JCTC Journal of Chemical Theory and Computation

pubs.acs.org/JCTC Article

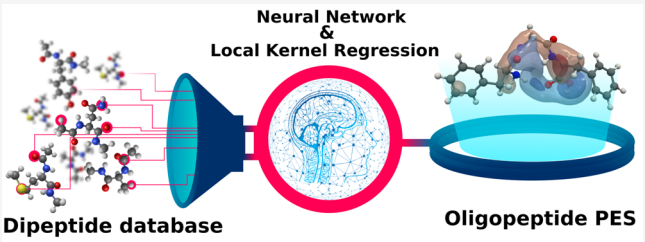
Local Kernel Regression and Neural Network Approaches to the Conformational Landscapes of Oligopeptides

Raimon Fabregat, Alberto Fabrizio, Edgar A. Engel, Benjamin Meyer, Veronika Juraskova, Michele Ceriotti, and Clemence Corminboeuf*

Cite This: J. Chem. Theory Comput. 2022, 18, 1467−1479 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: The application of machine learning to theoretical chemistry has made it possible to combine the accuracy of quantum chemical energetics with the thorough sampling of finite-temperature fluctuations. To reach this goal, a diverse set of methods has been proposed, ranging from simple linear models to kernel regression and highly nonlinear neural networks. Here we apply two widely different approaches to the same, challenging problem: the sampling of the conformational landscape of polypeptides at finite temperature. We develop a local kernel regression (LKR) coupled with a supervised sparsity method and compare it with a more established approach based on Behler-Parrinello type neural networks. In the context of the LKR, we discuss how the supervised selection of the reference pool of environments is crucial to achieve accurate potential energy surfaces at a competitive computational cost and leverage the locality of the model to infer which chemical environments are poorly described by the DFTB baseline. We then discuss the relative merits of the two frameworks and perform Hamiltonian-reservoir replica-exchange Monte Carlo sampling and metadynamics simulations, respectively, to demonstrate that both frameworks can achieve converged and transferable sampling of the conformational landscape of complex and flexible biomolecules with comparable accuracy and computational cost.



INTRODUCTION

Machine learning (ML) techniques have begun to supplement atomistic simulations by facilitating access to the potential energy surfaces (PES) with outstanding accuracy at a greatly reduced computational cost.[1−4] Behler and Parrinello's seminal work introduced one of the first condensed-phase potentials based on a neural network (NN). Using atom-centered symmetry functions to encode the molecular structures[1,5] and expressing the corresponding potential energy as a sum of the atomic contributions makes the potentials transferable and scalable. In recent years, several NN architectures for atomic-based potentials have been proposed, including SchNet[6−9] and PhysNet,[10,11] which predict energies, forces, and other properties (e.g., dipole moments or chemical potentials) of various chemical systems. Roitberg and co-workers also introduced the ANI-1[12] model, where single-atom atomic environment vectors (AEVs) are used to build deep NN potentials to approach the golden standard of CCSD(T)/CBS for reaction thermochemistry, isomerization, and drug-like molecular torsions.[13] Despite their widespread use, NNs have drawbacks: lack of interpretability, the nondeterministic and computationally demanding training, and the large amounts of training data required are some of them.

As an alternative to artificial NNs, kernel-based approaches such as kernel ridge regression (KRR) and gaussian process regression (GPR) overcome some of these limitations.[14] Kernel methods build a map between a target system and its properties by evaluating a similarity measure between the target and a set of known reference points. Gaussian approximation potentials (GAPs)[15,16] pioneered the use of kernels in molecular dynamic simulations and demonstrated that they can achieve results equivalent to NNs. Since then, they have been used to model bulk materials ranging from simple silicon,[17−21] to ternary $Ge_2Sb_2Te_5$.[22,23] In the wake of GAPs, numerous alternative kernel-based and linear methods have been proposed to predict PESs for atomistic simulations, including support vector machines (SVM),[24] the spectral neighbor analysis potentials (SNAPs),[25] general reproducing kernels models,[26,27] and coarse-grained approaches.[28] More recently, the symmetrized gradient-domain (sGDML) model has proven to yield nearly exact molecular dynamics simulations for small molecules based on coupled-cluster energies and forces.[2,29,30] However, despite the increasing number of kernel-based ML potentials, artificial NNs remain dominant for driving atomistic simulations.[5,31−39]

Received: August 11, 2021
Published: February 18, 2022

https://doi.org/10.1021/acs.jctc.1c00813
J. Chem. Theory Comput. 2022, 18, 1467−1479

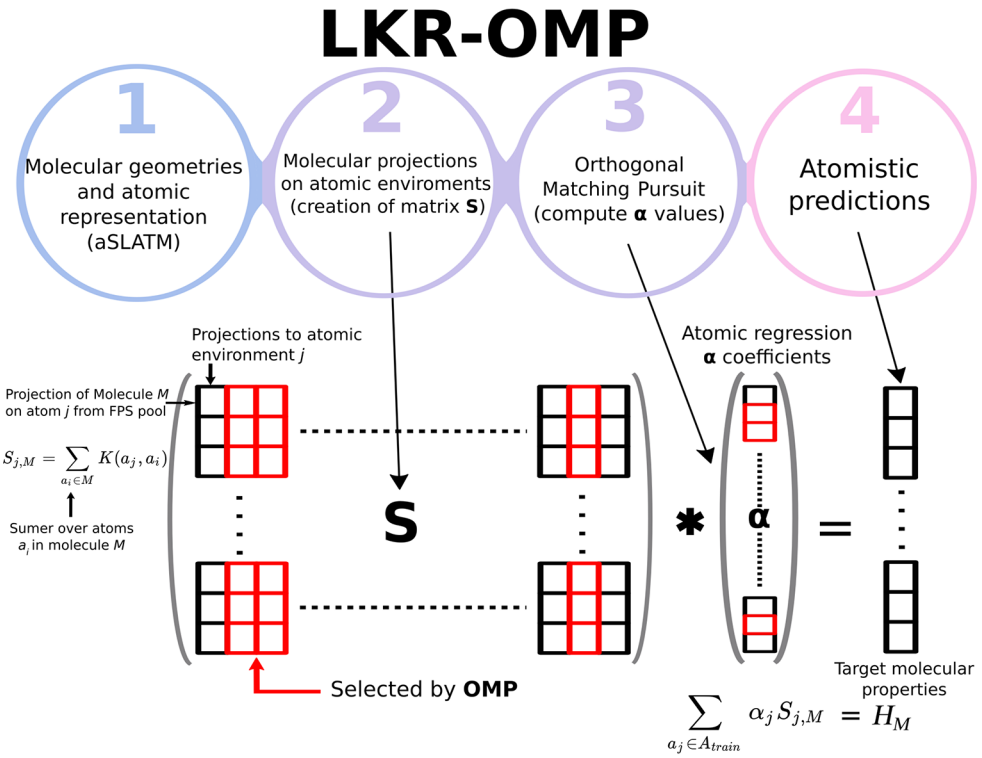


Figure 1. Workflow and schematic depiction of the LKR model.

When paired with global molecular representations (e.g., Coulomb matrix,[40] bag of bonds (BoB),[41] or the Spectral London Axilrod−Teller-Muto[42] (SLATM)), which encode the key physical information about the structure and composition of molecules as whole indivisible entities, kernel models are often lightweight, making them ideal for predicting molecular properties.[17,43−46] However, predictions made with these global representations are expected to be accurate only for molecules of similar size and composition with respect to those in the training set. This constraint limit severely the exploration and extrapolation to larger chemical and conformational spaces. Local representations (e.g., FCHL,[47,48] aSLATM,[42] and SOAP[49]), which describe molecules as a collection of atoms within their local environments, provide a greater transferability[50,51] but also significantly increase the computational cost of kernel-based methods, as the similarity between molecules is then computed as a function of the pairwise similarity between atoms.[52] To restore the data-efficiency typical of kernel-based methods and efficiently exploit the local representations, one can resort to sparse regression techniques. The simplest form amounts to sampling the entire set of atom-centered environments and retaining only the (*a priori*) most informative environments, assuming that substantial redundancy arises from recurring environments across training structures. The criteria for selection tend to be based on techniques such as farthest point sampling (FPS) or CUR matrix decomposition[53] that maximize the dissimilarity of the selected environments. While the environments sampled with FPS- or CUR-based methods represent the most varied set among the training instances, they are not necessarily the best for regressing the property of interest,[54] as the dissimilarity in the representation space does not necessarily correlate with dissimilarity in the property space.[55] (Δ-)[56]ML approaches represent a typical illustration of this issue, as the vast majority of chemical environments are well described by an approximated baseline model while the error is concentrated in localized areas of the feature space. This is particularly true when predicting PESs, where capturing the conformational changes (e.g., torsion of a single dihedral angle) is as crucial as capturing the dependence on chemical diversity.

In this work, our goal is to address the limitations of traditional unsupervised sparsification techniques and leverage the data-efficiency and transferability of local kernel models, by combining a local kernel regression (LKR) framework with a flexible orthogonal matching pursuit (OMP) algorithm. The efficiency of the resulting model is demonstrated by learning the PES of oligopeptides using a set of 52,000 conformations of dipeptides comprised of 26 amino acids. In this context, the OMP controls the sparsification process and selects (among tens of thousands of atom-centered environments present in the training set) the best possible reference pool for predicting the PES of any dipeptide. To increase the smoothness of the target energies, the model is baselined with density functional tight-binding (DFTB[57,58]) using a Δ-ML approach,[56] with the model improving the description of the PES in regions that are traditionally not accurately captured with the semiempirical baseline method (e.g., hydrogen atoms and polarized bonds). To further illustrate the transferability of LKR, we compare its performance with a state-of-the-art Behler-Parrinello type neural network, both on the dipeptide set and in an extrapolation test based on the Phe-Gly-Phe tripeptide. The two ML models are then used to drive enhanced sampling simulations to describe the free energy landscape of the tripeptide with DFT accuracy.

METHODS

Machine Learning Models. The ML potentials presented in this work correct a semiempirical baseline obtained from

https://doi.org/10.1021/acs.jctc.1c00813
J. Chem. Theory Comput. 2022, 18, 1467−1479

density functional tight-binding (DFTB) with the D3(BJ)[59] dispersion correction (shortened DFTB hereafter) to reproduce the target PBE[60]-dDsC[61−63] (shortened PBE hereafter) for DFT energetics. For each molecule in the data set, the property learned within the Δ-ML framework corresponds to the difference between the atomization energy evaluated at DFTB and PBE. For both levels, the atomization energies are computed using a two-step procedure. First, the contribution of each atom type to the total energy is evaluated by a multilinear regression (MLR) on the full data set (dressed-atom energies). Then, the difference between the computed total energy and the sum of the dressed-atom energies yields the atomization energy used herein. The following sections describe the two types of complementary ML architectures exploited in this work.

ML Model 1: Sparse Local Kernel Regression. The LKR inputs are the target molecular properties and the atomic representations of the corresponding molecular structures. In this case, we used the atomic spectral London Axilrod−Teller−Muto[42] representation (aSLATM) (see step 1 in Figure 1, upper panel), but other local atomic representations could be used. As it is standard procedure for local kernel-based atomistic models, LKR uses a selected pool of reference atomic environments taken from the training structures as the regression basis for predicting the target property. The structures available for the training are projected onto the pool of atomic environments using a Gaussian kernel to create the matrix **S**, effectively generating a new vectorial representation of the molecules (see step 2 in Figure 1, upper panel). By assuming a linear relationship between the features of **S** and the global molecular properties, LKR allows one to obtain the regression coefficients for each reference atomic environment without requiring an *a priori* decomposition of the target property, which is sometimes possible[64] but highly nontrivial for complex PES like the ones discussed here. If the pool of atomic environments is too large, prefiltering, which reduces the redundancy of the pool, is needed. Here, we use FPS,[53] which selects the most distinct environments in terms of their Euclidean distances.

For the final selection of the reference environments, the reduction of the training environments is commonly performed by constructing multiple models including a variable number of the FPS points, which is gradually increased until achieving a satisfying accuracy. It was already hypothesized[65] that some sort of supervision in the sparsification procedure would be desirable. Here, we rely on a supervised sparse regression model called orthogonal matching pursuit (OMP).[66] OMP is a greedy optimization algorithm that finds the best sparse choice of reference environments for a particular application (see step 3 in Figure 1, upper panel). The OMP algorithm searches greedily through the whole pool of atom-centered environments and selects at each time the specific environment that reduces the prediction error the most (i.e., the one with the highest inner product with the targeted property). At each iteration, the contributions from the previously selected environments to the global target property are subtracted and the search continues for the best match of the residual until convergence. With this procedure, OMP automatically identifies the most suitable, property-specific environment subset (i.e., best-matching basis) for the regression of the targeted molecular property in one shot. In the prediction step (see step 4 in Figure 1, upper panel), the similarity of each new atomic environment with respect to the reference pool is evaluated by computing a kernel sum with all the selected environments. The reader is referred to Figure 1 for a schematic depiction of the workflow and to the Supporting Information for a more detailed description of the model and procedure.

Overall, LKR-OMP combines the scalability and transferability of NNs, with the faster training and stability of kernel-based models. The addition and removal of training data also require minimal computational effort, as opposed to an NN, for which the procedure requires at best a partial retraining. This would be especially beneficial for active learning approaches,[67] when the training data evolves throughout the process. The counterpart is that the cost of the model scales linearly with the number of reference environments, while the cost of NNs is fixed by the architecture.

ML Model 2: Behler-Parrinello Neural Networks. To benchmark the LKR model against an established NN architecture, we further construct a Behler-Parrinello artificial NN.[1] For each atom, we describe the positions of all neighboring atoms inside a cutoff radius (its "atomic environment") by a set of atom-centered many-body symmetry functions (SF)[68] (see the Computational Details).

To allow for on-the-fly estimation of the uncertainties in the predictions, a committee of four Behler-Parinello NNs,[1,68] which only differ in the random initialization of the NN weights and the internal cross-validation splitting of the training data, was trained to reproduce the differences between the DFTB baseline and the target DFT energies and forces. This permits estimating the uncertainty associated with each committee prediction of the Δ-ML correction.[69] The uncertainty estimates were also used to modulate the application of the NN correction, using the weighted baseline scheme proposed by Imbalzano and co-workers.[70] This procedure minimizes the uncertainty in the total potential and ensures that it falls back to the baseline whenever the Δ-ML correction enters the extrapolative regime, thereby stabilizing the simulation. The total energy is calculated as the sum of the outputs of atomic NNs, and analytic gradients and thus forces are readily available. To train the NN models both energies and forces were used.

Training Data. The training set for the construction of the models described in the previous sections was built by selecting configurations from the 300 K replica of a DFTB-based temperature replica exchange (T-RE) simulation (with replicas at temperatures between 300 K and 1000 K) for each amino acid dipeptide. The most distinct 2000 configurations of each dipeptide were selected by means of FPS, using the Ramachandran plot[71] coordinates as the independent variables.

For a total of 26 amino acid dipeptides,[72] we obtained a pool of 52,000 conformations. Finally, to include the effects of side chain-side chain interactions into the model, the training set was enriched with an additional set of 3378 optimized peptide dimers from the BioFragment Database.[73] Single point computations were performed to obtain energy and forces at the target and baseline levels.

Enhanced Sampling Methods for the Tripeptide. We use the reservoir-Hamiltonian Replica Exchange (resH-RE)[3] technique to sample the canonical ensemble of the selected Phe-Gly-Phe tripeptide at 300 K with the LKR potential. ResH-RE is an enhanced Hamiltonian Replica Exchange[74] scheme, which serves to accelerate the sampling of the configurational space at a high level of theory using a canonical

https://doi.org/10.1021/acs.jctc.1c00813
J. Chem. Theory Comput. 2022, 18, 1467−1479

reservoir of structures generated with a less accurate but computationally cheaper potential energy. The replicas essentially help to capture the local diffusion in the phase space, whereas the most dramatic conformational changes, such as swaps between local minima and crossings of energy barriers, occur through coupling with the reservoir. By construction, the resH-RE simulation can be driven by molecular dynamics in the NVT ensemble but also by simpler Monte Carlo (MC) moves (i.e., random particle moves), which are otherwise largely inefficient for systems characterized by highly nonlinear PESs.[75] The possibility of using both molecular dynamics and Monte Carlo moves within resH-RE is especially advantageous given that the atomic forces are not readily available with the LKR model used here albeit, in principle, obtainable through computing the LKR energy derivatives[76] with respect to the nuclear coordinates.[77]

Considering that the forces are available, and actually needed to increase the robustness of the NN potential (*vide infra*), the NN-based sampling of the tripeptide was performed using the ATLAS metadynamics framework,[78] which employs a divide-and-conquer strategy to enable efficient biasing when working with several collective variables (CV). In ATLAS, the high-dimensional CV space is divided into basins, each of which is described by an automatically determined, low-dimensional subset of the CVs on which a local, well-tempered metadynamics-like bias is constructed. The local biases are translated into an effectively high-dimensional bias using indicator functions based on a Gaussian mixture model. Given the high dimensionality of the CV space of the Phe-Gly-Phe tripeptide, attempting convergence with conventional metadynamics would be futile. Meanwhile, the ATLAS framework, which was specifically designed to work in high dimensions, has already been tested on 6D spaces.[78] While alternative sampling techniques such as those based on temperature acceleration could have been used for the NN sampling,[79,80] overcoming the high energy barriers between basins would have required temperatures impractical for the problem at hand.

In this work, space is divided into five basins, identified by applying the PAMM framework[81] to an initial well-tempered metadynamics trajectory using the end-to-end distance of the backbone as the sole CV. Each basin is described and biased based on the two principal axes determined by performing a principal components analysis on the associated distributions of configurations in the six-dimensional CV space. The resultant metadynamics trajectories were unbiased using the ITRE scheme,[82] which makes efficient use of the entire trajectory and does not require the distribution to be evaluated on a grid, rendering it suitable for high-dimensional CV spaces.

Computational Details. All the baseline computations for the Δ-ML model were performed with DFTB3/3OB[57,58] in combination with the D3(BJ)[59] dispersion correction (DFTB), as implemented in the DFTB+ software.[83] The target potential was set at PBE[60]-dDsC[61−63] using the def2-TZVP basis set, as implemented in GAMESS-US.[84,85] Canonical sampling of each dipeptide was performed using T-RE simulations using the REMD@DFTB[86] protocol implemented in i-PI.[87] The simulations included 16 replicas with temperatures ranging from 300 K to 1000 K, equally spaced on a logarithmic scale. A time step of 0.75 fs was used in the dynamics, which ensured the stability and energy conservation of the dynamics (see Figure S8), with a Langevin thermostat to control the temperature. The simulations were run for two million steps, which ensured statistical convergence of the results (see Figure S10). The final batch of structures was split in two separate sets (70% (40,000) and 30% (15,378) of the molecules, respectively), which were used for training and testing of the models. The resH-RE simulations were run using the MORESIM python package.[3] They included four replicas with a potential linearly evolving from DFTB to DFTB + LKR. This choice resulted in an exchange acceptance probability of 40%. The resH-RE simulations were run for two million steps, which provided converged results. A global random displacement with a Gaussian distribution of standard deviation 0.001 Å was chosen as the Monte Carlo step, which resulted in a 50% acceptance rate.

All metadynamics simulations were performed by coupling the i-PI energy and force engine[88] to the open-source, community-developed PLUMED library[89] version 2.8.0-dev (git: 79bcb8947)[90] to apply a well-tempered bias and the DFTB+[83] and LAMMPS[91] codes to evaluate the baseline potential and Δ-learned correction, respectively. All metadynamics simulations employed a time-step of 0.5 fs to ensure the stability of the dynamics and the NN correction (see Figure S8) and a generalized Langevin equation (GLE) thermostat.[92,93]

The local kernel regression implementation (available on github[94]) relies on a Gaussian Kernel[40] and on the aSLATM representation, as provided in the QML-toolkit.[95] The width of the Gaussian kernel, the adimensional parameter σ, was chosen to be $\sigma = 4.5$ after a systematic grid search. We used FPS to preselect a first pool of 39,000 local atomic environments. The optimal number of reference environment selected by OMP can be obtained using a grid search optimization of this parameter (LKR-optimal), although the bigger the number the higher the cost of the model (Figure S1b). To achieve a converged statistical sampling (with resH-RE) at a reasonable computational cost (see the Supporting Information for a more detailed discussion on the computational cost), the size of the pool of reference environment is limited to 1,000. The relevance of this particular trade-off between accuracy and computational cost is shown in Figure S1b. The python library Sci-Kit Learn[96] was used to perform the OMP regression.

The NN models were trained using the N2P2 code.[97] Initial many-body symmetry functions (SF),[68] which describe the local atomic environment of each atom in a configuration and provide the inputs to the NNs, were generated following the protocol of Imbalzano et al.[53] and included G2 functions with $N = 12$ and cutoffs $r_c = 8$, 12, and 16 Bohr and G3 functions with $N = 4$, $r_c = 8$ Bohr, and $\zeta = 1$, 2, 4 and with $N = 2$, $r_c = 12$ Bohr, and $\zeta = 1$, 2. The cut-offs are long enough to describe the environment of the central atom substantially beyond its nearest neighbors in order to address the local differences between DFTB and DFT (long-range discrepancies between DFTB and DFT are also accounted for, albeit in a mean-field manner, through their effect on the local atomic environments). The 512 most informative among them were extracted using the semisupervised PCovCUR scheme;[98] a modification to the CUR approach, which uses a mixing parameter (here set to 0.5) to smoothly interpolate between a feature-covariance and a linear regression-like loss to identify features that reflect the (structural) variance of the data set while also correlating with the target property. Their values for a given atomic environment are concatenated into a feature vector and fed into the "atomic" NNs, which in the following consists of two fully connected, hidden layers with 24 nodes each. This

https://doi.org/10.1021/acs.jctc.1c00813
J. Chem. Theory Comput. 2022, 18, 1467−1479

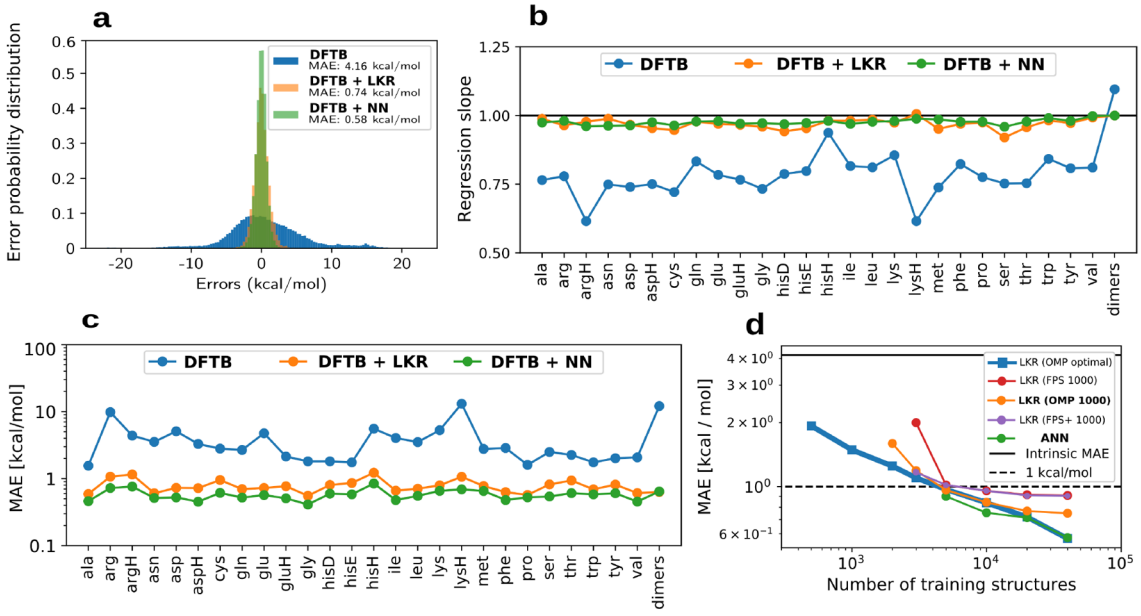


Figure 2. (a) Histogram of errors in test samples of the dipeptide data set. (b) Regression slopes between "bonding energies" of DFTB and PBE for each of the training dipeptides and for the dimers. (c) MAE achieved by the models in the test data for each dipeptite and for the peptide dimers. (d) Learning curves, i.e., achieved MAE vs number of structures used for the training. The different learning curves are LKR using OMP with the optimized number of atomic environments (blue), LKR exploiting OMP to select the best 1,000 environment (orange), the Behler-Parrinello-based NN (green), LKR using FPS to select the most distinct atomic environments, using 200 atoms per atom type (FPS 1000) (red), and LKR using FPS to select the most distinct atomic environments but with the same distribution as OMP (FPS+ 1000) (purple).

particular architecture has previously proven sufficiently flexible to describe molecular crystals containing up to four chemical species,[99,100] and multilayer perceptron networks with similar depths and widths have seen widespread success for a variety of molecular and condensed matter systems.[101,102]

RESULTS AND DISCUSSION

Performance of the Trained Machine Learning Models. The need for correcting DFTB to obtain reliable PESs for each amino acid dipeptide is made evident by Figure 2a, showing the histogram of the differences with respect to the target PBE (after removing the multilinear regression contribution). The inaccuracy of DFTB is also illustrated by the regression slopes between the atomization energies at the DFTB level with and without ML corrections (Figure 2b) and the PBE atomization energies. For each dipeptide, the slope between uncorrected DFTB and PBE is consistently smaller than unity, implying a systematic overstabilization of the most distorted configurations and an energy understabilization for the most stable ones (see Figure S2 for a more detailed analysis of the individual dipeptides). The flatter characteristic of the DFTB PESs has previously been discussed[3,103] and attributed to the limited amount of atomic overlap afforded by its minimal valence basis, which also affects the rotational barriers.[57] As shown in Figure 2b,c, the LKR and NN models correct for the systematic flattening of the PESs (slope ∼1, Figure 2b) and also decrease the absolute errors for each dipeptide. As shown by the learning curves (Figure 2d), the NN (0.58 kcal/mol, 40,000 training dipeptides) and LKR-OMP (optimal) (0.57 kcal/mol, 40,000 training dipeptides) predictions are equally accurate. The more computationally efficient LKR-OMP(1000) model discussed above achieves an accuracy of 0.74 kcal/mol. The relevance of using OMP for the selection of the reference environments instead of simpler algorithms is illustrated by comparing the accuracy of LKR-OMP(1000) and a ridge regression based on the same number of environments chosen by FPS. The LKR-OMP(1000) model (referred simply as "LKR" for the rest of the article) is significantly more accurate than the LKR based on FPS, which additionally highlights the importance of selecting atomic environments tailored for the specific target property. While the performance of the NN is slightly superior to the LKR in the training step, it must be noted that the latter model is only trained on energy data, whereas the NN uses both energies and forces (i.e., $3 \times N_{\text{atoms}}$ times more training scalar quantities). However, the mean absolute error for each individual dipeptide is consistently below 1 kcal/mol for both models. The learning rates of both approaches, defined as the error as a function of the number of training structures, are also both very similar and characterized by a decay exponent of −0.2 on a logarithmic scale.

The OMP algorithm provides insightful complementary information, allowing one to identify which atomic environment is associated with the largest difficulties in the learning procedure. This feature is unique to OMP and not available for standard kernel or NN-based approaches that do not rely on supervised sparsity methods. In particular, OMP identifies that only a few of the 39,000 atomic environments (as low as 300) are sufficient to reach the accuracy threshold (1 kcal/mol) for the predictions of the dipeptide atomization energies. The OMP selection within LKR is 45.1% C, 2.9% H, 18.6% O, 28.9% N, and 4.5% S atoms. For the sake of comparison, the atomic composition of the pool of dipeptide training structures is 29.5% C, 53% H, 8.5% O, 9% N, and 0.3% S atoms. Evidently, the optimal reference atomic environments selected by OMP do not follow the same atomic distribution as in the overall pool of structures. OMP does not only find an adequate percentage of atom types but also picks the most tailored

https://doi.org/10.1021/acs.jctc.1c00813
J. Chem. Theory Comput. 2022, 18, 1467−1479

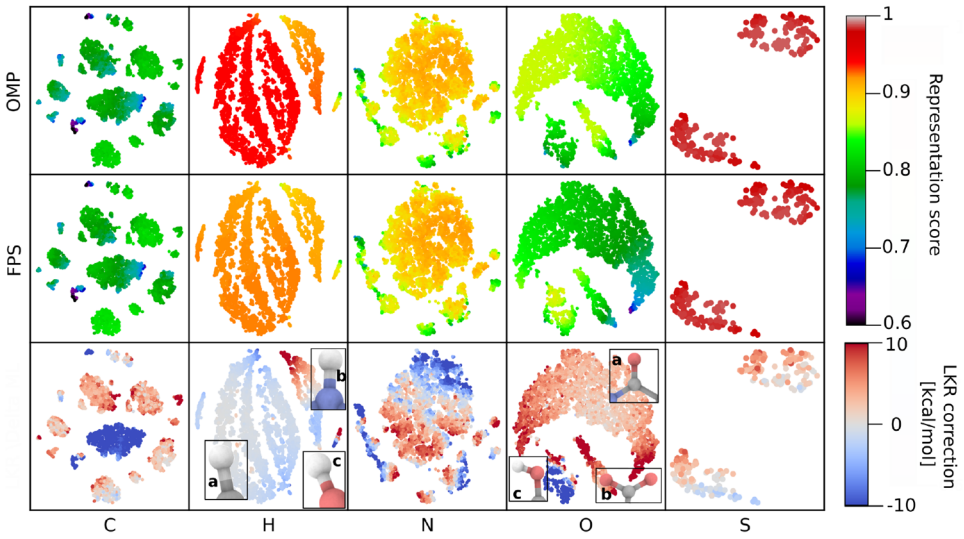


Figure 3. t-SNE maps constructed with the aSLATM representation as input for each atom type. Each point represents an atomic environment in the training data. The color code in the first two rows shows how well represented the training environments are by the reference environments chosen by OMP and FPS+. As representation score, we use the average "atomic Kernel Representation Score" (aKRS), the average value of the kernel similarity between each of the training atomic environments, and the selected reference environments of the same atom type. The color code in the last row shows the LKR correction on each of the training atomic environments.

atomic environments for the target property. In contrast, the FPS selection with the same enforced atomic distribution as OMP (FPS+ 1000) is not sufficient to achieve a MAE as low as OMP (Figure 2d). In fact, on average three times more atomic environments are needed for FPS+ to match OMP (see Figure S1b). This is further demonstrated by the 2D t-SNE (t-Stochastic Neighbor Embedding[104]) projection (Figure 3) of the training atomic environments (constructed using the aSLATM representation as input data for the t-SNE).

The first two rows of t-SNE maps are color-coded based on the average "atomic Kernel Representation Score" (aKRS), i.e., the average value of the kernel similarity between the training atomic environments and the selected reference ($\langle aKRS \rangle_j = \frac{1}{N_a}\sum_{i\in \mathrm{ref}} K(a_j, a_i)$, where j represents the index of an environment in the training data and i is runs over the N_a selected reference environments of each atom type). The score is computed for the reference environments selected by OMP (first row of Figure 3) and FPS+ (second row of Figure 3). This score, bound between zero and one, shows how well an atomic environment is represented by the selected reference environments. The most striking differences between OMP and FPS+ is in the selection of the oxygen and hydrogen atomic environments, whereas carbon, nitrogen, and sulfur are treated very similarly. In other words, the assumption behind the usage of FPS (the larger the variability in the reference environment, the higher the accuracy) is correct for carbon, nitrogen, and sulfur but not for hydrogen and oxygen. The oxygen maps are formed by one large smooth cluster, which represents the amide-bond oxygen atoms [O(a)], and two smaller regions regrouping the carboxylate [O(b)] and hydroxyl [O(c)] oxygen atoms, respectively. In comparison to FPS, OMP is placing more emphasis on the amide oxygens and the carboxylate groups but much less on the oxygen in the hydroxyl groups. For the hydrogen atoms, the large number of isolated clusters in the t-SNE is indicative of a large variability in the hydrogen environments, which could intuitively suggest that a high number of hydrogen reference atoms are necessary to get an accurate model. Yet, OMP only selects 2.9% of them. This result reinforces that the choice of tailored environments is the key to achieving a more robust regression model. Interestingly, OMP favors carbon-bonded hydrogen atoms lying in the central cluster rather than polar hydrogens (e.g., in a O−H bond). Since the model is constructed to capture the variations of the potential energy as a function of the molecule structural changes, the selection of more carbon-bonded hydrogens than any other type has to be attributed to the higher conformational variability of the environments surrounding a C−H bond.

Another useful analysis of how the model behaves involves comparing the choice of atomic environments by OMP with the magnitude of the ML correction in terms of atomic-contributions (last row of Figure 3). While one might expect a direct relationship between the atomic selection and the magnitude of the ML atomic error, this intuition is actually incorrect. In fact, a large DFTB error for a given atom type does not necessarily imply that the learning process would be improved by including more atom-environments of the same type. This is especially true if the electronic nature of the DFTB error is uniform across all the conformation available in the training set. This lack of correlation is evident while looking at the bottom panels of Figure 3. The DFTB errors are the largest for the hydroxyl functional groups [H(c) and O(c) in the figure], while only a small portion of carbonyl or amide oxygen atoms are characterized by similar errors of opposite sign. This trend is not reflected in the optimal OMP selection of reference atomic environments. Similarly, the most problematic carbon atoms (in terms of ML errors) are the oxygen-bonded carbons, which include the amide functions (the center cluster), as well as the carbons of the terminal guanidino group of arginine ($HNC(NH_2)_2$). However, OMP does not place special attention to these environments when selecting the best reference carbons. Nitrogen behaves similarly to carbon. The central cluster is the most well described, which is representative of C−NH−C nitrogens (mainly present in the amide bonds), while the outer clusters, including terminal

https://doi.org/10.1021/acs.jctc.1c00813
J. Chem. Theory Comput. 2022, 18, 1467−1479

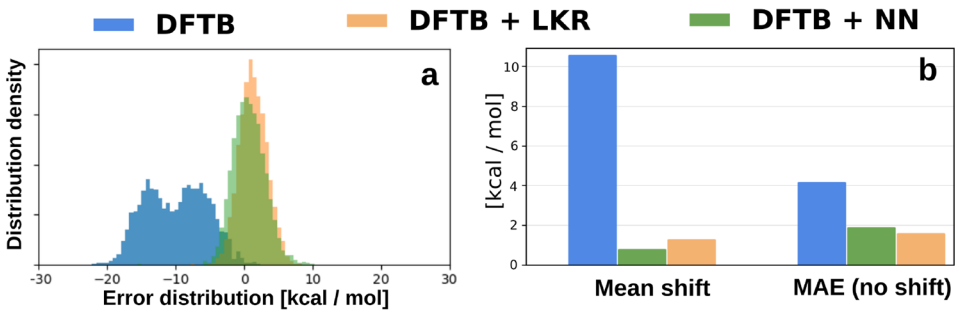


Figure 4. (a) Histogram of prediction errors made on the tripeptide test set. (b) Bar plots with the mean shifts of the error distributions and their MAE after being centered.

amines ($-NH_2$), the proline rings, and guanidino groups, are less sampled. In contrast to other atoms, the ML correction for nitrogen has similar magnitude in all the clusters. An interactive application to visualize and explore this data is available at https://atomic-environments-dipeptides.herokuapp.com, built with the Molecular Explorer Software.[105]

Extrapolation. The local nature of the two ML potentials can in principle be used to make predictions for any system containing no chemical species other than C, H, O, N, and S, although high accuracy is expected only for local environments similar to those present in the training set, i.e., in peptide chains or oligopeptides. Here we demonstrate the transferability of the two models by exploring the potential and free energy landscapes of the Phe-Gly-Phe tripeptide. The Phe-Gly-Phe tripeptide (in neutral form) is an appealing target to test the transferability of the ML models as it is one of the most suitable chemical systems to model noncovalent interactions in proteins.[106] Additionally, this tripeptide is not an adequate target for existing force fields, which are typically parametrized either for capped peptides or for charged forms. The Phe-Gly-Phe tripeptide is in the gas phase and uncapped, and contains a combination of neutral NH2 and COOH groups that are not stable in solution. As a result, many force fields (AMOEBA, AMBER) do not accept it as input or, alternatively, generate unstable dynamics (GAFF).

To assess the quality of the extrapolated energies, we compile two data sets of 1000 Phe-Gly-Phe structures subdivided into 900/100 subsets illustrative of the conformational landscape explored at 300 and 0 K, respectively (for further details, see Supporting Information), at both the baseline and target levels. The first set corresponds to 1,000 structures selected at random from the 300 K replica from the DFTB-based T-RE simulation. Out of these 1,000 structures, 100 are optimized at the same DFTB level (i.e., 0 K static optimization). The second set is a random selection of 1,000 structures taken from the 300 K sampling at the DFTB + LKR level (see the next section) out of which 100 are optimized with PBE.

The most striking difference between the error distributions of DFTB and the ML corrected versions (respectively the blue and orange/green histograms, Figure 4a) is the transition from a bimodal Gaussian distribution to the expected normal distribution centered at zero. The two peaks correspond to the DFTB energies of conformers generated using DFTB as underlying potential [DFTB//DFTB, overstabilized] and to the DFTB energies of conformers generated using a different potential [DFTB//PBE, understabilized]. The transition from a bimodal to a single Gaussian distribution upon application of the ML-corrections reveals that the DFTB-sampled conformational space (i.e., the set of visited structures) would be energetically disjoint from the reference-sampled space at PBE if we had to drive the dynamics using a DFTB potential. The ML-corrections allow concluding that this separation is spurious and that the DFTB structures from the PBE perspective [PBE//DFTB] are not peculiar. Interestingly, the slope between DFTB and PBE energies for all the tripeptide test structures combined is 0.96 (see Figures S4 and S6), which would suggest that systematic flattening of the PES by DFTB is not observed in this case. However, the correlation between DFT and DFTB breaks down when considering the 300 and 0 K conformations separately (in a clear example of the Simpson's paradox[107]), where the typical behavior of DFTB is recovered (slopes: 0.78 at 300 K and 0.73 at 0 K, see again Figure S6). Finite temperature effects offset the energies of the 300 K ensemble with respect to the 0 K, so that the joint distribution seems to correlate better with the DFT values.

The ML corrections (Figure 4 orange and green data) overcome all the issues present in the uncorrected DFTB potential. First, the mean bonding energy shift is reduced from 10.6 to 1.3 kcal/mol by the LKR and to 0.8 kcal/mol by the NN model (see Figure 4b). This error does not influence the conformational sampling of the molecule, as a constant shift in energy does not alter the relative probability of the conformers. Nevertheless, a decreased error is beneficial when comparing the electronic energies of different molecules. Most importantly, the average absolute deviation from the mean is reduced from 4.2 to 1.6 kcal/mol by the LKR model and to 1.9 kcal/mol by the NN (see Figure 4b). All the errors of the LKR model are below 8 kcal/mol, while the NN predictions on the tripeptide present two outliers of −15 and +26 kcal/mol. Additionally, the regression slope between the predictions and the target energies is also corrected to 0.99 for all the sets (see Figure S4b). These results are crucial since the standard deviation and the regression slope are the most important quantities for conformational sampling. Even a slight deviation from 1 in the regression slope causes significant changes in the resulting free energy surfaces. In particular, the observed regression slope between DFTB and PBE at 300 K (0.75) is roughly equivalent to perform sampling with a temperature 1.33 times higher (e.g., 400 K instead of 300 K). At the same time, outliers can lead to unstable dynamics and alter the results of sampling simulations.

Overall, while the NN model performs better on the dipeptide test structures, the LKR provides a more robust extrapolation (lower MAE, less outliers) for the Phe-Gly-Phe tripeptide. It must be noted that the superior stability of LKR

https://doi.org/10.1021/acs.jctc.1c00813
J. Chem. Theory Comput. 2022, 18, 1467−1479

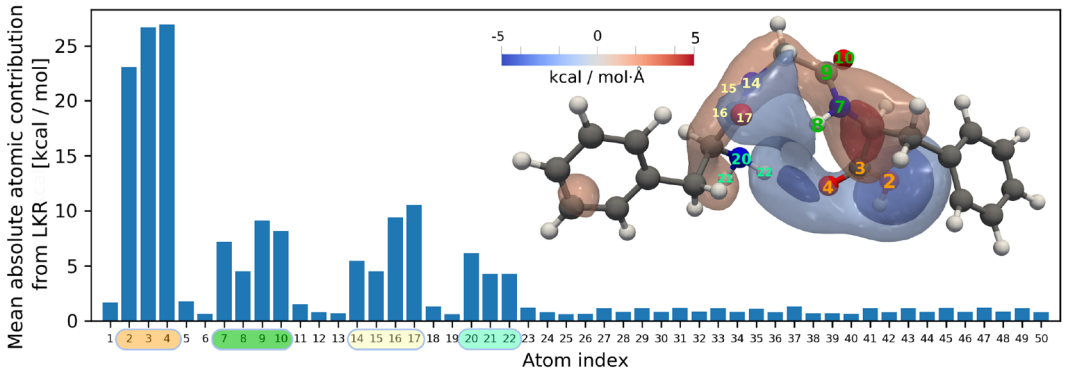


Figure 5. Histogram with the mean absolute atomic contribution to the LKR corrections for the tripeptide for the 2,000 test structures. The figure includes a particular conformation of the tripeptide with isosurfaces of a scalar field representing the localization of the ML correction. The scalar field was generated with the LKR atomic corrections to the energy for that structure, convoluted with the atomic positions and a Gaussian filter of width 1 Å. The isosurfaces correspond to the isovalues −5, −2, +2, and +5.

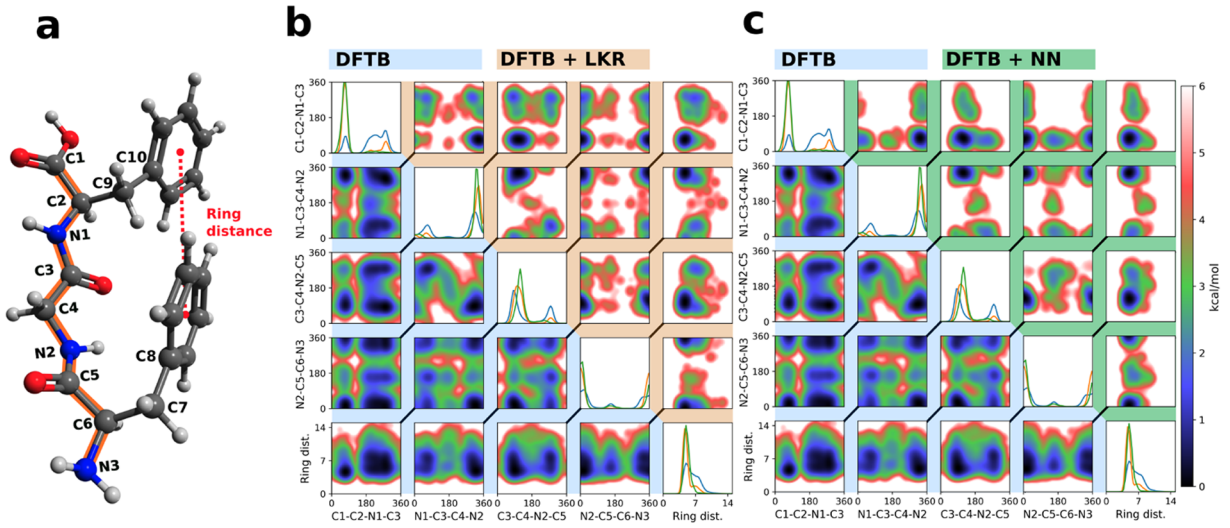


Figure 6. (a) Tripeptide Phe-Gly-Phe with highlighted atoms used for the collective variables in the analysis of the sampling simulations. (b and c) Grids with 2D free energy landscapes for each pair of the selected collective variables. The lower diagonals contain results from T-RE simulations using DFTB-D3(BJ). The upper diagonals contains the results of the resH-RE simulations using DFTB-D3(BJ) + LKR (b) and DFTB-D3(BJ) + NN (c). In the diagonal are the probability distributions of each collective variable for DFTB-D3(BJ)(blue), DFTB-D3(BJ) + LKR(orange), and DFTB-D3(BJ) + NN(green).

is not a consequence of using only energy data for the training. In fact, the NN model trained only with energies shows much poorer transferability and scalability capabilities (see Figure S5).

As shown in the previous section, the atomic decomposition of the ML correction naturally provides a measure of the error localization in the molecule. To visualize the error for the tripeptide, we constructed a scalar field using atomic-centered Gaussian functions scaled such as to match the LKR atomic predictions (see Figure 5). With the use of this procedure, it is possible to construct a real-space map highlighting the regions of the tripeptide where the DFTB potential deviates from the PBE reference. An example of these critical regions is identifiable between the oxygen and hydrogen atom forming an intramolecular hydrogen bond (e.g., between atoms 4 and 22 in Figure 5). In Figure 3, we have shown that the hydrogens bound to an oxygen or a nitrogen are the most difficult to describe at the DFTB level. Figure 5 shows the understabilization of the hydrogen bond between the NH_2 and the CO by DFTB, which is corrected by our models. However, this particular example does not imply that all hydrogen bonds are poorly described and in a systematic manner. For example, equivalent figures show that the OCO−H bond in the dipeptide of aspartate is actually overstabilized by DFTB, while the CO−HN in the protonated histidine is understabilized (see Figure S3). These inconsistencies have been shown to arise at the DFTB level due to a poor description of short-range electrostatic and polarization interactions arising from the use of a minimal valence basis.[108] While several empirical corrections to DFTB and more generally to semiempirical methods have been proposed,[108−112] the use of the D3H5 correction (the last of such corrections DFTB-D3H5[113]) does not change the performance of DFTB on the dipeptide set significantly (see Figure S7).

Furthermore, the analysis reported in Figure 5 shows that the description of the hydrogen-bond interactions are not the only limitation of DFTB. More generally, the highest absolute ML corrections appears whenever the bond between two atoms is polarized, such as in the region of the terminal carboxylic acid (atoms 2, 3, and 4 in Figure 5) and the amide

https://doi.org/10.1021/acs.jctc.1c00813
J. Chem. Theory Comput. 2022, 18, 1467−1479

moiety of the peptide bond (atoms 7, 8, 9, and 10 in Figure 5). In contrast to existing corrections, which are not meant to improve the description of these polarized bonds, the ML models guarantee by construction an equally accurate description for all the regions.

Free Energy Surface of Tripeptides. Having assessed the robustness of the ML models by evaluating the accuracy of the energy predictions on Phe-Gly-Phe tripeptide conformations and by providing comparisons with uncorrected DFTB, this section goes further and applies the ML corrections to sample the free-energy landscape of the tripeptide in the gas phase. As described in the Computational Details section, for the LKR model we use the resH-RE approach for a 300 K canonical sampling of the tripeptide generated with DFTB as a reservoir to accelerate the DFTB+LKR sampling without the need for high temperatures or bias potentials. The reservoir was generated using T-RE because in cases such as this where the barriers between conformers are not expected to be too high, and where relatively long trajectories are affordable, it allows one to obtain an explicit canonical distribution of conformers (that is needed for the reservoir of resH-RE simulations) without having to determine CVs and/or perform reweighting steps. The use of other accelerated sampling schemes (resH-RE or ATLAS) becomes necessary in the presence of a rougher PES. Figure 6a shows the set of characteristic collective variables (CVs) chosen to analyze the free-energy landscape. The set of CVs includes all the Ramachandran dihedral angles as well as the distance between the benzene rings at each end of the chain. To visually represent the resultant seven-dimensional FES, the 2D FESs for all pairs of CVs are obtained by marginalizing the seven-dimensional distribution. The DFTB-based 2D FES DFTB are shown in the lower triangle of Figure 6b,c, while their DFTB +LKR-based counterparts are shown in the upper triangle of Figure 6b. The C4−N2−C5−C6 and C2−N1−C3−C4 dihedral angles were excluded from the plot because their values remain constant throughout the sampling (see Figure S9).

To provide a complementary view, we further sample the same tripeptide FESs using the committee of NN models (upper diagonal of Figure 6c). We exploit the availability of forces to perform well-tempered metadynamics simulations and make use of the ability to assess the uncertainties in the predicted corrections to smoothly fall back onto the DFTB baseline when the NN predictions become uncertain. This suppresses instabilities in the dynamics due to unphysical NN corrections in areas of the PES, which are underrepresented in the training data. The metadynamics are biased in the six-dimensional CV space spanned by the six Ramachandran dihedral angles (for further information see the Computational Details section).

The comparison between the DFTB T-RE results (lower triangular portions of Figure 6) and the results of the ML potentials (upper triangular portions of Figure 6) shows the effects of correcting the flat PES on the final free energy landscape. In addition to increasing the free energy barriers, translated in very low populations in basin transition areas, the ML corrections dramatically affect the relative stability of the different basins, altering the qualitative dynamic behavior of the tripeptide at 300 K. These effects can be equally observed in both the sampling based on LKR and NN. The results obtained show good agreement. The single CV populations are nearly identical, and the lowest free energy minima are unequivocally determined. However, some disagreement in the free energy surfaces obtained by sampling using the two ML frameworks can be observed for the higher-energy portions of the FESs. Given the highly nontrivial nature of this exercise, it is not easy to pinpoint the source of the discrepancy. The entanglement between uncertainties arising from (i) finite statistics and (ii) possible discrepancies of the ML models complicates the analysis of their relative weight. As a benchmark, sampling results following the same methodologies applied to the alanine dipeptide show a very good agreement between the DFTB+LKR and DFTB+NN approaches (see Figure S11). Overall, it is clear that both ML-corrected frameworks predict a much sharper variation of the free energy compared with DFTB that instead predicts a very smooth landscape as a function of the dihedral angles. This qualitative difference is also clearly visible in a 2D Sketchmap[114,115] projection (Figure 7), which indicates that the more diffuse structural distribution at DFTB is a direct consequence of the flatness of the associated PES.

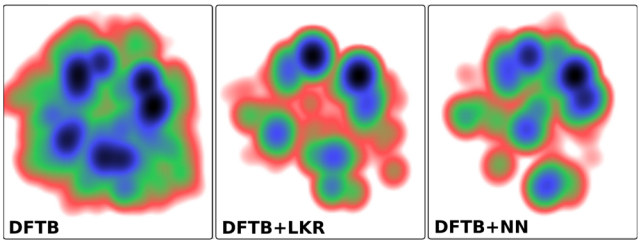


Figure 7. Sketchmap computed with DFTB-D3(BJ) (left) DFTB-D3(BJ) + LKR (middle) and DFTB-D3(BJ) + NN (right) sampling at 300 K using the selected CVs from Figure 6.

As a final note, it is important to stress that the generation of converged statistics using the target potential (PBE) would have been computationally unfeasible. While appealing, an alternative comparison with experimental results would require incorporating solvent effects, which is outside the scope of this work.

CONCLUSION

We introduced LKR-OMP, a local kernel regression model which exploits the supervised sparsity algorithm OMP and compared its performance along with that of a Behler-Parrinello neural network. LKR-OMP benefits from the straightforward training of kernel methods, combining it with the scalability and transferability of models based on neural networks.

We juxtapose the two approaches by applying them to the challenging task of learning the PES of oligopeptides at the PBE-dDsC level, using the semiempirical DFTB-D3(BJ) potential as a baseline and training on a combination of dipeptide structures and dimers of small organic fragments. To achieve comparable computational cost between sparse kernel regression and NNs, it is essential to select carefully the most representative environments. We show, both by comparing the final model accuracy and by combining the representation score with a 2D projection of the local atomic environments, that selection methods relying exclusively on structural information, such as FPS or CUR, are not always optimal and that substantial improvements can be achieved with the supervised strategy adopted in the LKR-OMP scheme.

https://doi.org/10.1021/acs.jctc.1c00813
J. Chem. Theory Comput. 2022, 18, 1467−1479

Using only energies for training, the LRK-OMP model achieves an accuracy and transferability compared to that of the NN-based model, that also uses forces to optimize its parameters. Thanks to the atom-centered construction of the ML correction, we can reveal the origin of the DFTB-D3(BJ) error relative to DFT, interpret in terms of chemical and atomic patterns, and demonstrate the relevance of relying upon a correction based on nonlinear regression techniques. As a final demonstration of the possibilities brought about by the use of ML corrections of the PES, we use them in combination with enhanced sampling approaches to explore the conformational energy landscape of the tripeptide Phe-Gly-Phe at an effective PBE-dDsC level. We use two different sampling strategies: resH-RE for LKR-OMP, which at present does not provide easy access to energy derivatives, and ATLAS metadynamics for the NN potential, that instead does. The free energy landscapes obtained with the two frameworks are consistent with each other, and show striking differences compared to the uncorrected baseline potential. This provides another example of the exaggerated smoothness of the DFTB potentials and highlights the dire need to make the accuracy of higher electronic structure levels accessible to the size and time scale that are necessary for free energy computations. In this respect, the fact that ML corrections have now become a mature, trustworthy approach to achieve this goal, with entirely different frameworks achieving comparable accuracy and efficiency, is very encouraging. The LKR-OMP model, in particular, offers a good compromise in terms of data-intensiveness, computational cost, generality and accuracy, in addition to providing unique analytical insight into the model performance.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jctc.1c00813.

Additional details of the local kernel regression model, the relationship between the accuracy of LKR and the number of reference environments, correlation plots between the predicted energies of each of the dipeptides versus the PBE-dDsC reference, the isosurfaces of the ML correction scalar field on the tripeptide, more details on the differences between the DFTB-D3(BJ) and PBE-dDsC tripeptide energies, and additional 2D free energy landscapes for each pair of the collective variable considered in this work (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Clemence Corminboeuf − *Laboratory for Computational Molecular Design, Institute of Chemical Sciences and Engineering and National Centre for Computational Design and Discovery of Novel Materials (MARVEL), École Polytechnique Fédérale de Lausanne, CH-1015 Lausanne, Switzerland;* orcid.org/0000-0001-7993-2879; Email: clemence.corminboeuf@epfl.ch

Authors

Raimon Fabregat − *Laboratory for Computational Molecular Design, Institute of Chemical Sciences and Engineering, École Polytechnique Fédérale de Lausanne, CH-1015 Lausanne, Switzerland*

Alberto Fabrizio − *Laboratory for Computational Molecular Design, Institute of Chemical Sciences and Engineering and National Centre for Computational Design and Discovery of Novel Materials (MARVEL), École Polytechnique Fédérale de Lausanne, CH-1015 Lausanne, Switzerland;* orcid.org/0000-0002-4440-3149

Edgar A. Engel − *Laboratory of Computational Science and Modeling, IMX, École Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland; National Centre for Computational Design and Discovery of Novel Materials (MARVEL), École Polytechnique Fédérale de Lausanne, CH-1015 Lausanne, Switzerland*

Benjamin Meyer − *Laboratory for Computational Molecular Design, Institute of Chemical Sciences and Engineering and National Centre for Computational Design and Discovery of Novel Materials (MARVEL), École Polytechnique Fédérale de Lausanne, CH-1015 Lausanne, Switzerland*

Veronika Juraskova − *Laboratory for Computational Molecular Design, Institute of Chemical Sciences and Engineering, École Polytechnique Fédérale de Lausanne, CH-1015 Lausanne, Switzerland*

Michele Ceriotti − *Laboratory of Computational Science and Modeling, IMX, École Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland; National Centre for Computational Design and Discovery of Novel Materials (MARVEL), École Polytechnique Fédérale de Lausanne, CH-1015 Lausanne, Switzerland;* orcid.org/0000-0003-2571-2832

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jctc.1c00813

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

R.F. and C.C. acknowledge funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (Grant 817977). A.F., B.M., and E.E. were funded by the National Centre of Competence in Research (NCCR) "Materials' Revolution: Computational Design and Discovery of Novel Materials (MARVEL)". V.J. acknowledges the Swiss National Science Foundation (SNSF Grant 200020_175496) for financial support.

■ REFERENCES

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

(2) Chmiela, S.; Tkatchenko, A.; Sauceda, H. E.; Poltavsky, I.; Schütt, K. T.; Müller, K.-R. Machine learning of accurate energy-conserving molecular force fields. *Sci. Adv.* **2017**, *3*, No. e1603015.

(3) Fabregat, R.; Fabrizio, A.; Meyer, B.; Hollas, D.; Corminboeuf, C. Hamiltonian-Reservoir Replica Exchange and Machine Learning Potentials for Computational Organic Chemistry. *J. Chem. Theory Comput.* **2020**, *16*, 3084−3094.

(4) 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.

(5) Behler, J. First Principles Neural Network Potentials for Reactive Simulations of Large Molecular and Condensed Systems. *Angew. Chem., Int. Ed.* **2017**, *56*, 12828−12840.

(6) Schütt, K. T.; Sauceda, H. E.; Kindermans, P.-J.; Tkatchenko, A.; Müller, K.-R. SchNet − A deep learning architecture for molecules and materials. *J. Chem. Phys.* **2018**, *148*, 241722.
(7) Schütt, K.; Kessel, P.; Gastegger, M.; Nicoli, K.; Tkatchenko, A.; Muller, K.-R. SchNetPack: A deep learning toolbox for atomistic systems. *J. Chem. Theory and Comput.* **2019**, *15*, 448−455.
(8) Schütt, K.; Gastegger, M.; Tkatchenko, A.; Müller, K.-R.; Maurer, R. J. Unifying machine learning and quantum chemistry with a deep neural network for molecular wavefunctions. *Nat. Commun.* **2019**, *10*, 1−10.
(9) Unke, O. T.; Chmiela, S.; Gastegger, M.; Schütt, K. T.; Sauceda, H. E.; Müller, K.-R. Spookynet: Learning force fields with electronic degrees of freedom and nonlocal effects. *Nat. Commun.* **2021**, *12*, 1.
(10) Unke, O. T.; Meuwly, M. A reactive, scalable, and transferable model for molecular energies from a neural network approach based on local information. *J. Chem. Phys.* **2018**, *148*, 241708.
(11) Unke, O. T.; Meuwly, M. *J. Chem. Theory Comput.* **2019**, *15*, 3678.
(12) 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.
(13) 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*, 2903.
(14) Hastie, T.; Friedman, J.; Tibshirani, R. *The Elements of Statistical Learning*, 2nd ed.; Springer, 2001; New York, USA.
(15) Bartók, A. P.; Payne, M. C.; Kondor, R.; Csányi, G. Gaussian Approximation Potentials: The Accuracy of Quantum Mechanics, without the Electrons. *Phys. Rev. Lett.* **2010**, *104*, 136403.
(16) Bartók, A. P.; Csányi, G. Gaussian approximation potentials: A brief tutorial introduction. *Int. J. Quantum Chem.* **2015**, *115*, 1051−1057.
(17) Bartók, A. P.; De, S.; Poelking, C.; Bernstein, N.; Kermode, J. R.; Csányi, G.; Ceriotti, M. Machine learning unifies the modeling of materials and molecules. *Sci. Adv.* **2017**, *3*, No. e1701816.
(18) Bartók, A. P.; Kermode, J.; Bernstein, N.; Csányi, G. Machine learning a general-purpose interatomic potential for silicon. *Phys. Rev. X* **2018**, *8*, 041048.
(19) Deringer, V. L.; Bernstein, N.; Bartók, A. P.; Cliffe, M. J.; Kerber, R. N.; Marbella, L. E.; Grey, C. P.; Elliott, S. R.; Csányi, G. Realistic Atomistic Structure of Amorphous Silicon from Machine-Learning-Driven Molecular Dynamics. *J. Phys. Chem. Lett.* **2018**, *9*, 2879−2885.
(20) Zhang, C.; Sun, Q. Gaussian approximation potential for studying the thermal conductivity of silicene. *J. Appl. Phys.* **2019**, *126*, 105103.
(21) Deringer, V. L.; Bernstein, N.; Csányi, G.; Mahmoud, C. B.; Ceriotti, M.; Wilson, M.; Drabold, D. A.; Elliott, S. R. Origins of structural and electronic transitions in disordered silicon. *Nature* **2021**, *589*, 59−64.
(22) Mocanu, F. C.; Konstantinou, K.; Lee, T. H.; Bernstein, N.; Deringer, V. L.; Csányi, G.; Elliott, S. R. Modeling the phase-change memory material, Ge2Sb2Te5, with a machine-learned interatomic potential. *J. Phys. Chem. B* **2018**, *122*, 8998−9006.
(23) Mocanu, F.; Konstantinou, K.; Elliott, S. Quench-rate and size-dependent behaviour in glassy Ge2Sb2Te5 models simulated with a machine-learned Gaussian approximation potential. *J. Phys. D: Appl. Phys.* **2020**, *53*, 244002.
(24) Balabin, R. M.; Lomakina, E. I. Support vector machine regression (LS-SVM)—an alternative to artificial neural networks (ANNs) for the analysis of quantum chemistry data? *Phys. Chem. Chem. Phys.* **2011**, *13*, 11710.
(25) Thompson, A.; Swiler, L.; Trott, C.; Foiles, S.; Tucker, G. Spectral neighbor analysis method for automated generation of quantum-accurate interatomic potentials. *J. Comput. Phys.* **2015**, *285*, 316−330.
(26) Ho, T.-S.; Rabitz, H. A general method for constructing multidimensional molecular potential energy surfaces from ab initio calculations. *J. Chem. Phys.* **1996**, *104*, 2584−2597.
(27) Unke, O. T.; Meuwly, M. Toolkit for the construction of reproducing kernel-based representations of data: Application to multidimensional potential energy surfaces. *J. Chem. Inf. Model.* **2017**, *57*, 1923−1931.
(28) Cendagorta, J. R.; Tolpin, J.; Schneider, E.; Topper, R. Q.; Tuckerman, M. E. Comparison of the performance of machine learning models in representing high-dimensional free energy surfaces and generating observables. *J. Phys. Chem. B* **2020**, *124*, 3647−3660.
(29) Chmiela, S.; Sauceda, H. E.; Müller, K.-R.; Tkatchenko, A. Towards exact molecular dynamics simulations with machine-learned force fields. *Nat. Commun.* **2018**, *9*, 3887 DOI: 10.1038/s41467-018-06169-2.
(30) Sauceda, H. E.; Chmiela, S.; Poltavsky, I.; Müller, K.-R.; Tkatchenko, A. Molecular force fields with gradient-domain machine learning: Construction and application to dynamics of small molecules with coupled cluster forces. *J. Chem. Phys.* **2019**, *150*, 114102.
(31) Artrith, N.; Urban, A. An implementation of artificial neural-network potentials for atomistic materials simulations: Performance for TiO2. *Comput. Matter. Sci.* **2016**, *114*, 135−150.
(32) Kobayashi, R.; Giofré, D.; Junge, T.; Ceriotti, M.; Curtin, W. A. Neural network potential for Al-Mg-Si alloys. *Phys. Rev. Mater.* **2017**, *1*, 053604.
(33) Ragoza, M.; Hochuli, J.; Idrobo, E.; Sunseri, J.; Koes, D. R. Protein−ligand scoring with convolutional neural networks. *J. Chem. Inf. Mod.* **2017**, *57*, 942−957.
(34) Janet, J. P.; Kulik, H. J. Predicting electronic structure properties of transition metal complexes with neural networks. *Chem. Sci.* **2017**, *8*, 5137−5152.
(35) Hellstrom, M.; Ceriotti, M.; Behler, J. Nuclear quantum effects in sodium hydroxide solutions from neural network molecular dynamics simulations. *J. Phys. Chem. B* **2018**, *122*, 10158−10171.
(36) Zhang, L.; Han, J.; Wang, H.; Car, R.; Weinan, E. Deep potential molecular dynamics: a scalable model with the accuracy of quantum mechanics. *Phys. Rev. Lett.* **2018**, *120*, 143001.
(37) Noé, F.; Tkatchenko, A.; Müller, K.-R.; Clementi, C. Machine learning for molecular simulation. *Annu. Rev. Phys. Chem.* **2020**, *71*, 361−390.
(38) Rossi, K.; Jurásková, V.; Wischert, R.; Garel, L.; Corminbœuf, C.; Ceriotti, M. Simulating solvation and acidity in complex mixtures with first-principles accuracy: the case of CH3SO3H and H2O2 in phenol. *J. Chem. Theory Comput.* **2020**, *16*, 5139−5149.
(39) Manzhos, S.; Carrington, T., Jr Neural network potential energy surfaces for small molecules and reactions. *Chem. Rev.* **2021**, *121*, 10187−10217.
(40) Rupp, M.; Tkatchenko, A.; Müller, K.-R.; von Lilienfeld, O. A. Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning. *Phys. Rev. Lett.* **2012**, *108*, 058301.
(41) 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.
(42) Huang, B.; von Lilienfeld, O. A. Quantum machine learning using atom-in-molecule-based fragments selected on the fly. *Nat. Chem.* **2020**, *12*, 945−951.
(43) Fabrizio, A.; Meyer, B.; Fabregat, R.; Corminboeuf, C. Quantum Chemistry Meets Machine Learning. *CHIMIA* **2019**, *73*, 983−989.
(44) Meyer, B.; Sawatlon, B.; Heinen, S.; von Lilienfeld, O. A.; Corminboeuf, C. Machine learning meets volcano plots: Computational discovery of cross-coupling catalysts. *Chem. Sci.* **2018**, *9*, 7069−7077.
(45) von Lilienfeld, O. A.; Müller, K.-R.; Tkatchenko, A. Exploring chemical compound space with quantum-based machine learning. *Nat. Rev. Chem.* **2020**, *4*, 347−358.

(46) Fabrizio, A.; Briling, K. R.; Girardier, D. D.; Corminboeuf, C. Learning on-top: Regressing the on-top pair density for real-space visualization of electron correlation. *J. Chem. Phys.* **2020**, *153*, 204111.
(47) Faber, F. A.; Christensen, A. S.; Huang, B.; von Lilienfeld, O. A. Alchemical and structural distribution based representation for universal quantum machine learning. *J. Chem. Phys. B* **2018**, *148*, 241717.
(48) Christensen, A. S.; Bratholm, L. A.; Faber, F. A.; von Lilienfeld, O. A. FCHL revisited: Faster and more accurate quantum machine learning. *J. Chem. Phys.* **2020**, *152*, 044107.
(49) Bartók, A. P.; Kondor, R.; Csányi, G. On representing chemical environments. *Phys. Rev. B* **2013**, *87*, 184115.
(50) 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.
(51) Käser, S.; Koner, D.; Christensen, A. S.; von Lilienfeld, O. A.; Meuwly, M. Machine Learning Models of Vibrating H_2CO: Comparing Reproducing Kernels, FCHL, and PhysNet. *J. Phys. Chem. A* **2020**, *124*, 8853−8865.
(52) De, S.; Bartók, A. P.; Csányi, G.; Ceriotti, M. Comparing molecules and solids across structural and alchemical space. *Phys. Chem. Chem. Phys.* **2016**, *18*, 13754−13769.
(53) Imbalzano, G.; Anelli, A.; Giofré, D.; Klees, S.; Behler, J.; Ceriotti, M. Automatic selection of atomic fingerprints and reference configurations for machine-learning potentials. *J. Chem. Phys.* **2018**, *148*, 241730.
(54) Helfrecht, B. A.; Cersonsky, R. K.; Fraux, G.; Ceriotti, M. Structure-property maps with Kernel principal covariates regression. *Mach. Learn.: Sci. Technol.* **2020**, *1*, 045021.
(55) Gallarati, S.; Fabregat, R.; Laplaza, R.; Bhattacharjee, S.; Wodrich, M. D.; Corminboeuf, C. Reaction-based machine learning representations for predicting the enantioselectivity of organocatalysts. *Chem. Sci.* **2021**, *12*, 6879−6889.
(56) 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.
(57) Gaus, M.; Goez, A.; Elstner, M. Parametrization and Benchmark of DFTB3 for Organic Molecules. *J. Chem. Theory Comput.* **2013**, *9*, 338−354.
(58) Gaus, M.; Lu, X.; Elstner, M.; Cui, Q. Parameterization of DFTB3/3OB for Sulfur and Phosphorus for Chemical and Biological Applications. *J. Chem. Theory Comput.* **2014**, *10*, 1518−1537.
(59) Grimme, S.; Ehrlich, S.; Goerigk, L. Effect of the damping function in dispersion corrected density functional theory. *J. Comput. Chem.* **2011**, *32*, 1456−1465.
(60) Perdew, J. P.; Burke, K.; Ernzerhof, M. Generalized Gradient Approximation Made Simple. *Phys. Rev. Lett.* **1996**, *77*, 3865−3868.
(61) Steinmann, S. N.; Corminboeuf, C. A System-Dependent Density-Based Dispersion Correction. *J. Chem. Theory Comput.* **2010**, *6*, 1990−2001.
(62) Steinmann, S. N.; Corminboeuf, C. Comprehensive Benchmarking of a Density-Dependent Dispersion Correction. *J. Chem. Theory Comput.* **2011**, *7*, 3567−3577.
(63) Steinmann, S. N.; Corminboeuf, C. A generalized-gradient approximation exchange hole model for dispersion coefficients. *J. Chem. Phys.* **2011**, *134*, 044117.
(64) 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*, 241725.
(65) Barker, J.; Bulin, J.; Hamaekers, J.; Mathias, S. *Sci. Comput. Algorithms Ind. Simulations*; Springer International Publishing: Cham, 2017; pp 25−42.
(66) Pati, Y. C.; Rezaiifar, R.; Krishnaprasad, P. S. Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition. *Proceedings of 27th Asilomar conference on signals, systems and computers* **1993**, 40−44.
(67) Sivaraman, G.; Krishnamoorthy, A. N.; Baur, M.; Holm, C.; Stan, M.; Csányi, G.; Benmore, C.; Vázquez-Mayagoitia, Á. Machine-learned interatomic potentials by active learning: amorphous and liquid hafnium dioxide. *npj Comput. Mater.* **2020**, *6*, 1−8.
(68) Behler, J. Neural network potential-energy surfaces in chemistry: a tool for large-scale simulations. *Phys. Chem. Chem. Phys.* **2011**, *13*, 17930−55.
(69) Musil, F.; Willatt, M. J.; Langovoy, M. A.; Ceriotti, M. Fast and Accurate Uncertainty Estimation in Chemical Machine Learning. *J. Chem. Theory Comput.* **2019**, *15*, 906.
(70) Imbalzano, G.; Zhuang, Y.; Kapil, V.; Rossi, K.; Engel, E. A.; Grasselli, F.; Ceriotti, M. Uncertainty estimation for molecular dynamics and sampling. *J. Chem. Phys.* **2021**, *154*, 074102.
(71) Richardson, J. S. *Advances in protein chemistry*; Elsevier, 1981; Vol. *34*; pp 167−339.
(72) Ropo, M.; Schneider, M.; Baldauf, C.; Blum, V. First-principles data set of 45,892 isolated and cation-coordinated conformers of 20 proteinogenic amino acids. *Sci. Data* **2016**, *3*, 160009.
(73) Burns, L. A.; Faver, J. C.; Zheng, Z.; Marshall, M. S.; Smith, D. G. A.; Vanommeslaeghe, K.; MacKerell, A. D.; Merz, K. M.; Sherrill, C. D. The BioFragment Database (BFDb): An open-data platform for computational chemistry analysis of noncovalent interactions. *J. Chem. Phys.* **2017**, *147*, 161727.
(74) Fukunishi, H.; Watanabe, O.; Takada, S. On the Hamiltonian replica exchange method for efficient sampling of biomolecular systems: Application to protein structure prediction. *J. Chem. Phys.* **2002**, *116*, 9058−9067.
(75) Frenkel, D.; Smit, B. *Understanding Molecular Simulation: From Algorithms to Applications*, 2nd ed.; Elsevier, 2001; London, UK.
(76) To obtain the LKR energy derivatives, it is necessary to derive both the kernel and the underlying molecular representation with respect to the nuclear coordinates. The SLATM representation used in this work has a rather cumbersome mathematical form, whose analytical derivatives are not readily obtained.
(77) Hu, D.; Xie, Y.; Li, X.; Li, L.; Lan, Z. Inclusion of Machine Learning Kernel Ridge Regression Potential Energy Surfaces in On-the-Fly Nonadiabatic Molecular Dynamics Simulation. *J. Phys. Chem. Lett.* **2018**, *9*, 2725−2732.
(78) Giberti, F.; Tribello, G. A.; Ceriotti, M. Global Free-Energy Landscapes as a Smoothly Joined Collection of Local Maps. *J. Chem. Theory Comput.* **2021**, *17*, 3292−3308.
(79) Abrams, J. B.; Tuckerman, M. E. Efficient and direct generation of multidimensional free energy surfaces via adiabatic dynamics without coordinate transformations. *J. Phys. Chem. B* **2008**, *112*, 15742−15757.
(80) Chen, M.; Cuendet, M. A.; Tuckerman, M. E. Heating and flooding: A unified approach for rapid generation of free energy surfaces. *J. Chem. Phys.* **2012**, *137*, 024102.
(81) Gasparotto, P.; Meißner, R. H.; Ceriotti, M. Recognizing Local and Global Structural Motifs at the Atomic Scale. *J. Chem. Theory Comput.* **2018**, *14*, 486−498.
(82) Giberti, F.; Cheng, B.; Tribello, G. A.; Ceriotti, M. Iterative unbiasing of quasi-equilibrium sampling. *J. Chem. Theory Comput.* **2020**, *16*, 100−107.
(83) Aradi, B.; Hourahine, B.; Frauenheim, T. DFTB+, a Sparse Matrix-Based Implementation of the DFTB Method. *J. Phys. Chem. A* **2007**, *111*, 5678−5684.
(84) Schmidt, M. W.; Baldridge, K. K.; Boatz, J. A.; Elbert, S. T.; Gordon, M. S.; Jensen, J. H.; Koseki, S.; Matsunaga, N.; Nguyen, K. A.; Su, S.; et al. General atomic and molecular electronic structure system. *J. Comput. Chem.* **1993**, *14*, 1347−1363.
(85) Gordon, M. S.; Schmidt, M. W. *Theory and Applications of Computational Chemistry*; Elsevier, 2005; pp 1167−1189.
(86) Petraglia, R.; Nicolaï, A.; Wodrich, M. D.; Ceriotti, M.; Corminboeuf, C. Beyond static structures: Putting forth REMD as a tool to solve problems in computational organic chemistry. *J. Comput. Chem.* **2016**, *37*, 83−92.

(87) Ceriotti, M.; More, J.; Manolopoulos, D. E. i-PI A Python interface for ab initio path integral molecular dynamics simulations. *Comput. Phys. Commun.* **2014**, *185*, 1019−1026.
(88) Kapil, V.; Rossi, M.; Marsalek, O.; Petraglia, R.; Litman, Y.; Spura, T.; Cheng, B.; Cuzzocrea, A.; Meißner, R. H.; M.Wilkins, D.; et al. i-PI 2.0: A universal force engine for advanced molecular simulations. *Comput. Phys. Commun.* **2019**, *236*, 214.
(89) Plumed Consortium. Promoting transparency and reproducibility in enhanced molecular simulations. *Nat. Methods* **2019**, *16*, 670−673.
(90) Tribello, G. A.; Bonomi, M.; Branduardi, D.; Camilloni, C.; Bussi, G. PLUMED2: New feathers for an old bird. *Comput. Phys. Commun.* **2014**, *185*, 604.
(91) Plimpton, S. Fast parallel algorithms for short-range molecular dynamics. *J. Compt. Phys.* **1995**, *117*, 1−19.
(92) Ceriotti, M.; Bussi, G.; Parrinello, M. Colored-noise thermostats à la carte. *J. Chem. Theory Comput.* **2010**, *6*, 1170−1180.
(93) Ceriotti, M.; Manolopoulos, D. E.; Parrinello, M. Accelerating the convergence of path integral dynamics with a generalized Langevin equation. *J. Chem. Phys.* **2011**, *134*, 084104.
(94) Fabregat, R.; Fabrizio, A.; Corminboeuf, C. *Local Kernel Regression*. 2021, DOI: 10.5281/zenodo.5172581.
(95) Christensen, A.; Faber, F.; Huang, B.; Bratholm, L.; Tkatchenko, A.; Müller, K.; von Lilienfeld, O. A. *QML: A Python Toolkit for Quantum Machine Learning*. 2017, DOI: 10.5281/zenodo.817332.
(96) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825−2830.
(97) Singraber, A.; Behler, J.; Dellago, C. A library-based LAMMPS implementation of high-dimensional neural network potentials. *J. Chem. Theory Comput.* **2019**, *15*, 1827.
(98) Cersonsky, R. K.; Helfrecht, B.; Engel, E. A.; Kliavinek, S.; Ceriotti, M. Improving sample and feature selection with principal covariates regression. *Mach. Learn.: Sci. Technol.* **2021**, *2*, 035038.
(99) Engel, E. A.; Kapil, V.; Ceriotti, M. Importance of Nuclear Quantum Effects for NMR Crystallography. *J. Phys. Chem. Lett.* **2021**, *12*, 7701−7707.
(100) Kapil, V.; Engel, E. A. A complete description of thermodynamic stabilities of molecular. *Proc. Natl. Acad. Sci. U.S.A.* **2022**, *119*, No. e2111769119.
(101) Cheng, B.; Engel, E. A.; Behler, J.; Dellago, C.; Ceriotti, M. Ab initio thermodynamics of liquid and solid water. *Proc. Natl. Acad. Sci. U.S.A.* **2019**, *116*, 1110−1115.
(102) Imbalzano, G.; Ceriotti, M. Modeling the Ga/As binary system across temperatures and compositions from first principles. *Phys. Rev. Mater.* **2021**, *5*, 063804.
(103) Huang, M.; Dissanayake, T.; Kuechler, E.; Radak, B. K.; Lee, T.-S.; Giese, T. J.; York, D. M. A multidimensional B-spline correction for accurate modeling sugar puckering in QM/MM simulations. *J. Chem. Theory Comput.* **2017**, *13*, 3975−3984.
(104) Van der Maaten, L.; Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579−2605.
(105) Fabregat, R.; Blaskovits, T.; Corminboeuf, C. *Molecular Explorer: A python based web app framework to visualize and share chemical data*. Zenodo, 2020; DOI: 10.5281/zenodo.4564039.
(106) Valdes, H.; Pluhackova, K.; Hobza, P. Phenylalanyl-Glycyl-Phenylalanine Tripeptide: A Model System for Aromatic-Aromatic Side Chain Interactions in Proteins. *J. Chem. Theory Comput.* **2009**, *5*, 2248−2256.
(107) Blyth, C. R. On Simpson's paradox and the sure-thing principle. *J. Am. Stat. Assoc.* **1972**, *67*, 364−366.
(108) Miriyala, V. M.; Řezáč, J. Description of non-covalent interactions in SCC-DFTB methods. *J. Comput. Chem.* **2017**, *38*, 688−697.
(109) Řezáč, J.; Fanfrlík, J.; Salahub, D.; Hobza, P. Semiempirical Quantum Chemical PM6Method Augmented by Dispersion and H-Bonding Correction Terms Reliably Describes Various Types of Noncovalent Complexes. *J. Chem. Theory Comput.* **2009**, *5*, 1749−1760.
(110) Korth, M.; Pitoňák, M.; Řezáč, J.; Hobza, P. A Transferable H-Bonding Correction for Semiempirical Quantum-Chemical Methods. *J. Chem. Theory Comput.* **2010**, *6*, 344−352.
(111) Korth, M. Third-Generation Hydrogen-Bonding Corrections for Semiempirical QM Methods and Force Fields. *J. Chem. Theory Comput.* **2010**, *6*, 3808−3816.
(112) Řezáč, J.; Hobza, P. Advanced Corrections of Hydrogen Bonding and Dispersion for Semiempirical Quantum Mechanical Methods. *J. Chem. Theory Comput.* **2012**, *8*, 141−151.
(113) Řezáč, J. Empirical Self-Consistent Correction for the Description of Hydrogen Bonds in DFTB3. *J. Chem. Theory Comput.* **2017**, *13*, 4804−4817.
(114) Ceriotti, M.; Tribello, G. A.; Parrinello, M. Simplifying the representation of complex free-energy landscapes using sketch-map. *Proc. Natl. Acad. Sci. U.S.A.* **2011**, *108*, 13023−13028.
(115) Tribello, G. A.; Ceriotti, M.; Parrinello, M. Using sketch-map coordinates to analyze and bias molecular dynamics simulations. *Proc. Natl. Acad. Sci. U.S.A.* **2012**, *109*, 5196−5201.

https://doi.org/10.1021/acs.jctc.1c00813
J. Chem. Theory Comput. 2022, 18, 1467−1479