Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning

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

[1]*Machine Learning Group, Technical University of Berlin, Franklinstr 28/29, 10587 Berlin, Germany*
[2]*Institute of Pure and Applied Mathematics, University of California Los Angeles, Los Angeles, California 90095, USA*
[3]*Fritz-Haber-Institut der Max-Planck-Gesellschaft, 14195 Berlin, Germany*
[4]*Argonne Leadership Computing Facility, Argonne National Laboratory, Argonne, Illinois 60439, USA*

(Received 15 June 2011; published 31 January 2012)

We introduce a machine learning model to predict atomization energies of a diverse set of organic molecules, based on nuclear charges and atomic positions only. The problem of solving the molecular Schrödinger equation is mapped onto a nonlinear statistical regression problem of reduced complexity. Regression models are trained on and compared to atomization energies computed with hybrid density-functional theory. Cross validation over more than seven thousand organic molecules yields a mean absolute error of ~10 kcal/mol. Applicability is demonstrated for the prediction of molecular atomization potential energy curves.

DOI: 10.1103/PhysRevLett.108.058301 PACS numbers: 82.20.Wt, 31.15.B−, 31.15.E−, 82.37.−j

Solving the Schrödinger equation (SE), $H\Psi = E\Psi$, for assemblies of atoms is a fundamental problem in quantum mechanics. Alas, solutions that are exact up to numerical precision are intractable for all but the smallest systems with very few atoms. Hierarchies of approximations have evolved, usually trading accuracy for computational efficiency [1]. Conventionally, the external potential, defined by a set of nuclear charges $\{Z_I\}$ and atomic positions $\{\mathbf{R}_I\}$, uniquely determines the Hamiltonian H of *any* system, and thereby the ground state's potential energy, by optimizing Ψ, [2] $H(\{Z_I, \mathbf{R}_I\}) \overset{\Psi}{\longmapsto} E$. For a diverse set of organic molecules, we show that one can use machine learning (ML) instead, $\{Z_I, \mathbf{R}_I\} \overset{\mathrm{ML}}{\longmapsto} E$. Thus, we circumvent the task of explicitly solving the SE by training once a machine on a finite subset of known solutions. Since many interesting questions in physics require us to repeatedly solve the SE, the highly competitive performance of our ML approach may pave the way to large-scale exploration of molecular energies in chemical compound space [3,4].

ML techniques have recently been used with success to map the problem of solving complex physical differential equations to statistical models. Successful attempts include solving Fokker-Planck stochastic differential equations [5], parametrizing interatomic force fields for fixed chemical composition [6,7], and the discovery of novel ternary oxides for batteries [8]. Motivated by these and other related efforts [9–12], we develop a nonlinear regression ML model for computing molecular atomization energies in chemical compound space [3]. Our model is based on a measure of distance in compound space that accounts for both stoichiometry and configurational variation. After training, energies are predicted for new (out-of-sample) molecular systems, differing in composition and geometry, at negligible computational cost, i.e., milliseconds instead of hours on a conventional CPU. While the model is trained and tested using atomization energies calculated at the hybrid density-functional theory (DFT) level [2,13,14], any other training set or level of theory could be used as a starting point for ML training. Cross validation on 7165 molecules yields a mean absolute error of 9.9 kcal/mol, which is an order of magnitude more accurate than counting bonds or semiempirical quantum chemistry.

We use a molecular generated database (GDB), a library of nearly 10^9 organic molecules that are stable and synthetically accessible according to organic chemistry rules [15–17]. While potentially applicable to any stoichiometry, as a proof of principle, we restrict ourselves to small organic molecules. Specifically, we define a controlled test bed consisting of all 7165 organic molecules from the GDB, with up to seven "heavy" atoms that contain C, N, O, or S, being saturated with hydrogen atoms. Atomization energies range from −800 to −2000 kcal/mol. Structural features include a rich variety of chemistry such as double and triple bonds, (hetero) cycles, carboxy, cyanide, amide, alcohol, and epoxy groups. For each of the many stoichiometries, many constitutional (differing chemical bonds) but no conformational isomers are part of this database. Based on the string representation of molecules in the database, we generated Cartesian geometries with OpenBabel [18]. Thereafter, the Perdew-Burke-Ernzerhof hybrid functional (PBE0) [19,20] approximation to hybrid DFT in a converged numerical basis, as implemented in the FHI-AIMS code [21] (tight settings/tier2 basis set), was used to compute reference atomization energies for training. Our choice of the PBE0 hybrid functional is motivated by small errors (< 5 kcal/mol) for thermochemistry data that include molecular atomization energies [22].

One of the most important ingredients for ML is the choice of an appropriate data representation that reflects prior knowledge of the application domain, i.e., a model of

the underlying physics. A variety of such "descriptors" is used by statistical methods for chem- and bioinformatics applications [23,24]. For modeling atomization energies, we use the same molecular information that enters the Hamiltonian for an electronic structure calculation, namely, the set of Cartesian coordinates, $\{\mathbf{R}_I\}$, and nuclear charges, $\{Z_I\}$. Our representation consists of atomic energies and the internuclear Coulomb repulsion operator; specifically, we represent *any* molecule by a "Coulomb" matrix $\mathbf{M}$,

$$M_{IJ} = \begin{cases} 0.5 Z_I^{2.4} & \text{for } I = J, \\ \frac{Z_I Z_J}{|\mathbf{R}_I - \mathbf{R}_J|} & \text{for } I \neq J. \end{cases} \quad (1)$$

Here, off-diagonal elements correspond to the Coulomb repulsion between atoms I and J, while diagonal elements encode a polynomial fit of atomic energies to nuclear charge.

Using ML, we attempt to construct a nonlinear map between molecular characteristics and atomization energies. This requires a measure of molecular (dis)similarity that is invariant with respect to translations, rotations, and the index ordering of atoms. To this end, we measure the distance between two molecules by the Euclidean norm of their diagonalized Coulomb matrices: $d(\mathbf{M}, \mathbf{M}') = d(\boldsymbol{\epsilon}, \boldsymbol{\epsilon}') = \sqrt{\sum_I |\epsilon_I - \epsilon'_I|^2}$, where $\boldsymbol{\epsilon}$ are the eigenvalues of $\mathbf{M}$ in order of decreasing absolute value. For matrices that differ in dimensionality, $\boldsymbol{\epsilon}$ of the smaller system is extended by zeros. Note that, by representing chemical compound space in this way, (i) any system is uniquely encoded because stoichiometry as well as atomic configuration are explicitly accounted for, (ii) symmetrically equivalent atoms contribute equally, (iii) the diagonalized $\mathbf{M}$ is invariant with respect to atomic permutations, translations, and rotations, and (iv) the distance is continuous with respect to small variations in interatomic distances or nuclear charges [25]. As discussed in Ref. [26], these are all crucial criteria for representing atomistic systems within statistical models.

In Fig. 1, relative atomization energies, as a function of $d(\mathbf{M}, \mathbf{M}')$, and a histogram of distances are shown for all pairs of molecules in our data set. The inset exemplifies the distances between three molecular species, pyrrol, thiophene, and ethanol: Within our measure of similarity, the nitrogen-containing aromatic heterocycle pyrrol is ~10 times farther away from its sulfur-containing analogue, thiophene, than from ethanol. This is due to the large difference in nuclear charges between atoms from different rows in the periodic table.

Within our ML model [27–29], the energy of a molecule $\mathbf{M}$ is a sum over weighted Gaussians,

$$E^{\text{est}}(\mathbf{M}) = \sum_{i=1}^{N} \alpha_i \exp\left[-\frac{1}{2\sigma^2} d(\mathbf{M}, \mathbf{M}_i)^2 \right], \quad (2)$$

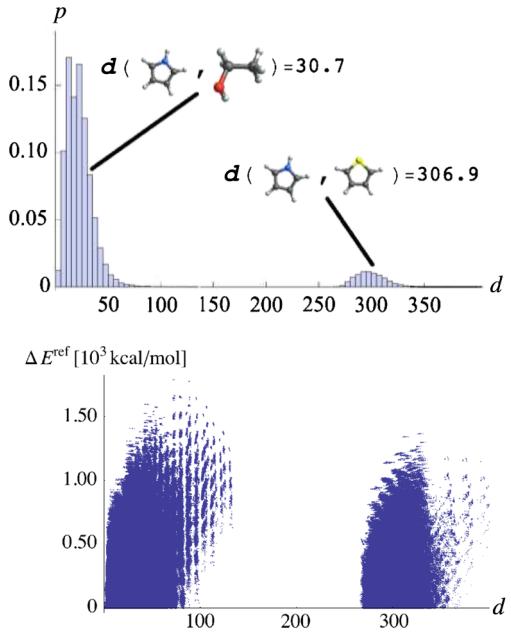

FIG. 1 (color online). Top: Distribution of distances, $d(\mathbf{M}, \mathbf{M}')$, for all molecular pairs occurring in the first 7165 small organic molecules from the GDB [15]. The inset exemplifies two distances, pyrrol/ethanol and pyrrol/thiophene (N: blue, O: red, S: yellow, C: black, and H: white). Bottom: Absolute differences in atomization energies between $\mathbf{M}$ and $\mathbf{M}'$ as a function of $d(\mathbf{M}, \mathbf{M}')$.

where i runs over all molecules $\mathbf{M}_i$ in the training set. Regression coefficients $\{\alpha_i\}$ and length-scale parameter σ are obtained from training on $\{\mathbf{M}_i, E_i^{\text{ref}}\}$. Note that each training molecule i contributes to the energy not only according to its distance, but also according to its specific weight α_i. The $\{E_i^{\text{ref}}\}$ were computed at the PBE0 DFT level of theory.

To determine $\{\alpha_i\}$, we used kernel ridge regression [28]. This regularized model limits the norm of regression coefficients, $\{\alpha_i\}$, thereby ensuring the transferability of the model to new compounds. For given length scale σ and regularization parameter λ, the explicit solution to the minimization problem,

$$\min_{\boldsymbol{\alpha}} \sum_i (E^{\text{est}}(\mathbf{M}_i) - E_i^{\text{ref}})^2 + \lambda \sum_i \alpha_i^2, \quad (3)$$

is given by $\boldsymbol{\alpha} = (\mathbf{K} + \lambda \mathbf{I})^{-1} \mathbf{E}^{\text{ref}}$, $K_{ij} = \exp[-d(\mathbf{M}_i, \mathbf{M}_j)^2 / (2\sigma^2)]$ being the kernel matrix of all training molecules and $\mathbf{I}$ denoting the identity matrix.

We used stratified [30] fivefold cross validation [28,29] for model selection and to estimate performance. Parameters λ and σ were determined in an inner loop of fivefold cross validation using a logarithmically scaling grid. This procedure is routinely applied in machine learning and statistics to avoid overfitting and overly optimistic error estimates.

The dependence of the cross-validated ML performance on the number of molecules in the training set, N, is illustrated in Fig. 2 (top). When increasing N from 500 to 7000, the mean absolute error (MAE) falls off from more than 17 kcal/mol to less than 10 kcal/mol. Furthermore, the width σ of the Gaussian kernel decreases from 460 to 25 on the distance scale of Fig. 1. Because of the discrete nature of chemical space (nuclear charges can only assume integer values), however, we do not expect continuous coverage for $N \to \infty$, implying that σ will converge to a small but finite value. The regularization hyperparameter λ remains small throughout, consistent with the fact that we model noise-free numerical solutions of the approximated Schrödinger equation. An asymptotic fit of the form $\sim 1/\sqrt{N}$, based on statistical theory [28,31], suggests that the MAE can be lowered to ~7.6 kcal/mol for $N \to \infty$. It is remarkable that, already for the here-presented, relatively small training set sizes, ML achieves errors of roughly 1% on the relevant scale of energies, outperforming bond counting or semiempirical quantum chemistry methods. The cross-validated performance for a training set size of $N = 1000$ is displayed in Fig. 2 (bottom). There is good correlation with the DFT data. For comparison, corresponding correlations are shown for bond counting [32] and semiempirical quantum chemistry (parametric method PM6 [33]) computed with MOPAC [34]. While the latter two methods exhibit a systematic shift in slope, the inset highlights that the ML correlation accurately reproduces clustering and a slope of one.

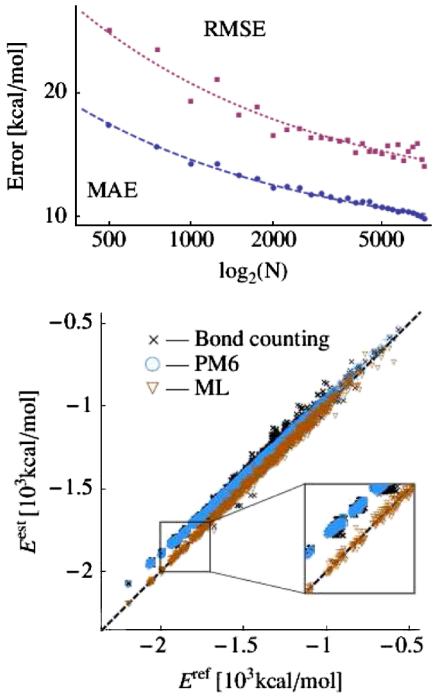

FIG. 2 (color online). Top: Cross-validated ML errors as a function of the number of molecules in the training set, N. Bottom: For $N = 1000$, correlation of DFT-PBE0 [19,20] results (E^{ref}) with ML (cross-validated) based estimates (E^{est}) of atomization energies. Correlations for bond counting [32] and semiempirical quantum chemistry (PM6 [33]) are also shown. Corresponding root mean square error (RMSE)/mean absolute error (MAE) for bond counting, PM6, and ML are 75.0/71.0, 75.1/73.1, and 30.1/14.9 kcal/mol, respectively.

Equation (2) implies that the energy of a query molecule $\mathbf{M}$ can be seen as an expansion in reference molecules $\{\mathbf{M}_i\}$. The regression weights $\{\alpha_i\}$ are scaled by the similarity between query and reference compound as measured by a Gaussian of the distance. Hence, α_i assigns a positive or negative weight for the energy contribution of the ith reference molecule. Since $\{\alpha_i\}$ are regression coefficients in a nonlinear model, i.e., after a nonlinear transformation of the training data, the resulting energy contributions are specific to the employed training set without general implications for other properties or regions of compound space. The locality of the model is measured by σ, enabling us to define a range outside of which reference molecules $\mathbf{M}_i$ can be neglected in their contribution to the energy. For a larger number N of training samples, a smaller σ is obtained and the model becomes more local in chemical space (see the Supplemental Materials for quantification [35]).

In order to assess transferability and applicability of our model to chemical compound space, we use a ML model trained on $N = 1000$ molecules (model 1k). The training set of model 1k contains all small molecules with 3 to 5 heavy atoms and a randomized stratified selection of larger compounds covering the entire energy range. The thousand Coulomb matrices corresponding to the OpenBabel configurations were included, as well as four additional Coulomb matrices per molecule. These additional matrices were scaled in order to represent the repulsive wall, the dissociative limit, and the energy minimum at $f = 1$ [36,37]. All predictions are made for molecules that were *not* used during training of the model.

For testing the transferability, we applied the 1k model to the remaining 6k molecules. The calculations yield errors that hardly change from the estimated performance in the training with a MAE of 15.2 kcal/mol. For the selected molecular subset of the seven thousand smallest molecules in the GDB [15], we therefore conclude that training on 15% of the molecules permits predictions of atomization energies for the remaining 85% with an accuracy of roughly 15 kcal/mol.

For probing its applicability, we investigated whether the 1k model can also be useful beyond the equilibrium geometries. Specifically, we calculated the functional dependence of atomization energies on scaling Cartesian geometries by a factor, f. From the 6k molecules (not used for training), we picked four chemically diverse species. Specifically, these molecules contain single bonds and branching only (C_7H_{16}), a double bond (C_6H_{12}), triple bonds including nitrogen (C_6NH_5), and a sulfur-containing

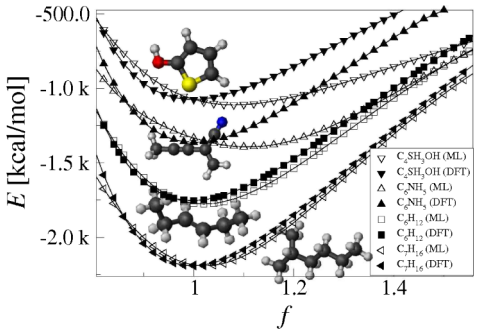

FIG. 3 (color online). Energy of atomization curves of four molecules containing single bonds and branching only (C_7H_{16}), a double bond (C_6H_{12}), triple bonds including nitrogen (C_6NH_5), and a sulfur-containing cycle with a hydroxy group (C_4SH_3OH). (From bottom to top in insets: black: Carbon; blue: Nitrogen; yellow: Sulfur; red: Oxygen; white: Hydrogen) (DFT-PBE0 and ML model 1k).

cycle with a hydroxy group (C_4SH_3OH). The resulting ML atomization energy curves (Fig. 3) correctly distinguish between the molecules, closely reproduce the DFT energy at $f = 1$, and appear continuous and differentiable throughout relevant bonding distances. For comparison, corresponding DFT potential curves are also displayed. While their well depth and position can be compared to the ML curve, their repulsive wall and dissociative limit was *not* used for training. Since DFT is a single-determinant theory, it does not reproduce the molecular dissociation limit properly and does not converge to the sum of atomic energies for large f. On the other hand, our ML model converges to the right dissociation limit by construction. Albeit significantly overestimating the position of the well depth in the case of C_4SH_3OH and C_6NH_5, the ML model is in very good agreement with the DFT data for the larger molecules. This might be due to the fact that, in the total set, larger molecules are more frequent than smaller molecules. Overall, the ML model is in good agreement with the correct physics (single and differentiable well depth of reasonable magnitude and position) as represented by the DFT potential curves. The four ML curves deviate from their dissociative and repulsive limit $[E(f = 2/3) = E(f = 3) = 0]$ used during training at most by 20 kcal/mol at $f = 2/3$ and by 8 kcal/mol at $f = 3$. We reiterate that, while the DFT curves had to be calculated explicitly for these four molecules, the ML curves correspond to analytical predictions based on a training set with 1000 *other* molecules.

We have developed a ML approach for modeling atomization energies across molecular compound space. For larger training sets, $N \geq 1000$, the accuracy of the ML model becomes competitive with mean-field electronic structure theory—at a fraction of the computational cost. We find good performance when making predictions for unseen organic molecules (transferability) and when predicting atomization energies for distorted equilibrium geometries. Our representation of molecules as Coulomb matrices is inspired by the nuclear repulsion term in the molecular Hamiltonian and free atom energies. Future extensions of our approach might be used for geometry relaxation, chemical reactions [38], molecular dynamics in various ensembles [39], or rational compound design applications [40–42]. Finally, our results suggest that the Coulomb matrix, or improvements thereof, could be of interest as a descriptor beyond the presented application.

We are thankful for helpful discussions with K. Burke, M. Cuendet, K. Hansen, J. E. Moussa, J.-L. Reymond, B. C. Rinderspacher, M. Rozgic, M. Scheffler, A. P. Thompson, M. E. Tuckerman, and S. Varma. All authors acknowledge support from the long program "Navigating Chemical Compound Space for Materials and Bio Design," IPAM, UCLA. This research used resources of the Argonne Leadership Computing Facility at Argonne National Laboratory, which is supported by the Office of Science of the U.S. DOE under Contract No. DE-AC02-06CH11357. M. R. and K.-R. M. acknowledge partial support by DFG (MU 987/4-2) and the EU (PASCAL2).

*anatole@alcf.anl.gov

[1] *Encyclopedia of Computational Chemistry*, edited by P. Ragué von Schleyer, N. Allinger, T. Clark, J. Gasteiger, P. Kollman, H. F. Schaefer III, and P. Schreiner (Wiley, New York, 1998).
[2] P. Hohenberg and W. Kohn, Phys. Rev. **136**, B864 (1964).
[3] P. Kirkpatrick and C. Ellis, Nature (London) **432**, 823 (2004).
[4] O. A. von Lilienfeld and M. E. Tuckerman, J. Chem. Phys. **125**, 154104 (2006).
[5] R. R. Coifman, I. G. Kevrekidis, S. Lafon, M. Maggioni, and B. Nadler, Multiscale Model. Simul. **7**, 842 (2008).
[6] A. P. Bartók, M. C. Payne, R. Kondor, and G. Csányi, Phys. Rev. Lett. **104**, 136403 (2010).
[7] C. M. Handley and P. L. A. Popelier, J. Chem. Theory Comput. **5**, 1474 (2009).
[8] G. Hautier, C. C. Fischer, A. Jain, T. Mueller, and G. Ceder, Chem. Mater. **22**, 3762 (2010).
[9] A. Brown, B. J. Braams, K. Christoffel, Z. Jin, and J. M. Bowman, J. Chem. Phys. **119**, 8790 (2003).
[10] S. Lorenz, A. Gross, and M. Scheffler, Chem. Phys. Lett. **395**, 210 (2004).
[11] J. Behler and M. Parrinello, Phys. Rev. Lett. **98**, 146401 (2007).
[12] J. Behler, R. Martonak, D. Donadio, and M. Parrinello, Phys. Rev. Lett. **100**, 185501 (2008).
[13] W. Kohn and L. J. Sham, Phys. Rev. **140**, A1133 (1965).
[14] A. D. Becke, J. Chem. Phys. **98**, 5648 (1993).
[15] L. C. Blum and J.-L. Reymond, J. Am. Chem. Soc. **131**, 8732 (2009).
[16] T. Fink, H. Bruggesser, and J.-L. Reymond, Angew. Chem., Int. Ed. Engl. **44**, 1504 (2005).

[17] T. Fink and J.-L. Reymond, J. Chem. Inf. Model. **47**, 342 (2007).
[18] R. Guha, M. T. Howard, G. R. Hutchison, P. Murray-Rust, H. Rzepa, C. Steinbeck, J. K. Wegner, and E. Willighagen, J. Chem. Inf. Model. **46**, 991 (2006).
[19] J. P. Perdew, M. Ernzerhof, and K. Burke, J. Chem. Phys. **105**, 9982 (1996).
[20] M. Ernzerhof and G. E. Scuseria, J. Chem. Phys. **110**, 5029 (1999).
[21] V. Blum, R. Gehrke, F. Hanke, P. Havu, V. Havu, X. Ren, K. Reuter, and M. Scheffler, Comput. Phys. Commun. **180**, 2175 (2009).
[22] B. J. Lynch and D. G. Truhlar, J. Phys. Chem. A **107**, 3898 (2003).
[23] G. Schneider, Nat. Rev. Drug Discov. **9**, 273 (2010).
[24] R. Todeschini and V. Consonni, *Handbook of Molecular Descriptors* (Wiley-VCH, Weinheim, 2009), 2nd ed.
[25] We compare two molecules using the n eigenvalues of their Coulomb matrices, n being the number of atoms, instead of all the $4n - 6$ internal degrees of freedom (3 from the coordinates, 1 from the atomic number). By consequence, this reduction in dimensionality results in a ML model that is undercomplete and that becomes invariant to certain geometrical changes. This issue could be addressed using an overcomplete metric of distance in chemical space. For example, comparing molecules via the Frobenius norm of the difference of two sorted Coulomb matrices yields a ML model with similar accuracy. Alternatively, multiple kernels could be used to more precisely control the desired number of degrees of freedom.
[26] J. Behler, J. Chem. Phys. **134**, 074106 (2011).
[27] B. Schölkopf and A. J. Smola, *Learning with Kernels* (MIT, Cambridge, MA, 2002).
[28] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (Springer, New York, 2009), 2nd ed.
[29] K.-R. Müller, S. Mika, G. Rätsch, K. Tsuda, and B. Schölkopf, IEEE Transactions on Neural Networks **12**, 181 (2001).
[30] Stratification was done by sorting energies of the training data, grouping corresponding sorted compounds into blocks of five compounds each, and, for each such block, randomly assigning one compound to one cross-validation fold. This procedure ensures that each fold covers the whole energy range.
[31] K. R. Müller, M. Finke, N. Murata, K. Schulten, and S. Amari, Neural Comput. **8**, 1085 (1996).
[32] http://www.wiredchemist.com; S. W. Benson, J. Chem. Educ. **42**, 502 (1965).
[33] J. J. P. Stewart, J. Mol. Model. **13**, 1173 (2007).
[34] James J. P. Stewart, MOPAC2009, Stewart Computational Chemistry, Colorado Springs, CO, 2008, http://openmopac.net.
[35] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevLett.108.058301 for details of model 1k and dependence of σ and λ on N.
[36] T.-C. Lim, Mol. Phys. **108**, 1589 (2010).
[37] For the repulsive wall, atomization energies for Coulomb matrices were set to zero at typical roots for covalent bonds, $f = 2/3$ [36]. For the minimum atomization energies, the finite difference derivative, $dE/df = 0$, was set to zero at $f = 1$, using $df = 0.005$. For the dissociative tail, we assume zero atomization energies at $f = 3$.
[38] G. Csányi, T. Albaret, M. C. Payne, and A. DeVita, Phys. Rev. Lett. **93**, 175503 (2004).
[39] M. E. Tuckerman, *Statistical Mechanics: Theory and Molecular Simulation* (Oxford University Press, New York, 2010).
[40] O. A. von Lilienfeld, R. D. Lins, and U. Rothlisberger, Phys. Rev. Lett. **95**, 153002 (2005).
[41] O. A. von Lilienfeld, J. Chem. Phys. **131**, 164102 (2009).
[42] D. Sheppard, G. Henkelman, and O. A. von Lilienfeld, J. Chem. Phys. **133**, 084104 (2010).