BAND NN: A Deep Learning Framework for Energy Prediction and Geometry Optimization of Organic Small Molecules

Siddhartha Laghuvarapu[†], Yashaswi Pathak[†], and U. Deva Priyakumar *

Recent advances in artificial intelligence along with the development of large data sets of energies calculated using quantum mechanical (QM)/density functional theory (DFT) methods have enabled prediction of accurate molecular energies at reasonably low computational cost. However, machine learning models that have been reported so far require the atomic positions obtained from geometry optimizations using high-level QM/DFT methods as input in order to predict the energies and do not allow for geometry optimization. In this study, a transferable and molecule size-independent machine learning model bonds (B), angles (A), nonbonded (N) interactions, and dihedrals (D) neural network (BAND NN) based on a chemically intuitive representation inspired by molecular mechanics force fields is presented. The model predicts the atomization energies of equilibrium and nonequilibrium structures as sum of energy contributions from bonds (B), angles (A), nonbonds (N), and dihedrals (D) at remarkable accuracy. The robustness of the proposed model is further validated by calculations that span over the conformational, configurational, and reaction space. The transferability of this model on systems larger than the ones in the data set is demonstrated by performing calculations on selected large molecules. Importantly, employing the BAND NN model, it is possible to perform geometry optimizations starting from nonequilibrium structures along with predicting their energies. © 2019 Wiley Periodicals, Inc.

DOI: 10.1002/jcc.26128

Introduction

Accurate estimation of molecular energies is important for reliable modeling of various chemical and biological phenomena in general. Quantum mechanical (QM) and density functional theory (DFT) methods are the methods of choices for the calculation of accurate molecular energies and physicochemical properties. However, application of these methods to molecular systems is computationally expensive and is impractical for large systems. For modeling such systems, one resorts to the use of molecular mechanics (MM) force fields methods which are computationally tractable.[1–3] Force fields provide the potential energy of a molecule as a function of nuclear positions and have empirical parameters that are derived based on their ability to reproduce certain experimental and QM data via a detailed optimization procedure.[4–6] Though the force field methods in general are widely used to model biological macromolecules to study their dynamics, structural, and thermodynamic properties, they are considered less accurate compared to ab initio or DFT methods.

In an attempt to develop new methods for predicting energies that are of DFT quality but are comparable to MM in terms of the computational cost, energy predictions have become an important application of supervised machine learning algorithms.[7–10] These algorithms have been shown to efficiently recognize patterns on training data, which can be applied on unseen data. Traditionally, various regression techniques using kernel-based methods[11] were used that convert three-dimensional coordinates of a molecule into fixed-length feature coordinates.[12–14] Recently, deep learning has become the sought after method for various supervised learning tasks due to their superior performance in several fields, primarily computer cision and natural language processing.[15–17] Various computational chemistry tasks[18] including QM property prediction,[19–22] protein structure prediction,[23–25] protein–protein interactions,[26] material property prediction,[7,27,28] retrosynthesis,[29] and drug discovery[30–33] have been the targets of the machine learning methods and more recently deep learning applications.[8]

In order to provide a molecule as an input to a supervised learning algorithm, accurate description of a molecule as a vector is required.[22,34] In other words, it is helpful to have a vector representation that captures as much chemical information as possible. The descriptor should precisely capture the atomic environment of each atom and should be sensitive to small changes in relative atomic positions. As hypothesized by Behler,[35] molecular descriptors should follow the properties such as rotational and translational invariance, invariance with respect to the permutation of atoms, and provide a unique description of the atomic positions. Molecular descriptors in general suffer from inconsistency in terms of the size of molecules since most supervised learning algorithms require a fixed-length representation of the input. Various approaches were

[a] *S. Laghuvarapu, Y. Pathak, U. D. Priyakumar*
Center for Computational Natural Sciences and Bioinformatics, International Institute of Information Technology, Hyderabad 500 032, India
E-mail: deva@iiit.ac.in
[†]*These authors contributed equally to this work.*
Contract Grant sponsor: Science and Engineering Research Board; Contract Grant number: EMR/2016/007697; Contract Grant sponsor: DST-SERB

proposed to tackle this problem. These approaches[12–14] extend the descriptor of every molecule in the set to the largest length descriptor by appending zeros at the end. These methods are not readily applicable to molecules larger than the ones trained with. Recent approaches have expressed total energy in terms of contributions from individual atoms[20,21,36] or have total energy broken down into contribution from individual bonds,[37] where the individual feature vectors have fixed sizes.

The recent machine learning (ML)-based methods generate DFT-level accurate potential energy surfaces, but their feature vectors are derived by transforming the nuclear coordinates of the constituent atoms, rather than explicit chemically intuitive terms. Smith et al.[36] used modification of symmetry functions originally developed by Behler and Parinello[38] to represent the local environment of each atom which are further used as inputs for the neural networks (NNs). Bartók et al. used smooth overlap of atomic positions to generate feature vectors.[39] Schütt et al. in their works[20,21] used nuclear charges (Z) and a matrix of interatomic distances as input to their model to find the energy of the molecule.

Although methods have been proposed that explicitly build feature vectors based on the bond topology of a molecule,[13,14] to the best of our knowledge they have not been demonstrated to generate potential energy surfaces or work on molecules larger than the ones present in the data set. In this study, we propose a novel molecular descriptor inspired by classical force fields terms[1]—bonds (B), angles (A), nonbonded (N) interactions, and dihedrals (D), which is named as BAND in this manuscript. A molecule is broken down into these terms, and energy contribution from each of these terms is measured through several feed-forward NNs. The sum of energies from each of the terms gives the total energy of the molecule. Through a series of studies that span over the conformational and configurational space, we show that our model can predict energies and potential energy surfaces accurate to DFT level. The applicability can be extended to molecules larger than the ones trained in the data set. We also demonstrate the ability of our model to perform geometry optimization of molecules to minimum energy when provided with an approximate structure over a defined bond topology. This is possible due to the nature of our molecular descriptor which is built taking into consideration the explicit bond topology of the molecule.

Theory

Deep learning[40] has been shown to learn complex nonlinear functions through artificial NNs. BAND NN proposed here uses feed-forward fully connected deep NNs. These consist of multiple layers of nodes—an input layer, one or more hidden layers, and an output layer. Each node is activated through weighted inputs from the previous layer and a nonlinear activation function. The "weights" are the optimizable parameters which can be trained through back propagation of derivatives of an objective function with respect to each of them. The objective or cost function is a measure of deviation of the predicted output from the ground truth. As mentioned earlier, NNs need a fixed-length input feature vector. This creates a fundamental problem of obtaining accurate feature vectors starting from typical molecular representations such as internal and Cartesian coordinates whose dimensions change with respect to the number of atoms. Such a fixed-length representation can further be used to train the NNs to predict molecular properties. The following sections describe the feature vector/molecular representation, their relationship with classical force fields, and the ML model used here.

BAND molecular descriptor

A molecular descriptor that captures the essence of typical MM force field equations is used here. Each molecule is broken down into bonded pairs (atoms that are adjacent) and nonbonded pairs (atom pairs that are not adjacent). From this, lists of angles identified as two consecutive bonds forming an angle and lists of dihedrals identified as three consecutive bonds forming a dihedral angle were created. Each atom is represented by an eight-dimensional feature vector: first four dimensions representing the atom name (the data set used here involves only four atoms C, N, O, and H) and the second four dimensions representing the atom type in terms of how many of the C, N, O, and H atoms are connected to it (Fig. 1) essentially capturing the atom type as referred to in force fields.[1–3] Each bond is represented by a 17-dimensional vector which is the concatenation of the vectors representing the two atoms (eight dimensions each) that form the bond followed the bond length. For the angle, it is the combination of the three atomic representations (24) followed by the bond angle and two bond lengths making it a 27-dimensional vector. Similarly for the dihedral angle, it is a 38-dimensional vector made by four atomic representations followed by the dihedral angle, two angles and three bond lengths as given in Figure 1. The nonbond pair representation is similar to bonds where the bond length is replaced by the internuclear distance.

Resemblance to classical force field equations

A typical force field[1] equation is represented as the sum of energy contributions from the bonded (E_{bonded}) and nonbonded terms ($E_{\text{nonbonded}}$). The E_{bonded} term usually involves energy as a function of bond lengths, bond angles, and dihedrals angles in addition to other terms like Urey–Bradley and improper dihedral terms depending on the force field, and the $E_{\text{nonbonded}}$ term is typically a combination of an electrostatic and Lennard-Jones terms.

$$E_{\text{total}} = E_{\text{bonded}} + E_{\text{nonbonded}}, \quad (1)$$

$$E_{\text{bonded}} = E_{\text{bonds}} + E_{\text{angles}} + E_{\text{dihedrals}}. \quad (2)$$

The molecular representation proposed here is inspired by the force field equations where the total energy is expressed as sum of individual contributions from the bonded (bonds, angles, and dihedrals) and nonbonded terms. In the force fields, the individual terms of the equation are expressed as a function of the nuclear coordinates in terms of bond lengths, angles, internuclear distances, and so on, along with their characteristic constants. For example, E_{bonds} is given as

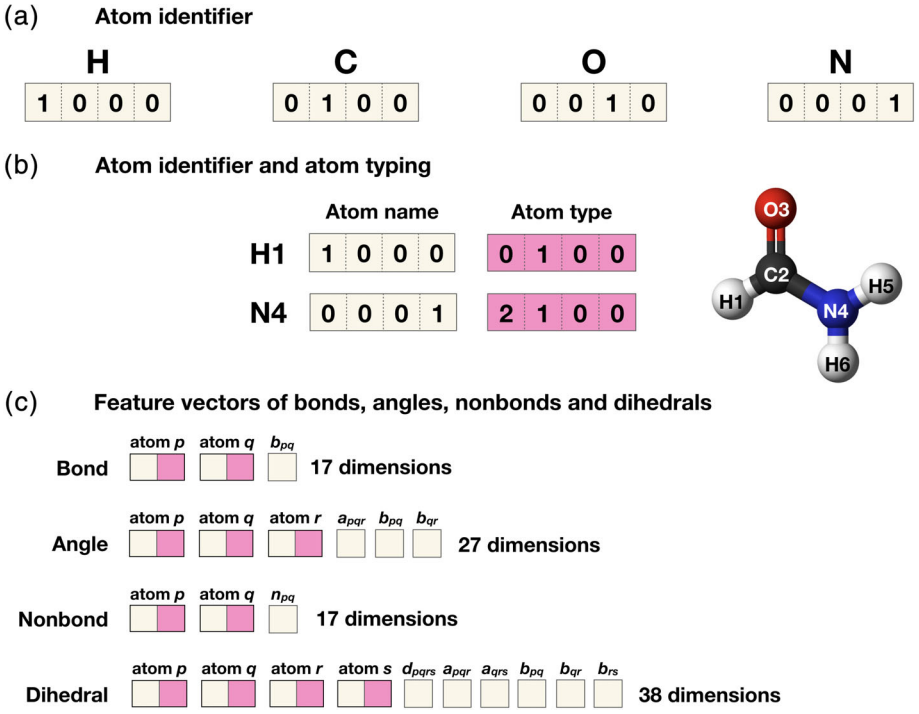

Figure 1. a) Four dimensional feature vector for the atom name. b) Eight dimensional feature vector for atom name and type. The atomic representation of two select atoms in formaldehyde is shown. c) Schematic representation of the feature vectors of bonds, angles, nonbonds and dihedrals. [Color figure can be viewed at wileyonlinelibrary.com]

$$E_{\text{bonds}} = \sum_{\text{bonds}} k_b (b - b_0)^2. \quad (3)$$

$$E = \sum_{\text{bonds}} E_B + \sum_{\text{angles}} E_A + \sum_{\text{nonbonds}} E_N + \sum_{\text{dihedrals}} E_D. \quad (4)$$

Here the constant k_b is the force constant that is characteristic of bond formed by the two participating atom types, b is the bond length, and b_0 is the equilibrium bond length. The atom type typically captures the nature of the atom, which comprises the atomic number and its connectivity. The molecular representation used here captures this by the eight-dimensional vector for each atom. One modification is the implicit consideration of coupling between stretching and bending, and stretching, bending, and rotation about single bonds akin to the class II force fields[41] (see Fig. 1).

The model

In this model, the atomization energy (difference between the molecular energy and that of the constituent atoms calculated at the DFT level) is expressed as the sum of the contributions from bonds, angles (coupled with bonds), dihedrals (coupled with bonds and angles), and nonbonds. More specifically, the contribution from each of these is estimated using a feed-forward fully connected NN. Four different models were trained for measuring contributions from bonds, angles, dihedrals, and nonbonded terms (see Fig. 2). Each of these bonds, angles, dihedrals, and nonbonded terms share the same weights, and different types of these are differentiated by their feature vectors. This allows the model to be scalable with the number of atoms (or other bonded/nonbonded terms) as the final energy is only expressed as sum of the individual contribution from each term as given below:

Methodology

Data selection

A subset of ANI-1 data set,[42] which is a large data set of non-equilibrium DFT total energy calculations for organic molecule with about 22 million molecular structures for 57,462 minimum energy structures, was used for developing the ML model. These molecules were picked from the GDB-11 data set[43,44] that has up to eight heavy atoms containing only H, C, N, and O. In addition to the equilibrium geometries obtained by performing geometry optimizations on ~57,000 molecules at the ωB97X/6-31G(d), Smith et al. have used normal-mode sampling to generate hundreds of nonequilibrium structures for each of the equilibrium structures resulting in ~22 million data points. Single-point energies of these configurations were calculated using the same method.[45] Although most methods use the QM-9 data set,[46] the conformation space is limited to equilibrium structures only and hence does not allow for calculating energies of nonequilibrium structures and hence geometry optimizations. All the equilibrium configurations along with each of their nonequilibrium structures whose relative energies with respect to the corresponding minimum energy structure are less than 30 kcal/mol were used for this study. The rationale are that (a) most of the structure generation software (such as Gaussview[47]) are able to give initial geometries that are not too far away from the minimum and (b) most of the drug

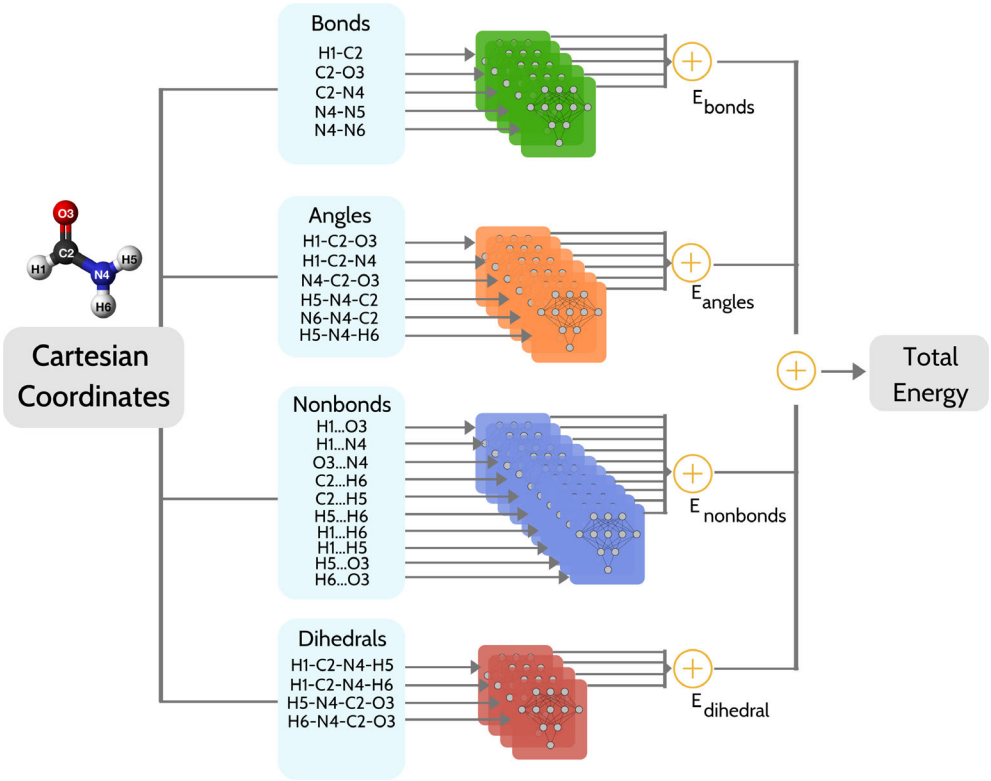

Figure 2. Schematic representation of the NN architecture used for BAND NN. As an example, the list of bonds, angles, nonbonds, and dihedrals for formaldehyde along with the number of NNs used to predict the energies of each of these are shown. [Color figure can be viewed at wileyonlinelibrary.com]

design/biomolecular simulations do not aim to model bond breaking/forming. Hence, optimization of structures generated using standard visualization software programs and for the purposes of such molecular modeling exercises, the chosen subset of the data set is deemed adequate.

Data preprocessing

Initial task is to make a list of all bonds, angles, nonbonds, and dihedral angles for each of the configurations in the data set for representation along the feature vectors proposed here. For a given molecule, the equilibrium structure was chosen to derive the molecular representation of its own and all its nonequilibrium structures. The list of bonds were generated using RDKit[48] based on the atomic coordinates of equilibrium structure which are extended to corresponding nonequilibrium configurations. Once the list of bonds were derived, the lists of angles were generated by taking all possible 1,3 neighbors that are connected to 2, and similarly all 1,4 neighbors where 2 and 3 are connected were taken as dihedrals. For the nonbonded lists, all pairs except 1,2 whose distances are less than 6 Å in the equilibrium structure were considered.

Training

Keras deep learning framework[49] with TensorFlow[50] backend was used for all training and validation purposes. Fully connected networks were used for bonds, angles, nonbonds, and dihedrals. Each network has an input layer, three hidden layers for each type, and an output layer that measures the energy contribution from that term. Table 1 gives the dimensions of bond, angle, nonbond, and dihedral networks used for BAND NN model. The output layer is a one-dimensional vector that predicts the energy contribution from that particular network. The total energy contribution is the sum of energy predictions from all the networks. A train–test–validation randomly split in the ratio of 80–10–10 was used in this work. This resulted in ~6.1 million data points in the training set and ~760,000 data points each in the test and validation sets. Adam optimizer was used for updating weights with $\beta_1 = 0.9$ and $\beta_2 = 0.999$ as suggested by Kingma and Ba.[51] Learning rate was set at 0.01 initially which was then gradually decreased to 10^{-5} by a factor of 10. All the intermediate layers were activated using the rectified linear unit (ReLU) activation function.[52] The objective minimization function is the mean squared error between the predicted and actual atomization energies. The training data were iterated for 20 epochs until no notable increase in validation accuracy was observed.

Geometry optimization

As mentioned earlier, the ANI-1 data set includes nonequilibrium structures that span over conformational and configurational space. This enables accurate prediction of energies at regions not limited to only minima on the potential energy surface but also higher energy structures. Geometry optimization involves finding

Table 1. Dimensions of the input and hidden layers of the network architecture of BAND NN.

Type of network	Input dimensions	Hidden layer dimensions
Bonds	17	128–256–128
Angles	27	128–350–128
Nonbonds	17	128–256–128
Dihedrals	38	128–512–128

Output dimension for each of the network is 1.

the least energy structure of a molecule (minimum on its potential energy surface) given an approximate structure over a defined bond topology. In this study, the suitability of the proposed BAND NN model to be used for geometry optimization is demonstrated. The optimization technique used is the Nelder–Mead's method,[53] which is a popularly used direct search method for nonlinear optimization. The method is initialized by construction of a simplex by randomly sampling points on the target surface. The method propagates through generation of a sequence of simplices by repeatedly replacing the worst point on the simplex with better ones. The algorithm terminates either when the working simplex is sufficiently small or when the differences in function values on the vertices of the simplex is less than a threshold. The implementation of Nelder–Mead's optimizer method in the Scikit-learn library with the default parameters[54] was used for the results reported in the study. Algorithm 1 (see below) describes the procedure followed for optimization of a molecule starting from its Cartesian coordinates and a defined bond topology.

The scripts along with example files for generation of molecular feature, training the model, and prediction of energies are provided at https://github.com/devalab/BAND-NN

Algorithm: Procedure for Geometry Optimization

Input: atomic coordinates, bond connectivity list
Initialise x to a z-matrix computed from atomic coordinates
Initialise $T \leftarrow 3$. This is a hyperparameter
Initialise *history* $\leftarrow$ [(∞, x), (∞, x), …T times], *terminate* $\leftarrow$ *False*
f is the function that takes z-matrix as the input and returns energy computed from BAND NN
while *terminate* = *False* **do**
  energy, x = *History*[0]
  Set x to a different representation of z-matrix randomly
  Minimize $f(x)$ using Nelder–Mead's optimization procedure.
  This step returns $energy'$, x' at minima of f
  Append ($energy'$, x') to *history* and sort *history*
  Set *worst_performer* to the last element of *history*
  if *worst_performer* = $energy'$, x' **then**
    Set *terminate* = *True*
  else
    Delete last element from history
  end
end.

Results and Discussion

In this section, the accuracy of the model to predict atomization energies of molecules in the data set and slightly larger molecules are presented. Following this, the ability of the BAND NN model to effectively learn the configurational and conformational space is demonstrated by predicting relative energies of isomers $C_{11}H_{22}$ and by performing potential energy scans on large drug molecules. This is followed by discussions on the predictive ability of the model for reaction energies of common organic reactions. Finally the importance of including the 3,4-body terms for accurate predictions and the capability of BAND NN model for utilization in geometry optimizations are presented.

Accuracy of the BAND NN model

As mentioned in the section above, all the conformers that were under 30 kcal/mol in the ANI-1 data set from the corresponding minimum energy structure were chosen for this study. This data set had about 7.6 million conformers, and a 80–10–10 split for training, testing, and validation was done on the data set. A mean absolute error of 1.45 kcal/mol on the test set was obtained, which is expected to be significantly better than the small molecule force fields in general. The distribution of the absolute errors calculated for the test set comprising about 700,000 structures is given in Figure 3a. Predicted atomization energies of about 75% of structures in the test data set are within 2 kcal/mol. To test the transferability of the BAND NN model to molecules with number of atoms more than that present in the training data set, energies of molecules and their high energy structures with 10 heavy atoms were calculated (calculations on much larger systems are discussed later). Smith et al. performed normal-mode sampling on 134 randomly chosen molecules with 10 heavy atoms from GDB-11 data set.[43,44] From these, we picked all structures whose relative energies are under 30 kcal/mol with respect to their corresponding minimum. This resulted in 1500 structures, and the mean absolute error of the atomization energies predicted using BAND NN for this set was found to be 2.1 kcal/mol, which demonstrate the transferability of the model to molecules larger than the ones trained with. The distribution of the absolute errors for this set of structures is given in Figure 3b.

BAND NN is based on a feature vector inspired from classical force field terms; direct comparisons is more appropriate with models based on comparable feature vectors. The Bag of Bonds approach reports a mean absolute error of 1.5 kcal/mol on 7000 molecules from GDB-7 and 2.0 kcal/mol on 30% of QM9.[13] Bonds-in-molecules NN reports a mean absolute error of 0.94 kcal/mol on the QM9 datA set.[37] Other recent approaches such as SchNet report a mean absolute error of 0.31 on QM9.[20] Recently, PhysNet model was proposed by Unke and Meuwly, which reports a mean absolute error of 0.19 kcal/mol.[55] It is to be noted that all of these methods have only been validated on data sets containing equilibrium structures. The ANAKIN-ME approach reports a root mean squared error of 1.3 kcal/mol when trained on the entire ANI-1 data

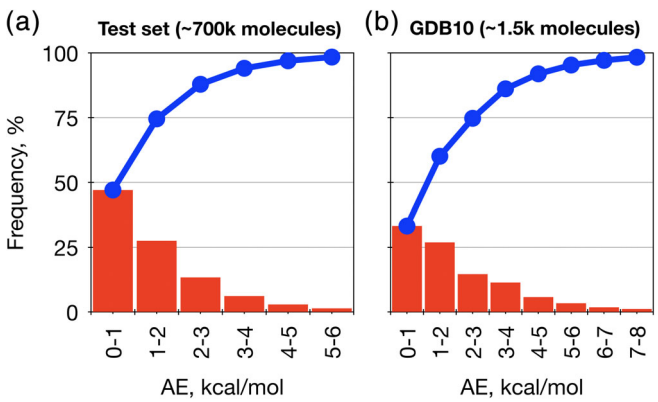

Figure 3. The histograms and the cumulative distributions of the absolute errors (in kcal/mol) calculated on a) test set and b) GDB-10 test set. [Color figure can be viewed at wileyonlinelibrary.com]

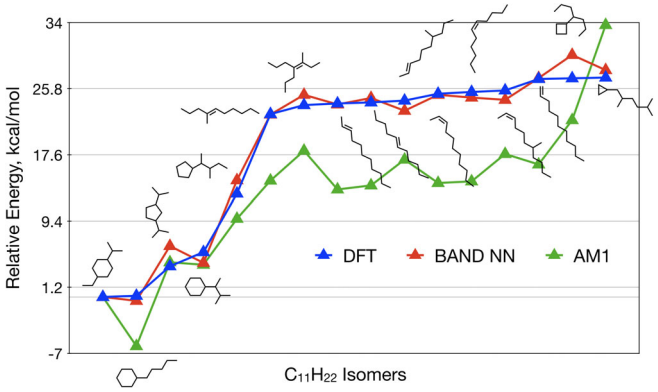

Figure 4. The relative energies (in kcal/mol) of select isomers of C_11H_22 relative to the least energy isomer calculated using the ωB97X/6-31G (d) level of theory, AM1 semiempirical method and using BAND NN. [Color figure can be viewed at wileyonlinelibrary.com]

set.[42] On the molecules from GDB-10 benchmark data set prepared by Smith et al., ANAKIN-ME reports mean absolute error of 0.83 kcal/mol for molecules with relative energies under 30 kcal/mol from their respective ground-state conformer.[36]

Structural and geometric isomers

The accuracy of the proposed model in satisfactorily predicting the relative energies of structural and geometric isomers is examined here. Several isomers of $C_{11}H_{22}$ spanning diverse structural and geometric space, namely, linear chains, *cis–trans* isomers, varying ring sizes (three to six), and so on, were chosen. The energies of the optimized geometries of these isomers were calculated using the ωB97X/6-31G(d) level of theory using the Gaussian 09 program.[56] Despite the diverse set of molecules considered for this evaluation, quantitative agreement between the DFT and BAND NN methods is observed (Fig. 4). It is also found that the NN model significantly outperforms the semiempirical QM AM1 method.[57] This further indicates that machine learning-based methods developed with molecular size invariant featurizations are capable of accurate modeling of molecular systems at the fraction of the computational expense that DFT or ab initio calculations would require.

Potential energy surfaces

From the above discussions, it is apparent that the BAND NN model is capable of prediction atomization energies of small organic molecules very well. However, it is also important that models such as the one proposed in this study are able to represent the potential energy surface of molecular systems and not just the energies for select points on the potential energy surface. Such a proper behavior of the model is necessary for it to be useful for performing energy minimizations, conformational analysis, and force calculations in molecular dynamics simulations. Potential energy scans with respect to bonds and angles were performed on molecules that are significantly larger than those in the training set. Figure 5 gives the potential energy surfaces corresponding to C—C and C—N bond lengths calculated using the ωB97X/6-31G(d) level and BAND NN. For both the bonds, the positions of the minima are predicted accurately, and the curves maintain a smooth curvature. Similarly, the potential energy scan for a C—C—C angle indicates very good agreement between the DFT results and the BAND NN data. To further show the chemical accuracy of the model, we performed conformational analysis for the central C—C bond of decane molecule and found very good agreement. The positions of the minima and maxima are predicted reasonably well along with the energies of different conformers and transition state with a mean absolute error of only 0.6 kcal/mol (Fig. 6).

Reaction energies

In this section, the ability of the BAND NN model to predict reaction energies of simple organic reactions is examined. Some of the most simple and common reactions in organic chemistry (conformational differences stabilized by intramolecular hydrogen bonds, hydrogenation, Diels–Alder reaction, aldol condensation,

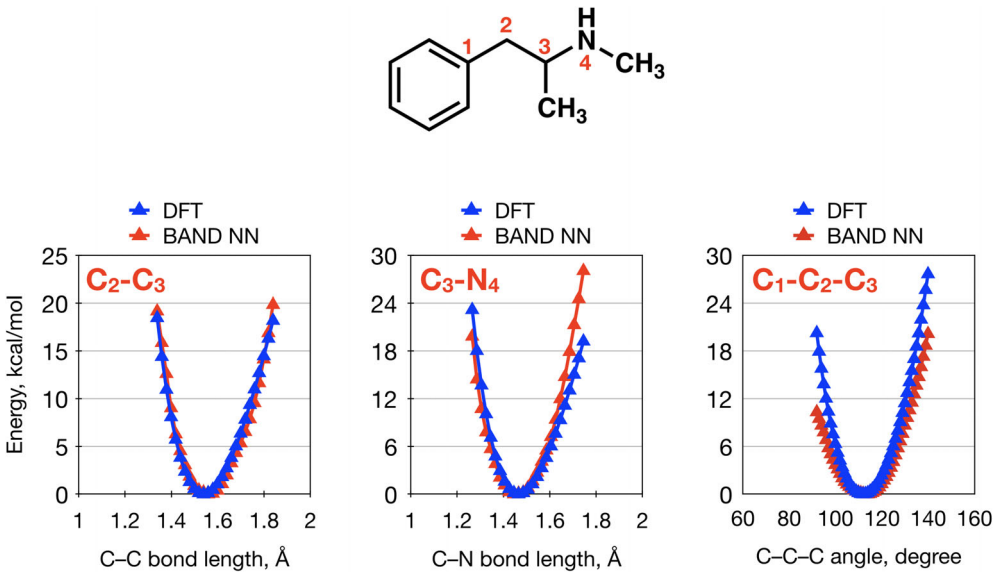

Figure 5. Potential energy surface (in kcal/mol) corresponding to C—C and C—N bond stretching and C—C—C angle bending of methamphetamine calculated using the ωB97X/6-31G(d) level of theory and BAND NN. The structure of the molecule along with the labels of atoms that were used for calculating the potential energies are given above the plots. [Color figure can be viewed at wileyonlinelibrary.com]

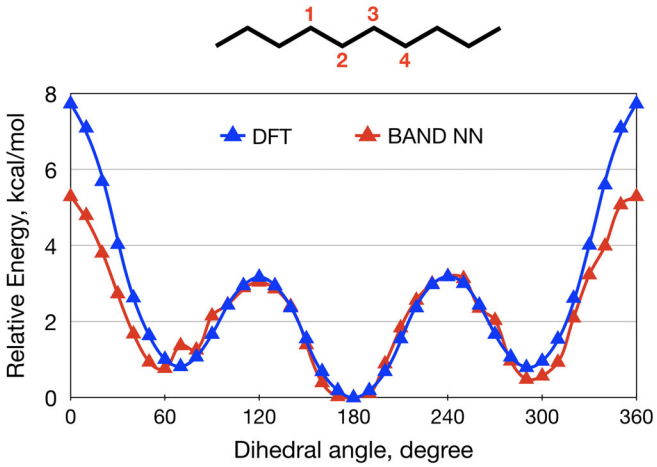

Figure 6. Potential energy surface (in kcal/mol) corresponding to the rotation about the central C—C single bond of *n*-decane calculated at the ωB97X/6-31G(d) level of theory and using BAND NN. [Color figure can be viewed at wileyonlinelibrary.com]

esterification, and electrocyclic ring closing reaction) were chosen for this analysis. The reaction energies calculated for these using the ωB97X/6-31G(d) level, AM1 method, and BAND NN model along with the schematic diagrams of the reactions are given in Figures 7. All the reaction energies obtained using the BAND NN model are comparable to the DFT results. Among the six reactions, largest difference between the DFT and the BAND NN model was observed for the hydrogenation reaction. Notably, no data pertaining to the H_2 system were present in the training data set. Similar to the prediction of relative energies of $C_{11}H_{22}$, the reaction energies computed using the BAND NN model outperfom the AM1 level of theory.

Importance of 3,4-body terms

Most of the machine learning models for QM/DFT energy predictions have been done by including only 2-body terms.[13,37] In this study, the energy is given as the sum of the energy contributions from all the bonds, angles, dihedral angles, and nonbonded pairs. Two other models, one excluding the dihedrals (referred to as BAN NN model) and another excluding the angles and dihedrals (referred to as BN NN model) were trained using the same procedure as the BAND NN model to investigate the importance of including the 3,4-body terms. The distributions of the absolute error obtained from these models are given in Figure 8. The atomization energies are predicted within 2 kcal/mol for only about 50 and 60% of the molecules in the data set in the BN NN and BAN NN models, respectively. The mean absolute errors are 2.7 and 2.4 kcal/mol (1.45 kcal/mol for the BAND NN model). The performances of these models are inferior compared to the BAND NN model. Previous studies that utilized "bag of bonds" feature involved the prediction of energies of molecules that are in their minimum energy states.[13] In other words, all the angles and dihedrals in these molecules are in their equilibrium values, and hence the variances of the angles and dihedrals in the data set are not large. In this study, we consider high energy configurations for each of the minimum energy structures for which the angles and dihedral angles are away from the minimum on the potential energy surface and hence sample a larger configurational/conformational space. This requires that the energy of the molecules is expressed as a function of angles and dihedral angles as well.

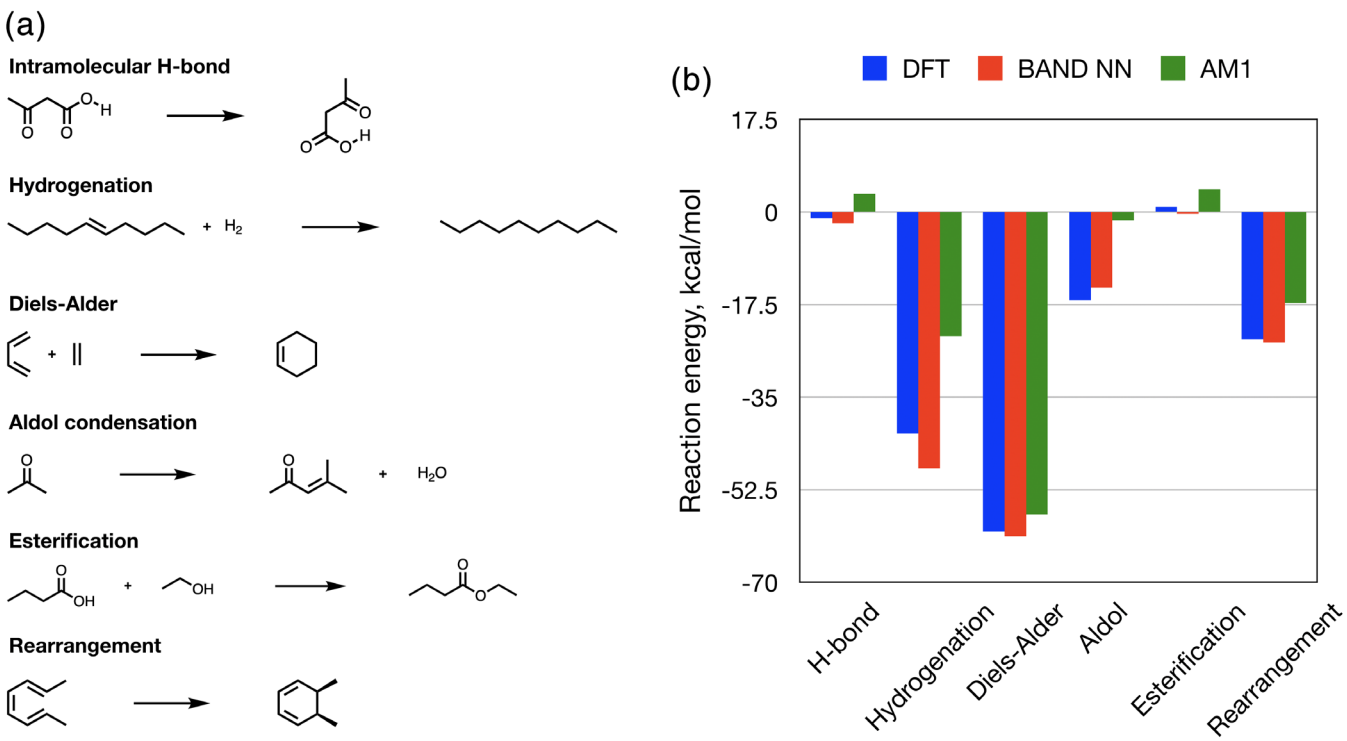

Figure 7. a) Select organic reactions chosen for the calculation of reaction energies. b) Reaction energies (kcal/mol) calculated using the ωB97X/6-31G(d) and AM1 levels of theory, and those predicted using BAND NN. [Color figure can be viewed at wileyonlinelibrary.com]

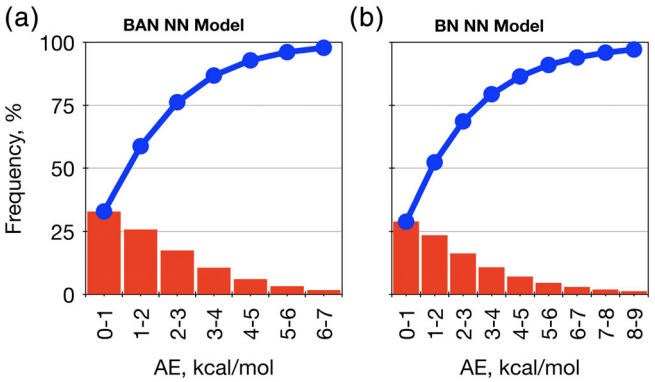

Figure 8. The histograms and the cumulative distributions of the absolute errors (in kcal/mol) calculated using the a) BAN NN and b) BN NN models. [Color figure can be viewed at wileyonlinelibrary.com]

Hence, the BAND molecular representation proposed in this manuscript is well suited for handling nonequilibrium structures compared to those that include only 2-body terms.

Geometry optimization

Though there have been quite a few ML models to predict atomization energies of small organic molecules have been published in the last 2 years, there are few shortfalls. Some of these models cannot be applied to molecules larger than the ones in the training set, most of them cannot be applied to structures that are not in their minima on the potential energy surface and they have not been used for geometry optimizations. The condition that the geometry optimized using the DFT level has to be provided for the ML model to predict the energy is not desirable, because the geometry optimization involves calculation of the DFT energy. The next useful step in applying machine learning for molecular systems is to be able to develop models that allow for geometry optimization such that one could start from a structure away from the minimum and use the model along with an optimization method to reach the minimum. BAND NN model has been trained on high energy structures with explicit topology of the molecule as defined by the featurization used here. Nelder–Mead's optimization method has been used for updating the geometric parameters starting from a nonequilibrium structure. Starting from a reasonable guess structure of ocatane and 2-methylprop-2-enol, geometry optimization was performed. Figure 9 gives the energy of these molecules with respect to the optimization step number. The energies of the two molecules gradually decrease with respect to the optimization step and reaches convergence. For another test, few structures were generated using the GaussView program[47] (as an acceptable way of generating initial geometries in electronic structure theory calculations), and optimizations were performed using the Nelder–Mead's optimization employing the BAND NN. The single-point energies of the initial and optimized geometries obtained using the ωB97X/6-31G(d) level are given in Table 2. In all the cases, the optimizer converged the molecules to structures whose energies are significantly lower than those of the initial structure. Though the results are not perfect for all the systems, it is clear

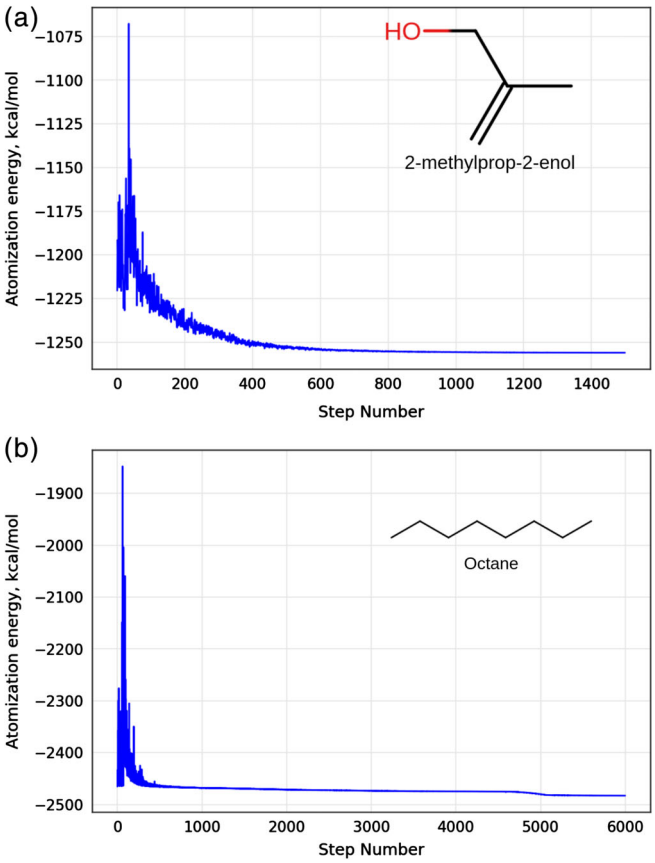

Figure 9. BAND NN atomization energies (kcal/mol) of 2-methylprop-2-enol and octane with respect to the optimization step number. [Color figure can be viewed at wileyonlinelibrary.com]

Table 2. Input structure: difference (kcal/mol) between the single point energies on the initial structure and the DFT-optimized structure obtained at the ωB97X/6-31G(d) level; BAND optimized: difference (kcal/mol) between the single-point energies on the BAND NN-optimized structure and the DFT-optimized structure obtained at the ωB97X/6-31G (d) level.

Molecule name	Input structure	BAND optimized
1	5.5	1.7
2	10.7	3.1
3	4.9	1.5
4	9.1	3.4
5	17.5	7.6

The structures of the molecules are given in Figure 10.

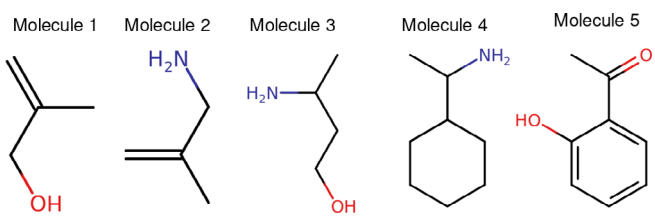

Figure 10. Molecules that were optimized starting from initial geometries generated using the GaussView program. Energies are presented in Table 2. [Color figure can be viewed at wileyonlinelibrary.com]

that it is possible to use an appropriate molecular representation that will allow for geometry optimizations and that optimal structures can be obtained from this method. Implementation of gradient-based methods may further improve the efficiency of the geometry optimization process.

Conclusions

A chemically intuitive molecular descriptor inspired from classical force field equation has been developed for the prediction of atomization energy of small organic molecules. BAND NN model was trained on a subset of ANI-1 data set by choosing molecules that were at most 30 kcal/mol higher than the corresponding minimum. It was shown to accurately predict atomization energies with a mean absolute error of 1.45 kcal/mol on the test set. It accurately predicted the atomization energies of molecules randomly sampled from GDB-10, which are larger than the molecules in the data set. The model was demonstrated to be sensitive to structural and geometric isomers, generate accurate potential energy surfaces, and predict reaction energies to DFT-level accuracy on larger molecules. These experiments demonstrate that the model is transferable to larger molecules. In recent years, several methods have been proposed to predict atomization energy for ground-state molecules, but for a model to be practically useful it should also be able to predict potential energy surfaces accurately. BAND NN model proposed in this work not only predicts the atomization energy for equilibrium and off-equilibrium structures but also can be used to perform geometry optimization. Further work in this area to develop robust transferable models using deep learning methods aimed at predicting accurate potential energy surfaces of molecular systems is expected to be more fruitful for state of the art problems in computational chemistry.

Acknowledgments

We thank the DST-SERB (grant no. EMR/2016/007697) for the financial support.

Keywords: machine learning · atomization energy · geometry optimization · conformational analysis · neural network

How to cite this article: S. Laghuvarapu, Y. Pathak, U. D. Priyakumar. *J. Comput. Chem*. **2019**, 9999, 1–10. DOI: 10.1002/jcc.26128

[1] A. D. MacKerell, Jr.., *J. Comput. Chem.* **2004**, *25*, 1584.
[2] S. A. Hollingsworth, R. O. Dror, *Neuron* **2018**, *99*, 1129.
[3] K. Vanommeslaeghe, O. Guvench, A. D. J. MacKerell, *Curr. Pharm. Des.* **2014**, *20*, 3281.
[4] R. B. Best, X. Zhu, J. Shim, P. E. Lopes, J. Mittal, M. Feig, A. D. MacKerell, Jr.., *J. Chem. Theory Comput.* **2012**, *8*, 3257.
[5] J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman, D. A. Case, *J. Comput. Chem.* **2004**, *25*, 1157.
[6] J. A. Lemkul, J. Huang, B. Roux, A. D. MacKerell, *Chem. Rev.* **2016**, *116*, 4983.
[7] K. T. Butler, D. W. Davies, H. Cartwright, O. Isayev, A. Walsh, *Nature* **2018**, *559*, 547.
[8] G. B. Goh, N. O. Hodas, A. Vishnu, *J. Comput. Chem.* **2017**, *38*, 1291.
[9] R. Ramakrishnan, O. A. von Lilienfeld, *Rev. Comput. Chem.* **2017**, *30*, 225.
[10] A. C. Mater, M. L. Coote, *J. Chem. Inf. Model.* **2019**, *59*, 2545.
[11] B. Schölkopf, A. J. Smola, Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond, Cambridge: MIT Press, **2002**.
[12] M. Rupp, A. Tkatchenko, K.-R. Müller, O. A. Von Lilienfeld, *Phys. Rev. Lett.* **2012**, *108*, 058301.
[13] K. Hansen, F. Biegler, R. Ramakrishnan, W. Pronobis, O. A. Von Lilienfeld, K.-R. Müller, A. Tkatchenko, *J. Phys. Chem. Lett.* **2015**, *6*, 2326.
[14] B. Huang, O. A. Von Lilienfeld, *J. Chem. Phys.* **2016**, *145*, 161102.
[15] A. Krizhevsky, I. Sutskever, G. E. Hinton, Advances in Neural Information Processing Systems, **2012**, NY, USA: Curran Associates Inc., p. 1097.
[16] Bahdanau, D.; Cho, K. and Bengio, Y., arXiv preprint arXiv:1409.0473, **2014**. https://arxiv.org/abs/1409.0473
[17] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Advances in Neural Information Processing Systems, NY, USA: Curran Associates Inc., **2014**, p. 2672.
[18] Z. Wu, B. Ramsundar, E. N. Feinberg, J. Gomes, C. Geniesse, A. S. Pappu, K. Leswing, V. Pande, *Chem. Sci.* **2018**, *9*, 513.
[19] J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, G. E. Dahl, Proceedings of the 34th International Conference on Machine Learning, Vol. *70*, **2017**, p. 1263. http://proceedings.mlr.press/v70/gilmer17a.html
[20] K. T. Schütt, H. E. Sauceda, P.-J. Kindermans, A. Tkatchenko, K.-R. Müller, *J. Chem. Phys.* **2018**, *148*, 241722.
[21] K. T. Schütt, F. Arbabzadah, S. Chmiela, K. R. Müller, A. Tkatchenko, *Nat. Commun.* **2017**, *8*, 13890.
[22] E. Swann, B. Sun, D. Cleland, A. Barnard, *Mol. Simulat.* **2018**, *44*, 905.
[23] J. Lyons, A. Dehzangi, R. Heffernan, A. Sharma, K. Paliwal, A. Sattar, Y. Zhou, Y. Yang, *J. Comput. Chem.* **2014**, *35*, 2040.
[24] Q. Jiang, X. Jin, S.-J. Lee, S. Yao, *J. Mol. Graph. Model.* **2017**, *76*, 379.
[25] J. Hanson, K. Paliwal, T. Litfin, Y. Yang, Y. Zhou, *Bioinformatics* **2018**, *35*, 2403.
[26] S. Romero-Molina, Y. B. Ruiz-Blanco, M. Harms, J. Münch, E. Sanchez-Garcia, *J. Comput. Chem.* **2019**, *40*, 1233.
[27] P. Raccuglia, K. C. Elbert, P. D. Adler, C. Falk, M. B. Wenny, A. Mollo, M. Zeller, S. A. Friedler, J. Schrier, A. J. Norquist, *Nature* **2016**, *533*, 73.
[28] L. Ward, A. Agrawal, A. Choudhary, C. Wolverton, *npj Comput. Mater.* **2016**, *2*, 16028.
[29] M. H. Segler, M. Preuss, M. P. Waller, *Nature* **2018**, *555*, 604.
[30] P. O. Dral, *J. Comput. Chem.* **2019**, *40*, 2339.
[31] Guimaraes, G. L.; Sanchez-Lengeling, B.; Outeiral, C.; Farias, P. L. C. and Aspuru-Guzik, A., arXiv preprint arXiv:1705.10843, **2017**. https://arxiv.org/abs/1705.10843
[32] M. H. Segler, T. Kogej, C. Tyrchan, M. P. Waller, *ACS Cent. Sci.* **2017**, *4*, 120.
[33] Y. Pathak, S. Laghuvarapu, S. Mehta, U. D. Priyakumar, *ChemRxiv* **2019**. https://doi.org/10.26434/chemrxiv.10282346.v2
[34] F. A. Faber, L. Hutchison, B. Huang, J. Gilmer, S. S. Schoenholz, G. E. Dahl, O. Vinyals, S. Kearnes, P. F. Riley, O. A. von Lilienfeld, *J. Chem. Theory Comput.* **2017**, *13*, 5255.
[35] J. Behler, *Int. J. Quantum Chem.* **2015**, *115*, 1032.
[36] J. S. Smith, O. Isayev, A. E. Roitberg, *Chem. Sci.* **2017**, *8*, 3192.
[37] K. Yao, J. E. Herr, S. N. Brown, J. Parkhill, *J. Phys. Chem. Lett.* **2017**, *8*, 2689.
[38] J. Behler, M. Parrinello, *Phys. Rev. Lett.* **2007**, *98*, 146401.
[39] A. P. Bartók, R. Kondor, G. Csányi, *Phys. Rev. B* **2013**, *87*, 184115.
[40] Y. LeCun, Y. Bengio, G. Hinton, *Nature* **2015**, *521*, 436.
[41] T. A. Halgren, *J. Comput. Chem.* **1996**, *17*, 490.
[42] J. S. Smith, O. Isayev, A. E. Roitberg, *Sci. Data* **2017**, *4*, 170193.
[43] T. Fink, H. Bruggesser, J.-L. Reymond, *Angew. Chem. Int. Ed.* **2005**, *44*, 1504.
[44] T. Fink, J.-L. Reymond, *J. Chem. Inf. Model.* **2007**, *47*, 342.
[45] J.-D. Chai, M. Head-Gordon, *J. Chem. Phys.* **2008**, *128*, 084106.
[46] R. Ramakrishnan, P. O. Dral, M. Rupp, O. A. Von Lilienfeld, *Sci. Data* **2014**, *1*, 140022.
[47] Dennington, R.; Keith, T.; Millam, J., Gaussview, Version 5, **2009**.
[48] Landrum, G., Rdkit: Open-source Cheminformatics.
[49] Chollet, F., Keras, **2015**.
[50] Abadi, M.; Agarwal, A.; Barham, P.; Brevdo, E.; Chen, Z.; Citro, C.; Corrado, G. S.; Davis, A.; Dean, J.; Devin, M.; Ghemawat, S.; Goodfellow, I.; Harp, A.; Irving, G.; Isard, M.; Jia, Y.; Jozefowicz, R.; Kaiser, L.; Kudlur, M.; Levenberg, J.; Mané, D.; Monga, R.; Moore, S.;

Murray, D.; Olah, C.; Schuster, M.; Shlens, J.; Steiner, B.; Sutskever, I.; Talwar, K.; Tucker, P.; Vanhoucke, V.; Vasudevan, V.; Viégas, F.; Vinyals, O.; Warden, P.; Wattenberg, M.; Wicke, M.; Yu, Y. and Zheng, X., TensorFlow: Large-scale Machine Learning on Heterogeneous Systems, **2015**.

[51] Kingma, D. P. and Ba, J., arXiv preprint arXiv:1412.6980, **2014**. https://arxiv.org/abs/1412.6980

[52] V. Nair, G. E. Hinton, Proceedings of the 27th International Conference on Machine Learning (ICML-10), **2010**, p. 807. Omnipress, 2010.

[53] J. A. Nelder, R. Mead, *Comput. J.* **1965**, *7*, 308.

[54] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, *J. Mach. Learn. Res* **2011**, *12*, 2825.

[55] O. T. Unke, M. Meuwly, *J. Chem. Theory Comput.* **2019**, *15*, 3678.

[56] Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenberg, J. L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery, Jr., J. A.; Peralta, J. E.; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers, E.; Kudin, K. N.; Staroverov, V. N.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, J. M.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Zakrzewski, V. G.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Dapprich, S.; Daniels, A. D.; Farkas, Ã.; Foresman, J. B.; Ortiz, J. V.; Cioslowski, J. and Fox, D. J., Gaussian 09 Revision E.01, **2009**.

[57] M. J. Dewar, E. G. Zoebisch, E. F. Healy, J. J. Stewart, *J. Am. Chem. Soc.* **1985**, *107*, 3902.

Received: 9 September 2019
Revised: 13 November 2019
Accepted: 21 November 2019