector. On the basis of the structure of the MFI precursor, a pair of atoms that defines the correct rotation axis can be easily selected.
2. In order to rotate a double unit, one first selects the reference frame that contains the double unit, and then one adds the rotation vector from the previous step to the selection. The menu function "Rotate about axis" first shows a popup dialogue that contains all of the rotation parameters (rotation axis, rotation center, and rotation angle). One only needs to give an input value for the rotation angle since the other parameters are filled in automatically. The rotation axis is completely determined by the selected vector. After clicking the "OK" button, the double unit is rotated.
3. A similar operation is carried out with the third double unit, but now a rotation over −120° is performed. The three double units can now be moved into a triangle in such a way that only at the corners are oxygen atoms overlapping. This is illustrated in Figure 10. A condensation reaction then makes a new structure. The conformation shown in the figure is not the unique one. There are probably others, but we gave preference to a regular triangle structure with an edge of approximately 2.0 nm.

The next operation is a periodic extension of the triangle of three double units to create the ultime hexagonal/triangular pattern of Zeotile-1. One can simply define the new unit cell vectors by drawing them in the model. Two half-unit cell vectors can be obtained by connecting the centers of the double units with a vector (see Figure 10). The lengths of these vectors can be doubled. The third unit cell vector is orthogonal to the first two, that is, parallel to the *c* direction in the MFI structure or the rotation axis defined above. It can by obtained by drawing a vector from the origin of the reference frame of one MFI precursor to the origin of the reference frame of a second MFI precursor that is translated along the *c* direction with respect to the first one. To finalize the model, one takes the following steps:

1. Select the three vectors that define the unit cell, and click "Object" > "Unit cell" > "Define unit cell vector(s)". They are indicated as "Arrow" in the tree display.
2. The initial building blocks (MFI precursors) are still stored in a hierarchy of separate reference frames, which is impractical for further computational applications. By choosing the menu option "Object" > "Arrange" > "Unframe" for all of these reference frames, all of the atoms are directly positioned in the global reference frame.
3. The overlapping oxygen atoms that connect the building blocks are merged with the menu option "Merge overlapping atoms".

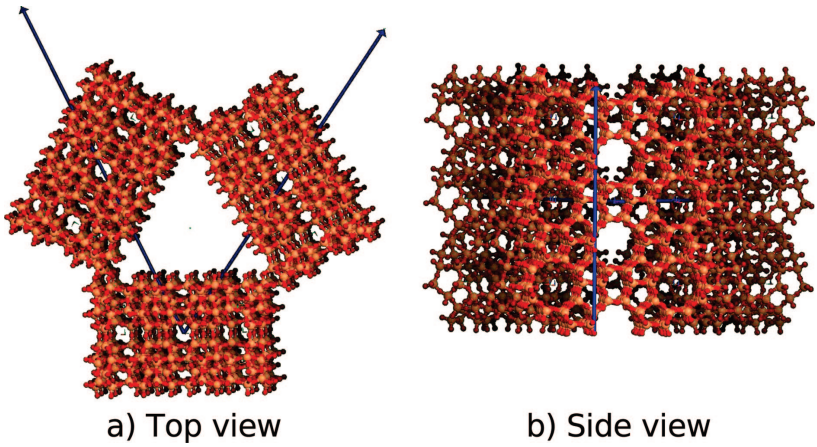


Figure 10. Two visualizations of the triangular structure, which is the basis for the zeotile unit cell. The cell vectors are shown as blue arrows.

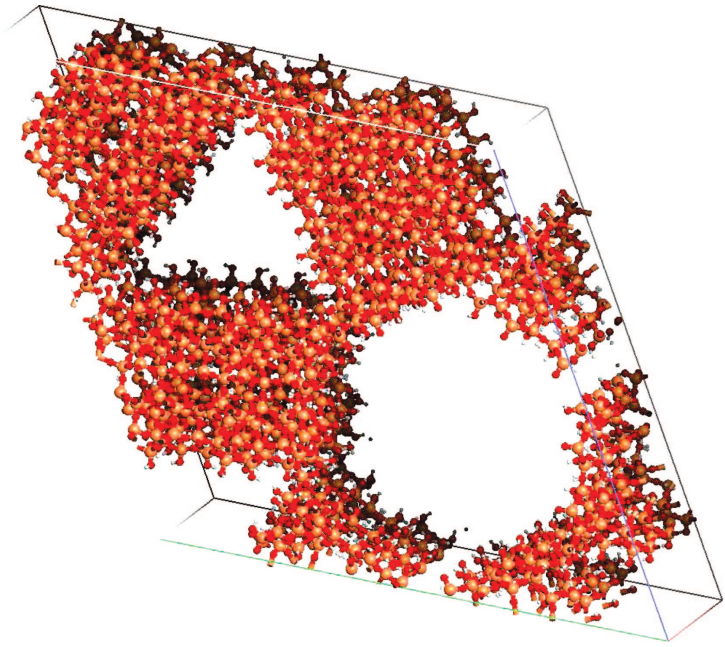


Figure 11. 3D view of Zeotile-1 constructed in example 1.

4. The final structure still contains many dangling oxygens. One can turn the model into a hydroxyl-terminated structure with the function "Saturate with hydrogens". In this way, a zeotile model has been built, and the result is seen in Figure 11.

4.2. Example 2: Aligning a Guest Molecule in a Zeolite Channel. This example will demonstrate how one can add a pentane molecule exactly in the center of the straight channel in a silicalite-1 model, with the pentane molecule aligned along the axis of the straight channel. One

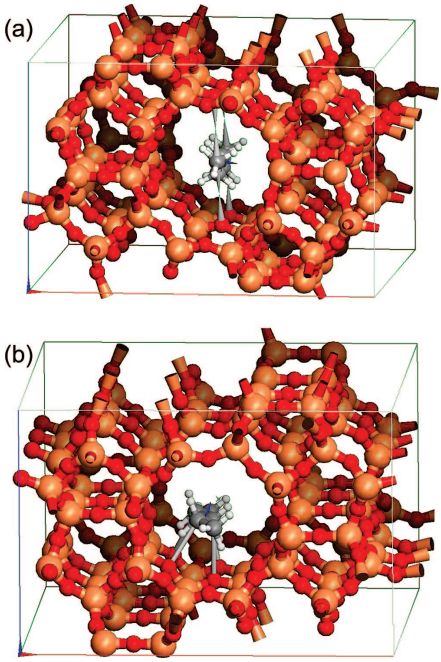


Figure 12. Embedding a pentane molecule in the MFI framework.

first loads the unit cell of Silicalite-1 (MFI) as in the preceding example. As shown in Figure 7, the straight channels of the MFI topology are located at the edges of the periodic box. It is instructive to carry out a translation of the reference frame to another frame with the origin at the center of one of the two straight channels in the Silicalite-1 unit cell.

1. Select the global reference frame.

2. Add a new point to the model with "Object" > "Add" > "Point". The point is automatically selected.

3. Modify the coordinates of the new point to (10 Å, 10 Å, 0 Å) with "Object" > "Properties". A dialogue box opens; choose the tab page Translation, and enter the translation coordinates. These coordinates approximately correspond with the origin of the new desired reference frame.

4. Apply the menu option "Object" > "Transform" > "Define origin". The display of the unit cell is now adapted to the new reference frame. The advantage of introducing a new origin of the reference frame is manifest. With the rotation tool, one easily rotates the model until the view coincides with the *ac* crystal plane. The 10T ring in the straight channel is clearly seen (see Figure 12).

The second step is to import the pentane molecule into the model with the menu function "File" > "Import". The molecule is loaded into a separate reference frame. This frame can be manually placed and oriented at a position where we want to embed the pentane molecule. This can be

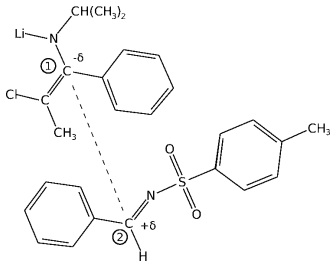


Figure 13. Schematic structure of the reaction wherein the carbon atom bonded to the nitrogen in a tosyl imine conducts an electrophilic attack on the carbon atom next to the chlorine in a lithiated 3-chloro-1-aza-allylic anion

carried out using the interactive rotation and translation tools. The second step makes use of the spring tool, as already outlined: the two ending C atoms of the pentane molecule are connected with oxygen atoms in the wall of the straight channel with springs, as shown in Figure 12a. One can easily introduce springs with the interactive sketch tool. Finally, one optimizes the springs with preset rest lengths. The location of the pentane molecule in the channels of the framework is not unique: one can connect the ending C atoms of the pentane molecule with different oxygen atoms in the zeolite channel, or one can specify different rest lengths, as shown in Figure 12b. All of these settings yield suitable geometrical structures that can be used as input for geometry optimizations in a variety of ab initio packages. This exercise leads to the quantitative description of the adsorption of pentane in the channels of a silica zeolite with MFI topology. One can also use these models as the initial geometry for a molecular dynamics simulation to study the diffusion of pentane in Silicate-1.

4.3. Example 3: Manipulating Organic Structures. The third example demonstrates how ZEOBUILDER can be applied in building molecular constructions beyond zeolite science. We have chosen an application that many computational chemists are confronted with when studying chemical kinetics. The search for the true transition state can be a time-consuming process, not only from the computational viewpoint. Also, the preparation of suitable input files with trial geometrical structures for the near transition state is a difficult task. In chemical reactions with complex reactants, the current graphical tools are not always appropriate for drawing starting geometries. The example below is an aldol-like reaction from the heterocyclic organic chemistry. More specifically, we build a near-transition-state structure of the reaction—displayed in Figure 13—wherein the carbon atom bonded to the nitrogen in a tosyl imine performs an electrophilic attack on the carbon atom next to the chlorine in a lithiated 3-chloro-1-aza-allylic anion. Once this model has been constructed, it can be used as the input geometry for a transition-state optimization in various quantum chemistry programs. The three active species are illustrated in Figure 14. The Li atom can coordinate with a lot of atoms from the two reactants. In principle, all combinations should be tried out. In this example, we take into consideration the situation where the Li atom is coordinated by one oxygen atom of the tosyl imine and by the nitrogen and chlorine

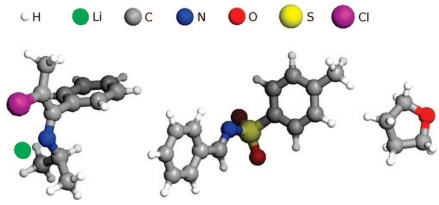


Figure 14. The three active species playing a role in the reaction under study. From left to right: the aza-allylic anion, the thosyl imine, and tetrahydrofuran (THF).

atoms in the aza-allylic anion. As input, we load the optimized structures for the lithiated 3-chloro-1-aza-allylic anion and the tosyl imine. It is essential that both molecules reside in their respective frames, so that they can be submitted to independent rotations/translations—each molecule keeping its internal structure. Good initial guesses for interatomic distances are 2 Å for the Li−O distances and 4 Å for the C−C distance of the forming bond. With the help of the interactive toolbar buttons, the two molecules are transformed in an acceptable position taking into account the estimates of coordination distances. But if too many constraints are imposed, it is more convenient to make use of the spring tools of ZEOBUILDER to facilitate the construction of a proper initial geometry. This is illustrated by the following steps:

1. Using the proper interactive tools, we bring the tosyl imine molecule into position with one of the two oxygens of the imine approximately at a coordination distance with the Li atom.

2. Then, we select the imine frame and consequently add the Li atom to the selection. This selection preserves the Li−O distance during an interactive rotation in the 3D view.

3. During the rotation of the imine frame, we bring the carbon atoms of the forming bonds closer to each other, but we prevent the overlap of the imine molecule with the aza-allylic anion. One gets a structure as displayed in Figure 15a.

4. We now fine-tune the relative orientation of the reacting species. First, we connect the atoms with the prescribed interatomic distances by springs. With the function "Object" > "Properties", we can configure the rest lengths of these springs and carry out a rigid body optimization. This procedure can be repeated until a satisfactory structure is obtained.

5. Other geometrical manipulations can be carried out, keeping all coordination distances unaltered and preventing many reoptimizations, by rotating about an axis formed by the coordinated oxygen and the C atom of the imine (displayed in Figure 15b). This is achieved by introducing an arrow between the two atoms. By rotating the imine about this axis, we can also fine-tune the relative orientation while the lengths of the springs remain constant. In this way, a near transition state is constructed that can serve as an input file in any ab initio molecular modeling package in search of the true transition state.

The case can even become more complex when taking into account solvent molecules. With ZEOBUILDER, the most complex configurations can be handled in an easy way. We first arrange the two reacting species into one new frame. This assures that their relative orientation will not change

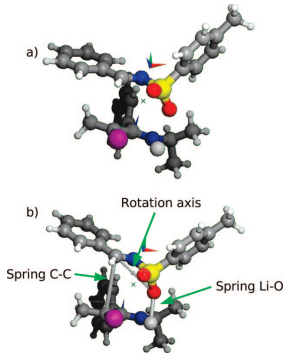


Figure 15. Two intermediate phases in the construction of the input geometry for a transition state optimization. The initial structure (a) is obtained through the interactive functions in ZEOBUILDER. The fine-tuning (b) is based on the spring optimization and a rotation that does not alter the lengths of the springs.

by further spring optimizations. Then, we import a THF molecule and connect the oxygen and the Li atom by a spring. We set the rest length of this spring to 2 Å, and we optimize the spring. The THF molecule can now be oriented with the interactive rotation tool, using the oxygen or the Li atom as a rotation center.

5. PROGRAM AVAILABILITY

Zeobuilder is distributed as open-source software under the conditions of the GNU General Public License, version 3. The software can be downloaded from the code Web site of the Center for Molecular Modeling: http://molmod.ugent.be/code. Documentation, tutorials, example files, installation instructions, and technical support are also available on this Web site.

6. SUMMARY AND OUTLOOK

The continuous advances in computational chemistry and zeolite modeling in particular have increased the demand for a molecular builder toolkit that fulfills all criteria of modern software. ZEOBUILDER is a strong answer to this real need for molecular editors that are used for the preparation of ab initio and molecular mechanics simulations.[43,44]

It is a common practice to describe complex molecular structures, for example, zeolite frameworks, in terms of polyatomic building units. ZEOBUILDER's hierarchical data structure of reference frames largely facilitates the construction and manipulation of complex models based on building blocks.

ZEOBUILDER is also an extensible framework for the development of novel and highly specialized builder algorithms. The spring optimization and the connection scanner, which have been thoroughly demonstrated in this paper, are two practical examples of builder tools that have been developed in the ZEOBUILDER framework. Despite our interest in theoretical zeolite modeling, we have carefully designed this framework with a broad range of applications in mind: no assumptions have been made that

could limit the transferability and the feasibility of ZEOBUILDER.

ZEOBUILDER, as presented in this paper, is only a snapshot from a dynamic development process. We plan diverse updates and extensions to ZEOBUILDER that will address some of the remaining challenges in the construction of sophisticated molecular models. A true geometry optimization, even at a low level of theory (e.g., UFF[41] or PM3[42]), would relieve the current limitation of purely rigid building units. For example, the spring optimization and the scanner assume that the building units are approximately rigid. This rigid-body assumption is adequate for the preparation of most molecular simulations, but it will give an incomplete picture of a synthesis process based on (partially) flexible building blocks. A next generation of scanning algorithms is in preparation to surmount these additional difficulties. Another attractive feature would be the stacking of secondary building units into zeolite crystals, optionally including stacking faults.[11,12] Other items on our schedule include an improved interoperability with external simulation codes and support for more file formats via OpenBabel.[18]

Since ZEOBUILDER is released as an open-source platform for the development of molecular builder algorithms, the authors look forward to collaborations with external researchers and developers in this area.

ACKNOWLEDGMENT

This work is supported by the Fund for Scientific Research - Flanders and the Institute for the Promotion of Innovation through Science and Technology in Flanders (IWT Vlaanderen).

REFERENCES AND NOTES

(1) *Cerius2 Modelling Environment*; Accelrys Software Inc.: San Diego, 2006.
(2) *Materials Studio*; Accelrys Software Inc.: San Diego, 2006.
(3) *CrystalMaker*; CrystalMaker Software Limited: Oxfordshire, UK, 2008.
(4) *Diamond*; Crystal Impact: Bonn, Germany, 2008.
(5) Pereira, C. General Atomistic Modelling Graphic Interface. http://www.gamgi.org/ (accessed Mar 25, 2008).
(6) DeLano, W. *The PyMOL Molecular Graphics System*; DeLano Scientific: Palo Alto, CA, 2002.
(7) Jmol: an open-source Java viewer for chemical structures in 3D. http://www.jmol.org/ (accessed Mar 25, 2008).
(8) Ali, S.; Braithwaite, R.; Bunt, J.; Curtis, D.; Hanwell, M.; Hutchison, G.; Jacob, B.; Margraf, T.; Niehaus, C.; Ochsenreither, S.; Vandermeersch, T. Avogadro. http://avogadro.sourceforge.net (accessed Mar 25, 2008).
(9) Agile Molecule. http://agilemolecule.com (accessed Mar 25, 2008).
(10) Smith, J. *Chem. Rev.* **1988**, *88*, 149–182.
(11) Lobo, R. F. *Introduction to the Structural Chemistry of Zeolites. In Handbook of Zeolite Science and Technology*, Auerbach, S., Carrado, K., Dutta, P., Eds.; Marcel Decker Inc.: New York, 2003; pp 65−89.
(12) Baerlocher, C.; McCusker, L.; Olson, D. *Atlas of Zeolite Framework Types*, 6th ed.; Elsevier: Amsterdam, The Netherlands, 2007; pp 1−398.
(13) Lewis, D. W.; Willock, D.; Catlow, C. R. A.; Thomas, J. M.; Hutchings, G. *Nature* **1996**, *382*, 604–606.
(14) Lewis, D. W.; Catlow, C. R. A.; Thomas, J. M. *Faraday Discuss.* **1997**, *106*, 451–471.
(15) Baerlocher, C.; Hepp, A.; Meier, W. *DLS-76*; Institut für Kristallographie und Petrographie: Zürich, Switzerland, 1976.
(16) Bail, A. L. *J. Appl. Crystallogr.* **2005**, *38*, 389–395.
(17) Sheldrick, G. *Acta Crystallogr.* **2008**, *A64*, 112–122.
(18) Guha, R.; Howard, M.; Hutchison, G.; Murray-Rust, P.; Rzepa, H.; Steinbeck, C.; Wegner, J.; Willighagen, E. *J. Chem. Inf. Model.* **2006**, *46*, 991–998.
(19) Ravishankar, R.; Kirschhock, C.; Knops-Gerrits, P.; Feijen, E.; Grobet, P.; Vanoppen, P.; De Schyver, F.; Miehe, G.; Schoeman, B.; Jacobs, P.; Martens, J. *J. Phys. Chem. B* **1999**, *103*, 4960–4964.
(20) Kirschhock, C.; Ravishankar, R.; Van Looveren, L.; Jacobs, P.; Martens, J. *J. Phys. Chem. B* **1999**, *103*, 4972–4978.
(21) Kirschhock, C.; Kremer, S.; Grobet, P.; Jacobs, P.; Martens, J. *J. Phys. Chem. B* **2002**, *106*, 4897–4900.
(22) Frisch, M. J. et al. *Gaussian 03*; Gaussian, Inc.: Wallingford, CT, 2004.
(23) te Velde, G.; Bickelhaupt, F.; van Gisbergen, C. F. G.; Baerends, E.; Snijders, J.; Ziegler, T. *J. Comput. Chem.* **2001**, *22*, 931–967.
(24) Guerra, C. F.; Snijders, J.; te Velde, G.; Baerends, E. *Theor. Chem. Acc.* **1998**, *99*, 391.
(25) Baerends, E. et al. *ADF*, 2007.01; SCM: Amsterdam, The Netherlands, 2007.
(26) *CPMD*, version 3.11; IBM Corp.: Armonk, NY, 1990−2006; MPI für Festkörperforschung Stuttgart: Stuttgart, Germany, 1997−2001.
(27) CP2K. http://cp2k.berlios.de (accessed on Mar 25, 2008).
(28) Vandevondele, J.; Krack, M.; Mohamed, F.; Parrinello, M.; Chassaing, T.; Hutter, J. *Comput. Phys. Commun.* **2005**, *167*, 103–128.
(29) Gale, J. *JCS Faraday Trans.* **1997**, *93*, 629.
(30) Janezic, D.; Brooks, B. *J. Comput. Chem.* **1995**, *16*, 1543.
(31) Lindahl, E.; Hess, B.; van der Spoel, D. *J. Mol. Modell.* **2001**, *7*, 306–317.
(32) Christen, M.; Hünenberger, P. H.; Bakowies, D.; Baron, R.; Bürgi, R.; Geerke, D. P.; Heinz, T. N.; Kastenholz, M. A.; Kräutler, V.; Oostenbrink, C.; Peter, C.; Trzesniak, D.; van Gunsteren, W. F. *J. Comput. Chem.* **2005**, *26*, 1719–1751.
(33) Plimpton, S. J. *J. Comput. Phys.* **1995**, *117*, 1–19.
(34) Phillips, J. C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R. D.; Kale, L.; Schulten, K. *J. Comput. Chem.* **2005**, *26*, 1781–1802.
(35) Lewis, G. V.; Catlow, C. R. A. *J. Phys. C: Solid State Phys.* **1985**, *18*, 1149–1161.
(36) Schröder, K. P.; Sauer, J. J. *Phys. Chem.* **1996**, *100*, 11043–11049.
(37) Sierka, M.; Sauer, J. *Faraday Discuss.* **1997**, *106*, 41–62.
(38) Kirschhock, C.; Buschmann, V.; Kremer, S.; Ravishankar, R.; Houssin, C.; Mojet, B.; Van Santen, R.; Grobet, P.; Jacobs, P.; Martens, J. *Angew. Chem., Int. Ed.* **2001**, *40*, 2637–2640.
(39) Kremer, S.; Kirschhock, C.; Aerts, A.; Vilani, K.; Martens, J.; Lebedev, O.; Van Tendeloo, G. *Adv. Mater.* **2003**, *15*, 1705–1707.
(40) Kirschhock, C.; Ravishankar, R.; Verspeurt, F.; Grobet, P.; Jacobs, P.; Martens, J. *J. Phys. Chem. B* **1999**, *103*, 4965–4971.
(41) Rappe, A.; Casewit, C.; Colwell, K.; Goddard, W.; Skiff, W. *J. Am. Chem. Soc.* **1992**, *114*, 10024–10035.
(42) Stewart, J. J. P. *J. Comput. Chem.* **1989**, *10*, 221–264.
(43) Hocevar, S.; Hodoscek, M.; Penca, M.; Janezic, D. *J. Comput. Chem.* **2008**, *29*, 122–129.
(44) Lesthaeghe, D.; Vansteenkiste, P.; Verstraelen, T.; Ghysels, A.; Kirschhock, C.; Martens, J.; Van Speybroeck, V.; Waroquier, M. *J. Phys. Chem. C* **2008**, *112*, in press.

CI8000748

Paper 2: "MD-TRACKS: A Productive Solution for the Advanced Analysis of Molecular Dynamics and Monte Carlo Simulations"

Toon Verstraelen, Marc Van Houteghem, Veronique Van Speybroeck, Michel Waroquier

Journal of Chemical Information and Modeling, **2008** , 48, 2414 - 2424

2414 *J. Chem. Inf. Model.* **2008,** *48,* 2414–2424

MD-TRACKS: A Productive Solution for the Advanced Analysis of Molecular Dynamics and Monte Carlo simulations

Toon Verstraelen, Marc Van Houteghem, Veronique Van Speybroeck, and Michel Waroquier*

Center for Molecular Modeling, Ghent University, Proeftuinstraat 86, 9000 Ghent, Belgium

Received July 11, 2008

In this paper, we present MD-TRACKS, an advanced statistical analysis toolkit for Molecular Dynamics and Monte Carlo simulations. The program is compatible with different molecular simulation codes, and the analysis results can be loaded into spreadsheet software and plotting tools. The analysis is performed with commands that operate on a binary trajectory database. These commands process not only plain trajectory data but also the output of other MD-TRACKS commands, which enables complex analysis work flows that are easily programmed in shell scripts. The applicability, capabilities, and ease of use of MD-TRACKS are illustrated by means of examples, that is, the construction of vibrational spectra and radial distribution functions from a molecular dynamics run is discussed in the case of tetrahydrofuran. These properties are compared with the experimental data available in the literature. MD-TRACKS is open-source software distributed at http://molmod.ugent.be/code/.

1. INTRODUCTION

Molecular dynamics (MD) and Monte Carlo (MC) simulations are well-established modeling techniques in diverse research fields, ranging from catalysis over biochemistry to solid-state physics.[1−5] Historically, most MD/MC simulation software relied on molecular mechanics models to cope with the cost of computing the potential energy and the interatomic forces.[6] However, during the past decade, the development of specialized algorithms and continuously progressing computer technology have made ab initio molecular dynamics[7−11] a viable and attractive alternative to the conventional molecular mechanics methods. Despite the ever-increasing computing power, molecular mechanics remains relevant for its scalability toward large systems, longer time scales, and the applications in hybrid QM/MM methods.[12] Today, a myriad of MD/MC simulation software is available (CPMD,[7,13] CP2K,[8,14] LAMMPS,[15,16] DL_POLY,[17] CHARMM,[18] NAMD,[19] GULP,[20] GROMACS,[21] GROMOS,[22] CERIUS2,[23]...) that enables one to generate a vast amount of trajectory data by integrating the equations of motion of the system under study. The trajectory data does not only contain the time-dependent atomic coordinates, but also many other quantities as a function of time, for example, atomic velocities, forces, cell parameters, pressure, temperature, partial charges, dipole moments, polarizability, different types of energies, orbitals, and so on. To translate the raw trajectory data into relevant scientific results, a proper statistical analysis is indispensable.

Several trajectory analysis solutions are readily available, for example, Visual Molecular Dynamics[24] (VMD), GROMACS,[21] PTRAJ,[25] SIRIUS,[26] and SMILYS.[27] They are certainly of high quality, and some of them offer impressive visualization functions for molecular dynamics simulations. Despite the presence of these valuable analysis tools, we felt the need for a more generic software program that enables complex analysis that goes beyond the standard functions present in the current tools. Ideally, a trajectory analysis program should also be flexible enough to be adapted for many different applications. In this paper, we present MD-TRACKS, a versatile, user-friendly, and freely available toolkit that addresses this challenge. We have tested the existing analysis software and prototypes of MD-TRACKS in foregoing studies[28−30] which has strongly influenced the software design. MD-TRACKS has some distinctive characteristics that make it suitable for a wide range of applications:

(1) MD-TRACKS is compatible with multiple MD/MC simulation codes. Prior to the actual analysis, the trajectory data is converted into a simple, manageable, fast, and cross-platform binary database. The current version of MD-TRACKS (0.003) has an interface to CP2K,[14] CPMD,[13] LAMMPS,[15,16] DL_POLY,[17] and CERIUS2,[23] and our software is easily extended to process trajectory data from other simulation codes.

(2) An analysis task is solved with a series of consecutive MD-TRACKS commands in a solution work flow. Each command loads only the actively used parts of the database in memory to analyze huge amounts of data without memory limitations.

(3) MD-TRACKS commands are orthogonal, that is, the output of each command is written to the database and can be used as input for any other command. MD-TRACKS has a built-in plotting function, but it is also possible to convert the analysis results from the binary database into plain text format that is supported by most spreadsheet applications and plotting tools.

(4) An MD-TRACKS programming library is provided to create custom Python scripts that can access the binary database. Once parts of the database are loaded in memory, efficient numerical operations are possible through NumPy.[31] If one can perform a very specific analysis task only partially

* To whom correspondence should be addressed. E-mail: michel.waroquier@ugent.be.

10.1021/ci800233y CCC: $40.75
Published on Web 11/18/2008

Table 1. Overview of the Most Important Modules in the MD-TRACKS Programming Library

module	functionality
`tracks.core`	low-level classes and auxiliary functions that efficiently read from or write to the binary MD-TRACKS database
`tracks.convert`	routines that convert the output of molecular dynamics and Monte Carlo programs into the binary format
`tracks.parse`	a set of functions that facilitate the interpretation of command line arguments
`tracks.vector tracks.cell`	tools for the manipulation of a collection of track files that represent time-dependent three-dimensional vectors or matrices

with the current MD-TRACKS commands, one can easily implement the remainder of the analysis in a specialized script using the MD-TRACKS programming library.

This paper describes both the implementation and the usage of MD-TRACKS. It is assumed that the reader has a basic knowledge of UNIX systems. The following section discusses the database format, the structure of a typical MD-TRACKS command, and an overview of the commands in the current MD-TRACKS version (0.003). In the third section, we give an impression of the capabilities of MD-TRACKS by showing how different types of vibrational spectra and radial distribution functions are easily computed from conventional trajectory data. In section four, we describe how MD-TRACKS can be obtained.

2. IMPLEMENTATION

2.1. MD-TRACKS Database Format. The MD-TRACKS toolkit stores all trajectory data and analysis output in a subdirectory `tracks` of the working directory where the molecular dynamics or Monte Carlo simulation software has written its output files. The commands that start with `tr-from-` convert the simulation output into binary files in the tracks directory. For example, `tr-from-cpmd-traj` converts a CPMD[13] trajectory file into binary data.

Each file in the directory `tracks` corresponds to a single time-dependent scalar, for example, `tracks/atom.pos.0000000.y` contains the values of the *Y* coordinate of the first atom for all iterations of the MD/MC simulation. This approach has the advantage that the database can be managed with simple UNIX commands like `ls`, `cp`, `rm`, and `mv`.

A single file in the database is called *a track*. Each track file consists of a fixed length header, followed by a binary data stream. The header contains a unique fingerprint and a format description of the binary data stream. The information in the header makes it possible to interpret the binary stream correctly on all computer architectures. The binary format has two major advantages, compared to conventional text files: (i) binary data is more efficient, both in terms of disk space and input/output performance, and (ii) binary data support random access, that is, one can read and write in track files at arbitrary positions without any overhead, while text files have to be read or written line by line.

MD-TRACKS follows these database-related conventions that facilitate the analysis process:

• All values in the MD-TRACKS database are stored in atomic units. The output files from MD/MC simulations are converted into atomic units by the `tr-from-*` scripts. Each MD-TRACKS command that produces human readable output, will present numerical data in the units specified by the user. All unit conversions are based on the CODATA basic constants and conversion factors.[32]

• Filenames in the MD-TRACKS database are reserved for a specific purpose. The most important reserved names are given below.

◦ `tracks/time`: This track contains the time axis, that is, the time in the simulated molecular system at each iteration step. In most cases, this is the integration time step multiplied by the step counter.

◦ `tracks/potential_energy`: This track stores the potential energies felt by the nuclei. It includes all interatomic interactions, that is, it is not limited to the Coulomb repulsion between the positively charged nuclei. In a molecular mechanics simulation, this also includes the valence interactions, Van Der Waals interactions, and so on. In the case of an ab initio simulation, this file contains the sum of the nuclei−nuclei, nuclei−electron and electron−electron interactions.

◦ `tracks/kinetic_energy`: This track contains the kinetic energy of the nuclei.

◦ `tracks/atom.pos.${index}.${c}`: These tracks hold the atomic positions, where `${index}`identifies the atom by an integer of seven characters and `${c}` is *x*, *y*, or *z*.

◦ `tracks/atom.vel.${index}.${c}`: These tracks contain the atom velocities, using the same conventions as in the previous item.

The proposed filename conventions are not strictly imposed but are strongly recommended when working with MD-TRACKS.

2.2. MD-TRACKS Commands. An overview of the MD-TRACKS commands (in version 0.003) is given in the Appendix. To execute a command, one enters its name at the UNIX command line shell, followed by options and arguments. The names of the commands are chosen to interact well with the tab-completion function in most popular UNIX shells. When the command line option `--help` is given as an argument to an MD-TRACKS command, the documentation for that command is printed on screen.

The MD-TRACKS commands are not meant to be entered manually, unless one explores new types of simulations or analysis techniques. In all other cases, it is much more efficient to collect a series of MD-TRACKS commands in a shell script that can be reused for the analysis of many MD/MC trajectories. This script-based approach automates the analysis process, but one still has the possibility to tune all parameters in this automated procedure.

Each command in the MD-TRACKS toolkit is based on the MD-TRACKS programming library which is a Python package that can be reused to write new MD-TRACKS commands for specialized applications. The most relevant modules in the MD-TRACKS programming library are listed in Table 1.

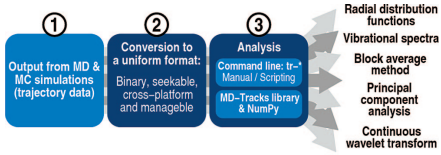


Figure 1. Workflow of an MD-TRACKS application.

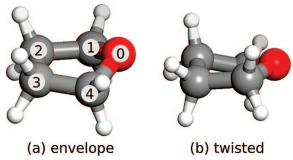


Figure 2. Two stable conformers of the THF molecule. On the basis of the molecular mechanics model for linear and cyclic ethers of Vorobyov et al.,[36] the twisted conformer is 2.5 kJ mol^{-1} lower in energy than the envelope conformer. The pseudorotation phase (as defined by Cremer and Pople[51]), is ±90° for the twisted and 0° or 180° for the envelope conformer.

3. EXAMPLE APPLICATION: VIBRATIONAL AND STRUCTURAL PROPERTIES OF TETRAHYDROFURAN

In this section, we present a selection of the various possibilities of MD-TRACKS by demonstrating how the vibrational and structural properties of tetrahydrofuran (THF) can be derived from a molecular dynamics simulation. This gives an idea of the usage pattern and the advantages of the MD-TRACKS software design. The typical MD-TRACKS workflow is illustrated in Figure 1. THF was chosen as a case study in view of its important role as structural unit in carbohydrates and biological systems and in the context of a general interest in the conformations and the ring-puckering of small-membered rings.[33,34]

A general consensus about the minimum energy structure of THF has not been achieved yet.[35,33] It is not the intention of this paper to resolve the question of the equilibrium geometry. Therefore, we will use a recently developed molecular mechanics force field for linear and cyclic ethers.[36] All conformational energies, vibrational frequencies and other analysis results mentioned below are obtained with this force field. The potential energy surface of THF in this molecular mechanics model has two local minima, corresponding to the twisted and envelope (or bended) conformer,[37,38] illustrated in Figure 2. The envelope conformer is 2.5 kJ mol^{-1} lower in energy. The vibrational frequencies of both conformers based on the harmonic oscillator approximation are listed in Table 2.

3.1. Vibrational Spectra of THF in the Gas Phase. In this example, we will apply MD-TRACKS to compute the infrared and inelastic neutron scattering (INS) spectrum of THF. In addition, we analyze the parts of the spectrum that are inherent to the pseudorotation motion, which is present in many puckered cyclic organic molecules.[39,40] The concept of pseudorotation was originally proposed by Kilpatrick et al. to explain the exceptionally high entropy of cyclopentane.[41] The IUPAC definition defines a psueodortation as a conformational change resulting in a structure that appears to have been produced by rotation of the entire initial molecule and is superimposable on the initial one, unless different positions are distinguished by substitution or isotopic labeling. No angular momentum is generated by this motion; this is the reason for the term. In the case of ring molecules, the conformational changes consist of puckering modes. The pseudorotation of tetrahydrofuran (THF) has been extensively investigated during the past decades, both in experimental and theoretical studies.[33,34,35,37,42,43]

Table 2. Vibrational Frequencies of the two THF Conformers (in cm^{-1}) Based on the Harmonic Oscillator Approximation Applied to the Molecular Mechanics Model for Linear and Cyclic Ethers by Vorobyov et al[36]

envelope	twisted
262	252
573	558
574	585
756	790
827	846
856	860
936	936
950	950
982	971
1052	1049
1067	1080
1120	1135
1166	1163
1193	1183
1287	1286
1321	1319
1334	1331
1350	1354
1451	1450
1465	1472
1472	1489
1487	1514
1580	1571
1595	1606
2851	2850
2851	2851
2885	2884
2888	2887
2897	2897
2903	2905
2920	2923
2931	2932

We performed an NVE molecular dynamics simulation of a single THF molecule at an average temperature of 300 Kelvin in the gas phase. The simulation time is 1 ns, and the integration time step is 1 fs. The simulation has been carried out with CP2K.[14] In the following paragraphs, we give step-by-step instructions for the analysis of the molecular dynamics simulation. As the analysis progresses, we unravel the relation between the harmonic frequencies of the two conformers, the vibrational density of states from the molecular dynamics simulation and the experimentally measured infrared spectrum.

The remainder of the text contains a transcript from the command line terminal and can be used by a potential user to reproduce this specific example. The output of the commands is hidden or reduced for reasons of clarity. All commands are preceded by a ">" sign and are printed in bold. The screen output generated by the MD-TRACKS commands is printed using a normal font weight. Long

commands are split over multiple lines because of the limited column width, but in practice, they should be entered without line breaks.

3.1.1. Setup of the MD-TRACKS Database. The output of the molecular dynamics simulation, as generated by CP2K (development version of June 27, 2008), is distributed over several files. The file `md-1.ener` contains the elementary energy terms as a function of time. The files `md-pos-1.xyz` and `md-vel-1.xyz` contain the atomic coordinates and velocities at each integration time step. The file `md-MM_DIPOLE-1.data` contains the time derivative of the dipole moment. The general output file md.out will not be used in this example. We load the relevant files in the binary MD-TRACKS database with the commands `tr-from-cp2k-ener`, `tr-from-xyz`, and `tr-from-txt`. If an MD-TRACKS database does not yet exist, it is automatically created. The conversion is done as follows. (The CP2K input files are not shown.)

```
> ls
md-1.ener
md-MM_DIPOLE-1.data
md-pos-1.xyz
md-vel-1.xyz
md.out
> tr-from-cp2k-ener md-1.ener
> tr-from-xyz md-pos-1.xyz pos
> tr-from-xyz md-vel-1.xyz vel -u au
> ls
...
tracks/
> ls tracks/
atom.pos.0000000.x
atom.pos.0000000.y
atom.pos.0000000.z
atom.pos.0000001.x
...
atom.pos.0000012.z
atom.vel.0000000.x
...
atom.vel.0000012.z
conserved_quantity
kinetic_energy
potential_energy
time
```

The command `tr-from-xyz` has two mandatory arguments: the XYZ trajectory filename and a suffix (e.g., pos or vel) that is used to generate the filenames in the binary tracks database. By default `tr-from-xyz` assumes that the text based XYZ file contains atom coordinates in angstroms. When the velocity file is converted, the option `-u au` is used to indicate that the XYZ file contains data in atomic units.

In the remainder of this example, we also need the time derivative of the dipole moment generated by the THF molecule. The conversion of the file `md-MM_DIPOLE-1.data` demonstrates how the generic `tr-from-txt` command converts ASCII data to the binary format when a specific `tr-from-*` command is not available. The records of interest in the file `md-MM_DIPOLE-1.data` have the following format (one line):

```
MM DIPOLE [NON-PERIODIC] DERIVATIVE
(A.U.)| 0.000226 −0.000289 −0.00070
```

The three numbers correspond to the *x*-, *y*-, and *z*-components of the time derivative of the dipole moment in atomic units. The file contains such a record for each time step. We can use the ubiquitous UNIX tools `grep` and `cut` to filter out the data of interest:

```
> grep DERIVATIVE md-MM_DIPOLE-1.data |
  cut -c 50-
0.000226 -0.000289 -0.000070
0.000226 -0.000289 -0.000070
0.000220 -0.000262 -0.000066
0.000186 -0.000186 -0.000060
0.000135 -0.000079 -0.000053
...
```

The `grep` command prints only the records from the file `md-MM_DIPOLE-1.data` that contain the word `DERIVATIVE`. The pipe symbol, `|`, prevents that the output of grep is printed on screen. Instead, the filtered records are redirected as input to the cut command, which discards the first 50 characters from each line. The command `tr-from-txt` reads text data formatted in columns from the standard input and writes this information into the binary database. We can use a second pipe symbol to redirect the output of the cut command to `tr-from-txt`.

```
> grep DERIVATIVE md-MM_DIPOLE-1.data |
  cut -c 50- | tr-from-txt
tracks/dipole.derivt.x
tracks/dipole.derivt.y
tracks/dipole.derivt.z
> ls tracks
...
dipole.derivt.x
dipole.derivt.y
dipole.derivt.z
...
```

So far, all given MD-TRACKS commands were specific for CP2K. From now on, the analysis is completely generic. One can replace the setup of the tracks database by specific commands for another molecular dynamics program and then continue with the instructions below to perform a similar analysis.

3.1.2. Standard Spectral Analysis. The infrared adsorption spectrum can be derived from a molecular dynamics simulation based on linear response theory. The classical approximation of the infrared adsorption spectrum is given by[44,45]

$$\alpha(\nu) \approx \lim_{\tau\to\infty} \frac{1}{\tau} \sum_{j=x,y,z} \left| \int_0^\tau \mathrm{d}t \, \exp(-i\nu t) \frac{\mathrm{d}\mu_j}{\mathrm{d}t} \right|^2 \qquad (1)$$

where ν is the frequency and μ_j are the Cartesian components of the dipole moment. This expression represents the power spectrum of the time derivative of the dipole moment. The command `tr-spectrum` is a generic tool to compute power spectra based on the numerical FFT algorithm.[46] When the time derivative of the dipole moment is given as input, the infrared spectrum is generated. When the atomic velocities are given as input, the inelastic neutron scattering (INS) spectrum or velocity power spectrum is computed. One could also compute the power spectrum of the time derivative of the polarizability, which leads to the Raman spectrum. The velocity power spectrum can be used as a classical approximation[47−49] of the vibrational density of states.[50]

Note that formula 1 is entirely equivalent to the Fourier transform of the autocorrelation function of the time-derivative of the dipole moment. The following MDTRACKS commands compute the infrared adsorption and the INS spectrum:

```
> tr-spectrum tracks/dipole.derivt.*
  tracks/time tracks/spectrum.ir
  --blocks=250
> ls tracks/
...
spectrum.ir.frequencies
spectrum.ir.wavenumbers
spectrum.ir.amplitudes
...
> tr-spectrum tracks/atom.vel.*
  tracks/time tracks/spectrum.vib
  --blocks=50
> ls tracks/
...
spectrum.vib.frequencies
spectrum.vib.wavenumbers
spectrum.vib.amplitudes
...
```

In the first command line, the three components of the time derivative of the dipole moment are given as input. The time axis is used to create a proper frequency and wavenumber axis. The output is written to files that start with `tracks/spectrum.ir`.

The option `--blocks=250` instructs the `tr-spectrum` command to divide the input data in 250 blocks of the same size. The spectrum is computed for each block and finally the average over all spectra is computed. A higher number of blocks improves the statistical accuracy of the final spectrum, but reduces the resolution on the wavenumber axis. For the infrared spectrum, the resolution on the wavenumber scale (X-axis) is 8 cm^{-1}, and the relative statistical error on the infrared adsorption (Y-axis) is 8%. The INS spectrum is obtained in a similar way. The latter is obtained with a resolution of 1.5 cm^{-1} and a relative statistical error of 9%. The INS spectrum is less sensitive to statistical noise because it is based on more input data: for each atom, there are three Cartesian velocity components, while the dipole moment has only three components in total.

The output files `tracks/spectrum.ir.frequencies` or `tracks/spectrum.ir.wavenumbers` can be used as the X-axis when plotting the spectrum. The corresponding amplitudes of the spectrum are stored in `tracks/spectrum.ir.amplitudes`.

In Figure 3, the INS spectrum is compared with the frequencies obtained within the harmonic oscillator approximation. A strict one-to-one correspondence between harmonic frequencies and peaks in the INS spectrum cannot be made. The main reason is that the molecular dynamics simulation takes into account finite temperature effects, while the harmonic oscillator approximation is only valid close to zero Kelvin. In the remainder of the text, we will show that at room temperature the THF molecule does not oscillate very long in one of the local minima of the potential energy surface. Instead THF continuously alters from the twisted to the envelope state and visits all intermediate structures. It is therefore impossible to assign peaks in the vibrational spectra to one of the conformers. It is clear however that

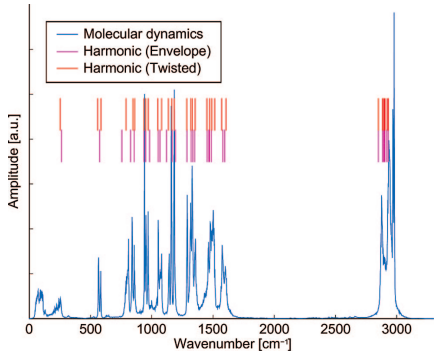


Figure 3. Vibrational frequencies in the THF molecule. The harmonic frequencies of the envelope (purple) and twisted (red) conformer are plotted as vertical lines. The vibrational density of states based on the molecular dynamics simulation is plotted in blue.

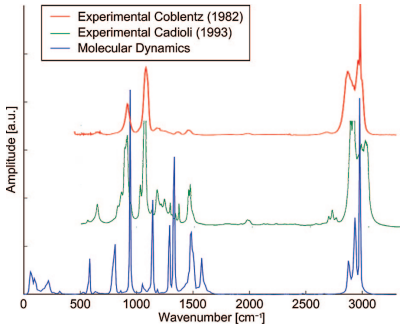


Figure 4. Infrared spectrum. The blue line represents the simulated spectrum. The green[58] and red[34] curves are experimental spectra from the literature.

clusters of harmonic frequencies correlate with clusters of peaks in the spectrum. The correlation fails visibly in two cases:

(1) In the region below 150 cm^{-1}, a band in the INS spectrum appears, which is completely absent in the harmonic oscillator approximation. This band represents a genuine vibration of the THF molecule and can not be attributed to coupling with rotational degrees of freedom. (The angular momentum is set to zero at the beginning of the molecular dynamics simulation and remains negligible.)

(2) The peaks around 3000 cm^{-1} are blue-shifted in the velocity power spectrum when compared to the harmonic frequencies. Figure 4 compares the simulated infrared spectrum with the experimental result. The molecular mechanics model approximates the experimental peak positions, but it fails to predict the infrared activity. In the remainder of this section, we will study the origin of adsorption band below 150 cm^{-1}.

3.1.3. Transitions between the THF Conformers. The distinct conformers of the THF molecule are well characterized by the ring puckering coordinates that can be computed

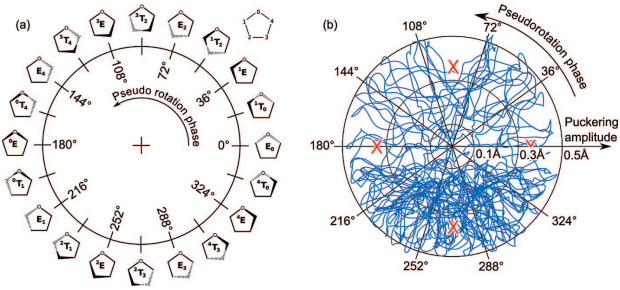


Figure 5. (a) Illustration of the pseudorotation phase, based on Figure 3 in the work of Altona and Sundaralingam.[59] (b) A polar plot of the puckering coordinates during the first 10 ps of the MD simulation. The radius is the puckering amplitude; the angle corresponds to the pseudorotation phase. The coordinates of the stable THF conformers are indicated with crosses.

with `tr-ic-puckering`. Our implementation relies on the general definition of puckering coordinates by Cremer and Pople.[51] In the case of a five-membered ring structure, there are two puckering coordinates: the puckering amplitude, q, which expresses the deviation from the planar ring structure and the pseudorotation phase, ϕ, which discriminates between all possible envelope and twisted geometries (see Figure 5a).

The time evolution of the ring puckering coordinates are computed with the following command:

```
> tr-ic-puckering 5
tracks/atom.pos.0000000
tracks/atom.pos.0000001
tracks/atom.pos.0000002
tracks/atom.pos.0000003
tracks/atom.pos.0000004 tracks/puck.pos
> ls tracks/
...
puck.pos.amplitude.0000002
puck.pos.phase.0000002
...
```

The first argument (`5`) stands for the number of atoms in the ring structure. The following five arguments are file prefixes that correspond to the five ring atoms (in consecutive order). The last argument is a filename prefix for the output files. The number of ring puckering coordinates is N-3 where N is the number of atoms in the ring. Conventionally, these coordinates are labeled with an integer index that starts from two.[51] For an eight-membered ring, one would obtain the following output files:

```
> ls tracks/
...
puck.pos.amplitude.0000002
puck.pos.amplitude.0000003
puck.pos.amplitude.0000004
puck.pos.phase.0000002
puck.pos.phase.0000003
...
```

During the molecular dynamics simulation, the THF molecule passes through all possible envelope and twisted conformers. This is demonstrated in Figure 5b where the time-dependent puckering coordinates during the first 10 ps are plotted as a solid line. The coordinates of the stable conformers are drawn as crosses. From this picture, it is clear the ring puckering motion can not be considered as a harmonic oscillation around a minimum on the potential energy surface. This suggests that the vibrational spectra from the MD simulation will exhibit features that are not present in the frequencies from the harmonic oscillator approximation. In other words, the band below 150 cm^{-1} is most likely related with the pseudorotation of the THF molecule. In the next part of this example application, this correspondence is unambiguously demonstrated.

3.1.4. Peak Assignment in Simulated Vibrational Spectra. For a proper understanding of the relation between the pseudorotation phase and the vibrational spectra, we should compute the contribution of the pseudorotation to the spectrum. A straightforward power spectrum of the time derivative of the pseudorotation phase can be misleading. The unit of the amplitude of this spectrum differs from the original INS spectrum, which disturbs a strict comparison. Alternatively, one can project the Cartesian velocity vector on the tangent of the internal coordinate of interest at each time step. The spectrum of these projected velocities only includes contributions of the motion along the selected internal coordinate and has the same unit as the original INS spectrum. The projected velocities are always smaller than or equal to the original velocities. The total spectrum will therefore be an (approximate) upper limit for the spectrum of the original velocities. When a peak in the *projected* spectrum coincides with its counterpart in the total spectrum, one can assume that there is no motion along other (orthogonal) coordinates that contributes to this peak. Such a peak is then completely resolved.

The following command will compute the ring puckering coordinates of the THF molecule, their time derivatives, and the projection of the Cartesian velocity vector on the tangent of each puckering coordinate:

```
> tr-ic-puckering 5
tracks/atom.pos.0000000
tracks/atom.pos.0000001
tracks/atom.pos.0000002
tracks/atom.pos.0000003
tracks/atom.pos.0000004
tracks/atom.vel.0000000
tracks/atom.vel.0000001
tracks/atom.vel.0000002
tracks/atom.vel.0000003
```

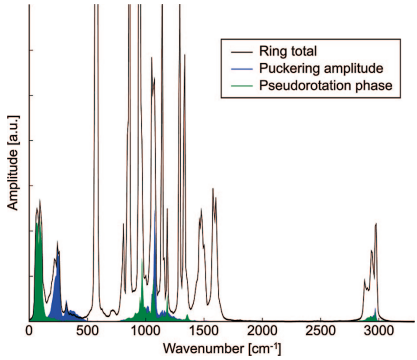


Figure 6. Analysis of the low-frequency bands in the velocity power spectrum of THF. The black line is the total velocity power spectrum of the five ring atoms. The blue curve represents the power spectrum of the velocity vector projected on the tangent of the puckering amplitude coordinate. The green curve is a similar spectrum that corresponds to the pseudorotation phase.

```
tracks/atom.vel.0000004
tracks/puck.pos tracks/puck.vel --project
> ls tracks/
...
atom.vel.0000000.x.proj.puck.vel.amplitude.0000002
atom.vel.0000000.x.proj.puck.vel.phase. 0000002
atom.vel.0000000.y.proj.puck.vel. amplitude.0000002
...
atom.vel.0000004.y.proj.puck.vel.amplitude.0000002
atom.vel.0000004.z.proj.puck.vel.amplitude.0000002
atom.vel.0000004.z.proj.puck.vel.phase.0000002
...
puck.pos.amplitude.0000002
puck.pos.phase.0000002
puck.vel.amplitude.0000002
puck.vel.phase.0000002...
```

Then we compute the power spectra of the projected velocities.

```
>tr-spectrum
tracks/atom.vel.*.proj.puck.vel.amplitude.0000002 tracks/time
tracks/spectrum.puck.amplitude
--blocks=250 > tr-spectrum
tracks/atom.vel.*.proj.puck.vel.phase.0000002
tracks/time tracks/spectrum.puck.phase --blocks=250
> ls tracks/
...spectrum.puck.amplitude.amplitudes
spectrum.puck.amplitude.wavenumbers
spectrum.puck.phase.amplitudes
spectrum.puck.phase.wavenumbers
...
```

Figure 6 gives an overview of the results. The black curve represents the total velocity power spectrum of the five ring atoms. The blue curve is the spectrum of the velocity vector projected on the tangent of the puckering amplitude coordinate, and the green curve corresponds to the pseudorotation phase. The contribution of the hydrogen atoms is not included in this figure because the ring puckering coordinates are not influenced by the hydrogen positions. The plot clearly reveals the origin of the lowest frequency band (from 20 to 150 cm^{-1}). It is entirely the result of the pseudorotation motion. The second-lowest band (from 150 to 280 cm^{-1}) is mainly caused by the puckering amplitude vibration. Apparently the projected velocities also correlate with bond stretch and bending angle modes at higher frequencies, mainly, because the tangents of the bond stretch, bending angle, and puckering coordinates are not orthogonal.

3.1.5. Final Remarks. This example stresses an important technical advantage of MD-TRACKS. The orthogonal design of the MD-TRACKS commands does not impose a predefined work flow during the analysis. In this case, the output of the command `tr-ic-puckering` is used as input for the command `tr-spectrum`. The puckering coordinates could have been used as an input for other commands, such as `tr-hist` or `tr-blav`. As demonstrated above, `tr-spectrum` also processes the output of many commands, for example, `tr-from-xyz`, `tr-from-txt`, `tr-ic-*`, etc. More advanced analysis tasks are carried out by a whole series of MD-TRACKS commands, grouped in a shell script. Each line in such a script processes the results from the previous commands. The final analysis results are converted to ASCII format with `tr-to-txt`, or they can be directly plotted with `tr-plot`.

3.2. Structural Properties of Liquid THF. In this section, we study radial distribution functions (RDF's) or pair distribution functions[52] of THF in the liquid phase. We consider both the center-of-mass RDF and the atom−atom RDF, which can be compared to neutron diffraction experiments.[53] The comparison between an experimental and a simulated RDF is a stringent test for the validity of the nonbonding interactions in the molecular dynamics simulation.

All radial distribution functions below are derived from an NVT molecular dynamics simulations of 64 THF molecules in a cubic box of 20.5 Å, using periodic boundary conditions. A Nosé−Hoover thermostat[54] with a relaxation time of 0.1 ps was applied to control the temperature of the system. The integration time step is 1 fs, and the total simulation time is 1 ns. The simulation has been carried out with CP2K.[14] The following subsections are organized in the same style as in the previous example.

3.2.1. Setup of the MD-TRACKS Database. This part is very similar to the previous example. We extract the time dependent atom positions with the following commands:

```
> ls
init.psf
md-1.ener
md-MM_DIPOLE-1.data
md-pos-1.xyz
md-vel-1.xyz
md.out
> tr-from-xyz md-pos-1.xyz pos
--slice=::1000
> ls tracks/
atom.pos.0000000.x
atom.pos.0000000.y
atom.pos.0000000.z
atom.pos.0000001.x
...
atom.pos.0000831.z
```

The option `--slice=::1000` instructs the script `tr-from-xyz` to read a frame from the trajectory file `md-pos-1.xyz` every 1000 time steps. The file `init.psf` is a CHARMM[18] topology file that will be used below to determine which atoms belong to the same molecule or to identify the chemical environment of an atom.

3.2.2. Center-of-Mass Radial Distribution Function. A center-of-mass RDF expresses the probability of finding two THF molecules at a certain distance apart in the liquid, relative to the probability of finding a pair of molecules that are homogeneously distributed at the same density. We will first derive the centers of mass of each molecule:

```
> tr-split-com tracks/atoms.pos pos
init.psf
> ls tracks/
...
com.pos.0000000.x
com.pos.0000000.y
...
com.pos.0000063.z
...
```

The arguments of `tr-split-com` are interpreted as follows: (i) the prefix for the track files that contain the atom coordinates, (ii) a tag for the output files, and (iii) the CHARMM topology file to identify the individual molecules. Consequently, we run the command that computes the radial distribution function based on these centers of mass:

```
> all_mol_prefixes=$(tr-select init.psf mol
  True --prefix=tracks/com.pos)
> echo $all_mol_prefixes
tracks/com.pos.0000000
tracks/com.pos.0000001...
> tr-rdf $all_mol_prefixes 20.5*A, 15*A 100
tracks/com.rdf
> ls
...
tracks/com.rdf.bins
tracks/com.rdf.hist
...
```

The first command `tr-select` lists molecules or atoms based on a filter expression. The first argument is the CHARMM topology file for the system under study. The second argument can be `at` or `mol` to indicate which objects one would like to select (atoms or molecules). The last argument is a filter expression that must evaluate to true for the atoms or molecules of interest. When `True` is literally given as filter expression, all atoms, or in this case molecules, are listed. The option `--prefix=tracks/com.pos` determines the format in which the list is printed. The construction `all_mol_prefixes=$(...)` assigns the output of the command `tr-select` to the variable `all_mol_prefixes`. The command `echo $all_mol_prefixes` prints the contents of this variable on screen. The command `tr-rdf` computes the actual radial distribution function for a list of time-dependent Cartesian coordinates. In this example, the centers of mass are used to compute the radial distribution function. The second argument specifies the box dimension so that the periodicity is properly taken into account when computing pair distances. The third argument is the maximum distance for which the RDF is computed. The fourth argument defines the number of bins in the histogram and the last argument is a prefix used for the output files. In this example, the file `tracks/com.rdf.hist` contains the *y*-values of the radial distribution function.

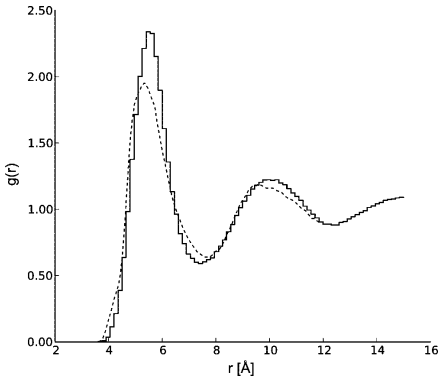


Figure 7. Center-of-mass radial distribution function of liquid THF at room temperature. The solid line represents the histogram derived from the molecular dynamics simulation. The dashed line is the experimentally observed radial distribution function by Bowron et al.[53]

The results are depicted in Figure 7, together with the experimental center-of-mass RDF by Bowron et al.[53] The overall correspondence is satisfactory, except for the maximum of the first peak, which is slightly overestimated in the simulated distribution. The area under the first peak of the simulated pair distribution reveals that (on average) there are 12.9 ± 0.5 molecules in the first shell that surrounds a given THF molecule. This compares very well to the experimental value of 12.6 ± 0.3.[53]

3.2.3. Atom−Atom Radial Distribution Functions. A more fine-grained picture of the relative position and orientation of THF molecules in the liquid phase is given by the atom−atom RDF, which expresses the probability of finding an atom of type A and B at a certain distance in the liquid, relative to the probability of finding these atoms at the same distance when they are homogeneously distributed. Analogously to the center-of-mass RDF, we first select the atoms for which we want to compute the RDF. In a second step, the actual RDF's are computed:

```
> O_prefixes=$(tr-select init.psf at
'a.symbol==″O″' --prefix=tracks/atom.pos)
> C1_prefixes=$(tr-select init.psf at
'a.symbol==″C″ and a.nsymbols=″O,C,H_2″
--prefix=tracks/atom.pos)
> C2_prefixes=$(tr-select init.psf at
'a.symbol==″C″ and a.nsymbols=″C_2,H_2″
--prefix=tracks/atom.pos)
> echo
tracks/atom.pos.0000001
tracks/atom.pos.0000004...
> tr-rdf $O_prefixes 25*A, 15*A 100
tracks/O.rdf
> tr-rdf $O_prefixes - $C1_prefixes 25*A,
15*A 100 tracks/OC1.rdf
> tr-rdf $O_prefixes - $C2_prefixes 25*A,
```

```
15*A 100 tracks/OC2.rdf
> ls tracks/
...
O.rdf.bins
O.rdf.hist
OC1.rdf.bins
OC1.rdf.hist
OC2.rdf.bins
OC2.rdf.hist
...
```

In the first three lines in the transcript above, three groups of atoms are defined with the command `tr-select`: the oxygen atoms (O), the carbon atoms that are directly bonded to an oxygen atom (group C1), and the carbons atoms that are not directly bonded to an oxygen atom (group C2). For a detailed description of the filter expressions, we refer to the documentation of `tr-select`, which can by consulted with the command `tr-select --help`. Consequently, the RDF's are computed with `tr-rdf`. The first example is based on distances between atoms in a single set, in this example, the set of oxygen atoms. The latter two RDF's consider the distances between the atoms in set A and B but not the distances within each set A or B. In this case, A is the set of oxygen atoms, and B is the set of C1 or C2 atoms.

The three radial distribution functions are plotted in Figure 8. The simulated distributions do not match perfectly with the experimental data. Mainly at short distances, the experimental RDF's show sharp peaks that are not present in their simulated counterparts. The experimental data suggest that our simulations underestimate the liguid structure in terms of relative orientation of neighboring THF molecules.

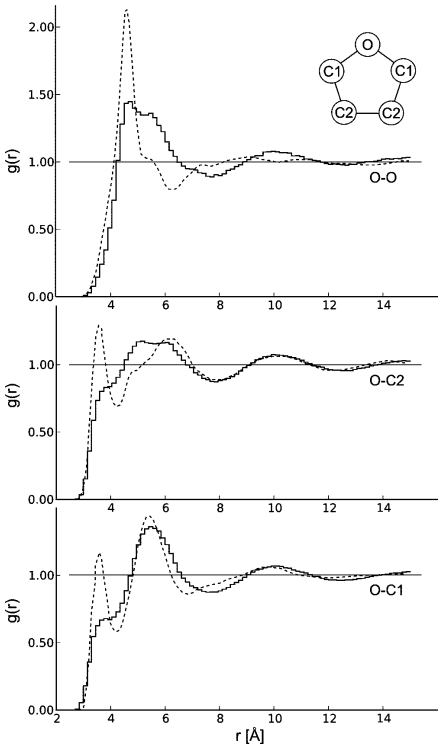


Figure 8. Atom−atom radial distribution functions for the pairs O−O, O−C1, and O−C2. C1 is the set of carbon atoms directly bonded to oxygen, and the C2 set contains to the remaining carbon atoms. The simulated RDF is plotted as a solid line, while the experimental curves are plotted as dashed lines.

4. PROGRAM AVAILABILITY

The MD-TRACKS toolkit is distributed as open source software under the conditions of the GNU General Public License, version 3. The software can be downloaded from the Code Web site of the Center for Molecular Modeling: http://molmod.ugent.be/code/. Documentation, installation instructions, and technical support are also available on this web site. In addition, there is a web-interface to the revision control systems that logs all changes in the source code. MD-TRACKS is released together with ZEOBUILDER,[55] which is a highly suitable GUI toolkit for the construction of initial molecular geometries for molecular dynamics simulations.

5. CONCLUSIONS

MD-TRACKS is a powerful and free molecular trajectory analysis toolkit. The MD-TRACKS toolkit consists of many commands that operate on an efficient and cross-platform binary trajectory database. There are three levels at which the MD-TRACKS toolkit can be used: one enters the individual MD-TRACKS commands at a UNIX command line shell, or one collects these commands in specialized shell scripts that automate the analysis work, or one creates new MD-TRACKS commands based on the MD-TRACKS programming library. Currently, the MD-TRACKS program has an interface CP2K,[14] CPMD[13] LAMMPS,[15,16] DL_POLY,[17] and Cerius2.[23]

The analysis of a molecular dynamics simulation of tetrahydrofuran with MD-TRACKS gives nontrivial insights in the vibrational and structural properties of this solvent. The transition of THF between the two (symmetric) twisted conformers is an anharmonic oscillation. It results in a broad band in the vibrational spectra between 20 and 150 cm^{-1}. The molecular mechanics model of Vorobyov et al.[36] leads in general to analysis results that correlate well with experimental observations. It should be clear that these examples merely cover a small part of the functionality of the MD-TRACKS toolkit. Even a full listing of the current MD-TRACKS commands does not reflect the continuous development of new features and improvements.

ACKNOWLEDGMENT

This work is supported by the Fund for Scientific Research, Flanders, and the Institute for the Promotion of Innovation through Science and Technology in Flanders (IWT Vlaanderen).

APPENDIX

The algorithms implemented in the commands below are discussed in standard text books on molecular dynamics and Monte Carlo simulations.[52,56]

- `tr-ac`: Computes the autocorrelation function.
- `tr-ac-error`: Computes the error on an autocorrelation function.
- `tr-angular-momentum`: Determines the angular momenta of one or more molecules.
- `tr-blav`: Applies the block-average method.
- `tr-calc`: Evaluates arbitrary mathematical functions on the data in track files
- `tr-corr`: Computes correlation coefficients
- `tr-cwt`: Computes the continuous wavelet transform.
- `tr-derive`: Numerically differentiates a function.
- `tr-fit-peaks`: Fits peaks in a spectrum.
- `tr-fluct`: Computes fluctuations.
- `tr-from-atrj`: Converts Cerius2 trajectory files (into the binary track files).
- `tr-from-cp2k-cell`: Converts CP2K unit cell data.
- `tr-from-cp2k-ener`: Converts CP2K energy files.
- `tr-from-cp2k-stress`: Converts CP2K stress tensor files.
- `tr-from-cpmd-ener`: Converts CPMD energy files.
- `tr-from-cpmd-traj`: Converts CPMD atom trajectory files.
- `tr-from-dlpoly hist`: Converts DL_POLY history files.
- `tr-from-dlpoly output`: Converts DL_POLY output files.
- `tr-from-lammps-dump`: Converts LAMMPS dump files.
- `tr-from-txt`: Reads data from a column-based plain text file and writes the data to binary track files.
- `tr-from-xyz`: Converts XYZ trajectory files.
- `tr-hist`: Computes histograms.
- `tr-ic-bend`: Computes a (time-dependent) bending angle.
- `tr-ic-dihed`: Computes a dihedral angle.
- `tr-ic-dist`: Computes an interatomic distance.
- `tr-ic-dtl`: Computes a distance to a line.
- `tr-ic-oop`: Computes an out-of-plane distance.
- `tr-ic-psf`: Enumerates and computes (a subset of) the internal coordinates based on the molecular topology.
- `tr-ic-puckering`: Computes the generalized puckering coordinates for an n-membered ring.
- `tr-integrate`: Numerically integrates a function.
- `tr-irfft`: Computes the inverse real Fourier transform.
- `tr-length`: Prints the length of a track file.
- `tr-mean-std`: Computes the (time-dependent) mean and the standard deviation.
- `tr-msd`: Derives the mean square displacement of a set of coordinates as a function of the time interval.
- `tr-msd-fit`: Derives the diffusion coefficient from the data obtained with tr-msd.
- `tr-norm`: Computes the time-dependent norm of a vector.
- `tr-pca`: Applies the principal component analysis method.
- `tr-plot`: Generates charts directly from data in the binary MD-TRACKS database.
- `tr-qh-entropy`: Computes the vibrational entropy, using the quasi-harmonic approximation.
- `tr-rdf`: Computes different types of radial distribution functions:
- `tr-reduce`: Reduces a data set with block averages.
- `tr-rfft`: Computes the forward real Fourier transform.
- `tr-select`: Prints atom or molecule indexes that fulfill a given filter expression.
- `tr-select-rings`: Prints atom indexes that belong to n-membered strong rings.[57]
- `tr-shortest-distance`: Computes the (time-dependent) shortest distance between two sets of atoms.
- `tr-slice`: Reduces a data set with subsampling.
- `tr-spectrum`: Computes various types vibrational spectra.
- `tr-split-com`: Computes the time dependent centers of mass of the molecules in the trajectory.
- `tr-to-txt`: Reads data from the binary database and convert it into plain text format.
- `tr-to-xyz`: Converts the trajectory data in the database to the XYZ format.

REFERENCES AND NOTES

(1) Rapaport, D.D.C. Introduction. In *The Art of Molecular Dynamics Simulationo*, 1st ed.; Cambridge University Press: Cambridge, U.K., 1997; pp 1−10.

(2) Tornroth-Horsefield, S.; Wang, Y.; Hedfalk, K.; Johanson, U.; Karlsson, M.; Tajkhorshid, E.; Neutze, R.; Kjellbom, P. Structural mechanism of plant aquaporin gating. *Nature.* **2006**, *439*, 688–694.

(3) Beckstein, O.; Tai, K.; Sansom, M. Not ions alone: Barriers to ion permeation in nanopores and channels. *J. Am. Chem. Soc.* **2004**, *126*, 14694–14695.

(4) Caleman, C.; van der Spoel, D. Picosecond melting of ice by an infrared laser pulse: A simulation study. *Angew. Chem. Int Ed.* **2008**, *47*, 1417–1420.

(5) Berendsen, H. J. C. *Simulating the Physical World: Hierarchical Modeling from Quantum Mechanics to Fluid Dynamics*, 1st ed.; Cambridge University Press: Cambridge, U.K., 2007; p 624.

(6) Allen, M. P.; Tildesley, D. J. Introduction. In *Computer Simulation of Liquids*; Oxford University Press: New York, 1989; pp 1−32.

(7) Car, R.; Parrinello, M. Unified approach for molecular dynamics and density-functional theory. *Phys. Rev. Lett.* **1985**, *55*, 2471.

(8) Vandevondele, J.; Krack, M.; Mohamed, F.; Parrinello, M.; Chassaing, T.; Hutter, J. Quickstep: Fast and accurate density functional calculations using a mixed Gaussian and plane waves approach. *Comput. Phys. Commun.* **2005**, *167*, 103–128.

(9) McGrath, M. J.; Siepmann, J. I.; Kuo, I. W.; Mundy, C. J.; VandeVondele, J.; Hutter, J.; Mohamed, F.; Krack, M. Isobaric−isothermal Monte Carlo simulations from first principles: Application to liquid water at ambient conditions. *Chem. Phys. Chem.* **2005**, *6*, 1894–1901.

(10) Kuhne, T. D.; Krack, M.; Mohamed, F. R.; Parrinello, M. Efficient and accurate Car−Parrinello-like approach to Born−Oppenheimer molecular dynamics. *Phys. Rev. Lett.* **2007**, *98*, 066401.

(11) Dal Peraro, M.; Ruggerone, P.; Raugei, S.; Gervasio, F. L.; Carloni, P. Investigating biological systems using first principles Car−Parrinello molecular dynamics simulations. *Curr. Opin. Struct. Biol.* **2007**, *17*, 149–156.

(12) Lin, H.; Truhlar, D. QM/MM: What have we learned, where are we, and where do we go from here? *Theor. Chem. Acc.* **2007**, *117*, 185–199.

(13) Parrinello, M.; Hutter, J.; Marx, D.; Focher, P.; Tuckerman, M.; Andreoni, W.; Curioni, A.; Fois, E.; Roetlisberger, U.; Gianozzi, P.; Deutsch, T.; Alavi, A.; Sebastiani, D.; Laio, A.; Vandevondele, J.; Seitsonen, A.; Billeter, S. *CPMD*, version 3.11.1; IBM Corp, MPI für Festkörperforschung Stuttgart: Stuttgart, Germany, 1990−2006, 1997−2007.

(14) CP2K. http://cp2k.berlios.de (accessed Mar 24, 2008).

(15) Plimpton, S. Fast parallel algorithms for short-range molecular dynamics. *J. Comput. Phys.* **1995**, *117*, 1–19.

(16) LAMMPS. http://lammps.sandia.gov/cite.html (accessed Jul 9, 2008).

(17) Smith, W.; Forester, T. R. DL_POLY_2.0: A general-purpose parallel molecular dynamics simulation package. *J. Mol. Graph.* **1996**, *14*, 136–141.

(18) Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem.* **1983**, *4*, 187–217.

(19) Phillips, J. C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R. D.; Kale, L.; Schulten, K. Scalable

molecular dynamics with NAMD. *J. Comput. Chem.* **2005**, *26*, 1781–1802.
(20) Gale, J. GULP—A computer program for the symmetry adapted simulation of solids. *J. Chem. Soc., Faraday Trans.* **1997**, *93*, 629.
(21) Lindahl, E.; Hess, B.; van der Spoel, D. GROMACS 3.0: A package for molecular simulation and trajectory analysis. *J. Mol. Model.* **2001**, *7*, 306–317.
(22) Christen, M.; Hünenberger, P. H.; Bakowies, D.; Baron, R.; Bürgi, R.; Geerke, D. P.; Heinz, T. N.; Kastenholz, M. A.; Kräutler, V.; Oostenbrink, C.; Peter, C.; Trzesniak, D.; Gunsteren, W. F. V. The GROMOS software for biomolecular simulation: GROMOS05. *J. Comput. Chem.* **2005**, *26*, 1719–1751.
(23) *Cerius2 Modelling Environment*; Accelrys Software Inc.: San Diego, CA, 2006.
(24) Humphrey, W.; Dalke, A.; Schulten, K. VMD: Visual molecular dynamics. *J. Mol. Graph.* **1996**, *14*, 33–38.
(25) Case, D.; Darden, T.; T. E. Cheatham, I. I. I.; Simmerling, C.; Wang, J.; Duke, R.; Luo, R.; Merz, K.; Pearlman, M.; Crowley, M.; Walker, R.; Zhang, W.; Wang, B.; Hayik, S.; Roitberg, A.; Seabra, G.; Wong, K.; Paesani, F.; Wu, X.; Brozell, S.; Tsui, V.; Gohlke, H.; Yang, L.; Tan, C.; Mongan, J.; Hornak, V.; Cui, G.; Beroza, P.; Mathews, D.; Schafmeister, C.; Ross, W.; Kollman, P.; *AMBER 9*; University of California: San Francisco, CA, 2006.
(26) Sirius. http://sirius.sdsc.edu/ (accessed Jul 10, 2008).
(27) Krüger, P.; Szameit, A. SIMLYS version 2.0. *Comput. Phys. Commun.* **1992**, *72*, 265–268.
(28) Pauwels, E.; Verstraelen, T.; Waroquier, M. Effect of temperature on the EPR properties of a rhamnose alkoxy radical: A DFT molecular dynamics study. *Spectrochim. Acta, Part A* **2008**, *69*, 1388–1394.
(29) Pauwels, E.; Verstraelen, T.; De Cooman, H.; Van Speybroeck, V.; Waroquier, M. Temperature study of a glycine radical in the solid state adopting a DFT periodic approach: Vibrational analysis and comparison with EPR experiments. *J. Phys. Chem. B.* **2008**, *112*, 7618–7630.
(30) Lesthaeghe, D.; Vansteenkiste, P.; Verstraelen, T.; Ghysels, A.; Kirschhock, C. E. A.; Martens, J. A.; Speybroeck, V. V.; Waroquier, M. MFI fingerprint: How pentasil-induced IR bands shift during zeolite nanogrowth. *J. Phys. Chem. C.* **2008**, *112*, 9186–9191.
(31) Oliphant, T. Numpy. http://www.numpy.org/ (accessed Jun 15, 2008).
(32) Mohr, P. J.; Taylor, B. N. CODATA recommended values of the fundamental physical constants: 1998. *Rev. Mod. Phys.* **2000**, *72*, 351.
(33) Legon, A. C. Equilibrium conformations of four- and five-membered cyclic molecules in the gas phase: determination and classification. *Chem. Rev.* **1980**, *80*, 231–262.
(34) Cadioli, B.; Gallinella, E.; Coulombeau, C.; Jobic, H.; Berthier, G. Geometric structure and vibrational spectrum of tetrahydrofuran. *J. Phys. Chem.* **1993**, *97*, 7844–7856.
(35) Duffy, P.; Sordo, J. A.; Wang, F. Valence orbital response to pseudorotation of tetrahydrofuran: A snapshot using dual space analysis. *J. Chem. Phys.* **2008**, *128*, 125102.
(36) Vorobyov, I.; Anisimov, V.; Greene, S.; Venable, R.; Moser, A.; Pastor, R.; Mackerell, A. Additive and classical drude polarizable force fields for linear and cyclic ethers. *J. Chem. Theory Comput.* **2007**, *3*, 1120–1133.
(37) Rayon, V. M.; Sordo, J. A. Pseudorotation motion in tetrahydrofuran: An ab initio study. *J. Chem. Phys.* **2005**, *122*, 204303.
(38) Harris, D. O.; Engerholm, G. G.; Tolman, C. A.; Luntz, A. C.; Keller, R. A.; Kim, H.; Gwinn, W. D. Ring Puckering in five-membered rings. I. General theory. *J. Chem. Phys.* **1969**, *50*, 2438–2445.
(39) Vansteenkiste, P.; VanSpeybroeck, V.; Verniest, G.; DeKimpe, N.; Waroquier, M. Applicability of the hindered rotor scheme to the puckering mode in four-membered rings. *J. Phys. Chem. A.* **2006**, *110*, 3838–3844.
(40) Vansteenkiste, P.; VanSpeybroeck, V.; Verniest, G.; DeKimpe, N.; Waroquier, M. Four-membered heterocycles with a carbon—heteroatom exocyclic double bond at the 3-position: puckering potential and thermodynamic properties. *J. Phys. Chem. A.* **2007**, *111*, 2797–2803.
(41) Kilpatrick, J. E.; Pitzer, K. S.; Spitzer, R. The thermodynamics and molecular structure of cyclopentane. *J. Am. Chem. Soc.* **1947**, *69*, 2483–2488.
(42) Coulombeau, C.; Jobic, H. Contribution to the vibrational normal modes analysis of d8-THF by neutron inelastic scattering. *THEOCHEM.* **1995**, *330*, 127–130.
(43) Yang, T.; Su, G.; Ning, C.; Deng, J.; Wang, F.; Zhang, S.; Ren, X.; Huang, Y. New diagnostic of the most populated conformer of tetrahydrofuran in the gas phase. *J. Phys. Chem. A.* **2007**, *111*, 4927–4933.
(44) Berens, P.; Wilson, K. Molecular dynamics and spectra. I. Diatomic rotation and vibration. *J. Chem. Phys.* **1981**, *74*, 4872–4882.
(45) McQuarrie, D. A. The time-correlation function formalism, I. In *Statistical Mechanics*; Rice, S., Ed.; Harper & Row: New York, 1976; pp 467—542.
(46) Cooley, J. W.; Tukey, J. W. An algorithm for the machine calculation of complex fourier series. *Math. Comput.* **1965**, *19*, 297–301.
(47) Dickey, J. M.; Paskin, A. Computer simulation of the lattice dynamics of solids. *Phys. Rev.* **1969**, *188*, 1407.
(48) Loong, C.; Vashishta, P.; Kalia, R. K.; Jin, W.; Degani, M. H.; Hinks, D. G.; Price, D. L.; Jorgensen, J. D.; Dabrowski, B.; Mitchell, A. W.; Richards, D. R.; Zheng, Y. Phonon density of states and oxygen-isotope effect in Ba1-*x*K*x*BiO_3. *Phys. Rev. B.* **1992**, *45*, 8052.
(49) Rahman, A.; Mandell, M. J.; McTague, J. P. Molecular dynamics study of an amorphous Lennard-Jones system at low temperature. *J. Chem. Phys.* **1976**, *64*, 1564–1568.
(50) McQuarrie, D. A. Crystals. In *Statistical Mechanics*; Rice, S., Ed.; Harper & Row: New York, 1976; pp 194—221.
(51) Cremer, D.; Pople, J. A. General definition of ring puckering coordinates. *J. Am. Chem. Soc.* **1975**, *97*, 1354–1358.
(52) Allen, M. P.; Tildesley, D. J. Statistical Mechanics. In *Computer Simulation of Liquids*; Oxford University Press: New York, 1989; pp 33—70.
(53) Bowron, D.; Finney, J.; Soper, A. The structure of liquid tetrahydrofuran. *J. Am. Chem. Soc.* **2006**, *128*, 5119–5126.
(54) Nose, S. A unified formulation of the constant temperature molecular dynamics methods. *J. Chem. Phys.* **1984**, *81*, 511–519.
(55) Verstraelen, T.; Van Speybroeck, V.; Waroquier, M. ZEOBUILDER: A GUI toolkit for the construction of complex molecular structures on the nanoscale with building blocks. *J. Chem. Inf. Model.* **2008**, *48*, 1530–1541.
(56) Frenkel, D.; Smit, B. "Monte Carlo Simulations" and "Molecular Dynamics Simulations". In *Understanding Molecular Simulation*, 2nd ed.; Academic Press: San Diego, CA, 2002; pp 23—108.
(57) O'Keeffe, M.; Hyde, B. Nets and infinite polyhedra. In *Crystal Structures I. Patterns and Symmetry*; Mineralogical Society of America: Washington, DC, 1996; pp 289—375.
(58) Smith, A. *The Coblentz Society Desk Book of Infrared Spectra*, 2nd ed.; Carver, C., Ed.; The Coblentz Society: Kirkwood, MO, 1982; pp 1—24.
(59) Altona, C.; Sundaralingam, M. Conformational analysis of the sugar ring in nucleosides and nucleotides. New description using the concept of pseudorotation. *J. Am. Chem. Soc.* **1972**, *94*, 8205–8212.

CI800233Y

Part 2: Model Development

Paper 3: "The Gradient Curves Method: An Improved Strategy for the Derivation of Molecular Mechanics Valence Force Fields from ab Initio Data"

Toon Verstraelen, Dimitri Van Neck, Paul W. Ayers, Veronique Van Speybroeck, Michel Waroquier

Journal of Chemical Theory and Computation , **2007**, 3, 1420 - 1434

1420 *J. Chem. Theory Comput.* **2007,** *3,* 1420−1434

JCTC Journal of Chemical Theory and Computation

The Gradient Curves Method: An Improved Strategy for the Derivation of Molecular Mechanics Valence Force Fields from ab Initio Data

T. Verstraelen,*,† D. Van Neck,† P. W. Ayers,‡ V. Van Speybroeck,† and M. Waroquier†

Center for Molecular Modeling, Ghent University, 9000 Gent, Belgium, and Department of Chemistry, McMaster University, Hamilton, ON L8S 4M1, Canada

Received June 23, 2006

Abstract: A novel force-field development strategy is proposed that tackles the well-known difficulty of parameter correlations arising in a conventional least-squares optimization. In the first step of the new gradient curves method (GCM), continuity criteria are imposed to transform the raw multidimensional ab initio training data to distinct sets of one-dimensional data, each associated with an individual energy term. In the second step, the transformed data suggest suitable analytical expressions, and the parameters in these expressions are fitted to the transformed data; that is, one does not have to postulate a priori analytical expressions for the force-field energy terms. This approach facilitates the derivation of valence terms. Benchmarks have been performed on a set of small molecules. The results show that the new method yields physically acceptable energy terms exactly when a conventional parametrization would suffer from parameter correlations, that is, when an increasing number of redundant internal coordinates is used in the force-field model. The generic treatment of parameter correlations in the proposed method facilitates an intuitive physical interpretation of the individual terms in the force-field expression, which is a prerequisite for the transferability of force-field models.

1. Introduction

The development of a molecular mechanics force field based on an ab initio parametrization is a tedious task plagued by model selection and parameter correlations, especially when one wants to extend its applicability to a broad range of molecular systems. The final goal of this study lies in the construction of an accurate all-atom zeolite-guest force field that is applicable both to unconstrained bulk zeolite structures and to unconstrained interfaces between zeolite nanoparticles and their environment. It is highly ambitious to assert that such a broad domain of applications can be covered by a single force-field model. Most of the force fields proposed in the literature can be used only for a subset of the applications mentioned above.[1−7] There are two reasons for the limited applicability of existing force fields. On the one hand, molecular mechanics models are limited, in general, to a specific domain of application due to the reduction of the full ab initio description of a molecule into a set of parametrized analytical energy terms. This failure is inherent to the nature of force-field models. On the other hand, the determination of reliable and transferable parameters for the analytical expressions in a force field is a nontrivial task. The main focus of this paper is the development of a reliable parametrization technique.

Parameter correlations, which are inherent to least-squares parametrization in general, represent the major difficulty in the development of force fields based on ab initio data. In the naive approach of an accurate force-field model, a large number of parameters should be introduced to describe all possible types of interactions. The optimization then usually leads to many degenerate solutions; that is, many disparate parameter sets have nearly the same goodness of fit. Only a

* Corresponding author e-mail: Toon.Verstraelen@UGent.be.
† Ghent University.
‡ McMaster University.

10.1021/ct6002093 CCC: $37.00
Published on Web 05/01/2007

very small number of these "good" fits are physically acceptable and transferable to molecules not belonging to the training set. In an attempt to fulfill these requirements, several techniques have been proposed in literature that select a physically meaningful and potentially transferable set of parameters yielding an acceptable goodness of fit.

(i) An intuitive procedure first parametrizes a coarse force field that only contains the most important energy terms, using a traditional least-squares method. Second, the residual error is further reduced by including corrective energy terms whose parameters are optimized without modifying the original coarse force field.[10] This approach keeps the contribution of the corrective energy terms small compared to the coarse force field, but the optimal goodness of fit is not reached. Moreover, only the correlations between parameters in the coarse force field and the corrective energy terms are treated. More generally, a global optimization is divided in smaller piecewise optimizations, to make the parametrization more tractable. A piecewise optimization only considers a subset of variable parameters, for example, the parameters associated with all bond-stretch terms. The global optimum is then approximated with a limited number of iterations in which each subset of parameters is optimized or reoptimized as to give an optimal fit with respect to the training data and the other subsets of parameters that are kept fixed.[8,9]

(ii) Another procedure avoids degeneracies in the first-order energy terms (i.e., correlations between first-order force constants and reference coordinates) by imposing constraints on their coefficients,[2] but degeneracies in higher-order terms are neglected.

(iii) The most systematic approach adds quadratic penalty functions to the χ^2 cost function.[11] With each parameter, a penalty function is associated that restrains this parameter to a physically acceptable value. This regularization technique is similar to restrained electrostatic potential fitting.[12] Unfortunately, one must choose the weight for each penalty function to be small enough so that the penalty functions only make small contributions to the total cost function but large enough so that the parameters are forced to retain a physically reasonable value. This "weight determination problem" is also ill-conditioned. Essentially, one has just replaced one ill-conditioned problem (parameter fitting) with another one (weight determination).

(iv) Parameter correlations can also be avoided by reducing the number of parameters in a force-field model.[1] One has to select carefully the energy terms that can be omitted and the analytical form of the retained terms. The disadvantage of this approach is that the absence of some molecular interaction terms in the force field will be compensated by biased parameters in the retained terms. Consequently, it is a common practice to exclude the atomic charges from the optimization procedure and to assign formal charges to these atoms instead; this prevents unphysical atomic charges.

(v) The most extreme approach in this comparative overview is represented by the rule-based force fields that do not contain fitted parameters.[13−15] All parameters are directly derived from semiempirical rules or are estimated on the basis of common sense. Such force fields sacrifice accuracy to achieve transferability.

Except for the second method, all the techniques mentioned above require additional subjective choices to tackle the problem of parameter correlations: the separation of coarse- and fine-grained components, a vast amount of weight factors, and so forth. Only the third method is truly systematic since it treats all parameter correlations, but it depends on a series of manually tuned weight factors.

This work aims to present a new force-field parametrization procedure—the gradient curves method (GCM)—which is innovative in its concept and which addresses the main concerns raised above. First, the method does not rely on subjective choices, for example, predefined analytical expressions for the energy terms, manually tuned parameters, repetitive parametrizations where at each iteration some parameters are included or excluded, and so forth. Second, the new method treats the problem of parameter correlations in a rigorous way. The only input is a set of ab initio training data and a list of the internal coordinates that will be used in the force-field model.

The gradient curves method is designed to extract the maximum amount of information from the ab initio training data set. A two-step procedure is used to achieve this objective. The first step encompasses a transformation of the raw multidimensional ab initio data into distinct one-dimensional data sets, each associated with a single energy term. During this transformation, a consistent treatment of parameter correlations guarantees a unique and physically acceptable series of transformed data sets. In this context, "physically acceptable" indicates that it is possible to give an intuitive physical interpretation to the individual transformed data sets. The analytical expressions enter the procedure only in the second step, where they can be easily estimated from the transformed data sets and may be modeled with nonlinear parameters without major difficulties.

For several reasons, the present version of the gradient curves method is less appropriate to parametrize long-range interactions. These interactions (i.e., the classical electrostatic and the dispersion interactions) obey well-known physical laws. Therefore, it would be highly inefficient to derive these long-range interactions without relying on their asymptotic behavior during the first step of the new method. Specific parametrization techniques for chemically accurate electrostatic models have already been actively studied during the past decades.[12,16−18] Due to the enormous computational cost of post-Hartree−Fock ab initio calculations that describe dispersion interactions properly,[19] it is more efficient to use such calculations specifically for the parametrization of dispersion interactions.[20,21]

Most of the ingredients of the gradient curves method are new, but the idea to express a multivariate function in terms of functions depending on a smaller number of variables is frequently applied. We refer to the high dimensional model representation[25] (HDMR) which has been applied in several fields, ranging from molecular modeling[26] to global atmospheric models.[27] This technique guarantees a unique multivariate expansion; that is, it treats parameter correlations, by imposing orthogonality constraints between all the

components in the expansion. HDMR is very efficient when the primary concern is only to reproduce a given set of training data. The end result is an efficient and reliable input−output model. At this point, our focus is different; that is, we would like to ensure that all the distinct energy terms are physically intuitive instead of orthogonal. Less popular black-box approaches where the expansion consists solely of one-dimensional functions[28,29] are based on Kolmogorov's solution[30] to Hilbert's 13th problem[31] and rely on nonsmooth component functions.

The applications in this paper are limited to a set of small molecules such as H_2O, NH_3, and CH_4. For the short-range aspects of interest, this is sufficient to illustrate and benchmark the new method. The aim of these examples is not to obtain transferable force-field parameters for these three molecules but rather to show how the prerequisites for transferability can be met. Additionally, it is not the intention to derive definitive force-field parameters for these three molecules that can be directly tested against experimental data, but we focus on the aspect of how well a reasonable force-field model can simulate a given set of ab initio calculations. We have intentionally generated ab initio training data for these molecules that include a significant portion of the anharmonic part of the potential energy surface, in order to test to what extent the gradient curves method is capable of parametrizing force fields that also reproduce the nonharmonic part of the potential energy surface of the three benchmark molecules. Work is in progress to extend the applicability of the gradient curves method to larger systems, taking into account long-range interactions.

The remainder of this article is organized as follows. In section 2, the new procedure is derived. The benchmark protocol that evaluates the merits of this new procedure is presented in section 3. Section 4 discusses the results obtained by the benchmarks. Finally, conclusions are given in section 5.

2. Gradient Curves Method

2.1. Outline. For the sake of simplicity, we limit ourselves to force fields of the class-I form:

$$E_{\mathrm{FF}} = \sum_{k=1}^{K} E_k(q_k) \tag{1}$$

where $K > 3N - 6$ and $N =$ the number of atoms. The force-field energy E_{FF} of a molecular geometry is expressed as a sum over functions E_k of only one internal coordinate q_k, where the q_k's are not restricted to the $(3N - 6)$ molecular degrees of freedom and may stand for a redundant set of internal coordinates. The redundancy originates from the observation that even a coarse valence force field[13,14] includes terms for all bond lengths, all bending angles, and some dihedral angles. Force fields that are accurate in the prediction of both structural and vibrational properties have to include cross terms $E_{k1,k2}(q_{k1},q_{k2})$ in the force-field expression.[22] In class-II force fields,[23] this is resolved by adding functions that depend on products of internal coordinates, that is, $q_{k1}q_{k2}$. We prefer to label products and other constructions of internal coordinates as new internal coordinates, which allows us to work with the class-I form in eq 1. This implies that for accurate force fields $K \gg 3N - 6$.

As a consequence of the redundancy, a direct fit of parametrized expressions for the E_k to a set of ab initio training data contains severe parameter correlations even when an abundant amount of training data is available. By selecting one arbitrary set of parameters that minimizes the residual errors, the resulting force field contains energy terms with an unphysical behavior and consequently lacks transferability.[2,11] Similar considerations about redundant internal coordinates in the theory of molecular vibrations have led to the canonical force-field concept, which is useful for the analysis of vibrational spectra.[24]

The detailed mathematical derivation of the gradient curves method will be presented in the next subsection. We now continue with a general outline of the method. The training data used in the gradient curves method are the ab initio calculated gradients for M different geometries of a given molecule

$$Y_i^{(m)} = \left(\frac{\partial E_{\mathrm{AI}}}{\partial x_i}\right)_{x=x^{(m)}} \tag{2}$$

where $m = 1...M$ and $x^{(m)}$ is the vector that contains all the Cartesian coordinates of the atoms in geometry m. For an energy surface of the class-I form in eq 1, one factorizes the Cartesian gradient for geometry m according to

$$G^{(m)} = J^{(m)} g^{(m)} \tag{3}$$

where the matrices in expression 3 are defined as

$$\begin{aligned} G_i^{(m)} &= \left(\frac{\partial E_{\mathrm{FF}}}{\partial x_i}\right)_{x=x^{(m)}} \\ g_k^{(m)} &= \left(\frac{\partial E_{\mathrm{FF}}}{\partial q_k}\right)_{q=q^{(m)}} = \left(\frac{dE_k}{dq_k}\right)_{q_k=q_k^{(m)}} \\ J_{i,k}^{(m)} &= \left(\frac{\partial q_k}{\partial x_i}\right)_{x=x^{(m)}} \end{aligned} \tag{4}$$

The convention for matrix notation in this article uses upper indexes to indicate different matrices and lower indexes to identify the matrix elements; for example, $G^{(m)}$ and $g^{(m)}$ are column matrices of dimension $3N$ and K, respectively, whereas $J^{(m)}$ is a rectangular matrix of dimensions $3N \times K$.

Since we want to find a suitable class-I representation of the true (ab initio) energy surface E_{AI} sampled in M geometries, we first identify the Cartesian gradient of the force-field energy in expression 1 with the ab initio training data

$$G_i^{(m)} \equiv Y_i^{(m)} \tag{5}$$

and try to solve the linear system

$$Y^{(m)} = J^{(m)} y^{(m)} \tag{6}$$

for the "ab initio gradient in internal coordinates", $y^{(m)}$. Due to the redundancy of the coordinates q_k, this equation has many solutions, that is, a particular solution plus an arbitrary

vector from the null space of $J^{(m)}$. Step I of the gradient curves method determines which vector from the null space must be taken for each geometry by an optimization procedure. In other words, the first step defines how the ab initio training data are transformed into one-dimensional data sets of the form $D_k = \{(q_k^{(m)}, y_{k(\mathrm{opt})}^{(m)}) | m = 1...M\}$. Through the identification $y_k^{(m)} \equiv g_k^{(m)}$, or $y_k^{(m)} \equiv (dE_k/dq_k)_{q_k=q_k^{(m)}}$, step II of the gradient curves method consists of proposing a functional form for the derivative of each energy term, (dE_k/dq_k), based on its corresponding transformed data set, D_k, and the expected asymptotic behavior. Finally, each functional form can be fitted to its corresponding data set with conventional fitting procedures.

The purpose of the transformation in step I of the gradient curves method is to make step II as successful as possible. This means that—for each geometry—the vector from the null space will be taken so as to optimize the continuity conditions of the data sets D_k. In practice, this is achieved by selecting the solutions of eq 6 for all geometries that minimize a cost function, Z, which is a measure for the continuity of the data sets D_k. In this work, continuity is measured by the goodness of fit of a generic high-order polynomial to a set of data points.

Unfortunately, this continuity requirement alone will in general not result in a uniquely defined transformation. In other words, the cost function, Z, as a function of the solutions of eq 6, can have a degenerate minimum. It will be shown in the next subsection that the transformation will always be ill-defined when the number of energy terms, K, is much larger than the number of independent internal degrees of freedom, $3N - 6$. To guarantee a unique minimum, we must introduce additional but subordinate criteria that will select from all the possible transformations to continuous data sets the one solution that corresponds optimally to what we expect from physical intuition. In this work "physical intuition" is interpreted as "having minimal forces along the internal coordinates". This prescription can be implemented as a least-norm criterion on the y values of the data sets D_k, in addition to the continuity criterion. Formally, such a least-norm criterion is implemented as an extra term in the cost function $Z^* = Z + \epsilon L$, where ϵ is a very small positive number and L is the contribution from the least-norm criterion. For small values of ϵ, the minimum of the new cost function approximately also minimizes the original cost function. This least-norm criterion is also known as zeroth-order regularization, and—as shown in the next subsection—it ensures that the transformation is always uniquely defined.

In order to understand the remainder of this paper, it is not strictly required to read the next subsection which describes the detailed mathematical derivation of the gradient curves method. Nevertheless, it is highly recommended for a deeper understanding, and mandatory when one is interested in implementing or extending the method.

2.2. Detailed Procedure. Since step II is a standard fitting procedure, we now concentrate on the details of step I. The general solution of the linear system (6) is given by

$$y^{(m)} = p^{(m)} + \mathcal{N}^{(m)} s^{(m)} \tag{7}$$

where $p^{(m)}$ is a particular solution, $\mathcal{N}^{(m)}$ is a matrix with orthogonal columns spanning the null space of the Jacobian $J^{(m)}$, and the vector $s^{(m)}$ contains arbitrary coefficients that determine which vector from the null space is added to the particular solution. One can derive the particular solution and the null space of a given linear system through the singular value decomposition algorithm.[32]

The coefficients $s^{(m)}$ are fixed by imposing continuity criteria: we select the $s^{(m)}$'s that minimize the sum of squared residual errors, obtained in a linear fit of a set of generic auxiliary functions, $f_n(q_k)$ (e.g., polynomials), to the "ab initio gradient in internal coordinates", $y_k^{(m)}$,

$$y_k^{(m)} \overset{\text{fit}}{\equiv} \sum_n a_n^{(k)} f_n(q_k^{(m)}) \tag{8}$$

The sum of the squared residual errors in the fit to the data set D_k is given by the expression

$$R_k^2 = \sum_m (\sum_n a_n^{(k)} f_n(q_k^{(m)}) - y_k^{(m)})^2 \tag{9}$$

In this equation, and in the following analysis, we find it convenient to switch to a notation where the different matrix quantities are labeled by the index of the internal coordinates under scrutiny, k, for example,

$$F_{m,n}^{(k)} = f_n(q_k^{(m)}) \qquad \tilde{y}_m^{(k)} = y_k^{(m)} \tag{10}$$

In the revised notation, the sum of squared residuals (using standard manipulations) is

$$R_k^2 = (F^{(k)}a^{(k)} - \tilde{y}^{(k)})^T (F^{(k)}a^{(k)} - \tilde{y}^{(k)}) \tag{11}$$

Minimizing this expression with respect to the expansion coefficients, $a_n^{(k)}$, allows one to discern how well the gradient information can be represented by a continuous function. The least-squares expansion coefficients from eq 8 are given by the expression

$$a_{(\mathrm{opt})}^{(k)} = [F^{(k)T}F^{(k)}]^{-1}F^{(k)T}\tilde{y}^{(k)} \tag{12}$$

and the residual error is

$$\min_{a_n^{(k)}} R_k^2 = \tilde{y}^{(k)T}[1 - F^{(k)}(F^{(k)T}F^{(k)})^{-1}F^{(k)T}]\tilde{y}^{(k)} = \tilde{y}^{(k)T}C^{(k)}\tilde{y}^{(k)} \tag{13}$$

which is indicative of the continuity of the data set D_k. Note that $C^{(k)}$ projects on the complement of the range of $F^{(k)}$. In analogy to eq 10, we can introduce relabeled matrix quantities

$$\tilde{p}_m^{(k)} = p_k^{(m)} \qquad \tilde{\mathcal{N}}_{m',\beta m}^{(k)} = \mathcal{N}_{k,\beta}^{(m)}\delta_{m',m} \qquad \tilde{s}_{\beta m} = s_\beta^{(m)} \tag{14}$$

in terms of which eq 7 can be rewritten as

$$\tilde{y}^{(k)} = \tilde{p}^{(k)} + \tilde{\mathcal{N}}^{(k)}\tilde{s} \tag{15}$$

This allows a compact expression for the desired cost function, which is a weighted sum of the continuity measures

of all the data sets D_k

$$Z(\tilde{s}) = \sum_k w_k^2 (\min_{a_n^{(k)}} R_k^2) = \sum_k (\tilde{p}^{(k)} + \tilde{\mathcal{N}}^{(k)}\tilde{s})^T w_k^2 C^{(k)} (\tilde{p}^{(k)} + \tilde{\mathcal{N}}^{(k)}\tilde{s}) \quad (16)$$

For the practical applicability of the gradient curves method, the weight factors w_k^2 which convert the R_k^2's to the dimension of an energy squared should be easy to obtain. A simple physical interpretation of $w_k R_k$ is illustrated in Figure 1. R_k is the RMS error obtained by fitting the auxiliary functions $f_n(q_k)$ to the optimized data set D_k. One obtains a tentative energy term by integrating the fitted function $\sum_n a_n^{(k)} f_n(q_k)$ over the physically relevant interval $[q_k^{(\min)}, q_k^{(\max)}]$. The error accumulated during this integration is equal to $(q_k^{(\max)} - q_k^{(\min)})R_k$. It always has the dimension of an energy, and it is a quality measure for the energy terms obtained by fitting functional forms for (dE_k/dq_k) to the data sets D_k. Therefore, it is both practical and acceptable to identify the conversion factor w_k with the width of the physically relevant interval of q_k. One can intuitively estimate w_k, or alternatively one can obtain these widths from the geometries in the training set if this training set is generated by a well-behaving and extensive sampling procedure. We have observed that the gradient curves method is insensitive to any reasonable changes in the values w_k, and that it is sufficient to estimate the correct order of magnitude.

The $\tilde{s}_{(\mathrm{opt})}$ that minimizes expression Z can be substituted back into expression 7, after reordering this solution into vectors $s_{(\mathrm{opt})}^{(m)}$. This yields the sets of data points D_k that are optimally continuous and thus slightly scattered around a continuous curve. The minimization of Z makes sure that this scattering is minimal. The selection of a suitable functional form for each E_k is easily accomplished by inspecting the scatter plots of the transformed data sets D_k.

Unfortunately, the solution $\tilde{s}_{(\mathrm{opt})}$ is in general not unique. Since Z is a quadratic expression, only one global minimum exists, although that minimum can still be degenerate. In the case of a degenerate minimum, there is a subspace S that contains all the arguments of Z that yield the minimum value. The dimension of S is equal to the dimension of the null space of the matrix

$$\begin{aligned}\mathscr{H} &= \sum_k \tilde{\mathcal{N}}^{(k)T} C^k \tilde{\mathcal{N}}^{(k)} \\ &= \begin{pmatrix} \tilde{\mathcal{N}}^{(1)} \\ \vdots \\ \tilde{\mathcal{N}}^{(K)} \end{pmatrix}^T \begin{pmatrix} w_1^2 C^{(1)} & & 0 \\ & \ddots & \\ 0 & & w_K^2 C^{(K)} \end{pmatrix} \begin{pmatrix} \tilde{\mathcal{N}}^{(1)} \\ \vdots \\ \tilde{\mathcal{N}}^{(K)} \end{pmatrix} \\ &= \tilde{\mathcal{N}}^T C \tilde{\mathcal{N}} \end{aligned} \quad (17)$$

This matrix is a projection of the singular matrix C on a lower-dimensional space. Note that the matrix $\tilde{\mathcal{N}}^T$ is a nonsquare full-rank matrix by construction. Therefore, a unique solution $\tilde{s}_{(\mathrm{opt})}$ will only be available if the intersection of the range of $\tilde{\mathcal{N}}^T$ and the null space of C is empty.

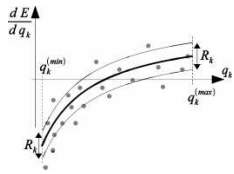


Figure 1. Schematic overview of how the weight factor w_k can be identified with the physically relevant interval of q_k. The fit of the auxiliary functions to the data set D_k is plotted, together with the RMS error on the fitted curve. The error on the integrated curve is approximated by $w_k R_k$.

Since

$$\tilde{\mathcal{N}} \in \mathbb{R}^{KM \times [K-(3N-6)]M} \quad (18)$$

with $N =$ the number of atoms and $K > 3N - 6$, one should expect $\mathscr{H}$ to be singular when $K \gg 3N - 6$, because then $\tilde{\mathcal{N}}$ is almost a square matrix. As stated in the introduction, an accurate force field always uses many more internal coordinates than independent coordinates. Consequently, for practical applications, a unique solution $\tilde{s}_{(\mathrm{opt})}$ will not be available, no matter how much training data are used. This is a reformulation of the parameter correlations that occur when conventional least-squares fitting is used to parametrize force-field models.

The degeneracy of the cost function gives us the opportunity to select a solution $\tilde{s}_{(\mathrm{opt})}$ that both minimizes Z and that will also result in a physically intuitive model. In this work, the physically intuitive character of a data set will be measured by a least-norm criterion: $\sum w_k^2 ||\tilde{y}^{(k)}||^2$. The lower this value, the smaller the forces along the internal coordinates in the resulting force-field model, and the more plausible the model. In general, $\mathscr{H}$ is much too large to store in any reasonable computer memory. It is therefore not feasible to perform a singular value decomposition of $\mathscr{H}$ in order to find the least-norm solution in S. Instead, a standard modification to the matrices C^k assures that Z has a unique solution that approximates the least-norm solution:

$$\begin{aligned} Z^*(\tilde{s}) &= Z(\tilde{s}) + \epsilon \sum_k w_k^2 (\tilde{y}^{(k)})^T \tilde{y}^k \\ &= \sum_k (\tilde{p}^{(k)} + \tilde{\mathcal{N}}^{(k)}\tilde{s})^T w_k^2 (C^{(k)} + \epsilon I)(\tilde{p}^{(k)} + \tilde{\mathcal{N}}^{(k)}\tilde{s}) \end{aligned} \quad (19)$$

where ϵ is a positive constant that is small compared to one. This approximation (of the least-norm solution) becomes exact in the limit of ϵ toward zero, but for numerical applications, the optimal value of ϵ depends on the floating point accuracy. The minimization of Z^* can now be accomplished by a conjugate gradient method and a sparse notation for all the matrices in expression 19.

For reasons of transparency, no restrictions on the functional dependencies of the different internal coordinates have been imposed in the above derivation, and we only considered geometries of a single molecule. When creating realistic force fields, the method is complicated by two practical aspects. First, a useful force field should describe the energy

Table 1. Overview of the a Priori Information Used by the Force-Field Models[a]

benchmark model	sets of equivalent internal coordinates	number of elements
Water_default	OH bond lengths	2
	HOH bending angles	1
	HOH span	1
Water_ext1	in addition to the internal coordinates of Water_default	
	(HOH bending cosine) × (OH bond lengths)	2
	(HOH span) × (OH bond lengths)	2
Water_ext2	in addition to the internal coordinates of Water_ext1	
	(OH1 bond length) × (OH2 bond length)	1
Ammonia_default	NH bond lengths	3
	HNH bending angles	3
	HNH spans	3
Ammonia_ext1	in addition to the internal coordinates of Ammonia_default	
	N(HHH) distance	1
	(N(HHH) distance) × (NH bond lengths)	3
Ammonia_ext2	in addition to the internal coordinates of Ammonia_ext1	
	(HNH bending cosines) × (NH bond lengths)	6
	(HNH spans) × (NH bond lengths)	6
Methane_default	CH bond lengths	4
	HCH bending angles	6
	HCH spans	6
Methane_ext1	in addition to the internal coordinates of Methane_default	
	(HCH bending cosines) × (CH bond lengths)	12
	(HCH spans) × (CH bond lengths)	12
Methane_ext2	in addition to the internal coordinates of Methane_ext1	
	(CH bond lengths) × (CH bond lengths)	6

[a]All internal coordinates that belong to the same set are modeled with the same function $E_k(q_k)$ (see eq 1).

dependence of equivalent internal coordinates with the same expression E_k. Second, for a good parametrization, one would sample geometries of different molecules. Because both extensions merely introduce more indexes in the derivation, the same method applies.

3. Benchmark Protocol

The comparison of our novel procedure with conventional force-field parametrizations follows a strict protocol that will be applied on three small benchmark molecules: H_2O, NH_3, and CH_4. The protocol consists of six steps: (i) the generation of training data by a sampling procedure that performs ab initio calculations on a set of different geometries of the given molecule, in addition to the generation of test data by a similar sampling procedure that covers a larger part of the potential energy surface, (ii) the selection of the internal coordinates that are used in the force-field model and the sets of equivalent internal coordinates q_k that are modeled with the same functional dependence, (iii) the gradient curves method presented in this paper, (iv) conventional force-field constructions, using the analytical expressions generated in the former step as input, and (v) the individual validation of each force-field model based on training and test data, and the comparison of all the force-field models.

3.1. Sampling Procedure. The sampling procedure starts with a geometry optimization of the given molecule. The optimized geometry is chosen as the origin of an equidistant $(3N - 6)$-dimensional grid. The training set is then extended iteratively, by selecting the neighboring grid point of the already calculated geometries that has the lowest estimated ab initio energy. For each benchmark molecule, 200 training samples and 200 test samples have been generated. The samples in the training data set span an energy range from 0 to 60 kJ mol^{-1} with respect to the optimized geometry (the origin), while the test samples have a higher upper limit of 100 kJ mol^{-1}. This sampling procedure is only appropriate for small molecules. For larger systems, Monte Carlo sampling should be used. Since our main aim is to test the gradient curves method (while the resulting parameters are of minor importance), a rather low level of theory (DFT/B3LYP) and a small basis set (3-21G*) were used. All ab initio calculations were performed with the MPQC program.[33]

3.2. Selection of Internal Coordinates. When developing a force field, one has to select sets of equivalent internal coordinates on which the force-field energy depends. For the gradient curves method, this is the only information that must be given in advance. In this work, nine benchmark force fields are extensively studied, using the different choices of coordinates described in Table 1. The default models for the three molecules use all the interatomic distances and all the cosines of the bending angles, as illustrated in Figure 2. These internal coordinates correspond to those in the well-known Urey−Bradley-type force field, but in this work, no quadratic functional dependencies are imposed. Additionally, two extended force fields are studied for each molecule. The products of internal coordinates in the extended models only contain products of different internal coordinates, and it is always assured that only products of related internal coordinates are considered; for example, a product of two bond

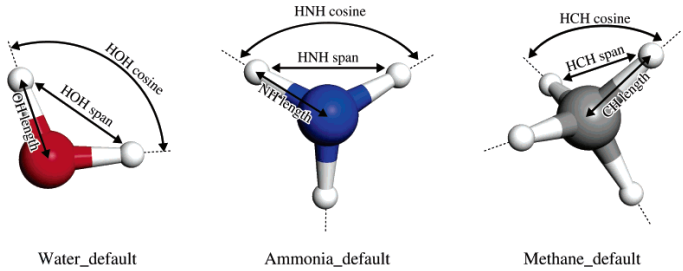


Figure 2. Schematic representation of the internal coordinates in the default models.

lengths will only be considered if the two bonds share exactly one atom. A detailed listing of which products have been used is given in the first section of the Supporting Information. Notice that the term "XYZ span" is defined as "the distance between the atoms X and Z that are both connected to the same atom Y", and the "A(BCD) distance" is defined as "the distance between an atom A and the plane that is defined by the atoms B, C, and D". The "XYZ span" is an internal coordinate initially introduced by Urey and Bradley[34] in their attempt to derive force fields for small molecules that show an improved reproduction of experimental vibrational frequencies. In their work, it is assumed that the corresponding energy term should be repulsive. We do not make this assumption a priori.

3.3. The Gradient Curves Method. For the auxiliary set of functions $f_n(q_k)$ in eq 8, polynomials up to the 11th order have been used. Two variants of the new gradient curves method are applied: **GCI** is the ill-conditioned variant of the new method, that is, without the least-norm criterion; **GCL** is the variant in which the least-norm correction is applied with $\epsilon = 10^{-6}$.

3.4. The Conventional Methods. In addition to the gradient curves method presented in this work, a series of conventional force-field parametrizations has been performed. They are conventional in the sense that the parameters have been obtained by directly minimizing a well-defined least-squares cost function, although in the literature, additional techniques are used to deal with parameter correlations. The different types of cost functions are listed below. Optionally, a constraint has been applied that compels the force field to reproduce the ab initio Hessian and the zero gradient for the ab initio optimized geometry.

CEU is an unconstrained minimization of the residual error on the energies[35]

$$Z_{\mathrm{CEU}} = \sum_{m=1}^{M} [(E_{\mathrm{AI}}^{(m)} - E_{\mathrm{AI}}^{(\mathrm{opt})}) - (E_{\mathrm{FF}}^{(m)} - E_{\mathrm{FF}}^{(\mathrm{opt})})]^2 \qquad (20)$$

where the sum over m contains all the molecules in the training set and corrections due to the difference in reference energies of the ab initio and the force-field model have been taken into account.

CEC is a minimization of the residual error on the energies constrained so that the ab initio Hessian and zero gradient are reproduced at the ab initio equilibrium geometry: $Z_{\mathrm{CEC}} = Z_{\mathrm{CEU}}$.

CGU is an unconstrained minimization of the residual error on the gradients[36]

$$Z_{\mathrm{CGU}} = \sum_{m=1}^{M} \sum_{i=1}^{3N} \left(\frac{\partial E_{\mathrm{AI}}^{(m)}}{\partial x_i} - \frac{\partial E_{\mathrm{FF}}^{(m)}}{\partial x_i} \right)^2 \qquad (21)$$

where i iterates over the Cartesian coordinates.

CGC is a minimization of the residual error on the gradients constrained such that the ab initio Hessian and zero gradient are reproduced at the ab initio equilibrium geometry: $Z_{\mathrm{CGC}} = Z_{\mathrm{CGU}}$.

CCU is an unconstrained minimization of the residual error on the energies and gradients of all the training geometries, as well as the Hessian of the optimized molecule where (i,j) iterates over all the pairs of the Cartesian

$$Z_{\mathrm{CCU}} = W_{\mathrm{CEU}} Z_{\mathrm{CEU}} + W_{\mathrm{CGU}} Z_{\mathrm{CGU}} + W_{\mathrm{CHU}} \sum_{i=1}^{3N} \sum_{j=i}^{3N} \left(\frac{\partial^2 E_{\mathrm{FF}}^{(\mathrm{opt})}}{\partial x_i \partial x_j} - \frac{\partial^2 E_{\mathrm{AI}}^{(\mathrm{opt})}}{\partial x_i \partial x_j} \right)^2 \qquad (22)$$

coordinates. The three contributions to the cost function have been weighted to ensure that they have a proportional influence on the obtained parameters. Alternative cost functions that combine ab initio energies, gradients, and/or Hessians have also been reported in the literature for the optimization of force-field parameters.[1,10,11]

The conventional parametrizations will serve as a reference for the results of the gradient curves method. To guarantee a fair comparison, the analytical expressions used in the conventional methods where obtained with GCL and these expressions only contain linear parameters.

3.5. Validation and Comparison. The generated force-field models are validated with three different criteria. (i) The standard deviation on $E_{\mathrm{FF}} - E_{\mathrm{AI}}$ for all geometries, defined as $\langle[(E_{\mathrm{FF}} - E_{\mathrm{AI}}) - \langle E_{\mathrm{FF}} - E_{\mathrm{AI}}\rangle]^2\rangle^{(1/2)}$, should be small. The standard deviation is not sensitive to the reference energies of both ab initio and force-field models, in contrast to the root mean square of $E_{\mathrm{FF}} - E_{\mathrm{AI}}$, given by $\langle(E_{\mathrm{FF}} - E_{\mathrm{AI}})^2\rangle^{(1/2)}$. (ii) The root mean square of $|\nabla E_{\mathrm{FF}} - \nabla E_{\mathrm{AI}}|$ should

be small where ∇ indicates the Cartesian gradient. (iii) At the ab initio optimized geometry, the ratios of the eigenvalues for matching eigenvectors of the force field and of the ab initio Hessian should be near unity.

The third quality criterion is calculated as follows. First, the ab initio Hessian and the force-field Hessian are calculated at the ab initio optimized geometry. The eigenmodes corresponding to the external degrees of freedom are removed by projecting both Hessians on the same basis of $3N - 6$ independent internal coordinates. Then, both projected matrices are diagonalized. The overlap matrix of the corresponding sets of eigenvectors shows clearly which two eigenvectors of the ab initio Hessian and the force-field Hessian correspond with each other. Significant mismatches have not been observed. Finally, the ratios of the eigenvalues associated with the corresponding eigenvectors are calculated.

The quality of the force fields will be compared by the three criteria defined above. In order to assess the robustness of the parametrization, validations i and ii are in addition applied to the set of test data. Finally, we have examined the possibility of giving a physical interpretation to the force-field expressions E_k obtained in the different models.

4. Results and Discussions

To illustrate the usage of the new procedure, we first discuss the three gradient curves generated by GCL applied to Water_default. For each geometry m, the Jacobian, $J^{(m)}$ (see eq 3) is a $N \times K$ matrix or 9×4 matrix of rank $3N - 6 = 3$. The matrix $\mathcal{N}^{(m)}$ describing the null space of such a Jacobian has the dimension $N \times K - (3N - 6)$ or 9×1. Consequently, given the 200 geometries in the training set, 200 unknown coefficients must be obtained by minimizing the cost function, Z^*. Although there are four distinct internal coordinates in this specific force-field model, the two transformed data sets corresponding to the OH-bond length have been merged into one; that is, their continuity is measured as a whole. Consequently, the data set associated with the bond length consists of 400 data points, while the two others contain 200 data points each. The continuity of each data set is measured by the goodness of fit of an auxiliary 11th-order polynomial. We used generic high-order polynomials to prevent any assumptions about the resulting energy terms being imposed by the continuity criterion; that is, these polynomials will not enforce specific features in the final energy terms. The results are depicted in Figure 3. The data sets D_k obtained by substituting $s_{(\mathrm{opt})}$ into eq 7 are plotted as black crosses. The minimization of Z^* guarantees that these data points lie on continuous curves. The (optimized) auxiliary polynomials that are used to measure the continuity are plotted as dashed lines. Their unphysical asymptotic behavior and the oscillations at the boundaries clarify that the auxiliary polynomials can only be regarded as a measure for the continuity and that they cannot be used as functional forms for the force-field model. In a next step, the analytical form of the derivative of E_k is estimated, on the basis of the data sets. For the energy curve of the OH stretch, a sixth-order polynomial in $1/r_{\mathrm{OH}}$ gives an accurate fit, and the resulting expression has the expected asymptotic

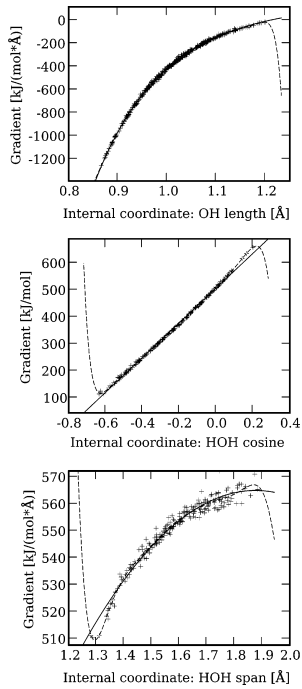


Figure 3. Gradient curves dE_k/dq_k (solid line) obtained for the Water_default model with the GCL method. The black crosses represent the transformed one-dimensional data (see text). The dashed curves are the fitted auxiliary functions for evaluating the continuity criterion.

behavior. The energy curves of the cosine of the bending angle and the interatomic HH distance are estimated to be quadratic and cubic, respectively. Finally, the parameters in the functional forms are optimized using one-dimensional least-squares optimization to the data sets D_k. The resulting curves, dE_k/dq_k, are plotted as solid lines in Figure 3, and the optimized parameters are given in Table 2. The GCL parameters for all nine benchmark models are included in the second section of the Supporting Information.

An overview of the quality criteria for each parametrization is given in Figure 4. The x axis shows the force-field models, and for each force-field model, the different parametrization methods (GCI, GCL, CEU, and so forth) are indicated with different colors. On the y axes, the quality criteria are plotted on a logarithmic scale. Figure 4a and b display respectively the standard deviation on ($E_{\mathrm{FF}} - E_{\mathrm{AI}}$) and the root mean square of $|\nabla E_{\mathrm{FF}} - \nabla E_{\mathrm{AI}}|$ for both training geometries (filled circles) and test geometries (open circles). Figure 4c gives an overview of the validation with the third criterion, represented by the ratios of corresponding Hessian eigenvalues (force-field over ab initio estimates) at the ab initio optimized geometry. It is clear that the overall quality of

Table 2. The Parameters for the Water Default Model Obtained with GCL[a]

OH bond length r		HOH bending cosine c		HOH span d	
terms	coefficients	terms	coefficients	terms	coefficients
r^{-1}	−4.608e−01	c	1.931e−01	d	9.898e−03
r^{-2}	5.210e−02	c^2	1.228e−01	d^2	2.933e−02
r^{-3}	3.578e−01			d^3	−2.758e−03
r^{-4}	3.988e−01				
r^{-5}	3.103e−01				
r^{-6}	1.943e−01				

[a] The functional form of each energy term, $E_k(q_k) = \sum_{i=1}^{I_k} c_k \mathcal{F}_i(q_k)$, is a linear combination of terms listed in the first column of each table. The corresponding coefficients in this linear combination are given in the second column. All parameters are given in atomic units.

the force fields constructed with GCL is comparable to that obtained by the conventional methods. Nevertheless, some interesting discrepancies appear, which will be discussed below.

The ammonia molecule serves as a good example of how to obtain relevant sets of internal coordinates. Initially, the new method was applied on the Ammonia_default model, which only contains the basic internal coordinates: bond lengths, interatomic distances, and bending angles. As shown in Figure 4c, the constructed force field predicts one eigenvalue of the Hessian that deviates significantly from the ab initio value. This eigenvalue corresponds to the inversion of the ammonia molecule. At the transition state of this umbrella inversion, the NH bond length increases due to the alteration from sp^3 to sp^2 hybridization. To describe the inversion more accurately, the extended ammonia model contains two extra sets of internal coordinates: the out-of-plane distance and the products of the out-of-plane distance with the bond lengths. It is striking to observe that the parametrization of the extended ammonia model results in a seriously improved reproduction of the eigenvalues. An attempt was made to avoid the inclusion of more internal coordinates, by constraining the parameters in order to reproduce the ab initio Hessian. This failed drastically for ammonia and methane, since these constraints led to unacceptable errors on the energies and gradients for both training and test data. The corresponding quality criteria falls out of the scope of Figure 4a and b. The performance of CCU in the parametrization of the Ammonia_default model manifestly suffers from the attempt to use information of the ab initio Hessian in the optimization.

The parametrization of ammonia demonstrates that, in some cases, the inclusion of additional redundant internal coordinates in a force-field model is indispensable. This is in agreement with previous studies where it was shown that a pure Urey−Bradley force field, that is, the default model in this work, is not sufficient for an adequate description of the ammonia molecule.[37,38] Unfortunately, the parametrization of a force field with a high number of internal coordinates ($K \gg 3N - 6$) is sensitive to parameter correlations, and a good treatment of these correlations is required to obtain a useful force field.

In the remainder of this section, we discuss the effect of increasing model complexity. The main effect of the extensions to the force-field models is visible in Figure 4. An improved reproduction of the energies, gradients, and the Hessian is obtained for all methods except the GCI method. This general trend is understandable: the more parameters a model contains, the further a cost function can be optimized. The poor performance of the GCI method for the extended models needs some explanation. Both GCI and GCL yield the same transformed data sets D_k for the default models. For these models, the cost function Z (see eq 16) has a unique solution, even without applying the least-norm correction. This is no longer true for the extended models. In these cases, the minimum of the cost function, Z, becomes highly degenerate, and GCI selects from this minimum an essentially random solution, in the sense that a small change in the training data would imply a very large change in the transformed data sets D_k. On average, such a random solution consists of transformed data sets D_k with very high ranges. This is unacceptable because the absolute errors from fitting energy terms to the transformed data sets (step II of the gradient curves method) scale with the range of D_k. Consequently, the absolute errors shown in Figure 4 are much higher for GCI when applied to the most extended models. We conclude that, of the new methods, GCL is to be preferred over GCI. Both are equivalent for a small number of internal coordinates, but GCL produces superior fits for the more extended models.

The most important trend noticed by increasing the complexity of the model is the behavior of the functions E_k, which is different for GCL as compared to all other methods (i.e., the conventional methods and GCI). Figures 5−7 display all the energy terms E_k, obtained with CCU and GCL, for the water, ammonia, and methane molecules, respectively. In these figures, CCU could have been replaced by any other method except GCL without generating significant differences in the global trends. Each row in these figures contains the plots of the energy terms that belong to a specific force-field parametrization, while every column corresponds to a specific set of equivalent internal coordinates. In what follows, we will first discuss the global trends in these figures, and consequently some more specific aspects will be discussed that are not applicable to all the results.

Figures 5a, 6a, and 7a show that CCU yields energy terms E_k with increasing amplitudes, when the force-field model is extended with extra internal coordinates. The conventional methods use the extra degrees of freedom to improve the accuracy, but this improvement is the result of a nonrobust cancellation of high-energy contributions. We have tested an implementation of the conventional methods that applies a singular value decomposition to the design matrix,[32,39] but a singular value cutoff that gives a good balance between accuracy and reasonable behavior of the functions E_k is not available. The reason is that a least-norm solution in the parameter space is not meaningful since the parameters have different units. A weighted least-norm solution, where the norm of dimensionless weighted parameters is minimal, would be more correct, but then one has to determine a weight value for each parameter as in the work of Ewig et al.[11] It is highly remarkable that, as depicted in Figures 5b,

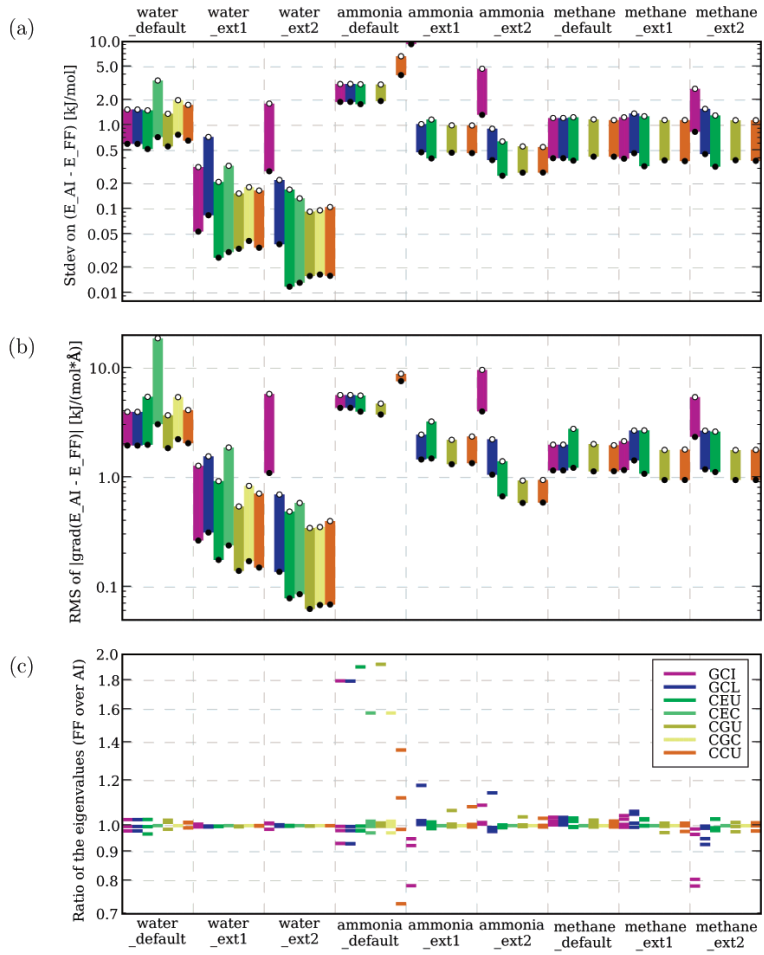


Figure 4. Overview of the force-field validations. Upper figure (a): Standard deviation of the energy differences. Middle figure (b): Root mean square of the gradient differences. Lower figure (c): Ratios of corresponding Hessian eigenvalues (force field over ab initio values), at the ab initio optimized geometry (see text). In parts a and b, the errors for the constrained methods applied on ammonia and methane are too large to fit in the scale of both plots.

6b, and 7b, GCL shows exactly the opposite trend from CCU: the ranges of the functions E_k are reduced in the extended force-field models, and for the ext2 models, it is even possible to give a physical interpretation to the important energy dependencies. For example, the minima of E_k correspond approximately to the internal coordinates of the ab initio optimized geometry. For the terms E_{OH}, E_{NH}, and E_{CH}, even a Morse-like behavior (i.e., the left side of the curve is steeper than the right side) is reproduced. It should be remarked that GCL does not depend on constraints, model selection, or ad hoc interventions to obtain physical force-field terms. When the gradient curves method will be applied on larger systems, we expect that the absence of cancellation effects will yield transferable and accurate force fields.

In addition to the global trends discussed above, some interesting specific features show up in the results. The most remarkable outcome is that the energy terms for the Ammonia_default model obtained with CCU are very reasonable, and at first instance, this appears to contradict the previous

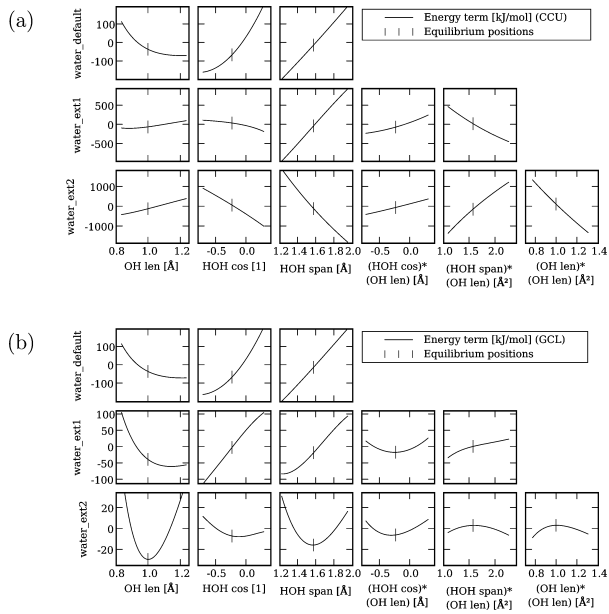


Figure 5. Energy terms E_k for the three different water models, generated (a) by CCU, a conventional parametrization method, and (b) by GCL, the gradient curves method with the least-norm correction. The values of the internal coordinates at the ab initio equilibrium geometry are marked by vertical lines.

paragraph where we stated that reasonable models could only be obtained with GCL. The explanation is that the Ammonia_default model with CCU parameters is indeed reasonable but less accurate compared to other parametrizations of Ammonia_default (see Figure 4a and b). The energy terms for the Ammonia_default model obtained with CGU (see Figure 6c) reveal that the incorporation of the ab initio Hessian of the optimized ammonia geometry in the CCU cost function forces the energy terms of the Ammonia_default model to behave reasonably.

A more subtle result is that the first row of Figure 5a contains virtually the same energy terms as the first row in Figure 5b. Similarly, the first row of both parts a and b of Figure 7 are virtually equal. This situation can be summarized as follows: CCU, a method that does not handle parameter correlations, yields the same energy terms as GCL, a method that does treat parameter correlations. The reason is that none of the parametrization methods in this paper suffer from parameter correlation problems in case of the default models. In the case of the Water_default or the Methane_default model, all the uniquely defined minima of the cost functions of CCU, CEU, CGU, GCL, and GCI even result in the same energy terms. As already discussed above, the different cost functions in the case of Ammonia_default have different—but each of them uniquely defined—optimal parameters. The absence of parameter correlations does however not imply reasonable energy terms. Actually, the sets of equivalent internal coordinates in the default models are too limited for an accurate reproduction of all the training data with reasonable energy terms. The OH-stretch term represents a repulsive interaction, whereas the energy terms for the HH distance and HOH cosine are both attractive interactions. Correct behavior is obtained only when the three energy terms are combined. For reasons of clarity, we note that the GCL curves in the default models are not supposed to coincide perfectly with the quadratic energy terms in a standard Urey−Bradley parametrization, which are fitted so as to reproduce experimental frequencies.[40−42] In the present case, the curves are fitted not only to molecular configurations near equilibrium but to higher-energy configurations as well. In fact, when the curves in the first row of Figures 5b and 7b are quadratically expanded around the equilibrium values, a fair correlation with the quadratic force constants and the minima in the work of Kuchitsu and Bartell[40,41] is observed.

At this point, we have shown how the gradient curves method is able to reconcile the accuracy and the physical interpretation of a force-field model. However, one could wonder how the energy terms, as shown in Figures 5b, 6b, and 7b, evolve when the force-field model is extended with even more additional sets of equivalent internal coordinates (higher-order products, cubic terms, etc.). In the HDMR approach,[25] orthogonality criteria are introduced to assert that the addition of higher-order terms does not have any

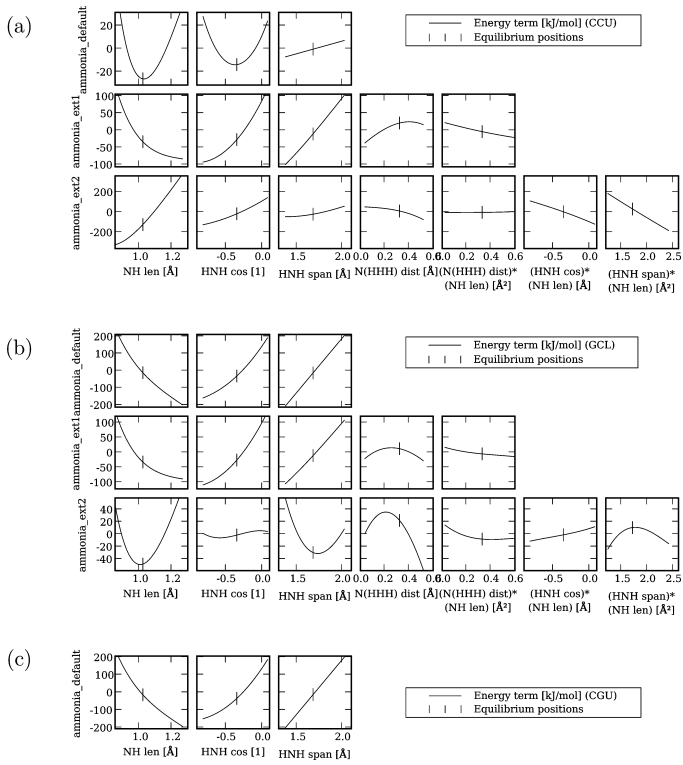


Figure 6. Energy terms E_k for the three different ammonia models, generated (a) by CCU, a conventional parametrization method, (b) by GCL, the gradient curves method with the least-norm correction, and (c) by CGU, a conventional parametrization method that only uses ab initio gradient training data. For part c, only the default model is shown. The values of the internal coordinates at the ab initio equilibrium geometry are marked by vertical lines.

influence on the lower-order terms in the model. The gradient curves method never relies on such orthogonality criteria; for example, this is the reason why the energy terms for the bond length of the three models in Figure 5b are different. There is no "theoretical guarantee" that modifications will not occur when the water model is extended with even more sets of equivalent internal coordinates. Additional energy terms make the continuity criterion extremely degenerate, and in such cases, the least-norm criterion might become an overly naive representation of our physical intuition. Figure 8 demonstrates the behavior of the energy terms for a series of additional extended water models. Similar plots for ammonia and methane are included in the third section of the Supporting Information. Except for the highest-order terms in the two most extended models for the water molecule, the modifications in the energy terms seem to converge once the model is extended enough to show a physically intuitive behavior. The inclusion of second-order derivatives of the ab initio energy in the training data and more sophisticated criteria for our physical intuition are viable candidates to cure the situation for the two most extended water models and are the subject of our current active research. Nevertheless, one should realize that also these additional measures would suffer from the same defects for the very hypothetical case of even more extended models.

5. Conclusions

This work shows how the gradient curves method can surmount several difficulties that are associated with the development of force fields using least-squares parametrization. Technically, the new method is a two-step procedure: in the first step, continuity criteria and subordinate least-norm criteria are imposed to transform the multidimensional training data into a series of separate one-dimensional data sets, each associated with an energy term of the proposed force field. In this work, the training data are the gradients of the ab initio energy for different molecular geometries.

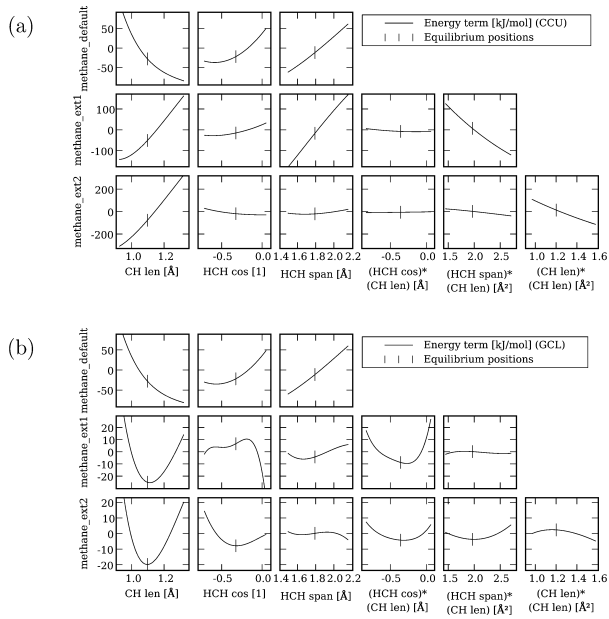


Figure 7. Energy terms E_k for the three different methane models, generated (a) by CCU, a conventional parametrization method, and (b) by GCL, the gradient curves method with the least-norm correction. The values of the internal coordinates at the ab initio equilibrium geometry are marked by vertical lines.

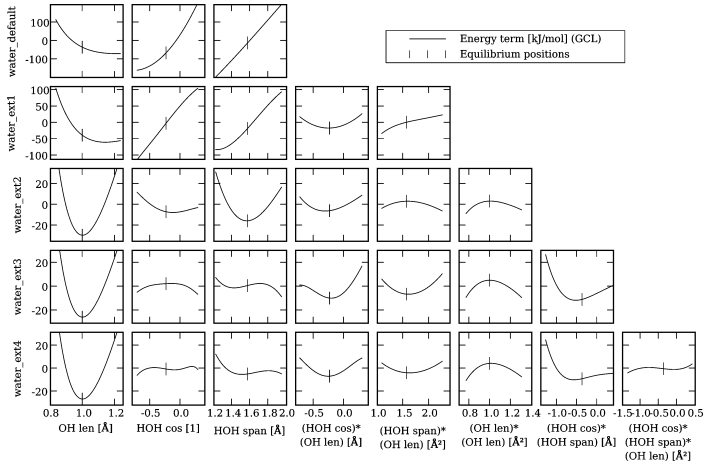


Figure 8. Overview of the energy terms for additional extended water models parametrized with GCL.

During the second step, the derivative of each energy term in the force field is fitted to the corresponding transformed data set.

The gradient curves method has several advantages. Only the internal coordinates have to be defined in advance, instead of a complete analytical ansatz of the force-field model. The

problem of parameter correlations that troubles the conventional force-field development is tackled during the transformation from the multidimensional training data to separate one-dimensional data sets. The continuity and least-norm criteria that are imposed do not only guarantee that the transformed data sets are unique but they also facilitate the physical interpretation of the energy terms fitted to these data sets. In fact, the least-norm criteria express the argument that a plausible force-field model should not contain large derivatives in the energy terms to acquire a marginal increase of accuracy. This prescription fixes all the parameter correlations that originate from the redundancy of the internal coordinates in the force-field model. Once the first step is completed, suitable analytical expressions for the energy terms can be easily proposed after analysis of the transformed data sets and taking into account the expected asymptotic behavior of these energy terms. Because the ability of interpreting the individual force-field terms is known to be a prerequisite for transferable force fields,[2,11] we expect this method to be very helpful when developing accurate and robust force-field models for larger systems.

The current research mainly focuses on an extended variation of the gradient curves method which is also capable of efficiently deriving the nonbonding interactions from ab initio training data. The primary application on a large system will be the construction of an accurate all-atom zeolite-guest force field. Other active areas include the extension of the gradient curves method to include the ab initio energy and Hessian in the training data, and a more sophisticated formalism for the intuitive character of the energy terms that will eventually supersede the least-norm criterion. We also expect a generalization of the gradient curves method (beyond the scope of force fields) to be useful whenever data parametrization is complicated by parameter correlations and the absence of theoretically supported analytical models.

Acknowledgment. T.V. would like to thank the Flemish organization IWT for its financial support. P.W.A. would like to thank NSERC for funding. V.V.S and M.W. thank the Fund for Scientific Research—Flanders and the Research Board of Ghent University.

Supporting Information Available: A listing of the internal coordinates, the GCL parameters, and an overview of additional extended models. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Schröder, K. P.; Sauer, J. *J. Phys. Chem.* **1996**, *100*, 11043−11049.

(2) Hill, J.; Sauer, J. *J. Phys. Chem.* **1995**, *99*, 9536−9550.

(3) Sierka, M.; Sauer, J. *Faraday Discuss.* **1997**, *106*, 41−62.

(4) Smirnov, K. S.; Bougeard, D. *Chem. Phys.* **2003**, *292*, 53−70.

(5) Ermoshin, V. A.; Engel, V. *J. Phys. Chem. A* **1999**, *103*, 5116−5122.

(6) Chandross, M.; Webb, E. B.; Grest, G. S.; Martin, M. G.; Thompson, A. P.; Roth, M. W. *J. Phys. Chem. B* **2001**, *105*, 5700−5712.

(7) Pascual, P.; Ungerer, P.; Tavitian, B.; Pernot, P.; Boutin, A. *Phys. Chem. Chem. Phys.* **2003**, *5*, 3684−3693.

(8) Allinger, N. L.; Chen, K.; Lii, J.-H. *J. Comput. Chem.* **1996**, *17*, 642−668.

(9) Halgren, T. A. *J. Comput. Chem.* **1996**, *17*, 490−519.

(10) Sun, H.; Rigby, D. *Spectrochim. Acta, Part A* **1997**, *53*, 1301−1323.

(11) Ewig, C.; Berry, R.; Dinur, U.; Hill, J.; Hwang, M.; Li, H.; Liang, C.; Maple, J.; Peng, Z.; Stockfisch, T.; Thacher, T.; Yan, L.; Ni, X.; Hagler, A. *J. Comput. Chem.* **2001**, *22*, 1782−1800.

(12) Bayly, C.; Cieplak, P.; Cornell, W.; Kollman, P. *J. Phys. Chem.* **1993**, *97*, 10269−10280.

(13) Mayo, S.; Olafson, B.; Goddard, W. *J. Phys. Chem.* **1990**, *94*, 8897−8909.

(14) Rappe, A.; Casewit, C.; Colwell, K.; Goddard, W.; Skiff, W. *J. Am. Chem. Soc.* **1992**, *114*, 10024−10035.

(15) Shi, S.; Yan, L.; Yang, Y.; Fisher-Shaulsky, J.; Thacher, T. *J. Comput. Chem.* **2003**, *24*, 1059−1076.

(16) Mortier, W.; Ghosh, S.; Shankar, S. *J. Am. Chem. Soc.* **1986**, *108*, 4315−4320.

(17) Rick, S. W.; Stuart, S. J.; Berne, B. J. *J. Chem. Phys.* **1994**, *101*, 6141−6156.

(18) Kaminski, G. A.; Stern, H. A.; Berne, B. J.; Friesner, R. A. *J. Phys. Chem. A* **2004**, *108*, 621−627.

(19) Chalasinski, G.; Szczesniak, M. *Chem. Rev.* **2000**, *100*, 4227−4252.

(20) Bordner, A. J.; Cavasotto, C. N.; Abagyan, R. A. *J. Phys. Chem. B* **2003**, *107*, 9601−9609.

(21) Giese, T.; York, D. *Int. J. Quantum Chem.* **2004**, *98*, 388−408.

(22) Maple, J.; Dinur, U.; Hagler, A. *Proc. Natl. Acad. Sci. U.S.A.* **1988**, *85*, 5350−5354.

(23) Maple, J. R.; Hwang, M. J.; Stockfisch, T. P.; Dinur, U.; Waldman, M.; Ewig, C. S.; Hagler, A. T. *J. Comput. Chem.* **1994**, *15*, 162−182.

(24) Martinez, E.; Lopez, J. J.; Vazquez, J. *J. Mol. Struct.* **2004**, *705*, 141−145.

(25) Rabitz, H.; Aliş, O.; Shorter, J.; Shim, K. *Comput. Phys. Commun.* **1999**, *117*, 11−20.

(26) Manzhos, S.; Carrington, T. *J. Chem. Phys.* **2006**, *125*, 084109.

(27) Shorter, J. A.; Ip, P. C.; Rabitz, H. A. *J. Phys. Chem.* A **1999**, *103*, 7192−7198.

(28) Gorban, A. N. *Appl. Math. Lett.* **1998**, *11*, 45−49.

(29) Frisch, H. L.; Borzi, C.; Ord, G.; Percus, J. K.; Williams, G. *Phys. Rev. Lett.* **1989**, *63*, 927−929.

(30) Kolmogorov, A. *Dokl. Akad. Nauk SSSR* **1957**, *114*, 679.

(31) Hilbert, D. *Bull. Am. Math. Soc.* **1902**, *8*, 461.

(32) Press, W. H.; Flannery, B. P.; Teukolsky, S. A.; Vetterling, W. T. Singular Value Decomposition. In *Numerical Recipes in C: The Art of Scientific Computing*; Cowles, L., Harvey, A., Hahn, R., Eds.; Press Syndicate of the University of Cambridge: Cambridge, United Kingdom, 2002; Chapter 2.6, pp 59−70.

(33) Janssen, C. L.; Nielsen, I. B.; Leininger, M. L.; Valeev, E. F.; Seidl, E. Y. *The Massively Parallel Quantum Chemistry Program (MPQC)*, version 2.3.0; Sandia National Laboratories: Livermore, CA, 2004.

(34) Urey, H. C.; Bradley, C. A. *Phys. Rev.* **1931**, *38*, 1969−1978.

(35) Kramer, G.; Farragher, N.; van Beest, B.; van Santen, R. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1991**, *43*, 5068−5080.

(36) Ercolessi, F.; Adams, J. *Europhys. Lett.* **1994**, *26*, 583−588.

(37) Pariseau, M.; Wu, E.; Overend, J. *J. Chem. Phys.* **1962**, *37*, 217−223.

(38) King, W. T. *J. Chem. Phys.* **1961**, *36*, 165−170.

(39) Press, W. H.; Flannery, B. P.; Teukolsky, S. A.; Vetterling, W. T. General Linear Least Squares. In *Numerical Recipes in C: The Art of Scientific Computing*; Cowles, L., Harvey, A., Hahn, R., Eds.; Press Syndicate of the University of Cambridge: Cambridge, United Kingdom, 2002; Chapter15.4, pp 671−681.

(40) Kuchitsu, K.; Bartell, L. S. *J. Chem. Phys.* **1962**, *36*, 2460−2469.

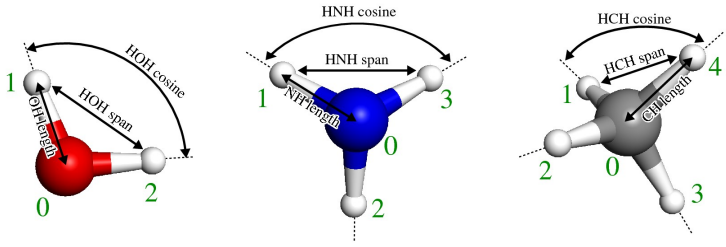
(41) Kuchitsu, K.; Bartell, L. S. *J. Chem. Phys.* **1962**, *36*, 2470−2481.

(42) Simanouthi, T. *J. Chem. Phys.* **1949**, *17*, 245−248.

CT6002093

1) Listing of the internal coordinates

Each set of equivalent internal coordinates is listed in the three tables below. Sets with products of internal coordinates only contain products of related internal coordinates. The atoms are indexed with numbers starting from zero as shown in the figure below.



Water_default Ammonia_default Methane_default

Water

Set of equivalent internal coordinates	Items in each set
OH length	bond length 0-2 bond length 0-1
HOH cosine	bending cosine 1-0-2
HOH span	Urey-Bradley 1-0-2
(HOH cosine)*(OH length)	bending cosine 1-0-2 * bond length 0-2 bending cosine 1-0-2 * bond length 0-1
(HOH span)*(OH length)	Urey-Bradley 1-0-2 * bond length 0-2 Urey-Bradley 1-0-2 * bond length 0-1
(OH length)*(OH length)	bond length 0-1 * bond length 0-2

Ammonia

Set of equivalent internal coordinates	Items in each set
NH length	bond length 0-2 bond length 0-3 bond length 0-1
HNH cosine	bending cosine 1-0-3 bending cosine 1-0-2 bending cosine 2-0-3
HNH span	Urey-Bradley 1-0-3 Urey-Bradley 1-0-2 Urey-Bradley 2-0-3
N(HHH) distance	out of plane distance 0-(2,3,1)
(N(HHH) distance)*(NH length)	out of plane distance 0-(2,3,1) * bond length 0-2 out of plane distance 0-(2,3,1) * bond length 0-3 out of plane distance 0-(2,3,1) * bond length 0-1

(HNH cosine)*(NH length)	bending cosine 1-0-3 * bond length 0-3 bending cosine 1-0-3 * bond length 0-1 bending cosine 1-0-2 * bond length 0-2 bending cosine 1-0-2 * bond length 0-1 bending cosine 2-0-3 * bond length 0-2 bending cosine 2-0-3 * bond length 0-3
(HNH span)*(NH length)	Urey-Bradley 1-0-3 * bond length 0-3 Urey-Bradley 1-0-3 * bond length 0-1 Urey-Bradley 1-0-2 * bond length 0-2 Urey-Bradley 1-0-2 * bond length 0-1 Urey-Bradley 2-0-3 * bond length 0-2 Urey-Bradley 2-0-3 * bond length 0-3

Methane

Set of equivalent internal coordinates	Items in each set
CH length	bond length 0-2 bond length 0-3 bond length 0-1 bond length 0-4
HCH cosine	bending cosine 1-0-3 bending cosine 1-0-4 bending cosine 1-0-2 bending cosine 3-0-4 bending cosine 2-0-3 bending cosine 2-0-4
HCH span	Urey-Bradley 1-0-3 Urey-Bradley 1-0-4 Urey-Bradley 1-0-2 Urey-Bradley 3-0-4 Urey-Bradley 2-0-3 Urey-Bradley 2-0-4
(HCH cosine)*(CH length)	bending cosine 1-0-3 * bond length 0-3 bending cosine 1-0-3 * bond length 0-1 bending cosine 1-0-4 * bond length 0-1 bending cosine 1-0-4 * bond length 0-4 bending cosine 1-0-2 * bond length 0-2 bending cosine 1-0-2 * bond length 0-1 bending cosine 3-0-4 * bond length 0-3 bending cosine 3-0-4 * bond length 0-4 bending cosine 2-0-3 * bond length 0-2 bending cosine 2-0-3 * bond length 0-3 bending cosine 2-0-4 * bond length 0-2 bending cosine 2-0-4 * bond length 0-4
(HCH span)*(CH length)	Urey-Bradley 1-0-3 * bond length 0-3 Urey-Bradley 1-0-3 * bond length 0-1 Urey-Bradley 1-0-4 * bond length 0-1 Urey-Bradley 1-0-4 * bond length 0-4 Urey-Bradley 1-0-2 * bond length 0-2 Urey-Bradley 1-0-2 * bond length 0-1 Urey-Bradley 3-0-4 * bond length 0-3 Urey-Bradley 3-0-4 * bond length 0-4 Urey-Bradley 2-0-3 * bond length 0-2 Urey-Bradley 2-0-3 * bond length 0-3 Urey-Bradley 2-0-4 * bond length 0-2 Urey-Bradley 2-0-4 * bond length 0-4

(CH length)*(CH length)	bond length 0-2 * bond length 0-1 bond length 0-2 * bond length 0-4 bond length 0-3 * bond length 0-2 bond length 0-3 * bond length 0-1 bond length 0-3 * bond length 0-4 bond length 0-4 * bond length 0-1

2) GCL Parameters

The tables below list the coefficients of the force-field models Water_default, Water_ext1, Water_ext2, Ammonia_default, Ammonia_ext1, Ammonia_ext2, Methane_default, Methane_ext1 and Methane_ext2 that have been obtained with the Gradient Curves Method with least norm criterion. Each energy term has the following functional form

$$E_k(x)=\sum_{t=1}^{T} c_t f_t(x) \quad ,$$

i.e. a linear combination of terms that are listed in the first row of each table below. The corresponding coefficients in this linear combination are given in the second row. All parameters are given in atomic units.

Water_default

OH length

Terms	x**(-1)	x**(-2)	x**(-3)	x**(-4)	x**(-5)	x**(-6)
Coefficients [a.u.]	-4.608e-01	5.210e-02	3.578e-01	3.988e-01	3.103e-01	1.943e-01

HOH cosine

Terms	x	x**2
Coefficients [a.u.]	1.931e-01	1.228e-01

HOH span

Terms	x	x**2	x**3
Coefficients [a.u.]	9.898e-03	2.933e-02	-2.758e-03

Water_ext1

OH length

Terms	x**(-1)	x**(-2)	x**(-3)	x**(-4)	x**(-5)	x**(-6)
Coefficients [a.u.]	8.325e-01	-2.626e+00	7.701e-01	2.864e+00	1.244e+00	-2.539e+00

HOH cosine

Terms	x	x**2	x**3	x**4	x**5	x**6
Coefficients [a.u.]	8.171e-02	-3.232e-02	-3.962e-02	5.269e-03	9.427e-03	-1.689e-02

HOH span

Terms	x	x**2	x**3	x**4
Coefficients [a.u.]	-1.413e+00	5.808e-01	-9.690e-02	5.629e-03

(HOH cosine)(OH length)*

Terms	x	x**2	x**3	x**4
Coefficients [a.u.]	1.497e-02	1.517e-02	-2.169e-04	1.760e-04

(HOH span)(OH length)*

Terms	x	x**2	x**3	x**4
Coefficients [a.u.]	1.055e-01	-2.128e-02	1.956e-03	-6.687e-05

Water_ext2

OH length

Terms	x**(-1)	x**(-2)	x**(-3)	x**(-4)	x**(-5)	x**(-6)
Coefficients [a.u.]	-2.354e+00	1.481e+00	1.095e+00	-1.499e-01	-3.707e-01	3.850e-01

HOH cosine

Terms	x	x**2	x**3	x**4	x**5	x**6
Coefficients [a.u.]	3.953e-03	1.211e-02	-2.093e-02	-1.280e-02	4.230e-03	1.191e-03

HOH span

Terms	x	x**2	x**3	x**4
Coefficients [a.u.]	-5.538e-01	1.541e-01	-1.364e-02	4.438e-06

(HOH cosine)(OH length)*

Terms	x	x**2	x**3	x**4
Coefficients [a.u.]	5.622e-03	3.240e-03	-1.435e-03	3.045e-04

(HOH span)(OH length)*

Terms	x	x**2	x**3	x**4
Coefficients [a.u.]	1.286e-02	-1.532e-03	4.624e-05	4.021e-08

(OH length)(OH length)*

Terms	x	x**2	x**3	x**4
Coefficients [a.u.]	2.073e-01	-6.700e-02	9.591e-03	-5.250e-04

Ammonia_default

NH length

Terms	x**(-1)	x**(-2)	x**(-3)	x**(-4)	x**(-5)	x**(-6)
Coefficients [a.u.]	6.211e+00	-1.008e+01	9.520e-01	8.645e+00	3.708e+00	-8.460e+00

HNH cosine

Terms	x	x**2	x**3	x**4
Coefficients [a.u.]	2.346e-01	1.236e-01	9.423e-03	-4.831e-03

HNH span

Terms	x
Coefficients [a.u.]	1.279e-01

Ammonia_ext1

NH length

Terms	x**(-1)	x**(-2)	x**(-3)	x**(-4)	x**(-5)	x**(-6)
Coefficients [a.u.]	2.377e+00	-5.030e+00	1.080e+00	4.869e+00	1.927e+00	-4.703e+00

HNH cosine

Terms	x	x**2	x**3
Coefficients [a.u.]	1.708e-01	9.383e-02	6.072e-03

HNH span

Terms	x	x**2	x**3
Coefficients [a.u.]	-7.185e-02	3.773e-02	-3.334e-03

N(HHH) distance

Terms	x	x**2	x**3
Coefficients [a.u.]	8.295e-02	-9.761e-02	1.603e-02

(N(HHH) distance)(NH length)*

Terms	x	x**2	x**3
Coefficients [a.u.]	-1.363e-02	6.292e-03	-1.287e-03

Ammonia_ext2

NH length

Terms	x**(-1)	x**(-2)	x**(-3)	x**(-4)	x**(-5)	x**(-6)
Parameters [a.u.]	-1.229e+01	1.473e+01	1.914e+00	-9.533e+00	-5.103e+00	9.321e+00

HNH cosine

Terms	x	x**2	x**3
Parameters [a.u.]	3.247e-01	9.382e-02	-9.637e-03

HNH span

Terms	x	x**2	x**3
Parameters [a.u.]	-3.897e-01	2.006e-02	7.639e-04

N(HHH) distance

Terms	x	x**2	x**3
Parameters [a.u.]	5.059e-02	-8.903e-02	-2.698e-03

(N(HHH) distance)(NH length)*

Terms	x	x**2	x**3
Parameters [a.u.]	-8.241e-03	4.386e-03	-4.104e-04

(HNH cosine)(NH length)*

Terms	x	x**2	x**3
Parameters [a.u.]	-1.257e-01	-9.516e-03	4.490e-04

(HNH span)(NH length)*

Terms	x	x**2	x**3
Parameters [a.u.]	6.782e-02	-6.137e-03	1.888e-04

Methane_default

CH length

Terms	x**(-1)	x**(-2)	x**(-3)	x**(-4)	x**(-5)	x**(-6)
Coefficients [a.u.]	2.595e+00	-5.826e+00	1.934e+00	5.521e+00	1.544e+00	-5.374e+00

HCH cosine

Terms	x	x**2
Coefficients [a.u.]	9.659e-02	8.328e-02

HCH span

Terms	x	x**2	x**3	x**4
Coefficients [a.u.]	-3.194e-03	-6.349e-03	4.911e-03	-5.515e-04

Model: methane_ext1

CH length

Terms	x**(-1)	x**(-2)	x**(-3)	x**(-4)	x**(-5)	x**(-6)
Coefficients [a.u.]	3.059e+00	-8.256e+00	3.076e+00	7.969e+00	1.925e+00	-8.233e+00

HCH cosine

Terms	x	x**2	x**3	x**4
Coefficients [a.u.]	-6.054e-02	-3.639e-01	-6.523e-01	-3.826e-01

HCH span

Terms	x	x**2	x**3	x**4
Coefficients [a.u.]	-5.592e-01	1.978e-01	-2.784e-02	1.300e-03

(HCH cosine)(CH length)*

Terms	x	x**2	x**3	x**4
Coefficients [a.u.]	2.841e-02	4.783e-02	3.399e-02	9.756e-03

(HCH span)(CH length)*

Terms	x	x**2	x**3	x**4
Coefficients [a.u.]	2.688e-02	-5.327e-03	4.142e-04	-1.130e-05

Model: methane_ext2

CH length

Terms	x**(-1)	x**(-2)	x**(-3)	x**(-4)	x**(-5)	x**(-6)
Coefficients [a.u.]	1.923e+00	-6.527e+00	3.012e+00	6.781e+00	1.461e+00	-7.140e+00

HCH cosine

Terms	x	x**2	x**3
Coefficients [a.u.]	1.996e-01	-7.029e-02	-1.052e-01

HCH span

Terms	x	x**2	x**3	x**4
Coefficients [a.u.]	-1.752e+00	5.895e-01	-8.437e-02	4.471e-03

(HCH cosine)(CH length)*

Terms	x	x**2	x**3	x**4
Coefficients [a.u.]	1.800e-03	3.992e-02	2.823e-02	8.156e-03

(HCH span)(CH length)*

Terms	x	x**2	x**3	x**4
Coefficients [a.u.]	2.601e-01	-3.780e-02	2.573e-03	-6.866e-05

(CH length)(CH length)*

Terms	x	x**2	x**3	x**4
Coefficients [a.u.]	-5.264e-01	1.114e-01	-1.181e-02	5.066e-04

3) Overview of additional extended models

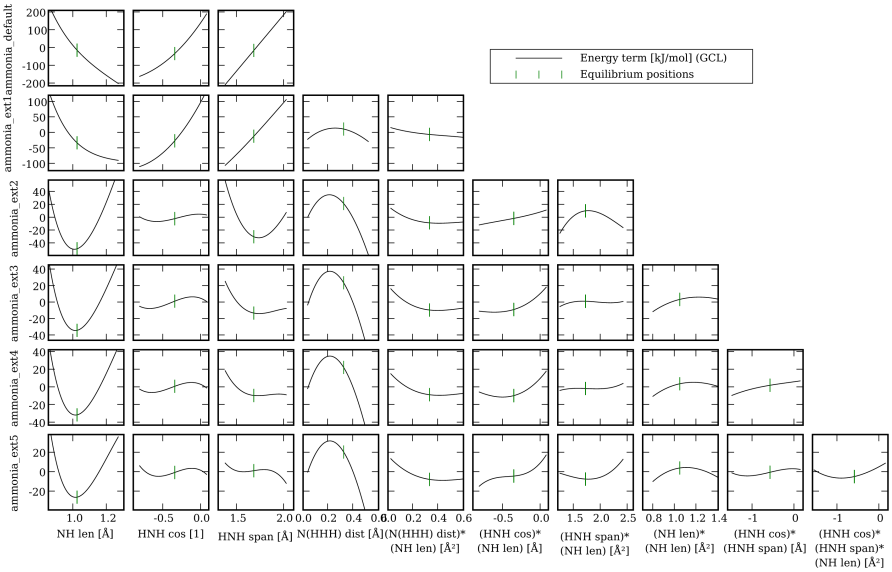


Figure 1: Overview of the energy terms for additional extended ammonia models parametrized with GCL.

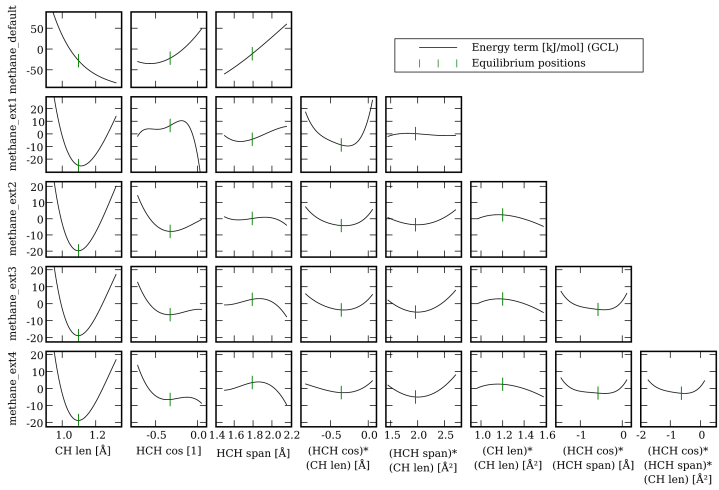


Figure 2: Overview of the energy terms for additional extended methane models parametrized with GCL.

Paper 4: "The Electronegativity Equalization Method and the Split Charge Equilibration Applied to Organic Systems: Parameterization, Validation and Comparison"

Toon Verstraelen, Veronique Van Speybroeck, Michel Waroquier

In Preparation

The Electronegativity Equalization Method and the Split Charge Equilibration Applied to Organic Systems: Parameterization, Validation and Comparison

Toon Verstraelen,[1,2] Veronique Van Speybroeck[1,2,a)] and Michel Waroquier[1,2]
[1] *Center for Molecular Modeling (CMM), Ghent University, 9000 Ghent, Belgium.*
[2] *QCMM - alliance Ghent-Brussels, Belgium.*

An extensive benchmark of the Electronegativity Equalization Method (EEM) and the Split Charge Equilibration (SQE) model on a very diverse set of organic molecules is presented. These models efficiently compute atomic partial charges, and are used in the development of polarizable force fields. The predicted partial charges depend on empirical parameters are calibrated to reproduce results from quantum mechanical calculations. Recently, SQE is presented as an extension of the EEM to obtain the correct size-dependence of the molecular polarizability. In this work, 12 parameterization protocols are applied to each model and the optimal parameters are benchmarked systematically. The training data for the empirical parameters comprises MP2/Aug-CC-pVDZ calculations on 500 organic molecules containing the elements H, C, N, O, F, S, Cl and Br. These molecules have been selected by an ingenious and autonomous protocol from an initial set of almost 500000 small organic molecules. It is clear that the SQE model outperforms the EEM in all benchmark assessments. When using Hirshfeld-I charges for the calibration, the SQE model optimally reproduces the molecular electrostatic potential from the ab initio calculations. Applications on chain molecules, i.e. alkanes, alkenes and alpha alanine helices, confirm that the EEM gives rise to a divergent behavior for the polarizability, while the SQE model shows the correct trends. We conclude that the SQE model is an essential component of a polarizable force field, showing several advantages over the original EEM.

I. Introduction

Molecular simulation is a very powerful toolbox in modern molecular modeling, and enables us to follow and understand structure and dynamics with extreme detail – literally on scales where motion of individual atoms can be tracked. The applications of computational chemistry are steadily growing towards larger molecular systems, reaching dimensions of several orders of the nanoscale. This evolution is not only due to steady advances in computer power, but also due to the continuous development of numerical and theoretical algorithms. Biomolecules are typical examples of systems containing a massive number of atoms (>10000). Molecular modeling of new materials is a second class of applications with vast system sizes.[1,2] The study of such large systems is only feasible with classical molecular mechanics using force fields (FF). These are analytical functions of the energy in terms of the atomic coordinates and depend on a set of parameters which are invariant during the course of the simulations. Particularly the non-bonding part of the force field is difficult to determine. Electrostatic fields dominate the long-range interactions between atoms, and hence play a crucial role in many biological processes such as protein folding, ligand docking, transport of ions across membranes, and so on.[3]

a) Corresponding author. E-mail: Veronique.VanSpeybroeck@UGent.be

The calculation of electrostatic interaction energies is often determined by “fixed” (invariant) atom-centered monopoles. The invariant nature of this representation prevents the response to an external electrostatic field, that is electronic polarization.[4] Polarization effects are known to yield an additional attractive intermolecular force. The error in the intermolecular interactions due to the absence of polarization effects can be compensated with an overestimate of the dispersion interaction and atomic partial charges, which must be fitted to experimental data. This compensation of errors is one of the reasons why non-bonding parameters for non-polarizable force fields obtained from high-level ab initio calculations on dimers in the gas phase are not simply transferable to condensed phase systems.[5-7] An explicit treatment of polarizability is indispensable for the transferability of force field parameters from the ab initio gas-phase reference data to large-scale biochemical simulations. The importance of this research field was highlighted in a special issue of the Journal of Chemical Theory and Computation in 2007,[8] discussing the current status of polarizable force fields.[9,10]

The oldest models to describe molecular polarization treat atoms as inducible dipoles.[11,12] The theoretical foundations to model interatomic charge transfer were first established by Mortier et al.,[13-16]who developed the Electronegativity Equalization Method (EEM) to predict the molecular charge distribution efficiently. The low computational cost of the EEM was the onset for the inclusion in molecular mechanics models to obtain a more accurate prediction of the electrostatic interactions.[17,18] In the late nineties, both concepts of inducible atomic dipoles and interatomic charge transfer were combined to model the linear response properties of small molecules.[5,19] A myriad of variations on the original EEM have been proposed during the past two decades under different names, including Charge Equilibration (QEq)[17], Chemical Potential Equalization (CPE or μEq)[20], Fluctuating Charges (FQ)[5,21]. Polarization energy terms in popular biochemical force fields such as CHARMM[22-24], AMBER[25,26] and OPLS[27,28] are currently under active development.

Although excellent parametrizations have been proposed in literature[5,19,23,24,26,28-40], the transferability towards larger systems is limited. Chelli et al.[41-43] have shown that several models based on the EEM predict an incorrect dependence of the polarizability as a function of the molecular size, i.e. an exaggerated overestimation of the polarizability for large molecules. Ab initio calculations and experimental results for n-alkanes show that the molecular polarizability should increase linearly with molecular size, while EEM predicts a cubic dependence.[44] An ad hoc remedy is the implementation of the EEM using the Constrained Charge Approximation (CCA)[45], which consists of a partitioning of the molecule into multiple subsystems (ensembles of atoms of the molecule) where net charge is constrained to a fixed value and charge transfer among subsystems is not allowed. In reference 43 Chelli et al. show that the CCA method can solve the polarizability catastrophe only in some cases. For aldehyde, nitro- and carboxylic acid series the problem remains unsolved.

Another solution to the polarizability problem is the Atom-Atom Charge Transfer (AACT) model proposed by Chelli et al.[41] The total energy in the AACT model is a second order expansion in terms of charges transfered between atoms. The total net charge of an atom is the sum of the charges transfered to that atom. Direct interatomic charge transfer can be limited to atom pairs that are covalently bonded, but also other atom pairs may be included. However, it is mandatory that charge transfer between very distant atoms is only possible as a superposition of local charge transfers. Recently Nistor and co-workers have introduced the Split Charge Equilibration (SQE), a generalization of the charge equilibration method for non-metallic materials.[46] The term "split charge" is a synonym for "atom atom charge transfer". The SQE model essentially combines the EEM and AACT model into one consistent scheme. As a consequence, the SQE model can describe polarization effects both in the metallic limit (EEM) and the dielectric limit (AACT). Warren and

coworkers have recently discussed the origin and control of the superlinear scaling of the polarizability in the EEM, and also benchmark all solutions in the literature on an abstract chain molecule.[44]

In this study parameters are estimated for the original EEM and the SQE model, which are applicable to a wide range of molecules. Twelve parameter calibration protocols are tested on each model and the performance of the optimal parameters is compared systematically.

The training data consists of MP2/Aug-CC-pVDZ calculations on a set of 500 organic molecules comprising the following elements: H, C, N, O, F, S, Cl and Br. The selection of the 500 molecules in the training set is carried out with an ingenious autonomous algorithm that constructs a representative subset of 431980 small molecules taken from the PubChem database.[47] Ab initio reference data is not only computed for the electronic ground states, but we also take into account the perturbation by point charges randomly positioned in the vicinity of the molecule. The latter is essential to obtain parameters that properly describe the linear response to an electrostatic perturbation.[5]

Reference values for the atomic partial charges are required for the parameterization. We benchmarked three population schemes: Mulliken charges[48], Natural charges[49] and Hirshfeld-I charges[50]. ESP (Electrostatic Potential) fitted charges are not included in this work due to their inherent statistical inaccuracy.[51] Several performance measures are used to compare the optimal parameters from different calibrations. Both the equilibrium partial charges and the changes in atomic charges due to a perturbation are compared with the ab initio data. We have also evaluated the ab initio equilibrium electrostatic potential and the changes in the potential due to a perturbation on a molecular grid, to assess the performance of each model when used for the computation of electrostatic interactions. Finally the change in ab initio energy due to a perturbation is also compared with the value predicted by the empirical models.

The major limitation of the charge transfer models in this work is their inability to describe polarization orthogonal to a chemical bond or polarization orthogonal to a planar molecule. One can introduce an inducible point dipole on each atom to surmount this limitation.[5,19] An extension of the models with inducible atomic dipoles is not considered in this paper to reduce the complexity of the parametrization procedure.

The remainder of this paper is structured as follows. In the next section we discuss briefly the EEM and SQE models, followed by a detailed account on the selection of molecules for the training set, and the parameterization and validation protocols used in this work. The third section discusses the quality of the different parameterizations with a variety of benchmarks. We end the third section with applications of the models on chain molecules. Conclusions are given in section 4.

II. Methods

A. EEM and SQE Model

The original Electronegativity Equalization Method (EEM)[14] is a second order expansion of the molecular energy in terms of partial charges:

$$E_{\mathrm{EEM}} = \sum_{i=1}^{N} \left[\chi_i q_i + \frac{1}{2}\eta_i q_i^2 + \left(\sum_{j=i+1}^{N} q_i q_j J_{ij} \right) + q_i J_{\mathrm{ext},i} \right]$$

The variables q_i are the atomic partial charges, χ_i and η_i are empirical parameters and J_{ij} represents the Coulomb interaction between two atoms i and j with unit charge. The last term represents the interaction energy due to an external electrostatic field. In this work, the external field is always generated by a single point charge, while atoms are treated as Gaussian charge distributions. The charge density of atom i is given by

$$\rho_i(\bar{r}) = q_i \left(R_i^2\pi\right)^{-\frac{3}{2}} \exp\left(-\frac{|\bar{r}-\bar{r}_i|^2}{R_i^2}\right)$$

where R_i is the width of the distribution and $\bar{r}_i$ is the Cartesian coordinate of atom i. In this case, the electrostatic interaction becomes

$$J_{ij} = \frac{1}{|\bar{r}_i - \bar{r}_j|}\mathrm{erf}\left(\frac{|\bar{r}_i - \bar{r}_j|}{\sqrt{R_i^2 + R_j^2}}\right)$$

and

$$J_{\mathrm{ext},i} = \frac{q_{\mathrm{ext}}}{|\bar{r}_i - \bar{r}_{\mathrm{ext}}|}\mathrm{erf}\left(\frac{|\bar{r}_i - \bar{r}_{\mathrm{ext}}|}{R_i}\right),$$

whe $\bar{r}_{\mathrm{ext}}$ is the position of the probe charge and q_{ext} is its amplitude. Gaussian charge distributions offer two advantages over conventional point charges: a Gaussian distribution is a more realistic model of an atom than a point charge, and the finite self-interaction energy of a Gaussian charge distribution facilitates the derivation of empirical parameters. (vide infra)

Recently, Nistor and coworkers have developed the Split Charge Equilibration scheme (SQE), which is an extension of the EEM method with additional energy terms that lead to a correct size dependence of the electronic polarizability for linear alkanes.[44,46] In addition to the traditional concept of partial charges, the SQE method also introduces so-called split charges, or charge transfers p_{ji}. They stand for the charge transferred from atom j to atom i and obey the following conditions:

$$q_i = \frac{q_{\mathrm{tot}}}{N} + \sum_{j=1}^{N} p_{ji} \quad \text{and} \quad p_{ji} = -p_{ij}$$

where q_{tot} is the total charge of the molecule and N is the number of atoms. In the present paper, direct charge transfer is only allowed between covalently bonded atoms. All other p_{ij} values are assumed to be zero. Charge transfer between more distant atoms must be a superposition of local charge transfers. Direct charge transfer through a hydrogen bond is not considered, because this is only a minor effect. The transformation from charge transfer variables to partial charges is trivial, but the inverse transformation is not always uniquely defined. In this work, the Moore-Penrose pseudo-inverse is used, in analogy with the work of Chen et al.[52,53] The following variant of the SQE scheme is used:

$$E_{\mathrm{SQE}} = E_{\mathrm{EEM}} + \sum_{i,j}^{\mathrm{bonded}} \left(\frac{1}{2}\zeta_{ij}p_{ij}^2 + \Delta\chi_{ij}(q_i - q_j)\right)$$

The sum only considers atoms i and j that are connected by a covalent bond. The first term associates a quadratic energy term with each charge transfer variable, where the bond hardness parameter ζ_{ij} represents the resistance against charge transfer through a chemical bond. It is an empirical parameter that depends on the type of the bond between atoms i and j, and on the types

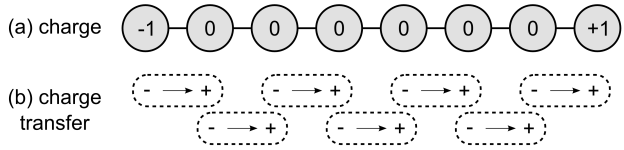


FIG. 1. A comparison of the schematic representation of a polarized molecule in terms atomic partial charges and in terms of charge transfer between bonded atoms.

of atom i and j. The second term introduces an perturbative bond-correction, which takes into account the effect of the direct chemical environment on the atomic electronegativity parameter. This correction parameter is antisymmetric: $\Delta\chi_{ij} = -\Delta\chi_{ji}$ and is zero when atoms i and j are of the same type. Nistor and coworkers proposed a more general set of perturbative corrections without imposing symmetry conditions, both for the atomic hardness and electronegativity parameter.[46] In this work only the antisymmetric electronegativity correction is present to reduce the number of empirical parameters. We expect that this correction is the most relevant: charge population schemes differ systematically in the way charge density between covalently bonded atoms is divided between the two atoms.[48] In this work population charges are used as reference data and the corrective term is introduced to capture the specific behavior of the different population schemes.

In order to stress the importance of the bond hardness term, we will illustrate its effect with a schematic example. Consider a linear chain of N identical atoms subject to an electrostatic field parallel to the chain. Figure 1 depicts two representations of the charge distribution in this linear molecule: (a) the atomic charge representation where the two most distant atoms carry an opposite charge due to an electrostatic field and (b) the charge transfer representation of the same molecule with the same charge distribution. In the limit of long chains, the Coulomb interaction between the two end points becomes negligible and we only have to consider the remaining energy terms. The linear energy terms can also be omitted since all atoms in the chain are of the same type. The antisymmetric corrections are zero and the atomic electronegativity parameter only affects the final equalized electronegativity, but not the charge distribution. Only the remaining quadratic energy terms, that are the atomic hardness and bond hardness terms, are relevant for this example. In the case of the original EEM formalism, the energy required to transfer one electron between the two endpoints does not depend on the chain length, which corresponds to the behavior of a metal. This approach is not suitable to describe molecular systems that behave like a dielectric (organic molecules, ceramic materials, ...), and the polarizability will be seriously overestimated in the limit of long chains. In the SQE formalism, one can only transfer a charge between the two end-points if there is an equal charge transfer on each bond. Each bond acts as an inducible dipole, which corresponds to the classical description of a dielectric. It is therefore expected that the SQE formalism will give a more reasonable picture of the polarizability, also for systems that mainly contain covalent bonds.

The hardness kernel, $\bar{\bar{\eta}}$, is an important concept in both models.[19] In this context the hardness kernel is defined as the matrix of second order derivatives of the total energy towards the atomic partial charges:

$$\eta_{ij} = \frac{\partial^2 E_{\mathrm{model}}}{\partial q_i \partial q_j}.$$

A set of empirical parameters is unphysical when it leads to negative eigenvalues in the hardness kernel for a certain molecule. When such a molecule can exchange electrons with a soft charge bath, the energy minimum is not bound. This issue must be taken into account in the parametrization procedure and will be discussed in detail below. A second important quantity is the polarizability tensor, $\bar{\bar{\alpha}}$. It describes the linear relation between a uniform external electrostatic field, $\bar{E}_{\mathrm{ext}}$, and the induced molecular dipole moment, $\bar{\mu}$:

$$\mu_i = \sum_{j=x,y,z} \alpha_{ij} E_{\mathrm{ext},j} \quad \text{for} \quad i = x, y, z.$$

An expression for the polarizability tensor can be derived from the hardness kernel. A detailed derivation is given in the work of York et al.[19]

B. Atom types

The empirical parameters in the EEM and SQE models are associated with atoms (χ_i and η_i) or bonds (ζ_{ij} and $\Delta\chi_{ij}$). We assign different parameters to each atom type or bond type. For large scale applications, it is crucial that atom types and bond types can be identified without a significant computational cost. In this work we propose two categories of atom and bond types.

The first approach to assign atom types is trivial: each element in the periodic system is considered as an atom type, disregarding the electronic behavior of the atom in its molecular environment. Each bond type is identified by the atom types of the two bonded atoms. We will refer to this scheme in the remainder of the text as "trivial atom types". There are in total 8 trivial atom types (H, C, N, O, F, S, Cl and Br) and 19 trivial bond types, listed in part I of the Supporting Information. Only the most relevant bond types, i.e. not all the theoretically possible combinations of atom types, are considered. The actual selection of bond types is determined by the molecules in the training set. (vide infra) This simple approach has an obvious disadvantage. It fails to distinguish between σ-bonds, π-bonds or conjugated systems, while the nature of bonds has a large impact on the molecular polarizability. For example, charge can displaced more easily in systems with delocalized orbitals.[41]

Ideally one should first compute the bond order of each covalent bond to assign proper bond types. This becomes a computationally demanding task for large molecular systems (>10000 atoms), which is in contrast to the final purpose of our work: the development of fast polarizable force field models. Therefore we propose a second scheme to assign atomic types and that can be regarded as a compromise between efficiency and accuracy. Each atom type consists of the atom symbol and its valence, that is the number of covalent bonds with neighboring atoms. Each bond type is again identified by the two atom types it connects. This approach will be referred to as "force-field atom types". There are 15 force-field atom types (H, C4, C3, C2, N4, N3, N2, N1, O2, O1, F, S2, S1, Cl, Br) and 56 force-field bond types, see part I of the supporting information. An sp3 carbon has atom type 'C4' and a sigma bond between two sp3 carbons is labeled with 'C4-C4'. Although this method will not always make a proper distinction between all hybridization and oxidation states, it works well for neutral carbon atoms and should therefore be a good compromise to describe organic molecules.

C. Training data

In order to calibrate all the empirical parameters, one needs a training data set of molecules for which the atomic partial charges are known from first principles following some properly selected population schemes. The selection of molecules depends on the type of modeling applications. For

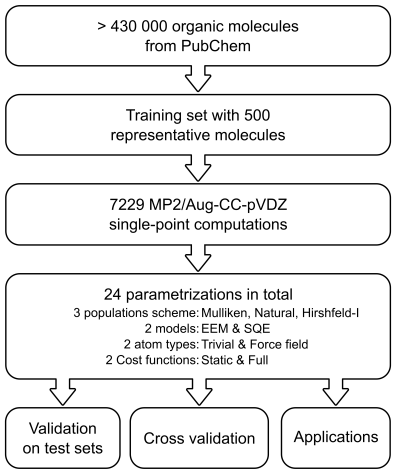


FIG. 2. Flowchart of the complete parametrization protocol, including the generation of training data and the validation of the models.

this work, we focus on a broad range of organic molecules to make a thorough comparison of the feasibility and performance of both the EEM and SQE model in biochemical applications. An overview of the entire parameterization and validation flow chart is given in figure 2.

Our first objective is the inception of an algorithm that generates a training set of 500 random molecules that are representative for a broad range of organic systems. This algorithm should be automatic, efficient, transparent, unbiased and should eventually take into account some specific constraints. We first download from the PubChem[47] database all molecules with no more than 12 non-hydrogen atoms and at least one hydrogen and one carbon atom. This selection leads to a data set of 1329823 unique molecular graphs stored in an SQLite[54] database. Histograms are constructed displaying the occurrence of each force-field atom and bond type. We then reduce the database by selecting only those molecules containing the 15 most prevalent force-field atom types and the 56 most prevalent force-field bond types (listed in part I of the Supporting Information). A further reduction is achieved by keeping only molecules that fulfill the following conditions:

(i) at least three non-hydrogen atoms, (ii) at least 60% of the atoms in the molecule should be hydrogen or carbon atoms, (iii) no more than two halogens (limited to fluorine, chlorine and bromine), (iv) no more than two sulfurs, (v) at most three nitrogens, (vi) at most four oxygens, (vii) the number of electrons should be even with a maximum of 80, and (viii) no more than one positively and/or negatively charged functional group.

These constraints are a good compromise between feasibility and representability, but still lead to 431980 molecules. It is not feasible, nor useful to perform ab initio calculations on all of them. We therefore developed an iterative algorithm to finally select a training set of 500 molecules from the large data set of 431980 molecules.

At each iteration, a new molecule is chosen randomly from the large set and is only added to the training set if it fulfills the conditions discussed below. These conditions guarantee the diversity of

the training set and are based on the number occurrences of each force-field atom and bond type in the training set. The first goal is to obtain 20 occurrences of each atom and bond type. Therefore a new molecule is only accepted when it increases the counters of the atom and bond types that are below 20. Once all atom and bond types occur at least twenty times in the training set, the new molecule is only accepted if the Shannon entropy of the histograms with atom types and bond types increases. This procedure is repeated until 500 molecules have been selected.

The algorithm is then continued applying the same rules but now a random molecule is first removed before the new molecule is taken up. After two million iterations, this procedure results into a diverse and representative training set of organic molecules, covering a wide range of functional groups. Each atom and bond type will occur at least twenty times, which guarantees a sufficient amount of reference data for each parameter in the model. Figure 3 depicts the final histogram with atom and bond types. The list of molecules is given in part II of the Supporting Information.

Equilibrium geometries are determined for all 500 molecules with a cascade of geometry optimizations, starting with a graph-based geometry optimization[55,56], followed by a geometry optimization with a force field, then by a geometry optimization at the HF/3-21G level, and fine-tuned with a geometry optimization at the MP2/CC-pVDZ level.[57-59] The quantum mechanical calculations were carried out with Gaussian03.[60]

For each geometry, multiple single point calculations were carried out at the MP2/Aug-CC-pVDZ level. The first calculation is performed on the reference state, which is sometimes an ion, depending on the functional groups present in the molecule. For example, when the molecule contains a reduced carboxyl group, the total charge of the reference state is -1 e. Afterwards a computation is carried out with one additional electron and with one electron less, which allows us to compute the Mulliken electronegativity and hardness. Finally, additional single point calculations are performed with distinct external electrostatic fields to probe the linear response function. For a given molecule with N atoms, N+1 perturbations are considered, each time generated by a point charge placed randomly at two times the Van Der Waals distance of the closest atom. The probe charge is limited to +0.5 e or -0.5 e, to reduce hyperpolarization effects. Calculations with very similar positions of the probe charges were avoided, by selecting the probe coordinates from a larger set with the Kennard-Stone algorithm.[61]

For each single point calculation, several population schemes were used to obtain atomic partial charges: Mulliken charges[48], Natural charges[49] and Hirshfeld-I charges[50]. We do not rely on charges that are fitted to reproduce the electrostatic potential around the molecule because they generally suffer from statistical inaccuracies.[51] This does not mean that the electrostatic potential around the molecule is an irrelevant quantity. For the development of a force-field model, one is in principle only interested in the reproduction of the electrostatic potential generated by the ab initio density. Only the latter will lead to correct electrostatic interactions. We evaluated, for each single point calculation, the ab initio electrostatic potential on a molecular grid to benchmark the performance of each parameterization. A two-dimensional schematic picture of the grid is given in figure 4. It is constructed as follows. First, 30 concentric spheres are placed around each atom. The minimum sphere radius is 1.5 times the radius of the noble gas core of the corresponding atom, the maximum radius is 30 times the cusp radius. The radii of intermediate spheres are equidistant on a logarithmic scale. On each sphere, we used randomly rotated 50-point Lebedev-Laikov grids.[62] (The random rotation avoids arbitrary preferred directions.) For this study, we only retained the grid points where the electron density is lower than 10^{-5} atomic units.

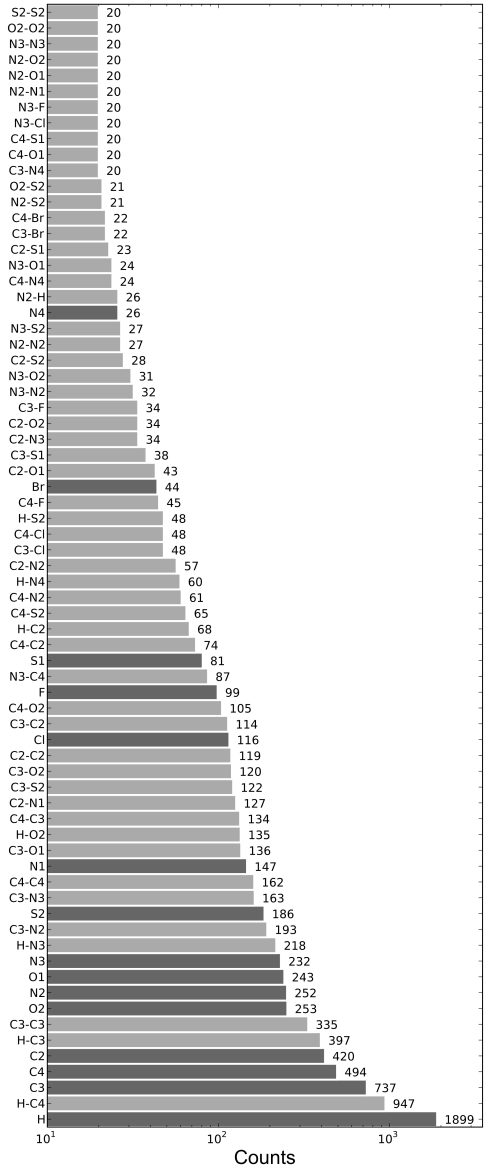


FIG. 3. Histogram of the force-field atom and bond types in the training set with 500 molecules. Bars corresponding to atom and bond types are in dark and light gray, respectively.

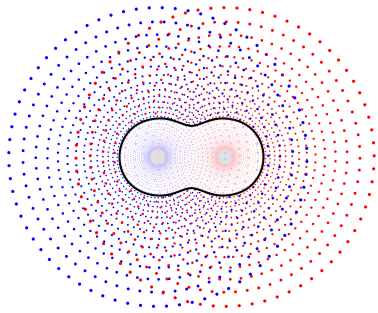


FIG. 4. Schematic representation of the grid points for a diatomic molecule used for the electrostatic potential benchmark. Grid points are constructed atom-centered spheres using Lebedev-Laikov grids.[66] Each spherical grid is rotated randomly to avoid arbitrary preferred directions in the benchmark. Grid points in the region where the density is larger than 10-5 atomic units, are excluded.

D. Parameterization

The empirical parameters in the EEM and SQE models must be calibrated to reproduce the training data. In practice, one defines an objective function or cost function that expresses the overall error between the charges from the empirical model and the ab initio training data for a given set of parameters. Consequently, a multidimensional minimization algorithm is applied to find the parameters that minimize the objective function. In this work, a conjugate gradient algorithm was used, with analytically computed gradients. Two types of cost functions have been examined. The first one is often used in foregoing studies:[34-36,46]

$$\mathrm{COST}_{\mathrm{static}}(\mathrm{parameters}) = \sum_{m=1}^{500} \sum_{i=1}^{N_m} w_{mi} \left(q_{mi}^{\mathrm{AI}} - q_{mi}^{\mathrm{MODEL}}(\mathrm{parameters})\right)^2$$

The first sum runs over all molecules in the training set and the second sum runs over each atom in molecule m. The constant q_{mi}^{AI} stands for a population charge of atom i in molecule m, as derived from the ab initio calculation on the reference state of molecule m. (This does not include charges from ab initio calculations with an external electrostatic field.) Similarly q_{mi}^{MODEL}, stands for the charge of the corresponding atom in an empirical model. The factor w_{mi} is a constant that determines the contribution of atom i in molecule m to the total cost. This is a generic description of the cost function and there are in practice 12 distinct ways to use it: each combination of population scheme (Mulliken, Natural or Hirshfeld-I), empirical model (EEM or SQE) and choice of atom and bond types (trivial or force-field) leads to a different parametrization.

There are several approaches to the weight constants in the cost function. When all weights $w_{m,i}$ are set to one, each atomic charge has an equal contribution to the cost function. Although this seems to be a democratic choice, the empirical parameters that minimize the cost function will be ill-defined. There are much more atoms with force-field type C4 compared to N4. In case of equal weights, the cost function will become relatively insensitive to errors in the charges of N4 atoms. We tested weights that are inversely proportional to the prevalence of the corresponding atom types, but this always leads to models that perform poor on the most prevalent atom types. Therefore we decided to use a compromise and set each weight to the inverse of the square root of the prevalence of the

corresponding atom type. This is the geometric mean of the two possibilities. An arithmetic mean would be of little use since most weights would then approximate 0.5.

The second cost function also includes data from the ab initio calculations with an external electrostatic field. It is a linear combination of the static cost function and a new term:

$$\mathrm{COST}_{\mathrm{full}}(\mathrm{parameters}) = \mathrm{COST}_{\mathrm{response}}(\mathrm{parameters}) + \lambda \mathrm{COST}_{\mathrm{static}}(\mathrm{parameters})$$

$$\mathrm{COST}_{\mathrm{response}}(\mathrm{parameters}) = \sum_{m=1}^{500} \sum_{p=1}^{P_m} \sum_{i=1}^{N_m} w_{mi} \left(\Delta q_{mpi}^{\mathrm{AI}} - \Delta q_{mpi}^{\mathrm{MODEL}}(\mathrm{parameters}) \right)^2$$

The first sum runs over all molecules in the training set, the second runs over each calculation on molecule m with a different external electrostatic field, and the third runs over each atom in molecule m. The constant $\Delta q_{mpi}^{\mathrm{AI}}$ stands for the difference in the population charge of atom i in molecule m between the reference and the perturbed state. $\Delta q_{mpi}^{\mathrm{MODEL}}$ is the same quantity predicted by the empirical model. Note that the response cost function measures the quality of the linear electrostatic response properties of a molecule in an empirical model. This part of the total cost function is only determined by parameters in the second order terms of the EEM or SQE model. The positive coefficient λ determines the relative importance of the static and the response part of the cost function. We propose to following procedure to determine this auxiliary parameter. At first instance λ is set to zero and we only optimize the parameters associated with second order terms. Consequently we perform an optimization of all parameters with the full cost function, with a non-zero value for lambda. The latter is tuned in such a way that the contribution of the response cost only increases by one percent, compared to the case where λ is zero. This approach has several advantages: (i) λ is obtained in a reproducible way, (ii) all the parameters calibrated with global final optimization instead of a piecewise optimization and (iii) the linear response cost is still very close to its minimum while the second order parameters are still allowed to vary. In analogy to the static cost function, we apply this full cost function in twelve different ways by considering all combinations of population schemes, atom types and empirical models.

If the parameters are optimized with the foregoing cost functions, there is no guarantee that the hardness kernel of the EEM or SQE model is positive definite for all molecules in the training set and all other molecules on which these models will applied. It is simply impossible to test the eigenvalues of all possible molecules and we must put forward general rules to avoid unphysical parameters. We experienced that it was not sufficient to test only the hardness kernels of the molecules in the training set. Test applications on proteins (> 6000 atoms) revealed in some cases an unbound minimum of the total energy.

One can define generic constraints for the parameters in the second order terms that guarentee a positive definite hardness kernel in all cases. We will first define constraints for the EEM case, and then extend them for the SQE model. Because the self-interaction energy of any charge distribution is always positive, we require that the diagonal elements of the hardness kernel are at least equal to the self-interaction of the corresponding Gaussian charge distributions:

$$\eta_i >= \frac{1}{\pi\sqrt{2}R_i} \quad \forall\ i$$

When the hardness parameters, η_i, are equal to their lower limits, the second order term of the EEM model expresses the electrostatic self energy of the sum of gaussian charge distributions representing the atoms. Higher values for the parameters η_i only make the hardness kernel more

positive definite. In the SQE model, the same condition is applied to all the atomic hardness parameters, and additionally all bond hardness parameters have to be positive. The latter guarantees that the quadratic bond hardness term is also positive definite. The constraints were imposed with a standard technique in the field of convex programming.[64] We added a logarithmic penalty term to the cost function for each constrained parameter x_c with minimum value $x_c^{(min)}$:

$$-K\left(\ln(x_c - x_c^{(min)}) - (x_c - x_c^{(min)})\right).$$

The logarithmic term imposes the lower bound for x_c and the linear term prevents constrained parameters from drifting to infinity during the optimization. First an optimization with a relatively high K value is carried out. This is followed by several iterations in which the strength K is divided by two and the parameters are optimized again. The latter is repeated until the contribution from the logarithmic penalties to the total cost function becomes negligible.

A proper choice for the atomic radii must take into account several different considerations. In principle, these radii should correspond to realistic data such as covalent radii or Van Der Waals radii. However, when the radii are very small, the minimum value for the atomic hardness parameters becomes too high and it is impossible to find reasonable empirical parameters. When the atomic radii become too large, the Coulomb interaction between nearby atoms is underestimated which leads to a poor performance of the empirical model. Ideally these radii are also considered as free parameters during the optimization procedure, but for this work we prefer to use fixed values that lead to a reasonable compromise between all the considerations discussed above. Recent covalent radii based on XRD data[63] are used as a starting point. For the hydrogen atom, three times the covalent bond radius was used, while the other covalent radii are multiplied by 1.5.

After each optimization, the Hessian of the cost function is computed in terms of all the empirical parameters, based on finite differentiation of analytically computed gradients. This matrix would be proportional to the covariance matrix of the parameters if each population charge had a statistical error inverse proportional to the square root of the corresponding weight in the cost function. Unfortunately, there is no measurement error in our reference data and we can not compute a true covariance matrix. The condition number of the Hessian is a measure for the stability of the parameters. Relatively low eigenmodes of the Hessian represent directions in the parameter hyper space along which the cost function is almost constant. The exact position of the optimal parameters along these directions is ill-defined. A high condition number reveals the presence of relatively low eigenvalues. Very small negative eigenvalues correspond to nearly zero eigenvalues plus an error due to the finite differences used to compute the Hessian. In such cases, the actual condition number is nearly singular.

Finally, we propose an alternative solution to fix the reference value for the electronegativity parameters. Usually, this is solved by arbitrarily fixing the electronegativity parameter of hydrogen (or any other atom type) to zero.[34-36] Nevertheless, we prefer to follow a different approach that introduces a physical absolute reference value for the atomic electronegativity parameters.

For the evaluation of the static cost function, we don't apply a strict charge constraint, but introduce an extended Lagrange multiplier instead:

$$E'_{\mathrm{MODEL}} = E_{\mathrm{MODEL}} + \chi_{\mathrm{bath}} q_{\mathrm{tot}} + \frac{1}{2}\eta_{\mathrm{bat}} q_{\mathrm{tot}}^2$$

The physical interpretation is that we allow the molecule to exchange charge with a surrounding bath with electronegativity χ_{bath} and hardness η_{bath}. For each molecule, the electronegativity of the charge bath is set to the Mulliken electronegativity derived from the ab initio calculations. The

TABLE I. The 24 acronyms that uniquely identify each parameterization in this work. Each acronynom consist of four capitals. The first one refers to the charge population scheme. The second stands for the empirical model. The third character is used for the atom type. The fourth one refers to the cost function.

Model	Atom types	Cost fn.	Mulliken	Natural	Hirshfeld-I
EEM	Trivial	Static	METS	NETS	HETS
EEM	Trivial	Full	METF	NETF	HETF
EEM	Force Field	Static	MEFS	NEFS	HEFS
EEM	Force Field	Full	MEFF	NEFF	HEFF
SQE	Trivial	Static	MSTS	NSTS	HSTS
SQE	Trivial	Full	MSTF	NSTF	HSTF
SQE	Force Field	Static	MSFS	NSFS	HSFS
SQE	Force Field	Full	MSFF	NSFF	HSFF

empirical parameters must lead to the correct molecular electronegativity, otherwise the molecule exchanges charge with the bath. The static cost function will penalize an incorrect total charge, and hence reasonable absolute values of the electronegativity parameters are enforced.

The hardness parameter of the bath controls the relative importance of this effect during the optimization procedure. In the limit of the bath hardness towards infinity, the extended Lagrange multiplier behaves like a conventional constraint, which results in a poor estimate of the absolute value of the atomic electronegativity parameters. When the bath hardness is set to zero, the empirical parameters will be biased to reproduce the Mulliken electronegativities of the 500 molecules in the training set at the cost of a good reproduction of atomic partial charges. As a compromise, the hardness of the bath term was set to 5 V e^{-1}.

E. Performance and Validation

We benchmarked the performance of 24 parametrizations by comparing different aspects of the empirical models with the original ab initio data. Each parameterization has a unique label as given in table I, which will be used for further reference. The benchmarks are performed on all molecules in the training set, and also on two test sets. The first test set contains all naturally occurring amino acids and the five nucleobases. The second test set consists of linear alkanes (up to eight carbon atoms) and linear conjugated alkenes (up to ten carbon atoms). Ab initio calculations for all the molecules in the test sets were carried out with the same protocol that was used for the training set.

We compared the reproduction of atomic partial charges, the electrostatic potential around the molecule and the change in partial charges, electrostatic potential and molecular energy due to the presence of an external electrostatic field. We also tested how well the charges from the three population schemes could reproduce the full ab initio electrostatic potential.

The performance of a model in the reproduction of property X is evaluated with the following standard performance measures:

$$\mathrm{RMSD} = \sqrt{\sum_{m=1}^{M}\sum_{k=1}^{K_m}\left(x_k^{\mathrm{AI}} - x_k^{\mathrm{MODEL}}\right)^2}$$

TABLE II. Overview of the performance measures used. The first column is a label for each performance, used for later reference. The second column is the specific property considered. The third column lists the corresponding summation index(es) used in the RMSD expressions.

	Property	Summation index(es)
$\mathbf{P}_{\mathrm{A}}$	Atomic partial charges in absence of an external electrostatic field	i: the atoms of molecule m
$\mathbf{P}_{\mathrm{B}}$	The change in atomic partial charges due to the external electrostatic field	p: the perturbed states, and i: the atoms of molecule m
$\mathbf{P}_{\mathrm{C}}$	Electrostatic potential in absence of an external electrostatic field	g: the grid points of molecule m
$\mathbf{P}_{\mathrm{D}}$	The change in the electrostatic potential generated by the molecule due to the external electrostatic field	p: the perturbed states, and g: the grid points of molecule m
$\mathbf{P}_{\mathrm{E}}$	The change in molecular energy due the external electrostatic field	p: the perturbed states

$$\text{Relative RMSD} = RMSD / \sqrt{\sum_{m=1}^{M} \sum_{k=1}^{K_m} \left(x_k^{\mathrm{MODEL}}\right)^2}$$

In these expressions, the first sum always runs over the molecules in the training or test set, while the second sum iterates over all indexes specific to property X in molecule m. Table II lists all properties that were benchmarked and also specifies the index used in the second summation of the performance measures.

In addition to the performance measures discussed above, we also conducted a cross validation test on one of the parameterizations. 100 representative subsets of the training set were constructed, each containing 400 molecules. Each subset is constrained to be representative, and comprises at least 16 occurrences of each atom or bond type. Parameters are estimated based on each subset and the performance measures are computed each time on the 100 remaining molecules. Finally the averages of each performance measure is computed over the 100 cross-validation runs. These cross-validation averages estimate the performance of the parameters on similar molecules that are not included in the training set.

F. Implementation

This entire benchmark study extensively relies on QFit, a computer program that has been especially developed for this work. QFit can easily handle all kinds of parameterization protocols and tremendously facilitates the comparison of different empirical models. The EEM and SQE forms are only two specific cases of the vast amount of possible models,[46] and all these variations are supported by QFit. The program is written in Python[65,66], and uses NumPy[67] for efficient computations and Matplotlib[68] for the visualization of the output. QFit also makes use of the MolMod library, a component of ZEOBUILDER.[69,70] The MolMod library provides data structures and algorithms to work with molecular graphs and geometries, to post-process Gaussian[60] calculations, to assign atom types, to construct initial geometries, to optimize empirical parameters, and so on. The software design of QFit is based on the object oriented formalism to guarantee the extensibility of the program. An extension for the parameterization of inducible atomic dipoles is foreseen in the near future. QFit is not available yet, but will be released jointly with MMFit, a

TABLE III. Overview of the eight models and the stability of the corresponding parameters. The number of empirical parameters depends only on the model and the atom types, not on the cost function. Optimizations with the static cost function often leads to negative eigenvalues in the Hessian of the cost function. The number of negative eigenvalues is given between brackets. The optimizations with the full cost function always results in positive definite Hessians.

Model	Atom types	Cost fn.	number of parameters	Mulliken	Natural	Hirshfeld-I
EEM	Trivial	Static	16	(1)	(3)	(3)
EEM	Trivial	Full	16	1.19E+3	8.09E+2	3.86E+2
EEM	Force Field	Static	30	(1)	(1)	(1)
EEM	Force Field	Full	30	1.88E+3	5.15E+3	2.01E+4
SQE	Trivial	Static	50	2.20E+5	(3)	(5)
SQE	Trivial	Full	50	3.15E+4	7.13E+3	3.62E+4
SQE	Force Field	Static	135	1.44E+7	8.05E+6	(4)
SQE	Force Field	Full	135	1.10E+5	2.09E+5	9.17E+4

computer program that implements the Gradient Curves Method.[71] The purpose of these two programs is the derivation of a complete force-field model from ab initio training data.

The computation of Hirshfeld-I charges and the construction of the molecular ESP grids is performed with HiPart, a computer program also developed in the context of this study. The source code of HiPart will be released in the near future. Just like QFit, HiPart is also written in Python[65] and uses NumPy[67] for number crunching.

III. Results and Discussion

A. Parametrization

In total 24 parameterizations (summarized in Table I) have been carried out and submitted to an in-depth investigation. For clarity each set of parameters is listed in part III of the Supporting Information. Before we discuss the benefits and the drawbacks of each case, it is interesting to discuss the robustness of the parameters in terms of the condition number of the Hessian of the cost function. A higher condition number is an indicator for less robust parameters, i.e. parameters that are more sensitive to irrelevant details in the training data. Table III lists the condition number of the Hessian and the number of parameters for each of the 24 parametrizations. When the condition number is nearly singular, the number of zero eigenvalues is given between parenthesis instead of the condition number.

There are two factors that have a major impact on the condition number: the number of parameters and the type of cost function. A higher number of parameters always implies an increased condition number (for the same training data and the same type of cost function). This corresponds to the intuitive observation that a higher number of parameters reduces the statistical accuracy on the parameters. The condition number also correlates with the cost function used to calibrate the parameters. In all cases, the full cost function yields a lower condition number, and hence more robust parameters, than the static cost function. This means that there is extra information contained in the linear response term of the full cost function that is not present in the static term.

TABLE IV. Performance in the reproduction of the electrostatic potential of the three population schemes.

RMSD [kJ/mol]	**Mulliken**	**Natural**	**Hirshfeld-I**
$\mathbf{P_C}$	23.3	9.4	4.8
$\mathbf{P_D}$	2.6	2.7	2.2

In principle, absolute values of the condition number have no straightforward interpretation. At first sight all the condition numbers in Table III are relatively high and it is therefore not possible to interpret individual parameters. However, even a parametrization with a high condition number still has predictive power when applied in practical applications. This will be demonstrated with the cross-validation test below.

B. Performance on the training set

All the results in this section are based on the 500 molecules in the training set. Before we discuss the performance of each model, it is instructive to test how well the three population schemes can reproduce the electrostatic potential generated by the full wave function. In table IV both performance measures P_C and P_D are computed, using population charges as input instead of charges generated by an empirical model. The results for P_C show that the Hirshfeld-I scheme yields population charges that reproduce the electrostatic potential with the smallest RMSD error. Natural charges yield a double error, while Mulliken charges show an RMSD error that is about five times as large. This result is in agreement with previous work of Van Damme et al.[72] The changes in the electrostatic potential generated by the wave function due to an external perturbation (P_D) is predicted with nearly the same accuracy by all three population schemes, although Hirshfeld-I is again the most reliable of the three schemes. This corresponds to the general observation that changes in partial charges are less dependent on a particular population scheme.[73]

The aim of this work is to contribute to the development of reliable polarizable force fields. One can only compute correct intermolecular interactions with a force-field model when the molecular electrostatic potential is reproduced. It is therefore appealing to rely on the Hirshfeld-I scheme to perform the parametrization. Hirshfeld-I has the additional advantage that it is easily extended to compute a complete multipole expansion of the charge density of each atom. In this work we only use atomic monopoles. One could argue that charges fitted to reproduce the electrostatic potential easily outperform Hirshfeld-I in this comparison. This is correct, but ESP-fitted partial charges are statistically ill-defined.[51] This would imply a noise on our training data, which would be reflected by noise on the empirical parameters.

Table V gives an overview of performance measure P_A, which corresponds to the reproduction of atomic partial charges by the empirical models. The general trend in table V is that Mulliken charges are harder to reproduce with any of the empirical models, compared to natural and Hirshfeld-I charges. The Mulliken population scheme has known unphysical artifacts such as a strong basis set dependence.[49] It is not surprising that the empirical models can not reproduce all these effects. Natural charges are reproduced slightly more accurately than Hirshfeld-I charges, although the difference is marginal. The static cost function results in parameters that are superior in the reproduction of the charges in absence of an external field, because the static cost function is very similar to performance measure P_A, while the full cost function also imposes other goodness-of-fit criteria. It is remarkable that this effect is much less pronounced in the case of the SQE model. The SQE model is designed to describe both linear response effects (polarization) and equilibrium charges at the same time. In the case of EEM, imposing a proper linear response

TABLE V. Performance measure P_A: the reproduction of atomic partial charges in absence of an electrostatic field.

RMSD [e]	**Model**	**Atom types**	**Cost fn.**	**Mulliken**	**Natural**	**Hirshfeld-I**
	EEM	Trivial	Static	0.30	0.09	0.11
	EEM	Trivial	Full	0.30	0.20	0.20
	EEM	Force Field	Static	0.25	0.08	0.08
	EEM	Force Field	Full	0.27	0.18	0.19
	SQE	Trivial	Static	0.26	0.08	0.11
	SQE	Trivial	Full	0.28	0.10	0.14
	SQE	Force Field	Static	0.15	0.04	0.05
	SQE	Force Field	Full	0.19	0.06	0.06
Relative RMSD [%]	**Model**	**Atom types**	**Cost fn.**	**Mulliken**	**Natural**	**Hirshfeld-I**
	EEM	Trivial	Static	68	22	31
	EEM	Trivial	Full	69	50	57
	EEM	Force Field	Static	57	20	21
	EEM	Force Field	Full	60	46	54
	SQE	Trivial	Static	58	19	30
	SQE	Trivial	Full	63	25	38
	SQE	Force Field	Static	34	10	13
	SQE	Force Field	Full	42	15	17

TABLE VI. Performance measure P_B: the reproduction of changes in atomic partial charges due to a perturbation with an external electrostatic potential.

RMSD [e]	**Model**	**Atom types**	**Cost fn.**	**Mulliken**	**Natural**	**Hirshfeld-I**
	EEM	Trivial	Static	0.022	0.016	0.020
	EEM	Trivial	Full	0.020	0.008	0.011
	EEM	Force Field	Static	0.022	0.015	0.019
	EEM	Force Field	Full	0.020	0.007	0.010
	SQE	Trivial	Static	0.022	0.012	0.020
	SQE	Trivial	Full	0.019	0.007	0.009
	SQE	Force Field	Static	0.023	0.008	0.011
	SQE	Force Field	Full	0.018	0.003	0.006
Relative RMSD [%]	**Model**	**Atom types**	**Cost fn.**	**Mulliken**	**Natural**	**Hirshfeld-I**
	EEM	Trivial	Static	89	119	107
	EEM	Trivial	Full	82	59	60
	EEM	Force Field	Static	90	113	101
	EEM	Force Field	Full	82	54	54
	SQE	Trivial	Static	91	92	104
	SQE	Trivial	Full	78	50	49
	SQE	Force Field	Static	94	62	58
	SQE	Force Field	Full	75	26	29

TABLE VII. Performance measure P_C: the reproduction of the electrostatic potential generated by the wavefunction in the absence of an external field.

RMSD [kJ/mol]	**Model**	**Atom types**	**Cost fn.**	**Mulliken**	**Natural**	**Hirshfeld-I**
	EEM	Trivial	Static	28.1	15.9	12.2
	EEM	Trivial	Full	28.5	29.6	17.2
	EEM	Force Field	Static	26.1	15.6	10.5
	EEM	Force Field	Full	26.0	27.8	17.5
	SQE	Trivial	Static	24.0	13.0	11.3
	SQE	Trivial	Full	21.5	15.6	12.6
	SQE	Force Field	Static	23.5	10.2	8.9
	SQE	Force Field	Full	20.2	12.7	10.1
Relative RMSD [%]	**Model**	**Atom types**	**Cost fn.**	**Mulliken**	**Natural**	**Hirshfeld-I**
	EEM	Trivial	Static	34	19	15
	EEM	Trivial	Full	34	35	20
	EEM	Force Field	Static	31	19	13
	EEM	Force Field	Full	31	33	21
	SQE	Trivial	Static	29	16	13
	SQE	Trivial	Full	26	17	15
	SQE	Force Field	Static	28	12	11
	SQE	Force Field	Full	24	15	12

TABLE VIII. Performance measure P_D: the reproduction of changes in the electrostatic potential generated by the wavefunction due to a perturbation with an external electrostatic field.

RMSD [kJ/mol]	**Model**	**Atom types**	**Cost fn.**	**Mulliken**	**Natural**	**Hirshfeld-I**
	EEM	Trivial	Static	2.6	2.9	3.4
	EEM	Trivial	Full	2.5	2.6	2.6
	EEM	Force Field	Static	2.6	2.8	3.3
	EEM	Force Field	Full	2.5	2.7	2.6
	SQE	Trivial	Static	3.6	2.3	2.8
	SQE	Trivial	Full	2.3	2.5	2.1
	SQE	Force Field	Static	3.8	2.8	2.1
	SQE	Force Field	Full	2.5	2.6	2.2
Relative RMSD [%]	**Model**	**Atom types**	**Cost fn.**	**Mulliken**	**Natural**	**Hirshfeld-I**
	EEM	Trivial	Static	56	63	74
	EEM	Trivial	Full	54	56	55
	EEM	Force Field	Static	55	61	72
	EEM	Force Field	Full	54	58	57
	SQE	Trivial	Static	78	50	61
	SQE	Trivial	Full	49	55	45
	SQE	Force Field	Static	81	61	45
	SQE	Force Field	Full	53	57	47

behavior will ruin the reproduction of equilibrium charges. The introduction of force field atom types only gives a marginal improvement in case of the EEM model, which is also observed by Bultinck and coworkers.[36] The gains are significant in the case of the SQE model. This is in line with the motivation for the force field atom types discussed in the previous section. It is also clear that the SQE model outperforms EEM in the reproduction of equilibrium charges. Although SQE is designed to show the proper linear response behavior in the limit of large molecules, also the reproduction of static charges improves. This is partially due to the perturbative electronegativity corrections in the SQE model, but even when the perturbations are disabled, the SQE model is more accurate than the EEM model.

Table VI lists the result of performance measure P_B, i.e. the quality of the reproduction of changes in atomic partial charges due to an electrostatic perturbation. In general, table VI offers insights that are analogous to the previous discussion. The change in partial charges due to an electrostatic perturbation is harder to predict in case of Mulliken charges, when compared to natural and Hirshfeld-I charges. Changes in natural charges are again reproduced slightly more accurately than changes in Hirshfeld-I charges. The full cost function results in parameters that reproduce the induced charges more accurately than the static cost function, due the linear response criteria that are only present in the full cost function. The introduction of force field atom types is only useful for the SQE models, in analogy with the results in table V. Again the SQE model outperforms the EEM model, which is expected because the extra term in the SQE model is introduced to improve the description of linear response properties.

Based on the results in table V and VI, one could conclude that the combination NSFF gives the best overall agreement. From the perspective of force-field parameterization, however, partial charges are not the property of interest. They are just a tool to obtain a correct molecular electrostatic potential (ESP). Table VII and VIII compare the reproduction of the molecular ESP in absence of an external field and changes in the ESP due to a perturbation, respectively. These tables are the ESP counterparts of table V and VI. As can be expected from table IV, the parameters based on Hirshfeld-I charges are now most favorable. There are a few minor anomalies in table VIII. In two cases the SQE model with parameters based on the static cost function give a better reproduction of the linear response than the parameters optimized with the full cost function. The two cases are NSTS versus NSTF and HSFS versus HSFF. Because the differences are small and these are exceptional cases, we expect that these anomalies are caused by a compensation of errors. Table IV shows clearly that an exact reproduction of induced Hirhsfeld-I charges results in an RMSD of 2.2 kJ/mol for peformance measure P_D. Fundamental improvements beyond this limit are only possible with atomic inducible dipoles. HSFS goes below this limit and has an RSMD of 2.1 kJ/mol for performance measure P_D. This is only possible by making errors with respect to the Hirshfeld-I charges. We conclude that the best overall performance is obtained with the combination HSFF.

Table IX compares the performance in terms of the reproduction of the change of the ab initio molecular energy due to an external perturbation. This table is mainly of interest for the development of molecular mechanics models, since it directly compares ab initio energy differences with empirical values. The trends in this table are completely analogous to table VII: the lowest RMSD error is obtained with the combination HSFS (7.7 kJ/mol), but it is closely followed by HSFF (8.8 kJ/mol). We prefer parameters obtained with the full cost function (HSFF) as to make sure that the model is also suitable for systems where polarization becomes more important.

Finally we analyze the polarizability tensors for the 500 molecules in the training set. The polarizability tensor characterizes the most important linear response effect: the change in dipole

TABLE IX. Performance measure P_E: the reproduction of changes in the molecular energy due to a perturbation with an external electrostatic field.

RMSD [kJ/mol]	**Model**	**Atom types**	**Cost fn.**	**Mulliken**	**Natural**	**Hirshfeld-I**
	EEM	Trivial	Static	22.2	12.5	10.0
	EEM	Trivial	Full	22.6	22.5	13.4
	EEM	Force Field	Static	20.5	12.3	8.6
	EEM	Force Field	Full	20.3	21.2	13.3
	SQE	Trivial	Static	19.1	10.6	9.4
	SQE	Trivial	Full	17.4	12.1	10.7
	SQE	Force Field	Static	18.6	8.8	7.7
	SQE	Force Field	Full	16.4	10.7	8.8
Relative RMSD [%]	**Model**	**Atom types**	**Cost fn.**	**Mulliken**	**Natural**	**Hirshfeld-I**
	EEM	Trivial	Static	37	21	17
	EEM	Trivial	Full	37	37	22
	EEM	Force Field	Static	34	20	14
	EEM	Force Field	Full	34	35	22
	SQE	Trivial	Static	32	17	16
	SQE	Trivial	Full	29	20	18
	SQE	Force Field	Static	31	15	13
	SQE	Force Field	Full	27	18	15

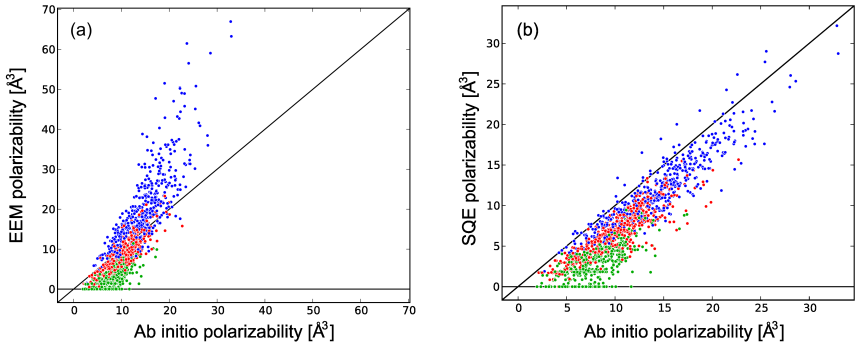


FIG. 5. Comparison of empirical versus ab initio eigenvalues of the polarizability tensor for the 500 molecules in the training set. Plot (a) contains the results of the HEFF parameterization and plot (b) the results for the HSFF parameterization. The lowest eigenvalue for each molecule is plotted in green, the second in red and the largest in blue.

moment due to a uniform electrostatic field. Note that the molecular monopole does not change due to the total charge constraint. For each molecule of the training set the three eigenvalues of the polarizability tensor resulting from the empirical models HEFF and HSFF are plotted against the equivalent eigenvalues resulting from the ab initio calculations in in figure 5a and in figure 5b, respectively. The only difference between both parameterizations is that the first one uses the EEM model, while the latter is based on the SQE approach. Both models have the intrinsic limitation that the polarizability orthogonal to a planar molecule is zero, which can be recognized in both plots: the

TABLE X. Performance of the HSFF model on two test sets ("bio" and "chain", see text for details) and on the training set. P_A to P_E are performance measures discussed in the text and in table II.

label	RMSD				relative RMSD [%]		
	bio	chain	train	unit	bio	chain	train
$\mathbf{P_A}$	0.049	0.058	0.065	e	11	40	18
$\mathbf{P_B}$	0.0037	0.0084	0.0055	e	26	55	29
$\mathbf{P_C}$	6.9	1.9	10.1	kJ/mol	18	45	12
$\mathbf{P_D}$	1.9	2.0	2.2	kJ/mol	38	40	47
$\mathbf{P_E}$	5.6	3.0	8.8	kJ/mol	22	64	15

lowest eigenvalue of the polarizability tensor of several molecules (green dots) is predicted to be zero, while the ab initio value is nonzero. This clearly points out that a complete empirical model for the molecular polarizability must include inducible atomic dipoles. The EEM model overestimates the largest polarizability eigenvalues. This issue is fixed with the SQE model, which exhibits the correct scaling in terms of molecular size. Figure 5b also reveals that SQE slightly underestimates all polarizability eigenvalues. This is again an indication that a contribution to the polarizability is missing, which can be cured by introducing inducible atomic dipoles in the model. The SQE model is a good starting point for such an extension, while EEM is clearly not. It is also important to realize that a large part of the polarizibility tensor is already captured by charge transfer alone. The contribution from inducible atomic dipoles should be relatively small. This issue will be subject of future research.

C. Performance on test sets

The most suitable parameterization, HSFF, has a similar performance on molecules not included in the training set. The first test set, labeled "bio", contains the 20 standard amino acids and the 5 nucleobases. The second test set, "chain", contains linear alkanes and linear conjugated alkenes up to 10 carbon atoms. Table X summarizes the five performance measures for the two test sets and the training set. In general both test sets have an even lower RMSD error in all cases except one: the changes in partial charges in the chain molecules induced by an external perturbation are reproduced less well compared to the training set. The general trend is mainly caused by the extreme variety of functional groups in the training set, while the test sets are modest in this perspective. The difficulty with the chain molecules is that they do not contain any polar functional groups and all atoms are close to neutral. Their electrostatic behavior in the ab initio reference data is largely determined by atomic dipoles. For the same reason, also the relative errors in the test set with chain molecules are large. For a proper description of pure alkanes and alkenes, an extension of the model with inducible dipoles is compulsory. However, the HSFF parameterization is readily transferable to more general organic molecules that do contain functional groups.

D. Cross validation

A cross-validation is performed on the HSFF parameterization to assess the transferability of the parameters to similar molecules that are not included in training set. The averages of the performance measures obtained from the cross-validation are summarized in table XI. Also the performance measures on the original HSFF parameterization are included to facilitate the comparison. All cross-validation results are nearly identical to the corresponding values from the

Table XI. Cross validation results for the HSFF parameterization. The performance measures of the cross validation are averages over 100 individual parameterizations. P_A to P_E are performance measures discussed in the text and in table II.

label	RMSD			Relative RMSD [%]	
	cross val.	train	unit	cross val.	train
$\mathbf{P_A}$	0.0654	0.0646	e	17.7	18.1
$\mathbf{P_B}$	0.00554	0.00550	e	28.5	28.9
$\mathbf{P_C}$	9.19	10.05	kJ/mol	12.5	12.0
$\mathbf{P_D}$	2.208	2.193	kJ/mol	48.2	47.3
$\mathbf{P_E}$	7.98	8.83	kJ/mol	15.2	14.5

original HSFF parameterization. The differences are irrelevant in magnitude and largely caused by the statistical error on the average computed over the 100 cross-validation parameterizations. This practically means that the HSFF parameterization is transferable to molecules similar to the ones included in the training set. Since the training set is very diverse in terms of functional groups, a wide range of organic molecules can be modeled with the HSFF parameterization. Additional validation is required to asses the transferability of the model to systems that are substantially larger than the molecules in the training set.

E. Applications to chain molecules

In this subsection we discuss some specific examples that show the differences between the EEM and SQE model in practical applications with extended systems, consisting of a larger number of atoms than the molecules in the training data (up to 24 atoms). In the first application the polarizability is computed for three different chain molecules, which can be systematically extended in size: a linear alkane, a linear conjugated alkene and a polyalanine alpha helix. Three parameterizations are used for comparison: HETS, HEFF and HSFF. In the second application, we analyze the dipole moment of the alpha helix as function of the chain length, using the HETS and HSFF parameterizations.

The geometries of the chain molecules are based on the Z-matrix of a single monomer. All monomer Z-matrices are listed in part IV of the supporting information. In contrast to a conventional Z-matrix, the references to previous atoms are relative and all the geometric parameters in the first three lines are fully defined. They refer to distances and angles with respect to the previous monomer. A straightforward concatenation of multiple Z-matrices results in the Z-matrix of a polymer. After the concatenation procedure, a few lines must be added to terminate the structure properly, e.g. for the alkane a hydrogen atom must be added to each terminal methyl group. In case of the alanine alpha helix, the dihedral angles in the first three rows correspond to the ψ, ω and φ angles of the protein backbone. They are set to -50°, 185° and -60° respectively to obtain the proper shape of the alpha helix. Figure 6 depicts an alanine alpha helix containing 10 amino acids or residues. For the remainder of the discussion, it is important to realize that the first internal hydrogen bond is formed when five residues are present. In principal all geometries should be optimized at the MP2/CC-vPDZ level before they are used in the applications below. The parameters are well-tested for equilibrium geometries, but they are not validated on non-equilibrium geometries. The required geometry optimizations are not feasible, and hence the geometries generated with the Z-matrices are not optimized. Consequently the results below are qualitative.

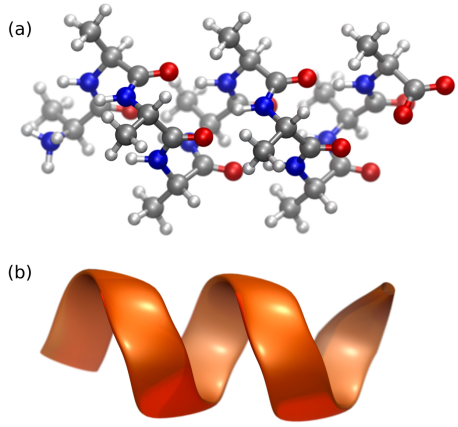


FIG. 6. (a) Ball and stick representation and (b) cartoon representation of the deca-alanine alpha helix.

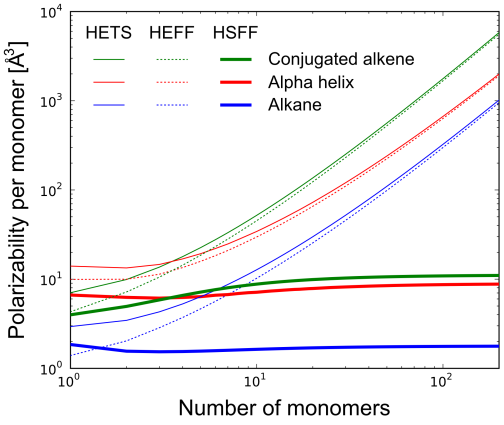


FIG. 7. The polarizability per monomer as a function of the chain length. Three chain molecules are considered: a linear conjugated alkene (green), a polyalanine alpha helix (red) and a linear alkene (blue). The polarizability along the axis of the chain molecule is computed with three different models: HETS (thin solid line), HEFF (thin dashed line) and HSFF (thick solid line).

The main purpose of the first application is to illustrate the different trends in polarizability between the EEM and SQE model. The polarizability per monomer is computed as a function of the chain length up to 200 monomers. The results are depicted in figure 7. The systematic error of the EEM model (HETS and HEFF) in the prediction of the polarizability as function of the chain length is manifest. In the macroscopic limit we expect that the polarizability of a dielectric material is an additive property. This means that for long chains each additional monomer should contribute a

constant amount to the polarizability. This behavior is only reproduced with the SQE model (HSFF). The EEM model predicts a divergent polarizability per monomer in the limit of long chains, which is manifestly unphysical. This error is inherent to the EEM model and can not be fixed by including linear response data in the cost function. (See HEFF results.) These observations are in line with results in the literature.[41-44] We show that the SQE model is a feasible solution, even for complicated systems with many different atom and bond types such as an alpha helix.

In the second application we also compute the dipole moment per monomer for two variations of the alpha helix: a normal alpha helix (with a charged C- and N-terminus) and one with charge compensating Na^+ and Cl^- ion at distance of about 5 angstrom from the end points of the helix. The ions are treated as points with fixed charges of +1e and -1e respectively. They represent a realistic electrostatic perturbation. The dipole moment is projected on the axis of the alpha chain and the contribution from the ions is not included. The results for the HETS and HSFF parameterization are plotted in figure 8a.

The two parameterizations HETS and HSFF behave differently but they show also some similarities. Both predict the same dipole moment for the single alanine molecule in the absence of a perturbation, which is in line with the performance of both parameterizations in table VII. The perturbation by the two ions increases the dipole moment in the two cases. There are however three important differences caused by the erroneous description of linear response in the EEM model. The first difference is trivial: the induced dipole due to the ions, is larger in the EEM model. Two other discrepancies are due to a combination of effects that will be discussed in detail below: (i) the SQE model predicts that the dipole moment per monomer increases again as soon as the first intramolecular hydrogen bonds are formed, while the EEM model predicts a further decrease, and (ii) in the limit towards very long chains, the dipole moment per monomer converges to a fixed value in the SQE case, while the EEM model shows very different values depending on the presence of charge compensating ions.

One gains more insight with a decomposition of the dipole moment into a contribution from the net charge on each residue and a contribution of the dipole moments of each residue. Since residues have a net charge different from zero, we have systematically taken the center of mass of each residue as the reference point to compute the corresponding dipole moment. The contribution from the residue dipoles is given in figure 8b and the contribution from the residue charges is given in figure 8c. In both plots, we recognize two regimes. The first regime is in the range of a few residues, when no hydrogen bonds are formed. The charge and dipole moment of each residue in a representative short peptide (4 monomers) are plotted and figure 9a and 9b, respectively. These plots contain the results of both the HETS and HSFF parameterization. The second regime covers the longer chains. Similar plots for a representative chain (30 monomers) are given in figure 9c and 9d.

The trend in the contribution from the residue dipoles is similar for all models. The dipole moment per monomer decreases when the first monomers are added because the terminal residues have the largest contribution, while the remaining residues only lower the average dipole moment per monomer. As soon as hydrogen bonds are formed, each pair of alanine residues connected by a hydrogen bond exerts a mutual polarization, resulting in an additional induced dipole moment along the direction of the alpha chain. This results in an increase of the average dipole moment per monomer.

The contribution from the net charge on the residues to the dipole moment per monomer is very different when comparing the SQE and the EEM model. In the limit of short chains, the dipole moment per monomer increases in all cases. The total charge of the terminal residues does not depend on the chain length in the SQE case. When the number of residues is x, the dipole moment

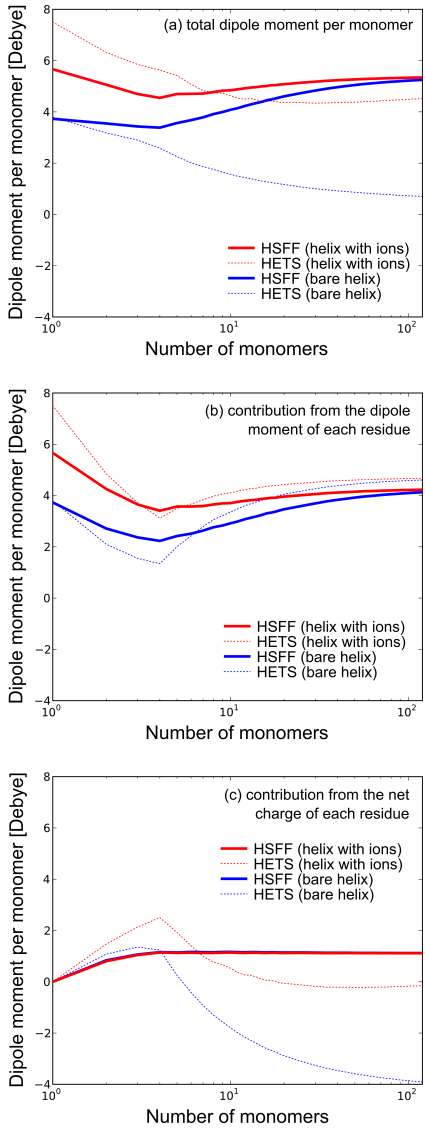


FIG. 8. The dipole moment per monomer of a polyalanine alpha helix as function of the chain length. The results are computed with two different parameterizations: HSFF (thick solid line) and HETS (thin dashed line). The results for a conventional helix are plotted in blue. We also considered a helix where the charged termini are compensated by Na^+ and Cl^- ions. Only the dipole moment of the chains is plotted; the contribution from the ions is not included. Plot (a) shows the total dipole moment. Plot (b) and (c) contain the contribution from the dipoles and the net charges of the individual residues, respectively.

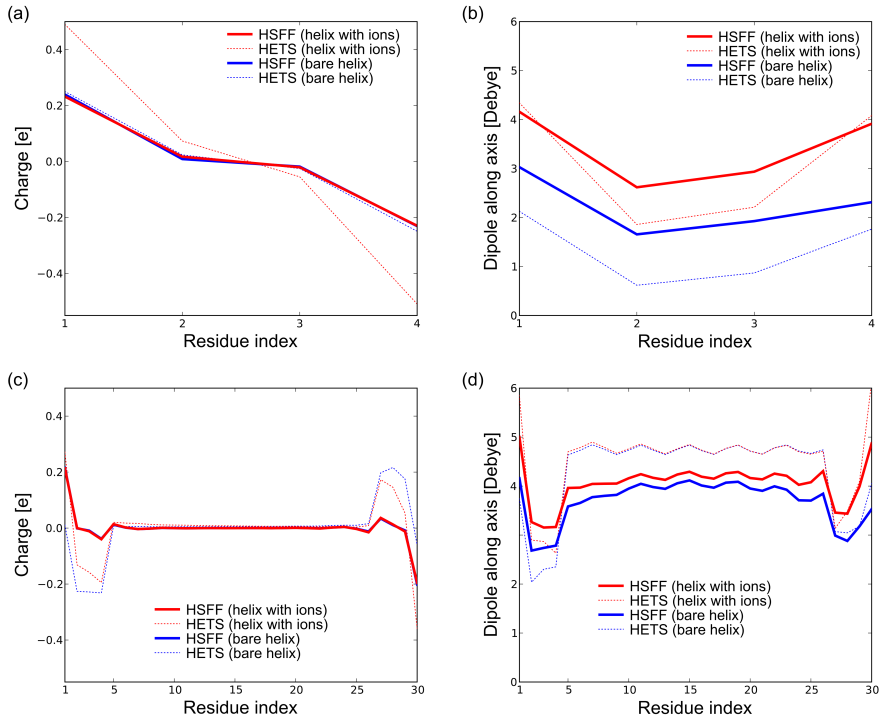


FIG. 9. The charge and dipole of each residue in a polyalanine alpha helix containing 4 and 30 monomers. (a): 4 monomers, charge of each residue. (b): 4 monomers, dipole of each residue. (c): 30 monomers, charge of each residue. (d): 30 monomers, dipole of each residue.

is proportional to $x - 1$ and the dipole moment per monomer is proportional to $\frac{x-1}{x}$. The results of the SQE model follow this trend faithfully. In case of the EEM model, the charges on the termini and the dipole moment of each residue induce a charge transfer between the two ends of the alpha helix. Figure 9c shows that the transfered charges accumulate in the residues next to the termini. This metallic type of induced charge transfer tries to cancel the static dipole moment. In the alpha helix without ions the dipole moment caused by the residue charges cancels to the dipole moment due to the residue dipoles in the limit of long chains. Figure 9c also shows that this metallic type of charge transfer is very sensitive to the perturbation of the ions: the total amount of charge in the "bumps" next to the termini changes sensitively due to the presence of the ions. The outcome of the EEM is rather unrealistic: the total dipole moment of an organic molecule should not depend on the presence of relatively small perturbations.

The atomic partial charges in a single residue obtained from the HETS and HSFF parameterizations are very similar. The differences between the EEM and SQE model only become apparent when the charge sum is taken over a molecular fragment such as an amino acid, e.g. the charges shown in figure 9c. Table XII compares the partial charges on residue number 3 of the alpha helix containing 30 residues. The numbering of the atoms is depicted in figure 10. The relative

Table XII. Side-by-side comparison of partial charges in residue 3 of the alpha helix with 30 residues.

Index	Atom type	HESF charge [e]	HSFF charge [e]
1	N3	-0.77	-0.66
2	C4	0.10	0.02
3	C3	0.80	0.81
4	O1	-0.67	-0.60
5	C4	-0.54	-0.49
6	H	0.11	0.12
7	H	0.28	0.32
8	H	0.14	0.17
9	H	0.16	0.17
10	H	0.16	0.14
	Sum	*-0.23*	*0.00*

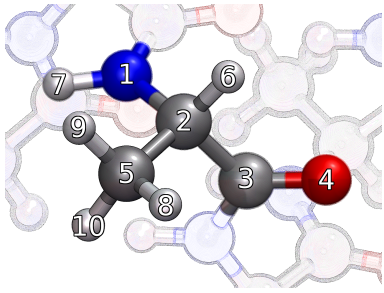


FIG. 10. The numbering of the atoms in the third residue of the alanine helix. The surrounding atoms of the alpha helix are drawn in low contrast.

differences are small compared to the absolute values of the charges. However, the relative difference between the total charges is large and significant for long range interactions. For example when two fragments are 5 Å apart from each other, the repulsive energy due to a net charge of 0.23 e on both fragments is about 15 kJ/mol.

The two applications show that the EEM overestimates the polarizability, also for complex organic systems such as an alpha helix. This is even relevant for simulations where the polarizability itself is not the quantity of interest. The dipole moment of an alpha helix in absence of an external perturbation is predicted incorrectly by the EEM model due to an overestimate of internal polarization effects. This problem is solved entirely with the SQE model. The latter model leads to the correct behavior in the limit of large system sizes.

IV. Conclusions

This papers presents an extensive comparison of the Electronegativity Equalization Method (EEM) and the Split Charge Equilibration (SQE) model for the empirical prediction of atomic partial charges of a broad range of organic molecules. Both models postulate a quadratic expansion of the molecular electronic energy as function of the atomic partial charges. One obtains the ground state charge distribution by minimizing this energy under a total charge constraint. The models can be extended with an energy term that takes into account an external electrostatic perturbation, and hence one can also describe polarization due to interatomic charge transfer. An essential defect of the EEM model, and many analogous approaches in the literature, is the incorrect size-dependence of the electronic polarizability. EEM predicts that the polarizability of chain molecules is a cubic function of the chain length, while both experimental and ab initio calculations reveal a linear dependence.[44] In the macroscopic limit one expects that polarizability is an additive property. This implies that the polarizability increases with a constant value when a monomer is added to the chain. The SQE model can be regarded as an extension of the EEM with an additional energy term that fixes this deficiency. In the EEM model, an electrostatic perturbation can cause significant charge transfer over an arbitrary distance, while the SQE energy is explicitly dependent on the length of the path - the number of bonds - along which charge is transfered.

The training data used for the parameter calibration consists of MP2/Aug-CC-vPDZ calculations on a set of 500 very distinct organic molecules. An autonomous algorithm has selected these molecules from a larger set of 431980 substances taken from the PubChem database.[47] The 500 molecules exhibit a wide variety of organic functional groups and contain the following elements: H, C, N, O, F, S, Cl and Br. The selection algorithm maximizes the diversity of the 500 molecules and guarantees that sufficient information is present in the training data for each parameter in the most extensive model in this work. Atomic partial charges were derived with three charge population schemes: Mulliken charges[48], Natural charges[49] and Iterative Hirshfeld charges[50]. Also the ab initio electrostatic potential is computed on a molecular grid for validation purposes. In line with the work of Van Damme et al, Hirshfeld-I charges reproduce the electrostatic potential with the highest accuracy.[72] For each molecule, also the linear response to the perturbation by randomly placed point charges is computed and this information is used for both parameter optimization and validation purposes.

Our benchmarks compare in total 24 parameterizations, based on three populations schemes to derive ab initio atomic reference charges, two ways to define atom types, two cost functions to optimize the parameters and two empirical models (EEM and SQE). The SQE model is not only superior in the reproduction of linear response data, but also reduces the RMSD error in the reproduction of atomic charges with a factor 2. Although both Natural charges and Hirshfeld-I charges can be reproduced accurately with the SQE model, the latter definition of atomic charges is preferred because it leads to a better reproduction of the molecular electrostatic potential. One obtains more robust parameters when linear response data is incorporated into the cost function and in case of the SQE model it is possible to have a good reproduction of both the molecular polarizibility and atomic partial charges. For the SQE model it is advantageous to introduce different parameters for the same atom in different chemical states, e.g. different hardness parameters for sp, sp^2 and sp^3 carbons. This is practically achieved with the definition of force-field atom types.

Applications on chain molecules illustrate the unphysical overestimation of the polarizability in the EEM model for increased chain lengths, while the SQE model shows the correct behavior.

The practical consequences of an overestimated polarizability are demonstrated with an analysis of the dipole moment of an alpha helix with both models. In large systems, the EEM reduces the dipole moment due to excessive internal polarization effects. Although individual atomic charges still seem reasonable in the EEM model, the total charge on molecular fragments, such as residues in the helix, reveal an unphysical long range charge transfer.

It is of utmost importance for the development of polarizable force fields that the empirical models are transferable from small molecules, typically used for the parameterization, to large molecular systems in state-of-the-art molecular dynamics simulations. This work shows that the SQE model is an essential component of such a polarizable force field. Interatomic charge transfer explains the larger part of the electronic polarizability and the SQE model is the only transferable model that describes this phenomenon effectively on massive molecular systems, both in terms of physical behavior and in terms of computational cost. The SQE model has the additional advantage that atomic partial charges are predicted twice as accurately as with the original EEM. The main difficulty of these models is a reliable parameterization, for which we offer the necessary methodology in this paper. A straightforward extension of the SQE model with inducible atomic dipoles will further refine the accuracy.

Acknowledgments

This work was performed in the frame of GOA (Research Board of the Ghent University), IDECAT, and CECAT projects, and BELSPO, in the frame of IAP 6/27. Christoph van Wuellen (Ruhr-Universitaet, Bochum, Germany) generated the Fortran routines to compute Lebedev-Laikov grids by translating the original C-routines kindly provided by Dmitri Laikov (Moscow State University, Moscow, Russia). We are in debt to Christoph van Wuellen and Dmitri Laikov for making these routines publicly available. Visual Molecular Dynamics (VMD) was used for figure 6 and 10. We would like to thank P. Bultinck (Ghent University, Ghent, Belgium), D. Van Neck (Ghent University, Ghent, Belgium) and P. W. Ayers (McMaster Univeristy, Hamilton, Canada) for fruitful discussions.

References

1. D. Lesthaeghe, P. Vansteenkiste, T. Verstraelen, A. Ghysels, C.E. Kirschhock, J.A. Martens, V. Van Speybroeck and M. Waroquier, J. Phys. Chem. C **112**, 9186 (2008).
2. T. Vertraelen, B.M. Szyja, D. Lesthaeghe, R. Declerck, V. Van Speybroeck, M. Waroquier, P.J. Jansen, A. Aerts, L.R.A. Follens, J.A. Martens, C.E.A. Kirschhock and R.A. van Santen, Top. Catal. in press (2009).
3. M.W. van der Kamp, K.E. Shaw, Christopher and A.J. Mulholland, J. R. Soc. Interaface **5**, 173 (2008).
4. T. Halgren and W. Damm, Curr. Opin. Struc. Biol. **11**, 236 (2001).
5. J.L. Banks, G.A. Kaminski, R. Zhou, D.T. Mainz, B.J. Berne and R.A. Friesner, J. Chem. Phys. **110**, 741 (1999).
6. J.B. Foresman and C.L. Brooks, J. Chem. Phys. **87**, 5892 (1987).
7. Y. Zhong, L.G. Warren and S. Patel, J. Comput. Chem. **29**, 1142 (2008).
8. W.L. Jorgensen, J. Chem. Theory Comp. **3**, 1877 (2007).
9. W. Xie, J. Pu, A.D. Mackerell and J. Gao, J. Chem. Theory Comput. **3**, 1878 (2007).
10. V.M. Anisimov, I.V. Vorobyov, B. Roux and A.D. Mackerell, J. Chem. Theory Comput. **3**, 1927 (2007).

11. B. Dick and A. Overhauser, Phys. Rev. **112**, 90 (1958).
12. J. Applequist, J.R. Carl and K.-K. Fung, J. Am. Chem. Soc. **94**, 2952 (1972).
13. W. Mortier, K. Van Genechten, J. Gasteiger, J. Am. Chem. Soc. **107**, 829 (1985).
14. W. Mortier, S. Ghosh and S. Shankar, J. Am. Chem. Soc. **108**, 4315 (1986).
15. G. O. Janssens, B. G. Baekelandt, H. Toufar, W.J. Mortier, R.A. Schoonheydt, J. Phys. Chem. **99**, 3251 (1995).
16. R. Heidler, G. Janssens, W. Mortier, R. Schoonheydt, Microporous Mat. **12**, 1 (1997).
17. A. Rappe and W. Goddard, J.Phys. Chem. **95**, 3358 (1991).
18. A. Rappe, C. Casewit, K. Colwell, W. Goddard and W. Skiff, J. Am. Chem. Soc. **114**, 10024 (1992).
19. D.M. York and W. Yang, J. Chem. Phys. **104**, 159 (1996).
20. R. Chelli and P. Procacci, J. Chem. Phys. **117**, 9175 (2002).
21. S.W. Rick, S.J. Stuart and B.J. Berne, J. Chem. Phys. **101**, 6141 (1994).
22. B.R. Brooks, R.E. Bruccoleri, B.D. Olafson, D.J. States, S. Swaminathan and M. Karplus, J. Comput. Chem. **4**, 187 (1983).
23. S. Patel and C.L. Brooks, J. Comput. Chem. **25**, 1 (2004).
24. S. Patel, A.D. Mackerell and J. Charles, Comput. Chem. **25**, 1504 (2004).
25. D.A. Case, T.E. Cheatham, T. Darden, H. Gohlke, R. Luo, K.M. Merz, A. Onufriev, C. Simmerling, B. Wang and R.J. Woods, J Comput. Chem. **26**, 1668 (2005).
26. Z.-X. Wang, W. Zhang, C. Wu, H. Lei, P. Cieplak and Y. Duan, J. Comput. Chem. **27**, 781 (2006).
27. W.L. Jorgensen, D.S. Maxwell and J. Tirado-Rives, J. Am. Chem. Soc. **118**, 11225 (1996).
28. J.R. Maple, Y. Cao, W. Damm, T.A. Halgren, G.A. Kaminski, L.Y. Zhang and R.A. Friesner, J. Chem. Theory Comput. **1**, 694 (2005).
29. K.A. Van Genechten, W.J. Mortier and P. Geerlings, J. Chem. Phys. **86**, 5063 (1987).
30. Z.Z. Yang and C.S. Wang, J. Phys. Chem. A **101**, 6315 (1997).
31. S.L. Njo, J. Fan and B. van de Graaf, J Mol. Cat. A **134**, 79 (1998).
32. J.N. Louwen and E.T. Vogt, J. Mol. Cat. A **134**, 63 (1998).
33. G. Menegon, K. Shimizu, J.P.S. Farah, L.G. Dias and H. Chaimovich, Phys. Chem. Chem. Phys. **4**, 5933 (2002).
34. P. Bultinck, W. Langenaeker, P. Lahorte, F. De Proft, P. Geerlings, M. Waroquier and J.P. Tollenaere, J. Phys. Chem. A **106**, 7887 (2002).
35. P. Bultinck, W. Langenaeker, P. Lahorte, F. De Proft, P. Geerlings, C. Van Alsenoy and J.P. Tollenaere, J. Phys. Chem. A **106**, 7895 (2002).
36. P. Bultinck, R. Vanholme, P.L.A. Popelier, F. De Proft and P. Geerlings, J. Phys. Chem. A **108**, 10359 (2004).
37. M.K. Gilson, H.S. Gilson and M.J. Potter, J. Chem. Inf. Comput. Sci. **43**, 1982 (2003).
38. Q. Yang and K.A. Sharp, J. Chem. Theory Comput. **2**, 1152 (2006).
39. R.S. Varekova, Z. Jirouskova, J. Vanek, S. Suchomel and J. Koca, Int. J. Mol. Sci. **8**, 572 (2007).
40. I. Berente, E. Czinki and G.J. Náray-Szabó, Comput. Chem. **28**, 1936 (2007).
41. R. Chelli, P. Procacci, R. Righini and S. Califano, J. Chem. Phys. **111**, 8569 (1999).
42. R. Chelli and P. Procacci, J. Chem. Phys. **118**, 1571 (2003).
43. R. Chelli and P. Procacci, J. Phys. Chem. B **108**, 16995 (2004).
44. L.G. Warren, J.E. Davis and S. Patel, J. Chem. Phys. **128**, 144110 (2008).

45. K. Shimizu, H. Chaimovich, J.P.S. Farah, L.G. Dias and D.L. Bostick, J. Phys. Chem. B **108**, 4171 (2004).
46. R.A. Nistor, J.G. Polihronov, M.H. Müser and N.J. Mosey, J. Chem. Phys. **125**, 94108 (2006).
47. NCBI, The PubChem Project, http://pubchem.ncbi.nlm.nih.gov/ (last accessed 16 March 2009).
48. R.S. Mulliken, J. Chem. Phys. **23**, 1833 (1955).
49. A.E. Reed, R.B. Weinstock and F. Weinhold, J. Chem. Phys. **83**, 735 (1985).
50. P. Bultinck, C. Van Alsenoy, P.W. Ayers and R.C. Dorca, J. Chem. Phys. **126**, 144111 (2007).
51. M.M. Francl and L.E. Chirlian, Rev. Comp. Chem. **14**, 1 (2000).
52. J. Chen and T.J. Martínez, Chem. Phys. Lett. **438**, 315 (2007).
53. J. Chen, D. Hundertmark and T.J. Martínez, J. Chem. Phys. **129**, 214113 (2008).
54. SQLite, http://www.sqlite.org/ (last accessed on 16 March 2009).
55. T.M.J. Fruchterman and E.M. Reingold, Softw.-Pract. Exper. **21**, 1129 (1991).
56. T. Kamada and S. Kawai, Inf. Process. Lett. **31**, 7 (1989).
57. C. Møller and M.S. Plesset, Phys. Rev. **46**, 618 (1934).
58. M. Headgordon, J. Pople and M. Frisch, Chem. Phys. Lett. **153**, 503 (1988).
59. D.E. Woon and T.H. Dunning, J. Chem. Phys. **98**, 1358 (1993).
60. M.J. Frisch, G.W. Trucks, H.B Schlege *et al.*, GAUSSIAN 03, Revision D1, Gaussian, Inc., Wallingford, CT, 2004
61. R.W. Kennard and L.A. Stone, Technometrics **11**, 137 (1969).
62. V.I. Lebedev and D.N. Laikov, Dokl. Math. **59**, 477 (1999).
63. B. Cordero, V. Gomez, A.E. Platero-Prats, M. Reves, J. Echeverria, E. Cremades, F. Barragan and S. Alvarez, Dalton Trans. 2832 (2008).
64. S. Boyd, L. Vandenberghe, in *Convex Optimization,* third edition, (Cambridge University Press, United Kingdom, 2006), Pt. III, Chap. 11.
65. The Python programming language, http://www.python.org/ (last accessed 16 March 2009)
66. P.F. Dubois, Comput. Sci. Eng. **9**, 7 (2007).
67. NumPy: Numerical extensions for Python, http://www.numpy.org/ (last accessed 16 March 2009).
68. Matplotlib: The Python plotting library, http://matplotlib.sourceforge.net/ (last accessed 16 March 2009).
69. T. Verstraelen, V. Van Speybroeck and M. Waroquier, J. Chem. Inf. Model. **48**, 1530 (2008).
70. T. Verstraelen, V. Van Speybroeck, M. Waroquier, ZEOBUILDER, http://www.zeobuilder.org/ (last accessed 16 March 2009).
71. T. Verstraelen, D. Van Neck, P.W. Ayers, V. Van Speybroeck and M. Waroquier, J. Chem. Theory Comput. **3**, 1420 (2007).
72. S. Van Damme, P. Bultinck and S. Fias, J. Chem. Theory Comput. **5**, 334 (2009).
73. K. Hemelsoet, D. Lesthaeghe, V. Van Speybroeck and M. Waroquier, J. Phys. Chem. C **111**, 3028 (2007).

Supporting Information for The Electronegativity Equalization Method and the Split Charge Equilibration Applied to Organic Systems: Parameterization, Validation and Comparison

Toon Verstraelen, Veronique Van Speybroeck, Michel Waroquier

Part I: Atom and bond types

Trivial atom types (8)

H, C, N, O, F, S, Cl, Br

Trivial bond types (19)

H-C, H-N, H-O, H-S, C-C, C-N, C-O, C-F, C-S, C-Cl, C-Br, N-N, N-O, N-F, N-S, N-Cl, O-O, O-S, S-S

Force field atom types (15)

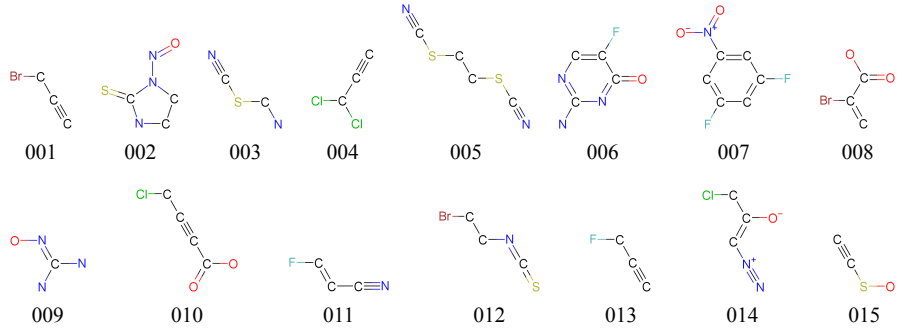
H, C4, C3, C2, N4, N3, N2, N1, O2, O1, F, S2, S1, Cl, Br

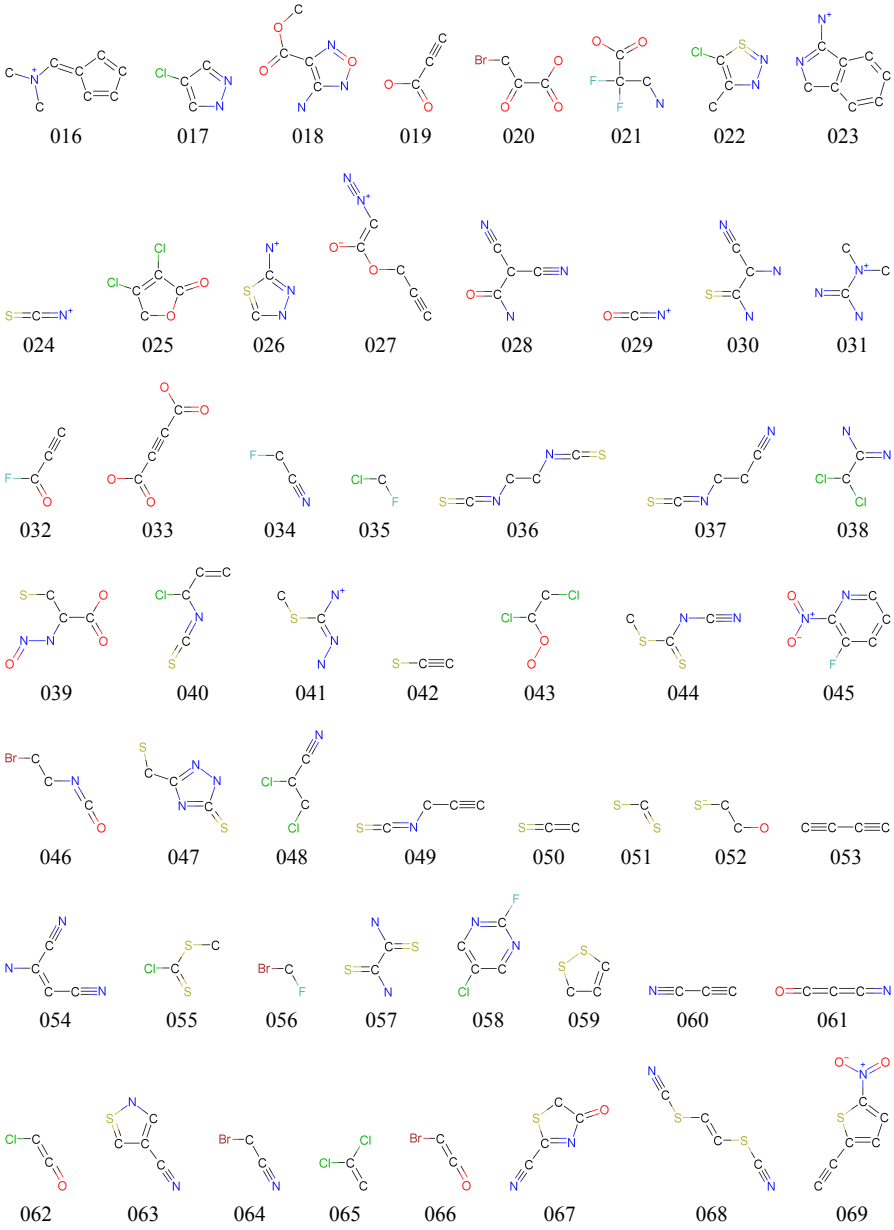
Force field bond types (56)

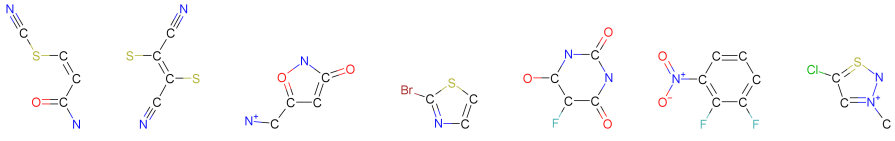
H-C4, H-C3, H-C2, H-N4, H-N3, H-N2, H-O2, H-S2, C4-C4, C4-C3, C4-C2, C4-N4, C4-N3, C4-N2, C4-O2, C4-O1, C4-F, C4-S2, C4-S1, C4-Cl, C4-Br, C3-C3, C3-C2, C3-N4, C3-N3, C3-N2, C3-O2, C3-O1, C3-F, C3-S2, C3-S1, C3-Cl, C3-Br, C2-C2, C2-N3, C2-N2, C2-N1, C2-O2, C2-O1, C2-S2, C2-S1, N3-N3, N3-N2, N3-O2, N3-O1, N3-F, N3-S2, N3-Cl, N2-N2, N2-N1, N2-O2, N2-O1, N2-S2, O2-O2, O2-S2, S2-S2

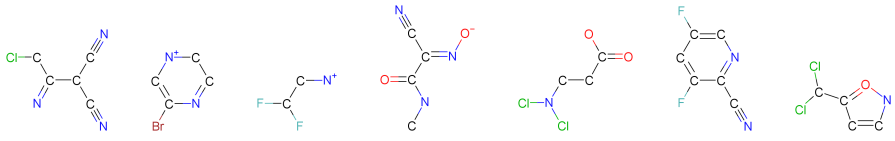
Part II: Molecules in the training set

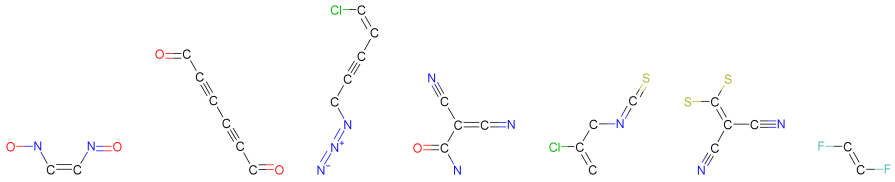
A. Lewis structures

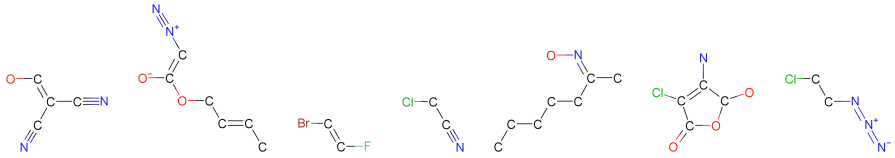


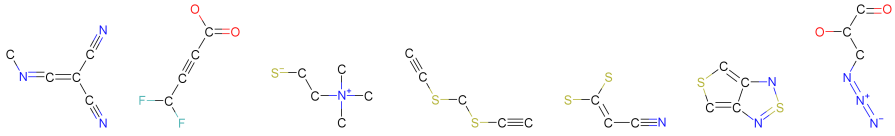

016
017
018
019
020
021
022
023
024
025
026
027
028
029
030
031
032
033
034
035
036
037
038
039
040
041
042
043
044
045
046
047
048
049
050
051
052
053
054
055
056
057
058
059
060
061
062
063
064
065
066
067
068
069

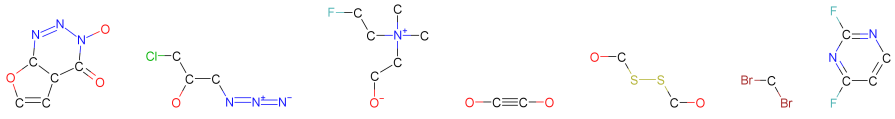

070 071 072 073 074 075 076

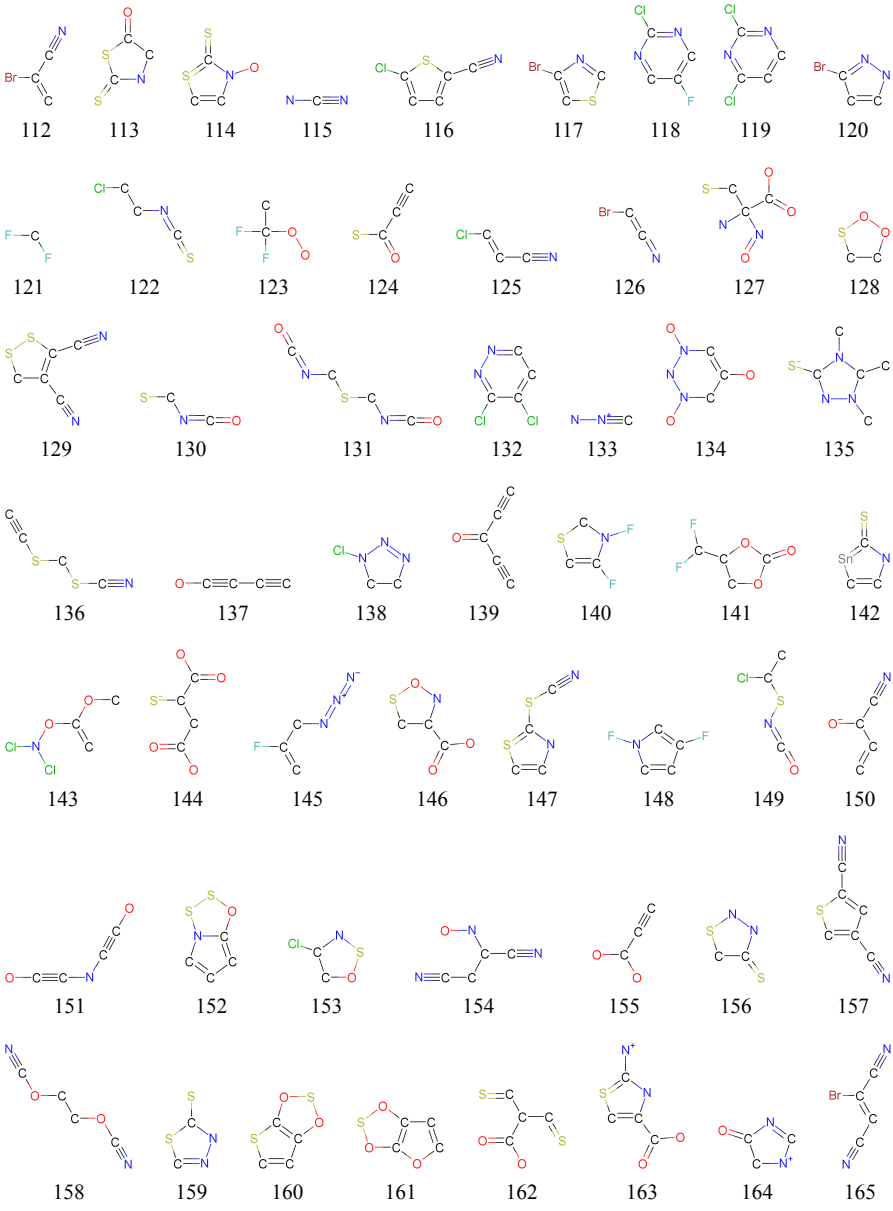

077 078 079 080 081 082 083

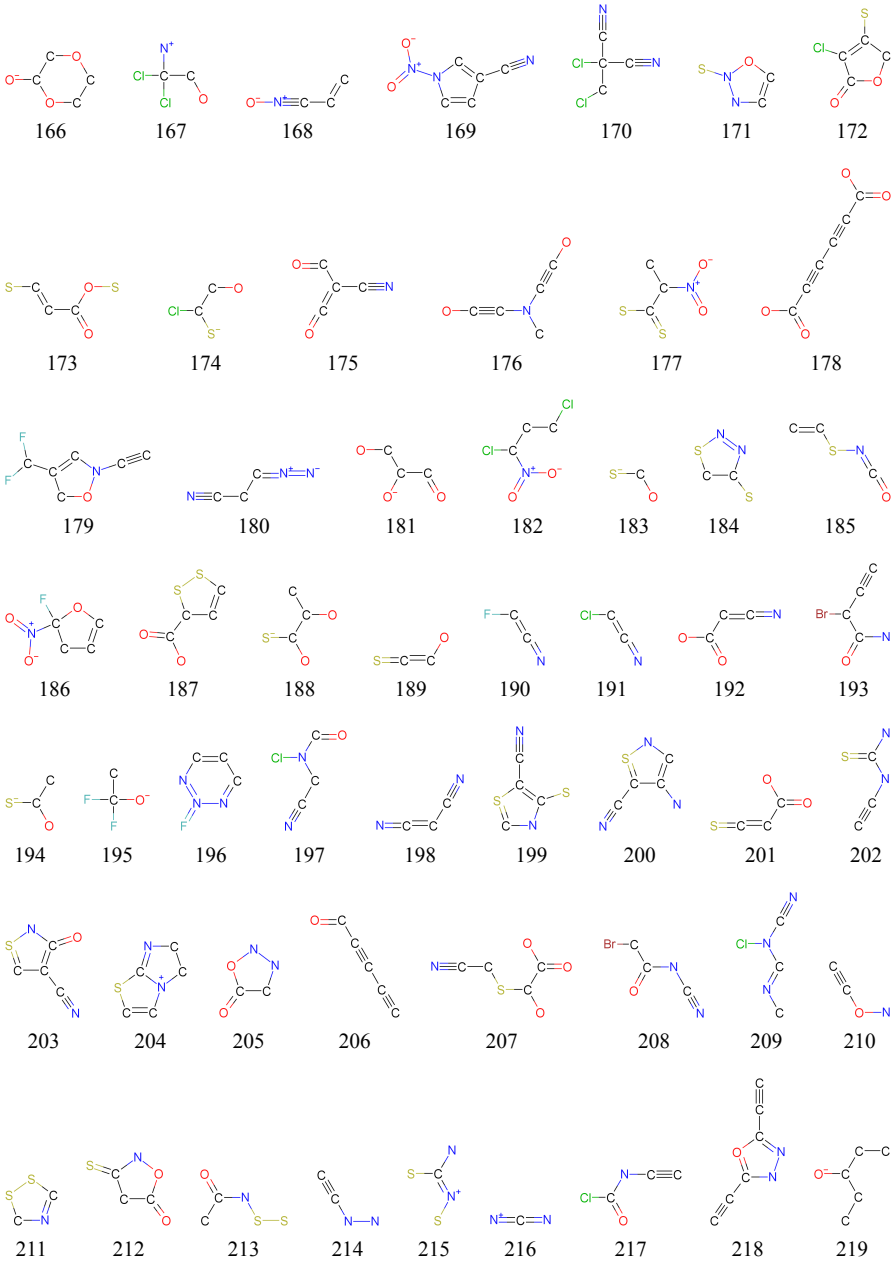

084 085 086 087 088 089 090

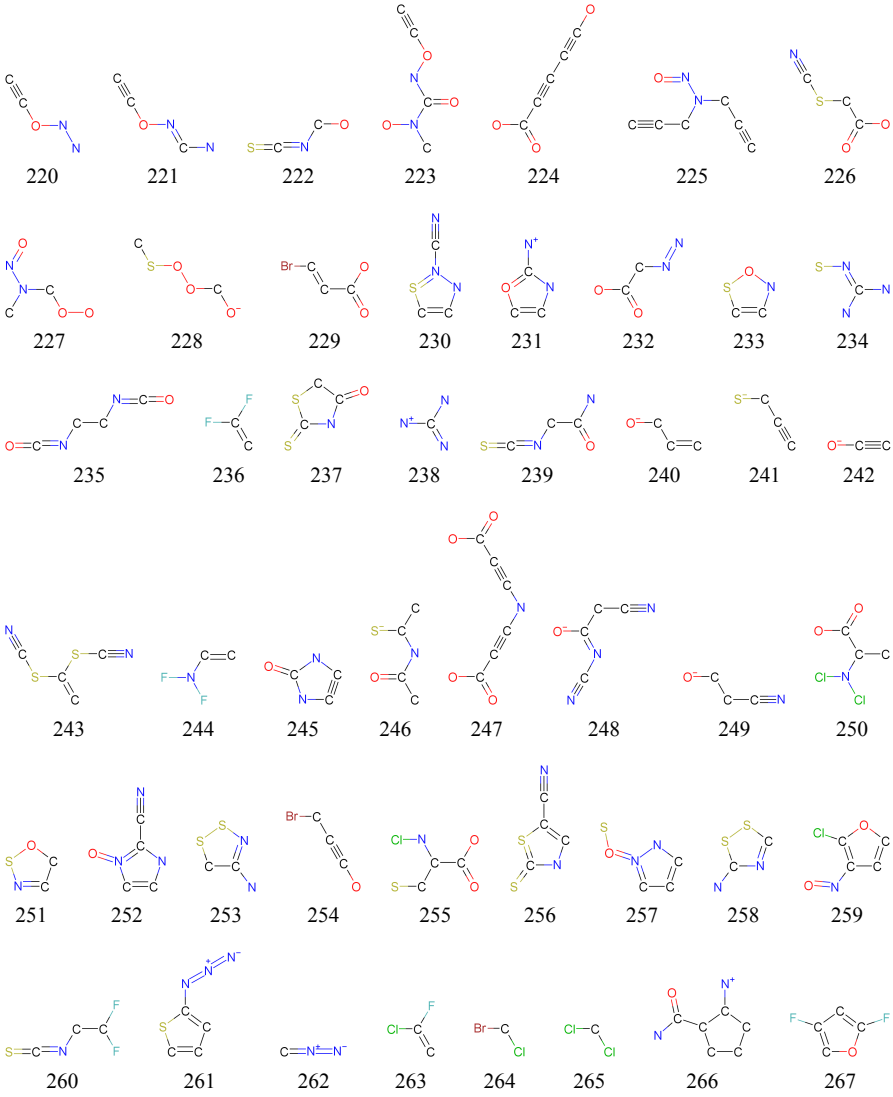

091 092 093 094 095 096 097

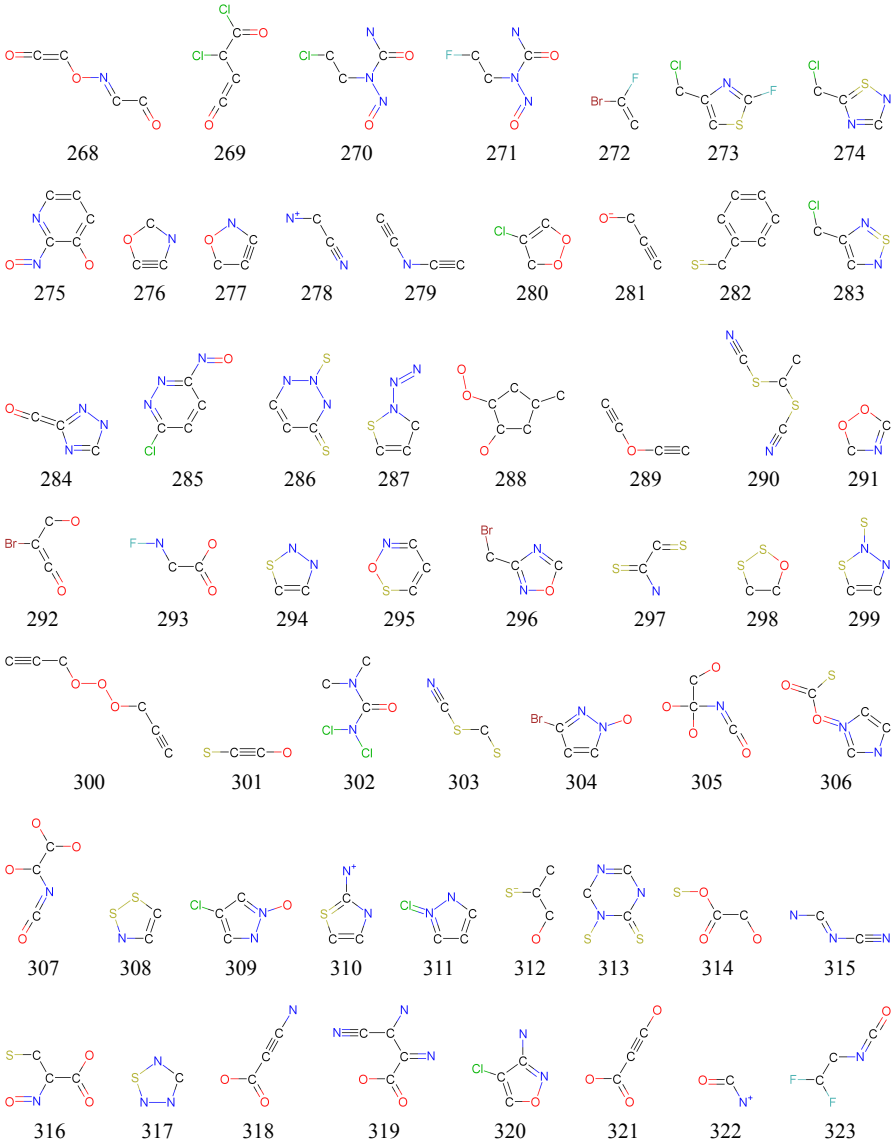

098 099 100 101 102 103 104

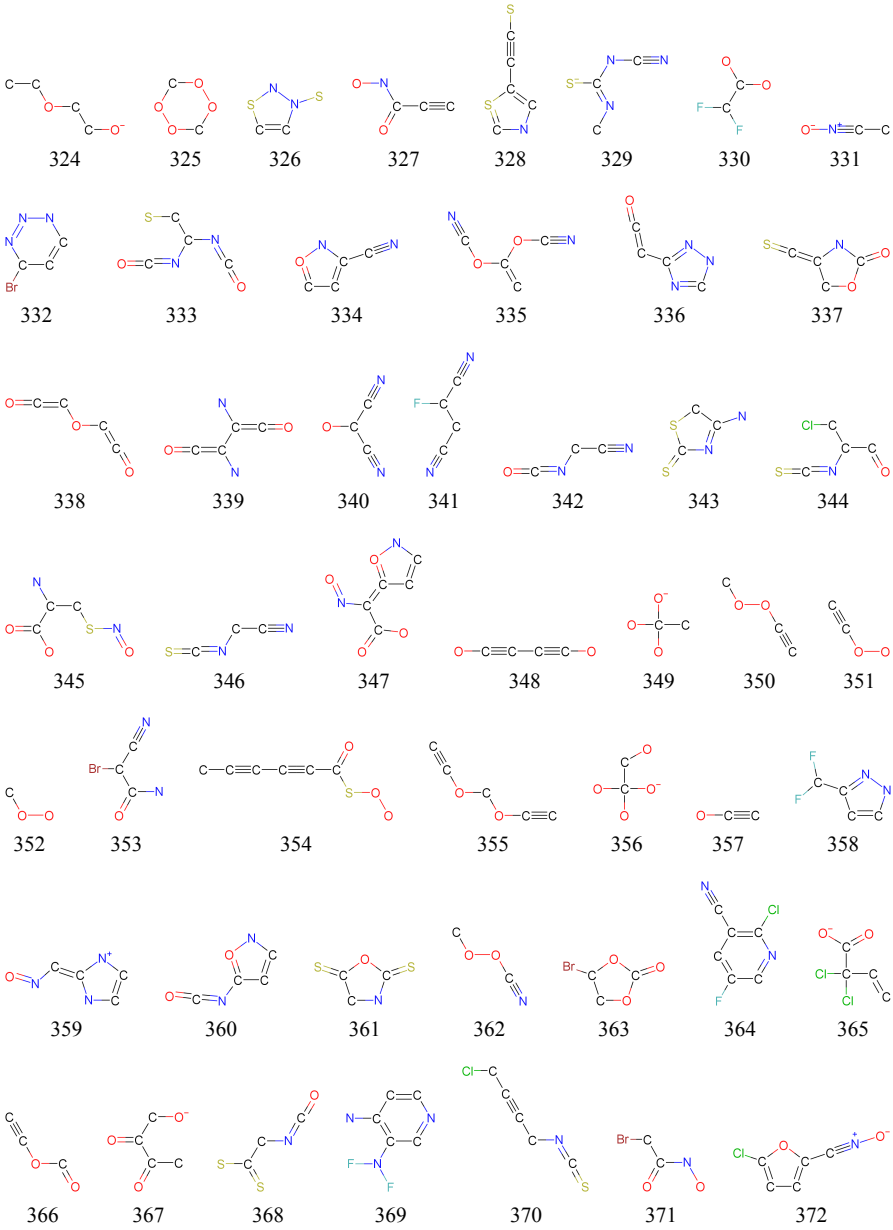

105 106 107 108 109 110 111

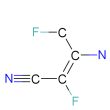

112
113
114
115
116
117
118
119
120
121
122
123
124
125
126
127
128
129
130
131
132
133
134
135
136
137
138
139
140
141
142
143
144
145
146
147
148
149
150
151
152
153
154
155
156
157
158
159
160
161
162
163
164
165

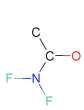

166
167
168
169
170
171
172
173
174
175
176
177
178
179
180
181
182
183
184
185
186
187
188
189
190
191
192
193
194
195
196
197
198
199
200
201
202
203
204
205
206
207
208
209
210
211
212
213
214
215
216
217
218
219

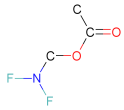

220
221
222
223
224
225
226
227
228
229
230
231
232
233
234
235
236
237
238
239
240
241
242
243
244
245
246
247
248
249
250
251
252
253
254
255
256
257
258
259
260
261
262
263
264
265
266
267

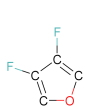

268
269
270
271
272
273
274
275
276
277
278
279
280
281
282
283
284
285
286
287
288
289
290
291
292
293
294
295
296
297
298
299
300
301
302
303
304
305
306
307
308
309
310
311
312
313
314
315
316
317
318
319
320
321
322
323

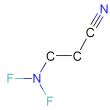

324
325
326
327
328
329
330
331
332
333
334
335
336
337
338
339
340
341
342
343
344
345
346
347
348
349
350
351
352
353
354
355
356
357
358
359
360
361
362
363
364
365
366
367
368
369
370
371
372

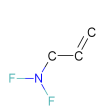

373

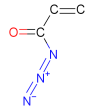

374

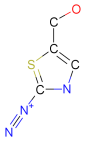

375

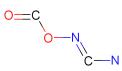

376

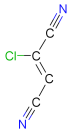

377


378


379

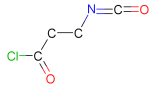

380

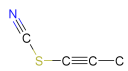

381


382

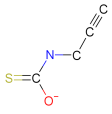

383

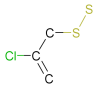

384

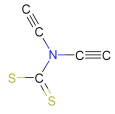

385


386

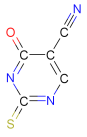

387

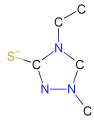

388

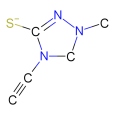

389


390


391


392


393

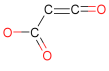

394

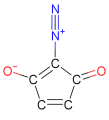

395


396


397

398


399

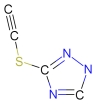

400

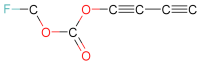

401

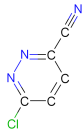

402


403

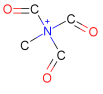

404


405


406

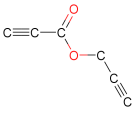

407


408


409


410


411


412

413

414

415

416

417

418

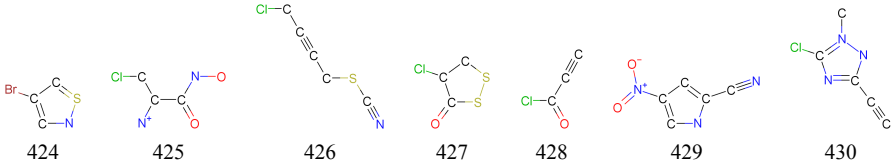
419

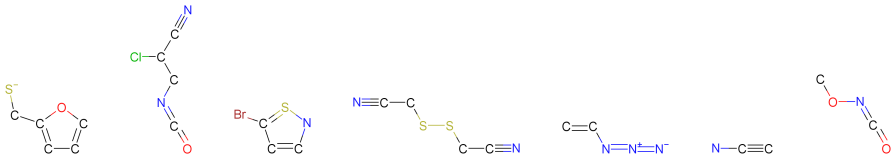
420

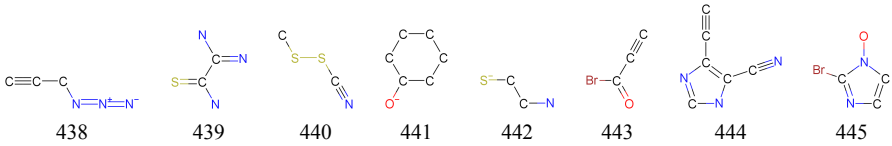
421

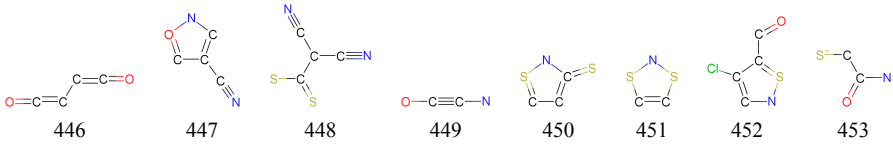
422

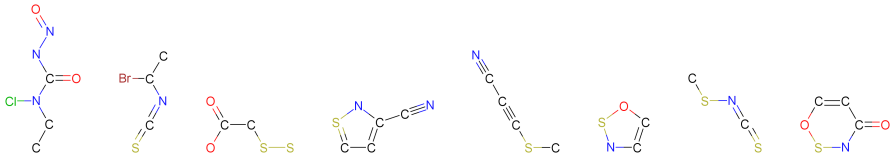
423

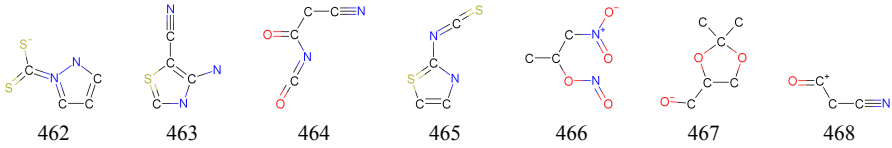

424
425
426
427
428
429
430

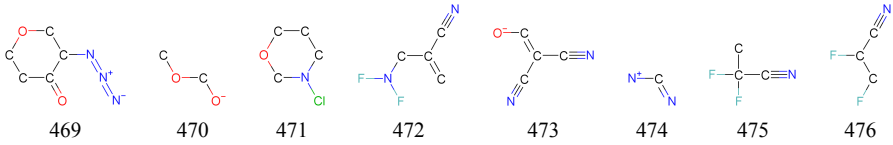

431
432
433
434
435
436
437

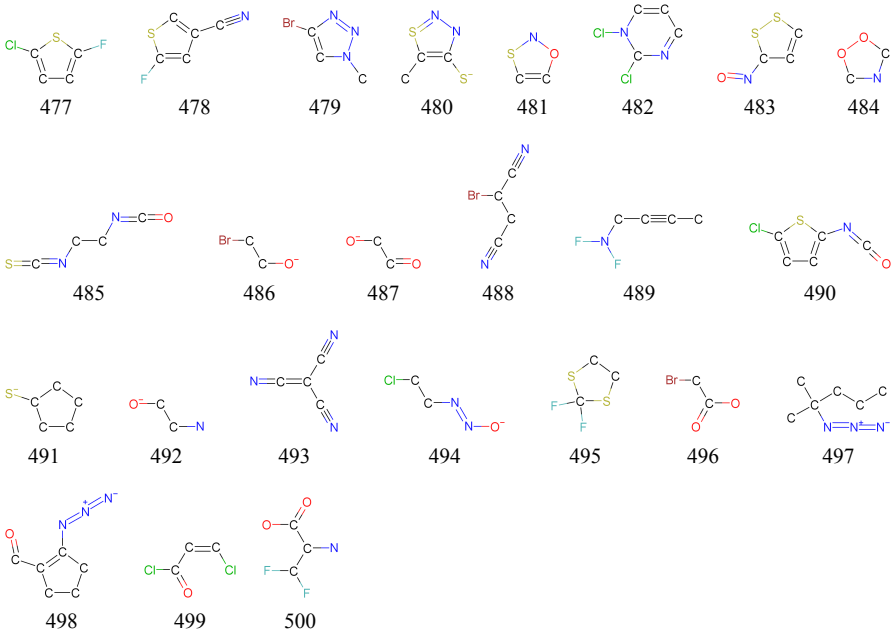

438
439
440
441
442
443
444
445


446
447
448
449
450
451
452
453


454
455
456
457
458
459
460
461


462
463
464
465
466
467
468


469
470
471
472
473
474
475
476



B. Smiles

BrCC#C
ClC(Cl)C#C
Fc1cc([N+](=O)[O-])cc(F)c1
ClCC#CC(=O)O
FCC#C
[NH+](\C=C1\C=CC=C1)(C)C
OC(=O)C#C
Clc1snnc1C
ClC1=C(Cl)C(=O)OC1
O=C(N)C(C#N)C#N
[NH+](C)(C)C(=N)N
FCC#N
S=C=NCCC#N
ClC(N=C=S)C=C
ClC(OO)CCl
BrCCN=C=O
S=C=NCC#C
[SH-]CCO
ClC(=S)SC
Clc1cnc(F)nc1
O=C=C=C=N
BrCC#N
S1CC(=O)N=C1C#N
S(\C=C/C(=O)N)C#N
Brc1sccn1
Clc1sn[n+](c1)C
FC(F)C[NH3+]
Fc1cc(F)cnc1C#N
S=C1N(N=O)CCN1
S(CCSC#N)C#N
BrC(=C)C(=O)O
F/C=C/C#N
ClCC(=C/[N+]#N)/[O-]
Clc1c[nH]nc1
BrCC(=O)C(=O)O
N1=C([NH3+])c2c(C1)cccc2
s1c([NH3+])nnc1
O=C=[NH2+]
FC(=O)C#C
ClCF
ClC(Cl)C(=N)N
S(\C(=N/N)[NH3+])C
S(C(=S)NC#N)C
SCc1[nH][nH]c(=S)n1
S=C=C
C(#C)C#C
BrCF
S1SC=CC1
ClC=C=O
ClC(=C)Cl
S(\C=C\SC#N)C#N
S\C(=C(\S)C#N)C#N
F[C@H]1[C@@H](O)NC(=O)NC1=O
ClCC(=N)C(C#N)C#N
O=C(NC)/C(=N/[O-])C#N
ClC(Cl)c1oncc1
S(CN)C#N
Fc1c(=O)[nH]c(nc1)N
O\N=C(\N)N
BrCCN=C=S
S(O)C#C
O(C(=O)c1nonc1N)C
FC(F)(CN)C(=O)O
S=C=[NH2+]
O(CC#C)\C(=C\[N+]#N)[O-]
S=C(N)C(N)C#N
OC(=O)C#CC(=O)O
S=C=NCCN=C=S
SCC(NN=O)C(=O)O
SC#C
Fc1c(nccc1)[N+](=O)[O-]
ClC(CCl)C#N
SC=S
N\C(=C\C#N)C#N
S=C(N)C(=S)N
N#CC#C
s1ncc(c1)C#N
BrC=C=O
s1c([N+](=O)[O-])ccc1C#C
o1[nH]c(=O)cc1C[NH3+]
Fc1c([N+](=O)[O-])cccc1F
BrC1=C[NH2+]CC=N1
ClN(Cl)CCC(=O)O
ON\C=C/N=O

O=CC#CC#CC=O
ClC(=C)CN=C=S
O/C=C(/C#N)C#N
Cl[13CH2][13C]#[15N]
ClCCN=[N+]=[NH-]
[SH-]CC[N+](C)(C)C
s1cc2nsnc2c1
ClC[C@@H](O)CN=[N+]=[NH-]
S(SCO)CO
BrC(=C)C#N
NC#N
Clc1ncc(F)cn1
F[CH]F
SC(=O)C#C
SCC(N)(N=O)C(=O)O
SCN=C=O
N[N+]#C
S(CSC#N)C#C
O=C(C#C)C#C
Sn1c(=S)[nH]cc1
FC(=C)CN=[N+]=[NH-]
Fn1cc(F)cc1
OC#CNC#CO
ONC(CC#N)C#N
s1c(cc(c1)C#N)C#N
s1c2OSOc2cc1
s1c([NH3+])nc(c1)C(=O)O
O1C([O-])COCC1
[O-][N+](=O)n1cc(cc1)C#N
ClC1=C(S)COC1=O
O=CC(=C=O)C#N
OC(=O)C#CC#CC(=O)O
OCC([O-])C=O
S1N=NC(S)C1
S1SC=CC1C(=O)O
FC=C=N
BrC(C(=O)N)C#C
Fn1[nH]cccn1
s1c(c(S)nc1)C#N
S=C(NC#C)N
O1NNCC1=O
BrCC(=O)NC#N
S1SC=NC1
N(N)C#C
ClC(=O)NC#C
O(NN)C#C
ON(C)C(=O)NOC#C
S(CC(=O)O)C#N
Br/C=C/C(=O)O
OC(=O)CN=N
O=C=NCCN=C=O
[NH3+]C(=N)N
[SH-]CC#C
FN(F)C=C
OC(=O)C#CNC#CC(=O)O
ClN(Cl)C(C)C(=O)O
S1SN=C(N)C1
s1c(c[nH]c1=S)C#N
Clc1occc1N=O
[N+](=[NH-])=C
ClCCl
O(\N=C\C=O)C=C=O
FCC[15N]([15N]=O)C(=O)N
Cl\C=C/C#CCN=[N+]=[NH-]
S\C(=C(\C#N)C#N)S
O(C\C=C\C)\C(=C\[N+]#N)[O-]
O\N=C(/CCCCC)C
N(=C=C(C#N)C#N)C
S(CSC#C)C#C
OC(CN=[N+]=[NH-])C=O
FCC[N+](CC[O-])(C)C
Br[CH]Br
S1C(=O)CNC1=S
Clc1sc(cc1)C#N
Clc1nc(Cl)ncc1
ClCCN=C=S
Cl/C=C/C#N
S1OOCC1
S(CN=C=O)CN=C=O
ON1NN(O)C=C(O)C1
OC#CC#C
S1CN(F)C(=C1)F
ClN(Cl)OC(=C)OC
S1ONC(C1)C(=O)O
ClC(SN=C=O)C
S1SOc2n1ccc2
OC(O)C#C
O(CCOC#N)C#N
S1Oc2c(occ2)O1
O=C1N=C[NH2+]C1
ClC(Cl)([NH3+])CO
ClC(CCl)(C#N)C#N
S/C=C/C(=O)OS
OC#CN(C)C#CO
FC(F)C1=CN(OC1)C#C
ClC([N+](=O)[O-])CCCl
S(N=C=O)C=C
[SH-]C(O)C(O)C
ClC=C=N
[SH-]C(O)C
ClN(CC#N)C=O
s1ncc(N)c1C#N
s1[nH]c(=O)c(c1)C#N
O=CC#CC#C
ClN(/C=N/C)C#N
S=C1NOC(=O)C1
S\C(=[NH+]/S)N
o1c(nnc1C#C)C#C
O(\N=C\N)C#C
OC(=O)C#CC#CO
O(O)CN(N=O)C
s1n([nH]cc1)C#N
S1ONC=C1
FC(=C)F
S=C=NCC(=O)N
[O-]C#C
O=C1NC#CN1
[O-]\C(=N/C#N)CC#N
S1OCC=N1
BrCC#CO
SOn1nccc1
S=C=NCC(F)F
ClC(=C)F
O=C(N)[C@H]1[C@H]([NH3+])CCC1
ClC(C=C=O)C(=O)Cl
BrC(=C)F
O=C(N)C(=C=N)C#N
F\C=C\F
[79Br]\C=C\F
ClC1=C(N)C(OC1=O)O
FC(F)C#CC(=O)O
S\C(=C\C#N)S
O1C2N=NN(O)C(=O)C2C=C1
OC#CO
Fc1nc(F)ncc1
s1c(=S)n(O)cc1
Brc1ncsc1
Brc1[nH]ncc1
FC(F)(OO)C
BrC=C=N
S1SC(=C(C1)C#N)C#N
Clc1ccnnc1Cl
[SH-]C1N(C(N(N1)C)C)C
ClN1N=NCC1
FC(F)C1OC(=O)OC1
[SH-]C(CC(=O)O)C(=O)O
s1c(SC#N)ncc1
[O-]C(C=C)C#N
ClC1NSOC1
S1NNC(=S)C1
S1C(S)NN=C1
S=CC(C(=O)O)C=S
Br\C(=C/C#N)C#N
[O-][N+]#CC=C
SN1OC=CN1
ClC([SH-])CO
SC(=S)C([N+](=O)[O-])C
[N+](=[NH-])=CCC#N
[SH-]CO
FC1(OC=CC1)[N+](=O)[O-]
S=C=CO
OC(=O)C=C=N
FC(F)([O-])C
N=C=CC#N
S=C=CC(=O)O
S1C2=NCC[NH+]2C=C1
S(C(O)C(=O)O)CC#N
O(N)C#C
S(S)NC(=O)C
[NH2+]=C=N
[O-]C(CC)CC
S=C=NCO
O=NN(CC#C)CC#C
S(OOC[O-])C
o1c([NH3+])ncc1
S\N=C(\N)N
S1CC(=O)NC1=S
[O-]CC=C
S(C(=C)SC#N)C#N
[SH-]C(NC(=O)C)C
[O-]CCC#N
On1c(ncc1)C#N
ClNC(CS)C(=O)O
S1SC=NC1N
s1c(N=[N+]=[NH-])ccc1
BrCCl
Fc1cc(F)oc1
ClCCN([15N]=O)C(=O)[15NH2]
ClCc1nc(sc1)F

ClCc1sncn1
O1NC#CC1
ClC1=COOC1
ClCc1nsnc1
SN1NC(=S)C=CN1
O(C#C)C#C
BrC(=C=O)CO
S1ON=CC=C1
S1SOCC1
SC#CO
Brc1nn(O)cc1
OC(N=C=O)C(O)O
s1c([NH3+])ncc1
SN1CN=CNC1=S
SCC(N=O)C(=O)O
OC(=O)C(=N)C(N)C#N
O=C[NH3+]
O1OCOOC1
s1c(cnc1)C#CS
[O-][N+]#CC
o1nc(cc1)C#N
S=C=C1NC(=O)OC1
OC(C#N)C#N
S1CC(=NC1=S)N
S=C=NCC#N
OC(O)([O-])C
O(O)C
O(COC#C)C#C
FC(F)c1[nH]ncc1
S=C1OC(=S)NC1
Clc1ncc(F)cc1C#N
O=C(C(=O)C)C[O-]
ClCC#CCN=C=S
FCC(=C(/F)C#N)/N
Fc1c(F)coc1
O=C(N=[N+]=[NH-])C=C
Cl\C(=C/C#N)C#N
ClC(=O)CCN=C=O
S=C([O-])NCC#C
O(OO)C
[SH-]C1=NN(CN1C#C)C
[SH-]C(CC)(C)C
[O-]C1=C([N+]#N)C(=O)C=C1
S(S)C#C
FCOC(=O)OC#CC#C
O=C[N+](C)(C=O)C=O
O(CC#C)C(=O)C#C
Br/C=C/CN=C=O
OC(=O)CNC(=C/[N+]#N)/[O-]
s1cc(C[NH3+])cc1
Brc1csnc1
ClC1CSSC1=O
Clc1n(nc(n1)C#C)C
Brc1sncc1
NC#C
S=C(N)C(=N)N
[SH-]CCN
Brc1n(O)ccn1
SC(=S)C(C#N)C#N
S1NSC=C1
ClN(CC)C(=O)NN=O
s1nc(cc1)C#N
S(N=C=S)C
Oc1cccnc1N=O
[NH3+]CC#N
[O-]CC#C
O=C=c1[nH][nH]cn1
S1N(N=N)CC=C1
S(C(SC#N)C)C#N
FNCC(=O)O
BrCc1ncon1
S1N(S)NC=C1
ClN(Cl)C(=O)N(C)C
OC(O)(N=C=O)CO
S1SC=CN1
Cln1nccc1
SOC(=O)CO
S1NCNN1
Clc1c(noc1)N
FC(F)CN=C=O
S1NN(S)C=C1
[SH-]\C(=N/C)NC#N
BrC1N=NNC=C1
O(C(=C)OC#N)C#N
O(C=C=O)C=C=O
FC(CC#N)C#N
ClCC(N=C=S)C=O
o1[nH]ccc/1=C(/N=O)C(=O)O
O(OC#C)C
BrC(C(=O)N)C#N
OC(O)([O-])CO
O=N/C=C1/[NH2+]C=CN1
O(OC#N)C
ClC(Cl)(C=C)C(=O)[O-]
SC(=S)CN=C=O
BrCC(=O)NO
FN(F)C(O)C
FN(F)CCC#N
s1c(CO)cnc1[N+]#N
ClC(Cl)([N+](=O)[O-])CC
S(C#CC)C#N
ClC(=C)CSS
S=c1[nH]c(=O)c(c[nH]1)C#N
[SH-]C(C)C
NC#CN
ClN(Cl)CCO
ClC(=O)CS
Clc1nnc(cc1)C#N
N#CCC#N
Br\C=C\Cl
OC[C@@H]([NH3+])[C@H](CC)C
S1SOC(=C1)C
O=C1NN(N=O)CC1
ClC[C@H]([NH3+])C(=O)NO
ClC(=O)C#C
[SH-]Cc1occc1
S(SCC#N)CC#N
O(N=C=O)C
S(SC#N)C
BrC(=O)C#C
O=C=CC=C=O
OC#CN
Clc1c(snc1)C=O
BrC(N=C=S)C
S(C)C#CC#N
S1OC=CC(=O)N1
O1CNC#C1
N(C#C)C#C
[SH-]Cc1ccccc1
Clc1[nH][nH]c(N=O)cc1
O(O)C1CC(CC1O)C
O1OC=NC1
S1NNC=C1
S=C(N)C=S
O(OOCC#C)CC#C
S(CS)C#N
SC(=O)On1ccnc1
Clc1cn(O)nc1
[SH-]C(CO)C
N/C=N/C#N
OC(=O)C#CN
OC(=O)C#CO
O(CC[O-])CC
ONC(=O)C#C
FC(F)C(O)O
SCC(N=C=O)N=C=O
O=C=Cc1[nH]ncn1
O=C=C(N)C(=C=O)N
O=C=NCC#N
S(N=O)C[C@@H](N)C(=O)O
OC#CC#CO
O(O)C#C
S(OO)C(=O)C#CC#CC
OC#C
o1nccc1N=C=O
BrC1OC(=O)OC1
O(C=O)C#C
FN(F)c1c(N)ccnc1
Clc1oc(cc1)C#[N+][O-]
FN(F)COC(=O)C
FN(F)CC=C
O(\N=C\N)C=O
S1SC(=N)CC1=N
Brc1[nH]cnc1N
SC(=S)N(C#C)C#C
[SH-]C1N(CN(N1)C)CC
[SH-]CC
OC(=O)C=C=O
ClNCC(=O)O
S(c1[nH]ncn1)C#C
S=C(NN)N
BrC([N+](=O)[O-])C
FC(F)(N=C=O)C
Cl/C=C/CONCl
ClC1=CC#CC=C1Cl
S1CC(=S)NC1=O
ClCC#CCSC#N
[O-][N+](=O)c1cc([nH]c1)C#N
ClC(CN=C=O)C#N
N(=[N+]=[NH-])C=C
N(=[N+]=[NH-])CC#C
[O-]C1CCCCC1
[nH]1c(c(nc1)C#N)C#C
o1ncc(c1)C#N
s1[nH]c(=S)cc1
[SH-]CC(=O)N
S(S)CC(=O)O
S1OC=CN1
[SH-]C(=S)n1nccc1

s1c(c(nc1)N)C#N
O(N=O)C(C[N+](=O)[O-])C
O1CC(N=[N+]=[NH-])C(=O)CC1
FN(F)CC(=C)C#N
FC(F)(C)C#N
s1cc(cc1F)C#N
S1NOC=C1
O1OCNC1
[O-]CC=O
Clc1sc(N=C=O)cc1
N=C=C(C#N)C#N
BrCC(=O)O
ClC(=O)/C=C\Cl
O=C(N=C=O)CC#N
O1C(OCC1C[O-])(C)C
O(C[O-])C
[O-]/C=C(/C#N)C#N
FC(CF)C#N
Brc1nnn(c1)C
ClC1N(Cl)C=CC=N1
S=C=NCCN=C=O
BrC(CC#N)C#N
[SH-]C1CCCC1
ClCC\N=N\[O-]
N(=[N+]=[NH-])C(CCC)(C)C
FC(F)C(N)C(=O)O
s1c(ncc1)N=C=S
O=[CH+]CC#N
ClN1CCCOC1
[NH3+]C=N
Clc1sc(F)cc1
s1nnc([SH-])c1C
S1SC=CC1N=O
BrCC[O-]
FN(F)CC#CC
[O-]CCN
S1CCSC1(F)F
O=CC1=C(N=[N+]=[NH-])CCC1

Part III: Parameters

METS: Mulliken charges, EEM model, Trivial atom types, Static cost function

Atomic electroneg. [V]		Atomic hardness [V e-1]	
H	7.15	H	19.50
C	2.16	C	12.65
N	9.06	N	19.85
O	28.44	O	40.05
F	26.48	F	39.91
S	2.58	S	9.87
Cl	24.57	Cl	32.08
Br	27.26	Br	33.64

METF: Mulliken charges, EEM model, Trivial atom types, Full cost function

Atomic electroneg. [V]		Atomic hardness [V e-1]	
H	15.06	H	27.41
C	8.27	C	18.76
N	7.96	N	18.75
O	10.15	O	21.75
F	23.16	F	36.60
S	5.36	S	12.65
Cl	12.83	Cl	20.34
Br	12.20	Br	18.59

MEFS: Mulliken charges, EEM model, Force-field atom types, Static cost function

Atomic electroneg. [V]		Atomic hardness [V e-1]	
H	9.81	H	22.16
C4	0.24	C4	10.73
C3	5.73	C3	16.23
C2	4.77	C2	15.26
N4	24.73	N4	35.52
N3	3.84	N3	14.62
N2	3.63	N2	14.42
N1	30.20	N1	40.98
O2	8.68	O2	20.29
O1	11.27	O1	22.87
F	26.18	F	39.62
S2	17.02	S2	24.32
S1	3.70	S1	11.00
Cl	5.90	Cl	13.40
Br	27.28	Br	33.66

MEFF: Mulliken charges, EEM model, Force-field atom types, Full cost function

Atomic electroneg. [V]		Atomic hardness [V e-1]	
H	14.28	H	26.63
C4	8.89	C4	19.39
C3	6.03	C3	16.52
C2	11.29	C2	21.78
N4	28.12	N4	38.91
N3	9.32	N3	20.11
N2	6.75	N2	17.54
N1	7.83	N1	18.62
O2	13.20	O2	24.80
O1	7.49	O1	19.09
F	20.99	F	34.43
S2	6.30	S2	13.60
S1	3.72	S1	11.01
Cl	11.73	Cl	19.24
Br	10.67	Br	17.06

MSTS: Mulliken charges, SQE model, Trivial atom types, Static cost function

Atomic electroneg. [V]		Atomic hardness [V e-1]		Bond hardness [V e-1]		Electroneg. correction [V]	
H	7.59	H	79.05	H-C	5.65	H-C	-4.45
C	-4.28	C	31.16	H-N	105.05	H-N	-0.15
N	20.93	N	47.48	H-O	1.96	H-O	-1.68
O	34.00	O	55.80	H-S	213.13	H-S	2.71
F	59.51	F	100.86	C-C	0.03	C-N	2.43
S	9.61	S	13.31	C-N	14.29	C-O	-3.50
Cl	20.72	Cl	39.38	C-O	28.17	C-F	-11.91
Br	-1.12	Br	45.21	C-F	45.68	C-S	7.15
				C-S	13.11	C-Cl	12.36
				C-Cl	9.42	C-Br	-10.71
				C-Br	1.26	N-O	-7.37
				N-N	8.97	N-F	3.65
				N-O	0.73	N-S	10.56
				N-F	6.94	N-Cl	-14.50
				N-S	41.56	O-S	9.84
				N-Cl	10.61		
				O-O	70.59		
				O-S	42.86		
				S-S	26.03		

MSTF: Mulliken charges, SQE model, Trivial atom types, Full cost function

Atomic electroneg. [V]		Atomic hardness [V e^{-1}]		Bond hardness [V e-1]		Electroneg. correction [V]	
H	8.16			H-C	9.57	H-C	-0.14
C	4.98	H	15.27	H-N	8.27	H-N	-0.28
N	8.22	C	11.22	H-O	8.49	H-O	0.01
O	12.56	N	12.40	H-S	7.25	H-S	-0.88
F	18.18	O	14.13	C-C	6.19	C-N	0.46
S	5.66	F	21.28	C-N	4.59	C-O	0.81
Cl	9.98	S	9.39	C-O	4.88	C-F	-0.16
Br	2.64	Cl	11.12	C-F	9.78	C-S	0.84
		Br	10.62	C-S	2.78	C-Cl	2.08
				C-Cl	5.16	C-Br	-0.67
				C-Br	4.23	N-O	-0.11
				N-N	4.78	N-F	-0.26
				N-O	6.89	N-S	1.38
				N-F	18.23	N-Cl	-2.44
				N-S	7.41	O-S	1.41
				N-Cl	8.05		
				O-O	21.18		
				O-S	12.38		
				S-S	6.68		

MSFS: Mulliken charges, SQE model, Force-field atom types, Static cost function

Atomic electroneg. [V]		Atomic hardness [V e-1]		Bond hardness [V e-1]		Electroneg. correction [V]	
H	6.36	H	21.68	H-C4	284.37	H-C4	-6.68
C4	5.15	C4	14.32	H-C3	10.03	H-C3	0.86
C3	-9.75	C3	19.70	H-C2	0.04	H-C2	3.18
C2	-6.99	C2	12.82	H-N4	26.99	H-N4	-0.88
N4	15.44	N4	47.84	H-N3	86.86	H-N3	-1.13
N3	14.91	N3	62.70	H-N2	62.41	H-N2	-0.79
N2	17.24	N2	35.20	H-O2	82.05	H-O2	-4.41
N1	24.19	N1	14.46	H-S2	15.42	H-S2	-13.24
O2	46.15	O2	77.53	C4-C4	23.66	C4-C3	0.44
O1	44.49	O1	61.43	C4-C3	62.20	C4-C2	5.80
F	55.58	F	100.86	C4-C2	107.37	C4-N4	-4.94
S2	-17.98	S2	43.61	C4-N4	57.35	C4-N3	-4.58
S1	9.21	S1	10.18	C4-N3	36.60	C4-N2	0.33
Cl	15.94	Cl	26.63	C4-N2	16.26	C4-O2	-7.06
Br	-32.54	Br	93.93	C4-O2	57.10	C4-O1	0.67
				C4-O1	3.15	C4-F	-9.78
				C4-F	56.68	C4-S2	3.82
				C4-S2	57.07	C4-S1	2.89
				C4-S1	34.70	C4-Cl	8.62
				C4-Cl	122.69	C4-Br	-3.46
				C4-Br	8.72	C3-C2	3.77
				C3-C3	5.27	C3-N4	2.21
				C3-C2	5.84	C3-N3	1.63
				C3-N4	64.59	C3-N2	2.96
				C3-N3	12.63	C3-O2	0.30
				C3-N2	7.04	C3-O1	3.28
				C3-O2	31.45	C3-F	-1.16
				C3-O1	24.99	C3-S2	4.58
				C3-F	16.67	C3-S1	5.64
				C3-S2	29.73	C3-Cl	9.99
				C3-S1	51.17	C3-Br	-7.04
				C3-Cl	5.79	C2-N3	7.44
				C3-Br	10.45	C2-N2	7.91
				C2-C2	0.00	C2-N1	9.84
				C2-N3	23.88	C2-O2	7.77
				C2-N2	5.70	C2-O1	6.49
				C2-N1	24.88	C2-S2	3.67
				C2-O2	47.60	C2-S1	5.96
				C2-O1	64.97	N3-N2	2.08
				C2-S2	18.35	N3-O2	-8.67
				C2-S1	3.21	N3-O1	-12.18
				N3-N3	51.46	N3-F	5.71
				N3-N2	8.77	N3-S2	10.91
				N3-O2	23.97	N3-Cl	-17.87
				N3-O1	2.84	N2-N1	14.53
				N3-F	10.97	N2-O2	-3.62
				N3-S2	52.05	N2-O1	-5.93
				N3-Cl	32.84	N2-S2	7.82
				N2-N2	85.21	O2-S2	9.74
				N2-N1	62.39		
				N2-O2	33.77		
				N2-O1	18.81		
				N2-S2	27.19		
				O2-O2	62.14		
				O2-S2	46.68		
				S2-S2	72.47		

MSFF: Mulliken charges, SQE model, Force-field atom types, Full cost function

Atomic electroneg. [V]		Atomic hardness [V e-1]		Bond hardness [V e-1]		Electroneg. correction [V]	
H	4.95	H	14.56	H-C4	13.68	H-C4	-0.24
C4	2.73	C4	10.96	H-C3	8.74	H-C3	0.72
C3	-1.44	C3	10.65	H-C2	15.67	H-C2	-3.95
C2	4.34	C2	11.05	H-N4	21.07	H-N4	1.73
N4	19.34	N4	35.20	H-N3	10.12	H-N3	-0.10
N3	5.54	N3	12.28	H-N2	10.86	H-N2	0.61
N2	8.79	N2	11.93	H-O2	9.81	H-O2	0.69
N1	14.89	N1	11.84	H-S2	9.05	H-S2	-1.64
O2	11.52	O2	14.35	C4-C4	7.51	C4-C3	0.46
O1	11.84	O1	12.70	C4-C3	5.86	C4-C2	-0.53
F	15.37	F	18.89	C4-C2	8.94	C4-N4	-0.47
S2	1.06	S2	8.18	C4-N4	13.96	C4-N3	-0.09
S1	5.96	S1	7.73	C4-N3	8.68	C4-N2	0.13
Cl	7.45	Cl	10.15	C4-N2	12.88	C4-O2	-0.24
Br	0.69	Br	9.35	C4-O2	14.31	C4-O1	1.75
				C4-O1	3.28	C4-F	-0.43
				C4-F	13.03	C4-S2	0.64
				C4-S2	11.19	C4-S1	1.92
				C4-S1	0.60	C4-Cl	2.08
				C4-Cl	7.59	C4-Br	-0.99
				C4-Br	6.54	C3-C2	-0.59
				C3-C3	6.85	C3-N4	1.55
				C3-C2	6.23	C3-N3	1.06
				C3-N4	12.39	C3-N2	1.90
				C3-N3	5.59	C3-O2	1.71
				C3-N2	5.45	C3-O1	3.03
				C3-O2	7.05	C3-F	1.47
				C3-O1	4.48	C3-S2	0.98
				C3-F	9.88	C3-S1	2.45
				C3-S2	7.22	C3-Cl	2.74
				C3-S1	2.51	C3-Br	-1.47
				C3-Cl	5.08	C2-N3	0.14
				C3-Br	4.46	C2-N2	1.66
				C2-C2	14.47	C2-N1	4.37
				C2-N3	5.63	C2-O2	1.69
				C2-N2	4.20	C2-O1	2.19
				C2-N1	3.44	C2-S2	-0.52
				C2-O2	7.58	C2-S1	1.11
				C2-O1	5.00	N3-N2	-0.17
				C2-S2	8.93	N3-O2	-0.08
				C2-S1	3.20	N3-O1	0.48
				N3-N3	16.45	N3-F	-0.42
				N3-N2	8.82	N3-S2	0.58
				N3-O2	10.78	N3-Cl	-2.15
				N3-O1	4.24	N2-N1	-1.53
				N3-F	15.85	N2-O2	-0.66
				N3-S2	9.85	N2-O1	-0.99
				N3-Cl	7.85	N2-S2	0.02
				N2-N2	7.92	O2-S2	-0.17
				N2-N1	1.45		
				N2-O2	19.59		
				N2-O1	10.83		
				N2-S2	8.79		
				O2-O2	20.46		
				O2-S2	11.27		
				S2-S2	9.12		

NETS: Natural charges, EEM model, Trivial atom types, Static cost function

Atomic electroneg. [V]		Atomic hardness [V e-1]	
H	7.01	H	19.36
C	0.97	C	11.47
N	1.39	N	12.18
O	3.32	O	14.92
F	25.88	F	39.32
S	0.94	S	8.24
Cl	25.58	Cl	33.09
Br	26.39	Br	32.77

NETF: Natural charges, EEM model, Trivial atom types, Full cost function

Atomic electroneg. [V]		Atomic hardness [V e-1]	
H	22.21	H	34.56
C	15.33	C	25.82
N	9.30	N	20.09
O	12.56	O	24.17
F	29.15	F	42.59
S	7.52	S	14.82
Cl	18.86	Cl	26.37
Br	16.17	Br	22.56

NEFS: Natural charges, EEM model, Force-field atom types, Static cost function

Atomic electroneg. [V]		Atomic hardness [V e-1]	
H	6.75	H	19.10
C4	1.14	C4	11.63
C3	0.85	C3	11.34
C2	1.22	C2	11.71
N4	3.47	N4	14.26
N3	1.29	N3	12.08
N2	0.77	N2	11.56
N1	3.66	N1	14.45
O2	1.77	O2	13.38
O1	3.24	O1	14.84
F	11.28	F	24.72
S2	1.43	S2	8.72
S1	1.65	S1	8.95
Cl	8.71	Cl	16.22
Br	14.82	Br	21.21

NEFF: Natural charges, EEM model, Force-field atom types, Full cost function

Atomic electroneg. [V]		Atomic hardness [V e-1]	
H	26.15	H	38.50
C4	65.55	C4	76.04
C3	11.20	C3	21.69
C2	6.55	C2	17.04
N4	45.42	N4	56.21
N3	17.36	N3	28.15
N2	8.14	N2	18.93
N1	4.76	N1	15.55
O2	19.54	O2	31.15
O1	8.90	O1	20.51
F	32.41	F	45.85
S2	8.79	S2	16.08
S1	5.55	S1	12.84
Cl	19.20	Cl	26.71
Br	15.83	Br	22.21

NSTS: Natural charges, SQE model, Trivial atom types, Static cost function

Atomic electroneg. [V]		Atomic hardness [V e-1]		Bond hardness [V e-1]		Electroneg. correction [V]	
H	0.22	H	21.98	H-C	5.45	H-C	-0.14
C	5.47	C	11.48	H-N	0.27	H-N	0.33
N	7.38	N	12.54	H-O	0.78	H-O	-0.29
O	9.37	O	16.56	H-S	11.41	H-S	0.21
F	19.53	F	41.06	C-C	1.54	C-N	-0.09
S	4.86	S	8.11	C-N	1.89	C-O	-0.29
Cl	4.53	Cl	12.03	C-O	0.00	C-F	-0.46
Br	1.80	Br	23.33	C-F	6.08	C-S	0.29
				C-S	1.06	C-Cl	0.01
				C-Cl	29.27	C-Br	-0.32
				C-Br	23.97	N-O	-0.40
				N-N	4.63	N-F	0.81
				N-O	0.17	N-S	0.49
				N-F	24.29	N-Cl	-0.01
				N-S	1.92	O-S	0.35
				N-Cl	3.94		
				O-O	31.54		
				O-S	0.28		
				S-S	25.37		

NSTF: Natural charges, SQE model, Trivial atom types, Full cost function

Atomic electroneg. [V]		Atomic hardness [V e-1]		Bond hardness [V e-1]		Electroneg. correction [V]	
H	1.49	H	16.67	H-C	13.47	H-C	1.15
C	6.40	C	17.74	H-N	12.08	H-N	-0.28
N	8.94	N	14.38	H-O	13.73	H-O	-0.43
O	10.82	O	15.63	H-S	8.25	H-S	1.28
F	14.27	F	24.48	C-C	6.45	C-N	-0.93
S	8.34	S	10.50	C-N	4.27	C-O	-2.05
Cl	6.68	Cl	13.08	C-O	6.24	C-F	-3.32
Br	6.78	Br	12.85	C-F	17.04	C-S	1.95
				C-S	3.12	C-Cl	0.53
				C-Cl	8.92	C-Br	-1.64
				C-Br	6.45	N-O	-1.04
				N-N	3.93	N-F	1.44
				N-O	6.41	N-S	2.44
				N-F	14.86	N-Cl	-0.64
				N-S	8.30	O-S	2.57
				N-Cl	9.58		
				O-O	15.57		
				O-S	13.24		
				S-S	7.33		

NSFS: Natural charges, SQE model, Force-field atom types, Static cost function

Atomic electroneg. [V]		Atomic hardness [V e-1]		Bond hardness [V e-1]		Electroneg. correction [V]	
H	-3.01	H	35.88	H-C4	12.82	H-C4	0.92
C4	8.77	C4	21.24	H-C3	11.59	H-C3	0.25
C3	8.13	C3	17.95	H-C2	1.42	H-C2	-0.94
C2	7.03	C2	15.46	H-N4	0.00	H-N4	-4.50
N4	1.59	N4	29.45	H-N3	3.74	H-N3	-3.30
N3	7.00	N3	22.20	H-N2	14.82	H-N2	-3.47
N2	8.39	N2	19.41	H-O2	6.48	H-O2	-5.18
N1	8.03	N1	14.42	H-S2	12.05	H-S2	-0.56
O2	9.31	O2	25.34	C4-C4	0.00	C4-C3	-0.45
O1	11.82	O1	19.93	C4-C3	0.00	C4-C2	-0.77
F	16.44	F	47.00	C4-C2	4.20	C4-N4	-2.92
S2	2.95	S2	12.64	C4-N4	0.00	C4-N3	-2.31
S1	7.39	S1	8.89	C4-N3	0.34	C4-N2	-2.45
Cl	5.22	Cl	17.97	C4-N2	10.07	C4-O2	-4.10
Br	3.45	Br	19.04	C4-O2	3.60	C4-O1	-4.11
				C4-O1	0.00	C4-F	-5.70
				C4-F	5.01	C4-S2	1.00
				C4-S2	28.24	C4-S1	1.34
				C4-S1	0.00	C4-Cl	-0.78
				C4-Cl	11.49	C4-Br	-0.22
				C4-Br	38.92	C3-C2	-0.58
				C3-C3	0.00	C3-N4	-2.81
				C3-C2	3.43	C3-N3	-2.03
				C3-N4	0.03	C3-N2	-1.79
				C3-N3	2.88	C3-O2	-3.20
				C3-N2	0.84	C3-O1	-5.03
				C3-O2	7.15	C3-F	-4.74
				C3-O1	9.11	C3-S2	0.85
				C3-F	5.72	C3-S1	1.36
				C3-S2	8.56	C3-Cl	-0.33
				C3-S1	0.44	C3-Br	-0.41
				C3-Cl	3.74	C2-N3	-2.17
				C3-Br	7.95	C2-N2	-2.76
				C2-C2	0.53	C2-N1	-1.38
				C2-N3	6.46	C2-O2	-3.62
				C2-N2	13.48	C2-O1	-7.18
				C2-N1	5.19	C2-S2	0.07
				C2-O2	19.13	C2-S1	1.63
				C2-O1	28.49	N3-N2	0.33
				C2-S2	4.84	N3-O2	-2.01
				C2-S1	2.44	N3-O1	-3.67
				N3-N3	1.69	N3-F	2.82
				N3-N2	2.76	N3-S2	2.93
				N3-O2	15.64	N3-Cl	-1.68
				N3-O1	15.80	N2-N1	6.25
				N3-F	12.96	N2-O2	-1.42
				N3-S2	15.84	N2-O1	-4.06
				N3-Cl	7.04	N2-S2	2.89
				N2-N2	12.45	O2-S2	3.71
				N2-N1	178.15		
				N2-O2	23.54		
				N2-O1	18.08		
				N2-S2	11.01		
				O2-O2	15.05		
				O2-S2	14.70		
				S2-S2	34.12		

NSFF: Natural charges, SQE model, Force-field atom types, Full cost function

Atomic electroneg. [V]		Atomic hardness [V e-1]		Bond hardness [V e-1]		Electroneg. correction [V]	
H	0.00	H	15.78	H-C4	10.80	H-C4	-0.26
C4	6.96	C4	13.92	H-C3	15.23	H-C3	-0.39
C3	8.06	C3	14.00	H-C2	20.54	H-C2	-0.58
C2	7.52	C2	12.59	H-N4	18.63	H-N4	-3.19
N4	9.99	N4	33.40	H-N3	15.57	H-N3	-0.89
N3	8.17	N3	14.79	H-N2	19.72	H-N2	-1.12
N2	9.61	N2	14.89	H-O2	17.56	H-O2	-1.62
N1	6.16	N1	11.80	H-S2	9.57	H-S2	0.10
O2	10.32	O2	16.28	C4-C4	16.90	C4-C3	-0.34
O1	9.12	O1	12.79	C4-C3	18.25	C4-C2	-0.50
F	13.89	F	21.06	C4-C2	20.39	C4-N4	-1.47
S2	3.69	S2	9.56	C4-N4	19.35	C4-N3	-1.00
S1	8.75	S1	8.14	C4-N3	17.61	C4-N2	-1.09
Cl	5.79	Cl	11.55	C4-N2	19.11	C4-O2	-1.78
Br	4.87	Br	10.78	C4-O2	17.12	C4-O1	-2.21
				C4-O1	6.79	C4-F	-2.27
				C4-F	18.02	C4-S2	0.35
				C4-S2	11.25	C4-S1	0.71
				C4-S1	5.76	C4-Cl	-0.19
				C4-Cl	9.78	C4-Br	-0.10
				C4-Br	7.24	C3-C2	-0.30
				C3-C3	3.93	C3-N4	-2.11
				C3-C2	10.09	C3-N3	-0.99
				C3-N4	20.17	C3-N2	-0.70
				C3-N3	7.48	C3-O2	-1.43
				C3-N2	3.27	C3-O1	-2.48
				C3-O2	10.29	C3-F	-2.41
				C3-O1	4.88	C3-S2	0.18
				C3-F	15.89	C3-S1	1.45
				C3-S2	6.68	C3-Cl	-0.24
				C3-S1	2.03	C3-Br	-0.15
				C3-Cl	9.32	C2-N3	-0.93
				C3-Br	7.92	C2-N2	-0.47
				C2-C2	0.19	C2-N1	-1.55
				C2-N3	10.26	C2-O2	-1.37
				C2-N2	0.92	C2-O1	-1.80
				C2-N1	2.16	C2-S2	0.11
				C2-O2	12.36	C2-S1	1.18
				C2-O1	4.39	N3-N2	0.01
				C2-S2	10.31	N3-O2	-0.81
				C2-S1	1.71	N3-O1	-1.47
				N3-N3	14.67	N3-F	0.70
				N3-N2	5.97	N3-S2	0.97
				N3-O2	14.20	N3-Cl	-0.52
				N3-O1	4.92	N2-N1	2.17
				N3-F	11.92	N2-O2	-0.59
				N3-S2	10.75	N2-O1	-2.13
				N3-Cl	9.03	N2-S2	0.73
				N2-N2	1.70	O2-S2	1.14
				N2-N1	1.77		
				N2-O2	9.03		
				N2-O1	4.94		
				N2-S2	6.39		
				O2-O2	12.56		
				O2-S2	11.42		
				S2-S2	8.80		

HETS: Hirshfeld-I charges, EEM model, Trivial atom types, Static cost function

Atomic electroneg. [V]		Atomic hardness [V e-1]	
H	3.18	H	15.54
C	0.25	C	10.74
N	0.14	N	10.93
O	1.24	O	12.85
F	24.93	F	38.37
S	0.57	S	7.86
Cl	24.38	Cl	31.89
Br	25.56	Br	31.94

HETF: Hirshfeld-I charges, EEM model, Trivial atom types, Full cost function

Atomic electroneg. [V]		Atomic hardness [V e-1]	
H	10.26	H	22.61
C	6.83	C	17.33
N	3.58	N	14.36
O	5.59	O	17.19
F	17.86	F	31.29
S	4.51	S	11.80
Cl	12.11	Cl	19.62
Br	11.95	Br	18.33

HEFS: Hirshfeld-I charges, EEM model, Force-field atom types, Static cost function

Atomic electroneg. [V]		Atomic hardness [V e-1]	
H	2.59	H	14.94
C4	0.30	C4	10.79
C3	0.15	C3	10.65
C2	0.29	C2	10.78
N4	0.01	N4	10.79
N3	0.07	N3	10.86
N2	0.03	N2	10.82
N1	1.25	N1	12.04
O2	0.05	O2	11.65
O1	2.13	O1	13.73
F	6.25	F	19.68
S2	1.25	S2	8.55
S1	1.95	S1	9.24
Cl	4.51	Cl	12.02
Br	13.12	Br	19.51

HEFF: Hirshfeld-I charges, EEM model, Force-field atom types, Full cost function

Atomic electroneg. [V]		Atomic hardness [V e-1]	
H	12.06	H	24.41
C4	69.36	C4	79.85
C3	5.13	C3	15.62
C2	1.35	C2	11.84
N4	41.05	N4	51.84
N3	6.20	N3	16.98
N2	2.34	N2	13.12
N1	1.88	N1	12.67
O2	6.65	O2	18.25
O1	3.96	O1	15.56
F	16.32	F	29.76
S2	4.63	S2	11.92
S1	3.96	S1	11.25
Cl	11.76	Cl	19.27
Br	11.01	Br	17.40

HSTS: Hirshfeld-I charges, SQE model, Trivial atom types, Static cost function

Atomic electroneg. [V]		Atomic hardness [V e-1]		Bond hardness [V e-1]		Electroneg. correction [V]	
H	6.18	H	15.71	H-C	2.13	H-C	-0.08
C	7.66	C	10.49	H-N	0.35	H-N	0.15
N	8.65	N	10.81	H-O	0.54	H-O	-0.23
O	9.53	O	13.28	H-S	28.55	H-S	-0.25
F	17.80	F	34.97	C-C	0.75	C-N	0.10
S	7.46	S	8.09	C-N	0.53	C-O	0.03
Cl	9.81	Cl	27.50	C-O	0.00	C-F	-0.19
Br	7.36	Br	31.46	C-F	18.50	C-S	0.06
				C-S	0.04	C-Cl	-0.21
				C-Cl	23.13	C-Br	0.11
				C-Br	26.43	N-O	-0.11
				N-N	0.00	N-F	0.36
				N-O	0.00	N-S	0.12
				N-F	30.61	N-Cl	0.22
				N-S	0.24	O-S	0.33
				N-Cl	28.85		
				O-O	30.11		
				O-S	7.59		
				S-S	26.46		

HSTF: Hirshfeld-I charges, SQE model, Trivial atom types, Full cost function

Atomic electroneg. [V]		Atomic hardness [V e-1]		Bond hardness [V e-1]		Electroneg. correction [V]	
H	6.46	H	14.13	H-C	7.09	H-C	0.52
C	7.30	C	12.36	H-N	5.49	H-N	-0.14
N	9.23	N	11.38	H-O	6.55	H-O	-0.54
O	9.48	O	12.67	H-S	4.67	H-S	0.37
F	8.90	F	18.35	C-C	5.79	C-N	-0.14
S	8.75	S	8.88	C-N	3.73	C-O	-0.70
Cl	7.18	Cl	10.73	C-O	4.48	C-F	-1.35
Br	7.23	Br	10.81	C-F	9.32	C-S	0.61
				C-S	3.10	C-Cl	-0.43
				C-Cl	6.08	C-Br	-0.06
				C-Br	4.41	N-O	-0.74
				N-N	3.66	N-F	1.21
				N-O	4.40	N-S	0.72
				N-F	8.01	N-Cl	0.60
				N-S	4.59	O-S	0.70
				N-Cl	5.44		
				O-O	7.46		
				O-S	6.53		
				S-S	3.91		

HSFS: Hirshfeld-I charges, SQE model, Force-field atom types, Static cost function

Atomic electroneg. [V]		Atomic hardness [V e-1]		Bond hardness [V e-1]		Electroneg. correction [V]	
H	3.16	H	16.80	H-C4	2.58	H-C4	0.15
C4	4.21	C4	11.42	H-C3	6.17	H-C3	0.09
C3	5.03	C3	10.55	H-C2	0.68	H-C2	-0.09
C2	5.20	C2	10.50	H-N4	0.01	H-N4	-1.78
N4	-1.14	N4	12.78	H-N3	1.55	H-N3	-1.09
N3	3.46	N3	11.98	H-N2	11.59	H-N2	-1.59
N2	5.40	N2	10.80	H-O2	4.04	H-O2	-1.59
N1	7.60	N1	11.48	H-S2	9.48	H-S2	-0.68
O2	5.38	O2	13.47	C4-C4	1.51	C4-C3	-0.18
O1	10.11	O1	13.89	C4-C3	3.14	C4-C2	-0.13
F	9.64	F	21.40	C4-C2	6.39	C4-N4	-0.80
S2	3.85	S2	9.18	C4-N4	3.55	C4-N3	-0.41
S1	6.44	S1	9.11	C4-N3	1.32	C4-N2	-0.44
Cl	4.71	Cl	13.43	C4-N2	7.70	C4-O2	-0.55
Br	4.62	Br	8.01	C4-O2	5.44	C4-O1	-0.15
				C4-O1	3.00	C4-F	-0.75
				C4-F	14.79	C4-S2	-0.01
				C4-S2	2.69	C4-S1	0.34
				C4-S1	0.00	C4-Cl	-0.38
				C4-Cl	1.12	C4-Br	0.17
				C4-Br	10.08	C3-C2	-0.06
				C3-C3	3.34	C3-N4	-0.93
				C3-C2	3.61	C3-N3	-0.45
				C3-N4	3.74	C3-N2	-0.34
				C3-N3	2.88	C3-O2	-0.47
				C3-N2	4.59	C3-O1	-0.55
				C3-O2	7.02	C3-F	-0.63
				C3-O1	8.06	C3-S2	-0.09
				C3-F	20.89	C3-S1	0.17
				C3-S2	2.24	C3-Cl	-0.28
				C3-S1	1.29	C3-Br	0.12
				C3-Cl	0.00	C2-N3	-0.42
				C3-Br	18.66	C2-N2	-0.28
				C2-C2	2.38	C2-N1	0.23
				C2-N3	2.94	C2-O2	-0.85
				C2-N2	5.19	C2-O1	-1.75
				C2-N1	2.09	C2-S2	0.02
				C2-O2	40.62	C2-S1	0.15
				C2-O1	18.70	N3-N2	-0.03
				C2-S2	4.57	N3-O2	-0.17
				C2-S1	3.70	N3-O1	-0.25
				N3-N3	2.87	N3-F	0.34
				N3-N2	4.28	N3-S2	0.53
				N3-O2	11.03	N3-Cl	-0.22
				N3-O1	11.10	N2-N1	2.26
				N3-F	32.74	N2-O2	0.02
				N3-S2	2.40	N2-O1	-0.94
				N3-Cl	12.56	N2-S2	0.49
				N2-N2	7.05	O2-S2	0.53
				N2-N1	30.22		
				N2-O2	2.28		
				N2-O1	20.55		
				N2-S2	5.20		
				O2-O2	16.49		
				O2-S2	4.54		
				S2-S2	15.91		

HSFF: Hirshfeld-I charges, SQE model, Force-field atom types, Full cost function

Atomic electroneg. [V]		Atomic hardness [V e-1]		Bond hardness [V e-1]		Electroneg. correction [V]	
H	3.49	H	13.60	H-C4	6.65	H-C4	0.11
C4	4.56	C4	10.59	H-C3	8.68	H-C3	0.21
C3	5.70	C3	11.57	H-C2	8.66	H-C2	0.06
C2	6.32	C2	11.12	H-N4	8.81	H-N4	-2.14
N4	-1.41	N4	12.43	H-N3	7.29	H-N3	-1.06
N3	3.90	N3	11.00	H-N2	7.42	H-N2	-0.99
N2	6.14	N2	11.57	H-O2	8.43	H-O2	-1.49
N1	6.17	N1	11.61	H-S2	5.62	H-S2	-0.28
O2	5.40	O2	12.13	C4-C4	12.62	C4-C3	-0.22
O1	6.62	O1	12.38	C4-C3	13.41	C4-C2	-0.17
F	7.00	F	17.07	C4-C2	15.30	C4-N4	-0.82
S2	4.47	S2	8.21	C4-N4	15.53	C4-N3	-0.45
S1	7.86	S1	8.08	C4-N3	11.89	C4-N2	-0.48
Cl	5.38	Cl	10.03	C4-N2	11.49	C4-O2	-0.60
Br	5.19	Br	9.47	C4-O2	10.68	C4-O1	-0.76
				C4-O1	4.10	C4-F	-0.74
				C4-F	9.44	C4-S2	0.04
				C4-S2	7.94	C4-S1	0.39
				C4-S1	3.97	C4-Cl	-0.29
				C4-Cl	6.71	C4-Br	0.09
				C4-Br	5.30	C3-C2	-0.13
				C3-C3	3.75	C3-N4	-1.29
				C3-C2	9.64	C3-N3	-0.62
				C3-N4	16.23	C3-N2	-0.47
				C3-N3	6.45	C3-O2	-0.71
				C3-N2	3.02	C3-O1	-1.33
				C3-O2	7.62	C3-F	-1.09
				C3-O1	2.84	C3-S2	-0.02
				C3-F	9.71	C3-S1	0.58
				C3-S2	5.15	C3-Cl	-0.35
				C3-S1	1.66	C3-Br	0.10
				C3-Cl	6.86	C2-N3	-0.59
				C3-Br	5.77	C2-N2	-0.40
				C2-C2	0.51	C2-N1	-0.99
				C2-N3	7.68	C2-O2	-0.67
				C2-N2	1.91	C2-O1	-1.39
				C2-N1	1.56	C2-S2	0.11
				C2-O2	8.81	C2-S1	0.44
				C2-O1	3.60	N3-N2	-0.07
				C2-S2	8.80	N3-O2	-0.16
				C2-S1	2.02	N3-O1	-0.20
				N3-N3	9.61	N3-F	0.08
				N3-N2	5.09	N3-S2	0.58
				N3-O2	9.20	N3-Cl	-0.28
				N3-O1	3.68	N2-N1	0.99
				N3-F	7.55	N2-O2	-0.21
				N3-S2	6.90	N2-O1	-0.97
				N3-Cl	5.69	N2-S2	0.29
				N2-N2	2.46	O2-S2	0.57
				N2-N1	1.63		
				N2-O2	6.12		
				N2-O1	3.65		
				N2-S2	3.88		
				O2-O2	7.95		
				O2-S2	6.83		
				S2-S2	5.26		

Part IV: Z-matrixes of the chain molecules

The Z-matrixes given below are convenient descriptions of chain molecules. Each line defines the position of a new atom relative to the previous atoms. The Z-matrices differ slightly from the conventional notation, that is the references to previous atoms are relative. The number of rows one has to go back is given as a negative number, e.g. -1 stands for the previous row, -2 stands for two rows earlier and so on. Another difference is that all the distances, angles and dihedral angles of the first three lines of a monomer are present. They fix the position of a monomer with respect to a previous monomer in the chain. Using these conventions, a straightforward concatenation of monomer Z-matrixes results in the Z-matrix of a chain molecule. A few additional atoms are required to terminate each molecule properly. The corresponding lines in the Z-matrix are also given. In the termination fragment, the symbol M stands for the number of monomers in the chain molecule.

Alkane

Fragment	Element	Distance [Å]	Reference 1	Angle [°]	Reference 2	Dihedral angle [°]	Reference 3
Monomer	C	1.53	-3	109.47	-1	-120.00	-2
	H	1.11	-1	109.47	-4	-180.00	-3
	H	1.11	-2	109.47	-1	120.00	-5
Termination	H	1.11	-3*M	109.47	-2	-120.00	-1
	H	1.11	-4	109.47	-3	120.00	-2

Conjugated alkene

Fragment	Element	Distance [Å]	Reference 1	Angle [°]	Reference 2	Dihedral angle [°]	Reference 3
Monomer	C	1.37	-2	123.00	-4	0.00	-3
	H	1.10	-1	118.50	-3	-180.00	-2
	C	1.37	-2	118.50	-1	-180.00	-3
	H	1.10	-1	118.50	-3	180.00	-2
Termination	H	1.10	-4*M	117.69	-3	-180.00	-2
	H	1.10	-3	117.69	-2	-180.00	-4

Alanine alpha helix

Fragment	Element	Distance [Å]	Reference 1	Angle [°]	Reference 2	Dihedral angle [°]	Reference 3
Monomer	N	1.33	-8	116.00	-9	-50.00	-20
	C	1.46	-1	122.00	-9	-175.00	-10
	C	1.53	-1	107.96	-2	-60.00	-10
	O	1.21	-1	123.47	-2	128.07	-3
	C	1.53	-3	113.04	-2	120.82	-4
	H	1.11	-4	108.25	-1	-115.41	-3
	H	1.02	-6	120.00	-5	-180.00	-14
	H	1.11	-3	113.24	-6	-49.50	-2
	H	1.10	-4	106.60	-1	-120.19	-7
	H	1.10	-5	107.97	-1	117.53	-2
Termination	H	1.02	-M	109.47	-9	-120.00	-4
	H	1.02	-M-1	109.47	-10	120.00	-5
	O	1.21	-10	120.00	-9	-180.00	-11

Paper 5: "Assessment of Polarizability Extent for Protein-Ligand Binding Through an Extended Electronegativity Equalization Model"

Toon Verstraelen, Ewald, Pauwels, Michel Waroquier, Veronique Van Speybroeck, authors ALGC

In preparation

Ligand-protein binding

DOI: 10.1002/anie.200((will be filled in by the editorial staff))

Assessment of polarizability extent for protein-ligand binding through an Extended Electronegativity Equalization model

Toon Verstraelen, Ewald Pauwels, Michel Waroquier, Veronique Van Speybroeck et authors ALGC*

The functions of many proteins involved in signal transduction, transport, and catalysis rely upon the specific recognition of small ligands. These proteins must differentiate between the molecule of physiological interest and a myriad of other similar species.[1] Various properties of candidate molecules contribute to ligand recognition such as the shape, charge and polarity. Even though the capacities of computational chemistry are considerably expanded, the calculation of electronic polarization of biomolecules is still a challenging task, as several models show an exaggerated overestimation of this property in terms of the molecular size.[2-5] In this communication, we present an extended charge equilibration model which is able to predict polarizabilities of biomolecules. The proof of the concept is shown by calculating polarizabilities for a range of pharmaceutically relevant protein binding sites. This work serves as a step-stone for developing polarizable molecular mechanics force fields.

Molecular modelling of biomolecules has become within reach the last years due the steady increase of computational power but also due to the ongoing improvement of force fields.[6] These functions describe the interaction energy potential of the molecular system in terms of the atomic coordinates.[7] To describe however subtle effects such as docking of competing drugs or inhibitors in the recognition site of a protein, the force fields need to describe correctly polarization, i.e. the response of the molecule to an external field.[8-10] One promising solution is the use of variable atom-centred charges instead of fixed ones.[11-14]

The charge equilibration methods (CE) or electronegativity equalization method (EEM) provide a means for introducing such variable charges as they redistribute the charge within a molecule upon response to conformational and environmental changes.[15-17] However the up scaling and transferability towards massively large systems, found in biochemistry, has been prevented by a serious caveat of EEM methods, as super linear polarizability scaling was detected.[18,3-4] The underlying reason for this behaviour relies on the fact that EEM methods treat molecules more as conductors rather than insulators, which causes charge to flow freely among the entire molecular system as response to an external perturbation. Various methods have been proposed to cure the problem as explained in detail in reference.[3] The extended EEM method used in this communication is very similar to the method introduced by Nistor and co-workers, and relies on the introduction of split charges.[19-20] The effective charge on an atom is expressed as the sum of charge flows of the given atom with its neighbours. For details on the implementation of our method we refer to a forthcoming publication.[21]

Before shifting to the pharmaceutically relevant biomolecules, the scaling behaviour of the polarizability is investigated for a simple model system i.e. an α-helical alanine polypeptide consisting of an increasing number of L-alanine residues. The largest value of the polarizability tensor, which in this case points along the long axis of the helix, is plotted in terms of the number of alanine residues in Figure 1 both for the original EEM model as with our model (SQE).

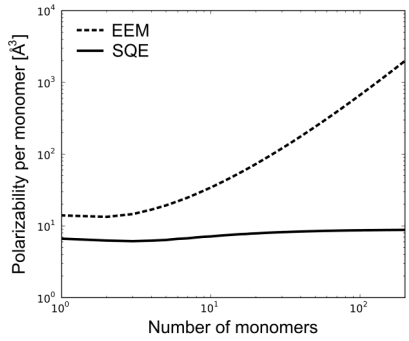


Figure 1. Scaling for the largest value of the polarizibility tensor along alanine helix axis for the original EEM model and our model (SQE).

As earlier found the molecular polarizability scales cubically in terms of the number of monomers for the standard EEM method. The extended version predicts a asymptotic saturation of the per monomer polarizability, which is in line with the behaviour of real conjugated polymers and molecular chains as predicted by ab initio results.[5,22-24]

At this point, we can assess the importance of polarizability in drug-ligand interactions. For this purpose we have selected 5

[*] Ir. Toon Verstraelen, Dr. Ewald Pauwels, Prof. Dr. Michel Waroquier and Prof. Dr. Ir. Veronique Van Speybroeck, Center for Molecular Modeling, Proeftuinstraat 86, 9000 Gent, Belgium, Email :
Fax: (+) 32-9-2646697
E-mail: Veronique.vanspeybroeck@ugent.be
Homepage: http://molmod.ugent.be

Authors ALGC

[**] This work was supported by the Fund for Scientific Research Flanders (FWO), the Research Board of the Ghent University (BOF) and BELSPO, in the frame of network IAP 6/27. TV also acknowledges the Institute for the Promotion of Innovation through Science and Technology in Flanders (IWT-Vlaanderen) for funding this strategic basic research (SBO).

Supporting information for this article is available on the WWW under http://www.angewandte.org or from the author. (Please delete if not appropriate)

proteins which can be categorized into 3 superfamilies. In an effort to explore the landscape of diverse protein binding sites, the selection of proteins was based on a divergent and functionally distinct set of protein superfamilies. The selected proteins belong to the following classes with varying functional activity: the nuclear hormone receptors, serine proteases and aspartyl proteases as tabulated in Table 1.

Table 1. Studie proteins from the Protein Data Bank (PDB)

PDB	Protein	Ligand/substrate	Reference
1I37	Androgen Receptor	Dihydrotestosterone	Sack, 2001
1CVW	Coagulation factor viia	Chloromethyl ketone	Kemball-Cook 1999
2BPV	HIV-1 protease	L-738, 317	Munshi, 1998
1HSG	HIV-1 protease	L-735, 524	Chen, 1994
1HRN	Renin	BILA 980	Tong 1995

The polarizability tensor is a molecular property but the binding capacity of a ligand in a protein is expected to be governed by the local properties of the constituent amino acids of the binding site. Here we evaluate the polarization energy of the binding site as follows. An external perturbation is introduced by placing a gaussian charge distribution with a net total charge +1 at the center of mass of the ligand. As a response all atomic charges are influenced and induced charges are superimposed on the equilibrium charge distribution. The polarization energy is defined as the electrostatic binding energy between the induced charges and the gaussian probe minus the energy required to generate the induced charges.This property is plotted in Figure 2 in terms of the width of the Gaussian distribution function. It is found that the ranking of the polarization capacities of the binding sites is unaltered for widths larger than four Angstrom.

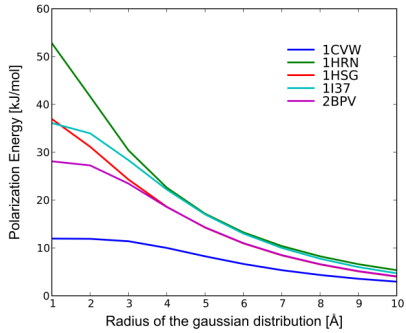


Figure 2 Polarization energies in terms of the with of the probe charge for the five selected proteins.

Before shifting to an in depth interpretation of the results, the electrostatic potential and polarization energy is visualized on the cartoon representation of the protein both in the original EEM model and our model for the HIV-1 protease (1HSG). These properties are obtained similarly as before but the width of the probe charge was set to 4 angstrom to obtain local effects at the scale of individual amino acids. The surfaces are shown in Figure 3. Within the EEM model the electrostatic potential doesn't show substantial variations, the values are all concentrated in a very small region around 240 kJ/mol. This is typical for the EEM model, since it strives for equalization of the electrostatic potential. In the SQE model the electrostatic potential varies substantially in terms of the various regions of the protein (with values ranging from 160 to 430 kJ/mol). The potential is more positive around the ligand docking site, whereas it is less positive at the outside of the protein. The values are always positive since in this case the protein has a global positive charge.

The polarization energy within the EEM model is seriously overestimated giving values up to 110 kJ/mol whereas in the SQE model the maximum lies around 60 kJ/mol. Moreover the protein is about as polarizable in all regions of space. Within the SQE model the polarization energy reaches maximum values in the vicinity of the protein where the ligand binds. These results clearly show the incapability of the EEM model to predict a qualitative correct picture on polarization. The whole system behaves as a conductor where charge can flow freely along all parts of the protein.

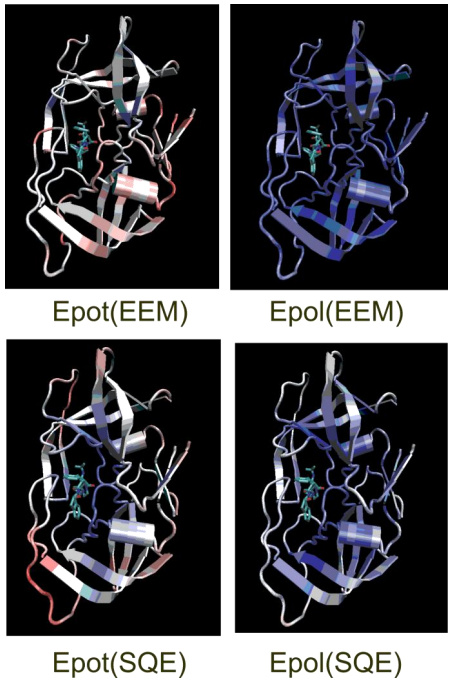


Figure 2 Electrostatic potential and polarization energy visualized on the cartoon of the HIV-1 protease in the original EEM model and our model (SQE).

At this point, we are able to discuss the importance of polarization on the binding of the ligand for the various proteins. As mentioned earlier, ligand recognition relies on various factors. The proposed

method enables to elucidate on the importance of polarization. Our method gives a means for evaluating the importance of polarization very quickly.

The most polarizable binding site is found for the androgen receptor 1I37. In general nuclear hormone receptors bind to fairly unpolarizable hydrophobic ligands. The origin of the high polarization energy must be traced back to the nature amino acids in the binding site. There are an abnormally high number of methionines, several phenylalanines and an extremely large number of leucines which are characterized by polarizabilities of respectively 14.51; 17.95 and 13.38. The magnitude of these values must be evaluated taking into account the overall range of amino acids polarizabilities from 6.37 for glycine to 22.04 for Tryptophan. These values were calculated at the MP2/CC-pVDZ level of theory. Thus not only amino acids with aromatic groups or delocalized electrons contribute to the polarizability of the binding site.

The least polarizable site is a serine protease, 1CVW, because it is composed of fairly non or medium polarizable amino acids such as aspartic acids (10.82), serine(8.78) glutamide (13.36), glycine (6.37) and valine (11.67). In this case the ligand is covalently bound to glutamide and is charged. Moreover the ligand is positioned at the outside (border) of the protein. The results on the polarization point out that the ligand docking is not particularly governed by polarization effects.

All aspartyl proteases (2BPV, 1HSG, 1HRN) are characterized by slightly higher polarizibilities because they also have a fairly large amount of isoleucines and leucines not present in the serine proteases. In particular 2BPV has a large amount of isoleucines (13.41) and other low polarizable amino acids such as proline (10.96), aspartic acid (10.82) and a large amount of glycines (6.37). 2BPV contains a non-polarizing ligand and is itself only medium polarizable. We expect a small contribution of polarizability to the ligand binding.

At first sight 1HSG and 1HRN are very similar, belonging to the same class with similar-sized binding sites. Although their polarizability is fairly different. From our results, it is expected that polarization is more important in the case of 1HRN than in 1HSG.

Summarizing our findings, we have proposed an extended electronegativity equalization method based on a versatile and broad training set, which accounts for correct charge flows in large systems. Our results show primarily that the proposed method shows the correct scaling in terms of the molecular size and secondly that it shows the correct qualitative behavior of electrostatic potential and polarization in biomolecules. The last statement is based on the fact that both previous mentioned properties show distinct variations among the ligand binding site, which is not the case in the standard EEM model. The proposed method enables to evaluate very quickly the impact of polarization on ligand-protein recognition.

Computational Section

All psf-files were generated using the psfgen subroutine in VMD.[25] Optimizations were performed with the aid of the CP2K software[26] using the CHARMM22 all-hydrogen force field for proteins (without CMAP corrections), version c35b1.[27-28]

Received: ((will be filled in by the editorial staff))
Published online on ((will be filled in by the editorial staff))

Keywords: ligand binding · polarization · force fields · electrostatic equilibration

[1] M.L. Quillin, W.A. Breyer, I.J. Griswold, B.W. Matthews, *J. Mol. Biol.* **2000**, *302*, 955-977.
[2] R. Chelli, P. Procacci, *J. Phys. Chem. B* **2004**, *108*, 16995-16997.
[3] G.L. Warren, J.E. Davis, S. Patel, *J. Chem. Phys.* **2008**, *128*, 144110.
[4] D.M. York, W. Yang, *J. Chem. Phys.* **1996**, *104*, 159-172.
[5] P. Mori-Sánchez, Q. Wu, W. Yang, *J. Chem. Phys.* **2003**, *119*, 11001-11004.
[6] A. Nayeem, S. Krystek Jr., T. Stouch, *Biopolymers* **2002**, *70*, 201-211
[7] E.R. Lindahl, in *Methods in Molecular Biology*, Vol. 443 (ed: A. Kukol), Humana Press, Totowa, **2008**, pp. 3-24.
[8] N. Gresh, G.A. Cisneros, T.A. Darden, J.P. Piquemal, *J. Chem. Theory Comput.* **2007**, *3*, 1960-1986.
[9] N. Gresh, *Current Pharmaceutical Design* **2006**, *12*, 2121-2158.
[10] J. Chen, D. Hundertmark, T.J. Martinez, *J. Chem. Phys.* **2008**, *129*, 214113.
[11] A.K. Rappé, W.A. Goddard III, *J. Phys. Chem.* **1991**, *95*, 3358-3363.
[12] Y.P. Liu, K. Kim, B.J. Berne, R.A. Friesner, S.W. Rick, *J. Chem. Phys.* **1998**, *108*, 4739-4755.
[13] J.L. Banks, G.A. Kaminski, R. Zhou, D.T. Mainz, B.J. Berne, R.A. Friesner, *J. Chem. Phys.* **1999**, *110*, 741-754.
[14] R. Chelli, P. Procacci, *J. Chem. Phys.* **2002**, *117*, 9175-9189.
[15] W.J. Mortier, K. van Genechten, J. Gasteiger, *J. Am. Chem. Soc.* **1985**, *107*, 829-835.
[15] S.W. Rick, S.J. Stuart, B.J. Berne, *J. Chem. Phys.* **1994**, *101*, 6141-6156.
[17] S.W. Rick, S.J. Stuart, *Rev. Comp. Chem.* **2002**, *18*, 89-146.
[18] R. Chelli, P. Procacci, *J. Phys. Chem. B* **2004**, *108*, 16995-16997.
[19] R.A. Nistor, J.G. Polihronov, M.H. Müser, N.J. Mosey, *J. Chem. Phys.* **2006**, *125*, 094108.
[20] D. Mathieu, *J. Chem. Phys.* **2007**, *127*, 224103.
[21] T. Verstraelen, *J. Chem. Theory Comput*. to be submitted.
[22] M. van Faassen, P.L. de Boeij, R. van Leeuwen, J.A. Berger, J.G. snijders, *Phys. Rev. Lett.* **2002**, *88*, 186401.
[23] T.J. Giese, D.M. York, *J. Chem. Phys.* **2004**, *120*, 9903-9906.
[24] B. Kirtman, J.L. Toto, K.A. Robins, M. Hasan, *J. Chem. Phys.* **1995**, *102*, 5350-5356.
[25] W. Humphrey, A. Dalke, K. Schulten, *J. Molec. Graphics* **1996**, *14*, 33-38.
[26] CP2K Molecular Dynamics program, http://cp2k.berlios.de.
[27] MacKerell, A.D., Jr,. Feig, M., Brooks, C.L., III, *J. Comput. Chem.* **2004**, *25*, 1400-1415.
[28] A. D. MacKerell, D. Bashford, M. Bellott, R.L. Dunbrack, *J. Phys. Chem. B* **1998**, *102*, 3586-3616.

Part 3: Modeling the Synthesis of Zeolites

Paper 6: "MFI fingerprint: How Pentasil-Induced IR Bands Shift During Zeolite Nanogrowth"

David Lesthaeghe, Peter Vansteenkiste, Toon Verstraelen, An Ghysels, Christine E.A. Kirschhock, Johan A. Martens, Veronique Van Speybroeck, Michel Waroquier

Journal of Physical Chemistry C, **2008**, 112, 9186 - 9191

9186 *J. Phys. Chem. C* **2008,** *112,* 9186–9191

MFI Fingerprint: How Pentasil-Induced IR Bands Shift during Zeolite Nanogrowth

David Lesthaeghe,[†] Peter Vansteenkiste,[†] Toon Verstraelen,[†] An Ghysels,[†] Christine E. A. Kirschhock,[‡] Johan A. Martens,[‡] Veronique Van Speybroeck,*[,†] and Michel Waroquier[†]

Center for Molecular Modeling, Ghent University, Proeftuinstraat 86, 9000 Ghent, Belgium, and Centre for Surface Chemistry and Catalysis, Katholieke Universiteit Leuven, Kasteelpark Arenberg 23, 3001 Heverlee, Belgium

Received: December 7, 2007; Revised Manuscript Received: March 20, 2008

Silicalite-1 zeolite exhibits a characteristic pentasil framework vibration around 540−550 cm^{-1}. In the initial stages of zeolite synthesis, however, this band is observed at much higher wavenumbers: literature shows this vibration to depend on particle size and to shift over 100 cm^{-1} with increasing condensation. In this work, the pentasil vibration frequency was derived from theoretical molecular dynamics simulations to obtain the correct IR band assignments for important nanoparticles. The IR spectroscopic fingerprint of oligomeric five-ring containing precursors proposed in the literature was computed and compared with experimental data. Our theoretical results show that, while isolated five-membered rings show characteristic vibrational bands around 650 cm^{-1}, the combination of five-membered rings in the full MFI-type structure readily generates the bathochromic shift to the typical pentasil vibration around 550 cm^{-1}. As opposed to what was previously believed, the IR band does not shift gradually as nanoparticle size increases, but it is highly dependent on the specific way structural units are added. The most important feature is the appearance of an additional band when double five-membered rings are included, which allows for a clear distinction between the key stages of early zeolite nucleation. Furthermore, the combination of the simulated spectra with the experimental observation of this spectral feature in nanoparticles extracted from silicalite-1 clear solutions supports their structured nature. The theoretical insights on the dependency of pentasil vibrations with the degree of condensation offer valuable support toward future investigations on the genesis of a zeolite crystal.

Introduction

The literature of IR studies on the early stages of zeolite synthesis is exclusively experimental and severely fragmented. We will, therefore, first give a structured overview of the existing literature before addressing the specific issues that arise.

Zeolites are nanoporous aluminosilicate materials, often used in industrial applications for their unique properties in both catalysis and molecular separation.[1] The development of various new synthesis procedures in the past 2 decades has made it possible to design a wide variety of microporous, mesoporous, and hierarchical materials, with varying pore sizes and channel structures. As the number of new materials increases, the field is expanding to all kinds of new applications, e.g., controlled release, sensors, optics, and electronic components.[2] In parallel with the expansion of the nanoporous materials family, fundamental understanding of the molecular and supramolecular mechanisms of the polymerization of specific porous silicate materials is also steadily improving. However, even though significant advances have been made in understanding the role of template molecules, there is still a glaring lack of insight into the details of elementary steps in silica organization. The process of new material discovery and upscaling of syntheses would benefit greatly from submicroscopic insight into the discrete steps, from initial nucleation to full crystal growth.

A landmark discovery in zeolite synthesis was the clear-solution technique of silicalite-1,[3,4] which has the MFI topology and is the full silica version of the industrially important aluminosilicate ZSM-5.[5] The clear-solution mixture is composed of tetrapropylammonium hydroxide ((TPA)OH), water, and tetraethyl orthosilicate (TEOS) at a rather low concentration. When this solution is heated for a few hours, it forms sub-micrometer-sized high-quality crystals of silicalite-1, with its channel intersections occupied by the TPA cations. Even though other synthesis procedures for silicalite-1 exist, the clear-solution technique provides a major advantage for experimental researchers, as silicate particles can be readily extracted and freeze-dried. Furthermore, the clear-solution technique allows for the use of all kinds of in situ diagnostics that require optically transparent and dilute media, such as, for example, dynamic light scattering (DLS). Because the clear-solution synthesis technique is also highly reproducible, it makes analysis of the initial synthesis steps much more amenable for fundamental research in zeolite nucleation and crystal growth.

During the clear-solution synthesis of silicalite-1, a stable suspension of nanometer-sized silica particles has been unequivocally observed, prior to the observation of zeolite crystals characterized by Bragg scattering of X-rays.[6,7] However, the precise structure and role of these nanoparticles during the first steps has been highly debated ever since. Various nanosized structure models have been proposed, ranging from amorphous silicate particles to highly defined framework fragments resembling Lego building blocks.[8] The growth process along which these nanoparticles assemble is also widely debated: do structured silicate nanoparticles self-assemble piece-by-piece in a

* To whom correspondence should be addressed. Telephone: +32 9 264 65 58. Fax: +32 9 264 66 97. E-mail: veronique.vanspeybroeck@ugent.be.
† Ghent University.
‡ Katholieke Universiteit Leuven.

10.1021/jp711550s CCC: $40.75
Published on Web 05/30/2008

multicluster aggregation process?[8–10] Or does zeolite growth occur more in analogy to an Ostwald ripening process, where small nanoparticles sacrifice monomers to feed steadily growing larger particles?[11,12] To add further complexity to this issue, it has even been suggested that one mechanism may dominate the earlier stages, while the other takes over later on.[13]

The major problem in identification of these first stages is that, even with the clear-solution technique, experimental characterization of these embryonic nanoparticles remains difficult.[14] Zeolite nucleation and growth occurs at length scales just above the NMR window but also just below the diffraction regime.[11,15] Furthermore, the template (TPA)OH and the co-solvent EtOH give a very high Raman yield with the consequence that the comparably weak Raman signals of the silicate are unobservable. Because of a lack of other suitable tools, Fourier transform infrared spectroscopy (FTIR) has been one of the key methods used to investigate the solid samples.[16,17] The most important fingerprint in these studies has always been the observed band at approximately 540−550 cm^{-1}, which demonstrates the presence of condensed five-membered structures.[9,18,19] This band is considered to be the spectroscopic signature of MFI-type zeolite,[20,21] but it is also a typical feature of similar structures built from five-membered rings, like the MOR and FER frameworks. Most importantly, this absorbance band is completely absent for amorphous silica particles, which makes its presence highly sensitive to the locally structured nature of the material. In combination with a neighboring band around 450 cm^{-1}, the ratio of band intensities is often used as an approximate assessment of the degree of crystallinity of the observed material.

A handful of papers have pointed out a special evolution of this signature band: as MFI-structured nanoparticles grow, the band red-shifts from initially higher values to the frequency associated with fully crystalline MFI structures.[22–24] In-situ IR spectra of (TPA)OH−TEOS−H_2O mixtures initially display IR absorption around 650 cm^{-1}, shifting and broadening to 600 cm^{-1} with time, as shown in Figure 1A.[22] Kirschhock et al.[22] managed to isolate two different populations of particles and to record their IR absorption spectra ex situ. The smaller species was named the "precursor", containing between 33 and 36 Si atoms. The larger species was called the silicalite "nanoslab" and is built from the aggregation of 12 precursor species, containing at least 400 Si atoms. Comparison of the IR spectra of the precursor, the nanoslab, and crystalline silicalite-1 revealed an aggregation-induced red shift. More specifically, for the precursor sample, Kirschhock et al. observed an absorption band around 590 cm^{-1}, while for the sample of nanoslabs this band was red-shifted over 20 cm^{-1}. An additional shift was observed after these slabs condensed into Bragg−crystalline silicalite-1 structure, where the band was located around the expected 550 cm^{-1}, illustrated in Figure 1B. Other researchers have observed similar results but have not gone so far as identifying the nanoparticles. Hsu et al.[24] found that young precursor-type species show a broad band at 570 cm^{-1}, and as the aging time increased, the band red-shifted toward 550 cm^{-1} and became narrower and more intense. In a similar study Serrano and van Grieken observed a broad vibration band around 560 cm^{-1} in extracted subcolloidal particles from clear-solution synthesis.[23] Titanium-substituted specimens have been shown to exhibit similar spectroscopic changes depending on particle size.[25]

Even though it has been only occasionally exploited to date, this IR shift seems to be a powerful tool toward identification of intermediates in the synthesis stage. However, as several

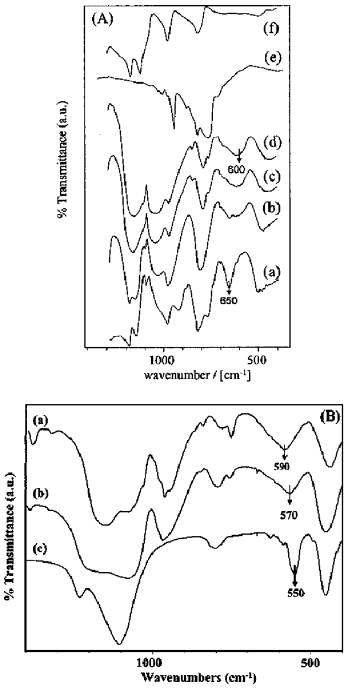


Figure 1. (A) In situ IR spectra of (TPA)OH−TEOS mixture after time intervals of (a) 1, (b) 10, (c) 20, and (d) 46 min, respectively, and (e) TEOS and (f) (TPA)OH. (B) IR spectra of extracted solids: (a) precursor, (b) nanoslab, and (c) micrometer size silicalite-1. Reproduced with permission from ref 22. Copyright 1999 American Chemical Society.

important questions still need to be answered, measuring the shift is not yet as widespread as measuring the intensity of the MFI fingerprint IR band in final materials. Can such a shift really be assigned to nanoparticle assembly? At what stage or stages does a shift occur and why?

In this paper, we use modeling techniques to calculate the IR shifts for pentasil-induced bands from the initial stages of self-organization through to nanosized silicalite-1 crystals. By identifying a detailed link between IR band and crystallinity, this contribution aims to aid experimental scientists in this field. However, as this is a static approach on extracted solids, it does not elucidate the dynamics along which the nanoparticles are formed. The IR shift alone cannot, therefore, distinguish between precursor self-assembly or an Ostwald ripening process in which the zeolite crystal grows at the expense of other nanoparticles. However, since the insights provided in this paper will simplify the identification of intermediate structures and intermediate stages, these results will strongly assist future experimental studies on this issue.

Theoretical Basis. In this paper we make use of an in-house-developed force field for zeolites $GCMFF_{SiOH}$ (version 0.2). This force field was calibrated at the post-Hartree−Fock MP2/6-311+g(d,p) level of theory, with the gradient curves method

(GCM).[26,27] This is a novel technique that facilitates the development of transferable force-field models. It makes extensive use of regularization techniques[28] and of generic energy terms based on series expansions to obtain an optimal bias−variance tradeoff during the fitting procedure. The force-field parameters can be found in the Supporting Information. The IR spectrum was derived from molecular dynamics simulations based on this force field. During an initialization run of 2 ps, the molecular geometries were brought into equilibrium with a thermostat at 300 K, using a velocity-scaling algorithm. The equilibrated geometry was then used as the starting point for 10 consecutive NVT molecular dynamics simulations of 100 ps at a temperature of 300 K, using the Nosé−Hoover-chains method. At the beginning of each of the 10 simulations, the velocities were randomly sampled from the Maxwell distribution. The trajectories were divided into five intervals of 20 ps each, and the average IR spectrum was calculated from these 50 individual intervals.

Scaling of the theoretical results needs to be considered since, as was demonstrated by Scott and Radom,[29] the MP2 level of theory severely underestimates low-frequency vibrations. For full ab initio calculations, MP2 frequency scaling factors have been obtained, ranging from 1.01 to 1.05, depending on the basis set used.[29] As no such scaling factor has yet been obtained for the MP2-based classical force field, we suggest the use of 1.04, which is the ratio between the experimental band at 650 cm^{-1} and our simulated result for small oligomers. It is reassuring to note that, even though just the initial stage was fitted to experimental data, the exact same scaling factor would independently be derived by fitting the final stages. For completeness, both the scaled and unscaled values are reported in Table 1. Figure 3 depicts the unscaled IR spectra, while in Figure 4 and in the remainder of this text we will always refer to the scaled values.

IR Spectrum. The IR adsorption cross-section is given by[30]

$$\alpha(\omega) = \frac{4\pi^2\omega[1-\exp(\hbar\beta\omega)]}{3\hbar cn} I(\omega)$$

where $\beta = k_B T$ and

$$I(\omega) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \mathrm{d}t\, e^{-i\omega t}\langle \mathrm{d}\vec{\mu}(0)\, \vec{\mu}(t)\rangle$$

In these expressions, k_B is the Boltzmann constant, c is the speed of light, n is the refractive index of the medium, $\hbar$ is the reduced Planck constant, and $\langle \vec{\mu}(0)\, \vec{\mu}(t)\rangle$ is the time autocorrelation function of the dipole moment. Using standard manipulations, one obtains in the classical limit

$$\alpha(\omega) \sim \lim_{\tau\to\infty}\frac{1}{\tau}\left|\int_0^{\tau} \mathrm{d}t\, e^{-i\omega t}\,\frac{\mathrm{d}\vec{\mu}(t)}{\mathrm{d}t}\right|^2$$

where $\mu(t)$ is the time-dependent dipole moment from the molecular dynamics simulation. When the velocity and the charge of the atom i are given by $v_i(t)$ and q_i, respectively, the time derivative of the dipole moment can be approximated as

$$\frac{\mathrm{d}\vec{\mu}(t)}{\mathrm{d}t} = \sum_{i=1}^{N} q_i \vec{v}_i(t)$$

where N is the number of atoms.

Results and Discussion

In this section, we have grouped the results and discussion into several subsections corresponding to key synthesis stages.

TABLE 1: MFI Fingerprint Peak Position of Simulated IR Spectra for Complete Range of Structures[a]

		theoretical band	scaled × 1.04
small oligomers	5R	624	649
	8T (2 × 5)	625	650
	11T (3 × 5)	628	653
	7T (2 × 5)	603	627
	8T (4 × 5)	595	619
	12T (4 × 5)	578	601
	10T (2 × 5)	625	650
		578	601
precursors	22T	595	619
		553	575
	36T	548	570
growth along *a*-axis	2a	549	571
	2a′	545	567
	3a	549	571
growth along *b*-axis	2b	541	563
	2b′	543	565
	3b	540	562
	4b	538	560
	4b′	540	562
	5b	539	561
growth along *c*-axis	2c	538	560
	3c	536	557
	4c	536	557
half-slab	2b3c	531	552
	2b′3c	532	553
slab	4b3c	534	555
	4b′3c	534	555
block	2a2b2c	533	554
	2a2b′2c	533	554

[a] For combining two precursors along the *a*- or the *b*-direction, two different connections are possible: a and a′ or b and b′, respectively. The frequency scaling factor of 1.04 has been applied to correct for the underestimation of the MP2-based force field.

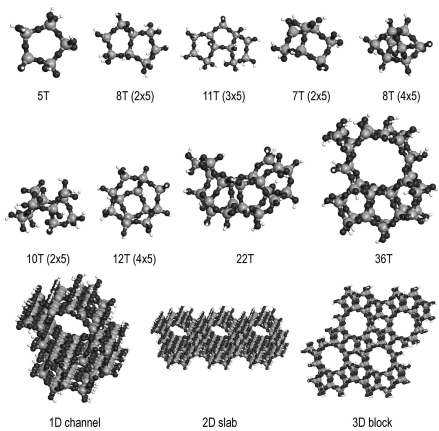


Figure 2. Nomenclature for the structures used in the simulations, ranging from small oligomers to an almost full MFI crystal. The figures were constructed with the in-house-developed Zeobuilder program.[27]

Within each subsection we will discuss the theoretical results, followed immediately by an interpretation of the implications for that particular stage. Figure 2 gives an overview of the structures used (some of which are key components of the MFI structure and some of which might just be formed during the

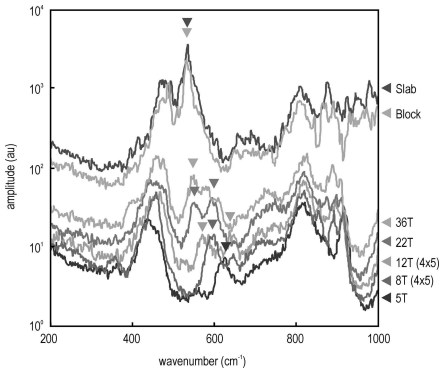


Figure 3. Several key simulated IR spectra, ranging from the isolated five-membered ring (5T), through the quadruple five-membered ring (8T), the pentacyclic dodecamer (12T), the 22T MFI-structured unit (22T), and the 36T MFI precursor (36T), to the 3 × 4 silicalite nanoslab (slab), and 2 × 2 × 2 nanoblock (block).

early stages), accompanied by the nomenclature used in this paper. In Figure 3 several key IR spectra are shown, while the fingerprint IR bands for all structures are summarized in Table 1. Figure 4 presents a summarizing timeline (chronological order from right to left) in which each structure is linked to the position of the corresponding IR band.

Pentasil Ring and Small Oligomeric Units. The elementary building block in this study is the five-membered ring (5T), also called the pentasil ring, as shown in Figure 2. Combinations of this ring are deemed responsible for the typical IR-active vibrations observed at 540−550 cm^{-1}.[31] Our simulations of an isolated five-membered ring show a much higher theoretical band at 650 cm^{-1} (illustrated in Figure 3 and Table 1), matched to experimental observations.[22] The value obtained is too high to correspond with the MFI fingerprint around 540−550 cm^{-1}, which means that, during crystal growth, this value undergoes a significant red shift of almost 100 cm^{-1}.

By connecting pentasil rings, slightly larger building blocks can be constructed. One possible first step is the construction of two and three sideways annealed five-membered rings (called 8T (2 × 5) and 11T (3 × 5), respectively) by successively adding three T-atoms (as shown in Figure 2 and Figure 4). The resulting vibrational frequencies do not differ substantially from the single five-membered ring though and show a similar IR peak around 650 cm^{-1} (Table 1). The mere connection of five-membered rings, sharing just two silicon sites, has no major influence on the frequency of the IR band. This situation changes if, instead, a double five-membered ring is constructed in such a way that two adjacent sides are communal; i.e., three T-atoms are shared by the five-membered rings, forming an additional six-membered ring, referred to as the 7T (2 × 5) species in Figure 2. A sudden shift of approximately 20 cm^{-1} is observed (as shown in Table 1). An even more condensed structure is the silica octamer, referred to as 8T (4 × 5), in which four five-membered rings form a tiny cagelike structure. Here, opposing T-atoms in annealed double five-membered rings are connected by an extra bond, resulting in a highly symmetric species containing four identical Q^3 and four identical Q^2 Si atoms This extremely condensed structure shows an even more significant shift in the infrared peak to a value of approximately 620 cm^{-1}.

Increasing the number of five-membered rings can thus have two different effects on the frequency of the typical band assigned to pentasil vibrations: if the structures remain only loosely connected, sharing at most two silicon atoms between separate five-membered rings, the typical IR band around 650 cm^{-1} remains stationary and there is no observable shift. It would, therefore, be impossible to distinguish between these small oligomeric units from IR bands alone. However, when a similar number of five-membered rings are interconnected to form building blocks with a higher degree of condensation, a significant red shift over 20−50 cm^{-1} occurs. This effect is an important step toward interpretation of the experimentally observed red shift during early oligomeric stages of MFI synthesis,[22] but it is only part of the full story. The simulated IR bands are still too far away from the characteristic MFI bands observed for larger precursor MFI structures and fully crystalline silicalite-1.

This picture undergoes a major change if the double five-membered ring units suggested by Jacobs et al.[20] are simulated. This species is referred to as 10T (2 × 5) in Figure 2. As the five-membered rings themselves do not share any oxygen bridges, the typical band around 650 cm^{-1} remains present. However, a second band appears at 600 cm^{-1}. Even though the five-membered rings do not share any adjacent sides, the appearance of a new peak provides an instant jump of approximately 50 cm^{-1}. This trend is also obvious for the silica dodecamer structure 12T (4 × 5), which was proposed as one early building unit in silicalite-1 precursor formation.[32] This caplike species shows just a single strongly red-shifted band around 600 cm^{-1}.

MFI Precursors. The role of multiple IR bands becomes even more pronounced if several such units combine to form the small cages that are typically encountered in MFI. In this study we have focused on an MFI-structured 22T precursor, which forms a crucial part of the firm 10-ring channel walls. The simulations demonstrate how the multiple IR bands show up in our region of interest upon increasing degree of 5R condensation. As clearly shown in Figure 3 and Figure 4, the primary peak is observed at 620 cm^{-1}, after which it loses intensity and is joined by a new rising secondary band around 575 cm^{-1}. In a next step, the 22T precursor is extended to form the typical 10-membered ring (36T precursor). These essential features will finally lead to the straight and sinusoidal channels that are typical for the materials that have the MFI topology. Furthermore, the 36T structure is the actual precursor on which the aggregation model for zeolite synthesis, proposed by Kirschhock et al.,[8] was based. The original peak around 620 cm^{-1} loses even more intensity and becomes difficult to distinguish. The new peak increases in intensity and undergoes an additional small red shift toward 570 cm^{-1}. However, as is shown in Figure 3, the IR spectrum of the 36T precursor could just as well be perceived as a single broad band around 600 cm^{-1}.

From our results, this seems to be the crucial stage in both the observed IR bands and the synthesis of silicalite-1. In the previous stages the IR band red-shifted gradually, depending on the way the five-membered rings were connected. However, these types of structures can still be expected to be found in amorphous silica particles as well as in many other zeolite structures. A major jump of 50 cm^{-1} occurs only once substantial characteristics of the true MFI framework topology are defined in the double five-membered rings that are present in nanosize precursors. This jump does not correspond to the typical gradual shift that seems to accompany nanogrowth but

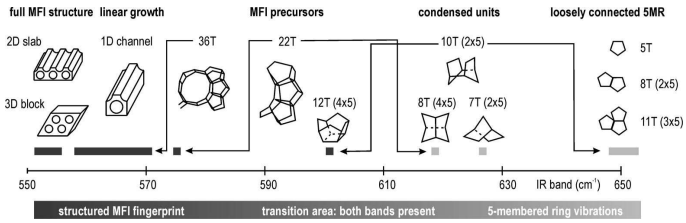


Figure 4. Different regions in silicalite-1 nanogrowth with their corresponding shift in IR band.

arises from the increased intensity of a new band smothering the higher-lying band. The presence of these two bands thus fully explains why Hsu et al.[24] observed a broad band around 570 cm^{-1} for young precursor-type particles. Kragten et al. also observed a broad weak-to-medium band at 565 cm^{-1} for nanoparticles, which they believed to be similar to an extremely weak and very broad band for amorphous silica in the same region.[18] Ravishankar et al., however, succeeded in the observation of a split band, reporting a doublet of 555 and 570 cm^{-1} for crystalline nanosized silicalite-1 as compared to a single band at 550 cm^{-1} for micrometer-sized silicalite-1.[33]

The theoretically simulated IR results thus not only correspond with earlier experimental work but also provide new insights into the fundamental nature of nanoparticle growth. As shown in Figure 4, the synthesis of small MFI precursors marks the boundary area between which two different IR bands are observed. Smaller particles will exhibit the vibrations typical for isolated five-membered rings, while larger more condensed structures in nanoparticles will show the MFI fingerprint even in the absence of Bragg crystallinity.

One-Dimensional Growth along Crystal Axes. Our simulations show that, once several of these precursors are attached to each other, the higher-lying band disappears completely, resulting in a single, much narrower band, in accordance with the observation of Hsu et al.[24] Furthermore, the 570 cm^{-1} band exhibits a further bathochromic shift. Quite remarkably, this shift is slightly dependent on the direction along which growth occurs (Table 1). Along the *a*-axis, the direction of the sinusoidal channels, only four connections are necessary for linking this type of precursor and hardly any shift occurs. For the growth along the *b*- and *c*-axes, six bonds are formed, which results in significant shifts of −10 cm^{-1} to 560 cm^{-1}. The shift appears to be slightly more pronounced for growth along the *c*-axis, which results in a very condensed network between two precursors. The number of connections and the resulting degree of condensation along the *b*-axis (i.e., direction of straight channels), on the other hand, is lower because of the presence of the channel crossings with the sinusoidal channels in the *a*-direction. When looking at linear growth by adding the precursors one-by-one, it is clear that the main shift already occurs when just two precursors are attached along the *b*- or *c*-axis. The effect of nanogrowth on the IR band diminishes rapidly if more building blocks are successively attached in the same direction.

Two-Dimensional and Three-Dimensional Growth to Full Crystal. When growth is considered in two dimensions, to create the nanoslabs as proposed by Kirschhock et al.,[8] the IR band will shift to even lower values. Already for half a nanoslab (containing two 36T precursor units in the *b*-direction and three 36T precursor units along the *c*-axis), the IR band is located near 555 cm^{-1}. Enlargement to a full nanoslab does not fundamentally alter the vibrational frequencies, and the value hovers around 555 cm^{-1}. Theoretically, a combination of 36T building blocks into three-dimensional blocks gives the maximal constraints on the lattice, yet, compared to the two-dimensional nanoslabs, no further shift is observed. The IR band shift of the MFI nanoblock remains at 555 cm^{-1}, close to the value observed in the fully periodic MFI lattice. From these results it seems that, from slabs and blocks onward, the silicalite-1 crystal is well-defined in terms of pentasil vibrations and the MFI fingerprint is observable at full strength.

On the basis of the available IR spectra of the silicalite-1 crystallization intermediate,[22,33] there is some discrepancy between computed and experimentally determined FTIR signatures in the 540−600 cm^{-1} region. Extracted precursors and nanoslabs show IR absorption mainly around 590 cm^{-1} and 570 cm^{-1},[22] respectively, while theoretically the absorption should already be almost entirely red-shifted to 555 cm^{-1}. Even 18−100 nm sized silicalite-1 nanophase material still shows a splitting of the pentasil vibration into 555 and 570 cm^{-1} wavenumbers.[33] This discrepancy can have several causes, both from the experimental and from the computational points of view. The samples might still contain a fraction of smaller entities leading to overlapped signals, or the precursor and larger units may not yet assume the ideal connectivity and degree of condensation assumed in the present computation. Computed values, on the other hand, should mainly be treated as a guide for a qualitative trend. Further investigation is needed to clarify these issues and better exploit the IR tool.

Conclusions

Because experimental characterization of silica nanoparticles is particularly difficult, shifts in IR bands of freeze-dried samples from clear solution have been used to define certain stages in the synthesis procedure of silicalite-1. In this paper, we have used molecular modeling techniques to verify whether such a shift should indeed be expected and to identify the position of the IR peaks for a complete range of structures. We have provided a strong foundation from a modeling viewpoint for the experimental observation of a red-shifting band, which starts out broad but narrows as the structure becomes more defined and condensed. This is not a continuous process as nanoparticle size increases, but it is highly dependent on the specific way monomers have been added. Since we have included simulation of the smallest possible five-membered rings, this shift spans a range of 100 cm^{-1}. However, the narrow MFI fingerprint peak around 550 cm^{-1} becomes strongly present only once the true MFI structure is defined in nanoslabs and nanoblocks. The strongest red shift occurs in the broad transition area, where small precursor-type particles exhibit both the typical five-ring vibrations and the MFI fingerprint band, caused by the vibrations

of a double five-membered ring as proposed by Jacobs et al.[20] Our simulations clearly reveal that pentasil rings at the particle boundary vibrate at higher wavenumbers than more embedded ones. Therefore, it can be expected that the IR spectrum will be sensitive to both particle size and morphology.

From these results it is clear that FTIR is a powerful tool in the characterization of silica nanoparticles. In combination with other techniques to determine size and shape, FTIR should be able to give a more defined picture of the initial stages in zeolite synthesis.

Acknowledgment. This work was sponsored by the Belgian government through an Interuniversity Attraction Pole program (IAP-PAI). The authors acknowledge support from the Fund for Scientific Research Flanders (FWO-Vlaanderen). The authors also acknowledge the Institute for the Promotion of Innovation through Science and Technology in Flanders (IWT-Vlaanderen) for funding this strategic basic research (SBO).

Supporting Information Available: Force field expressions and parameters (pdf). This material is available free of charge via the Internet at http://pubs.acs.org.

References and Notes

(1) Guisnet, M.; Gilson, J.-P. *Catalytic Science Series*, Vol. 3; Imperial College Press: London, 2002.
(2) Davis, M. E. *Nature* **2002**, *417*, 813–821.
(3) Schoeman, B. J.; Sterte, J.; Otterstedt, J. E. *Zeolites* **1994**, *14*, 110–116.
(4) Persson, A. E.; Schoeman, B. J.; Sterte, J.; Ottesstedt, J. E. *Zeolites* **1994**, *14*, 557–567.
(5) Flanigen, E. M.; Bennett, J. M.; Grose, R. W.; Cohen, J. P.; Patton, R. L.; Kirchner, R. M.; Smith, J. V. *Nature* **1978**, *271*, 512–516.
(6) Schoeman, B. J. *Microporous Mesoporous Mater.* **1998**, *22*, 9–22.
(7) Ravishankar, R.; Kirschhock, C. E. A.; Knops-Gerrits, P. P.; Feijen, E. J. P.; Grobet, P. J.; Vanoppen, P.; De Schryver, F. C.; Miehe, G.; Fuess, H.; Schoeman, B. J.; Jacobs, P. A.; Martens, J. A. *J. Phys. Chem. B* **1999**, *103*, 4960–4964.
(8) Kirschhock, C. E. A.; Ravishankar, R.; Van Looveren, L.; Jacobs, P. A.; Martens, J. A. *J. Phys. Chem. B* **1999**, *103*, 4972–4978.
(9) Watson, J. N.; Iton, L. E.; Keir, R. I.; Thomas, J. C.; Dowling, T. L.; White, J. W. *J. Phys. Chem. B* **1997**, *101*, 10094–10104.
(10) de Moor, P.; Beelen, T. P. M.; van Santen, R. A. *J. Phys. Chem. B* **1999**, *103*, 1639–1650.
(11) Auerbach, S. M.; Ford, M. H.; Monson, P. A. *Curr. Opin. Colloid Interface Sci.* **2005**, *10*, 220–225.
(12) Provis, J. L.; Vlachos, D. G. *J. Phys. Chem. B* **2006**, *110*, 3098–3108.
(13) Van Erp, T. S.; Caremans, T. P.; Kirschhock, C. E. A.; Martens, J. A. *Phys. Chem. Chem. Phys.* **2007**, *9*, 1044–1051.
(14) Schüth, F. *Curr. Opin. Solid State Mater. Sci.* **2001**, *5*, 389–395.
(15) Wu, M. G.; Deem, M. W. *J. Chem. Phys.* **2002**, *116*, 2125–2137.
(16) Jacobs, P. A.; Derouane, E. G.; Weitkamp, J. *J. Chem. Soc., Chem. Commun.* **1981**, 591–593.
(17) Coudurier, G.; Naccache, C.; Vedrine, J. C. *J. Chem. Soc., Chem. Commun.* **1982**, 1413–1415.
(18) Kragten, D. D.; Fedeyko, J. M.; Sawant, K. R.; Rimer, J. D.; Vlachos, D. G.; Lobo, R. F.; Tsapatsis, M. *J. Phys. Chem. B* **2003**, *107*, 10006–10016.
(19) Mohamed, R. M.; Aly, H. M.; El-Shahat, M. F.; Ibrahim, I. A. *Microporous Mesoporous Mater.* **2005**, *79*, 7–12.
(20) Jacobs, P. A.; Beyer, H. K.; Valyon, J. *Zeolites* **1981**, *1*, 161–168.
(21) Scarano, D.; Zecchina, A.; Bordiga, S.; Geobaldo, F.; Spoto, G.; Petrini, G.; Leofanti, G.; Padovan, M.; Tozzola, G. *J. Chem. Soc., Faraday Trans.* **1993**, *89*, 4123–4130.
(22) Kirschhock, C. E. A.; Ravishankar, R.; Verspeurt, F.; Grobet, P. J.; Jacobs, P. A.; Martens, J. A. *J. Phys. Chem. B* **1999**, *103*, 4965–4971.
(23) Serrano, D. P.; van Grieken, R. *J. Mater. Chem.* **2001**, *11*, 2391.
(24) Hsu, C. Y.; Chiang, A. S. T.; Selvin, R.; Thompson, R. W. *J. Phys. Chem. B* **2005**, *109*, 18804–18814.
(25) Ravishankar, R.; Kirschhock, C.; Schoeman, B. J.; De Vos, D.; Grobet, P. J.; Jacobs, P. A.; Martens, J. A. *Proceedings of the 12 th International Zeolite Conference*, Vol. III; Treacy, M. M. J., Marcus, B. K., Bisher M. E., Jiggins, J. B., Eds.; Materials Research Society: Warrendale, PA, 1999; p 1825.
(26) Verstraelen, T.; Van Neck, D.; Ayers, P. W.; Van Speybroeck, V.; Waroquier, M. *J. Chem. Theory Comput.* **2007**, *3*, 1420–1434.
(27) Verstraelen, T.; Van Speybroeck, V.; Waroquier, M. *J. Chem. Inf. Model.*, in press.
(28) Wood, S. N. J. *R. Stat. Soc. B* **2000**, *62*, 413–428.
(29) Scott, A. P.; Radom, L. *J. Phys. Chem.* **1996**, *100*, 16502–16513.
(30) Berens, P. H.; Wilson, K. R. *J. Chem. Phys.* **1981**, *74*, 4872–4882.
(31) Jansen, J. C.; Vandergaag, F. J.; Vanbekkum, H. *Zeolites* **1984**, *4*, 369–372.
(32) Kirschhock, C. E. A.; Kremer, S. P. B.; Grobet, P. J.; Jacobs, P. A.; Martens, J. A. *J. Phys. Chem. B* **2002**, *106*, 4897–4900.
(33) Ravishankar, R.; Kirschhock, C.; Schoeman, B. J.; Vanoppen, P.; Grobet, P. J.; Storck, S.; Maier, W. F.; Martens, J. A.; De Schryver, F. C.; Jacobs, P. A. *J. Phys. Chem. B* **1998**, *102*, 2633–2639.

JP711550S

GCM-SiOH-0.2 Force-field model

Applicability

Pure silica. Both periodic systems and hydroxyl terminated clusters.

Atom types

Si: Any silica atom

Ob: An oxygen atom between a silica and a hydrogen atom

Ot: An oxygen atom between to silica atoms

O: Any oxygen atom (Both Ob and Ot)

H: Any hydrogen atom

Interaction types

In this section and the sections below, x is used as a generic symbol for an internal coordinate of the molecular system. It's actual interpretation depends on the interaction type. The force field uses four interaction types:

Bond stretch: x = the distance between two bonded atoms (a-b)

Bending: x = the cosine of the angle formed by two bonds that share one atom (a-b-c)

Urey-Bradley: x = the distance between the two atoms that are bonded to one common atom (a-b-c)

Nonbond: x = the distance between two atoms that are not bonded to each other and that are not bonded with a common atom (a b)

Analytical expressions

The force field uses three types of analytical expressions for the energy terms: PN, RPN and RPN_CUT. Each energy term depends only on one internal coordinate, x.

[EQ]V_{PN(m)}(x) = \sum_{i=1}^m a_i x^i[/EQ]

$$V_{PN(m)}(x) = \sum_{i=1}^{m} a_i x^i$$

[EQ]V_{RPN(m)}(x) = \sum_{i=1}^m a_i*x^{-i}[/EQ]

$$V_{RPN(m)}(x) = \sum_{i=1}^{m} a_i x^{-i}$$

[EQ]
V_{RPN_CUT(m,n,x_c)}(x) = \left\{\begin{matrix}
\sum_{i=1}^m a_i \left(x^{-i}+ \sum_{j=0}^n b_{i,j}(x_c) (x-x_c)^j\right) & \mbox{when} & 0 < x < x_c \\
0 & \mbox{when} & x \geq x_c
\end{matrix}\right.
[/EQ]

$$V_{RPN\_CUT(m,n,x_c)}(x) = \begin{cases} \sum_{i=1}^{m} a_i \left(x^{-i} + \sum_{j=0}^{n} b_{i,j}(x_c)(x - x_c)^j \right) & \text{when} \quad 0 < x < x_c \\ 0 & \text{when} \quad x \geq x_c \end{cases}$$

The constants b_(i,j)(x_c) were determined before the actual fitting procedure. They guarantee that each term in the expression V_{RPN_CUT(m,n,x_c)} is continuous up to the n'th order derivative for x > 0:

[EQ]b_{i,j}(x) = -\frac{1}{j!}\frac{\partial x^{-i}}{\partial x^j}[/EQ]

$$b_{i,j}(x) = -\frac{1}{j!}\frac{\partial x^{-i}}{\partial x^j}$$

Parameters

All parameters are given in atomic units.

x_c = 37.794522498

Atom types	Interaction type	Functional form	a1	a2	a3	a4	a5
Si-O	Bond stretch	RPN(4)	3.0414E+1	-1.6019E+2	3.3354E+2	-2.3782E+2	
O-H	Bond stretch	RPN(4)	-2.4334E+0	-7.4849E-1	5.9958E+0	-3.3248E+0	
Si-O-Si	Bending	PN(4)	5.3619E-3	1.0321E-2	-3.9335E-4	-2.3325E-3	
Si-O-Si	Urey-Bradley	PN(2)	-2.6132E-1	2.0308E-2			

O-Si-O	Bending	PN(4)	3.8454E-2	3.2970E-2	2.4373E-2	8.8445E-3	
O-Si-O	Urey-Bradley	PN(4)	-7.5795E-1	1.6002E-1	-1.5720E-2	6.0512E-4	
Si-O-H	Bending	PN(4)	6.6420E-3	1.7401E-2	-3.8780E-3	-7.4695E-4	
Si-O-H	Urey-Bradley	PN(2)	-1.4153E-1	1.5129E-2			
Si Si	Nonbond	RPN_CUT(5.1,x_c)	2.8580E-1	-3.5895E+0	3.8572E+1	-1.4207E+2	3.5772E+2
Si O	Nonbond	RPN_CUT(5.1,x_c)	1.6341E+0	-2.6363E+1	1.9974E+2	-7.6311E+2	1.0983E+3
Si H	Nonbond	RPN_CUT(5.1,x_c)	1.3181E-2	4.1222E+0	-4.2238E+1	1.7505E+2	-2.2748E+2
O O	Nonbond	RPN_CUT(5.1,x_c)	-9.9697E-1	1.6578E+1	-1.2463E+2	4.3203E+2	-4.8719E+2
Ob H	Nonbond	RPN_CUT(5.1,x_c)	7.9068E-2	-2.6556E+0	1.9408E+1	-6.0318E+1	6.3437E+1
Ot H	Nonbond	RPN_CUT(5.1,x_c)	7.0999E-2	-2.5981E+0	1.9146E+1	-6.2609E+1	7.0652E+1
H H	Nonbond	RPN_CUT(5.1,x_c)	-2.6607E-2	1.1037E-1	1.6637E+0	-5.1527E+0	5.3852E+0

Paper 7: "Multi-level modeling of silica-template interactions during initial stages of zeolite synthesis"

Toon Verstraelen, Bartek M. Szyja, David Lesthaeghe, Reinout Declerck, Veronique Van Speybroeck, Michel Waroquier, Antonius P. J. Jansen, Alexander Aerts, Lana R. A. Follens, Johan A. Martens, Christine E. A. Kirschhock, Rutger A. van Santen

Topics in Catalysis, **2008**, accepted

Multi-level modeling of silica-template interactions during initial stages of zeolite synthesis

Toon Verstraelen,[1] Bartłomiej M. Szyja,[2,3] David Lesthaeghe,[1] Reinout Declerck,[1]

Veronique Van Speybroeck,[1,] Michel Waroquier,[1] Antonius P. J. Jansen,[2]*

Alexander Aerts,[4] Lana R. A. Follens,[4]

Johan A. Martens,[4] Christine E. A. Kirschhock,[4] Rutger A. van Santen[2,]*

[1] Ghent University, Center for Molecular Modeling, Proeftuinstraat 86, 9000 Gent, Belgium

[2] Eindhoven University of Technology, Department of Chemical Engineering and Chemistry, Den Dolech 2, 5612 AZ Eindhoven, The Netherlands

[3] Wrocław University of Technology, Faculty of Chemistry, Department of Fuels Chemistry and Technology, Gdańska 7/9, 50-344 Wrocław, Poland

[4] K. U. Leuven, Centre for Surface Chemistry and Catalysis, Kasteelpark Arenberg 23, 3001 Leuven, Belgium

veronique.vanspeybroeck@ugent.be, r.a.v.santen@tue.nl

Abstract

Zeolite synthesis is driven by structure-directing agents, such as tetrapropyl ammonium ions (TPA^+) for Silicalite-1 and ZSM-5. However, the guiding role of these organic templates in the complex assembly to highly ordered frameworks remains unclear, limiting the prospects for advanced material synthesis. In this work, both static ab initio and dynamic classical modeling techniques are employed to provide insight into the interactions between TPA^+ and Silicalite-1 precursors. We find that as soon as the typical straight 10-ring channel of Silicalite-1 or ZSM-5 is formed from smaller oligomers, the TPA^+ template is partially squeezed out of the resulting cavity. Partial retention of the template in the cavity is, however, indispensable to prevent collapse of the channel and subsequent hydrolysis.

KEYWORDS Zeolites, TPA template, structure-directing agent, ZSM-5, Silicalite-1, precursors, nucleation, molecular dynamics, density functional theory

BRIEFS Silica-water-template interactions during the initial stages of zeolite synthesis

Introduction

This study aims to unravel the elementary interactions and driving forces behind Silicalite-1 formation, whose aluminosilicate counterpart ZSM-5 is a commonly used catalyst in the petrochemical industry.[1] Silicalite-1 provides a textbook case study: it has been the object of countless investigations on zeolite formation and is formed through the best understood zeolite synthesis procedure to date. [2-26] Colloidal Silicalite–1 is synthesized from 'clear solutions', from which it is obtained by hydrolysis of tetraethylorthosilicate (TEOS) as a monomeric silica source in aqueous tetrapropylammonium hydroxide (TPAOH) at room temperature. The 'clear solution' is actually a clear suspension of subcolloidal nanoparticles smaller than 10 nm that forms spontaneously upon mixing the reagents at ambient temperature. The nature of these nanoparticles and their role in the nucleation and growth process is currently subject to considerable discussion. Some believe that the silica nanoparticles do not participate in nucleation directly, but dissolve and serve as nutrients during crystallization. [8,27,28] thers assume the direct incorporation of formed nanoparticles into the growing crystals. This could be accomplished via an aggregation mechanism [5,16,23,29-32] in which these nanoparticles either already resemble the MFI structure [29-31] or exhibit a different silicon connectivity beforehand. [16,21-23] For the particular case of Silicalite-1 formation in presence of tetrapropylammonium cations (TPA^+), evidence in favor of the aggregation type mechanisms is growing. [16,22,23,33]

The exact role of organic cations like TPA^+ as structure-directing agents (SDAs) for zeolite synthesis is also still controversial. It is not clear whether they act as true templates shaping silica around them, [4,34,35] whether they act as external 'scaffolds', organizing the solvent and stabilizing the hetero-network of oligomers, [36-38] or whether they form a shell around the silica-rich, negatively charged core of the nanoparticles. [21,40] The first function implies direct silicate organization by the template inside the nanoparticles on a molecular level, [39] while the second and third hypotheses assume supramolecular organization, during which silica nanoparticles are shielded from excessive hydrolysis.

This study will focus on these early stages of zeolite formation from a modeling perspective. In particular we consider aggregation type mechanisms, given the growing evidence for their role as mentioned above. In an aggregation type mechanism, silica-template interactions convert smaller oligomers into nano-sized precursor species. [41-43] Many such species have been identified in the synthesis mixture. Our analysis is based primarily on the Si_{33} precursor species (constructed from three Si_{11} undecamers). This precursor with MFI-like connectivity has been proposed to participate actively in Silicalite-1 formation as it strongly resembles a fragment of the future crystal. [41-43] The following specific issues will be investigated: (i) What is the nature of the interaction between TPA^+ and the Si_{33} precursor. (ii) Where is the template located relative to the Si_{33} precursor. (3) Which potential role can the template play in the aggregation of multiple Si_{33} precursors ?

Recently, comparison between simulated and experimental IR patterns has illustrated how the silica enclosed in the colloidal nanoparticles evolves in time, leading from small five-ring oligomers towards successively more condensed 5-ring species. [44] We will, therefore, not only focus on the Si_{33} precursor, but also on possibly preceding Si_{22} intermediates (formed from two Si_{11} undecamers), which are the smallest oligomers able to create an initial section of the straight 10-ring channels that are present in the final Silicalite-1 structure. The Si_{11} undecamer units, based on 5-membered rings, are the elementary building blocks for the simulated species. For these S_{11} units, Kirschhock et al. have proposed three possible structures: the capped double five ring, the tetracyclic undecamer and the tricyclic

undecamer, [41,42] for which only the latter can be combined to a Si_{22} nanoparticle. As observed from the splitting of the chromatographic peak corresponding to the Si_{22} species, [43] such a construction can proceed via two routes, for which only one forms the 10-ring channel. Addition of a third Si_{11} unit to either of these double units can lead to the formation of the proposed Si_{33} precursor, as shown in Figure 1. [43]

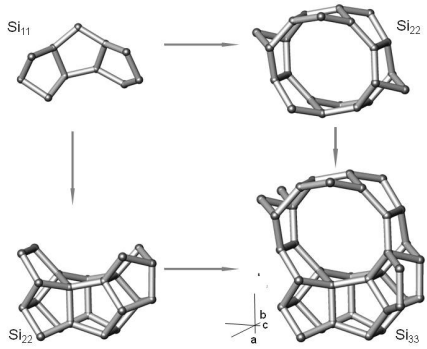


Figure 1: Assembly of Si_{11} units to Si_{22} and Si_{33} precursor nanoparticles.

For such a complex system, many different variables need to be addressed and many different techniques can be used, which is why theoretical modeling of zeolite synthesis has proven to be a challenging task. [45-47] Nevertheless, several attempts have already been made, using both quantum chemical and classical molecular mechanics techniques. [48] Major contributions have been given by Catlow and co-workers using a variety of different modeling techniques including both static and dynamic approaches. [49-53] From classical molecular dynamics simulations of silica precursors and a structure-directing agent, they found long range electrostatic interactions to be of crucial importance, since without these interactions the investigated complexes tended to dissociate rather than agglomerate. Rao and Gelb, [54] on the other hand, studied earlier stages of silica polymerization using the reactive forcefield developed by Feuston and Garofalini. [55] They observed that at time scales shorter than 0.5 ns four-membered rings will be most common, while at time scales longer than 1 ns five-membered rings will dominate. Very recently, Mora-Fonz, Catlow and Lewis have shown the importance of solvent and pH to control specific oligomerization and cyclization processes in the nucleation of microporous silicas. [56-57] Wu and Deem have shown that the nucleation barrier of silica without templates at high pH (~12) lies around 50 T atoms. [58] Jorge, Auerbach and Monson performed large scale simulations on silica-template mixtures in a coarse-grained model and were able to reproduce the experimental observation of the stabilizing role of the template in the growth of metastable silica nanoparticles. [59]

The major drawback for theoretical simulations is the fact that there are many different variables which, each to a specific yet unknown extent, might all influence the interaction between silica and template molecules. In this paper, we have tried to address this shortcoming through a multi-level approach. Among the most crucial parameters that can be

varied are the various ways solvent can be treated, from polarizable continuum methods to the explicit treatment of individual molecules. The silica-template interaction can furthermore be described at numerous levels of theory, ranging from fast yet approximate force fields to highly accurate but extremely time-demanding *ab initio* methods. To reduce the enormous task of performing a multitude of different simulations tailored to each individual parameter, while at the same time maximizing the benefits of various approaches, we have performed just two sets of calculations that complement each other well, one static and one dynamic. Within these two sets of simulations, we have accommodated as many different variables as possible: an overview of the major differences between the two sets is given in Table 1. A more detailed discussion of the various contributions will be given in the following section.

Table 1: Comparison of the treatment of crucial variables for the two complementary sets of calculations.

	Static approach	**Dynamic approach**
Level of theory	Hartree-Fock Density Functional Theory (DFT)	Universal Forcefield (UFF)
Solvent treatment	dielectric continuum model (COSMO)	explicit molecules
Solvent type	no solvent water ethanol	water + additional TPA^+
Charge on silica nanoparticle	neutral Al defect	SiO^-
Total charge on nanoparticle	0 -1	-1 on Si_{33} -3 on Si_{11} -6 on Si_{22}
Nanoparticle size	Si_{33}	Si_{11} Si_{22} Si_{33}

Theoretical basis

Ab initio static calculations

In our first type of calculations, the system was treated using static ab initio calculations with the GAUSSIAN03 and CP2K/QUICKSTEP packages. [60-61] We used the 33T precursor as proposed by Kirschhock et al., [41] which was also modeled in previous work. [44,62] With GAUSSIAN03, the silica cluster and the template were optimized at the ab initio HF/3-21g level of theory, after which HF/6-31+g(d) and density functional theory (DFT) B3LYP/6-31+g(d) single-point energies were calculated. [63-64] Throughout the manuscript, these two level methods will be designated as HF/6-31+g(d)//HF/3-21g and B3LYP/6-31+g(d)//HF/3-21g respectively. All initial optimizations were performed in the gas-phase, while the solvent was taken into account by using the COSMO model as implemented in GAUSSIAN03.[65-66] Both water and ethanol solvents were considered, as these are typical solvents for a clear solution synthesis procedure. We further corroborated these results with a full DFT treatment (i.e. also for the geometry relaxation), by employing the Gaussian and plane-wave (GPW) density functional method with periodic boundary conditions, as implemented in the CP2K/QUICKSTEP program package. [61] A PBE gradient-corrected functional was used throughout, [67]

together with a TZ2VP-PSP basis set, [68] a 320 Ry cutoff for the auxiliary plane wave grid, and pseudopotentials developed by Goedecker *et al.* [69-70] This method is further referenced in the paper by GPW/PBE/TZ2VP-PSP.

A single tetrapropylammonium (TPA^+) template molecule was considered for each precursor species. Since the TPA^+ molecule is positively charged, the entire system's charge neutrality was maintained by incorporating an Al atom substituting a Si atom, thus creating a net negative charge on the Si_{33} precursor. In addition, similar calculations with a neutral precursor were performed as well by treating the entire system (precursor + TPA^+) as net positively charged to evaluate the effect of the electrostatic interactions. In the remainder of this article, the following nomenclature is used for the precursor-template structures : Si_{33}-TPA^+ and $Si_{32}Al^-$-TPA^+ for the net positively charged and the neutral system, $Si_{32}Al^-$-TPA^+ / water and $Si_{32}Al^-$-TPA^+ /ethanol for the neutral system that is additionally embedded in a dielectric medium that characterizes water and ethanol. Initial structures for all complexes were generated with the in house developed software package ZEOBUILDER, which is specifically designed for building molecular architectures starting from elementary building blocks. [71]

Classical molecular dynamics calculations

In our second type of calculations, the system was simulated using classical molecular dynamics (MD) with periodic boundary conditions. The unit cell was cubic in shape with an edge length of 25 Å. The content of the unit cell was varied to capture different stages during synthesis. A first set of simulations was performed on a unit cell containing 3 Si_{11} oligomers, two of which were connected to form the Si_{22} structure (shown in Figure 2). While the third Si_{11} oligomer was also present, it remained a spectator species during the entire simulation. In addition, the environment contained 9 TPA^+ molecules and 250 explicit water molecules. To study the influence of charged configurations on the system, we applied the following classical approach: all nitrogen atoms belonging to TPA^+ were assigned a positive charge (+1). Formation of a Si-O-Si bridge requires some of the oxygen atoms to be ionized, which is illustrated by the high pH required for condensation reactions to occur. [56,72] Therefore, three oxygen atoms in each of the silica oligomers were deprotonated, and charged negatively (-1) to counter-balance the TPA^+ ions. Since the total number of positive and negative charges was kept equal, the whole system could be treated as electrostatically neutral. This approach still requires a dielectric constant correction to properly account for bulk solvent effects. [73] This relative dielectric constant was set to 60, which corresponds to the dielectric constant of water under simulation conditions.

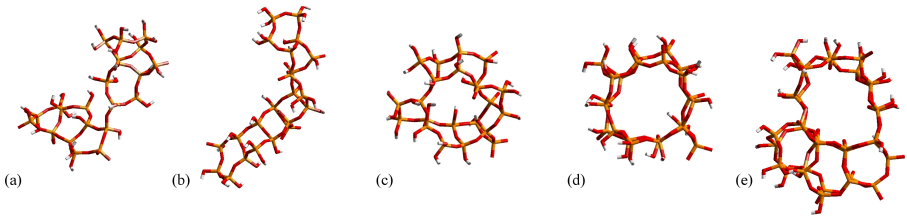


Figure 2: Five silica models used in simulations. (a) Si_{22} with one connection (oxygen bridge) between two Si_{11} oligomers; (b) Si_{22} with 2 connections; (c) Si_{22} with 3 connections; (d) Si_{22} with 4 connections (fully formed channel); (e) Si_{33} precursor

In a subsequent set of calculations, the unit cell contained the fully formed Si_{33} precursor with only one positively charged TPA template and 333 water molecules. As before, the nitrogen of the TPA^+ ion was assigned a positive charge (+1), while the precursor was deprotonated at one oxygen to maintain charge neutrality in the simulated box. As noted above, the dielectric constant was set to the value of 60. These simulations with explicit water molecules serve as an ideal comparative set to the static simulations.

The MD simulations were carried out in the NVT ensemble. The total time was set to 1000 ps with a timestep of 1 fs, while the temperature was set to 350 K, controlled by the temperature damping thermostat described by Berendsen et al. [74]

The selection of an adequate potential is crucial for every classical molecular simulation. There are many reactive and non-reactive forcefields optimized for zeolitic systems available: Feuston-Garofalini, [55,75] ReaxFF, [76-79] BKS, [80] CVFF [81] or the Catlow library-collection of potentials. [82-83] All of them suffer from important disadvantages related to the system described here. Both reactive forcefields (Feuston-Garofalini and ReaxFF) allow creation or breaking of chemical bonds during simulation, but they do not include interactions between the silica oligomers and TPA. Similarly, the BKS forcefield, which is limited only to the atom types Si, Al, P and O, also fails to do so. Among the aforementioned forcefields, only CVFF could provide the required parameters for the MD simulation. However, the quality of the results obtained with CVFF were well below expectations: even for long time-scale simulations we only observed minor thermal vibrations, and the system did not significantly evolve during the simulation.

Therefore, the Universal Forcefield (UFF) was applied in the present work, despite the absence of a hydrogen bonding term. [84] Results obtained using this forcefield are qualitatively consistent with the quantum chemical calculations.

The molecular dynamics runs were analysed using the in house developed software package MD-TRACKS, allowing efficient analysis of molecular dynamics and Monte Carlo runs and generation of physical properties such as radial distribution functions along the run. [85]

Models used in simulations

Due to the fact that a non-reactive forcefield was applied in the classical simulations, it was necessary to manually create the bonds connecting two Si_{11} oligomers before starting new simulations. We created four different models of Si_{22} oligomers, and carried out Molecular Dynamics simulations for each of them. These models differ in number of connections (oxygen bridges) between two Si_{11} oligomers. The most flexible structure is the model with only one connection between the Si_{11} units, as shown in Fig. 2(a). The Si_{22} structure with 4 connections, representing a segment of the straight channel in an MFI type zeolite, 2(d), is the most rigid.

The MD simulations represent two stages in the formation of Si_{22} – models (a) and (b) represent the initial stage, where only one end of the Si_{22} structure is connected, while models (c) , (d) and (e) represent the final stage where a silicalite channel fragment is formed. For the computationally more demanding ab initio calculations only one model was used: the Si_{33} precursor, as shown in figure 1(e), which is obtained when an additional Si_{11} structure is added. This structure fully corresponds to the Silicalite-1 precursor species proposed in literature. [41]

Results and discussion

In this section, the stability of the Si_{33} precursor structure and its interaction with the template is discussed first, after which the positioning of the template with the preceding Si_{22} species will be investigated. Finally, a possible scenario for subsequent aggregation to a full crystal is put forward.

Ab initio static calculations: precursor-template interaction

An initial important observation applies to the neutral Si_{33} cluster without a TPA^+ molecule interacting with it. When optimizing this structure to a global minimum, the channel of this silica precursor collapses completely, as shown in figure 3. This observation is in accordance with earlier results by Lewis et al.: [53] a template prevents the decrease in surface area of the open-structured fragment. Absence of this stabilization will lead to reduction of the surface area, increasing the likelihood of subsequent hydrolysis. Without TPA^+, the precursor structure is not favored nor can it lead to a fully crystalline zeolite.

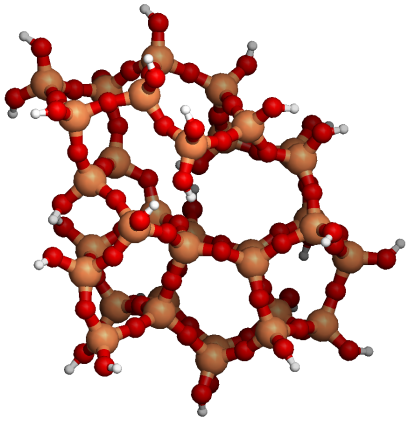


Figure 3: Structure of the optimized neutral Silica cluster Si_{33} without a TPA^+ at the B3LYP/6-31+g(d) level of theory.

In presence of the organic template, however, the neutral precursor does not fully collapse. In agreement with earlier results of Magusin et al., [62] encapsulation of the TPA cation within the 10-ring of the Si_{33} precursor corresponds to a stable minimum on the potential energy surface (Si_{33}-TPA^+ is similar in structure to $Si_{32}Al$-TPA^+, which is labeled as structure B in Figure 4). However, two additional stable structures were located (their aluminosilica counterparts are labeled as structures A and C in Figure 4), for which TPA^+ is adsorbed slightly off-centre to the precursor. These 'half in- half out' structures are considerably more stable than structure B, where TPA is fully located inside the precursor: up to 205 and 164 kJ/mol at the GPW/PBE/TZ2VP-PSP level of theory, depending on the side along which the TPA template molecule moves out of the precursor. The relative energies of these calculated structures at various levels of theory are given in Table 2.

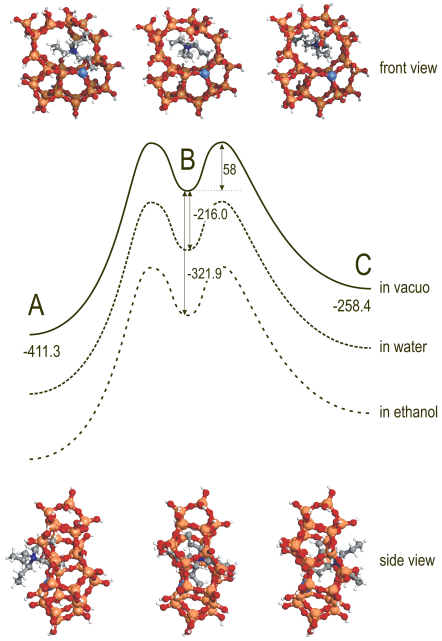


Figure 4: Relative energies in kJ/mol between the three stable structures ($Si_{32}Al^-$ precursor + TPA^+) at the GPW/PBE/TZ2VP-PSP level of theory.

Table 2: Relative energies [kJ/mol] for the precursor and a single TPA^+ molecule.

	Level of theory	structure A[a]	structure B[a]	structure C[a]	Solvation energy
Si_{33}-TPA^+	HF/3-21g//HF-3-21g	-188.4	0.0	-180.1	
	GPW/PBE/TZ2VP-PSP.	-205.2	0.0	-163.6	
$Si_{32}Al^-$-TPA^+	HF/3-21g // HF-3-21g	-407.8	0.0	-279.2	
	HF/3-21g//HF/6-31+g(d)	-464.2	0.0	-346.8	
	HF/3-21g//B3LYP/6-31+g(d)	-413.1	0.0	-293.6	
	GPW/PBE/TZ2VP-PSP.	-411.3	0.0	-258.4	
$Si_{32}Al^-$-TPA^+/water	HF/3-21g//B3LYP/6-31+g(d)	-353.4	0.0	-278.2	-216.0
$Si_{32}Al^-$-TPA^+/ethanol	HF/3-21g//B3LYP/6-31+g(d)	-373.8	0.0	-282.4	-321.9

[a] In structure B the TPA cation is fully encapsulated within the 10-ring of the 33T precursor, in structures A and C the TPA cation is positioned off-centre of the precursor as shown in figure 4. [b] The solvation energy is defined as the energy difference between gas phase energies and the energies obtained by including a continuous dielectric medium.

As long as TPA^+ fully resides within the cavity, it prevents any collapse of the structure. In this configuration Coulombic interactions between template and framework are of considerable importance [53] and the positively charged template will mostly interact with the diffuse electron cloud of the surrounding framework oxygen atoms. [86-87] For the

structures in which TPA is 'half in- half out', this interaction is reduced and the silicate structure is allowed to slightly relax, yet without collapsing completely. The interaction between the propyl 'arms' of the template and the hydrophobic inner surface of the precursor channel remains strong enough to stabilize an open precursor structure.

When the Si_{33} precursor is negatively charged with an Al defect to compensate for the charged template ($Si_{32}Al^{-}$-TPA^{+}), the qualitative picture remains identical, while the energy differences between the various structures in Figure 4 are more pronounced. For the GPW/PBE/TZ2VP-PSP level of theory this amounts to 411 kJ/mol and 258 kJ/mol (Table 2), depending not only on the side along which the TPA template molecule moves out of the precursor, but also on the position of the Al defect. Structure A, for which the positively charged TPA is located closest to the negative Al defect, is substantially more favored over structure C, where the template is on the opposite side of the Al substitution. In both cases there is a significant reduction in energy with the template positioned 'half in - half out' of the precursor. The precursor is partially relaxed along one side (i.e. on the side facing away from the template) to account for this energy difference.

To assess the barrier between these various structures, an energy profile was constructed along a pathway during which TPA moves out of the precursor. The energy profile resembles a possible transition path connecting structure B to structure A. [88] The barrier for escape of the TPA from the precursor was estimated at 58 kJ/mol, which should be considered as an upper limit since other, more preferable, escape paths might exist. This value is relatively small, indicating that such a transition should easily occur.

In a next step, the influence of solvation was investigated by applying the COSMO model in Gaussian03, since this methodology has already been tested for this type of systems by Mora-Fonz *et al*. [56] Solvation leads to an increased stability of the $Si_{32}Al^{-}$-TPA^{+} complex by 216 and 322 kJ/mol for water and ethanol respectively. The solvation energy as shown in Table 2 originates from enclosing the complex in a cavity within a dielectric medium and is often referred to as the dielectric solvation energy (DSE). [73,89-90] The DSE values are larger for ethanol than for water, in accordance with experimental observations by Kirschhock et al., [41,43] stating that the precursor species are long-lived in ethanol, or at least more so than in water. The qualitative picture on the relative stability of the various structures remains unaltered by including the solvent, while the energy difference between structures A and C is slightly reduced.

The structures for which the template is positioned partially out of the 10-ring fragment hint at a supramolecular structure-directing role of TPA. The template is responsible for stabilizing tiny segments of the straight MFI channel, and one could speculate about the tendency of supramolecular precursor-TPA complexes to organize into larger aggregates. At first instance, we only investigated this possibility from a purely geometrical point of view, by using our optimized precursor-TPA complexes to construct nano-crystalline structures as proposed earlier in literature. [33,41-43] This was achieved by using our in-house developed software ZEOBUILDER, which allows construction of complex molecular architectures starting from elementary building blocks, [71] similar to building toy structures with Lego® blocks. The program uses a condensation algorithm that searches for optimal connections between sets of oxygen pairs in order to form new oxygen bridges. The procedure employed here, was based purely on geometrical constraints and the obtained structures were not further optimized. If serious geometrical obstacles would be encountered via this procedure, this would mean that the proposed structures can definitely not organize into larger subunits. However, by combining both A and C structures from Figure 4, it was possible to generate a small nano-crystallite in which the template resides close to the channel

intersections (as shown in Figure 5). The template molecule is expected to move more or less freely into the relatively larger space available, but additional molecular dynamics simulations of TPA in the channel intersections would be necessary to further support this proposal, by investigating the flexibility of the template in the channel intersections. To date, this type of calculation at a solid level of theory is beyond current computer capabilities. Even ab initio geometry optimizations of the illustrated nano-crystal (containing approximately 2200 atoms) are beyond what is feasible today. For such an optimized structure, there would be approximately 9 TPA molecules inside a 12-block nano-crystal. [30]

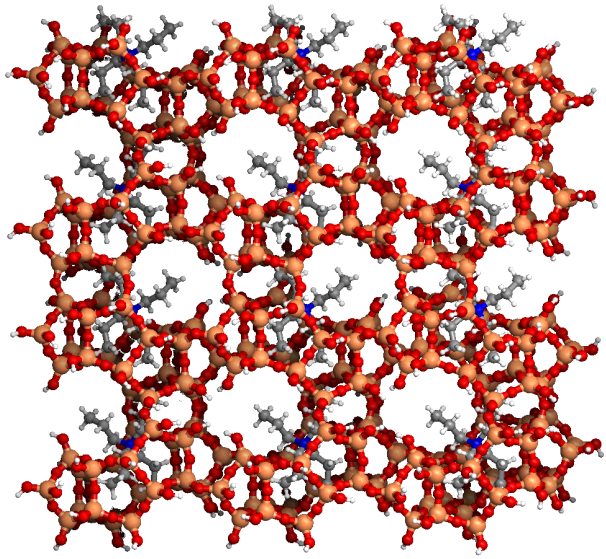


Figure 5: MFI-type nano-crystal containing 12 aggregated Si_{33} precursors, constructed from stable structures A and C from Figure 4.

From the quantum chemical results, it seems that only the hydrophobic interaction between the alkyl groups of the template and the inner pore surface of the precursor are of crucial importance for the stabilization of an MFI precursor on a molecular level. Together with the exceptional lattice stability of the MFI topology, [91] the partial role of TPA in a 'half in – half out' structure indicates why, next to TPA, also many other template molecules are capable of creating MFI-like structures.

Classical molecular dynamics calculations: template positioning at elevated temperatures

In the previous section, the static ab initio calculations showed the template to partially exit the straight channel fragment in the precursor. In this section, molecular dynamics calculations using a classical force field are presented to complement these findings and to provide additional insight into the position of the template during the preceding stages of precursor synthesis. Therefore, various Si_{22} intermediates are taken into consideration, where the number of bonds is gradually increased to mimic the initial assembly of silica into a local

MFI connectivity, culminating in the formation of a rigid fragment of the straight 10-ring channel.

Figure 6 shows representative snapshots during the molecular dynamics runs for the various Si_{22} oligomers. The structures shown in (a) and (b), in which the channel is not yet formed, offer a wide enough gap between the Si_{11} parts to fit a TPA^+ ion. This is confirmed by geometry analysis which shows that the template molecule is located in between separate ends of the silica oligomer during the entire molecular dynamics run. The template's position changes when additional oxygen bridges are formed to obtain structures (c) and (d). Although the TPA^+ ion might be able to fit inside the channel of system (c), its preferred position is outside the channel. The same holds for the fully formed 10-ring (d). Just as in the static calculations, the template moves out of the channel fragment into the direction of the later to be formed channel cross-section in MFI.

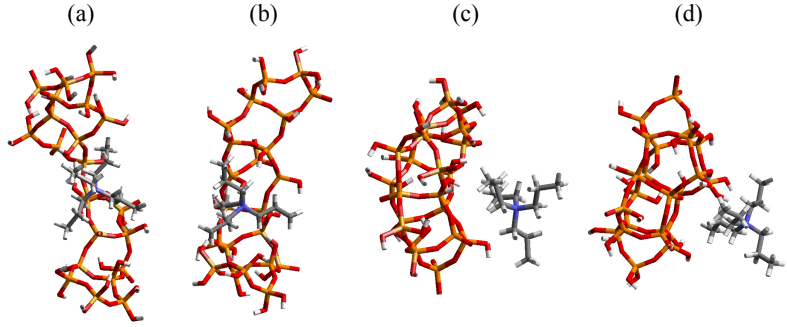


Figure 6: Position of TPA molecules relative to Si_{22} oligomers with 1 bond (a), 2 bonds (b), 3 bonds (c) and 4 bonds (d). These representative snapshots from the simulations demonstrate how the template molecule can be initially enclosed by the Si_{22} structure, but will be pushed out as the channel is formed.

As opposed to the continuum solvent model used in the static ab initio calculations, the current simulations provide insight in explicit silica-water interactions. The closest water molecules are arranged in a single layer around the silica species, while the template shields the inner hydrophobic region from the water layer, which is in accordance with simulations of Catlow et al. [53] Since temperature is also accounted for in the MD simulations, the template can move further apart from the nano-particle compared to the static calculations. Despite the template exiting, the six negative charges on the oxygen atoms of the Si_{22} structure induce strong electrostatic repulsive interactions, which prevent further collapse of the 10-membered ring. It needs to be pointed out that under experimental conditions the charge of the silica species decreases with increasing condensation due to release of hydroxyls. In order to assess the dependence of the results on the assigned charges, we repeated the simulations for the investigated systems, where only 2 or 4 out of total 6 negative charges were present on the Si_{22} cluster and the other charges were transferred to hydroxide anions in solution. During these simulations, the repulsive interactions between closely positioned oxygen atoms on the silica nanocluster are reduced, but the qualitative picture on the relative position of TPA with respect to the Si_{22} cluster is maintained.

In order to assess the effect of explicit solvation and temperature on the position of the template with respect to the Si_{33} precursor, molecular dynamics calculations were also

performed on a single Si_{33} unit (only once deprotonated) and one TPA^+ cation further surrounded by 333 water molecules. A representative snapshot taken during the simulation is shown in Figure 7. Just as in the static calculations, the template moves out of the 10-ring into the direction of the future channel cross-section.

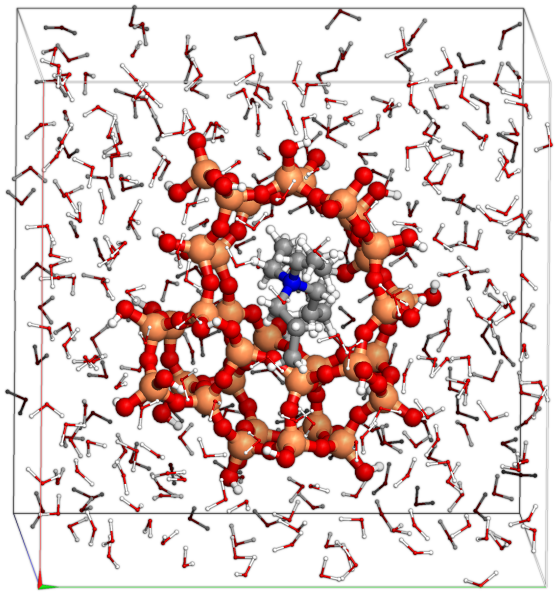


Figure 7: Snapshot from the classical molecular dynamics run of the Si_{33} precursor with one TPA cation, solvated by 333 explicit water molecules.

Further insight into the position of the template with respect to the silica precursor is obtained by analyzing a variety of distances along the trajectory and by calculating a histogram of the nitrogen-oxygen distances from the canonical (NVT) molecular dynamics calculations. This histogram, as shown in Figure 8, describes the frequency with which various distances were observed during the molecular dynamics run. Distinctions between the deprotonated oxygen (circle in figure 8 (a)), the oxygens in the hydrophobic 10-membered rings (which are highlighted in figure 8 (a)), and the other oxygens are made.

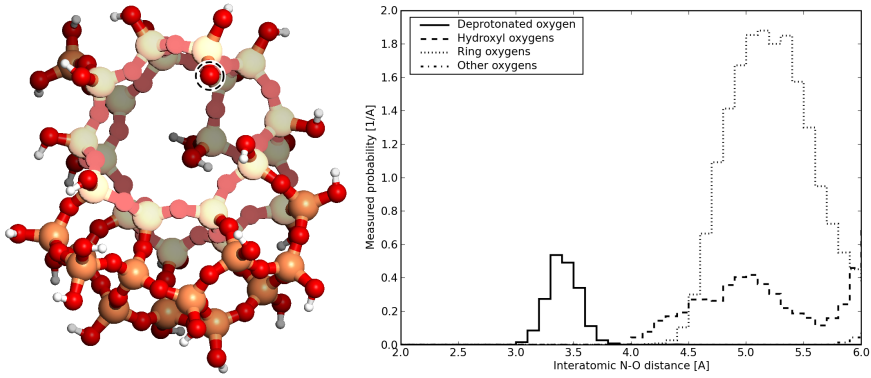


Figure 8: (a) Labeling of various oxygen atoms in the precursor (b) Histogram of the oxygen-nitrogen distances calculated from the canonical (NVT) molecular dynamics calculations

According to the histogram for the oxygen-nitrogen distance, a first probability peak is situated around 3.5 Å, which corresponds to the contact between TPA^+ and the deprotonated oxygen atom, primarily governed by electrostatic Coulomb interactions. A second broader peak around 4.5 Å, originates from coordination of the TPA to the oxygen bridges of the 10-membered rings forming the inner hydrophobic layer. This interaction works independently from the oxygen charge defect, and is similar to what was observed in the ab initio simulations, where the diffuse electron cloud of the oxygen atoms provided additional stabilization of the TPA cation.

The molecular dynamics calculations, allowing elevated temperatures and presence of explicit water molecules, fully corroborate the static ab initio calculations. TPA^+ is pushed out of the straight channel but remains coordinated with the Si_{33} precursor, despite inclusion of temperature effects.

Conclusions

Structure directing agents or templates are the key components in zeolite synthesis as they steer the system to a certain framework topology, yet their specific function on a molecular level remains vague. For Silicalite-1 (whose aluminosilicate counterpart is ZSM-5) a clear solution synthesis technique is based on tetrapropyl ammonium ions (TPA^+) as templates. To gain more insight into the interactions between TPA^+ and possible Silicalite-1 precursors, we have performed two complementary sets of calculations. Both static and dynamic simulations were performed which allowed the investigation of a variety of parameters, that are temperature, charge compensation, implicit/explicit solvation models, and quantum mechanical versus classical force field treatment of the system. All our simulations reveal that, as soon as the typical straight 10-ring channel of Silicalite-1 is formed from smaller oligomers, the TPA^+ template is partially squeezed out of the resulting cavity. This result is found for charged as well as neutral silicate species, indicating that the interaction between the template molecule and a precursor with local zeolite connectivity is not purely based on electrostatic contributions with a localized negative charge. Partial retention of the template is, however, indispensable to prevent collapse of the channel and subsequent hydrolysis. In solvation, the template will also shield the hydrophobic inner region of the channel: without

an organic template, open framework-like building blocks would succumb under the influence of water. The 'half in – half out' adsorption of the template shows that only the hydrophobic appendages are necessary to prevent collapse of the precursor nano-particle. This contribution not necessarily needs to be given by TPA^+, but might be provided by a whole range of organic molecules, which may explain why the MFI structure is observed for many other templates as well.

After this initial phase, the final 'half in – half out' position of the template hints at the possibility of a supramolecular organization of precursors to larger aggregates and maybe even the final crystalline product. The template has already shifted closer to what should be its final position at the channel intersections, which suggests that it could take on a different role, this time in supramolecular organization. Such a hypothesis would fit nicely in the gap between the three major theories regarding the role of the organic cations. This would result in two different functions for TPA^+: the hydrophobic appendages are vital to stabilize any kind of 10-ring containing precursor species, while the structure of TPA^+ fits the channel intersections during and after aggregation. The mobility of TPA^+ during these various stages holds the key to a fundamental understanding of this synthesis process.

Acknowledgment

Calculations have been carried out in the Wrocław Centre of Networking and Supercomputing (PWr), in the Department of Chemical Engineering and Chemistry (TU/e) and at the Center for Molecular Modeling (UGent). Software used for simulations was Cerius2, Materials Studio, Gaussian03, Zeobuilder, CP2K and TRACKS. DL, RD, MW, VVS, JAM and CEAK acknowledge support from the Fund for Scientific Research Flanders (FWO-Vlaanderen), ESA and the Belgian Prodex office. TV also acknowledges the Institute for the Promotion of Innovation through Science and Technology in Flanders (IWT-Vlaanderen) for funding this strategic basic research (SBO).

References

[1] A. Corma, Chem. Rev. 97 (1997) 2373-2420.
[2] A.E. Persson, B.J. Schoeman, J. Sterte, J.E. Otterstedt, Zeolites 14 (1994) 557-567.
[3] T.A.M. Twomey, M. Mackey, H.P.C.E. Kuipers, R.W. Thompson, Zeolites 14 (1994) 162-168.
[4] S.L. Burkett, M.E. Davis, Chem. Mater. 7 (1995) 920-928.
[5] W.H. Dokter, H.F. Van Garderen, T.P.M. Beelen, R.A. van Santen, W. Bras, Angew. Chem. Int. Ed. 34 (1995) 73-75.
[6] R. W. Corkery, B.W. Ninham, Zeolites 18 (1997) 379-386.
[7] R. Gougeon, L. Delmotte, P. Reinheimer, B. Meurer, J.M. Chezeau, Mag. Res. Chem. 36 (1998) 415-421.
[8] B.J. Schoeman, Microporous Mesoporous Mater. 22 (1998) 9-22.
[9] C.S. Tsay, A.S.T. Chiang, Microporous Mesoporous Mater. 26 (1998) 89-99.
[10] J.N. Watson, A.S. Brown, L.E. Iton, J.W. White, J. Chem. Soc. Faraday Trans 94 (1998) 2181-2186.
[11] P.P.E.A. de Moor, T.P.M. Beelen, R.A. Van Santen, L.W. Beck, M.E. Davis, J. Phys. Chem. B 104 (2000) 7600-7611.
[12] Q. Li, B. Mihailova, D. Creaser, J. Sterte, Microporous Mesoporous Mater. 43 (2001) 51-59.
[13] S. Mintova, N.H. Olson, J. Senker, T. Bein, Angew. Chem. Int. Ed. 41 (2002) 2558-2561.

[14] S. Yang, A. Navrotsky, Chem. Mater. 16 (2004) 3682-3687.
[15] C.H. Cheng, D.F. Shantz, J. Phys. Chem. B 110 (2006) 313-318.
[16] T.M. Davis, T.O. Drews, H. Ramanan, C. He, J. Dong, H. Schnablegger, M.A. Katsoulakis, E. Kokkoli, A.V. McCormick, L.R. Penn, M. Tsapatsis, Nat. Mater. 5 (2006) 400-408.
[17] M. Haouas, F. Taulelle, J. Phys. Chem. B 110 (2006) 3007-3014.
[18] C.T.G. Knight, J. Wang, S.D. Kinrade, Phys. Chem. Chem. Phys. 8 (2006) 3099-3103.
[19] J.D. Rimer, J.M. Fedeyko, D.G. Vlachos, R.F. Lobo, Chem. Eur. J. 12 (2006) 2926-2934.
[20] R.A. van Santen, Nature 444 (2006) 46-47.
[21] A. Aerts, L.R.A. Follens, M. Haouas, T.P. Caremans, M.A. Delsuc, B. Loppinet, J. Vermant, B. Goderis, F. Taulelle, J.A. Martens, C.E.A. Kirschhock, Chem. Mater. 19 (2007) 3448-3454.
[22] Kirschhock, C. E. A.; Aerts, A.; Martens, J. A. Stud. Surf. Sci. Catal. 2007, 170, 1473-1478.
[23] S. Kumar, T.M. Davis, H. Ramanan, R.L. Penn, M. Tsapatsis, J. Phys. Chem. B 111 (2007) 3398-3403.
[24] S.A. Pelster, R. Kalamajka, W. Schrader, F. Schüth, Angew. Chem. Int. Ed. 46 (2007) 2299-2302.
[25] A. Patis, V. Dracopoulos, V. Nikolakis, J. Phys. Chem. C 111 (2007) 17478-17484.
[26] C.A. Fyfe, R.J. Darton, C. Schneider, F. Scheffler, J. Phys. Chem. C 112 (2008) 80-88.
[27] C.S. Cundy, M. Henty, R. Plaisted, Zeolites 15 (1995) 353-372.
[28] C.S. Cundy, P.A. Cox, Microporous Mesoporous Mater. 82 (2005) 1-78.
[29] J.N. Watson, L.E. Iton, R.I. Keir, J.C. Thomas, T.L. Dowling, J.W. White, J. Phys. Chem. B 101 (1997) 10094-10104.
[30] C.E.A. Kirschhock, R. Ravishankar, P.A. Jacobs, J. A. Martens, J. Phys. Chem. B 103 (1999) 11021-11027.
[31] C.J.Y. Houssin, C.E.A. Kirschhock, P.C.M.M. Magusin, B.L. Mojet, P.J. Grobet, P.A. Jacobs, J.A. Martens, R. A. van Santen, Phys. Chem. Chem. Phys. 5 (2003) 3518-3524.
[32] P.-P.E.A. de Moor, T.P.M. Beelen, B.U. Komanschek, L.W. Beck, P. Wagner, M.E. Davis, R.A. van Santen, Chem. Eur. J. 5 (1999) 2083-2088.
[33] D. Liang, L.R.A. Follens, A. Aerts, J.A. Martens, G. VanTendeloo, C.E.A. Kirschhock, J. Phys. Chem. C 111 (2007) 14283-14285.
[34] S.L. Burkett, M.E. Davis, Chem. Mater. 7 (1995) 1453-1463.
[35] D. Lewis, D. Willock, C. Catlow, J.M. Thomas, G. Hutchings, Nature 382 (1996) 604-606.
[36] S.D. Kinrade, C.T.G. Knight, D.L. Pole, R.T. Syvitski, Inorg. Chem. 37 (1998) 4272-4277.
[37] S. Caratzoulas, D.G. Vlachos, J. Phys. Chem. B 112 (2008) 7-10.
[38] Caratzoulas, S.; Vlachos, D. G.; Tsapatsis, M. J. Am. Chem. Soc. 2006, 128, 16138-16147.
[39] C.E.A. Kirschhock, S.P.B. Kremer, P.J. Grobet, P.A. Jacobs, J.A. Martens, J. Phys. Chem. B 106 (2002) 4897-4900.
[40] J.M. Fedeyko, J. D. Rimer, R.F. Lobo, D.G. Vlachos, J. Phys. Chem. B 108 (2004) 12271-12275.
[41] C.E.A. Kirschhock, R. Ravishankar, F. Verspeurt, P.J. Grobet, P.A. Jacobs, J. A. Martens, J. Phys. Chem. B 103 (1999) 4965-4971.
[42] R. Ravishankar, C.E.A. Kirschhock, P.P. Knops-Gerrits, E.J.P. Feijen, P.J. Grobet, P. Vanoppen, F.C. De Schryver, G. Miehe, H. Fuess, B.J. Schoeman, P.A. Jacobs, J.A. Martens, J. Phys. Chem. B 103 (1999) 4960-4964.

[43] C.E.A. Kirschhock, R. Ravishankar, L.V. Looveren, P.A. Jacobs, J. A. Martens, J. Phys. Chem. B 103 (1999) 4972-4978.
[44] D. Lesthaeghe, P. Vansteenkiste, T. Verstraelen, A. Ghysels, C.E.A. Kirschhock, J.A. Martens, V.V. Speybroeck, M. Waroquier, J. Phys. Chem. C 112 (2008) 9186-9191.
[45] C.R.A. Catlow, A.K. Cheetham A. K., in: New Trends in Materials Chemistry, eds. NATO Scientific Affairs Division (Kluwer Academic, London, 1997).
[46] C.R.A. Catlow;, R.A. van Santen R. A., Computer Modelling of Microporous Materials (Academic Press, London, 2004).
[47] S. Yip, Handbook of materials modeling (Springer, London, 2005).
[48] S.M. Auerbach, M.H. Ford, P.A. Monson, Curr. Opin. Colloid Interface Sci. 10 (2005) 220-225.
[49] C.R.A. Catlow, D.S. Coombes, D.W. Lewis, J.C.G. Pereira, Chem. Mater. 10 (1998) 3249-3265.
[50] J.C.G. Pereira, C.R.A. Catlow, G.D. Price, J. Phys. Chem. A 103 (1999) 3252-3267.
[51] J.C.G. Pereira, C.R.A. Catlow, G.D. Price, J. Phys. Chem. A 103 (1999) 3268-3284.
[52] J.C.G. Pereira, C.R.A. Catlow, G.D. Price, J. Phys.Chem. A 106 (2002) 130-148.
[53] D.W. Lewis, C.R.A. Catlow, J.M. Thomas, Faraday Discuss. 106 (1997) 451-471.
[54] N.D. Rao, L.D. Gelb, J. Phys. Chem. B 108 (2004) 12418-12428.
[55] B.P. Feuston, S.H. Garofalini, J. Phys. Chem. 94 (2004) 5351-5356.
[56] M.J. Mora-Fonz, C.R.A. Catlow, D.W. Lewis, Angew. Chem. Int. Ed. 44 (2005) 3082-3086.
[57] M.J. Mora-Fonz, S. Hamad, C.R.A. Catlow, Molecular Physics, 105 (2007) 177-187.
[58] M.G. Wu, M.W.J. Deem, Chem. Phys. 116 (2002) 2125
[59] M. Jorge, S.M. Auerbach, P.A. Monson, J. Am. Chem. Soc. 127 (2005) 14388
[60] M.J. Frisch, et al., Gaussian 03, Revision C.02 (Gaussian Inc., Wallingford CT, 2004).
[61] http://cp2k.berlios.de
[62] P.C.M.M. Magusin, V.E. Zorin, A. Aerts, C.J.Y. Houssin, A.L. Yakovlev, C.E.A. Kirschhock, J.A. Martens, R.A. van Santen, J. Phys. Chem. B 109 (2005) 22767-22774.
[63] A.D. Becke, J. Chem. Phys. 98 (2005) 5648-5652.
[64] S.A. Zygmunt, R.M. Mueller, LA. Curtiss, L.E. Iton, J. Mol. Struct. 430 (1998) 9-16.
[65] V. Barone, M. Cossi, J. Phys. Chem. A 102 (1998) 1995-2001.
[66] M. Cossi, N. Rega, G. Scalmani, V. Barone, Comp. Chem. 24 (2003) 669-681.
[67] J.P. Perdew, K. Burke, M. Ernzerhof, Phys. Rev. Lett. 77 (1996) 3865-3868.
[68] G. Lippert, J. Hutter, P. Ballone, M. Parrinello, J. Phys. Chem. 100 (1996) 6231-6235.
[69] S. Goedecker, M. Teter, J. Hutter, Phys. Rev. B 54 (1996) 1703-1710.
[70] C. Hartwigsen, S. Goedecker, J. Hutter, Phys. Rev. B 58 (1998) 3641-3662.
[71] T. Verstraelen, V. Van Speybroeck, M. Waroquier, J. Chem. Inf. Model. 48 (2008) 1530-1541
[72] T.T. Trinh, A.P.J. Jansen, R.A. van Santen, J. Phys. Chem. B 110 (2006) 23099-23106.
[73] V. Van Speybroeck, K. Moonen, K. Hemelsoet, C.V. Stevens, M. Waroquier, J. Am. Chem. Soc. 128 (2006) 8468-8478.
[74] H.J.C. Berendsen, J.P.M. Postma, W.F. van Gunsteren, A. DiNola, J.R. Haak, J. Chem. Phys. 81 (1984) 3684-3690.
[75] S.H. Garofalini, G. Martin, J. Phys. Chem. 98 (1994) 1311-1316.
[76] A.C.T. van Duin, S. Dasgupta, F. Lorant, W.A. Goddard III, J. Phys. Chem. A 105 (2001) 9396-9409.
[77] K.D. Nielson, A.C.T. van Duin, J. Oxgaard, W.-Q. Deng, W.A. Goddard III, J. Phys. Chem. A 109 (2005) 493-499.
[78] A. Strachan, E.M. Kober, A.C.T. van Duin, J. Oxgaard, W.A. Goddard III, J. Chem. Phys. 122 (2005) 54501-54510.

[79] A. Strachan, A.C.T. van Duin, D. Chakraborty, S. Dasgupta, W.A. Goddard III, Phys. Rev. Lett. 91 (2003) 98301-98304.
[80] B.W.H. van Beest, G.J. Kramer, R.A. van Santen, Phys. Rev. Lett. 64 (1990) 1955-1958.
[81] P. Dauber-Osguthorpe, V.A. Roberts, D.J. Osguthorpe, J. Wolff, M. Genest Hagler, Proteins: Structure, Function and Genetics 4 (1988) 31-47.
[82] K.P. Schroder, J. Sauer, M. Leslie, C.R.A. Catlow, J.M. Thomas, Chem. Phys. Lett. 188 (1992) 320-325.
[83] J.D. Gale, N.J. Henson, J. Chem. Soc. Faraday Trans. 90 (1994) 3175-3179.
[84] A.K. Rappe, C.J. Casewit, K.S. Colwell, W.A. Goddard III, W.M. Skiff, J. Am. Chem. Soc. 114 (1992) 10024-10035.
[85] T. Verstraelen, V. Van Speybroeck, M. Waroquier, J. Chem. Inf. Model. 48 (2008) 2414-2424
[86] R. Fricke, H. Kosslick, G. Lischke, M. Richter, Chem. Rev. 100 (2000) 2303-2405.
[87] D. Lesthaeghe, B. De Sterck, V. Van Speybroeck, G.B. Marin, M. Waroquier, Angew. Chem. Int. Ed. 46 (2007) 1311-1314.
[88] This trajectory was approximated roughly by applying harmonic restraints [i.e. adding $E=1/2\ k\ (x - x_0)^2$ to the total energy term, with k = 46.9 kJ/(mol Å²)] on equidistant points of a collective variable. This collective variable was defined as the component in channel direction of the vector between the geometric centers of the framework and the template.
[89] C.P. Kelly, C.J. Cramer, D.G.J. Truhlar, Chem. Theory Comput. 1 (2005) 1133-1152.
[90] L.M. Pratt, A. Streitwieser, J. Org. Chem. 68 (2003) 2830-2838.
[91] A. Navrotsky, Current Opinion in Colloid & Interface Science 10 (2005) 195– 202

Part 4: Other Modeling Applications

Paper 8: "Ab initio calculation of entropy and heat capacity of gas-phase n-alkanes with hetero elements O and S: ethers/alcohols and sulfides/thiols"

Peter Vansteenkiste, Toon Verstraelen, Veronique Van Speybroeck, Michel Waroquier

Chemical Physics, **2006**, 328, 251 - 258

Available online at www.sciencedirect.com
ScienceDirect
Chemical Physics 328 (2006) 251–258
Chemical Physics
www.elsevier.com/locate/chemphys

Ab initio calculation of entropy and heat capacity of gas-phase *n*-alkanes with hetero-elements O and S: Ethers/alcohols and sulfides/thiols

P. Vansteenkiste, T. Verstraelen, V. Van Speybroeck, M. Waroquier *

Center of Molecular Modeling, Laboratory of Theoretical Physics, Ghent University, Proeftuinstraat 86, B-9000 Ghent, Belgium

Received 13 April 2006; accepted 5 July 2006
Available online 12 July 2006

Abstract

In this paper, the performance of the one-dimensional hindered rotor approach (1D-HR) is evaluated for *n*-alkanes with hetero-elements O or S. The internal rotations in these molecules show a behavior distinct from those in *n*-alkanes, for which 1D-HR is a cost-efficient method to describe the thermochemical features (entropy and heat capacity). It turns out that also for ethers, alcohols, sulfides and thiols this approach gives a satisfactory experimental agreement. This work confirms earlier results, and consolidates the assumption that the 1D-HR model is highly suitable for reproducing thermodynamic properties of single chain molecules, and that multi-dimensional coupled hindered rotor approaches (*n*D-HR) are not necessarily required for attaining high accuracy. Moreover, it seems that the 1D-HR results are almost independent of the details of the level of theory.

Keywords: Hindered rotor; Entropy; Heat capacity

1. Introduction

The microscopic evaluation of thermodynamic properties of stable species and kinetic data for chemical reactions has now found widespread use in physical chemistry. For molecules containing various single bonds, the one-dimensional hindered rotor (1D-HR) treatment has become an essential tool for an accurate ab initio evaluation of chemical properties. Various works in the literature have shown that the standard harmonic oscillator approach (HO) is inadequate for the treatment of large amplitude vibrations and a correct description of microscopic partition functions and deduced thermochemical and kinetic quantities, and should be corrected with a hindered rotor treatment of the internal rotations. There are different ways to implement the HR concept:

(i) the simplest corrections can be obtained from the tabulated values proposed by Pitzer in the early days [1],
(ii) later interpolating formulae between harmonic oscillator (HO) and free rotor (FR) treatments were proposed by Truhlar and co-workers [2],
(iii) and more recently, a variety of 'full' treatments are proposed by various groups [3–8] in which calculated potentials are used and moments of inertia are calculated from the optimized geometries.

In a previous paper of the authors [7], it was found that the one-dimensional model (1D-HR) was able to reproduce the thermodynamic features – more specifically, third law entropy and heat capacity – of *n*-alkanes quite well (within the level of theory B3LYP/6-311g**).

This hindered rotor study was further elaborated in a more recent paper, where the origin of this good behavior of the 1D model was unravelled [9]. Implementation of the more advanced, coupled 2D-HR method on pentane and the 3D-HR method on hexane, did not result in an improvement

* Corresponding author. Tel.: +32 9 264 65 59.
E-mail address: michel.waroquier@ugent.be (M. Waroquier).

0301-0104/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.chemphys.2006.07.006

of the predictions of entropy, but confirmed the results obtained with the 1D-HR approach. It turned out that the effects of two approximative ingredients inherent to the 1D-HR model – construction of the 1D rotational potential energy and the use of a constant reduced moment of inertia – are systematically cancelled. The conclusions made on *n*-alkanes may not a priori be extended to other single chain molecules.

The goal of this study is a verification whether the excellent performance of the 1D-HR for *n*-alkanes holds for compounds such as ethers and alcohols or sulfides and thiols where a CH_2 fragment is substituted by the hetero-elements O or S. Sufficient experimental data are available to validate the various approximative schemes to handle internal rotors. As already suggested by East and Radom [3] it is not excluded that the standard procedure coinciding with 1D-HR does no longer achieve the desired accuracy for some of these molecules, especially since a conformational study showed that these molecules have a distinct behavior from *n*-alkanes for internal rotations close to the hetero element [10]. We therefore focus on those specific rotations.

The first step in any ab initio study is the selection of an appropriate level of theory. Since there are not always experimental data to benchmark the theoretical data, one has to rely on general studies of similar molecules. This does not give any guaranties. It is therefore important to validate the sensitivity of the HR-corrections to the description of the potential energy. From Ref. [10] one learns that within the B3LYP method, the basis sets 6-31+g* (B1) and 6-311g** (B2) select different geometries as most stable conformer, with only small potential energy differences. Experimental verification suggests that the B1 basis produces the correct results for ethers and alcohols. However, for sulfides and thiols no definite conclusion may be drawn.

Use of the correlation consistent Dunning basis sets (cc-pVDZ and aug-cc-pVDZ) [11] on dipropylether learns that the inclusion of diffuse functions is the crucial ingredient in discriminating between the two types of results on conformational energies obtained by the two basis sets B1 and B2.

We will calculate the entropy and heat capacity of ethers and alcohols with both B3LYP/B1 and B3LYP/B2 levels of theory in the 1D-HR *and* 2D-HR approaches, in order to draw some conclusions about the sensitivity of the HR approach to these differences.

2. Theory

For the evaluation of thermodynamic properties such as entropy and heat capacity directly from the molecular properties, the molecular partition function is needed. In the present case, we use a mixed Harmonic Oscillator/Hindered Rotor model in which all internal motions, except for the internal rotations, are approximated as independent harmonic oscillators, without any additional scaling factors.

The partition function belonging to the internal rotations depends critically upon the way both potential and kinetic energy contributions of the system are calculated.

In the one-dimensional approach (1D-HR) the usual approximation for the potential energy is used: for each internal rotation a one-dimensional potential energy curve (1D-PES) is calculated, and the total multi-dimensional potential energy surface (*m*D-PES) is assumed to be the sum of these one-dimensional contributions:

$$V^{m\mathrm{D}}(\phi_1,\ldots,\phi_m) \approx \sum_{i=1}^{m} V_i^{1\mathrm{D}}(\phi_i). \quad (1)$$

In the 2D-HR method, multiple two-dimensional potential energy surfaces (2D-PES) are constructed and the final potential energy is obtained by subtracting all 1D-PES which are counted twice [12]:

$$V^{m\mathrm{D}}(\phi_1,\ldots,\phi_m) \approx \sum_{i=1}^{m-1} V_{i(i+1)}^{2\mathrm{D}}(\phi_i,\phi_{i+1}) - \sum_{i=2}^{m-1} V_i^{1\mathrm{D}}(\phi_i). \quad (2)$$

The calculation of the moments of inertia is also performed according to well-established methods described in previous papers of the authors [6,7,9,12]: fixed reduced moments of inertia for each single internal rotation for the 1D-HR scheme, in contrast to variable moments of inertia of both global and internal rotations for the 2D-HR model. The calculation of the kinetic energy matrix $A(\phi_1,\ldots,\phi_m)$ is needed for the latter method [13,9,14].

3. Labelling convention for conformers

In order to unambiguously describe the different (reference) conformers, a consistent labelling system must be introduced. This was already done in a previous study on alcohols, thiols, ethers and sulfides [10], and a short overview is outlined in this paragraph.

Individual conformations about an internal rotation are defined as *t*, *g*− and *g*+, corresponding to a *trans* or a *gauche* ∓ orientation. When multiple internal rotations within a molecule are considered, the individual conformation of each rotation has to be assigned. The appropriate labelling convention for a sequence of internal rotations is illustrated in Fig. 1. For alcohols and thiols (Fig. 1a), the first internal rotation (with dihedral angle ϕ_{l1}) is about the CX bond (X = O or S). The other rotations are labelled as ϕ_{lx} where *x* indicates the position of the CC rotation axis in reference to the CX bond. Also the position of the hydroxyl top is written explicitly. For example, the HO*g*+*tg*−*t*

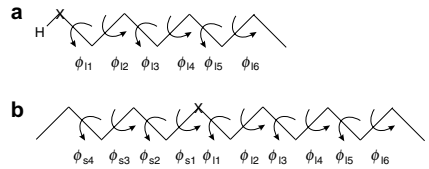


Fig. 1. Labels used to identify specific internal rotations (with their dihedral angles) in: (a) primary alcohols (X = O) and thiols (X = S) and (b) in ethers (X = O) and sulfides (X = S).

conformer of 1-pentanol (or HSg+*tg*−*t* for 1-pentane thiol) stands for $\phi_{l1} = g+$, $\phi_{l2} = t$, $\phi_{l3} = g-$, and $\phi_{l4} = g-$, where l4 indicates the ethyl torsion, l3 the propyl torsion, etc.

For ethers and sulfides, the same convention applies, but one has to distinguish between the two alkyl fragments (Fig. 1b). The 'l' subscript indicates that the rotation is situated in the longest alkyl top on the hetero-element, and the 's' refers to the shortest alkyl branch. We refer to a conformer by specifying the individual conformations in the order $\phi_{s\max} \ldots \phi_{s1} X \phi_{l1} \ldots \phi_{l\max}$, with X = O or S. For example, in ethyl propyl ether *t*Og−*g*+ stands for $\phi_{s1} = t$, $\phi_{l1} = g-$, and $\phi_{l2} = g+$, while the same configuration in ethyl propyl sulfide is referred to as *t*Sg−*g*+.

4. Potential energy profiles

Because the hetero compounds under study in this work exhibit several (very) low energy conformers [10], one could expect multi-dimensional potential energy surfaces to be different than those obtained in pure *n*-alkanes [9,12]. We calculate some two-dimensional energy surfaces (2D-PES) that can serve as examples for a whole series of compounds. For *alcohols and thiols*, 1-hexanol and 1-hexanethiol are selected, and all 2D-surfaces are calculated. These plots are displayed in Fig. 2. While the –OH and –*SH* rotations exhibit a very specific behavior, the other surfaces are only slightly different than corresponding plots in *n*-alkanes. Note that for increasing ϕ_{lx} the *g*−*g*+ double minimum becomes more pronounced. Less subtle differences are found in *ethers and sulfides*, for which we studied the $\phi_{s1}\phi_{l1}$ and $\phi_{l1}\phi_{l2}$ surfaces, illustrated in methyl propyl ether (MPE) and sulfide (MPS), di-ethyl ether (DEE) and sulfide (DES), and finally di-propyl ether (DPE) and sulfide (DPS).

The $\phi_{l1}\phi_{l2}$ potential energy surfaces (in MPE, DPE, MPS and DPS: Fig. 3) show, compared to a typical *n*-alkane 2D-PES, mainly one deviating coupling effect: in ethers the typical alkane double minima around the *g*−*g*+ geometry is completely destroyed (compare with *e.g.* the $\phi_{l4}\phi_{l5}$ surfaces of hexanol and hexanethiol). The $\phi_{s1}\phi_{l1}$ potential energy surfaces (Fig. 4) are very instructive, and show very distinct features for ethers and sulfides. Not only is the behavior around the *g*−*g*+ geometry different (again), but there are also more minima on the sulfide surfaces. These numerous minima are connected by broad, low energy valleys. Even the energy barriers are substantially lower than in ethers (note the different scale of the contour plots in Fig. 4). The influence of the level of theory on the 2D-profiles is not very pronounced. We display the 2D-PES obtained with the two basis sets B1 and B2 and we scarcely notice any discrepancy on the qualitative level. Still, both levels of theory exhibit a different global minimum, which will result in different one-dimensional paths to be used in the 1D-HR approach.

In particular, the largest deviations are noticed in DPE and to illustrate the variations of the potential energy surface we give the various conformational energies of DPE for the two basis sets B1 and B2 in Table 1. We also include

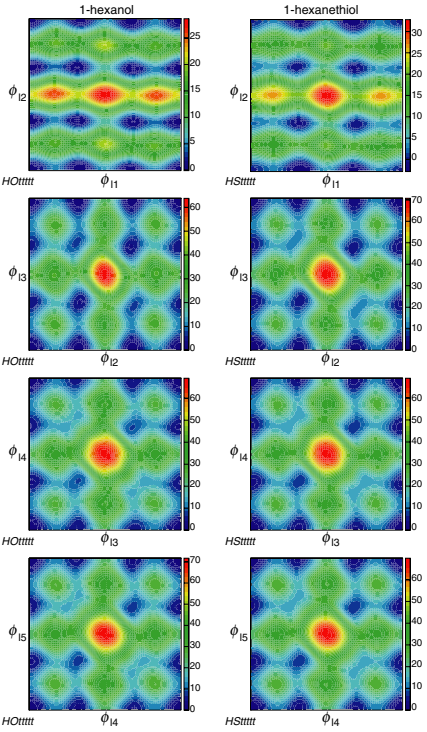


Fig. 2. Two-dimensional potential energy profiles of 1-hexanol (B3LYP/B1) and 1-hexane thiol (B3LYP/B2). The dihedral variation is relative to the all-*trans* conformer: HO*ttttt* and HS*ttttt*.

the predictions for two correlation consistent Dunning basis sets [11]. The similarity between the values obtained within the two basis sets involving diffuse functions [6-31+g* (B1) and aug-cc-pVDZ] is striking, and we may conclude that the deviations noticed between the B1 and B2 results are directly linked to the presence or not of diffuse functions in the basis set under consideration. The one-dimensional cuts starting from the global minimum conformation will clearly be different in the two types of bases, giving rise to a completely different 1D-HR description, and causing strongly deviating results.

Both in 1D-HR and 2D-HR, the methyl top rotations are described one-dimensionally. Their energy profiles are determined unambiguously by the barrier height. For *n*-alkanes this barrier converges about the value of 12.1 kJ/mol for long enough chains [7]. In *n*-alkanes with substitution of a CH_2 fragment by an hetero element O or S, the methyl barrier height depends on the position of the hetero

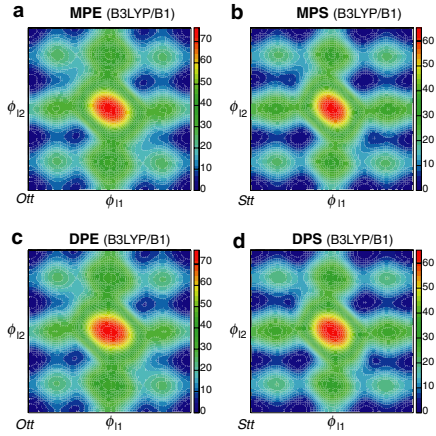


Fig. 3. Two-dimensional potential energy profiles (relative to the all-*trans* conformer: O*tt* and S*tt*) of the $\phi_{l1}\phi_{l2}$ energy surfaces in: (a) methyl propyl ether (MPE), (b) methyl propyl sulfide (MPS), (c) di-propyl ether (DPE) and (d) di-propyl sulfide (DPS) on the B3LYP/B1 level of theory.

element in the chain with respect to the ending methyl top. The largest variations take place in the smallest molecules: C_2H_6O may represent ethanol as well as dimethyl ether (DME) but the barrier varies from 11 kJ/mol (in DME) to 15.5 kJ/mol (ethanol) (Fig. 5). The variation of the barriers decreases for longer hetero substituted alkanes and converges to the pure *n*-alkane value of 12.1 kJ/mol. The lowest barrier is noticed in molecules, where the methyl top is nearest to the hetero element and its magnitude largely depends on the nature of the hetero element: 6 kJ/mol for methyl tops close to S, which may be the result of the (very) long bond length of S–C.

The methyl barrier heights are almost independent of the choice of basis, with differences less than 0.2 kJ/mol.

5. Thermodynamic properties

5.1. Oxygen compounds

Table 2 gives a summary of the results obtained in the considered set of alcohols and ethers[1]. The entropy and heat capacity values are calculated at different levels of theory: B3LYP/6-31+g*(B1) and B3LYP/6-311g**(B2), and also with different calculation schemes: the harmonic oscillator model (HO), and the two hindered rotor schemes 1D-HR and 2D-HR. For transparency, we also report the results from the works from Guthrie [16] and Chen [17] when available. The reference values are taken from Ref. [15].

[1] Supplementary material can be found, online on ScienceDirect at doi:10.1016/j.chemphys.2006.07.006 (http://www.sciencedirect.com).

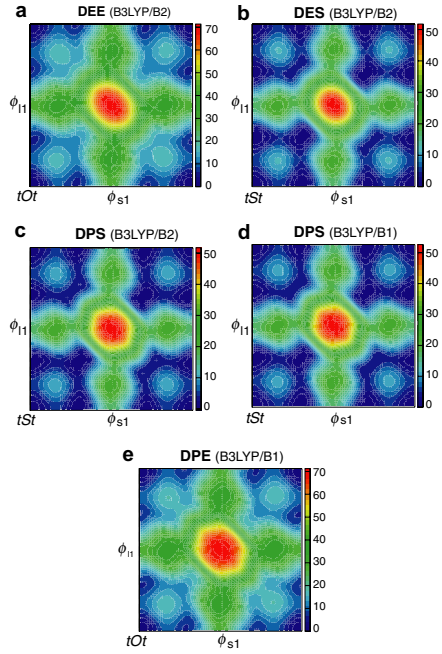


Fig. 4. Two-dimensional potential energy profiles of the $\phi_{s1}\phi_{l1}$ energy surfaces in di-ethyl ether (DEE), di-propyl ether (DPE) and di-ethyl sulfide (DES), di-propyl sulfide (DPS): (a) DEE on the B3LYP/B2 level of theory, (b) DES on the B3LYP/B2 level of theory, (c) DPS on the B3LYP/B2 level of theory, (d) DPS on the B3LYP/B1 level of theory and (e) DPE on the B3LYP/6-31+g*(B1) level of theory.

Table 1
Influence of the basis set on the relative energy of some conformers of di-propyl ether, and on the entropy and heat capacity calculated in the HO approximation

kJ/mol	6-31+g(d) (B1)	6-311g(d, p) (B2)	cc-pVDZ	aug-cc-pVDZ
*tt*O*tt*	0.00	2.51	3.36	0.24
*tt*O*tg*	0.22	1.28	1.66	0.11
*tt*O*gt*	6.24	8.60	8.69	6.24
*tg*O*gt*	11.69	13.36	12.93	11.99
*gt*O*tg*	0.54	0.00	0.00	0.00
J/mol K	6-31+g(d) (B1)	6-311g(d, p) (B2)	cc-pVDZ	aug-cc-pVDZ
S	394.03	388.58	389.17	392.06
C	144.38	143.96	144.46	145.29

The entropy values are given at 298.15 K. All HO predictions systematically underestimate the experimental data. The discrepancies are of the order of 10% of the total magnitude. The full ab initio corrections arising from taking into account 1D internal rotors bring the theoretical

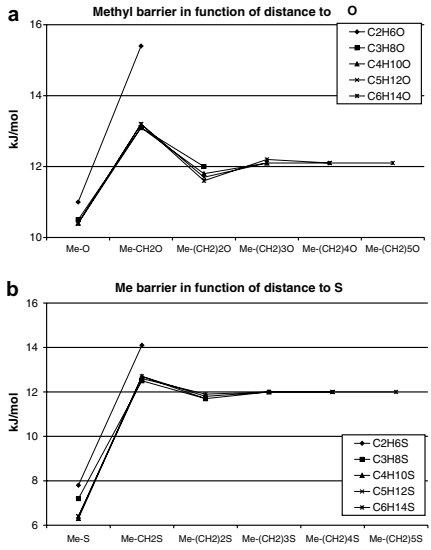


Fig. 5. Methyl top barrier height in function of chain length and position of the hetero element in the chain, calculated at the B3LYP/B2 level of theory. Note that for *n*-alkanes the barrier amounts to about 12.1 kJ/mol.

predictions to values which are very close to the experimental estimates (within a few J/mol K) independent of the level of theory B1 or B2. In some ethers, the agreement is even spectacular, especially on the B3LYP/6-311g**(B2) level of theory. The 2D-HR entropy values are somewhat higher than the 1D-HR results, and the discrepancy with the experimental values increases.

This excellent quality of the 1D-HR results was also noted for *n*-alkanes [7]. Further research revealed that, for those molecules, a cancellation of errors (potential versus kinetic energy) in the construction of 1D-HR partition function [9] lies at the origin of its success.

For some selected molecules the 2D-HR procedure has also been applied. It turns out that the agreement with experiment is not improving. In most of the cases, the 2D-HR model overestimates the entropy. To illustrate with an example: in dipropyl ether (DPE) the 2D-HR correction with respect to the HO estimate amounts to 40.5 J/mol K, compared to 35.7 J/mol K in the 1D-HR model. They both overestimate the experimental value with, respectively, 12.1 and 7.2 J/mol K. This result is rather surprising as it does not happen frequently that a more elaborated method turns out to be less adequate than the more simplified one. It is probably due to some hazardous cancellation of errors as already emphasized in previous works (Refs. [9,14]). In addition the calculated value of the entropy for dipropyl ether strongly depends on the basis set chosen

Table 2
Entropy and heat capacity of alcohols and ethers at 298.15 K

	Basis	S (298.15 K)	C (298.15 K)
Ethanol			
Ref. values [15]		282.59	66.17
HO	B1	269.68	64.65
	B2	268.78	64.63
1D-HR	B1	278.11	64.88
	B2	282.33	65.38
Dimethyl ether			
Ref. values [15]		266.69	66.03
HO	B1	264.47	62.43
	B2	264.48	62.68
1D-HR	B1	267.28	65.04
	B2	269.15	65.26
1-Propanol			
Ref. values [15]		324.72	93,48
HO	B1	301.81	85.03
	B2	298.93	84.47
1D-HR	B1	324.25	86.60
	B2	322.05	86.84
Methyl ethyl ether			
Ref. values [15]		310.62	92.04
HO	B1	302.05	82.99
	B2	302.26	83.32
1D-HR	B1	309.88	92.94
	B2	311.89	92.77
Chen [17]		312.63	92.47
1-Butanol			
Ref. values [15]		362.75	111.91
HO	B1	333.82	105.22
	B2	331.28	104.78
1D-HR	B1	365.44	112.09
	B2	362.36	111.56
Guthrie [16]		361.00	
Methyl propyl ether			
Ref. values [15]		349.13	112.38
HO	B1	334.60	103.34
	B2	332.56	103.29
1D-HR	B1	351.13	114.22
	B2	349.71	112.75
2D-HR	B1	352.71	114.12
Guthrie [16]		352.21	
Diethyl ether			
Ref. values [15]		341.00	118.26
HO	B1	327.74	103.53
	B2	328.56	103.99
1D-HR	B1	341.96	120.64
	B2	341.61	119.80
2D-HR	B2	347.90	119.66
Guthrie [16]		345.97	
1-Pentanol			
Ref. values [15]		402.50	133.70
HO	B1	364.55	125.32
	B2	362.57	125.10
1D-HR	B1	408.39	137.48
	B2	402.81	136.21
Guthrie [16]		401.45	
Methyl butyl ether			
Ref. values [15]		390.10	
HO	B1	366.80	123.57
	B2	364.74	123.70

(*continued on next page*)

Table 2 (*continued*)

	Basis	S (298.15 K)	C (298.15 K)
1D-HR	B1	389.05	139.87
	B2	389.46	138.61
Ethyl propyl ether			
Ref. values [15]		388.10	134.49
HO	B1	366.56	123.94
	B2	365.01	123.99
1D-HR	B1	384.83	141.82
	B2	386.91	139.85
1-Hexanol			
Ref. values [15]		441.50	156.97
HO	B1	398.31	145.84
	B2	395.26	145.57
1D-HR	B1	442.06	163.17
	B2	445.57	161.02
2D-HR	B1	446.32	161.97
Guthrie [16]		441.83	
Methyl pentyl ether			
Ref. values [15]			
HO	B1	398.95	143.89
	B2	397.28	144.11
1D-HR	B1	421.17	165.18
	B2	417.12	163.74
Ethyl butyl ether			
Ref. values [15]		429.00	157.75
HO	B1	398.70	144.12
	B2	396.45	144.32
1D-HR	B1	424.12	167.46
	B2	427.72	165.93
Dipropyl ether			
Ref. values [15]		422.50	153.41
HO	B1	394.03	144.38
	B2	388.58	143.96
1D-HR	B1	429.74	163.08
	B2	420.92	160.23
2D-HR	B1	434.57	163.49
Guthrie [16]		428.53	

and can differ substantially (approximately 9 J/mol K). It has already been observed that the presence of diffuse functions in the basis set lies on the origin of these discrepancies. This is confirmed by the estimates made in the more advanced cc-pVDZ and aug-cc-pVDZ basis sets (see Table 1) within the full HO approximation. The two basis sets B1 and B2 predict HO estimates differing by 5.5 J/mol K (calculated with respect to the reference conformers *tt*O*tt* and *gt*O*tg*, respectively). This suggests that the vibrational modes strongly depend on the structure, and assuming their harmonic description fixed for other geometries than the reference conformer may generate large errors. In a recent paper, we have shown that even minor changes in the vibrational modes can lead to different predictions of thermodynamic quantities [14]. In this work, a more elaborated 'extended hindered rotor' (EHR) model has been developed and could be regarded as the 'exact' reference model instead of the 2D-HR approach. However, EHR needs the calculation of the complete *m*D-PES, and is therefore not the most appropriate method for *e.g.* dipropyl ether ($n = 4$).

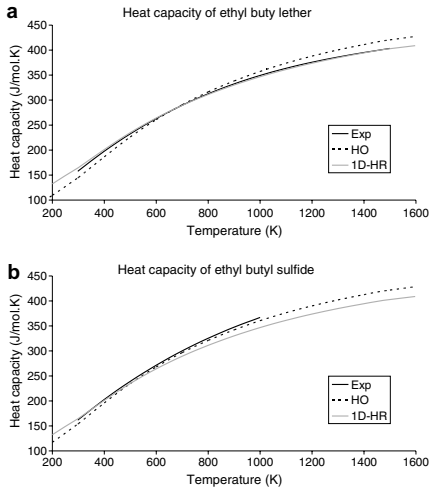


Fig. 6. Heat capacity in J/(mol.K) of: (a) ethyl butyl ether and (b) ethyl butyl sulfide.

The reproduction of the heat capacity is of the same order of accuracy as observed in *n*-alkanes, and almost independent of the specific HR scheme (1D or 2D). The calculated values are somewhat too high for lower temperatures as a result of the classical implementation of our hindered rotor treatments [7]. On the other hand for medium and higher temperatures we achieve a very satisfactory agreement. In Fig. 6a, the temperature behavior of the heat capacity in ethyl butyl ether is plotted. It confirms our conclusions.

5.2. Sulfur compounds

Entropy and heat capacity values for the selected set of thiols and sulfides are presented in Table 3[2]. The reference values are taken from Ref. [15]. When available, the results of Guthrie's work [16] are also reported.

The agreement of the calculated entropy values with experiment is satisfactory but not of the same level as for alcohols and ethers. While the HO results still underestimate the experimental values significantly, the 1D-HR predictions are close to experiment, except for some smaller sulfides. For dimethyl sulfide, methyl ethyl sulfide and diethyl sulfide the HO values are already reproducing the entropy quite well, and the 1D-HR and 2D-HR predictions now exceed the experimental data. Both models predict

[2] Supplementary material can be found online on ScienceDirect at doi:10.1016/j.chemphys.2006.07.006 (http://www.sciencedirect.com).

almost the same entropy values, the differences are small compared with the corrections obtained in alcohols and ethers. Anyway, as before, the best overall agreement is given by the 1D-HR approach.

The analysis of the heat capacities reveals some interesting features. For higher temperatures, the HO approximation will normally provide an upper limit for the heat capacities, as the contribution for each activated mode will be R, while for the HR modes it will tend to $R/2$. This rule indicates that the HO models are expected to overestimate the heat capacity at high temperatures. This rule is not systematically respected in Table 3. For the larger molecules (such as 1-butanethiol) the experimental heat capacities are larger than those predicted in the HO approximation!

As the HR model lowers the heat capacity values by $R/2$ (at high temperatures) for each internal rotation present, the 1D and 2D-HR predictions will be worse than the HO values. This situation is shown in Fig. 6b. This failure in reproducing the correct behavior of the heat capacity in sulfides lies in the presence of low vibrational modes. Contrary to ethers and *n*-alkanes, there are low temperature bending modes resulting from the heavy S atom and the long CS bonds. They are responsible for higher moments of inertia ($I \sim r^2 = |CS|^2$) and hence lower HO frequencies ($\sim (\frac{V''}{I})^{1/2}$; with V'' the second-order derivative of the potential energy along the bending mode). As the vibrational temperatures originating from these bending modes are of the same magnitude of the vibrational temperatures of internal rotations, they are heavily mixed. This is confirmed by visual inspection of the normal modes.

Low and medium vibrational modes contribute strongly to the heat capacity value. Therefore, their frequency should be determined very accurately. This requires a high level of theory.

6. Summary

In this work the thermodynamic properties – entropy and heat capacity – of alcohols/thiols and ethers/sulfides have been the subject of a thorough investigation. In these molecules, several internal rotations are present. These rotations have a tendency to generate multiple conformers with similar energies, and may exhibit, depending on the level of theory, a global potential energy minimum at a conformation different from the all-*trans* geometry. Both properties are different from *n*-alkanes in which there is a distinct energy minimum at the all-*trans* conformation. It was therefore difficult to predict if the 1D-HR approach would be able to reproduce the experimental data on entropy and heat capacity with the same kind of accuracy as for *n*-alkanes.

The 1D-HR method is indeed capable of reproducing entropy values close to experiment. The 2D-HR method, although theoretically more evolved, seems to slightly overestimate the entropy.

The prediction of the heat capacity on the other hand is more problematic: the reproduction of heat capacities in

Table 3
Entropy and heat capacity of thiols and sulfides at 298.15 K

	Basis	S (298.15 K)	C (298.15 K)
Ethanethiol			
Ref. values [15]		296.02	74.37
HO	B2	284.63	70.94
1D-HR	B2	293.45	72.54
Dimethyl sulfide			
Ref. values [15]		285.85	75.22
HO	B2	284.56	72.18
1D-HR	B2	292.79	73.18
1-Propanethiol			
Ref. values [15]		336.50	96.86
HO	B2	317.39	91.63
1D-HR	B2	335.34	95.92
Methyl ethyl sulfide			
Ref. values [15]		333.15	95.04
HO	B2	323.62	92.44
1D-HR	B2	341.42	93.91
1-Butanethiol			
Ref. values [15]		375.20	120.58
HO	B2	348.69	111.84
1D-HR	B2	373.17	120.79
Methyl propyl sulfide			
Ref. values [15]		371.68	117.29
HO	B1	355.85	112.15
	B2	353.75	113.10
1D-HR	B1	375.44	117.21
	B2	374.91	117.76
2D-HR	B1	376.67	117.09
	B2	376.21	116.84
Diethyl sulfide			
Ref. values [15]		368.00	120.11
HO	B2	351.59	112.75
1D-HR	B2	378.55	114.49
2D-HR	B2	378.29	114.56
1-Pentanethiol			
Ref. values [15]		415.39	141.78
HO	B2	380.35	132.16
1D-HR	B2	417.48	146.06
Methyl butyl sulfide			
Ref. values [15]		411.90	140.67
HO	B2	385.95	133.43
1D-HR	B2	414.98	142.99
Guthrie [16]		414.47	
Ethyl propyl sulfide			
Ref. values [15]		414.12	139.02
HO	B2	387.43	133.34
1D-HR	B2	415.95	139.00
Guthrie [16]		405.76	
1-Hexanethiol			
Ref. values [15]		454.70	167.69
HO	B2	411.60	152.51
1D-HR	B2	457.18	170.88
2D-HR	B2	460.64	168.85
Guthrie [16]		457.73	
Methyl pentyl sulfide			
Ref. values [15]		451.18	163.50
HO	B2	418.80	153.85
1D-HR	B2	450.18	168.29
Guthrie [16]		453.67	

(*continued on next page*)

Table 3 (*continued*)

	Basis	S (298.15 K)	C (298.15 K)
Ethyl butyl sulfide			
Ref. values [15]		453.39	161.85
HO	B2	418.56	153.62
1D-HR	B2	453.89	164.11
Dipropyl sulfide			
Ref. values [15]		448.80	161.10
HO	B1	417.17	152.75
	B2	412.56	154.06
1D-HR	B1	444.99	161.25
	B2	451.89	162.75
2D-HR	B1	452.33	160.49
	B2	449.76	160.59
Guthrie [16]		450.91	

alcohols and ethers is satisfactory, but the hindered rotor *and* harmonic oscillator models fail to predict reliable values in thiols and sulfides.

The general conclusion is that the 1D-HR approach is a very satisfactory model to describe the entropy of the presented single chain molecules: primary alcohols and thiols, and ethers and sulfides. The calculated values are rather insensitive to the details of the description: both B3LYP/6-31+g* and B3LYP/6-311g** perform very well, although the basis set B3LYP/6-31+g* including diffuse functions (and in analogy the basis set aug-cc-pVDZ) should be the appropriate level of theory to describe the energy differences between the conformers.

This is a very useful property of 1D-HR, since the time-consuming step of the level of theory study can be reduced considerably.

Acknowledgements

This work is supported by the Fund for Scientific Research – Flanders (FWO) and the Research Board of Ghent University.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.chemphys.2006.07.006.

References

[1] (a) K.S. Pitzer, J. Chem. Phys. 8 (1940) 711;
(b) K.S. Pitzer, W.D. Gwin, J. Chem. Phys. 10 (1942) 428;
(c) K.S. Pitzer, J. Chem. Phys. 14 (1946) 239;
(d) J.E. Kilpatrick, K.S. Pitzer, J. Chem. Phys. 17 (1949) 1064;
(e) D.R. Herschbach, H.S. Johnston, K.S. Pitzer, R.E. Powell, J. Chem. Phys. 25 (1956) 736;
(f) J.C.M. Li, K.S. Pitzer, J. Phys. Chem. 60 (1956) 466.
[2] D.G. Truhlar, J. Comp. Chem. 12 (1991) 266.
[3] A.L.L. East, L. Radom, J. Chem. Phys. 106 (1997) 6655.
[4] (a) E.E. Aubanel, S.H. Robertson, D.M. Wardlaw, J. Chem. Soc. 87 (1991) 2291;
(b) S.H. Robertson, D.M. Wardlaw, Chem. Phys. Lett. 199 (1992) 391;
(c) S.H. Robertson, A.F. Wagner, D.M. Wardlaw, J. Chem. Phys. 103 (1995) 2917;
(d) S.H. Robertson, A.F. Wagner, D.M. Wardlaw, Faraday Discuss. Chem. Soc. 102 (1995) 65;
(e) J. Gang, M.J. Pilling, S.H. Robertson, J. Chem. Soc., Faraday Trans. 92 (1996) 3509;
(f) J. Gang, M.J. Pilling, S.H. Robertson, J. Chem. Soc. 93 (1997) 1481;
(g) S.H. Robertson, A.F. Wagner, D.M. Wardlaw, J. Chem. Phys. 113 (2000) 2648;
(h) S.H. Robertson, A.F. Wagner, D.M. Wardlaw, J. Phys. Chem. A 106 (2002) 2598;
(i) S.H. Robertson, A.F. Wagner, D.M. Wardlaw, J. Chem. Phys. 117 (2002) 593.
[5] J.T. Vivian, S.A. Lehn, J.H. Frederick, J. Chem. Phys. 107 (1997) 6646.
[6] V. Van Speybroeck, D. Van Neck, M. Waroquier, S. Wauters, M. Saeys, G.B. Marin, J. Phys. Chem. A 104 (2000) 10939.
[7] P. Vansteenkiste, V. Van Speybroeck, G.B. Marin, M. Waroquier, J. Phys. Chem. A 107 (2003) 3139.
[8] M. Tafipolsky, R. Schmid, J. Comp. Chem. 26 (2005) 1579.
[9] V. Van Speybroeck, P. Vansteenkiste, D. Van Neck, M. Waroqiuer, Chem. Phys. Lett. 402 (2005) 479.
[10] P. Vansteenkiste, E. Pauwels, V. Van Speybroeck, M. Waroquier, J. Phys. Chem. A 109 (2005) 9617.
[11] (a) T.H. Dunning Jr., J. Chem. Phys. 90 (1989) 1007;
(b) D.E. Woon, T.H. Dunning Jr., J. Chem. Phys. 98 (1993) 1358.
[12] P. Vansteenkiste, V. Van Speybroeck, E. Pauwels, M. Waroquier, Chem. Phys. 314 (2005) 109.
[13] M.L. Eidinoff, J.G. Aston, J. Chem. Phys. 3 (1946) 379.
[14] P. Vansteenkiste, D. Van Neck, V. Van Speybroeck, M. Waroquier, J. Chem. Phys. 124 (2006) (Art. No. 04431).
[15] C.L. Yaws, Chemical Properties Handbook, McGraw-Hill Handbooks, 1999.
[16] J.P. Guthrie, J. Phys. Chem. A 105 (2001) 8495.
[17] C. Chen, J.W. Bozzelli, J. Phys. Chem. A 107 (2003) 4531.

Supplementary material of 'Ab initio calculation of entropy and heat capacity of gas-phase n-alkanes with hetero elements O and S: ethers/alcohols and sulfides/thiols'

P. Vansteenkiste, T. Verstraelen, V. Van Speybroeck, and M. Waroquier [1]

Center of Molecular Modeling, Laboratory of Theoretical Physics, Ghent University, Proeftuinstraat 86, B-9000 Ghent, Belgium

Abstract

In this paper, the performance of the one-dimensional hindered rotor approach (1D-HR) is evaluated for n-alkanes with hetero elements O or S. The internal rotations in these molecules show a behavior distinct from those in n-alkanes, for which 1D-HR is a cost-efficient method to describe the thermochemical features (entropy and heat capacity). It turns out that also for ethers, alcohols, sulfides and thiols this approach gives a satisfactory experimental agreement. This work confirms earlier results, and consolidates the assumption that the 1D-HR model is highly suitable for reproducing thermodynamic properties of single chain molecules, and that multi-dimensional coupled hindered rotor approaches (nD-HR) are not necessarily required for attaining high accuracy. Moreover, it seems that the 1D-HR results are almost independent of the details of the level of theory.

[1] To whom all correspondence should be addressed. Tel: 32 (0)9 264 65 59. E-mail: michel.waroquier@ugent.be

(Table 1)

	basis	S(298.15K)	C(100K)	C(200K)	C(298.15K)	C(400K)	C(600K)	C(800K)	C(1000K)	C(1500K)
ethanol										
ref. values [1]		282,59	39,10	52,45	66,17	80,41	106,41	127,71	143,35	167,56
HO	B1	269,68	39,31	51,64	64,65	80,34	108,22	129,22	145,13	170,53
	B2	268,78	38,46	51,14	64,63	80,58	108,62	129,60	145,44	170,70
1D-HR	B1	278,11	41,05	53,03	64,88	79,31	104,94	124,37	139,27	163,44
	B2	282,33	41,25	53,15	65,38	80,19	106,14	125,48	140,19	163,96
dimethyl ether										
ref. values [1]		266,69	43,18	54,00	66,03	79,18	104,56	126,40	142,82	165,15
HO	B1	264,47	41,25	51,89	62,43	76,79	104,88	127,13	144,10	170,71
	B2	264,48	41,24	51,94	62,68	77,27	105,58	127,87	144,80	171,25
1D-HR	B1	267,28	42,09	54,34	65,04	77,98	102,84	122,97	138,66	163,79
	B2	269,15	42,03	54,22	65,26	78,61	103,78	123,91	139,53	164,41
1-propanol										
ref. values [1]		324,72	53,81	74,56	93,48	111,63	143,05	169,23	190,58	225,92
HO	B1	301,81	48,40	66,41	85,03	107,31	146,58	175,90	197,90	232,63
	B2	298,93	46,38	64,91	84,47	107,40	147,13	176,49	198,43	232,96
1D-HR	B1	324,25	53,08	69,36	86,60	107,19	143,08	169,86	190,14	222,61
	B2	322,05	56,90	69,19	86,84	108,17	144,56	171,21	191,25	223,24
methyl ethyl ether										
ref. values [1]		310,62	53,51	72,80	92,04	111,70	147,38	176,87	198,87	227,63
HO	B1	302,05	49,65	66,18	82,99	104,31	143,85	174,27	197,21	232,97
	B2	302,26	49,73	66,28	83,32	104,91	144,68	175,12	198,01	233,56
1D-HR	B1	309,88	51,26	74,65	92,94	111,01	143,86	170,26	190,81	223,68
	B2	311,89	51,51	74,86	92,77	111,06	144,48	171,10	191,68	224,38
Chen [3]		312.63	52.13	74.85	92.47	110.33	143.01	168.91	188.91	220.83
1-butanol										
ref. values [1]		362,75		81,64	111,91	139,76	185,05	219,82	246,12	286,09
HO	B1	333,82	58,31	81,10	105,22	134,14	184,90	222,61	250,75	294,81
	B2	331,28	56,08	79,59	104,78	134,42	185,68	223,42	251,47	295,27
1D-HR	B1	365,44	65,16	90,82	112,09	137,29	182,00	215,71	241,26	281,97
	B2	362,36	72,73	90,20	111,56	137,82	183,34	216,98	242,32	282,56
Guthrie [2]		361,00								
methyl propyl ether										
ref. values [1]		349,13			112,38	138,66	183,40	218,82	244,63	
HO	B1	334,60	59,92	81,16	103,34	131,23	182,18	220,97	250,03	295,13
	B2	332,56	58,23	80,22	103,29	131,82	183,24	222,06	251,04	295,84
1D-HR	B1	351,13	62,11	90,92	114,22	138,22	181,43	215,37	241,47	282,81
	B2	349,71	62,14	88,22	112,75	139,05	184,12	218,24	244,12	284,75
2D-HR	B1	352,71	62,04	90,23	114,12	138,74	182,10	215,90	241,90	283,17
Guthrie [2]		352,21								
diethyl ether										
ref. values [1]		341,00		92,06	118,26	143,95	188,79	225,01	252,40	292,77
HO	B1	327,74	58,87	80,53	103,53	131,79	182,78	221,39	250,31	295,22
	B2	328,56	59,12	80,72	103,99	132,57	183,81	222,41	251,24	295,89
1D-HR	B1	341,96	61,18	94,94	120,64	143,85	184,87	217,62	243,07	283,67
	B2	341,61	61,92	95,71	119,80	142,96	184,99	218,26	243,88	284,41
2D-HR	B2	347,90	62,89	99,59	119,66	140,82	182,85	216,66	242,68	283,84
Guthrie [2]		345,97								

(Table 1 continued)

	basis	S(298.15K)	C(100K)	C(200K)	C(298.15K)	C(400K)	C(600K)	C(800K)	C(1000K)	C(1500K)
1-pentanol										
ref. values [1]		402,50		96,60	133,70	168,19	224,45	266,71	297,07	341,91
HO	B1	364,55	68,01	95,69	125,32	160,86	223,13	269,25	303,54	356,96
	B2	362,57	65,97	94,33	125,10	161,44	224,23	270,34	304,50	357,57
1D-HR	B1	408,39	79,10	112,70	137,48	167,23	220,86	261,55	292,38	341,31
	B2	402,81	86,83	111,42	136,21	167,35	222,18	262,93	293,57	342,03
Guthrie [2]		401,45								
methyl butyl ether										
ref. values [1]		390,10								
HO	B1	366,80	70,02	95,90	123,57	158,09	220,55	267,71	302,91	357,33
	B2	364,74	68,07	95,04	123,70	158,94	221,87	269,03	304,11	358,17
1D-HR	B1	389,05	75,23	112,73	139,87	168,43	220,44	261,30	292,64	342,20
	B2	389,46	76,31	110,17	138,61	169,55	223,33	264,25	295,32	344,11
ethyl propyl ether										
ref. values [1]		388,10			134,49	168,18	223,89	266,70	298,34	345,65
HO	B1	366,56	69,51	95,63	123,94	158,76	221,17	268,13	303,16	357,40
	B2	365,01	67,86	94,71	123,99	159,52	222,40	269,37	304,29	358,19
1D-HR	B1	384,83	72,41	111,32	141,82	170,95	222,40	262,72	293,72	342,81
	B2	386,91	72,67	108,96	139,85	171,07	224,69	265,43	296,33	344,79
1-hexanol										
ref. values [1]		441,50		113,32	156,97	197,57	263,66	312,88	347,68	398,02
HO	B1	398,31	78,93	110,84	145,84	187,97	261,68	316,13	356,52	419,21
	B2	395,26	76,42	109,29	145,57	188,58	262,88	317,33	357,58	419,91
1D-HR	B1	442,06	93,16	134,76	163,17	197,47	259,93	307,53	343,60	400,72
	B2	445,57	102,11	132,68	161,02	197,07	261,12	308,91	344,83	401,48
2D-HR	B1	446,32	95,89	133,51	161,97	197,52	260,93	308,77	344,91	401,95
Guthrie [2]		441,83								
methyl pentyl ether										
ref. values [1]										
HO	B1	398,95	80,36	110,80	143,89	185,01	258,93	314,48	355,80	419,54
	B2	397,28	78,31	109,88	144,11	186,04	260,49	316,01	357,18	420,50
1D-HR	B1	421,17	89,36	134,62	165,18	198,30	259,27	307,14	343,77	401,56
	B2	417,12	90,42	131,79	163,74	199,47	262,42	310,35	346,69	403,63
ethyl butyl ether										
ref. values [1]		429,00			157,75	197,83	263,12	312,73	349,31	403,45
HO	B1	398,70	79,63	110,36	144,12	185,56	259,47	314,83	356,01	419,59
	B2	396,45	77,56	109,45	144,32	186,54	260,95	316,28	357,31	420,49
1D-HR	B1	424,12	85,76	133,19	167,46	201,08	261,29	308,56	344,84	402,17
	B2	427,72	86,59	130,81	165,93	201,82	264,06	311,55	347,61	404,18
dipropyl ether										
ref. values [1]		422,50			153,41	196,75	263,55	312,67	349,10	
HO	B1	394,03	80,41	110,84	144,38	185,72	259,53	314,85	356,01	419,57
	B2	388,58	76,50	108,67	143,96	186,43	260,97	316,29	357,30	420,46
1D-HR	B1	429,74	84,12	127,97	163,08	198,04	259,85	307,77	344,33	401,92
	B2	420,92	85,61	121,37	160,23	200,09	265,06	312,92	348,91	405,12
2D-HR	B1	434,57	84,72	129,88	163,49	197,64	259,50	307,70	344,53	402,56
Guthrie [2]		428,53								

Table 1
Thermodynamic properties of alcohols and ethers

(Table 2)

	basis	S(298.15K)	C(100K)	C(200K)	C(298.15K)	C(400K)	C(600K)	C(800K)	C(1000K)	C(1500K)
ethanethiol										
ref. values [1]		296,02	53,33	62,93	74,37	87,26	112,27	133,05	147,53	168,32
HO	B2	284,63	41,94	56,21	70,94	87,09	114,13	134,29	149,59	173,75
1D-HR	B2	293,45	45,49	59,71	72,54	87,04	111,50	129,90	144,07	166,84
dimethyl sulfide										
ref. values [1]		285,85		62,10	75,22	88,51	112,45	132,23	147,24	169,79
HO	B2	284,56	46,11	58,45	72,18	87,78	114,06	133,69	148,71	172,86
1D-HR	B2	292,79	47,57	61,14	73,18	86,47	109,72	127,83	142,05	165,32
1-propanethiol										
ref. values [1]		336,50		76,98	96,86	116,84	152,34	180,97	202,02	232,71
HO	B2	317,39	52,60	71,60	91,63	114,38	152,83	181,28	202,64	236,04
1D-HR	B2	335,34	61,77	78,95	95,92	115,96	150,16	175,74	195,23	226,22
methyl ethyl sulfide										
ref. values [1]		333,15			95,04	116,64	152,18	179,70	200,51	
HO	B2	323,62	55,44	73,09	92,44	114,64	152,36	180,36	201,51	235,00
1D-HR	B2	341,42	57,27	76,15	93,91	113,79	147,70	173,26	192,90	224,51
1-butanethiol										
ref. values [1]		375,20		95,35	120,58	146,44	193,67	233,32	263,87	310,80
HO	B2	348,69	62,43	86,20	111,84	141,30	191,30	228,14	255,61	298,30
1D-HR	B2	373,17	76,53	100,35	120,79	145,62	188,94	221,56	246,35	285,59
methyl propyl sulfide										
ref. values [1]		371,68			117,29	145,28	191,66	227,95	255,74	
HO	B1	355,85	65,50	87,73	112,15	140,68	189,73	226,17	253,56	296,63
	B2	353,75	64,28	87,82	113,10	142,10	191,26	227,49	254,66	297,32
1D-HR	B1	375,44	73,05	96,19	117,21	142,03	185,33	218,12	243,21	283,34
	B2	374,91	69,79	94,35	117,76	143,57	187,07	219,56	244,36	284,02
2D-HR	B1	376,67	72,85	95,65	117,09	142,11	185,56	218,41	243,52	283,65
	B2	376,21	70,76	94,42	116,84	142,57	186,59	219,50	244,53	284,38
diethyl sulfide										
ref. values [1]		368,00		95,66	120,11	145,36	191,60	230,12	259,29	304,20
HO	B2	351,59	64,84	87,76	112,75	141,58	190,75	227,10	254,37	297,17
1D-HR	B2	378,55	67,57	90,86	114,49	141,21	185,95	218,97	244,00	283,87
2D-HR	B2	378,29	67,86	91,21	114,56	141,23	186,07	219,19	244,30	284,23

(Table 2 continued)

	basis	S(298.15K)	C(100K)	C(200K)	C(298.15K)	C(400K)	C(600K)	C(800K)	C(1000K)	C(1500K)
1-pentanethiol										
ref. values [1]		415,39			141,78	176,13	233,85	279,75	314,90	360,05
HO	B2	380,35	72,63	100,99	132,16	168,29	229,82	275,03	308,62	360,60
1D-HR	B2	417,48	91,15	122,22	146,06	175,56	227,93	267,55	297,62	345,05
methyl butyl sulfide										
ref. values [1]		411,90			140,67	175,04	232,69	278,55	314,08	
HO	B2	385,95	74,64	102,60	133,43	169,11	229,80	274,40	307,68	359,62
1D-HR	B2	414,98	84,75	116,19	142,99	173,42	225,94	265,47	295,57	343,46
Guthrie [2]		414,47								
ethyl propyl sulfide										
ref. values [1]		414,12			139,02	173,84	232,43	279,18	315,44	
HO	B2	387,43	73,50	102,37	133,34	168,97	229,60	274,19	307,48	359,47
1D-HR	B2	415,95	81,74	110,60	139,00	171,23	225,37	265,31	295,51	343,42
Guthrie [2]		405,76								
1-hexanethiol										
ref. values [1]		454,70		130,54	167,69	205,10	271,82	326,01	366,24	425,84
HO	B2	411,60	82,83	115,81	152,51	195,34	268,39	321,96	361,65	422,89
1D-HR	B2	457,18	105,62	143,46	170,88	205,22	266,77	313,45	348,81	404,46
2D-HR	B2	460,64	113,25	140,64	168,85	204,98	267,72	314,68	350,14	405,74
Guthrie [2]		457,73								
methyl pentyl sulfide										
ref. values [1]		451,18			163,50	204,14	271,71	324,83	365,62	
HO	B2	418,80	85,19	117,53	153,85	196,20	268,39	321,35	360,73	421,93
1D-HR	B2	450,18	99,44	137,92	168,29	203,47	265,05	311,56	346,91	402,96
Guthrie [2]		453,67								
ethyl butyl sulfide										
ref. values [1]		453,39			161,85	202,96	271,46	325,47	367,00	
HO	B2	418,56	83,73	117,12	153,62	195,93	268,09	321,07	360,48	421,76
1D-HR	B2	453,89	96,29	132,22	164,11	201,03	264,23	311,22	346,72	402,86
dipropyl sulfide										
ref. values [1]		448,80			161,10	202,29	272,16	328,47	372,46	
HO	B1	417,17	85,39	117,36	152,75	194,46	266,36	319,49	359,12	420,86
	B2	412,56	82,17	117,07	154,06	196,50	268,54	321,36	360,65	421,80
1D-HR	B1	444,99	99,21	130,68	161,25	198,17	261,90	309,36	345,21	401,92
	B2	451,89	96,47	129,34	162,75	200,89	264,80	311,79	347,19	403,12
2D-HR	B1	452,33	100,59	129,38	160,49	197,99	262,25	309,93	345,93	402,77
	B2	449,76	96,71	127,27	160,59	199,14	263,84	311,37	347,13	403,52
Guthrie [2]		450,91								

Table 2
Thermodynamic properties of thiols and sulfides

References

[1] Yaws, C. L. *Chemical Properties Handbook*, McGraw-Hill Handbooks, **1999**

[2] Guthrie, J.P. *J. Phys. Chem. A* **2001**, *105*, 8495-8499

[3] Chen, C.; Bozzelli, J.W. *J. Phys. Chem. A* **2003**, *107*, 4531-4546

Paper 9: "Vibrational Modes in partially optimized molecular systems"

An Ghysels, Dimitri Van Neck, Veronique Van Speybroeck, Toon Verstraelen, Michel Waroquier

Journal of Chemical Physics, **2007**, 126, 224102

THE JOURNAL OF CHEMICAL PHYSICS **126**, 224102 (2007)

Vibrational modes in partially optimized molecular systems

A. Ghysels, D. Van Neck, V. Van Speybroeck, T. Verstraelen, and M. Waroquier[a)]
Center for Molecular Modeling, Laboratory of Theoretical Physics, Ghent University, Proeftuinstraat 86, B-9000 Gent, Belgium

(Received 27 February 2007; accepted 12 April 2007; published online 12 June 2007)

In this paper the authors develop a method to accurately calculate localized vibrational modes for partially optimized molecular structures or for structures containing link atoms. The method avoids artificially introduced imaginary frequencies and keeps track of the invariance under global translations and rotations. Only a subblock of the Hessian matrix has to be constructed and diagonalized, leading to a serious reduction of the computational time for the frequency analysis. The mobile block Hessian approach (MBH) proposed in this work can be regarded as an extension of the partial Hessian vibrational analysis approach proposed by Head [Int. J. Quantum Chem. **65**, 827 (1997)]. Instead of giving the nonoptimized region of the system an infinite mass, it is allowed to move as a rigid body with respect to the optimized region of the system. The MBH approach is then extended to the case where several parts of the molecule can move as independent multiple rigid blocks in combination with single atoms. The merits of both models are extensively tested on ethanol and di-*n*-octyl-ether. © *2007 American Institute of Physics*.
[DOI: 10.1063/1.2737444]

I. INTRODUCTION

The applications of molecular modeling nowadays focus more and more on extended systems, in which numerous atoms are involved. Examples are polymer chains, supramolecular assemblies, systems embedded in a solvent, or (macro)molecules adsorbed within porous materials. For the description of the electronic part of the system, one usually resorts to a hybrid model, in which the chemically active part where bonds may be formed or broken is described at a quantum mechanical level, whereas the outer region is described at a lower molecular mechanics level. The previous models are often referred to as quantum mechanical/molecular mechanical methods.[1–4] The basic idea originated from the observation that normally only a portion of the atoms in a reaction is directly involved in bond breaking and forming or in changing of bond order. These methods can be used either in a cluster or a periodic based approach.

In many cases, only part of the system is optimized to restrict the computational cost or to prevent unphysical deformations of the border of periodic systems. Good examples are reactions occurring in porous materials, where the border of the system is kept fixed during the geometry optimization to prevent unphysical deformations due to the neglect of the whole periodic structure. Another example is the simulation of defects embedded in a crystal lattice. In such cases, it is common practice to cut out a cluster around the defect and to keep the cluster border atoms fixed during the geometry optimization.[5] After such a constrained geometry optimization, residual forces remain on the fixed border atoms and the partially optimized structure corresponds to a nonequilibrium state. The usual normal mode analysis (NMA) equations can be applied to this nonequilibrium configuration using the full Hessian of all atoms in the structure, i.e., the matrix of second derivatives of the potential energy surface with respect to all nuclear coordinates. Such a procedure, however, has some serious defects. The Hessian will have only three eigenvalues equal to zero instead of six, implying that the rotational invariance of the potential surface is not manifest. In addition, spurious imaginary normal frequencies appear, suggesting that the partially optimized system resides in a transition state, even when this is obviously not the case.

Whereas for the determination of the energy the partitioning into chemically active and passive areas is common practice, it is far less applied for the determination of the Hessian of such extended systems. The calculation of the Hessian is one of the most expensive steps in the calculation of free energies, so a partitioning scheme for the Hessian would also seriously reduce the computational cost.

Within this respect, the work of Head and co-workers is especially interesting.[6,7] They introduced a strategy by which only the frequencies of part of a chemical system are computed. In 2002, Li and Jensen introduced the name partial Hessian vibrational analysis (PHVA) and extended the method for the calculation of vibrational enthalpy and entropy changes for chemical reactions.[8] Recently, Besley and Metcalf applied the partial Hessian approximation to calculate the amide I band of polypeptides and proteins.[9] Within this methodology, however, the normal modes are calculated for the system with the fixed atoms frozen at their reference positions as if they were given an infinite mass and only the relaxed atoms can participate in the small amplitude vibrations. Head and co-workers have also developed a more sophisticated partial Hessian method,[10–12] where frequencies are corrected in lowest order perturbation theory for the cou-

a)Author to whom correspondence should be addressed. Electronic mail: michel.waroquier@ugent.be

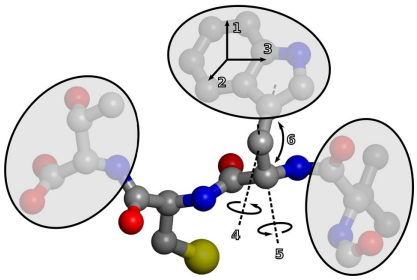


FIG. 1. Schematic representation of the basic idea behind the MBH method. The shaded blocks symbolize the parts of the molecule of which the internal geometry is kept fixed during the partial geometry optimization. In the MBH approach, they are described as rigid bodies with six degrees of freedom (translations and rotations).

pling between PHVA modes and modes in the fixed part. This method has been very useful for the calculation of localized vibrations of adsorbates on surfaces.

In this paper we propose an extended version of the PHVA method, in which the atoms that were kept fixed during the optimization can participate in small amplitude vibrations of the system, with the restriction that only coherent movements as a single block are allowed. The block of the fixed atoms can rotate and translate as a rigid body, while the internal geometry of the block is kept fixed. This model is referred hereafter to as the mobile block Hessian (MBH) approach. A schematic representation of the basic idea behind the MBH method is given in Fig. 1.

It can be expected that both methods would give similar results in those cases where the nonoptimized part of the system is quite rigid. The example of a defect in a crystal lattice is a case in point. For systems in which the molecular environment is more flexible, such as reactions occurring in solvents or polymer chains, the motions of the surroundings cannot *a priori* be neglected. It needs to be investigated to what extent the two methods coincide for such applications and whether such a partial optimization can give an accurate description of the localized modes in the optimized region. Localized modes are characterized by the fact that during the small amplitude vibrations, changes in the geometry only occur in a restricted region of the molecular system. In order to investigate this question, one should be able to construct the normal modes and frequencies of the partially optimized system in a rigorous way and compare them with frequencies of the fully optimized system.

The computational cost of the PHVA and MBH methods is similar, and the NMA equations can be rephrased in terms of a Hessian of reduced dimension in both cases. This leads to a significantly reduced load compared to a full Hessian frequency calculation. Related with a less expensive treatment of the Hessian is the work of Lin *et al.*[13] They have proposed efficient methods for calculating the Hessian in the optimization procedure of the multiconfigurational molecular mechanics method. The partial Hessian vibrational analysis and also the mobile block Hessian approach could be useful in combination with such methods as they could additionally speed up the computationally expensive task of determining the Hessian for reactions on extended systems.

In Sec. II the problem of determining normal modes in nonequilibrium systems is discussed. The partial Hessian vibrational analysis is revised, and a detailed theoretical derivation is given of the mobile block Hessian approach, which we propose as an extension of the PHVA method for extended systems which are quite flexible. In Sec. III the two methods are compared for the example of the ethanol molecule. This very simple molecule has been chosen as a test case, as it enables one to address various phenomenological issues of the PHVA and MBH methods. The MBH approach can be easily extended to the case where several parts of the molecule can move as independent rigid bodies in combination with single atoms. The framework of this "multiple" mobile block Hessian approach is worked out in Sec. IV and is applied to di-*n*-octyl-ether in Sec. V. Finally, the results are summarized in Sec. VI, and future applications of the MBH model are discussed.

II. BACKGROUND AND THEORETICAL DEVELOPMENT

A. Normal modes in nonequilibrium configurations

Consider a molecule with N masses m_A $(A=1,\ldots,N)$, the positions of which are described by Cartesian coordinates $\mathbf{r}_A \equiv \{r_{A\mu}\}_{\mu=x,y,z}$, with respect to a space-fixed frame. The energy of the system reads

$$E = \frac{1}{2}\sum_{A\mu} m_A \dot{r}_{A\mu}^2 + V(\{\mathbf{r}_A\}), \tag{1}$$

where $\dot{r}_{A\mu}$ is a time derivative and V is the potential energy. Expanding V around a reference configuration $\{r_{A\mu}^0\}$ gives

$$V(\{\mathbf{r}_A\}) = V_0 + \sum_{A\mu}\left(\frac{\partial V}{\partial r_{A\mu}}\right)_0 \Delta_{A\mu} + \frac{1}{2}\sum_{A\mu;B\nu}\left(\frac{\partial^2 V}{\partial r_{A\mu}\partial r_{B\nu}}\right)_0 \Delta_{A\mu}\Delta_{B\nu} + \ldots \tag{2}$$

in terms of the displacement coordinates $\Delta_{A\mu}=r_{A\mu}-r_{A\mu}^0$. By collecting the coordinates in one vector $x_i=r_{A\mu}$, with $i \equiv A\mu=1,\ldots,3N$, one can expand the energy in matrix form as

$$E - V_0 = \frac{1}{2}\dot{\Delta}^T M\dot{\Delta} + \Delta^T G + \frac{1}{2}\Delta^T H\Delta \tag{3}$$

up to second order in the displacements $\Delta_i=x_i-x_i^0$. In Eq. (3), the mass matrix M is a diagonal matrix containing the masses, G is the gradient vector defined by the gradients of the potential energy at the reference point, and H is the Hessian or second-derivative matrix in the reference point. The NMA equation determining the eigenmodes v and normal frequencies $\lambda^{1/2}$ reads

$$Hv = \lambda M v, \tag{4}$$

representing a generalized eigenvalue problem.

Solution schemes of these equations are implemented in various standard *ab initio* molecular modeling packages, but if the whole molecular system is not in its equilibrium state, the reference configuration is not gradient-free and some of the resulting frequencies are completely unphysical. An additional problem is that a discrimination of the unphysical frequencies from the physical values is not obvious.

We want to make two comments in the situation of nonequilibrium configurations (gradient vector $G \neq 0$). First, one can show that the eigenvalues of Eq. (4) are coordinate-dependent: a second-order expansion of the potential energy expressed in curvilinear coordinates, which are nonlinearly related to the Cartesian coordinates, leads to different normal mode frequencies. This phenomenon is "well known" and is related to the difference between ordinary and covariant derivatives in a nontrivial metric space.[14] Its importance and unpleasant consequences have recently been emphasized by several authors.[15,16]

A second comment deals with the invariance of the potential energy surface under overall rotations. When $G \neq 0$, the Cartesian Hessian will only generate three zero eigenvalues (those related to the global translation). The three eigenvalues associated with the global rotation are different from zero,[17] and the absence of these Goldstone modes[18] is due to the use of the second-order expansion in rectilinear Cartesian coordinates, which breaks the rotational symmetry of V. This defect is simply cured by taking coordinates that respect the symmetries of V, i.e., internal coordinates.

With any choice of $3N-6$ internal coordinates $\{\theta_I\}$, and a body frame whose origin lies at the center of mass $\mathbf{r}_{\mathrm{c.m.}}$, the energy in Eq. (1) can be rewritten in a standard way[19–21] as the sum of the potential energy $W(\{\theta_I\})$ expressed in internal coordinates and kinetic energy terms arising from the center-of-mass motion, global rotation, Coriolis coupling, and internal motions:

$$E = \frac{1}{2}\mathcal{M}\dot{\mathbf{r}}_{\mathrm{c.m.}}^2 + \frac{1}{2}\boldsymbol{\omega}\bar{\bar{\mathbf{I}}}\boldsymbol{\omega} + \boldsymbol{\omega}\sum_I \mathbf{A}_I\dot{\theta}_I + \frac{1}{2}\sum_{IJ} B_{IJ}\dot{\theta}_I\dot{\theta}_J + W(\{\theta_I\}). \tag{5}$$

Here $\mathcal{M}$ is the total mass and ω the angular velocity vector of the body frame. The inertial tensor $\bar{\bar{\mathbf{I}}}$, the Coriolis coupling $\mathbf{A}_I$ between the body-frame rotation and internal velocity $\dot{\theta}_I$, and the B_{IJ} matrix are all functions of the internal coordinates.

The potential energy $W(\{\theta_I\})$, expanded to second order about a reference geometry $\{\theta_I^0\}$, reads

$$W(\{\theta_I\}) = V_0 + \sum_I \left(\frac{\partial W}{\partial \theta_I}\right)_0 \Delta_I + \frac{1}{2}\sum_{IJ}\left(\frac{\partial^2 W}{\partial\theta_I\partial\theta_J}\right)_0 \Delta_I\Delta_J. \tag{6}$$

The set of new coordinates now consists of the center-of-mass coordinates $\mathbf{r}_{\mathrm{c.m.}}$, three angles to specify the orientation of the body frame (e.g., Euler angles), and the $3N-6$ internal coordinates θ_I. The NMA equations using internal coordinates show a block structure:

$$\begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & H^{(ii)} \end{pmatrix}\begin{pmatrix} v^{(c)} \\ v^{(r)} \\ v^{(i)} \end{pmatrix} = \lambda \begin{pmatrix} M^{(cc)} & 0 & 0 \\ 0 & M^{(rr)} & M^{(ri)} \\ 0 & (M^{(ri)})^T & M^{(ii)} \end{pmatrix}\begin{pmatrix} v^{(c)} \\ v^{(r)} \\ v^{(i)} \end{pmatrix}, \tag{7}$$

where $v^{(c)}$, $v^{(r)}$, and $v^{(i)}$ have dimensions 3, 3, and $3N-6$, respectively, and the mass matrix entries read $M_{\mu\nu}^{(cc)} = \mathcal{M}\delta_{\mu\nu}$, $M_{\mu\nu}^{(rr)} = I_{\mu\nu}(\{\theta_I^0\})$, $M_{\mu I}^{(ri)} = A_{\mu I}(\{\theta_I^0\})$, and $M_{IJ}^{(ii)} = B_{IJ}(\{\theta_I^0\})$. If the gradient is nonzero, the NMA equations in Eq. (7) will give different eigenvalues from those of Eq. (4). In particular, the eigenvalue equations (7) always generate six zero eigenvalues due to its construction, as the Hessian has vanishing matrix elements in the (c) and (r) subspaces, while in Eq. (4) the presence of six zero eigenvalues is not ensured.

In a completely general situation it is, in principle, not possible, due to the coordinate dependence, to define meaningful normal modes in a nonequilibrium point. However, the situations of interest for the present paper are not completely general but arise from physical considerations.

We consider cases where the reference point is obtained by optimizing the energy with respect to a *subset* of coordinates, keeping the remainder fixed during the optimization. The fixed coordinates correspond, e.g., to the geometry of a part of the molecule that is expected to influence the localized mode of interest only slightly. The system corresponding to the subset of coordinates that have been optimized is actually in equilibrium, even though the complete gradient is nonzero.

A detailed theoretical analysis of the coordinate (in)dependence and symmetry properties of the NMA equations in nonequilibrium systems, as well as the link between descriptions using Cartesian and internal coordinates, will be given in a separate publication. In this paper we focus on the comparison of two practical methods for treating such cases.

B. Partial Hessian vibrational analysis

One assumes a certain reference structure $\{\mathbf{r}_{A_F}^0\}$ for part of the molecule consisting of N_F atoms (labeled $A_F, B_F, \ldots$). Keeping $\{\mathbf{r}_{A_F}^0\}$ fixed in space, the positions of the remaining $N_E = N - N_F$ atoms (labeled $A_E, B_E, \ldots$) are optimized, resulting in an "equilibrium" configuration $\{\mathbf{r}_{A_E}^0\}$. Obviously, the full gradient is nonzero, since $(\partial V/\partial r_{A_E\mu})_0 = 0$, but $(\partial V/\partial r_{A_F\mu})_0 \neq 0$. Solutions of the NMA equations taking into account the full Hessian and mass matrix will therefore lead to unphysical results.

In a first model, known as partial Hessian vibrational analysis, one assumes that the fixed atoms A_F do not participate in the small amplitude vibrations. Hence their displacements $\Delta_{A_F\mu}$ and velocities $\dot{\Delta}_{A_F\mu}$ are set to zero. This is consistent with a situation in which infinite masses m_{A_F} are associated with the fixed atoms. With this assumption, the second-order energy of Eq. (3) reduces to

$$E - V_0 = \frac{1}{2}\sum_{A_E\mu} m_{A_E}\dot{\Delta}^2_{A_E\mu} + \frac{1}{2}\sum_{A_E\mu;B_E\nu}\left(\frac{\partial^2 V}{\partial\Delta_{A_E\mu}\partial\Delta_{B_E\nu}}\right)_0 \Delta_{A_E\mu}\Delta_{B_E\nu} + \cdots = \frac{1}{2}\dot{\Delta}'^T M'\dot{\Delta}' + \frac{1}{2}\Delta'^T H'\Delta', \quad (8)$$

where Δ', $\dot{\Delta}'$, M', and H' are the displacements, velocities, mass matrix, and Hessian restricted to the reduced $3N_E$-dimensional subsystem of the nonfixed atoms with mass m_{A_E}. Note that the reduced gradient ($G'=0$) vanishes in this subspace, so the $3N_E$ system is in equilibrium at the reference configuration and provides coordinate-independent normal modes.

In practice, one simply needs to disregard the rows and columns related to the coordinates of the fixed atoms in the full (Cartesian) Hessian and mass matrix. The NMA equation reduces to a $3N_E \times 3N_E$ generalized eigenvalue problem:

$$\sum_{B_E\nu}\left(\frac{\partial^2 V}{\partial r_{A_E\mu}\partial r_{B_E\nu}}\right)_0 v_{B_E\nu} = \lambda m_{A_E} v_{A_E\mu}. \quad (9)$$

The reduced mass matrix M' is positive definite, and the presence of negative solutions λ in Eq. (9) gives indication that the reduced system resides in a transition state rather than a minimum. The number of zero eigenvalues depends on the remaining symmetry of the reduced system. Usually no zeros will occur, since the fixed atoms act as an external field breaking translational and rotational invariance for the nonfixed atoms.

C. The mobile block Hessian approach

In the MBH approach, the fixed atoms are allowed to participate in the small amplitude vibrations by moving as a rigid body (block). This is a physically different situation from the previous approach, since one now takes into account the finite masses of the fixed atoms $\{m_{A_F}\}$.

This approach is easily implemented by a suitable choice of the $3N-6$ internal coordinates $\{\theta_I\}$, which is always possible. In the Z-matrix formalism, for instance, one can determine the first N_F atoms by $3N_F-6$ Z coordinates (distances, angles, and dihedral angles), and the next N_E atoms can consecutively be described by $3N_E$ Z coordinates.

The $3N_F-6$ internal coordinates, describing the geometry of the fixed atoms A_F, are labeled $\{\theta_{I_F}\}$, and the remaining $3N_E$ internal coordinates are labeled $\{\theta_{I_E}\}$. The imposed reference structure $\{\mathbf{r}^0_{A_F}\}$ of the fixed atoms then determines the values of the $3N_F-6$ internal coordinates, $\{\theta^0_{I_F}\}$. Optimizing the energy with fixed $\{\theta^0_{I_F}\}$ and varying $\{\theta_{I_E}\}$ give the reference structure $\{\theta^0_{I_E}\}$. Again, the full gradient is nonzero, since $(\partial W/\partial\theta_{I_E})_0=0$ but $(\partial W/\partial\theta_{I_F})_0\neq 0$. As a result, the NMA equation (7) with the full Hessian will not give correct frequencies (though one does have six zero eigenvalues).

Allowing the fixed block to move as a rigid body, keeping its internal geometry, can be imposed by putting the displacements Δ_{I_F} and corresponding velocities $\dot{\Delta}_{I_F}$ equal to zero in the second-order expansion of Eq. (5). The part of the molecule described by the $3N_E$ internal coordinates $\{\theta_{I_E}\}$ has been relaxed during the geometry optimization and hence is in equilibrium. The gradient term in Eq. (5) thereby vanishes.

The corresponding NMA equations are obtained by omitting the rows and columns related to the $\{\theta_{I_E}\}$ variables from the Hessian and mass matrix in Eq. (7). The resulting reduced eigenvalue problem of dimension $3N_E+6$ still has six zero eigenvalues corresponding to overall translation and rotation. These can be decoupled in the usual way[19–21] by a congruent transformation, $\tilde{v}^{(c)}=v^{(c)}$, $\tilde{v}^{(i)}=v^{(i)}$, and $\tilde{v}^{(r)}=v^{(r)}+(M^{(rr)})^{-1}M'^{(ri)}v^{(i)}$, yielding the final $3N_E$-dimensional NMA equation,

$$H'^{(ii)}\tilde{v}^{(i)} = \lambda[M'^{(ii)} - (M'^{(ri)})^T(M^{(rr)})^{-1}M'^{(ri)}]\tilde{v}^{(i)}. \quad (10)$$

The primed matrices do not contain the components related to $\{\theta_{I_F}\}$. The transformed mass matrix in the right hand side of Eq. (10) is positive definite and the presence of negative eigenvalues λ will now unambiguously indicate that the reduced system is in a transition state.

D. Discussion: PHVA versus MBH

Both PHVA and MBH models offer a vibrational analysis in which only a subblock of the Hessian matrix is diagonalized to produce vibrational frequencies for partially optimized systems. In both descriptions, the molecular system is composed of a rigid body with a number N_F of fixed atoms and N_E atoms that are free to relax in a partial geometry optimization. Conceptually, the difference between PHVA and MBH lies in the use of near-infinite masses in the former approach for the atoms in the rigid body, while in the MBH model, reduced masses are used.

A more transparent and fundamental comparative study between the PHVA and MBH models can be made when using a specific choice of internal coordinates: for the set of $3N_E$ internal coordinates $\{\theta_{I_E}\}$ describing the geometry of the nonfixed atoms, one can take the Cartesian coordinates $\{\mathbf{r}^{\mathrm{RB}}_{A_E}\}$ in a frame attached to the rigid block (RB) (e.g., the frame constructed by the principal axes of the fixed atoms). The potential energy then becomes

$$W(\{\theta_{I_F},\theta_{I_E}\}) \equiv V(\{\mathbf{r}^{\mathrm{RB}}_{A_F},\mathbf{r}^{\mathrm{RB}}_{A_E}\}). \quad (11)$$

Since the $\{\mathbf{r}^{\mathrm{RB}}_{A_F}\}$ are constant in the fixed geometry of the block, one has

$$\left(\frac{\partial^2 W}{\partial\theta_{I_E}\partial\theta_{I_E}}\right)_0 \equiv \left(\frac{\partial^2 V}{\partial r_{A_E\mu}\partial r_{A_E\mu}}\right)_0. \quad (12)$$

Here it is assumed that at the reference point the frame attached to the rigid body coincides with the space-fixed frame, which can always be done. The reduced Hessian $H'^{(ii)}$ of MBH now exactly coincides with the Cartesian H' of the PHVA. This gives evidence that the discrepancy between PHVA and MBH only lies in a different approach of handling the mass matrix.

The difference between both approaches is best illustrated by the example of two masses m_1 and m_2 moving in one dimension and joined by a spring characterized by a

spring constant K. Following the method of Sec. II B (PHVA), the mass m_1 is considered fixed in space and the normal frequency of the system (vibration of mass m_2) is given by $\omega^2=K/m_2$. In the method of Sec. II C (MBH), on the other hand, the mass m_1 is allowed to participate in the vibration and one simply has the normal frequency of the free spring, $\omega^2=K/\mu$, with μ the reduced mass. In the limit of $m_1 \gg m_2$, the reduced mass tends to m_2 and the eigenvalues of the two approaches converge to the same value.

Generally, the partial Hessian vibrational analysis model can be considered as a limit case of the mobile block Hessian approach. When the mass and the inertial moments of the rigid body increase with respect to the N_E free masses, the mass matrix of the MBH in the limit becomes equal to the mass matrix of the PHVA. Thus the higher the fixed atom masses, the more the two models converge to each other.

A more stringent comparison is obtained by inspecting the mass term taken up in Eq. (10). Following the various mass definitions given in Sec. II A, one sees that the fixed atom masses m_{A_F} only occur in the matrix $M^{(rr)}$, which coincides with the inertial matrix, while $M'^{(ri)}$ is independent of the m_{A_F}. The larger the number of atoms in the fixed part of the molecule, the more the total mass of the rigid body increases. As mentioned above, only the elements of the matrix $M^{(rr)}$ are influenced, and the second term in the right hand side of Eq. (10) will tend to zero. With the specific choice of internal coordinates as stated in Eq. (11), it is clear that the PHVA frequencies form the limits of the MBH frequencies.

Another (non-negligible) aspect concerns the computational cost of both models. Since both models require the numerical evaluation of the reduced Hessian H' or $H'^{(ii)}$, no difference is expected in computing time. The calculation of the corresponding mass matrices is also straightforward. Since most of the required computation time goes to the evaluation of the second derivatives of the potential energy surface, the reduction of the dimension of the full Hessian to $3N_E$ is essential to get a serious reduction of the computation time.

The applicability of both methods depends on the type of the simulated system under consideration. It can be expected that the MBH method is better suited to describe molecules in the gas phase, because they move quasifreely in space and can be considered as isolated systems. Assigning an infinite mass to the nonoptimized part of the system may be too crude an approximation. The potential energy surface should be invariant under global translation and rotation, and these global modes should effectively decouple from the internal vibrations.

However, when modeling a lattice using the cluster *in vacuo* approach and fixing the border atoms, the situation is different. The cluster cannot move freely and there is no reason why the description of the cluster should be translationally or rotationally invariant. In fact, the border atoms are more or less pinned down at their positions by the presence of the surrounding infinite lattice, which was, however, left out of the simulation cluster. So it is quite a realistic picture to assume that the border atoms are really fixed by external forces generated by the surrounding lattice. In other words, the rigid body of fixed border atoms may be assumed to have a fixed position and orientation, and thus, the PHVA approach is the appropriate way to calculate eigenvalues.

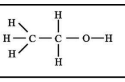
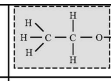
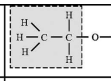
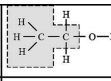
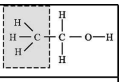
In what follows, both the PHVA and MBH methods are numerically validated and benchmarked against the normal modes extracted from a fully optimized geometry. Independent of the specific application under consideration, some general remarks are useful. Inherent to the PHVA and MBH methods is the introduction of a rigid part of the system that is not optimized, and thus, the number of PHVA and MBH frequencies is always smaller than the number of benchmark frequencies. The benchmark modes generated by atoms belonging to the rigid body are *a priori* not reproduced. Additionally, modes where the displacements are spread out over nonfixed as well as fixed atoms are expected to be very badly described by both the PHVA and MBH. On the other hand, normal modes that are completely localized in the relaxed molecular region with nearly vanishing fixed atom displacements, or normal modes where the fixed atoms move collectively with respect to the optimized region, can be expected to result from the MBH approach.

Note that Head and co-worker investigated the coupling between the PHVA modes and the omitted modes occurring in the fixed atom region.[10–12] It may be possible to apply similar methods to the MBH modes.

III. APPLICATION TO THE ETHANOL MOLECULE

The PHVA and MBH methods are, in principle, developed to evaluate the frequencies in a large molecular system which cannot be optimized entirely at a high level of theory as the size increases, but in which only part (the active site) of the molecule is optimized at the high level and the remaining part at a substantially lower level of theory. However, in order to understand the advantages and disadvantages of both methods, it is instructive to validate them first on a small molecule, where the exact frequencies are readily available and can be compared with those predicted by PHVA and MBH in partially optimized geometries. To meet this purpose, we have chosen ethanol containing a well localized O–H stretch. The entire molecule has been optimized at a so-called high level [B3LYP/6-31+$g(d)$], and this optimized geometry will be used further on as the reference. Frequencies of all present normal modes are obtained by solving the mass-weighted eigenvalue problem [Eq. (4)] of the full 27×27 Hessian, and are tabulated in the first column of Table I. They coincide with the values that are obtained from the standard analytical frequency calculation in GAUSSIAN03,[22] and will serve as benchmark values for further comparative studies. An exact treatment should generate six eigenvalues exactly equal to zero, corresponding to the global translation and rotation. In practice, the values differ slightly from zero (varying between -1 and 8 cm^{-1}). Translational frequencies are sensitive to numerical errors in the construction of the Hessian. Rotational frequencies are, in addition, affected by the small residual forces due to the finite convergence criteria. The effect of the almost zero fre-

TABLE I. Normal mode frequencies (in cm^{-1}) of ethanol derived from the benchmark geometry, which corresponds to the geometry optimization obtained at B3LYP/6-31+$g(d)$. The rigid body is composed of the atoms in the shaded region. In the left column, translational and rotational frequencies from the full Hessian calculation are plotted before and after projection. Vibrational frequencies are not affected by this projection. The PHVA and MBH frequencies were ordered according to the maximum overlap with the benchmark modes.

									
Full	Projected	PHVA	MBH	PHVA	MBH	PHVA	MBH	PHVA	MBH
-1	0	-	-	-	-	72	-	59	-
-1	0	-	-	-	-	242	-	94	-
0	0	-	-	-	-	678	-	211	-
1	0	-	-	-	-	-	-	355	-
2	0	-	-	-	-	-	-	-	-
8	0	-	-	-	-	-	-	-	-
internal rotation:									
244		-	-	-	-	-	249	-	244
295		275	289	299	289	279	297	292	295
C-C-O stretch:									
417		-	-	-	491	-	458	-	419
mixed modes:									
825		-	-	272	807	823	992	1024	831
902		-	-	-	-	-	1011	-	906
1037		-	-	-	-	-	-	748	1040
1104		-	-	-	1031	-	-	1058	1111
1185		-	-	-	-	-	-	-	1195
1270		1164	1208	1224	1245	1210	1218	1226	1277
1305		-	-	715	-	1254	1269	1262	1308
1419		-	-	-	-	-	-	-	-
1461		-	-	-	-	1317	1378	1445	1452
1505		-	-	-	-	-	-	-	-
1521		-	-	-	-	-	-	-	-
C-H stretch:									
1544		-	-	-	-	1501	1534	1541	1541
2995		-	-	245	-	2928	2977	2995	2995
3021		-	-	-	-	2877	3015	3023	3023
3049		-	-	-	-	-	-	-	-
3116		-	-	-	-	-	-	-	-
3124		-	-	-	-	-	-	-	-
O-H stretch:									
3756		3645	3708	3755	3756	3755	3756	3756	3756

quencies on the other 21 frequencies is negligible here: projecting out the overall translation and rotation, which is implemented in most of the program packages such as GAUSSIAN03, gives six eigenmodes exactly equal to zero and does not affect the vibrational frequencies. Some of the modes have a clear interpretation: the two lowest frequencies (244 and 295 cm^{-1}) correspond to internal rotations of the methyl top and the hydroxyl group, while the highest (3756 cm^{-1}) is associated with the highly localized O–H stretching mode.

We applied the PHVA and MBH methods to two different cases: first to the fully optimized structure and then to several partially optimized structures. The PHVA equations (9) are constructed with the submatrix H' of the Hessian that contains the second derivatives of the potential energy with respect to the Cartesian coordinates of the free atoms. The MBH frequencies corresponding to Eq. (10) are calculated, following the discussion in Sec. II D, with the same submatrix H' [see Eq. (12)] but with a different mass matrix.

The ordering of the calculated frequencies in the tables is determined by the maximum of the square of the overlap $|\langle \nu_{\text{bench}}| \nu\rangle|^2$ between the benchmark mode with frequency ν_{bench} and the calculated mode with frequency ν. For modes without a pronounced maximum overlap, this is of course rather arbitrary.

A. PHVA and MBH applied to the equilibrium structure

Frequencies are calculated for the fully optimized geometry, while the part of the Hessian chosen to be included in the vibrational analysis is varied. In this case, there are no remaining residual forces and the difference between the two models (PHVA or MBH) can easily be studied. The results are given in Table I for various rigid body sizes. The shaded box indicates the part of the molecule that is not included in the calculation of the Hessian. For instance, the second column of the table reports the three frequencies corresponding to the modes generated by a rigid body and one single atom that can vibrate. The only nonfixed atom in the vibrational analysis is the hydrogen atom of the hydroxyl group. In the third column, the whole hydroxyl group can vibrate, while in the last column we display the situation with only the methyl group fixed at its reference geometry.

Only three PHVA and MBH modes are calculated in the case where only the H atom of the hydroxyl group is taken into account (second column of Table I). Visualization of the three modes revealed that they all qualitatively correspond to a normal mode of the benchmark frequency spectrum, but that their values are underestimated (benchmark values are 295, 1270, and 3756 cm^{-1}). In the PHVA method, the underestimation is even more pronounced due to the use of infinite masses. This is a general conclusion: the PHVA fails in accurately reproducing the benchmark values, especially in the small and medium frequency regions, while the quantitative agreement of the MBH predictions of localized modes—taking place in the nonfixed region—with the benchmark values is manifestly present. The agreement is even very close in the last case, where the rigid body consists of a fixed methyl group. The O–H stretch mode is always present in the frequency spectrum, only in the case of one relaxed atom (the hydrogen) is the obtained frequency slightly underestimated.

Looking at lower frequencies, the MBH approach gives consistently better results than the PHVA, because the reduced mass effect is taken into account. As the total mass of the fixed block decreases, it is obvious that the PHVA induces spurious low frequency modes of the order of 60–100 cm^{-1}, corresponding to translations/rotations of the nonfixed atoms in the field of their environment (the fixed atoms). Summarizing, when working with molecules in the gas phase, the use of the MBH model is highly recommended as soon as the total mass of the fixed block becomes of the same order as the total mass of the relaxed atoms.

B. PHVA and MBH applied to partially optimized structures

We introduce partially optimized geometries and verify whether the relevant calculated frequencies in the active site can be reproduced by both the PHVA and MBH models in an accurate way with respect to the benchmark frequencies. Initially, the ethanol molecule has been optimized at the lower HFO/STO-3g level. Then the system was partially optimized at the higher B3LYP/6-31+$g(d)$ level while keeping the atoms of the rigid block fixed at their initial HF/STO-3g positions. Results are collected in Table II. The atoms belonging to the shaded box were not optimized at the high level but were kept fixed at their low level geometries. For each case, the frequencies resulting from a full Hessian [calculated at B3LYP/6-31+$g(d)$] diagonalization, i.e., from a standard normal mode analysis in Gaussian, as well as the normal modes resulting from the PHVA and MBH methods are tabulated. The standard frequency analysis always gives a number of spurious imaginary frequencies, as could be expected since residual forces on the nonoptimized atoms disturb the evaluation of the frequencies. In addition, the positive frequencies deviate substantially from the benchmark values given in Table I. This means that the normal frequencies generated by standard procedures in program packages such as GAUSSIAN03,[22] in molecules whose atomic positions have not been optimized at the same level of theory, are far from being accurate.

By applying the PHVA or the MBH model, the unphysical imaginary frequencies disappear, but the resulting frequencies, however, differ significantly between the two methods. The MBH results converge rapidly to the benchmark values, highlighting the efficiency of the proposed MBH model. A striking resemblance is even observed in the last column where the fixed body is restricted to the ending methyl group. The low frequency spectrum of the PHVA method, however, deviates largely from the benchmark values. This is entirely due to the reduced mass effect, inducing spurious unphysical modes.

While the full Hessian frequencies are very sensitive to the exact molecular geometry, it is remarkable that PHVA and MBH frequencies from the partially optimized geometry are very close to those obtained with the benchmark geometry with the same block size. This indicates that the PHVA and MBH models are less sensitive to the exact internal geometry of the fixed block. This could be important for future applications in large systems.

It is also interesting to compare the thermodynamical quantities associated with the vibrational part of the partition function of the system. In Fig. 2 the vibrational contribution to the entropy S and the free enthalpy G are given for the different partially optimized ethanol configurations at $T=298.15$ K. The values calculated with the benchmark frequencies are also indicated. It is clear that reducing the num-

TABLE II. Normal mode frequencies (in cm^{-1}) of ethanol derived on the basis of partially optimized geometries at the B3LYP/6-31+$g(d)$ level of theory. The rigid body is composed of atoms in the shaded region and its geometry is originated from a geometry optimization of the whole molecule at the low HF/STO-3g level. Benchmark frequencies are given in the left column for comparison. The PHVA and MBH frequencies were ordered according to the maximum overlap with the benchmark modes.

												
bench	Full	PHVA	MBH	Full	PHVA	MBH	Full	PHVA	MBH	Full	PHVA	MBH
-1	-258	-	-	-264	-	-	-254	71	-	-251	60	-
-1	-130	-	-	-132	-	-	-80	243	-	-90	95	-
0	-62	-	-	-75	-	-	-61	678	-	-66	212	-
1	-1	-	-	-2	-	-	-1	-	-	-1	355	-
2	-1	-	-	0	-	-	1	-	-	1	-	-
8	1	-	-	1	-	-	1	-	-	1	-	-
244	91	-	-	42	245	-	124	-	248	124	-	246
295	294	281	295	284	296	285	289	280	298	289	294	297
417	403	-	-	396	-	490	397	-	458	397	-	419
825	760	-	-	758	272	813	793	824	992	793	746	839
902	873	-	-	877	-	-	880	-	-	883	-	909
1037	1021	-	-	1025	-	1032	1025	-	1013	1029	1025	1041
1104	1070	-	-	1096	-	-	1097	-	-	1098	1058	1116
1185	1147	-	-	1156	-	-	1177	-	-	1177	-	1201
1270	1257	1161	1205	1261	1224	1244	1266	1210	1218	1267	1225	1280
1305	1267	-	-	1277	715	-	1298	1252	1268	1300	1262	1313
1419	1384	-	-	1385	-	-	1388	-	-	1390	-	-
1461	1422	-	-	1429	-	-	1452	1315	1376	1454	-	-
1505	1489	-	-	1488	-	-	1488	-	-	1488	1444	1452
1521	1505	-	-	1506	-	-	1507	-	-	1508	-	-
1544	1527	-	-	1527	-	-	1544	1502	1535	1544	1542	1542
2995	3140	-	-	3137	-	-	2995	2877	2978	2994	2995	2995
3021	3162	-	-	3162	-	-	3023	2929	3016	3022	3023	3023
3049	3177	-	-	3173	-	-	3162	-	-	3162	-	-
3116	3242	-	-	3243	-	-	3243	-	-	3242	-	-
3124	3253	-	-	3252	-	-	3251	-	-	3251	-	-
3756	3751	3640	3703	3755	3755	3755	3756	3756	3756	3756	3756	3756

ber of modes by introducing the rigid block affects both the vibrational entropy and free enthalpy. However, the MBH values tend consistently to the benchmark value when the block size decreases, whereas the PHVA values do not. This is to be expected, since PHVA assumes infinite mass for the fixed block and for the present application the MBH is physically more relevant.

IV. EXTENSION: MULTIPLE MOBILE BLOCKS

The application field of MBH can be easily extended to multiple mobile blocks. Suppose the molecule is decomposed into K rigid blocks and N_E freely relaxed atoms. The whole molecule is optimized at a low level of theory determining the position of each atom in each of the rigid blocks. In a second step, a higher level of theory is used and one optimizes the positions of the N_E free atoms, as well as the positions and orientations of the rigid blocks, keeping their original internal geometry. As a consequence, residual forces remain between the fixed atoms within a block, thus the full gradient is nonzero, and a normal frequency analysis along the lines of Eq. (7) with the full Hessian will therefore produce unphysical frequencies.

In the extended method, one assumes that each block is allowed to participate in the small amplitude vibrations, moving as a rigid body. To implement this approach, a suitable choice of internal coordinates is necessary, e.g., the Z-matrix formalism. It is, in fact, sufficient to choose the numbering of the atoms in the Z-matrix construction such that the atoms in each fixed block are numbered consecutively.

As in Sec. II C, the geometry of each block b is described by $3N_{F,b}-6$ internal coordinates $\{\theta_{I_{F,b}}\}$. The imposed reference structure $\{\mathbf{r}^0_{A_{F,b}}\}$ of the fixed atoms of block b then determines the values of the $3N_{F,b}-6$ internal coordinates $\{\theta^0_{I_{F,b}}\}$. The remaining $3N_E+6(K-1)$ internal coordinates will

be labeled $\{\theta_{I_E}\}$. Optimizing the energy with fixed $\{\theta^0_{I_{F,b}}\}$ and varying $\{\theta_{I_E}\}$ yield the reference structure $\{\theta^0_{I_E}\}$.

To impose the fixed internal geometry of each block during the vibrational analysis, the displacements $\Delta_{I_{F,b}}$ and corresponding velocities $\dot{\Delta}_{I_{F,b}}$ are set equal to zero in the second-order expansion of Eq. (6). The gradient term in Eq. (6) thereby vanishes, so the reduced system of $3N_E+6(K-1)$ internal coordinates $\{\theta_{I_E}\}$ is in equilibrium.

The corresponding NMA equations are obtained by omitting the rows and columns related to the $\{\theta^0_{I_{F,b}}\}$ variables from the Hessian and mass matrix in Eq. (7). The resulting reduced eigenvalue problem of dimension $3N_E+6K$ has six zero eigenvalues corresponding to overall translation and rotation. These can be decoupled as in the mobile block Hessian approach with a congruent transformation similar to Eq. (10).

The discussion of this multiple MBH approach is similar to the single MBH with only one rigid body. First, the disturbing negative eigenvalues due to residual forces are eliminated, and imaginary frequencies will only occur if the reduced system of relaxed atoms and K blocks is actually in a transition state. The presence of six zeros shows that the translational and rotational invariance are respected in the second-order expansion of the potential energy of the reduced system. Concerning the mass effect, the finite mass of each block is taken into account. Moreover, the method provides a reduction in computer time, because for each block b with $N_{F,b}$ atoms, only derivatives with respect to six coordinates instead of $3N_{F,b}$ have to be calculated.

Of course, the main question is whether the multiple MBH approach is capable of reproducing the normal mode frequencies of the fully optimized benchmark structure. Generally, two types of benchmark modes are well reproduced by the multiple MBH in analogy with the single MBH version: (i) modes localized in the relaxed part of the molecule (ii) and modes where the fixed atoms move as a whole. Some other modes present in the benchmark, on the other hand, are obviously not reproduced by the multiple MBH. These are modes in which the internal geometry of the rigid blocks is changed. Such modes are not localized in the chemically active part of the molecule and are therefore not relevant for the present study. Modes where the displacements are spread out over the relaxed as well as fixed atoms are mostly not reproduced in the multiple MBH model, except those where the atoms of the rigid bodies move coherently [modes of type (ii)]. Some examples of such modes are shown in Fig. 1. For instance, the rigid bodies can rotate about the single bonding axis connecting them with the active part of the molecule or they can contribute to a stretching, a bending motion, etc.

For practical purposes, it is important to select the freely relaxed region of the molecule attentively. One of the main advantages of this extension to multiple MBH is the computational profit as a consequence of the reduced number of Hessian matrix elements to be evaluated, but unfortunately this option is not (yet) present in most of the standard program packages.

V. APPLICATION TO DI-*N*-OCTYL-ETHER

A suitable example to validate the multiple MBH approach is di-*n*-octyl-ether ($C_8H_{17}-O-C_8H_{17}$). Ethers are known to have a C–O stretching band that falls in the fingerprint region at 1050–1260 cm^{-1}. This vibration is a localized mode and very characteristic for all ethers. The ability of the various models in reproducing this stretching band is a strong validation for the MBH method.

The geometry of the molecule is first fully optimized at the B3LYP/6-31+$g(d)$ level of theory, and a normal mode

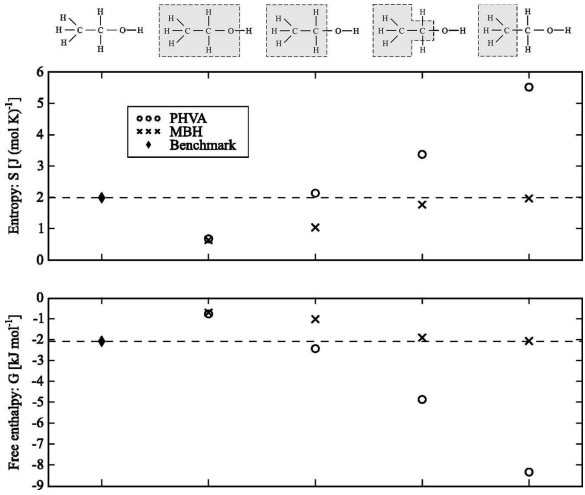


FIG. 2. Vibrational contribution to the entropy S and the free enthalpy G calculated with PHVA (○) and MBH (×) frequencies are given for the different partially optimized ethanol configurations at $T=298.15$ K. Benchmark values are indicated by the dashed lines. The fixed block in the MBH calculation consists of the atoms in the shaded box.

TABLE III. Frequencies $\lambda^{1/2}$ (in cm^{-1}) of di-*n*-octyl-ether of various partially optimized configurations defined in Fig. 3 are compared with the benchmark frequencies of the fully optimized geometry (left column). Three approaches are used: the full Hessian calculation (Full), the PHVA method, and the MBH approach. The size of the rigid bodies is defined by the configuration label. The Full/PHVA/MBH frequencies are ordered according to the maximum overlaps $|\langle \nu_{\text{bench}} | \nu \rangle|^2$ with benchmark eigenmodes, which are given by the values between parentheses (in %).

Bench	Configuration 8						Configuration 5						Configuration 2					
	Full		PHVA		MBH		Full		PHVA		MBH		Full		PHVA		MBH	
0	0	(100)	-	(-)	−1	(100)	0	(100)	29	(33)	−1	(100)	0	(100)	12	(56)	−1	(100)
0	0	(100)	-	(-)	0	(100)	0	(100)	-	(-)	−1	(100)	0	(100)	134	(73)	0	(100)
0	0	(100)	-	(-)	0	(100)	0	(100)	-	(-)	0	(100)	0	(100)	22	(58)	0	(100)
1	−8	(95)	-	(-)	1	(100)	7	(92)	-	(-)	1	(100)	9	(88)	228	(17)	0	(100)
1	−12	(96)	-	(-)	0	(100)	−12	(96)	-	(-)	1	(100)	−9	(98)	-	(-)	−1	(100)
4	−156	(61)	-	(-)	4	(97)	−124	(29)	-	(-)	4	(99)	-	(-)	34	(67)	3	(100)
11	9	(65)	-	(-)	16	(95)	27	(54)	-	(-)	12	(99)	34	(53)	228	(17)	11	(100)
19	−15	(98)	-	(-)	24	(98)	−15	(99)	50	(40)	19	(100)	−9	(99)	-	(-)	19	(100)
27	−154	(40)	-	(-)	42	(67)	−124	(32)	-	(-)	20	(99)	−36	(29)	63	(49)	27	(100)
33	53	(37)	-	(-)	64	(54)	−52	(43)	-	(-)	35	(98)	64	(34)	34	(59)	33	(100)
51	36	(96)	-	(-)	114	(78)	37	(95)	132	(44)	55	(99)	40	(97)	53	(60)	51	(100)
53	−104	(34)	215	(30)	283	(52)	−53	(41)	-	(-)	60	(94)	37	(53)	-	(-)	53	(100)
65	-	(-)	-	(-)	-	(-)	57	(50)	57	(39)	79	(86)	-	(-)	63	(42)	65	(100)
84	−96	(25)	-	(-)	-	(-)	61	(32)	79	(30)	101	(71)	75	(66)	-	(-)	84	(100)
92	-	(-)	-	(-)	-	(-)	90	(48)	101	(33)	117	(63)	-	(-)	95	(78)	92	(100)
96	84	(98)	261	(22)	-	(-)	86	(97)	-	(-)	114	(93)	87	(98)	102	(72)	96	(100)
113	77	(32)	-	(-)	-	(-)	90	(34)	-	(-)	-	(-)	109	(51)	95	(51)	113	(100)
129	90	(57)	-	(-)	-	(-)	118	(76)	-	(-)	362	(54)	153	(34)	133	(87)	130	(100)
133	110	(41)	244	(50)	300	(47)	116	(52)	144	(48)	-	(-)	132	(68)	122	(62)	133	(100)
138	137	(100)	-	(-)	180	(88)	138	(100)	-	(-)	145	(99)	138	(100)	276	(56)	138	(100)
150	-	(-)	-	(-)	-	(-)	139	(83)	228	(44)	362	(51)	169	(53)	149	(62)	150	(100)
154	143	(98)	-	(-)	-	(-)	144	(99)	-	(-)	198	(79)	145	(99)	167	(71)	154	(100)
160	127	(44)	-	(-)	-	(-)	152	(80)	-	(-)	149	(54)	158	(98)	157	(68)	160	(100)
162	128	(46)	-	(-)	-	(-)	150	(80)	130	(36)	143	(55)	162	(73)	154	(76)	162	(100)
178	160	(83)	125	(29)	143	(41)	175	(95)	230	(57)	169	(78)	182	(82)	178	(95)	178	(100)
220	209	(99)	-	(-)	336	(27)	210	(99)	243	(55)	209	(56)	212	(99)	233	(83)	220	(100)
...																		
1149	1146	(99)	1146	(94)	1147	(95)	1148	(100)	1148	(100)	1148	(100)	1149	(100)	1149	(100)	1149	(98)
1461	1457	(99)	1442	(83)	1442	(83)	1461	(100)	1461	(100)	1461	(100)	1461	(100)	1461	(100)	1461	(100)
1549	1547	(98)	1545	(88)	1545	(88)	1549	(100)	1548	(100)	1548	(100)	1549	(100)	1549	(100)	1549	(100)
2954	2953	(100)	2953	(99)	2953	(99)	2954	(100)	2954	(100)	2954	(100)	2954	(100)	2954	(100)	2954	(100)
2965	2965	(99)	2965	(99)	2965	(99)	2965	(100)	2965	(100)	2965	(100)	2965	(100)	2965	(100)	2965	(100)
2983	2983	(99)	2984	(98)	2984	(98)	2983	(100)	2983	(100)	2983	(100)	2983	(100)	2983	(100)	2983	(100)
2985	2985	(99)	2986	(98)	2986	(98)	2985	(100)	2985	(100)	2985	(100)	2985	(100)	2985	(100)	2985	(100)

analysis at the same level, constructed with analytical second derivatives, provides the whole spectrum of frequencies, which serve as benchmark values. Some eigenfrequencies are listed in the left column of Table III. The full Hessian frequency analysis on this completely optimized geometry generates six (almost) zero eigenvalues, as it should be. In the *low* frequency spectrum, typical modes are found that involve large parts of the molecule. In this class, bending and torsionlike modes are found where large massive blocks of the molecule are involved. Also internal rotations and collective accordionlike motions are present. For instance, the three lowest benchmark frequencies (11, 19, and 27 cm^{-1}) are identified as the relative rotation/torsion of the two octyl chains around the three axes of inertia of the molecule. In view of later application in which the frequencies are used as input for the construction of the partition function in the harmonic oscillator approximation, this low frequency spectrum is extremely important as they give the largest contribution to the vibrational partition function. In the *high* frequency range ($\sim$3000 cm^{-1}), localized C–H stretches are found. Some specific modes are interesting as they are localized near the central O atom: 2985, 2983, 2965, and 2954 cm^{-1} correspond to C–H stretching modes—symmetrically or antisymmetrically—of the CH_2 moieties attached to the middle oxygen, 1461 cm^{-1} is mainly a bending of these two CH_2 units with respect to each other and 1149 cm^{-1} is the typical C–O stretch for ethers.

In what follows, the multiple MBH model is applied to a set of partially optimized structures at the B3LYP/6-31 $+g(d)$ level of theory, while part of the molecule is kept fixed at the geometry optimized at the lower HF/STO-3*g* level. The central part of the molecule is always allowed to relax. Two nonoptimized blocks are systematically taken into account, located at both sides of the central part. They differ in the length of the chain that is considered in the nonoptimized

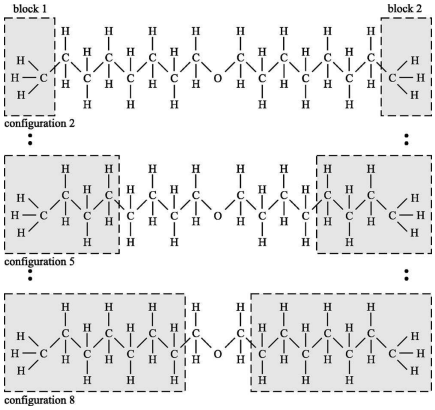


FIG. 3. Specification of the various configurations of di-*n*-octyl-ether with rigid bodies indicated as shaded regions. Atoms in shaded boxes are fixed at HF/STO-3*g* positions during the partial geometry optimization at the B3LYP/6-31+*g*(*d*) level.

block, as schematically depicted in Fig. 3. The influence of varying chain length on the reproduction of the frequency spectrum will be investigated. The fully optimized case is referred to as configuration 1 (the benchmark geometry), whereas in configuration 8, two heptyl chains are constrained during the optimization.

The low energy part of the spectra obtained by diagonalizing the full Cartesian Hessian (as would be done in the common normal mode analysis) is represented in Fig. 4. The same deficiencies related to nonequilibrium geometries as in the ethanol example are noticed: the appearance of non-negligible negative (imaginary) frequencies of the order of $-225\ \mathrm{cm}^{-1}$ and the lack of three eigenvalues equal to zero belonging to the global rotation. Moreover, the number of spurious negative frequencies can vary depending on the number of atoms belonging to the fixed block and is thus unpredictable. The repercussion of the negative frequencies is mainly found in the low frequency modes, where large parts of the molecule, including some nonoptimized atoms, are involved. Values for a selected number of modes are given in Table III. The reproduction of the three lowest eigenfrequencies is very poor and even negative values are found. It is, nevertheless, remarkable that the previously mentioned localized modes are very well reproduced even within the presence of some negative frequencies. For instance, the localized C–O band, situated at 1149 cm^{-1}, is well reproduced, probably owing to the fact that only optimized atoms are involved in this internal motion.

This numerical example illustrates the necessity to strictly reduce the dimension of the Hessian, following the procedure explained in the (multiple) MBH approach, in order to get rid of the disturbing negative frequencies. The low frequency part of the resulting spectra is plotted in Fig. 5 for the different configurations. Thereby all negative eigenvalues are eliminated, six eigenvalues are equal to zero as could be expected, and the other frequencies are all physically significant. Figure 5 allows qualitative comparison between the MBH normal mode frequencies and the benchmark values. Obviously, the MBH results belonging to the largest fixed blocks (configuration 8) differ substantially from the benchmark values. As the number of relaxed atoms increases, the agreement improves (configuration 2). Quantitative values are given for a selected number of modes in Table III, where the MBH modes were sorted according to their resemblance to the benchmark modes (see caption of table). The PHVA method was applied as well, combining the two individual fixed blocks into one fixed block, as shown in Fig. 6. The PHVA method is capable of reproducing localized modes, but frequencies in the lower spectrum are very poorly reproduced or are absent. Apparently, the MBH model is able to reproduce accurately not only the previously mentioned localized modes, but more importantly, the low frequency modes have much more realistic values compared to the full Hessian values or the PHVA values.

A more pronounced study can be made by evaluating the overlap between the calculated multiple MBH modes and

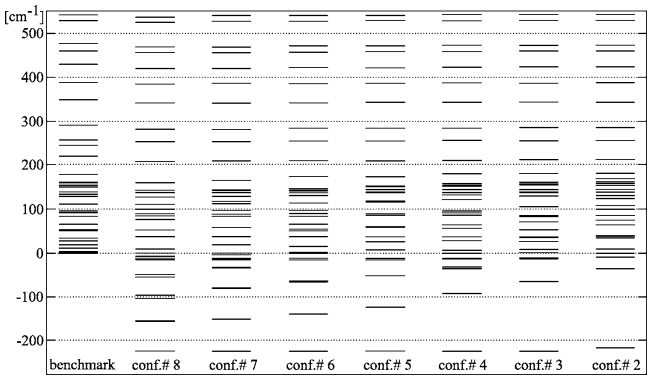


FIG. 4. Lowest frequencies $\lambda^{1/2}$ (in cm^{-1}) of di-*n*-octyl-ether based on the full Cartesian Hessian belonging to the various partially optimized configurations defined in Fig. 3. Partial optimization at the B3LYP/6-31+*g*(*d*) level. Plot on the left displays the exact normal mode frequencies (full geometry optimization) that serve as benchmark.

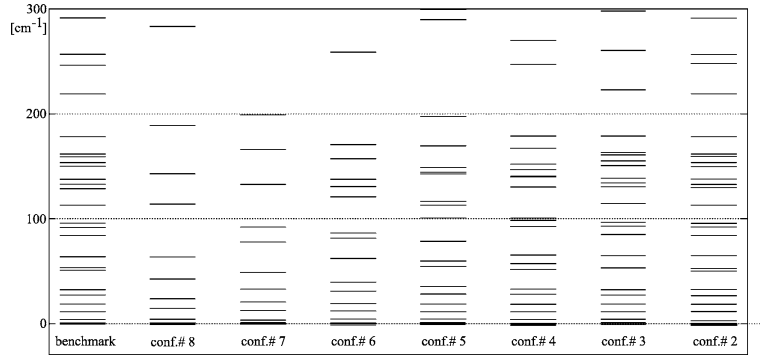


FIG. 5. Lowest frequencies $\lambda^{1/2}$ (in cm^{-1}) of di-*n*-octyl-ether based on the multiple MBH model belonging to the various partially optimized configurations defined in Fig. 3. Partial geometry optimization at the B3LYP/6-31+$g(d)$ level. Plot on the left displays the exact normal mode frequencies (full geometry optimization) that serve as benchmark for the other plots where two rigid bodies (defined by the configuration label) are taken into account in the frequency analysis.

benchmark modes. The square of the overlap $|\langle \nu_{\text{bench}}| \nu_{\text{MBH}}\rangle|^2$ gives a measure of how strong the benchmark mode with frequency ν_{bench} is involved in the specific MBH mode with frequency ν_{MBH}. Due to the completeness of the basis of benchmark modes, the sum of these numbers over all benchmark frequencies is equal to 1. The results for configurations 8, 5, and 2 are given in a scatter plot in Fig. 7. All values above a certain limit (we take 20% throughout this work) are indicated by a circle, and the darker the fill intensity of the circle, the larger the magnitude of the overlap. The most ideal case is a black filled circle on the diagonal, as it means that the multiple MBH model has reproduced the benchmark mode in an excellent way. If the strength of a specific MBH mode is spread out over various benchmark modes, a larger scattering of circles is noticed off the diagonal, which is indeed the case for configuration 8.

By enlarging the optimized region, the discrepancy with the benchmark almost disappears. In fact, already from configuration 5 results have essentialy converged (see Fig. 7). By retaining only the ending methyl groups in the fixed boxes (configuration 2), all relevant frequencies are excellently reproduced. In the low frequency region (below 1500 cm^{-1}), they are identical within a margin of 1 cm^{-1}. Only a few nonrelevant modes that are localized in the methyl tops are absent.

In Table III the maximum overlap $|\langle \nu_{\text{bench}}| \nu\rangle|^2$ between benchmark modes $|\nu_{\text{bench}}\rangle$ and the selected full Hessian, PHVA, and MBH modes $|\nu\rangle$ is added between brackets. The overlap for the localized modes is consistently excellent. The overlap for the full Hessian or PHVA modes is quite poor for the low frequencies, whereas the MBH modes show very reasonable overlaps even in this region.

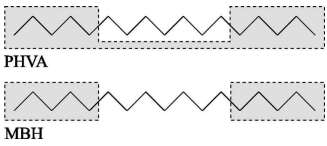


FIG. 6. The PHVA method implies the introduction of one block. For the multiple MBH method, two rigid blocks were used.

VI. SUMMARY AND CONCLUSIONS

In this paper a new method, referred to as the MBH approach, was introduced to calculate the normal modes of partially optimized molecular systems. The MBH approach is an extension of the previously introduced PHVA method in the sense that the nonoptimized regions of the system are allowed to move as rigid blocks with respect to the optimized part of the system. Both the MBH and PHVA methods eliminate the imaginary frequencies that result from the common full Hessian normal mode analysis on a partially optimized structure. In cases where the surrounding nonoptimized part of the system effectively are kept immobile by external forces at its reference position, such as in a lattice, the PHVA and MBH methods perform equally well, but for a flexible surrounding medium the MBH method is more appropriate. The various methods were outlined in two examples, i.e., ethanol and di-*n*-octyl-ether. It was found that the localized modes in the optimized part of the system are always well reproduced, irrespective of the applied method. Even a full Hessian normal mode analysis gives quite accurate values for the frequencies of localized modes. However, for normal modes that involve a larger part of the molecular system, the MBH method performs better, since the nonoptimized blocks can move coherently with respect to the chemically active part of the system.

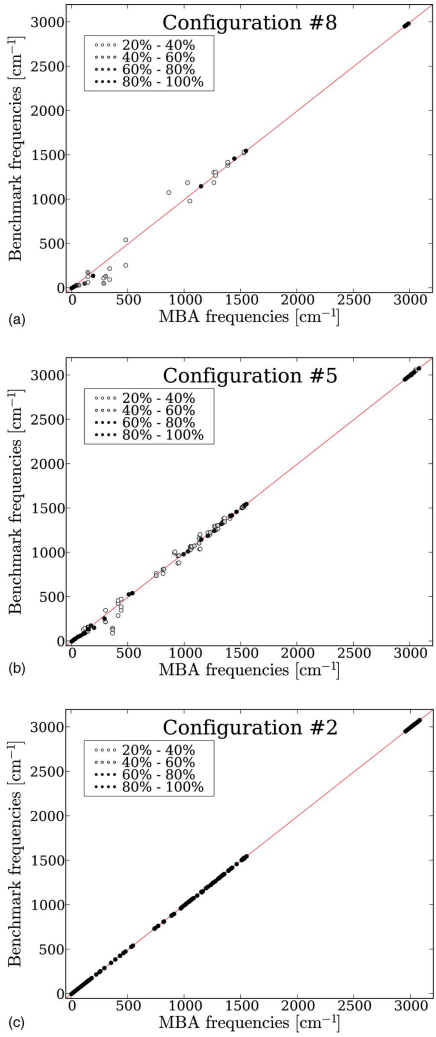


FIG. 7. Square of the overlap between the MBH normal modes and the benchmark normal mode frequencies of di-*n*-octyl-ether. The sum of the strengths is always normalized to 1 for each MBH frequency.

The MBH method does not only eliminate spurious negative frequencies but implicates also a serious reduction of the computational cost for large molecular systems, as the calculation of the Hessian is the most expensive part after a geometry optimization. Molecular modeling focuses more and more on these extended molecular systems, and hence, such efficient techniques are indispensable. The multiple MBH model looks highly suited for use in systems in which the molecular environment is rather flexible, such as reaction in solvents. Most of the solvent molecules can be regarded as rigid bodies moving in all directions with respect to the optimized central part of the system. Their participation to the normal modes can be simulated by the MBH model. The application of the MBH approach for the calculation of partition functions and derived quantities will be further investigated in the future.

ACKNOWLEDGMENTS

This work is supported by the Fund for Scientific Research-Flanders and the Research Board of Ghent University. This work was partly performed within the framework of the SBO-BIPOM program of the Institute for the Promotion of Innovation through Science and Technology in Flanders (IWT-Vlaanderen).

[1] A. Warshel and M. Levitt, J. Mol. Biol. **103**, 227 (1976).
[2] X. Assfeld and J. L. Rivail, Chem. Phys. Lett. **263**, 100 (1996).
[3] J. L. Gao, P. Amara, C. Alhambra, and M. J. Field, J. Phys. Chem. A **102**, 4714 (1998).
[4] Y. K. Zhang, T. S. Lee, and W. T. Yang, J. Chem. Phys. **110**, 46 (1999).
[5] F. Stevens, H. Vrielinck, V. Van Speybroeck, E. Pauwels, F. Callens, and M. Waroquier, J. Phys. Chem. B **110**, 8204 (2006).
[6] S. Q. Jin and J. D. Head, Surf. Sci. **318**, 204 (1994).
[7] M. D. Calvin, J. D. Head, and S. Q. Jin, Surf. Sci. **345**, 161 (1996).
[8] H. Li and J. Jensen, Theor. Chem. Acc. **107**, 211 (2002).
[9] N. A. Besley and K. A. Metcalf, J. Chem. Phys. **126**, 7 (2007).
[10] J. D. Head, Int. J. Quantum Chem. **65**, 827 (1997).
[11] J. D. Head and Y. Shi, Int. J. Quantum Chem. **75**, 815 (1999).
[12] J. D. Head, Int. J. Quantum Chem. **77**, 350 (2000).
[13] H. Lin, J. Z. Pu, T. V. Albu, and D. G. Truhlar, J. Phys. Chem. A **108**, 4112 (2004).
[14] A. Tachibana and K. Fukui, Theor. Chim. Acta **49**, 321 (1978).
[15] D. J. Wales, J. Chem. Phys. **113**, 3926 (2000).
[16] R. Murry, J. T. Fourkas, L. Wu-Xiong, and T. Keyes, J. Chem. Phys. **110**, 10410 (1999).
[17] I. Kolossvary and C. McMartin, J. Math. Chem. **9**, 359 (1992).
[18] J. Goldstone, A. Salam, and S. Weinberg, Phys. Rev. **127**, 965 (1962).
[19] V. Van Speybroectk, D. Van Neck, and M. Waroquier, J. Phys. Chem. A **104**, 10939 (2000).
[20] V. Van Speybroeck, P. Vansteenkiste, D. Van Neck, and M. Waroquier, Chem. Phys. Lett. **402**, 479 (2005).
[21] P. Vansteenkiste, D. Van Neck, V. Van Speybroeck, and M. Waroquier, J. Chem. Phys. **124**, 044314 (2006).
[22] M. J. Frisch, G. W. Trucks, H. B. Schlegel *et al.*, GAUSSIAN 03, Revision C.02, Gaussian, Inc., Wallingford, CT, 2004.

Paper 10: "Calculating Reaction Rates with Partial Hessians: Validation of the MBH Approach"

An Ghysels, Veronique Van Speybroeck, Toon Verstraelen, Dimitri Van Neck , Michel Waroquier

Journal of Chemical Theory and Computation , **2008**, 4, 614 - 625

614 *J. Chem. Theory Comput.* **2008,** *4,* 614–625

JCTC Journal of Chemical Theory and Computation

Calculating Reaction Rates with Partial Hessians: Validation of the Mobile Block Hessian Approach

A. Ghysels, V. Van Speybroeck, T. Verstraelen, D. Van Neck, and M. Waroquier*

Center for Molecular Modeling, Ghent University, Proeftuinstraat 86, B-9000 Gent, Belgium

Received October 24, 2007

Abstract: In an earlier paper, the authors have developed a new method, the mobile block Hessian (MBH), to accurately calculate vibrational modes for partially optimized molecular structures [*J. Chem. Phys.* **2007,** *126* (22), 224102]. The proposed procedure remedies the artifact of imaginary frequencies, occurring in standard frequency calculations, when parts of the molecular system are optimized at different levels of theory. Frequencies are an essential ingredient in predicting reaction rate coefficients due to their input in the vibrational partition functions. The question arises whether the MBH method is able to describe the chemical reaction kinetics in an accurate way in large molecular systems where a full quantum chemical treatment at a reasonably high level of theory is unfeasible due to computational constraints. In this work, such a validation is tested in depth. The MBH method opens a lot of perspectives in predicting accurate kinetic parameters in chemical reactions where the standard full Hessian procedure fails.

1. Introduction

Ab initio prediction of reaction rate constants of chemical reactions has a high computational cost, especially when large (bio)molecular systems are involved. An accurate description of chemical kinetics of reactions in gas phase is nowadays perfectly practicable for moderate-sized molecules, but once the molecular environment comes into play, one has to adapt the level of theory in such way to make the computation feasible.[1] This puts a heavy burden on the accuracy of the numerical results. Chemical kinetics in static approaches is still widely based on transition state theory (TST).[2–5] Key parameters are the reaction energy barrier between the reactants and activated complex (the transition state) and the vibrational frequencies, which serve as an input in the partition functions, and their accurate computation is essential. In the molecular-statistical formulation of TST, they completely determine the equilibrium constant by the use of partition functions.

In the harmonic oscillator approximation, the molecular partition function is factorized in a translational, rotational, and vibrational contribution, where the latter is completely determined by the eigenfrequencies. Frequencies are usually computed by a normal-mode analysis (NMA). This is the main bottleneck in ab initio predictions of chemical kinetics in large molecular systems, since frequency calculations are computationally very demanding even if analytical second derivatives are employed, rather than numerical ones. If a molecular mechanics (MM) force field is used instead of a quantum mechanics (QM) or hybrid (QM/MM)[6–9] description, the frequency calculation becomes less problematic, though even at the full MM level other issues, such as the storage and manipulation of the huge Hessian matrices associated with very large systems, can become prohibitive in real applications. Anyway, chemical reactions inherently involve bond breaking and charge transfer; so, it is essential to provide a QM description for (at least) the reactive region and a full MM description is usually no option.

In addition, there are computational limitations in the geometry optimization of extended systems at a high level of theory (LOT). Very often one goes over to a partial optimization: the interesting region containing the active site is optimized at a high LOT, while the environment is kept fixed at a low LOT geometry. This approach permits one to obtain an ab initio description of the chemically active site

* Corresponding author. E-mail: michel.waroquier@UGent.be.

10.1021/ct7002836 CCC: $40.75
Published on Web 02/27/2008

in large molecular systems, but at the same time, it creates several new problems. One of them is the extraction of accurate frequencies for the relevant vibrational modes. All partially optimized systems are nonequilibrium structures, and as a consequence of the residual gradients on the potential energy surface (PES), the standard full Hessian normal-mode analysis may show some unphysical results, e.g., spurious imaginary frequencies may appear. A frequency analysis in terms of a subset of coordinates that are optimized, i.e. a partial Hessian method, can avoid these problems.

The authors have succeeded recently in deriving a method that is able to calculate physical frequencies. The main idea is to group the atoms that were kept fixed during the partial optimization into one or more blocks that are able to move as rigid bodies with respect to the relaxed molecular part in the vibrational analysis.[10] This mobile block Hessian (MBH) method has shown to be very efficient for an accurate evaluation of relevant frequencies of vibrational modes. The proposed procedure remedies the artifact of imaginary frequencies occurring in standard frequency calculations for partially optimized systems. In addition, only a subblock of the Hessian matrix has to be constructed and diagonalized, leading to a serious reduction of the computation time for the frequency analysis.

MBH can be regarded as an extension of the partial Hessian vibrational analysis approach (PHVA). Only part of the cartesian Hessian has been retained, excluding all the atoms of the passive site of the molecule that is kept fixed during the optimization. This methodology was first introduced and developed by Head and co-workers[11–14] and was further investigated by Li and Jensen[15] and Besley and Metcalf.[16] It comes to giving an infinite mass to the fixed atoms so that they are frozen at their initial position. Only the relaxed atoms can participate in the small amplitude vibrations.

The novelty of MBH with respect to PHVA lies in the fact that, in the former, the finite mass of each block is taken into consideration in the NMA, instead of giving an infinite mass to the fixed atoms. Six degrees of freedom are attributed to each block to describe its position and orientation with respect to the fully optimized part, and the global translational/rotational invariance of the potential energy surface (PES) is fully respected. Moreover, the PHVA is always limited to the case of one immobile block with infinite mass, whereas in the MBH model, parts of the molecular system can be ranged in multiple blocks which can move as rigid bodies with respect to the relaxed part of the molecule. In ref 10, both PHVA and MBH methods are submitted to a tough comparative study, while in ref 17, attention is given to the practical implementation of the MBH model and the interface with molecular modeling program packages.

One of the main applications that can be deduced from the knowledge of accurate normal-mode frequencies, is the prediction of chemical kinetics, as already mentioned. By means of the partition functions and a molecular-statistical formulation of transition state theory, the reaction rate constant k of a chemical reaction can be determined.[2–5] A somewhat different approach is proposed by the group of Lin et al.[18] Basic assumption is that the Hessian elements that involve only the atoms of the active site might be more critical than the other Hessian elements. The less critical elements are approximated following some interpolation procedure, mainly for elements at the nonstationary points on the potential energy surface which are not consistently constructed by the same level (dual level scheme). Other related papers suggest proper methods to predict accurate QM/MM kinetics by incorporating quantum mechanical effects by treating vibrational motions quantum mechanically and applying multidimensional tunneling approximations into reaction rate calculations.[19,1] Recently, more sophisticated techniques concerning transition state theory have been developed including tunneling effects, quantum dynamical effects and multiple pathways (we refer to ref 20 for a review of all modern developments), but in view of the goal of this paper to validate the MBH approach in predicting kinetics, conventional TST largely suffices and tunneling and other effects will not be incorporated.

In principle, the expression of k includes all normal vibrational modes in reactants and activated complex. It is inherent to both MBH and PHVA approaches that the number of frequencies is always smaller than in a standard frequency calculation. The question arises whether this reduction has a significant influence on the reaction rate constant. Here lies the scope of this work: we will demonstrate that the normal modes which disappear when defining fixed blocks have little influence on the chemical kinetics. This work aims at promoting MBH as a suitable and highly efficient tool for predicting accurate chemical kinetics parameters in large extended molecular systems where the standard full Hessian procedure fails.

Applications of MBH are numerous. They can be classified in various categories:

(i) Large biosystems consisting of thousands of atoms require a hybrid quantum mechanical/molecular mechanical (QM/MM) approach.[6–9] The whole MM region can be taken up in one or multiple blocks.

(ii) A cluster description of zeolites or other periodic systems, such as lattices, requires fixed positions of border atoms to prevent collapse of the molecule during optimization.[21–23] This represents a particular situation of partial optimization.

(iii) Reactions in solvents often require an approach with a chemical reactive site and various layers treated at different levels of theory (QM/MM or QM/QM′). The whole can even be circumvented by a bulk solvent described by a polarizable continuum model (PCM).[24,25] In MBH, the various solvent molecules are all regarded as mobile blocks which can translate and rotate freely around the active site. Only the internal structure of each solvent molecule is held fixed.

The structure of the paper is as follows. In section 2, a short outline of the theoretical methodology is given, and in section 3, the computational details are summarized. Section 4 is devoted to the validation of MBH as an adequate method to predict rate constants. Different reactions with various block choices are taken up in the test set for validation with the benchmark values (full optimization before frequency calculation). The test set includes a prototype substitution reaction, a hydrogen transfer reaction, as well as several

radical addition reactions, since these have a localized reactive site. The effectiveness of the multiple MBH has been illustrated with a more extended aminophosphonate system in section 4.3, for which solvent molecules are taken into account. The results of the MBH have been compared to those of the PHVA approach as well, and based on theoretical considerations, a modified PHVA method is presented in section 4.4, hereafter referred to as PHVA*. Finally in section 5, some conclusions are drawn.

2. Theoretical Background

2.1. Partition functions. Within the harmonic oscillator approximation, the $3N$ degrees of freedom of a N-atom system can be decoupled into three groups of independent motions—3 translational, 3 rotational, and $3N - 6$ vibrational motions—that all contribute to the total partition function Q:

$$Q = qq_{\mathrm{elec}} \tag{1}$$

where

$$q = q_{\mathrm{trans}}q_{\mathrm{rot}}q_{\mathrm{vib}} \tag{2}$$

The translational partition function reads

$$q_{\mathrm{trans}} = \left(\frac{2\pi MkT}{h^2}\right)^{3/2} V \tag{3}$$

M stands for the total mass of the system, T is the temperature, k_B is the Boltzmann constant, h is the Planck constant, and V is the volume. If I_1, I_2, and I_3 denote the moments of inertia of the system and σ is the symmetry number, the rotational partition function reads

$$q_{\mathrm{rot}} = \frac{8\pi^2}{\sigma}\left(\frac{2\pi k_B T}{h^2}\right)^{3/2}\sqrt{I_1 I_2 I_3} \tag{4}$$

Each vibration with frequency ν_i, gives a contribution

$$q_{(\nu_i)} = \frac{\mathrm{e}^{-h\nu_i/2k_BT}}{1-\mathrm{e}^{-h\nu_i/k_BT}} \tag{5}$$

to the total vibrational partition function

$$q_{\mathrm{vib}} = \prod_i q_{(\nu_i)} \tag{6}$$

The electrons also contribute to the partition function, but when the first electronic excitation energy is much greater than k_BT, the first and higher excited states are assumed to be inaccessible. If E_0 is the energy of the ground-state level, this assumption simplifies the electronic partition function to

$$q_{\mathrm{elec}} = \mathrm{e}^{-E_0/k_BT} \tag{7}$$

Note that the zero point energy contribution $\mathrm{e}^{-h\nu_i/2k_BT}$ in the numerator of eq 5 is frequently left out from the vibrational partition function and incorporated in the electronic partition function.

Ab initio molecular calculations can be used to generate the molecular properties required for the evaluation of the above partition functions, such as the geometry (for the moments of inertia I_i), the Hessian matrix (for the vibrational frequencies ν_i), and the electronic ground-state energy E_0.

2.2. Reaction Rate Constant within Conventional Transition State Theory (TST). Transition state theory has been proved to be very useful to determine the reaction rate constants.[2–5] It supposes that the transition state or activated complex is in equilibrium with the reactants, although, strictly speaking, this hypothesis is not valid since the transition state corresponds to a saddle point rather than a minimum on the PES. Within this assumption the rate constant is completely determined by the microscopic partition functions and the reaction barrier at 0 K.

For a unimolecular reaction, $A \rightarrow A^{\ddagger} \rightarrow B$ or $A \rightarrow A^{\ddagger} \rightarrow B + C$ (with the $\ddagger$ superscript indicating the activated complex) the rate constant k is given by:

$$k(T) = \frac{k_BT}{h}\frac{q(\ddagger)/V}{q(A)/V}\mathrm{e}^{-\Delta E_0/k_BT} \tag{8}$$

ΔE_0 represents the molecular energy difference at 0 K between the activated complex and the reactants. The transition state frequency is assumed not to be included in the partition function $q(\ddagger)$ of the activated complex. k is expressed in units s^{-1}.

For a bimolecular reaction $A + B \rightarrow (AB)^{\ddagger} \rightarrow C$ or $A + B \rightarrow (AB)^{\ddagger} \rightarrow C + D$, the expression for the rate constant becomes

$$k(T) = \frac{k_BT}{h}\frac{q(\ddagger)/V}{q(A)/V\, q(B)/V}\mathrm{e}^{-\Delta E_0/k_BT} \tag{9}$$

expressed in units of cubic meters per mole second.

2.3. MBH and PHVA. Partially optimized geometries are nonequilibrium structures. The usual normal-mode analysis (NMA) equations $H\nu = \omega^2 M\nu$, with H being the full cartesian Hessian and M the Cartesian diagonal mass matrix, could be solved to obtain the frequencies, but this procedure shows some serious defects. The Hessian H is the second derivative matrix of the potential energy with respect to all the Cartesian coordinates. At nonequilibrium geometries, it has only three zero-eigenvalues instead of six, implying that the rotational invariance of the potential energy surface is not manifest anymore.[26] Spurious imaginary frequencies appear. Moreover, the eigenvalues of the Hessian depend on the choice of coordinates.[27,28]

In the partial Hessian vibrational analysis (PHVA),[13,15] these defects are surmounted by giving the fixed atoms an infinite mass. The normal mode equations are then restricted to the relaxed atoms only, by taking a submatrix of the Hessian and the mass matrix:

$$H_E\nu = \omega^2 M_E\nu \tag{10}$$

The mobile block Hessian (MBH) model has been proposed recently by the authors[10] as an improvement of the PHVA. In the MBH model the fixed part is considered as a rigid body that is allowed to participate in the small amplitude vibrations, thus taking into account the finite mass of the fixed block. The spurious frequencies and the coordinate dependence are avoided since the system composed of optimized atoms plus block is in equilibrium. Relying on the global translational and rotational invariance, it is possible[10] to write the single block MBH normal mode equations in terms of the same submatrix H_E of the Hessian,

while the corresponding mass matrix is adapted because of the finite block mass:

$$H_E \nu' = \omega'^2 \tilde{M}' \nu' \quad (11)$$

with

$$\tilde{M}' = M_E - M_E D_E S^{-1} D_E^T M_E \quad (12)$$

The matrix D_E is constructed in terms of the coordinates of the free atoms with respect to a space fixed frame. The matrix S contains the information on the mass distribution, i.e. the total mass and the moments of inertia of the molecule. Details can be found in the Appendix and more extensively in ref 17.

The usefulness and applicability of the MBH approach are seriously increasing in case of extension to several mobile blocks. The multiple MBH takes into account the finite mass of each block, by including six parameters per block describing its position and orientation into the NMA equations and by mass weighting with the appropriate block mass and moments of inertia.

The multiple MBH method is for instance extremely useful when simulating chemical reactions in a solvent. Solvent molecules can easily be associated to rigid blocks with a fixed internal structure. They can move freely with respect to each other and with respect to the active site of the molecule.

At first sight the MBH is similar to the united atom concept in force fields, since groups of atoms are treated there also as a single entity.[29] However, in spite of this resemblance, the MBH is essentially different. In the MBH blocks each atom keeps its identity and continues to contribute individually to, e.g., moments of inertia, Hessian elements, steric hindrance, etc. Coarse-grained or united atom methods reduce the number of atoms and the initial all-atom potential energy surface is approximated by a parametrizized PES of lower dimension. The MBH on the other hand does not simplify the potential energy surface but freezes certain degrees of freedom when performing the vibrational analysis.

3. Computational Details

In order to validate both MBH and PHVA methods in their performance in reproducing accurate chemical kinetics, we compare the MBH and PHVA predictions for the reaction rate constant with benchmark values k.

Benchmark structures and frequencies are generated with a full geometry optimization at a high level of theory (DFT/B3LYP/6-311 g**) with tight convergence criteria such that the residual gradients on the PES are negligibly small. Consequently, a frequency calculation is carried out at the same level of theory for the whole molecular system. These equilibrium geometries permit to calculate the reaction rate with the full cartesian Hessian frequencies.

In a first analysis, frequencies and rate constants are calculated for the fully optimized geometry, while the block size is varied in the vibrational analysis. For each reaction under study, we take into consideration various choices of fixed blocks, or, various submatrices H_E of the Hessian. The normal mode equations, eqs 10 and 11, are constructed and solved using the same geometry, and thereby, any perturbation resulting from geometry differences is excluded in this particular treatment. This comparative study is thus highly appropriate to investigate the influence on the rate constants of exclusion of parts of the Hessian in the frequency calculation, i.e. limiting the NMA to a partial Hessian.

In a second analysis, partial geometry optimization is performed and consequently followed by a frequency calculation. For the MBH model, the position/orientation (six degrees of freedom) of each block are optimized, in contrary to PHVA, where the atoms in the single block are kept fixed in space. Therefore, a partial optimization with multiple blocks produces a better structure than one with a single block. We remind that PHVA is always limited to a single block, whereas MBH is very suitable to treat multiple blocks.

The partial optimization is performed as follows. First, one optimizes the system at a low level of theory (HF/STO-3g) to find a plausible starting structure. Then the rigid blocks are introduced and the system is partially optimized at a high level of theory (DFT/B3LYP/6-311 g**), while keeping the rigid blocks fixed at their initial internal geometry. All calculations were carried out with the Gaussian03 software package.[30] Next, a frequency calculation is performed with the second derivatives of the potential energy using the same high level of theory. The standard full Hessian frequency analysis would give unphysical results due to the residual forces present in the partial optimized structures, as mentioned in the Introduction. Instead, the PHVA or MBH normal mode eqs 10 and 11 are constructed, as these yield physical frequencies. Obviously, the same rigid blocks are chosen as those considered in the precedent partial optimization.

A partial Hessian method such as the MBH or PHVA approach, however, reduces the number of calculated frequencies. The difference in the number of degrees of freedom between reactants and transition state determines the temperature dependence of the reaction rate, as can be easily seen by inspecting eqs 8 and 9. This difference should not change when introducing the MBH blocks. It is therefore obvious that the chosen blocks must consist of the same atoms in reactant(s) and transition state. Note that strictly speaking the internal rigid block geometry might differ between reactants and transition state, because of the first step, i.e. optimization at the low level of theory before the actual partial optimization.

Finally, we made a selection of various chemical reactions for the validation. Most of them are radical addition reactions, but also one prototype substitution reaction (S_N2) and the hydrogen abstraction of one of the ending carbons are included (R6 and R7, respectively). In these reactions the reactive site (the radical center) is well localized. We choose addition reactions of ethene to a large variety of radicals with different substituents. It enables us to select various types of blocks (large and heavy blocks, substituents with ring structure(s), etc) and to give some recommendations in choosing the fixed blocks and the relaxed molecular region.

An overview of the different reactions under study is depicted in Figure 1. The reactions are labeled as R1, R2, etc. and the reactants and products are numbered. The block choices in the reactants are indicated and labeled in Figures

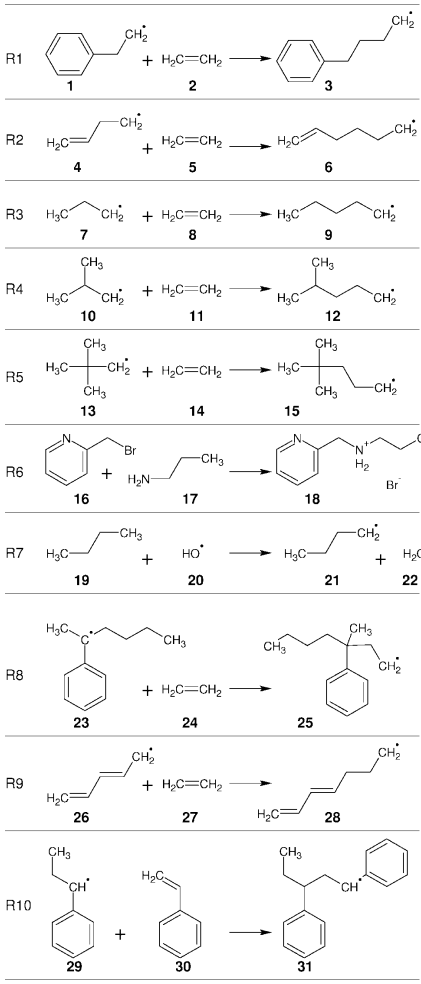


Figure 1. Overview of studied reactions.

2 and 3, and the transition state and product are assumed to contain the same block(s). Blocks can be classified in various types: they can include the reactive center (which at first view appears to be a surprising choice), they can directly be connected with the reactive center by a single bond, or they are separated by more than one bond. It is also possible to combine blocks to the case of multiple blocks. In bimolecular reactions, rigid blocks can be introduced in each of the two reactants; in the activated complex and product, they form multiple blocks, hindering in principle the application of the PHVA method. To illustrate, in reaction R10, one can combine blocks a and b in the description of the two

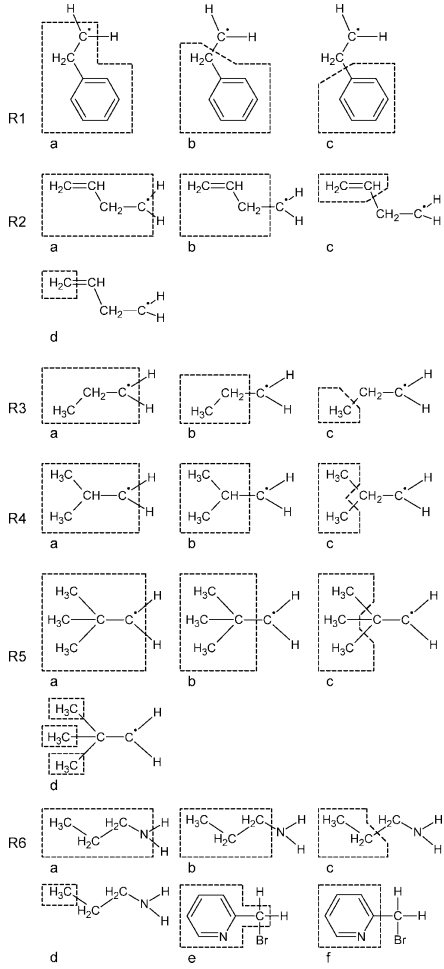


Figure 2. Numbering and choice of the different blocks.

reactants. This case will be denoted as a−b and only makes sense when using MBH.

4. Discussion

4.1. MBH with a Single Block. In the MBH (PHVA) approach, the total partition function of eq 1 is used to calculate the reaction rate, but the vibrational partition function q_{vib} is constructed with the MBH (PHVA) frequencies:

$$q^{MBH} = q_{trans} q_{rot} q_{vib}^{MBH} \quad (13)$$

$$q^{PHVA} = q_{trans} q_{rot} q_{vib}^{PHVA} \quad (14)$$

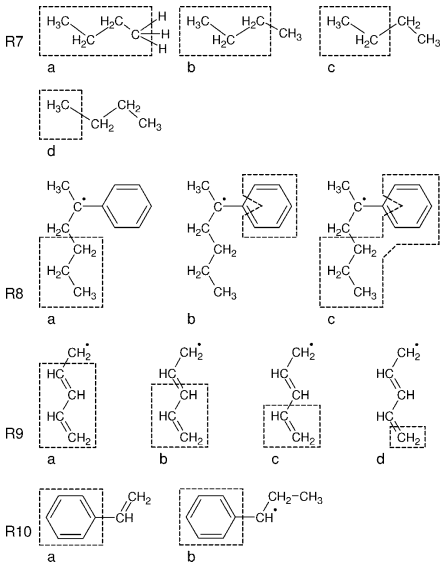


Figure 3. Numbering and choice of the different blocks—continuation.

The chosen test set of chemical reactions allows an exhaustive investigation of the influence on the rate constant of the position of the rigid block, the block's mass, its distance to the reactive center, and the stochiometry of the reaction.

In Tables 1 and 2, the rate constants at $T = 300$ K are listed for the several reactions in units of cubic meters per mole second (bimolecular reactions) or inverse seconds (unimolecular reactions). In the first column, the benchmark values k, calculated with the full Hessian frequencies of the equilibrium structure, are tabulated for comparison. The benchmark is only available for the fully optimized structure and is calculated in absence of any block. The block size in the MBH or PHVA approach applied on a fully optimized structure is indicated by a, b, c etc. A prime is added if the geometry was obtained by partial optimization, e.g. a′, b′, c′, etc.

In a first step, we concentrate on the results obtained with the fully optimized structures. In the next step, the influence of the partial optimization will be discussed.

As can be seen in Tables 1 and 2, the overall agreement of the MBH rate constants with the benchmark values is remarkably good. The reaction rate constants are reproduced to within a factor of 2, apart from a few cases in Table 2, which are discussed further. This observation holds for a variety of reactions: for unimolecular and bimolecular reactions, for radical and nonradical reactions, and for heavy or small block masses. The deviation is within acceptable limits and is smaller than corrections induced by the level of theory,[31] internal rotations,[32,33] tunnel effects, and other factors.[34,35]

The apparent agreement of the MBH predictions with the benchmark implies that the contribution of the omitted normal modes, inherent to the MBH method, is of the same magnitude for both the transition states and reactants. Apparently the omitted modes are not essential in the determination of the rate constant. These rather unimportant modes are localized in the fixed block or spread out over the fixed block and the optimized region. The more interesting modes are located in the optimized region that contains the active site, and are well reproduced by the MBH approach. The coupling of the MBH modes with the modes localized in the fixed block is left out in the model, but a logical choice of the blocks makes this coupling irrelevant for the rate constant.

When a block is chosen too close to the active site, the coupling between MBH modes and the omitted modes is not always irrelevant anymore. In reaction R1 the rigid block a includes the reactive center, and the border of block b crosses the bond connecting the radical center. The reaction rate constant k^{MBH} indeed overestimates the benchmark value. Block c is a better choice because it is not directly connected to the reactive site.

In some particular cases, e.g. reaction R4 with block a or b, the MBH approach reproduces k fairly well even with a direct bond between active site and block. However, one should not rely on such coincidences, and anyway, a more suitable choice of a block further away from the radical center still improves the rate estimate. As a general rule, hereafter referred to as the bond-distance rule, it is recommended not to bring the block region too close to the active site.

The mass of the rigid block does not play a crucial role in the validation of MBH in reproducing rate constants. This is best illustrated by comparing reactions R1 and R2. In R1, the fixed block c contains a phenyl group, while block c in R2 consists of an ethyl group. Results are comparable for both the forward and reverse reactions.

When we finally consider the results of the partial optimization, it is clear that the effect is rather moderate. We concentrate on the forward reaction R1 for a detailed study (Table 3). The partial optimization affects the geometry, because the rigid block conserves its initial internal geometry. This will cause differences with the benchmark geometry. In this simple example, this induces quite slight changes (some C−C distances are increased by 0.03 Å), but in more complex systems, the low level of theory geometry and partial optimized geometry may differ substantially. Or, the full optimization at the low level of theory should give a plausible internal geometry for the blocks, but the exact position/orientation of the blocks and the positions of the relaxed atoms are less important, since these are optimized during the consecutive partial optimization at the high level of theory, giving a plausible geometry of the whole system.

The ground-state configuration of a partially optimized system is obviously less bound than the fully optimized system. However, the energy increase of 2 kJ/mol, noticed in the ethylbenzene radical, is mostly compensated by a similar increase of the binding energy of the TS, hence resulting in an almost equal reaction barrier. For instance,

Table 1. Calculated Rate Constants at $T = 300$ K, for Reactions R1–R4 of the Test Set[a]

	forward					backward				
reaction	k	block	k^{PHVA}/k	k^{PHVA*}/k	k^{MBH}/k	k	block	k^{PHVA}/k	k^{PHVA*}/k	k^{MBH}/k
R1	3.46E−02	a	5.36	1.23	1.36	1.86E−06	a	0.39	0.35	0.48
		b	7.44	1.71	1.74		b	0.91	0.83	0.83
		c	4.33	0.99	1.02		c	1.13	1.02	0.95
		a′	5.97	1.37	1.52		a′	0.37	0.34	0.46
		b′	7.58	1.73	1.76		b′	0.99	0.90	0.90
		c′	4.21	0.96	0.99		c′	1.15	1.04	0.97
R2	2.85E−02	a	10.00	1.00	1.12	1.78E−06	a	0.43	0.36	0.65
		b	14.68	1.47	1.44		b	1.03	0.86	0.88
		c	8.91	0.89	0.93		c	1.09	0.91	0.96
		d	9.62	0.96	1.00		d	1.16	0.96	1.00
		a′	8.31	0.84	0.94		a′	0.40	0.34	0.61
		b′	10.83	1.09	1.07		b′	1.06	0.89	0.92
		c′	8.41	0.84	0.88		c′	1.07	0.89	0.96
		d′	9.62	0.96	1.00		d′	1.16	0.96	1.00
R3	1.47E−02	a	14.68	0.98	1.23	1.51E−06	a	0.44	0.35	0.69
		b	19.84	1.32	1.33		b	1.00	0.81	0.85
		c	14.78	0.98	0.99		c	1.18	0.95	0.99
		a′	16.56	1.10	1.39		a′	0.41	0.33	0.65
		b′	21.08	1.40	1.41		b′	1.06	0.85	0.89
		c′	14.73	0.98	0.99		c′	1.22	0.98	1.03
R4	1.99E−03	a	7.21	0.82	1.10	4.93E−06	a	0.34	0.28	0.61
		b	9.14	1.04	1.09		b	0.75	0.62	0.70
		c	8.66	0.98	0.97		c	1.27	1.05	1.08
		a′	7.52	0.86	1.14		a′	0.28	0.23	0.49
		b′	8.52	0.97	1.01		b′	0.76	0.63	0.71
		c′	8.57	0.97	0.96		c′	1.38	1.14	1.16

[a] The forward rate constants are expressed in units of cubic meters per mole second (bimolecular), and the backward rate constants are in units of inverse seconds (unimolecular). The benchmark value k is given for comparison. Rate constants k^{MBH} (k^{PHVA}, k^{PHVA*}) are calculated with the MBH (PHVA, PHVA*) frequencies, for several block choices. The ratios reflect the influence of the MBH (PHVA, PHVA*) treatment with respect to the benchmark value. A block without a prime indicates a fully optimized structure, and a block with a prime indicates a partially optimized structure.

in the case study, a suitable choice of the fixed block (block c′) predicts a reaction barrier that is hardly different (by 0.04 kJ/mol) from the benchmark value (see Table 3). Significant changes of reaction barriers alter the reaction rate constant to a large extent, but apparently the various reactions R1–R5 of the test set give no indication of this behavior, if one respects sufficient distance between the fixed blocks and the reactive site.

In Table 3, the kinetic parameters A and E_a, determined within the temperature range 300−700 K, are also given. Activation energies remain almost unaffected as could be expected. Potential deviations of k^{MBH} originate from the pre-exponential factor, which is mainly determined by the vibrational contribution to the partition function.

The above discussion validates the use of the MBH model to predict the rate constant on an accurate level. A plausible choice of the fixed block is the only essential ingredient a potential user of MBH should take into account to get adequate predictions of chemical kinetics. MBH is computationally attractive, makes quantum chemical calculations feasible in extended molecular systems, and preserves the true reaction mechanism.

4.2. MBH with Multiple Blocks. The ability of MBH to choose multiple blocks freely moving but conserving their internal structure makes it a powerful tool to a broad range of applications. This is demonstrated in reactions R5, R8, and R10 of the test set, and the results are tabulated in Tables 1 and 2.

In reaction R5, the effect of multiple blocks compared to a single block (block choice d versus c) is moderate. Reaction R8 describes a more complex system. Two individual blocks a and b can be merged to one solid block c, or they can be considered as two mobile blocks a−b. Here, the multiple MBH implies a significant improvement with respect to the single block treatment c. Block c yields ratios 3.86 and 0.44 for the forward and backward reaction, respectively, while the multiple blocks a−b give values of 1.67 and 0.76. Inspection of the MBH values obtained with the individual blocks a and b shows that the global effect of multiple blocks is mostly given by the following multiplication rule:

$$\frac{k_{a-b}^{MBH}}{k} \approx \frac{k_{a}^{MBH}}{k} \times \frac{k_{b}^{MBH}}{k} \tag{15}$$

This seems to be true for the forward and backward reaction.

A third example is reaction R10 where we choose a block in each reactant. The TS will then contain two blocks, which are treated within the multiple MBH. The overall factor is indeed fairly well reproduced by the multiplication rule (14). At least it gives an indication on the global error induced by the presence of multiple blocks. A plausible block choice is always of importance to keep the error within the limits. Unphysical block choices are for example b in R8 and c in R9. In both cases the block's border crosses a bond that is part of a delocalized system, and therefore, the k^{MBH} ratios are badly reproduced, even when the bond-distance rule is respected.

Table 2. Calculated Rate Constants at $T = 300$ K, for Reactions R5–R10 of the Test Set[a]

	forward					backward				
reaction	k	block	k^{PHVA}/k	k^{PHVA^*}/k	k^{MBH}/k	k	block	k^{PHVA}/k	k^{PHVA^*}/k	k^{MBH}/k
R5	1.72E−03	a	5.46	0.84	1.01	2.94E−06	a	0.42	0.36	0.60
		b	6.57	1.01	1.04		b	0.91	0.77	0.85
		c	6.28	0.96	0.99		c	1.03	0.88	0.95
		d			1.04		d			0.98
		a′	6.12	0.94	1.13		a′	0.43	0.36	0.61
		b′	6.47	0.99	1.03		b′	1.07	0.91	0.99
		c′	6.46	0.99	1.01		c′	1.19	1.01	1.08
		d′			1.16		d′			1.23
R6	9.93E−12	a	1060.01	1.74	1.76	1.41E−13	a	0.51	0.49	0.69
		b	1002.56	1.64	1.71		b	0.81	0.78	0.78
		c	794.07	1.30	1.33		c	0.89	0.86	0.85
		d	657.28	1.08	1.10		d	1.11	1.08	1.00
		e	2.61	0.29	0.85		e	0.17	0.16	0.56
		f	8.70	0.96	1.00		f	0.83	0.80	0.80
R7	2.85E+07	a	4.15	0.43	0.95	1.24E−04	a	10.67	2.03	1.73
		b	4.66	0.49	0.99		b	8.37	1.59	1.56
		c	4.86	0.51	1.05		c	5.50	1.04	1.02
		d	4.79	0.50	1.03		d	5.41	1.03	1.02
		a′	5.06	0.53	1.16		a′	16.69	3.19	2.71
		b′	5.22	0.55	1.11		b′	9.74	1.84	1.82
		c′	5.03	0.52	1.08		c′	5.58	1.06	1.04
		d′	4.85	0.51	1.04		d′	5.48	1.04	1.04
R8	8.04E−11	a	2.14	1.14	1.11	2.97E+02	a	1.31	1.25	0.99
		b	3.01	1.61	1.50		b	0.89	0.84	0.77
		c	5.91	3.15	3.86		c	0.48	0.46	0.44
		a−b			1.67		a−b			0.76
R9	1.83E−09	a	21.89	2.53	2.64	3.27E+01	a	1.60	1.31	2.27
		b	14.00	1.62	1.50		b	0.99	0.81	1.29
		c	14.40	1.66	1.63		c	2.29	1.87	5.14
		d	10.24	1.18	0.98		d	0.69	0.57	2.04
R10	3.02E−06	a	43.51	0.89	0.96					
		b	30.18	1.34	1.59					
		a−b			1.53					

[a] See the footnote of Table 1 for more details.

Table 3. MBH for Reaction R1 with Fully and Partially Optimized Structures[a]

	full optim			partial optim		
	a	b	c	a′	b′	c′
k^{MBH}/k	1.36	1.74	1.02	1.52	1.76	0.99
A^{MBH}/A	1.29	1.59	1.03	1.38	1.70	1.01
$E_a^{MBH} - \Delta E_a$	−0.14	−0.21	0.01	−0.22	−0.02	+0.06
$\Delta E_0^{MBH} - \Delta E_0$	0	0	0	−0.13	+0.13	+0.04

[a] The rate constant is given at 300 K, and kinetic parameters are fitted in the temperature range 300–700 K. k and A are in cubic meters per mole second, and energies are in kilojoules per mole. Benchmark values: $k = 3.46 \times 10^{-2}$ m^3 mol^{-1} s^{-1}, $A = 67.22 \times 10^2$ m^3 mol^{-1} s^{-1}, $E_a = 36.53$ kJ/mol, $\Delta K_0 = 24.66$ kJ/mol.

4.3. MBH for Modeling Solvents. Finally, we have tested the concept of multiple blocks on a more realistic and more extended example, where several explicit solvent molecules are taken into account in the computation. In this case, blocks can be chosen within the reacting molecule and/or the solvent molecules can be treated as blocks, and moreover, the influence of solvent species on both reaction barrier and frequencies, i.e. pre-exponential factor, can be tested.

We have chosen the cyclization of functionalized aminophosphonates, as a representative reaction occurring in an organic solvent (reaction R11, see Figure 4). The choice of this reaction was inspired by a recent combined experimental and theoretical study on the formation of β-lactams by some

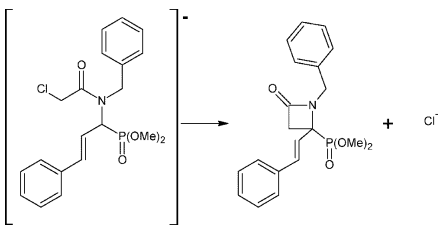


Figure 4. Reaction R11.

of the authors.[36] It was found that, starting from an ambient allylic anion, ring closure occurred exclusively by 4-ring formation, without any trace of 6-ring lactams. At that time, pre-exponential factors were not calculated due to the high computational cost. This is thus an ideal example to validate the approach.

The studied system consists of the aminophosphonate anion together with a sodium counterion and solvated in dimethyl ether solvent molecules (DME). The latter were taken as model molecules for tetrahydrofuran. Three cases are considered: the reaction in the absence of explicit solvent molecules (R11), in the presence of one DME (R11 + 1DME), and in the presence of two DMEs (R11 + 2DME). The benchmark values of k, A, E_a, and ΔE_0 are given in Table 4 for $T = 300$ K. The ratios (for k and A) and differences

Table 4. Benchmark Results for Reaction R11 without and with 1 and 2DME[a]

	R11	R11 + 1DME		R11 + 2DME	
k	7.62E−15	6.79E−14	(8.91)	2.09E−12	(274.63)
A	2.10E+13	8.43E+12	(0.40)	5.64E+13	(2.69)
E_a	157.73	149.99	(−7.73)	146.19	(−11.54)
ΔE_0	159.85	151.78	(−8.07)	148.23	(−11.63)

[a] The rate constant is given at 300 K, and kinetic parameters are fitted in the temperature range 300–700 K. *k* and *A* are in inverse seconds, and energies are in kilojoules per mole. Ratios (*k* and *A*) and differences (E_a and ΔE_0) between solvated and nonsolvated values are given between brackets.

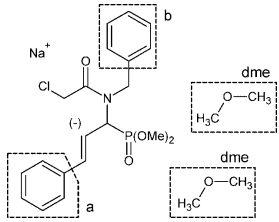


Figure 5. Definition of blocks, reaction R11.

Table 5. Calculated at 300 K[a]

block	R11	R11 + 1DME	R11 + 2DME
a	1.71	1.62	1.43
b	0.98	0.97	0.88
dme		1.07	
b−dme		1.04	
dme−dme			0.94
b−dme−dme			0.83

[a] Several blocks choices are taken up.

(for energies) given between brackets indicate the effect of the solvation. The presence of one or two solvent molecules indeed increases the reaction rate constant by a factor of 8.91 or 274.63, respectively, with respect to the nonsolvated situation.

The relevant question is whether the MBH model is capable of reproducing the enhancement of *k* due to the solvent. Several block choices are depicted in Figure 5, including the case of blocks within the reactant (a, b), as well as blocks consisting of solvent molecules (dme). Table 5 shows the ratios between the MBH estimates and the benchmark values of the rate constant. Block a is clearly not a good choice, which is easily understood when noting that the block's border cuts through a delocalized bond. Therefore, possible combinations of a with blocks b or dme are not considered in the table. Block b and block dme on the other hand are excellent block choices: since the ratios are close to 1.0, the *k* enhancement 1:8.91:274.63 as reported by the benchmark is maintained and the MBH is thus clearly capable of reproducing the solvation effect. Multiple block combinations such as b−dme, dme−dme, and b−dme−dme reproduce the rate constant very well, which is in agreement with the multiplication rule as stated in eq 14. Resuming, the multiple MBH has proven to be extremely useful and effective in predicting reaction rates, both with blocks belonging to the reactant or with blocks coinciding with solvent molecules.

4.4. PHVA and PHVA*. Conceptually, the difference between MBH and PHVA is mainly a mass effect. In the MBH, the finite mass of the blocks is taken into consideration, while in the PHVA approach, infinite masses are associated with the atoms in the rigid body. As a result, an extension to multiple blocks has no physical meaning in PHVA. When two blocks with infinite mass are present within one molecule, the system of free atoms and blocks will behave as if the two blocks were one big block with infinite mass. The case of one block in each of the reactants of a bimolecular reaction must also be excluded. The transition state itself would have two blocks with infinite mass. Thereby, six degrees of freedom describing the relative position and orientation of the two blocks will be lost in the transition state, leading to a completely wrong temperature dependence of the reaction rate constant. From a physical point of view, it is also hard to imagine how two reactants, each containing a block with infinite mass, could ever approach each other to form the transition state. The following discussion is therefore limited to the case of a single block with fixed geometry.

In contrast to the MBH, the PHVA cannot be extended to treat multiple blocks. PHVA is thus only applicable within the single block approximation. An overview of the various PHVA reaction rates in Tables 1 and 2 shows that unimolecular reactions are reasonably well described using PHVA frequencies. On the other hand, bimolecular reaction rates are poorly reproduced and significant deviancies are noticed. The systematic overestimation of the reaction rate finds its origin in the appearance of spurious low frequency modes in the PHVA approach. A profound investigation of these spurious modes reveals that they represent slow translation/rotationlike movements of the whole group of free atoms. This collective motion encompasses a lot of mass, explaining why (through the mass weighting in the NMA analysis) these frequencies are low. They give a significant contribution to the vibrational partition functions, while the translational/rotational degrees of freedom, however, are already taken into account in the total partition function. The larger the total mass of the free atoms with respect to the mass of the fixed block, the more pronounced is this double counting. Hence, in unimolecular reactions, the enhancement of the vibrational partition functions due to this double counting effect is nearly similar for reactant and transition state, and the enhancement factor is canceled (see eq 8). In bimolecular reactions on the other hand, the double counting is much more prominent for the transition state than for the reactants, thus leading to an overestimated reaction rate.

In order to prevent this double counting effect, we present a corrected version of the PHVA method. In Figure 6, the ratio $q_{vib}^{MBH}/q_{vib}^{PHVA}$ between the MBH and PHVA vibrational partition functions for the reactants, TS, and products of reactions R1–R6 is plotted against a mass related factor *t* given by

$$t = \sqrt{\frac{M_F^3 I_{F1} I_{F2} I_{F3}}{M^3 I_1 I_2 I_3}} \tag{16}$$

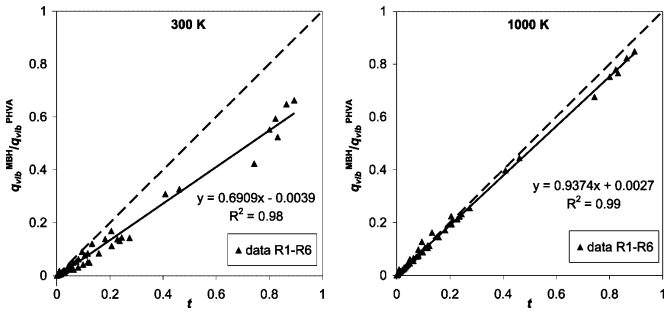


Figure 6. Ratio $q_{vib}^{MBH}/q_{vib}^{PHVA}$ for reactants, TS, and products of reactions R1– R6 plotted against the mass related factor t at 300 and 1000 K. The linear regression line (full) is fitted to the data with the least-squares method. The diagonal (dashed line) is added for comparison.

where M is the total mass and I_i ($i = 1, 2, 3$) are the moments of inertia of the molecule, while M_F and I_{Fi} ($i = 1, 2, 3$) are the total mass and moments of inertia of the fixed block. For higher temperatures, an almost linear behavior is observed.

$$\frac{q_{vib}^{MBH}}{q_{vib}^{PHVA}} \approx t \tag{17}$$

On the basis of eq 17, we now propose the following corrected PHVA partition function, hereafter referred to as PHVA*,

$$q^{PHVA*} = q_{trans} q_{rot} q_{vib}^{PHVA} \sqrt{\frac{M_F^3 I_{F1} I_{F2} I_{F3}}{M^3 I_1 I_2 I_3}} \tag{18}$$

It is not surprising that the ratio $q_{vib}^{MBH}/q_{vib}^{PHVA}$ depends only on mass properties, because the essential difference between the MBH and PHVA approach is a reduced mass effect. The plot in Figure 6 is so overwhelming that a mathematical proof of eq 16 should be on hand and that a close similarity between PHVA* and MBH predictions for the reaction rate constants is expected. The latter is indeed confirmed by the corrected PHVA* estimates taken up in Tables 1 and 2. Concluding, PHVA* and MBH perform equally well, but the main advantage of MBH, i.e. enabling the extension the procedure to multiple blocks, still holds.

In the following, a mathematical derivation of eq 17 will be presented. In the high temperature limit ($k_B T >> h\nu$, $\forall \nu$), it is possible to relate q_{vib}^{PHVA} and q_{vib}^{MBH} as a simple expression containing ratios of the masses and moments of inertia. The contribution of a vibration with frequency ν to the partition function is given by eq 5 and can be approximated by $k_B T/h\nu$ if the temperature is high with respect to the vibrational temperature $h\nu/k_B$. Since the number of MBH and PHVA frequencies is equal, the ratio is independent of temperature.

$$\frac{q_{vib}^{MBH}}{q_{vib}^{PHVA}} = \frac{\Pi_\nu \nu^{PHVA}}{\Pi_\nu \nu^{MBH}} \tag{19}$$

The product of frequencies coincides with the square root of the determinant of the matrix in the NMA normal mode equations. In the PHVA model (eq 10), this is

$$\prod_\nu \nu^{PHVA} = \sqrt{\det(M_E^{-1/2} H_E M_E^{-1/2})} \tag{20}$$

while in the MBH model (eq 11), this is

$$\prod_\nu \nu^{MBH} = \sqrt{\det(\tilde{M}'^{-1/2} H_E \tilde{M}'^{-1/2})} \tag{21}$$

The ratio becomes

$$\frac{q_{vib}^{MBH}}{q_{vib}^{PHVA}} = \sqrt{\frac{\det \tilde{M}'}{\det M_E}} \tag{22}$$

We now introduce the matrix S, defined in the Appendix, which contains the mass information of the complete system. Eigenvalues are the total mass M and the moments of inertia I_i. Similarly we introduce the matrix S_F for the fixed atoms, with eigenvalues M_F and I_{Fi}. Using the properties described in ref 17, the ratio can be rewritten (see the Appendix):

$$\frac{q_{vib}^{MBH}}{q_{vib}^{PHVA}} = \sqrt{\frac{\det S_F}{\det S}} \tag{23}$$

which is equivalent to expression 17 and which proves the PHVA* correction factor of eq 18.

Numerically, we find that eq 18 is not only valid for high temperatures but that its validity holds quite well for lower temperatures (300 K); see Figure 6.

An interesting property is that the mass related factor t of eq 16 is also equal to the ratio of the translational/rotational partition functions of the fixed block versus global molecule

$$\sqrt{\frac{M_F^3 I_{F1} I_{F2} I_{F3}}{M^3 I_1 I_2 I_3}} = \frac{q_{F,trans} q_{F,rot}}{q_{trans} q_{rot}} \tag{24}$$

The subscript F refers to the fixed block atoms. Thus, an alternative formulation of the PHVA* approach is presented, where only the vibrational partition function is taken into account. The total translational and rotational partition

function of the molecule are omitted as if it were to avoid the double counting effect.

$$q^{\mathrm{PHVA}^*} = q_{\mathrm{vib}}^{\mathrm{PHVA}} \tag{25}$$

Expression 25 is different from the one proposed in eq 17, but it amounts to the same result when calculating reaction rates. The factor $M_F^{3/2}$ is the same in both TS and reactants because the same block atoms are chosen, and thus, this factor is canceled in the numerator and denominator in the expression of k. The factor $\sqrt{I_{F1}I_{F2}I_{F3}}$ might slightly differ between TS and reactants, if the internal geometry of the blocks is not completely identical for the TS and reactants, but in good approximation, it is canceled as well. The factor $M^{3/2}$ cancels with the translational partition function and $\sqrt{I_1I_2I_3}$ with the vibrational partition function. Therefore, eq 25 will lead to (almost) identical results as eq 17.

5. Conclusion

In this work, the MBH method has been shown to act as an accurate method for the prediction of chemical kinetics in large extended molecular systems. In contrast to the PHVA approach, the MBH method also performs fairly well in bimolecular reactions. An adapted version of PHVA is presented correcting for the double counting effect of global rotation and translation inherent to the PHVA method. The surplus value of MBH with regard to PHVA* lies in the flexibility of MBH to introduce multiple rigid blocks which are freely moving with respect to each other but keeping their initial internal structure. This facility gives a lot of new perspectives in predicting chemical kinetics in very complex systems, where the introduction of one single fixed block is a too crude approximation. Partial optimization is necessary to make quantum chemical computations feasible. The possibility to introduce multiple blocks, each still having six degrees of freedom, makes an accurate reproduction of kinetics to the possibilities.

Most promising application field of MBH would be the description of chemical reactions in a solvent. Each solvent molecule may be regarded as a fixed block, keeping its internal structure, but still enabling to translate/rotate freely with respect to the chemically active part of the system. All ab initio program packages can be used on the condition that the built-in optimization routine allows constraints on internal degrees of freedom. The computational advantage of the MBH method can be exploited when the program package has the ability to calculate partial Hessians. If both features are implemented, MBH could be regarded as a groundbreaking model in the treatment of complex reactions where environment plays a crucial role.

Acknowledgment. This work is supported by the Fund for Scientific Research−Flanders and by the BOF funds of Ghent University.

Appendix

Consider a molecule with N masses m_A, $A = 1, \ldots, N$. The positions are described by Cartesian coordinates $\mathrm{r}_A \equiv \{r_{A\mu}\}_{\mu=x,y,z}$, with respect to a space-fixed frame. We will treat the case of one MBH block, consisting of N_F atoms. The remaining $N_E = N - N_F$ atoms are in equilibrium due to the partial optimization. An index E (F) will be used to indicate quantities where only the free (fixed) atoms are considered.

We will focus on the normal mode equations for PHVA and for MBH and, in particular, on the difference in the mass matrices, in order to study the transition from eq 22 to eq 23. The PHVA mass matrix is simply given by M_E [see ()]. The original (not yet transformed to (11)) MBH normal mode equations read

$$\tilde{H}\nu = \lambda\tilde{M}\nu \tag{26}$$

where $\tilde{M}$ and $\tilde{H}$ are the MBH mass matrix and Hessian [see ref 17].

Define now a $3N \times 6$ matrix D with components

$$D_{A\mu,\alpha} = \begin{cases} \delta_{\mu,x} & \alpha = 1 \\ \delta_{\mu,y} & \alpha = 2 \\ \delta_{\mu,z} & \alpha = 3 \\ \Sigma_\lambda \epsilon_{\lambda\mu x} & \alpha = 4 \\ \Sigma_\lambda \epsilon_{\lambda\mu y} & \alpha = 5 \\ \Sigma_\lambda \epsilon_{\lambda\mu z} & \alpha = 6 \end{cases} \tag{27}$$

With M as the diagonal $3N \times 3N$ mass matrix, the matrix $S = D^{\mathrm{T}}MD$ is introduced, and similarly, $S_F = D_F{}^{\mathrm{T}}M_FD_F$. The MBH mass matrix is then given by the block diagonal matrix

$$\tilde{M} = \begin{pmatrix} S_F & 0_{6\times d} \\ 0_{d\times 6} & M_E \end{pmatrix} \tag{28}$$

with $d = 3N_E$. The normal mode equations are transformed by simultaneous block diagonalization of $\tilde{H}$ and $\tilde{M}$. The required transformation matrices are given by

$$T_1 = \begin{pmatrix} 1_{6\times 6} & 0_{6\times d} \\ x & 1_{d\times d} \end{pmatrix}; \quad T_2 = \begin{pmatrix} 1_{6\times 6} & y \\ 0_{d\times 6} & 1_{d\times d} \end{pmatrix} \tag{29}$$

with $x = D_E$ and $y = -S^{-1}D_E{}^{\mathrm{T}}M_E$. The transformed MBH mass matrix and Hessian directly lead to eq 11:

$$T_2^{\mathrm{T}}T_1^{\mathrm{T}}\tilde{H}T_1T_2 = \begin{pmatrix} 0_{6\times 6} & 0_{6\times d} \\ 0_{d\times 6} & H_E \end{pmatrix}, \quad T_2^{\mathrm{T}}T_1^{\mathrm{T}}\tilde{M}T_1T_2 = \begin{pmatrix} S & 0_{6\times d} \\ 0_{d\times 6} & \tilde{M}' \end{pmatrix} \tag{30}$$

with $\tilde{M}' = M_E - M_ED_ES^{-1}D_E{}^{\mathrm{T}}M_E$. Or, the relevant mass matrix is $\tilde{M}'$ for MBH.

Since by construction $\det T_1 = \det T_2 = 1$, it is obvious that the following relations between determinants hold:

$$\det \tilde{M} = \det S_F \det M_E \tag{31}$$

$$\det (T_2^{\mathrm{T}}T_1^{\mathrm{T}}\tilde{M}T_1^{\mathrm{T}}T_2^{\mathrm{T}}) = \det \tilde{M} = \det S \det \tilde{M}' \tag{32}$$

or

$$\frac{\det \tilde{M}'}{\det M_E} = \frac{\det S_F}{\det S} \tag{33}$$

This proves the transition between eqs 22 and 23.

References

(1) Gao, J. L.; Truhlar, D. G. *Annu. Rev. Phys. Chem.* **2002**, *53*, 467–505.

(2) Eyring, H. *J. Chem. Phys.* **1935**, *3*, 107.

(3) Evans, M. G.; Polanyi, M. *Trans. Faraday Soc.* **1935**, *31*, 875.

(4) Laidler, K. J. *Chemical Kinetics*; Harper Collins Pulbishers, Inc.: New York, 1987, 87–138..

(5) Mc Quarrie, D. A.; Simon, J. D. *Physical Chemistry - a molecular approach*; University Science Books: Sausalito, CA, 1997; pp 1075–1079..

(6) Warshel, A.; Levitt, M. J. *Mol. Biol.* **1976**, *103* (2), 227–249.

(7) Assfeld, X.; Rivail, J. L. *Chem. Phys. Letters* **1996**, *263* (1–2), 100–106.

(8) Gao, J. L.; Amara, P.; Alhambra, C.; Field, M. J. *J. Phys. Chem. A* **1998**, *102* (24), 4714–4721.

(9) Zhang, Y. K.; Lee, T. S.; Yang, W. T. *J. Chem. Phys.* **1999**, *110* (1), 46–54.

(10) Ghysels, A.; Van Neck, D.; Van Speybroeck, V.; Verstraelen, T.; Waroquier, M. *J. Chem. Phys.* **2007**, *126* (22), 224102.

(11) Jin, S. Q.; Head, J. D. *Surf. Sci.* **1994**, *318* (1–2), 204–216.

(12) Calvin, M. D.; Head, J. D.; Jin, S. Q. *Surf. Sci.* **1996**, *345* (1–2), 161–172.

(13) Head, J. D. *Int. J. Quantum Chem.* **1997**, *65* (5), 827–838.

(14) Head, J. D. *Int. J. Quantum Chem.* **2000**, *77* (1), 350–357.

(15) Li, H.; Jensen, J. H. *Theor. Chem. Acc.* **2002**, *107*, 211–219.

(16) Besley, N. A.; Metcalf, K. A. *J. Chem. Phys.* **2007**, *126* (3), 035101.

(17) Ghysels, A.; Van Neck, D.; Waroquier, M. *J. Chem. Phys.* **2007**, *127*, 164108.

(18) Lin, H.; Pu, J. Z.; Albu, T. V.; Truhlar, D. G. *J. Phys. Chem. A* **2004**, *108* (18), 4112–4124.

(19) Garcia-Viloca, M.; Alhambra, C.; Truhlar, D. G.; Gao, J. *J. Chem. Phys.* **2001**, *114* (22), 9953–9958.

(20) Fernandez-Ramos, A.; Miller, J. A.; Klippenstein, S. J.; Truhlar, D. G. *Chem. Rev.* **2006**, *106* (11), 4518–4584.

(21) Stevens, F.; Vrielinck, H.; Van Speybroeck, V.; Pauwels, E.; Callens, F.; Waroquier, M. *J. Phys. Chem. B* **2006**, *110* (16), 8204–8212.

(22) Lesthaeghe, D.; Delcour, G.; Van Speybroeck, V.; Marin, G.; Waroquier, M. *Microporous Mesoporous Mater.* **2006**, *96*, 350−356.

(23) Lesthaeghe, D.; De Sterck, B.; Van Speybroeck, V.; Marin, G. B.; Waroquier, M. *Angew. Chem., Int. Ed.* **2007**, *46* (8), 1311–1314.

(24) Cui, Q. *J. Chem. Phys.* **2002**, *117* (10), 4720.

(25) Tomasi, J.; Mennucci, B.; Cammi, R. *Chem. Rev.* **2005**, *105* (8), 2999–3093.

(26) Tachibana, A.; Fukui, K. *Theor. Chim. Acta* **1978**, *49* (4), 321–347.

(27) Murry, R.; Fourkas, J. T.; Wu-Xiong, L.; Keyes, T *J. Chem. Phys.* **1999**, *110*, 10410–10422.

(28) Wales, D. J. *J. Chem. Phys.* **2000**, *113*, 3926–3927.

(29) Yang, L. J.; Tan, C. H.; Hsieh, M. J.; Wang, J. M.; Duan, Y.; Cieplak, P.; Caldwell, J.; Kollman, P. A.; Luo, R. *J. Phys. Chem. B* **2006**, *110* (26), 13166–13176.

(30) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A., Jr.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A. Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E. ; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. *Gaussian 03*, revision C.02; Gaussian, Inc.: Wallingford, CT, 2004.

(31) Van Speybroeck, V.; Van Cauter, K.; Coussens, B.; Waroquier, M. *Chemphyschem* **2005**, *6* (1), 180–189.

(32) Van Speybroeck, V.; Van Neck, D.; Waroquier, M. *J. Phys. Chem. A* **2000**, *104* (46), 10939–10950.

(33) Vansteenkiste, P.; Van Neck, D.; Van Speybroeck, V.; Waroquier, M. *J. Chem. Phys.* **2006**, *124* (4), 044314.

(34) Sabbe, M. K.; Saeys, M.; Reyniers, M. F.; Marin, G. B.; Van Speybroeck, V.; Waroquier, M. *J. Phys. Chem. A* **2005**, *109* (33), 7466–7480.

(35) Sabbe, M. K.; Vandeputte, A. G.; Reyniers, M. F. O.; Van Speybroeck, V.; Waroquier, M.; Marin, G. B. *J. Phys. Chem. A* **2007**, *111* (34), 8416–8428.

(36) Van Speybroeck, V.; Moonen, K.; Hemelsoet, K.; Stevens, C. V.; Waroquier, M. *J. Am. Chem. Soc.* **2006**, *128* (26), 8468–8478.

CT7002836

Paper 11: "Effect of temperature on the EPR properties of a rhamnose alkoxy radical: a DFT molecular dynamics study"

Ewald Pauwels, Toon Verstraelen and Michel Waroquier

Spectrochimica Acta Part A - Molecular and Biomolecular Spectroscopy, **2008**, 69, 1388 - 1394

Available online at www.sciencedirect.com
ScienceDirect
Spectrochimica Acta Part A 69 (2008) 1388–1394
SPECTROCHIMICA ACTA PART A
www.elsevier.com/locate/saa

Effect of temperature on the EPR properties of a rhamnose alkoxy radical: A DFT molecular dynamics study

Ewald Pauwels *, Toon Verstraelen, Michel Waroquier

Center for Molecular Modeling, Ghent University, Proeftuinstraat 86, B-9000 Gent, Belgium

Received 9 August 2007; accepted 17 September 2007

Abstract

It has been shown previously that two distinctive variants (called RHop and RO4) exist of the radiation-induced rhamnose alkoxy radical. Density functional theory (DFT) calculations of the electron paramagnetic resonance (EPR) properties were found to be consistent with two separate measurements at different temperatures [E. Pauwels, R. Declerck, V. Van Speybroeck, M. Waroquier, Radiat. Res., in press]. However, the agreement between theory and experiment was only of a qualitative nature, especially for the latter radical. In the present work, it is examined whether this residual difference between theoretical and experimental spectroscopic properties can be explained by explicitly accounting for temperature in DFT calculations. With the aid of ab-initio molecular dynamics, a temperature simulation was conducted of the RO4 variant of the rhamnose alkoxy radical. At several points along the MD trajectory, *g* and hyperfine tensors were calculated, yielding time (and temperature) dependent mean spectroscopic properties. The effect of including temperature is evaluated but found to be within computational error.

Keywords: DFT; EPR; Solid state; Alkoxy radical; α-Rhamnose; Periodic calculations; Molecular dynamics; Hyperfine coupling; *g* tensor

1. Introduction

One of the species that is commonly generated in irradiated crystalline sugars is an alkoxy radical. These oxygen-centered species readily decay under the influence of temperature or light and are generally considered to be primary radiation products. As such, they have attracted considerable attention. Since sugars have some features that are also present in more complex biomolecules (e.g. the deoxyribose unit in DNA), they effectively present ideal test systems to investigate initial radiation-induced events. Crystalline α-L-rhamnose is a representative system in this respect, since both primary oxidation and reduction processes have been identified in this sugar.

The leading technique to examine alkoxy radicals is electron paramagnetic resonance (EPR) spectroscopy at very low temperatures (both irradiation and measurement). The lack of thermal energy limits the conversion of primary radiation products in secondary reactions and this allows thorough spectroscopic characterization. Two EPR studies have been undertaken of the alkoxy radical in rhamnose. Samskog and Lund performed a Q-band EPR measurement at 77 K and determined the *g* tensor, along with two hyperfine coupling constants [2]. In a later study by Budzinski and Box, EPR and ENDOR (electron nuclear double resonance) were used to thoroughly characterize this species at 4.2 K: the *g* tensor and seven hyperfine tensors were distinguished [3]. Even though the same radical structure was proposed in both studies, the two experimental data sets are strikingly dissimilar, as noted by other authors [4]. The 77 K measurement revealed two hyperfine tensors with isotropic couplings of 112 and 39 MHz. At 4.2 K conversely, a myriad of smaller hyperfine interactions was detected along with two major isotropic couplings of 67 and 54 MHz. But both studies also differed in the *g* tensor: a maximum anisotropic *g* tensor component of 2.0456 was reported by Samskog and Lund, whereas Budzinski and Box found 2.0202.

In a recent theoretical study by the authors a basic reconciliation was made of both EPR studies [1]. The monohydrate crystal of rhamnose was simulated using density functional theory (DFT) in a periodic approach and several reaction steps were examined leading up to alkoxy radical formation. The proposed mechanism is summarized in Fig. 1. Resulting from oxidation, a primary rhamnose cation (PrimCat+) ejects a proton that migrates first to a nearby crystal water and then further on into the lattice. Molecular modeling provided

* Corresponding author. Fax: +32 9 264 66 97.
E-mail address: ewald.pauwels@UGent.be (E. Pauwels).

1386-1425/$ – see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.saa.2007.09.034

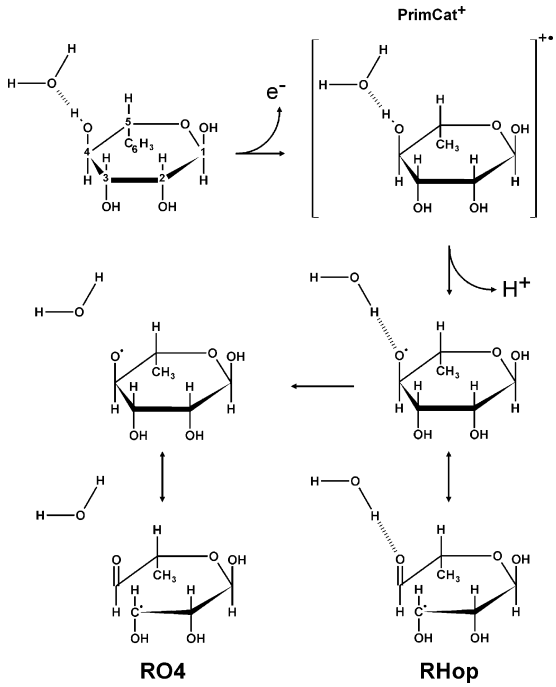


Fig. 1. Atom numbering scheme and summary of reaction steps leading to alkoxy radical formation in α-L-rhamnose. Oxygens and hydrogens are numbered according to the carbon to which they are bound.

evidence that this migration is in fact a proton hopping process – similar to the Grotthuss mechanism in solutions [5] – along an 'infinite' hydrogen bond chain parallel to the ⟨*b*⟩ axis of the crystal. The resultant of this proton removal is an alkoxy radical (dubbed RHop), for which calculated EPR properties were in close agreement with the 4.2 K EPR measurements. Yet, the subsequent proton hops alter the hydrogen bonds, also in the vicinity of the radical. The initial state of the H-bonds can be restored, by rotating several hydroxyl groups and crystal waters to their original orientation in the undamaged crystal. In this way the more stable RO4 variant of the alkoxy radical was obtained and calculated EPR properties were in agreement with the 77 K EPR measurements. Hence, the main difference between the two radicals observed in the experiments was not the radical structure itself, but rather its environment.

Yet, although the theoretical EPR predictions for RO4 and RHop succeeded in clarifying the differentiation between the 4.2 and 77 K measurements on a qualitative level, the reproduction on the quantitative level of the individual hyperfine and *g* tensor values with their experimental counterparts could be subject to some improvement. In particular for RO4, relatively large differences were found [1] for the maximum anisotropic *g* tensor component (2.0263 instead of 2.0456 experimentally) and – to a smaller extent – for the hyperfine couplings (50.4 and 87.1 MHz instead of 39.2 and 112.1 MHz). One could ascribe these residual differences between theory and experiment to temperature effects not taken into account in the theoretical description of the radical system. Static calculations simulate the system at a temperature of 0 K. One could argue that such an approach is acceptable for a comparison between the calculated spectroscopic properties of RHop and the 4.2 K measurements, but is insufficient when evaluating the theoretical EPR properties of RO4 with regard to those measured at a higher temperature of 77 K. In this work, the validity of this statement is assessed and calculations are presented on the RO4 alkoxy radical in which temperature is duly accounted for. To this end, molecular dynamics simulations have been performed of the radical structure at 77 K, adopting a periodic approach to include the molecular environment.

2. Computational details

In the previous study [1], the structure of α-L-rhamnose monohydrate [6] was consistently described in a periodic approach, exploiting the translational symmetry of the crystalline state. The original unit cell of the crystal was doubled in all directions, to ensure that the radical was well separated from its periodic images. However, the results of the calculations within this $\langle 2a\,2b\,2c\rangle$ supercell – containing 416 atoms – were found to be virtually on the same level as within an $\langle a\,2b\,c\rangle$ cell, obtained by only doubling the unit cell in the $\langle b\rangle$ direction. Therefore, the latter supercell was chosen in the present work as it considerably reduces the computational burden. An initial structure for the radical was obtained by removing the HO4 hydroxy proton from the model space (see Fig. 1 for the atom numbering scheme).

All periodic calculations were performed using the ab-initio Car–Parrinello approach [7] as implemented in the CPMD software package [8]. The BP86 gradient-corrected density functional [9] was used, together with a plane wave basis set (cutoff 25 Ry) and ultra-soft pseudopotentials of the Vanderbilt type to describe the electron–ion interaction [10]. First, the global minimum of the radical within the $\langle a\,2b\,c\rangle$ supercell was determined in a (static) geometry optimization. No constraints were imposed on any of the atoms in the supercell. Molecular dynamics simulations were then initiated starting from the completely optimized structure, with a simulation temperature of 77 K. Initial calculations were performed to quickly equilibrate the temperature of the system by rescaling ionic velocities whenever the instantaneous temperature differed more than 50 K from the target temperature (1000 time steps). Subsequently, a Nosé–Hoover thermostat was activated for the ionic and electronic systems. A characteristic frequency of 3000 cm^{-1} was chosen for the ionic thermostat, with a target temperature of 77 K. The electron thermostat was set to a frequency of 10,000 cm^{-1} and an average target kinetic energy of 0.012430 a.u. was determined for the electronic system from separate NVE molecular dynamics simulations. The MD timestep was 5 a.u. (0.12 fs) and the fictitious electronic mass was set to 400 a.u. The Nosé–Hoover run was equilibrated for 2000 time steps, after which 8000 further steps were considered as production quality, amounting to about 0.97 ps.

To determine the temperature dependence of the hyperfine and g tensors, EPR properties were calculated for 200 sampled snapshots along the resulting MD trajectory of 8000 time steps (one snapshot every 40 time steps). Spectroscopic parameters were determined consistent with the computational protocol adopted in the previous work [1]. For each snapshot, a cluster was cut out of the periodic system to contain the radical and all the molecules that are hydrogen bound to it (seven rhamnose and eight water molecules). Hyperfine tensors were calculated with the aid of the Gaussian03 software suite [11], using the B3LYP functional [12] and a 6-311G** basis set [13] for all atoms within the cluster. However, this level of theory is prohibitively expensive from a computational point of view for the calculation of g tensors. Consequently, the 6-311G** basis set was only maintained for the atoms of the central radical along with those of two nearby water molecules. The other atoms of the cluster were still included in the calculation at the B3LYP level of theory, but were considered at the much smaller 3-21G basis set [14].

3. Results and discussion

Since the smaller $\langle a\,2b\,c\rangle$ supercell was adopted in the present work, the validity of this choice was evaluated by benchmarking with previous calculations using a $\langle 2a\,2b\,2c\rangle$ supercell approach [1]. The $\langle a\,2b\,c\rangle$ optimized structure was compared with the part of the $\langle 2a\,2b\,2c\rangle$ supercell matching the size of the smaller supercell and containing the radical site. This resulted in an RMSD of 0.009 Å, indicating that both structures are essentially similar. EPR properties were also calculated for the optimized $\langle a\,2b\,c\rangle$ supercell and compared with the results obtained earlier. An overview of calculated as well as measured EPR properties is presented in Table 1. Comparing sections (b) and (c), it is clear that the supercell size reduction does not really have a noticeable effect on the calculated EPR properties of the alkoxy radical. A slight change is only noticed in the H4 isotropic coupling, which rises to 90.4 MHz and is now somewhat closer to the actual experimental value (see Table 1(a)). Yet, structurally and electronically it appears that a $\langle a\,2b\,c\rangle$ supercell is virtually identical to a $\langle 2a\,2b\,2c\rangle$ one, endorsing the chosen computational approach.

Exploring the geometrical changes of the radical and its environment during the 77 K molecular dynamics run, only small fluctuations from the statically optimized $\langle a\,2b\,c\rangle$ structure were registered. The mean 77 K geometry – obtained by averaging all 8000 geometries in the MD run – differs by only 0.008 Å from the static geometry. On the other hand, large variations are noticed for the EPR properties calculated at 200 snapshots along the trajectory. This is illustrated in Fig. 2, where

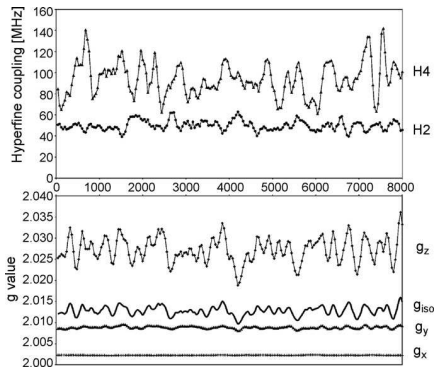


Fig. 2. Isotropic hyperfine couplings (in MHz) and (an)isotropic g tensor values as a function of time step (0.12 fs) in the molecular dynamics simulation. g_x, g_y and g_z refer to the minor, intermediate and maximum anisotropic g tensor components, respectively.

Table 1
Overview of calculated and measured EPR data

	A_{iso}/g_{iso}	A_{aniso}/g_{aniso}	Direction cosines versus $\langle a^* b\, c\rangle$			Ψ
(a) Experimental EPR properties[a]						
H2	39.2					
H4	112.1					
		2.0032	−0.020	−0.982	0.189	
g	2.0184	2.0064	−0.698	0.149	0.700	
		2.0456	−0.716	−0.118	−0.688	
(b) Static EPR calculation on optimized $\langle 2a\, 2b\, 2c\rangle$ supercell[b]						
		−3.5	0.692	−0.528	−0.492	
H2	50.4	−0.3	0.579	−0.001	0.815	
		3.8	0.431	0.849	−0.305	
		−5.9	0.443	0.881	0.163	
H4	87.1	−4.6	−0.209	−0.075	0.975	
		10.5	0.872	−0.466	0.151	
		2.0023	−0.316	−0.839	0.443	24
g	2.0125	2.0087	−0.547	0.542	0.637	25
		2.0263	−0.775	−0.041	−0.631	6
(c) Static EPR calculation on optimized $\langle a\, 2b\, c\rangle$ supercell						
		−3.4	0.695	−0.523	−0.494	
H2	49.6	−0.2	0.549	−0.059	0.834	
		3.6	0.465	0.851	−0.246	
		−6.1	0.448	0.882	0.146	
H4	90.4	−4.7	−0.198	−0.061	0.978	
		10.8	0.872	−0.468	0.148	
		2.0023	−0.314	−0.872	0.375	21
g	2.0126	2.0088	−0.572	0.489	0.659	21
		2.0267	−0.758	−0.008	−0.652	7
(d) Dynamic EPR calculation on $\langle a\, 2b\, c\rangle$ supercell at 77 K						
		−3.4	0.692	−0.526	−0.495	
H2	49.9	−0.2	0.551	−0.058	0.832	
		3.6	0.467	0.848	−0.250	
		−6.2	0.454	0.884	0.108	
H4	94.0	−4.7	−0.184	−0.025	0.983	
		10.9	0.872	−0.466	0.152	
		2.0023	−0.308	−0.877	0.369	20
g	2.0128	2.0089	−0.581	0.480	0.657	20
		2.0272	−0.753	−0.012	−0.658	7
(e) Static EPR calculation on optimized cluster[c]						
		−3.0	0.705	−0.521	−0.481	
H2	40.5	−0.3	0.488	−0.136	0.862	
		3.3	0.515	0.843	−0.159	
		−6.5	0.470	0.870	−0.151	
H4	100.4	−4.4	−0.152	0.249	0.956	
		11.0	0.870	−0.426	0.249	
		2.0022	−0.251	−0.885	0.392	19
g	2.0189	2.0090	−0.690	0.448	0.569	19
		2.0456	−0.679	−0.128	−0.723	3

Hyperfine couplings are presented in MHz. The Ψ angle (in degrees) indexes the deviation in orientation between the calculated eigenvectors and their experimental counterparts.
[a] Ref. [2].
[b] Ref. [1].
[c] Ref. [15].

the isotropic hyperfine and (anisotropic) *g* tensor values are plotted as a function of time step in the molecular dynamics simulation. Especially the H4 hyperfine coupling and the g_z (major) anisotropic *g* tensor component vary significantly. This reveals that the EPR properties of the rhamnose alkoxy radical are quite sensitive to even minute changes in the geometry.

Based on the 200 samples of the EPR properties along the trajectory, it is now possible to calculate the thermal average of these properties. For isotropic values, this can simply be performed by averaging the results of the 200 snapshot calculations. To determine the anisotropic components and the corresponding eigenvectors, however, it is necessary to first calculate the mean, non-diagonalized *g*- or hyperfine matrices and then diagonalize

this matrix. The resulting dynamic EPR properties are presented in Table 1(d). Overall, the eigenvectors and anisotropic hyperfine couplings are least affected by the dynamics, since they are nearly identical to those in Table 1(c). For the H4 isotropic hyperfine coupling, the isotropic g value and the maximum anisotropic g component (g_z), the differences are small, but discernible: the H4 coupling increases to 94 MHz and g_z rises by about 450 ppm (to 2.0272). Although these changes do seem to enhance the agreement between theory and experiment, the effect is only minor and too small to classify it as an improvement. In other words, the temperature effect does not really account for the remaining discrepancies between theoretical and experimental EPR properties of the RO4 alkoxy radical. Of course, the time scale over which the thermal average has been taken (∼1 ps) is relatively limited and the effect of molecular motions with longer periods is not accounted for. One could improve this by increasing the time step and/or total simulation time in the MD, although calculations at the nanosecond scale are prohibitively expensive at the ab-initio level and still far below the typical time scale of the CW-EPR experiment (μs).

Apart from the previous conclusions, it is worth noticing that the calculated EPR properties for rhamnose, reported earlier by some of the authors [15], are in much better agreement with experiment than the values determined in the current MD simulation. In that study, the RO4 alkoxy radical has been investigated within a cluster approach. For comparison, these results are summarized in Table 1(e). Although large similarities exist in the reproduction of the eigenvectors and anisotropic couplings, the isotropic hyperfine splittings vary substantially (almost by 9 MHz as compared to the dynamic calculation) and are in closer agreement with experiment. Even more spectacularly, the g_z anisotropic component exactly equals the experimental value of 2.0456! And yet, from a theoretical point of view, several aspects in the earlier approach were essentially inferior to the methodology adopted in the current work:

(i) Because a cluster approach was followed for the geometry optimization, the molecular environment of the radical was finite, contrary to the current periodic approach. But more importantly, constraints were imposed during optimization on all atoms except those of the central radical in the cluster. This effectively prevented any relaxation in the molecular environment of the radical away from the ideal crystal structure. Nevertheless, the resulting cluster optimized geometry is very similar to the structure of the $\langle a\,2b\,c\rangle$ supercell optimization in the present work. The RMSD with the cluster cut out of the periodic calculation – which was used for the calculation of the EPR properties – amounts only to 0.005 Å. Hence, since the same computational protocol was used in [15] and the current work to determine the hyperfine coupling constants (B3LYP functional with 6-311G** basis set for all atoms of the cluster), it must be concluded that the differences in (isotropic) hyperfine couplings between Table 1(c) and (e) are solely due to slight geometrical changes. This is consistent with the observation that in Fig. 2, the EPR properties can vary significantly to even minute changes in the geometry (see above).

(ii) In the previous work [15], a different computational protocol was used for the calculation of the g tensor. A 'single molecule' approach was followed and g tensor properties were calculated solely on the radical itself, neglecting any interaction with its molecular environment. Although this approach seemed to work amazingly well in the case of the g_z component, in retrospect it is probable that it did so for the wrong reasons. A g tensor calculation of a single rhamnose alkoxy radical, without taking into account the molecular environment of the radical, showed in principle the same values for the RO4 and RHop radical variants. The incorporation of surrounding molecules, as in reference [1], yields g tensors that are distinct for these two species, although in worse quantitative agreement with experiment. Hence, it is conceivable that error cancellation effects in the single molecule calculations have coincidentally resulted in an identical reproduction of the measured g_z anisotropic g tensor component. But, for the moment it remains unclear what the real origin is of the apparent discrepancy between the current, time-averaged g tensor calculations and the experiment. This paper shows that temperature effects are too moderate to remove this failure.

Regardless, the molecular dynamics simulation reveals that some large variations of the relevant spectroscopic properties take place (see Fig. 2). The H4 coupling may vary from 60 to 140 MHz and the experimental value of 112 MHz belongs to that interval. These variations are the result of geometry changes generated by the MD simulations. It could be interesting to search for a correlation between the calculated EPR properties and geometry. For this purpose, linear regression analyses were performed with the main EPR properties as dependent variables (g_z value, H2 and H4 isotropic couplings), and all internal coordinates of the radical and nearby water molecules as explanatory variables (27 bond lengths, 26 bond angles and 25 torsional angles). This is actually a limited set of structural parameters since other coordinates could be chosen involving more molecules of the supercell. Nevertheless, the subset proved more than sufficient and, moreover, a variance of up to 0.8 was readily obtained for all spectroscopic properties by including only eight internal coordinates in the regression. However, it is difficult to understand the variation of the hyperfine or g tensor values in terms of all these parameters from a chemical point of view. Hence, the parameter set was further reduced in an attempt to explain the variance of the EPR properties by considering only one or two of the dominant predictor variables. Fig. 3 plots the variance of the g_z, H2 and H4 isotropic coupling values with respect to the selected variables.

(i) g_z value

The variation of g_z was found to mainly depend on two geometric coordinates: the C3–C4 bond length in the alkoxy radical, and the length of a hydrogen bond between oxygen O4 and one of the protons of a crystal water. A lin-

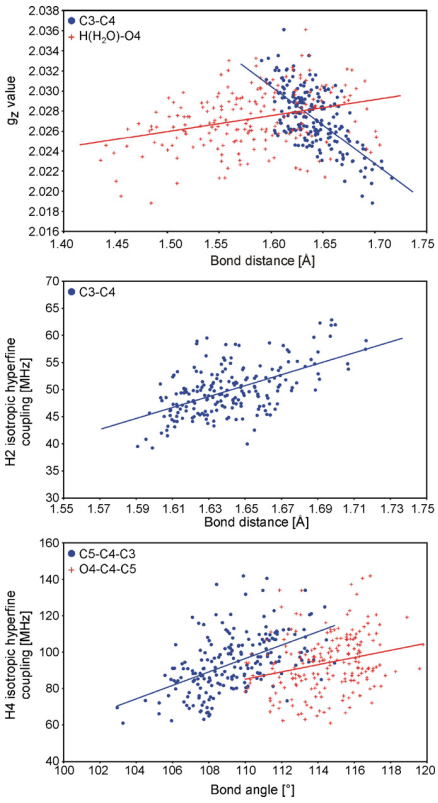


Fig. 3. Dependence of g_z and isotropic H2/H4 coupling values with respect to several internal coordinates within the rhamnose alkoxy radical. H(H_2O) refers to one of the protons within the ⟨a 2b c⟩ supercell that is hydrogen bound to O4.

ear regression model including only these two explanatory variables already accounted for 56% of g_z variance. As is apparent in Fig. 3, the inverse correlation with C3–C4 is quite strong (−0.67), which can be attributed to the existence of resonance in both RHop and RO4 variants of the rhamnose alkoxy radical. In the alternative resonance structure, the unpaired electron is mainly localized on C3 instead of on O4, as is illustrated in the bottom of Fig. 1. The larger the C3–C4 bond length, the more this resonance structure will contribute and the less spin density will be localized on O4, giving rise to a decrease in g_z. A similar interpretation is valid for the H(H_2O)–O4 hydrogen bond length. As this hydrogen bond elongates, the spin density will become more isolated on O4 and g_z will increase.

(ii) H2

C3–C4 bond elongation also explains the variation in the H2 hyperfine coupling to a large extent (correlation 0.57). This is immediately clear when considering the resonance structure in Fig. 1. The more spin density is localized on C3, the larger the isotropic H2 hyperfine coupling will be. In several previous works [1,3] this relation was already suggested.

(iii) H4

The variation of the isotropic H4 coupling proved quite difficult to interpret with only a limited number of structural parameters. The consideration of the two most important internal coordinates led to an explained variance of only 0.38. Apparently, the coupling is sensitive to changes in the bond angles of the rhamnose pyranose ring (C5–C4–C3) and the relative position of the C4–O4 axis with respect to the ring (reflected by the O4–C4–C5 bond angle). The H4 coupling does not depend on the C3–C4 bond distance, however. For both resonance structures of Fig. 1, H4 remains in a beta position with respect to the main site of unpaired spin density.

The importance of resonance in the rhamnose alkoxy radical is further exemplified in a vibrational analysis of the periodic molecular system. In a molecular dynamics simulation, atomic positions as well as velocities are determined. By using the latter and calculating the Fourier transform (FFT) of the velocity autocorrelation function, a mass-weighted power spectrum can be obtained yielding vibrational frequencies. The resulting spectra at 77 K are shown in Fig. 4 for the entire supercell (dashed line) and only for velocity components of atoms belonging to the radical (solid line). In principle, these spectra could be compared with infrared measurements. However, because changes in dipole moment were not determined throughout the MD simulation, the computed intensities are not really meaningful (they could well be IR-silent for instance). Nevertheless the calculated spectra do reveal what kind of vibrational modes are present in the system. Particular attention was paid to modes that only occur

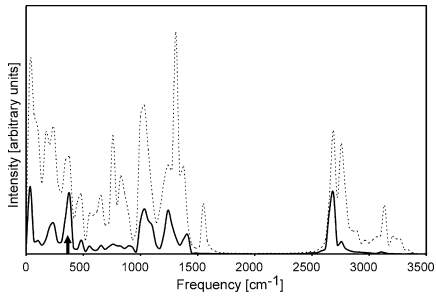


Fig. 4. Vibrational frequency spectrum for the radical (black line) and all atoms of the ⟨a 2b c⟩ supercell (dashed line).

in the radical and not in the 'undamaged' rhamnose molecules within the supercell, and vice versa. It is clear from the figure that the spectrum for the entire supercell reveals several peaks (e.g. at 1550, 3135 cm^{-1}) that are not present in the spectrum of the radical. These modes could be attributed to the relative movement of crystal water. The majority of peaks in the spectrum for the radical are also present in the spectrum for the entire supercell. Only one mode is predominantly due to vibrations of the radical: the peak at 378 cm^{-1} (indicated by an arrow in Fig. 4). Structural analysis reveals that it specifically involves strong C3–C4 bond length vibrations, combined with less intense rotational modes of the alkoxy pyranose ring. Hence, one could state that for this low vibrational mode, the RO4 rhamnose alkoxy radical rapidly switches between both resonance structures as depicted in Fig. 1.

4. Conclusions

In this article, a temperature simulation of the RO4 variant of the rhamnose alkoxy radical was conducted with the aid of ab-initio molecular dynamics. It was examined whether these temperature simulations could account for the residual difference between theoretical and experimental spectroscopic properties. At several points along the MD trajectory, g and hyperfine tensors were calculated and averaged, resulting in temperature dependent mean values. The effect of including temperature was found to be within computational error and no improvement was obtained between theory and experiment.

Additionally, the dynamics of the alkoxy radical were further explored. Using linear regression models, the variance of the g_z value, H2 and H4 hyperfine couplings was considered in relation to several geometric internal coordinates of the radical structure. It was found that g_z and H2 are considerably affected by the C3–C4 bond distance. This could be ascribed to the existence of a resonance structure of the alkoxy radical, through which spin density gets delocalized on the C3 carbon center. In a vibrational analysis of the molecular dynamics, a low frequency mode was furthermore identified with C3–C4 bond vibration as main characteristic. Apparently, the structure of the RO4 rhamnose alkoxy radical is alternating between resonance structures at a finite temperature.

Acknowledgements

This work is supported by the Fund for Scientific Research – Flanders (FWO). The authors kindly acknowledge Dr. R. Raymaekers for assisting with the linear regression analyses.

References

[1] E. Pauwels, R. Declerck, V. Van Speybroeck, M. Waroquier, Radiat. Res, in press.
[2] P.O. Samskog, A. Lund, Chem. Phys. Lett. 75 (1980) 525–527.
[3] E.E. Budzinski, H.C. Box, J. Chem. Phys. 82 (1985) 3487–3490.
[4] E. Sagstuen, M. Lindgren, A. Lund, Radiat. Res. 128 (1991) 235–242.
[5] C.J.T. de Grotthuss, Ann. Chim. 58 (1806) 54–74;
C.J.T. de Grotthuss, Biochim. Biophys. Acta 1757 (2006) 871–875 (English translation by R. Pomès);
N. Agmon, Chem. Phys. Lett. 244 (1995) 456–462.
[6] S. Takagi, G.A. Jeffrey, Acta Cryst. B 34 (1978) 2551–2555.
[7] P. Car, M. Parrinello, Phys. Rev. Lett. 55 (1985) 2471–2474.
[8] CPMD V3.11 Copyright IBM Corp. 1990–2006, Copyright MPI fuer Festkoerperforschung Stuttgart 1997–2001.
[9] J.P. Perdew, Phys. Rev. B 33 (1986) 8822–8824;
A.D. Becke, J. Chem. Phys. 96 (1992) 2155–2160.
[10] D. Vanderbilt, Phys. Rev. B 41 (1990) 7892–7895.
[11] Gaussian 03, Revision B.03, M.J. Frisch, G.W. Trucks, H.B. Schlegel, G.E. Scuseria, M.A. Robb, J.R. Cheeseman, J.A. Montgomery, Jr., T. Vreven, K.N. Kudin, J.C. Burant, J.M. Millam, S.S. Iyengar, J. Tomasi, V. Barone, B. Mennucci, M. Cossi, G. Scalmani, N. Rega, G.A. Petersson, H. Nakatsuji, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, M. Klene, X. Li, J. E. Knox, H.P. Hratchian, J.B. Cross, C. Adamo, J. Jaramillo, R. Gomperts, R.E. Stratmann, O. Yazyev, A. J. Austin, R. Cammi, C. Pomelli, J.W. Ochterski, P.Y. Ayala, K. Morokuma, G.A. Voth, P. Salvador, J.J. Dannenberg, V.G. Zakrzewski, S. Dapprich, A.D. Daniels, M.C. Strain, O. Farkas, D.K. Malick, A.D. Rabuck, K. Raghavachari, J.B. Foresman, J.V. Ortiz, Q. Cui, A.G. Baboul, S. Clifford, J. Cioslowski, B.B. Stefanov, G. Liu, A. Liashenko, P. Piskorz, I. Komaromi, R.L. Martin, D.J. Fox, T. Keith, M.A. Al-Laham, C.Y. Peng, A. Nanayakkara, M. Challacombe, P.M.W. Gill, B. Johnson, W. Chen, M.W. Wong, C. Gonzalez, J.A. Pople, Gaussian, Inc., Wallingford, CT, 2004.
[12] A.D. Becke, J. Chem. Phys. 104 (1996) 1040–1046.
[13] R. Krishnan, J.S. Binkley, R. Seeger, J.A. Pople, J. Chem. Phys. 72 (1980) 650–654;
A.D. McLean, G.S. Chandler, J. Chem. Phys. 72 (1980) 5639–5648.
[14] J.S. Binkley, J.A. Pople, W.J. Hehre, J. Am. Chem. Soc. 102 (1980) 939–947;
M.S. Gordon, J.S. Binkley, J.A. Pople, W.J. Pietro, W.J. Hehre, J. Am. Chem. Soc. 104 (1982) 2797–2803.
[15] E. Pauwels, V. Van Speybroeck, M. Waroquier, J. Phys. Chem. A 110 (2006) 6504–6513.

Paper 12: "Temperature Study of a Glycine Radical in the Solid State Adopting a DFT Periodic Approach: Vibrational Analysis and Comparison with EPR Experiments"

Ewald Pauwels, Toon Verstraelen, Hendrik De Cooman, Veronique Van Speybroeck, Michel Waroquier

Journal of Physical Chemistry B, **2008**, 112, 7618 - 7630

7618 *J. Phys. Chem. B* **2008,** *112,* 7618–7630

Temperature Study of a Glycine Radical in the Solid State Adopting a DFT Periodic Approach: Vibrational Analysis and Comparison with EPR Experiments

Ewald Pauwels,* Toon Verstraelen, Hendrik De Cooman, Veronique Van Speybroeck, and Michel Waroquier

Center for Molecular Modeling, Ghent University, Proeftuinstraat 86, B-9000 Gent, Belgium

Received: December 21, 2007; Revised Manuscript Received: March 20, 2008

The major radiation-induced radical in crystalline glycine is examined using DFT calculations, in which both molecular environment and temperature are accounted for. This is achieved by molecular dynamics simulations of the radical embedded in a supercell under periodic boundary conditions. At 100 and 300 K, a vibrational analysis is performed based on Fourier transformation of the atomic velocity autocorrelation functions. By the use of a novel band-pass filtering approach, several vibrational modes are identified and associated with experimental infrared and Raman assignments. Decomposition of the calculated spectra in terms of radical motion reveals that several vibrational modes are unique to the radical, the most prominent one at 702 cm^{-1} corresponding to out-of-plane motion of the paramagnetic center, inversely coupled with similar motion of the carboxyl carbon. A hybrid periodic/cluster scheme is used to evaluate the EPR properties of the glycine radical along the MD trajectories resulting in temperature dependent magnetic properties. These are compared with available experimental data conducted at 77 K and room temperature. Ground state or low temperature calculations yield very good agreement with 77 K experimental EPR properties. From the 300 K simulations, an important improvement is achieved on the isotropic hyperfine coupling of the ^{13}C tensor, which becomes closer to the value measured at room temperature. It is established that this is the result of a nonlinear relation between the planarity of the radical center and the isotropic couplings of the nuclei bound to it. Finally, a critical reevaluation of the experimental ^{14}N hyperfine tensor data strongly suggests that an erroneous tensor was reported in literature. It is convincingly shown that from the same experimental data set a different tensor can be derived, which is in substantially better agreement with all calculations.

1. Introduction

The radiation-induced radicals of biomolecules have been the subject of numerous investigations, due to their impact and often nocuous effect in fundamental biochemical processes. Both from theoretical and experimental perspective, efforts have been made to gain a better understanding of how they are generated. Particularly vital in this respect is to identify the precise structure of the induced radicals. Spectroscopic techniques based on electron paramagnetic resonance (EPR)[1] accurately probe the structure, but sometimes it remains difficult to deduce a radical model solely based on a set of spectroscopic properties. In recent years, EPR experiments are therefore often complemented by ab initio calculations based on density functional theory (DFT).[2] These methods allow the explicit optimization of a radical structure and the calculation of its EPR properties, [3–5] which can then serve to verify the experimental models and assumptions. Typically, static calculations are performed on a computational model space that only accounts for the radical itself (isolated molecule approach). From a methodological point of view, this amounts to simulating the radical in vacuo at 0 K. In reality, of course, the radical is surrounded by a molecular environment (e.g., a solvent or a crystal), and experimental measurements on this system occur at a finite temperature.

In this work, a series of calculations is presented on one of the radiation-induced radicals of crystalline α-glycine, in which both the molecular environment and temperature are accounted for. This is achieved by doing molecular dynamics simulations of the glycine radical embedded within the crystal lattice under periodic boundary conditions. Since the structure and spectroscopic properties of the radical are quite well-known from several EPR experiments, the system is ideally suited to evaluate the effect of explicit temperature treatment on the determination of magnetic properties in molecular crystals.

The radical species under study, $^{+}NH_3-{}^{\bullet}CH-CO_2^{-}$, is the major component in the EPR spectrum of crystalline glycine irradiated at room temperature. It is a fairly simple hydrogen abstraction product and is thought to arise from a sequence of intra- and intermolecular rearrangements following reduction or oxidation.[6] First identified in 1959,[7] its structure was later verified in several other studies [8–10] on single crystals, the most recent encompassing an elaborate range of EPR-derived techniques.[11] Several of these works report on the temperature-dependence of the EPR properties for this radical; at low temperatures the individual proton hyperfine couplings of the amino group are observable, whereas they are averaged out at room temperature, supposedly due to a rotational averaging motion of the amino group about the C−N axis.[6,9]

This glycine radical has also been the subject of several theoretical studies. Barone et al.,[12] Ban et al., [13] and Rega et al.[14] initially performed isolated molecule calculations on this species. To prevent intramolecular proton transfer within the vacuum from the amino group to one of the oxygen atoms, constraints were imposed in the first study.[12] In the latter two papers,[13,14] intermolecular interactions of the glycine radical with its environment were implicitly considered by adopting solvent

* To whom correspondence should be addressed. E-mail: ewald.pauwels@ugent.be.

10.1021/jp711997y CCC: $40.75
Published on Web 05/30/2008

simulation models surrounding the radical with a uniform dielectric continuum[15] or with more advanced PCM models,[16] respectively. The effect of explicitly incorporating the molecular environment of the radical in the computational model was recently examined by the authors in a comparative study between isolated molecule, cluster, and periodic approaches.[17] In a cluster (or supermolecule) method, part of the crystal lattice is modeled by placing discrete molecules around the target radical, in accordance with the crystal structure. A periodic approach, on the other hand, places the radical within a complete crystalline environment, with periodic boundary conditions accounting for the periodicity of the crystal. In that study, it was established that the explicit inclusion of the environment (using either approach) can have a considerable impact not only on the determination of the radical geometry but also on the calculated EPR properties.

Up till now, all theoretical studies of the glycine radical were mostly restricted to a static treatment of the system. In the works by Barone et al.[12] and Rega et al.,[14] temperature effects on the EPR properties of the radical were to some extent accounted for by including vibrational averaging effects within a perturbational approach. This method, however, is less generally applicable as it can only be employed when a single large amplitude motion dominates the dynamics of the molecular system, and this is hardly the case in a cluster or periodic model, where several other molecular species are present beside the radical. The most natural way of improving on this point is to explicitly consider the dynamics of the system by performing ab initio molecular dynamics simulations. A computationally feasible approach uses the Car−Parrinello approach.[18] Within the canonical (NVT) regime, this method allows the rapid sampling of radical conformations within the crystal lattice at a finite temperature. In literature, this methodology has been used to examine both paramagnetic substances (e.g., ref 19) and crystals (e.g., ref 20).

Atomic velocities that are obtained from the molecular dynamics give the opportunity to calculate vibrational spectra, which can be compared with experimental data from infrared (IR) or Raman spectroscopy. By decomposition, the spectrum due to the radical can be separated from the undamaged molecules in the periodic cell of the calculation. This offers the possibility to examine the vibrational degrees of freedom of the radical in contrast with those of the undamaged molecules. In a number of recent experimental studies, similar comparisons have been made with the aid of Fourier transform IR or Raman spectroscopy. Discriminant analysis was used to evaluate the effect of γ-irradiation damage on carbohydrates[21] and lipids[22] and also to assess the cellular damage in bacteria and human cells.[23] Even though this method does not provide as much information as, for instance, EPR experiments on the chemical structure of the radical, it is a viable alternative to establish the irradiation dose response of molecular systems.

2. Computational Details

α-glycine crystals have $P2_1/n$ space group symmetry and four glycine molecules make up the monoclinic unit cell. From X-ray diffraction studies,[24] the lattice constants have been determined as $a = 5.087$ Å, $b = 11.773$ Å, $c = 5.460$ Å with $\beta = 111.99°$. Starting from the crystallographic data, a supercell was constructed by doubling the unit cell in the a and c directions. One of the resulting 16 molecules was then transformed into a radical by removing one of the hydrogens on C_α. The cell obtained in this way is illustrated in Figure 1. This supercell approach is essential for a proper simulation of the glycine radical in the crystal lattice within a periodic approach, since it ensures that radical defects in adjacent cells are well separated from each other. The necessity and validity of such an approach was extensively shown in our previous paper.[17]

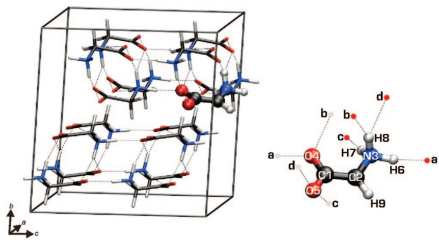


Figure 1. The glycine radical within the periodic supercell and atom numbering scheme. Dashed lines and labels indicate the different hydrogen bond interactions.

All periodic calculations were performed using the ab initio Car−Parrinello approach[18] as implemented in the CPMD software package.[25] Periodic boundary conditions were applied on the above-mentioned supercell (with constant cell dimensions) but without constraints on the individual atoms. This is in contrast with our previous study, where all but the atoms of the radical were constrained at their crystallographic positions.[17] The gradient-corrected BP86 density functional[26] was used with a plane wave basis set (cutoff 25.0 Ry) and ultra soft pseudopotentials of the Vanderbilt type.[27] First, the minimum of the supercell was determined in a (static) geometry optimization, followed by a number of constrained optimizations to investigate several degrees of freedom in the glycine radical. Only for the amino group rotation, several isolated molecule calculations were performed in addition to the periodic ones. From the constrained supercell conformations, the radical was selected and placed within an empty periodic box, $(10\ \text{Å})^3$ in size. Energies were then calculated for these isolated[28] systems using the Martyna and Tuckerman Poisson solver.[29]

A series of molecular dynamics simulations were initiated starting from the completely optimized supercell structure, with temperatures of approximately 100 and 300 K. In both cases, initial calculations were performed to equilibrate the temperature of the system by rescaling ionic velocities whenever the instantaneous temperature differed more than 50 K from the target temperature. Subsequently, Nosé-Hoover thermostats were activated for the ionic and electronic systems. A characteristic frequency of 3000 cm^{-1} was chosen for the ionic thermostat, with a target temperature of either 100 or 300 K. The electron thermostat was set to a frequency of 10 000 cm^{-1}, and an average target kinetic energy of 0.008 au (at 100 K) and 0.028 au (at 300 K) were determined for the electronic system from separate NVE molecular dynamics simulations. The MD time step was 5 au (0.12 fs), and the fictitious electronic mass was set to 400 au. Subsequent to the equilibration, production runs of 20 000 time steps were obtained, giving a total simulation time of 2.4 ps each. Additionally, an MD simulation was run with an ionic thermostat at 500 K (with 0.048 au as target kinetic energy for the electronic thermostat). Although 20 000 time steps were also collected for this calculation, the system was not sufficiently equilibrated. As such, the calculation was not taken up in most of the analyses.

When calculating the EPR properties of a radical embedded within a solid-state matrix, it is imperative that its molecular

TABLE 1: Overview of Selected Geometrical Features of the Glycine Radical in Comparison with Crystal Structure Data[a]

		static		dynamic	
	crystal structure (undamaged)	constrained	optimized	average (100 K)	average (300 K)
bond lengths					
C1−C2	1.527	1.450	1.466	1.467	1.471
C1−O4	1.257	1.290	1.283	1.284	1.288
C1−O5	1.259	1.291	1.287	1.286	1.286
C2−N3	1.482	1.437	1.441	1.443	1.446
C2−H9	1.080/1.090	1.083	1.088	1.089	1.093
N3−H6	1.036	1.076	1.081	1.082	1.087
N3−H7	1.024	1.078	1.081	1.082	1.082
N3−H8	1.024	1.051	1.053	1.054	1.053
dihedral angles					
O4−C1−C2−N3	19.5	15.2	15.9	13.6	10.3
O5−C1−C2−N3	−161.3	−163.6	−162.4	−164.8	−168.6
H6−N3−C2−C1	177.3	178.0	179.5	180.7	181.5
H7−N3−C2−C1	−60.6	−57.1	−56.8	−54.7	−54.2
H8−N3−C2−C1	57.9	59.3	59.9	61.5	62.2
improper torsion angles					
C2−O4−O5−C1	0.6	−0.8	−1.2	−1.0	−0.7
C1−N3−H9−C2	−37.5/36.0	1.5	3.5	2.9	2.1
hydrogen bonding distances					
O4-Ha	1.748	1.707	1.644	1.651	1.686
O4-Hb	2.387	2.417	2.507	2.417	2.336
O5-Hc	1.821	1.739	1.725	1.732	1.776
O5-Hd	2.040	2.088	1.958	2.020	2.225
H6-Oa	1.748	1.686	1.628	1.632	1.669
H7-Oc	1.821	1.685	1.632	1.661	1.713
H8-Od	2.040	2.091	1.933	2.090	2.203
H8-Ob	2.387	2.226	2.409	2.274	2.222

[a] The "optimized" and "constrained" data refer to static calculations from this work.[17] Bond lengths are given in Å, and angles are shown in degrees. The atomic numbering scheme is given in Figure 1.

environment is also included in the calculation. This can be achieved in a periodic scheme [30,31] or by adopting a cluster approach.[32] The latter approach was chosen in this work, requiring the construction of a cluster based on the atomic coordinates from the periodic supercell. For each snapshot or optimized structure, therefore, a cluster was cut out of the periodic system to contain the radical and all glycine molecules that are hydrogen bound to it (6 in total). Hyperfine tensors were calculated with the aid of the Gaussian03 software suite[33] using the B3LYP functional[34] and a 6−311G** basis set[35] for all atoms within the cluster. Although somewhat intricate, this hybrid periodic/cluster scheme has the benefit that EPR properties can be consistently determined with cluster methods while still maintaining the structural information obtained from a periodic approach. It has been successfully applied to examine radiation-induced processes in other molecular crystals.[36,37] Furthermore, the adopted cluster for the EPR calculation was also used in the previous study on glycine.[17]

The hybrid scheme was applied to calculate the EPR properties on the global minimum of the supercell, as determined from (static) geometry optimization. To examine the effect of the basis, only for this structure calculations were carried out with the EPR-III basis set[3] in addition to the 6−311G** set. The hybrid periodic/cluster scheme was also used to determine temperature-dependent hyperfine tensors for the molecular dynamics simulations. From the entire 100 and 300 K MD trajectories, one snapshot was taken every 50 time steps. For these 400 supercell geometries, the corresponding cluster model was constructed and EPR properties calculated (using the 6−311G** basis). Thermal averages were then determined for the hyperfine tensors along the trajectories. For isotropic values, this is merely the average of the 400 snapshot EPR calculations. For the anisotropic components, the nondiagonalized hyperfine matrices along the trajectory have to be averaged, after which diagonalization yields the correct thermal average of eigenvalues and eigenvectors. The sampling rate of 1/50 that was used for the calculation of the EPR properties corresponds to ∼5500 cm^{-1} and is therefore sufficiently high to incorporate all major molecular fluctuations.

3. Results and Discussion

Structural Aspects. As mentioned, the computational protocol is slightly altered from that in a previous work.[17] Constraints are no longer imposed on the atoms of the molecules surrounding the radical within the supercell. To assess the effect of this modification, first a comparison is made between optimized geometries using either method. In Table 1, selected structural properties of the glycine radical along with hydrogen bond lengths are presented for both methods (marked "constrained" and "optimized", respectively). The numbering scheme is presented in Figure 1. It is easily observed that the optimized conformation of the glycine radical is virtually independent of the methodology used; the bond distances and reported dihedrals hardly change. The glycine radical assumes a nearly planar conformation (represented by the improper dihedral angle $C_1-N_3-H_9-C_2$), in accordance with earlier calculations.[12−14,17] Furthermore (but not apparent from the table), the overall orientation of the glycine radical within the supercell is almost unaltered. This makes sense, as the method adopted earlier ("constrained" in the table) only constrained the radical environment but not the radical itself.

The hydrogen bond distances between the radical and the surrounding molecules, on the other hand, are altered because the geometries of those molecules have changed in the "optimized" calculation. When comparing the H-bond distances with those of the (undamaged) crystal structure (Table 1), it is clear that the "optimized" distances differ by some 0.1 Å on

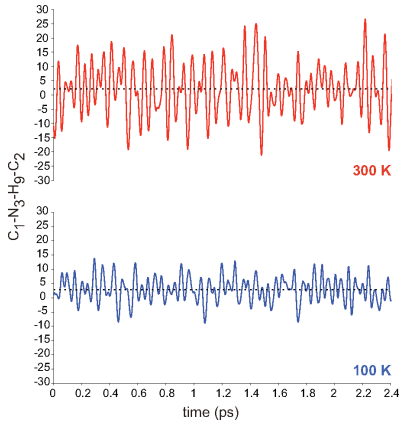


Figure 2. Time evolution of the radical planarity (indexed by the improper dihedral angle $C_1-N_3-H_9-C_2$) in the 100 and 300 K MD simulations. Dashed lines indicate the average conformation.

average. This is a larger difference than for the "constrained" geometry. Apart from O_4-H_b and H_8-O_b, which are formally van der Waals contacts, all hydrogen bonds in the "optimized" structure are shorter than those in the crystal balanced by somewhat longer N−H bonds. In other words, the attractive interaction between the oxygens and the amino hydrogens is overestimated to some extent, causing the hydrogens to slightly shift away from the nitrogens toward the oxygens. Most likely, this can be ascribed to the use of the BP86 density functional, which is known to overbind in some cases.[38] In the "constrained" calculations, this effect was also present but for the better part masked, due to the imposed constraints. Nevertheless, the H-bond lengths never appear to be spurious and overall resemble those of the intact crystal. Furthermore, since the geometry of the central radical is virtually unaltered, the computational methodology can be maintained in the context of the present work.

It is now interesting to examine several averaged structural features obtained from the 100 and 300 K dynamics trajectories, since the glycine radical and intact molecules in the periodic supercell are in fact quite flexible. Logically, atomic displacement amplitudes are considerably larger in the high temperature simulation. In Table 1, selected average distances and dihedral angles from the MD runs are presented. Again, the geometry of the radical barely changes with respect to the static optimization. Mean bond distances just slightly elongate with rising temperature (static → 100 K → 300 K), due to increased thermal excitation. On average, the glycine radical center maintains an essentially planar conformation. However, the variation in this property is much larger in the 300 K simulation. This is illustrated in Figure 2, where the planarity is plotted as a function of time. In the 100 K regime, the planarity varies between −10° and +15°, whereas it oscillates between −20° and +25° at 300 K. Dashed lines indicate the average planar conformation in both plots. There is a minor tendency toward a more pronounced planarity with rising temperature. Going from 0 K (static calculations) over 100 to 300 K, the CO_2 group rotates toward the plane of the radical (dihedrals $O_4-C_1-C_2-N_3$ and $O_5-C_1-C_2-N_3$), contributing to the overall planarity of the radical backbone. This is a result of the increased thermal freedom. The main effect of introducing temperature in the simulation is however not reflected in the radical conformation itself but rather in its immediate environment. The strong (short) hydrogen bonds, type a and c, moderately elongate with increasing temperature, but the weak (longer) H-bonds, type b and d, become virtually equal in length. The latter interactions extend throughout the crystal along the *a* axis, connecting every H_8 atom with O_4 and O_5 atoms of neighboring glycine molecules. In the static approach, interaction d is somewhat stronger than interaction b. But as the molecules become more and more thermally activated, vibrations overcome the initial distinction, rendering the interactions b and d virtually equivalent.

Vibrational Properties of the Supercell. For both the 100 and 300 K MD trajectories, vibrational spectra are obtained by Fourier transformation (FFT) of the atomic velocity autocorrelation functions:

$$F(\omega) = \sum_{\alpha} \int_{-\infty}^{+\infty} \langle \bar{v}_\alpha(t) \cdot \bar{v}_\alpha(0) \rangle e^{-i\omega t} dt \qquad (1)$$

where α runs over all atoms of the supercell. The power spectra for all atoms of the supercell are shown in Figure 3. To unambiguously discern the vibrations due to the radical, separate power spectra are given in the top of the figure, calculated by restraining the sum over α in the above expression to the atoms of the radical only.

The overall shape of the supercell vibrational spectra is comparable with experimental FT-IR[39] or Raman spectra [40,41] for (nonirradiated) solid-state glycine. Yet, for a closer comparison between theory and experiment it is preferable to consider the spectra in terms of the molecular vibrational modes that are associated with the frequency bands. The identification of these modes from MD spectra is, however, usually not straightforward. One approach is to decompose the spectrum in individual (atomic) contributions. However, this procedure is not able to extract the precise nature of the various vibrational modes. Calculating the power spectrum for particular internal coordinates (e.g., a bond distance) is another viable alternative, provided that the motion associated with the vibrational mode is not dependent on more than one internal coordinate. More systematic approaches have been developed to extract the normal modes from molecular dynamics, such as (among many others) principal mode analysis[42] or, more recently, through localization of Fourier transformed velocity time-correlation functions.[43]

In the present work, a more intuitive approach is adopted based on band-pass filtering. Frequencies within a certain range (bandwidth) are passed, while frequencies outside that range are attenuated. First, the Fourier transform spectra are determined for the Cartesian coordinate vectors of all atoms ($\bar{x}_\alpha(t)$) during the trajectory. This data is then passed through a band-pass filter centered on a vibrational frequency φ for which one wants to identify the normal mode motion. A bandwidth $\Delta\varphi$ is introduced to incorporate the whole signal. This is generated by a masking function $M^{\varphi,\Delta\varphi}(\omega)$, defined as:

$$M^{\varphi,\Delta\varphi}(\omega) = 1 \text{ for } \varphi - \frac{\Delta\varphi}{2} < |\omega| < \varphi + \frac{\Delta\varphi}{2} \qquad (2)$$

The masking function vanishes for all other values of the frequency except for $\omega = 0$ to keep the mean atomic positions in place. A reverse Fourier transform now yields a time-dependent trajectory but only contains the molecular/atomic motions that are associated with the frequency φ:

$$\bar{x}_\alpha^{\sigma,\Delta\varphi}(t) = \int_{-\infty}^{+\infty} S^\sigma(\omega) M^{\varphi,\Delta\varphi}(\omega) [\int_{-\infty}^{+\infty} \bar{x}_\alpha(t) e^{-i\omega t} \mathrm{d}t] e^{i\omega t} \mathrm{d}\omega \tag{3}$$

An optional scaling function $S^\sigma(\omega)$ is added to artificially enhance the molecular motion by a factor σ, facilitating identification of the normal mode:

$$S^\sigma(\omega) = 1 \text{ for } \omega = 0 \qquad \sigma \text{ else} \tag{4}$$

The method is demonstrated in Figure 4. A frequency window of width 30 cm^{-1} is selected, centered on the 702 cm^{-1} peak that is encountered in the 300 K velocity autocorrelation function. By applying the band-pass method on the coordinate trajectory, the associated normal mode vibration is easily identified from the obtained filtered trajectory, for which several snapshots are shown in the top of the figure.

This rather basic method is not generally applicable due to a number of limitations. First, because all filtered molecular motions are determined relative to the average geometry of the dynamics simulation, this geometry has to make sense from a physical point of view. The occurrence of (nearly) free rotors in the MD, for instance, would collapse the average structure of these rotating groups. Yet, this is not the case in the 100 or 300 K simulations, as it was already established earlier that the average structural parameters are meaningful. Second, the use of a nonzero bandwidth ($\Delta\varphi$) in the masking function restricts the accuracy by which vibrational modes can be distinguished. However, it is not the intention of this work to disentangle between all normal modes in the glycine vibrational spectra. The proposed method is merely used to make a fairly accurate characterization of the molecular vibrations that are associated with certain frequency bands.

In Table 2, an overview is presented of the vibrational modes that are identified from the MD simulations using this approach and a comparison is made with available experimental data.[41] With respect to the modes in the supercell vibrational spectra, the agreement between experiment and theory is quite reasonable; most prominent features in the spectra are successfully assigned and matched with measured Raman frequency bands. Some ambiguity remains in the identification of the NH_3 torsional modes (440−620 cm^{-1}) and it is difficult to unambiguously specify the nature of the normal modes that are associated with the intense peaks between 1200 and 1400 cm^{-1}. The low frequency modes below 300 cm^{-1} are not further examined, as molecular vibrations below this frequency are insufficiently sampled in an MD run of 2.4 ps (a 300 cm^{-1} vibration occurs only about 20 times in such a time span). All high-frequency stretch modes are red-shifted by some 200 cm^{-1} with respect to the experiment. This can been attributed to the use of Vanderbilt type pseudopotentials, which cause a softening of vibrational modes involving H atoms.[38] Nevertheless, the qualitative character of the high-frequency part of the spectrum is fairly well maintained. In particular, the difference in frequency between the symmetric and asymmetric CH_2 stretch modes is nicely reproduced. Slight frequency shifts are discerned between several modes of the 100 and 300 K MD spectra. This is the result of the essentially anharmonic nature of the

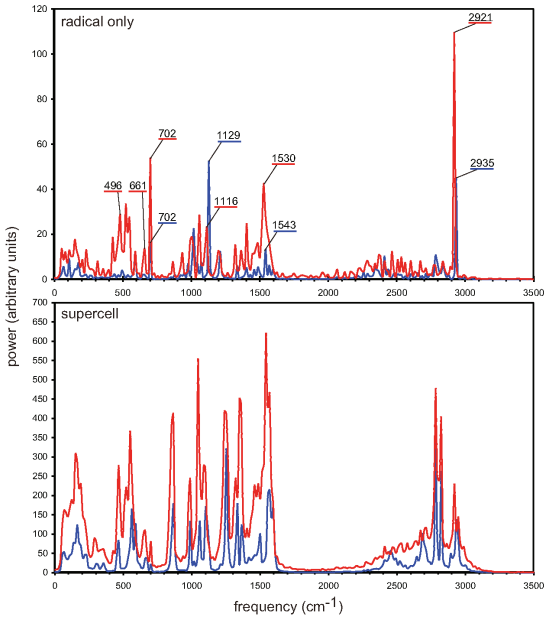


Figure 3. Vibrational frequency spectra as determined from velocity autocorrelation functions for the 100 K (blue) and 300 K (red) simulations. The lower spectra are determined for all atoms of the supercell and the upper spectra indicate the contribution of the radical.

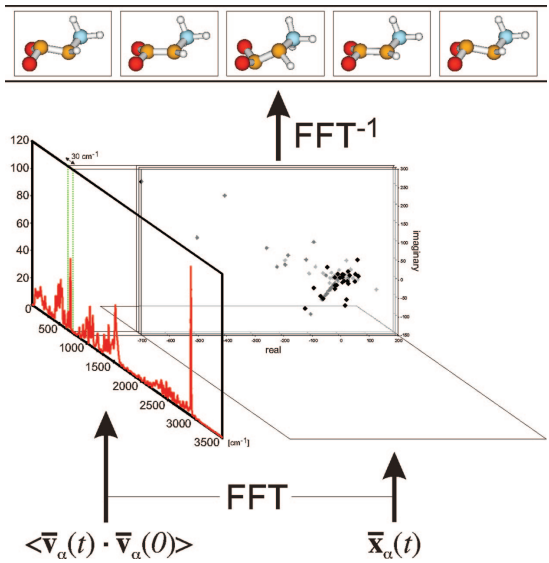


Figure 4. Illustration of the band-pass filtering method used in this work to make an approximate identification of the vibrational mode associated with a particular frequency.

TABLE 2: Characterized Vibrational Bands (in cm^{-1}) and Assignments for the 100 and 300 K Molecular Dynamic Simulations[a]

Supercell Modes			
experiment	100 K MD	300 K MD	assignment
491	440−620	440−620	NH_3 torsion
694	675	661	CO_2 bend
894	868	854	C−C stretch
1035	992	992	C−N stretch
1108, 1133	1061, 1102	1047, 1102	NH_3 rock
1574	1570	1543	NH_3 deformation
2976	2783	2783	CH_2 stretch (symmetric)
3008	2824	2824	CH_2 stretch (asymmetric)
3150	2935	2921	N−H stretch
	Characteristic Radical Modes		
		496	NH_3 torsion
		661	CO_2 bend
	702	702	C• out of plane
	1129	1116	NH_3 rock
	1543	1530	NH_3 deformation
	2935	2921	N−H and C•-H stretch

[a] Experimental data are taken from ref 41.

vibrational modes, which is more pronounced in the high temperature simulation.

Vibrational Properties of the Radical and Associated Energy Change. When comparing the supercell spectra with the vibrational spectra due to the radical only (top of Figure 3), it is clear that the latter feature several very intense peaks, some of which are unique and not present for the intact molecules of the supercell. At the high frequency end of the radical spectrum, the C•−H stretching mode cannot be disentangled from the N−H stretch in the radical. However, band-pass analysis clearly shows that the former motion accounts for most of the intensity. Even though this radical mode is slightly red-shifted from 2935 cm^{-1} in the 100 K simulation to 2921 cm^{-1} at 300 K, it always partly overlaps with the N−H stretching band of the undamaged glycine molecules in the supercell. As such, it accounts for roughly half the observed intensity in the supercell spectra. The peaks due to NH_3 rocking and deformation motion are even less distinguishable from the other molecules in the supercell. Furthermore, whereas these features are quite narrow at 100 K, they rapidly become broad and less distinguishable at a higher temperature. The most prominent feature of the radical spectra, though, is the peak at 702 cm^{-1}. This mode is particularly

intense (comparable to one-fifth of the intensity of the C−N stretching band of all molecules in the supercell) and never overlaps with other modes due to undamaged molecules in the supercell. On the basis of these calculations, the contrast and intensity for this radical mode seem sufficient to warrant discriminatory detection of glycine radicals with the aid of Raman or IR spectroscopy. However, to the best of our knowledge, no experimental studies have specifically examined glycine radicals so far. Hence, an independent experimental verification of the 702 cm^{-1} radical mode would be of particular scientific interest, given the importance of glycine radicals, e.g., in enzyme catalysis.[44]

Visualization of the 702 cm^{-1} normal mode reveals that it corresponds with inversion of the $C_2^{\bullet}$ radical center, inversely coupled with out-of-plane movement of C_1. The motion is illustrated in the top of Figure 4. In previous theoretical works on the isolated glycine radical,[12,14] out-of-plane displacements were already found to be important, but an inverse relation between C_1 and $C_2^{\bullet}$ was not established. This inversion mode gives rise to a degree of freedom that may affect the spectroscopic properties of the radical to a large degree, as will be discussed later. A graphic representation of the out-of-plane motion of $C_2^{\bullet}$ is given in Figure 5a, along with the associated energy potential. The latter is obtained by performing several constrained (static) geometry optimizations within the periodic supercell. At the ground state, the glycine radical is nearly planar and for maximum deviations of 20°, the energy change amounts to about 7 kJ/mol.

The NH_3 rotational vibration mode of the glycine radical is only clearly visible in the 300 K simulation, a broad band centered at 496 cm^{-1}. The corresponding energy profile for rotation of the amino group is given in Figure 5b. Since intermolecular interactions are duly accounted for at each point (obtained by constrained geometry optimizations within the periodic supercell), this potential (solid line) is quite asymmetric in shape and certainly not as smooth as would be expected from isolated molecule calculations on the radical. Rotation of the amino group is controlled by a substantial barrier of 45 kJ/mol, mainly caused by the hydrogen bonds between the amino protons and carboxy oxygens of the lattice. A Newman projection along the N_3-C_2 axis in Figure 5c reveals the orientations of the different groups at one of the three equivalent minima. It is clear that in any such minimum, the amino protons are oriented toward the oxygen atoms of the molecular environment. Yet, in the absence of the lattice, such an orientation would virtually lead to a maximum in energy! This is also substantiated in Figure 5b, where the rotational potential for the isolated radical conformations are plotted (dotted line). The periodic minima are shifted by some 60° with respect to the isolated molecule. At a value close to 0° for the $H-N_3-C_2-C_1$ dihedral angle, one amino proton can engage in an additional intramolecular hydrogen bond with the carboxy group, as corroborated in other isolated molecule studies.[12–14] Also, since the isolated molecule barrier height for rotation is one-half of that of the periodic calculations, it is evident that the 45 kJ/mol barrier for NH_3 rotation originates for ∼50% from intermolecular hydrogen bonding. A comparable barrier for rotation of the amino group (∼40 kJ/mol) was obtained in earlier calculations on alanine using a cluster approach.[45] The isolated molecule minimum in that work was also shifted by about 60° from the cluster minimum.

Because the NH_3 rotational potential is so steep and high, sufficient thermal energy is required to excite the corresponding rotational vibration mode and, eventually, surpass the barrier.

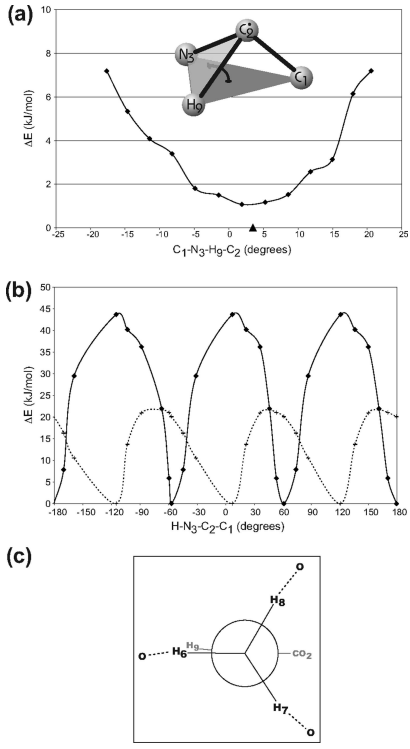


Figure 5. Energy change associated with main degrees of freedom in the glycine radical. (a) Potential associated with out-of-plane movement of the radical center embedded within the supercell. ▲ indicates the position of the global minimum (absolute energy −903.2763 au). (b) Energy profile for rotation of the amino group about the C_2-N_3 axis. The solid line (♦) is the result of several constrained geometry optimizations within the periodic supercell. The dotted line (+) is obtained by performing additional energy calculations within an isolated molecule approach on the periodic geometries (see computational details). (c) Newman projection along the N_3-C_2 axis indicating the orientations of the N−H bonds in the optimized geometry of the supercell (ground state). Dashed lines indicate hydrogen bonds.

When that happens, the amino protons undergo rotational exchange as the NH_3 group rotates over 120°. From EPR experiments, it was determined that the amino group of the radical should rotate freely at room temperature, whereas this motion was observed to be freezed out at temperatures below 100 K.[6,9] However, even though the 496 cm^{-1} rotational vibration mode is present at 300 K, rotational exchange of the amino protons in the radical never takes place during the 2.4 ps of the simulation. In fact, none of the amino groups in the supercell undergoes a 120° rotation at 100 or 300 K. This is apparent in Figure 6, where histograms display the number of conformations as a function of the rotation angle (for all amino groups in the supercell). For the 100 K trajectory (blue), isolated sharp peaks are obtained indicating that the NH_3 group vibrates moderately about the C_2-N_3 axis but never makes a transition

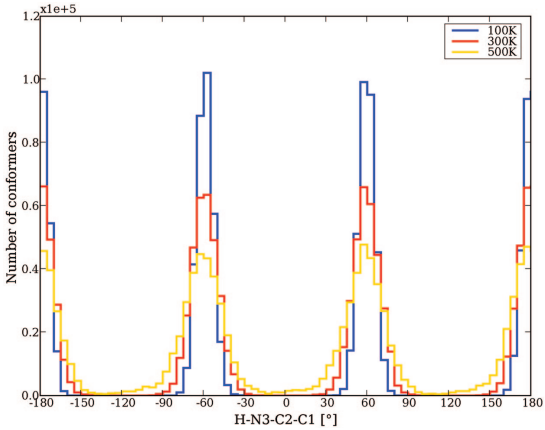


Figure 6. Histograms of the number of conformations with a certain H−N_3−C_2−C_1 dihedral angle for all amino groups in the supercell. The step of the histogram is 5°.

and rotates over 120°. Assuming that the potential for amino group rotation of an undamaged glycine molecule is similar to that established for the radical (Figure 5b), it is clear that only a very small part of the potential well is sampled at 100 K, accounting for the low intensities of this mode in the vibrational spectra. Hence, the amino protons (of glycine molecules and radical) remain distinguishable and merely vibrate about their average positions. But even at 300 K (red), rotational exchange of the amino protons never occurs. The histogram is broader, though, and conformational states that are 30° away from the minima can now be attained. This implies that the rotational potential in Figure 5b is excited for roughly 30 kJ/mol. It is not implausible that the remainder of the barrier is crossed as a result of tunneling effects, as was suggested in the original papers describing the EPR experiments.[6,9,11] These effects are, however, not taken into account in the simulations. An alternative explanation is that the calculated barrier for rotation is too high. With the aid of solid-state nuclear magnetic resonance (NMR) spectroscopy, this parameter was estimated to be 24 ± 2 kJ/mol for (not-irradiated) α-glycine crystals.[46] This is lower than what is determined from the calculations and might further indicate that the BP86 density functional overestimates the strength of the hydrogen bonds.[38] It is possible to promote rotational exchange of the amino groups by raising the temperature of the MD simulation, e.g., to 500 K. As can be seen in Figure 6 (yellow), NH_3 rotation does occur at this temperature for some of the glycine molecules in the supercell (but not for the radical). However, as was mentioned in the computational details, further analysis of the 500 K MD trajectory shows that the system is still not sufficiently equilibrated after 2.4 ps. Other methods of sampling the conformational space (e.g., constrained molecular dynamics or metadynamics [47]) are therefore more suitable to thoroughly examine the amino group rotation, but are beyond the scope of this work.

Temperature Dependence of EPR Properties. A summary of available experimental EPR data for the glycine radical is given in Table 3. From several studies (see ref 6 for an overview), altogether seven hyperfine coupling tensors are identified and assigned to nuclei within the glycine radical (the assignments in the table refer to the atom numbering scheme of Figure 1). At room temperature, one ^{13}C and a ^{14}N hyperfine tensor are found, along with two proton tensors. The H_N tensor is assigned to the three magnetically equivalent amino protons, ascribed to free rotation of the NH_3 group at 300 K. At lower temperatures, these protons are locked in position and can be determined independently. In the 1966 investigation by Collins and Whiffen (at 77 K),[9] all three tensors could be derived, whereas the tensor with the smallest coupling could not be unambiguously distinguished in the 1998 study by Sanderud and Sagstuen (at 100 K).[11] Since the $H_{exp}(\beta 2)$ and $H_{exp}(\beta 3)$ tensors slightly differ between these works, the theoretical EPR parameters of H_7 and H_8 are compared with the results of both experimental studies.

Previous EPR calculations on the structure of the glycine radical obtained within a static, periodic approach resulted in an already good agreement with experiment.[17] It could therefore be anticipated that the minor adjustments to the computational protocol for geometry optimization in the present work have no big impact on this experiment either. In Table 4, the results are presented of the EPR calculation on the optimized structure of the supercell (indicated static/6−311G**). As expected, the accordance between these new results and experiment is quite good, apart from some persisting errors, which were already reported earlier.[17] The ^{13}C and ^{14}N tensors, in particular, are unsatisfactory. In the former, the isotropic coupling is underestimated by 50 MHz but the eigenvectors are more or less aligned with the measurements. This is confirmed by Ψ values of about 20° or less, which indicate the angles (in degrees) between predicted and corresponding experimental eigenvectors. For the ^{14}N tensor, the situation is reversed; the isotropic coupling is well reproduced, but the eigenvectors are quite poorly reproduced. For both nuclei, the anisotropic eigenvalues of the hyperfine tensors are only reproduced on a qualitative level. Other differences between theory and experiment are less dramatic; the anisotropic eigenvalues of H_6 are not entirely consistent with those of $H_{exp}(\beta 2)$. For H_6 to H_9, on the other hand, the isotropic and anisotropic values as well as the eigenvectors are very well reproduced by the calculations; differences in A_{iso} amount to only 5 MHz, and Ψ is kept below 15°.

TABLE 3: Summary of Available Experimental EPR Data for the $^{+}NH_3-{}^{\bullet}CH-CO_2^{-}$ Glycine Radical[a]

	A_{iso}	T_{aniso}	principal axes vs $\langle a^*bc\rangle$			A_{iso}	T_{aniso}	principal axes vs $\langle a^*bc\rangle$			temperature	ref	assignment
C_{exp}	126.7	−90.0	0.902	−0.080	−0.424						300 K	*a*	C_2
		−36.7	0.360	−0.404	0.841								
		126.8	0.239	0.911	0.335								
N_{exp}	−8.7	−1.0	−0.412	−0.293	0.863						300 K	*b*	N_3
		−0.8	0.885	0.097	0.456								
		1.7	−0.217	0.951	0.219								
$H_{exp}(\beta 1)$	3.3	−7.3	0.620	−0.680	0.391						77 K	*c*	H_6
		−1.8	0.614	0.731	0.296								
		9.2	−0.488	0.057	0.871								
$H_{exp}(\beta 2)$	62.0	−6.0	0.005	−0.810	0.587	62.0	−6.4	−0.077	−0.976	0.206	77 K/100 K	*c/d*	H_7
		−3.2	0.146	0.581	0.801		−3.8	0.031	0.204	0.978			
		9.3	0.989	−0.082	−0.121		10.2	0.997	−0.081	−0.015			
$H_{exp}(\beta 3)$	82.0	−6.1	0.751	0.361	0.551	83.1	−7.1	0.725	0.387	0.570	77 K/100 K	*c/d*	H_8
		−4.0	−0.371	−0.460	0.807		−3.5	−0.386	−0.457	0.801			
		10.1	0.545	−0.811	−0.212		10.6	0.570	−0.801	−0.182			
H_N	49.1	−2.9	0.548	0.700	0.458						300 K	*d*	H_{6-8}
		−2.1	0.222	−0.649	0.727								
		5.0	0.807	−0.297	−0.511								
$H_{exp}(\alpha)$	−63.7	−33.8	−0.507	0.141	0.851						300 K	*d*	H_9
		1.9	0.330	0.943	0.040								
		31.9	0.797	−0.301	0.524								

[a] The assignment of the tensors has been elaborately discussed in refs 6 and 17. All principal axes (or eigenvectors) are given with respect to the $\langle a^*bc\rangle$ reference axis system (*a* = ref 8; *b* = ref 10; *c* = ref 9; *d* = ref 11).

TABLE 4: Overview of EPR Data Calculated using the Hybrid Periodic/Cluster Scheme as Described in the Computational Details[a]

	static						dynamic					
	6−311G**			EPR-III			100 K			300 K		
	A_{iso}	T_{aniso}	Ψ	A_{iso}	T_{aniso}	Ψ	A_{iso}	T_{aniso}	Ψ	A_{iso}	T_{aniso}	Ψ
C_2	73.4	−73.3	22	83.0	−76.4	22	78.7	−72.2	37	95.1	−70.8	32
		−71.9	10		−74.7	10		−67.5	31		−61.3	26
		145.1	22			151.1	22		139.8	21	132.0	20
N_3	−7.0	−1.7	80	−6.4	−1.7	88	−6.9	−1.5	68	−6.5	−1.4	57
		−1.3	81		−1.4	88		−1.2	70		−1.1	59
		3.0	26		3.1	26		2.8	24		2.5	23
H_6	0.5	−4.9	3	0.6	−4.9	2	2.4	−4.8	2	5.6	−4.7	1
		−4.1	10		−4.0	10		−4.2	9		−4.3	8
		9.0	10		8.9	10		9.0	8		9.0	7
H_7	62.4	−7.3	10/14	66.9	−7.4	14/11	59.8	−7.0	11/13	59.0	−6.5	9/15
		−3.6	11/14		−3.6	15/11		−3.7	13/13		−3.8	10/15
		10.9	6/1		10.9	7/1		10.6	6/2		10.3	5/4
H_8	84.9	−7.4	4/4	90.2	−7.4	0/2	85.5	−7.1	6/5	80.3	−6.7	8/8
		−3.9	6/5		−3.9	5/5		−3.9	6/5		−3.8	8/8
		11.3	5/5		11.4	5/5		11.0	2/2		10.5	1/1
H_9	−58.3	−35.0	5	−60.7	−34.8	6	−58.0	−34.5	4	−56.1	−33.3	4
		−3.1	10		−1.2	11		−2.1	8		−0.6	7
		38.1	9		36.0	10		36.6	7		34.0	6

[a] The dynamic results are obtained with the 6-311G** basis set only. Ψ indicates the angle (in degrees) between predicted and corresponding experimental eigenvectors (see Table 3). For H_7 and H_8, Ψ is determined with respect to the experimental data of Collins and Whiffen[9] and Sanderud and Sagstuen.[11]

As outlined in the computational details, thermal averages are also determined for the hyperfine tensors of glycine along the trajectories of the molecular dynamics simulations. The results are given in Table 4 for the 100 and 300 K MD runs. At the low temperature, the differences between the static and dynamic calculated EPR properties are only minimal. The main effect of introducing temperature becomes apparent in the 300 K simulation, where clearly the isotropic hyperfine coupling of ^{13}C increases by more than 20 MHz. This is an obvious improvement in the direction of the C_{exp} value (126.7 MHz) that was measured at room temperature. But also the anisotropic eigenvalues are better reproduced. The near-degeneracy of the minor and intermediate anisotropic coupling in the static calculation is gradually removed in the temperature simulations. The 300 K tensor has a more anisotropic shape, which better matches the shape of C_{exp}. The anisotropic parts of N_3 and H_9 improve in a similar way, although the adjustments are less pronounced. The individual H_6-H_8 proton couplings should be compared with the $H_{exp}(\beta 1)-H_{exp}(\beta 3)$ tensors, which were measured at a temperature of 77 or 100 K. The dynamic calculations are consistent in this respect, since the 100 K results agree much better with the experimental amino proton couplings than those of the 300 K simulation.

The improvement of the isotropic ^{13}C coupling in the 300 K simulation is the most notable effect in accounting for the temperature. Yet, the calculated value still differs from the experimental coupling by 30 MHz. This residual difference may partly be attributed to a basis set effect. It is well-known that

TABLE 5: Experimentally Determined ^{14}N Tensor as Reported by Hedberg and Ehrenberg in ref 10 and by a Reanalysis Done by the Authors Using an (a → −a) Symmetry Operation on the Original Tensor in the ⟨*abc⟩ Reference Axis System**

			principal axes			Ψ			
	A_{iso}	T_{aniso}	*a*	*b*	*c**	static 6−311G**	static EPR-III	dynamic 100 K	dynamic 300 K
N_{exp} (as reported)	−8.7	−1.0	−0.705	−0.293	0.646	80	88	68	57
		−0.8	0.650	0.097	0.754	81	88	70	59
		1.7	−0.283	0.951	0.121	26	26	24	23
N_{exp} (a → −a)	−8.7	−1.0	0.705	−0.293	0.646	26	19	35	45
		−0.7	−0.650	0.097	0.754	23	11	33	44
		1.7	0.283	0.951	0.121	18	18	16	14

isotropic hyperfine couplings depend on the local quality of the wave function at the nucleus, due to the presence of a Fermi contact term in its expression (see ref 5). Special basis sets have been constructed for a more accurate determination of this parameter, e.g., the EPR-III set by Barone.[3] Unfortunately, the recalculation of the thermal averages of the EPR properties in Table 4 (which are determined with the 6−311G** basis) would impose a considerable computational burden and are therefore not performed. Still, an estimate of this effect can be obtained by considering an additional EPR calculation on the static geometry, in which this enhanced basis set is applied. The results are presented in Table 4 (labeled static/EPR-III). With respect to the former static results, the C_2 coupling increases by 10 MHz and H_7, H_8 by about 7 MHz; H_9 reduces by 2 MHz. Summarizing, an extended basis set can cause a 5−10% change in the size of the coupling constant, leaving the eigenvectors unaffected. On the contrary, it does not necessarily improve the agreement with experiment for all nuclei; although the ^{13}C coupling rises in the direction of the experiment, H_7 and H_8 are now overestimated.

Irrespective of temperature, level of theory, or basis set influences, substantial differences in orientation persist between the calculated eigenvectors of the ^{14}N tensor and those determined from experiments. Especially, the principal directions associated with the two smaller anisotropic hyperfine couplings are misaligned, as is apparent from the Ψ angles (Table 4), which are consistently higher than 57° for all calculations. However, the accuracy and reliability of this tensor are limited. A first indication is the discrepancy between various experimental studies in literature. Several works (unfortunately all older than 1968) report the ^{14}N hyperfine tensor with varying precision and completeness. Whereas the isotropic and anisotropic couplings are more or less comparable, the eigenvectors clearly are not. Ghosh and Whiffen[7] only report a rough estimate of the shape of the tensor, whereas Collins and Whiffen[9] only succeed in approximating the eigenvector associated with the maximum anisotropic coupling. The tensor reported by Hedberg and Ehrenberg[10] is the most complete and is therefore selected as reference in this work. These authors applied an iterative fitting procedure to six resolution-enhanced EPR spectra, measured for six different orientations of the magnetic field B with respect to the glycine crystal: for B aligned along the *a*-, *b*- and *c**-axis and along the three bisectors of this reference frame. This method suffers from a rather limited accuracy, and regardless of experimental inaccuracies, data points for only six directions yield a number of mathematically valid but physically irrelevant solutions. More specifically, changing the sign of one or more off-diagonal elements A_{ij} (as well as A_{ji} with $i \neq j$) in the hyperfine tensor (determined in the ⟨*abc**⟩ reference frame) yields an essentially different tensor that equally well fits the data points. In the case at hand, a much better agreement with the calculated eigenvectors is obtained when a sign change is applied on A_{12} and A_{13} (and on A_{21} and A_{31}). This corresponds to an ($a \rightarrow -a$) symmetry operation on the eigenvectors while the eigenvalues remain unchanged. The original as well as the new hyperfine tensor (in this reference frame) are shown in Table 5 along with Ψ angles with respect to the principal directions of all calculated ^{14}N tensors. Further corroboration of this tensor form comes from the ENDOR experiment by Collins and Whiffen.[9] The only eigenvector that was determined in that work (corresponding to the maximum anisotropic coupling) is in better agreement with the corresponding N_{exp} ($a \rightarrow -a$) eigenvector.

In the above discussion on the glycine radical vibrations, it is ascertained that the amino group does not rotate freely during the MD simulations. Yet, in the original papers describing the EPR experiments,[6,9,11] it was clearly attested that the amino group is a free rotor, causing the amino protons to be equivalent. Hence, it could be anticipated that the experimental H_N tensor at room temperature would be ill reproduced by the temperature simulations. If full rotational exchange would occur in an MD simulation, the average hyperfine tensor for each of the individual amino protons would become identical. In the absence of such exchange, the H_N tensor can be estimated by calculating the mean of the individual amino proton tensors (H_6−H_8). This is shown in Table 6 for the dynamic as well as the static calculations. For reference purposes, the average tensor is also calculated using the inequivalent $\beta1$−$\beta3$ amino tensors as determined by Collins and Whiffen at 77 K, where rotational averaging does not occur.[9] All but one of the calculated tensors are qualitatively similar to this avg($\beta1$−3) result. Isotropic and anisotropic values are in good correspondence with the H_N(exp) tensor (measured at 300 K). With respect to the tensor eigenvectors, only the one corresponding to maximum anisotropic value agrees to within 10° from the measured principal direction. This vector is aligned with the rotational C_2−N_3 axis, as is clear from the table where the unit vector along this direction in the (undamaged) crystal is given. The eigenvectors with minor and intermediate anisotropic eigenvalues are more troublesome and cannot be determined on the basis of static properties. This is evidenced in the mean avg($\beta1$−3) tensor, for which these eigenvectors deviate both by 76° from the H_N data. Also in the 100 and 300 K dynamics calculations, the Ψ angles for these components never drop below 40°. In the 500 K simulation, however, a drastic change takes place, pushing the Ψ angles below 15°, in good agreement with the H_N data. This is somewhat surprising, since even in this MD run, the amino group of the radical does not undergo rotational exchange. Presumably, enough high-energy conformers along the rotational barrier (sketched in Figure 5b) are attained at 500 K to improve the time-averaged mean amino proton hyperfine coupling.

Correlation between Dynamics and EPR Properties. From the vibrational analysis it is already clear that inversion at the C_2 center (at 702 cm^{-1}) is an important dynamic feature of the radical structure. These modifications in the planarity of the radical center also have a significant impact on the EPR

TABLE 6: Comparison of the Measured H_N Tensor with the Mean of the Three Amino Proton Tensors Predicted in Static and Dynamic Calculations[a]

	A_{iso}	T_{aniso}	principal axes vs $\langle a^*bc \rangle$			Ψ
H_N (exp)	49.1	−2.9	0.548	0.700	0.458	
		−2.1	0.222	−0.649	0.727	
		5.0	0.807	−0.297	−0.511	
avg(β1−3)	49.1	−3.2	0.345	−0.460	0.818	76
		−2.2	−0.536	−0.812	−0.230	76
		5.4	0.771	−0.359	−0.527	4
static/6−311G**	49.3	−3.2	0.381	0.885	−0.268	45
		−2.4	0.501	0.046	0.864	45
		5.5	0.777	−0.464	−0.426	11
static/EPR-III	52.6	−3.2	0.425	0.884	−0.195	40
		−2.2	0.464	−0.028	0.886	40
		5.5	0.777	−0.467	−0.422	11
100 K	49.3	−3.2	0.354	0.904	−0.240	44
		−2.4	0.499	0.034	0.866	44
		5.5	0.791	−0.426	−0.439	9
300 K	48.3	−3.1	0.355	0.915	−0.191	42
		−2.5	0.500	−0.013	0.866	41
		5.5	0.790	−0.403	−0.462	7
500 K	50.8	−2.6	0.359	0.841	0.404	14
		−2.4	0.450	−0.535	0.715	15
		5.0	0.818	−0.075	−0.571	13
	C2−N3		0.754	−0.456	−0.472	10

[a] The avg(β1−3) hyperfine tensor is determined on the basis of the individual β1−β3 amino tensors as measured by Collins and Whiffen at 77 K.[9] For reference purposes, the unit vector is given along the C_2−N_3 axis in the undamaged glycine crystal.

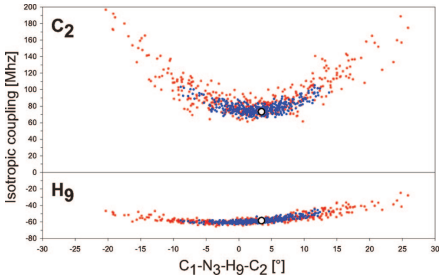


Figure 7. Plots of the C_2 and H_9 isotropic hyperfine coupling during the 100 K (blue) and 300 K (red) simulations as a function of radical planarity (indexed by the dihedral angle C_1−N_3−H_9−C_2). White circles indicate the corresponding couplings of the static/6−311G** calculation.

properties of the atoms that are directly associated with the center (C_2 and H_9 in glycine). This can be easily inferred by considering that the energy potential associated with the out-of-plane motion (Figure 5a) is activated in both 100 and 300 K simulations. In the ground state (C_1−N_3−H_9−$C_2 \approx 0°$), it would appear that the unpaired electron on the radical resides in a p-type orbital of C_2, perpendicular to the $\langle H_9C_1N_3 \rangle$ plane. However, by explicitly introducing temperature, higher energy conformers in the potential can be attained: up to ± 15° at 100 K (∼4 kJ/mol) and ± 25° at 300 K (∼7 kJ/mol). At a finite temperature, therefore, the spin density in reality switches between sp^3-like orbitals above and below the $\langle H_9C_1N_3 \rangle$ plane of the radical. The isotropic hyperfine coupling constants of C_2 and H_9 are directly related to the spin density at the nucleus and will be equally influenced by the inversion motion. This relation is apparent in Figure 7, where the 400 calculated isotropic couplings during the sampled 100 and 300 K MD trajectories are plotted as a function of the radical planarity. The effect of the inversion is much more pronounced for the carbon than for H_9, but both plots show a sharp increase in the coupling when the glycine radical deviates from planarity. Since the dependence on the out-of-plane movement is nonlinear, time averaging over all points will always result in an increase for both nuclei. This increase is apparent in the dynamic EPR results of Table 4, but is more pronounced for the 300 K simulation because the thermal activation of the potential (Figure 5a) is higher. So, although the radical at 300 K, on average, also assumes a virtually planar conformation, the average isotropic hyperfine couplings for H_9 and C_2 are considerably higher than those determined only at the ground state. For reference, the latter are marked by white dots in Figure 7.

Of course, this dependency has already been reported earlier. The isotropic hyperfine coupling of alfa protons in organic π-type radicals, for instance, is found to depend on radical planarity.[48] But more importantly, the group of Barone et al. has determined an equivalent relation as in Figure 7.[12,14] In a number of studies on the (zwitterionic) glycine, vibrational averaging effects[49] were accounted for to approximate the isotropic hyperfine couplings of the radical at a finite temperature. This required the (static) calculation of hyperfine couplings for several out-of-plane distortions of the radical center, yielding a relation between planarity and isotropic value much like that of Figure 7. Because only a very limited number of calculations sufficed to complete the relation in this way, temperature averages of the magnetic properties could be determined at much higher levels of theory or with large basis sets. Using an adapted triple-ζ basis set at the UQCISD level, Barone et al.[12] calculated 127.8 MHz for the ^{13}C coupling, whereas Rega et al.[14] estimated it at 120.5 MHz using DFT (B3LYP functional) and the EPR-II basis set.[3] Intriguingly, even though the influence of the environment was not taken into account in the former calculation, the correspondence with experiment was nearly perfect. In the latter study, intermolecular interactions of the radical were implicitly described with the aid of a solvent simulation model (CPCM) resulting in an equally fair correspondence. More importantly, though, a temperature of 77 K was taken as reference for the vibrational

averaging in both works. Given that the expected value for the isotropic coupling increases with rising temperature, it is not impossible that both methods would actually overshoot the experimental value of 126.7 MHz at room temperature.

Despite restrictions regarding the computational cost, the method adopted in this work is more generally applicable. On the basis of the dynamics, the relation between geometric features and magnetic properties can be easily established.[37] Moreover, it is not compulsory that these relations (if present) are recognized to determine a temperature average, provided that the MD simulation is long enough to ensure (quasi) ergodicity. In addition, possible anharmonicities or couplings with other degrees of freedom, for e.g., the inverse coupling between out-of-plane motion for the radical center and the CO_2 group, are automatically taken into account.

4. Conclusions

In this work, the results are presented of periodic supercell calculations on the $^{+}NH_3-^{\bullet}CH-CO_2^{-}$ radical of glycine in the solid state. Static as well as dynamic properties are considered for the radical in interaction with its environment. This allows the evaluation of temperature effects, which are determined on the basis of molecular dynamics simulations at several temperatures.

The structure of the glycine radical remains largely unaltered by explicitly introducing temperature. The radical center assumes an, on average, planar structure, corresponding to static calculations of the ground state. The potential energy surface for this structural feature is determined (Figure 5a) and is adequately sampled in the 100 and 300 K simulations. The corresponding surface for rotation of the radical amino group about the C_2-N_3 axis is also established, but in the time frame of the MD simulations, rotational exchange does not occur at 300 K. This is apparently in contrast with EPR experiments, revealing that the amino group undergoes a rotational averaging motion at room temperature.[6,9] The discrepancy can be attributed to the small simulation time (2.4 ps) since free rotation of other amino groups in the supercell is observed in tentative molecular dynamics at 500 K.

For the 100 and 300 K simulations, a vibrational analysis is performed derived from Fourier transformation of the atomic velocity autocorrelation functions. Using an approximation based on band-pass filtering, several vibrational modes are identified and brought in correspondence with experimental infrared and Raman assignments for (unirradiated) glycine crystals. Decomposition of the calculated spectra in terms of radical motion reveals that several vibrational modes are unique to the paramagnetic species. The most prominent feature is a particularly intense mode at 702 cm^{-1}, corresponding to out-of-plane motion of the radical center inversely coupled with similar motion of the CO_2 carbon. Since this frequency band is well separated from any other bands due to intact molecules, it might be applicable for discriminatory detection of glycine radicals with the aid of Raman spectroscopy, as was recently accomplished in other biomolecules.[21,22]

A hybrid periodic/cluster scheme is used to evaluate the EPR properties of the glycine radical along the (sampled) MD trajectories and in the ground state. This method ensures that structural information regarding the molecular environment of the radical obtained from a periodic approach can be employed in a subsequent determination of its magnetic properties. As such, calculations on the ground-state structure of the glycine radical yield an already good agreement with experimental EPR properties, and only a modest improvement is achieved by explicitly introducing temperature in the simulations. The most evident effect is on the isotropic hyperfine coupling of the ^{13}C tensor, which increases to 95.1 MHz in a thermal average determined on the 300 K molecular dynamics run. This constitutes an important improvement in the direction of the experimental value (126.7 MHz). Residual differences with the experiment are attributed to the use of a basis set that is not extensive enough. Additionally, the hyperfine tensors for the individual amino protons, as determined from low temperature ENDOR measurements, are found to match consistently better with the thermal averages at 100 K than at 300 K.

Since none of the dynamics simulations succeeds in reproducing free amino group rotation for the radical, theoretical estimates of the H_N tensor rely on the mean of the individual amino protons. Even though this tensor could only be determined at room temperature in experiments, the calculated mean tensors always correspond quite well, virtually independent from the simulation temperature. Only for the tentative 500 K simulation, a significant further improvement is obtained for the eigenvectors corresponding to minor and intermediate anisotropic hyperfine couplings.

Critical reevaluation of the experimental data used by Hedberg and Ehrenberg[10] indicates that the ^{14}N tensor reported by these authors is most likely a mathematically valid but physically irrelevant solution of the fitting procedure that was applied. The implementation of an $(a \rightarrow -a)$ symmetry operation on the N_{exp} eigenvectors results in a substantially better agreement with the calculated magnetic properties. Further, corroborated by (albeit incomplete) ENDOR data from another experiment, it is concluded that this adjustment is probably legitimate.

Finally, from the molecular dynamics a nonlinear relation is verified between the planarity of the radical center and the isotropic couplings of C_2 and H_9. This dependency was established earlier from vibrational averaging approaches and is found to be comparable with the present work. It is argued that the method adopted in this work is more generally applicable, since no prior knowledge is required of any relation between magnetic and structural properties to derive a reasonable estimate of the temperature dependent EPR properties.

Note Added in Proof

Shortly after submission of the manuscript, a paper was published in which similar static calculations were presented on the glycine radical in the solid state.[50]

Acknowledgment. This work is supported by the Fund for Scientific Research − Flanders (FWO) and the Research Board of the Ghent University.

References and Notes

(1) (a) Atherton, N. M. *Principles of Electron Spin Resonance*; Ellis Horwood: New York, 1993. (b) Weil, J. A.; Bolton, J. R.; Wertz, J. E. Electron Paramagnetic Resonance: Elementary Theory and Practical Applications; Wiley-Interscience: New York, 1994.

(2) For an example of a reference work, see Parr, R. G.; Yang, W. *Density-Functional Theory of Atoms and Molecules*; Oxford University Press: New York, 1989.

(3) Barone, V. In *Recent Advances in Density Functional Methods, Part I*; Chong, D. P., Ed.; World Scientific Publishing Co.: Singapore, 1995; Chapter *8.*

(4) Malkin, V. G.; Malkina, O. L.; Eriksson, L. A.; Salahub, D. R. In *Modern Density Functional Theory: A Tool for Chemistry*; Pulitzer, P., Seminario, J. M. Eds.; Elsevier: Amsterdam, 1995; Chapter *9.*

(5) Kaupp, M.; Bühl, M.; Malkin, V. G. *Calculation of NMR and EPR Parameters: Theory and Applications*; Wiley-VCH: Weinheim, 2004.

(6) Sagstuen, E.; Sanderud, A.; Hole, E. O. *Radiat. Res.* **2004**, *162*, 112.
(7) Ghosh, D. K.; Whiffen, D. H. *Mol. Phys.* **1959**, *2*, 285.
(8) Morton, J. R. *J. Am. Chem. Soc.* **1964**, *86*, 2325.
(9) Collins, M. A.; Whiffen, D. H. *Mol. Phys.* **1966**, *10*, 317.
(10) Hedberg, A.; Ehrenberg, A. *J. Chem. Phys.* **1968**, *48*, 4822.
(11) Sanderud, A.; Sagstuen, E. *J. Phys. Chem. B* **1998**, *102*, 9353.
(12) Barone, V.; Adamo, C.; Grand, A.; Subra, R. *Chem. Phys. Lett.* **1995**, *242*, 351.
(13) Ban, F.; Gauld, J. W.; Boyd, R. J. *J. Phys. Chem. A* **2000**, *104*, 5080.
(14) Rega, N.; Cossi, M.; Barone, V. *J. Am. Chem. Soc.* **1998**, *120*, 5723.
(15) (a) Onsager, L. *J. Am. Chem. Soc.* **1936**, *58*, 1486. (b) Wong, M. W.; Frisch, M. J.; Wiberg, K. B. *J. Am. Chem. Soc.* **1991**, *113*, 4776. (c) Wong, M. W.; Wiberg, K. B.; Frisch, M. J. *J. Chem. Phys.* **1991**, *95*, 8991. (d) Wong, M. W.; Wiberg, K. B.; Frisch, M. J. *J. Am. Chem. Soc.* **1992**, *114*, 523. (e) Wong, M. W.; Wiberg, K. B.; Frisch, M. J. *J. Am. Chem. Soc.* **1992**, *114*, 1645.
(16) (a) Cossi, M.; Barone, V.; Cammi, R.; Tomasi, J. *Chem. Phys. Lett.* **1996**, *255*, 327. (b) Barone, V.; Cossi, M.; Tomasi, J. *J. Comput. Chem.* **1998**, *19*, 407. (c) Barone, V.; Cossi, M. *J. Phys. Chem.* **1998**, *102*, 1995.
(17) Pauwels, E.; Van Speybroeck, V.; Waroquier, M. *J. Phys. Chem. A* **2004**, *108*, 11321.
(18) Car, P.; Parrinello, M. *Phys. Rev. Lett.* **1985**, *55*, 2471.
(19) (a) Asher, J. R.; Doltsinis, N. L.; Kaupp, M. *J. Am. Chem. Soc.* **2004**, *126*, 9854. (b) Rothlisberger, U.; Carloni, P. *Int. J. Quantum Chem.* **1999**, *73*, 209.
(20) Morrison, C. A.; Siddick, M. M.; Camp, P. J.; Wilson, C. C. *J. Am. Chem. Soc.* **2005**, *127*, 4042.
(21) (a) Kizil, R.; Irudayaraj, J.; Seetharaman, K. *J. Agric. Food Chem.* **2002**, *50*, 3912. (b) Kizil, R.; Irudayaraj, J. *J. Sci. Food Agric.* **2007**, *87*, 1244.
(22) Kinder, R.; Ziegler, C.; Wessels, J. M. *Int. J. Radiat. Biol.* **1997**, *71*, 561.
(23) (a) Melin, A. M.; Perromat, A.; Deleris, G. *Arch. Biochem. Biophys.* **2001**, *394*, 265. (b) Gault, N.; Lefaix, J. L. *Radiat. Res.* **2003**, *160*, 238. (c) Gault, N.; Rigaud, O.; Poncy, J. L.; Lefaix, J. L. *Int. J. Radiat. Biol.* **2005**, *81*, 767.
(24) Destro, R.; Roversi, P.; Barzaghi, M.; Marsh, R. E. *J. Phys. Chem. A* **2000**, *104*, 1047.
(25) CPMD V3.11 Copyright IBM Corp 1990−2006. Copyright MPI fuer Festkoerperforschung Stuttgart 1997−2001.
(26) (a) Perdew, J. P. *Phys. Rev. B* **1986**, *33*, 8822. (b) Becke, A. D. *J. Chem. Phys.* **1992**, *96*, 2155.
(27) Vanderbilt, D. *Phys. Rev. B* **1990**, *41*, 7892.
(28) Hockney, R. W. *Methods Comput. Phys.* **1970**, *9*, 136.
(29) Martyna, G. J.; Tuckerman, M. E. *J. Chem. Phys.* **1999**, *110*, 2810.
(30) Declerck, R.; Pauwels, E.; Van Speybroeck, V.; Waroquier, M. *Phys. Rev. B* **2006**, *74*, 245103.
(31) Declerck, R.; Pauwels, E.; Van Speybroeck, V.; Waroquier, M. *J. Phys. Chem. B* **2007**, *112*, 1508.
(32) Pauwels, E.; Van Speybroeck, V.; Waroquier, M. *Spectrochim. Acta, Part A* **2006**, *63*, 795.
(33) *Gaussian 03, Revision B.03*, Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A., Jr.; Vreven, T.; Kudin, K. N. et al. Gaussian, Inc.: Wallingford CT, 2004.
(34) Becke, A. D. *J. Chem. Phys.* **1996**, *104*, 1040.
(35) (a) Krishnan, R.; Binkley, J. S.; Seeger, R.; Pople, J. A. *J. Chem. Phys.* **1980**, *72*, 650. (b) McLean, A. D.; Chandler, G. S. *J. Chem. Phys.* **1980**, *72*, 5639.
(36) Pauwels, E.; Van Speybroeck, V.; Waroquier, M. *J. Phys. Chem. A* **2006**, *110*, 6504.
(37) Pauwels, E.; Verstraelen, T.; Waroquier, M. *Spectrochim. Acta, Part A* **2008**, *69*, 1388.
(38) Sprik, M.; Hutter, J.; Parrinello, M. *J. Chem. Phys.* **1996**, *105*, 1142.
(39) Chernobai, G. B.; Chesalov, Y. A.; Burgina, E. B.; Drebushchak, T. N.; Boldyreva, E. V. *J. Struct. Chem.* **2007**, *48*, 332.
(40) (a) Murli, C.; Thomas, S.; Venkateswaran, S.; Sharma, S. M. *Physica B* **2005**, *364*, 233. (b) Otero, E.; Urquhart, S. G. *J. Phys. Chem. A* **2006**, *110*, 12121.
(41) Shi, Y.; Wang, L. *J. Phys. D: Appl. Phys.* **2005**, *38*, 3741.
(42) (a) Wheeler, R. A.; Dong, H.; Boesch, S. E. *Chem. Phys. Chem* **2003**, *4*, 382. (b) Schmitz, M.; Tavan, P. *J. Chem. Phys.* **2004**, *121*, 12233. (c) Schmitz, M.; Tavan, P. *J. Chem. Phys.* **2004**, *121*, 12247.
(43) Martinez, M.; Gaigeot, M. P.; Borgis, D.; Vuilleumier, R. *J. Chem. Phys.* **2006**, *125*, 44106.
(44) Stubbe, J.; van der Donk, W. A. *Chem. Rev.* **1998**, *98*, 705.
(45) Pauwels, E.; Van Speybroeck, V.; Lahorte, P.; Waroquier, M. *J. Phys. Chem. A* **2001**, *105*, 8794.
(46) Gu, Z.; Ebisawa, K.; McDermott, A. *Solid State Nucl. Magn. Reson.* **1996**, *7*, 161.
(47) Iannuzzi, M.; Laio, A.; Parrinello, M. *Phys. Rev. Lett.* **2003**, *90*, 238302.
(48) Erling, P. A.; Nelson, W. H. *J. Phys. Chem. A* **2004**, *108*, 7591.
(49) Improta, R.; Barone, V. *Chem. Rev.* **2004**, *104*, 1231.
(50) Barone, V.; Causà, M. *Chem. Phys. Lett.* **2008**, *452*, 89.

JP711997Y

Paper 13: "Insight into the Solvation and Isomerization of 3-Halo-1-Azaallylic Anions from ab Initio Metadynamics Calculations and NMR Experiments"

Reinout Declerck, Bart De Sterck, Toon Verstraelen, Guido Verniest, Sven Mangelinckx, Jan Jacobs, Norbert De Kimpe, Michel Waroquier, Veronique Van Speybroeck

Chemistry - A European Journal, **2008**, 15, 580 - 584

CHEMISTRY
A EUROPEAN JOURNAL

DOI: 10.1002/chem.200800948

Insight into the Solvation and Isomerization of 3-Halo-1-azaallylic Anions from Ab Initio Metadynamics Calculations and NMR Experiments

Reinout Declerck,[a] Bart De Sterck,[a] Toon Verstraelen,[a] Guido Verniest,[b] Sven Mangelinckx,[b] Jan Jacobs,[b] Norbert De Kimpe,*[b] Michel Waroquier,[a] and Veronique Van Speybroeck*[a]

In organic synthesis, it is observed experimentally that the nature of the solvent can influence tremendously the reactivity and overall product selectivity. In this communication, we report on the *E*/*Z* isomerization of a typical solvated species, that is, the stable lithiated 3-chloro-3-methyl-1-azaallylic anion readily accessible from the *N*-isopropylimine of α-chloropropiophenone, from a theoretical and a subsequent NMR study. To date, the configurational properties of these 3-chloro-3-methyl-1-azaallylic anions are poorly understood. Our main emphasis is devoted to the monomeric species as these are believed to be the most important for further reactivity studies (see below). It will be shown that the investigated species is a particular example in which the inclusion of the solvent in the modeling study is of the utmost importance to determine the proper chemical behavior.

The 3-chloro-3-methyl-1-azaallylic anion was chosen as a model compound to obtain a deeper insight into the structural features of 3-halo-1-azaallylic anions and to get a better understanding, and eventually a better control, of the stereochemical outcome of the reactions in which these anions are involved. Since their first use in the early 1960s,[1–3] non-halogenated 1-azaallylic anions have gained a predominant role in organic synthesis due to their ability to form new C–C bonds with a lack of side products.[4] The chemistry of 1-azaallylic anions leads to basic heterocyclic systems such as aziridines, azetidines, pyrrolidines, pyrroles, piperidines, oxiranes, oxolanes, and higher functionalized ring systems, currently of interest for pharmaceutical chemistry and agrochemistry. The application of certain halogenated counterparts, that is, the 3-chloro-3-methyl-1-azaallylic anions, in particular by the group of De Kimpe and more general by the group of Florio, which incorporated the 3-chloro-3-methyl-1-azaallylic moiety into heterocyclic structures, has led to the synthesis of various important classes of compounds such as cyclopropanes,[5] tetrahydrofurans,[6] tetrahydropyrans,[6c] oxiranes,[7] aziridines,[7b,d,8] chloroimines,[9] pyrroles and pyridines,[10] steroids,[11] alkenylheterocycles,[12] and oxazetidines.[13] As mentioned, 3-chloro-3-methyl-1-azaallylic anions can be used for the synthesis of functionalized oxiranes and aziridines since the former anions behave as nucleophiles in Darzens- and aza-Darzens-type reactions with carbonyl compounds and imines.[7,8] One of the determining factors in the stereochemical outcome of these Darzens-type reactions is the *E*/*Z* stereochemistry of the starting 1-azaallylic anion.[14] Therefore, it is important to know and understand the configurational properties of 3-chloro-3-methyl-1-azaallylic anions in order to perform aldol- and Mannich-type reactions with these intermediates in a stereocontrolled manner.

NMR investigation and semiempirical calculations on the stereochemistry of (2-(α-chloroethyl)benzothiazolyl)lithium and (4,4-dimethyl-2-(α-chloroethyl)oxazolinyl)lithium have demonstrated that internal coordination between lithium and chlorine stabilizes the corresponding isomer with nitrogen and chlorine at the same side of the C=C double bond.[15] The lack of such structural investigations on 3-chloro-3-methyl-1-azaallylic anions in sensu stricto urged us to study their stereochemical properties.

The *E*/*Z* isomerism for non-halogenated 1-azaallylic anions has been observed and investigated quite frequently. The facile carbon–carbon bond rotation in simple lithiated 1-azaallylic anions was investigated using ^{1}H NMR spectros-

[a] Dr. R. Declerck, B. De Sterck, T. Verstraelen, Prof. Dr. M. Waroquier, Prof. Dr. V. Van Speybroeck
Center for Molecular Modeling, Ghent University
Proeftuinstraat 86, 9000 Gent (Belgium)
Fax: (+32) 9-264-6697
E-mail: Veronique.VanSpeybroeck@UGent.be

[b] Dr. G. Verniest, Dr. S. Mangelinckx, J. Jacobs, Prof. Dr. N. De Kimpe
Department of Organic Chemistry, Faculty of Bioscience Engineering
Ghent University, Coupure links 653, 9000 Gent (Belgium)
Fax: (+32) 9-264-6243
E-mail: Norbert.DeKimpe@UGent.be

Supporting information for this article is available on the WWW under http://dx.doi.org/10.1002/chem.200800948.

copy.[16] The rotational activation free energy was found to be (74.1 ± 1.3) kJ mol^{-1} at 313 K. *E*/*Z* isomerization was also observed upon deprotonation of ketimines of 2-butanone at room temperature.[17] For closely related non-chlorinated analogues of the species **1**–**2** in Figure 1, that is, the lithiated anion derived from the *N*-phenylimine of propiophenone, isomerization from the kinetically favored *E* isomer with the methyl group and the phenyl group at the same side of the C2=C3 double bond (like in *Z*-isomer **1**) to the thermodynamically most stable *Z* isomer with the methyl group and nitrogen at the same side of the C2=C3 double bond (like in *E*-isomer **2**), was observed.[18] Therefore, it was assumed that the lithiated 3-chloro-1-azaallylic anions of the present work could also undergo a similar type of isomerization.

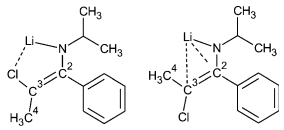


Figure 1. *Z*-isomer **1** and *E*-isomer **2** of the lithiated (2-chloro-1-phenylprop-1-en-1-yl)isopropylamide anion.

At first instance, the lithiated 3-chloro-3-methyl-1-azaallylic anion was generated by deprotonation of *N*-(2-chloro-1-phenylpropylidene)isopropylamine with lithium diisopropylamide (LDA) in [D$_8$]THF at 273 K and analyzed by ^{1}H and ^{13}C NMR spectroscopy. The α-chloropropiophenone imine was deprotonated under these conditions to a single stereoisomer as demonstrated by the presence of a single set of characteristic ^{1}H NMR chemical shifts of the methyl group on the double bond (s, $\delta = 1.77$ ppm), methine function (septet, 2.96 ppm) and isopropyl methyl groups (d, 0.83 ppm). Also a single set of characteristic ^{13}C NMR chemical shifts of the lithiated 3-chloro-1-azaallylic anion were observed (see Supporting Information). The stereochemistry of the *Z* anion **1** was determined by off-resonance ROESY spectroscopy showing ROE effects between the methyl group and the *ortho*-protons of the phenyl ring positioned at the same side of the carbon–carbon double bond (Figure 2). Furthermore, the observed ROE effects between the *N*-isopropyl substituent and the phenyl group support the *anti* stereochemistry of the 3-chloro-1-azaallylic anion, that is, the *N*-isopropyl group is oriented anti with respect to the C2=C3 double bond. In contrast to the non-chlorinated species,[19] the aforementioned NMR experiments indicate that in THF only the *Z*/*anti* isomer **1** occurs and that both amide and C=C double bond rotations are inhibited. This particular behavior of chlorinated 1-azaallylic anions demanded a theoretical interpretation. Before elaborating on the theoretical results, it is important to focus on the tendency of 1-azaallylic anions to form higher aggregates in solution. From the extensive work that has been performed on

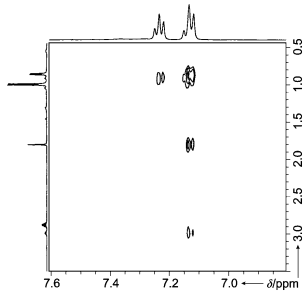


Figure 2. Details of the off-resonance ROESY spectrum of the *Z*-isomer **1** of the lithiated 3-chloro-1-azaallylic anion ([D$_8$]THF). The resonance signals at 0.96 ppm (doublet) and 2.84 ppm (septet) are from diisopropylamine formed upon protonation of LDA.

lithium 1-azaallylic anions, higher aggregates are also expected here.[19] It must, however, be stressed that the experimental NMR spectroscopy data gives by no means any information on the solution aggregation number of the title compounds. According to a series of NMR spectroscopic studies and colligative measurements in THF, both monomeric and dimeric species will occur at low concentration of the lithium 1-azaallyl anions.[20] On the other hand, the monomeric lithium 1-azaallylic anions retain in most cases their structural properties in the dimer form;[21] the stereochemical preferences in reactions can readily be explained with the properties of the monomeric form.[22] There are also several indications that the monomeric forms are the most likely reactive forms of these molecules. Studies by Streitwieser and co-workers have shown that the reactive form of lithiated compounds, present in solution for the larger extent as higher aggregates, can be the monomeric form due to a kinetic advantage because of lower-energy transition states as compared to the less reactive higher aggregates.[23] Based on these arguments and the expensive computational resources needed for treating higher arguments, the theoretical part of this communication will focus only on the configuration of the monomeric species, being aware that these might be part of higher aggregates under experimental conditions.

Despite the huge amount of theoretical studies that appeared the last years, modeling of complex phenomena such as chemistry in liquids remains a challenge as standard optimization techniques and ab initio molecular dynamics methods are often not suitable.[24,25] The first set of methods is routinely performed nowadays, but for our systems in which the solvent participates actively, a single optimized structure does not resemble the configurational distribution at finite temperature. First-principle molecular dynamics simulations are often restricted by short simulation times. As such, interesting regions of phase space are often so high in free energy that their sampling during a standard MD simulation is a rare event. Enhanced sampling techniques have become

CHEMISTRY
A EUROPEAN JOURNAL

an active research domain.[25] The relatively new metadynamics method has particularly attracted our attention. It was first proposed by Laio and Parrinello and enables an enhanced sampling of separated regions in phase space, simultaneously mapping the underlying free-energy landscape as a function of a limited number of collective variables.[26] The particular implementation is based on the work of Iannuzi et al.[27]

Prior to the modeling of the 1-azaallylic anions, we modeled the liquid structure of pure THF by using first-principle molecular dynamics calculations. The liquid structure of THF was recently assessed via hydrogen/deuterium isotopic substitution neutron-diffraction techniques by Bowron, Finney, and Soper.[28] A periodic cubic simulation cell was filled with 64 THF molecules. This choice represents an optimal compromise between computational cost and a proper embedding of the solute in the solvent. The simulation cell size was chosen to correspond with the experimental density of 0.88 kg dm^{-3}.[29] The performance of the THF model was validated by calculating the radial distribution function (RDF) of the molecular centers, which was found to be in excellent agreement with the benchmark RDF reported in reference [28] (see Supporting Information). Moreover, the MD simulations yielded a conformational distribution of 59% twisted and 41% oxygen envelope, indicating a thorough sampling of the system.[28]

After the THF model had been successfully assessed, it was applied to study the degree of coordination of the 3-chloro-1-azaallylic anions in solution. The coordination number for lithium enolates in ethereal solvents is rather difficult to establish but four-coordinate lithium cations have been clearly recognized in NMR studies of solvent separated ion pairs.[30] For contact ion pairs, coordination is expected less important because of the electrostatic effect of the counter ion. Theoretically the structures of a variety of organic lithium compounds were determined in the gas phase and in solvation using microsolvation with explicit ethereal ligands and/or continuum models.[31] For the 1-azaallylic anions as encountered here which are subject to large steric crowding, the degree of coordination is not a priori clear and can not be deduced straightforwardly from the experimental data. Isothermal–isobaric (NPT) molecular dynamics simulations during a period of 2.5 ps show that the *Z*-isomer **1** is monocoordinated whereas the *E*-isomer **2** features a two-fold coordination with THF (illustrated in Figure 3). In the *E*-isomer **2** the halogen–lithium coordination is not present which allows a second THF molecule to coordinate with the counter ion.

In order to obtain insight into the occurrence of only one stereoisomer in case of 3-chloro-3-methyl-1-azaallylic anions **1** and **2**, we decided to construct the free-energy landscape connecting the basins of the two isomers. To this end we applied the metadynamics method in which the dihedral angles Cl-C3-C2-N and C4-C3-C2-N were chosen as collective variables. This choice guarantees the independent movement of the methyl and chlorine substituents. The resulting free-energy landscape as a function of the two dihedral angles is displayed in Figure 4. The Gibbs free energy barriers for *E*-to-*Z* and *Z*-to-*E* isomerization amount to (107.1 ± 12.1) kJ mol^{-1} and (128.6 ± 12.1) kJ mol^{-1}, respectively.[32] These barriers are high, preventing isomerization at the experimental temperature. The *Z*-isomer **1** is more stable than the *E*-isomer **2** by $\Delta G_{Z\text{-}E} = (21.5 \pm 12.1)$ kJ mol^{-1}, which indicates that the experimentally observed *Z*-isomer **1** is thermodynamically favored. Within a static cluster approach using a combined explicit/implicit solvent model we were unable to determine the transition state for *E*/*Z* isomerization as the coordination number varies during the chemical transformation. Moreover, the stability of the *Z*-isomer **1** with respect to the *E*-isomer **2** was 20 kJ mol^{-1} too high compared to the metadynamics calculations. By capturing the movement of both dihedral angles, we were able to observe the sp^2 to sp^3 hybridization transition of the C3 carbon atom upon rotation, a well-known feature of rotations about allylic bonds. This is reflected in the fact that the saddle point, denoted as $(E\text{–}Z)^+$, does not lie on the linear pathway connecting both isomers, which confirms a posteriori the importance of capturing the movement of both dihedral angles.

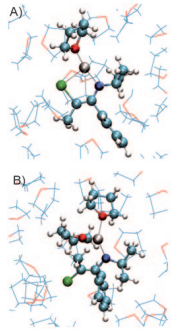


Figure 3. Characteristic snapshot of the MD simulation of the *Z*-isomer **1** (A) and the *E*-isomer **2** (B) solvated in THF.

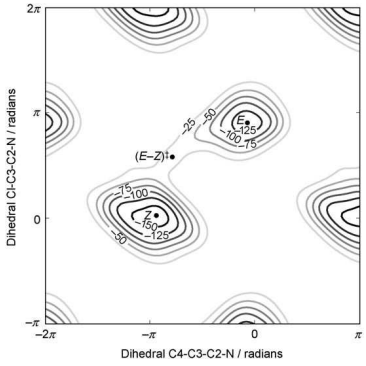


Figure 4. Gibbs free energy profile (in kJ mol^{-1}) governing the *E*–*Z* isomerization of the lithiated 3-chloro-1-azaallylic anion in THF. The positions of both stable isomers *E* (**2**) and *Z* (**1**) and the saddle point $(E\text{–}Z)^+$ are added. Note that the two collective variables feature a 2π periodicity.

Finally, we infer from both first principle metadynamics and NMR experiments that the *Z* isomer is the only configuration formed upon the deprotonation of the starting imine to the lithium 3-chloro-1-azaallylic anion. The interaction between the counter ion and the halogen, an effect that is not present in the non-halogenated 1-azaenolates, stabilizes the *Z* isomer by 21 kJ mol^{-1}. Moreover, the transition from the *Z* to the *E* isomer is very highly activated and features a change in coordination for the lithium cation as the broken lithium-chlorine interaction is replaced by a lithium–THF interaction. These effects can only be seen because of the explicit inclusion of the large THF model in the QM simulations. These results show that the stereochemistry of 3-chloro-3-methyl-1-azaallylic anions is manifestly different compared to their non-chlorinated counterparts and is the result of their configurational stability which should be beneficial during their synthetic use as functionalized intermediates in stereoselective reactions.

Experimental Section

All molecular dynamics calculations were performed within the cp2k/quickstep code,[33] employing the Gaussian and plane-wave (GPW) density functional method and periodic boundary conditions. A BLYP[34] gradient-corrected functional was used throughout, together with a TZVP-PSP[35] basis set, a 400 Ry cutoff for the auxiliary plane wave grid, and pseudopotentials developed by Goedecker and co-workers.[36] Isothermal-isobaric (NPT) MD simulations of both isomers were conducted. The species were properly embedded in the THF solvent model by determining, using atomic Pauling radii, the volume associated with their solvent accessible surface.[37] As the volume of THF is 2.96 times smaller compared to the volume of the 3-chloro-1-azaallylic anion, three THF molecules in the simulation cell were replaced by the 3-chloro-1-azaallylic species. An equilibration time of 2.5 ps has been respected to allow the solvent to accommodate to the presence of the solute and vice versa, followed by a 40 ps metadynamics run. Accurate metadynamics parameter values were determined from Gibbs free energy barrier predictions of the lithiated 3-chloro-1-azaallylic anion in the gas phase (using a supercell approach), including only one THF molecule to impose the limited freedom of the lithium cation. The set of parameter values $w=2.0$ kJ mol^{-1}, $s=0.33$ rad and $G=50$ fs (notation: see ref. [26]) yielded an energy barrier within 1.0 kJ mol^{-1} of the convergence limit. According to ref. [32] the estimated error using these parameters is $\varepsilon=6.1$ kJ mol^{-1}.

Lithiated 3-chloro-1-azaallylic anion 1: To a stirred solution of diisopropylamine (0.056 g, 0.55 mmol) in [D$_8$]THF (1 mL), *n*BuLi (0.22 mL, 0.55 mmol, 2.5 N in hexanes) was added slowly at 273 K. After 30 min of stirring at 273 K, the solution was evaporated in vacuo to dryness, after which, [D$_8$]THF (0.5 mL) was added and a solution of *N*-(2-chloro-1-phenylpropylidene)isopropylamine (0.11 g, 0.5 mmol) in [D$_8$]THF (0.5 mL) was dropped to the LDA solution at 273 K and stirring was continued for 1 h. Subsequently, the reaction mixture was allowed to reach room temperature during 15 min. ^{1}H NMR (300 MHz) and ^{13}C NMR (75 MHz) spectra were taken from the prepared 3-chloro-1-azaallylic anion **1** at room temperature. ^{1}H NMR (300 MHz, [D$_8$]THF): $\delta=0.83$ (d, $J=6.33$ Hz, 6H; $(CH_3)_2$CH), 1.77 (s, 3H; CH_3), 2.96 (septet, $J=6.1$ Hz, 1H; $(CH_3)_2CH$), 7.08–7.14 (m, 3H; *o*-CH$_{ar}$ and *p*-CH$_{ar}$), 7.18–7.24 ppm (m, 2H; *m*-CH$_{ar}$); ^{13}C NMR (75 MHz, [D$_8$]THF): $\delta=22.3$, 28.1, 48.6, 83.6, 125.9, 127.6, 130.0, 144.0, 155.0 ppm.

Acknowledgements

The authors thank Professor J. Martins (Department of Organic Chemistry, NMR and Structure Analysis, Ghent University) for recording the ROESY spectra, and Dr. D. T. Bowron (STFC Rutherford Appleton Laboratory) for providing the EPSR-derived RDF data. This work was supported by the Fund for Scientific Research-Flanders and the Research Board of Ghent University.

Keywords: azaallylic anions · isomerization · metadynamics · NMR spectroscopy · solvent effects

[1] G. Stork, S. R. Dowd, *J. Am. Chem. Soc.* **1963**, *85*, 2178–2180.
[2] G. Wittig, H. D. Fommeld, P. Suchanek, *Angew. Chem.* **1963**, *75*, 978–979.
[3] G. Wittig, H. Reiff, *Angew. Chem.* **1968**, *80*, 8–15.
[4] S. Mangelinckx, N. Giubellina, N. De Kimpe, *Chem. Rev.* **2004**, *104*, 2353–2400 and references therein.
[5] a) N. De Kimpe, P. Brunet, R. Verhé, N. Schamp, *J. Chem. Soc. Chem. Commun.* **1988**, 825–827; b) W. Aelterman, K. Abbaspour Tehrani, W. Coppens, T. Huybrechts, N. De Kimpe, D. Tourwé, J.-P. Declercq, *Eur. J. Org. Chem.* **1999**, 239–250; c) V. Capriati, S. Florio, R. Luisi, M. T. Rocchetti, *J. Org. Chem.* **2002**, *67*, 759–763; d) S. Florio, F. M. Perna, R. Luisi, J. Barluenga, F. Rodríguez, F. J. Fañanás, *J. Org. Chem.* **2004**, *69*, 5480–5482.
[6] a) N. De Kimpe, E. Stanoeva, N. Schamp, *Tetrahedron Lett.* **1988**, *29*, 589–592; b) N. De Kimpe, W. Aelterman, K. De Geyter, J.-P. Declercq, *J. Org. Chem.* **1997**, *62*, 5138–5143; c) W. Aelterman, N. Giubellina, E. Stanoeva, K. De Geyter, N. De Kimpe, *Tetrahedron Lett.* **2004**, *45*, 441–444.
[7] a) P. Sulmon, N. De Kimpe, N. Schamp, J.-P. Declercq, B. Tinant, *J. Org. Chem.* **1988**, *53*, 4457–4462; b) V. Capriati, L. Degennaro, S. Florio, R. Luisi, C. Tralli, L. Troisi, *Synthesis* **2001**, 2299–2306; c) F. M. Perna, V. Capriati, S. Florio, R. Luisi, *J. Org. Chem.* **2002**, *67*, 8351–8359; d) F. Bona, L. De Vitis, S. Florio, L. Ronzini, L. Troisi, *Tetrahedron* **2003**, *59*, 1381–1387; e) L. Troisi, L. De Vitis, C. Granito, P. Metrangolo, T. Pilati, L. Ronzini, *ARKIVOC* **2004**, *xiv*, 61–73.
[8] a) R. Luisi, V. Capriati, S. Florio, R. Ranaldo, *Tetrahedron Lett.* **2003**, *44*, 2677–2681; b) L. De Vitis, S. Florio, C. Granito, L. Ronzini, L. Troisi, V. Capriati, R. Luisi, T. Pilati, *Tetrahedron* **2004**, *60*, 1175–1182; c) R. Luisi, V. Capriati, S. Florio, P. Di Cunto, B. Musio, *Tetrahedron* **2005**, *61*, 3251–3260; d) L. Troisi, C. Granito, C. Carlucci, F. Bona, S. Florio, *Eur. J. Org. Chem.* **2006**, 775–781.
[9] P. Sulmon, N. De Kimpe, N. Schamp, *Synthesis* **1989**, 8–12.
[10] W. Aelterman, N. De Kimpe, V. Tyvorskii, O. Kulinkovich, *J. Org. Chem.* **2001**, *66*, 53–58.
[11] L. Troisi, S. Florio, C. Granito, *Steroids* **2002**, *67*, 687–693.
[12] V. Capriati, L. Degennaro, S. Florio, R. Luisi, *Eur. J. Org. Chem.* **2002**, 2961–2969.
[13] R. Luisi, V. Capriati, S. Florio, E. Piccolo, *J. Org. Chem.* **2003**, *68*, 10187–10190.
[14] V. K. Aggarwal, D. M. Badine, V. A. Moorthie in *Aziridines and Epoxides in Organic Synthesis* (Ed.: A. K. Yudin), Wiley-VCH, Weinheim, **2006**, pp. 1–35.
[15] A. Abbotto, S. Bradamante, S. Florio, V. Capriati, *J. Org. Chem.* **1997**, *62*, 8937–8940.
[16] J. Y. Lee, T. J. Lynch, D. T. Mao, D. E. Bergbreiter, M. Newcomb, *J. Am. Chem. Soc.* **1981**, *103*, 6215–6217.
[17] J. K. Smith, D. E. Bergbreiter, M. Newcomb, *J. Org. Chem.* **1981**, *46*, 3157–3158.
[18] R. Knorr, P. Low, *J. Am. Chem. Soc.* **1980**, *102*, 3241–3242.
[19] a) H. Ahlbrecht, E. O. Düber, D. Enders, H. Eichenauer, P. Weuster, *Tetrahedron Lett.* **1978**, *19*, 3691–3694; b) P. von Ragué Schleyer, R. Hacker, H. Dietrich, W. Mahdi, *J. Chem. Soc. Chem. Commun.* **1985**, 622–624; c) R. Glaser, C. M. Hadad, K. B. Wiberg, A. Streit-

CHEMISTRY
A EUROPEAN JOURNAL

wieser, *J. Org. Chem.* **1991**, *56*, 6625–6637; d) R. Knorr, H. Dietrich, W. Mahdi, *Chem. Ber.* **1991**, *124*, 2057–2063; e) P. B. Hitchcock, M. F. Lappert, D.-S. Liu, *J. Chem. Soc. Chem. Commun.* **1994**, 2637–2638; f) A. Abbotto, A. Streitwieser, P. von Ragué Schleyer, *J. Am. Chem. Soc.* **1997**, *119*, 11255–11268; Schleyer, *J. Am. Chem. Soc.* **1997**, *119*, 11255–11268; g) L. M. Pratt, T. E. Hogen-Esch, I. M. Khan, *Tetrahedron* **1995**, *51*, 5955–5970; h) L. R. Liou, A. J. McNeil, A. Ramirez, G. E. S. Toombes, J. M. Gruver, D. B. Collum, *J. Am. Chem. Soc.* **2008**, *130*, 4859–4868.

[20] a) R. A. Wanat, D. B. Collum, G. Van Duyne, J. Clardy, R. T. DePue, *J. Am. Chem. Soc.* **1986**, *108*, 3415–3422; b) N. Kallman, D. B. Collum, *J. Am. Chem. Soc.* **1987**, *109*, 7466–7472; c) L. M. Jackman, L. M. Scarmoutzos, B. D. Smith, P. G. Williard, *J. Am. Chem. Soc.* **1988**, *110*, 6058–6063.

[21] a) R. R. Fraser, N. Chuaqui-Offermanns, K. N. Houk, N. G. Rondan, *J. Organomet. Chem.* **1981**, *206*, 131–138; b) R. Glaser, A. Streitwieser, *J. Org. Chem.* **1991**, *56*, 6612–6624; c) R. Glaser, C. M. Hadad, K. B. Wiberg, A. Streitwieser, *J. Org. Chem.* **1991**, *56*, 6625–6637.

[22] a) R. Glaser, A. Streitwieser, *J. Am. Chem. Soc.* **1987**, *109*, 1258–1260; b) R. Glaser, A. Streitwieser, *Pure Appl. Chem.* **1988**, *60*, 195–204; c) R. Glaser, A. Streitwieser, *J. Am. Chem. Soc.* **1989**, *111*, 7340–7348; d) R. Glaser, A. Streitwieser, *J. Am. Chem. Soc.* **1989**, *111*, 8799–8809; e) R. Glaser, A. Streitwieser, *J. Org. Chem.* **1989**, *54*, 5491–5502; f) R. Glaser, A. Streitwieser, *J. Mol. Struct.* **1988**, *163*, 19–50.

[23] A. Streitwieser, *J. Mol. Model.* **2006**, *12*, 673–680.

[24] a) A. V. Marenich, R. M. Olson, A. C. Chamberlin, C. J. Cramer, D. G. Truhlar, *J. Chem. Theory Comput.* **2007**, *3*, 2055–2067; b) C. P. Kelly, C. J. Cramer, D. G. Truhlar, *J. Chem. Theory Comput.* **2005**, *1*, 1133–1152; c) J. Tomasi, M. Persico, *Chem. Rev.* **1994**, *94*, 2027–2094.

[25] B. Ensing, M. De Vivo, Z. Liu, P. Moore, M. L. Klein, *Acc. Chem. Res.* **2006**, *39*, 73–81.

[26] A. Laio, M. Parrinello, *Proc. Natl. Acad. Sci. USA* **2002**, *99*, 12562–12566.

[27] M. Iannuzzi, A. Laio, M. Parrinello, *Phys. Rev. Lett.* **2003**, *90*, 238302.

[28] a) D. T. Bowron, J. L. Finney, A. K. Soper, *J. Am. Chem. Soc.* **2006**, *128*, 5119–5126; b) M. Štrajbl, J. Florián, *Theor. Chem. Acc.* **1998**, *99*, 166–170; c) V. M. Rayón, J. A. Sordo, *J. Chem. Phys.* **2005**, *122*, 204303.

[29] a) D. R. Lide in *CRC Handbook of Chemistry and Physics*, 84th ed., CRC Press, Wiley, New York, **2003**; b) C. Carvajal, K. J. Tölle, J. Smid, M. Szwarc, *J. Am. Chem. Soc.* **1965**, *87*, 5548–5553.

[30] a) L. M. Pratt, N. Văn Nguyên, B. Ramachandran, *J. Org. Chem.* **2005**, *70*, 4279–4283; b) L. M. Pratt, B. Ramachandran, J. D. Xidos, C. J. Cramer, D. G. Truhlar, *J. Org. Chem.* **2002**, *67*, 7607–7612; c) P. L. Fast, M. L. Sanchez, D. G. Truhlar, *Chem. Phys. Lett.* **1999**, *306*, 407–410; d) J. A. Pople, M. Head-Gordon, K. Raghavachari, *J. Chem. Phys.* **1987**, *87*, 5968–5975.

[31] a) L. M. Pratt, A. Streitwieser, *J. Org. Chem.* **2003**, *68*, 2830–2838; b) V. Van Speybroeck, K. Moonen, K. Hemelsoet, C. V. Stevens, M. Waroquier, *J. Am. Chem. Soc.* **2006**, *128*, 8468–8478.

[32] A. Laio, A. Rodriguez-Fortea, F. L. Gervasio, M. Ceccarelli, M. Parrinello, *J. Phys. Chem. B* **2005**, *109*, 6714–6721.

[33] http://cp2k.berlios.de.

[34] a) A. D. Becke, *Phys. Rev. A* **1988**, *38*, 3098–3100; b) C. Lee, W. Yang, R. G. Parr, *Phys. Rev. B* **1988**, *37*, 785–789.

[35] G. Lippert, J. Hutter, P. Ballone, M. Parrinello, *J. Phys. Chem.* **1996**, *100*, 6231–6235.

[36] a) S. Goedecker, M. Teter, J. Hutter, *Phys. Rev. B* **1996**, *54*, 1703–1710; b) C. Hartwigsen, S. Goedecker, J. Hutter, *Phys. Rev. B* **1998**, *58*, 3641–3662.

[37] a) J. Tomasi, B. Mennucci, R. Cammi, *Chem. Rev.* **2005**, *105*, 2999–3093; b) Y. Takano, K. N. Houk, *J. Chem. Theory Comput.* **2005**, *1*, 70–77.

Received: May 18, 2008
Revised: August 29, 2008
Published online: November 7, 2008

Conclusions and Perspectives

A complete understanding of the molecular mechanisms that lead to the formation of zeolites is imperative for the continued progress in the synthesis and development of zeolite catalysts. Such insights facilitate a rational optimization of synthesis recipes and are indispensable to design zeolite catalysts that are tailor-made for a specific industrial application. Moreover, the zeolite precursors and the nanocrystals - these are the molecular fragments that precede the actual macroscopic zeolite crystals - are versatile ingredients for the synthesis of micro- and mesoporous materials. Detailed insights in the different roles of template molecules at the molecular level would be especially advantageous. Templates are mainly responsible for the selective synthesis of zeolites with cavities that have a shape complementary to the template, but they also have other functions during the synthesis.

In this thesis, the synthesis of MFI-structured zeolites based on the tetrapropylammonium (TPA) template is investigated with molecular modeling. In particular special attention has been devoted to the materials synthesized at the Center for Surface Chemistry and Catalysis (COK) of the KULeuven. This collaboration COK-CMM, originated in a strategic basic research project (SBO), is still growing and opens a lot of perspectives for the design of new materials. Computer simulations on molecular systems reveal precisely those insights that are inaccessible to experimental research due to the small scale. The ultimate goal, that is an exhaustive analysis of the molecular mechanisms in the synthesis of zeolites, is not achieved yet, but this PhD thesis brings it closer. The modeling puzzle is attacked from different angles, covering both the development of new software tools and the conception novel theoretical models.

The development of specialized software is vital for state-of-the-art molecular simulations and is an important part of this thesis. Both the preparation of input files and the analysis of the output of simulations are complex tasks that are infeasible without the appropriate software tools. ZEOBUILDER is the most prominent example of such a tool in this work. The program provides a user-friendly graphical interface for the construction of atomic models of micro- and mesoporous materials. Such models are a part of the input for molecular simulations. The current version of ZEOBUILDER is also highly usable in studies that do not concern zeolites. MD-TRACKS is the second important computer program in this thesis. It is a powerful analysis toolkit for molecular dynamics simulations. These two projects are distributed through the CMM Code website: http://molmod.ugent.be/code.

Several other software projects will be released in the near future. TAMKin computes kinetic and thermodynamic parameters based on the second order approximation of the molecular potential energy surface, and incorporates advanced features to treat partially optimized systems. HIPart is a program for the Hirshfeld partitioning of molecular electron densities and related schemes.

QFit and MMFit are used for the derivation empirical force-field parameters. All these computer codes are designed as object oriented frameworks that are easily extended with new features, which facilitates the implementation of new theoretical developments.

The silica-template interactions that direct the growth of zeolite crystals, can in principle be studied with molecular dynamics simulations using molecular mechanics models. However, the applicable models that are available in the literature are of poor quality. Mainly the inadequate description of the electrostatic interactions prohibits reliable simulations. We established well-defined protocols that facilitate the derivation of a next generation of force-field models. The Gradient Curves Method is an algorithm to obtain transferable valence force fields. We also developed and benchmarked parametrization techniques for the SQE model, i.e. an empirical electrostatics models suitable for molecular mechanics simulations that correctly describes electronic polarization.

The foregoing technological and methodological work is applied in computational studies of the key steps in the precursor-based synthesis scheme of MFI-structured materials developed at the COK of the KULeuven. A force-field model based on the Gradient Curves Method was used for the prediction of IR adsorption spectra of various silica oligomers, zeolite precursors and zeolite nanocrystals. This work confirms the hypothesis based on experimental observations that a shift of the so-called MFI-fingerprint in the IR spectrum can be associated with the formation of nanoscopic pentacyclic zeolite crystals. The properties of the TPA template in contact with 33T and 22T MFI-precursors is examined with a multi-level modeling approach. Both static quantum mechanical and classical molecular dynamics simulations unambiguously point out that the position of the TPA template with respect to these precursors already corresponds to their final position in macroscopic zeolite crystals. In the absence of TPA, the zeolite precursors collapse due to the formation of internal hydrogen bonds.

Although this work already provides a theoretical underpinning of the elementary properties of zeolite nanoparticles, there are plenty of opportunities for future molecular modeling research. The zeolite synthesis phenomenon is a complex network of physico-chemical processes that can transform relatively simple ingredients into fascinating micro- and mesoporous materials. Many of these constituent processes remain to be discovered and understood. The continued development a polarizable force-field model that can describe the relevant molecular systems, is essential to make such research possible and it will be a principal milestone in future work.

Bibliography

1. Cronstedt, A. F.; *Akad. Handl. Stockholm* **1756**, *18*, 120
2. Baerlocher, Ch.: McCusker, L.B.; Olson, D.H. *Atlas of Zeolite Framework Types*, Elsevier: 6th ed.; 2007
3. Cejka, J.; van Bekkum, H.; Corma, A.; Schueth, F. *Introduction to Zeolite Molecular Sieves, Volume 168, Third Edition (Studies in Surface Science and Catalysis);* Elsevier: 3rd ed.; 2007.
4. Auerbach, S. M.; Carrado, K. A.; Dutta, P. K. *Handbook of Zeolite Science and Technology;* CRC: 1 ed.; 2003
5. Cundy, C. S.; Cox, P. A. *Chem. Rev.* **2003**, *103*, 663-701
6. Barrer, R. M.; Denny, P. J. *J. Chem. Soc.* **1961**, 971.
7. Kerr, G. T.; Kokotailo, G. T. *J. Am. Chem. Soc.* **1961**, *83*, 4675.
8. Argauer, R. J.; Landolt, G. R. *U.S. Patent 3,702,886*, **1972**
9. Grose, R. W.; Flanigen, E. M. *Belgian Patent (BE) 851,066*, **1977**
10. Flanigen, E.; Bennett, J.; Grose, R.; Cohen, J.; Patton, R.; Kirchner, R. *Nature* **1978**, *271*, 512-516.
11. Wilson, S. T.; Lok, B. M.; Messina, C. A.; Cannan, T. R.; Flanigen, E. M. *J. Am. Chem. Soc.* **1982**, *104*, 1146-1147
12. Davis, M. E.; Saldarriaga, C.; Montes, C.; Garces, J.; Crowdert, C. *Nature* **1988**, *331*, 698-699
13. Taramasso, M.; Perego, G.; Notari, B. *U.S. Patent 4,410,501*, **1983**.
14. Yanagisawa, T.; Shimizu, T.; Kuroda, K.; Kato, C. *Bull. Chem. Soc. Jpn.* **1990**, *63*, 988-992
15. Kresge, C. T.; Leonowicz, M. E.; Roth, W. J.; Vartuli, J. C.; Beck, J. S. *Nature* **1992**, *359*, 710-712
16. Beck, J. S.; Vartuli, J. C.; Roth, W. J.; Leonowicz, M. E.; Kresge, C. T.; Schmitt, K. D.; Chu, C. T. W.; Olson, D. H.; Sheppard, E. W. a. *J. Am. Chem. Soc.* **1992**, *114*, 10834-10843
17. Tao, Y.; Kanoh, H.; Abrams, L.; Kaneko, K. *Chem. Rev.* **2006**, 106, published online
18. Ravishankar, R.; Kirschhock, C.; Schoeman, B. J.; Vanoppen, P.; Grobet, P. J.; Storck, S.; Maier, W. F.; Martens, J. A.; De Schryver, F. C.; Jacobs, P. A. *J. Phys. Chem. B* **1998**, *102*, 2633-2639
19. Ravishankar, R.; Kirschhock, C. E. A.; Knops-Gerrits, P. P.; Feijen, E. J. P.; Grobet, P. J.; Vanoppen, P.; De Schryver, F. C.; Miehe, G.; Fuess, H.; Schoeman, B. J.; Jacobs, P. A.; Martens, J. A. *J. Phys. Chem. B* **1999**, *103*, 4960-4964.

20. Kirschhock, C. E. A.; Ravishankar, R.; Verspeurt, F.; Grobet, P. J.; Jacobs, P. A.; Martens, J. A. *J. Phys. Chem. B* **1999**, *103*, 4965-4971

21. Kirschhock, C. E. A.; Ravishankar, R.; Looveren, L. V.; Jacobs, P. A.; Martens, J. A. *J. Phys. Chem. B* **1999**, *103*, 4972-4978

22. Kirschhock, C. E. A.; Ravishankar, R.; Jacobs, P. A.; Martens, J. A. *J. Phys. Chem. B* **1999**, *103*, 11021-11027

23. Kirschhock, C. E. A.; Buschmann, V.; Kremer, S.; Ravishankar, R.; Houssin, C. J. Y.; Mojet, B. L.; van Santen, R. A.; Grobet, P. J.; Jacobs, P. A.; Martens, J. A. *Angew. Chem. Int. Ed.* **2001**, *40*, 2637-2640

24. Kirschhock, C. E. A.; Kremer, S. P. B.; Grobet, P. J.; Jacobs, P. A.; Martens, J. A. *J. Phys. Chem. B* **2002**, *106*, 4897-4900

25. Kremer, S. P. B.; Kirschhock, C. E. A.; Aerts, A.; Villani, K.; Martens, J. A.; Lebedev, O. I.; Van Tendeloo, G. *Adv. Mater.* **2003**, *15*, 1705-1707

26. Kirschhock, C. E. A.; Kremer, S. P. B.; Vermant, J.; Van Tendeloo, G.; Jacobs, P. A.; Martens, J. A. *Chem. Eur. J.* **2005**, *11*, 4306-4313.

27. Press, W. H.; Flannery, B. P.; Teukolsky, S. A.; Vetterling, W. T. *Numerical Recipes in C: The Art of Scientific Computing*; Cambridge University Press: 1992.

28. Stroustrup, B. *The C++ Programming Language (3rd Edition)*; Addison-Wesley Professional: 1997

29. Magusin, P. C. M. M.; Zorin, V. E.; Aerts, A.; Houssin, C. J. Y.; Yakovlev, A. L.; Kirschhock, C. E. A.; Martens, J. A.; Vansanten, R. A. *J. Phys. Chem. B* **2005**, *109*, 22767-22774

30. Kragten, D. D.; Fedeyko, J. M.; Sawant, K. R.; Rimer, J. D.; Vlachos, D. G.; Lobo, R. F.; Tsapatsis, M. The *J. Phys. Chem. B* **2003**, *107*, 10006-10016

31. Liang, D.; Follens, L. R.; Aerts, A.; Martens, J. A.; Van Tendeloo, G.; Kirschhock, C. E. *J. Phys. Chem. C* **2007**, *111*, 14283-14285.

32. Davis, T. M.; Drews, T. O.; Ramanan, H.; He, C.; Dong, J.; Schnablegger, H.; Katsoulakis, M. A.; Kokkoli, E.; Mccormick, A. V.; Penn, L. R.; Tsapatsis, M. *Nat. Mater.* **2006**, *5*, 400-408.

33. Kumar, S.; Wang, Z.; Penn, L. R.; Tsapatsis, M. *J. Am. Chem. Soc.* **2008**, *130*, 17284-17286.

34. The Python Programming Language, http://www.python.org/

35. Dubois, P.F. *Comput. Sci. Eng.* **2007**, 9, 7

36. Baerlocher, Ch.; Hepp, A.; Meier, W.M. DLS-76, a program for the simulation of crystal structures by geometric refinement. Institut fuer Kristallo- graphie und Petrographie, ETH Zuerich.

37. Gale, J.D.; Rohl, A.L. *Mol. Simul.*, **2003**, *29*, 291

38. Verstraelen, T.; Van Speybroeck, V.; Waroquier, M. *J. Chem. Inf. Model.* **2008**, *48*, 1530-1541

39. Verstraelen, T.; Van Houteghem, M.; Van Speybroeck, V.; Waroquier, M. *J. Chem. Inf. Model.* **2008**, *48*, 2414-2424

40. Van Houteghem, M. "A Study of Liquids on an Atomic Scale: Molecular Dynamics on Organic Solvents", **2008**, june 18, Center for Molecular Modeling, Ghent Univeristy.

41. Standardisation and databasing of ab-initio and classical simulations, *Cecam workshop* **2008**, September 18, CECAM-ETHZ, Zurich, Switzerland

42. Bultinck, P.; Van Alsenoy, C.; Ayers, P. W.; Dorca, R. C. *J. Chem. Phys.* **2007,***126,* 144111

43. Bultinck, P.; Ayers, P. W.; Fias, S.; Tiels, K.; Van Alsenoy, C. *Chem. Phys. Lett.* **2007**, *444*, 205-208.

44. Van Damme, S.; Bultinck, P.; Fias, S. *J. Chem.l Theory Comp.* **2009**, *5*, 334-340.

45. Gaussian 03, Frisch, M. J. et al. Gaussian, Inc., Wallingford CT, **2004**

46. Hirshfeld, F. L. *Theor. Chim. Acta* **1977**, *44*, 129

47. Lillestolen, T. C.; Wheatley, R. *J. Chem. Commun.* **2008**, 5909-5911

48. Van Speybroeck, V.; Van Neck, D.; Waroquier, M. *J. of Phys. Chem. A* **2002**, *106*, 8945-8950

49. Van Speybroeck, V.; Reyniers, M.-F.; Marin, G. B.; Waroquier, *M. Chem. Phys. Chem.* **2002**, *3*, 863-870

50. Vansteenkiste, P.; Van Speybroeck, V.; Marin, G. B.; Waroquier, M. *J. Phys. Chem. A* **2003**, *107*, 3139-3145

51. Speybroeck, V.; Vansteenkiste, P.; Neck, D.; Waroquier, M. *Chem. Phys. Lett.* **2005**, *402*, 479-484

52. Vansteenkiste, P.; Pauwels, E.; Van Speybroeck, V.; Waroquier, M. *J. Phys. Chem. A* **2005**, *109*, 9617-9626

53. Vansteenkiste, P.; Van Neck, D.; Van Speybroeck, V.; Waroquier, M. *J. Chem. Phys.* **2006**, *124*, 044314

54. Vansteenkiste, P.; Van Speybroeck, V.; Verniest, G.; De Kimpe, N.; Waroquier, M. *J. Phys. Chem. A* **2006**, *110*, 3838-3844

55. Vansteenkiste, P.; Verstraelen, T.; Van Speybroeck, V.; Waroquier, M. *Chem. Phys.* **2006**, *328*, 251-258.

56. Ghysels, A.; Van Neck, D.; Van Speybroeck, V.; Verstraelen, T.; Waroquier, M. *J. Chem. Phys.* **2007**, *126*, 224102

57. Ghysels, A.; Van Speybroeck, V.; Verstraelen, T.; Van Neck, D.; Waroquier, M. *J. Chem. Theory Comp.* **2008**, *4*, 614-625

58. Mortier, W.; Van Genechten, K.; Gasteiger, J. *J. Am. Chem. Soc.* **1985**, *107*, 829-835

59. Nistor, R. A.; Polihronov, J. G.; Müser, M. H.; Mosey, N. J. *J. Chem. Phys.* **2006**, *125*, 094108

60. Rappe, A.; Casewit, C.; Colwell, K.; Goddard, W.; Skiff, W. *J. Am. Chem. Soc.* **1992**, *114*, 10024-10035

61. Rimer, J. D.; Fedeyko, J. M.; Vlachos, D. G.; Lobo, R. F. *Chem. Eur. J.* **2006**, *12*, 2926-2934

62. de Moor, P.-P. E. A.; Beelen, T. P. M.; Komanschek, B. U.; Beck, L. W.; Wagner, P.; Davis, M. E.; van Santen, R. A. *Chem. Eur. J.* **1999**, *5*, 2083-2088

63. Watson, J. N.; Iton, L. E.; Keir, R. I.; Thomas, J. C.; Dowling, T. L.; White, J. W. *J. Phys. Chem. B* **1997**, *101*, 10094-10104

64. Discover, Molecular Simulations Inc., San Diego, CA, 1995.

65. Lewis, D. W.; Catlow,; Thomas, J. M. *Faraday Disc.* **1997**, *106*, 451-471

66. Morse, P. M. *Phys. Rev.* **1929**, *34*, 57-64

67. Urey, H. C.; Bradley, C. A. *Phys. Rev.* **1931**, *38*, 1969-1978

68. Warshel, A.; Levitt, M.; Lifson, S. *J. Mol. Spectrosc.* **1970**, *33*, 84–99.

69. Allinger, N. L. *J. Am. Chem. Soc.* **1977**, *99*, 8127-8134

70. Allinger, N.; Yuh, Y.; Lii, J. *J. Am. Chem. Soc.* **1989**, *111*, 8551-8566

71. Allinger, N. L.; Chen, K.; Lii, J.-H. *J. Comp. Chem.* **1996**, *17*, 642-668

72. Halgren, T. *J. Comp. Chem.* **1996**, *17*, 490-519

73. Schuler, L. D.; Daura, X.; van Gunsteren, W. F. *J. Comp. Chem.* **2001**, *22*, 1205-1218

74. Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. *J. Comp. Chem.* **1983**, *4*, 187-217

75. Case, D. A.; Cheatham, T. E.; Darden, T.; Gohlke, H.; Luo, R.; Merz, K. M.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. J. *J. Comput. Chem.* **2005**, *26*, 1668-1688

76. Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. *J. Am. Chem. Soc.* **1996**, *118*, 11225-11236

77. Woodcock, L. V.; Angell, C. A.; Cheeseman, P. The *J. Chem. Phys.* **1976**, *65*, 1565-1577

78. Sanders, M. J.; Leslie, M.; Catlow, C. R. A. *J. Chem. Soc., Chem. Commun.* **1984**, *19*, 1271-1273

79. Kramer, G.; Farragher, N.; van Beest, B.; van Santen, R. *Phys. Rev. B* **1991**, *43*, 5068-5080

80. Hill, J.; Sauer, J. *J. Phys. Chem.* **1994**, *98*, 1238-1244

81. Hill, J.; Sauer, J. *J. Phys. Chem.* **1995**, *99*, 9536-9550

82. Schroder, K. P.; Sauer, J. *J. Phys. Chem.* **1996**, *100*, 11043-11049

83. Sierka, M.; Sauer, J. *Faraday Disc.* **1997**, *106*, 41-62

84. Blake, N. P.; Weakliem, P. C.; Metiu, H. *J. Phys. Chem. B* **1998**, *102*, 67-74

85. Bussai, C.; Hannongbua, S.; Fritzsche, S.; Haberlandt, R. *Chem. Phys. Lett.* **2002**, *354*, 310-315

86. van Duin, A. C. T.; Strachan, A.; Stewman, S.; Zhang, Q.; Xu, X.; Goddard, W. A. *J. Phys. Chem. A* **2003**, *107*, 3803-3811

87. Fuchs, A. H.; Cheetham, A. K. The *J. Phys. Chem. B* **2001**, *105*, 7375-7383

88. Smit, B.; Maesen, T. L. *Chem. Rev.* **2008**, *108*, 4125-4184.

89. Scott, A. P.; Radom, L. *J. Phys. Chem.* **1996**, *100*, 16502-16513.

90. Ewig, C.; Berry, R.; Dinur, U.; Hill, J.; Hwang, M.; Li, H.; Liang, C.; Maple, J.; Peng, Z.; Stockfisch, T.; Thacher, T.; Yan, L.; Ni, X.; Hagler, A. *J. Comp. Chem.* **2001**, *22*, 1782-1800

91. Norrby, P.-O.; Brandt, P. *Coord. Chem. Rev.* **2001**, *212*, 79-109

92. Ganda-Kesuma, F. S.; Miller, K. J. *J. Comp. Chem.* **1994**, *15*, 1291-1301

93. Verstraelen, T.; Van Neck, D.; Ayers, P. W.; Van Speybroeck, V.; Waroquier, M. *J. Chem. Theory Comput.* **2007**, *3*, 1420-1434

94. van der Kamp, M. W.; Shaw, K. E.; Christopher,; Mulholland, A. J. *J. R. Soc. Interface* **2008**, *5*, 173-190

95. Wang, Z.-X.; Zhang, W.; Wu, C.; Lei, H.; Cieplak, P.; Duan, Y. *J. Comp. Chem.* **2006**, *27*, 781-790

96. Patel, S.; Brooks, C. L. *J. Comp. Chem.* **2004**, *25*, 1-16

97. Banks, J. L.; Kaminski, G. A.; Zhou, R.; Mainz, D. T.; Berne, B. J.; Friesner, R. A. The *J. Chem. Phys.* **1999**, *110*, 741-754

98. Maple, J. R.; Cao, Y.; Damm, W.; Halgren, T. A.; Kaminski, G. A.; Zhang, L. Y.; Friesner, R. A. *J. Chem. Theory Comput.* **2005**, *1*, 694-715.

99. Patel, S.; Mackerell, A. D.; Charles, *J. Comp. Chem.* **2004**, *25*, 1504-1514

100. Foresman, J. B.; Charles, *J. Chem. Phys.* **1987**, *87*, 5892-5894

101. Zhong, Y.; Warren, L. G.; Patel, S. *J. Comp. Chem.* **2008**, *29*, 1142-1152

102. Lewis, G. V.; Catlow, C. R. A. *J. Phys. C* **1985**, *18*, 1149-1161

103. Dick, B.; Overhauser, A. *Phys. Rev.* 1958, 112, 90-102

104. Applequist, J.; Carl, J.; Fung, K. *J. Am. Chem. Soc.* **1972**, *94*, 1952-2960

105. York, D. M.; Yang, W. The *J. Chem. Phys.* **1996**, *104*, 159-172

106. Chelli, R.; Procacci, P. *J. Chem. Phys.* **2002**, *117*, 9175-9189

107. Van Genechten, K. A.; Mortier, W. J.; Geerlings, P. *J. Chem. Phys.* **1987**, *86*, 5063-5071

108. Rappe, A.; Goddard, W. *J. Phys. Chem.* **1991**, *95*, 3358-3363

109. Yang, Z. Z.; Wang, C. S. *J. Phys. Chem. A* **1997**, *101*, 6315-6321

110. Njo, S. L.; Fan, J.; van de Graaf, B. *J. Mol. Cat. A* **1998**, *134*, 79-88

111. Louwen, J. N.; Vogt, E. T. *J. Mol. Cat. A* **1998**, *134*, 63-77

112. Menegon, G.; Shimizu, K.; Farah, J. P. S.; Dias, L. G.; Chaimovich, H. *Phys. Chem. Chem. Phys.* **2002**, *4*, 5933-5936

113. Bultinck, P.; Langenaeker, W.; Lahorte, P.; De Proft, F.; Geerlings, P.; Waroquier, M.; Tollenaere, J. P. *J. Phys. Chem. A* **2002**, *106*, 7887-7894.

114. Bultinck, P.; Langenaeker, W.; Lahorte, P.; De Proft, F.; Geerlings, P.; Van Alsenoy, C.; Tollenaere, J. P. *J. Phys. Chem. A* **2002**, *106*, 7895-7901.

115. Gilson, M. K.; Gilson, H. S.; Potter, M. J. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1982-1997

116. Bultinck, P.; Vanholme, R.; Popelier, P. L. A.; De Proft, F.; Geerlings, P. *J. Phys. Chem. A* **2004**, *108*, 10359-10366

117. Yang, Q.; Sharp, K. A. *J. Chem. Theory Comput.* **2006**, *2*, 1152-1167

118. Varekova, R. S.; Jirouskova, Z.; Vanek, J.; Suchomel, S.; Koca, J. *Int. J. Mole. Sci.* **2007**, *8*, 572-582

119. Berente, I.; Czinki, E.; Náray-Szabó, G. *J. Comp. Chem.* **2007**, *28*, 1936-1942

120. Chelli, R.; Procacci, P.; Righini, R.; Califano, S. *J. Chem. Phys.* **1999**, *111*, 8569-8575

121. Warren, L. G.; Davis, J. E.; Patel, S. *J. Chem. Phys.* **2008**, *128*, 144110

122. Bauer, B. A.; Patel, S. J. Mol. Liq. **2008**, 142, 32-40

123. Giese, T. J.; York, D. M. *J. Chem. Phys.* **2004**, *120*, 9903-9906

124. Chelli, R.; Schettino, V.; Procacci, P. *J. Chem. Phys.* **2005**, *122*, 234107

125. CP2K: General Program to Perform Molecular Dynamics Simulations, http://cp2k.berlios.de/

126. Schoeman, B. J.; Sterte, J.; Otterstedt, J. E. *Zeolites* **1994**, *14*, 110-116

127. Persson, A. E.; Schoeman, B. J.; Sterte, J.; Otterstedt, J. E. *Zeolites* **1995**, *15*, 611-619

128. Jacobs, P. A.; Derouane, E. G.; Weitkamp, J. *J. Chem. Soc., Chem. Commun.* **1981**, 591-593

129. Coudurier, G.; Naccache, C.; Vedrine, J. C. *J. Chem. Soc., Chem. Commun.* **1982**, 1413-1415

130. Mohamed, R.; Aly, H.; Elshahat, M.; Ibrahim, I. *Microporous Mesoporous Mat.* **2005**, 79, 7-12

131. Serrano, D. P.; Grieken, R. *J. Mater. Chem.* **2001**, *11*, 2391-2407

132. Hsu, C.-Y.; Chiang, A. S.; Selvin, R.; Thompson, R. W. *J. Phys. Chem. B* **2005**, *109*, 18804-18814

133. Lesthaeghe, D.; Vansteenkiste, P.; Verstraelen, T.; Ghysels, A.; Kirschhock, C. E.; Martens, J. A.; Speybroeck, V. V.; Waroquier, M. *J. Phys. Chem. C* **2008**, *112*, 9186-9191

134. Berens, P. H.; Wilson, K. R. *J. Chem. Phys.* **1981**, *74*, 4872-4882

135. Trinh, T. T.; Jansen, A. P.; van Santen, R. A. *J. Phys. Chem. B* **2006**, *110*, 23099-23106

136. Trinh, T. T.; Jansen, A. P.; van Santen, R. A.; Meijer, E. J. *J. Phys. Chem. C* **2009**, *113*, 2647-2652

137. Mora-Fonz, M. J.; Richard,; Lewis, D. W. *Ang. Chem. Int. Ed.* **2005**, *44*, 3082-3086

138. Dey, B. K.; Ayers, P. W. *J. Math. Chem.* **2009**, *45*, 981-1003.

139. Dey, B. K.; Bothwell, S.; Ayers, P. W. *J. Math. Chem.* **2007**, *41*, 1-25.

140. Dey, B. K.; Janicki, M. R.; Ayers, P. W. *J. Chem. Phys.* **2004**, *121*, 6667-6679.

141. Dellago, C.; Bolhuis, P. G.; Csajka, F. S.; Chandler, D. *J. Chem. Phys.* **1998**, *108*, 1964-1977

142. van Erp, T. S.; Moroni, D.; Bolhuis, P. G. *J. Chem. Phys.* **2003**, *118*, 7762-7774

List of Publications

Updated March 2009

Publications in International Peer-Reviewed Journals

1. Verstraelen T., Szyja B.M., Lesthaeghe D., Declerck R., Van Speybroeck V., Waroquier M., Jansen A. P. J., Aerts A., Follens L.R.A., Martens J. A., Kirschhock C.E.A. and van Santen R.A.

   Multi-level modeling of silica-template interactions during initial stages of zeolite synthesis

   Top. Catal., accepted, **2009**

2. Declerck R., De Sterck B., Verstraelen T., Verniest G., Mangelinckx S., Jacobs J., De Kimpe N., Waroquier M. and Van Speybroeck V.

   Insight into the solvation and isomerization of 3-halo-1-azaallylic anions from ab initio metadynamics calculations and NMR experiments.

   Chem. Eur. J., *15*, **2009**, 580-584

3. Verstraelen T., Van Houteghem M., Van Speybroeck V. and Waroquier M.

   MD-TRACKS: A productive solution for the advanced analysis of Molecular Dynamics and Monte Carlo simulations.

   J. Chem. Inf. Mod., *48*, **2008**, 2414-2424

4. Verstraelen T., Van Speybroeck V. and Waroquier M.

   ZEOBUILDER: a GUI toolkit for the construction of complex molecules on the nanoscale with building blocks.

   J. Chem. Inf. Mod., *48*, **2008**, 1530-1541

5. Pauwels E., Verstraelen T., De Cooman H., Van Speybroeck V. and Waroquier M.

   Temperature study of a glycine radical in the solid state adopting a DFT periodic approach: vibrational analysis and comparison with EPR experiments,

   J. Phys. Chem. B, *112*, **2008**, 7618-7630

6. Lesthaeghe D., Vansteenkiste P., Verstraelen T., Ghysels A., Kirschhock C.E.A., Martens J.A., Van Speybroeck V. and Waroquier M.
   MFI fingerprint: How pentasil-induced IR bands shift during zeolite nano-growth
   J. Phys. Chem. C, *112*, **2008**, 9186-9191

7. Pauwels E., Verstraelen T. and Waroquier M.
   Effect of temperature on the EPR properties of a rhamnose alkoxy radical: a DFT molecular dynamics study
   Spectroc. Acta Pt. A-Molec. Biomolec. Spectr., *69*, **2008**, 1388 - 1394

8. Ghysels A., Van Speybroeck V., Verstraelen T., Van Neck D. and Waroquier M.
   Calculating reaction rates with partial Hessians: validation of the MBH approach
   J. Chem. Theory Comput., *4*, **2008**, 614 - 625

9. Verstraelen T., Van Neck D., Ayers P.W., Van Speybroeck V. and Waroquier M.
   The Gradient Curves Method: An improved strategy for the derivation of molecular mechanics valence force fields from ab initio data
   J. Chem. Theory Comput., *3*, **2007**, 1420 - 1434

10. Ghysels A., Van Neck D., Van Speybroeck V., Verstraelen T. and Waroquier M.
    Vibrational Modes in partially optimized molecular systems.
    J. Chem. Phys., *126*, **2007**, 224102

11. Vansteenkiste P., Verstraelen T., Van Speybroeck V. and Waroquier M.
    Ab initio calculation of entropy and heat capacity of gas-phase n-alkanes with hetero elements O and S: ethers/alcohols and sulfides/thiols.
    Chem. Phys., *328*, **2006**, 251 - 258

Oral Contributions

1. Van Speybroeck V., Verstraelen T., Vansteenkiste P., Lesthaeghe D., Aerts A., Kirschhock C.E.A., Martens J.A. and Waroquier M.
   Formation mechanisms for new zeolite materials from a molecular modeling perspective
   NCCC, Noordwijkerhout, The Netherlands, 3-5 March 2008

2. Verstraelen T., Ayers P.W., Van Neck D., Van Speybroeck V. and Waroquier M.
   New force-field models for biporous zeolites with guest molecules
   Advanced micro- and mesoporous materials, Varna, Bulgaria, 6 - 9 September 2007

3. Vansteenkiste P., Verstraelen T., Van Speybroeck V. and Waroquier M.
   Nanoslab formation from MFI precursors with interacting TPAOH
   Advanced micro- and mesoporous materials, Varna, Bulgaria, 6 - 9 September 2007

4. Verstraelen T., Van Neck D., Ayers P.W., Van Speybroeck V. and Waroquier M.
   New methods in force-field development.
   NCCC, Noordwijkerhout, The Netherlands, 5 - 7 March 2007

5. Verstraelen T., Van Neck D., Ayers P.W., Van Speybroeck V. and Waroquier M.
   The Gradient Curves Method: An improved strategy for the derivation of molecular mechanics valence force fields from ab initio data.
   ICCMSE, Chania, Greece, 27 October - 1 November 2006

6. Verstraelen T., Van Neck D., Ayers P.W., Van Speybroeck V. and Waroquier M.
   The Gradient Curves Method: An improved strategy for the derivation of molecular mechanics valence force fields from ab initio data
   CECAM Workshop: Computational aspects of building blocks, nucleation, and synthesis of porous materials, Lyon, 28 - 31 August 2006

7. Verstraelen T., Van Speybroeck V. and Waroquier M.
   Zeobuilder: A GUI-toolkit with building algorithms for the construction of complex zeolite models
   Third EFCATS School on catalysis, Poland, 20 - 26 September 2004

Poster Presentations

Research results were presented as poster presentations on several international conferences.