Rupp et al. Reply:

Matthias Rupp, Alexandre Tkatchenko, Klaus-Robert Müller, and O. Anatole von Lilienfeld

Phys. Rev. Lett. **109**, 059802 — Published 3 August 2012

DOI: 10.1103/PhysRevLett.109.059802

Reply to Comment on "Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning"

Matthias Rupp,[1] Alexandre Tkatchenko,[2] Klaus-Robert Müller,[3,4] and O. Anatole von Lilienfeld[5,*]

[1] *Institute of Pharmaceutical Sciences, ETH Zurich, 8093 Zürich, Switzerland*
[2] *Fritz-Haber-Institut der Max-Planck-Gesellschaft, 14195 Berlin, Germany*
[3] *Machine Learning Group, Technical University of Berlin, Franklinstr 28/29, 10587 Berlin, Germany*
[4] *Department of Brain and Cognitive Engineering, Korea University, Anam-dong, Seongbuk-gu, Seoul 136-713, Korea*
[5] *Argonne Leadership Computing Facility, Argonne National Laboratory, Argonne, Illinois 60439, USA*

In his comment [1], J. E. Moussa (JEM) raises concerns regarding the accuracy of our recently published Machine Learning (ML) model [2]. Our performance estimates, based on cross-validated Kernel Ridge Regression, amount to less than 10 kcal/mol mean absolute error (MAE) with respect to DFT-PBE0 [3, 4] predictions of atomization energies, using a training set of more than 7000 small organic molecules from the GDB-13 data set [5]. As such, the ML model achieves an accuracy similar to generalized gradient DFT, and significantly exceeds that of Hartree-Fock or local density approximated DFT [6].

In our Letter we presented numerical evidence that ML models can be built using (i) sufficient examples and (ii) a molecular representation based on Cartesian coordinates and elemental composition *without* explicitly accounting for the electronic degrees of freedom. Therefore, performance of our ML model should exclusively be assessed with respect to methods that perform similar maps, i.e. $\{Z_I, \mathbf{R}_I\} \mapsto E$. In order to place our performance estimates into the general context of atomistic simulation, however, our Letter also provides results for semi-empirical methods, namely bond counting (BC) (MAE 71 kcal/mol) and PM6 (73 kcal/mol), along with ML model results (15 kcal/mol).

Since pre-conceived knowledge about underlying chemical bonding is exploited, BC and PM6 differ from our ML model. Obviously, explicit fitting of BC and PM6 parameters to atomization energies of GDB molecules, instead of enthalpies of other data sets, will improve their performance. It is only after introducing knowledge about covalent bond distances *and* order (single, double, triple) that the MAE of BC decreases to the 10 kcal/mol quoted by JEM. Furthermore, and unlike BC, the ML model can be used for estimating binding curves [2]. Semi-empirical models, such as PM6, result from decades of parameterization, and it is not surprising that they can be reparameterized to improve atomization energies. By contrast, the virtue of our ML approach is that it is not only accurate and fast but general, i.e. it can be trained and used *without* electronic structure knowledge.

JEM discusses the remaining error of our ML model. For acetylene, the effect of coarse graining is illustrated for one of the degrees of freedom that can be chosen such that the Coulomb-matrix' eigenvalues remain constant. When using instead the Frobenius norm as a measure of distance between Coulomb matrices (footnote 25 [2]), and after cross-validated training on acetylene geometries supplied by JEM, the ML model yields out-of-sample estimates that reproduce DFT-PBE0 energies with a MAE of 0.24 kcal/mol (Fig. 1). According to JEM,the Frobenius norm producing identical coordinates for "homometric molecules" [7], aka. enantiomers, might be another origin of error. We believe this to be desirable since the employed DFT potentials conserve parity, i.e. particle interaction invariance under space inversion at the molecular origin of geometry. Electroweak quantum chemistry results would be required to account for parity violation in molecules [8, 9]. Finally, JEM blames perceived lack of size-consistency for the error residual of our ML model. We have statistically accounted for the effect of size-consistency on atomization energies by imposing atomic dissociation at interatomic distances three times larger than in equilibrium (footnote 37 [2]). Regarding the scaling properties mentioned by JEM, we believe conclusive statements to be premature.

To improve the ML model we propose the following: (i) coverage of molecular space for training; increase number of constitutional and conformational isomers. (ii) flexibility in kernel function space, e.g. multiple kernel learning [10]. (iii) molecular representation; see our Letter [2] for requirements. (iv) explore various distance metrics between Coulomb matrices. We conclude that our ML model is capable of yielding fast and accurate atomization energy estimates out of sample, without *any* prior knowledge about electronic structure effects such as covalent bonding or electronic configuration.

Discussions with K. Hansen are greatly acknowledged. This research used resources of the Argonne Leadership Computing Facility at Argonne National Laboratory, and is supported by the U.S. Department of Energy, Basic Energy Sciences, Office of Science, under contract # DE-AC02-06CH11357. K.-R. M. acknowledges partial support by DFG (MU 987/4-2) and EU (PASCAL2). M. R. acknowledges support by FP7 programme of the European Community (Marie Curie IEF 273039). This work was also supported by the World Class University Pro-

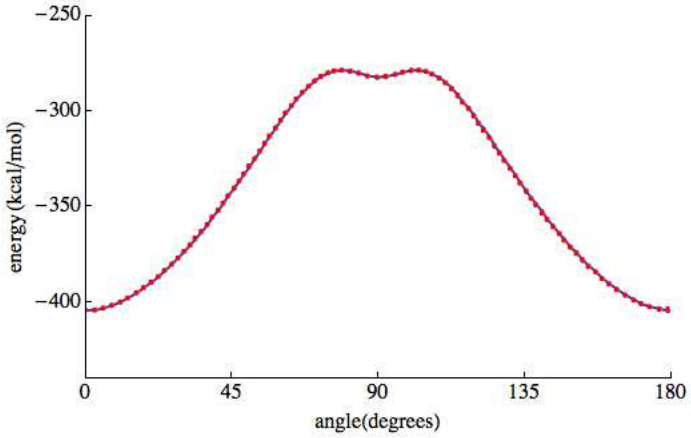

FIG. 1: (Color online) Blue line: PBE0. Red dots: ML model using Frobenius norm of, and trained on, Coulomb matrices of geometries corresponding to JEM's example.

gram through the National Research Foundation of Korea funded by the Ministry of Education, Science, and Technology, under Grant R31-10008.

* Electronic address: anatole@alcf.anl.gov

[1] J. E. Moussa, Phys. Rev. Lett. (2012), to be published.
[2] M. Rupp, A. Tkatchenko, K.-R. Müller, and O. A. von Lilienfeld, Phys. Rev. Lett. **108**, 058301 (2012).
[3] J. P. Perdew, M. Ernzerhof, and K. Burke, J. Chem. Phys. **105**, 9982 (1996).
[4] M. Ernzerhof and G. E. Scuseria, J. Chem. Phys. **110**, 5029 (1999).
[5] L. C. Blum and J.-L. Reymond, J. Am. Chem. Soc. **131**, 8732 (2009).
[6] W. Koch and M. C. Holthausen, *A Chemist's Guide to Density Functional Theory* (Wiley-VCH, 2002).
[7] A. L. Patterson, Nature **143**, 939 (1939).
[8] M. Quack and J. Stohner, CHIMIA **59**, 530 (2005).
[9] A. Bakasov, T.-K. Ha, and M. Quack, J. Chem. Phys. **109**, 7263 (1998).
[10] S. Sonnenburg, G. Rätsch, C. Schäfer, and B. Schölkopf, J. Mach. Learn. Res. **7**, 1531 (2006).