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

In a recent Letter [1], the authors construct a machine learning (ML) model of molecular atomization energies, which they compare to bond counting (BC) and the PM6 semiempirical method [2]. However, their ML model was trained and tested on density functional theory (DFT) *energies* while BC and PM6 are fit to *standard enthalpies*. For fair comparison, bond energies are refit to DFT data and PM6 is converted to an electronic energy using per-atom corrections [3]. BC and PM6 both perform better than the ML model and are free of large outliers in their error distributions as shown in Fig. 1.

As noted in [25] of [1], some ML model error may originate from the coordinate system choice. The n eigenvalues of the Coulomb matrix correspond to an equienergy $2n$-dimensional space of n-atom molecules rather than one molecule. For $n = 3$, this corresponds to the 3 translations and 3 rotations that naturally preserve the energy of an isolated molecule. For $n > 3$, the space includes unphysical molecular deformations that destroy structural rigidity. Figure 2 shows this with a distortion of acetylene (C_2H_2) that preserves its ML energy and coordinate, (53.058, 21.149, 0.290, 0.219).

It is suggested in [25] of [1] that the n^2 sorted entries of a Coulomb matrix might be utilized instead of its n eigenvalues as a ML coordinate system. This eliminates the dimensional deficiency, but produces identical coordinates for homometric molecules [5] that do not necessarily have equal energies. A computationally expensive alternative is the equivalence class of permuted Coulomb matrices with distance metric

$$d(\mathbf{M}, \mathbf{M}') = \min_{\mathbf{P}} \|\mathbf{M} - \mathbf{P}^T\mathbf{M}'\mathbf{P}\|_F \qquad (1)$$

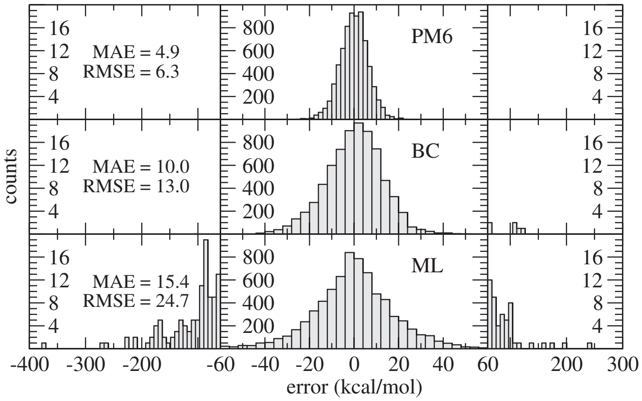

FIG. 1. Error histograms ($E_{\mathrm{DFT}} - E_{\mathrm{model}}$), mean absolute errors (MAE), and root-mean-square errors (RMSE) for PM6, BC, and ML models compared to DFT on the 7169 molecules of the GDB-13 set [8] with the formulas $C_vH_wN_xO_yS_z$ for $3 \le v + x + y + z \le 7$.

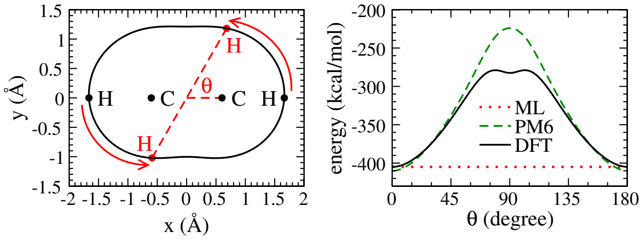

FIG. 2 (color online). A continuous deformation of acetylene. (left) Hydrogen atoms follow the closed curve with the line connecting them fixed to the origin. Carbon atoms remain near their equilibrium positions. (right) Atomization energy as a function of the H-origin-C angle.

for Coulomb matrices $\mathbf{M}$ and $\mathbf{M}'$, permutation matrices $\mathbf{P}$, and the Frobenius matrix norm.

Another possible source of ML model error is its lack of size-consistency. Even if the energy of two molecules A and B are accurately modeled in isolation, there are no guarantees that the well-separated pair of molecules $A + B$ will be similarly accurate. This requires explicitly filling the chemical compound space with a sufficiently dense set of training molecules, which likely leads to an $O(\alpha^n)$ computational complexity for n atoms ($\alpha > 1$). While benchmarks are favorable for $n \le 7$, the ML model cannot scale favorably compared to $O(n)$ classical force fields or $O(n^3)$ DFT or semiempirical methods. Alternative ML methods [6] enforce size consistency by modeling an intensive quantity, per-atom energy, rather than directly modeling the extensive total energy and control costs by exploiting nearsightedness [7].

Sandia National Laboratories is a multiprogram laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

Jonathan E. Moussa*
Sandia National Laboratories
Albuquerque, New Mexico 87185, USA

Received 6 February 2012; published 3 August 2012
DOI: 10.1103/PhysRevLett.109.059801
PACS numbers: 82.20.Wt, 31.15.B−, 31.15.E−, 82.37.−j

*godotalgorithm@gmail.com

[1] M. Rupp, A. Tkatchenko, K.-R. Müller, and O. A. von Lilienfeld, Phys. Rev. Lett. **108**, 058301 (2012).
[2] J. J. P. Stewart, J. Mol. Model. **13**, 1173 (2007).
[3] DFT energies are recomputed using the 6–311$G(3df, 2p)$ basis set and PBE0 functional. The ML model is trained on the 5000 Coulomb matrices of "model 1k" [1] and tuned ($\sqrt{\lambda} = 0.0021$ and $\sigma = 24.0$) to minimize test set

MAE. Prescribed PM6 corrections [2] and bond energies are least-squares fit to the test set. Geometries and bond orders are determined by OpenBabel 2.3.1 [4].

[4] N. M. O'Boyle, M. Banck, C. A. James, C. Morley, T. Vandermeersch, and G. R. Hutchison, J. Cheminformatics **3**, 33 (2011).

[5] A. L. Patterson, Nature (London) **143**, 939 (1939).

[6] A. P. Bartók, M. C. Payne, R. Kondor, and G. Csányi, Phys. Rev. Lett. **104**, 136403 (2010).

[7] W. Kohn, Phys. Rev. Lett. **76**, 3168 (1996).

[8] L. C. Blum and J.-L. Reymond, J. Am. Chem. Soc. **131**, 8732 (2009).

Rupp *et al.* Reply: 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 semiempirical methods, namely, bond counting (BC) (MAE 71 kcal/mol) and PM6 (73 kcal/mol), along with ML model results (15 kcal/mol).

Since preconceived 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]. Semiempirical 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 ([25] in [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 property to be desirable in this context 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 3 times larger than in equilibrium ([37] in [2]). Regarding the scaling properties mentioned by JEM, we believe conclusive statements to be premature.

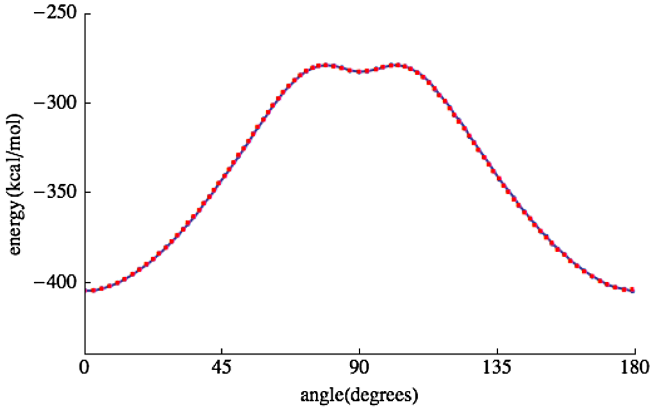

FIG. 1 (color online). Blue line: PBE0. Red dots: ML model using Frobenius norm of, and trained on, Coulomb matrices of geometries corresponding to JEM's example.

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 No. 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 Program through the National Research Foundation of Korea funded by the Ministry of Education, Science, and Technology, under Grant R31-10008.

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

Received 30 May 2012; published 3 August 2012
DOI: 10.1103/PhysRevLett.109.059802
PACS numbers: 82.20.Wt, 31.15.B−, 31.15.E−, 82.37.−j

*anatole@alcf.anl.gov

[1] J. E. Moussa, preceding Comment, Phys. Rev. Lett. **109**, 059801 (2012).

[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, Weinheim, Germany, 2002).

[7] A. L. Patterson, Nature (London) **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), [http://jmlr.csail.mit.edu/papers/v7/sonnenburg06a.html].