JCTC Journal of Chemical Theory and Computation

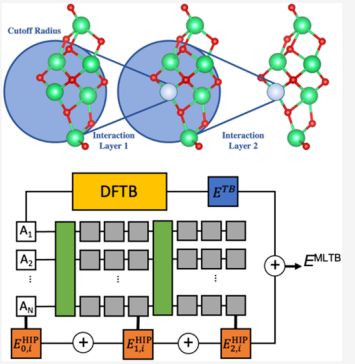
pubs.acs.org/JCTC Article

MLTB: Enhancing Transferability and Extensibility of Density Functional Tight-Binding Theory with Many-body Interaction Corrections

Daniel J. Burrill,[‖] Chang Liu,[‖] Michael G. Taylor, Marc J. Cawkwell, Danny Perez,* Enrique R. Batista,* Nicholas Lubbers, and Ping Yang*

Cite This: https://doi.org/10.1021/acs.jctc.4c00858 Read Online

ABSTRACT: We present a hybrid semiempirical density functional tight-binding (DFTB) model with a machine learning neural network potential as a correction to the repulsive term. This hybrid model, termed machine learning tight-binding (MLTB), employs the standard self-consistent charge (SCC) DFTB formalism as a baseline, enhanced by the HIP-NN potential as an effective many-body correction for short-range pairwise repulsive interactions. The MLTB model demonstrates significantly improved transferability and extensibility compared to the SCC-DFTB and HIP-NN models. This work provides a practical computational framework for developing reliable SCC-DFTB models with additional many-body corrections that more closely approach the DFT level of accuracy. We illustrate this method with the development of an accurate model for the thorium−oxygen system, applied to the study of its nanocluster structures $(ThO_2)_n$.



1. INTRODUCTION

Discovering fast and accurate approximations for quantum mechanical calculations has provided a host of different methods to gain insights into processes otherwise not approachable with more computationally intensive theories. In particular, Kohn−Sham density functional theory (KS-DFT) provides a good step in that direction with a reasonable compromise between accuracy, computational speed, and electronic structure for predicting energetics and properties of molecular and bulk systems.[1] However, even at state-of-the-art supercomputer facilities, the computational cost of DFT still presents a barrier to its wide applicability for the simulation of large systems, such as those needed for studying realistic solution environments, defects in solids, and nanostructures, to name only a few. Fast semiempirical models of the electronic structure of molecules and materials, which typically include a parametrized two-center Hamiltonian with a linear combination of atomic orbitals (LCAO) basis set, have enjoyed widespread use in computational chemistry and materials science for decades.[2−8] The tight-binding bond model (TBBM)[3] and the density functional tight-binding (DFTB) framework[9] sought to put tight-binding (TB) models on a more solid theoretical footing by making explicit connections to Kohn−Sham DFT. Later approaches, where the Kohn−Sham DFT total energy was expanded in terms of charge fluctuations, led to the development of self-consistent TB models,[8,10] such as self-consistent charge (SCC) DFTB theory, which enabled the inclusion of charge transfer, long-range electrostatics, and/or spin polarization alongside the formation of covalent or metallic bonds.[11,12]

Interatomic forces and cohesion in TB models have fundamentally many-body characters through the action of the many-body density matrix, P, even though the TB Hamiltonian itself is constructed from sums of two-center (i.e., pairwise) terms. The density matrix is computed from the electronic occupancy and eigenvectors of the TB Hamiltonian. In addition, more complex many-body interactions are captured by nonorthogonal TB models, like DFTB, when the generalized eigenvalue problem is transformed to a regular eigenvalue problem.[13,14] Nevertheless, despite the many-body, quantum mechanical nature of interatomic bonding that is captured naturally within TB models, the short-range repulsion between atoms that arises from Pauli exclusion has often been described by a sum of atom-centered pair potentials that provide no additional many-body character. The TBBM model[3] provided the first theoretical justification of why the approximation of pairwise short-range repulsion works as well as it does, and it has been shown that augmenting the TB total energy by many-body repulsive terms[15,16] or embedding

Received: July 4, 2024
Revised: January 10, 2025
Accepted: January 13, 2025

functions[17] can provide significant improvements in accuracy and transferability, albeit at the expense of a more complex parametrization procedure. In this work, we have augmented the SCC-DFTB total energy by a many-body machine-learned neural network (NN) potential. This term is parametrized numerically to the differences between the binding energies and forces obtained from a baseline SCC-DFTB model and those obtained from high-quality DFT reference data. The NN potential is computationally fast with respect to both DFT and DFTB calculations, adding only 0.7% to the computational time in a 96-atom cluster, and it is shown, when applied to model the interatomic forces in ThO_2 nanoclusters, to yield significant improvements in the accuracy and transferability with respect to both standalone DFTB and standalone NN models.

Since DFTB relies on careful parametrization,[18,19] it can be computationally costly and time-consuming to generate training data and perform parameter optimization. Three key challenges for DFTB parametrizations come from (1) applying the parameters to local chemical environments not explicitly included in the training data (transferability), (2) on nonlocal environments in systems larger than those in the training set (extensibility), and (3) insufficient flexibility in the underlying functional form, such as a lack of many-body short-range repulsion. Recently, there have been a number of solutions proposed to improve DFTB through the incorporation of machine learning (ML). For example, ML models were used to incorporate many-body effects in water clusters.[20] That study showed that ML models performed nearly the same as an explicit many-body formalism when trained to a data set spanning a relevant potential energy surface (PES). This result begs the question of how best to use machine learning to incorporate additional many-body effects into the DFTB. A separate study approached the correction of semiempirical methodologies by investigating the effect of incorporating delta-learning models to improve property prediction between levels of theory.[21] That study found improvement in the prediction of atomization energies in organic molecules using a semiempirical (PM7) delta-learning model compared with the ML model alone. This demonstrated that physics-based levels of theory can be effectively used in concert with ML models. Beyond the PM7 example, a number of schemes have been developed to augment DFTB with machine learning models.[22−31] These tend to focus either on improving the DFTB theory as a whole or on improving just the repulsive potential. A Behler−Parinello neural network (BPNN) used to introduce corrections to DFTB energies for the glycine molecule in a fashion similar to refs 20,22. That study found that the BPNN delta models were able to more accurately predict rotational energy barriers compared with DFTB alone.

Improvements to the repulsive potential of DFTB3 were sought using a deep tensor neural network (DTNN),[23,32] obtaining an improvement over the existing DFTB3 parametrization for molecular atomization and relative energies. More recently an open-source software package was introduced, TBMaLT,[33] with PyTorch-native implementations of DFTB methods, allowing for simultaneous co-optimization and implementation of both ML and DFTB methods. However, each of these investigations melding DFTB with ML interatomic potentials limited the chemical space to small organic molecules. Beyond small organic molecules, the many-body effects found in the realm of nanoclusters, such as those in the transition from molecular to bulk interactions represents an active area of investigation for DFTB[34] but have yet to be explored with DFTB + ML methods.

Herein, we examine how the delta-learning scheme can be applied to the improvement of DFTB by using a machine learning model to incorporate many-body interactions. We use the hierarchically interacting particle neural network (HIP-NN), which is a type of message-passing neural network, which was developed for modeling molecular energies and potential energy surfaces with an intrinsically many-body approach.[35,36] The hierarchical structure of the HIP-NN model allows one to regularize the long-range many-body predictions in the model, improving the accuracy of the model.[35] HIP-NN was subsequently augmented with a tensor sensitivity strategy for including angles into its message-passing function, which improved the performance significantly and showed very strong transferability and extensibility using the ANI-1x data set of molecular energies.[36] To the best of our knowledge, this is the first time the HIP-NN model is used to augment a DFTB model. Partially optimized DFTB parameters were coupled with HIP-NN to create a general framework for improving DFTB through additional many-body interactions in a scheme called machine learning DFTB tight-binding (MLTB).

The methodology is applied to predictions of ThO_2 nanoclusters in the range from small molecules, to nanostructures, to solid ThO_2. Although we applied the methodology to the Th−O system as an example, the parametrization procedure, analysis, and results are general enough that studies of other systems can also benefit from them. The choice of studying an actinide system was not arbitrary but picked in a region of the periodic table where DFTB tends to underperform and the augmentation is more justified. This approach is found to not only yield better accuracy but also save development time in the parametrization process.

2. COMPUTATIONAL METHODS

Exploring the space of possible configurations of large heavy-metal systems such as nanoparticles is an inherently expensive process. Besides the computational cost imposed by the number of atoms, the large number of valence electrons with multiple angular momenta yields a potential energy surface (PES) with many local minima, a number that grows significantly as the radius of the nanoparticle increases.[37] In this work, we propose MLTB as a computationally efficient method for such complicated systems, fast enough to search for global minima and accurate at the DFT level of accuracy.

2.1. SCC-DFTB Model and Parameterization. Since the general DFTB theory has been extensively reviewed in the literature,[9,11,38−45] we will only outline the main ideas of DFTB that are closely related to this work. The expression for SCC-DFTB total energy can be derived by expanding the DFT total energy in terms of atomic charges and truncating up to the second order terms.[11] The resulting total energy expression is

$$u = 2\mathrm{Tr}[(\mathrm{P} - \mathrm{P}_0)\mathrm{H}^0] + \frac{1}{2}\sum_{i=1}^{N}\sum_{j\neq i=1}^{N}\gamma_{ij}q_iq_j + \frac{1}{2}\sum_{i=1}^{N}\sum_{j\neq i=1}^{N}\Phi_{ij}(R_{ij}) \quad (1)$$

The first term of eq 1 is the bond energy,[46] which depends on the Slater−Koster Hamiltonian, H^0, the density matrix for noninteracting atoms, P_0, and the self-consistent density matrix P. The diagonal elements of H^0 are the energies of the valence

https://doi.org/10.1021/acs.jctc.4c00858
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

orbitals, known as the onsite energies, for free atoms, $H^0_{i\alpha,i\alpha} = \epsilon_{il}$, where i is the atom index and $l = s, p, d, f$ denote the angular momentum of the orbital α. The off-diagonal terms are constructed from the Slater−Koster transformation of the two-center bond integrals, $h_{ll'\tau}(R_{IJ})$, for the given pair of angular momentum ($l, l' = s, p, d, f$) and different bonding types ($\tau = \sigma, \pi, \delta, \phi$), and we employ a nonorthogonal basis of atom-centered orbitals.[2,47]

The second term of eq 1, the Coulomb energy, depends on the Mulliken charge on each atom, q_i, and the screened Coulomb potential γ_{ij}, which depends on the Hubbard U parameter for different elements.[11] The last term of eq 1 is the sum of empirical repulsive pair potentials, Φ_{ij}, between atom pairs. The pair potential term accounts for all contributions to the binding energy missing from the previous two terms, particularly repulsion arising from orbital overlap.

All DFTB calculations were carried out with the Los Alamos Transferable Tight-binding for Energetics (LATTE) package using the table parameter format.[48] With the table format, the two-center Slater−Koster bond integrals, overlap integrals, and pair potentials are represented by cubic splines, while the remaining one-center parameters are represented by floating point numbers.

In this work, three sets of DFTB parameters from our previous work for ThO_2 nanoclusters were employed as the baseline for the MLTB model.[49] The three sets, which are labeled as LANL-ThO-i, LANL-ThO-p, and LANL-ThO-f, correspond to the parameters from the initial guess, preoptimization (stage 1), and full optimization (stage 5). Details for the generation of these parameters are described in ref 49. We will describe only the major difference about the three parameter sets here.

The LANL-ThO-i set has Slater−Koster bond integrals fixed at the original DFT values, while the pair potentials were optimized against a training set. The training set employed by the LANL-ThO-i is the same as LANL-ThO-p, where 1109 structures up to Th_3O_6 are included. The LANL-ThO-p and LANL-ThO-f sets have both Slater−Koster bond integrals and pair potentials optimized. The parametrization of LANL-ThO-p requires less computational resources than LANL-ThO-f because the latter employed a larger training set (6741 structures up to Th_6O_{12}) and a more flexible bond integral model with more tunable parameters (31 vs 71). Details about the DFTB parameter generation is beyond the scope for this work. Readers can consult ref 49 for details. In general, the series LANL-ThO-i, LANL-ThO-p, and LANL-ThO-f provides baseline models that improve systematically.

2.2. Correcting with HIP-NN Potentials. DFTB has been employed with success for many applications.[45] However, the approximation that the short-range repulsion is purely pairwise in character may be insufficiently flexible to reproduce the DFT training data with the required accuracy. There have been several recent investigations into correcting this by including a many-body correction to DFTB as discussed above,

$$E_{\mathrm{MBDFTB}} = E_{\mathrm{DFTB}} + E_{\mathrm{MBCorrection}} \quad (2)$$

where the many-body correction can be included at any suitable level of theory. Machine learning corrections through the use of Δ-learning have been employed, where the DFTB energies are supplemented with learned differences from a reference level of theory. This work follows such an approach, introducing corrections to the DFTB functional via the HIP-NN potentials to improve the predictions of the atomization energy,

$$E_{\mathrm{MLTB}} = E_{\mathrm{DFTB}} + E_{\mathrm{HIP-NN}} \quad (3)$$

The details of HIP-NN can be found in the literature,[35,36] so a brief overview will be presented here for completeness. In order to introduce corrections to DFTB, many models have been focusing on local contributions to molecular energy with the goal of identifying what constitutes an appropriate atomic local environment.[50−52] While many methods focus on explicit functional forms of the interactions between atoms, HIP-NN takes the approach that the features can be learned through a series of interaction steps, as shown in Figure 1. Beginning as a

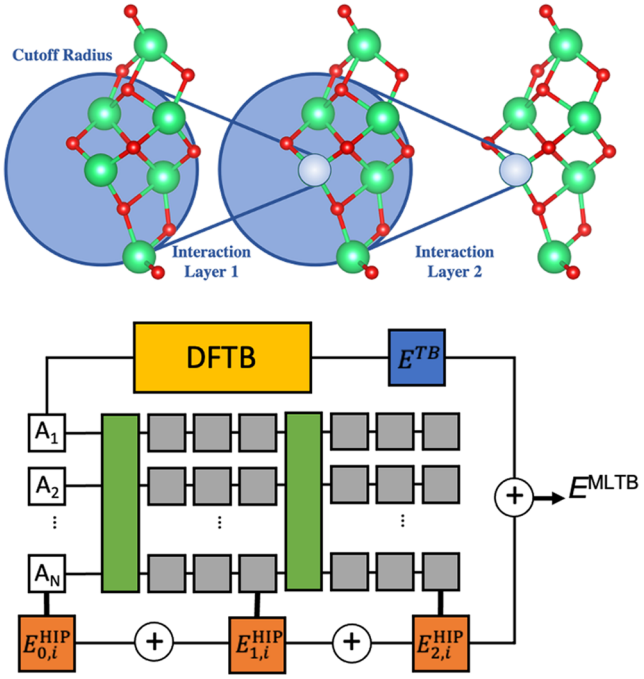


Figure 1. Schematic of HIP-NN interaction layers[35] for ThO_2 nanoclusters. Onsite atomic features are illustrated by the light blue circles in the middle and right figures. Interactions layers incorporate atomic features from all atoms within the cutoff radius to provide effective many-body interactions. Multiple interaction layers are employed within HIP-NN, which transfer atomic information from outside of the cutoff radius back to the onsite atomic feature with diminished influence due to its distance. Figure adapted with permission from J. Phys. Chem. 2018, 148, 21,7475. Copyright American Institute of Physics (2018).

one-hot-encoded vector of an atomic number, denoted in the A_n boxes, the features of a single atom are updated by calculating the interaction with other feature vectors within a given radius (cutoff radius). The novelty of HIP-NN is that it incorporates interactions through two main mechanisms: the onsite (gray boxes) and interaction layers (green boxes). Onsite layers learn how the features of a particular atom can be altered to improve the prediction, while interaction layers learn how atoms in the vicinity change the features. By iterating through onsite and interaction layers, the features of any given atom are influenced by atoms further away and outside of the radius of the initial interaction layer. The metric by which these components are fitted is the atomization energy. Before each interaction layer, the atomic contributions to the atomization energy are summed to compute the many-body

https://doi.org/10.1021/acs.jctc.4c00858
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

terms, as denoted by the orange boxes. The first energy term corresponds to single body, atomic contributions, while each subsequent box incorporates a higher order many-body correction. Many-body interactions are therefore included through the use of multiple interaction layers with a diminishing influence at longer distances.

Here, the HIP-NN models are trained to the error between DFT and DFTB energies to enable combined Δ-learning models. One advantage of using HIP-NN to provide corrections in Δ-learning is that its computational cost scales linearly with the number of atoms. This opens the possibility of applying a data-driven approach to improving the accuracy of DFTB calculations that otherwise would either not be accessible through many-body interactions or require time-consuming tuning of the DFTB parameters to accommodate all of the interactions in the PES. In total, for this work, this amounts to generating a set of minimally optimized DFTB parameters (circumventing most of the computational effort of optimizing all DFTB parameters), followed by Δ-learning with HIP-NN. To mitigate the potential for overfitting, the size of the training set is increased when training the HIP-NN models to ensure that the number of parameters remains smaller than the number of data points and to better describe the full PES and decrease the amount of extrapolation performed by HIP-NN. This approach retains knowledge of the quantum interactions among atoms from DFTB with improved energies and forces from HIP-NN.

As mentioned above, three DFTB models were employed as the baseline for the MLTB models. We will label the three MLTB models as MLTB-ThO-i, MLTB-ThO-p, and MLTB-ThO-f for MLTB models with LANL-ThO-i, LANL-ThO-p, and LANL-ThO-f as baselines. As will be discussed below and in the Supporting Information, the MLTB-ThO-p provides the best balance between computational cost for parametrization and accuracy. As a result, we refer to MLTB-ThO-p simply as MLTB in the following text. For comparison with pure ML methodology without electronic structure, a pure HIP-NN model was parametrized using the same training set. This model will be labeled as HIP-NN.

2.3. Computational Details. In combining the DFTB methodology with the HIP-NN, we sought to take advantage of the flexibility and relatively computationally cheap Δ-learning step, by minimizing the amount of fitting of the DFTB parameters. In application to modeling the PES of ThO_2 nanoparticles, this yields a methodology of low computational cost as required to search for global minima. Reference data was generated using the ADF code[53] with the Perdew−Burke−Ernzerhof (PBE) functional[54] and large core triple-ζ basis set.[55] Relativistic corrections were included at the ZORA level of theory.[56] This combination of methodology has been shown to perform well for geometries and energies of heavy-element systems.[57−63] The set was built around testing two main factors of the parameters: transferability and extensibility. Transferability refers to the ability of the parameters to show low error when applied to structures similar to those trained on but not explicitly trained on, while extensibility is demonstrated by applying the parameters to larger clusters than that exists in the training set. A data set with 22,859 structures was generated which includes DFT and DFTB optimized structures and distorted structures from those minima with cluster sizes ranging from 1 to 8, 12, and 16 ThO_2 units. The electronic state of $(ThO_2)_n$ was restricted to the singlet state. We are cognizant that under some circumstances, such as Th with a peroxo motif (Th−O−O), the electronic state may switch to triplet. However, in all cases, the electronic state of DFTB was kept consistent to that of DFT, even if it represented an excited state. For each minimum energy structure, five structures were generated by displacing the *x*,*y*,*z*-coordinates of each atom with displacements drawn from a random number generator with a uniform distribution between −0.5 Å and +0.5 Å. A DFT single-point calculation on these structures followed to obtain the energy and forces for structures out of equilibrium and sample the potential energy surface close to the local minima. Structures with relative energy less than 10 eV are included in the data set. This data set was split into DFTB training, ML training, ML validation, and two MLTB test sets as indicated in Table 1 below.

Table 1. Data Set Split[a]

data partition	# of Th per cluster	number of structures
ML training	1−8	16,859
DFTB training (subset of ML training)	1−8	3315
ML validation	1−8	2000
test (transferability)	1−8	2000
test (extensibility)	12,16	2000

[a]The data set was split into five parts. Density functional tight-binding (DFTB) repulsive potential training was performed with smaller numbers of clusters. Machine Learning (ML) training and model selection were performed with the training and validation sets. The accuracy of the transferability and extensibility were measured with data from within the same n = 1−8 clusters or n = 12, 16, respectively.

Optimization of the HIP-NN model was performed with an objective function consisting of an equal weighting between mean absolute error (MAE) and root-mean-squared error (RMSE) divided by the standard deviation of DFT values for both the atomization energy and atomic force components. Details of the HIP-NN parameters are given in Table 2.

Table 2. HIP-NN Parameters

property	value
L_2 regularization	1e-6
hier regularization	1e-4
dist. soft min.	1.0
dist. soft max	5.5
dist. hard max	6.0
num. features	20
num. sensitivities	30
num. interaction layers	2
num. atomic layers	3

3. RESULTS AND DISCUSSION

3.1. Model Accuracy. Figure 2 shows the comparison between DFTB and MLTB results computed over the (n = 1−8) testing data set. The large plots show the correlation between DFT (reference) and predicted (DFTB or MLTB) per-atom energies. While both show good correlation, MLTB is shown to reduce the root-mean-squared error (RMSE) by about 3 times compared to DFTB. Additionally, bias from the predictions is removed with MLTB as indicated by the reduced mean signed error (MSE): all the definitions of the statistical metrics here presented to assess the accuracy of the models are

https://doi.org/10.1021/acs.jctc.4c00858
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

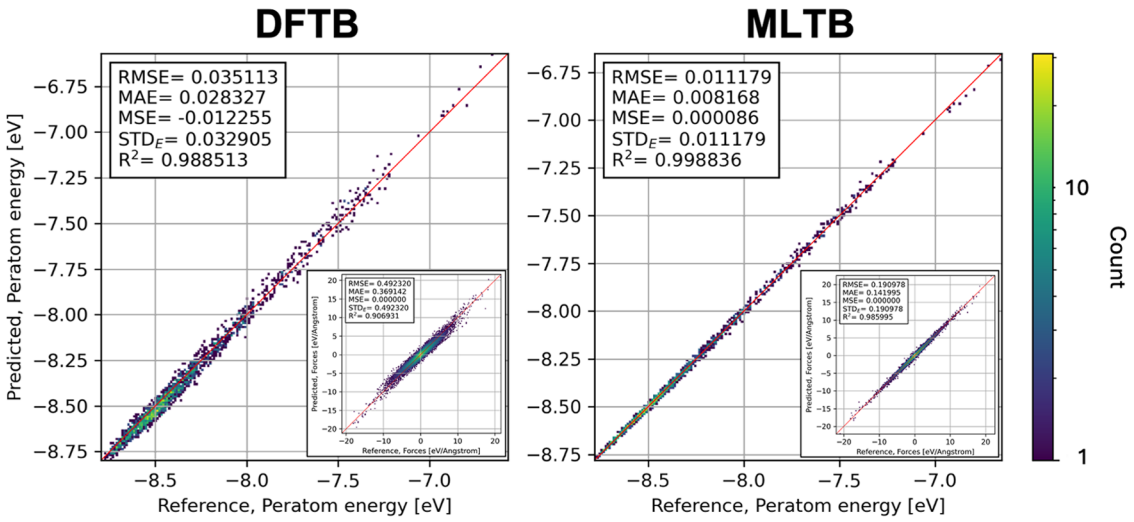


Figure 2. Correlation between DFTB (predicted) and DFT (reference) data for the testing set consisting of cluster sizes from 1 to 8 ThO_2 units. Comparisons between the per-atom energy correlations are shown for DFTB (left) and MLTB (right). Force component correlations are insets in the figures. Results indicate that the MLTB shows a better correlation with respect to the DFT results.

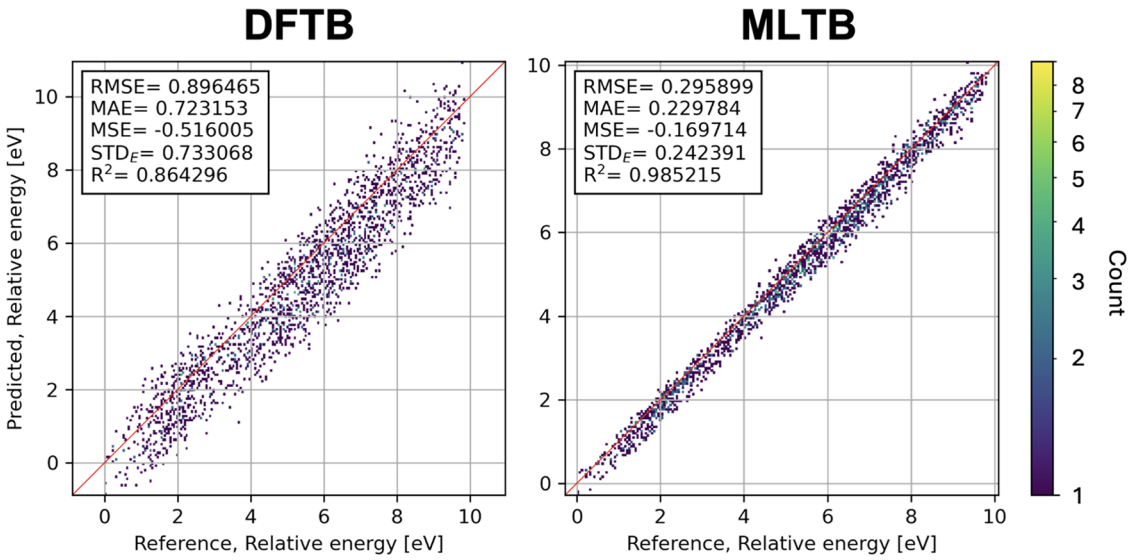


Figure 3. Correlation between DFTB (predicted) and DFT (reference) data for the testing set consisting of cluster sizes from 1 to 8 ThO_2 units. Comparisons between the relative energy correlations are shown for DFTB (left) and MLTB (right). The relative energies give a better understanding of the transferability of the particular model since structures that are further away from equilibrium on the potential energy surface are indicated by higher energies. MLTB is shown to be superior to DFTB alone when predicting higher energy isomers.

described in the SI document, Section S4. Note that the statistical measures reported are the root-mean-square error (RMSE), mean absolute error (MAE), standard deviation of the error (STD_E), coefficient of determination (R^2), and mean signed error (MSE) that indicates systematic deviations. Insets of these figures are the results for the force components. As we see from the DFTB results, the predictions are slightly biased meaning that the DFTB model is introducing a small systematic error. Similar to the energy results, MLTB reduced the RMSE on the forces by 2.5 times compared with DFTB and reflected in the improved correlation.

Figure 3 shows the testing set (n = 1−8) comparison between relative energies for both the DFTB and MLTB models. Relative energies are computed as the energy difference between a given structure and the lowest energy ground state conformer. Since the data set was generated by sampling the PES about the local minima for different cluster sizes, this metric more directly gives us a measure of the transferability of the models. The higher the energy, the further from equilibrium the structure resides and the more difficult it is to make an accurate prediction. Indeed, the higher energy conformers get poorer descriptions from approximated models for the following two reasons: the higher energy conformers usually have more complicated electronic structures, whose chemical environment is not easy to describe, and the high energy local structures are under represented in the training set due to their high order of distortion from local minima. The results from these plots indicate that MLTB again outperforms the pure DFTB model by reducing the RMSE by almost a factor of 3 compared to DFTB and improving the correlation from R^2 = 0.86 to 0.98.

Both Figures 2 and 3 show that MLTB provides an overall performance increase over DFTB alone in terms of error reduction across the range of energies studied. The improved

https://doi.org/10.1021/acs.jctc.4c00858
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

transferability is provided by HIP-NN. However, the testing data was similar to the training data set in that it covers the same chemical space in terms of cluster types and sizes. Next, the extensibility of clusters beyond the training set must be examined.

3.2. Bond Integral Optimization. Figure 4 shows a comparison of the RMSE values between the atomization

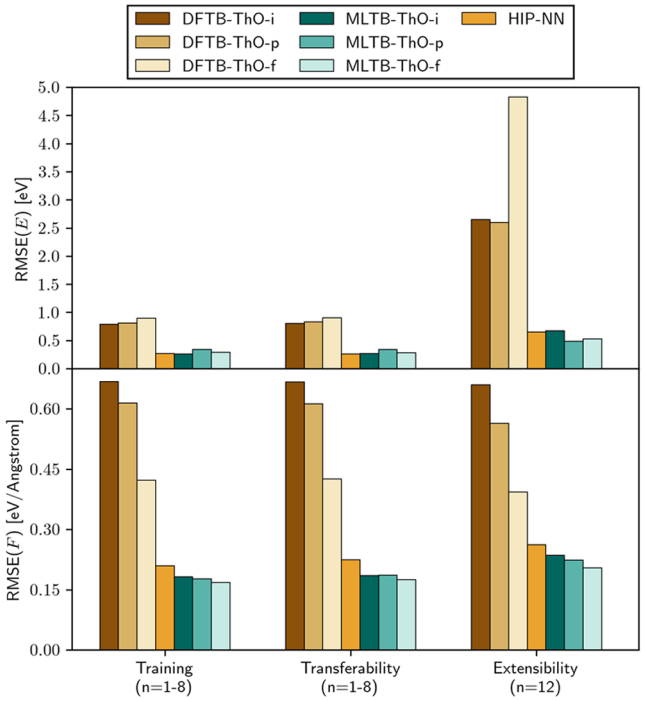


Figure 4. Comparison between DFTB, HIP-NN, and MLTB RMSE for training, transferability, and extensibility test sets. The value of n indicates the cluster size. The top panel focuses on the RMSE of the cluster energies, and the lower panel displays the error in the forces. The raw data for this figure has been included in the SI file for the reference of the reader.

energies and forces for pure DFTB, MLTB, and HIP-NN models applied to the training, testing, and extensibility data sets. The results on the training data show that pure DFTB has the largest error for both optimization cases. The error on both the atomization energy and forces is significantly reduced when using only HIP-NN, which gives an upper-bound for the training error that should be expected from combining HIP-NN and DFTB. Note that optimizing the bond integrals in addition to the repulsive potential does further decrease energy errors. Also note that similar results are seen for the transferability case in Figure 4 where the isolated HIP-NN model shows a larger error on the forces, which is corrected when MLTB is applied.

The extensibility results in Figure 4 show an interesting deviation from the trends observed in the training and transferability sets. The error on the atomization energy from isolated DFTB grows considerably compared to the other cases, and there is little relative improvement with the optimization of the bond integrals. The HIP-NN model yields similar results with overall lower RMSEs, while MLTB atomization energies perform worse when just the repulsive potential is optimized. When the bond integrals are also optimized, MLTB performs better than both DFTB and HIP-NN highlighting the importance of optimizing the bond integrals.

The stabilization energy of successively larger nanoclusters from the LANL-ThO-f DFTB model is shown in Figure 5,

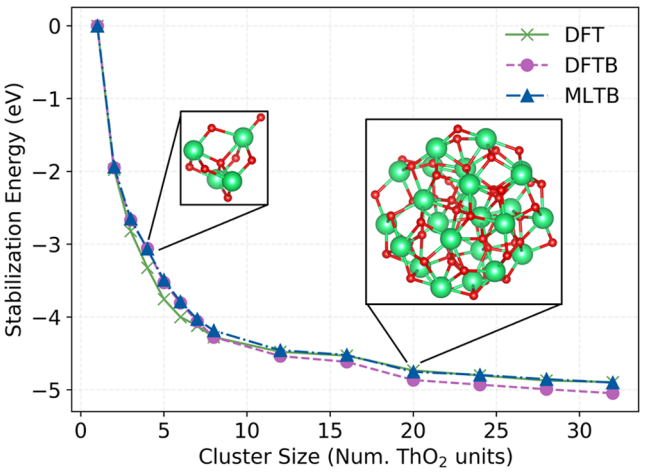


Figure 5. Stabilization energy of ThO_2 nanoclusters computed with DFT, DFTB, and MLTB.

along with the DFT and MLTB results. It was found that the n = 12 and 16 nanoclusters fall into a phase transition between small molecule-like clusters and bulk ThO_2. This means that the DFTB parameters are being applied to chemical environments that were not included in the training set with small clusters ($n \leq 8$), therefore explaining the larger errors. However, the fact that the bond integrals need to be optimized for MLTB to show improvement indicates something more fundamental is occurring during the DFTB parameter optimization process. Despite the two-center nature of DFTB bond integral parameters (Slater−Koster tables), the many-body interactions found in DFT training set can be included by tuning their values because the procedure for finding the density matrix from the DFTB Hamiltonian mixes the information from different matrix elements. As a result, when only the pair potential is optimized (as in the LANL-ThO-i case), the many-body interactions of the baseline model are not optimal. For smaller clusters, this does not pose a problem, but as the clusters grow and the transition from cluster to bulk behavior begins to occur, the chemical environment begins to resemble the interface between a bulk and surface layer of ThO_2. The optimized repulsive potential alone cannot coherently account for these interactions and creates higher errors when using MLTB. When the many-body interactions are better represented in the DFTB by optimizing the bond integrals, MLTB is then able to pick up on trends to provide its own many-body corrections to DFTB.

3.3. Th−O Potential Energy Curve. Structural changes can be further examined by investigating the terminal oxygen through the potential energy curve (PEC) of the Th−O bond distance. Since DFTB is fit to DFT energies, the ideal case would have both the minimum distance and energy reproduced as close as possible. Figure 6 shows the PEC of the terminal Th−O bond for $(ThO_2)_{10}$ computed with DFT, DFTB, and MLTB near equilibrium. To highlight the differences in relative energies, all of the curves have been shifted to zero even though the total energy of the MLTB curve was already in much better agreement with DFT than

https://doi.org/10.1021/acs.jctc.4c00858
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

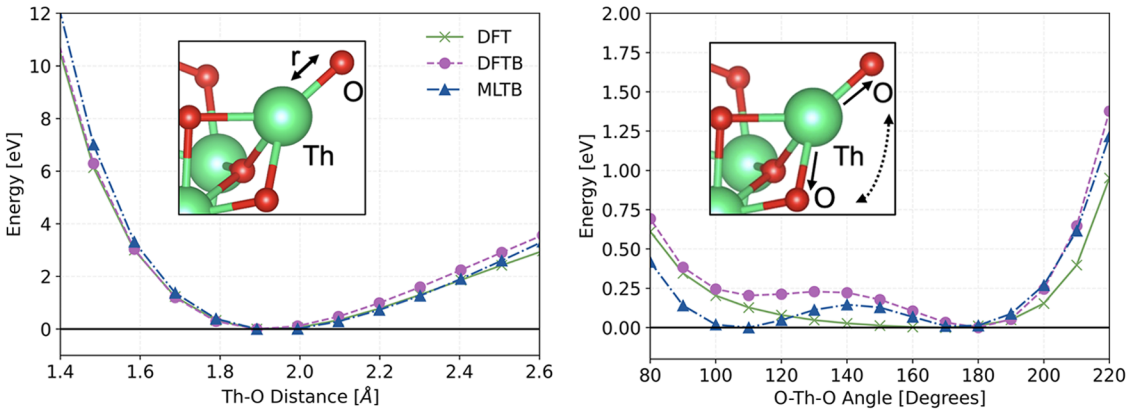


Figure 6. Potential energy curve of a terminal oxygen on the surface of a $(ThO_2)_{10}$ cluster as a function of bond distance (left) and the O−Th−O bend angle (right). Note that each curve was shifted so that the lowest energy value for each method would be at zero. The shifts applied, relative to the DFT curve, are 1.29 and −0.17 eV, for DFTB and MLTB, respectively. This shows that MLTB provides overall significant improvement in the prediction of the total energy and, as seen in the graph, the relative energy is also improved except at short distances, which represent nonphysically compressed bonds under most reasonable chemical situations.

the DFTB curve. It is noticeable that the DFTB and MLTB potential energy surfaces introduce an extra maximum at around 140°, leading to two minima along the direction in which the bond was explored. This particular cluster size was chosen because it is not part of the training set and represents a unique extensibility test for the models. Both DFTB and MLTB show nearly the same equilibrium distance as DFT although the total energy predicted by DFTB is off not only in magnitude but also in curvature; for example, there is a 2 eV difference at 1.5 Å and 1 eV at 2.4 Å. The angular component between thorium and oxygen, which is arguably more sensitive to many-body effects, shows more qualitative differences as a function of the angle. Since this component of the interaction is governed by bond integrals, MLTB with trained bond integrals is able to correct these types of errors.

Furthermore, the DFTB shows a considerable energy shift with respect to the DFT values. As shown in Figure 2, there is a bias in the DFTB per-atom energies. As the number of atoms in the cluster increases, the bias becomes more pronounced. Since MLTB was only trained on clusters up to size 8, the bias is removed for those structures, but once the model is applied to larger clusters, it must rely on the DFTB extrapolations, meaning energy bias re-enters from DFTB.

4. CONCLUSIONS

In this work, we have developed a machine learning tight-binding (MLTB) model with enhanced accuracy and extensibility using a Δ-learning model to include many-body interaction corrections. In particular, ThO_2 nanoclusters were used to investigate the ability of the DFTB and MLTB models to capture many-body effects as the clusters undergo a phase transition from small molecules to incorporating bulk-like chemical environments. Performance, compared to DFT, over the testing set showed that MLTB was superior to DFTB in all aspects, including transferability as indicated by relative energy error metrics.

Many-body corrections are found to play a significant role in correctly predicting the properties of larger clusters (extensibility). When only the repulsive potential of the DFTB parameters is optimized, both DFTB and MLTB struggle to reproduce their energies. When the bond integrals are optimized, the accuracy of MLTB improves significantly. Overall, MLTB is found to outperform HIP-NN alone indicating that the hybrid nature of the Δ-model provided insight that could not be gained with just the neural network.

The potential energy surface from each model was also examined through the optimized structures. MLTB outperforms DFTB in general when using RMSD to compare against reference DFT structures. In particular, the DFTB model was found to be deficient for Th−O angles at the surface of a $(ThO_2)_{10}$ cluster. MLTB was able to correct this deficiency and almost exactly reproduced the DFT reference.

In this work, it is shown that MLTB models can enhance the performance of the complex energy landscape of ThO_2 nanoclusters by including many-body corrections. Although ThO_2 nanoclusters typically reside in the singlet spin state, there is no fundamental barrier in theory to applying this formalism to other actinide molecules and clusters that possess more complex electronic states with unpaired electrons. Future work is being pursued in this direction in an effort to expand the chemistry space that can be examined with this fast and accurate method. From a methodology improvement standpoint, we notice that a variety of approaches are available for building many-body corrections with machine learning,[64,65] and further work might explore which approach demonstrates the most improvement as the ML component of the MLTB scheme introduced here, or whether the best architecture might rather be problem-specific.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jctc.4c00858.

Optimized parameters and the xyz files for the training and testing structures (ZIP)

Tabulated accuracy of reported models; their correlation with DFT data (PDF)

■ AUTHOR INFORMATION

Corresponding Authors

Danny Perez − *Theoretical Division, Los Alamos National Laboratory, Los Alamos, New Mexico 87545, United States;* orcid.org/0000-0003-3028-5249; Email: danny_perez@lanl.gov

https://doi.org/10.1021/acs.jctc.4c00858
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

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

Daniel J. Burrill − *Theoretical Division, Los Alamos National Laboratory, Los Alamos, New Mexico 87545, United States;* Present Address: Lincoln Laboratory, Massachusetts Institute of Technology, Lexington, Massachusetts 02421, United States

Chang Liu − *Theoretical Division, Los Alamos National Laboratory, Los Alamos, New Mexico 87545, United States;* orcid.org/0000-0001-6008-7513

Michael G. Taylor − *Theoretical Division, Los Alamos National Laboratory, Los Alamos, New Mexico 87545, United States;* orcid.org/0000-0003-4327-2746

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

Nicholas Lubbers − *Computer, Computational and Statistical Sciences Division, Los Alamos National Laboratory, Los Alamos, New Mexico 87545, United States;* orcid.org/0000-0002-9001-9973

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jctc.4c00858

Author Contributions

[‖]D.J.B. and C.L. contributed equally to this work.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We acknowledge the support by the U.S. Department of Energy, Office of Science, Office of Basic Energy Sciences, Separation Program (KC0302031) (D.J.B., D.P., and P.Y.) and Heavy Element Chemistry Program (C.L. and E.R.B.) at Los Alamos National Laboratory (LANL). M.J.C. and N.L. acknowledge the support from the LANL LDRD program under project 20220059ER. D.J.B. acknowledges the support of the CNLS Postdoc Fellowship from the Center for Nonlinear Studies at LANL. LANL is operated by Triad National Security, LLC, for the National Nuclear Security Administration of U.S. Department of Energy (contract no. 89233218CNA000001). The computation works were supported by LANL Institutional Computing.

■ REFERENCES

(1) Kohn, W.; Sham, L. J. Self-Consistent Equations Including Exchange and Correlation Effects. *Phys. Rev.* **1965**, *140*, No. A1133.

(2) Slater, J. C.; Koster, G. F. Simplified LCAO method for the periodic potential problem. *Phys. Rev.* **1954**, *94*, No. 1498.

(3) Sutton, A. P.; Finnis, M. W.; Pettifor, D. G.; Ohta, Y. The tight-binding bond model. *J. Phys. C: Solid State Phys.* **1988**, *21*, 35−66.

(4) Goringe, C. M.; Bowler, D. R.; Hernandez, E. Tight-binding modelling of materials. *Rep. Prog. Phys.* **1997**, *60*, 1447−1512.

(5) Harrison, W. A. *Electronic Structure and the Properties of Solids - the Physics of the Chemical Bond*; Dover, 1988.

(6) Sutton, A. P. *Electronic Structure of Materials*; Oxford University Press, 1993.

(7) Pettifor, D. G. *Bonding and Structure of Molecules and Solids*; Oxford University Press, 1995.

(8) Finnis, M. *Interatomic Forces in Condensed Matter*; Oxford University Press, 2003.

(9) Porezag, D.; Frauenheim, T.; Köhler, T.; Seifert, G.; Kaschner, R. Construction of tight-binding-like potentials on the basis of density-functional theory: Application to carbon. *Phys. Rev. B* **1995**, *51*, No. 12947.

(10) Fabris, S.; Paxton, A. T.; Finnis, M. W. Relative energetics and structural properties of zirconia using a self-consistent tight-binding model. *Phys. Rev. B* **2000**, *61*, No. 6617.

(11) 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*, No. 7260.

(12) Frauenheim, T.; Seifert, G.; Elstner, M.; Hajnal, G.; Jungnickel, H. G.; Porezag, D.; Suhai, S.; Scholz, R. A self-consistent charge density functional based tight bindi ng method for predictive materials simulations in physics, chemistry, and biol ogy. *Phys. Stat. Sol. B* **2000**, *217*, 41−62.

(13) Szabo, A.; Ostlund, N. S. *Modern Quantum Chemistry*; Dover, 1989.

(14) Niklasson, A. M. N. A note on the Pulay force at finite electronic temperatures. *J. Chem. Phys.* **2008**, *129*, No. 244107.

(15) Znam, S.; Nguyen-Manh, D.; Pettifor, D. G.; Vitek, V. Atomistic modelling of TiAl I. Bond-order potentials with environmental dependence. *Philos. Mag.* **2003**, *83*, 415−438.

(16) Nguyen-Manh, D.; Vitek, V.; Horsfield, A. P. Environmental dependence of bonding: A challenge for modelling of intermetallics and fusion materials. *Prog. Mater. Sci.* **2007**, *52*, 255−298.

(17) Subramanyam, A. P. A.; Jenke, J.; Ladines, A. N.; Drautz, R.; Hammerschmidt, T. Parametrization protocol and refinement strategies for accurate and transferable analytic bond-order potentials: Application to Re. *Phys. Rev. Mater.* **2024**, *8*, No. 013803.

(18) Krishnapriyan, A.; Yang, P.; Niklasson, A.; Cawkwell, M. 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.

(19) Aguirre, N. F.; Morgenstern, A.; Cawkwell, M. J.; Batista, E. R.; Yang, P. Development of density functional tight-binding parameters using relative energy fitting and particle swarm optimization. *J. Chem. Theory Comput.* **2020**, *16*, 1469−1481.

(20) Nguyen, T. T.; Székely, E.; Imbalzano, G.; Behler, J.; Csányi, G.; Ceriotti, M.; Götz, A. W.; Paesani, F. Comparison of permutationally invariant polynomials, neural networks, and Gaussian approximation potentials in representing water interactions through many-body expansions. *J. Chem. Phys.* **2018**, *148*, No. 241725.

(21) 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.

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

(23) Stöhr, M.; Sandonas, L. M.; 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.

(24) Fernandez, M.; Bilić, A.; Barnard, A. S. Machine learning and genetic algorithm prediction of energy differences between electronic calculations of graphene nanoflakes. *Nanotechnology* **2017**, *28*, No. 38LT03.

(25) Panosetti, C.; Anniés, S. B.; Grosu, C.; Seidlmayer, S.; Scheurer, C. DFTB Modeling of Lithium-Intercalated Graphite with Machine-Learned Repulsive Potential. *J. Phys. Chem. A* **2021**, *125*, 691−699.

(26) Babaei, M.; Azar, Y. T.; Sadeghi, A. Locality meets machine learning: Excited and ground-state energy surfaces of large systems at the cost of small ones. *Phys. Rev. B* **2020**, *101*, No. 115132.
(27) Stocks, R.; Barnard, A. S. Enhancing classical gold nanoparticle simulations with electronic corrections and machine learning. *J. Phys.: Condens. Matter* **2021**, *33*, No. 324003.
(28) 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.
(29) 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.
(30) Li, H.; Collins, C.; Tanha, M.; Gordon, G. J.; Yaron, D. J. A density functional tight binding layer for deep learning of chemical Hamiltonians. *J. Chem. Theory Comput.* **2018**, *14*, 5764−5776.
(31) Kandy, A. K. A.; Wadbro, E.; Aradi, B.; Broqvist, P.; Kullgren, J. Curvature constrained splines for DFTB repulsive potential parametrization. *J. Chem. Theory Comput.* **2021**, *17*, 1771−1781.
(32) Schütt, K. T.; Arbabzadah, F.; Chmiela, S.; Müller, K. R.; Tkatchenko, A. Quantum-chemical insights from deep tensor neural networks. *Nat. Commun.* **2017**, *8*, No. 13890.
(33) McSloy, A.; Fan, G.; Sun, W.; Hölzer, C.; Friede, M.; Ehlert, S.; Schütte, N.-E.; Grimme, S.; Frauenheim, T.; Aradi, B. TBMaLT, a flexible toolkit for combining tight-binding and machine learning. *J. Chem. Phys.* **2023**, *158*, No. 034801.
(34) Lourenço, M. P.; Galvão, B. R. L.; Herrera, L. B.; Hostaš, J.; Tchagang, A.; Silva, M. X.; Salahub, D. R. A new active learning approach for global optimization of atomic clusters. *Theor. Chem. Acc.* **2021**, *140*, No. 62.
(35) Lubbers, N.; Smith, J. S.; Barros, K. Hierarchical modeling of molecular energies using a deep neural network. *J. Chem. Phys.* **2018**, *148*, No. 241715.
(36) Chigaev, M.; Smith, J. S.; Anaya, S.; Nebgen, B.; Bettencourt, M.; Barros, K.; Lubbers, N. Lightweight and effective tensor sensitivity for atomistic neural networks. *J. Chem. Phys.* **2023**, *158*, No. 184108.
(37) Andreeva, N. A.; Chaban, V. V. Global minimum search via annealing: Nanoscale gold clusters. *Chem. Phys. Lett.* **2015**, *622*, 75−79.
(38) Seifert, G.; Joswig, J.-O. Density-functional tight binding—an approximate density-functional theory method. *WIREs Comput. Mol. Sci.* **2012**, *2*, 456−465.
(39) Gaus, M.; Cui, Q.; Elstner, M. Density functional tight binding: application to organic and biological molecules. *WIREs Comput. Mol. Sci.* **2014**, *4*, 49−61.
(40) Elstner, M.; Seifert, G. Density functional tight binding. *Philos. Trans. R. Soc., A* **2014**, *372*, No. 20120483.
(41) Koskinen, P.; Mäkinen, V. Density-functional tight-binding for beginners. *Comput. Mater. Sci.* **2009**, *47*, 237−253.
(42) Frauenheim, T.; Seifert, G.; Elstner, 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.
(43) Seifert, G.; Joswig, J.-O. Density-functional tight binding—An approximate density-functional theory method. *WIREs Comput. Mol. Sci.* **2012**, *2*, 456−465.
(44) Seifert, G.Tight-Binding Density Functional Theory: DFTB. In *Handbook of Solid State Chemistry*; Wiley, 2017.
(45) Spiegelman, F.; Tarrat, N.; Cuny, J.; Dontot, L.; Posenitskiy, E.; Martí, C.; Simon, A.; Rapacioli, M. Density-functional tight-binding: basic concepts and applications to molecules and clusters. *Adv. Phys.: X* **2020**, *5*, No. 1710252.
(46) Sutton, A. P.; Finnis, M. W.; Pettifor, D. G.; Ohta, Y. The tight-binding bond model. *J. Phys. C: Solid State Phys.* **1988**, *21*, 35−66.
(47) Elena, A. M.; Meister, M. Automatic generation of matrix element derivatives for tight binding models. *Phys. Rev. B* **2005**, *72*, No. 165107.
(48) 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.
(49) Liu, C.; Aguirre, N.; Cawkwell, M.; Batista, E. R.; Yang, P. Efficient Parameterization of Density Functional Tight-Binding for 5f-Elements: A Th-O Case Study. *J. Chem. Theory Comput.* **2024**, *20*, 5923−5936.
(50) Behler, J.; Parrinello, M. Generalized neural-network representation of high-dimensional potential-energy surfaces. *Phys. Rev. Lett.* **2007**, *98*, No. 146401.
(51) Behler, J.; Lorenz, S.; Reuter, K. Representing molecule-surface interactions with symmetry-adapted neural networks. *J. Chem. Phys.* **2007**, *127*, No. 014705.
(52) Behler, J. Atom-centered symmetry functions for constructing high-dimensional neural network potentials. *J. Chem. Phys.* **2011**, *134*, No. 074106.
(53) Te Velde, G.; Bickelhaupt, F. M.; Baerends, E. J.; Guerra, C. F.; Van Gisbergen, S. J. A.; Snijders, J. G.; Ziegler, T. Chemistry with ADF. *J. Comput. Chem.* **2001**, *22*, 931−967.
(54) Perdew, J. P.; Burke, K.; Ernzerhof, M. Generalized Gradient Approximation Made Simple. *Phys. Rev. Lett.* **1996**, *77*, 3865−3868.
(55) Van Lenthe, E.; Baerends, E. J. Optimized Slater-type basis sets for the elements 1−118. *J. Comput. Chem.* **2003**, *24*, 1142−1156.
(56) Van Lenthe, E.; Ehlers, A.; Baerends, E.-J. Geometry optimizations in the zero order regular approximation for relativistic effects. *J. Chem. Phys.* **1999**, *110*, 8943−8953.
(57) Batista, E. R.; Martin, R. L.; Hay, P. J.; Peralta, J. E.; Scuseria, G. E. Density Functional Investigations of the Properties and Thermochemistry of UF_6 and UF_5 Using Valence-Electron and All-Electron Approaches. *J. Phys. Chem. A* **2004**, *121*, 2144−2150.
(58) Gaunt, A. J.; Reilly, S. D.; Enriquez, A. E.; Scott, B. L.; Ibers, J. A.; Sekar, P.; Ingram, K. I. M.; Kaltsoyannis, N.; Neu, M. P. Experimental and Theoretical Comparison of Actinide and Lanthanide Bonding in $M[N(EPR_2)_2]_3$ Complexes (M = U, Pu, La, Ce; E = S, Se, Te; R = Ph, iPr, H). *Inorg. Chem.* **2008**, *47*, 29−41.
(59) Kelley, M. P.; Deblonde, G. J.-P.; Su, J.; Booth, C. H.; Abergel, R. J.; Batista, E. R.; Yang, P. Bond Covalency and Oxidation State of Actinide Ions Complexed with Therapeutic Chelating Agent 3,4,3-LI(1,2-HOPO). *Inorg. Chem.* **2018**, *57*, 5352−5363.
(60) Kelley, M. P.; Bessen, N. P.; Su, J.; Urban, M.; Sinkov, S. I.; Lumetta, G. J.; Batista, E. R.; Yang, P.; Shafer, J. C. Revisiting complexation thermodynamics of transplutonium elements up to einsteinium. *Chem. Commun.* **2018**, *54*, 10578−10581.
(61) Kelley, M. P.; Popov, I. A.; Jung, J.; Batista, E. R.; Yang, P. δ and ϕ back-donation in An^{IV} metallacycles. *Nat. Commun.* **2020**, *11*, No. 1558.
(62) Aguirre, N. F.; Jung, J.; Yang, P. Unraveling the structural stability and the electronic structure of ThO_2 clusters. *Phys. Chem. Chem. Phys.* **2020**, *22*, 18614−18621.
(63) Aebersold, L. E.; Wilson, A. K. Considering Density Functional Approaches for Actinide Species: The An66 Molecule Set. *J. Phys. Chem. A* **2021**, *125*, 7029−7037.
(64) Unke, O. T.; Chmiela, S.; Sauceda, H. E.; Gastegger, M.; Poltavsky, I.; Schütt, K. T.; Tkatchenko, A.; Müller, K.-R. Machine learning force fields. *Chem. Rev.* **2021**, *121*, 10142−10186.
(65) Kulichenko, M.; Smith, J. S.; Nebgen, B.; Li, Y. W.; Fedik, N.; Boldyrev, A. I.; Lubbers, N.; Barros, K.; Tretiak, S. The rise of neural networks for materials and chemical dynamics. *J. Phys. Chem. Lett.* **2021**, *12*, 6227−6243.

https://doi.org/10.1021/acs.jctc.4c00858
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX