JCTC Journal of Chemical Theory and Computation

Article

Cite This: J. Chem. Theory Comput. 2018, 14, 2991−3003

pubs.acs.org/JCTC

Many-Body Descriptors for Predicting Molecular Properties with Machine Learning: Analysis of Pairwise and Three-Body Interactions in Molecules

Wiktor Pronobis,[†] Alexandre Tkatchenko,[*,‡] and Klaus-Robert Müller[*,†,¶,§]

[†]Machine Learning Group, Technische Universität Berlin, 10587 Berlin, Germany

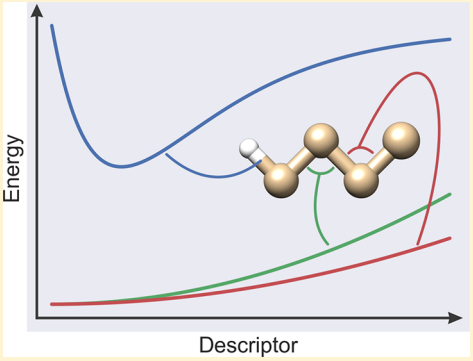
[‡]Physics and Materials Science Research Unit, University of Luxembourg, Luxembourg L-1511, Luxembourg

[¶]Max Planck Institute for Informatics, 66123 Saarbrücken, Germany

[§]Department of Brain and Cognitive Engineering, Korea University, Seoul 136-713, South Korea

Supporting Information

ABSTRACT: Machine learning (ML) based prediction of molecular properties across chemical compound space is an important and alternative approach to efficiently estimate the solutions of highly complex many-electron problems in chemistry and physics. Statistical methods represent molecules as descriptors that should encode molecular symmetries and interactions between atoms. Many such descriptors have been proposed; all of them have advantages and limitations. Here, we propose a set of general two-body and three-body interaction descriptors which are invariant to translation, rotation, and atomic indexing. By adapting the successfully used kernel ridge regression methods of machine learning, we evaluate our descriptors on predicting several properties of small organic molecules calculated using density-functional theory. We use two data sets. The GDB-7 set contains 6868 molecules with up to 7 heavy atoms of type CNO. The GDB-9 set is composed of 131722 molecules with up to 9 heavy atoms containing CNO. When trained on 5000 random molecules, our best model achieves an accuracy of 0.8 kcal/mol (on the remaining 1868 molecules of GDB-7) and 1.5 kcal/mol (on the remaining 126722 molecules of GDB-9) respectively. Applying a linear regression model on our novel many-body descriptors performs almost equal to a nonlinear kernelized model. Linear models are readily interpretable: a feature importance ranking measure helps to obtain qualitative and quantitative insights on the importance of two- and three-body molecular interactions for predicting molecular properties computed with quantum-mechanical methods.



INTRODUCTION

Recently, machine learning has been ubiquitously used in the industry and sciences. The possibility of parallel implementations using GPU cards in addition to new deep learning architectures has enabled powerful learning machines which reach and even surpass human performance in a variety of applications. From imperfect information games like heads-up no-limit Texas hold'em poker[1] over real-time strategy games like StarCraft,[2] the program AlphaGo Zero[3] has been trained without human knowledge and is arguably the strongest Go player in history. ML approaches reach human performance in human interaction tasks like speech recognition,[4] image recognition,[5] and speech generation.[6]

In this work, we follow one of the very intriguing applications of ML in sciences: the prediction of highly complex properties of quantum mechanical systems. Specifically, we are interested in the prediction of the properties of intermediate size molecules and the analysis of the pairwise and three-body interactions. Before proceeding, we put our work in the context of existing literature on machine learning of molecular properties.

Recently, machine learning has been successfully used to predict the atomization energies of small molecules[7−13] and molecular dynamics simulations[14−18] as well as for studying properties of quantum-mechanical densities.[19,20] Descriptors of molecules are constructed to provide an invariant, unique, and efficient representation as input to ML models,[21−33] e.g. for the atomization energy, a popular molecular descriptor is the bag-of-bonds (BOB) model,[34] which is an extension of the Coulomb matrix (CM) approach[7] and groups the pairwise distances according to pairs of atom types.

Shapeev et al.[35,36] introduce systematically improvable interatomic potential descriptors based on invariant polynomials. These moment tensor potentials are invariant with respect to permutation, rotation, and reflection and have the advantage that the computational complexity of computing these polynomials scales like $O(n)$, where n is the number of atoms. One possible limitation is that these potentials treat all atoms as chemically equivalent. Shapeev et al. suggest a future extension to alleviate this issue, namely to let the radial basis functions depend on the types of atoms.

Faber et al.[10] studied a representation using the histogram of distances, angles, and dihedral angles with kernel ridge regression

Received: February 2, 2018
Published: May 11, 2018

DOI: 10.1021/acs.jctc.8b00110
J. Chem. Theory Comput. 2018, 14, 2991−3003

which achieves a mean absolute error of 0.58 kcal/mol on the GDB-9 set, when trained on 118000 molecules. An angle representation based on molecular atomic radial angular distributions (MARAD) achieves a MAE of 1.2 kcal/mol with kernel ridge regression and 4.0 kcal/mol with the linear Bayesian ridge regression model when trained on 118000 molecules.

The recently introduced BAML (bonds angles machine learning) representation[21] can be viewed as a many-body extension of BOB and constructs arbitrary distance functions between pairwise distances. BAML reaches a MAE of 1.15 kcal/mol on the GDB-7 set trained on 5000 molecules[24] and a MAE of 1.2 kcal/mol on the GDB-9 set when trained on 118000 molecules.[10]

Huo et al.[24] introduced a many-body tensor representation which improves on the histogram descriptors of Faber et al. by "smearing" the histograms of given many-body features. For one of their best models, a MAE of 0.60 kcal/mol on GDB-7 using Gaussian kernel ridge regression and a MAE of 0.74 kcal/mol using a linear model (with many-body interactions) have been reported.

Recently, even more accurate models for predicting the atomization energy have been introduced,[37,38] which reach an accuracy of 0.26 kcal/mol[37] on 100000 training samples and 0.45 kcal/mol[38] on 110000 training samples, respectively.

Most of the above approaches use explicit three-body (e.g., angle) or four-body (e.g., dihedral angle) features to construct the respective representation. In this work, we propose novel translational, rotational, and atom indexing invariant molecular descriptors which build on the success of inverse pairwise distances for predicting the atomization energy.[7−9,11,23,34] In particular, we construct many-body interaction features of *arbitrary order* from inverse pairwise distances which helps to alleviate sorting challenges encountered in e.g. CM. Accordingly, our model learns e.g. a three-body interatomic potential, which is not necessarily a function of angle. Our novel descriptors allow for construction of an invariant two-body and many-body interaction representation at a *fixed* descriptor size. Note that fixed sized molecular descriptors are useful in practice as they can be easily used in combination with kernel ridge regression or deep neural networks or other models that expect fixed size input data. Also, such fixed size representations are generally extensible to large molecules and solids, while incorporating informative higher-order interaction terms. While missing long-range interactions (H-bond, van der Waals, etc.), those can be easily built on top of our proposed short-range models.[39,40] Clearly, any such combination of short-range and long-range models for interatomic potentials will have to carefully avoid double-counting effects. Furthermore, when using these novel descriptors we observe that linear models perform only slightly worse than the nonlinear methods. The latter is helpful in practice as linear models allow to simply and easily analyze the importance of the proposed two-, three-, or many-body interaction features for predicting atomization energies of the molecules. This allows for extracting insights from the learned model.

We view our new descriptors as an optimal compromise that allows high-throughput calculations of extensive molecular properties for equilibrium geometries throughout chemical space. Our many-body model is complementary to recently developed deep neural networks and nonlinear kernel methods for estimating molecular properties.[10,11,24,34]

The paper is structured as follows. The next section defines the invariant two-body and three-body molecular descriptors. The following section details the data sets as well as the learning task and the prediction of several properties of small molecules. This is followed by the analysis of the importance of the two- and three-body molecular features and the conclusion.

INVARIANT MANY-BODY INTERACTION DESCRIPTORS

We represent a molecule or material by the respective finite set from which the molecule or unit cell is constructed.

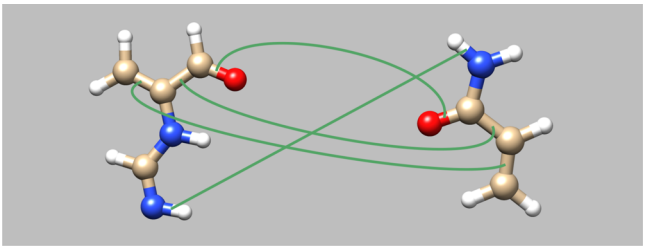


Figure 1. Illustration of the bag-of-bonds molecule similarity. The distance between two atoms of the left molecule gets directly compared to an arbitrary distance of the right molecule corresponding to the same atom types composing the pairwise interaction.

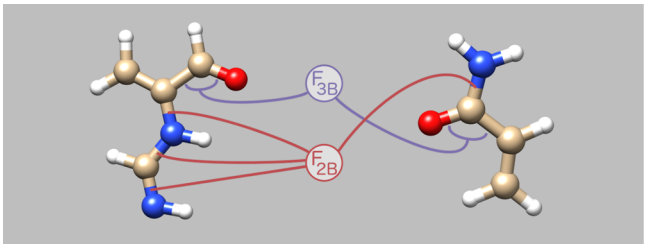


Figure 2. Illustration of the F_{2B} and F_{3B} molecule similarity. For F_{2B}, the pairwise distances of the left molecule corresponding to a fixed pair of atom types are computed into a feature entry, which gets compared to the same feature entry of the right molecule composed identically for the same pair of atom types. Similarly, F_{3B} compares three bonded atoms which have an angle.

Table 1. Prediction Errors of the PBE0 Atomization Energy of the Molecules of the Set GDB-7 by Various ML Models with Random 5K Train Molecules and the Remaining 1868 Molecules as Test Set[a]

method	features	MAE	RMSE	MAX. DEV
mean		174	219	1166
RR	CM	25	33	134
RR	BOB	23	30	144
RR	F_{2B}	4.9	12	350
RR	F_{2B} + F_{3B}	1.0	8.3	327
KNN	CM	80	104	461
KNN	BOB	70	102	424
KNN	F_{2B}	49	73	230
KNN	F_{2B} + F_{3B}	10	28	306
KRR (Gauss)	CM	8.6	15	433
KRR (Laplace)	CM	3.7	5.8	89
KRR (Gauss)	BOB	7.6	10	99
KRR (Laplace)	BOB	1.8	3.9	103
KRR (Gauss)	F_{2B}	1.9	4.7	155
KRR (Laplace)	F_{2B}	4.2	6.1	62
KRR (Gauss)	F_{2B} + F_{3B}	**0.8**	**1.5**	**28**
KRR (Laplace)	F_{2B} + F_{3B}	2.4	3.8	51

[a]The errors are given in kcal/mol. The models used are ridge regression (RR), kernel ridge regression (KRR), and k-nearest neighbors (KNN).

DOI: 10.1021/acs.jctc.8b00110
J. Chem. Theory Comput. 2018, 14, 2991−3003

Table 2. Mean Absolute Errors of Predicting Several Ground- and Excited-State Properties by Kernel Ridge Regression Trained on 5000 Random Molecules and Tested on the Remaining 1868 Molecules of the GDB-7 Data Set[a]

property	CM	BOB	F_{2B}	$F_{2B} + F_{3B}$	unit	description
ae-pbe0	3.7	1.8	1.9	**0.8**	kcal/mol	atomization energy (DFT/PBE0)
homogw	0.212	0.138	0.167	**0.128**	eV	highest occuppied molecular orbital (GW)
lumo-gw	0.187	**0.142**	0.155	0.147	eV	lowest unoccupied molecular orbital (GW)
homo-pbe0	0.202	0.130	0.156	**0.120**	eV	highest occupied molecular orbital (DFT/PBE0)
lumo-pbe0	0.174	**0.108**	0.133	**0.108**	eV	lowest unoccupied molecular orbital (DFT/PBE0)
homozindo	0.279	0.144	0.173	**0.132**	eV	highest occupied molecular orbital (ZINDO/s)
lumo-zindo	0.252	0.134	0.168	**0.112**	eV	lowest unoccupied molecular orbital (ZINDO/s)
p-pbe0	0.130	**0.083**	0.103	0.088	Ångström^3	polarizability (DFT/PBE0)
p-scs	0.065	0.042	0.061	**0.032**	Ångström^3	polarizability (self-consistent screening)
e1-zindo	0.37	0.19	0.21	**0.15**	eV	first excitation energy (ZINDO)
ea-zindo	0.29	0.15	0.18	**0.13**	eV	electron affinity (ZINDO/s)
imax-zindo	0.084	**0.067**	0.074	0.071	au	excitation energy at maximal absorption (ZINDO)
emax-zindo	1.47	**1.20**	1.29	1.26	eV	maximal absorption intensity (ZINDO)
ip-zindo	0.32	**0.18**	0.21	**0.18**	eV	ionization potential (ZINDO/s)

[a]The best performing models are marked in bold.

Table 3. Mean Absolute Errors of Predicting Several Properties Calculated at the B3LYP/6-31G(2df,p) Level of Quantum Chemistry and Predicted by Kernel Ridge Regression Trained on 5000 Random Molecules and Tested on the Remaining 126722 Molecules of the GDB-9 Data Set[a]

property	CM	BOB	F_{2B}	$F_{2B} + F_{3B}$	unit	description
U0	7.9	4.0	4.8	**1.5**	kcal/mol	internal energy at 0 K
U	7.9	4.0	4.8	**1.5**	kcal/mol	internal energy at 298.15 K
H	7.9	4.0	4.8	**1.5**	kcal/mol	enthalpy at 298.15 K
G	7.9	4.0	4.8	**1.5**	kcal/mol	free energy at 298.15 K
HOMO	5.8	4.3	4.7	**3.6**	kcal/mol	energy of highest occupied molecular orbital
LUMO	8.9	5.7	6.0	**5.1**	kcal/mol	energy of lowest occupied molecular orbital
gap	11	6.8	7.9	**6.2**	kcal/mol	gap, difference between LUMO and HOMO
alpha	1.00	0.63	0.72	**0.49**	Bohr^3	isotropic polarizability
mu	0.77	0.65	0.67	**0.61**	Debye	dipole moment
r2	16	8.5	**7.3**	9.0	Bohr^2	electronic spatial extent
zpve	0.33	0.20	0.18	**0.10**	kcal/mol	zero point vibrational energy
A	0.42	**0.37**	0.40	0.42	GHz	rotational constant A
B	0.12	**0.10**	0.12	0.13	GHz	rotational constant B
C	0.052	**0.045**	0.046	0.050	GHz	rotational constant C
cv	0.38	0.20	0.21	**0.12**	cal/(mol K)	heat capacity at 298.15 K

[a]The best performing descriptors are marked in bold.

Let $\mathbf{r}_i$ denote the position of the atom i with atomic number Z_i in three-dimensional coordinate space. Then, a physical system with N atoms is defined by $S = \{Z_i, \mathbf{r}_i\}_{i=1}^{N}$. From this physical system S, we propose the many-body interaction descriptors

$$f_{\bar{\mathbf{Z}},p}(S) = \sum_{(j_1,\cdots,j_k)\in G(k,N)} \delta_{\bar{\mathbf{Z}}}(\mathbf{Z})\cdot p(\mathbf{r}_{j_1}, Z_{j_1}, \cdots, \mathbf{r}_{j_k}, Z_{j_k}) \quad (1)$$

where $\mathbf{Z} := (Z_{j_1},\cdots,Z_{j_k})$, $\bar{\mathbf{Z}}$ is a given k-tuple of k atomic numbers with $k \leq N$, p is a k-body interaction term, and the partial permutations set $G(k, N)$ consists of the sequences without repetition of k elements from the set $\{1,2,\cdots,N\}$ and the Kronecker delta $\delta_{\bar{\mathbf{Z}}}(\mathbf{Z})$, which equals 1 if and only if the two k-tuples $\bar{\mathbf{Z}}$ and $\mathbf{Z}$ are equal and zero otherwise. The number of elements of the k-permutations of N set $G(k, N)$ is $\frac{N!}{(N-k)!}$.

The descriptors in eq 1 are intrinsically invariant to the indexing of the atoms comprising the system S, as we sum over all elements of the k-permutations of N set $G(k, N)$. If the k-body interactions term p satisfies invariance to the translation and rotation of the atoms of S, this carries over to the descriptors $f_{\bar{\mathbf{Z}},p}(S)$. In the following, we propose a set of translational and rotational invariant two-body and three-body interaction terms p, which will define our invariant many-body interaction descriptors.

Invariant Two-Body Interaction Descriptors F_{2B}. We define the set of translational and rotational invariant two-body interaction terms

$$p_m^{2B}(\mathbf{r}_1, Z_1, \mathbf{r}_2, Z_2) := \|\mathbf{r}_1 - \mathbf{r}_2\|^{-m} \quad (2)$$

where $m \in \mathbb{N}^+$. For a given set of n different atomic numbers $A_n := \{Z_i\}_{i=1}^{n}$ with $Z_i \neq Z_j \,\forall\, i,j \in \{1,\cdots,n\}$, let S_{2B} denote the set of all tuples (Z_i, Z_j) with $Z_i \leq Z_j$ and $Z_i, Z_j \in A_n$. Let M_{2B} denote the set $M_{2B} := \{1, 2, \cdots, 15\}$. For a given physical system S, the two-body interaction descriptors F_{2B} are now given by

$$F_{2B} := \{f_{\bar{\mathbf{Z}},p_m^{2B}}(S)\}_{m\in M_{2B},\bar{\mathbf{Z}}\in S_{2B}} \quad (3)$$

Typically, the set S_n contains the atomic numbers present in the data set. The dimension of the two-body interaction descriptors is $15 \cdot n \cdot (n + 1)/2$.

Invariant Three-Body Interaction Descriptors F_{3B}. We define the set of translational and rotational invariant three-body interaction terms

DOI: 10.1021/acs.jctc.8b00110
J. Chem. Theory Comput. 2018, 14, 2991−3003

Table 4. Prediction Errors of the B3LYP/6-31G(2df,p) Atomization Energy of the Molecules of the Set GDB-9 by Various ML Models with Random 5K Train Molecules and the Remaining 126722 Molecules as Test Set[a]

method	features	MAE	RMSE	MAX. DEV
mean		185	235	1544
RR	CM	235	308	1289
RR	BOB	89	134	653
RR	F_{2B}	6.8	10	462
RR	$F_{2B} + F_{3B}$	1.6	**2.8**	88
KNN	CM	239	279	898
KNN	BOB	231	272	758
KNN	F_{2B}	151	177	556
KNN	$F_{2B} + F_{3B}$	25	42	358
KRR (Gauss)	CM	17	22	181
KRR (Laplace)	CM	7.9	10	129
KRR (Gauss)	BOB	11	16	253
KRR (Laplace)	BOB	4.0	6.0	132
KRR (Gauss)	F_{2B}	4.8	6.4	**45**
KRR (Laplace)	F_{2B}	8.2	11	190
KRR (Gauss)	$F_{2B} + F_{3B}$	**1.5**	**2.8**	96
KRR (Laplace)	$F_{2B} + F_{3B}$	4.5	6.4	147

[a]The errors are given in kcal/mol. The models used are ridge regression (RR), kernel ridge regression (KRR), and k-nearest neighbors (KNN).

$$p^{3B}_{m_1,m_2,m_3}(\mathbf{r}_1, Z_1, \mathbf{r}_2, Z_2, \mathbf{r}_3, Z_3) := \frac{1}{\|\mathbf{r}_{12}\|^{m_1}\|\mathbf{r}_{13}\|^{m_2}\|\mathbf{r}_{23}\|^{m_3}} \cdot\theta(Z_1, Z_2, Z_3, \|\mathbf{r}_{12}\|, \|\mathbf{r}_{13}\|, \|\mathbf{r}_{23}\|) \quad (4)$$

where $m_1, m_2, m_3 \in \mathbb{N}^+$, $\mathbf{r}_{ij} := \mathbf{r}_i - \mathbf{r}_j$ for $i, j = \{1,2,3\}$, and the bond angle indicator

$$\theta(\cdot) := \begin{cases} 1, & d_{12} < B(Z_1, Z_2) \wedge d_{13} < B(Z_1, Z_3) \\ 1, & d_{13} < B(Z_1, Z_3) \wedge d_{23} < B(Z_2, Z_3) \\ 1, & d_{12} < B(Z_1, Z_2) \wedge d_{23} < B(Z_2, Z_3) \\ 0, & \text{otherwise} \end{cases} \quad (5)$$

where $B(Z_1, Z_2) := 1.1 \cdot L(Z_1, Z_2)$, and the values for the bond length function L are given in Table 6. For a given set of n different atomic numbers $A_n := \{Z_i\}_{i=1}^n$ with $Z_i \neq Z_j \; \forall \; i, j \in \{1, \cdots, n\}$, let S_{3B} denote the set of all 3-tuples (Z_i, Z_j, Z_k) with $Z_i \leq Z_k$ and $Z_i, Z_j, Z_k \in A_n$. Let M_{3B} be the set of partial permutations $G(3, 6)$ as defined above. For a given physical system S, the three-body interaction descriptors F_{3B} are now given by

$$F_{3B} := \{f_{\bar{\mathbf{Z}}, p^{3B}_{m_1,m_2,m_3}}(S)\}_{(m_1,m_2,m_3)\in M_{3B}, \bar{\mathbf{Z}}\in S_{3B}} \quad (6)$$

The dimension of the of the three-body interaction descriptors is $n^2 \cdot (n+1)/2 \cdot \frac{6!}{(6-3)!}$.

The difference between the molecular descriptors BOB, F_{2B}, and F_{3B} is illustrated in Figures 1 and 2, respectively. The bag-of-bonds model compares arbitrary pairwise distances with each other, while for the proposed $F_{2B} + F_{3B}$ descriptors, two- and three-body features are computed, and corresponding features are compared with each other.

TESTS ON MOLECULAR DATA SETS

We use the following two reference data sets for the evaluation of the predictive power of ML models with our proposed invariant many-body interaction descriptors.

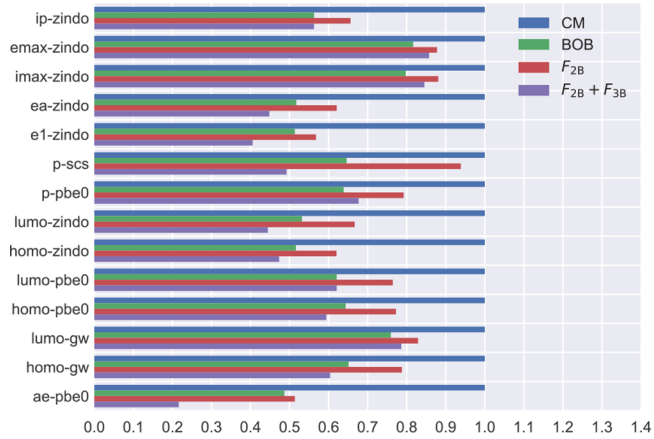


Figure 3. Mean absolute errors of several electronic ground- and excited-state properties of the molecules of the set GDB-7 predicted with KRR using the descriptors CM, BOB, F_{2B}, and $F_{2B} + F_{3B}$. For CM and BOB, the Laplace kernel has been used; for F_{2B} and $F_{2B} + F_{3B}$, the Gauss kernel has been used. The MAEs are normalized by the MAE of the KRR-CM model.

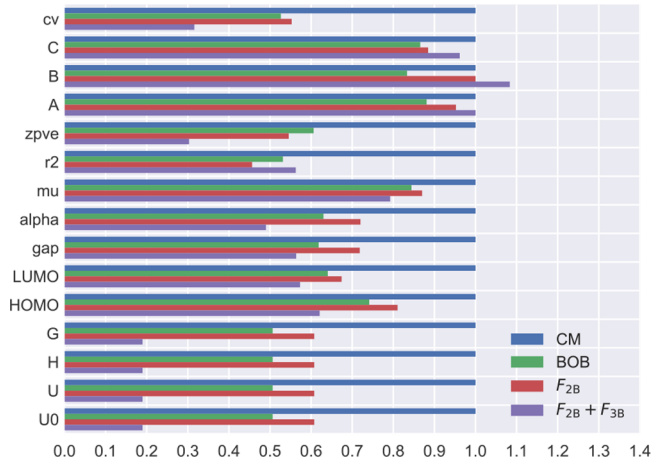


Figure 4. Mean absolute errors of several properties of the molecules of the set GDB-9 calculated at the B3LYP/6-31G(2df,p) level of quantum chemistry and predicted with KRR using the descriptors CM, BOB, F_{2B}, and $F_{2B} + F_{3B}$. For CM and BOB, the Laplace kernel has been used; for F_{2B} and $F_{2B} + F_{3B}$, the Gauss kernel has been used. The MAEs are normalized by the MAE of the KRR-CM model.

GDB-7. The GDB-7 data set is a subset of the freely available small molecule database GDB-13[41] with up to seven heavy atoms CNO. For this data set, electronic ground- and excited-state properties have been calculated. Hybrid density functional theory with the Perdew−Burke−Ernzerhof hybrid functional approximation (PBE0)[42,43] has been used to calculate the atomization energy of the molecules. The electron affinity, ionization potential, excitation energies, and maximal absorption intensity have been obtained from ZINDO.[44−46] For the static polarizability, PBE0 and self-consistent screening (SCS)[47] have been used. The frontier orbital (HOMO and LUMO) eigenvalues have been calculated using PBE0, SCS, and Hedin's GW approximation.[48] The SCS, PBE0, and GW calculations have been performed using FHI-AIMS[49] (tight settings/tier2 basis set), and ZINDO/s calculations are based on the ORCA[50] code.

GDB-9. The GDB-9 data set is a subset of the chemical universe database GDB-17[51] of 166 billion organic small molecules. The subset contains molecules with up to nine heavy

DOI: 10.1021/acs.jctc.8b00110
J. Chem. Theory Comput. 2018, 14, 2991−3003

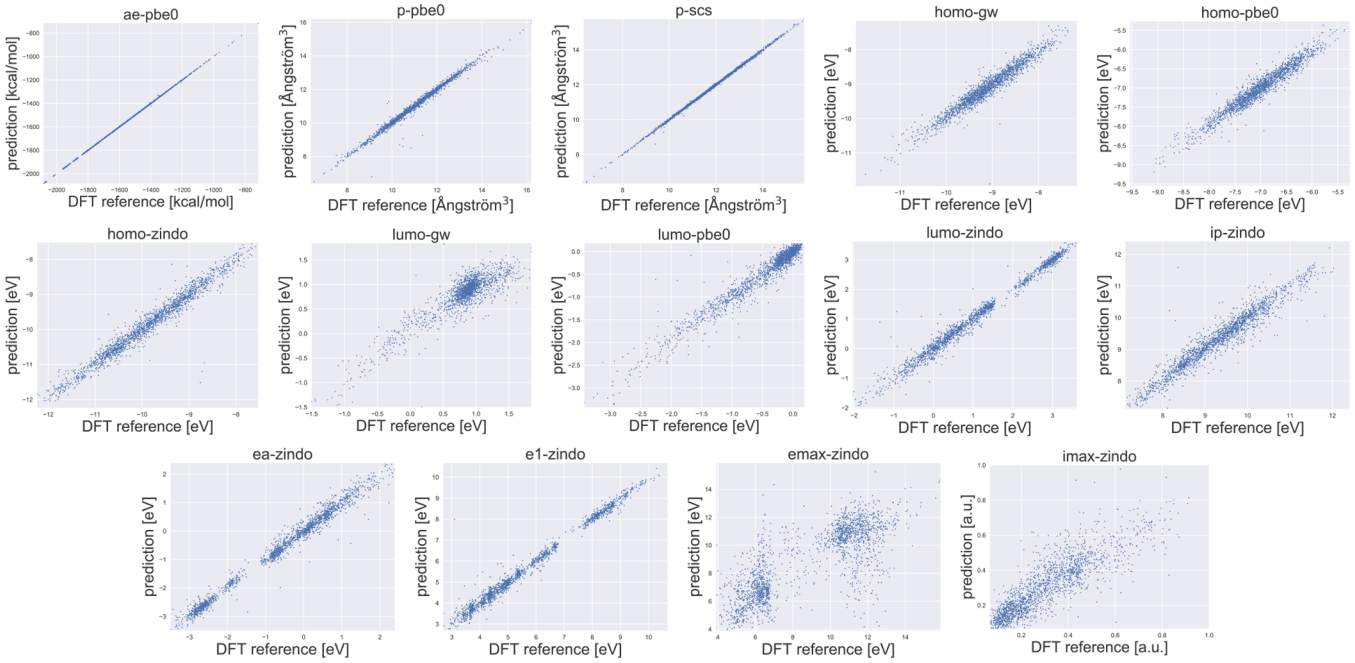


Figure 5. Prediction versus DFT reference of several electronic ground- and excited-state properties of the molecules of the set GDB-7 predicted with KRR using the Gauss kernel and the descriptors F_{2B} + F_{3B}.

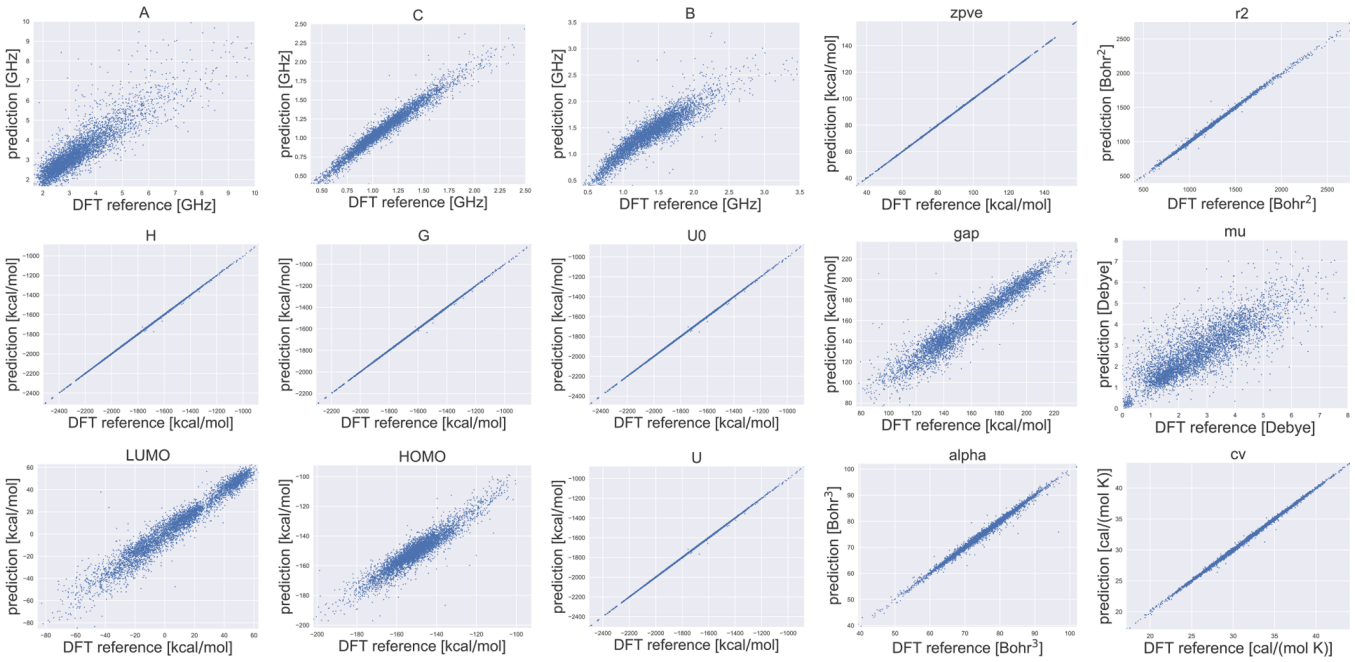


Figure 6. Prediction versus DFT reference of several properties of the molecules of the set GDB-9 calculated at the B3LYP/6-31G(2df,p) level of quantum chemistry and predicted with KRR using the Gauss kernel and the descriptors F_{2B} + F_{3B}.

atoms CNO with corresponding harmonic frequencies, dipole moments, and polarizabilities, along with energies, enthalpies, and free energies of atomization, all calculated at the B3LYP/6-31G(2df,p) level of quantum chemistry.[52]

We evaluate the performance of predicting the properties of the molecules of these two data sets by using our proposed invariant many-body interaction descriptors F_{2B} and F_{2B} + F_{3B}. Additionally, we computed the sorted Coulomb matrices (CM)[8] and the popular bag-of-bonds (BOB)[34] molecular representations. For the atomization energy, we use the models kernel ridge regression (KRR) (see e.g. Hansen et al.[8] and Müller et al.[53]), ridge regression (RR),[54] k-nearest neighbors (KNN),[55] and the mean predictor (MEAN), see Appendix A. For the other properties, we use kernel ridge regression with the Laplace kernel for CM and BOB which works better compared to the Gauss kernel for these descriptors[8] and the Gauss kernel in combination with the F_{2B} and F_{2B} + F_{3B} descriptors, respectively. To fit the model parameters (hyperparameters), we use 10-fold cross-validation,[56] see ref 8 for details. Unless otherwise noted, the models are trained on 5000 random molecules. The performance is evaluated on the remaining molecules of the respective set, by the mean absolute error (MAE), the root-mean-square

DOI: 10.1021/acs.jctc.8b00110
J. Chem. Theory Comput. 2018, 14, 2991−3003

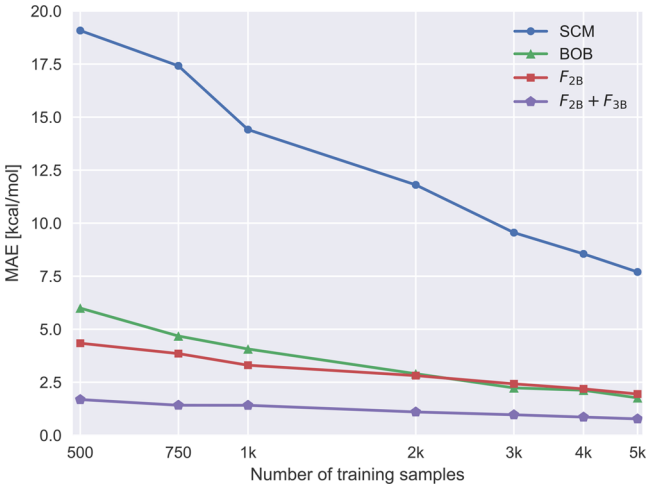


Figure 7. Mean absolute error of predicting the PBE0 atomization energy of the molecules of the set GDB-7 with KRR in dependence of the number of training samples. The errors are given in kcal/mol. For CM and BOB, the Laplace kernel has been used; for F_{2B} and $F_{2B} + F_{3B}$, the Gauss kernel has been used. The model hyperparameters have been determined by 10-fold cross-validation.

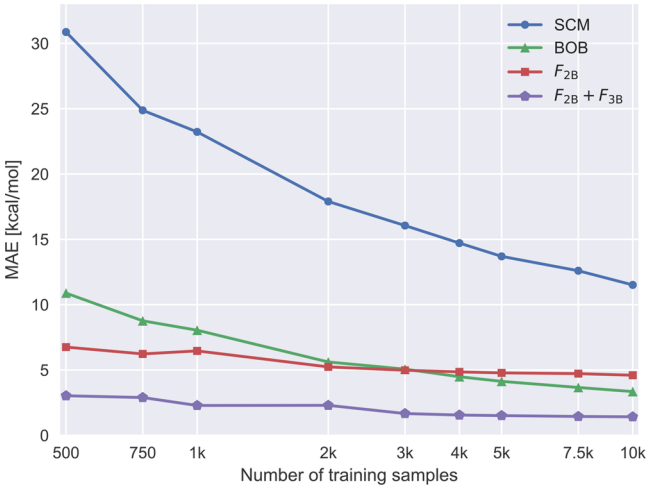


Figure 8. Mean absolute error of predicting the B3LYP/6-31G(2df,p) atomization energy of the molecules of the set GDB-9 with KRR in dependence of the number of training samples. The errors are given in kcal/mol. For CM and BOB, the Laplace kernel has been used; for F_{2B} and $F_{2B} + F_{3B}$, the Gauss kernel has been used. The model hyperparameters have been determined by 10-fold cross-validation.

error (RMSE), and the maximum deviation (MAX. DEV), respectively.

For the atomization energy, the results of the ML models are given in Tables 1 and 4. The results for predicting diverse quantum mechanical properties are given in Tables 2 and 3, respectively. Figures 3 and 4 show the MAE of the models normalized by the mean absolute error of the CM model. From these results it is not clear how good a specific molecular property can be predicted. Therefore, Figures 5 and 6 show the predicted properties relative to the reference results. The MAE in dependence of the number of training samples is shown in Figures 7 and 8, respectively.

The $F_{2B} + F_{3B}$ model outperforms the BOB descriptor in the prediction of the static polarizability computed with self-consistent screening (20% improvement), the first excitation energy (20% improvement), and the atomization energy

Table 5. Mean Absolute Errors of Predicting the B3LYP/6-31G(2df,p) Atomization Energy of 3000 Random Small and Large Molecules with Kernel Ridge Regression[a]

train/test set	CM	BOB	$F_{2B} + F_{3B}$
small/small	7.2	3.5	**1.4**
small/large	733	493	**6.2**
large/small	793	797	**50**
large/large	3.2	1.2	**0.5**

[a]The models are trained on sets of 3000 random small and large molecules, respectively. The errors are given in kcal/mol. Best results are marked bold.

Table 6. Bond Lengths in Ångström for All Combinations of the Elements C, H, N, and O[a]

bond-type	(Z_1, Z_2)	$L(Z_1, Z_2)$
H−H	(1, 1)	0.74
H−C	(1, 6)	1.08
H−O	(1, 8)	0.96
H−N	(1, 7)	1.01
C−C	(6, 6)	1.51
C−O	(6, 8)	1.43
C−N	(6, 7)	1.47
O−O	(8, 8)	1.48
O−N	(8, 7)	1.40
N−N	(7, 7)	1.45

[a]Used for computing the three-body interaction descriptors F_{3B}.

(50% improvement) of the molecules of the GDB-7 set. Additionally, the prediction errors of the electron affinity and the HOMO eigenvalues are improved by 5%. The largest correlation between prediction and reference is achieved for the static polarizability computed with SCS as well as the atomization energy.

The $F_{2B} + F_{3B}$ model outperforms the BOB descriptor in the prediction of the heat capacity (40% improvement), the zero point vibrational energy (50% improvement), the isotropic polarizability (30% improvement), and the atomization energies (60% improvement) of the molecules of the GDB-9 set. Additionally, the prediction errors of the HOMO and LUMO eigenvalues as well as the gap are improved by 15%, 10%, and 9%, respectively. The largest correlation between prediction and reference is achieved for the electronic spatial extent, zero point vibrational energy, the heat capacity, the isotropic polarizability, and the atomization energies.

The three-body descriptors F_{3B} are local in the sense that they include pairs of bonded atoms which have an angle. This locality suggests an applicability of our descriptors to predict quantum mechanical properties of much larger molecules. In a first attempt to justify such transferability, we conducted an additional experiment where we predict the atomization energy of a set of large molecules when trained on a set of small molecules and vice versa. To this end, we select a set of small molecules composed of 3000 random molecules of the GDB9 set with a number of atoms smaller than 14. Similarly, the set of larger molecules consists of 3000 random molecules of the GDB9 set with a number of atoms larger than 22. The results of training and testing a kernel ridge regression model in combination with the CM, BOB, and $F_{2B} + F_{3B}$ descriptors using all combinations of small and large molecule sets are shown in Table 5. The CM and BOB models show poor performance when predicting the atomization energy of the larger molecule set from the small molecule set and vice

DOI: 10.1021/acs.jctc.8b00110
J. Chem. Theory Comput. 2018, 14, 2991−3003

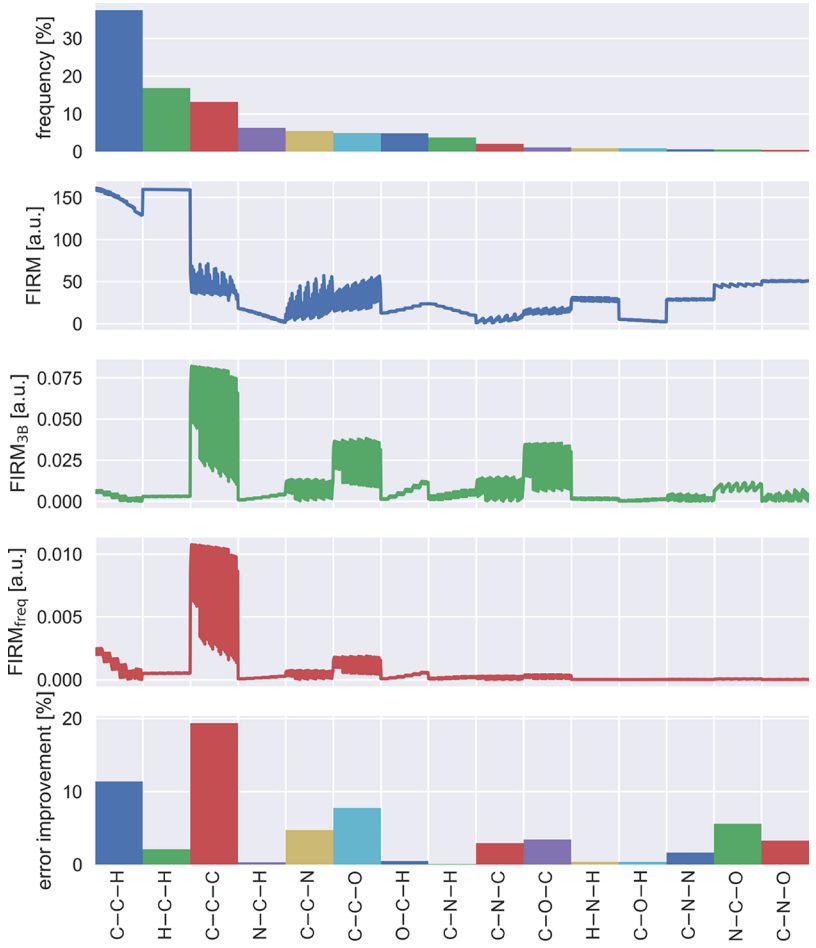


Figure 9. FIRM by eq 7 (second from top), $FIRM_{3B}$ by eq 8 (third from top), and $FIRM_{freq}$ by (9) (fourth from top) for the F_{3B} descriptors of the molecules of the set GDB-7. Additionally, the frequency of the corresponding bond-type (top) and the error improvement by using KRR with the F_{2B} features in combination with the bond-type subset of the F_{3B} descriptors (bottom) are shown.

versa. Using our F_{2B} + F_{3B} descriptors significantly improves the prediction accuracy of learning the atomization energy of large molecules when trained on the set of small molecules. As kernel ridge regression intrinsically suffers from differing data distributions of the training and test sets, more accurate models or those compensating for nonstationarity such as covariate shift[57] using our many-body descriptors can potentially improve these transferability results.

The prediction of the atomization energy by using the linear RR model is comparable to the KRR model. This makes the F_{2B} + F_{3B} descriptors interesting candidates for alternative linear regression models such as Bayesian linear regression,[58] partial least-squares,[59] or generalized least-squares.[60] In this work, we will utilize this fact to compute a feature ranking measure in the next section.

FEATURE IMPORTANCE OF THE INVARIANT MANY-BODY INTERACTION DESCRIPTORS

The inclusion of the three-body descriptors F_{3B} increases the predictive power of the KRR model by more than 50% over using the two-body descriptors F_{2B} for both data sets GDB-7 and GDB-9. Due to the nonlinear kernels used, it is not obvious how the three-body features improve the performance. The frequencies of the bond-types corresponding to three bonded atoms which have an angle (Figure 9 and Figure 10 top) suggest the top three most important connections C−C−H, H−C−H, and C−C−C, respectively. On the other hand, using the F_{2B} descriptors in combination with the H−C−H subset of F_{3B} features (Figure 9 and Figure 10 bottom) shows a negligible decrease of the mean absolute error of the KRR model as compared to the inclusion of the C−C−H and C−C−C subsets.

There are a number of ways to define feature importance[61−64] respectively to explain nonlinear models.[65−71] Here, we use the feature importance ranking measure (FIRM),[72] which defines the feature importance according to the standard deviation of a conditional expected output of the learner. FIRM can be applied to a broad family of learning machines, and the measure is robust with respect to perturbation of the problem and invariant with respect to irrelevant transformations. In general, the computation of the exact FIRM is infeasible. For the unregularized linear regression model and normally distributed input features, the FIRM of a feature f can be computed analytically[72] by

$$\mathrm{FIRM}(f) := \frac{1}{n}\cdot\frac{1}{\sigma(f)}\cdot\mathrm{cov}(f, y) \tag{7}$$

where n is the number of samples, $\sigma(\cdot)$ is the standard deviation, y denotes the labels, and $\mathrm{cov}(\cdot)$ is the covariance. In the above formula, FIRM is computed for each feature independently. To capture the importance of the inclusion of the three-body descriptors F_{3B}, we propose to use FIRM on the signed deviation of labels and prediction of the KRR model with the two-body features F_{2B}

DOI: 10.1021/acs.jctc.8b00110
J. Chem. Theory Comput. 2018, 14, 2991−3003

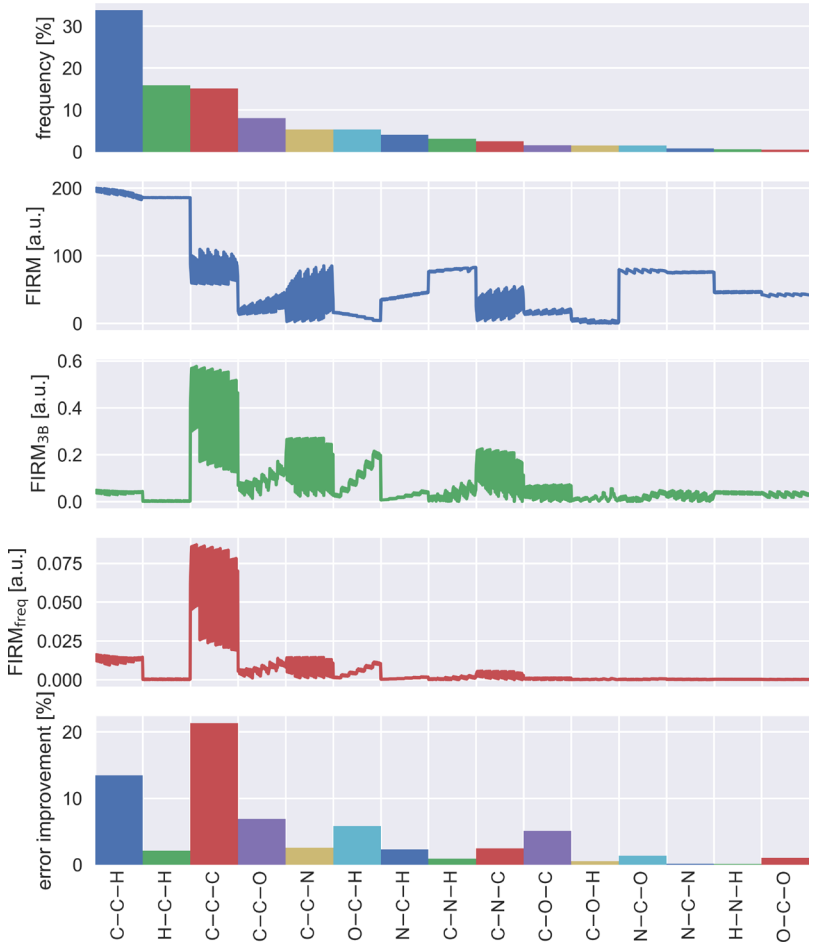


Figure 10. FIRM by eq 7 (second from top), $FIRM_{3B}$ by eq 8 (third from top), and $FIRM_{freq}$ by (9) (fourth from top) for the F_{3B} descriptors of the molecules of the set GDB-9. Additionally, the frequency of the corresponding bond-type (top) and the error improvement by using KRR with the F_{2B} features in combination with the bond-type subset of the F_{3B} descriptors (bottom) are shown.

Table 7. Frequency of the Bond-Type and the Error Improvement by Using KRR with the F_{2B} Features in Combination with the Bond-Type Subset of the F_{3B} Descriptors for the Set GDB-7

bond-type	freq [%]	error improvement [%]
C−C−H	38	11.4
H−C−H	17	2.1
C−C−C	13	19.3
N−C−H	6.3	0.3
C−C−N	5.4	4.7
C−C−O	4.9	7.8
O−C−H	4.8	0.5
C−N−H	3.7	0.07
C−N−C	2.0	2.9
C−O−C	1.1	3.4
H−N−H	0.9	0.4
C−O−H	0.8	0.4
C−N−N	0.6	1.6
N−C−O	0.5	5.6
C−N−O	0.4	3.3

Table 8. Frequency of the Bond-Type and the Error Improvement by Using KRR with the F_{2B} Features in Combination with the Bond-Type Subset of the F_{3B} Descriptors for the Set GDB-9

bond-type	freq [%]	error improvement [%]
C−C−H	34	13.5
H−C−H	16	2.1
C−C−C	15	21.3
C−C−O	8.0	6.9
C−C−N	5.3	2.5
O−C−H	5.3	5.8
N−C−H	4.1	2.3
C−N−H	3.1	0.9
C−N−C	2.5	2.5
C−O−C	1.5	5.1
C−O−H	1.5	0.5
N−C−O	1.5	1.3
N−C−N	0.8	0.2
H−N−H	0.6	0.1
O−C−O	0.4	1.0

$$\mathrm{FIRM}_{3B}(f) := \frac{1}{n}\cdot\frac{1}{\sigma(f)}\cdot\mathrm{cov}(f, y - p_{2B}) \tag{8}$$

where p_{2B} is the prediction of the KRR model using the F_{2B} descriptors, see also ref 73. Additionally, we compute the product of the above $FIRM_{3B}$ of the feature f with the frequency of its corresponding bond-type

$$\mathrm{FIRM}_{freq}(f) := \mathrm{freq}(f)\cdot\frac{1}{n}\cdot\frac{1}{\sigma(f)}\cdot\mathrm{cov}(f, y - p_{2B}) \tag{9}$$

DOI: 10.1021/acs.jctc.8b00110
J. Chem. Theory Comput. 2018, 14, 2991−3003

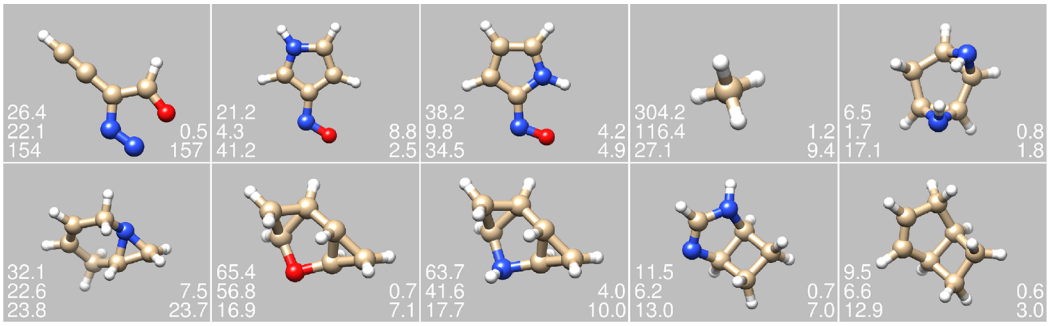


Figure 11. Molecules of the set GDB-7 with the largest difference of the mean absolute error of KRR using the F_{2B} and $F_{2B} + F_{3B}$ descriptors, respectively. The MAE in kcal/mol is shown for CM (top left), BOB, (left middle), F_{2B} (left bottom), $F_{2B} + F_{3B}$ (top right), and the F_{2B} + C−C−C subset of F_{3B} (right bottom), respectively.

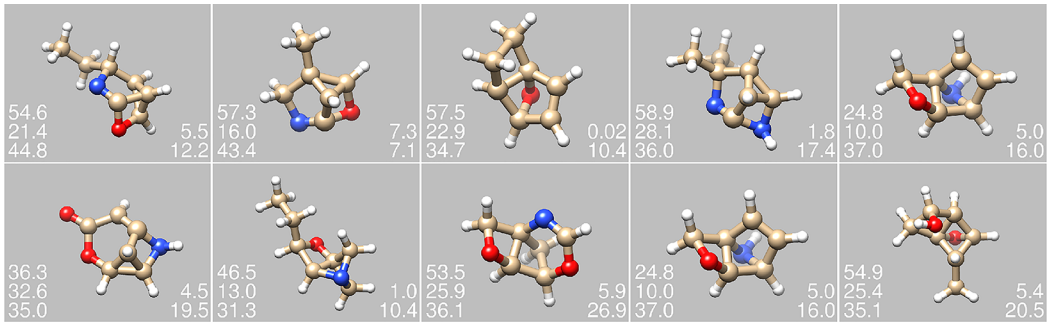


Figure 12. Molecules of the set GDB-9 with the largest difference of the mean absolute error of KRR using the F_{2B} and $F_{2B} + F_{3B}$ descriptors, respectively. The MAE in kcal/mol is shown for CM (top left), BOB, (left middle), F_{2B} (left bottom), $F_{2B} + F_{3B}$ (top right) and the F_{2B} + C−C−C subset of F_{3B} (right bottom), respectively.

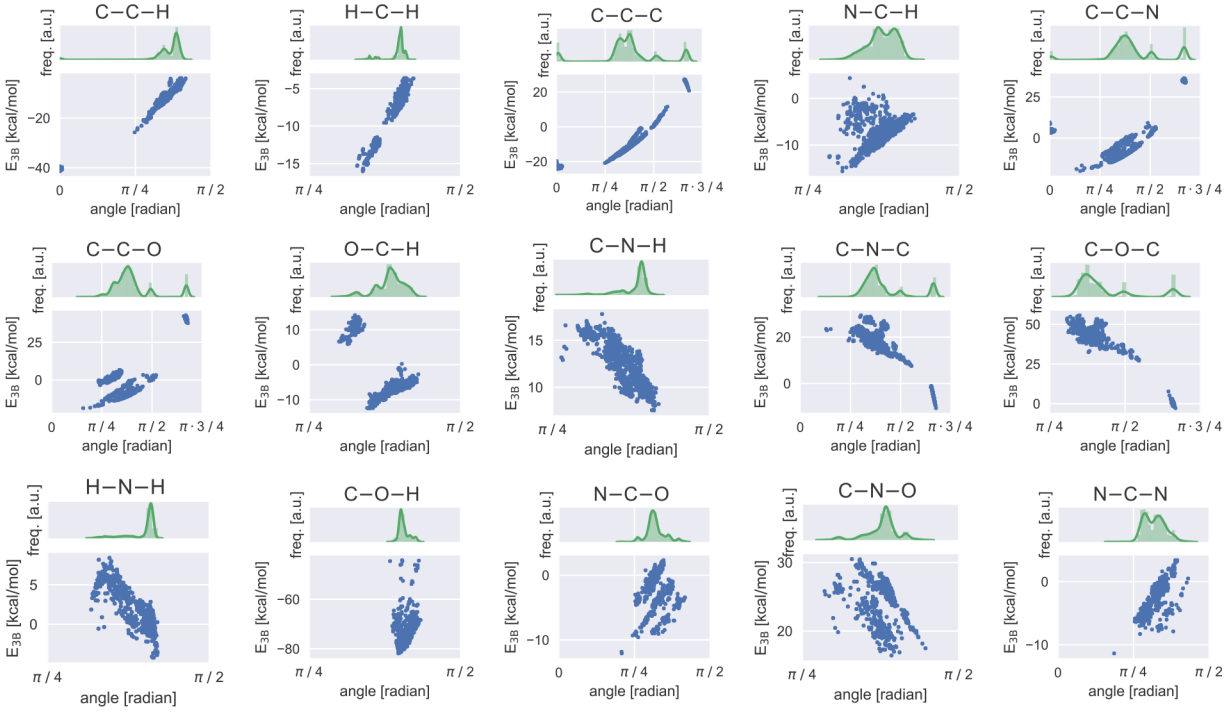


Figure 13. E_{3B} by eq 10 in dependence of the bond angle for the dominant bond-types of the molecules of the set GDB-7 along with the distribution of angles.

where freq(f) is the frequency of the bond-type corresponding to the feature f. Figure 9 and Figure 10 show the FIRM, FIRM_{3B}, and $\text{FIRM}_{\text{freq}}$ for the three-body descriptors F_{3B} for both data sets GDB-7 and GDB-9. Additionally, we show the frequency of the bond-type corresponding to the feature f and the error improvement of using the KRR model with the F_{2B} features augmented with the corresponding subset of three-body features F_{3B}. The frequencies and error improvements are shown in Tables 7 and 8, respectively.

DOI: 10.1021/acs.jctc.8b00110
J. Chem. Theory Comput. 2018, 14, 2991−3003

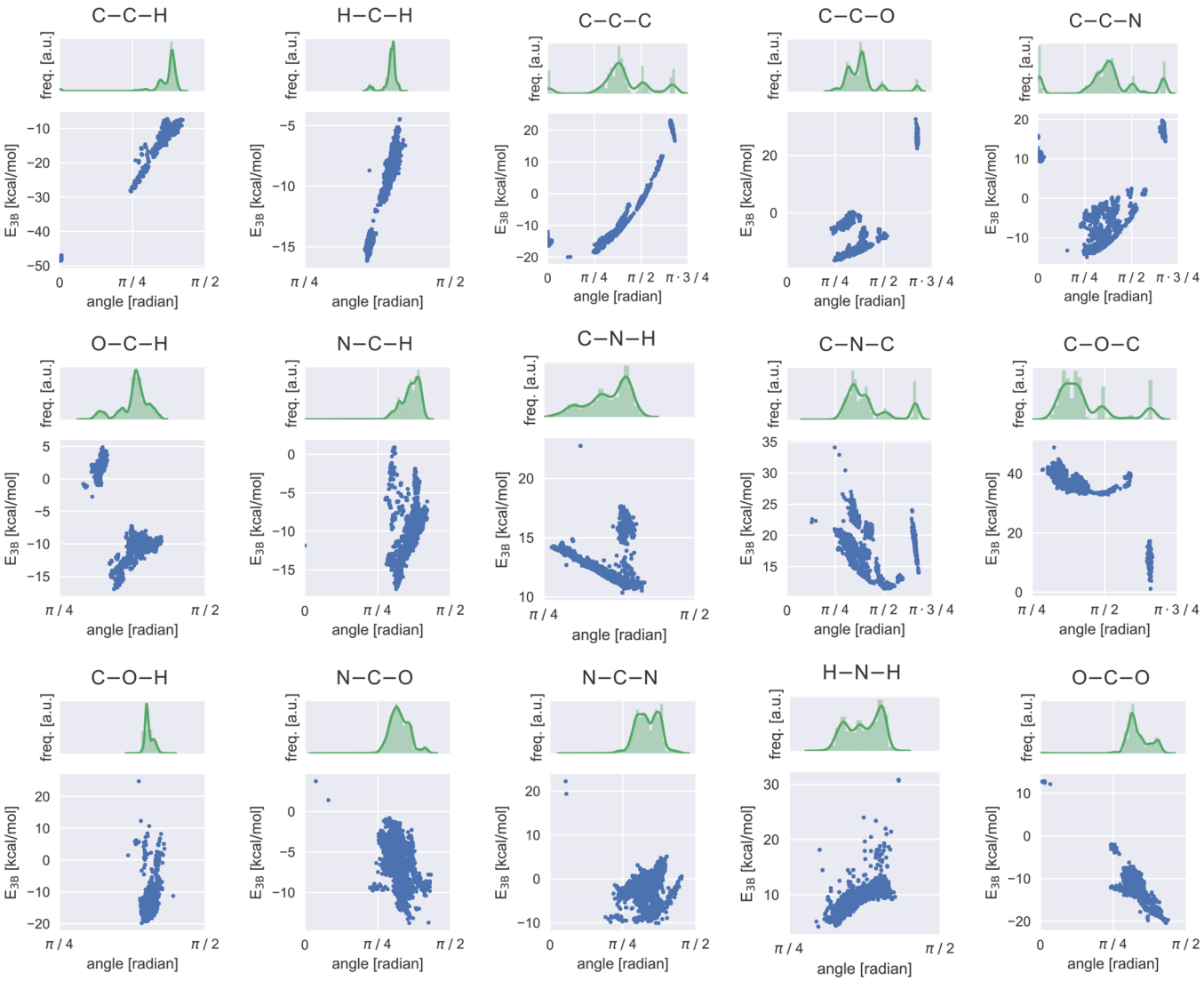


Figure 14. E_{3B} by eq 10 in dependence of the bond angle for the dominant bond-types of the molecules of the set GDB-9 along with the distribution of angles.

The $FIRM_{3B}$ indicates low importance of the H−C−H and increased importance of the C−C−C features, which correlates with the error improvement by using these features in combination with the F_{2B} descriptors. This indicates that three-body interactions relevant for prediction improvement are more dominant for the C−C−C bond-type as compared to the H−C−H bond-type, where the correlation with the atomization energy can be captured by using the corresponding two-body features F_{2B}. Figure 11 and Figure 12 show the molecules with the largest difference of absolute errors of the KRR F_{2B} and F_{2B} + F_{3B} models. For these molecules, using the combination of the F_{2B} descriptors with the subset of C−C−C three-body features significantly improves the predictive performance of the KRR model using the F_{2B} features.

The measure $FIRM_{3B}$ reduces the importance of the hydrogen type bonds in favor of the non-hydrogen features, as compared to FIRM. The correlation of a molecular descriptor with the target (atomization energy) is not necessarily a good predictor variable in the presence of other features. In this sense, $FIRM_{3B}$ captures the importance of the three-body descriptors F_{3B} in the presence of the two-body interactions modeled by the two-body descriptors F_{2B}. For the non-hydrogen type three-body features, FIRM indicates approximately equal importance of the C−C−C, C−C−N, and N−C−O bonds, in contrast to $FIRM_{3B}$, which lifts the C−C−C importance. This shows, that for non-hydrogen bonds, our set of descriptors is better able to capture three-body interactions of the C−C−C type as compared to the other bond-types. In view of the 5 times lower frequency of the N−C−O bond compared to C−C−N, both, the error improvement and $FIRM_{3B}$ show approximately equal importance of these three-body interactions.

For the three-body features, we can use the parameters of the linear RR model to compute the energy of a given bond-type

$$E_{3B}(b) := \sum_{i=1}^{N} \delta_b(\mathrm{bond}(i)) \cdot c_i \cdot f_i \tag{10}$$

where c_i are the coefficients of the trained RR model, f_i are the three-body features, N is the number of three-body features, b is the bond-type under examination, and bond(i) indicates the bond-type corresponding to the feature f_i. Figure 13 and Figure 14 show E_{3B} in dependence of the bond angle for the 15 predominant bond-types of the GDB-7 and GDB-9 set, respectively.

DOI: 10.1021/acs.jctc.8b00110
J. Chem. Theory Comput. 2018, 14, 2991−3003

Physically, these results indicate, that for intermediate size molecules, the interaction of the hydrogen atom with all other atoms (of type C, N, O) can be captured effectively by pairwise interactions. In fact, if we use the F_{2B} features in combination with the non-hydrogen subset of F_{3B}, we get a mean absolute error of 0.9 kcal/mol for the GDB-7 set and 1.8 kcal/mol for the GDB-9 set on the rest of the molecules, respectively. In view of the fact that the hydrogen atom constitutes by far the dominant atom type for both data sets, the errors degrade by 13% and 20% as compared to the full $F_{2B} + F_{3B}$ descriptors, respectively. This intriguing result lets us formulate the following conjecture:

For the accurate prediction of the atomization energy of intermediate size molecules, the interaction potential of the hydrogen atom with all other atoms can be effectively approximated as a pairwise interaction potential.

The interatomic interaction between non-hydrogen atoms goes beyond pairwise interactions. Interestingly, for the C−C−C bond-type, the energy E_{3B} shows a clear dependence of the bond angle, as compared to the other bond-types. This result indicates that there is a simple relation between the angle at the C atom of the C−C−C bond-type and the atomization energy. Between the angles $\pi/4$ and $\pi/2$, there exist two branches of the dependence of the atomization energy of the angle. This indicates, that for C−C−C, our model learns two angle-type functions, distinguishing single−double and single−single C−C−C bonds, see the C−C−C angle dependence of E_{3B} in Figures 13 and 14, respectively.

CONCLUSION

We developed a new set of translation, rotation, and atomic indexing invariant many-body interaction descriptors which avoid the perhaps somewhat artificial sorting of the feature entries. With our two- and three-body invariant molecular representation, the atomization energy and polarizability of small organic molecules can be accurately predicted. Using our descriptors, the performance of the linear regularized ridge regression model is comparable to the nonlinear kernel methods. Applying a feature importance ranking measure, we show that the C−C−C bond-type is - as to be expected - the most important three-body interaction for predicting the atomization energy. In addition, the bonds involving hydrogens and heavier atoms can be captured effectively by pairwise interatomic potentials.

APPENDIX A: MACHINE LEARNING MODELS

Ridge Regression (RR)

Given N samples of dimension D, the data can be represented by the design matrix $\mathbf{X} \in \mathbb{R}^{N\times D}$. In ridge regression, the predictions are given by

$$\boldsymbol{p} = \mathbf{X}\boldsymbol{w} \tag{11}$$

The weights $\boldsymbol{w}$ are computed by minimizing the objective

$$(\boldsymbol{y} - \mathbf{X}\boldsymbol{w})^2 + \lambda\cdot\boldsymbol{w}^{\mathrm{T}}\boldsymbol{w} \tag{12}$$

with the labels $\boldsymbol{y}$. The solution for $\boldsymbol{w}$ is given by

$$\boldsymbol{w} = (\lambda\cdot\mathbf{I} + \mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\boldsymbol{y} \tag{13}$$

where λ is the regularization parameter, which can be determined by grid search and cross-validation to prevent overfitting.

Kernel Ridge Regression (KRR)

In kernel ridge regression, a kernel is used as similarity measure between two molecules. From this similarity measure, the prediction of the atomization energy of a new molecule $\boldsymbol{x}$ is obtained by

$$E = \sum_{i=1}^{N} \alpha_i\cdot K(\boldsymbol{x}_i, \boldsymbol{x}) \tag{14}$$

where the α_i denote the weighting coefficients obtained by training the model, and N is the number of training molecules $\boldsymbol{x}_i$. Training the model involves a set of molecules with known labels $\{\boldsymbol{x}_i, E_i\}$ from which the $\boldsymbol{\alpha}$ is obtained by solving a regularized system of linear equations

$$(\lambda\cdot\mathbf{I} + \mathbf{K})\cdot\boldsymbol{\alpha} = \boldsymbol{E} \tag{15}$$

where $\mathbf{K}_{ij} := K(\boldsymbol{x}_i, \boldsymbol{x}_j)$, and λ is the regularization parameter.

Popular choices of kernels include the Gaussian kernel

$$K_{\mathrm{Gauss}}(\boldsymbol{x}, \boldsymbol{y}) := e^{-\|\boldsymbol{x}-\boldsymbol{y}\|_2^2/(2\sigma^2)} \tag{16}$$

where the distance measure is the scaled squared Euclidean l_2-distance between the features of the molecules, and the Laplace kernel

$$K_{\mathrm{Laplace}}(\boldsymbol{x}, \boldsymbol{y}) := e^{-\|\boldsymbol{x}-\boldsymbol{y}\|_1/\sigma} \tag{17}$$

where the distance measure is the scaled absolute difference. Both of these kernels contain a scaling hyperparameter σ, which, together with the regularization parameter λ, we determine by grid search and cross-validation to prevent overfitting.

k-Nearest Neighbors (KNN)

In KNN regression, the output of a sample $\boldsymbol{x}$ is the average of the values of its k-nearest neighbors, where k is a hyperparameter which can be determined by cross-validation. In analogy to the KRR case, we use the l_1-distance for BOB and CM and the l_2-distance for our F_{2B} and F_{3B} descriptors.

Mean Predictor (MEAN)

For the MEAN predictor, the output is constant for all samples in the test set. This constant is given by the average value of the output of the samples in the training set.

APPENDIX B: BOND LENGTHS

Table 6 lists the bond lengths in Ångström for all combinations of the elements C, H, N, and O used to compute the three-body interaction descriptors F_{3B}.

APPENDIX C: FREQUENCIES OF THREE-BODY BONDS

Table 7 and Table 8 list the frequency of the three-body bond-type and the error improvement by using KRR with the F_{2B} features in combination with the bond-type subset of the F_{3B} descriptors for the set GDB-7 and GDB-9, respectively.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jctc.8b00110.

Pseudocode for calculating two- and three-body descriptors (Algorithms 1 and 2) (PDF)

AUTHOR INFORMATION

Corresponding Authors

*E-mail: alexandre.tkatchenko@uni.lu (A.T.).
*E-mail: klaus-robert.mueller@tu-berlin.de (K.-R.M.).

ORCID

Wiktor Pronobis: 0000-0003-4849-7151
Alexandre Tkatchenko: 0000-0002-1012-4854

DOI: 10.1021/acs.jctc.8b00110
J. Chem. Theory Comput. 2018, 14, 2991−3003

Funding

This work was supported by the Federal Ministry of Education and Research (BMBF) for the Berlin Big Data Center BBDC (01IS14013A). Additional support was provided by the DFG (MU 987/20-1), from the European Unions Horizon 2020 research and innovation program under the Marie Sklodowska-Curie grant agreement NO 657679, the BK21 program funded by Korean National Research Foundation grant (No. 2012-005741), and the Institute for Information & Communications Technology Promotion (IITP) grant funded by the Korea government (no. 2017-0-00451).

Notes

The authors declare no competing financial interest.

REFERENCES

(1) Moravčík, M.; Schmid, M.; Burch, N.; Lisý, V.; Morrill, D.; Bard, N.; Davis, T.; Waugh, K.; Johanson, M.; Bowling, M. DeepStack: Expert-level artificial intelligence in heads-up no-limit poker. *Science* **2017**, *356*, 508−513.

(2) Ontanon, S.; Synnaeve, G.; Uriarte, A.; Richoux, F.; Churchill, D.; Preuss, M. A Survey of Real-Time Strategy Game AI Research and Competition in StarCraft. *IEEE Transactions on Computational Intelligence and AI in Games* **2013**, *5*, 293−311.

(3) Silver, D.; et al. Mastering the game of Go without human knowledge. *Nature* **2017**, *550*, 354−359.

(4) Hinton, G.; Deng, L.; Yu, D.; Dahl, G.; Mohamed, A.; Jaitly, N.; Senior, A.; Vanhoucke, V.; Nguyen, P.; Sainath, T.; Kingsbury, B. Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups. *IEEE Signal Processing Magazine* **2012**, *29*, 82−97.

(5) He, K.; Zhang, X.; Ren, S.; Sun, J. *Deep Residual Learning for Image Recognition*. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016; pp 770−778, DOI: 10.1109/CVPR.2016.90.

(6) van den Oord, A.; Dieleman, S.; Zen, H.; Simonyan, K.; Vinyals, O.; Graves, A.; Kalchbrenner, N.; Senior, A. W.; Kavukcuoglu, K. *WaveNet: A Generative Model for Raw Audio*. 2016, arXiv preprint arXiv:1609.03499.

(7) 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.

(8) 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.

(9) 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.

(10) 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. Prediction Errors of Molecular Machine Learning Models Lower than Hybrid DFT Error. *J. Chem. Theory Comput.* **2017**, *13*, 5255−5264.

(11) 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.

(12) Schütt, K. T.; Kindermans, P.-J.; Sauceda Felix, H. E.; Chmiela, S.; Tkatchenko, A.; Müller, K.-R. SchNet: A continuous-filter convolutional neural network for modeling quantum interactions. *Advances in Neural Information Processing Systems 30* **2017**, 992−1002.

(13) Schütt, K. T.; Sauceda, H. E.; Kindermans, P.-J.; Tkatchenko, A.; Müller, K.-R. SchNet−A deep learning architecture for molecules and materials. *J. Chem. Phys.* **2018**, *148*, 241722.

(14) Chmiela, S.; Tkatchenko, A.; Sauceda, H. E.; Poltavsky, I.; Schütt, K. T.; Müller, K.-R. Machine learning of accurate energy-conserving molecular force fields. *Science Advances* **2017**, *3* (5), e1603015.

(15) Noé, F.; Clementi, C. Kinetic Distance and Kinetic Maps from Molecular Dynamics Simulation. *J. Chem. Theory Comput.* **2015**, *11*, 5002−5011.

(16) Gastegger, M.; Behler, J.; Marquetand, P. Machine learning molecular dynamics for the simulation of infrared spectra. *Chemical Science* **2017**, *8*, 6924−6935.

(17) Bartók, A. P.; Gillan, M. J.; Manby, F. R.; Csányi, G. Machine-learning approach for one- and two-body corrections to density functional theory: Applications to molecular and condensed water. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2013**, *88*, 054104.

(18) Mardt, A.; Pasquali, L.; Wu, H.; Noé, F. VAMPnets for deep learning of molecular kinetics. *Nat. Commun.* **2018**, *9*, 5.

(19) Snyder, J. C.; Rupp, M.; Hansen, K.; Müller, K.-R.; Burke, K. Finding Density Functionals with Machine Learning. *Phys. Rev. Lett.* **2012**, *108*, 253002.

(20) Brockherde, F.; Vogt, L.; Li, L.; Tuckerman, M. E.; Burke, K.; Müller, K.-R. Bypassing the Kohn-Sham equations with machine learning. *Nat. Commun.* **2017**, *8*, 872.

(21) Collins, C. R.; Gordon, G. J.; von Lilienfeld, O. A.; Yaron, D. J. *Constant Size Molecular Descriptors For Use With Machine Learning*. 2017, arXiv preprint arXiv:1701.06649.

(22) 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.

(23) Montavon, G.; Hansen, K.; Fazli, S.; Rupp, M.; Biegler, F.; Ziehe, A.; Tkatchenko, A.; Lilienfeld, A. V.; Müller, K.-R. Learning Invariant Representations of Molecules for Atomization Energy Prediction. *Advances in Neural Information Processing Systems* **2012**, *25*, 440−448.

(24) Huo, H.; Rupp, M. *Unified Representation of Molecules and Crystals for Machine Learning*. 2017, arXiv preprint arXiv:1704.06439.

(25) Behler, J.; Parrinello, M. Generalized Neural-Network Representation of High-Dimensional Potential-Energy Surfaces. *Phys. Rev. Lett.* **2007**, *98*, 146401.

(26) Bartók, A. P.; De, S.; Poelking, C.; Bernstein, N.; Kermode, J. R.; Csányi, G.; Ceriotti, M. Machine learning unifies the modeling of materials and molecules. *Science Advances* **2017**, *3* (12), e1701816.

(27) 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.

(28) González-Díaz, H.; Herrera-Ibatá, D. M.; Duardo-Sánchez, A.; Munteanu, C. R.; Orbegozo-Medina, R. A.; Pazos, A. ANN Multiscale Model of Anti-HIV Drugs Activity vs AIDS Prevalence in the US at County Level Based on Information Indices of Molecular Graphs and Social Networks. *J. Chem. Inf. Model.* **2014**, *54*, 744−755.

(29) Romero-Durán, F. J.; Alonso, N.; Yañez, M.; Caamaño, O.; García-Mera, X.; González-Díaz, H. Brain-inspired cheminformatics of drug-target brain interactome, synthesis, and assay of TVP1022 derivatives. *Neuropharmacology* **2016**, *103*, 270−278.

(30) Schneider, G. Virtual screening: an endless staircase? *Nat. Rev. Drug Discovery* **2010**, *9*, 273.

(31) Todeschini, R.; Consonni, V. *Handbook of molecular descriptors*; John Wiley & Sons: 2008; Vol. *11*, DOI: 10.1002/9783527613106.

(32) Müller, K.-R.; Rätsch, G.; Sonnenburg, S.; Mika, S.; Grimm, M.; Heinrich, N. Classifying 'Drug-likeness' with Kernel-Based Learning Methods. *J. Chem. Inf. Model.* **2005**, *45*, 249−253.

(33) Faber, F. A.; Christensen, A. S.; Huang, B.; von Lilienfeld, O. A. Alchemical and structural distribution based representation for universal quantum machine learning. *J. Chem. Phys.* **2018**, *148*, 241717.

(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) Shapeev, A. V. Moment Tensor Potentials: A Class of Systematically Improvable Interatomic Potentials. *Multiscale Model. Simul.* **2016**, *14*, 1153−1173.

(36) Podryabinkin, E. V.; Shapeev, A. V. Active learning of linearly parametrized interatomic potentials. *Comput. Mater. Sci.* **2017**, *140*, 171−180.

DOI: 10.1021/acs.jctc.8b00110
J. Chem. Theory Comput. 2018, 14, 2991−3003

(37) Lubbers, N.; Smith, J. S.; Barros, K. *Hierarchical modeling of molecular energies using a deep neural network*. 2017, arXiv preprint arXiv:1710.00017.
(38) Gilmer, J.; Schoenholz, S. S.; Riley, P. F.; Vinyals, O.; Dahl, G. E. *Neural Message Passing for Quantum Chemistry*. 2017, arXiv preprint arXiv:1704.01212.
(39) Bereau, T.; DiStasio, R. A.; Tkatchenko, A.; von Lilienfeld, O. A. Non-covalent interactions across organic and biological subsets of chemical space: Physics-based potentials parametrized from machine learning. *J. Chem. Phys.* **2018**, *148*, 241706.
(40) Yao, K.; Herr, J. E.; Parkhill, J. The many-body expansion combined with neural networks. *J. Chem. Phys.* **2017**, *146*, 014106.
(41) Blum, L. C.; Reymond, J.-L. 970 Million Druglike Small Molecules for Virtual Screening in the Chemical Universe Database GDB-13. *J. Am. Chem. Soc.* **2009**, *131*, 8732−8733.
(42) Perdew, J. P.; Ernzerhof, M.; Burke, K. Rationale for mixing exact exchange with density functional approximations. *J. Chem. Phys.* **1996**, *105*, 9982−9985.
(43) Ernzerhof, M.; Scuseria, G. E. Assessment of the Perdew−Burke−Ernzerhof exchange-correlation functional. *J. Chem. Phys.* **1999**, *110*, 5029−5036.
(44) Ridley, J.; Zerner, M. An intermediate neglect of differential overlap technique for spectroscopy: Pyrrole and the azines. *Theoretica chimica acta* **1973**, *32*, 111−134.
(45) Bacon, A. D.; Zerner, M. C. An intermediate neglect of differential overlap theory for transition metal complexes: Fe, Co and Cu chlorides. *Theoretica chimica acta* **1979**, *53*, 21−54.
(46) Zerner, M. C. *Reviews in Computational Chemistry*; John Wiley & Sons, Inc.: 2007; pp 313−365, DOI: 10.1002/9780470125793.ch8.
(47) Tkatchenko, A.; DiStasio, R. A.; Car, R.; Scheffler, M. Accurate and Efficient Method for Many-Body van der Waals Interactions. *Phys. Rev. Lett.* **2012**, *108*, 236402.
(48) Hedin, L. New Method for Calculating the One-Particle Green's Function with Application to the Electron-Gas Problem. *Phys. Rev.* **1965**, *139*, A796−A823.
(49) Blum, V.; Gehrke, R.; Hanke, F.; Havu, P.; Havu, V.; Ren, X.; Reuter, K.; Scheffler, M. Ab initio molecular simulations with numeric atom-centered orbitals. *Comput. Phys. Commun.* **2009**, *180*, 2175−2196.
(50) Neese, F. The ORCA program system. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **2012**, *2*, 73−78.
(51) 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.
(52) 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.
(53) Müller, K. R.; Mika, S.; Rätsch, G.; Tsuda, K.; Schölkopf, B. An introduction to kernel-based learning algorithms. *IEEE Transactions on Neural Networks* **2001**, *12*, 181−201.
(54) Hoerl, A. E.; Kennard, R. W. Ridge Regression: Applications to Nonorthogonal Problems. *Technometrics* **1970**, *12*, 69−82.
(55) Altman, N. S. An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression. *Am. Stat.* **1992**, *46*, 175−185.
(56) Kohavi, R. *A Study of Cross-validation and Bootstrap for Accuracy Estimation and Model Selection*. Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 2. San Francisco, CA, USA, 1995; pp 1137−1143.
(57) Sugiyama, M.; Krauledat, M.; Müller, K.-R. Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research* **2007**, *8*, 985−1005.
(58) Hill, B. M. Bayesian Inference in Statistical Analysis. *Technometrics* **1974**, *16*, 478−479.
(59) Wold, S.; Sjöström, M.; Eriksson, L. PLS-regression: a basic tool of chemometrics. *Chemom. Intell. Lab. Syst.* **2001**, *58*, 109−130.
(60) Aitken, A. C. On Least Squares and Linear Combination of Observations. *Proc. R. Soc. Edinburgh* **1936**, *55*, 42−48.
(61) Tibshirani, R. Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society, Series B* **1996**, *58*, 267−288.
(62) Bennett, K. P.; Mangasarian, O. L. Robust linear programming discrimination of two linearly inseparable sets. *Optimization Methods and Software* **1992**, *1*, 23−34.
(63) Friedman, J. H. Greedy function approximation: A gradient boosting machine. *Ann. Statist.* **2001**, *29*, 1189−1232.
(64) Sonnenburg, S.; Zien, A.; Philips, P.; Rätsch, G. POIMs: positional oligomer importance matrices—understanding support vector machine-based signal detectors. *Bioinformatics* **2008**, *24*, i6−i14.
(65) Baehrens, D.; Schroeter, T.; Harmeling, S.; Kawanabe, M.; Hansen, K.; Müller, K.-R. How to Explain Individual Classification Decisions. *J. Mach. Learn. Res.* **2010**, *11*, 1803−1831.
(66) Bach, S.; Binder, A.; Montavon, G.; Klauschen, F.; Müller, K.-R.; Samek, W. On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. *PLoS One* **2015**, *10*, e0130140.
(67) Montavon, G.; Lapuschkin, S.; Binder, A.; Samek, W.; Müller, K.-R. Explaining nonlinear classification decisions with deep Taylor decomposition. *Pattern Recognition* **2017**, *65*, 211−222.
(68) Montavon, G.; Samek, W.; Müller, K.-R. Methods for interpreting and understanding deep neural networks. *Digital Signal Processing* **2018**, *73*, 1−15.
(69) Zeiler, M. D.; Fergus, R. *Visualizing and Understanding Convolutional Networks*. Computer Vision -- ECCV 2014. Cham, 2014; pp 818−833, DOI: 10.1007/978-3-319-10590-1_53.
(70) Simonyan, K.; Vedaldi, A.; Zisserman, A. *Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps*. 2013, arXiv preprint arXiv:1312.6034.
(71) Lenc, K.; Vedaldi, A. *Understanding image representations by measuring their equivariance and equivalence*. 2014, arXiv preprint arXiv:1411.5908.
(72) Zien, A.; Krämer, N.; Sonnenburg, S.; Rätsch, G. *The Feature Importance Ranking Measure*. Machine Learning and Knowledge Discovery in Databases. Berlin, Heidelberg, 2009; pp 694−709.
(73) Haufe, S.; Meinecke, F.; Görgen, K.; Dähne, S.; Haynes, J.-D.; Blankertz, B.; Bießmann, F. On the interpretation of weight vectors of linear models in multivariate neuroimaging. *NeuroImage* **2014**, *87*, 96−110.

DOI: 10.1021/acs.jctc.8b00110
J. Chem. Theory Comput. 2018, 14, 2991−3003