JCTC Journal of Chemical Theory and Computation

pubs.acs.org/JCTC Article

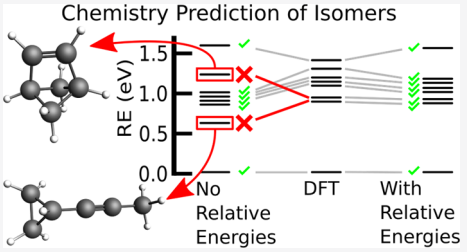
Development of Density Functional Tight-Binding Parameters Using Relative Energy Fitting and Particle Swarm Optimization

Néstor F. Aguirre, Amanda Morgenstern, M. J. Cawkwell, Enrique R. Batista,* and Ping Yang*

Cite This: *J. Chem. Theory Comput.* 2020, 16, 1469−1481 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: We provide a strategy to optimize density functional tight-binding (DFTB) parameterization for the calculation of the structures and properties of organic molecules consisting of hydrogen, carbon, nitrogen, and oxygen. We utilize an objective function based on similarity measurements and the Particle Swarm Optimization (PSO) method to find an optimal set of parameters. This objective function considers not only the common DFTB descriptors of binding energies and atomic forces but also incorporates relative energies of isomers into the fitting procedure for more chemistry-driven results. The quality in the description of the binding energies and atomic forces is measured based on the Ballester similarity index and relative energies through a similarity index induced by the Levenshtein edit distance to quantify the correct energetic order of isomers. Training and testing datasets were created to include all relevant chemical functional groups. The accuracy of this strategy is assessed, and its range of applicability is discussed by comparison against our previous parameterization [A. Krishnapriyan, et al., *J. Chem. Theory Comput. 13*, 6191 (2017)]. The improved performance of the new DFTB parameterization is validated with respect to the density functional theory large datasets QM-9 [R. Ramakrishnan, et al., *Sci. Data 1*, 140022 (2014)] and ANI-1 [J. S. Smith, et al., *Sci. Data 4*, 170193 (2017)], where excellent agreement is found between the structures and properties available in these datasets, and the ones obtained with DFTB.



1. INTRODUCTION

One of the greatest challenges in computational chemistry today is the development of methods capable of modeling large systems in a reasonable amount of time that are also able to include electronic degrees of freedom. *Ab initio* methods, such as density functional theory (DFT), provide accurate results, but *ab initio* molecular dynamics (AIMD) can only be run realistically for relatively small systems, reaching sizes of several hundreds of atoms and time scales on the order of hundreds of ps (see, e.g., refs 1 and 2). Alternatively, classical molecular dynamic (CMD) simulations have been performed on systems with hundreds of thousands of atoms with timescales of several nanoseconds or even microseconds (see, e.g., refs 3−7) but have other limitations, including less accurate energies than AIMD and the inability to form and break chemical bonds. Density functional tight binding (DFTB) is an attractive alternative to DFT to simulate chemical processes, as it is a semiempirical method based on quantum mechanics, leading to accurate results, but with some approximations with respect to DFT and empirical parameters. For medium sized systems (≤100 atoms), DFTB is roughly 3 orders of magnitude faster than DFT (with a medium sized basis set) and 3 orders of magnitude slower than CMD methods (see, e.g., refs 8−10). For large-scale electronic structure calculations, even when DFTB scales $O(N^3)$ with the size of the system as DFT, further algorithmic improvements for special cases demonstrated a linear-scaling performance closer to CMD.[11−13]

As in any parameterized method, the accuracy that can be reached by DFTB calculations is related to the quality of parameterization. The parameters included in DFTB simulations vary between different implementations, as does the method for optimizing the parameters. A simultaneous optimization of pair potentials and bond integrals, represented by analytical functions, has recently been reported that resulted in accurate and transferable DFTB models for systems containing H, C, N, and O atoms. Versions of these parameters have been used in molecular dynamics simulations to calculate properties of explosive materials, including Hugoniot loci and shock temperatures.[14,15]

In this work, we strive to create a parameterization protocol of general applicability for optimizing both pair potentials and bond integrals. We begin this process with the DFTB parameterization for molecules containing H, C, N, and O atoms only (HCNO). We chose this set of elements as our test case, as there are many DFTB parameterizations[16−21] and datasets that are limited to these atoms, providing us with ample resources for optimization and validation of parameters. In addition to the parameterization methods developed here, we also created new datasets of molecules, containing HCNO

Received: September 2, 2019
Published: February 20, 2020

https://dx.doi.org/10.1021/acs.jctc.9b00880
J. Chem. Theory Comput. 2020, 16, 1469−1481

with both optimized and distorted geometries (stretching bonds and angles to get forces) for training and testing the DFTB parameters.

Typical parameterization schemes focus on single-molecule properties, such as atomic forces and atomization (binding) energies. In this work, we include additional chemical information provided by the relative energies of isomers in the optimization of parameters. We developed an objective function that includes the energy differences between molecules with the same stoichiometry and places extra emphasis on the correct identification of the most stable isomer. In this way, the objective function is more chemistry-driven, as determining relative energies of isomers is important for predicting and understanding the products of chemical reactions. Furthermore, this objective function is based on similarity indices (see, e.g., refs 22−25), whereas most DFTB parameterizations have been based on least squares methods. Although square objective-based strategies are simple, they have limitations. One disadvantage is that they have the tendency to be dominated by outliers. The final sum tends to be the result of a few particularly large values, rather than an expression of the average. Thus, it tends to penalize outliers excessively,[26] as opposed to objective functions based on similarity measurements. Finally, we parameterized our model using a global optimization method, particle swarm optimization (PSO), as opposed to many optimization methods that search for local minima in the basin of the initial guess. This approach has a higher chance of finding acceptable parameters even if the initial guess is not of high quality.

This paper is structured as follows. Section 2 outlines the methodological approach with emphasis on the plan adopted to design the global optimization strategy. Section 3 presents the methodological background necessary for its application to the optimization of DFTB parameterization, and results are shown and discussed in Section 4. Section 5 closes with a summary and suggests possible directions for future applications and extensions.

2. OPTIMIZATION STRATEGY

In this section, we first define the objective function based on similarity indices developed for this work. Then, we briefly summarize the PSO method, including the specific adaptations developed. Finally, we describe the transferability tests that we used to identify over-fitting situations and improve the predictive capability of the model.

2.1. Objective Function. Similarity measures are applied to quantify how similar objects are with respect to one another. In the field of computational chemistry, this concept has been used to group molecules according to their biological effects or physicochemical properties, and this concept has found extensive use in drug discovery techniques (see, e.g., refs 22−25). We are interested in comparing how different properties of a set of molecules compare when calculated with the approximate DFTB method with respect to a reference method like DFT. A formal measure of similarity between molecules is essential and depends on the underlying molecular representation. In general, this measure is based on a vectorized form representation of molecular descriptors (e.g., binding energy, dipole moment, highest occupied molecular orbital−lowest unoccupied molecular orbital gap, relative energies, etc.) that include the information to describe the chemical properties of interest using the parameterized DFTB model.

We introduce two different sets of molecular descriptors. The first set is computed from a methodology based solely on the structure of individual molecules. The second set includes properties that emerge from a collection of molecules. In this work, we used binding energies (defined as the minimum energy required to disassemble the molecule into its atoms in the ground-state) and atomic forces for the first set and the success in the determination of the lowest energy isomer and the relative order of all isomers as the second set.

To formalize these concepts consider the following two sets: first, a set of N molecules calculated with the approximated method (i.e., DFTB) $\mathbf{M} = \{M_1, M_2, ..., M_N\}$, and second the same molecules calculated with the reference method (i.e., DFT) $\hat{\mathbf{M}} = \{\hat{M}_1, \hat{M}_2, ..., \hat{M}_N\}$. Note that these molecules are not necessarily minimum energy configurations unless specifically stated. We distinguish between "local minima" and distortions when the atomic forces are not equal to zero.

In general, the properties of a molecule are represented by a vector containing n_p molecular descriptors

$$\mathbf{P}(M_a) = (P_1^a, P_2^a, ..., P_{n_p}^a) \tag{1}$$

The first set of descriptors is numerical values assigned to a specific molecular structure M_a. For the binding energy (BE) and the atomic forces (F), each molecule is represented as follows

$$P(M_a) = P_b(M_a) \oplus P_f(M_a) \tag{2}$$

where

$$P_b(M_a) = (\mathrm{BE}^a),$$
$$P_f(M_a) = \frac{1}{3n_a}(F_{1x}^a, F_{1y}^a, F_{1z}^a, ..., F_{nx}^a, F_{ny}^a, F_{nz}^a) \tag{3}$$

and n_a is the number of atoms in molecule M_a. Other properties can be included in a straightforward way by further appending the vector, $P(M_a)$. The weighting factor in the forces, $1/3n$, is included to give equal weight to all the descriptions based on the dimensionality of $P_f(M_a)$. Properties are normalized using the preprocessing step described in the Supporting Information, Section I, to remove the bias induced by the different units of different descriptors.

We compare molecular descriptors on the basis of similarity measure proposed by Ballester and Richards.[27] For the case $M_a \in \mathbf{M}$ and $\hat{M}_a \in \hat{\mathbf{M}}$, the similarity function is defined as

$$S(M_a, \hat{M}_a) = \frac{1}{1 + \frac{1}{n_p}\sum_{i=1}^{n_p} |P_i^a - \hat{P}_i^a|} \tag{4}$$

and the similarity measure for the full set of molecules is defined as

$$S_p(M, \hat{M}) = \frac{1}{N}\sum_{a=1}^{N} S(M_a, \hat{M}_a) \tag{5}$$

Notice that $0 < S_p \leq 1$ by definition. To define the similarity measurement for the second set of descriptors in regard to relative energies of isomers, we chose those molecules from $\mathbf{M}$ with the same chemical composition s (e.g., H_2C_4) as follows

$$\mathcal{M}_s := \left\{ M_i \in M: \begin{array}{l} M_i \text{ has the stoichiometries} \\ M_i \text{ is a local minimum} \end{array} \right\} \tag{6}$$

https://dx.doi.org/10.1021/acs.jctc.9b00880
J. Chem. Theory Comput. 2020, 16, 1469−1481

Note that distortions are not considered in this definition. The relative energies vector $\Delta\mathbf{E}$ is defined as the collection of the differences in the energies of all molecular species (e.g., $E_{a_i^s}$) with the same stoichiometry (isomers) relative to the lowest energy configuration ($E_{a_1^s}$)

$$\Delta\mathbf{E}(\mathcal{M}_s) = (\Delta E_{a_1^s}, \Delta E_{a_2^s}, \Delta E_{a_3^s}, ..., \Delta E_{a_{m_s}^s}) \tag{7}$$

where $\Delta E_{a_i^s} = E_{a_i^s} - E_{a_1^s}$, and its elements are sorted such that $\Delta E_{a_1^s} = 0 < \Delta E_{a_2^s} < \cdots < \Delta E_{a_{m_s}^s}$, and m_s is the number of isomers with the stoichiometry s. Equivalently for the reference set of molecules we have

$$\Delta\mathbf{E}(\hat{\mathcal{M}}_s) = (\widehat{\Delta E}_{b_1^s}, \widehat{\Delta E}_{b_2^s}, \widehat{\Delta E}_{b_3^s}, ..., \widehat{\Delta E}_{b_{m_s}^s}) \tag{8}$$

Notice that the two set of indices

$$\mathbf{a}^s = (a_1^s, a_2^s, a_3^s, ..., a_{m_s}^s)$$
$$\mathbf{b}^s = (b_1^s, b_2^s, b_3^s, ..., b_{m_s}^s) \tag{9}$$

contain the same elements but may be in a different order. This is due to the fact that the approximate model only ideally can reproduce the order of the reference method. Generally, one says that if these two vectors are similar, the chemistry is well described by the approximate model. Therefore, it is necessary to introduce a measurement to quantify their similarity. The relevant information is in the order rather than the magnitude of their elements because the magnitude is already included in the binding energy component of S_p. To balance these two factors, equal weight was applied in the objective function for: the determination of the lowest energy configuration (l) and the determination of the relative order of the structures higher in energy (o).

For fitting the lowest energy configuration, we defined the objective function as the fraction of success in the determination of the lowest energy state

$$S_l(\mathrm{M}, \hat{\mathrm{M}}) = \frac{1}{N_\Pi}\sum_{s\in\Pi}\delta_{a_1^s, b_1^s} \tag{10}$$

where Π is the set containing all possible stoichiometries in the sets $\mathbf{M}$ and $\hat{\mathbf{M}}$, N_Π the number of elements of Π (number of different stoichiometries), and δ the Kronecker delta function, which is 1 only if $a_1^s = b_1^s$ and 0 otherwise. In other words, S_l counts how many times the lowest energy configuration is predicted correctly.

We used the edit distance proposed by Levenshtein,[28] for determining the relative order of energies, in order to measure how close two sequences are. This edit distance is based on the minimum number of edit operations needed to convert one sequence of elements into another by using the following operations: removal, insertion, and substitution. For example, the Levenshtein distance between the words "global" and "local" is two (one substitution, b → c and one removal, g).

For two sequences $\mathbf{a}$, $\mathbf{b}$, we denote this measure as $\Gamma(\mathbf{a}^s, \mathbf{b}^s)$. Thus, the similarity measurement to quantify the relative order of the isomers based on this edit distance was defined as

$$S_o(\mathrm{M}, \hat{\mathrm{M}}) = \frac{1}{N_\Pi}\sum_{s\in\Pi}\left(1 - \frac{\Gamma(a^s, b^s)}{N_{a^s}}\right) \tag{11}$$

where $N_{\mathbf{a}}^s$ is the number of elements of the sequence $\mathbf{a}^s$ and is also equal to the maximum value of the Levenshtein distance. Notice that $0 \le S_o \le 1$.

Finally, by taking into account the three contributions, the total similarity S_t between the set of molecules and the reference is written as

$$S_t = \frac{2}{3}S_p + \frac{1}{6}(S_l + S_o) \tag{12}$$

where two weighting factors were added to ensure the equal contribution from all properties depending of the number of considered properties in S_p. To generalize for further applicability if more properties are included, the equivalent expression is

$$S_t = \frac{n_p}{n_p + 1}S_p + \frac{1}{2(n_p + 1)}(S_l + S_o) \tag{13}$$

for example, if considered properties are binding energy, forces, relative energies, and dipoles, the n_p value is 4. Notice that S_t also satisfies $0 < S_t \le 1$.

The optimization process consists of maximizing the similarity S_t, or, equivalently, minimizing the dissimilarity $1 - S_t$, which is the objective function, by varying the parameters $\mathbf{v}$, as follows

$$\hat{\mathbf{v}} = \arg\min_{v}\{1 - S_t(\mathbf{M}(\mathbf{v}), \hat{\mathbf{M}})\} \tag{14}$$

where $\hat{\mathbf{v}}$ is the solution vector.

Notice that eqs 10 and 11 are not continuous and therefore neither is S_t, which makes the optimization process more difficult to converge. To overcome this challenge, we carried out the optimizations in two steps: first by optimizing the molecular properties based on the continuous similarity index S_p (see eq 5) and then by optimizing the total similarity S_t starting from the former solution. Hereafter, we refer to these two steps as $\mathrm{opt}_{\mathrm{BE,F}}$ and $\mathrm{opt}_{\mathrm{RE}}$, respectively.

2.2. PSO and Transferability. We used a PSO method[29] to minimize the objective function eq 14. PSO uses a swarm of n particles that represent n points in the $\mathbf{v}$ parameter space. Each particle retains information of the best position (lowest objective function) that the individual has reached in the space (cognitive component) and also has information of the best position reached by the entire swarm population (social component). By combining these two pieces of information the swarm of particles is able to search through a complex space, testing different funnels and eventually converging upon a single optimal point. We also used the accelerated PSO algorithm (APSO)[30−32] to speed up the initial optimization procedure ($\mathrm{opt}_{\mathrm{BE,F}}$) before performing a full PSO within a smaller search space for $\mathrm{opt}_{\mathrm{RE}}$. The APSO algorithm is a simpler version of the PSO algorithm which uses the global best only without the cognitive component of each particle. APSO optimizations begin with a normal PSO optimization, then the particles gradually lose the cognitive component in order to speed up the convergence. Acceleration is activated once less than 10% of the particles are far away from the swarm (objective function is larger than two standard deviations from the mean). In all PSO simulations, the number of particles, or population of the swarm, is $n \ge 10 + [2\sqrt{d}]$ particles where d is the dimensionality of the system.[33] Further details of these methods are available in the Supporting Information, Section II.

The PSO algorithm produces a single solution, which is possibly different in consecutive runs of the algorithm (see, e.g., ref 34). A method for choosing only one of these solutions (the one with the minimum overfitting) is through evaluations

https://dx.doi.org/10.1021/acs.jctc.9b00880
J. Chem. Theory Comput. 2020, 16, 1469−1481

of the objective function on the testing set (see Figure 1). If a set of parameters fits the training dataset and also fits the test

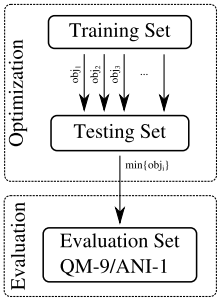


Figure 1. Schematic diagram of the optimization and evaluation process.

dataset, we have a minimal overfitting situation. On the contrary, a better fitting of the training dataset than to the test dataset indicates an overfitting situation. The testing set should be chosen independently of the training dataset but should have similar chemical information as the training dataset. To test the transferability of the optimized DFTB parameters, we used a training set of small molecules including the most representative chemical functional groups and, as a testing set, larger systems with combinations of different functional groups. This strategy allows for validation in which the subsets are assigned nonrandomly but based on a relevant factor such as the chemical complexity. Thus, the best set of parameters is chosen as the one with the highest predictive ability, based on the testing sets.

Finally, the performance of the final parameterization was evaluated with respect to the published large datasets QM-9 and ANI-1 (see Figure 1).

Additionally, we used the k-fold cross-validation method, with $k = 5$, where one-fifth of the original training data is kept for model testing and the rest is kept as the training dataset. In this way, all data are iteratively used for both training and testing (see, e.g., refs 35 and 36). These groups are made by a random selection of the molecules from the initial set. Using a reliable initial guess, the model is then trained and evaluated under the five situations. Thus, this cross-validation step allows us to quantify the sensitivity of the optimized parameters to the training set (sensitivity analysis) and to determine the error of the model.

3. DFTB PARAMETERIZATION

In this section, we describe the DFTB parameterization method we followed. First, we introduce the main concepts of the DFTB approximation including the specific parameters to be optimized and their physical meaning (Section 3.1). Second, we describe a method for obtaining initial guesses of these parameters that are based on DFT calculations in order to guarantee a reasonable starting point and reduce computational effort during optimization (Section 3.2). Then, we introduce training and testing datasets of molecules and properties based on DFT calculations that we use as a reference to optimize the parameters and evaluation datasets used to monitor the performance and transferability of the new parameterization (Section 3.3). Finally, we introduce the optimization protocol used, which is designed to increase both the chance of finding the global minimum and the transferability of the model (Section 3.4).

3.1. DFTB. In this work, we used the self-consistent charge density functional tight binding (SCC-DFTB) method, which includes terms to handle the energetics of charge fluctuations due to interatomic interactions. The equations behind the SCC-DFTB method have been very didactically reviewed in several previous works (see, e.g., refs 8−10). Here, we only outline the main expressions to understand the physical meaning of each optimized parameter. More details to unify notations are included in the Supporting Information, Section III.

The fundamental ingredients of a SCC-DFTB parameterization consist of the Hamiltonian $H^{0,ab}_{ll'\mathrm{s}}(R_{ab}) = \langle \phi^a_l | \hat{H}^0 | \phi^b_{l'} \rangle$ and overlap $S^{ab}_{ll'\mathrm{s}}(R_{ab}) = \langle \phi^a_l | \phi^b_{l'} \rangle$ matrix elements, the pair repulsion potentials $\Phi_{ab}(R_{ab})$, and the Hubbard U parameters U_a. Here, $\phi^a_l(R_a)$ and $\phi^b_{l'}(R_b)$ are atomic orbitals centered on the atoms a and b, respectively (l and l' represent their angular momentum). The pair repulsion potential $\Phi_{ab}(R_{ab})$ mainly contains the repulsive interaction between the ions, but in practice, it contains any electronic contribution that is not captured by the other terms. The Hubbard U parameter U_a modulates the charge transfer as a result of different electronegativities (atomic chemical hardness). The functions of $H^{0,ab}_{ll'\mathrm{s}}(R_{ab})$, $S^{ab}_{ll'\mathrm{s}}(R_{ab})$, and $\Phi_{ab}(R_{ab})$ are traditionally evaluated numerically as a function of the distance between the atoms a and b (R_{ab}) and stored in tables (Slater−Koster tables).

We used SCC-DFTB as is implemented in LATTE[37] program. In this implementation, the bond and overlap integrals and repulsive potentials are represented with analytical expressions controlled by a small number of parameters, which reduces significantly the computational effort during the optimization process.[14] These analytical expressions are as follows

$$
\begin{aligned}
H^{0,ab}_{ll'\mathrm{s}}(R_{ab}) &= \varepsilon^{0,a}_l, \qquad \text{if } a = b,\\
H^{0,ab}_{ll'\mathrm{s}}(R_{ab}) &= A^{ab,0}_{ll'\mathrm{s}} \exp\left\{\sum_{k=1}^{4} A^{ab,k}_{ll'\mathrm{s}} (R_{ab} - R_0^{ab})^k\right\},\\
&\qquad \text{if } a \neq b,\\
S^{ab}_{ll'\mathrm{s}}(R_{ab}) &= B^{ab,0}_{ll'\mathrm{s}} \exp\left\{\sum_{k=1}^{4} B^{ab,k}_{ll'\mathrm{s}} (R_{ab} - R_0^{ab})^k\right\},\\
\Phi_{ab}(R_{ab}) &= C^{ab,0} \exp\left\{\sum_{k=1}^{4} C^{ab,k} (R_{ab} - R_0^{ab})^k\right\}
\end{aligned}
\tag{15}
$$

Here, $\varepsilon^{0,a}_l$ is the "on-site energies" and represents the orbital energies of the isolated atoms that guarantee the right asymptotic limit. The latter expressions are valid for $R < R_1$, where R_1 is chosen as the maximum radius where the value of the overlap integrals is close to zero. For larger distances $R_1 < R < R_{\mathrm{cut}}$, a polynomial cutoff tail is used to assure convergence to zero at $R = R_{\mathrm{cut}}$.[14,37] The analytical expressions shown in eq 15 facilitate the parameterization process as described later. In this work, we optimized the parameters associated with the Hamiltonian matrix elements $A^{ab,k}_{ll'\mathrm{s}}$, pair repulsion potentials $C^{ab,k}$, and the Hubbard U_a. As in our previous work, the parameters describing the radial dependencies of the overlap integrals $B^{ab,k}_{ll'\mathrm{s}}$, which were fitted to overlap integrals obtained

https://dx.doi.org/10.1021/acs.jctc.9b00880
J. Chem. Theory Comput. 2020, 16, 1469−1481

from minimal basis DFT calculations, were not optimized with the bond integrals and remained fixed for all calculations. The numerical optimization of the overlap integrals can easily lead to unphysical, nonpositive definite overlap matrices. On-site energies $\varepsilon_l^{0,a}$ were also fixed to the initial guess because tests indicated that there are no significant improvements in the precision of the molecular properties by optimizing them.

3.2. Initial Guesses. Initial guesses for the bond integrals and pair potentials were taken from ref 14 with small modifications. The bond integral and overlap integral radial dependencies were derived from minimal basis DFT calculations using the PLATO package[38−40] with the confining potential used in ref 15

$$V_{\text{conf}}(R) = \left(\frac{R}{2R_{\text{cov}}}\right)^2 \tag{16}$$

where R_{cov} is the covalent radius of the atom.

Equation 15 for $H_{ll'\text{s}}^{0,ab}(R_{ab})$ and $S_{ll'\text{s}}^{ab}(R_{ab})$ were parameterized by a least squares fit using the PLATO data. The radial cutoff values were obtained as follows: R_0^{ab} values were defined as the average of covalent radii (i.e., single, double, and triple bonds) for each atom in the pair (taken from ref 41). For each pair, for example, *ab* the bond integrals and overlaps were cut-off at R_1^{ab}, defined by the distance where the overlap integrals reach a value of 10^{-3} and R_{cut}^{ab} was defined as R_1^{ab} + 0.5 Å.

Pair potential $\Phi_{ab}(R_{ab})$ initial guesses were refitted from the analytical functions with $k = 4$ used in ref 14 to $k = 2$. Tests indicated no improvement in binding energy or force fitting using $k = 4$. Therefore, two terms were used to reduce the number of parameters. R_1^{ab} was also defined by the distance where the function reached the value of 10^{-3} eV. Some values were increased as needed post-optimization to ensure that the pair potentials decayed to at least 10^{-1} at the R_1 values. Table 1 shows all *R* cutoff values for HCNO pairs.

Table 1. *R* Cutoff Values in Å for Hamiltonian *H*, Overlap *S*, and Pair Potential Φ Analytical Functions[a]

		H and *S*		Φ	
pair	R_0	R_1	R_{cut}	R_1	R_{cut}
H−H	0.64	3.0	3.5	1.0	1.5
H−C	0.99	3.6	4.1	1.3	1.8
H−N	0.94	3.5	4.0	1.4	1.9
H−O	0.90	3.3	3.8	1.4	1.9
C−C	1.35	4.0	4.5	1.7	2.2
C−N	1.29	4.2	4.7	1.7	2.2
C−O	1.25	3.7	4.2	1.5	2.0
N−N	1.23	4.0	4.5	1.7	2.2
N−O	1.19	3.9	4.4	1.5	2.0
O−O	1.15	3.4	3.9	1.7	2.2

[a]See eq 15.

3.3. Training, Testing, and Evaluation Sets. The training and testing sets were built from molecules obtained from the NIST[42] and ChemSpider[43,44] databases for chemical structures containing H, C, N, and O. They contain 300 and 157 molecules, respectively. Tables of molecules included in training and testing sets are shown in the Supporting Information, Section VII. The criteria for choosing molecules for each dataset was to include a variety of bonding interactions (e.g., single, double, triple); a wide range of functional groups; various placements of each bonded atom (e.g., primary, secondary, tertiary carbons); combinations of linear, branched, and cyclic structures; and conjugated, aromatic, and anti-aromatic systems. This spans all chemical moieties with as few molecules as possible for accurate yet fast parameterization. Each dataset was designed to focus on the specific pair-wise interactions being optimized in each step described in Section 3.4.

The geometries of all molecules were optimized using the NWChem package[45] with the B3LYP[46−49] functional and cc-pVTZ[50] basis set. Frequency calculations were performed to verify that geometries are local minima (no imaginary frequencies). From each relaxed structure, three distorted geometries were generated in order to sample atomic forces and regions far from equilibrium for the binding energies. Geometry distortions were created using random displacements of the atoms, up to 0.2 Å. We carried out tests by increasing the number of distortions, and we found that using more than three distortions does not significantly affect the results. See the Supporting Information, Section V, for convergence tests using 10 or 20 distortions.

As evaluation sets, we used two large datasets from the literature: QM-9 and ANI-1. QM-9 includes 134k optimized molecules, including several constitutional isomers, and ANI-1 contains more than 20 million distorted conformations of 57k molecules. Both sets were calculated at the DFT level of theory.

3.4. Optimization Protocol. Optimizations of DFTB parameters were performed in five steps: atomic interactions were optimized in groups, in the order of (1) H−H, (2) H−C, C−C, (3a) H−N, C−N, N−N, (3b) H−O, C−O, O−O, (4) N−O, and (5) H−H, H−C, H−N, H−O, C−C, C−N, C−O, N−N, N−O, O−O. Steps 3a and 3b were performed independently. Step 5 includes all interactions. A figure with the detailed workflow is available in the Supporting Information, Section III.

In general, the bond integrals and pair potentials were optimized for all atom pairs, and the overlap integrals were kept fixed to the initial guesses. Notice that our initial guesses are not random, and they already contain the right physical behavior from DFT calculations. Thus, during all optimization steps, the space boundaries of the parameters were chosen as a certain percentage of variation (100 or 50%) from the initial guesses. In some cases, optimizations resulted in variables that were within 10% of the boundaries. To ensure that the variables were at optimal values (and not simply stuck at the artificial boundaries we created), these optimizations were restarted from the best result with new boundaries defined by 50% variation around the new starting point. Each optimization was run five times, and the best parameters were selected as the model that produced the lowest testing set objective function (see Figure 1).

Because of the scarcity of H−H bonding interactions in small molecules, a unique optimization procedure was used for H−H parameters. A specific training set was created using 50 H_2 molecules with bond lengths between 0.62 and 1.05 Å, and no testing set was necessary. Hubbard *U* values were not optimized for H−H, H−C, or C−C during steps 1−2, as there is little charge transfer in these interactions and preliminary tests showed that the quality of results was not sensitive to Hubbard *U* values. For the final HCNO optimization, a training set was created by combining all training sets from

https://dx.doi.org/10.1021/acs.jctc.9b00880
J. Chem. Theory Comput. 2020, 16, 1469−1481

Table 2. Final LANL-2019 HCNO Bond Integral and Overlap Integral Variables

	bond integrals $H_{ll's}$			overlap integrals $S_{ll's}$				
$ab,ll's$	$A_{ll's}^{ab,0}$	$A_{ll's}^{ab,1}$	$A_{ll's}^{ab,2}$	$B_{ll's}^{ab,0}$	$B_{ll's}^{ab,1}$	$B_{ll's}^{ab,2}$	$B_{ll's}^{ab,3}$	$B_{ll's}^{ab,4}$
HH ssσ	−10.6824	−1.0771	−0.4629	0.6525	−1.2700	−0.8434	0.0929	−0.0140
CC ssσ	−8.4755	−1.5058	−0.8548	0.3696	−1.4361	−0.4904	0.0405	0.0001
CC spσ	8.1834	−1.2278	−0.1615	−0.4107	−0.9183	−0.8204	0.2782	−0.0661
CC ppσ	7.9596	−0.6177	−0.6436	−0.3376	0.4664	−2.9836	2.0947	−0.5725
CC ppπ	−4.4916	−1.6630	−0.0347	0.2296	−1.8144	−0.3917	0.0592	−0.0157
NN ssσ	−11.2780	−1.4731	−0.4652	0.3478	−1.6401	−0.5556	0.0889	−0.0119
NN spσ	11.2069	−1.3556	−1.1013	−0.3974	−1.0636	−0.9129	0.3634	−0.0923
NN ppσ	9.2542	−0.8638	−0.4716	−0.3145	0.4626	−3.2864	2.4079	−0.6784
NN ppπ	−4.8706	−1.9166	0.0151	0.2213	−1.9561	−0.4000	0.0755	−0.0218
OO ssσ	−12.0283	−2.5182	−0.1399	0.3052	−1.9138	−0.6219	0.1441	−0.0293
OO spσ	14.2369	−1.3485	−1.0472	−0.3652	−1.2999	−0.9770	0.4429	−0.1252
OO ppσ	9.4779	−1.3516	−0.3675	−0.3006	0.1418	−3.2579	2.6054	−0.8017
OO ppπ	−4.1269	−2.7645	0.0327	0.1949	−2.1667	−0.4093	0.0875	−0.0286
HC ssσ	−9.3619	−1.0888	−0.6629	0.4708	−1.3639	−0.6537	0.0835	−0.0102
HC spσ	9.2814	−0.9450	−0.3714	−0.5295	−0.7149	−1.1039	0.3581	−0.0743
HN ssσ	−13.7512	−1.8405	−0.2968	0.4702	−1.5006	−0.7004	0.1076	−0.0148
HN spσ	10.1808	−1.1661	−0.3445	−0.5047	−0.7396	−1.1793	0.3939	−0.0812
HO ssσ	−16.0493	−1.6240	−0.7196	0.4614	−1.6260	−0.7535	0.1240	−0.0180
HO spσ	10.5828	−1.3188	−0.3153	−0.4736	−0.7769	−1.2551	0.4244	−0.0882
CN ssσ	−10.1000	−1.8842	−0.5677	0.3574	−1.5268	−0.5247	0.0624	−0.0049
CN spσ	10.0812	−1.3750	−0.4459	−0.3817	−0.9191	−0.8991	0.3324	−0.0800
NC spσ	12.1990	−1.1023	−0.6279	−0.4221	−1.0429	−0.8385	0.3065	−0.0769
CN ppσ	7.3783	−0.5508	−0.6130	−0.3233	0.4762	−3.1412	2.2498	−0.6234
CN ppπ	−4.4121	−1.8866	−0.0501	0.2243	−1.8792	−0.3970	0.0667	−0.0185
CO ssσ	−10.5524	−2.4086	−0.2484	0.3304	−1.6334	−0.5608	0.0840	−0.0107
CO spσ	8.9346	−1.2399	−0.3399	−0.3420	−0.9688	−0.9637	0.3852	−0.0972
OC spσ	12.4937	−1.2311	−0.7853	−0.4153	−1.1853	−0.8424	0.3254	−0.0878
CO ppσ	8.5873	−1.0291	−0.4842	−0.3104	0.3186	−3.1294	2.3580	−0.6867
CO ppπ	−4.4908	−2.1020	−0.0030	0.2056	−1.9700	−0.4027	0.0712	−0.0212
NO ssσ	−8.9005	−1.7812	−0.0168	0.3284	−1.7587	−0.5917	0.1153	−0.0196
NO spσ	9.9676	−1.2092	−0.8456	−0.3637	−1.1154	−0.9842	0.4214	−0.1108
ON spσ	11.7848	−1.6071	−0.3458	−0.4002	−1.2082	−0.9227	0.3886	−0.1049
NO ppσ	9.1714	−1.0313	−0.5102	−0.3053	0.3452	−3.3233	2.5305	−0.7403
NO ppπ	−4.5271	−2.4941	0.0038	0.2090	−2.0510	−0.4062	0.0819	−0.0250

steps 1−4, and the final testing set was created in the same way.

A five-fold cross validation test was performed using the final set of parameters LANL-2019 as the initial guess and allowing all 136 variables to change by 20%. The average variable precision was 0.8% with the maximum change in any variable being only 8.4%, well within the boundaries of the optimization. This demonstrates that we obtained a stable HCNO parameter set from the PSO. Final variables of the parameterization (hereafter called LANL-2019 parameters) are shown in Tables 2 and 3.

4. RESULTS AND DISCUSSION

In this section, we first discuss the impact of the relative energy component in the objective function. Then, we analyze the transferability of the parameterization on the testing set. Finally, we benchmark the optimized parameters for two large evaluation sets available in the literature, QM-9, and ANI-1, which have been used to estimate the global performance of other parametrized methodologies (see, e.g., refs 51−59).

4.1. Effect of the Relative Energy Fitting. One of the novel parts of the objective function we propose in this work is the fitting of the relative energies of a set of molecules with the same stoichiometry (isomers). In order to highlight this property, we test the performance of the objective functions for the $opt_{BE,F}$ and opt_{RE} steps. Because each step in the optimization process uses parameterization from the relative energy fitting of the previous steps, the pure effect of the opt_{RE} step can only be fully studied during the optimization of the HC parameters. Determining the relative energies of chemical compounds with the same composition of atoms but different connectivity is essential for predicting the ratios of various products of chemical reactions. Especially important is knowing which of these isomers has the lowest energy. For this reason, we tested the ability of our optimization procedure to place isomers in the correct relative energy order and to predict the lowest energy isomer. We collectively refer to both of these fitting properties as the relative energy (RE) fit.

For analysis purposes, it is convenient to establish a threshold to determine whether or not a set of molecules are in the same energetic order. When carrying out single-point energy calculations (B3LYP/cc-pVTZ) of a representative set of molecules from the QM-9 dataset, we observed that the standard deviation of the differences of the energy was approximately 0.1 eV (∼2 kcal/mol) compared to the reference data B3LYP/6-31G(2df,p). Thus, we use this as a

https://dx.doi.org/10.1021/acs.jctc.9b00880
J. Chem. Theory Comput. 2020, 16, 1469−1481

Table 3. Final LANL-2019 HCNO Pair Potential Variables and Atomic Parameters (On-Site Energies $\varepsilon_l^{0,a}$ and Hubbard U Values U_a)

	pair potentials Φ^{ab}		
Ab	$C^{ab,0}$	$C^{ab,1}$	$C^{ab,2}$
HH	0.9654	−7.5207	−3.6591
CC	1.1728	−8.8151	−6.3396
NN	1.6131	−7.2998	−1.1035
OO	1.8199	−3.3321	−5.1864
HC	0.8538	−9.2285	−5.9352
HN	2.0855	−7.6988	−5.9034
HO	2.2054	−7.0789	−1.4573
CO	1.0725	−10.0993	−5.5292
CN	1.6105	−7.4270	−2.5958
NO	0.9630	−11.7569	−12.7556
	atomic parameters		
a	$\varepsilon_s^{0,a}$	$\varepsilon_P^{0,a}$	U_a
H	−6.4835		12.7337
C	−13.7199	−5.2541	14.3245
N	−18.5565	−7.0625	15.2986
O	−23.9377	−9.0035	12.9027

tolerance level, meaning that the order of isomers is considered to be correct if molecules can be moved to their correct energy order by changing their energy by ≤0.1 eV.

While improving the RE fit, we wanted to ensure that we were not sacrificing the binding energy (BE) and force (F) fits. Table 4 compares the BE, F, and RE fits for the best HC optimization from both steps ($opt_{BE,F}$ and opt_{RE}) evaluated on the HC training set, testing set, and the QM-9 evaluation set. Only molecules containing hydrogen and carbon atoms were considered. The testing set does not contain enough isomers to show the effect of RE optimization. The RE fit is 100% for even the first optimization step. Contrarily, the QM-9 dataset was purposely created to include many sets of isomers, which allows us to illustrate the real effect of the RE fitting. The RE fit improves for both the order of isomers and determination of the lowest energy isomer in the training set when moving from $opt_{BE,F}$ to opt_{RE}. As expected, when adding relative energies into the objective function, the accuracy in the forces decreases, but only by 0.06 eV/Å for the training set. We demonstrate below with geometry optimizations using the final HCNO parameters that this magnitude of change in the forces does not result in significant changes in molecular geometries. The SDR of binding energies have negligible change, as shown with a change of 0.002 eV/atom for the training set. Additionally, we observed an improvement of 0.006 eV/atom in the BE SDR for the QM-9 evaluation set.

Table 4. Comparison of Results from the First Optimization Step of the HC Parameter Optimization, Only Fitting to Binding Energies and Forces ($opt_{BE,F}$), and the Second Step, Which Also Includes Relative Energies (opt_{RE})[a]

	$opt_{BE,F}$	opt_{RE}	change
Training Set			
lowest (%)	70	100	+30
order (%)	80	98	+28
BE SDR (eV/atom)	0.010	0.012	+0.002
F SDR (eV/Å)	0.251	0.309	+0.058
Testing Set			
lowest (%)	100	100	0
order (%)	100	100	0
BE SDR (eV/atom)	0.016	0.014	−0.002
F SDR (eV/Å)	0.258	0.305	+0.047
QM-9 Evaluation Set			
lowest (%)	74	91	+17
order (%)	62	71	+9
BE SDR (eV/atom)	0.012	0.006	−0.006
F SDR (eV/Å)	0.319	0.334	+0.015

[a]Percentages of RE results for the training set are the true values used in the objective function during optimization while a relative energy tolerance of 0.1 eV is used for analysis both the testing set and QM-9 evaluation dataset. Binding energy and force results are reported as the standard deviation of the residuals (SDR) (residuals are the difference between the DFTB prediction and DFT).

In the QM-9 dataset, there is an improvement of 17% for the correct identification of the lowest energy isomer and 9% for the order of isomers in the evaluation set. Note that there are several isomers for the same stoichiometry, in some cases more than 1000, which reflects the high quality of the sampling of the energy landscape and the good performance of the new parameterization. It is important to highlight that an improvement of 9% in the correct determination of lowest energy isomers for QM-9 indicates an improved chemical description of 13,000 molecules.

To further illustrate the effect of the opt_{RE} optimization step, we show an energy diagram for the set of isomers with the molecular formula H_8C_6 from the QM-9 evaluation (10 molecules) in Figure 2. This figure shows DFT relative energies compared with those obtained in the two steps: $opt_{BE,F}$ and opt_{RE}. In both cases, the lowest energy isomer is well described. However, the rest of the isomer energies for opt_{RE} compared to $opt_{BE,F}$ are both closer to DFT values and in the right order (using a tolerance of 0.1 eV).

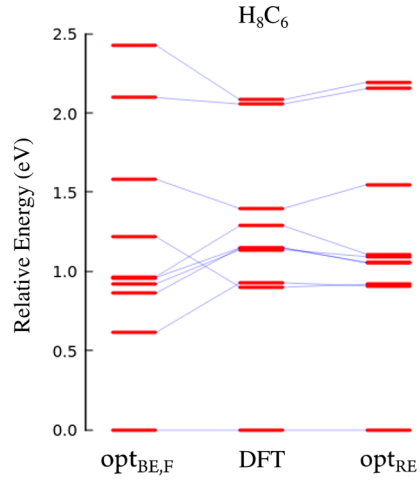


Figure 2. Energetic comparison for DFT and the two steps $opt_{BE,F}$ and opt_{RE} of the DFTB parameterization for all available isomers of H_8C_6 in the QM-9 dataset. Each red line represents a geometry optimized structure of one of the isomers and blue lines connect the same structure optimized with the different models. The success in the determination of the order of the isomers $opt_{BE,F}$ and opt_{RE} is 80% and 100%, respectively.

4.2. Testing Set. To assess the accuracy of the LANL-2019 DFTB parameters, we first used the testing set with molecular

https://dx.doi.org/10.1021/acs.jctc.9b00880
J. Chem. Theory Comput. 2020, 16, 1469−1481

properties calculated at the same level of theory as the training set. We evaluated the SDR in the binding energies and forces and the percentage of correct determination of both the lowest energy and order of the isomers. This information is shown in Table 5, where we also show the results obtained for the

Table 5. SDR for Binding Energies and Forces of HCNO Molecules in the Testing Sets and QM-9 Dataset Calculated Using New Parameters LANL-2019 and the Reference Parameters *lanl1*[14a]

	testing		QM-9	
	LANL-2019	*lanl1*	LANL-2019	*lanl1*
BE (eV)	0.019	0.031	0.014	0.030
F (eV/Å)	0.365	0.335	0.340	0.515
lowest %	97	90	83	67
order %	97	93	62	57

[a]Percent correct lowest energy and order of isomers were determined with 0.1 eV RE tolerance level.

reference parameterization from ref 14 (hereafter called *lanl1* parameters). Additionally, a correlation diagram of the binding energies for all molecules in the testing set for both the LANL-2019 and reference parameters, compared to the DFT results, is shown in Figure 3. These results indicate that the new

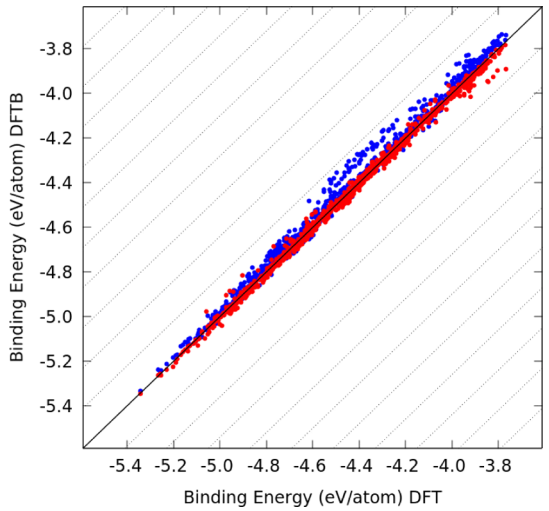


Figure 3. HCNO binding energies for the testing set. The dashed lines represent intervals of ±10% of the spread of data. Red: this work (LANL-2019), blue: *lanl1* parameters.[14]

parameterization more accurately predicts the binding energies with an improvement of 0.012 eV/atom in the SDRs. Furthermore, the reference parameters show a broader distribution and a systematic off-set from DFT results, which is not present in the new LANL-2019 parameterization. However, the SDR of the forces is slightly higher in our new parameterization (+0.030 eV/Å). The percentage of the correct determination of the lowest energy isomers and order of isomers is improved using LANL-2019 on the testing set, bringing both percentages to 97%.

The maximum deviation in the binding energies is observed for the 1-nitroso-3,5-dinitro-hexahydro-1,3,5-triazine (CAS: 5755-27-1) and its distortions (0.12 eV/atom). This molecule is complex, as it includes four functional groups. Other molecules in the set including the same functional groups do not show similar deviations. Maximum deviations in forces correspond to short triple C−C bonds (bond distances ∼1.1 Å), 3-pentyne-2-one (CAS: 7299-55-0), and 2-penten-4-yn-1-ol (CAS: 5557-88-0).

These results show that we are improving the binding energy and relative energies at the expense of sacrificing precision in the forces, which could have adverse effects for geometry optimizations and molecular dynamics simulations. In order to determine the magnitude of error caused by a decrease in precision of forces, we carried out geometry optimizations for all of the molecules in the testing dataset. Figure 4 shows the

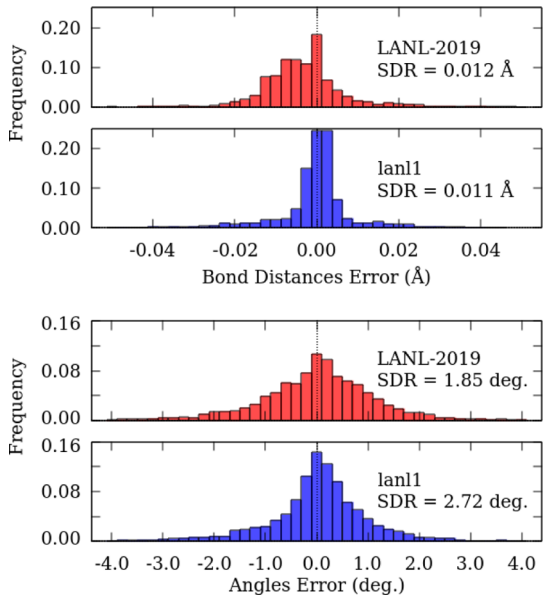


Figure 4. Histograms representing the error in the determination of the bond distances and angles of the testing set. Red bars: this work (LANL-2019), blue bars: *lanl1* parameters.[14]

error in the bond distances and angles after the geometry optimization for both the reference and new parameters compared to the DFT geometries. These results show that the slight increase of force deviation has negligible effects of stable structures. Overall, the histograms of distance and angle errors are similar for both sets of parameterizations. Despite the distributions of *lanl1* looking narrower than for LANL-2019, the presence of outliers (out of the figure's range) yields a smaller SDR value for LANL-2019. The error in the bond distances using current parameterization (SDR = 0.012 Å) is similar to the reference parameters (SDR = 0.011 Å), and for angles, we perform better improving from SDR = 2.71° with *lanl1* to SDR = 1.85° with LANL-2019. The maximum deviation in the bond distances is observed for the 4,5-dihydro-3-nitro-isoxazole-2-oxide (CAS: 4122-45-6) and the maximum deviation in the angles for the 4-methoxy-N-oxide benzonitrile (CAS: 15500-73-9). In both cases, the common motif is R−C−N−O, where the C−N distance is overestimated (0.22 Å) and the R−C−N angle changes significantly, from a linear configuration (180°) to slightly bent (152°). This does not happen in the *lanl1* parameterization.

4.3. Evaluation Sets. In order to assess the accuracy of our LANL-2019 parameterization, we carried out DFTB calculations on two benchmark datasets available in the literature, QM-9[60] and ANI-1.[56] We used each dataset to evaluate different properties of interest.

https://dx.doi.org/10.1021/acs.jctc.9b00880
J. Chem. Theory Comput. 2020, 16, 1469−1481

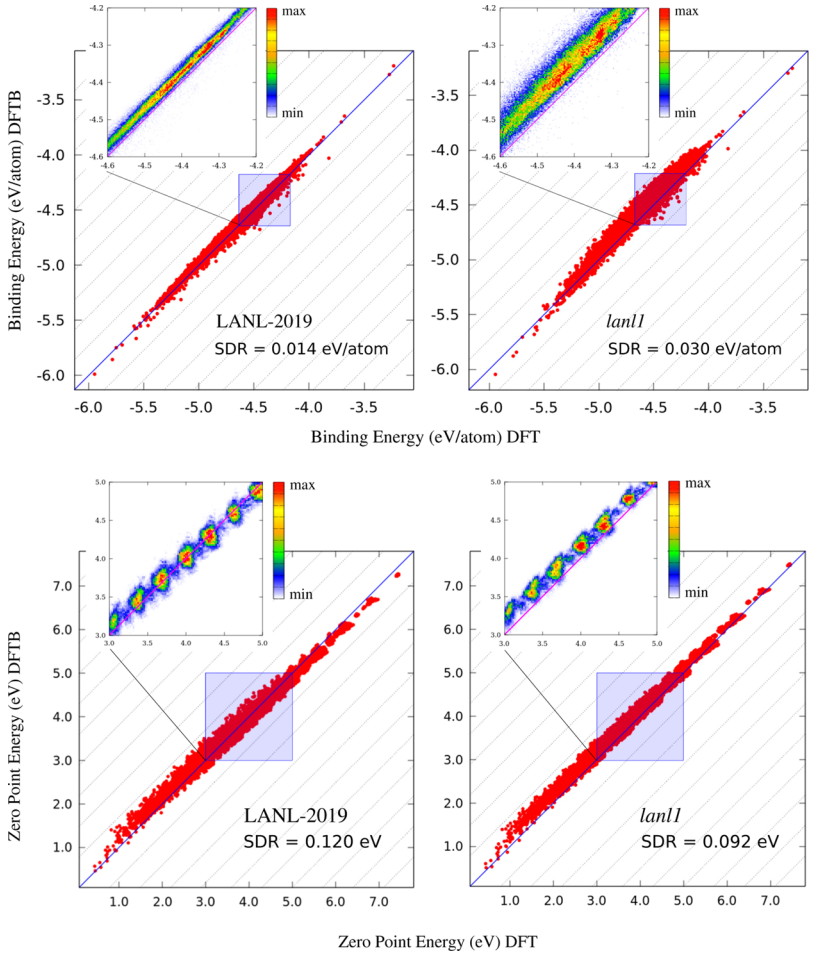


Figure 5. Binding energy (top panels) and zero point vibrational energy (bottom panels) correlation plots for the QM-9 dataset comparing results from DFTB geometry optimizations using the parameterization developed in this work (left) and *lanl1* (right). The dashed lines represent intervals of ±10% for the spread of data from the diagonal. The inset plots show the density of the points in the most populated region.

4.3.1. QM-9 Evaluation Set. This dataset contains equilibrium molecular structures (no distortions) for the first 133,657 molecules of the GDB-17 database[61] with HCNOF elements and molecular properties such as energies, dipole moments, frontier orbital eigenvalues, isotropic polarizabilities, and harmonic frequencies. The dataset corresponds to the GDB-9 subset of all neutral molecules with a singlet electronic ground state containing up to nine heavy atoms. The QM-9 dataset was created using Gaussian09[62] at the B3LYP/6-31G(2df,p) level of theory. The 134k molecules include 620 stoichiometries with several constitutional isomers, for example, 6,095 isomers for $H_{10}C_7O_2$. Filtering for atomic interactions analyzed in this study, 4,907 HC molecules, 45,770 HCO molecules, 14,192 HCN molecules, and 66,591 HCNO molecular geometries remain. Even though QM-9 was created with a different level of theory compared to current parameterization, all calculated properties should show linear correlation trends and likely a small shift in the binding energies. It is important to point out that 233 molecules were removed from the QM-9 dataset because they did not have a connected molecular graph. Two atoms were considered bonded/linked if they were separated by a distance of less than the sum of their covalent radii. The list of removed molecules is provided in the Supporting Information, Section VI.

The QM-9 dataset is used to verify that geometry optimizations with our new parameterization can reproduce both geometries and energetics of all molecules. Moreover, because of a large number of isomers contained in the QM-9 dataset, this dataset provides the opportunity to verify the improvement of our parameterization in determining the lowest energy state and the energetic order of the isomers.

As shown in Table 5, LANL-2019 improves the description of the binding energies, forces, and relative energies relative to *lanl1* for the QM-9 dataset. It is interesting to note that the forces improve for QM-9 molecules, but they did not improve using our testing set. This may be due to the fact that all of the molecules in the QM-9 dataset are equilibrium geometries, whereas the testing set included distorted geometries.

As a complementary analysis, we present correlation plots of the binding energy per atom for the QM-9 dataset after geometry optimization with DFTB (see Figure 5 top panels). Although there is a good correlation between the reference and new parameterization binding energies, there is a systematic off-set from DFT results using both sets of DFTB parameters, which is larger using the reference *lanl1* parameterization than the new LANL-2019; 0.070 and 0.031 eV/atom, respectively. The fact that this systematic off-set is not present in the training set (see Figure 3) suggests that this variation is caused by the different basis sets used during the optimization of the DFTB model and the one used in the QM-9 dataset. The reference parameters also produce a broader distribution in the binding energies than LANL-2019 parameters compared to

https://dx.doi.org/10.1021/acs.jctc.9b00880
J. Chem. Theory Comput. 2020, 16, 1469−1481

DFT results (SDR = 0.030 eV/atom and SDR = 0.014 eV/atom, respectively). Again, distributions of *lan1* look narrower, but due to outliers (out of the figure's range), the SDR values are smaller for LANL-2019. Regarding the geometric parameters shown in Figure 6, the geometry optimizations of QM-9 molecules result in a similar error in bond lengths and angles as with the testing set (Figure 4).

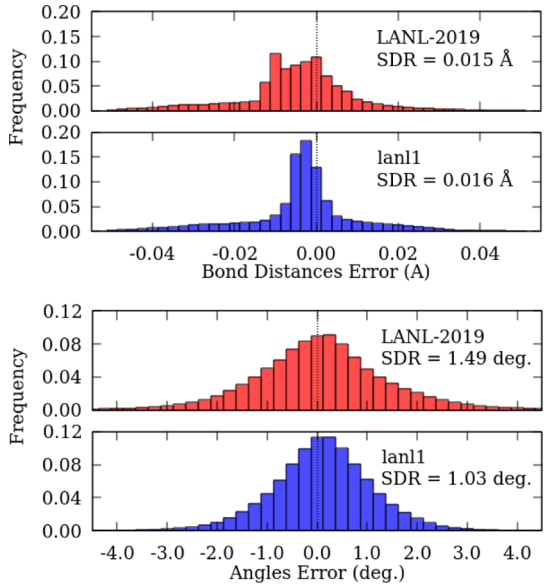


Figure 6. Histograms representing the error in the determination of the bond distances and angles of the QM-9 dataset. Red bars: this work, blue bars: *lanl1* parameters.

While all molecules in the HCNO testing set retained their DFT geometry connectivity after optimization with DFTB, 230 molecules in the QM-9 dataset changed their connectivity after DFTB optimization. These molecules are available in the Supporting Information, Section VIII. The majority of these molecules contained an N−N−O motif in the DFT geometries that broke apart during DFTB optimization. We reoptimized a subset of these molecules using NWChem with the same level of theory as was used for creating of our training and testing sets, starting from the QM-9 geometries. After this DFT optimization, about half of the molecules changed their connectivity to the DFTB predicted geometries. The other half of the molecules maintained the original QM-9 geometries. When performing the same optimization with NWChem, but starting from the DFTB geometries we obtained, all DFTB geometry connectivity remained, and the majority of the DFTB geometries were lower in the energy that the energies starting from the QM-9 geometries. This indicated that these 230 molecules from QM-9 were in metastable states with very shallow energy minima. Thus, these set of molecules are not trustworthy to evaluate the transferability of the parameterization.

Finally, the accuracy of current parameterization was tested on harmonic frequencies using the QM-9 dataset, by computing the zero-point vibrational energies (ZPE). It is important to highlight that vibrational information was not included in the objective function. Therefore, this analysis provides information about the transferability of current parameterization to nonfitted properties, specifically the calculation of the second derivatives of the energy with respect to the nuclear coordinates. There is excellent agreement in the ZPE of the QM-9 dataset using both the reference and the new DFT parameters. Results are shown in Figure 5 (bottom). The reference parameters are more precise in the determination of the ZPE than LANL-2019 (SDR = 0.092 eV and SDR = 0.120 eV, respectively). However, the reference parameters show a systematic error of about 0.2 eV that is not present in the current parameterization.

4.3.2. ANI-1 Evaluation Set. In contrast to the QM-9 dataset, ANI-1 explores the conformational space by reporting several distorted geometries. This dataset contains DFT structures and energies for more than 20 million distorted conformations of 57,462 small organic molecules. All of the electronic structure calculations were carried out with Gaussian-09[62] and the ωB97× functional in combination with the 6-31G(d) basis set. The ANI-1 dataset was built from a subset of the GDB-11 database containing molecules with between one and eight heavy atoms and limiting the atomic species to H, C, N, and O. All molecules are neutral and have a singlet electronic ground state. The nonequilibrium conformations or distortions were created by a uniform sampling along the normal modes coordinates for a certain temperature (see original paper for details ref 56).

Given the large number of distortions per molecule offered by this dataset, ANI-1 gives the opportunity to analyze the

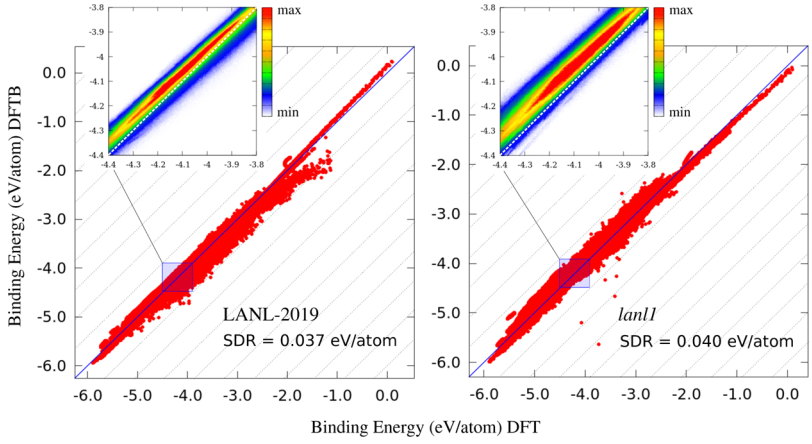


Figure 7. Binding energy correlation plots for the ANI-1 dataset comparing (a) LANL-2019 and (b) *lanl1* parameterizations. The dashed lines represent intervals of ±10% of error from the diagonal. The inset plots show the density of the points in the most populated region.

https://dx.doi.org/10.1021/acs.jctc.9b00880
J. Chem. Theory Comput. 2020, 16, 1469−1481

performance of our DFTB parameterization on extended regions farther away from the local minima and thus to determine if we get a correct reproduction of the energy landscapes.

Figure 7 shows the correlation plots of the binding energy per atom for the ANI-1 dataset from single-point energy calculations. Overall a good correlation for both the reference and new parameterization is evident. However, there are two main aspects to highlight: (1) *lanl1* shows a broader distribution in the binding energies compared to current parameterization (SDR = 0.037 eV/atom and SDR = 0.040 eV/atom, respectively); (2) there is a systematic off-set from DFT results that is smaller for current parameterization versus *lanl1* (0.044 and 0.076 eV/atom, respectively). This is similar to what we observed for the QM-9 dataset.

The maximum deviation in binding energy is observed for certain distortions of hydrazine $H_2N{-}NH_2$ (binding energy ∼−1.0 eV) and other compounds with the formulas H_3C_6N and $H_2C_4N_2O_2$ (binding energy ∼−4.0 and ∼−3.5 eV, respectively). In all these cases, the common factor is a short bond distance. The outliers of hydrazine show N−H bond distances ranging from 0.61 Å up to 0.99 Å (N−H single bond), and the other compounds show bond distances from 0.87 Å up to 1.16 Å (N ≡ N triple bond). These results indicate that repulsive barriers in bonds involving nitrogen are slightly less repulsive in current parameterization compared with the *lanl1* parameters and DFT results.

5. SUMMARY AND CONCLUSIONS

We proposed a strategy to parametrize SCC-DFTB models based on similarity measurements that quantify both the values of molecular properties (e.g., binding energy and atomic forces) and fitting to relative energies of isomers. The former utilizes the Ballester similarity index, and the latter takes advantage of the Levenshtein distance to quantify the correct order of the isomers. A cross-validation scheme is applied to detect overfitting situations and verify and quantify the transferability of the model. All of the above study was carried out using PSO as the global optimization engine.

This new strategy allowed us to improve our previous parameterization of the approximate SCC-DFTB for molecules containing hydrogen, carbon, oxygen, and nitrogen elements, considering the importance of these elements in organic chemistry. Using this proposed parameterization approach we developed the LANL-2019 HCNO parameter set, which improves the binding energies and relative energies at the expense of sacrificing precision in the forces. However, we also found that the change in the forces is not significant given that the studied molecules do not differ significantly in either bond distances or angles with respect to the DFT reference data after geometry optimization (average errors of ∼0.05 Å and ∼4.0°, respectively).

The new parameters were tested on a diverse set of molecules of chemical and biological relevance. Three datasets were used for this evaluation, namely: (1) our own testing set, containing a set of molecules chosen by hand that includes a variety of functional groups, which were calculated at the same level of theory as the training set, (2) QM-9 standard dataset,[60] which includes ∼134k geometry optimized molecules, (3) ANI-1 standard dataset, which includes 22M distorted conformations of 57k molecules.[56] We focused on the geometries, binding energies, atomic forces, and relative energies of isomers in these molecular datasets. In all cases reported here, our proposed parameterization is capable of reproducing the reference data with an accuracy that is better than previous DFTB parameterizations.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jctc.9b00880.

Details on the data preprocessing procedure, PSO method, DFTB equations, optimization protocol, molecular distortions included in DFTB optimization, list of molecules removed from QM-9 dataset due to nonconnected molecular graphs, training and testing sets used, and molecules found to have nonstable geometries in the QM-9 dataset (PDF)

■ AUTHOR INFORMATION

Corresponding Authors

Enrique R. Batista − *Theoretical Division, Los Alamos National Laboratory, Los Alamos, New Mexico 87545, United States;* orcid.org/0000-0002-3074-4022; Email: erb@lanl.gov

Ping Yang − *Theoretical Division, Los Alamos National Laboratory, Los Alamos, New Mexico 87545, United States;* orcid.org/0000-0003-4726-2860; Email: pyang@lanl.gov

Authors

Néstor F. Aguirre − *Theoretical Division, Los Alamos National Laboratory, Los Alamos, New Mexico 87545, United States;* orcid.org/0000-0001-7901-0479

Amanda Morgenstern − *Theoretical Division, Los Alamos National Laboratory, Los Alamos, New Mexico 87545, United States;* orcid.org/0000-0001-9279-0588

M. J. Cawkwell − *Theoretical Division, Los Alamos National Laboratory, Los Alamos, New Mexico 87545, United States;* orcid.org/0000-0002-8919-3368

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jctc.9b00880

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was supported by the US Department of Energy through the Los Alamos National Laboratory. Los Alamos National Laboratory is operated by Triad National Security, LLC, for the National Nuclear Security Administration of U.S. Department of Energy (contract no. 89233218CNA000001). P.Y., and E.R.B. were sponsored by the US Department of Energy, Chemical Sciences, Geosciences, and Biosciences Division, Heavy Element Chemistry program, under contract DE-AC52-06NA25396. N.F.A, A.M., and M.C were supported by the Laboratory Research and Development Program (LDRD) at LANL, under project 20170198ER. N.F.A. was also supported by a Seaborg postdoctoral fellowship through the G. T. Seaborg Institute of Los Alamos National Laboratory (LANL). We acknowledge the allocation of computer time for this project from the Environmental Molecular Sciences Laboratory of the Pacific Northwest National Laboratory in the Cascade supercomputer.

REFERENCES

(1) Hassanali, A. A.; Cuny, J.; Verdolino, V.; Parrinello, M. Aqueous solutions: state of the art in ab initio molecular dynamics. *Philos. Trans. R. Soc., A* **2014**, *372*, 20120482.

(2) González, M. A. Force fields and molecular dynamics simulations. *Ecol. Thematique Soc. Fr. Neutronique* **2011**, *12*, 169−200.

(3) Hansson, T.; Oostenbrink, C.; van Gunsteren, W. Molecular dynamics simulations. *Curr. Opin. Struct. Biol.* **2002**, *12*, 190−196.

(4) Vlachakis, D.; Bencurova, E.; Papangelopoulos, N.; Kossida, S. Current State-of-the-Art Molecular Dynamics Methods and Applications. *Advances in Protein Chemistry and Structural Biology*; Academic Press, 2014; Vol. *94*, pp 269−313.

(5) Hollingsworth, S. A.; Dror, R. O. Molecular Dynamics Simulation for All. *Neuron* **2018**, *99*, 1129−1143.

(6) Karplus, M.; McCammon, J. A. Molecular dynamics simulations of biomolecules. *Nat. Struct. Biol.* **2002**, *9*, 646.

(7) Shaw, D. E.; Dror, R. O.; Salmon, J. K.; Grossman, J. P.; Mackenzie, K. M.; Bank, J. A.; Young, C.; Deneroff, M. M.; Batson, B.; Bowers, K. J.; Chow, E.; Eastwood, M. P.; Ierardi, D. J.; Klepeis, J. L.; Kuskin, J. S.; Larson, R. H.; Lindorff-Larsen, K.; Maragakis, P.; Moraes, M. A.; Piana, S.; Shan, Y.; Towles, B. Millisecond-scale Molecular Dynamics Simulations on Anton. *Proceedings of the Conference on High Performance Computing Networking, Storage and Analysis. New York, NY, USA*, 2009; pp 65:1−65:11.

(8) Frauenheim, T.; Seifert, G.; Elsterner, M.; Hajnal, Z.; Jungnickel, G.; Porezag, D.; Suhai, S.; Scholz, R. A Self-Consistent Charge Density-Functional Based Tight-Binding Method for Predictive Materials Simulations in Physics, Chemistry and Biology. *Phys. Status Solidi B* **2000**, *217*, 41−62.

(9) Koskinen, P.; Mäkinen, V. Density-functional tight-binding for beginners. *Comput. Mater. Sci.* **2009**, *47*, 237−253.

(10) Elstner, M.; Seifert, G. Density functional tight binding. *Philos. Trans. R. Soc., A* **2014**, *372*, 20120483.

(11) Niklasson, A. M. N.; Mniszewski, S. M.; Negre, C. F. A.; Cawkwell, M. J.; Swart, P. J.; Mohd-Yusof, J.; Germann, T. C.; Wall, M. E.; Bock, N.; Rubensson, E. H.; Djidjev, H. Graph-based linear scaling electronic structure theory. *J. Chem. Phys.* **2016**, *144*, 234101.

(12) Djidjev, H. N.; Hahn, G.; Mniszewski, S. M.; Negre, C. F. A.; Niklasson, A. M. N.; Sardeshmukh, V. B. Using Graph Partitioning for Scalable Distributed Quantum Molecular Dynamics. *Algorithms* **2019**, *12*, 187.

(13) Ghale, P.; Kroonblawd, M. P.; Mniszewski, S.; Negre, C. F. A.; Pavel, R.; Pino, S.; Sardeshmukh, V.; Shi, G.; Hahn, G. Task-based Parallel Computation of the Density Matrix in Quantum-based Molecular Dynamics using Graph Partitioning. *SIAM J. Sci. Comput.* **2017**, *39*, C466−C480.

(14) Krishnapriyan, A.; Yang, P.; Niklasson, A. M. N.; Cawkwell, M. J. Numerical Optimization of Density Functional Tight Binding Models: Application to Molecules Containing Carbon, Hydrogen, Nitrogen, and Oxygen. *J. Chem. Theory Comput.* **2017**, *13*, 6191−6200.

(15) Cawkwell, M. J.; Perriot, R. Transferable density functional tight binding for carbon, hydrogen, nitrogen, and oxygen: Application to shock compression. *J. Chem. Phys.* **2019**, *150*, 024107.

(16) 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: Condens. Matter Mater. Phys.* **1998**, *58*, 7260−7268.

(17) Gaus, M.; Goez, A.; Elstner, M. Parametrization and Benchmark of DFTB3 for Organic Molecules. *J. Chem. Theory Comput.* **2013**, *9*, 338−354.

(18) Wahiduzzaman, M.; Oliveira, A. F.; Philipsen, P.; Zhechkov, L.; van Lenthe, E.; Witek, H. A.; Heine, T. DFTB Parameters for the Periodic Table: Part 1, Electronic Structure. *J. Chem. Theory Comput.* **2013**, *9*, 4006−4017.

(19) Oliveira, A. F.; Philipsen, P.; Heine, T. DFTB Parameters for the Periodic Table, Part 2: Energies and Energy Gradients from Hydrogen to Calcium. *J. Chem. Theory Comput.* **2015**, *11*, 5209−5218.

(20) Grimme, S.; Bannwarth, C.; Shushkov, P. A Robust and Accurate Tight-Binding Quantum Chemical Method for Structures, Vibrational Frequencies, and Noncovalent Interactions of Large Molecular Systems Parametrized for All spd-Block Elements (Z = 1−86). *J. Chem. Theory Comput.* **2017**, *13*, 1989−2009.

(21) Bannwarth, C.; Ehlert, S.; Grimme, S. GFN2-xTB—An Accurate and Broadly Parametrized Self-Consistent Tight-Binding Quantum Chemical Method with Multipole Electrostatics and Density-Dependent Dispersion Contributions. *J. Chem. Theory Comput.* **2019**, *15*, 1652−1671.

(22) Carbó-Dorca, R.; Mezey, P. G. *Advances in Molecular Similarity*; JAI Press: Stanford, CT, 1998; Vol. *2*.

(23) Bero, S. A.; Muda, A. K.; Choo, Y. H.; Muda, N. A.; Pratama, S. F. Similarity Measure for Molecular Structure: A Brief Review. *J. Phys.: Conf. Ser.* **2017**, *892*, 012015.

(24) Bender, A.; Glen, R. C. Molecular similarity: a key technique in molecular informatics. *Org. Biomol. Chem.* **2004**, *2*, 3204.

(25) Maggiora, G. M.; Shanmugasundaram, V. Molecular Similarity Measures. In *Chemoinformatics and Computational Chemical Biology*; Bajorath, J., Ed.; Humana Press: Totowa, NJ, 2011; Vol. *672*, pp 39−100.

(26) Rosasco, L.; Vito, E. D.; Caponnetto, A.; Piana, M.; Verri, A. Are Loss Functions All the Same? *Neural Comput.* **2004**, *16*, 1063−1076.

(27) Ballester, P. J.; Richards, W. G. Ultrafast shape recognition to search compound databases for similar molecular shapes. *J. Comput. Chem.* **2007**, *28*, 1711−1723.

(28) Levenshtein, V. I. Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Sov. Phys. Dokl.* **1966**, *10*, 707.

(29) Kennedy, J.; Eberhart, R. Particle swarm optimization. *Proceedings of ICNN'95—International Conference on Neural Networks*, 1995; pp 1942−1948.

(30) Yang, X.-S. Particle Swarm Optimization. *Engineering Optimization: An Introduction with Metaheuristic Applications*; John Wiley & Sons, Inc.: Hoboken, NJ, USA, 2010; pp 203−211.

(31) Yang, X. Swarm Optimization. *Nature-Inspired Metaheuristic Algorithms*; Luniver Press, 2010; pp 63−69.

(32) Yang, X.-S.; Deb, S.; Fong, S. Accelerated Particle Swarm Optimization and Support Vector Machine for Business Optimization and Applications. In *Networked Digital Technologies. NDT 2011. Communications in Computer and Information Science*; Fong, S., Ed.; Springer: Berlin, Heidelberg, 2011; Vol. *136*, pp 53−66.

(33) Clerc, M. Confinements and Biases in Particle Swarm Optimisation. 2006, Technical Report hal-00122799, Open archive HAL. http://hal.archives-ouvertes.fr/.

(34) 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.* **2016**, *12*, 53−64.

(35) Refaeilzadeh, P.; Tang, L.; Liu, H. Cross-Validation. *Encyclopedia of Database Systems*, 2nd ed.; Springer, 2018.

(36) 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.

(37) Bock, N.; Cawkwell, M. J.; Coe, J. D.; Krishnapriyan, A.; Kroonblawd, M. P.; Lang, A.; Liu, C.; Saez, E. M.; Mniszewski, S. M.; Negre, C. F. A.; Niklasson, A. M. N.; Sanville, E.; Wood, M. A.; Yang, P. LATTE. 2008, https://github.com/lanl/LATTE.

(38) Kenny, S. D.; Horsfield, A. Plato: A localised orbital based density functional theory code. *Comput. Phys. Commun.* **2009**, *180*, 2616−2621.

(39) Kenny, S.; Horsfield, A.; Fujitani, H. Transferable atomic-type orbital basis sets for solids. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2000**, *62*, 4899−4905.

(40) Horsfield, A. P. Efficient ab initio tight binding. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1997**, *56*, 6594−6602.

https://dx.doi.org/10.1021/acs.jctc.9b00880
J. Chem. Theory Comput. 2020, 16, 1469−1481

(41) Pyykkö, P. Additive Covalent Radii for Single-, Double-, and Triple-Bonded Molecules and Tetrahedrally Bonded Crystals: A Summary. *J. Phys. Chem. A* **2015**, *119*, 2326−2337.
(42) Linstrom, P. *NIST Chemistry WebBook, NIST Standard Reference Database 69*, 1997; http://webbook.nist.gov/chemistry/.
(43) Pence, H. E.; Williams, A. ChemSpider: An Online Chemical Information Resource. *J. Chem. Educ.* **2010**, *87*, 1123−1124.
(44) Ayers, M. ChemSpider: The Free Chemical Database2012312-ChemSpider: The Free Chemical Database. URL: www.chemspider.-com: Royal Society of Chemistry Last visited April 2012. Gratis. *Ref. Rev.* **2012**, *26*, 45−46.
(45) Valiev, M.; Bylaska, E. J.; Govind, N.; Kowalski, K.; Straatsma, T. P.; Van Dam, H. J. J.; Wang, D.; Nieplocha, J.; Apra, E.; Windus, T. L.; de Jong, W. A. NWChem: A comprehensive and scalable open-source solution for large scale molecular simulations. *Comput. Phys. Commun.* **2010**, *181*, 1477−1489.
(46) Becke, A. D. Densityfunctional thermochemistry. III. The role of exact exchange. *J. Chem. Phys.* **1993**, *98*, 5648−5652.
(47) 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: Condens. Matter Mater. Phys.* **1988**, *37*, 785−789.
(48) Vosko, S. H.; Wilk, L.; Nusair, M. Accurate spin-dependent electron liquid correlation energies for local spin density calculations: a critical analysis. *Can. J. Phys.* **1980**, *58*, 1200−1211.
(49) Stephens, P. J.; Devlin, F. J.; Chabalowski, C. F.; 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.
(50) Dunning, T. H. Gaussian basis sets for use in correlated molecular calculations. I. The atoms boron through neon and hydrogen. *J. Chem. Phys.* **1989**, *90*, 1007−1023.
(51) Gastegger, M.; Behler, J.; Marquetand, P. Machine learning molecular dynamics for the simulation of infrared spectra. *Chem. Sci.* **2017**, *8*, 6924−6935.
(52) Ragoza, M.; Hochuli, J.; Idrobo, E.; Sunseri, J.; Koes, D. R. Protein−Ligand Scoring with Convolutional Neural Networks. *J. Chem. Inf. Model.* **2017**, *57*, 942−957.
(53) Butler, K. T.; Davies, D. W.; Cartwright, H.; Isayev, O.; Walsh, A. Machine learning for molecular and materials science. *Nature* **2018**, *559*, 547−555.
(54) 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.
(55) Ramakrishnan, R.; Dral, P. O.; Rupp, M.; von Lilienfeld, O. A. Big Data Meets Quantum Chemistry Approximations: The Δ-Machine Learning Approach. *J. Chem. Theory Comput.* **2015**, *11*, 2087−2096.
(56) 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.
(57) 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.
(58) 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.
(59) Gómez-Bombarelli, R.; Wei, J. N.; Duvenaud, D.; Hernández-Lobato, J. M.; Sánchez-Lengeling, B.; Sheberla, D.; Aguilera-Iparraguirre, J.; Hirzel, T. D.; Adams, R. P.; Aspuru-Guzik, A. Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. *ACS Cent. Sci.* **2018**, *4*, 268−276.
(60) 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.
(61) 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.
(62) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenberg, J. L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery, J. A., Jr.; Peralta, J. E.; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers, E.; Kudin, K. N.; Staroverov, V. N.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, J. M.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Zakrzewski, V. G.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Dapprich, S.; Daniels, A. D.; Farkas, Ö.; Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fox, D. J. *Gaussian09* Revision E.01; Gaussian Inc.: Wallingford, CT, 2009.

https://dx.doi.org/10.1021/acs.jctc.9b00880
J. Chem. Theory Comput. 2020, 16, 1469−1481