Machine Learning of Parameters for Accurate Semiempirical Quantum Chemical Calculations

Journal:	*Journal of Chemical Theory and Computation*
Manuscript ID:	ct-2015-001414.R1
Manuscript Type:	Article
Date Submitted by the Author:	24-Mar-2015
Complete List of Authors:	Dral, Pavlo; Max-Planck-Institut für Kohlenforschung, von Lilienfeld, Anatole; University of Basel, Institute of Physical Chemistry Thiel, Walter; Max-Planck-Institut fuer Kohlenforschung, Theorie

Machine Learning of Parameters for Accurate Semiempirical Quantum Chemical Calculations

Pavlo O. Dral,*,† O. Anatole von Lilienfeld,‡,¶ and Walter Thiel*,†

Max-Planck-Institut für Kohlenforschung, Kaiser-Wilhelm-Platz 1, 45470 Mülheim an der Ruhr, Germany, Institute of Physical Chemistry and National Center for Computational Design and Discovery of Novel Materials, Department of Chemistry, University of Basel, Klingelbergstrasse 80, CH-4056 Basel, Switzerland, and Argonne Leadership Computing Facility, Argonne National Laboratory, Argonne, Illinois 60439, USA

E-mail: dral@kofo.mpg.de; thiel@kofo.mpg.de

Abstract

We investigate possible improvements in the accuracy of semiempirical quantum chemistry (SQC) methods through the use of machine learning (ML) models for the parameters. For a given class of compounds, ML techniques require sufficiently large training sets to develop ML models that can be used for adapting SQC parameters to reflect changes in molecular composition and geometry. The ML-SQC approach allows the automatic tuning of SQC parameters for individual molecules, thereby improving the accuracy without deteriorating transferability to molecules with molecular descriptors very different from those in the training set. The performance of this approach is

*To whom correspondence should be addressed

†Max-Planck-Institut für Kohlenforschung, Kaiser-Wilhelm-Platz 1, 45470 Mülheim an der Ruhr, Germany

‡Institute of Physical Chemistry and National Center for Computational Design and Discovery of Novel Materials, Department of Chemistry, University of Basel, Klingelbergstrasse 80, CH-4056 Basel, Switzerland

¶Argonne Leadership Computing Facility, Argonne National Laboratory, Argonne, Illinois 60439, USA

demonstrated for the semiempirical OM2 method using a set of 6095 constitutional isomers $C_7H_{10}O_2$, for which accurate *ab initio* atomization enthalpies are available. The ML-OM2 results show improved average accuracy and a much reduced error range compared with standard OM2 results, with mean absolute errors in atomization enthalpies dropping from 6.3 to 1.7 kcal/mol. They are also found to be superior to the results from specific OM2 reparameterizations (rOM2) for the same set of isomers. The ML-SQC approach thus holds promise for fast and reasonably accurate high-throughput screening of materials and molecules.

1 Introduction

In the field of *de novo in silico* materials and drug design, fast and accurate methods are required for high-throughput screening of a wide range of systems.[1–6] Density functional theory (DFT) methods are widely used[1,3,5,7] as they are robust, often sufficiently accurate, *and* universally applicable for most of the less exotic materials that can be composed of main group and transition metal elements. Typically, their computational cost is significantly smaller than that of high-level correlated *ab initio* methods.

Semiempirical quantum chemistry (SQC) and machine learning (ML) methods are much faster than DFT and may thus become viable alternatives to DFT for high-throughput screening. In fact, SQC methods have already been used for such studies, as in the search for the selective kinase inhibitors.[8] However, they may often not be accurate enough for this purpose. Their usefulness could be improved significantly by enhancing their average accuracy and transferability, and especially by reducing the number of severe outliers in the calculated properties. Parameters in SQC methods are usually fitted in a global way to reproduce available experimental observables or highly accurate quantum chemistry (QC) reference values for a broad range of reference molecules.[9,10] While this general-purpose strategy often provides acceptable average accuracy in a statistical sense,[9,10] SQC calculations may be quite inaccurate for particular compounds.[11,12]

SQC parameters are sometimes refitted specifically for some class of compounds (e.g., fullerenes[13]), for certain reactions (e.g., to study kinetic isotopic effects[14]) or for intermolecular interactions (e.g., in water[15]). The resulting special-purpose SQC approaches can achieve high accuracy by closely reproducing experimental or high-level QC data for the target systems. Of course, such special-purpose SQC methods are accurate only for the types of compounds, reactions, and properties, for which they have been reparametrized, but generally not for other targets.

ML methods can be used in computational chemistry to infer properties of new molecules through interpolation in chemical compound space.[16–18] They employ simple but flexible ad-hoc models for interpolation that are trained on a sufficiently large set of compounds and can then be used to predict the properties of related target compounds. Evaluation of an ML model is generally orders of magnitude faster than an SQC calculation, but due to the lack of rigorous physical approximations, more and larger outliers can be expected.[19,20] Obviously, the accuracy and transferability of ML methods depends dramatically on the compound diversity present in the training set.

Here we explore a novel use of ML methods in SQC. Instead of applying ML methods to directly compute certain properties within a large class of related compounds, we use them to determine optimum SQC parameters for individual molecules within the class of target compounds. In both cases, there is an initial training step for calibrating the ML model on a reference subset of compounds, followed by production runs that yield predictions for the other target compounds. The basic idea is to employ ML techniques to optimize the SQC parameters for individual molecules (within a class of related compounds) such that subsequent SQC calculations are as accurate as possible for predicting the properties of interest. Hence, we introduce a hybrid ML-SQC approach where ML is used as an automatic parametrization tool (APT) to determine on-the-fly optimal *individual* semiempirical parameters as a function of atomic configuration and composition.

In this article, we first explain the chosen APT approach. Thereafter we present an illus-

trative application of the ML-SQC method for a set of 6095 constitutional isomers $C_7H_{10}O_2$, for which accurate thermochemical reference data from G4MP2[21] calculations are available.[22] These molecules were drawn from the chemical universe database GDB-17 that covers many drug-like molecules and contains 166.4 billion molecules with up to 17 non-hydrogen atoms.[23] As SQC method, we use the semiempirical OM2 (orthogonalization model 2) approach.[12,24,25] In this proof-of-concept study, we evaluate the accuracy that can be achieved by the ML-OM2 method for the chosen target set, and we compare the ML-OM2 results with those obtained using the standard OM2 parameters as well as special-purpose OM2 parameters from reparametrizations for the same target compounds.

2 Automatic parametrization technique

2.1 Overview

As outlined above, the idea behind APT relies on the use of ML to "locally" improve upon the global SQC parameter values. To this end we have implemented the following procedure:

1 Find optimal corrections to parameter values for each individual molecule in the training set.

2 Train ML model on the parameter corrections from the previous step.

3 Use ML model to predict corrections to parameters for target molecules.

4 Carry out SQC calculations with corrected parameter values for target molecules.

In this procedure, one may in principle apply any combination of appropriate parameter optimization and machine learning techniques. In the following we present the chosen hybrid ML-SQC approach in detail.

2.2 Technical Details

Step 1

Here, we only vary one of the many OM2 parameters at a time. More specifically, we tune a given parameter to minimize the error in the atomization enthalpy for each molecule in the training set using the Levenberg–Marquardt optimization algorithm.[26,27] Generally, convergence to complete error depletion was reached after few iterations such that an OM2 calculation with the resulting parameter gives an error-free atomization enthalpy for each molecule (in its standard OM2 geometry) of the training set. Systematic application of this procedure yields a set of changes (ΔP^{opt}) to the standard OM2 parameter values for each molecule in the training set. Failures of this procedure were encountered in a vanishingly small number of cases, which were ignored since they do not affect the overall performance. These minimizations were carried out successively for all OM2 parameters, which are listed in Table 1 in standard notation.[12,24,25]

2.2.1 Step 2

The corrections $\{\Delta P^{opt}\}$ of the parameter values for each molecule in the training set (obtained at its standard OM2 geometry) are used to train the ML model. We apply an ML approach introduced in 2012,[16] which has been described in detail in the literature.[16–18] Therefore we give only a brief outline of the procedure and refer to the original publications for further information.

We employ kernel ridge regression with a Laplacian kernel. In this approach, the default parameter correction ΔP for molecule M is estimated by summing over all N_{train} molecules $\{M_i\}$ in the training set.

$$\Delta P\left(M\right) = \sum_{i=1}^{N_{train}} \alpha_i e^{-\|M-M_i\|_1/\sigma}, \tag{1}$$

where α_i is the regression coefficient for molecule M_i, σ is the length-scale hyperparameter (same value for any pair of molecules M and M_i), and $\|M - M_i\|_1$ is the 1-norm calculated

from the vectorized molecular descriptor $\mathbf{X}$ of size N_x by summing the absolute differences between the elements of $\mathbf{X}(M)$ and $\mathbf{X}(M_i)$:

$$\|M - M_i\|_1 = \sum_a^{N_x} |X_a(M) - X_a(M_i)| \tag{2}$$

As molecular representation, we choose the Coulomb matrix $\mathbf{C}$.[16–18] It is an atom-by-atom matrix with the elements:

$$C_{IJ} = \begin{cases} 0.5 Z_I^{2.4} & \text{if } I = J, \\ \dfrac{Z_I Z_J}{|\mathbf{R}_I - \mathbf{R}_J|} & \text{if } I \neq J, \end{cases} \tag{3}$$

where internuclear distances $|\mathbf{R}_I - \mathbf{R}_I|$ are measured between the atomic coordinates $\mathbf{R}$ (in Bohr) and all nuclear charges Z_I are in e. When one molecule is larger than the other, we extend the Coulomb matrix of the smaller molecule by zeros. The Coulomb matrix is a unique yet non-stereospecific representation of a molecule, and it can thus distinguish diastereomers but not enantiomers. It is translationally and rotationally invariant. In order to also achieve atom-index invariance we sort all atom indices by the norm of their Coulomb matrix row. Sorted Coulomb matrices are used to calculate the norm $\|M - M_i\|_1$ according to eq. 2, where X_a is an element C_{IJ} of the corresponding Coulomb matrix and the sum runs over all Coulomb matrix elements, with N_x being the square of the number of atoms of the largest molecule.

Training the ML model outlined above requires solving the minimization problem:

$$\min_\alpha \sum_i^{N_{train}} \left[\Delta P(M_i) - \Delta P^{opt}(M_i)\right] + \lambda \cdot \sum_{i,j}^{N_{train}} \alpha_i K_{ij} \alpha_j \tag{4}$$

The analytical solution involves the following matrix transformations:[16–18]

$$\boldsymbol{\alpha} = (\mathbf{K} + \lambda \mathbf{I})^{-1} \Delta \mathbf{P}^{\mathbf{opt}} \tag{5}$$

where $\mathbf{I}$ is identity matrix, $\Delta\mathbf{P}^{\mathbf{opt}}$ is the vector with corrections to the standard parameter value, and λ is a so-called regularization parameter that ensures the transferability of the model to new compounds.[16,17] The elements K_{ij} of the kernel matrix $\mathbf{K}$ are defined by:

$$K_{ij} = e^{-\|M_i - M_j\|_1/\sigma} \tag{6}$$

We determined optimal values of hyperparameters by five-fold cross-validation within the training sets, following a previously reported procedure.[18] The training set was sorted according to the values of the parameter correction. It was then divided into buckets with only five items each. Thereafter, five splits were created by successively taking out a single item from each bucket. Four of these stratified samples were used to train the ML model, and the fifth out-of-sample split was used to estimate the error of the ML model. All five possible such folds were generated. The error in the out-of-sample split was minimized by varying the hyperparameters σ and λ. Optimal σ and λ values were found for each fold by a simple logarithmic grid search. These hyperparameter values were averaged over five folds to train our final ML model on the entire training set.

2.2.2 Step 3

The ML model trained in the previous step was employed to predict corrections to the OM2 parameters for other molecules (outside the training set) according to eq. 1, using geometries optimized with default OM2 parameters.

2.2.3 Step 4

The corrections to the OM2 parameters predicted in the previous step were added to their OM2 default values, and the resulting parameters were used in a subsequent OM2 calculation of the atomization enthalpy.

3 Computational details

All OM2 calculations with default and modified parameters were carried out with our locally modified `MNDO2005`[28] program. The SCF energy convergence criterion was set to 10^{-8} eV. In addition, the diagonal elements of the density matrix were converged to less than 10^{-8}. Geometry optimizations were considered converged when the Cartesian gradient norm dropped below 0.1 kcal/(mol·Å). No cut-offs were applied for the three-center orthogonalization corrections in the OM2 calculations.

4 Results and discussion

The G4MP2 atomization enthalpies at $T = 298$ K of the 6095 constitutional isomers $C_7H_{10}O_2$[22] that can be extracted from GDB-17[23] (see above) served as reference data for ML.

4.1 Application of APT

In an initial screening of all OM2 parameters, we estimated the potential improvements in accuracy that one might expect from a hybrid ML-SQC approach. ML-OM2 calculations were performed at OM2 geometries (denoted as ML-OM2//OM2). In the screening of the 61 OM2 parameters for hydrogen, carbon, and oxygen, we used 1000 randomly taken molecules for training and the remaining 5095 molecules for testing (APT Step 1). The resulting mean absolute errors (MAEs) in the atomization enthalpies for the training and test sets are given in Table 1. In addition, mean absolute deviations (MADs) of the individually optimized parameters (ATP Step 1) from the standard OM2 parameters are also listed.

The MAEs in the atomization enthalpies calculated with ML-OM2//OM2 for the test set typically improve from 6.3 kcal/mol (standard OM2 for these molecules) down to 2.80–3.00 kcal/mol in 38 out of 61 cases when a single OM2 parameter is adjusted individually through ML. Thus, in principle any of these 38 parameters could be used for APT. Here, we have chosen to develop ML models for corrections to the ζ parameter (orbital exponent)

Table 1: Mean absolute deviations (MAD) of parameter values optimized in APT Step 1 from the standard OM2 values and mean absolute errors (MAEs) in atomization enthalpies from ML-OM2//OM2 calculations at OM2 geometries for 1000 $C_7H_{10}O_2$ molecules (drawn at random) in the training set and 5095 $C_7H_{10}O_2$ molecules in the test set (remainder). MADs in %, MAEs in kcal/mol. Standard OM2 yields a MAE of 6.30 kcal/mol for these molecules.

Parameter	Hydrogen			Carbon			Oxygen		
	MAD, %	MAE, kcal/mol		MAD, %	MAE, kcal/mol		MAD, %	MAE, kcal/mol	
		training	test		training	test		training	test
One-center one-electron terms									
U_{ss}	1.20	0.00	2.89	0.10	0.00	2.83	4.10	0.51	3.50
U_{pp}				0.10	0.00	2.84	0.30	0.00	2.84
Orbital exponent									
ζ	1.10	0.00	2.85	0.40	0.00	2.82	1.20	0.00	2.88
Resonance integrals									
β_s	1.20	0.00	2.82	1.50	0.00	2.87	13.40	0.00	3.09
β_p				0.90	0.00	2.84	2.50	0.00	3.04
β_π				3.90	0.00	3.77	9.80	0.00	3.78
β_s(X–H)				2.30	0.00	2.86	117.80	0.44	6.27
β_p(X–H)				1.40	0.00	2.84	35.60	0.08	6.69
α_s	2.50	0.00	2.82	1.30	0.00	2.84	9.40	0.00	2.99
α_p				0.90	0.00	2.84	2.90	0.00	3.27
α_π				2.50	0.00	3.49	6.60	0.00	3.33
α_s(X–H)				4.40	0.00	2.88	203.20	1.37	6.01
α_p(X–H)				4.70	0.00	2.99	47.40	0.24	6.28
Orthogonalization factors									
F_1	4.20	0.00	2.82	0.70	0.00	2.82	1.60	0.00	2.84
F_2	5.40	0.00	2.86	8.60	0.00	2.84	4.70	0.00	2.86
G_1	40.10	0.64	3.57	17.00	0.00	3.04	215.50	0.18	5.52
G_2	26.30	0.00	2.80	11.90	0.00	2.84	223.30	0.11	4.22
Effective core potentials									
ζ_α				0.40	0.00	2.88	4.80	0.00	3.12
$F_{\alpha\alpha}$				1.60	0.00	2.88	13.90	0.00	2.86
β_α				6.50	0.00	2.86	116.00	0.00	3.08
α_α				4.10	0.00	2.87	250.50	1.40	25.40
One-center two-electron integrals									
g_{ss}	7.40	0.46	3.49	0.30	0.00	2.83	4.50	0.00	2.85
g_{pp}				0.70	0.00	2.83	1.60	0.00	2.84
g_{sp}				1.50	0.10	3.18	1.30	0.00	2.89
g_{p2}				0.20	0.00	2.83	0.60	0.00	2.84
h_{sp}				11.80	0.02	3.13	11.40	0.02	3.06

of carbon. This choice is motivated by the fact that tuning ζ apparently leads to minimal changes in parameter value combined with maximal changes in the computed property (see Table 1). More specifically, optimizing the ζ parameter of carbon for the individual molecules in the training set leads to a mean absolute change of only 0.4% in the parameter value, while the MAE for the atomization enthalpy is reduced from 6.3 kcal/mol to 2.82 kcal/mol in the test set. Such small changes of parameters can be considered fine tuning, rather than drastic reparametrization to populate significantly different regions of parameter space. In addition, it seems natural to use an OM2 parameter of carbon for fine tuning since our present application deals exclusively with organic molecules. The OM2 parameters for the effective core potential might offer a promising alternative for fine tuning, because they yield MAEs of less than 3 kcal/mol for parameter changes in the single-digit percentage range.

Next we studied the effect of the size of the chosen training on the ML-OM2 results (using again the ζ parameter of carbon for ML). We considered N_{train} = 10, 100, 1000, 2000, 3000, 4000, and 5000, with molecules being drawn at random from the full set of 6095 isomers $C_7H_{10}O_2$. After applying the ML model to the training set, the remaining fitting errors were vanishingly small in all cases (*i.e.*, for all N_{train} values). Not surprisingly, the accuracy for the out-of-sample test set improved systematically with increasing N_{train}. The results are summarized in Table 2 and shown in Figure 1. The 5k ML-OM2//OM2 model (N_{train} = 5000, MAE = 1.72 kcal/mol) has a substantially improved accuracy when compared to standard OM2 (MAE = 6.30 kcal/mol), and even approaches the highly coveted target of "chemical accuracy" (1 kcal/mol), at an overall computational cost of about six cpu hours. We also note that a 2 kcal/mol accuracy for atomization enthalpies is on par with (if not better than) many of the more advanced DFT methods.[29]

4.2 Special-purpose reparametrization of OM2

For the sake of comparison, we performed a conventional reparametrization of all OM2 parameters for the same set of 6095 $C_7H_{10}O_2$ isomers using the same reference atomization

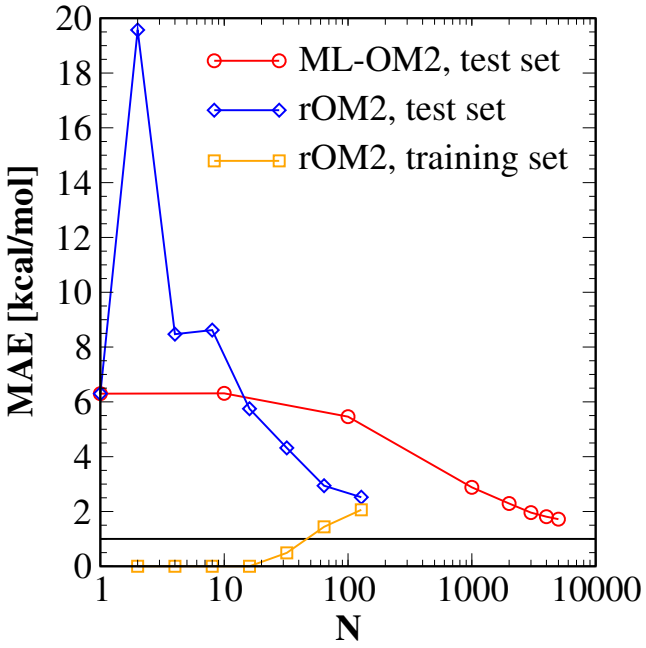

Figure 1: Mean absolute errors (MAEs) in the predicted atomization enthalpies for the out-of-sample test set of molecules with $C_7H_{10}O_2$ stoichiometry for ML-OM2//OM2 and rOM2 (see text). MAEs for the training set are only shown for rOM2 (vanishingly small for ML-OM2//OM2). The MAEs are plotted as a function of the training set size (N_{train}, logarithmic scale). The horizontal line at 1.0 kcal/mol indicates the onset of chemical accuracy.

Table 2: Mean absolute errors (MAEs) in the predicted atomization enthalpies of the constitutional isomers of $C_7H_{10}O_2$ from OM2 ($N_{train} = 0$) and ML-OM2//OM2 calculations at OM2 geometries (see text). MAEs in kcal/mol for N_{train} molecules in the training sets and for $6095 - N_{train}$ molecules in the test sets.

N_{train}	Training set	Test set
0		6.30
10	0.00	6.31
100	0.00	5.46
1000	0.00	2.88
2000	0.00	2.29
3000	0.00	1.96
4000	0.00	1.81
5000	0.00	1.72

enthalpies. The first 2^n (n = 1–7) molecules from the randomly ordered set served as training sets for the reparametrization. All OM2 parameters were reoptimized using a modified implementation of the Subplex method[30] based on the NLopt library,[31] without imposing any limits or constraints. The standard OM2 parameters were taken as starting values. The accuracy of the resulting series of reparametrized OM2 (rOM2) methods was evaluated on the corresponding test sets consisting of the remaining $6095 - 2^n$ molecules. For a fair comparison with the APT approach, the rOM2 calculations on the test molecules were done at geometries optimized at the standard OM2 level (designated as rOM2//OM2).

For small training sets (N_{train} = 2–8) the reparameterization results in overfitting, as indicated by MAEs for the test set that are larger than the MAE of standard OM2 (Table 3). For larger training sets, the MAE for the training set grows monotonically, while the MAE for test set decreases monotonically. With increasing size of the training set size, the MAEs for both sets converge to the same range, reaching values of 2.06–2.52 kcal/mol for $N_{train} = 128$ (Figure 1).

Table 3: Mean absolute errors (MAEs) in atomization enthalpies from OM2 ($N_{train} = 0$) and rOM2//OM2 calculations at OM2 geometries for N_{train} molecules in the training sets and $6095 - N_{train}$ in the test sets. MAEs in kcal/mol.

N_{train}	Training set	Test set
0		6.30
2	0.00	19.57
4	0.00	8.47
8	0.00	8.62
16	0.00	5.75
32	0.49	4.32
64	1.44	2.94
128	2.06	2.52

Obviously, the MAE for the test set is generally higher than that for the training set. Since the MAE for training set continually increases with the size of the set, it is safe to assume that the MAEs for both sets will be slightly larger than 2.06 kcal/mol for larger N_{train} values.

4.3 Comparison of ML-OM2 and rOM2 results

We now analyze the distribution of errors for both reparametrization approaches. For a set of 1095 randomly drawn constitutional isomers of $C_7H_{10}O_2$, Figure 2 displays the error distributions of their atomization enthalpies obtained from the 5k ML-OM2//OM2 model as well as rOM2 and standard OM2 calculations. In the error distribution of standard OM2, there is a systematic shift (*i.e.*, an underestimation of atomization enthalpies) and a substantial skew. The rOM2 reparametrization overcomes both these problems, yielding a more normal distribution centered at zero. However, in the case of 5k ML-OM2//OM2, the error distribution is more narrow, suggesting a higher degree of fidelity and a lower number of outliers. The 5k ML-OM2 model has the lowest MAE (1.72 kcal/mol). The worst outlier has an error of more than 26 kcal/mol in OM2, which is reduced to 9.8 and 8.2 kcal/mol in rOM2 and ML-OM2//OM2, respectively.

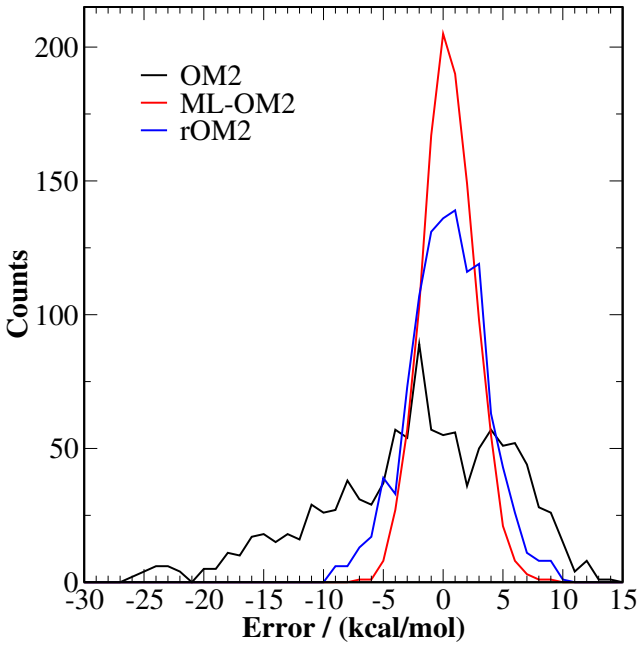

Figure 2: Error histogram for OM2, 5k ML-OM2//OM2, and rOM2//OM2 for a test set of 1095 molecules (see text).

We have already noted that the conventional reparametrization of OM2 appears to have a lower bound for MAEs for the test set, in our case 2.06 kcal/mol (see above), presumably

due to the fixed functional form dictated by the use of OM2. The ML-based APT approach, on the other hand, is highly flexible because of the use of an expansion in non-linear terms, which can be systematically improved by adding more examples to the training set. However, conventional reparametrization schemes do have the advantage of providing rapid improvements even for small training sets, whereas APT typically requires thousands of reference data points. Therefore, APT is particularly suited to problems involving big data sets. Another advantage of APT over conventional reparametrizations is due to the fact that its kernel inversion is convenient and computationally less demanding, while in the case of conventional reparametrizations complex multidimensional optimization problems must be solved.

Yet another important issue arises when it comes to transferability. More specifically, one might wonder what happens when we attempt to compute properties of molecules that differ substantially from those present in the training set. Such molecules will normally not be well represented by the modified parameters, for obvious reasons, and one may thus expect huge errors. In the case of APT, by contrast, the ML model will predict vanishing corrections to the individual parameters for molecules that are very different, and consequently, the results will be close to those obtained with the standard parameters. In this sense, the APT-ML model is well-tempered and transferable. It can only improve—and will never deteriorate—performance, regardless whether we consider molecules structurally similar to or different from the species in the training set.

For a more quantitative study of this aspect, we performed a comparative error analysis of the OM2, 5k ML-OM2//OM2, and rOM2//OM2 results on a validation set of 100 molecules drawn at random from the big database of ca. 134,000 molecules,[22] which covers all organic molecules with up to nine heavy atoms (not counting hydrogens). We compared the three computed sets of atomization enthalpies to the reference G4MP2 values.[22] Many molecules in the validation set have N or F atoms, not present in $C_7H_{10}O_2$ isomers, which poses a severe challenge. In the case of rOM2, we combined the modified parameters for H, C,

and O (see above) with the standard OM2 parameters for N and F, whereas the trained 5k ML-OM2 model employed only one modified parameter (ζ for C) together with the standard values of all other OM2 parameters in a given molecule. Consequently, it is not too surprising that rOM2//OM2 yields dramatic errors that may exceed 400 kcal/mol (see Table 4). In the case of ML-OM2//OM2, the MAE is drastically reduced to ∼20 kcal/mol, with a maximum outlier of ∼50 kcal/mol. As expected, these ML-OM2//OM2 results are not too far off from the corresponding standard OM2 results (MAE ∼10 kcal/mol, maximal error ∼40 kcal/mol). This confirms that the ML-OM2//OM2 approach is fairly robust even in difficult cases.

Table 4: Mean absolute errors (MAEs) of atomization enthalpies of 100 molecules drawn at random from GDB-17.[23] Results are given for OM2 with default parameters, rOM2//OM2 reparametrized using 128 $C_7H_{10}O_2$ isomers, and the ML-OM2//OM2 model trained on 5k $C_7H_{10}O_2$ isomers. MAEs in kcal/mol.

Method	MAE, kcal/mol	Range of errors, kcal/mol
OM2	10.94	−39.57...9.6
ML-OM2//OM2	21.58	−52.02...0.87
rOM2//OM2	145.39	−414.15...484.33

5 Conclusions

We have introduced an automatic parametrization tool that augments semiempirical parameters for *any* new molecule. This tool is based on machine learning models of parameters as a function of molecular structure (requiring as input only the identities of the constituent atoms and their coordinates). After training the model on sufficiently large training sets (yielding pre-calculated corrections to parameters), it can be applied to other new molecules for predicting molecule-specific corrections to the parameters that allow semiempirical quantum chemical calculations with improved accuracy.

For numerical demonstration, we chose the OM2 method which has a mean absolute error of 6.3 kcal/mol in atomization enthalpy for the 6095 constitutional isomers of $C_7H_{10}O_2$ stoichiometry in calculations with standard OM2 parameters. After individually adjusting

the parameters in the ML-OM2 approach for the largest training set of 5000 isomers, the mean absolute error for the remaining 1095 isomers in the test set can be reduced from 6.3 kcal/mol (standard OM2, same value as for the full set) to 1.7 kcal/mol, and the largest error can be lowered from 26.3 to 8.2 kcal/mol, respectively. Furthermore, ML-OM2 has a narrower error distribution than OM2, or than a conventionally reparametrized variant of OM2 (rOM2), and it is found to be quite robust even when screening structures that differ substantially from those present in the training set.

To summarize, we have presented numerical evidence that the ML-APT approach can significantly improve the predictive accuracy of well-established semiempirical quantum-chemical methods for large sets of molecules, without increasing the computational burden beyond the need of having a reference data base at disposal. We emphasize that due to its general nature, the APT idea may be useful for any method of fixed functional form that depends on parameters, *e.g.* in DFT or in the "Learning-On-The-Fly" approach to *ab initio* molecular dynamics.[32]

Acknowledgement

The authors thank to Xin Wu for kindly providing the parametrization program. WT is grateful to the European Research Council (ERC) for financial support through an ERC Advanced Grant. OAvL acknowledges support from the Swiss National Science foundation (No. PP00P2_138932). This research used resources of the Argonne Leadership Computing Facility at Argonne National Laboratory, which is supported by the Office of Science of the U.S. DOE under contract DE-AC02-06CH11357

References

(1) Hachmann, J.; Olivares-Amaya, R.; Atahan-Evrenk, S.; Amador-Bedolla, C.; Sánchez-Carrera, R. S.; Gold-Parker, A.; Vogt, L.; Brockway, A. M.; Aspuru-Guzik, A. *J. Phys. Chem. Lett.* **2011**, *2*, 2241–2251.

(2) Yang, L.; Ceder, G. *Phys. Rev. B* **2013**, *88*, 224107.

(3) Jain, A.; Ong, S. P.; Hautier, G.; Chen, W.; Richards, W. D.; Dacek, S.; Cholia, S.; Gunter, D.; Skinner, D.; Ceder, G.; Persson, K. A. *APL Mater.* **2013**, *1*, 011002.

(4) Jorgensen, W. L. *Science* **2004**, *303*, 1813–1818.

(5) Nørskov, J. K.; Bligaard, T.; Rossmeisl, J.; Christensen, C. H. *Nat. Chem.* **2009**, *1*, 37–46.

(6) Schneider, G. *Nat. Rev. Drug Discovery* **2010**, *9*, 273–276.

(7) Hautier, G.; Miglio, A.; Ceder, G.; Rignanese, G.-M.; Gonze, X. *Nat. Comms.* **2013**, *4*, 2292.

(8) Zhou, T.; Caflisch, A. *ChemMedChem* **2010**, *5*, 1007–1014.

(9) Thiel, W. *WIREs: Comput. Mol. Sci.* **2014**, *4*, 145–157, *see also references therein*.

(10) Clark, T.; Stewart, J. J. P. MNDO-Like Semiempirical Molecular Orbital Theory and Its Application to Large Systems. In *Computational Methods for Large Systems: Electronic Structure Approaches for Biotechnology and Nanotechnology*; Reimers, J. R., Ed.; John Wiley & Sons, Inc.: Hoboken, NJ, USA., 2011; pp 259–286.

(11) Jensen, F. *Introduction to Computational Chemistry*; John Wiley & Sons Ltd: Baffins Lane, Chichester, West Sussex PO19 1UD, England, 1999; pp. 81–97.

(12) Korth, M.; Thiel, W. *J. Chem. Theory Comput.* **2011**, *7*, 2929–2936.

(13) Tseng, S.-P.; Shen, M.-Y.; Yu, C.-H. *Theor. Chim. Acta* **1995**, *92*, 269–280.

(14) Gonzalez-Lafont, A.; Truong, T. N.; Truhlar, D. G. *J. Phys. Chem.* **1991**, *95*, 4618–4627.

(15) Wu, X.; Thiel, W.; Pezeshki, S.; Lin, H. *J. Chem. Theory Comput.* **2013**, *9*, 2672–2686.

(16) Rupp, M.; Tkatchenko, A.; Müller, K.-R.; von Lilienfeld, O. A. *Phys. Rev. Lett.* **2012**, *108*, 058301.

(17) von Lilienfeld, O. A. *Int. J. Quantum Chem.* **2013**, *113*, 1676–1689.

(18) Hansen, K.; Montavon, G.; Biegler, F.; Fazli, S.; Rupp, M.; Scheffler, M.; von Lilienfeld, O. A.; Tkatchenko, A.; Müller, K.-R. *J. Chem. Theory Comput.* **2013**, *9*, 3404–3419.

(19) Moussa, J. E. *Phys. Rev. Lett.* **2012**, *109*, 059801.

(20) Rupp, M.; Tkatchenko, A.; Müller, K.-R.; von Lilienfeld, O. A. *Phys. Rev. Lett.* **2012**, *109*, 059802.

(21) Curtiss, L. A.; Redfern, P. C.; Raghavachari, K. *J. Chem. Phys.* **2007**, *127*, 124105.

(22) Ramakrishnan, R.; Dral, P. O.; Rupp, M.; von Lilienfeld, O. A. *Sci. Data* **2014**, *1*, 140022.

(23) Ruddigkeit, L.; van Deursen, R.; Blum, L.; Reymond, J.-L. *J. Chem. Inf. Model.* **2012**, *52*, 2684.

(24) Weber, W. Ein neues semiempirisches NDDO-Verfahren mit Orthogonalisierungskorrekturen: Entwicklung des Modells, Parametrisierung und Anwendungen. Ph.D. thesis, Universität Zürich, 1996.

(25) Weber, W.; Thiel, W. *Theor. Chem. Acc.* **2000**, *103*, 495–506.

(26) Levenberg, K. *Quart. Appl. Math.* **1944**, *2*, 164–168.

(27) Marquardt, D. W. *SIAM J. Appl. Math.* **1963**, *11*, 431–441.

(28) Thiel, W. MNDO2005, version 7.0. Max-Planck-Institut für Kohlenforschung: Mülheim an der Ruhr, Germany, 2005.

(29) Cohen, A. J.; Mori-Sánchez, P.; Yang, W. *Chem. Rev.* **2012**, *112*, 289–320.

(30) Rowan, T. Functional Stability Analysis of Numerical Algorithms. Ph.D. thesis, Department of Computer Sciences, University of Texas at Austin, 1990.

(31) Johnson, S. G. The NLopt nonlinear-optimization package. `http://ab-initio.mit.edu/nlopt`, accessed on February 11, 2015.

(32) Csányi, G.; Albaret, T.; Payne, M. C.; Vita, A. D. *Phys. Rev. Lett.* **2004**, *93*, 175503.

Graphical TOC Entry