High Accuracy Semi-Empirical Quantum Models Based on a Reduced Training Set

Cong Huy Pham,[*,†] Rebecca K. Lindsey,[†] Laurence E. Fried,[†] and Nir Goldman[†]

†*Physical and Life Sciences Directorate, Lawrence Livermore National Laboratory, Livermore, California 94550, United States*

‡*Department of Chemical Engineering, University of California, Davis, California 95616, United States*

E-mail: pham20@llnl.gov

Abstract

There exists a great need for computationally efficient quantum simulation approaches that can achieve an accuracy similar to high-level theories while exhibiting a wide degree of transferability. In this regard, we have leveraged a machine-learned force field based on Chebyshev polynomials to determine Density Functional Tight Binding (DFTB) models for organic materials. The benefit of our approach is two-fold: (1) many-body interactions can be corrected for in a systematic and rapidly tunable process, and (2) high-level quantum accuracy for a broad range of compounds can be achieved with ∼0.3% of data required for one advanced deep learning potential (ANI-1x). In addition, the total number of data points in our training set is less than one half of that used for a recent DFTB-neural network model (trained on a separate dataset). Validation tests of our DFTB model against energy and vibrational data for gas-phase molecules for additional quantum datasets shows strong agreement with reference data from either hybrid density-functional theory, coupled-cluster calculations, or experiments. Preliminary testing on graphite and diamond successfully reproduce condensed phase structures. The models developed in this work, in principle, can retain most of the accuracy of quantum-based methods at any level of theory with relatively small training sets. Our efforts can thus allow for high throughput physical and chemical predictions with up to coupled-cluster accuracy for materials that are computationally intractable with standard approaches.

Atomistic simulations are an essential tool for studies in physics, chemistry, biology, and materials science. These calculations can provide atomic level detail of physical phenomena and chemical reactions that can help to understand experimental observations. With modern computing advances, computationally intensive simulations using quantum mechanical methods such as hybrid density-functional theory (DFT), Møller-Plesset second-order perturbation theory (MP2), or coupled-cluster approaches can be performed to provide accurate descriptions of electronic and vibrational states, as well as thermodynamic quantities for a diverse set of systems. The high computational expense of these approaches, however, generally limits their application to static gas-phase clusters or small, sub-nanometer systems sizes. This can be far below the spatial scales of most experimental studies, which frequently probe nanometers or beyond and involve dynamic measurements. For example, the "gold standard" CCSD(T) (coupled-cluster considering single, double, and perturbative triple excitations) method scales as $\mathcal{O}(N^7)$, where N is the number of basis functions involved in the calculation. Consequently, it is generally only applicable to systems with tens of atoms or less, precluding its use for larger biomolecules or condensed phases. Neural network (NN) approaches can be used to completely parameterize the quantum mechanical interactions,[1–3] bypassing the need for direct electronic state calculation and thus yielding improved scaling and efficiency. However, NNs tend to require extremely large training sets and can perform poorly outside of their training regime.[1] This need for large data sets, in particular, makes such approaches challenging to implement and use effectively. Therefore, there is a widespread need for the development of computational methods that can maintain the accuracy of high level approaches while yielding substantial gains in training set size, computational costs, and scaling.

Semi-empirical methods, for example Density Functional Tight Binding (DFTB),[4–7] are a potential strong alternative quantum approach. The DFTB Hamiltonian is derived directly from an expansion of the Kohn-Sham DFT total energy, yielding a good balance between approximate quantum mechanics and empiricism. This can result in calculations requiring

only a small fraction of the computational cost compared to DFT or other high-level quantum approaches. Here, the DFTB total energy is written as:

$$E_{\mathrm{DFTB}} = E_{\mathrm{BS}} + E_{\mathrm{Coul}} + E_{\mathrm{rep}}, \tag{1}$$

where E_{BS} corresponds to the band structure energy, E_{Coul} is the charge fluctuation term, and E_{rep} is the repulsive energy. E_{BS} is calculated as a sum over occupied electronic states from the DFTB Hamiltonian. In practice, DFTB Hamiltonian matrix elements are computed from pre-tabulated Slater-Koster tables derived from reference calculations with a minimal basis set. The repulsive energy, E_{rep}, corresponds to ion-ion repulsions, as well as Hartree and exchange-correlation double counting terms. This term can be expressed as an empirical function where parameters are fit to reproduce high-level quantum or experimental reference data. A pairwise potential energy function is often used for the repulsive energy term,[8,9] though many-body interaction terms are required in some cases.[10,11] DFTB is approximately three orders of magnitude more efficient than DFT calculations and exhibits $\mathcal{O}(N^3)$ scaling. Its combination of approximate quantum mechanics with empirical functions can allow for a high degree of flexibility in terms of optimization approaches, desired accuracy, and transferability across element types and diverse conditions.[12–14] DFTB models have been created for a broad range of materials, though the repulsive energy largely has been tuned to relatively low-level DFT data for condensed phases.[15–19]

Recent efforts have been made to enhance the accuracy of DFTB through creation of more sophisticated and systematic approaches for determining the repulsive energy term.[14,20–22] Kranz *et al.* added general polynomial forms to the standard DFTB repulsive energy to describe different bond types and various chemical environments.[20] By training to energies and atomic forces of $\sim$ 150,000 structures of 2100 unique C, H, N, O, and F containing molecules in both equilibrium and distorted configurations, their model can reproduce reference atomization energies of an $\sim$ 130,000 molecule test set. Gaussian Process Regression[21]

and the Curvature Constrained Splines methodology[14] have been used to create strictly pair-wise additive repulsive energies for several organic and inorganic systems. However, these methods can struggle for systems where greater than two-body interactions in E_{rep} are needed.[14] NNs have been proposed as a promising method to include many-body interactions into the DFTB repulsive energy.[22,23] The resulting DFTB-NN models have the capability of predicting molecular properties for a wide range of compounds and element types. However, the drawback of utilizing NNs for E_{rep} is similar to their use for classical force fields in that they require large amounts of training data and can have slow parameterization due to the presence of quasi-degenerate local minima,[24] making their development exceedingly challenging.

Here, we explore the possibility of creating DFTB models that can leverage the relative simplicity of linear regression machine learning in the recently developed Chebyshev Interaction Model for Efficient Simulation (ChIMES) method. ChIMES is a many-body force field based on linear combinations of Chebyshev polynomials.[25] It has been shown that ChIMES models yield good agreement with DFT reference method for a wide range of properties and materials under both ambient and extreme conditions.[26–28] The main advantage of ChIMES is that it is completely linear in fitted coefficients, allowing for rapid parameterization to a global minimum. The reliance on Chebyshev polynomials, which are orthogonal, allows the complexity of a ChIMES model to be systematically tuned to an arbitrary degree of accuracy and transferability, while also providing straightforward methods for regularization to minimize overfitting.[16] In this study, we determine an optimal DFTB/ChIMES model for C, H, N, O-containing systems using high level quantum chemical reference data. We use an iterative scheme to systematically expand our training set where at each iteration, a small fraction of the force configurations with largest deviation in our validation set are included in the next training set iteration. The accuracy and transferability of the resulting model are investigated for a wide variety of gas-phase clusters as well as some carbon solids. We find that use of a small fraction of our chosen data set ($\sim$0.3% of similar NN efforts) yields

DFTB/ChIMES models that maintain close to hybrid functional, coupled-cluster, and/or experimental accuracy for the gas-phase clusters studies here, and compares favorably to previous DFTB-NN efforts for similar systems.

For our DFTB/ChIMES models, the total energy is determined as the sum of the standard DFTB energy with an additional ChIMES contribution:

$$E_{\mathrm{DFTB/ChIMES}} = E_{\mathrm{BS}} + E_{\mathrm{Coul}} + E_{\mathrm{rep}} + E_{\mathrm{ChIMES}}, \tag{2}$$

For this work, DFTB calculations are performed using the 3ob-3-1 parameter set, which contains a third-order expansion about the charges and is considered an optimal DFTB starting point for most organic system.[22,29] The ChIMES energy is written as a many-body expansion:

$$\begin{aligned} E_{\mathrm{ChIMES}} = &\sum_{i=1}^{n_a} E_i + \sum_{i=1}^{n_a}\sum_{j=1}^{i-1} E_{ij} + \sum_{i=1}^{n_a}\sum_{j=1}^{i-1}\sum_{k=1}^{j-1} E_{ijk} \\ &+ \sum_{i=1}^{n_a}\sum_{j=1}^{i-1}\sum_{k=1}^{j-1}\sum_{l=1}^{k-1} E_{ijkl} + \cdots, \end{aligned} \tag{3}$$

where n_a is the number of atoms in the system. The atomic energies E_i are constants used to match energies from reference data, and two-body (pairwise) energies are expressed as linear combinations of Chebyshev polynomials of the first kind.[30,31] Higher-bodied interactions are determined through products of a cluster's constituent pair-wise polynomials.[32] ChIMES parameters are determined by fitting to the difference between the reference energies and atomic forces and those computed from DFTB alone using the following objective function:

$$F_{\mathrm{obj}} = \sum_{i=1}^{n_{\mathrm{g}}} \left[\sum_{j=1}^{n_{\mathrm{a}}^{(i)}} \sum_{k=1}^{3} w_{\mathrm{F}_{ijk}}^2 \left(\Delta F_{ijk}\right)^2 + w_{\mathrm{E}_i}^2 \left(\Delta E_i\right)^2 \right] / N_d, \tag{4}$$

where N_d is the total number of data entries, given by $N_d = n_{\mathrm{g}} + \sum_{i=1}^{n_{\mathrm{g}}} 3n_{\mathrm{a}}^{(i)}$. Here the number of gas phase molecular conformations in the training set is given by n_{g} and $n_{\mathrm{a}}^{(i)}$ indicates the number of atoms in the configuration i. F_{ijk} is the k^{th} Cartesian component of the force acting

on atom j in configuration i, while E_i is the energy per atom of configuration i. The term ΔX is equal to $X_{\text{ref}} - X_{\text{DFTB}} - X_{\text{ChIMES}}$ (where $X \equiv F_{ijk}, E_i$). The subscripts "ref", "DFTB", and "ChIMES" indicate the predicted quantity X from reference method, DFTB, and the present ChIMES correction, respectively. Further details about our DFTB calculations, the ChIMES functional form, the fitting procedure and ChIMES hyper-parameter selection (including radial ranges, polynomial orders, and other pertinent details) can be found in the Supporting Information.

The dataset used to develop the DFTB/ChIMES model here is a subset of the ANI-1x dataset which we will refer to as 'sub_ANI-1x'. It contains only molecular conformations from ANI-1x computed using CCSD(T)/CBS, wB97X/def2-TZVPP, and wB97X/6-31G* levels of theory.[33] This corresponds to ∼10% of the full ANI-1x data set, and resulted in 459,464 molecular conformations of 1895 unique molecules, including transition states of some chemical reactions. Since there are no data for atomic forces at the CCSD(T)/CBS level of theory in the 'sub_ANI-1x', for our fitting purposes we use the wB97X/def2-TZVPP reference data only. We note here that fitting a DFTB/ChIMES model using a whole 'sub_ANI-1x' set directly would utilize ∼19M data points, resulting in a slow parameterization. In addition, it is of great benefit to create semi-empirical quantum approaches that do not require the traditionally vast amounts of training data needed by most machine learning approaches. As a result, our first objective is to determine how much data is needed from 'sub_ANI-1x' to train a good DFTB/ChIMES model. Starting from this dataset, we randomly selected only 1% of the total possible configurations (4594 molecular geometries) with the remainder kept for validation purposes. The performance of this iteration is labeled "circle 0" in figure 1(panel (a) and (d) for the comparison of energies and forces, respectively). The largest deviations in energies and forces come from molecular configurations where short-ranged interactions are not adequately sampled, i.e., the smallest atom pair distances in non-equilibrium molecular conformations are somewhat poorly sampled in the initial training set. We then used an iterative fitting scheme to generate the next set of

DFTB/ChIMES models.

At each "circle", an additional 1% of configurations from the validation set with the highest force deviations is added to the previous training set to generate our next iteration of the model. Model performance evaluation (based on the MAE/RMSE for energies and forces) is done on the new validation set. The process is repeated until changes in the MAE/RMSE converge. Surprisingly, we found that our approach converges quickly, since at "circle 2" the changes in MAE/RMSE for energies and forces are relatively small (less than 0.1 kcal/mol for energy and 0.1 kcal/mol-Å for force, see Supporting Information). This means that by using only 3% of the 'sub_ANI-1x' data, DFTB/ChIMES is able to reproduce the reference data in the remaining 97% of the configurations. The required training set size for this DFTB/ChIMES model (only 0.3% of ANI-1x dataset) is therefore significantly less than that of previous ANI potentials,[1–3] which generally have used at least 5M molecular conformations in the training data. This can be attributed in part to the relative accuracy of the approximate quantum mechanics in DFTB as well as the flexibility of ChIMES. Our method also leverages previous efforts in generating high accuracy and highly diverse quantum-mechanical ANI-1x dataset.[34] We note here that our final DFTB/ChIMES model contains approximately two orders of magnitude fewer parameters than a similar NN effort[2] (5546 for DFTB/ChIMES vs. 389376 for ANI-1x), also helping to reduce the needed amount of training data. The final total size of our training set (∼372K data points) is also less than the ∼800K data points used for a recent DFTB-NN model with deep tensor neural networks.[22]

To further test the transferability of our model, we compute comparisons from DFTB/ChIMES to three additional high-level quantum chemical data sets for organic materials, with reference energies/forces at the level of ωB97X (the same as in the training data) or coupled-cluster calculations (higher level of accuracy). The GDB-10to13 data set[2] consists of the molecular energies and forces at the DFT (ωB97X) level of nearly 3000 molecules containing 10-13 C, N, or O atoms. For each molecule in this data set, 12 to 24 non-equilibrium geome-

tries are generated from displacements along normal modes, yielding a total number of 47,670 configurations. Table 1 provides MAE and RMSE for DFTB and DFTB/ChIMES models on the GDB-10to13 benchmark. The standard DFTB method predicts relatively large values for MAE/RMSE. Our DFTB/ChIMES model exhibits a 60% and 45% decrease in MAE/RMSE of the energies and forces, respectively, over standard DFTB. We note that the predicted MAE/RMSE using DFTB/ChIMES is similar to values from the ANI-1 and ANI-1x potentials,[2] though here we have only trained our model on ∼14k molecular geometries compared to 22M and 5M molecular conformations used to train ANI-1 and ANI-1x, respectively. Our computed MAE/RMSE (3.57/4.72 kcal/mol for energies and 3.62/5.33 kcal/mol-Å for forces) are smaller than the variations between wB97X-DFT itself and higher levels such as CCSD(T) and MP2 (4.9/5.9 kcal/mol for energies and 4.6/5.9 kcal/mol-Å for forces).[3]

The performance of DFTB/ChIMES in comparison to coupled-cluster reference data is also provided in Table 1. Here, we have selected the ISO34 data set[35] which consists of computed energies only of 34 isomers containing the elements C, H, N, and O. The reference isomerization energies are at the CCSD(T) level of theory or contain experimental reaction enthalpies with the removal of vibrational and thermal effects. This data set has been widely used for benchmarking different computational methods, including DFT,[36] DFTB[37] and DFTB-NN_{rep} (DFTB-NN with deep tensor neural networks).[22] One can see that the accuracy of DFTB/ChIMES is much better than that for standard DFTB, is slightly improved over that from DFTB-NN_{rep}, and approaches the PBE0 data given in Reference 22. To test the performance of our model on high accuracy force data specifically, we compare DFTB/ChIMES with the CCSD(T)/cc-pVTZ data for 2000 configurations of ethanol in the GDML data set[38] (54,000 data points total). Again our DFTB/ChIMES gives an improvement over standard DFTB as MAE and RMSE are both reduced by ∼40%. A direct force comparison to DFTB-NN_{rep} or the ISO34 reference was unavailable.

To probe the smoothness in the potential energy surface from DFTB/ChIMES, we have also computed the potential energy profile for rotation around the dihedral angles in alka-

nes. The torsional profile for $n-$butane is shown in Figure 2 as a test case. The reference data are at the level of wB97X/def2-TZVPP, where the C-C-C-C dihedral angles are fixed at corresponding values and the other degrees of freedom are fully optimized. Comparison was then made to DFTB or DFTB/ChIMES single point calculations at the DFT optimized geometries. As shown, both DFTB and DFTB/ChIMES predict the relative energy of the metastable minimum (at $\pm 70^\circ$) in good agreement with wB97X reference data with deviations of less than 0.5 kcal/mol. DFTB, however, underestimates the torsional barriers by 0.9 and 1.7 kcal/mol for the lower-energy (at $\pm 120^\circ$) and main barrier (at 0°), respectively. DFTB/ChIMES is more accurate overall, with deviations of less than 0.5 kcal/mol for predicting all energy barriers discussed here.

Next, we compare the the vibrational frequency predictions of DFTB and DFTB/ChIMES on 342 gas phase molecules from the Computational Chemistry Comparison and Benchmark Database or CCCBDB (https://cccbdb.nist.gov/). The reference data is at the MP2/cc-pVTZ level of theory. We also make comparisons with several DFT methods. The functionals chosen here are wB97XD,[39] which is the same as wB97X with an additional dispersion correction, and the Perdew-Burke-Ernzerhof (PBE) functional.[40] The predicted vibrational frequencies for those DFT functionals are also taken from CCCBDB. Figure 3 shows the distribution of frequencies for each computational method. wB97XD gives good agreement with MP2 reference data with MAE/RMSE = 20/36 cm^{-1}. DFTB and PBE underestimate the vibrational stretching frequencies by about 100 cm^{-1} on average, where the MAE/RMSE are 77/114 and 61/79 cm^{-1}, respectively. DFTB/ChIMES yields smaller errors in the frequency prediction with MAE/RMSE = 36/61 cm^{-1}, showing notable improvement over PBE and comparable accuracy to wB97XD.

Though the DFTB/ChIMES model developed here is trained on molecular (gas phase) data, we have also tested its performance in reproducing the structural properties of graphite and diamond. These systems were chosen due to the fact that they contain a single element only while still probing different types of chemical bonds. Figure 4 shows the potential-

energy landscape in the a-c plane of graphite ($a = b$) with DFTB (top) and DFTB/ChIMES (bottom). The standard DFTB method is insufficient for obtaining the experimental cell geometry and the potential energy surface is very flat along the c-direction, indicative of the severe underestimation of dispersion interactions. The ChIMES correction, which was fitted to hybrid DFT reference data, captures the interlayer interactions of graphite, leading to a minimum that is in excellent agreement with experiment. The predicted density and lattice parameters of graphite and diamond using DFTB, DFTB/ChIMES, and PBE-DFT in comparison with experiment are provided in Table 2. For graphite, all computational models considered here give an accurate description of the in-plane lattice parameters. DFTB and PBE overestimate the interlayer separation ($c/2$) by over 25% and 30%, respectively, and therefore underestimate the density by over 20%. DFTB/ChIMES predicts the lattice parameters and density in excellent agreement with the experimental value, with a deviation of less than 1%. For diamond, the computed values using DFTB, DFTB/ChIMES, and PBE-DFT differ by $\sim$1% for lattice parameters and $\sim$3% for density from experimental values.

In conclusion, we have shown that ChIMES can be used to extend DFTB to hybrid functional accuracy or greater. ChIMES parameters are determined rapidly through linear optimization, creating a beyond-pairwise interaction potential for DFTB. DFTB/ChIMES has the capability of reproducing vast quantities of high-level reference data while requiring only a small fraction of it for training. The accuracy of DFTB/ChIMES is discussed for total energies, atomic forces, isomerization energies, and vibrational frequencies across the vast conformational diversity of organic molecules in several popular datasets, as well as for the dihedral rotation energy profile of $n-$butane. Preliminary testing on solid carbon allotropes at ambient conditions show that DFTB/ChIMES is able to reproduce the experimental structure of graphite (a well-known challenge for standard DFT) as well as bulk diamond properties, while having been determined from gas-phase cluster data, only. On the basis of the results presented here, DFTB/ChIMES represents a promising direction for

developing general purpose quantum models that are applicable to a wide range of materials and conditions. The small training set required by our approach, as shown in this study, could yield significant advantages for future development of computational models with a coupled cluster accuracy, significantly improved scaling, and high efficiency. The utility and ease of parameterization of DFTB/ChIMES allows for high-level quantum theory accuracy in the systems where traditional methods are far too computationally intensive for use.

Acknowledgement

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344. NG and CHP initiated and designed the project. Simulations reported here were conducted by CHP. RKL, LEF, and NG developed the ChIMES parameter fitting code and force evaluation routines. RKL and NG implemented the interface to ChIMES code in a development version of the DFTB+ package. RKL and LEF contributed valuable discussion and insights. CHP and NG wrote the manuscript. The manuscript was revised by all authors. ChIMES potential parameters are available upon request.

Supporting Information Available

References

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

(2) Smith, J. S.; Nebgen, B.; Lubbers, N.; Isayev, O.; Roitberg, A. E. Less is more: Sampling chemical space with active learning. *J. Chem. Phys.* **2018**, *148*, 241733.

(3) Smith, J. S.; Nebgen, B. T.; Zubatyuk, R.; Lubbers, N.; Devereux, C.; Barros, K.;

Tretiak, S.; Isayev, O.; Roitberg, A. E. Approaching coupled cluster accuracy with a general-purpose neural network potential through transfer learning. *Nat. Commun.* **2019**, *10*, 1–8.

(4) 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.

(5) 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.

(6) Aradi, B.; Hourahine, B.; Frauenheim, T. DFTB+, a sparse matrix-based implementation of the DFTB method. *J. Phys. Chem. A* **2007**, *111*, 5678–5684.

(7) Hourahine, B.; Aradi, B.; Blum, V.; Bonafé, F.; Buccheri, A.; Camacho, C.; Cevallos, C.; Deshaye, M. Y.; Dumitrică, T.; Dominguez, A. DFTB+, a software package for efficient approximate density functional theory based atomistic simulations. *J. Chem. Phys.* **2020**, *152*, 124101.

(8) Goldman, N.; Fried, L. E.; Koziol, L. Using force-matched potentials to improve the accuracy of density functional tight binding for reactive conditions. *J. Chem. Theory Comput.* **2015**, *11*, 4530–4535.

(9) Goldman, N.; Fried, L. E. Extending the density functional tight binding method to carbon under extreme conditions. *J. Phys. Chem. C* **2012**, *116*, 2198–2204.

(10) Goldman, N.; Goverapet Srinivasan, S.; Hamel, S.; Fried, L. E.; Gaus, M.; Elstner, M. Determination of a density functional tight binding model with an extended basis set and three-body repulsion for carbon under extreme pressures and temperatures. *J. Phys. Chem. C* **2013**, *117*, 7885–7894.

(11) Srinivasan, S. G.; Goldman, N.; Tamblyn, I.; Hamel, S.; Gaus, M. A density functional tight binding model with an extended basis set and three-body repulsion for hydrogen under extreme thermodynamic conditions. *J. Phys. Chem. A* **2014**, *118*, 5520–5528.

(12) 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.

(13) Hellström, M.; Jorner, K.; Bryngelsson, M.; Huber, S. E.; Kullgren, J.; Frauenheim, T.; Broqvist, P. An SCC-DFTB repulsive potential for various ZnO polymorphs and the ZnO–water system. *J. Phys. Chem. C* **2013**, *117*, 17004–17015.

(14) Ammothum Kandy, A. K.; Wadbro, E.; Aradi, B.; Broqvist, P.; Kullgren, J. Curvature Constrained Splines for DFTB Repulsive Potential Parametrization. *J. Chem. Theory Comput.* **2021**, *17*, 1771–1781.

(15) Koziol, L.; Fried, L. E.; Goldman, N. Using force matching to determine reactive force fields for water under extreme thermodynamic conditions. *J. Chem. Theory Comput.* **2017**, *13*, 135–146.

(16) Goldman, N.; Kweon, K. E.; Sadigh, B.; Heo, T. W.; Lindsey, R. K.; Pham, C. H.; Fried, L. E.; Aradi, B.; Holliday, K.; Jeffries, J. R. Semi-Automated Creation of Density Functional Tight Binding Models through Leveraging Chebyshev Polynomial-Based Force Fields. *J. Chem. Theory Comput.* **2021**, *17*, 4435–4448.

(17) Miró, P.; Cramer, C. J. Water clusters to nanodrops: a tight-binding density functional study. *Phys. Chem. Chem. Phys.* **2013**, *15*, 1837–1843.

(18) Vuong, V. Q.; Madridejos, J. M. L.; Aradi, B.; Sumpter, B. G.; Metha, G. F.; Irle, S. Density-functional tight-binding for phosphine-stabilized nanoscale gold clusters. *Chem. Sci.* **2020**, *11*, 13113–13128.

(19) Lindsey, R. K.; Bastea, S.; Goldman, N.; Fried, L. E. Investigating 3, 4-bis (3-nitrofurazan-4-yl) furoxan detonation with a rapidly tuned density functional tight binding model. *J. Chem. Phys.* **2021**, *154*, 164115.

(20) Kranz, J. J.; Kubillus, M.; Ramakrishnan, R.; von Lilienfeld, O. A.; Elstner, M. Generalized density-functional tight-binding repulsive potentials from unsupervised machine learning. *J. Chem. Theory Comput.* **2018**, *14*, 2341–2352.

(21) Panosetti, C.; Engelmann, A.; Nemec, L.; Reuter, K.; Margraf, J. T. Learning to use the force: Fitting repulsive potentials in density-functional tight-binding with gaussian process regression. *J. Chem. Theory Comput.* **2020**, *16*, 2181–2191.

(22) Stöhr, M.; Medrano Sandonas, L.; Tkatchenko, A. Accurate Many-Body Repulsive Potentials for Density-Functional Tight Binding from Deep Tensor Neural Networks. *J. Phys. Chem. Lett.* **2020**, *11*, 6835–6843.

(23) Zhu, J.; Vuong, V. Q.; Sumpter, B. G.; Irle, S. Artificial neural network correction for density-functional tight-binding molecular dynamics simulations. *MRS Commun.* **2019**, *9*, 867–873.

(24) Behler, J. Perspective: Machine learning potentials for atomistic simulations. *J. Chem. Phys.* **2016**, *145*, 170901.

(25) Lindsey, R. K.; Fried, L. E.; Goldman, N. ChIMES: A Force Matched Potential with Explicit Three-Body Interactions for Molten Carbon. *J. Chem. Theory Comput.* **2017**, *13*, 6222–6229.

(26) Lindsey, R. K.; Fried, L. E.; Goldman, N. Application of the ChIMES Force Field to Nonreactive Molecular Systems: Water at Ambient Conditions. *J. Chem. Theory Comput.* **2018**, *15*, 436–447.

(27) Lindsey, R. K.; Goldman, N.; Fried, L. E.; Bastea, S. Many-body reactive force field development for carbon condensation in C/O systems under extreme conditions. *J. Chem. Phys.* **2020**, *153*, 054103.

(28) Pham, C. H.; Lindsey, R. K.; Fried, L. E.; Goldman, N. Calculation of the detonation state of HN_3 with quantum accuracy. *J. Chem. Phys.* **2020**, *153*, 224102.

(29) Gaus, M.; Goez, A.; Elstner, M. Parametrization and benchmark of DFTB3 for organic molecules. *J. Chem. Theory Comput.* **2013**, *9*, 338–354.

(30) Lindsey, R. K.; Kroonblawd, M. P.; Fried, L. E.; Goldman, N. *Computational Approaches for Chemistry Under Extreme Conditions*; Springer, 2019; pp 71–93.

(31) Armstrong, M. R.; Lindsey, R. K.; Goldman, N.; Nielsen, M. H.; Stavrou, E.; Fried, L. E.; Zaug, J. M.; Bastea, S. Ultrafast shock synthesis of nanocarbon from a liquid precursor. *Nat. Commun.* **2020**, *11*, 1–7.

(32) Lindsey, R. K.; Fried, L. E.; Goldman, N.; Bastea, S. Active learning for robust, high-complexity reactive atomistic simulations. *J. Chem. Phys.* **2020**, *153*, 134117.

(33) Chai, J.-D.; Head-Gordon, M. Systematic optimization of long-range corrected hybrid density functionals. *J. Chem. Phys.* **2008**, *128*, 084106.

(34) Smith, J. S.; Zubatyuk, R.; Nebgen, B.; Lubbers, N.; Barros, K.; Roitberg, A. E.; Isayev, O.; Tretiak, S. The ANI-1ccx and ANI-1x data sets, coupled-cluster and density functional theory properties for molecules. *Sci. Data* **2020**, *7*, 1–10.

(35) Grimme, S.; Steinmetz, M.; Korth, M. How to compute isomerization energies of organic molecules with quantum chemical methods. *J. Org. Chem.* **2007**, *72*, 2118–2126.

(36) Huenerbein, R.; Schirmer, B.; Moellmann, J.; Grimme, S. Effects of London dispersion on the isomerization reactions of large organic molecules: a density functional benchmark study. *Phys. Chem. Chem. Phys.* **2010**, *12*, 6940–6948.

(37) Gruden, M.; Andjeklović, L.; Jissy, A. K.; Stepanović, S.; Zlatar, M.; Cui, Q.; Elstner, M. Benchmarking density functional tight binding models for barrier heights and reaction energetics of organic molecules. *J. Comput. Chem.* **2017**, *38*, 2171–2185.

(38) Sauceda, H. E.; Gastegger, M.; Chmiela, S.; Müller, K.-R.; Tkatchenko, A. Molecular force fields with gradient-domain machine learning (GDML): Comparison and synergies with classical force fields. *J. Chem. Phys.* **2020**, *153*, 124109.

(39) Chai, J.-D.; Head-Gordon, M. Long-range corrected hybrid density functionals with damped atom–atom dispersion corrections. *Phys. Chem. Chem. Phys.* **2008**, *10*, 6615–6620.

(40) Perdew, J. P.; Burke, K.; Ernzerhof, M. Generalized gradient approximation made simple. *Phys. Rev. Lett.* **1996**, *77*, 3865.

(41) Zhao, Y. X.; Spain, I. L. X-ray diffraction data for graphite to 20 GPa. *Phys. Rev. B* **1989**, *40*, 993.

(42) Bucko, T.; Hafner, J.; Lebegue, S.; Angyán, J. G. Improved description of the structure of molecular and layered crystals: ab initio DFT calculations with van der Waals corrections. *J. Phys. Chem. A* **2010**, *114*, 11814–11824.

(43) Occelli, F.; Loubeyre, P.; LeToullec, R. Properties of diamond under hydrostatic pressures up to 140 GPa. *Nat. Mater.* **2003**, *2*, 151–154.

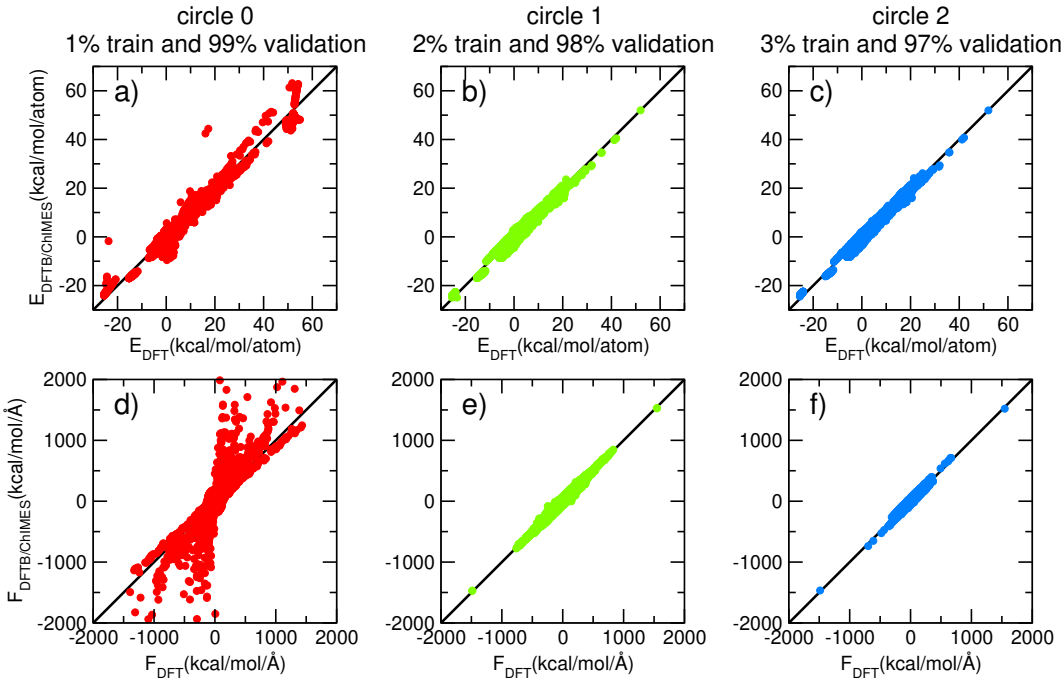


Figure 1: Comparison of energies per atom (top panels) and forces (bottom panels) predicted by DFT (ωB97X) and DFTB/ChIMES for all configurations in the validation set. The dataset used here is 'sub_ANI-1x', $\sim$10% of the full ANI-1x. The black line shows perfect correlation.

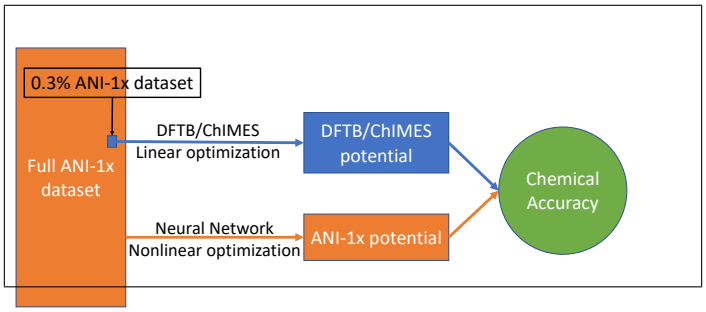
Table 1: Performance of DFTB and DFTB/ChIMES in predicting reference energies and/or atomic forces in the GDB-10to13, ISO34, and GDML data set. $MAE_E/RMSE_E$ and $MAE_F/RMSE_F$ are in kcal/mol and kcal/mol-Å, respectively. Reference molecular energies and atomic forces in the GDB-10to13 data set are at the ωB97X/6-31G* level of theory. Isomerization energies in the ISO34 data set are a mixture of experimental- and CCSD(T) extrapolation energies. The CCSD(T)/cc-pVTZ atomic forces of 2000 configurations of ethanol in the GDML data set are used for comparison.

	GDB-10to13		ISO34	GDML
method	$MAE_E/RMSE_E$	$MAE_F/RMSE_F$	$MAE_E/RMSE_E$	$MAE_F/RMSE_F$
DFTB	9.10/11.70	6.34/9.85	3.69/4.96	4.52/6.12
DFTB/ChIMES	3.57/4.72	3.62/5.33	2.06/2.56	2.72/3.61
ANI-1[2]	3.12/4.74	3.96/7.09	-	-
ANI-1x[2]	2.30/3.21	3.67/6.01	-	-
DFTB-NN_{rep}[22]	-	-	2.21/3.30	-
PBE0[22]	-	-	1.82/2.48	-

Table 2: Comparison of predicted density and lattice parameters of graphite and diamond for DFTB, DFTB/ChIMES, PBE-DFT with experimental data.

phase	method	density (g/cm^3)	a(Å)	$c/2$(Å)
graphite	Expt.[41]	2.26	2.462	3.356
	PBE-DFT[42]	1.71	2.470	4.420
	DFTB/ChIMES	2.25	2.461	3.379
	DFTB	1.77	2.474	4.248
diamond	Expt.[43]	3.51	3.567	
	PBE-DFT[9]	3.48	3.580	
	DFTB/ChIMES	3.42	3.600	
	DFTB	3.42	3.600	

Graphical TOC Entry



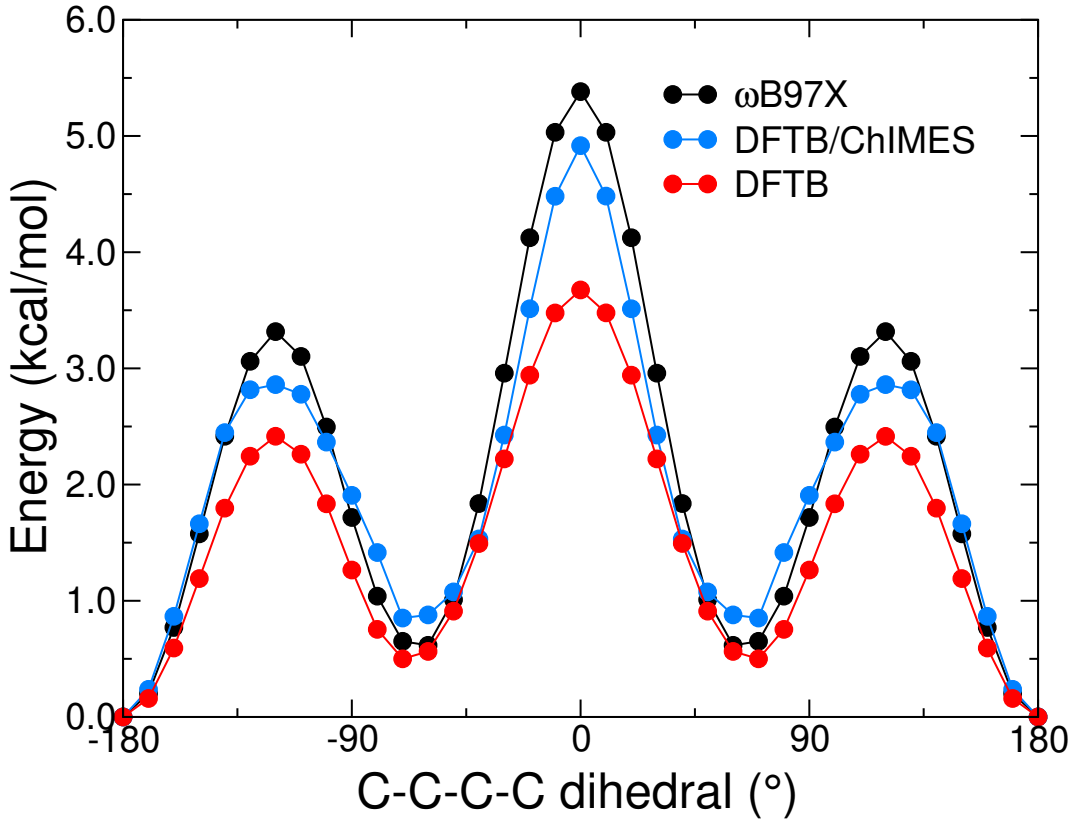


Figure 2: One dimensional potential energy profile for C-C-C-C dihedral angles of $n-$butane. The molecular configurations are fully relaxed at fixed dihedral angles in the DFT (wB97X) calculations. The single point energy calculations at DFT optimized geometries are performed for DFTB and DFTB/ChIMES.

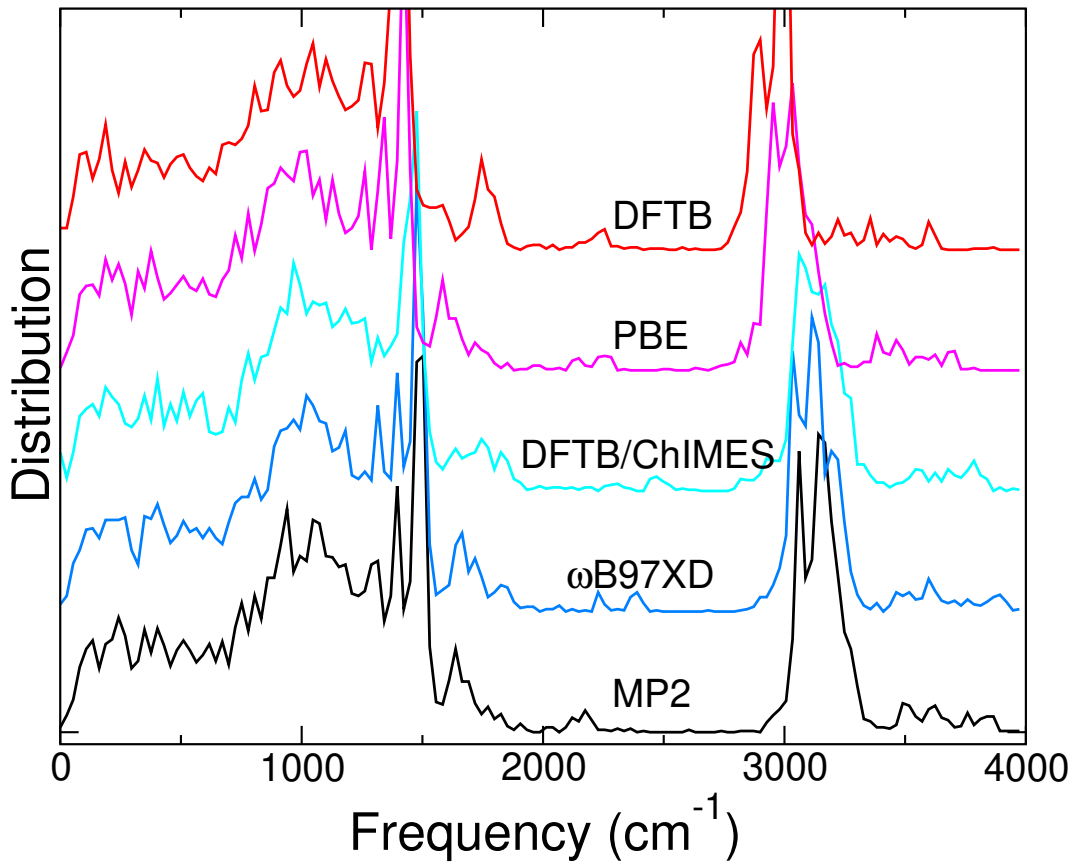


Figure 3: The distribution of the calculated frequency values using DFTB and DFTB/ChIMES for 342 neutral molecules taken from the CCCBDB database. The MP2 and DFT (PBE and wB97XD) calculations using cc-pVTZ basis set in the CCCBDB are selected for comparison.

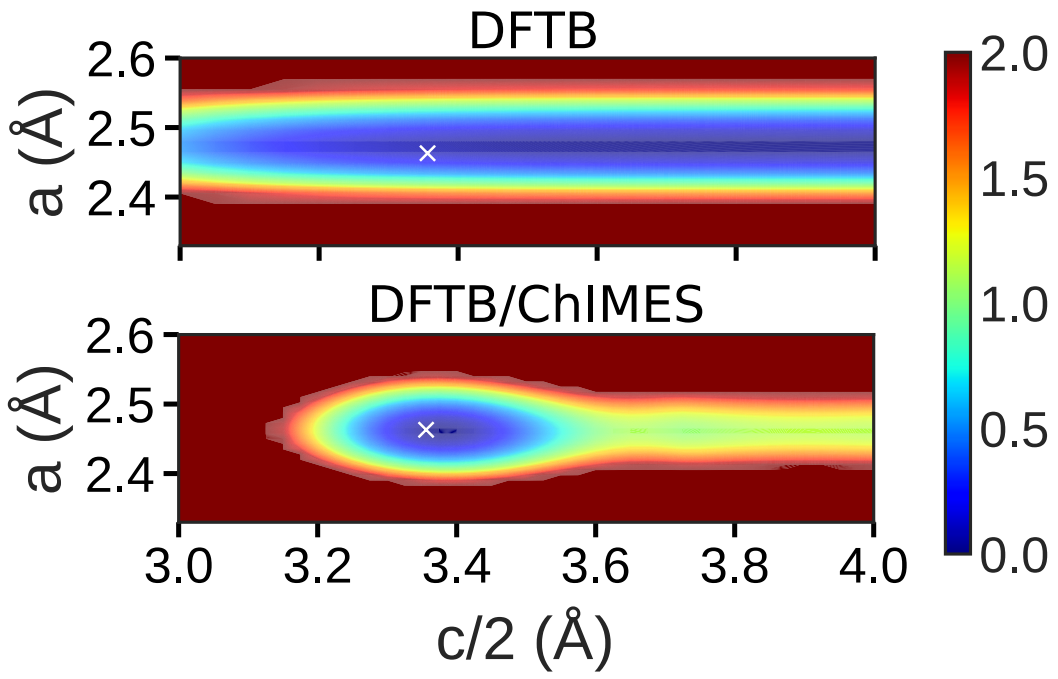


Figure 4: Potential-energy surfaces (in kcal/mol) for the *a*-*c* plane of graphite obtained using: DFTB (top) and DFTB/ChIMES (bottom). Experimental lattice parameters are marked by a cross.