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, Germany
[b]Institute of Physical Chemistry, Department of 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

Inductive, supervised-learning approaches from machine learning (ML) have received increasing interest from computational chemists and material scientists over the last years. ML models 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 makes them obvious candidates for exploring relevant areas of the huge chemical compound space. One molecular descriptor appropriate for such ML studies that satisfies many requirements and has been successfully applied in previous studies consists of the “Coulomb matrix" with $0.5Z_i^{2.4}$ diagonal elements and Z_iZ_j/R_{ij} off-diagonal elements, where Z is the nuclear charge and R is the internuclear distance between atoms i and j.[1] For sufficiently large training sets and good ML algorithm choices, the mean absolute error (MAE) of atomization energies of small organic molecules using the Coulomb matrix can be lower than 10 kcal/mol [2] and even lower than the corresponding MAE of PM6. 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. We find that by combining ML and fast QM methods molecular properties can be calculated with chemical accuracy (MAE ≈ 1 kcal/mol). The computation time is essentially determined by the QM methods [3]. We thus recommend the use of semiempirical QM, or fast DFT methods with small basis sets, augmented by a ML model to correct their errors. These QM/ML methods hold great 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.

[1] a) M. Rupp, A. Tkatchenko, K.-R. Müller, and O. A. von Lilienfeld, *Phys. Rev. Lett.* **108**, 058301 (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.* (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,** 3543–3556 (2013).
[3] R. Ramakrishnan, P. Dral, M. Rupp, O. A. von Lilienfeld, submitted (2014)