Learning Molecular Representations for Thermochemistry Prediction of Cyclic Hydrocarbons and Oxygenates

Maarten R. Dobbelaere[1], Pieter P. Plehiers[1], Ruben Van de Vijver[1], Christian V. Stevens[2],

Kevin M. Van Geem[1,*]

[1]Laboratory for Chemical Technology, Department of Materials, Textiles and Chemical Engineering, Ghent University, Technologiepark 125, 9052 Gent, Belgium
[2]SynBioC Research Group, Department of Green Chemistry and Technology, Faculty of Bioscience Engineering, Ghent University, Coupure Links 653, 9000 Gent, Belgium

[*] Corresponding author: Kevin.VanGeem@UGent.be, Technologiepark 125, 9052 Gent, Belgium; Tel: +32 9 264 55 97

Keywords: Machine learning, Reaction engineering, Molecular representation, Cyclic molecules

Abstract:

Accurate thermochemistry estimation of polycyclic molecules is crucial for kinetic modeling of chemical processes that use renewable and alternative feedstocks. In kinetic model generators, molecular properties are estimated rapidly with group additivity, but this method is known to have limitations for polycyclic structures. This issue has been resolved in our work by combining a geometry-based molecular representation with a deep neural network trained on *ab initio* data. Each molecule is transformed into a probabilistic vector from its interatomic distances, bond angles and dihedral angles. The model is tested on a small experimental dataset (200 molecules) from literature, a new medium-sized set (4000 molecules) with both open-shell and closed-shell species, calculated at CBS-QB3 level with empirical corrections, and a large G4MP2-level QM9-based dataset (40000 molecules). Heat capacities between 298.15 K and 2500 K are calculated in the medium set with an average deviation of about 1.5 J mol^{-1} K^{-1} and the standard entropy at 298.15 K is predicted with an average error below 4 J mol^{-1} K^{-1}. The standard enthalpy of formation at 298.15 K has an average out-of-sample error below 4 kJ mol^{-1} on a QM9 training set size of around 15k molecules. By fitting NASA polynomials, the enthalpy of formation at higher temperatures can be calculated with the same accuracy as the standard enthalpy of formation. Uncertainty quantification by means of the ensemble standard deviation is included to indicate when molecules are evaluated that are on the edge or outside of the application range of the model.

Acknowledgements:

The authors thank Hans-Heinrich Carstensen for the quantum chemical calculations. The computational resources and services used in this work were provided by the VSC (Flemish Supercomputer Center), funded by the Research Foundation – Flanders (FWO) and the Flemish Government – department EWI. PPP and RVDV acknowledge financial support respectively from a doctoral (grant number 1150817N) and a postdoctoral (grant number 3E013419) fellowship from the Research Foundation - Flanders (FWO). The authors acknowledge funding from the European Research Council under the European Union's Horizon 2020 research and innovation programme / ERC grant agreement n° 818607.

1. Introduction

Estimating molecular properties is a crucial component of many applications in the fields of chemistry, chemical engineering and materials science. In chemical reaction engineering, the creation of kinetic models requires accurate thermodynamic data, such as enthalpies of formation, entropies and heat capacities. A typical kinetic model can contain thousands of species and tens of thousands of reactions [1, 2], with the model size increasing exponentially with the number of heavy (non-hydrogen) atoms in the reactant molecule [3]. Due to the transition to alternative feedstocks, such as lignin, plastic waste or heavy oils, new processes and hence new kinetic models must be created to understand how these feedstocks are converted into useful chemicals and chemical building blocks [4, 5]. Some of these feedstocks contain many large molecules and are converted using pyrolytic processes, in which many radical species and polycyclic compounds are generated [6]. Since determining all thermochemical properties experimentally is unfeasible, computational methods are required. In most methods, a trade-off is made between accuracy and time. In computational chemistry methods, the term "chemical accuracy" is used to indicate a level of accuracy that matches experimental accuracy [7, 8]. The most commonly used values in literature are 4 kJ mol^{-1} or 1 kcal mol^{-1} (4.184 kJ mol^{-1}) [4, 9]. Another definition for chemical accuracy is 1.4 kcal mol^{-1} (5.9 kJ mol^{-1}), which corresponds to the change in rate-determining free energy of activation at room temperature that is needed to change the rate of a chemical reaction by one order of magnitude [10]. Ruscic [9] imposes the strictest definition by stating that the 95% confidence interval should be lower than 1 kcal mol^{-1} (4.184 kJ mol^{-1}). A similar error of 1 cal mol^{-1} K^{-1} on the entropy will also cause an error of about a factor two on the rate coefficient.

High-level *ab initio* quantum chemistry methods are commonly employed to calculate thermochemical properties of core species since they are able to obtain chemically accurate results [11]. However, calculating all properties with high-level methods remains too computationally demanding. A first solution is to use density functional theory (DFT) methods. Popular functionals such as B3LYP [12, 13] are less computationally expensive than high-level *ab initio* methods, but far less accurate and hence not recommended for creating highly accurate kinetic models [14, 15]. New, more accurate functionals have been created, but at a higher computational cost [16]. Automatic kinetic network generators, such as RMG [17] and Genesys [18], make use of Benson group additivity schemes [19] to make fast and accurate predictions of thermochemical properties, without having to determine molecular geometries. Despite being able to reach chemical accuracy at a much smaller cost than *ab initio* methods [20], this method also has several disadvantages. Its applicability range is limited to molecules for which group additive values (GAV) have been calculated and determining new GAV requires time and experience. In addition, extra ring-strain corrections must be included when dealing with cyclic species. This causes extra difficulties as it is impossible to calculate ring-strain corrections for all possible ring structures. Even with extended algorithms property predictions of polycyclic species lack high accuracy [21].

In the last decade, machine learning methods have been used to predict a wide range of molecular properties. As early as 2003, neural networks were applied to improve the performance of density functional theory predictions of enthalpies of formation Regarding the data, molecular representations and models – the three major requirements for machine learning in chemical engineering – there are some variations. Because large amounts of experimental or high-level *ab initio* data are hard to find, it is common to train models on large open datasets with molecular properties calculated at a lower level-of-theory. QM9 [22] is the best-known

quantum chemical benchmark, containing geometries and properties of 134k drug-like molecules, calculated at B3LYP/6-31G(2df,p) level-of-theory. Experimental data is less available and usually scattered across the literature. As an illustration, Yalamanchi *et al.* [23] composed a small dataset of only 192 experimental datapoints for predicting enthalpies of formation with machine learning.

In machine learning applications, molecules are, in many cases, translated into a numerical vector of fixed size, which is called the molecular representation. A molecular representation can either be fixed or learned, depending on whether the algorithm will always return the same vector for a molecule (fixed) or will learn a task-specific, database-dependent vector (learned) [24]. Representations can also be distinguished based on the molecular information that is needed to create the vector. Topology-based representations only require the two-dimensional connectivity, typically with some additional atomic features. Message-passing neural networks [25], which are popular in property prediction [24, 26], use learned topology-based representations inspired by the fixed extended-connectivity fingerprints (ECFP) [27]. Geometry-based representations, on the other hand, are created from the three-dimensional molecular coordinates [28-31]. These representations are used for regression models, such as support vector regression, kernel ridge regression or artificial neural networks.

In this study, the fixed Histograms of Distances, Angles and Dihedrals (HDAD) representation by Faber *et al.* [30] is modified into a learned representation, named Gaussian Learned (GauL) HDAD. This representation is used to predict the standard enthalpy of formation, standard entropy and the heat capacity at 46 temperatures of cyclic and polycyclic hydrocarbons and oxygenates. The performance of the model on two literature datasets is presented together with the performance on a new dataset, which also contains radical species. Radical species are usually not included in datasets for machine learning, but play a crucial role in modeling of

many chemical processes such as pyrolysis or combustion. The relation between the learned GauL HDAD representation and the target property is also discussed. Finally, a comparison is made with directed message-passing neural networks [24], the state-of-the-art in machine learning-based property prediction.

2. Methods

2.1. Datasets

A dataset, named "Lignin QM", containing the standard enthalpy of formation, the standard entropy and heat capacities at 46 temperatures of 3926 cyclic and polycyclic hydrocarbons and oxygenates is constructed. Part of the data has been published by Ince [32-34], Khandavilli [35, 36] and Vermeire [37, 38]. The remainder is unpublished work by Carstensen. All properties are calculated with the CBS-QB3 method [39, 40], to which spin-orbit corrections (SOC) [41] and empirical bond additive corrections (BAC) [42] are added for calculating the enthalpy of formation. With SOC and BAC, the calculated standard enthalpy of formation is chemically accurate, as well as the standard entropy and heat capacities [43]. The standard entropy contains symmetry corrections for internal and external rotations. Lignin QM includes heat capacity values which are calculated at 298.15 K and from 300 to 2500 K with an interval of 50 K. All molecules in Lignin QM have 0 or 1 free radical, at least 1 and maximum 5 rings, ring sizes between 3 and 10 atoms, and 3 to 24 heavy (non-hydrogen) atoms. Species with atoms other than hydrogen, carbon or oxygen are not included as data availability of molecules with other heteroatoms is limited. The geometries optimized at B3LYP/6-31G(2df,p) level are available for all species, along with SMILES [44] and InChI [45].

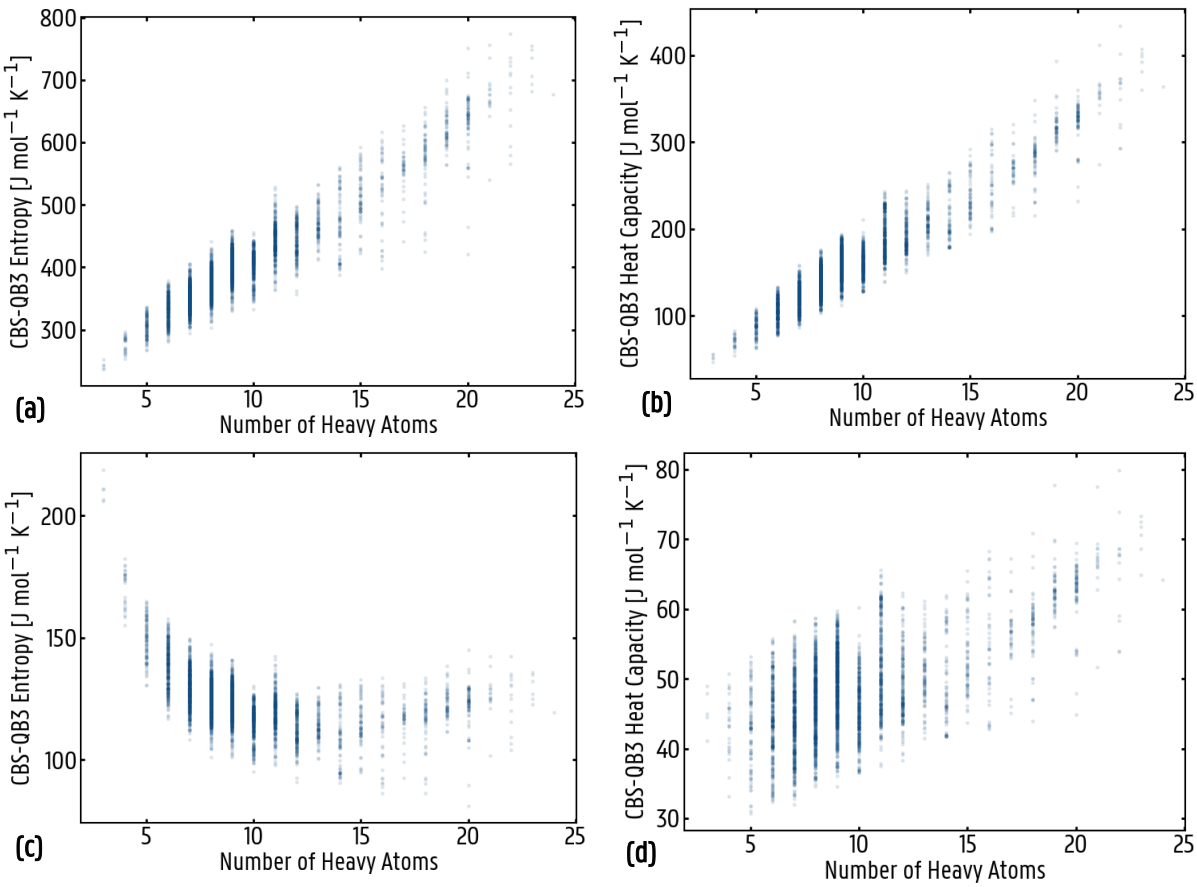


Figure 1: Distribution of the standard entropy (a) and the heat capacity at 298.15 K (b) as function of the number of heavy atoms in the Lignin QM dataset. The distributions in (c) and (d) are the normalized counterparts of respectively (a) and (b).

Figure 1a and b show the distribution of the standard entropy and the heat capacity at 298.15 K as a function of the number of heavy atoms. Both properties increase with increasing number of heavy atoms, in contrast to the standard enthalpy of formation, shown in **Figure 2**a, which does not exhibit such behavior. In machine learning it is common to rescale data to a certain range. Before training on entropy or heat capacity data, a normalization factor based depending on n_{HA}, the number of heavy atoms of the corresponding molecule, was sought to reduce the increasing behavior of the properties. This scaling factor was taken as $\ln(n_{HA})^{\frac{3}{2}}$, inspired by the

logarithmic contribution of the partition function to the entropy. The normalized distributions, which do not exhibit a strong linear trend, are illustrated in **Figure 1**c and d.

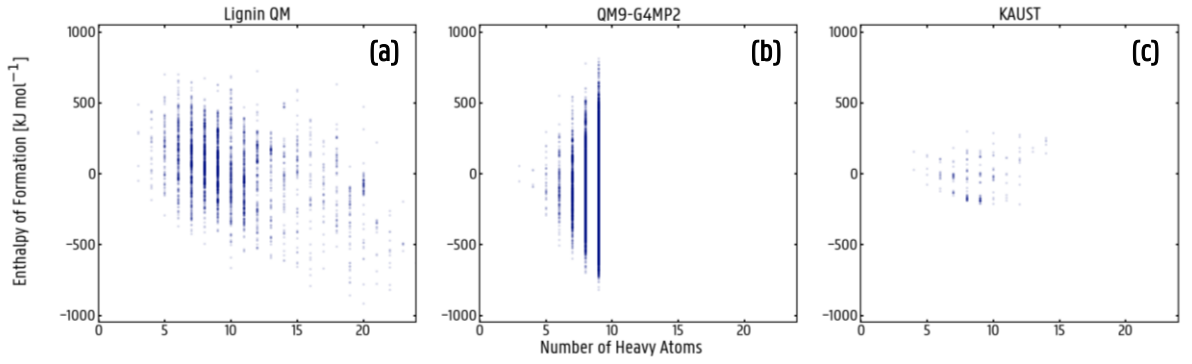


Figure 2: Distribution of enthalpies of formation as function of the number of heavy atoms in the (a) Lignin QM, (b) QM9-G4MP2 and (c) KAUST datasets

Two datasets containing enthalpies of formation of cyclic molecules are taken from literature to compare the performance of the GauL Histograms with other property prediction methods. Yalamanchi *et al*. [23] use a small dataset, suited for modeling combustion processes, to which we will refer as the KAUST dataset. It contains experimental values of 192 cyclic hydrocarbons with 4 to 14 carbon atoms, collected from the work of Ghahremanpour *et al*. [46], the CRC Handbook of Chemistry and Physics [47], and Minenkov *et al*. [48]. This dataset does not include radical species nor three-membered rings.

QM9 [22] is the most commonly used database for molecular property prediction, originally created for drug discovery purposes from the GDB-17 space of small drug-like molecules [49]. Li *et al.* [50] have trained machine learning models on QM9 subsets for kinetic modeling purposes. However, since it contains many highly strained structures that are of low importance in combustion and pyrolysis, and lacks reactive intermediates (e.g. radical species), QM9 should only be used, in this context, for testing machine learning models and for transfer learning [51]. Curtiss and co-workers [52] calculated the enthalpy of formation for all 133296 molecules in QM9 with the high level-of-theory G4MP2 method [53]. From this QM9-G4MP2

dataset, 42161 cyclic and polycyclic hydrocarbons and oxygenates are extracted to evaluate the learning curve of GauL Histograms. Figure 2 shows the distribution of enthalpies of formation as function of the number of heavy (non-hydrogen) atoms in the molecule. Lignin QM contains more large molecules than the other data sets and covers an enthalpy range of over 1700 kJ mol^{-1}, which is similar to the range of QM9-based datasets. The enthalpy values in the KAUST dataset are spread over a much smaller range, within the application range of Lignin QM.

2.2. Representation

The original Histograms of Distances, Angles and Dihedrals (HDAD) are created by Faber and Hutchison, and outperformed other methods for predicting atomization energies in the QM9 dataset when used in combination with kernel ridge regression [30]. New methods have outperformed HDAD [31, 54-58], but due to the simplicity of the method and the possibility to physically understand the representation, HDAD is chosen as a starting point. However, this representation also has some disadvantages. It is a fixed, geometry-based molecular representation which requires manual selection of the important features. In this study, HDAD is upgraded into **Gau**ssian **L**earned (GauL) HDAD – an automated, learned representation that incorporates on-the-fly geometry generation. The workflow for creating a molecular representation is illustrated in Figure 3.

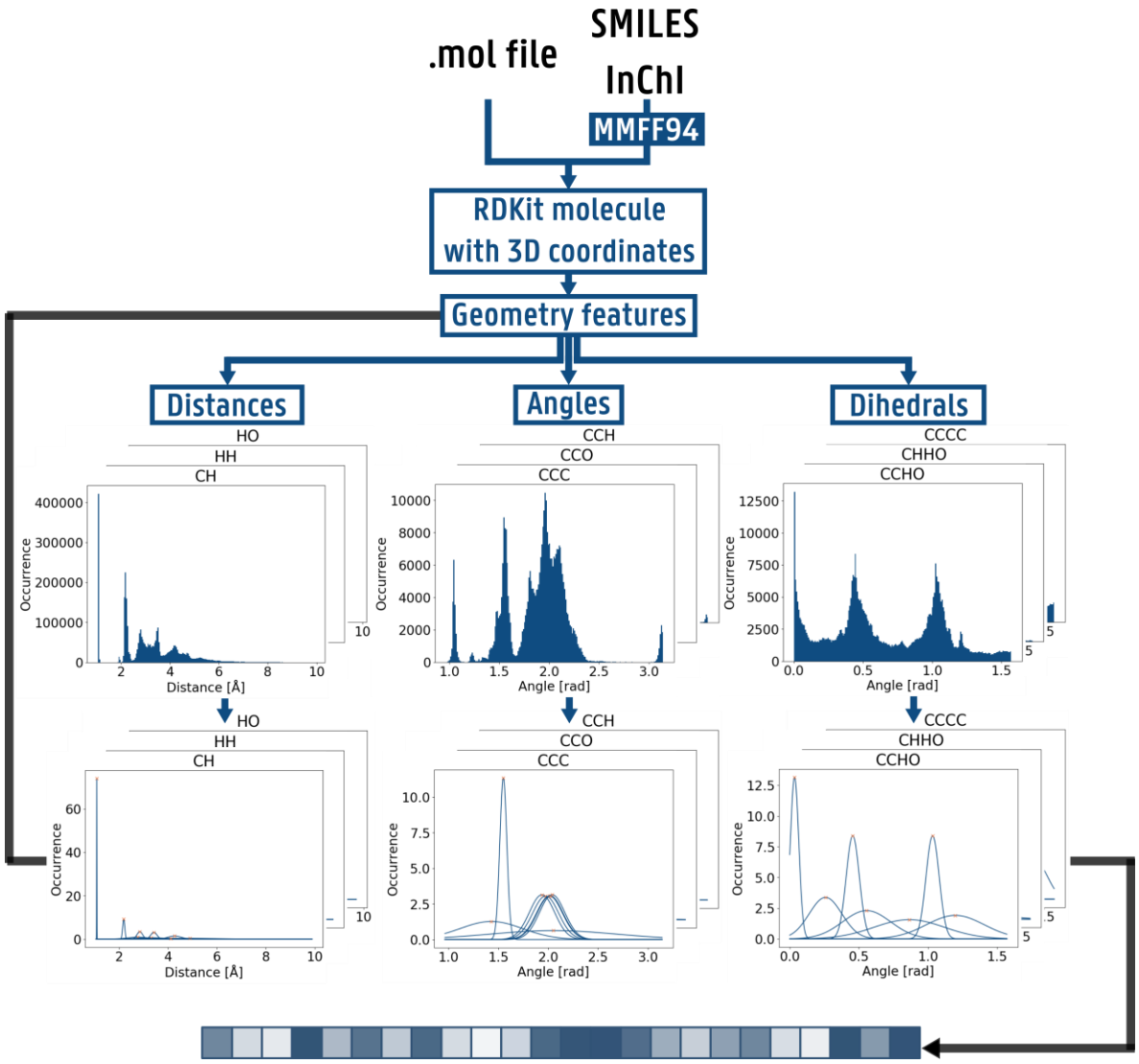


Figure 3: Workflow for creating a GauL HDAD molecular representation from a string-based identifier or a 3D molecular geometry.

2.2.1. Input

The first step consists of converting the molecule into an RDKit [59] object. Three input formats are accepted: 3D molecular coordinates saved in `.mol` files, SMILES [44] and InChI [45]. When a string-based identifier (SMILES or InChI) is used, 3D coordinates are generated in RDKit by embedding [60] and optimized with the MMFF94 forcefield [61].

2.2.2. Geometry Features

All interatomic distances, bond angles and dihedral angles in each molecule are collected and labeled. For a given molecule with *n* atoms, the interatomic distance features are the Euclidean distances between two atoms a_i $(i = 1, \ldots, n-1)$ and a_j $(j = i + 1, \ldots, n)$. Notice that this includes many more distances than just strict bond lengths. As a result, interactions between non-nearest neighbor atoms are included. Distances are labeled by the symbols of the two atoms, sorted alphabetically. A distance between a carbon atom and a hydrogen atom is, thus, labeled CH. Due to the importance of CC and CO distances in hydrocarbons and oxygenates, these bonds are further divided for better coverage in the Gaussian mixture models. This is explained in detail in the supporting information.

The angular features are the bond angles formed between two consecutive bonds. Faber *et al.* calculate the angles as "the principal angles formed by the two vectors spanning from each atom a_i to every subset of 2 of its 3 neighboring atoms, a_j and a_k" [30]. This leads to large number of angles smaller than 1 radian, which have no physical meaning. In this study, these non-physical angles are not calculated. The angles are labeled in a similar way as the distances: an angle between a carbon, a hydrogen, and an oxygen atom, is labeled CHO.

Dihedral angles are the third type of geometry features collected. Dihedrals or torsion angles are the angles between the planes, formed by two consecutive bonds. The two planes have two atoms in common. Because the dihedral value can be calculated in various ways, only the positive, acute angle is taken. Therefore, the dihedral value lies between 0 and $\frac{\pi}{2}$. The labeling is done analogous to the distances and the angles. As an example, a dihedral between two carbon atoms, a hydrogen atom, and an oxygen atom, is a CCHO dihedral.

2.2.3. Histograms

The geometry features of all molecules are combined and grouped by their label type. Figure 3 shows the histograms for the carbon-hydrogen distance (CH), the bond angle between two carbon-carbon bonds (CCC) and the dihedral angle for two carbon atoms, a hydrogen atom and an oxygen atom (CCHO) in the lignin QM dataset. The carbon-hydrogen bond length is clearly recognized as a single peak in the first histogram. The other peaks represent distances between non-neighboring carbon and hydrogen atoms. In the CCC histogram, several peaks can be distinguished: a peak between 1 and 1.1 rad for the CCC angle in three-membered rings, around 1.55 rad a peak for CCC angles in four-membered rings and the two main peaks in the middle are attributed to sp^3 hybridization (1.91 rad) and sp^2 hybridization (2.1 rad). Since these histograms are for cyclic species, the peaks are broadened due to strain. The CCHO dihedral histogram shows three large peaks, of which the largest, at 0 rad, is for 4 atoms in the same plane.

2.2.4. Gaussian Mixture Models

The molecular representation is created from the histograms by first selecting the important peaks in each histogram. In Faber *et al.* [30], this is done by manually selecting relevant peaks. Because this is a cumbersome task which requires knowledge about the features, GauL Histograms use Gaussian mixture models to identify the major peaks. Counting the number of peaks (K) in a histogram remains a manual step. K is a value between 1 and 10 for each histogram. Although this step can be automated by e.g. Bayesian optimization, this is not done to limit the required computational resources.

Gaussian mixture modeling is an unsupervised machine learning technique that uses the expectation-maximization (EM) algorithm [62] to identify different clusters of unlabeled data,

in this case the different values for each distance, angle or dihedral. The input is a file with the number of peaks per histogram (i.e. per geometry label). For each label, K gaussian fits are initialized with an arbitrary standard deviation $\sigma_k = 0.5$ and mean $\mu_k = \frac{k-1}{K} \cdot r_{max} (k = 1, \dots, K)$, with r_{max} the largest value per label. The gaussian distribution is defined in eq (1).

$$g_k(r) = \frac{1}{\sigma_k \sqrt{2\pi}} \exp\left(-\frac{(r-\mu_k)^2}{2\sigma_k^2}\right) \tag{1}$$

Every iteration step j, there is an expectation (E) step and a maximization (M) step. In the expectation step, the probability w_{ik} that a distance, angle or dihedral value r_i $(i = 1, \dots, N)$ in the label dataset with size N is located at the k-th gaussian is calculated as shown in eq (2).

$$w_{ik} = \frac{g_k(r_i)}{\sum_{m=1}^{K} g_m(r_i)} \tag{2}$$

Calculating equation (2) for all values per label yields an $N \times K$ matrix $\boldsymbol{W}$. The sum of every column N_k is then the effective value assigned to mixture component k, given by eq (3).

$$N_k = \sum_{i=1}^{N} w_{ik} \tag{3}$$

After every E-step, the log-likelihood $\log \ell(\boldsymbol{\Theta})$ is calculated as a measure for the goodness-of-fit, with $\boldsymbol{\Theta}$ the complete set of parameters μ_k and σ_k. The log-likelihood is defined in eq (4).

$$\log \ell(\boldsymbol{\Theta}) = \sum_{i=1}^{N} \log \sum_{k=1}^{K} g_k(r_i) \tag{4}$$

The iterations stop when the stopping criterion, shown in eq (5), is satisfied. The tolerance value is taken as 10^{-5}. A smaller value did not improve the model performance any further.

$$\left| \frac{\log \ell(\boldsymbol{\Theta})_{new}}{\log \ell(\boldsymbol{\Theta})_{old}} - 1 \right| < 10^{-5} \tag{5}$$

When the log-likelihood has not converged, the algorithm proceeds to the M-step with recalculation of the means and standard deviations, as given by eqs (6) and (7).

$$\mu_k = \frac{1}{N_k} \sum_{i=1}^{N} w_{ik} \cdot r_i \tag{6}$$

$$\sigma_k = \frac{1}{N_k} \sum_{i=1}^{N} w_{ik} \cdot (r_i - \mu_k)^2 \tag{7}$$

If convergence is not reached after 100 iterations, the parameters at the last iteration are taken as final parameters, since the EM algorithm increases monotonically in likelihood.

2.2.5. Feature Vector

The molecular representation that is used as input in the machine learning model is a fixed-size feature vector of which the size equals the total number of gaussian fits for all histograms. Fixed-size indicates that every molecule in the same dataset is represented by a vector of equal length. The size of this vector can be different when using a different dataset, due to the prevalence of different distances, angles and dihedrals. For every distance, angle and dihedral in a molecule, w_{ik} is calculated with eq (2). This value is the probability that a feature value is found under a certain gaussian. For each molecule, the vectors of all features are condensed to a single feature vector by summing all contributions to each bin. An extra value is added to that vector indicating whether or not the molecule has a free radical, which is the eventual molecular representation.

2.3. Machine Learning Models

The fixed-size feature vector is used as input for an artificial neural network (ANN). All ANN models are implemented in Python using the deep learning framework Keras [63], with

TensorFlow [64] as backend. The complete code is available as open-source software on the GitHub repository https://github.com/mrodobbe/GauL-HDAD.

The ANN architecture is optimized with the Hyperband algorithm [65], as implemented in the Keras-tuner package [66]. All ANN have a depth of 5 with the middle layer being significantly smaller than the other layers. The hidden layers have bias, have no dropout layers and are connected with each other using a leaky ReLU [67] activation function. Depending on the dataset, the hidden layer sizes are different. The different architectures are included in the supporting information.

Nested cross-validation, with 10 folds in the outer loop, is used for splitting the datasets, with fixed random seed for reproducibility. Within each of the 10 outer folds, another k-fold cross-validation is run to determine the best internal training and validation set out of the 90% training data. In the inner loop, the number of folds is set to 9, which corresponds to an eventual training/validation/test ratio of 80/10/10. The inner loop model with the lowest validation root-mean-square error (RMSE) and the inner loop ensemble are chosen to test the test set of the corresponding outer fold.

2.4. Directed Message-Passing Neural Networks

Message-passing neural networks (MPNN) are topology-based (i.e. starting from the molecular graph) property prediction models that have gained a lot of interest because of their accuracy [68-70]. In this work, we use the state-of-the-art directed MPNN that is implemented in the open-source software package chemprop [24], which is available at https://github.com/chemprop/chemprop. For general information about MPNN, the reader is referred to Gilmer *et al.* [25], and for details about chemprop, we refer to the original paper by Yang *et al.* [24]. The performance on the Lignin QM dataset is tested with 10-fold cross

validation, considering the same splits as for the GauL HDAD model. Bayesian optimization [71] with the python package Hyperopt [72], as implemented in chemprop, is performed to determine the appropriate hyperparameters.

3. Results and Discussion

3.1. Lignin QM

3.1.1. Prediction Accuracy

The Lignin QM prediction results are listed in **Table 1** by means of the mean absolute error (MAE) and root-mean-square error (RMSE) of the ensembles, averaged over all folds with their respective standard deviation. For all properties, GauL HDAD outperforms chemprop, with a difference in RMSE of nearly a factor 2. The difference in MAE is smaller, which indicates that there are some large outliers present in chemprop predictions. Neither of the methods is able to reach chemical accuracy for enthalpy predictions. Standard entropy and heat capacities are predicted with absolute errors below 4 J mol^{-1} K^{-1}. To meet the strictest definition of thermochemical accuracy, the 95% confidence interval – approximately twice the RMSE – must be below 1 kcal mol^{-1} (4.184 kJ mol^{-1}) or 1 cal mol^{-1} K^{-1}. For none of the properties, this definition is met. In addition, the error on the initial *ab initio* calculations must be taken into account, as well. Therefore, this method should be seen as an additional tool next to group additivity and not as a replacement for all thermochemistry estimation methods. The machine learning model is a valuable tool for calculating properties of less important species for which group additive values or corrections are missing or not accurate. However, it also meets a sufficient accuracy for all species in a network so that key species in kinetic networks can be

identified via e.g. sensitivity analyses. It is recommended that properties of these key species are still calculated with high-level-of-theory *ab initio* methods.

Table 1: Nested 10-fold cross-validation test set performance of the enthalpy of formation ($\Delta H°_{298}$), standard entropy ($S°_{298}$) and heat capacity ($c_{p,avg}$, averaged over 46 temperatures) for the Lignin QM dataset, evaluated with GauL HDAD and chemprop. [1]For chemprop, the $c_{p,298}$ prediction accuracy is reported.

$\Delta H°_{298}$ [kJ mol^{-1}]	MAE	RMSE
GauL HDAD	9.34 ± 0.39	15.89 ± 1.22
chemprop	15.43 ± 1.54	29.67 ± 4.34
$S°_{298}$ [J mol^{-1} K^{-1}]	MAE	RMSE
GauL HDAD	3.86 ± 0.18	5.32 ± 0.35
chemprop	5.90 ± 0.51	10.79 ± 1.58
$c_{p,avg}$ [J mol^{-1} K^{-1}]	MAE	RMSE
GauL HDAD	1.47 ± 0.05	2.59 ± 0.64
chemprop[1]	3.10 ± 0.42	5.33 ± 1.82

Figure 4 shows the parity plots for the GauL HDAD model tested in the second fold, which is representative for the other folds. The heat capacities of only 6 molecules have an absolute error over 20 J mol^{-1} K^{-1}. All of these molecules count at least 16 heavy atoms, which means that these species are larger than the average molecule in the dataset. In **Figure 1** and **Figure 2**, it was illustrated that the data for molecules with over 15 heavy atoms is rather scarce. The heat capacities that are poorly predicted by chemprop are large molecules too, often the same species that are predicted poorly by GauL HDAD. However, when one very similar molecule (e.g. cis-trans isomers) is available, GauL HDAD is far more accurate than chemprop. A possible explanation is that GauL HDAD can better estimate the size of a molecule, while chemprop – a convolutional model – focuses more on the functional groups in the molecule.

The entropy predictions are in many respects similar to heat capacity predictions. Most of the poor predictions from GauL HDAD correspond to large molecules in scarce data regions. However, nearly all the molecules with the largest errors for entropies predicted by chemprop count five or fewer heavy atoms. This behavior can be linked with **Figure 1**c. Due to the normalization of entropies by $\ln(n_{HA})^{\frac{3}{2}}$, the normalized value for small species is higher than for larger molecules. The result is a systematic underestimation for small molecules by chemprop. It is suggested by the entropy and heat capacity results that GauL HDAD is the better option when predicting properties that are strongly geometry-related.

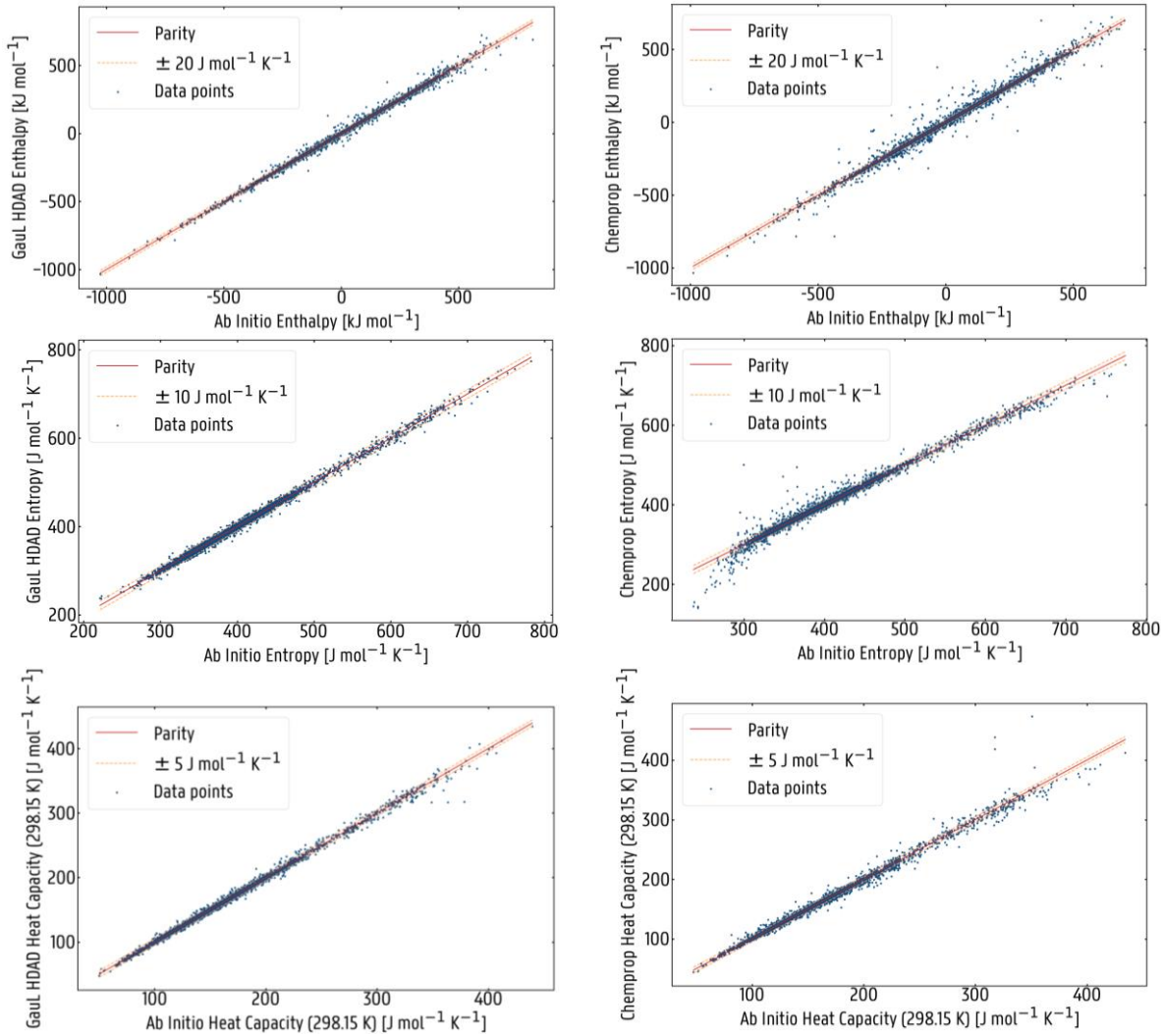


Figure 4: Parity plots for GauL HDAD (left) and chemprop (right) predictions of Lignin QM standard enthalpy of formation (top), standard entropy (middle) and heat capacity at 298.15 K (bottom) with error bars of respectively 20 kJ mol^{-1}, 10 J mol^{-1} K^{-1} and 5 J mol^{-1} K^{-1}.

3.1.2. Enthalpy Prediction at Higher Temperatures

Most chemical processes are not performed at room temperature and therefore thermochemistry at higher temperatures is needed. In combustion processes, the engineering enthalpy of formation $\Delta H_f(T)$ is calculated from the standard enthalpy of formation $\Delta H^0_{f,298.15\,K}$ and the heat capacity c_p as given by eq (8).

$$\Delta H_f(T) = \Delta H^0_{f,298.15\,K} + \int_{298.15\,K}^{T} c_p(T)dT \quad (8)$$

The heat capacity function is obtained by fitting the heat capacity at different temperatures to the found NASA polynomial coefficients [73], as shown in eq (9), where *R* is the ideal gas constant (8.314 J mol^{-1} K^{-1}) .

$$c_p(T) = R(a_1 + a_2T + a_3T^2 + a_4T^3 + a_5T^4) \quad (9)$$

The prediction errors on the engineering enthalpy of formation at a selection of temperatures is given in

Table 2, by means of the MAE and the RMSE. Notice that the value at 298.15 K is, by definition, equal to the standard enthalpy of formation, reported in **Table 1**. There is only a small increase in error noticed at higher temperatures, smaller than the error on single heat capacity values. Since the heat capacity error is almost removed by polynomial fitting, the error on higher-temperature enthalpy of formation values can be approximated by the standard enthalpy of formation error.

Table 2: Prediction error on the engineering enthalpy at different temperatures. All results in kJ mol^{-1}.

T[K]	298.15	600	1000	1500	2000	2500
MAE	9.34	9.41	9.47	9.52	9.57	9.67
RMSE	15.89	15.95	16.00	16.02	16.04	16.11

3.1.3. Interpretability of Enthalpy Predictions

Standard entropy and heat capacity are molecular properties that are strongly related to the size of a molecule and their prediction errors are related to data scarcity for molecules with more than 15 heavy atoms. Enthalpy of formation is a molecular property that does not increase with the number of heavy atoms (see also **Figure 2**) and prediction errors cannot simply be assigned to this data scarcity. The first explanation for larger deviations for enthalpy of formation predictions is the range of the values. While standard entropy and heat capacity values span a range of around $500\,\mathrm{J\,mol^{-1}\,K^{-1}}$, this is around $1700\,\mathrm{kJ\,mol^{-1}}$ for enthalpy values. The second explanation of why some molecules have poorer estimations is related to the GauL HDAD representation and how the neural network learns to predict an enthalpy from this representation. The artificial neural network has an architecture with five hidden layers. Since the size of the third hidden layer is significantly smaller, it is regarded as the actual learned molecular representation.

In **Figure 5**, the middle layer of the GauL HDAD model, trained on the Lignin QM Enthalpy dataset, is graphically represented. The middle layer dimension is first reduced to 9 with singular value decomposition, for computational reasons, and then from 9 to 2 by t-distributed stochastic neighbor embedding (t-SNE) [74]. Each circle in the figure is a molecule from the training or internal validation set and the triangles denote test set molecules. The circles are

colored by the molecule's true label (in this case the *ab initio* enthalpy of formation) and the triangles by the absolute prediction error on a logarithmic (ln) scale. Although absolute distances in t-SNE plots do not have a physical meaning, due to the stochastic nature of t-SNE, the molecules are clearly clustered into smaller groups of molecules with a similar output value. Poorly predicted molecules are usually found "within" a cluster with molecules that are not related or maximally resemble the molecules graphically.

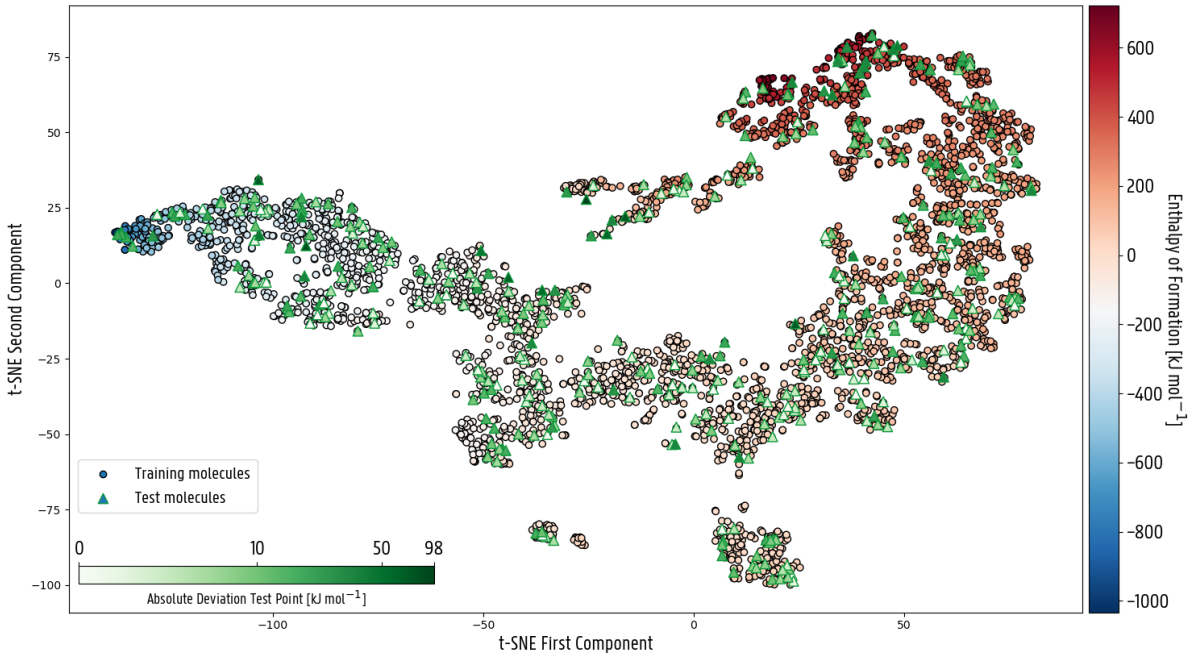


Figure 5: First and second t-SNE component for the learned molecular representation (i.e. the middle hidden layer in the ANN) for enthalpy of formation prediction in the Lignin QM dataset. The learned features clearly depend on the target property.

One such example is a 2-(hydroperoxymethyl)-5-methylfuran radical, which is the poorest prediction in this fold. **Figure 6** zooms in on this molecule's neighborhood in **Figure 5**. It is easy to recognize that most of the molecules in this small cluster contain a phenyl radical, with two substituents and two oxygen atoms. The 2-(hydroperoxymethyl)-5-methylfuran radical does not fit in this group and is possibly just an outlier, also since only few species with peroxide

bonds are included in the dataset. It is not possible to explain from this plot why a certain value is predicted. Nevertheless, these plots have been used for fine-tuning the number of gaussians in a histogram by inspecting the differences between a poorly predicted molecule and the molecules around. It should be remarked that although t-SNE is a stochastic algorithm, which will return different plots in different runs, the clusters remain the same, but with different coordinates.

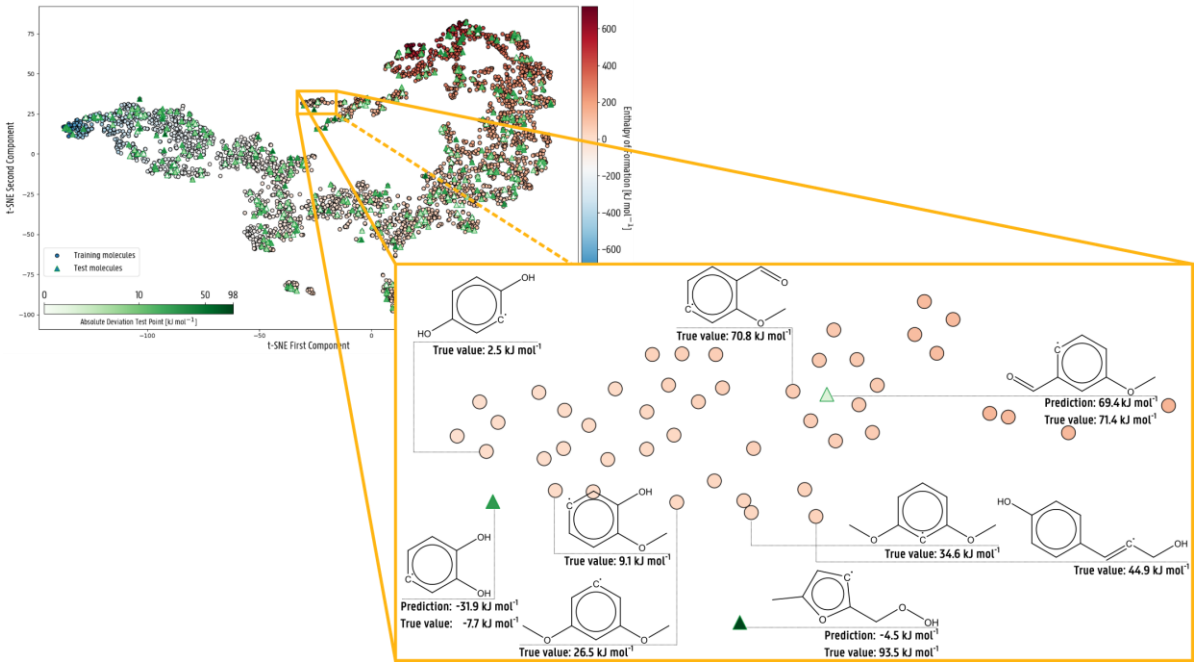


Figure 6: Zooming in on the neighborhood of the poorest predicted value in Figure 5.

3.1.4. Uncertainty Estimation

Ensemble averaging helps discovering outliers by not only averaging the predictions, but also returning the standard deviation on these predictions. The standard deviation is in first place a measure for how well the models agree with each other. However, molecules with a larger standard deviation on the predictions tend to have a larger average error too. This is also seen when taking a random cut-off at 7 kJ mol^{-1}. About 21% of the Lignin QM molecules have a standard deviation over 7 kJ mol^{-1} and the prediction RMSE for these molecules is 28.8 kJ mol^{-}

[1], compared to 15.8 kJ mol^{-1} overall. The other 79% have an RMSE of 10.0 kJ mol^{-1}. This uncertainty estimation method is ideal for active learning, as was also shown by Li *et al.* [50].

3.1.5. Input Format

GauL HDAD currently accepts molecules in three input formats: stored as a SMILES or InChI string, or as the explicit geometry stored in an individual `.mol` file. **Figure 7** illustrates that the way a molecule is stored has an effect on the eventual prediction accuracy. For these results, no internal cross-validation was used and these predictions are thus not ensemble averaged. As already mentioned above, when the input is a string-based identifier, the three-dimensional coordinates are generated by embedding and optimized with a force field. Computationally there are no differences – RDKit can optimize geometries of hundreds of molecules in less than a minute. However, the performance of forcefield optimized geometries is inferior to DFT optimized geometries. This can be due to the selection of different conformers, in which DFT is more consistent than forcefields, but also because forcefield-optimized geometries are less accurate than DFT optimized geometries. More surprising is the significant difference in enthalpy prediction accuracy between SMILES and InChI, which is an RDKit issue. The role of the molecular geometry is less important in entropy and heat capacity values than for the prediction of the enthalpy of formation. When considering different conformers, consistent and accurate geometry calculation is crucial.

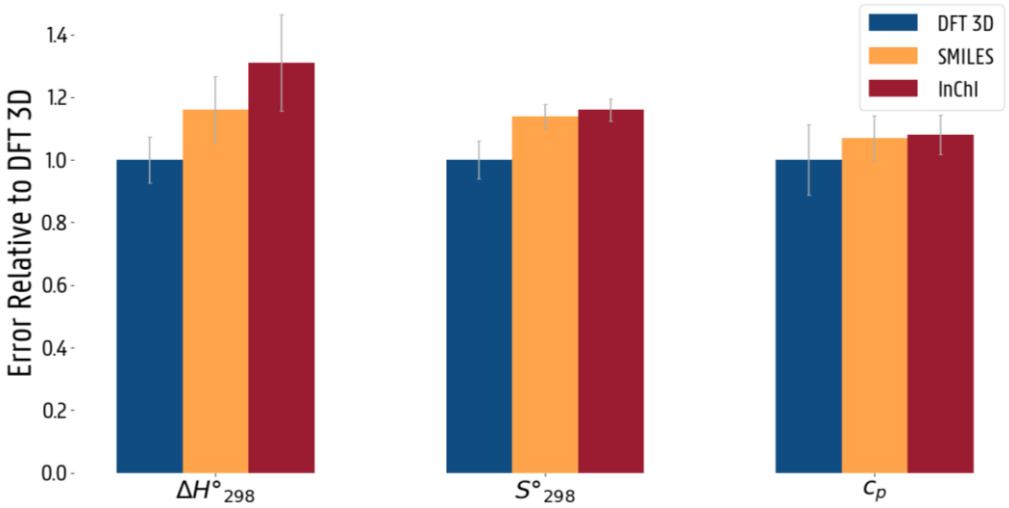
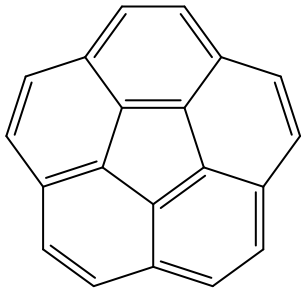


Figure 7: Comparison of the three GauL HDAD input modes, tested on the Lignin QM dataset. Lower is better.

Corannulene (IUPAC name: dibenzo[*ghi,mno*]fluoranthene [75]) is a polycyclic aromatic hydrocarbon, which can be visualized as the hydrogen-terminated cap of buckminsterfullerene (C_{60}) [76]. **Figure 8** shows its molecular structure. This $C_{20}H_{10}$ isomer is predicted poorly with all input methods and has the largest standard deviation when using ensemble averaging.



dibenzo[ghi,mno]fluoranthene

Figure 8: The 2D molecular structure of corannulene with its IUPAC name.

However, the absolute error when using DFT optimized coordinates is far lower than when geometries are embedded and optimized using forcefields. Corannulene is an example of where the forcefield approach fails. This structure has a bowl-shaped curvature [77], but due the aromatic rings, the whole structure is embedded in the xy-plane. The lack of a z-axis gradient

leads to an "optimized" planar structure. Since there are no similar molecules in the dataset, it is – even with correct geometry – predicted inaccurately.

A numerous number of species has a nearly identical geometry (*i.e.* equal distances, angles and dihedrals) when optimized with forcefields or with DFT. Yet, the prediction for one input type is significantly less accurate than for another. Since neural networks are black-box models, retracing what the model learns, is a nearly impossible task. However, the histograms (see **Figure 3**) can be compared for all input modes and there, small difference are noticed, especially for "rare" bond lengths. Inconsistent calculation of these distances can lead to different histograms and hence different Gaussian mixture models. Since the feature vector representation is directly impacted by the mixture models, a similar geometry can be represented differently, depending on how the distances, angles and dihedrals are calculated.

However, this does not explain yet why the forcefield-optimized geometries differ when using SMILES or InChI. Nearly all molecules for which there is a large difference in prediction between SMILES and InChI are either (oxygenated) polycyclic aromatic hydrocarbons or radical species with unsaturated bonds. These species have resonance structures and delocalized electrons. InChI identifiers do not take mesomerism into account [45] and the structure is systematically deviating from the SMILES. Both the SMILES and InChI were, however, defined in RDKit based on the `.mol` files in the dataset. For radical species, this leads to the radical being assigned to another atom, which results in different distances, angles and dihedrals. Polycyclic aromatic hydrocarbons, such as butalene (visualized as two fused cyclobutadiene molecules) are seen as non-aromatic species. When using InChI input, the cross-ring bond is seen as a rather short bond of about 1.309 Å, while for SMILES this bond length is 1.521 Å and B3LYP calculates it as 1.570 Å, which is also found by Warner and Jones [78]. All other bond lengths in this molecule were underestimated when using InChI input, which

makes it a different molecule for GauL HDAD. This problem is related to the InChI definition and not related to RDKit or the GauL algorithm. When handling radical and aromatic species, it is recommended to make use of SMILES instead of InChI, which are not appropriate for models involving mesomeric radicals.

3.2. Literature Datasets

3.2.1. QM9-G4MP2

The GauL HDAD algorithm is trained and tested on the standard enthalpy of formation values of (poly)cyclic hydrocarbons and oxygenates in QM9-G4MP2, to evaluate the model's learning curve. Using the same approach as above and after optimizing hyperparameters, an MAE of 2.52 ± 0.06 kJ mol^{-1} and an RMSE of 4.36 ± 0.46 kJ mol^{-1} are obtained for a training set size of 37945 molecules and using B3LYP-level geometries.

Figure 9 presents the prediction errors (MAE and RMSE) as a function of the training set size. The training set size is varied between 100 and 37945 molecules (90% of the dataset) and the test set is always the remaining molecules in the dataset; the test set size ranges from 42061 to 4216 molecules. In **Figure 9**, the shaded area is defined as the average plus/minus one times the standard deviation over the different folds. The MAE drops below 4.184 kJ mol^{-1} at a training set size of around 15000 molecules. The model systematically improves with increasing the training set size and the errors exhibit a linear decay on a log-log scale. Von Lilienfeld *et al.* [79] reported similar learning curves for different representations, evaluated on atomization energies in the original QM9.

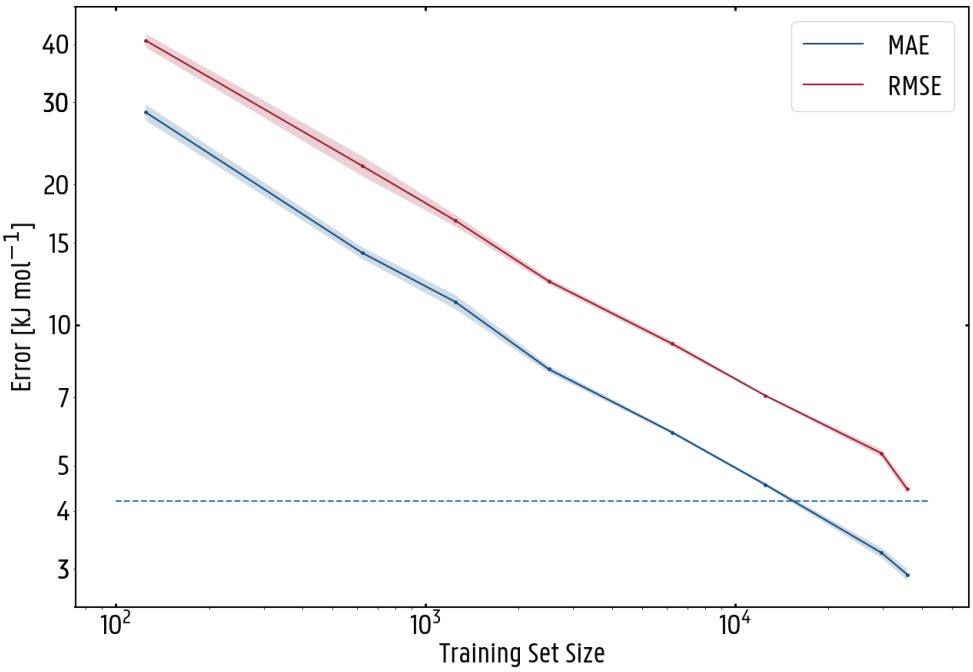


Figure 9: Learning curve for the enthalpy of formation of QM9-G4MP2 molecules (42161 cyclic and polycyclic hydrocarbons and oxygenates) predicted with the GauL HDAD method. Mean absolute error (MAE) and root-mean-square error (RMSE) as a function of the training set size, the test set being all other molecules in the dataset. The shaded area is defined by the standard deviation on the predictions. The dashed horizontal line corresponds to chemical accuracy.

3.2.2. KAUST

Yalamanchi *et al.* [23] trained a support vector regression model on a set of 192 experimentally measured enthalpies of formation [46-48]. This dataset of cyclic hydrocarbons is used to evaluate the performance on small datasets. Two molecules of which the SMILES could not be parsed by RDKit are omitted, reducing the dataset to 190. The results are reported in **Table 3**, by means of the mean absolute error (MAE), the root-mean-square error (RMSE) and the maximal absolute error (maxAE). Six models are presented in **Table 3**, of which KAUST contains the results as they were reported by Yalamanchi *et al.* [23]. The GauL HDAD algorithm is trained on the KAUST dataset, following the same procedure as above. Since the

geometries of the molecules are not available in the dataset, the SMILES identifier is used as input for the algorithm. There is no significant difference between the MAE of the KAUST algorithm and GauL HDAD, but the GauL RMSE is smaller, as well as the maximal error.

Since the molecules in the KAUST dataset are located inside of the Lignin QM range, they are tested without any training on the KAUST dataset (Lignin in **Table 3**). Remarkably, the accuracy is almost exactly the same value as when trained on the KAUST dataset. Two important notes must be considered: the molecules are tested on a model that used SMILES as input and 30 molecules are overlapping in both datasets, albeit with different labels. Although the model that uses DFT coordinates is more accurate, the test accuracy is much poorer. The reason for the performance difference that the consistency in the calculation of geometry features is crucial, as discussed above. The complete Lignin QM consists of about 2000 oxygenated species, while the KAUST dataset contains hydrocarbons only. For that reason, the oxygenates are excluded and the KAUST dataset is tested on a model trained on hydrocarbons from Lignin QM only (Lignin HC in **Table 3**). This does, however, not improve, but worsen the accuracy. It is suggested that the model trained on the complete Lignin QM dataset performs better, because it simply contains more geometric information as oxygenated species consist mainly out of carbon and hydrogen.

The test on Lignin QM models does not reflect the true model performance since the Lignin QM data are calculated at CBS-QB3 level and the KAUST dataset contains experimental data. One way to get rid of this error is by using Δ-machine learning [80], but this is not possible since not all molecules in the KAUST dataset are calculated at CBS-QB3 level. Another method is using transfer learning [51, 70], where the model trained on one dataset is used to train another dataset for a limited number of epochs, 50 in this work (Transfer Lignin and Transfer Lignin HC in **Table 3**). For the training of the KAUST molecules, 10-fold cross validation is used

again. The results improve, compared to the direct training. It is believed that the prediction error does not further drop, because the KAUST dataset is very small, compared to the Lignin QM dataset. More epochs led to overtraining and results comparable to the direct testing.

Table 3: Prediction accuracies on the KAUST dataset as reported by Yalamanchi *et al.* [23] (KAUST), predicted directly by GauL HDAD, tested after training on the full Lignin QM dataset (Lignin) and the hydrocarbons in Lignin QM (Lignin HC), using transfer learning with models trained on the full Lignin QM dataset (Transfer Lignin) and the hydrocarbons in Lignin QM (Transfer Lignin HC)

[kJ mol^{-1}]	MAE	RMSE	maxAE
KAUST	9.77	15.00	118.23
GauL HDAD	9.60	12.90	51.58
Lignin	9.81	12.97	48.65
Lignin HC	10.14	14.09	65.34
Transfer Lignin	8.20	11.19	44.43
Transfer Lignin HC	9.29	12.46	45.87

4. Conclusions

A neural network-based method is presented for high-accuracy predictions of the standard enthalpy of formation, standard entropy and heat capacity of cyclic and polycyclic hydrocarbons and oxygenates. The prediction accuracy is determined by an interplay of the training data, the molecular representation and the neural network architecture. The backbone is a dataset named “Lignin QM”, containing the standard enthalpy of formation, the standard entropy and heat capacities at 46 temperatures of 3926 cyclic and polycyclic hydrocarbons and oxygenates, calculated with high-level-of-theory *ab initio* methods. In order to capture the difference correctly between an open-shell molecule and its closed-shell equivalent, a molecular representation named GauL HDAD is developed that makes use of the three-dimensional

molecular geometry. Using a geometry that is optimized at density functional theory level leads to significantly more accurate machine learning models and predictions than when using forcefields.

GauL HDAD outperforms message-passing neural networks for predicting enthalpy, entropy and heat capacity values. Entropy and heat capacity, which are strongly geometry-related, are predicted with a mean absolute error lower than 4 J mol^{-1} K^{-1}. The standard enthalpy of formation is less influenced by the size of a molecule and needs either denser datasets or more training to become chemically accurate. Evaluation on the cyclic and polycyclic hydrocarbons and oxygenates in QM9 showed that a mean absolute error below 4 kJ mol^{-1} is obtainable with a training size of about 10k species. Outlier analysis shows that molecules with a poorly predicted enthalpy usually contain multiple distances, angles or dihedrals that are unique in the dataset. As a warning sign for using the model outside of the application range, ensemble averaging is used and the standard deviation is used as uncertainty estimation on the predictions. The high accuracies reported for thermochemical properties in a small, medium and large-sized dataset, with and without radicals, show that the GauL HDAD model is a reliable and promising prediction tool. However, the predictions are not chemically accurate and are based on *ab initio* calculations that have some deviation, too. Crucial species in kinetic models still require high-level calculations but machine learning models offer a new tool, next to group additivity, in automatic network generation tools. More open datasets containing high-level-of-theory *ab initio* data or experimental data combined with active learning and Δ-machine learning might pave the way towards chemically accurate learned thermochemistry predictions.

References

1. Lu, T. and Law, C.K., *Toward accommodating realistic fuel chemistry in large-scale computations.* Progress in Energy and Combustion Science, 2009. **35**(2): p. 192-215.
2. Battin-Leclerc, F., Blurock, E., Bounaceur, R. *et al.*, *Towards cleaner combustion engines through groundbreaking detailed chemical kinetic models.* Chemical Society Reviews, 2011. **40**(9): p. 4762-4782.
3. Van de Vijver, R., Vandewiele, N.M., Bhoorasingh, P.L. *et al.*, *Automatic Mechanism and Kinetic Model Generation for Gas- and Solution-Phase Processes: A Perspective on Best Practices, Recent Advances, and Future Challenges.* International Journal of Chemical Kinetics, 2015. **47**(4): p. 199-231.
4. Miller, J.A., Sivaramakrishnan, R., Tao, Y. *et al.*, *Combustion chemistry in the twenty-first century: Developing theory-informed chemical kinetics models.* Progress in Energy and Combustion Science, 2021. **83**: p. 100886.
5. Ranzi, E., Frassoldati, A., Stagni, A. *et al.*, *Reduced Kinetic Schemes of Complex Reaction Systems: Fossil and Biomass-Derived Transportation Fuels.* International Journal of Chemical Kinetics, 2014. **46**(9): p. 512-542.
6. Van Geem, K.M., *Chapter 6 - Kinetic modeling of the pyrolysis chemistry of fossil and alternative feedstocks*, in *Computer Aided Chemical Engineering*, T. Faravelli, F. Manenti, and E. Ranzi, Editors. 2019, Elsevier. p. 295-362.
7. Bauschlicher, C.W. and Langhoff, S.R., *Quantum Mechanical Calculations to Chemical Accuracy.* Science, 1991. **254**(5030): p. 394.
8. Pople, J.A., *Nobel Lecture: Quantum chemical models.* Reviews of Modern Physics, 1999. **71**(5): p. 1267-1274.
9. Ruscic, B., *Uncertainty quantification in thermochemistry, benchmarking electronic structure computations, and Active Thermochemical Tables.* International Journal of Quantum Chemistry, 2014. **114**(17): p. 1097-1101.
10. Cramer, C.J., *Essentials of computational chemistry: theories and models*. 2013: John Wiley & Sons.
11. Klippenstein, S.J., Harding, L.B., and Ruscic, B., *Ab Initio Computations and Active Thermochemical Tables Hand in Hand: Heats of Formation of Core Combustion Species.* The Journal of Physical Chemistry A, 2017. **121**(35): p. 6580-6602.
12. Becke, A.D., *Density-functional thermochemistry. III. The role of exact exchange.* The Journal of Chemical Physics, 1993. **98**(7): p. 5648-5652.
13. Stephens, P.J., Devlin, F.J., Chabalowski, C.F. *et al.*, *Ab Initio Calculation of Vibrational Absorption and Circular Dichroism Spectra Using Density Functional Force Fields.* The Journal of Physical Chemistry, 1994. **98**(45): p. 11623-11627.
14. Saeys, M., Reyniers, M.-F., Marin, G.B. *et al.*, *Ab Initio Calculations for Hydrocarbons: Enthalpy of Formation, Transition State Geometry, and Activation Energy for Radical Reactions.* The Journal of Physical Chemistry A, 2003. **107**(43): p. 9147-9159.
15. Klippenstein, S.J. and Cavallotti, C., *Chapter 2 - Ab initio kinetics for pyrolysis and combustion systems*, in *Computer Aided Chemical Engineering*, T. Faravelli, F. Manenti, and E. Ranzi, Editors. 2019, Elsevier. p. 115-167.
16. Goerigk, L., Hansen, A., Bauer, C. *et al.*, *A look at the density functional theory zoo with the advanced GMTKN55 database for general main group thermochemistry,*

kinetics and noncovalent interactions. Physical Chemistry Chemical Physics, 2017. **19**(48): p. 32184-32215.
17. Gao, C.W., Allen, J.W., Green, W.H. *et al., Reaction Mechanism Generator: Automatic construction of chemical kinetic mechanisms.* Computer Physics Communications, 2016. **203**: p. 212-225.
18. Vandewiele, N.M., Van Geem, K.M., Reyniers, M.-F. *et al., Genesys: Kinetic model construction using chemo-informatics.* Chemical Engineering Journal, 2012. **207-208**: p. 526-538.
19. Benson, S.W., Cruickshank, F.R., Golden, D.M. *et al., Additivity rules for the estimation of thermochemical properties.* Chemical Reviews, 1969. **69**(3): p. 279-324.
20. Sabbe, M.K., Saeys, M., Reyniers, M.-F. *et al., Group Additive Values for the Gas Phase Standard Enthalpy of Formation of Hydrocarbons and Hydrocarbon Radicals.* The Journal of Physical Chemistry A, 2005. **109**(33): p. 7466-7480.
21. Han, K., Jamal, A., Grambow, C.A. *et al., An Extended Group Additivity Method for Polycyclic Thermochemistry Estimation.* International Journal of Chemical Kinetics, 2018. **50**(4): p. 294-303.
22. Ramakrishnan, R., Dral, P.O., Rupp, M. *et al., Quantum chemistry structures and properties of 134 kilo molecules.* Scientific Data, 2014. **1**(1): p. 140022.
23. Yalamanchi, K.K., Monge-Palacios, M., van Oudenhoven, V.C.O. *et al., Data Science Approach to Estimate Enthalpy of Formation of Cyclic Hydrocarbons.* The Journal of Physical Chemistry A, 2020. **124**(31): p. 6270-6276.
24. Yang, K., Swanson, K., Jin, W. *et al., Analyzing Learned Molecular Representations for Property Prediction.* Journal of Chemical Information and Modeling, 2019. **59**(8): p. 3370-3388.
25. Gilmer, J., Schoenholz, S.S., Riley, P.F. *et al., Neural message passing for quantum chemistry.* arXiv preprint arXiv:1704.01212, 2017.
26. Coley, C.W., Barzilay, R., Green, W.H. *et al., Convolutional Embedding of Attributed Molecular Graphs for Physical Property Prediction.* Journal of Chemical Information and Modeling, 2017. **57**(8): p. 1757-1772.
27. Rogers, D. and Hahn, M., *Extended-Connectivity Fingerprints.* Journal of Chemical Information and Modeling, 2010. **50**(5): p. 742-754.
28. Rupp, M., Tkatchenko, A., Müller, K.-R. *et al., Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning.* Physical Review Letters, 2012. **108**(5): p. 058301.
29. Hansen, K., Montavon, G., Biegler, F. *et al., Assessment and Validation of Machine Learning Methods for Predicting Molecular Atomization Energies.* Journal of Chemical Theory and Computation, 2013. **9**(8): p. 3404-3419.
30. Faber, F.A., Hutchison, L., Huang, B. *et al., Prediction Errors of Molecular Machine Learning Models Lower than Hybrid DFT Error.* Journal of Chemical Theory and Computation, 2017. **13**(11): p. 5255-5264.
31. Christensen, A.S., Bratholm, L.A., Faber, F.A. *et al., FCHL revisited: Faster and more accurate quantum machine learning.* The Journal of Chemical Physics, 2020. **152**(4): p. 044107.
32. Ince, A., Carstensen, H.-H., Reyniers, M.-F. *et al., First-principles based group additivity values for thermochemical properties of substituted aromatic compounds.* AIChE Journal, 2015. **61**(11): p. 3858-3870.

33. Ince, A., Carstensen, H.-H., Sabbe, M. *et al.*, *Group additive modeling of substituent effects in monocyclic aromatic hydrocarbon radicals.* AIChE Journal, 2017. **63**(6): p. 2089-2106.
34. Ince, A., Carstensen, H.-H., Sabbe, M. *et al.*, *Modeling of thermodynamics of substituted toluene derivatives and benzylic radicals via group additivity.* AIChE Journal, 2018. **64**(10): p. 3649-3661.
35. Khandavilli, M.V., Vermeire, F.H., Van de Vijver, R. *et al.*, *Group additive modeling of cyclopentane pyrolysis.* Journal of Analytical and Applied Pyrolysis, 2017. **128**: p. 437-450.
36. Khandavilli, M.V., Djokic, M., Vermeire, F.H. *et al.*, *Experimental and Kinetic Modeling Study of Cyclohexane Pyrolysis.* Energy & Fuels, 2018. **32**(6): p. 7153-7168.
37. Vermeire, F.H., De Bruycker, R., Herbinet, O. *et al.*, *Experimental and kinetic modeling study of the pyrolysis and oxidation of 1,5-hexadiene: The reactivity of allylic radicals and their role in the formation of aromatics.* Fuel, 2017. **208**: p. 779-790.
38. Vermeire, F.H., Carstensen, H.-H., Herbinet, O. *et al.*, *Experimental and modeling study of the pyrolysis and combustion of dimethoxymethane.* Combustion and Flame, 2018. **190**: p. 270-283.
39. Montgomery, J.A., Frisch, M.J., Ochterski, J.W. *et al.*, *A complete basis set model chemistry. VI. Use of density functional geometries and frequencies.* The Journal of Chemical Physics, 1999. **110**(6): p. 2822-2827.
40. Montgomery, J.A., Frisch, M.J., Ochterski, J.W. *et al.*, *A complete basis set model chemistry. VII. Use of the minimum population localization method.* The Journal of Chemical Physics, 2000. **112**(15): p. 6532-6542.
41. Curtiss, L.A., Raghavachari, K., Redfern, P.C. *et al.*, *Assessment of Gaussian-2 and density functional theories for the computation of enthalpies of formation.* The Journal of Chemical Physics, 1997. **106**(3): p. 1063-1079.
42. Petersson, G.A., Malick, D.K., Wilson, W.G. *et al.*, *Calibration and comparison of the Gaussian-2, complete basis set, and density functional methods for computational thermochemistry.* The Journal of Chemical Physics, 1998. **109**(24): p. 10570-10579.
43. Paraskevas, P.D., Sabbe, M.K., Reyniers, M.-F. *et al.*, *Group Additive Values for the Gas-Phase Standard Enthalpy of Formation, Entropy and Heat Capacity of Oxygenates.* Chemistry – A European Journal, 2013. **19**(48): p. 16431-16452.
44. Weininger, D., *SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules.* Journal of Chemical Information and Computer Sciences, 1988. **28**(1): p. 31-36.
45. Heller, S., McNaught, A., Stein, S. *et al.*, *InChI - the worldwide chemical structure identifier standard.* Journal of Cheminformatics, 2013. **5**(1): p. 7.
46. Ghahremanpour, M.M., van Maaren, P.J., Ditz, J.C. *et al.*, *Large-scale calculations of gas phase thermochemistry: Enthalpy of formation, standard entropy, and heat capacity.* The Journal of Chemical Physics, 2016. **145**(11): p. 114305.
47. Rumble, J.R., Lide, D.R., and Bruno, T.J., *CRC Handbook of Chemistry and Physics*. 2017.
48. Minenkov, Y., Wang, H., Wang, Z. *et al.*, *Heats of Formation of Medium-Sized Organic Compounds from Contemporary Electronic Structure Methods.* Journal of Chemical Theory and Computation, 2017. **13**(8): p. 3537-3560.

49. Ruddigkeit, L., van Deursen, R., Blum, L.C. *et al., Enumeration of 166 Billion Organic Small Molecules in the Chemical Universe Database GDB-17.* Journal of Chemical Information and Modeling, 2012. **52**(11): p. 2864-2875.
50. Li, Y.-P., Han, K., Grambow, C.A. *et al., Self-Evolving Machine: A Continuously Improving Model for Molecular Thermochemistry.* The Journal of Physical Chemistry A, 2019. **123**(10): p. 2142-2152.
51. Grambow, C.A., Li, Y.-P., and Green, W.H., *Accurate Thermochemistry with Small Data Sets: A Bond Additivity Correction and Transfer Learning Approach.* The Journal of Physical Chemistry A, 2019. **123**(27): p. 5826-5835.
52. Narayanan, B., Redfern, P.C., Assary, R.S. *et al., Accurate quantum chemical energies for 133 000 organic molecules.* Chemical Science, 2019. **10**(31): p. 7449-7455.
53. Curtiss, L.A., Redfern, P.C., and Raghavachari, K., *Gaussian-4 theory.* The Journal of Chemical Physics, 2007. **126**(8): p. 084108.
54. Schütt, K.T., Sauceda, H.E., Kindermans, P.J. *et al., SchNet – A deep learning architecture for molecules and materials.* The Journal of Chemical Physics, 2018. **148**(24): p. 241722.
55. Bartók, A.P., De, S., Poelking, C. *et al., Machine learning unifies the modeling of materials and molecules.* Science Advances, 2017. **3**(12): p. e1701816.
56. Huang, B. and von Lilienfeld, O.A., *The" DNA" of chemistry: Scalable quantum machine learning with" amons".* arXiv preprint arXiv:1707.04146, 2017.
57. Pronobis, W., Tkatchenko, A., and Müller, K.-R., *Many-Body Descriptors for Predicting Molecular Properties with Machine Learning: Analysis of Pairwise and Three-Body Interactions in Molecules.* Journal of Chemical Theory and Computation, 2018. **14**(6): p. 2991-3003.
58. Eickenberg, M., Exarchakis, G., Hirn, M. *et al., Solid harmonic wavelet scattering for predictions of molecule properties.* The Journal of Chemical Physics, 2018. **148**(24): p. 241732.
59. Landrum, G. *RDKit: Open-source cheminformatics.* 2020 - Available from: https://www.rdkit.org.
60. Havel, T.F., Kuntz, I.D., and Crippen, G.M., *The theory and practice of distance geometry.* Bulletin of Mathematical Biology, 1983. **45**(5): p. 665-720.
61. Halgren, T.A., *Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94.* Journal of Computational Chemistry, 1996. **17**(5-6): p. 490-519.
62. Moon, T.K., *The expectation-maximization algorithm.* IEEE Signal Processing Magazine, 1996. **13**(6): p. 47-60.
63. Chollet, F., *keras*. 2015.
64. Abadi, M., Barham, P., Chen, J. *et al. Tensorflow: A system for large-scale machine learning*.
65. Li, L., Jamieson, K., DeSalvo, G. *et al., Hyperband: a novel bandit-based approach to hyperparameter optimization.* J. Mach. Learn. Res., 2017. **18**(1): p. 6765–6816.
66. O'Malley, T., Bursztein, E., Long, J. *et al. Keras Tuner*. 2019 - Available from: https://keras-team.github.io/keras-tuner/.
67. Maas, A.L., Hannun, A.Y., and Ng, A.Y., *Rectifier nonlinearities improve neural network acoustic models*, in *ICML Workshop on Deep Learning for Audio, Speech and Language Processing*. 2013.

68. Grambow, C.A., Pattanaik, L., and Green, W.H., *Deep Learning of Activation Energies.* The Journal of Physical Chemistry Letters, 2020. **11**(8): p. 2992-2997.
69. Stokes, J.M., Yang, K., Swanson, K. *et al.*, *A Deep Learning Approach to Antibiotic Discovery.* Cell, 2020. **180**(4): p. 688-702.e13.
70. Vermeire, F.H. and Green, W.H., *Transfer learning for solvation free energies: from quantum chemistry to experiments.* arXiv:2012.11730, 2020.
71. Pelikan, M., Goldberg, D.E., and Cantú-Paz, E. *BOA: The Bayesian optimization algorithm*. in *GECCO'99: Proceedings of the 1st Annual Conference on Genetic and Evolutionary Computation*.
72. Bergstra, J., Komer, B., Eliasmith, C. *et al.*, *Hyperopt: a Python library for model selection and hyperparameter optimization.* Computational Science & Discovery, 2015. **8**(1): p. 014008.
73. Gardiner, W.C. and Burcat, A., *Combustion chemistry*. 1984: Springer.
74. van der Maaten, L. and Hinton, G., *Visualizing data using t-SNE.* Journal of machine learning research, 2008. **9**(Nov): p. 2579-2605.
75. Barth, W.E. and Lawton, R.G., *Dibenzo[ghi,mno]fluoranthene.* Journal of the American Chemical Society, 1966. **88**(2): p. 380-381.
76. Nestoros, E. and Stuparu, M.C., *Corannulene: a molecular bowl of carbon with multifaceted properties and diverse applications.* Chemical Communications, 2018. **54**(50): p. 6503-6519.
77. Wu, Y.-T. and Siegel, J.S., *Aromatic Molecular-Bowl Hydrocarbons: Synthetic Derivatives, Their Structures, and Physical Properties.* Chemical Reviews, 2006. **106**(12): p. 4843-4867.
78. Warner, P.M. and Jones, G.B., *Butalene and Related Compounds: Aromatic or Antiaromatic?* Journal of the American Chemical Society, 2001. **123**(42): p. 10322-10328.
79. von Lilienfeld, O.A., Müller, K.-R., and Tkatchenko, A., *Exploring chemical compound space with quantum-based machine learning.* Nature Reviews Chemistry, 2020. **4**(7): p. 347-358.
80. Ramakrishnan, R., Dral, P.O., Rupp, M. *et al.*, *Big Data Meets Quantum Chemistry Approximations: The Δ-Machine Learning Approach.* Journal of Chemical Theory and Computation, 2015. **11**(5): p. 2087-2096.