Supporting Information for 'ParAMS: Parameter Optimization for Atomistic and Molecular Simulations'

Leonid Komissarov,[†,‡] Robert Rüger,[‡] Matti Hellström,[‡] and Toon Verstraelen[*,†]

†*Center for Molecular Modeling (CMM), Ghent University, Technologiepark-Zwijnaarde 46, B-9052, Ghent, Belgium*

‡*Software for Chemistry & Materials (SCM) B.V., De Boelelaan 1083, 1081 HV Amsterdam, The Netherlands*

E-mail: toon.verstraelen@ugent.be

S1 The Parameterization Problem

This section provides a mathematical framework for the parameterization problem. We assume that the training data can be defined as a set of physico-chemical properties for a number of isolated or periodic systems. Examples for relevant properties are energy differences, nuclear gradients or system geometries. In the context of ParAMS, we define an arbitrary property P that can be expressed as the output of a computational job. When working with multiple jobs as part of a training set, a job function can be defined as

$$J(R_j, S_j, M) = \left(R'_j, P^n_j\right) \quad \forall\, j \in \{1 \ldots N_{\text{job}}\},\ n \in \{1 \ldots N_{\text{prop}}(j)\}, \tag{1}$$

calculating the output geometry R'_j and all properties P^n_j of a job j. The input for every job in J consists of the input geometry R_j, the job settings S_j (*e.g.* geometry optimization and

frequencies) and the computational model M. Note that a parametric model is additionally a function of the parameter vector $\boldsymbol{x}$, in which case the outputs of the above equation can be denoted with the hat operator (*i.e.* $\hat{R}'_j, \hat{P}^n_j$), as to distinguish between reference properties and properties predicted by the parametric model. Training set entries can be constructed, for example, from a linear combination of multiple properties

$$y_i = \sum_{k=1}^{N_{\mathrm{lc}}(i)} c_{i,k} P^{n(i,k)}_{j(i,k)} \quad \forall\, i \in \{1 \dots N_{\mathrm{data}}\}, \tag{2}$$

where $c_{i,k}$ is the coefficient for term k of training set entry i and $N_{\mathrm{lc}}(i)$ is the total number of terms per entry. Non-linear combinations of properties to construct y_i are also possible. Such a formulation offers a high degree of flexibility for the construction of a training set. One example is the combination of multiple system energies into one reaction energy. It should be noted that a training set entry, as defined in Eq. 2, does not have to originate from the results of computational jobs. The reference value can instead be provided directly, making it easy to work with experimental or external data.

While training set entries $\boldsymbol{y}$ have to be defined only once, their predicted counterpart $\hat{\boldsymbol{y}}$ has to be re-calculated every time the model parameters change. For this purpose, we introduce a Data Set function

$$DS(\boldsymbol{x}|\boldsymbol{y}) = \boldsymbol{y} - \hat{\boldsymbol{y}}, \tag{3}$$

which extracts all properties needed for the calculation of $\hat{\boldsymbol{y}}$ based on a parameter set $\boldsymbol{x}$ and returns the respective vector of residuals. A metric in the form of a loss function $L\Big((\boldsymbol{y}-\hat{\boldsymbol{y}})\boldsymbol{w}\Big)$ is then applied to the residuals for a qualitative measure of how close reference and predicted values are. The additional weights vector vector $\boldsymbol{w}$ can be used to balance possibly different orders of magnitude in the data set or make certain entries more relevant for the fitting process than others.

Finally, the optimization algorithm can be defined as a function that minimizes L with

respect to the parameters

$$O(\boldsymbol{x}_0, L) = \arg\min_{\boldsymbol{x}} L = \boldsymbol{x}^*, \tag{4}$$

finding an optimal solution $\boldsymbol{x}^*$ from an initial point $\boldsymbol{x}_0$.

S2 Additional Display Items

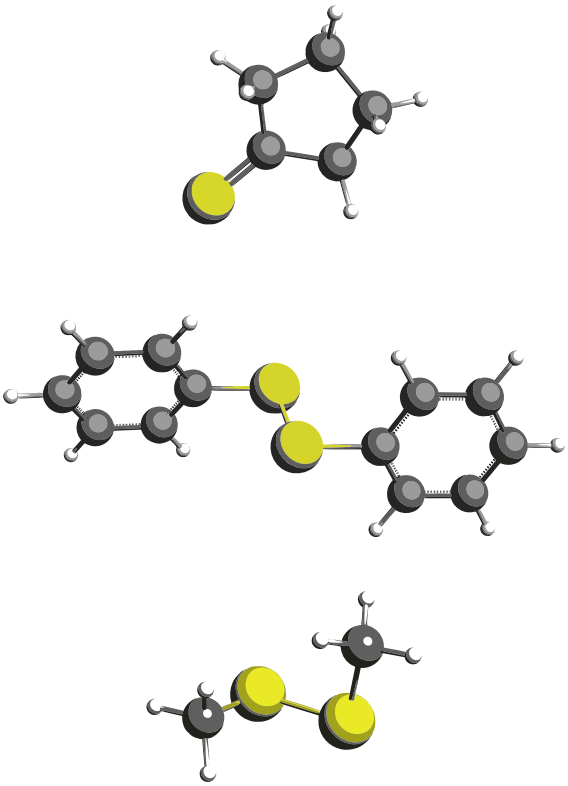

Figure S1: From top to bottom: Example structures of cyclopentathione, diphenyl disulfide and dimethyl disulfide, containing S (yellow), C (black), and H (white), included in the data provided by Müller and Hartke.[1] The fitted properties include bond distances, angles, relative energies and atomic forces.

Table S1: Composition of the reference data published by Müller and Hartke,[1] split by the computational tasks Single Point (SP) and Geometry Optimization (GO). For each of the two sets, the upper part describes the chemical systems, while the lower breaks down the individual entries in the training and validation sets. Note that some entries might be a function of multiple chemical systems, meaning that the sum of SP+GO is not necessarily equal to the total number of entries for that row (*cf.* Sec. 3.1 in the main text).

Training Set	**SP**	**GO**	**Total**
Number of systems	222	9	231
Mean system size (atoms)	6.6	11.4	6.8
Std. dev. (atoms)	2.9	7.7	3.3
Total number of entries	4620	317	4875
Energies	219	62	219
Forces	4401	0	4401
Atomic distances	0	94	94
Angles	0	85	85
Dihedrals	0	76	76
Validation Set			
Number of systems	200	24	224
Mean system size (atoms)	24.0	12.7	22.8
Std. dev. (atoms)	0.0	5.9	4.0
Total number of entries	199	771	970
Energies	199	0	199
Forces			0
Atomic distances	0	281	281
Angles	0	257	257
Dihedrals	0	233	233

Table S2: Summary of relevant ParAMS settings used for the re-parameterization of Mue2016.

Setting	**Value**
Number of optimizations	9
Number of parameters to optimize	35
Lower / upper parameter bounds	$\boldsymbol{x}_0 \pm 0.2\lvert\boldsymbol{x}_0\rvert$
Optimization timeout	24 hours
CMA-ES population size	36
CMA-ES sigma	0.3
Loss function	sum of squared errors
Early stopping patience	6000 evaluations
Constraints	$r_0^\sigma \geq r_0^\pi$ and $r_0^\pi \geq r_0^{\pi\pi}$

References

(1) Müller, J.; Hartke, B. ReaxFF Reactive Force Field for Disulfide Mechanochemistry, Fitted to Multireference ab Initio Data. *J. Chem. Theory Comput.* **2016**, *12*, 3913–3925.