Combining Semiempirical Quantum Mechanics with Machine Learning: Towards Hybrid Quantum Mechanics/Machine Learning (QM/ML)

Pavlo O. Dral,[a] Raghunathan Ramakrishnan,[b] Matthias Rupp,[b] Walter Thiel,[a] O. Anatole von Lilienfeld[b,c]

[a]Max-Planck-Institut für Kohlenforschung, Kaiser-Wilhelm-Platz 1, 45470 Mülheim an der Ruhr. [b]Institute of Physical Chemistry, University of Basel, Klingelbergstraße 80, CH-4056 Basel, Switzerland. [c]Argonne Leadership Computing Facility, Argonne National Laboratory, 9700 S. Cass Avenue, Lemont, IL 60439, USA

Introduction

Inductive, supervised-learning approaches from machine learning (ML) have the advantage of being able to predict various electronic structure properties of different materials in short computation times given a reference data set.[1] This can be used for exploring relevant areas of the huge chemical compound space. The mean absolute error (MAE) of atomization energies of small organic molecules can be lower than 10 kcal/mol.[2] However, approximate QM methods, based on rigorous physics arguments, usually have much less severe outliers.[1] *Here, we combine both ML and QM methods into a hybrid QM/ML approach exploiting advantages of both methods while trying to eliminate their disadvantages.*

Molecular Descriptor

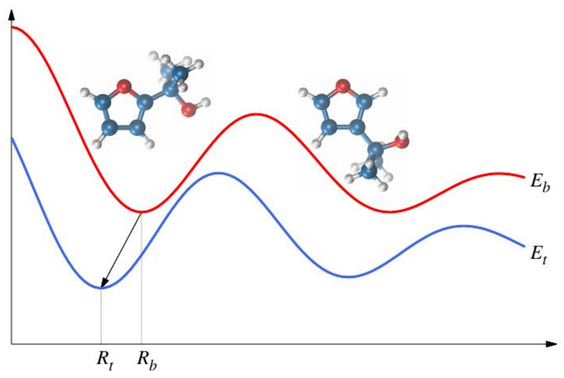
One molecular descriptor appropriate for ML studies that satisfies many requirements and has been successfully applied in previous studies is the "Coulomb matrix". Its elements are defined by

$$C_{ij} = \begin{cases} 0.5 Z_i^{2.4}, & for\ i = j \\ \dfrac{Z_i Z_j}{\left|\mathbf{R}_i - \mathbf{R}_j\right|}, & for\ i \neq j \end{cases}$$

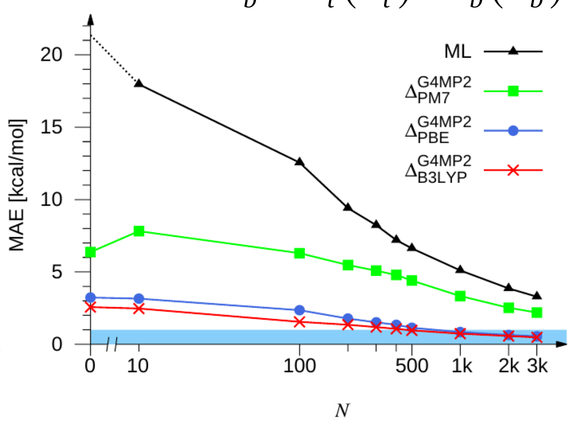
where Z is the nuclear charge and $|\mathbf{R}_i - \mathbf{R}_j|$ is the internuclear distance between atoms i and j.[1]

Δ-QM/ML

Δ-QM/ML Ansatz[4] is based on the simple idea to eliminate errors (Δ's) of less accurate *baseline* QM methods by estimating them with ML. In practice it is done by training ML model on the set of errors in properties Y_b calculated with baseline methods with respect to values Y_t of a more accurate target-line method: $\Delta_b^t = Y_t(R_t) - Y_b(R_b)$.



Δ-QM/ML fixes errors of the baseline method relative to the target-line method in energies ($E_t - E_b$) , even though molecular geometries optimized at base- and target-line methods differ (R_b and R_t, respectively)



Improving accuracy in atomization energies calculated at different Δ-QM/ML levels of theory with increased training set size. For comparison, the errors of pure ML and QM methods are marked (at N = 0).

The Δ-QM/ML Ansatz allows us to improve the accuracy of PM7 significantly: its MAE drops from 5.5 kcal/mol to ca. 2 kcal/mol for the 3k training set, which is below the MAE of PBE/6-31G(2df,p) and B3LYP/6-31G(2df,p).

Machine Learning Algorithm

We used in our studies a kernel ridge regression approach with a Laplacian kernel. Some property Y^{est} of a molecule a is predicted by

$$Y_a^{est} = \sum_{i=1}^{N_{train}} \alpha_i k_{i,a}$$

where M is the molecular descriptor, N_{train} is number of molecules in the training set, k is the kernel function, σ is the hyperparameter defining the length scale.[1,2]

$$k_{i,a} = \exp\left(-\frac{\left|M_i - M_a\right|_1}{\sigma}\right)$$

The coefficients are found by solving a minimization problem:

$$\min_{\alpha} \sum_{i}^{N_{train}} \left(Y_i^{est} - Y_i^{ref}\right)^2 + \lambda \sum_{i,j}^{N_{train}} \alpha_i \exp\left(-\frac{\left|M_i - M_j\right|_1}{\sigma}\right)\alpha_j$$

with reference values Y^{ref} of the property and the regularization hyperparameter λ ensuring the transferability of the model to new compounds.[1,2] The solution in matrix form is very simple:[1,2]

$$\boldsymbol{\alpha} = \left(\mathbf{K} + \lambda \mathbf{I}\right)^{-1} \mathbf{Y}^{ref}$$

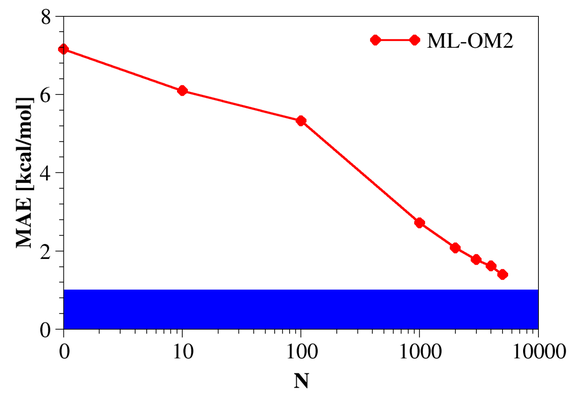
Data Set

We have chosen for our studies the set of 6095 constitutional isomers of small molecules with the sum formula $C_7H_{10}O_2$. Recently we performed accurate calculations of electronic and thermochemical properties for this set using the G4MP2 protocol.[3] The accuracy of G4MP2 is close to 1.0 kcal/mol.

Semiempirical QM Parameter Learning

Parameters of semiempirical QM methods (SQM) are optimized for broad classes of compounds, and their average accuracy is far from chemical accuracy. We demonstrate for heats of formation predicted by the OM2 method[5] that ML can be used to estimate the optimal value(s) of SQM parameter(s) for individual molecules.[6] We call the resulting method ML-OM2.

Property values Y^{OM2} calculated at OM2 depend on the set of parameters $\{P_{OM2}\}$. We find such sets $\{P_{OM2}^{opt}\}$ that $Y^{OM2opt} \approx Y^{ref}$ for all molecules in the training set. Then we use ML to predict $\{P_{OM2}^{est}\}$ for unknown molecules and calculate respective $Y^{ML\text{-}OM2}$ values.



Improving accuracy in heats of formation at 298 K calculated at ML-OM2 with increased training set size for ML. Error of pure OM2 method is indicated at N = 0.

ML-OM2 accuracy is improved from 7.16 for pure OM2 to 1.40 kcal/mol for the 5k training set. We carefully monitored the parameters to be estimated by ML so that their values changed by less than 3% on average from the default OM2 values.

References

[1] a) M. Rupp, A. Tkatchenko, K.-R. Müller, and O. A. von Lilienfeld, *Phys. Rev. Lett.* **108**, 058301; J. E. Moussa **109**, 059801; M. Rupp, A. Tkatchenko, K.-R. Müller, and O. A. von Lilienfeld **109**, 059802 (2012) b) G. Montavon, M. Rupp, V. Gobre, A. Vazquez-Mayagoitia, K. Hansen, A. Tkatchenko, K.-R. Müller, and O. A. von Lilienfeld, *New J. Phys.* **15**, 095003 (2013).
[2] K. Hansen, G. Montavon, F. Biegler, S. Fazli, M. Rupp, M. Scheffler, O. A. von Lilienfeld, A. Tkatchenko, and K.-R. Müller, *J. Chem. Theory Comput.* **9**, 3404–3419 (2013).
[3] R. Ramakrishnan, P. O. Dral, M. Rupp, O. A. von Lilienfeld, *Sci. Data* **1**, 140022 (2014).
[4] R. Ramakrishnan, P. Dral, M. Rupp, O. A. von Lilienfeld, *submitted* (2014)
[5] a) Weber W. Ein neues semiempirisches NDDO-Verfahren mit Orthogonalisierungs-korrekturen: Entwicklung des Modells, Parametrisierung und Anwendungen. PhD Thesis, Universität Zürich (1996) b) W. Weber W, W. Thiel *Theor. Chem. Acc.* **103**, 495–506 (2000).
[6] P. O. Dral, W. Thiel, O. A. von Lilienfeld, *in preparation*.

Conclusions and Outlook

We find that, by combining ML and semiempirical QM methods, molecular properties can be calculated with accuracy better than that of low-cost DFT methods: MAE is less than 2.0 kcal/mol and approaching 1.0 kcal/mol for very large training sets. The computation time is essentially determined by the SQM calculations [3]. These QM/ML methods hold promise for high-throughput screening of large data sets of molecules, for which the use of more accurate but computationally expensive QM methods is desirable yet not feasible. Our current efforts address the simultaneous improvement of several properties *via* the SQM parameter learning Ansatz.