Generalized DFTB repulsive potentials from unsupervised machine learning

Journal:	*Journal of Chemical Theory and Computation*
Manuscript ID	ct-2017-00933j.R2
Manuscript Type:	Article
Date Submitted by the Author:	14-Mar-2018
Complete List of Authors:	Kranz, Julian ; Institute of Physical Chemistry, KIT, Kubillus, Maximilian; Karlsruhe Institute of Technology, Department of Chemistry and Biosciences Ramakrishnan, Raghunathan; Tata Institute of Fundamental Research, Centre for Interdisciplinary Sciences von Lilienfeld, O. Anatole; University of Basel, Chemistry Department Elstner, Marcus; Karlsruhe Institute of Technology, Physical Chemisrty

Generalized DFTB repulsive potentials from unsupervised machine learning

Julian J. Kranz,[†] Maximilian Kubillus,[†] Raghunathan Ramakrishnan,[‡,§] O. Anatole von Lilienfeld,[*,‡] and Marcus Elstner[*,¶]

[†]*Institute of Physical Chemistry, Karlsruhe Institute of Technology, Germany*

[‡]*Institute of Physical Chemistry and National Center for Computational Design and Discovery of Novel Materials (MARVEL), Department of Chemistry, University of Basel, Klingelbergstrasse 80, CH-4056 Basel, Switzerland*

[¶]*Institute of Physical Chemistry and Institute of Biological Interfaces (IBG-2), Karlsruhe Institute of Technology, Germany*

[§]*Present address: Tata Institute of Fundamental Research, Centre for Interdisciplinary Sciences, 21 Brundavan Colony, Narsingi, Hyderabad 500075, India*

E-mail: anatole.vonlilienfeld@unibas.ch; m.elstner@kit.edu

Abstract

We combine the approximate Density-Functional Tight-Binding (DFTB) method with unsupervised machine learning . This allows to improve transferability and accuracy, make use of large quantum chemical data sets for the parametrization, and to efficiently automatize the parametrization process of DFTB. For this purpose, generalized pair-potentials are introduced, where the chemical environmental is included during the learning process leading to more specific effective two-body potentials. We train on energies and forces of equilibrium and non-equilibrium structures of 2100 molecules, and test on $\sim$ 130.000 organic molecules containing O, N, C, H and F atoms. Atom-

ization energies of the reference method can be reproduced within an error of $\sim$ 2.6 kcal/mol, indicating drastic improvement over standard DFTB.

Introduction

Despite major advances in computer hardware, computational efficiency is still a major factor in the application of quantum chemical methods. Accurate Kohn-Sham Density Functional Theory (DFT) methods[1,2] face limitations when it comes to system size, number of molecules or molecular clusters to be evaluated, or sampling efficiency in Molecular Dynamics (MD) simulations.[3] Empirical force field (Molecular Mechanics: MM) methods, on the other hand, are extremely powerful tools in terms of computational efficiency and are quite accurate for the class of problems they have been parametrized for. But they are also limited by high parametrization effort, lack of parameter transferability, and explicit functional form restricting their flexibility to adapt to changing chemical environments. In terms of computational efficiency, accuracy and parameter transferability, semi-empirical (SE) methods are midway between ab initio/DFT and empirical force field methods. They are roughly three orders of magnitude faster than DFT methods using medium sized basis sets and roughly three orders of magnitude slower than empirical force field methods.

SE methods can be derived from Hartree-Fock (HF) or DFT, and recent years have shown major improvements in accuracy concerning covalent[4–6] and non-bonding interactions.[7] SE models approach the accuracy of DFT methods with medium sized basis sets, e.g. 'double-zeta plus polarization' (DZP) basis sets, which yield sufficient accuracy for many problems of interest. Despite many improvements in the last years, it seems that SE methods in general and the Density Functional Tight Binding (DFTB) method in particular have been converged in terms of accuracy due to their inherent inflexibility to reflect varying chemical situations with relatively fixed variants of SE-Hamiltonians. For example, heats of formations and reaction energies for standard test sets show errors of 3-7 kcal/mol, depending

on the focus of the parametrization strategy. DFTB already includes computed or experimental data in a fitting process, in the form of the so called repulsive potentials, which are one part of the DFTB total energy and can be based completely on empirical data or quantum chemical calculations. The other part of the total energy, the electronic part, can be completely computed from DFT traditionally using GGA functionals, but recently also long-range corrected functionals have been implemented, hence improving on typical DFT errors.[8–11]

With the availability of large amounts of reference data, data driven approaches become interesting alternatives to physical model potentials and approximate solutions of the Schrödinger equation. Lately, artificial intelligence techniques have become increasingly popular in molecular modeling, quantum chemistry, and condensed matter physics.[12–19] Several applications of machine learning techniques[14,15,18,19] and neural networks[16,17,20] to traditional quantum chemical problems show the great promise of this approach. A typical feature of data driven methods is its interpolative nature. Extrapolations beyond the data set are difficult, and convergence beyond a certain accuracy can be slow if poor choices are made among the many representations, similarity and regressor options.

Therefore, it has been suggested to combine the efficient SE methods with ML approaches,[21,22] since the former contributes important chemical information "easy" to capture, while the latter may improve on the accuracy by overcoming the limited flexibility of SE methods due to their inherent approximations, such as minimal basis sets, integral approximations, or use of atomic charges in the Hamilton. One possibility is to augment an SE method with a machine learning approach in the so called ΔML[21] method, correcting results based on a description of the entire molecule, or alternatively parameters of an SE Hamiltonian matrix[22] may trained. Both approaches lead to significant improvements in accuracy.

In this work, we combine the semi-empirical Density-Functional Tight-Binding (DFTB) method with ML to improve the prediction of thermochemical data and molecular struc-

tures. In contrast to other SE methods, the DFTB Hamilton matrix elements are computed in a 2-center approximation and are not derived by fitting to experimental or computed data. However, the repulsive potential, which is bond-specific rather than atom or molecule-specific, is fitted to reproduce molecular energies and structures. The repulsive potential is a natural target for a data driven expansion of DFTB. First of all because it already is an empirical term, so no existing rigor is taken away by moving to a different model for the fit. Second, because it is an inherently local, that is, spatially confined property, acting across bonds, and as such ought to be well described by a local model. Many current ML developments employ local representations in their molecular descriptors, to allow for transferability, for example to larger molecules than those included in the training set,[15,19,20,23] which is difficult if the descriptor depends on the molecular size. In that case, long-ranged interactions, such as between charges or molecular dipole moments, are not easily accounted for, but in a DFTB based approach those are treated at the level of the electronic terms, not the repulsive potentials. Hence, DFTB provides a platform for a model scalable to large molecules, because only local properties are fit, while including intermolecular and environmental interactions. Also, the set of potential bond topologies that may occur is much smaller than the set of all possible molecules. Working with repulsive potentials for bonds thus reduces the expected amount of required training data significantly. Finally, the repulsive potential model treats equilibrium and perturbed geometries on equal footing, providing a good foundation for the description of entire potential energy surfaces, eventually to include, as we hope, transition state structures. Since information about bond angles is already provided by the electronic part, only distance dependent information needs to be fit, and this also reduces the amount of training data required.

The machine learning trend in molecular modeling brings about ever larger sets of molecular data. So far the DFTB methodology did not benefit from this development, since parametrizations were created manually,[24] although progress has been made on automatizing the process.[25,26] Yet, since the number of free parameters is small, still only limited and

handpicked data can be included in the fit.

In this work, we therefore propose a generalization of the DFTB repulsive potentials, which will depend on a quantitative notion of the bond topology, rather than on atom types. Hence, we intend to overcome the limits imposed on DFTB by the small number of parameters, and tackle those errors that cannot be reduced anymore by physically motivated extensions of the method, and that are mostly of spatially localized nature.

The method is designed to require as little user input as possible. It is meant to scale to large training data sets, hence rendering the growing amount of available data useful for DFTB parametrization. Yet, we intend the method to also work with more limited amounts of training data, to be applicable in cases where the amount of data is insufficient for pure ML models to be feasible.

The paper is structured as follows: We first describe DFTB and its repulsive potentials briefly and then introduce the generalized repulsive potentials, for which we provide a proof of principle implementation. Then we analyze its performance. Finally, we discuss some technical details necessary for practical implementation and draw conclusions.

DFTB Background

DFTB

The DFTB methodology consists of a series of computational models, which are derived as an approximation to DFT. The total energy $E[\rho]$ is expanded at a reference electron density ρ_0, which is taken as the sum of contracted free atom densities. The expansion may be truncated at the first, second or third order and the corresponding models are known as DFTB1,[27] DFTB2[28] and DFTB3.[29] Introducing Kohn-Sham orbitals ϕ_i the energy functional expansion

in the DFTB2 case reads:

$$
\begin{aligned}
E[\rho] \approx \sum_i f_i \langle \phi_i | -\frac{1}{2}\nabla^2 + \int d^3\mathbf{r}' \, \frac{\rho_0(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|} + \frac{\delta E_{\mathrm{xc}}[\rho_0]}{\delta\rho(\mathbf{r})} + V_{\mathrm{ext}}(\mathbf{r}) | \phi_i \rangle \\
+ \frac{1}{2}\int d^3\mathbf{r} d^3\mathbf{r}' \left(\frac{1}{|\mathbf{r}-\mathbf{r}'|} + \frac{\delta^2 E_{\mathrm{xc}}[\rho_0]}{\delta\rho(\mathbf{r})\delta\rho(\mathbf{r}')} \right) \delta\rho(\mathbf{r})\delta\rho(\mathbf{r}') \\
+ E_{\mathrm{xc}}[\rho_0] + E_{\text{nuc-nuc}} - \frac{1}{2}\int d^3\mathbf{r} d^3\mathbf{r}' \, \frac{\rho_0(\mathbf{r})\rho_0(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|} - \int d^3\mathbf{r} \, \frac{\delta E_{\mathrm{xc}}[\rho_0]}{\delta\rho(\mathbf{r})}\rho_0(\mathbf{r}).
\end{aligned}
\tag{1}
$$

The f_i are orbital occupations, V_{ext} includes electron-nuclei and external field interaction, $E_{\text{nuc-nuc}}$ denotes the inter-nuclear interaction, and E_{xc} refers to the exchange-correlation functional, where DFTB employs the gradient-corrected PBE functional. The terms in the last line depend only on the reference density ρ_0 and the nuclear repulsion, and form the so-called repulsive potential V_{rep}. This is the focal point of our method and will be discussed in more detail in the next section. The linear and second order terms in eq. 1 in the first and second line are further approximated and expressed as:

$$
E^{(1)}_{\mathrm{DFTB}} = \sum_{ij}\sum_{\mu\nu} c_{\mu i} c_{\nu j} H^{(0)}_{\mu\nu} \tag{2}
$$

$$
E^{(2)}_{\mathrm{DFTB}} = \frac{1}{2}\sum_{A,B} \Delta q_A \Delta q_B \gamma_{AB}, \tag{3}
$$

where the Δq_A are the differences between the Mulliken charges of atom A and the corresponding neutral atom, and the $c_{\mu i}$ are the expansion coefficients of the Kohn-Sham orbital $\phi_i = \sum c_{\mu i}\chi_\mu$ in the basis $\{\chi_i\}$ that consists of a minimal basis of Slater-type orbitals confined to the valence shell. The zeroth order Hamiltonian

$$
H^{(0)}_{\mu\nu} = \langle \chi_\mu | -\frac{1}{2}\nabla^2 + \int d^3\mathbf{r}' \, \frac{\rho_0(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|} + \frac{\delta E_{\mathrm{xc}}[\rho_0]}{\delta\rho(\mathbf{r})} + V_{\mathrm{ext}}(\mathbf{r}) | \chi_\nu \rangle
$$

is pre-calculated in a two-center approximation, where ρ_0 is the sum of atomic densities around the atoms on which the basis functions χ_μ and χ_ν are centered. At second order, the shape of the local density around atom A is assumed to be well described by a spherical

function Φ_A, so that the Coulomb integrals $\gamma_{AB} = \int d^3\mathbf{r} d^3\mathbf{r}' \left(\frac{1}{|\mathbf{r}-\mathbf{r}'|} + \frac{\delta^2 E_{\mathrm{xc}}[\rho_0]}{\delta(\mathbf{r})\delta\mathbf{r}'}\right) \Phi_A(\mathbf{r})\Phi_B(\mathbf{r}')$ can be evaluated analytically. In the DFTB3 model, resulting from a third order expansion, eq. 1 is augmented by an extra term as follows:

$$E^{(3)}_{\mathrm{DFTB}} = \frac{1}{3}\sum_{AB} \Delta q_A^2 \Delta q_B \Gamma_{AB} \tag{4}$$

The off-diagonal terms Γ_{AB} are analytic representations of third order integrals and the diagonal terms can be calculated as atomic hardness derivatives.

Repulsive Potential

The last row of eq. 1 contains the core-core repulsion and those energy contributions, that depend on the reference density ρ_0 only. They are grouped together into a single term called the repulsive potential V_{rep}, which is approximated as a sum of two-center repulsions,

$$V_{\mathrm{rep}} = \frac{1}{2}\sum_{A,B} V_{AB}(|\mathbf{R}_A - \mathbf{R}_B|). \tag{5}$$

These two-body potentials are fitted to the difference of the total energy of a reference calculation and the DFTB electronic energy,

$$V_{\mathrm{rep}}(|\mathbf{R}_A - \mathbf{R}_B|) = E_{\mathrm{ref}}(|\mathbf{R}_A - \mathbf{R}_B|) - E_{\mathrm{DFTB}}(|\mathbf{R}_A - \mathbf{R}_B|),$$

with the energy contributions eqns. 2, 3 and 4,

$$E_{\mathrm{DFTB}} = \sum_i E^{(i)}_{\mathrm{DFTB}}.$$

The DFTB pairwise potentials V_{AB} depend only on the atom types of A and B, in contrast e.g. to force field models, where different bonding environments are encoded by different bonding parameters. Since the terms from eq. 1 that are grouped into the repulsive

potential contribution depend on the reference density only, the adaption to different bonding situations is, in principle, governed by the DFTB electronic energy contributions.

Atomization and reaction energies can form part of the reference energies, as well as forces, in particular at equilibrium structures.[6] In previous work, the repulsive potential contributions have been fitted to minimize the errors for atomization energies, geometries and vibrational frequencies for the G2/97 reference set.

However, since the electronic terms are subjected to approximations as well – using a minimal basis set, applying a monopole approximation, neglecting three-center contributions, to name only the most prominent ones – the transferability is limited in practice. This becomes apparent, for example, in an optimization conflict for atomization energies and vibrational frequencies. To reach a reasonable accuracy for both properties, two distinct parametrizations had to be generated[6] because of the limited transferability of the parameters between different hybridization states, that is, single, double and triple bonds. This is due to a number of reasons: (i) For good vibrational frequencies, the repulsive potentials need to have certain curvatures at the equilibrium distances, for atomization energies certain absolute values are needed, and these two conditions can not be fulfilled simultaneously when a certain accuracy is targeted. (ii) Further, a different degree of over-binding is found for single, double and triple bonds, leading to a relative shift of the potentials between the binding regimes, which is not possible to integrate into a single repulsive potential function. (iii) Finally, the repulsive potentials have to vanish before the second neighbor distances in order to avoid spurious long-range effects, which put additional restraints on the optimization for the bonding properties.

Therefore, entangling theses different issues in a more adaptive repulsive energy scheme should lead to an overall improvement in accuracy.

Method description

Generalized repulsive potentials

In the standard DFTB approach only one repulsive potential

$$V_{AB} = V_{t(A)t(B)},$$

is used to connect two atoms A and B of certain atom types, denoted by $t(A)$ and $t(B)$.

In contrast, we now introduce a variable number of different potentials

$$V_{AB}(R) = V_{t(A)t(B)}(R) + \Delta V_{b(A,B)}(R), \tag{6}$$

called generalized repulsive potentials, which depend also on the bond type $b(A, B)$, to be defined later. They are generated automatically and in a scalable way and augment the element pair repulsive potential $V_{t(A)t(B)}(R)$, which comes from any existing DFTB parametrization. In this work, we use the repulsive parameters from 3OB,[6,30] while $\Delta V_{b(A,B)}(R)$ is a correction to this potential that can incorporate environment-specific information not grasped by the electronic parts of DFTB. By fitting corrections, rather than entirely new potentials, the existing potentials continue to serve as a fall-back for very unusual bonds, while for known bonds the correction term will improve the description. As $b(A, B)$ will denote bonds much more specific than the element pair $t(A)t(B)$, in practical applications there is a chance to encounter bond topologies for which no specific repulsive potential has been fitted yet. For example, it is possible to assign different repulsive potentials to different bond types (e.g. single, double, triple), but also to be more specific and distinguish various chemical environments for each type. A carbon-carbon single bond may be subject to change when, for instance, neighboring electronegative atoms withdraw electrons, compared to the situation in pure hydrocarbons.

For the functional form of $\Delta V(R)$ we chose to use polynomials of degree k:

$$\Delta V_b(R) = \sum_{i=0}^{k} a_i^{(b)} R^i. \tag{7}$$

Other forms, such as splines, are possible as well, but at present we find simple polynomials to be sufficient. Note that if the forms of $V_{t(A)t(B)}$ and $\Delta V_{b(A,B)}$ agree, to linearly fit a correction potential ΔV_b is equivalent to fitting parameter corrections $\Delta a_i^{(b)}$. This holds for polynomials, splines, and other models linear in the parameters. Repulsive potentials should be short ranged and therefore tend to zero at large distances. To impose such asymptotic behavior, a cut-off R_c can be introduced at which $\Delta V_b(R)$ is smoothly set to zero. At present, we only run tests on geometries near equilibrium. Therefore, the asymptotic behavior is not relevant in this context.

Eventually, the full generalized repulsive potential for a given molecular geometry reads

$$V_{\text{rep}} = \frac{1}{2} \sum_{AB} \left(V_{t(A)t(B)}(R_{AB}) + \Delta V_{b(A,B)}(R_{AB}) \right), \tag{8}$$

$R_{AB} = |\mathbf{R}_A - \mathbf{R}_B|$ is the distance between atom A and B and $b(A, B)$ adds corrections for a set of bond types much larger than in traditional DFTB.

The reason we opt to merely generalize the repulsive potentials, rather than to build a direct ML model for them, is that we find the required amount of data for the approach to work to be much lower than for a pure ML model. To construct a model beyond equilibrium geometries, scans of the potential energy surface are required, but if bonds reoccur with different lengths in various molecules, this information can in part be taken from different molecules to reduce the number of data points required for individual molecules. Moreover, the repulsive potentials only need to entail distance dependent information, with angular information coming from the electronic contributions, further reducing data requirements. For example the ANI-1 neural network based model[20] is remarkably successful as a direct ML model for the potential energy surfaces of isolated organic molecules, and parametrized

from a very similar training set, but in training uses vastly more extensive scans. The underlying data set has recently been made publicly available,[31] and in future work it will be insightful to also explore a combination of such a model with DFTB as a direct ML model for the repulsive potentials. Yet, we believe models with less data requirements will remain useful, at least in the foreseeable future. The ANI-1 training data, like most comparable data sets, is based on double-zeta quality DFT calculations that often lack the required accuracy, but are used because the computational cost associated with larger basis sets, or even higher level methods, is very significant. However, replacing a training set of the size used in this work to parametrize our method is absolutely feasible. We further intend to apply the method to systems such as transition metal complexes, where suitable training data is harder to produce.

Bond descriptor

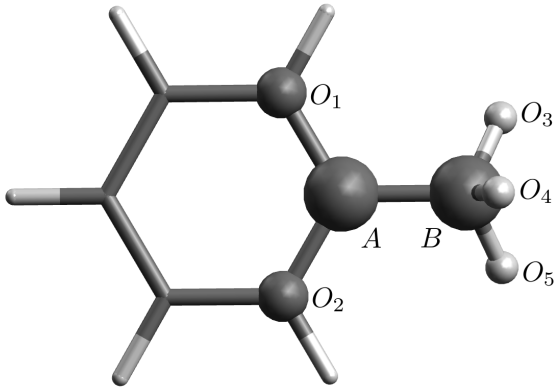

Figure 1: Example of a potential bond descriptor. The two large carbon atoms form the bond, all atoms displayed as balls are included in the descriptor. Atoms are labeled as they appear in the descriptor.

To define $b(A, B)$, bond descriptors have to be introduced, which allow the recognition of certain bonds in molecular structures. This information is basically encoded in the geometrical arrangement of atoms in the immediate vicinity of the bond. For instance, single and double bonded atoms will have a different number of nearest neighbor contacts, determining their hybridization state. The atom type of the neighbors can indicate certain properties

of the local electronic structure (like local electron density). Lately, machine learning techniques for the prediction of molecular properties have become popular and accordingly much research on molecular descriptors has been undertaken, providing a wealth options of varying sophistication, see e.g. Refs. 32,33. Here, we started out with the rather simple Coulomb matrix descriptor,[12] which turns out to work satisfactorily for the purpose of this work. Many known descriptors perform much better in ML models than the Coulomb matrix, so this choice requires some justification. In common ML models of molecular properties,[34] in particular regression models, the descriptor appears directly in the mathematical expressions used for property predictions, and hence its precise form can have a strong impact on the quality of the model. However, in the clustering step for bond identification that will later follow, most of the descriptor's nuanced behavior is lost, so that the effect of the descriptor on the present application is much weaker, and good performance in ML models needs not transfer directly to this application. The dependence on molecular size is also not a problem here because the number of atoms that can appear near a bond in physically meaningful systems is limited, so that bond identification is effectively a constant size problem.

The bond geometry is represented by a matrix with diagonal terms identifying and atom type and off-diagonal terms are given by the nuclear Coulomb repulsion of the respective atom pairs. The atoms are ordered unambiguously and the descriptor respects all the important symmetries like translational and rotational invariance of the bond.

The bond descriptor requires two parameters and is defined as follows: Two atoms A and B are considered bonded for the purpose of the repulsive potentials if their distance $R_{AB} = |\mathbf{R}_A - \mathbf{R}_B|$ is smaller than an element dependent cut-off $R_{AB} < R^c_{t(A)t(B)}$. A second parameter R^c_b defines a volume within which all atoms O are included in the description of the chemical environment of the bond between A and B (see Fig. 1). Specifically, an atom O is included in the descriptor if

$$\min_{C=A,B} |\mathbf{R}_O - \mathbf{R}_C| < R^c_b. \tag{9}$$

Then the bond descriptor $b(A,B)$ is defined as

$$b(A,B) = \begin{pmatrix} \frac{1}{2}\eta Z_A^{2.4} & \frac{Z_A Z_B}{|\mathbf{R}_A - \mathbf{R}_B|} & \frac{Z_A Z_{O_1}}{|\mathbf{R}_A - \mathbf{R}_{O_1}|} & \cdots \\ \frac{Z_A Z_B}{|\mathbf{R}_A - \mathbf{R}_B|} & \frac{1}{2}\eta Z_B^{2.4} & \frac{Z_B Z_{O_1}}{|\mathbf{R}_B - \mathbf{R}_{O_1}|} & \cdots \\ \frac{Z_{O_1} Z_A}{|\mathbf{R}_{O_1} - \mathbf{R}_A|} & \frac{Z_{O_1} Z_B}{|\mathbf{R}_{O_1} - \mathbf{R}_B|} & \frac{1}{2} Z_{O_1}^{2.4} & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}. \tag{10}$$

Attention has to be paid to the order of the atoms to make the descriptor unambiguous. The two atoms of the bond, A and B come first and their order is determined by their norm, that is

$$\left(\frac{1}{2}\eta Z_A^{2.4}\right)^2 + \sum_{C \neq A} \left(\frac{Z_A Z_C}{|\mathbf{R}_A - \mathbf{R}_C|}\right)^2 \geq \left(\frac{1}{2}\eta Z_B^{2.4}\right)^2 + \sum_{C \neq B} \left(\frac{Z_B Z_C}{|\mathbf{R}_B - \mathbf{R}_C|}\right)^2,$$

where C runs over all atoms in the descriptor. Likewise, the other atoms $O_1, O_2, ...$, following A, B are ordered according to the norms of their rows.

$b(A,B)$ basically constitutes the Coulomb matrix of the bond environment, except for the special role of the first two rows that always contain the information about the bonded atoms and the extra factor η that scales the diagonal entries for atoms A and B. η should be larger than one to give particular weight to the atom types of the bonded atoms and to ensure that bonds involving different elements are always further apart than bonds involving the same elements in the space of bonds spanned by the bond descriptors. There is no sensitive dependence on the precise value of η.

Finally, we need to define a notion of distance between two bonds $b^{(1)}$ and $b^{(2)}$. We use the 2-norm

$$d(b^{(1)}, b^{(2)}) = \sqrt{\sum_{ij} \left(b_{ij}^{(1)} - b_{ij}^{(2)}\right)^2}, \tag{11}$$

where b_{ij} are the entries of the descriptor matrix. The 2-norm provides the practical advantage that certain libraries can be used directly in the implementation of the method, which do not support the 1-norm that is often used for estimating similarity of structures when

using the Coulomb matrix.[12] However, there is no significant difference between the two choices for our application. All Coulomb matrices need to have the same dimension, and for chemical environments characterized by a smaller number of participating atom the matrix is filled with zeros. The zeros may be thought of as atoms infinitely removed from the bond.

The parameter R_b^c critically affects the specificity of the bond descriptor. If it is smaller than the shortest typical bond length, only the two bonded atoms will be included, and nothing is gained over the existing DFTB repulsive potentials since no information about the environment enters the description. Using values larger than typical bond lengths, the nearest neighbors of the bonding atoms will be included, which is the minimal representation of the chemical environment and already leads to very good results. Further increasing the magnitude of the parameter, non-local information can be included as well. Hence, the method can take in ever more information as the amount of training data grows. This is a desired feature of the approach. In principle, one can go up to the limit where the entire molecule forms the descriptor and one has a molecule specific fit. Such descriptors are used for example in the Δ machine-learning approach.[21] Of course, with increasing specificity more and more training data is required.

Bond clustering

The next step is the automatic identification of bonds from data, that is, the clustering of relevant bond-types from a large training set of molecular structures according to the the descriptor defined above. Every cluster, or bond type, will define one generalized repulsive potential. In each of the molecular structures contained in the training set, bonds and their respective environments are identified according to the two cut-off criteria, and the Coulomb-matrices are then set up. Similar bonds yield very similar descriptors, although they are not exactly identical due to slightly varying interatomic distances. Hence, bonds form clusters of a finite, but narrow width in the high dimensional feature space spanned by the bond descriptors, where different clusters correspond to different bond types. The

dimension of the feature space is determined by the size of the largest Coulomb matrix and depends on the value of R_b^c. When R_b^c is such that nearest neighbors are included in the descriptor, the dimension is bounded from above by about $8^2 = 64$ because there are 8 atoms in the descriptor of a C-C single bond, and it is not chemically possible to gather many more atoms in such a small volume. Recall that this ensures we deal with a problem of fixed dimensionality, and the size dependence of the Coulomb matrix descriptor is therefore of no concern.

Identifying bond types now becomes a clustering problem. Such problems are commonplace in unsupervised machine learning,[35] and many methods have been proposed for their solution. The choice of an appropriate algorithm, however, turned out to be not completely straight forward. In particular, the highly unbalanced number of data points in different clusters was problematic. Since some bond types are far more abundant than others, the clustering algorithm has to be insensitive to the number of cluster members. For example, the popular k-means[36,37] algorithm was found to be unsuitable for this reason, as our training set contained, among others, far more, C-H than C-F bonds and k-means would only produce a large amount of C-H, but no C-F cluster.

The mean-shift algorithm[38,39] that was originally developed for image processing applications, in contrast, turned out to work very well. Here, we give only brief description of the algorithm; more detail can be found in original literature.[38,39] Let

$$m(b) = \frac{\sum_i b_i K\left(\frac{d(b,b_i)}{h}\right)}{\sum_i K\left(\frac{d(b,b_i)}{h}\right)} \tag{12}$$

be the mean-shift vector at position b. b is a general bond descriptor, the b_i are the bond descriptors in the set to be clustered, the real number $h > 0$ is the kernel width, and the function K is the kernel function. K can be any positive function integrating to one, but for

the purpose of this work we adopt a flat kernel:

$$K(x) = \begin{cases} 1 \text{ for } 0 \leq x \leq 1 \\ 0 \text{ otherwise} \end{cases} . \tag{13}$$

The flat kernel is a simple choice and performs well for our application. With the mean-shift vector clusters are now identified through the iteration

$$b^{(t+1)} = m(b^{(t)}),\ t = 0, 1, ... \tag{14}$$

that will converge to a value b^*, the centroid of a cluster. One scans sufficiently many initial values $b^{(0)}$ to find all centroids and thus all clusters.

The algorithm in this form appears rather abstract, but it has an intuitive interpretation. If we assume the data points b_i to be samples from a continuous density $\rho(b)$, then with a smoothing kernel function K the smooth density can be approximated as $\rho(b) \approx \frac{1}{Nh^d} \sum_i K\left(\frac{d(b,b_i)}{h}\right)$, where N is the number of samples and the exponent d is the dimension. It can be shown that $\nabla\rho(b) \propto m(b)$. Consequently, the mean-shift algorithm can be thought of as a steepest-descent optimization to find the local maxima of ρ, and clusters can be identified with blobs in the density. Although the flat kernel is not a smooth one, it can be approximated arbitrarily well by a smooth kernel and, therefore, the same argument still holds.

The mean-shift algorithm is available as part of the scikit-learn Python module,[40] which we employ in our implementation of the method.

The kernel width parameter h is of crucial importance, since it sets the minimal distance of two points at which they are still regarded as members of the same cluster. Therefore, it also determines how many different bond types will be identified. Because the number of clusters M is more intuitive and tangible, we will classify the resulting repulsive potentials by M, rather than h. Yet, even though we refer to M for clarity, h remains the fundamental

variable.

M can be increased as the amount of training data grows. Hence, the method can scale to large training sets by fitting many different potentials. In the limit of very large sets, every individual bond could be represented by its own fitted potential.

At last, after bond types have been identified through clustering, any new bond b not in the training set must be assigned the cluster or bond type it belongs to. Alternatively, it may also happen that no existing bond type describes b well. This case must be recognized, too. A mean-shift iteration started from b will converge to the centroid b^* of the cluster best describing b, thus identifying the bond type. For simplicity's sake, we assume b^* to be the centroid closest to b according to the metric $d(b, b^*)$, an assumption we found well justified. Then, b^* can be identified simply as

$$b^* = \underset{\tilde{b} \in \text{centroids}}{\arg\min}\ d(b, \tilde{b}).$$

To rule out the cases where b^* does not describe b well, we demand that the centroid and bond are closer to each other than a certain tolerance distance w^*:

$$d(b, b^*) < w^*.$$

A well chosen value of w^* depends on whether the cluster of b^* is narrow or wide. Therefore, we calculate the cluster width σ^* from the training data:

$$\sigma^* = \sqrt{\sum_{\tilde{b} \text{ belongs to } b^*} d(\tilde{b}, b^*)^2}.$$

The sum runs over all bond descriptors belonging to centroid b^* in the training set. Hence, we set

$$w^* = \tau \cdot \sigma^*,$$

with a tolerance factor τ to be chosen manually.

Potential fit

In the next step, the repulsive potentials $\Delta V_b(R)$ for all new bond types b are fitted in the same way as the standard DFTB repulsive parameters. The $a_i^{(b)}$ in eq. 7 are determined such as to minimize a fitness function $f(a_0^{(1)}, a_1^{(1)}, ..., a_0^{(2)}, ...)$, which contains molecular atomization energies E_{at} and forces $\mathbf{F}$ for a set of equilibrium and perturbed molecular geometries as target properties:[1]

$$\begin{aligned} f(a_0^{(1)}, a_1^{(1)}, ..., a_0^{(2)}, ...) = & \sum_{m \in \mathrm{equi}} \left(E_{\mathrm{at},m}^{\mathrm{ref}} - E_{\mathrm{at},m}^{\mathrm{DFTB}} - \sum_{b \in m} \Delta V_b(R_b) \right)^2 \\ & + f_{\mathrm{opt}} \sum_{m \in equi} \frac{1}{3N_{\mathrm{at,m}}} \left(\mathbf{F}_{\mathrm{at},m}^{\mathrm{ref}} - \mathbf{F}_{\mathrm{at},m}^{\mathrm{DFTB}} + \sum_{b \in m} \frac{\mathbf{R}_b}{R_b} \frac{\partial}{\partial R} \Delta V_b(R_b) \right)^2 \\ & + e_{\mathrm{pert}} \sum_{m \in \mathrm{pert}} \left(E_{\mathrm{at},m}^{\mathrm{ref}} - E_{\mathrm{at},m}^{\mathrm{DFTB}} - \sum_{b \in m} \Delta V_b(R_b) \right)^2 \\ & + f_{\mathrm{pert}} \sum_{m \in \mathrm{opt}} \frac{1}{3N_{\mathrm{at,m}}} \left(\mathbf{F}_{\mathrm{at},m}^{\mathrm{ref}} - \mathbf{F}_{\mathrm{at},m}^{\mathrm{DFTB}} + \sum_{b \in m} \frac{\mathbf{R}_b}{R_b} \frac{\partial}{\partial R} \Delta V_b(R_b) \right)^2 . \end{aligned} \tag{15}$$

Perturbed geometries are created from equilibrium geometries through displacement along normal mode coordinates. The fitness function is generated by summation of all equilibrium and perturbed molecular geometries, and by computation of the energy and force contributions resulting from the repulsive potentials, which sum over all bond types b. The potentials can be written as

$$\Delta V_b(R_b) = \left(1 \, R_b \, R_b^2 \, ...\right) \cdot \left(a_0^{(b)} \, a_1^{(b)} \, ...\right)^T ,$$

[1]The atomization energy is the difference between the molecular energy and the sum of free atom energies of the atoms constituting the molecule. To the DFTB free atom energies spin polarization terms are added to account for the lack of direct spin polarization.[6]

and therefore the optimization procedure is a least squares problem of the form $\min_{\mathbf{x}} |\mathbf{y} - A\mathbf{x}|^2$, with given vector $\mathbf{y}$ and matrix A, where the parameter vector $\mathbf{x}$ has to be determined. Many tools exist to solve this problem. We use the Numpy[41] least-squares function that utilizes a singular value decomposition.

The weight factor $\frac{1}{3N_{\mathrm{at},m}}$ is added since for each geometry only one energy but $3N_{\mathrm{at},m}$ force conditions have to be fulfilled. In that way energies and forces are given the same initial weight. The additional weight factors $f_{\mathrm{opt}}, e_{\mathrm{pert}}$, and f_{pert} control the relative weights of energies and forces for equilibrium and perturbed geometries, respectively. They must be set manually.

Application

The approach is applied to a large set of molecules, which features structures and energies computed at the B3LYP/6-31G(2df,p) level of theory. For a final reparametrization a higher level of theory is desirable. Therefore, we use these data for a proof of concept approach in order to evaluate the procedure suggested in this work.

Data set

To test the method we use a molecular data set created by Ramakrishnan et al.,[42] which provides optimized structures and properties for small organic molecules from the GDB-17[43] set, containing the elements C, H, O, N, and F with up to nine non-hydrogen atoms. The set contains 133,885 molecules, geometry optimized at the B3LYP/6-31G(2df,p) level of theory.[44–46]

The set of molecules is separated into a training set composed of the first 2100 molecules and a test set containing the rest. The training set is supplemented with non-equilibrium geometries generated as follows: Starting from relaxed geometries, all coordinates are displaced in both directions of all those normal modes which affect bond lengths. The amount

of displacement is chosen such that the energies vary only on the order of 1 $\frac{\text{kcal}}{\text{mol}}$ with respect to the equilibrium energy. For the distorted structures, energies and forces are computed using B3LYP/6-31G(2df,p), to be consistent with the other reference data. Eventually the training set contains about 150000 relaxed and unrelaxed structures in total. For every molecule in the test set we run DFTB calculations with 3OB parameters and the full third order formalism[6,29] that yield the DFTB base line of electronic and repulsive potential contributions.

A training set of 2100 different molecules is small by ML standards, where often at least several ten thousands to a hundred thousand different molecules are used for training. The training set size for now is limited by the need to perform potential energy surface scans for each molecule and the computational resources available to us. As more large data sets involving perturbed geometries will become publicly available in the future, the underlying data can be improved. By the standards of DFTB parametrization 2100 molecules is a very large number of training molecules, orders of magnitudes more than what is, and can be, used for conventional repulsive potential parametrization.

We supplement the test set with some external molecules to benchmark transferability that will be discussed when results are presented.

Clustering and fit

The cut-off parameters used for the 3OB repulsive potentials[6] are chosen as the cut-off parameters $R^c_{t(A)t(B)}$ that determine atom pairs connected by generalized repulsive potentials. The cut-off parameter R^c_b that determines which atoms enter into the Coulomb-matrix descriptor is set to $R^c_b = 1.8\,\text{Å}$. For the molecular structures considered in the present work this includes nearest neighbors of the bonded atoms. The parameter η, which gives special weight to the bonded atoms, is chosen as $\eta = 5$ and tests showed that the results are not very sensitive to the precise value of η, once $\eta > 2$. Lastly, the tolerance factor τ is put to $\tau = 3$, and again results are not very sensitive to the precise value of the parameter

within reasonable bounds. Mean-shift clustering is performed on the set of all bonds found in the equilibrium geometries of the training set molecules for these parameter values. As a precaution to prevent over-fitting, we drop bond types that are not present in at least three different molecules. Especially when two bonds occur only once and in the same molecule, their potentials can cancel each other and therefore assume arbitrary values. To solve this problem, one can simply generate more data from geometry perturbations for bonds that do not occur often enough, but we refrained from doing so at the present exploratory stage. Hence, sets with $M = 10, 25, 42, 84, 111, 200$, and 259 different bond types are created. Tab. 1 gives values of h for each M for reproducibility. M does not vary continuously with h, but tends to jump, so that there is no exact one-to-one correspondence.

Table 1: Values for the Kernel width h resulting in certain numbers of potentials M. Most widths h were in terms calculated as the qth percentile of pairwise distances of data points, explaining their odd values. The percentiles q are then also given.

M	10	25	42	84	111	200	259
h	93.26	75.525	58.627	43.402	36.089	27.130	23.4
q	15%	9%	5%	3%	2%	1.3%	-

Varying numbers of generalized repulsive potentials allow us to investigate whether the performance of the method indeed improves with growing numbers, and at what point performance saturates.

For each set of bond types, we fit generalized repulsive potentials $\Delta V_b(R)$ as polynomials of degree $k = 6$, that is, with 7 free parameters. The weight of equilibrium forces f_{opt} is set to $f_{\text{opt}} = 100$, non-equilibrium force weight is put to $f_{\text{pert}} = 1$, and non-equilibrium atomization energies carry weight $e_{\text{pert}} = 1$. We found those values by trial and error. They have not been properly optimized yet.

Results

At first, we investigate the clustering step that is pivotal for the method, which stands and falls with the meaningful identification of bond types as a foundation for the generalized

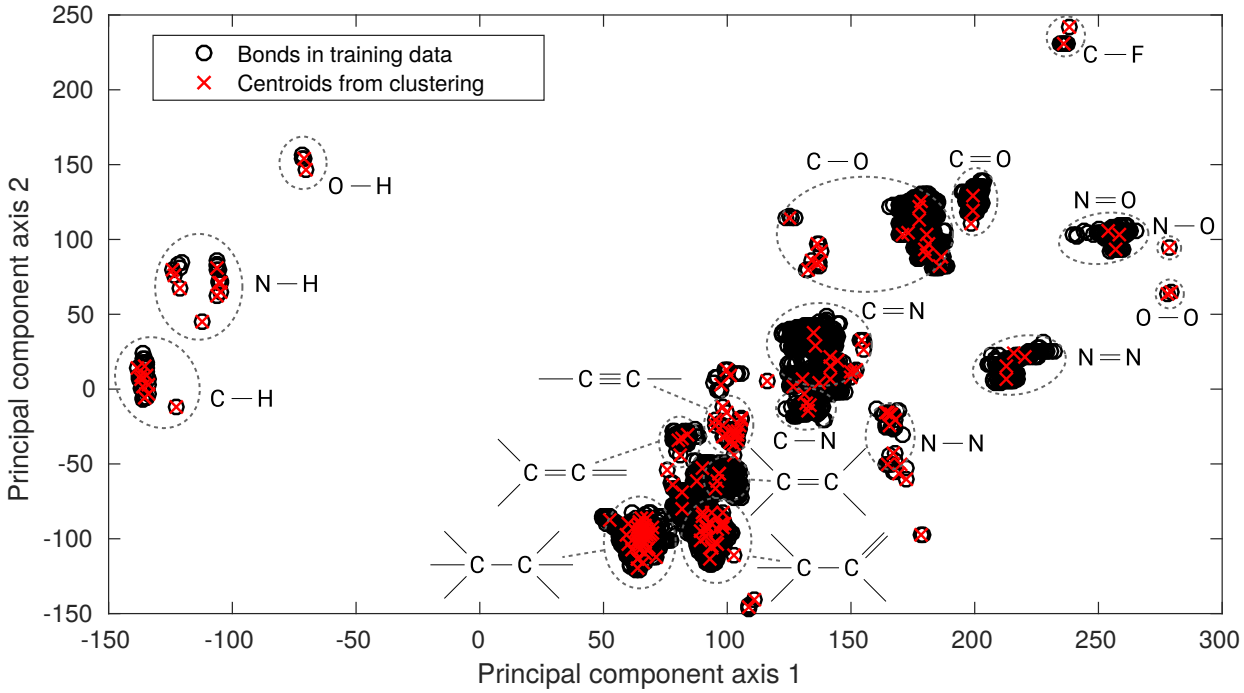

Figure 2: Visualization of the bond descriptors from the training set and $M = 200$ centroids from mean-shift clustering in two dimensions. The high dimensional descriptor space has been projected to two dimensions by a principal component analysis.[47,48] Note that the large clusters decompose into many smaller clusters corresponding to different chemical environments. The mean-shift algorithm covers all clusters well regardless of the number of data points in them.

repulsive potentials. Fig. 2 shows two dimensional projections of all the bond descriptors in the training set, generated by a principal component analysis[47,48] (PCA). PCA identifies the two dimensional subspace in which the variance of the data is maximal. Various clusters of different sizes appear already in two dimensions that in higher dimensions decompose in terms into more, smaller clusters. Clusters correspond to different kinds of bonds, many of which are indicated in the figure. Also displayed in Fig. 2 are the centroids of the clustering with $M = 200$ clusters. For every centroid there is one generalized repulsive potential. We find that all clusters are covered with centroids, and the large clusters carry many centroids because of the many smaller clusters they contain, but small, isolated clusters are covered too. This is important, and, for example, k-means failed to achieve this. Altogether, Fig. 2 shows a reasonable clustering according to bond topology and supports that the Coulomb matrix based bond descriptor, together with mean-shift clustering, can indeed identify meaningful bond types.

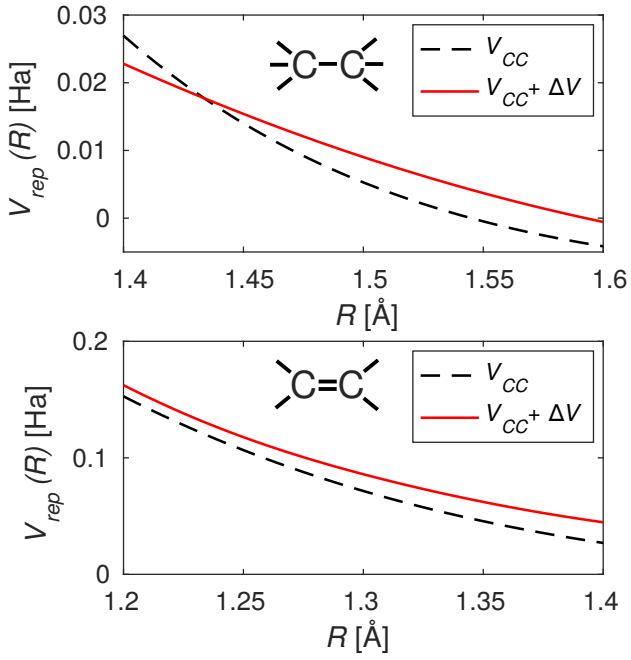

Figure 3: Sample repulsive potentials for a C-C single and double bond with $M = 111$ generalized repulsive potentials. Magnitude, slope and curvature are altered by the potential corrections.

Next, we examine the fitted potentials ΔV. Higher order polynomials can oscillate sig-

nificantly in the case of over-fitting. By visual inspection the potentials are confirmed to be well behaved. Fig. 3 shows two representative sample potentials for a single and double C-C bond. Position, slope and curvature of the repulsive potentials are altered, but the functions remain monotonously decaying without spurious behavior. Of course, because no boundary conditions were applied, this is only true for distances sufficiently close to the respective bond lengths, but in the present study we only work with such geometries. It is also inter-

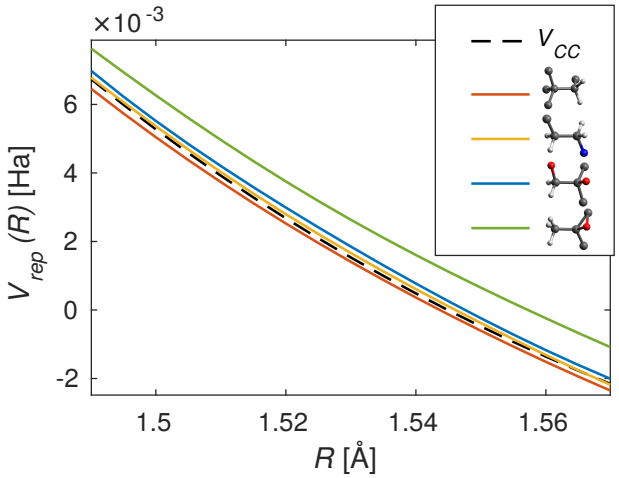

Figure 4: Generalized repulsive potentials for different types of C-C single bonds. The atoms that make up the descriptor are displayed in the legend. Potentials differ, particularly by a vertical displacement, with slopes more similar.

esting to see how different potentials for the same bond topology, such as, for example, C-C single bonds, can occur. Fig. 4 shows various potentials for C-C single bonds. While most of them remain close to the uncorrected potential V_{CC}, there is a considerable variation for some of them. They appear to be up- and down-shifted, while slopes remain similar. There is a tendency that for chemical environments with more electronegative substituents on the bonded atoms, the potentials are shifted upwards.

Table 2: Mean absolute (MAE) and root mean squared (RMSE) error in atomization energy taken over all test set molecules.

M	0	10	25	42	84	111	200	259
MAE [kcal/mol]	7.38	7.34	4.32	3.71	3.01	2.97	2.89	2.64
RMSE [kcal/mol]	9.31	9.74	5.65	5.46	3.94	3.95	4.00	3.82

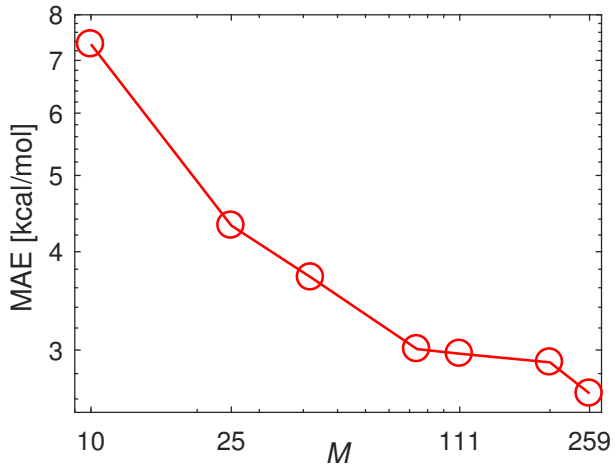

Figure 5: Mean absolute error in atomization energy for the molecules in the test set for a growing number M of generalized repulsive potentials. As the number of potentials grows, the error decreases quickly and significantly, demonstrating that increasing the number of repulsive potentials indeed improves the method.

Ultimately, the method's usefulness is determined by its quantitative performance. For a benchmark, we applied it to all of the $\sim 130,000$ molecules in the test set. Fig. 5 shows the mean absolute errors (MAEs) in atomization energy with varying numbers of generalized repulsive potentials; Tab. 2 displays MAEs and root mean squared errors (RMSE). The MAEs monotonously decay as the number of potentials grows. The first step is very small, then the error decreases rapidly. The improvement brought about by the addition of generalized potentials is clear. With $M = 259$ generalized repulsive potentials the remaining error is $\Delta E_{259} = 2.64\,\frac{\text{kcal}}{\text{mol}}$, down from $\Delta E_0 = 7.34\,\frac{\text{kcal}}{\text{mol}}$ with the original 3OB repulsive potentials. Therefore, the error is significantly reduced to about a third of the original error. Figs. 6 and 7 show histograms of absolute errors in atomization energies and force per atom, respectively. The distribution of errors in the atomization energy narrows significantly, demonstrating a systematic improvement, already reflected in the lower MAEs. Forces improve too: initially the histogram shows two peaks, a large and a much smaller one at a higher error. The second peak indicates a small, systematic error in the predicted geometries, and this is removed after the addition of generalized repulsive potentials. However, the large peak is hardly moved. DFTB already predicts geometries well, and most of

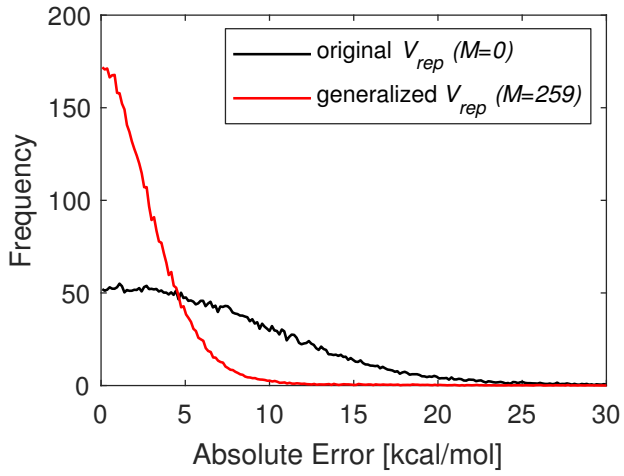

Figure 6: Normalized histogram of absolute errors in atomization energy for the $\sim$ 130000 molecules in the test set with $M = 259$ generalized repulsive potentials.

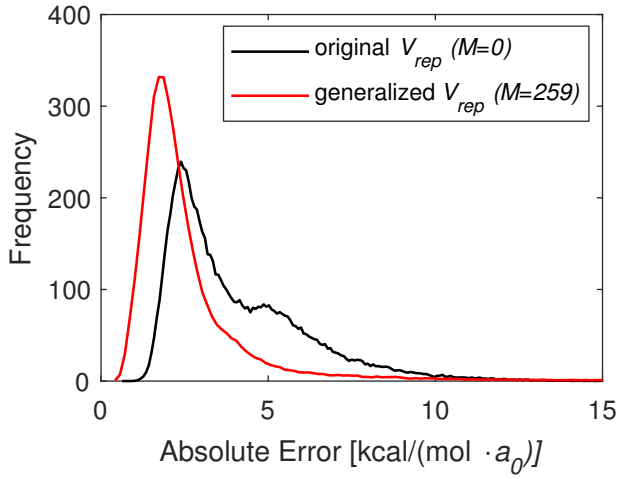

Figure 7: Normalized histogram of the magnitudes of errors in force per atom for the molecules in the test set with $M = 259$ generalized repulsive potentials.

the error is in the bond angles, not lengths. Because repulsive potentials only yield forces along bond axes, the generalized potentials cannot reduce those errors.

Overall, a clear improvement of the performance of the method by the addition of generalized repulsive potentials is apparent.

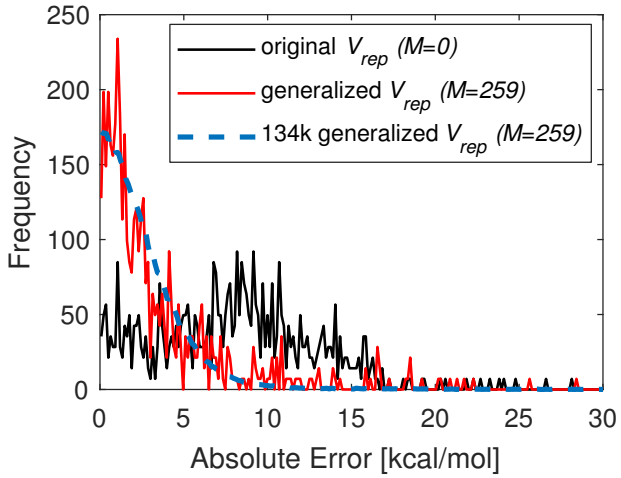

Figure 8: Normalized histogram of absolute errors in atomization energy for the 622 molecules in the Jorgsen test set with $M = 259$ generalized repulsive potentials. The histogram for the large test set is also shown for comparison.

Beyond testing on the large test set of 130k molecules, we conduct some further benchmarks. First, we also analyze MAEs for molecules from a smaller test set introduced by Jorgsen et al.[49] that has been previously used to benchmark DFTB.[6] The set contains 622 molecules containing C, H, N, and O. Most molecules of the Jorgsen set are contained in the large test set, but the smaller set is commonly used for quantum chemical benchmarks, and therefore it is illustrative to compare performance on the well known subset with the whole. For a fair assessment, we did not use results reported in the literature, but rather created data at our own reference level of theory, B3LYP/6-31G(2df,p). Fig. 8 shows error histograms with and without generalized repulsive potentials. A clear improvement is visible. Results look very similar to the large test set results, albeit more oscillatory due to sparcity of data. The MAEs of $\Delta E_0 = 8.26\,\frac{\text{kcal}}{\text{mol}}$ with the original DFTB and $\Delta E_{259} = 3.64\,\frac{\text{kcal}}{\text{mol}}$ are somewhat larger because of a few outliers, which are less frequent in the large set. That may be regarded as an example that good performance on average not necessarily implies good

performance for every specific problem.

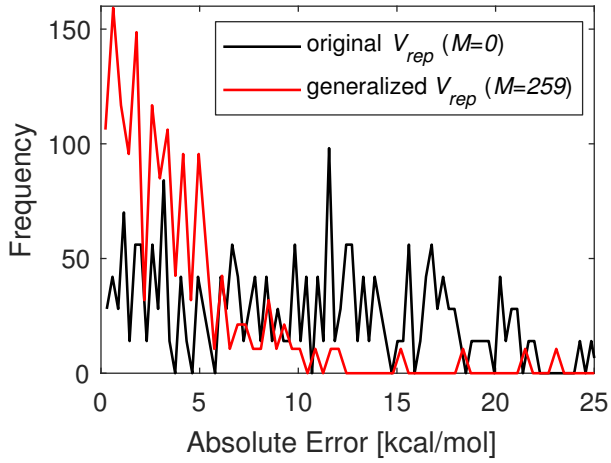

Figure 9: Normalized histogram of absolute errors in atomization energy for 155 random molecules from the GDB-17 database with 17 heavy atoms. $M = 259$ generalized repulsive potentials are used.

To demonstrate transferability to larger molecules, we sample 155 random molecules with 17 C, N, O, or F atoms from the GDB-17 database, for which we created geometries following the same protocol as before for the smaller molecules. Fig. 9 displays a histogram of absolute errors in atomization energy, and the visible improvement is comparable to that in the set of smaller molecules. The MAE reduces from $\Delta E_0 = 10.31\,\frac{\text{kcal}}{\text{mol}}$ to $\Delta E_{256} = 4.17\,\frac{\text{kcal}}{\text{mol}}$, by a factor of about 1/2. At this point, note also that the vast majority of training molecules contained only 8 or fewer heavy atoms, whereas the benchmark molecules were composed of 9 heavy atoms, so that the large benchmark already constituted an, although limited, transfer to larger molecules. We also test the method on a set of drug molecules we have previously used for benchmarks of other methods.[19] We test on the 18, out of 24, drug molecules listed in Fig. 3 of Ref. 19 that contain only H, C, N, and O atoms. This set contains molecules such as aspirin and vitamin C, with up to 113 atoms. The MAE drops from $\Delta E_0 = 12.42\,\frac{\text{kcal}}{\text{mol}}$ to $\Delta E_{256} = 7.90\,\frac{\text{kcal}}{\text{mol}}$, in line with the other results.

To assess predictions at non-equilibrium geometries molecular dynamics (MD) sampling, lasting 10 ps at a temperature of $T = 300$ K, was performed for 100 random molecules from the test set. Of course, thus only the local vicinities of the molecular potential energy

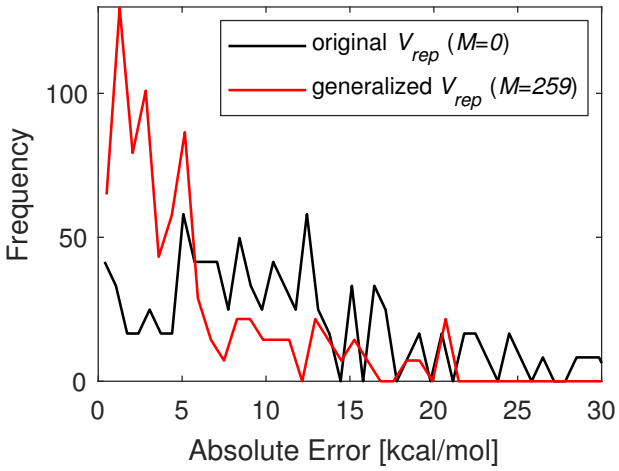

Figure 10: Normalized histogram of absolute errors in atomization energy for 100 random molecules from the test set at perturbed geometries sampled from MD simulations at $T = 300$ K.

surfaces are sampled, but no extended regions involving, for example, reaction trajectories, which the method does not improve at present. Fig. 10 shows a reduction in errors; MAEs improve from $\Delta E_0 = 11.23\,\frac{\text{kcal}}{\text{mol}}$ to $\Delta E_{256} = 5.66\,\frac{\text{kcal}}{\text{mol}}$ by a margin comparable to the other results.

Table 3: Comparison of errors in atomization energies for $\sim$ 130k molecules with results obtained from smaller training sets.

training set size	M	h	MAE $[\frac{\text{kcal}}{\text{mol}}]$	RMSE $[\frac{\text{kcal}}{\text{mol}}]$
2100	200	27.13	2.89	4.0
2100	111	36.09	2.97	3.95
1000	156	24.71	3.52	5.52

Finally, in Tab. 3 we investigate the effects of training set size. A clustering and fit was performed with only 1000, rather than 2100, molecules in the training set. MAEs of atomization energies are about 0.5 $\frac{\text{kcal}}{\text{mol}}$ larger than for comparable parametrizations with the full training set. A training set of 1000 molecules is already large enough to be useful, but clearly there is still room for improvement with more training data.

Conclusion

We have introduced generalized repulsive potentials for the DFTB method, where the traditional atom-type potentials are substituted by bond-type potentials. Bond types are determined by automatic clustering, leading to a description which reflects the chemical environment of the bond. Due to the automatic bond identification, repulsive potentials can be parametrized to fit large data sets and are not limited by the number of parameters that can be scaled up as required. This brings DFTB into the age of data driven approaches. We presented a preliminary implementation of the method that clearly demonstrates the potential of the method to push forward DFTB development, offering significantly improved quantitative performance. It is more difficult to compare the method directly to pure ML approaches, rather than to conventional DFTB, because most employ much more training data, and results with only a few thousands molecules in the training set are rarely reported. In one recent study, Kernel ridge regression with a constant size molecular descriptor,[50] capable of reaching chemical accuracy with bigger training sets, yields an MAE of $\Delta E_{\mathrm{KKR}} = 3.64\,\frac{\mathrm{kcal}}{\mathrm{mol}}$ for the test set of 130k molecules with 2000 molecules in the training set, against an error of $\Delta E_{256} = 2.64\,\frac{\mathrm{kcal}}{\mathrm{mol}}$ with extended DFTB and a training set size of 2100. Extrapolating the learning curves reported in Ref. 34, where multiple ML methods and descriptors were compared, indicates a similar MAE of about $3.5\,\frac{\mathrm{kcal}}{\mathrm{mol}}$ with 2000 training molecules for the best performing method considered in this study. When we apply the ΔML method[21] with a DFTB base line and 2000 training molecules, we find an MAE of $\Delta E_{\Delta\mathrm{ML}} = 3.4\,\frac{\mathrm{kcal}}{\mathrm{mol}}$. Overall, DFTB with generalized repulsive potentials seem to compare favorably to these examples, given the limited training set size here considered. Some recent direct ML models, however, perform better even with small training sets. In Ref. 18, transferability of information from one chemical element to another is exploited to reach chemical accuracy of $\Delta E = 1\,\frac{\mathrm{kcal}}{\mathrm{mol}}$ with only 2000 training molecules. The SOAP-GAP method achieves the same feat[15] with about a 1000 training molecules. Finally, in Ref. 19, 200 training samples suffice to reach chemical accuracy. Here, however, the samples are preselected from a much larger reference

set. This idea, in particular, should also be applicable to the repulsive potential approach, with bond types selected based on their performance for equilibrium geometries in a large set, and geometry sampling only performed afterwards. It must also be born in mind that DFTB calculations are usually at least one order of magnitude slower than pure ML models. But at least presently, most ML models require preoptimized structures at the SE level, or better, destroying any advantage in computational cost.

In conclusion, what this work achieves is a clear demonstration that simple data based extensions can lead to significantly improved performance of DFTB, and likely similar SE methods, and it thus helps guide the future development of such methods. What the role of SE models in the long run will be in light of the impressive development of direct ML models is to be seen.

As a next step, a larger training set and the incorporation of recent ideas from direct ML models will certainly help to further improve the method. Also, since this paper reports a proof of concept, further developments are required before routine application to quantum chemical problems: (i) The reference data should be be computed using a higher level method than used here. Now that the principal functionality of the method has been established, a smaller test set for verification will be sufficient, which allows to compute using higher level approaches. (ii) Using better reference data, a test of different descriptors should be performed in order to evaluate, whether the final performance can be improved. The literature offers a rich choice of options and in future work we will try to identify the best solution. Because bond clustering is conceptually quite different from direct, supervised ML models, we do not necessarily expect performance data reported in the literature to directly transfer to this application. (iii) Finally, a smooth switching between different bond-potentials has to be enabled, in particular when this scheme shall be applied for molecular dynamics simulations. Up to now, hard cut-offs are assigned, which have to be substituted switching

functions f. The generalized repulsive potential becomes

$$\Delta V_{AB}(R_{AB}) = \sum_{\tilde{b}} f(d(b(A,B),\tilde{b})\Delta V_{\tilde{b}}(R_{AB}),$$

where $\tilde{b}$ runs over all bond types and f interpolates smoothly between $f(0) = 1$ and $f(\infty) = 0$. One possibility would be the use of an error-function. (iv) Further, the description of chemical reactions can be improved by adding transition state geometries to the training set.

Acknowledgement

The authors thank V. M. Pérez Wohlfeil for help with creating Fig. 2 and critical reading of the manuscript. This work was supported through grant DFG-SFB 1249 'N-Heteropolyzyklen als Funktionsmaterialien'. We acknowledge computational resources provided by the state of Baden-Württemberg through bwHPC and the DFG through grant no INST 40/467-1 FUGG. OAvL acknowledges support from the Swiss National Science foundation (No. PP00P2_138932). This research was partly supported by the NCCR MARVEL, funded by the Swiss National Science Foundation.

References

(1) Hohenberg, P.; Kohn, W. Inhomogeneous electron gas. *Phys. Rev.* **1964**, *136*, B864.

(2) Kohn, W.; Sham, L. J. Self-consistent equations including exchange and correlation effects. *Phys. Rev.* **1965**, *140*, A1133.

(3) Cui, Q.; Elstner, M. Density functional tight binding: values of semi-empirical methods in an ab initio era. *Phys. Chem. Chem. Phys.* **2014**, *16*, 14368–14377.

(4) Repasky, M. P.; Chandrasekhar, J.; Jorgensen, W. L. PDDG/PM3 and PDDG/MNDO: improved semiempirical methods. *J. Comp. Chem.* **2002**, *23*, 1601–1622.

(5) Korth, M.; Thiel, W. Benchmarking semiempirical methods for thermochemistry, kinetics, and noncovalent interactions: OMx methods are almost as accurate and robust as DFT-GGA methods for organic molecules. *J. Chem. Theory Comput.* **2011**, *7*, 2929–2936.

(6) Gaus, M.; Goez, A.; Elstner, M. Parametrization and benchmark of DFTB3 for organic molecules. *J. Chem. Theory Comput.* **2012**, *9*, 338–354.

(7) Christensen, A. S.; Tomas, K.; Cui, Q.; Elstner, M. Semiempirical quantum mechanical methods for noncovalent interactions for chemical and biochemical applications. *Chem. Rev.* **2016**, *116*, 5301–5337.

(8) Humeniuk, A.; Mitrić, R. Long-range correction for tight-binding TD-DFT. *J. Chem. Phys.* **2015**, *143*, 134120.

(9) Lutsker, V.; Aradi, B.; Niehaus, T. A. Implementation and benchmark of a long-range corrected functional in the density functional based tight-binding method. *J. Chem. Phys.* **2015**, *143*, 184107.

(10) Humeniuk, A.; Mitric, R. DFTBaby: A software package for non-adiabatic molecular dynamics simulations based on long-range corrected tight-binding TD-DFT (B). *arXiv:1703.04049* **2017**,

(11) Kranz, J. J.; Elstner, M.; Aradi, B.; Frauenheim, T.; Lutsker, V.; Garcia, A. D.; Niehaus, T. A. Time-dependent extension of the long-range corrected density functional based tight-binding method. *J. Chem. Theory Comput.* **2017**, *13*, 1737–1747.

(12) 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.

(13) 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.

(14) 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.

(15) Bartok, A. P.; De, S.; Poelking, C.; Bernstein, N.; Kermode, J. R.; Csanyi, G.; Ceriotti, M. Machine learning unifies the modeling of materials and molecules. *Sci, Adv.* **2017**, *3*, e1701816.

(16) Schütt, K. T.; Arbabzadah, F.; Chmiela, S.; Müller, K. R.; Tkatchenko, A. Quantum-chemical insights from deep tensor neural networks. *Nature Comm.* **2017**, *8*, 13890.

(17) Gilmer, J.; Schoenholz, S. S.; Riley, P. F.; Vinyals, O.; Dahl, G. E. Neural message passing for quantum chemistry. *arXiv:1704.01212* **2017**,

(18) Faber, F. A.; Christensen, A. S.; Huang, B.; von Lilienfeld, O. A. Alchemical and structural distribution based representation for improved QML. *arXiv:1712.08417* **2017**,

(19) Huang, B.; von Lilienfeld, O. A. Chemical space exploration with molecular genes and machine learning. *arXiv:1707.04146* **2017**,

(20) Smith, J. S.; Isayev, O.; Roitberg, A. E. ANI-1: an extensible neural network potential with DFT accuracy at force field computational cost. *Chem. Science* **2017**, *8*, 3192–3203.

(21) Ramakrishnan, R.; Dral, P. O.; Rupp, M.; von Lilienfeld, O. A. Big data meets quan-

tum chemistry approximations: the Δ-machine learning approach. *J. Chem. Theory Comput.* **2015**, *11*, 2087–2096.

(22) Dral, P. O.; von Lilienfeld, O. A.; Thiel, W. Machine learning of parameters for accurate semiempirical quantum chemical calculations. *J. Chem. Theory Comput.* **2015**, *11*, 2120–2125.

(23) Behler, J. Atom-centered symmetry functions for constructing high-dimensional neural network potentials. *J. Chem. Phys.* **2011**, *134*, 074106.

(24) Elstner, M.; Porezag, D.; Jungnickel, G.; Elsner, J.; Haugk, M.; Frauenheim, T.; Suhai, S.; Seifert, G. Self-consistent-charge density-functional tight-binding method for simulations of complex materials properties. *Phys. Rev. B* **1998**, *58*, 7260.

(25) Knaup, J. M.; Hourahine, B.; Frauenheim, T. Initial Steps toward Automating the Fitting of DFTB $E_{rep}(r)$. *J. Phys. Chem. A* **2007**, *111*, 5637–5641.

(26) Chou, C.-P.; Nishimura, Y.; Fan, C.-C.; Mazur, G.; Irle, S.; Witek, H. A. Automatized Parameterization of DFTB Using Particle Swarm Optimization. *J. Chem. Theory Comput.* **2015**, *12*, 53–64.

(27) Porezag, D.; Frauenheim, T.; Köhler, T.; Seifert, G.; Kaschner, R. Construction of tight-binding-like potentials on the basis of density-functional theory: Application to carbon. *Phys. Rev. B* **1995**, *51*, 12947.

(28) Elstner, M.; Porezag, D.; Jungnickel, G.; Elsner, J.; Haugk, M.; Frauenheim, T.; Suhai, S.; Seifert, G. Self-consistent-charge density-functional tight-binding method for simulations of complex materials properties. *Phys. Rev. B* **1998**, *58*, 7260.

(29) Gaus, M.; Cui, Q.; Elstner, M. DFTB3: extension of the self-consistent-charge density-functional tight-binding method (SCC-DFTB). *J. Chem. Theory Comput.* **2011**, *7*, 931–948.

(30) Kubillus, M.; Kubar, T.; Gaus, M.; Rezac, J.; Elstner, M. Parameterization of the DFTB3 method for Br, Ca, Cl, F, I, K, and Na in organic and biological systems. *J. Chem. Theory Comput.* **2014**, *11*, 332–342.

(31) Smith, J. S.; Isayev, O.; Roitberg, A. E. ANI-1, A data set of 20 million calculated off-equilibrium conformations for organic molecules. *Sci. Data* **2017**, *4*, 170193.

(32) Bartok, A. P.; Kondor, R.; Csanyi, G. On representing chemical environments. *Phys. Rev. B* **2013**, *87*, 184115.

(33) 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*, 161102–161102.

(34) 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.

(35) Friedman, J.; Hastie, T.; Tibshirani, R. *The elements of statistical learning*; Springer series in statistics New York, 2001; Vol. 1.

(36) Steinhaus, H. Sur la division des corp materiels en parties. *Bull. Acad. Polon. Sci* **1956**, *1*, 801.

(37) MacQueen, J. Some methods for classification and analysis of multivariate observations. Proceedings of the fifth Berkeley symposium on mathematical statistics and probability. 1967; pp 281–297.

(38) Fukunaga, K.; Hostetler, L. The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Trans. Inf. Theory* **1975**, *21*, 32–40.

(39) Cheng, Y. Mean shift, mode seeking, and clustering. *IEEE Trans. Pattern Anal. Mach. Intell.* **1995**, *17*, 790–799.

(40) Pedregosa, F. et al. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **2011**, *12*, 2825–2830.

(41) Walt, S. v. d.; Colbert, S. C.; Varoquaux, G. The NumPy array: a structure for efficient numerical computation. *Comput. Sci. Eng.* **2011**, *13*, 22–30.

(42) 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.

(43) 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. Mod.* **2012**, *52*, 2864–2875.

(44) Lee, C.; Yang, W.; Parr, R. G. Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density. *Phys. Rev. B* **1988**, *37*, 785.

(45) Becke, A. D. Density-functional exchange-energy approximation with correct asymptotic behavior. *Phys. Rev. A* **1988**, *38*, 3098.

(46) Stephens, P.; Devlin, F.; Chabalowski, C.; Frisch, M. J. Ab initio calculation of vibrational absorption and circular dichroism spectra using density functional force fields. *J. Phys. Chem.* **1994**, *98*, 11623–11627.

(47) Pearson, K. LIII. On lines and planes of closest fit to systems of points in space. *Philos. Mag.* **1901**, *2*, 559–572.

(48) Hotelling, H. Analysis of a complex of statistical variables into principal components. *J. Ed. Psycho.* **1933**, *24*, 417.

(49) Sattelmeyer, K. W.; Tirado-Rives, J.; Jorgensen, W. L. Comparison of SCC-DFTB and NDDO-based semiempirical molecular orbital methods for organic molecules. *J. Phys. Chem. A* **2006**, *110*, 13551–13559.

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

Graphical TOC Entry