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

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

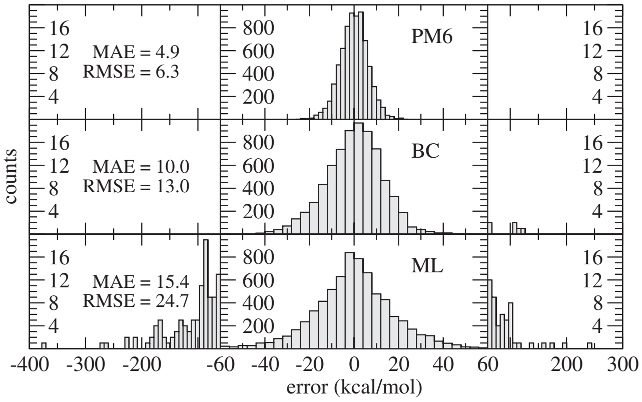

FIG. 1. Error histograms ($E_{\mathrm{DFT}} - E_{\mathrm{model}}$), mean absolute errors (MAE), and root-mean-square errors (RMSE) for PM6, BC, and ML models compared to DFT on the 7169 molecules of the GDB-13 set [8] with the formulas $C_vH_wN_xO_yS_z$ for $3 \leq v + x + y + z \leq 7$.

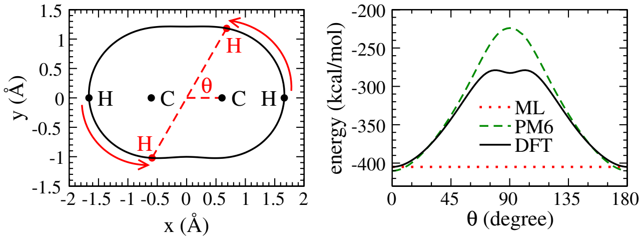

FIG. 2 (color online). A continuous deformation of acetylene. (left) Hydrogen atoms follow the closed curve with the line connecting them fixed to the origin. Carbon atoms remain near their equilibrium positions. (right) Atomization energy as a function of the H-origin-C angle.

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 $n \leq 7$, 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).