THE JOURNAL OF PHYSICAL CHEMISTRY Letters

Letter

pubs.acs.org/JPCL

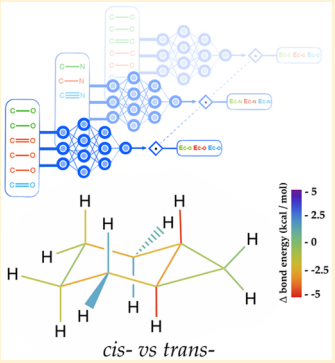
# Intrinsic Bond Energies from a Bonds-in-Molecules Neural Network

Kun Yao, John E. Herr, Seth N. Brown, and John Parkhill*

Department of Chemistry and Biochemistry, The University of Notre Dame du Lac, Notre Dame, Indiana 46556, United States

Supporting Information

**ABSTRACT:** Neural networks are being used to make new types of empirical chemical models as inexpensive as force fields, but with accuracy similar to the ab initio methods used to build them. In this work, we present a neural network that predicts the energies of molecules as a sum of intrinsic bond energies. The network learns the total energies of the popular GDB9 database to a competitive MAE of 0.94 kcal/mol on molecules outside of its training set, is naturally linearly scaling, and applicable to molecules consisting of thousands of bonds. More importantly, it gives chemical insight into the relative strengths of bonds as a function of their molecular environment, despite only being trained on total energy information. We show that the network makes predictions of relative bond strengths in good agreement with measured trends and human predictions. A Bonds-in-Molecules Neural Network (BIM-NN) learns heuristic relative bond strengths like expert synthetic chemists, and compares well with ab initio bond order measures such as NBO analysis.




Neural networks (NN) make accurate models of high dimensional functions, with chemical applications such as density functionals,[1−4] potential energy surfaces[5−32] and molecular properties.[33−60] With modern general-purpose graphical processing unit (GP-GPU) computing, NN's are inexpensive to train and evaluate, with a cost that lies much closer to a force-field than ab initio theory. To model energy extensively, most black-box NN schemes partition the energy of a molecule into fragments.[10,15,27,61] Separate networks are often used for qualitatively different energy contributions to maximize accuracy and efficiency. Besides balancing accuracy with complexity, a decomposition of the energy can yield chemical insights, and give the heuristic principles of chemistry reproducible, quantitative rigor. In this paper we use an NN to express the cohesive energy of a molecule as a sum of bond energies. Besides offering a very accurate and inexpensive decomposition of the total energy, this method, Bonds-in-Molecules Neural Network (BIM-NN) produces an instant estimate of the embedded unrelaxed bond strengths in a molecule. BIM-NN responds quantitatively to molecular geometry in a way that textbook tabulated values cannot, and updates a textbook concept used by all chemists, making it a quantative tool.

Atoms have been a popular choice of NN energy decomposition since the pioneering contributions of Behler, Khaliullin, and Parinello.[5,10,11,17] Within the atom scheme separate networks are trained for each element. Another reasonable choice for noncovalent aggregates explored in our own work is a many-body expansion.[61] These different fragmentations of the energy navigate a trade-off between breadth and accuracy. One atom's contribution to the total energy is difficult to learn because it varies significantly when engaged in different bonds. One sort of three-body interaction[27,30,61] is comparatively easy to learn, but there are too many three-body combinations in chemistry to learn them all. The purpose of this paper is to explore the advantages of bonds as a decomposition unit.

Bonds have served as the central abstraction throughout the history of chemistry.[62,63] They are a popular building block for NN inputs,[34] but to our knowledge no NN has been used to predict total energies as a sum of bond energies as we are about to describe. One reason for this might be the large, but manageable, number of bond networks required to make a general model chemistry. A complex neural network is required for BIM-NN with many bond-branches that must be dynamically learned and evaluated. We developed a general open-source software framework for producing NN models[64] of molecules, TensorMol, in which we have implemented BIM-NN, that simplifies the process of creating the bond-centered network. The complete source allowing readers to reproduce and extend this work is publicly available in the TensorMol repository.[65]

The left panel of Figure 1 schematically describes how BIM-NN is trained and evaluated. A molecule is broken down into overlapping bonds, such as C−H, C−C, C−O, etc. An optimal choice of descriptor describing the chemical environment of the bond is crucial for neural networks to reach their best performance.[11,19,33,52,66,67] This work uses our own version of the Bag of Bonds (BoBs)[33,34] descriptor, which is similar to BALM[68] for this purpose. Bonds are defined simply with an element specific distance cutoff such that only covalent bonds are included. The descriptor contains the bond length of the target bond, and lengths and angles of bonds attached to it in order of connectivity. Each type of bond has its own branch consisting of three fully connected hidden layers with 1000 neurons in each layer summed to the bond energy. The

**Received:** May 3, 2017
**Accepted:** June 2, 2017
**Published:** June 2, 2017





DOI: 10.1021/acs.jpclett.7b01072
J. Phys. Chem. Lett. 2017, 8, 2689−2694

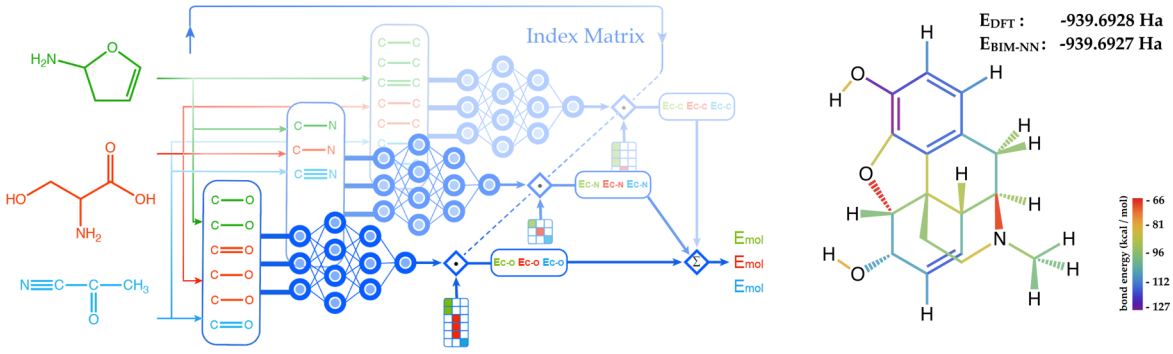
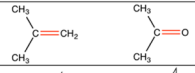
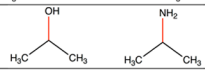
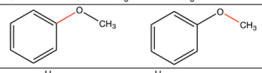
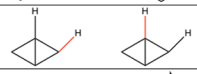
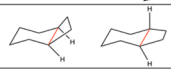



**Figure 1.** Left panel: A scheme of how a BIM-NN computational graph is trained. A batch of molecules is shattered into bonds. Bond types within the batch are fed into a type-specific subnetwork, and a linear transformation reassembles molecular energies. Right panel: Morphine, a molecule not included in the training set, with its bonds color-coded by their bond energies as calculated using BIM-NN. The total energy as a sum of bonds is accurate within 1 $mE_h$.

**Table 1. Relative Strengths of the Bond Highlighted in Red, As Predicted by a Chemist's Intuition, BIM-NN, and NBO Analysis**[a]

| Case # | | Chemist | Neural Network | NBO |
|---|---|---|---|---|
| 1 |  | ≪ | −40.5 | −0.0294 |
| 2 |  | ≪ | −17.2 | −0.0096 |
| 3 |  | ≫ | −16.5 | −0.0103 |
| 4 |  | > | 14.2 | 0.0020 |
| 5 |  | > | 10.7 | −0.0016 |
| 6 |  | < | 6.1 | −0.0030 |
| 7 |  | ≫ | 3.8 | 0.0024 |
| 8 |  | > | 3.0 | 0.0070 |
| 9 |  | > | 1.8 | 0.0016 |
| 10 |  | < | −1.7 | −0.0057 |

[a]≫ and ≪ means the chemist expected the difference to be larger than 10 kcal/mol, > and < means the difference to be between 1−10 kcal/mol. The BIM-NN column shows the predicted relative bond strength in kcal/mol. The relative bond strength is calculated by take the bond energy of left-hand bond and subtracting the bond energy of the right-hand bond. The number in the NBO column is the difference of the occupation number predicted by NBO.

energies of all bonds are summed up at the last layer to give a predicted molecular total energy, which is the training target. Back-propagation of errors[69] to the previous layers allows the bond branches to learn the strengths of their specific bond types. These errors propagate back through a linear transformation matrix that maps the many bond energies coming from different molecules produced by a type branch back onto the molecules. We use the popular GDB9 database[57,70] including H, C, N, and O for our training data, which is more than 130,000 molecules in total. GDB9 is a subset of GDB17[71] generated by using atoms (N, O, etc.) to substitute carbon atoms in neutral hydrocarbons, which eliminates molecules with certain functional groups ($R-NO_2$, R−NC). It only contains neutral isolated molecules, and implies that this iteration of our method is inapplicable to charged or noncovalently interacting molecules, although that can be changed with additional training data. Before training, 20% percent of the database was chosen randomly as a test set and kept independent. Our results also examine larger molecules, which are not a part of GDB9. Further methodological details are presented in the Supporting Information.

Our results try to answer three questions we had about BIM-NN: How well does BIM-NN predict total energies? How well does BIM-NN predict bond strength trends? What is the relationship between BIM-NN bond energies and observed chemical reactivity? We were especially excited about BIM-NN as a method to systematically and quantitatively reproduce chemical heuristics, and appealed to expert synthetic chemists to make predictions of relative bond strengths to test BIM-NN. Because BIM-NN predicts the energy of a molecule relative to



DOI: 10.1021/acs.jpclett.7b01072
J. Phys. Chem. Lett. 2017, 8, 2689−2694

free neutral atoms, we expect bond strengths in BIM-NN to correlate best with unrelaxed, homolytic dissociation energies and some evidence to this effect is presented later on. However, no specific products are implied by BIM-NN bond energies. They depend only on the equilibrium geometry. Some sorts of bonds are strong near equilibrium but can kinetically access their products which are relatively more stable, or electronically relax, and in these cases a chemist's intuition may be at variance with BIM-NN.

The training mean absolute error (vs B3LYP/6-31G(2df,p)) of the molecular energies of our BIM-NN reaches 0.86 kcal/mol, and the independent test mean absolute error is 0.94 kcal/mol. The test error is close to the training error, suggesting that the our model is neither overfitting nor under-fitting. The accuracy of our model is competitive with the state-of-art sub-1 kcal/mol accuracy on this database.[7,34,50,52,67,68,72,73] The bond-wise nature of our model makes it transferable to large molecules. The right panel of Figure 1 shows BIM-NN predicted total energy and the calculated DFT total energy of morphine molecule, which is not in our training set and contains 21 heavy atoms. We also tested BIM-NN on vitamins D2 and B5. The difference between BIM-NN and DFT energies of these two molecules are 0.6 and 1.2 kcal/mol, respectively. All these errors are small relative to the inherent errors of the B3LYP model chemistry used to produce BIM-NN. Our model can be trivially trained on higher quality chemical data as it becomes available. Correct reproduction of the bond trends described later on depends sensitively on the accuracies of the total energies. Before completion of the training process when total energies are in error by roughly 5 kcal/mol, more than 40% of the examples in Table 1 and Table S1 are answered incorrectly by BIM-NN. This is due to the fact that errors can accumulate in bonds and cancel in the molecular energy unless the NN is tightly trained on a broad sample of chemical space.

Our scheme not only calculates the total energy of a molecule, but also predicts the strengths of bonds individually. Figure 1 shows the bonds in the morphine molecule drawn with colors keyed to the bond strengths predicted by BIM-NN. The slight energy differences predicted by BIM-NN between bonds with identical bonding but different environment are perceptible. BIM-NN can suggest how the stress is distributed in an unstable isomer compared with the stable one. Figure 3 shows the geometry of *cis*- and *trans*-bicyclo[4.1.0]heptane. The DFT calculation shows that the *cis*- structure is 26.3 kcal/mol more stable than the *trans*- structure. BIM-NN predicts that the energy difference is 24.6 kcal/mol, in good agreement for this molecule outside GDB9. Figure 3 also shows how the stress is distributed in the *trans*- structure. Within the carbon framework, the strain is distributed rather equally among the bonds in the cyclohexane ring. Some C−H bonds become weaker, especially those attached to at the ring junction carbons, and some become stronger.

Figure 2 shows the bond energies of all the carbon−carbon bonds in the GDB9 database with respect to their bond lengths. Different bond types (single, double, triple, and conjugated versions of those) are indicated by color. The spread of each clouds is caused by the different chemical environments of the bonds. Bond strength correlates predictably with bond order, but general chemical trends that are less obvious are also revealed. For example, the variance of single bonds energies caused by chemical environments is much larger than those of double or triple bonds. BIM-NN also predicts that bond

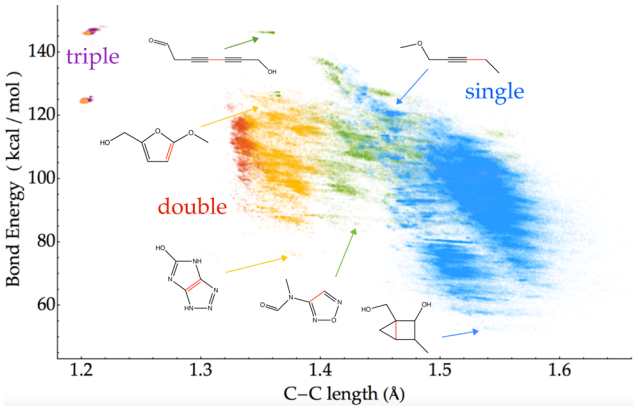



**Figure 2.** BIM-NN-predicted bond energies of the 720 000 C−C bonds in GDB9 plotted as a function of bond length. Single, conjugated single, double, conjugated double, triple, and conjugated triple bonds are shown in blue, green, red, yellow, orange, and purple, respectively. Exemplary strong and weak bonds are annotated. The roundness of each region shows that bond length is roughly as significant as all the other factors that modulate the strength of bonds combined.

lengths are roughly as relevant to their energies as all other factors combined.

BIM-NN predicts that a bicyclic C−C single bond shared by a three-membered carbon ring and a four-membered carbon ring is extremely weak, consistent with chemical intuition. The strongest C−C single bond predicted by BIM-NN is the bond that is connected with C−C triple bonds, in agreement with textbook bond dissociation energies.[74,75] BIM-NN bond energies reproduce several other established chemical rules of thumb, for example, that the strength of a CH bond decreases in the following series: methyl carbon, primary carbon in ethane, secondary carbon in propane, and tertiary carbon in isobutane. The bond energies of these four types of carbon−hydrogen bonds predicted by BIM-NN are 105.2, 105.0, 95.2, and 89.8 kcal/mol, respectively, whereas the experimental values are 105.1, 98.2, 95.1, and 93.2 kcal/mol, respectively.[74] BIM-NN also predicts that the C−C single bond in pyrrole is 1.2 kcal/mol more stable than the C−C single bond in furan, which agrees with greater bond delocalization in pyrrole than in furan, consistent with its larger NICS aromaticity.[76,77]

We asked a synthetic colleague for a quiz that consists of 19 pairs of bond strength comparisons. We compare the predictions of relative bond strength made by the chemist, natural bond orbital analysis (NBO),[78−80] and BIM-NN. Table 1 and Table S1 shows that the problems in this quiz range in difficulty from pairs separated by >10 kcal/mol to subtle differences on the order of $k_bT$, which challenges the density functional data used to produce BIM-NN. NBO makes predictions disagreeing with the chemist in five cases, while BIM-NN disagrees with the chemist in two cases: case 3 and case S6. Both of these two cases are comparisons of C−H bond strength where the carbon atom is connected to an oxygen atom and NBO also dissents with the chemist in these cases. We believe this disagreement is due to the fact that both NBO and the BIM-NN schemes do not relax the electronic structure of a molecule following bond cleavage, while a chemist takes into account the stabilization of the dissociated radicals.

To further corroborate the interpretation of BIM-NN bond energies as unrelaxed homolytic bond dissociation energies (BDEs), we have directly compared BIM-NN bond energies


DOI: 10.1021/acs.jpclett.7b01072
J. Phys. Chem. Lett. 2017, 8, 2689−2694

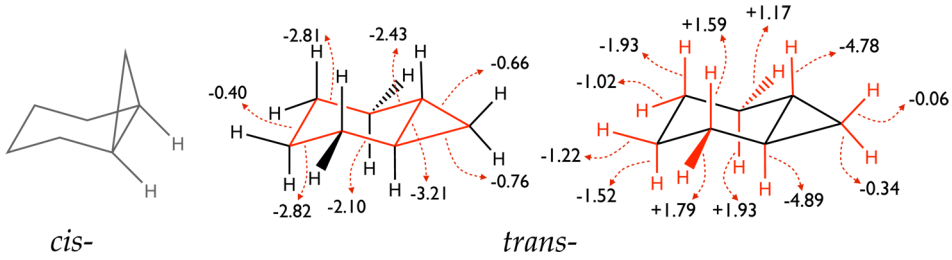



**Figure 3.** Bond energies (kcal/mol) of *trans*-bicyclo[4.1.0]heptane, given as changes relative to the *cis*-isomer show how strain energy is distributed mostly over the cyclohexane ring.

with experiment and ΔDFT. The mean absolute difference of BDEs relative to experimental best estimates from BIM-NN, geometrically unrelaxed, and relaxed DFT are 12.8, 17.2, and 8.3 kcal/mol respectively (Table S2). The unrelaxed DFT BDEs use the geometry of the molecule, but relax the electronic wave function. Please note that, by design, BIM-NN does not include either electronic or geometric relaxation energy. The fact that BIM-NN outperforms unrelaxed DFT may indicate a cancellation of electronic and geometric relaxation energies.

BIM-NN bond energies are very close to experimental BDEs for simple alkanes which do not undergo significant electronic or geometric relaxation (methyl, cyclopentyl, ethyl, cyclopropyl, *t*-butyl, H−CH($CH_3$)CN, etc.), even closer to experiment than DFT. The BIM-NN bond energy is systematically larger than the experimental BDE when the product radical is electronically stabilized (allyl, tolyl, cyclopentadienyl, 1,4 cyclohexyldienyl, and C−H bonds $\alpha$ to oxygen). The difference in these cases can be interpreted as the electronic relaxation energy of the radical when the difference between relaxed and unrelaxed DFT is small.

To investigate the electronic relaxation effect, we performed embedded DFT calculations of the relative CH BDEs in the tetrahydrofuran and actealdehyde-propene examples from Table 1, and estimated the relaxation effect by comparing the difference of the fragment reaction energies frozen at their electronic configuration in the molecule.[81] In both cases, the frozen-embedded calculation predicts the same ordering as BIM-NN and NBO. BIM-NN successfully learns many important classes of chemical heuristics such as bond types (case 1, case 4), geometric stresses (case 7, case 8), and conjugation (case 2). The interpretation consistent with these results is that BIM-NN produces intrinsic bond energies, which can be used to separate the stabilization of products from the intrinsic stabilization of a bond.

We have presented a method to cheaply and accurately sum-up a molecular total energy as an ensemble of bond-energies, BIM-NN. The method can be thought of as a quantitative version of the textbook bond energy table all chemists think about and rely upon to understand molecules. Chemists could use BIM-NN to study how small changes in geometry might affect the strength of a bond, and produce quantitative numbers to match their qualitative intuition. The method also produces accurate total energies without any sort of sophisticated nonlocal interaction between the bonds. This shows that bond networks could contribute to a general neural network model chemistry. This ambitious long-term goal merits future work, for example, extending our bond decomposition with additional noncovalent contributions that describe correct long-range forces.[82] We will soon report on further decompositions which partition electrostatic energy contributions and generalize away from equilibrium. Because of TensorMol's flexibility, these extensions fit easily within our decomposition scheme. A general model chemistry also requires more diverse sampling of chemical space and a broader swath of the periodic table, and this work is underway in our laboratory. Accurate decomposition of bonds that always occur in pairs, for example the C−C and C−H bonds in a terminal alkyne, will benefit significantly from additional nonequilibrium geometrical data. Interested parties may download our source and train their own generalized and improved BIM-NN models.

## ■ ASSOCIATED CONTENT

### Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jpclett.7b01072.

Details of neural network structure and training parameters; BIM-NN predicted bond energies of other bonds in GDB9 plotted as a function of bond length; C−H bond energies comparison between measured experimental values, BIM-NN predictions, and DFT results; timing of BIM-NN model; script that uses BIM-NN to generate the total energy of a molecule and all bond energies in the molecule (PDF)

## ■ AUTHOR INFORMATION


### Corresponding Author

*E-mail: jparkhil@nd.edu.

### ORCID

Kun Yao: 0000-0003-2032-7441

Seth N. Brown: 0000-0001-8414-2396


### Notes

The authors declare no competing financial interest.

## ■ ACKNOWLEDGMENTS

The authors would like to thank Prof. Xavier Creary for valuable discussions.

## ■ REFERENCES


(1) Snyder, J. C.; Rupp, M.; Hansen, K.; Blooston, L.; Müller, K.-R.; Burke, K. Orbital-free bond breaking via machine learning. *J. Chem. Phys.* **2013**, *139*, 224104.

(2) Snyder, J. C.; Rupp, M.; Hansen, K.; Müller, K.-R.; Burke, K. Finding density functionals with machine learning. *Phys. Rev. Lett.* **2012**, *108*, 253002.

(3) Yao, K.; Parkhill, J. Kinetic energy of hydrocarbons as a function of electron density and convolutional neural networks. *J. Chem. Theory Comput.* **2016**, *12*, 1139−1147.

(4) Kolb, B.; Lentz, L. C.; Kolpak, A. M. Discovering charge density functionals and structure-property relationships with PROPhet: A general framework for coupling machine learning and first-principles methods. *Sci. Rep.* **2017**, *7*, 1192.




DOI: 10.1021/acs.jpclett.7b01072
*J. Phys. Chem. Lett.* 2017, 8, 2689−2694

(5) Behler, J. Neural network potential-energy surfaces in chemistry: a tool for large-scale simulations. *Phys. Chem. Chem. Phys.* **2011**, *13*, 17930−17955.
(6) Handley, C. M.; Popelier, P. L. Potential energy surfaces fitted by artificial neural networks. *J. Phys. Chem. A* **2010**, *114*, 3371−3383.
(7) Schütt, K. T.; Arbabzadah, F.; Chmiela, S.; Müller, K. R.; Tkatchenko, A. Quantum-chemical insights from deep tensor neural networks. *Nat. Commun.* **2017**, *8*, 13890.
(8) Chmiela, S.; Tkatchenko, A.; Sauceda, H. E.; Poltavsky, I.; Schütt, K. T.; Müller, K.-R. Machine learning of accurate energy-conserving molecular force fields. *Sci. Adv.* **2017**, *3*, e1603015.
(9) Tian, Y.; Yan, X.; Saha, M. L.; Niu, Z.; Stang, P. J. Hierarchical self-assembly of responsive organoplatinum (II) metallacycle−TMV complexes with turn-on fluorescence. *J. Am. Chem. Soc.* **2016**, *138*, 12033−12036.
(10) Behler, J.; Parrinello, M. Generalized neural-network representation of high-dimensional potential-energy surfaces. *Phys. Rev. Lett.* **2007**, *98*, 146401.
(11) Behler, J. Atom-centered symmetry functions for constructing high-dimensional neural network potentials. *J. Chem. Phys.* **2011**, *134*, 074106.
(12) Shakouri, K.; Behler, J.; Meyer, J.; Kroes, G.-J. Accurate neural network description of surface phonons in reactive gas-surface dynamics: N2+ Ru (0001). *J. Phys. Chem. Lett.* **2017**, *8*, 2131−2136.
(13) Gastegger, M.; Kauffmann, C.; Behler, J.; Marquetand, P. Comparing the accuracy of high-dimensional neural network potentials and the systematic molecular fragmentation method: A benchmark study for all-trans alkanes. *J. Chem. Phys.* **2016**, *144*, 194110.
(14) Morawietz, T.; Singraber, A.; Dellago, C.; Behler, J. How van der Waals interactions determine the unique properties of water. *Proc. Natl. Acad. Sci. U. S. A.* **2016**, *113*, 8368.
(15) Bartók, A. P.; Payne, M. C.; Kondor, R.; Csányi, G. Gaussian approximation potentials: The accuracy of quantum mechanics, without the electrons. *Phys. Rev. Lett.* **2010**, *104*, 136403.
(16) Mones, L.; Bernstein, N.; Csanyi, G. Exploration, sampling, and reconstruction of free energy surfaces with Gaussian process regression. *J. Chem. Theory Comput.* **2016**, *12*, 5100−5110.
(17) Khaliullin, R. Z.; Eshet, H.; Kühne, T. D.; Behler, J.; Parrinello, M. Nucleation mechanism for the direct graphite-to-diamond phase transition. *Nat. Mater.* **2011**, *10*, 693−697.
(18) Smith, J. S.; Isayev, O.; Roitberg, A. E. ANI-1: An extensible neural network potential with DFT accuracy at force field computational cost. *Chem. Sci.* **2017**, *8*, 3192−3203.
(19) Jiang, B.; Guo, H. Permutation invariant polynomial neural network approach to fitting potential energy surfaces. *J. Chem. Phys.* **2013**, *139*, 054112.
(20) Zhang, Z.; Zhang, D. H. Effects of reagent rotational excitation on the H+ CHD3 → H2+ CD3 reaction: A seven dimensional time-dependent wave packet study. *J. Chem. Phys.* **2014**, *141*, 144309.
(21) Shao, K.; Chen, J.; Zhao, Z.; Zhang, D. H. Communication: Fitting potential energy surfaces with fundamental invariant neural network. *J. Chem. Phys.* **2016**, *145*, 071101.
(22) Li, J.; Guo, H. Permutationally invariant fitting of intermolecular potential energy surfaces: A case study of the Ne-C2H2 system. *J. Chem. Phys.* **2015**, *143*, 214304.
(23) Li, J.; Chen, J.; Zhao, Z.; Xie, D.; Zhang, D. H.; Guo, H. A permutationally invariant full-dimensional ab initio potential energy surface for the abstraction and exchange channels of the H+ CH4 system. *J. Chem. Phys.* **2015**, *142*, 204302.
(24) Kolb, B.; Luo, X.; Zhou, X.; Jiang, B.; Guo, H. High-dimensional atomistic neural network potentials for molecule-surface interactions: HCl scattering from Au (111). *J. Phys. Chem. Lett.* **2017**, *8*, 666.
(25) Deringer, V. L.; Csányi, G. Machine learning based interatomic potential for amorphous carbon. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2017**, *95*, 094203.
(26) Medders, G. R.; Götz, A. W.; Morales, M. A.; Bajaj, P.; Paesani, F. On the representation of many-body interactions in water. *J. Chem. Phys.* **2015**, *143*, 104102.
(27) Medders, G. R.; Babin, V.; Paesani, F. A critical assessment of two-body and three-body interactions in water. *J. Chem. Theory Comput.* **2013**, *9*, 1103−1114.
(28) Conte, R.; Qu, C.; Bowman, J. M. Permutationally invariant fitting of many-body, non-covalent interactions with application to three-body methane−water−water. *J. Chem. Theory Comput.* **2015**, *11*, 1631−1638.
(29) Manzhos, S.; Carrington, T. A random-sampling high dimensional model representation neural network for building potential energy surfaces. *J. Chem. Phys.* **2006**, *125*, 084109.
(30) Manzhos, S.; Dawes, R.; Carrington, T. Neural network-based approaches for building high dimensional and quantum dynamics-friendly potential energy surfaces. *Int. J. Quantum Chem.* **2015**, *115*, 1012−1020.
(31) Manzhos, S.; Yamashita, K.; Carrington, T., Jr. Fitting sparse multidimensional data with low-dimensional terms. *Comput. Phys. Commun.* **2009**, *180*, 2002−2012.
(32) Malshe, M.; Raff, L.; Hagan, M.; Bukkapatnam, S.; Komanduri, R. Input vector optimization of feed-forward neural networks for fitting ab initio potential-energy databases. *J. Chem. Phys.* **2010**, *132*, 204103.
(33) Rupp, M.; Tkatchenko, A.; Müller, K.-R.; Von Lilienfeld, O. A. Fast and accurate modeling of molecular atomization energies with machine learning. *Phys. Rev. Lett.* **2012**, *108*, 058301.
(34) Hansen, K.; Biegler, F.; Ramakrishnan, R.; Pronobis, W.; Von Lilienfeld, O. A.; Müller, K.-R.; Tkatchenko, A. Machine learning predictions of molecular properties: Accurate many-body potentials and nonlocality in chemical space. *J. Phys. Chem. Lett.* **2015**, *6*, 2326−2331.
(35) Hansen, K.; Biegler, F.; Ramakrishnan, R.; Pronobis, W.; von Lilienfeld, O. A.; Müller, K.-R.; Tkatchenko, A. Machine Learning Predictions of Molecular Properties: Accurate Many-Body Potentials and Nonlocality in Chemical Space. *J. Phys. Chem. Lett.* **2015**, *6*, 2326−2331.
(36) Montavon, G.; Rupp, M.; Gobre, V.; Vazquez-Mayagoitia, A.; Hansen, K.; Tkatchenko, A.; Müller, K.-R.; von Lilienfeld, O. A. Machine learning of molecular electronic properties in chemical compound space. *New J. Phys.* **2013**, *15*, 095003.
(37) Pilania, G.; Wang, C.; Jiang, X.; Rajasekaran, S.; Ramprasad, R. Accelerating materials property predictions using machine learning. *Sci. Rep.* **2013**, *3*, 2810.
(38) Ghasemi, S. A.; Hofstetter, A.; Saha, S.; Goedecker, S. Interatomic potentials for ionic systems with density functional accuracy based on charge densities obtained by a neural network. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2015**, *92*, 045131.
(39) Schütt, K.; Glawe, H.; Brockherde, F.; Sanna, A.; Müller, K.; Gross, E. How to represent crystal structures for machine learning: Towards fast prediction of electronic properties. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2014**, *89*, 205118.
(40) Olivares-Amaya, R.; Amador-Bedolla, C.; Hachmann, J.; Atahan-Evrenk, S.; Sánchez-Carrera, R. S.; Vogt, L.; Aspuru-Guzik, A. Accelerated computational discovery of high-performance materials for organic photovoltaics by means of cheminformatics. *Energy Environ. Sci.* **2011**, *4*, 4849−4861.
(41) Ma, X.; Li, Z.; Achenie, L. E.; Xin, H. Machine-learning-augmented chemisorption model for CO2 electroreduction catalyst screening. *J. Phys. Chem. Lett.* **2015**, *6*, 3528−3533.
(42) Ediz, V.; Monda, A. C.; Brown, R. P.; Yaron, D. J. Using molecular similarity to develop reliable models of chemical reactions in complex environments. *J. Chem. Theory Comput.* **2009**, *5*, 3175−3184.
(43) Lopez-Bezanilla, A.; von Lilienfeld, O. A. Modeling electronic quantum transport with machine learning. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2014**, *89*, 235411.
(44) Cuny, J.; Xie, Y.; Pickard, C. J.; Hassanali, A. A. Ab initio quality NMR parameters in solid-state materials using a high-dimensional neural-network representation. *J. Chem. Theory Comput.* **2016**, *12*, 765−773.
(45) Hachmann, J.; Olivares-Amaya, R.; Atahan-Evrenk, S.; Amador-Bedolla, C.; Sánchez-Carrera, R. S.; Gold-Parker, A.; Vogt, L.;


DOI: 10.1021/acs.jpclett.7b01072
*J. Phys. Chem. Lett.* 2017, 8, 2689−2694

Brockway, A. M.; Aspuru-Guzik, A. The Harvard clean energy project: large-scale computational screening and design of organic photovoltaics on the world community grid. *J. Phys. Chem. Lett.* **2011**, *2*, 2241−2251.
(46) Hachmann, J.; Olivares-Amaya, R.; Jinich, A.; Appleton, A. L.; Blood-Forsythe, M. A.; Seress, L. R.; Roman-Salgado, C.; Trepte, K.; Atahan-Evrenk, S.; Er, S.; et al. Lead candidates for high-performance organic photovoltaics from high-throughput quantum chemistry−the Harvard Clean Energy Project. *Energy Environ. Sci.* **2014**, *7*, 698−704.
(47) Simon, V.; Gasteiger, J.; Zupan, J. A combined application of two different neural network types for the prediction of chemical reactivity. *J. Am. Chem. Soc.* **1993**, *115*, 9148−9159.
(48) Hautier, G.; Fischer, C. C.; Jain, A.; Müller, T.; Ceder, G. Finding nature's missing ternary oxide compounds using machine learning and density functional theory. *Chem. Mater.* **2010**, *22*, 3762−3767.
(49) Ediz, V.; Lee, J. L.; Armitage, B. A.; Yaron, D. Molecular engineering of torsional potentials in fluorogenic dyes via electronic substituent effects. *J. Phys. Chem. A* **2008**, *112*, 9692−9701.
(50) Barker, J.; Bulin, J.; Hamaekers, J.; Mathias, S. Localized Coulomb descriptors for the Gaussian approximation potential. 2016, arXiv:1611.05126. arXiv.org e-Print archive. https://arxiv.org/abs/1611.05126 (accessed May 31, 2017).
(51) Chen, L.; Gasteiger, J. Knowledge discovery in reaction databases: Landscaping organic reactions by a self-organizing neural network. *J. Am. Chem. Soc.* **1997**, *119*, 4033−4042.
(52) Collins, C. R.; Gordon, G. J.; von Lilienfeld, O. A.; Yaron, D. J. Constant size molecular descriptors for use with machine learning. 2017, arXiv:1701.06649. arXiv.org e-Print archive. https://arxiv.org/abs/1701.06649 (accessed May 31, 2017).
(53) Wu, Z.; Ramsundar, B.; Feinberg, E. N.; Gomes, J.; Geniesse, C.; Pappu, A. S.; Leswing, K.; Pande, V. MoleculeNet: A benchmark for molecular machine learning. 2017, arXiv:1703.00564. arXiv.org e-Print archive. archive. https://arxiv.org/abs/1703.00564 (accessed May 31, 2017).
(54) Isayev, O.; Fourches, D.; Muratov, E. N.; Oses, C.; Rasch, K.; Tropsha, A.; Curtarolo, S. Materials cartography: representing and mining materials space using structural and electronic fingerprints. *Chem. Mater.* **2015**, *27*, 735−743.
(55) Rupakheti, C.; Virshup, A.; Yang, W.; Beratan, D. N. Strategy to discover diverse optimal molecules in the small molecule universe. *J. Chem. Inf. Model.* **2015**, *55*, 529−537.
(56) Bodor, N.; Harget, A.; Huang, M. J. Neural network studies. 1. Estimation of the aqueous solubility of organic compounds. *J. Am. Chem. Soc.* **1991**, *113*, 9480−9483.
(57) Virshup, A. M.; Contreras-García, J.; Wipf, P.; Yang, W.; Beratan, D. N. Stochastic voyages into uncharted chemical space produce a representative library of all possible drug-like compounds. *J. Am. Chem. Soc.* **2013**, *135*, 7296−7303.
(58) Qu, X.; Latino, D. A.; Aires-de Sousa, J. A big data approach to the ultra-fast prediction of DFT-calculated bond energies. *J. Cheminf.* **2013**, *5*, 34.
(59) Nelson, T.; Fernandez-Alberti, S.; Chernyak, V.; Roitberg, A. E.; Tretiak, S. Nonadiabatic excited-state molecular dynamics: Numerical tests of convergence and parameters. *J. Chem. Phys.* **2012**, *136*, 054108.
(60) Kvasnicka, V.; Sklenak, S.; Pospichal, J. Neural network classification of inductive and resonance effects of substituents. *J. Am. Chem. Soc.* **1993**, *115*, 1495−1495.
(61) Yao, K.; Herr, J. E.; Parkhill, J. The many-body expansion combined with neural networks. *J. Chem. Phys.* **2017**, *146*, 014106.
(62) Tully, J. C. Diatomics-in-molecules potential energy surfaces. I. First-row triatomic hydrides. *J. Chem. Phys.* **1973**, *58*, 1396−1410.
(63) Tully, J. C. Diatomics-in-molecules potential energy surfaces. II. Nonadiabatic and spin-orbit interactions. *J. Chem. Phys.* **1973**, *59*, 5122−5134.
(64) Abadi, M.; Agarwal, A.; Barham, P.; Brevdo, E.; Chen, Z.; Citro, C.; Corrado, G. S.; Davis, A.; Dean, J.; Devin, M. et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. 2016, arXiv:1603.04467. arXiv.org e-Print archive. https://arxiv.org/abs/1603.04467 (accessed May 31, 2017).
(65) Yao, K.; Herr, J.; Parkhill, J. TensorMol: A Statistical Model of Molecular Structure. 2017; https://jparkhill.github.io/TensorMol/ (accessed May 31, 2017).
(66) Bartók, A. P.; Kondor, R.; Csányi, G. On representing chemical environments. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2013**, *87*, 184115.
(67) Faber, F. A.; Hutchison, L.; Huang, B.; Gilmer, J.; Schoenholz, S. S.; Dahl, G. E.; Vinyals, O.; Kearnes, S.; Riley, P. F.; von Lilienfeld, O. A. Fast machine learning models of electronic and energetic properties consistently reach approximation errors better than DFT accuracy. 2017, arXiv:1702.05532. arXiv.org e-Print archive. https://arxiv.org/abs/1702.05532 (accessed May 31, 2017).
(68) Huang, B.; Von Lilienfeld, O. A. Communication: Understanding molecular representations in machine learning: The role of uniqueness and target similarity. *J. Chem. Phys.* **2016**, *145*, 161102.
(69) Rumelhart, D. E.; Hinton, G. E.; Williams, R. J. Learning representations by back-propagating errors. *Nature* **1986**, *323*, 533−536.
(70) Ramakrishnan, R.; Dral, P. O.; Rupp, M.; Von Lilienfeld, O. A. Quantum chemistry structures and properties of 134 kilo molecules. *Sci. Data* **2014**, *1*, 140022.
(71) Ruddigkeit, L.; Van Deursen, R.; Blum, L. C.; Reymond, J.-L. Enumeration of 166 billion organic small molecules in the chemical universe database GDB-17. *J. Chem. Inf. Model.* **2012**, *52*, 2864−2875.
(72) Hansen, K.; Montavon, G.; Biegler, F.; Fazli, S.; Rupp, M.; Scheffler, M.; Von Lilienfeld, O. A.; Tkatchenko, A.; Müller, K.-R. Assessment and validation of machine learning methods for predicting molecular atomization energies. *J. Chem. Theory Comput.* **2013**, *9*, 3404−3419.
(73) Kearnes, S.; McCloskey, K.; Berndl, M.; Pande, V.; Riley, P. Molecular graph convolutions: moving beyond fingerprints. *J. Comput.-Aided Mol. Des.* **2016**, *30*, 595−608.
(74) McMillen, D. F.; Golden, D. M. Hydrocarbon bond dissociation energies. *Annu. Rev. Phys. Chem.* **1982**, *33*, 493−532.
(75) Speight, J. G. *Lange's Handbook of Chemistry*; McGraw-Hill: New York, 2005.
(76) Horner, K. E.; Karadakov, P. B. Chemical bonding and aromaticity in furan, pyrrole, and thiophene: a magnetic shielding study. *J. Org. Chem.* **2013**, *78*, 8037−8043.
(77) Schleyer, P. v. R.; Maerker, C.; Dransfeld, A.; Jiao, H.; Hommes, N. J. R. v. E. Nucleus-Independent chemical shifts: a simple and efficient Aromaticity Probe. *J. Am. Chem. Soc.* **1996**, *118*, 6317−6318.
(78) Reed, A. E.; Weinstock, R. B.; Weinhold, F. Natural population analysis. *J. Chem. Phys.* **1985**, *83*, 735−746.
(79) Foster, J.; Weinhold, F. Natural hybrid orbitals. *J. Am. Chem. Soc.* **1980**, *102*, 7211−7218.
(80) Weinhold, F. Natural bond orbital analysis: a critical overview of relationships to alternative bonding perspectives. *J. Comput. Chem.* **2012**, *33*, 2363−2379.
(81) Manby, F. R.; Stella, M.; Goodpaster, J. D.; Miller, T. F., III A simple, exact density-functional-theory embedding scheme. *J. Chem. Theory Comput.* **2012**, *8*, 2564−2568.
(82) Tkatchenko, A.; DiStasio, R. A., Jr; Car, R.; Scheffler, M. Accurate and efficient method for many-body van der Waals interactions. *Phys. Rev. Lett.* **2012**, *108*, 236402.


DOI: 10.1021/acs.jpclett.7b01072
*J. Phys. Chem. Lett.* 2017, 8, 2689−2694