JCTC Journal of Chemical Theory and Computation

pubs.acs.org/JCTC Article

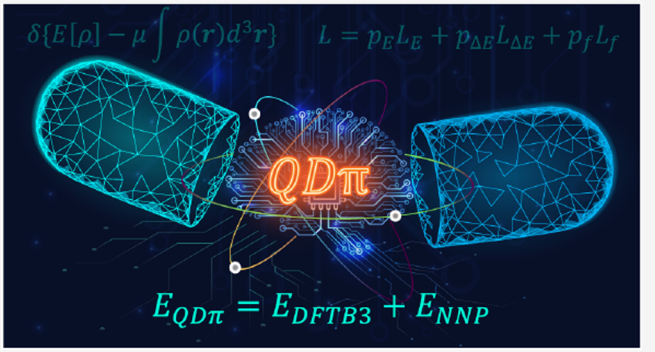
QDπ: A Quantum Deep Potential Interaction Model for Drug Discovery

Jinzhe Zeng, Yujun Tao, Timothy J. Giese, and Darrin M. York*

Cite This: J. Chem. Theory Comput. 2023, 19, 1261−1275 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: We report QDπ-v1.0 for modeling the internal energy of drug molecules containing H, C, N, and O atoms. The QDπ model is in the form of a quantum mechanical/machine learning potential correction (QM/Δ-MLP) that uses a fast third-order self-consistent density-functional tight-binding (DFTB3/3OB) model that is corrected to a quantitatively high-level of accuracy through a deep-learning potential (DeepPot-SE). The model has the advantage that it is able to properly treat electrostatic interactions and handle changes in charge/protonation states. The model is trained against reference data computed at the ωB97X/6-31G* level (as in the ANI-1x data set) and compared to several other approximate semiempirical and machine learning potentials (ANI-1x, ANI-2x, DFTB3, MNDO/d, AM1, PM6, GFN1-xTB, and GFN2-xTB). The QDπ model is demonstrated to be accurate for a wide range of intra- and intermolecular interactions (despite its intended use as an internal energy model) and has shown to perform exceptionally well for relative protonation/deprotonation energies and tautomers. An example application to model reactions involved in RNA strand cleavage catalyzed by protein and nucleic acid enzymes illustrates QDπ has average errors less than 0.5 kcal/mol, whereas the other models compared have errors over an order of magnitude greater. Taken together, this makes QDπ highly attractive as a potential force field model for drug discovery.



1. INTRODUCTION

Computational methods that enable the prediction of the binding affinity and selectivity of small molecule drugs to protein or nucleic acid targets are essential tools for drug discovery.[1−4] Among the most powerful of these methods are so-called "alchemical free energy" (AFE) simulations: physics-based approaches that strive to rigorously calculate the absolute and/or relative binding free energy (ABFE and RBFE, respectively) through atomistic simulations.[5] The accuracy of such predictions depends critically on the quality and robustness of the underlying potential energy model from which atomic forces are derived.[6]

ABFE and RBFE simulations require the construction of thermodynamic cycles whereby ligands must dynamically sample phase space in "unbound" (aqueous solution) and "target-bound" (e.g., protein complexed) environments. Electrostatic interactions in these environments can differ substantially, and hence a desirable feature of the ligand potential energy model is the ability to explicitly polarize in order to electronically respond to these changes.[7] Further, roughly 25% of potential drug molecules can exist in alternative tautomeric forms, and almost all of them can have multiple ionizable protonation states. These states are important as they are sensitive to their environment (e.g., pH, ionic conditions, aqueous versus membrane, etc.) and can change upon binding.[8−10] In order to accommodate these changes, it is advantageous to have a "universal" potential energy model that is not restricted to a specific predetermined bonding pattern or protonation state within the same simulation, unlike conventional molecular mechanical (MM) force fields (including polarizable force fields). High-level ab initio quantum mechanical (QM) models are universal in this sense and also have been demonstrated to be robust and accurate,[11] but these methods require tremendous computational resources making them intractable for routine simulations. Approximate "semi-empirical" QM models, on the other hand, are orders of magnitude faster and can be routinely applied in simulations where the QM region is limited to up to a few hundred atoms (which encompasses most drug molecules); however, these models typically do not have the quantitative accuracy that real-world drug discovery applications demand.[12]

Received: November 21, 2022
Published: January 25, 2023

https://doi.org/10.1021/acs.jctc.2c01172
J. Chem. Theory Comput. 2023, 19, 1261−1275

Table 1. Data Sets Used in the Current Work[a]

usage	data set	# of data points E	F	ΔE	ref
training	ANI-1x	4,956,005	227,101,443		36, 37
	ANI-1xm	2,641,429	130,421,121		36, 37
	S66×8	528	31,536	462	36, 40, 41
	HB375×10	3750	192,690	3375	39
	AEGIS:BP	32	1953	10	60
	AEGIS:TAUT	37	1668	25	60
	Tautobase	700	39,216	350	61, 62
	AAMC	50	1527	25	43
	NAMC	68	3018	53	43
	PA26	34	1137	17	42
	RegioSQM20 (95%)	1088	84,576	544	63
testing	COMP5m	64,667	5,215,848		36, 64−66
	TAUT15	21	831	13	42
	RegioSQM20 (5%)	50	4023	25	63

[a]Data sets are described in the Methods section. In the current work, all reference DFT data is computed at the ωB97X/6-31G* level of theory for consistency.

An alternative approach is to develop machine learning (ML) potentials that are both fast and accurate within the scope of their training.[13−18] To date, many such models have been developed and more continue to emerge.[19−35] So-called "pure" ML potentials face many challenges for use in free energy simulations. They must be able to model a wide range of intra- and intermolecular interactions,[36−38] including relative conformational energies,[38] hydrogen bonding,[39] π stacking, London dispersion, and mixed interactions,[40,41] in addition to different tautomers[8] and protonation states[10,42,43] as mentioned previously. The models must be able to distinguish variable electron number (charge) and spin (multiplicity). Finally, the models not only need to be trained to give back accurate energies and forces for the regions of configurational space expected to be sampled under relevant temperature and pressure conditions but also must be trained to avoid inaccessible regions of configurational space.

Among the first and most widely recognized ML potentials are the ANI[30,36,44,45] class of models that to a large degree formed the inspiration for the current work. These pure ML models are both robust and computationally efficient. The ANI models take as basic arguments the positions and identity of atoms in order to return an energy and through derivative relations a set of forces. However, challenges remain for these models to distinguish different charge and/or spin states and properly treat electrostatic interactions (although there has been recent progress to determine atomic charges[33]). This is currently a serious limitation, as it has been estimated that up to 95% of drug molecules contain ionizable groups that can cause variations in the charge state and greatly alter electrostatic interactions.[46]

In the present work, we develop a Quantum Deep-learning Potential Interaction (QDπ) model that uses a fast third-order self-consistent density-functional tight-binding (DFTB3/3OB) model[47,48] that is corrected to a quantitatively high-level of accuracy through a range-corrected deep-learning potential (DPRc).[49,50] In this way, the QDπ model developed here is the form of a quantum mechanical/machine learning potential correction (QM/Δ-MLP).[35,49−54] The use of DFTB3 as a robust QM base model has several important advantages. First, it provides a reasonable description of the conformational potential energy landscape, greatly reducing the requirement to explicitly train the MLP to avoid inaccessible high-energy regions. Second, DFTB3 uses polarizable atomic charge densities that are easily integrated into an efficient particle-mesh Ewald[55] framework to capture long-range electrostatic interactions in condensed phase QM/MM[56] and QMFF[57−59] simulations. Third, DFTB3 is able to model changes in the charge, protonation, and spin states in a size-consistent manner. QDπ is developed and validated with respect to a number of existing and new databases (DBs).[36,37,39−43,60−66] Special emphasis is placed on developing a universal model that is able to quantitatively predict tautomers[8] and protonation states.[10] The present work develops QDπ for internal ligand energetics. This advance sets the stage for intermolecular interactions to be fully developed through the quantum mechanical/molecular mechanical (QM/MM) Δ-MLP. This would enable alchemical free energy simulations for drug discovery to be made using the QDπ model through the use of indirect MM → QDπ free energy "book-ending" methods.[67−70]

2. METHODS

This paper brings together several facets in order to develop the QDπ model for drug discovery. The first is the collection and curation of several existing molecular databases of structures, energies, and forces. The second is the generation of new data sets that fill needed gaps in training and/or testing data. Third, we develop new tools within DeePMD-kit[71,72] that enable more general flexible forms of the loss function (including relative energies) used in training of the neural networks. Fourth, we create computational infrastructure for consistent comparison of a wide array of existing potential energy models. Each of these is described in detail below.

2.1. Preparation for Data Sets. The purpose of this first-generation QDπ model is to create a highly robust universal potential that can accurately model drug-like molecules containing H, C, N, and O atoms as relevant for binding to biological targets. Important properties for consideration include the following: relative conformational energies, a wide range of intermolecular interactions, as well as relative energies associated with different tautomers and protonation states. While ultimately this model can be extended to predict

https://doi.org/10.1021/acs.jctc.2c01172
J. Chem. Theory Comput. 2023, 19, 1261−1275

covalent binding (irreversible inhibition), the initial focus here is on noncovalent binding.

We prepared several data sets for training and (benchmark) testing of the QDπ model. These are summarized in Table 1 and described in more detail below. As a general theme, we endeavored to be consistent with the ANI-1x data set that was generated using the ωB97X/6-31G* level of theory. Toward that end, where needed, we recalculated the energy and forces and performed geometry optimizations at a consistent ωB97X/6-31G* level of theory[73] (all using Gaussian 16[74]) in order to train evolving versions of the QDπ model. The QDπ model is trained to be a QM/Δ-MLP; i.e., a nonelectronic DPRc "correction" to the DFTB3/3OB[75] QM model potential energy similar to previous work.[49,50]

2.1.1. Broad Data Sets: ANI-1xm and COMP5m. These data sets contain a diverse range of bio and drug-like molecules at equilibrium and nonequilibrium conformations and contain structures, potential energies, and forces. Generally, previous chemical space data sets[76−78] are usually derived from the GDB databases[64,65,79] that contains billions of SMILES strings[80] for organic small molecules. Herein, we use modified versions of the public ANI-1x[37] and COMP6[36] databases as follows.

ANI-1xm. The ANI-1x data set is an open-source chemical space data set proposed by Smith et al.[37] that includes ωB97X/6-31G* energies and forces generated by diverse normal mode sampling (DNMS). We examined the ANI-1x data set and observed that the DNMS procedure would in some cases generate free radicals by breaking covalent bonds (which were still computed with a singlet spin state in the reference data set), and this led to problems in QDπ training (an example is provided in Section 1 of the Supporting Information). Thus, we curated a subset of the ANI-1x data to create a modified data set we refer to as ANI-1xm by analyzing and removing such predicted free radicals in addition to a few other select outlier molecules through the procedure described below.

As an example, the DFTB3/3OB base QM model is known to have rare anomalous outlier energies for some inorganic molecules such as cyanogen[75] that we did not consider as highly relevant for drug discovery. In other cases, it has been reported that some reference values in the ANI-1x data set are not reliable.[51] We thus used the following outlier detection criteria to remove 50 points that satisfy the condition

$$\frac{\|E_k - \overline{E}\|}{\sigma(E)} \geq 8 \tag{1}$$

where E_k is the energy difference between ωB97X/6-31G* and DFTB3, and $\overline{E}$ and $\sigma(E)$ are the mean and standard deviation of the energy differences for all molecules with the same chemical formula. The threshold is taken from the TorchANI program.[81] After curating the ANI-1xm data set in this way, we obtained a total of 2,641,429 points (a 46.7% reduction from the original ANI-1x data set).

COMP5m. The COmprehensive Machine-learning Potential (COMP6)[36] benchmark is a chemical space data set that was built from six separate data sets: 1) the original S66×8 benchmark,[40,41] 2) ANI-MD,[36] 3) GDB7to9,[64] 4) GDB10to13,[65] 4) Tripeptide,[36] and 5) DrugBank[66] data sets. We begin by extracting the S66×8 data set, which we will analyze separately, and we refer to the truncated data set as "COMP5". The COMP5 data set, like ANI-1x, used the DNMS procedure which cleaved covalent bonds in some instances; therefore, we applied the same outlier detection procedure described above to arrive at a modified COMP5m data set containing 64,667 data points (a 35.9% reduction from the original COMP5 data set).

2.1.2. Intermolecular Data Sets. Intermolecular data sets contain dimers at multiple separations. Each dimer was geometry optimized at the reference theoretical level and the intermonomeric distances without altering the internal geometries.[40] The ωB97X/6-31G* and DFTB3 energies and forces were evaluated at the reference geometries.

For the S66×8 data set and the HB375×10 data set, we compute the relative energies as follows

$$\Delta E = E - E_{\min} \tag{2}$$

where $E_{\min}$ is the minimum energy of the dimer (at the most favorable intermolecular separation). The QDπ model was trained using all available dimer separations. Because most of the interaction energies are quite small, we report only the ΔE of the most separated dimer configuration. The following intermolecular data sets are used in the current work.

S66×8. The S66×8 data set[41] is a nonbonded interaction data set containing 66 noncovalent pairs at 8 separations. The ωB97X/6-31G* energy and forces are directly taken from the COMP6 benchmark.[36] The 8 relative energies are computed for each of the 66 dimers.

HB375×10. The HB375×10 data set[39] is an S66×8-like nonbonded interaction data set containing 375 hydrogen bonding pairs. We use geometry provided by the data set to compute the ωB97X/6-31G* energy and forces.

AEGIS:BP. The AEGIS:BP data set is a subset of 10 hydrogen-bonded nucleic acid base pairs (BPs) within the artificially expanded genetic information system (AEGIS)[60,82] database. The entire list of BPs is given in Figure S2 of the Supporting Information. The structures were generated by Open Babel and optimized at the ωB97X/6-31G* level.

2.1.3. Tautomerization Data Sets. The tautomerization data sets described below are used to evaluate the relative energy of the tautomeric configurations. The relative energy between A and B is the difference between their total energies (E_A and E_B).

$$\Delta E = E_A - E_B \tag{3}$$

Each tautomer is optimized from the initial geometry at the ωB97X/6-31G* level. The relative energies and force corrections between ωB97X/6-31G* and DFTB3 at the reference geometries are used in the neural network training. When tabulating the results to compare different methods, we report the mean absolute errors (maEs) and root mean squared errors (rmsEs) of the relative energies; these values evaluate the energies upon geometry optimizing the structures with each method.

Tautobase. The Tautobase data set[61] is a broad tautomer data set. A subset of the Tautobase data set was constructed by Wieder et al.[62] that includes 354 tautomer pairs with C, H, O, and N elements. Each pair involves the relocation/bonding of a hydrogen atom. The initial geometry is generated by Open Babel[83] and optimized at the ωB97X/6-31G* level.

TAUT15. TAUT15 is a tautomer data set from the GMTK55 database[42] containing 15 relative energies. The initial geometry is provided by the data set and further optimized at the ωB97X/6-31G* level.

AEGIS:TAUT. The AEGIS:TAUT data set is a subset of 25 tautomeric (TAUT) equilibria extracted from the AEGIS[60]

https://doi.org/10.1021/acs.jctc.2c01172
J. Chem. Theory Comput. 2023, 19, 1261−1275

database. The entire list of TAUT is given in Figure S3 of the Supporting Information. The initial geometry is provided by the data set and further optimized at the ωB97X/6-31G* level.

2.1.4. Protonation Energy Data Sets. These data sets are intended to be reflective of titratable sites in biological and ligand/drug-like molecules. We depart from public data sets that provide SMILES strings of absolute deprotonation energies AH → A^- + H^+; however, we train and test the methods against relative protonation energies AH + B^- → A^- + BH. Each reactant or product is optimized. For a generic chemical reaction

$$\nu_A A + \nu_B B + \ldots \rightarrow \nu_X X + \nu_Y Y + \ldots \tag{4}$$

the relative energy for a reaction (donated ΔE_{rxn}) is defined as the potential energy difference between total reactants and total products

$$\Delta E_{rxn} = \sum_{\text{products,p}} \nu_p E - \sum_{\text{reactants,r}} \nu_r E \tag{5}$$

where ν_p and ν_r are the stoichiometric coefficients of each product p and reactant r.

Amino Acid Model Compounds (AAMCs). The AAMC data set contains 21 O−H and N−H bond-containing molecules (OHNH) with deprotonation energies, including amino acid model compounds from ref 43. The entire list of compounds is given in Section 5.1 of the Supporting Information. The initial geometries were generated by Open Babel[83] and optimized at the ωB97X/6-31G* level.

Nucleic Acid Model Compounds (NAMCs). The NAMC data set contains 53 deprotonation energies of nucleic acid (DNA and RNA bases) model compounds introduced in ref 43. The entire list of compounds is given in Section 5.2 of the Supporting Information. The initial geometry is generated by Open Babel[83] and optimized at the ωB97X/6-31G* level.

PA26. The PA26 data set is a subset of the GMTKN55 database[42] containing 26 adiabatic proton affinities. The initial geometry is provided by the data set and optimized at the ωB97X/6-31G* level.

RegioSQM20. We selected a subset of the RegioSQM20[63] database containing C, H, O, and N elements. The subset was randomly divided into the training and test sets. Some outliers were removed using the procedure described above for the ANI-1x and COMP5m data sets. Then, there are 544 and 25 deprotonation energies in the training and test sets, respectively. The entire list of compounds is given in Section 5.3 of the Supporting Information. The initial geometry is generated by Open Babel[83] (or RDKit[84] for some compounds to get better structures) and optimized at the ωB97X/6-31G* level.

2.2. QDπ (v1.0). In this work, we develop a general QDπ model as a Δ-MLP correction[54] to DFTB3/3OB. The correction is parametrized to reproduce target energies and forces for closed-shell bio and drug-like organic molecules and ions composed of C, H, O, and N elements. The QDπ energy is the sum of DFTB3 and neural network potential (NNP) model energies

$$E_{QD\pi} = E_{DFTB3} + E_{NNP} \tag{6}$$

where E_{DFTB3} is the DFTB3/3OB energy, and E_{NNP} is the Δ-MLP correction using the Deep Potential-Smooth Edition (DeepPot-SE) functional form.[28] DFTB3 was chosen as the base model because it is robust and internally consistent and has been reported[36] to have better overall accuracy for the ANI-1x data set compared to PM6.

2.2.1. DeepPot-SE. The QDπ model parametrizes a Deep Potential (DP) using the DeepPot-SE descriptor[28] used as a Δ-MLP correction. The functional form of the DP and DeepPot-SE descriptor has been previously described,[85] and additional details can be found in Section 2 of the Supporting Information. DeepPot-SE is a popular descriptor implemented in the DeePMD-kit package[71,72] which has seen use in over 100 works[86] since its proposal in 2018. It also serves as the foundation for the DPRc[49] Δ-MLP. The DPRc correction includes corrections for QM/MM interactions. Although the present work does not involve QM/MM interactions, the common framework between the DPRc and DP potentials affords the opportunity to extend the QDπ model to QM/MM applications using the DPRc potential.

A recent work[18] has compared the theories of different NNPs, including ANI-1 and DeepPot-SE. NNPs that use atomic-centered symmetry functions (ACSFs),[19] such as ANI-1, have fixed descriptors that must be determined before training. In contrast, the descriptors used in the SchNet[24] and DeepPot-SE NNPs are trained to improve accuracy. It was found that NNPs with trainable descriptors require more computational effort to train because the descriptor includes additional parameters (and thus the additional parameter gradients must be evaluated during training).[87] To address this issue, a model compression scheme has been introduced that can freeze and compress the DeepPot-SE descriptor to improve performance (either during or after training).[87] In this work, we apply this model compression scheme in the latter part of training (see below).

2.2.2. Relative Energy Loss. In this work, we introduce a new component to the loss function to train relative energies. It is common for ML training algorithms to update the neural network parameters using a subset (a "batch") of the available training data. A "loss function", L, is evaluated using the data contained within the batch, the neural network parameters are updated, and a new batch is created for the next optimization step by randomly selecting another subset of data.[88] In the past, a batch consisted of molecules whose total energies and forces are to be trained. In which case, the loss function consisted of two components: errors arising from the total energies L_E and forces L_f.

$$L = p_E L_E + p_f L_f \tag{7}$$

In the present work, we allow relative energies to be included within a batch

$$L = p_E L_E + p_{\Delta E} L_{\Delta E} + p_f L_f \tag{8}$$

where $L_{\Delta E}$ is the relative energy loss

$$L_{\Delta E} = \frac{1}{\mathcal{B}_{\Delta E}} \sum_{k=1}^{\mathcal{B}_{\Delta E}} \frac{1}{N_k} (\Delta E_k - \Delta E_k^*)^2 \tag{9}$$

where $\mathcal{B}_{\Delta E}$ is the number of relative energies within the batch. N_k is the total number of atoms for system k (the sum of all product and reactant atoms in the case of reaction energy). ΔE_k and ΔE_k^* are the model and reference relative energies, respectively. L_E and L_f are defined in the same way

$$L_E = \frac{1}{\mathcal{B}_E} \sum_{k=1}^{\mathcal{B}_E} \frac{1}{N_k} (E_k - E_k^*)^2 \tag{10}$$

https://doi.org/10.1021/acs.jctc.2c01172
J. Chem. Theory Comput. 2023, 19, 1261−1275

Table 2. Data Sets and Neural Network Optimization Steps Used in Training Different QDπ Model Versions

iter	data set	steps	descriptor	output model
1	ANI-1x (E,F)	10,000,000	normal	
2	ANI-1xm (E,F)	60,000,000	compressed	QDπ v0.0
3	ANI-1xm (E,F), Tautobase (ΔE, F), AAMC (ΔE, F)	31,000,000	compressed	QDπ v0.1
4	ANI-1xm (E,F), Tautobase (ΔE, F), AAMC (ΔE, F), NAMC (ΔE, F), S66×8 (ΔE, F)	50,000,000	compressed	QDπ v0.2
5	ANI-1xm (E,F), Tautobase (ΔE, F), AAMC (ΔE, F), NAMC (ΔE, F), S66×8 (ΔE, F), PA26 (ΔE, F), RegioSQM20 (95%) (ΔE, F), HB375×10 (ΔE, F)	47,000,000	compressed	QDπ v0.3
6	ANI-1xm (E,F), Tautobase (ΔE, F), AAMC (ΔE, F), NAMC (ΔE, F), S66×8 (ΔE, F), PA26 (ΔE, F), RegioSQM20 (95%) (ΔE, F), HB375×10 (ΔE, F), AEGIS:BP (ΔE, F), AEGIS:TAUT (ΔE, F)	43,800,000	compressed	QDπ v1.0

$$L_f = \frac{1}{\mathcal{B}_f}\sum_{k=1}^{\mathcal{B}_f}\frac{1}{3N_k}(f_k - f_k^*)^2 \tag{11}$$

where $\mathcal{B}_E$ and $\mathcal{B}_f$ are the number of relative energies and forces within the batch, respectively. E_k and E_k^* are the model and reference energy components, and f_k and f_k^* are the model and reference force components, respectively. p_E, $p_{\Delta E}$, and p_f are weights assigned to energy, relative energy, and force contributions to the loss function. The weights are linearly updated with the learning rate, α:

$$p_E(t) = p_E^0\frac{\alpha(t)}{\alpha(0)} + p_E^\infty\left(1 - \frac{\alpha(t)}{\alpha(0)}\right) \tag{12}$$

$$p_{\Delta E}(t) = p_{\Delta E}^0\frac{\alpha(t)}{\alpha(0)} + p_{\Delta E}^\infty\left(1 - \frac{\alpha(t)}{\alpha(0)}\right) \tag{13}$$

$$p_f(t) = p_f^0\frac{\alpha(t)}{\alpha(0)} + p_f^\infty\left(1 - \frac{\alpha(t)}{\alpha(0)}\right) \tag{14}$$

The learning rate decays exponentially with the training step, t

$$\alpha(t) = \alpha_0\lambda^{t/\tau} \tag{15}$$

where α_0 is the initial learning rate, λ is the decay rate, and τ is the decay steps. If energy, relative energy, or force is not available in a batch, p_E, $p_{\Delta E}$, or p_f will be set to zero to disable the corresponding loss contribution. In this work, we set $p_E^0 = 2$, $p_E^\infty = 20$, $p_{\Delta E}^0 = 2$, $p_{\Delta E}^\infty = 20$, $p_f^0 = 100$, $p_f^\infty = 0.1$, $\alpha_0 = 0.0001$, $\lambda = 0.99$, and $\tau = 400$. It is worth mentioning that direct training to the relative energies typically will not improve the accuracy of the absolute atomic energies (the energy of an atom in a vacuum). We have implemented the relative energy loss contributions into the DeePMD-kit package.[71,72]

2.2.3. Training Process. The QDπ model was trained using the DeePMD-kit software package. As shown in Table 2, we performed 6 training iterations with different data sets and training properties. In the first two iterations, the model was trained to reproduce the total energies and forces of the molecules contained within the relative protonation energy data set. After the first iteration, the DP Compress[87] algorithm was applied to freeze the model descriptor $\mathcal{D}$ to improve performance. All subsequent iterations restart the training from the previous iteration. Starting from iteration 3, the loss function was changed to train against relative energies rather than molecular total energies. All of the tables in the main text show results only for QDπ-v1.0, but the Supporting Information contains extended tables that have results for each version to compare.

2.3. Energy/Force Calculation and Geometry Optimizations. This section describes the various potentials compared, in addition to the basic methods used for performing geometry optimizations. Additional details for relaxed 2D potential energy surface scans are provided in Section 3 of the Supporting Information.

ωB97X/6-31G.* We used Gaussian 16[74] to evaluate ωB97X/6-31G* energies and perform geometry optimizations.[73]

Semiempirical methods. The AMBER 20[89] SQM module[90] was used to perform geometry optimizations and evaluate the energies and forces of several semiempirical models, including DFTB3[91,92] (3OB parameters[75]), MNDO/d,[93] AM1,[94] and PM6.[95] The DFTB+[96] package was used to validate DFTB3 results from AMBER.

The DFTB+[96] package was used to calculate GFN1-xTB[97] and GFN2-xTB[98] energies and forces. For these models, the ASE package[99] was used to optimize the geometries with the Limited-memory Broyden-Fletcher-Goldfarb-Shanno (LBFGS) algorithm.[100]

QDπ Models. The QDπ energy is the sum of the DFTB3 and the DP contributions. The DP contribution was directly evaluated within the DeePMD-kit program.[71] The ASE package[99] was used to optimize the QDπ structures.

ANI Models. The TorchANI program[81] was used to provide energies and forces of the ANI-1x[36] and ANI-2x[45] models. Each ANI model consists of 8 independent parameter sets. We only use the first model (index 0) for benchmarking. It has been suggested that using an average of multiple models will improve the accuracy[44] but at additional computational cost.[81] We performed the geometry optimizations with the ASE package.[99]

3. RESULTS AND DISCUSSION

For consistency, all the reference data used to train and test the QDπ model was performed at the ωB97X/6-31G* level, as in the ANI databases[30,36,44,45] used to train the ANI-1x and ANI-2x models. In making the comparison with the target reference data, we report various error metrics such as mean absolute and root-mean-square errors (maEs and rmsEs, respectively). Other models compared in this work have been trained to different reference data (levels of theory and data sets). Hence, it should not be concluded that deviation of these other models from the reference data used in this work implies they are necessarily less accurate in the theoretical electronic structure limit (which cannot be practically obtained for any of the data considered here). Thus, a comparison of results from other models is not meant to be critical but rather provide a broader context with respect to variation from a well-defined reference. In our view, the real litmus test for a drug discovery force field is the accuracy of binding free energy predictions

https://doi.org/10.1021/acs.jctc.2c01172
J. Chem. Theory Comput. 2023, 19, 1261−1275

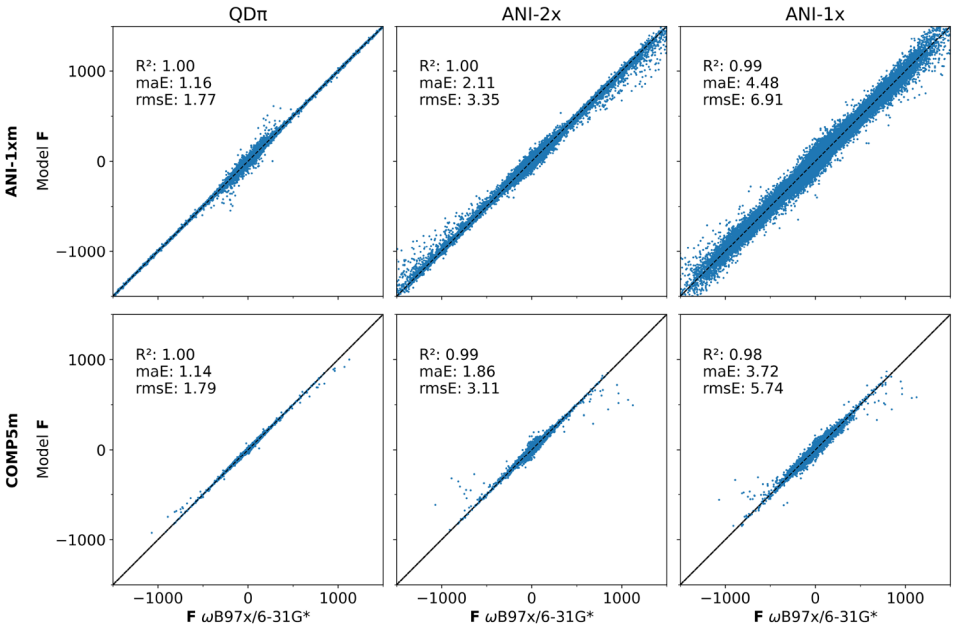


Figure 1. Relation between forces in kcal/(mol·Å) calculated by ωB97X/6-31G* and QDπ, ANI-2x, and ANI-1x, respectively, for the ANI-1xm and COMP5m data sets.

Table 3. Mean Absolute Error (maE) and Root Mean Square Error (rmsE) of Energies in kcal/mol and Forces in kcal/(mol·Å) for the ANI-1xm Training and COMP5m Testing (Marked with an "*") Data Sets (Table 1)[a]

	ANI-1xm				COMP5m*			
	energy		force		energy		force	
model	maE	rmsE	maE	rmsE	maE	rmsE	maE	rmsE
QDπ v1.0	0.83	1.22	1.16	1.77	1.48	2.44	1.14	1.79
ANI-1x	1.48	2.07	4.48	6.91	1.96	3.33	3.72	5.74
ANI-2x	1.07	1.58	2.11	3.35	1.67	2.66	1.86	3.11
DFTB3			7.58	12.45			5.46	8.76
MNDO/d[b]			15.14	24.53			11.52	17.77
AM1			14.95	24.29			12.13	18.07
PM6			12.96	23.63			9.33	14.30
GFN1-xTB			4.69	7.02			3.68	5.40
GFN2-xTB			5.81	8.65			4.33	6.33

[a]Models and data sets are described in the Methods section. [b]Some points (<0.1%) fail to converage and are removed.

from rigorous and precise AFE simulations,[5,101] which is beyond the scope of this work. The current work focuses on developing a robust and internally consistent internal energy model from drug-like molecules that provide a foundation from which to extend to accurate intermolecular interactions as a Δ-MLP to the QM/MM potential.

3.1. Performance for Internal (Intramolecular) Potential Energy. The majority of data used to train the QDπ model was from the ANI-1xm data set using energies and forces and validated against the COMP5m data set. The ANI-1x model was trained to the energies only, whereas the ANI-2x was trained to both energy and forces. Comparison of the force component errors for these models is illustrated in Figure 1, and error results for QDπ and all of the established comparison models are summarized in Table 3. Not surprisingly, the ANI-2x model performs best among the established models compared, having mean absolute errors (maEs) of 1.07 kcal/mol and 2.11 kcal/(mol·Å) for the energy and forces, respectively. Moreover, these errors are transferable to COMP5m, for which neither the ANI nor QDπ models were trained (maEs of 1.67 kcal/mol and 1.86 kcal/(mol·Å) for energy and force, respectively). The other models ranged in maE in forces of 4.69−15.14 kcal/(mol·Å) for ANI-1xm and 3.68−12.13 kcal/(mol·Å) for COMP5m (energies were not compared for the semiempirical models as the zero of total energy uses a different reference than for ωB97X/6-31G*). The worst models overall are the NDDO-based semiempirical models (MNDO/d, AM1, and PM6), which require additional fixes such as orthogonalization corrections[102] or other empirical terms[103] in order to accurately reproduce relative conformational energies. The tight-binding models that explicitly account for orthogonalization through the inclusion of an overlap matrix in the generalized eigenvalue problem perform generally better, with the GFN models slightly outperforming DFTB3/3OB. QDπ performs exceptionally well on both the ANI-1xm training and COMP5m testing data sets, having an maE of 0.83 and 1.48 kcal/mol for the energy, respectively, and 1.16 and 1.14 kcal/(mol·Å) for the forces, respectively.

https://doi.org/10.1021/acs.jctc.2c01172
J. Chem. Theory Comput. 2023, 19, 1261−1275

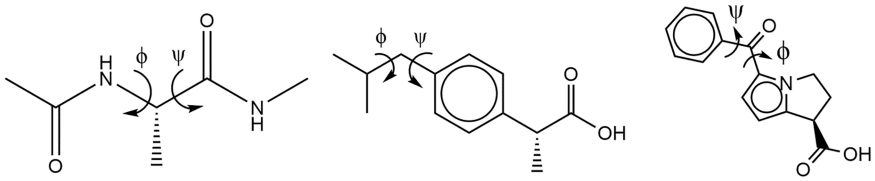


Figure 2. Model (a) alanine dipeptide; (b) ibuprofen; and (c) ketorolac. ϕ and ψ represent the 2D torsion angles.

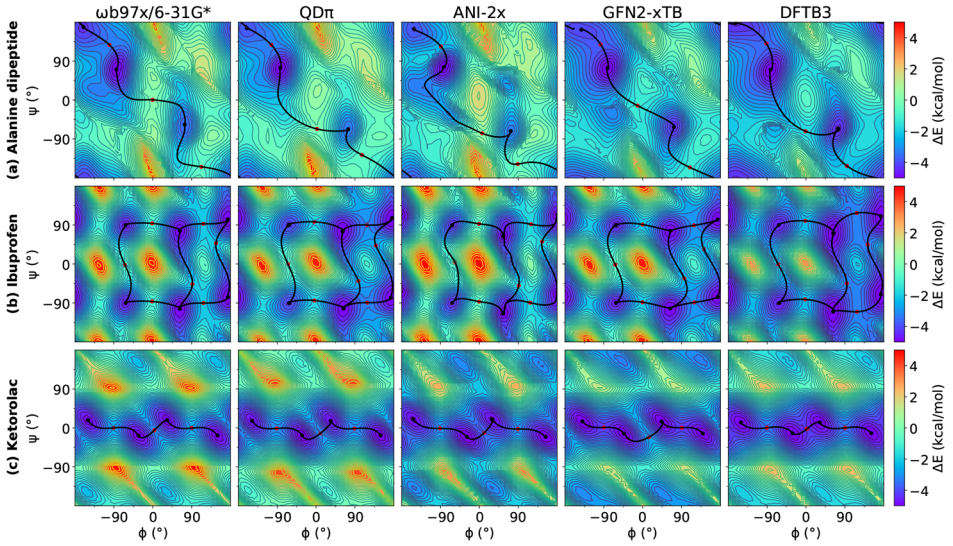


Figure 3. Relaxed 2D torsion profiles for (a) alanine dipeptide; (b) ibuprofen; and (c) ketorolac. Each molecule was computed using ωB97X/6-31G*, QDπ, DFTB3, ANI/2x, and GFN2-xTB, respectively. ωB97X/6-31G* is the reference potential, and the other potentials are compared with ωB97X/6-31G*. The color bars represent the potential energy (with respect to the minimum energy) ωB97X/6-31G* in kcal/mol. The black and red points represent the minima and the transition states, respectively, and the black curves represent the transition paths between minima.

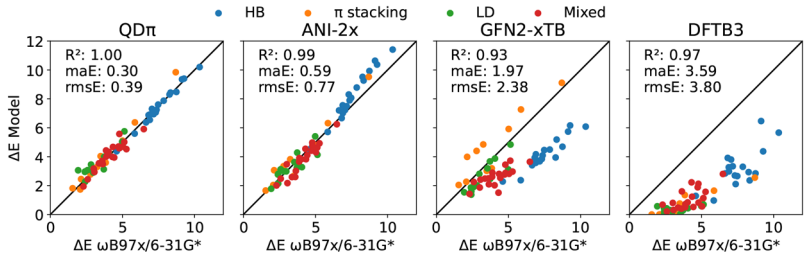


Figure 4. Relation between relative energies in kcal/mol calculated by ωB97X/6-31G* and QDπ, ANI-2x, GFN2-xTB, and DFTB3, respectively, for the S66×8 data set. Relative energies consist of the difference between the optimized structure and the structure with the furthest distance in each of the 66 dimer pairs.

In order to illustrate the degree to which the QDπ model can reproduce conformational energy landscapes, we examined relaxed 2D torsion profiles for three systems: the alanine dipeptide and the drug molecules ibuprofen and ketorolac illustrated in Figure 2. Figure 3 compares the potential energy surface for 2D torsion scans at the ωB97X/6-31G*, QDπ, ANI-2x, GFN2-xTB, and DFTB3 levels. All of the models qualitatively predict the correct trends. Overall, QDπ and ANI-2x are quite similar and have the closest agreement with ωB97X/6-31G*. While none of the models is able to reproduce conformational energy barriers below 1 kcal/mol in all cases (see Table S4 of the Supporting Information), the DFTB3 model errors are the largest and most systematic in their underestimation of the barriers between minima.

3.2. Performance for Intermolecular Interactions.

Despite the focus being on training a QDπ internal energy model, we felt it important to include training and testing data to intermolecular interaction DBs (S66×8[36,40,41] and HB375×10[39]) as some large, drug-like molecules can form similar interactions (e.g., intramolecular hydrogen bonds). Figure 4 and Table 4 compare intermolecular interactions (ΔE values) for S66×8. Overall, QDπ has the smallest maE (0.30 kcal/mol) relative to the other models that ranged in maE from 0.59 kcal/mol (ANI-2x) to 3.59 kcal/mol for DFTB3. A more detailed breakdown of the errors into hydrogen bonding (HB), π stacking, London dispersion (LD), and mixed interactions is provided in Table 4. QDπ has maE values that range from 0.21 to 0.41 kcal/mol (the largest is the LD subset), whereas ANI-2x ranges from 0.40 to 0.91 kcal/mol

https://doi.org/10.1021/acs.jctc.2c01172
J. Chem. Theory Comput. 2023, 19, 1261−1275

Table 4. Mean Absolute Error (maE) and Root Mean Square Error (rmsE) of Relative Energies (ΔE) in kcal/mol for Hydrogen Bonding (HB), π Stacking, London Dispersion (LD), and Mixed Influence (Mixed) Subsets of the S66×8 and HB375×10 Training Data Sets (Table 1)[a]

	S66×8 ΔE (subsets)								HB375×10	
	HB		π stacking		LD		mixed			
	ΔE		ΔE		ΔE		ΔE		ΔE	
model	maE	rmsE	maE	rmsE	maE	rmsE	maE	rmsE	maE	rmsE
QDπ v1.0	0.21	0.26	0.35	0.47	0.41	0.51	0.31	0.38	0.90	1.18
ANI-1x	2.01	3.12	1.34	1.49	0.78	0.98	1.42	1.70	1.54	1.97
ANI-2x	0.91	1.11	0.49	0.57	0.40	0.48	0.40	0.49	1.62	3.49
DFTB3	4.64	4.75	3.14	3.39	2.93	3.07	3.06	3.15	4.12	4.34
MNDO/d	5.37	5.73	2.68	2.83	2.49	2.58	2.66	2.75	6.44	7.04
AM1	7.07	7.63	3.76	4.21	3.04	3.26	3.74	3.83	5.28	5.74
PM6	9.86	10.87	3.83	4.33	3.28	3.42	4.12	4.24	3.92	4.23
GFN1-xTB	3.31	3.45	0.88	1.09	0.90	1.03	1.95	2.03	2.52	2.68
GFN2-xTB	3.53	3.59	0.96	1.12	0.65	0.82	1.55	1.69	2.66	2.89

[a]Models and data sets are described in the Methods section. Relative energies consist of the difference between the optimized structure and the structure with the furthest distance in each of the dimer pairs.

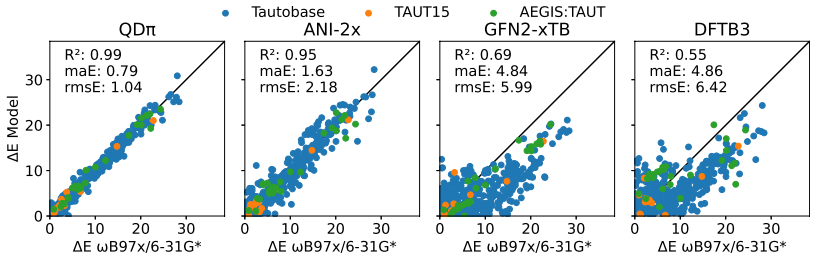


Figure 5. Relation between tautomerization energies in kcal/mol calculated by ωB97X/6-31G* and QDπ, ANI-2x, GFN2-xTB, and DFTB3, respectively, for the TAUT15 data set and the artificially expanded genetic information system: Tautomer (AEGIS:TAUT) data set.

(the largest being for the HB subset). Closer examination of the more extensive HB735 data set indicates that QDπ has the lowest maE (0.44 kcal/mol), whereas ANI-2x, GFN2-xTB, and DFTB3 have maE values of 1.40, 0.85, and 1.17 kcal/mol, respectively. In the case of ANI-2x, the distribution of errors has a considerably larger variance as indicated by the rmsE value of 3.84 which is over 2.5 times larger than the maE.

3.3. Performance for Tautomers. Figure 5 compares the tautomer energies for the Tautobase[61] data subset and TAUT15[42] and AEGIS:TAUT data sets. The QDπ model is the only model that has quantitative (less than 1 kcal/mol) accuracy for tautomer energies (maE values that range from 0.70 to 0.82 kcal/mol for data subsets listed in Table 5 and 0.79 kcal/mol overall). Of the other models, ANI-2x performs the best but still has an maE roughly twice as large (maE 1.63 kcal/mol, with data subset values that range from 1.00 to 1.76 kcal/mol). The GFN2-xTB and DFTB3 models have maE values exceeding 4.5 kcal/mol overall with the largest contributions coming from the Tautobase data subset (maE values for GFN2-xTB and DFTB3 both approximately 5.5 kcal/mol). Both QDπ and ANI-2x have a high linear correlation with ωB97X/6-31G* (0.99 and 0.95, respectively), whereas GFN-xTB and DFTB3 have moderate correlation (0.69 and 0.55). Hence, the QDπ model is the only model that consistently provides tautomer energies with errors below 1 kcal/mol relative to the reference.

3.4. Performance for Relative Protonation States. Figure 6 compares the differences in protonation energies for a series of amino acid and nucleic acid model compounds,[43] as well as molecules in the PA26[42] data set. Changes in the charge

Table 5. Mean Absolute Error (maE) and Root Mean Square Error (rmsE) of Energies and Relative Energies (ΔE) in kcal/mol for the Tautobase/AEGIS:TAUT Training and TAUT15 Testing (Marked with an "*") Data Sets (Table 1)[a]

	Tautobase		TAUT15*		AEGIS:TAUT	
	ΔE		ΔE		ΔE	
model	maE	rmsE	maE	rmsE	maE	rmsE
QDπ v1.0	0.82	1.09	0.70	0.89	0.71	0.97
ANI-1x	1.73	2.42	1.63	1.83	1.54	2.08
ANI-2x	1.76	2.39	1.00	1.20	1.41	1.94
DFTB3	5.45	6.93	3.65	4.60	5.25	6.12
MNDO/d	9.69	11.35	7.78	9.55	8.20	9.43
AM1	5.01	6.34	3.99	5.85	3.88	4.66
PM6	4.90	6.12	5.60	7.11	7.39	8.85
GFN1-xTB	5.23	6.51	5.32	6.53	5.61	6.58
GFN2-xTB	5.68	6.81	2.84	3.62	3.16	3.59

[a]Models and data sets are described in the Methods section.

state are notoriously challenging for minimal valence basis models[104] and particularly problematic for the ANI-2x model that cannot distinguish molecules from ions and thus breaks down. Both the GNF2-xTB and DFTB3 models have a high correlation (0.99) with ωB97X/6-31G*, owing mainly to the large range of values that are clustered into two sets (0−75 and 125−225 kcal/mol), and large maE values that are 7.78 and 10.24 kcal/mol, respectively. The QDπ model performs exceptionally well with an maE of 0.17 kcal/mol and almost perfect correlation.

https://doi.org/10.1021/acs.jctc.2c01172
J. Chem. Theory Comput. 2023, 19, 1261−1275

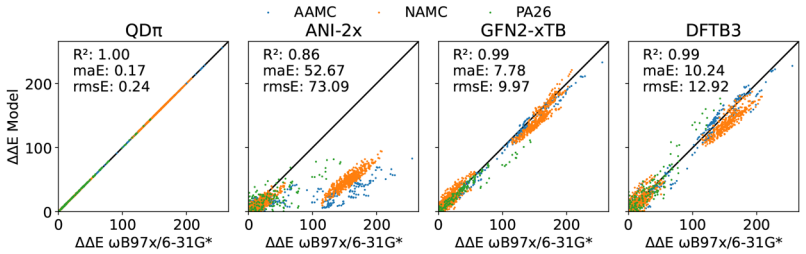


Figure 6. Relation between relative protonation energies (AH + B^- → A^- + BH) in kcal/mol calculated by ωB97X/6-31G* and QDπ, ANI-2x, GFN2-xTB, and DFTB3, respectively, for AAMC, NAMC, and PA26 data sets.

Table 6. Mean Absolute Error (maE) and Root Mean Square Error (rmsE) of Energies and Relative Energies (Δ*E*) in kcal/mol and Forces in kcal/(mol·Å) for Relative Protonation Energies for the AAMC, NAMC, and PA26 Training and RegioSQM* Testing (Validation) Data Sets (Table 1)[a]

	AAMC		NAMC		PA26		RegioSQM*	
	Δ*E*		Δ*E*		Δ*E*		Δ*E*	
model	maE	rmsE	maE	rmsE	maE	rmsE	maE	rmsE
QDπ v1.0	0.09	0.14	0.17	0.22	0.39	0.49	2.53	3.19
ANI-1x	86.95	112.31	52.68	71.74	43.02	62.28	16.85	22.15
ANI-2x	70.52	89.39	52.48	72.33	23.80	30.54	13.64	17.24
DFTB3	8.63	11.12	10.85	13.33	12.54	15.84	4.59	5.74
MNDO/d	11.71	14.13	11.29	14.08	13.07	16.10	5.18	6.29
AM1	4.43	5.49	7.32	9.10	13.51	20.89	4.13	5.11
PM6	11.23	13.86	11.03	13.58	17.84	34.36	5.30	6.57
GFN1-xTB	5.00	6.07	11.73	35.39	4.43	5.39	4.10	5.07
GFN2-xTB	5.77	7.14	8.45	10.40	7.35	11.73	4.12	4.96

[a]Models and data sets are described in the Methods section.

A closer examination of the data subsets is listed in Table 6. The maE values between AAMC, NAMC, and PA26 data subsets have ranges 0.09−0.39 (QDπ), 5.77−8.45 (GFN2-xTB), 8.63−12.54 (DFTB3), and 23.8−70.5 kcal/mol (ANI-2x). Examination of errors from the RegioSQM* data set that was not used in training reveals more similar maE values: 2.53 (QDπ), 4.12 (GFN2-xTB), 4.59 (DFTB3), and 13.6 kcal/mol (ANI-2x). The reason for this may be due in part to the design of the database to predict the regioselectivities of electrophilic aromatic substitution reactions from the calculation of proton affinities, but this is not fully clear.

3.5. Example Application: Acid/Base Reactions Important in Enzyme Catalyzed RNA Cleavage. Although the current QDπ model has been designed with the intent ultimately for applications to drug discovery, it is nonetheless instructive to consider well-studied examples where protonation/deprotonation events are of biological significance. One such example presents itself in the chemistry of RNA strand cleavage[105−107] that is catalyzed by protein ribonucleases[108−110] as well as small self-cleaving ribozymes[111,112] and several artificially engineered DNAzymes.[113−116] In this reaction, the 2′-OH group of an RNA nucleotide is activated by deprotonation by a general base. The resulting activated nucleophile makes an in-line attack to the adjacent phosphorus of the scissile phosphate, and the reaction proceeds through a pentacovalent transition state followed by departure of the 5′-O leaving group that is facilitated by donation of a proton from a general acid. In the case of protein ribonucleases, the general base−acid catalysis is thought to be carried out by active site histidine residues,[108−110] although early work speculated that a functionally important lysine residue might also be a plausible candidate.[117] In the case of small self-cleaving ribozymes and DNAzymes, the general acid−base catalysis is carried out by nucleobases and metal ions or in some cases assisted by the 2′-OH of the ribose sugar moiety. Considering the nucleobase candidates, the general base is often an active site guanine (at the N1 position), whereas the general acid can be either a cytosine (at the N3 position) or else an adenine (either N1 or N3 positions).

We hence consider the energetics of reactions that involve the relative protonation/deprotonation of the general acid and base models with respect to the 2′-OH nucleophile (a secondary alcohol modeled as isopropyl alcohol, iPrOH) and the 5′OH leaving group (a primary alcohol modeled as ethanol, EtOH). These are listed in Table 7. The model reactions where iPrOH is deprotonated to form $iPrO^-$ is meant to represent a model system for general base activation of the 2′-OH nucleophile, whereas the reactions where EtO^- is protonated to form EtOH is meant to represent a model system for general acid stabilization of the 5′-O^- leaving group. The values shown in Table 7 show the relative energetics of the noninteracting molecular and ionic reaction species. In the gas phase, the formation of neutral molecules from noninteracting ions from neutral molecules is highly exothermic; although for interacting systems in an enzyme environment, the differences are expected to be much smaller. Nonetheless, the inherent energetics associated with the relative protonation/deprotonation events still is a major factor that regulates reactivity. For the protein enzyme model reactions (top block, Table 7), QDπ performs exceptionally well with errors all less than 0.5 kcal/mol, whereas the DFTB3 and GFN2-xTB errors range from −11.33 to 11.72 and −7.02 to 9.66 kcal/mol, respectively. For the nucleobase reaction models (middle block, Table 7), QDπ has larger errors that

https://doi.org/10.1021/acs.jctc.2c01172
J. Chem. Theory Comput. 2023, 19, 1261−1275

Table 7. Selected Relative Protonation/Deprotonation Energies from ωB97X/6-31G* and the Model Error (kcal/mol) Relevant to Acid/Base Catalysis in RNA Cleavage Reactions[a]

	ωB97X/6-31G*	QDπ	DFTB3	ANI-2x	GFN2-xTB
protonation pair	ΔE	Err	Err	Err	Err
Lys:NH_2 + iPrOH → Lys:NH_3^+ + $iPrO^-$	167.76	0.00	6.11	−115.04	0.04
His:N_ϵ + iPrOH → His:$N_\epsilon H^+$ + $iPrO^-$	158.33	0.02	−11.33	−126.62	−7.02
His:$N_\epsilon H^+$ + EtO^- → His:N_ϵ + EtOH	−160.25	0.04	11.71	137.70	9.66
G:N_1^- + iPrOH → G:N_1H + $iPrO^-$	43.06	−1.11	−8.63	−28.62	−2.69
A:N_1H^+ + EtO^- → A:N_1 + EtOH	−165.06	1.25	15.21	137.24	10.02
A:N_3H^+ + EtO^- → A:N_3 + EtOH	−190.89	1.21	16.00	143.42	11.40
C:N_3H^+ + EtO^- → C:N_3 + EtOH	−160.33	0.89	4.66	145.20	6.58
A:N_1H^+ + G:N_1^- → A:N_1 + G:N_1H	−120.07	0.08	6.20	97.55	4.69
A:N_3H^+ + G:N_1^- → A:N_3 + G:N_1H	−145.91	0.04	6.99	103.73	6.07
C:N_3H^+ + G:N_1^- → C:N_3 + G:N_1H	−115.34	−0.27	−4.35	105.50	1.25
maE		0.49	9.12	114.06	5.94
rmsE		0.71	9.96	118.72	6.96

[a]Models and data sets are described in the Methods section. Shown are model reactions for protein enzymes (top block) and nucleic acid enzymes (middle block). Additionally, the relative acid and base protonation/deprotonation energies for different nucleobases are provided (bottom block).

range from −1.11 to 1.25 kcal/mol, whereas the corresponding ranges are −8.63 to 16.00 and −2.69 to 11.40 for DFTB3 and GFN2-xTB, respectively. For the relative errors between nucleobase general acid and base (bottom block, Table 7), again QDπ is in very close agreement with the reference values (maximum error −0.27 kcal/mol), where the DFTB3 and GFN2-xTB models have considerably larger errors (maximum errors 6.99 and 6.07 kcal/mol, respectively). Overall, the average error for QDπ is below 0.5 kcal/mol. As mentioned earlier, the ANI-2x model breaks down for a system that has a varying charge, with average errors of over 100 kcal/mol. These results provide an example in a biological context that emphasizes the importance of modeling relative protonation/deprotonation events with quantitative accuracy. For drug discovery, these will be especially important, as it has been estimated that over 95% of drug molecules have ionizable sites, many of which may potentially change upon binding to a biological target.

3.6. Current Limitations and Future Directions. A key aspect of this project was to create a first-generation potential energy model trained against broad data all computed at the same level of theory (and where possible, even using the same software package). At the time this project was initiated, the largest such data set was ANI-1x DB[37] at the ωB97X/6-31G* level that only contained compounds with elements H, C, N, and O. This chemical space is incomplete, as many drug molecules contain phosphorus, sulfur, and halogen atoms, and some contain metal ions.[118−120] The ANI-2x model was extended to include S, F, and Cl,[45] but the full data set, including the important reference energies and forces at the ωB97X/6-31G* level, to our knowledge, has not yet become publicly available. Currently, there are a number of recent data sets that include compounds that contain phosphorus, sulfur, and halogens at various levels of theory[121−125] as well as metal ions.[118] Among them, only the SPICE data set[124] includes forces at the ωB97M-D3BJ/def2-TZVPPD level and currently includes over 420K phosphorus, 520K sulfur, 750K halogen, and 8K metal-containing structures.

Hence, current limitations of the QDπ-v1.0 model include restricted chemical space (molecules containing H, C, N, and O) and the ωB97X/6-31G* reference level of theory. The ωB97X/6-31G* reference level, like the DFTB3/3OB base QM model, lacks dispersion corrections and also does not include counterpoise corrections and complete basis set extrapolations that are important for intermolecular interactions. Further, this reference level of theory is not ideal for all molecular properties, including ionization energies and in some cases proton affinities of anions that may be sensitive to inclusion of diffuse basis functions. So while it is important to start with an established and consistent reference level of theory and chemical scope, ultimately as higher-level data sets become more complete and made publicly available, QDπ and other machine learning potentials can continue to evolve.

The next step of future work will involve developing an intermolecular QM/MM interaction potential as a new range-corrected deep-learning potential.[49,50] The full (internal and intermolecular interaction) QDπ model is designed to be a correction to the QM/MM potential energy using DFTB3/3OB and the latest AMBER FF19SB for proteins,[126] OL3/OL15 for nucleic acids,[127−129] OPC model for water,[130,131] and 12−6−4 ion models.[132−134] Once the intermolecular interaction component of the QDπ model has been developed and validated in alchemical free energy simulations,[5] next steps will be to extend the chemical space of drug molecules to include P, S, F, and Cl atoms. With GPU acceleration, QDπ is typically less than 10% overhead relative to a traditional QM/MM energy/force evaluation with DFTB3/3OB.[49] With this design, QDπ could in principle also be used to modify the internal energy of protein residues and/or solvent molecules in contact with the drug, but this would incur greater cost as the size of the QM region grows larger. Should treatment of these surrounding residues with a neural network correction potential be deemed necessary, an alternative strategy would be to extend the model such that it can directly correct the MM potential rather than the QM/MM potential.

4. CONCLUSION

We report QDπ-v1.0 for modeling the internal energy of drug molecules. The development of this model required the following: 1) collection and curation of several existing molecular databases of structures, energies, and forces; 2) generation of new data sets at the ωB97X/6-31G* level that fill needed gaps in training and/or testing data; 3) development of new tools within DeePMD-kit that enable more general flexible forms of the loss function used in training of the neural networks; and 4) creation of computational infrastructure for consistent comparison of a wide array of existing potential energy models. The QDπ model has the advantage that it is able to properly treat electrostatic interactions and handle changes in charge/protonation states. The QDπ model is demonstrated to be accurate for a wide range of intra- and

https://doi.org/10.1021/acs.jctc.2c01172
J. Chem. Theory Comput. 2023, 19, 1261−1275

intermolecular interactions (despite its intended use as an internal energy model) and has shown to perform exceptionally well for relative protonation/deprotonation energies and tautomers. Comparison with several other approximate semiempirical and machine learning potentials (ANI-1x, ANI-2x, DFTB3, MNDO/d, AM1, PM6, GFN1-xTB, and GFN2-xTB) indicates QDπ agrees much more closely with training and testing data at the reference ωB97X/6-31G* level. An example application to model reactions involved in RNA strand cleavage catalyzed by protein and nucleic acid enzymes further illustrates the QDπ accuracy in a biological context. This makes QDπ highly attractive as a potential force field model for drug discovery.

■ ASSOCIATED CONTENT

Data Availability Statement

QDπ-v1.0 is openly available in our GitLabrepository at https://gitlab.com/RutgersLBSR/qdpi.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jctc.2c01172.

Removal of radicals, Deep Potential-Smooth Edition (DeepPot-SE) model, details for 2D potential energy surface scans (QDπ-Scan), structures of AAMC, NAMC, RegioSQM20, and AEGIS, Table S1: data sets, Table S2: data sets and neural network optimization steps, Table S3: maE and rmsE of energies and forces, Table S4: relative energies for minima and transition state, Table S5: selected stationary points for minima and transition state, Table S6: maE and rmsE of relative energies of S66x8 and HB375x10 training data sets, Table S7: maE and rmsE of energies and relative energies for Tautobase/AEGIS:TAUT training and TAUT15 testing data sets, Table S8: maE and rmsE of energies and relative energies and forces for AAMC, NAMC, and PA26 training and RegioSQM testing data sets, Figure S1: relationship between energies at different levels and length of one of C−H bonds on methane system, Figure S2: structures for AEGIS:BP data set, and Figure S3: structures for AEGIS:TAUT data set (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Darrin M. York − *Laboratory for Biomolecular Simulation Research, Institute for Quantitative Biomedicine and Department of Chemistry and Chemical Biology, Rutgers University, Piscataway, New Jersey 08854, United States;* orcid.org/0000-0002-9193-7055; Email: Darrin.York@rutgers.edu

Authors

Jinzhe Zeng − *Laboratory for Biomolecular Simulation Research, Institute for Quantitative Biomedicine and Department of Chemistry and Chemical Biology, Rutgers University, Piscataway, New Jersey 08854, United States;* orcid.org/0000-0002-1515-8172

Yujun Tao − *Laboratory for Biomolecular Simulation Research, Institute for Quantitative Biomedicine and Department of Chemistry and Chemical Biology, Rutgers University, Piscataway, New Jersey 08854, United States;* orcid.org/0000-0002-4520-941X

Timothy J. Giese − *Laboratory for Biomolecular Simulation Research, Institute for Quantitative Biomedicine and Department of Chemistry and Chemical Biology, Rutgers University, Piscataway, New Jersey 08854, United States;* orcid.org/0000-0002-0653-9168

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jctc.2c01172

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors thank Dr. Han Wang for the code review of DeePMD-kit. The authors are grateful for the financial support provided by the National Institutes of Health (No. GM107485 to D.M.Y.). Computational resources were provided by the Office of Advanced Research Computing (OARC) at Rutgers, The State University of New Jersey; the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by the National Science Foundation grant ACI-1548562 (supercomputer Expanse at SDSC through allocation CHE190067); and by the Texas Advanced Computing Center (TACC) at the University of Texas at Austin (supercomputer Longhorn through allocation CHE20002).

■ REFERENCES

(1) Jorgensen, W. L. Efficient drug lead discovery and optimization. *Acc. Chem. Res.* **2009**, *42*, 724−733.
(2) Abel, R.; Wang, L.; Harder, E. D.; Berne, B. J.; Friesner, R. A. Advancing Drug Discovery through Enhanced Free Energy Calculations. *Acc. Chem. Res.* **2017**, *50*, 1625−1632.
(3) Cournia, Z.; Allen, B.; Sherman, W. Relative Binding Free Energy Calculations in Drug Discovery: Recent Advances and Practical Considerations. *J. Chem. Inf. Model.* **2017**, *57*, 2911−2937.
(4) Cournia, Z.; Allen, B. K.; Beuming, T.; Pearlman, D. A.; Radak, B. K.; Sherman, W. Rigorous Free Energy Simulations in Virtual Screening. *J. Chem. Inf. Model.* **2020**, *60*, 4153−4169.
(5) Lee, T.-S.; Allen, B. K.; Giese, T. J.; Guo, Z.; Li, P.; Lin, C.; McGee, T. D., Jr.; Pearlman, D. A.; Radak, B. K.; Tao, Y.; Tsai, H.-C.; Xu, H.; Sherman, W.; York, D. M. Alchemical Binding Free Energy Calculations in AMBER20: Advances and Best Practices for Drug Discovery. *J. Chem. Inf. Model.* **2020**, *60*, 5595−5623.
(6) Cole, D. J.; Horton, J. T.; Nelson, L.; Kurdekar, V. The future of force fields in computer-aided drug design. *Future Med. Chem.* **2019**, *11*, 2359−2363.
(7) Vanommeslaeghe, K.; MacKerell, A., Jr. CHARMM additive and polarizable force fields for biophysics and computer-aided drug design. *Biochim. Biophys. Acta* **2015**, *1850*, 861−871.
(8) Martin, Y. C. Let's not forget tautomers. *J. Comput.-Aided Mol. Des.* **2009**, *23*, 693−704.
(9) Milletti, F.; Storchi, L.; Sforna, G.; Cross, S.; Cruciani, G. Tautomer Enumeration and Stability Prediction for Virtual Screening on Large Chemical Databases. *J. Chem. Inf. Model.* **2009**, *49*, 68−75.
(10) Manallack, D. T. The pK(a) Distribution of Drugs: Application to Drug Discovery. *Perspect Medicin Chem.* **2007**, *1*, 25−38.
(11) Rathore, R. S.; Sumakanth, M.; Siva Reddy, M.; Reddanna, P.; Rao, A. A.; Erion, M. D.; Reddy, M. R. Advances in Binding Free Energies Calculations: QM/MM-Based Free Energy Perturbation Method for Drug Design. *Curr. Pharm. Des.* **2013**, *19*, 4674−4686.
(12) Kříž, K.; Rezáč, J. Benchmarking of Semiempirical Quantum-Mechanical Methods on Systems Relevant to Computer-Aided Drug Design. *J. Chem. Inf. Model.* **2020**, *60*, 1453−1460.
(13) Behler, J. Perspective: Machine learning potentials for atomistic simulations. *J. Chem. Phys.* **2016**, *145*, 170901.

(14) Butler, K. T.; Davies, D. W.; Cartwright, H.; Isayev, O.; Walsh, A. Machine learning for molecular and materials science. *Nature* **2018**, *559*, 547−555.
(15) Noé, F.; Tkatchenko, A.; Müller, K.-R.; Clementi, C. Machine Learning for Molecular Simulation. *Annu. Rev. Phys. Chem.* **2020**, *71*, 361−390.
(16) Pinheiro, M., Jr; Ge, F.; Ferré, N.; Dral, P. O.; Barbatti, M. Choosing the right molecular machine learning potential. *Chem. Sci.* **2021**, *12*, 14396−14413.
(17) Manzhos, S.; Carrington, T., Jr Neural Network Potential Energy Surfaces for Small Molecules and Reactions. *Chem. Rev.* **2021**, *121*, 10187−10217.
(18) Zeng, J.; Cao, L.; Zhu, T.Neural network potentials. In *Quantum Chemistry in the Age of Machine Learning*; Dral, P. O., Ed.; Elsevier: 2022; Chapter 12, pp 279−294, DOI: 10.1016/B978-0-323-90049-2.00001-9.
(19) Behler, J.; Parrinello, M. Generalized Neural-Network Representation of High-Dimensional Potential-Energy Surfaces. *Phys. Rev. Lett.* **2007**, *98*, 146401−146404.
(20) 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.
(21) Behler, J. Atom-centered Symmetry Functions for Constructing High-dimensional Neural Network Potentials. *J. Chem. Phys.* **2011**, *134*, 074106.
(22) Gastegger, M.; Schwiedrzik, L.; Bittermann, M.; Berzsenyi, F.; Marquetand, P. wACSF—Weighted atom-centered symmetry functions as descriptors in machine learning potentials. *J. Chem. Phys.* **2018**, *148*, 241709.
(23) 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*, 1603015.
(24) Schütt, K. T.; Arbabzadah, F.; Chmiela, S.; Müller, K. R.; Tkatchenko, A. Quantum-chemical insights from deep tensor neural networks. *Nat. Commun.* **2017**, *8*, 13890.
(25) Schütt, K.; Sauceda, H.; Kindermans, P.; Tkatchenko, A.; Müller, K. SchNet - A Deep Learning Architecture for Molecules and Materials. *J. Chem. Phys.* **2018**, *148*, 241722.
(26) Chen, X.; Jørgensen, M. S.; Li, J.; Hammer, B. Atomic Energies from a Convolutional Neural Network. *J. Chem. Theory Comput.* **2018**, *14*, 3933−3942.
(27) Zhang, L.; Han, J.; Wang, H.; Car, R.; E, W. Deep potential molecular dynamics: a scalable model with the accuracy of quantum mechanics. *Phys. Rev. Lett.* **2018**, *120*, 143001.
(28) Zhang, L.; Han, J.; Wang, H.; Saidi, W.; Car, R.; E, W.End-to-end Symmetry Preserving Inter-atomic Potential Energy Model for Finite and Extended Systems. In *Advances in Neural Information Processing Systems 31*; Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R., Eds.; Curran Associates, Inc.: 2018; pp 4436−4446.
(29) Zhang, Y.; Hu, C.; Jiang, B. Embedded Atom Neural Network Potentials: Efficient and Accurate Machine Learning with a Physically Inspired Representation. *J. Phys. Chem. Lett.* **2019**, *10*, 4962−4967.
(30) 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.
(31) Unke, O.; Meuwly, M. PhysNet: A Neural Network for Predicting Energies, Forces, Dipole Moments, and Partial Charges. *J. Chem. Theory Comput.* **2019**, *15*, 3678−3693.
(32) Glick, Z. L.; Metcalf, D. P.; Koutsoukas, A.; Spronk, S. A.; Cheney, D. L.; Sherrill, C. D. AP-Net: An atomic-pairwise neural network for smooth and transferable interaction potentials. *J. Chem. Phys.* **2020**, *153*, 044112.
(33) Zubatiuk, T.; Isayev, O. Development of Multimodal Machine Learning Potentials: Toward a Physics-Aware Artificial Intelligence. *Acc. Chem. Res.* **2021**, *54*, 1575−1585.
(34) Khajehpasha, E. R.; Finkler, J. A.; Kühne, T. D.; Ghasemi, S. A. CENT2: Improved charge equilibration via neural network technique. *Phys. Rev. B* **2022**, *105*, 144106.
(35) Pan, X.; Yang, J.; Van, R.; Epifanovsky, E.; Ho, J.; Huang, J.; Pu, J.; Mei, Y.; Nam, K.; Shao, Y. Machine-Learning-Assisted Free Energy Simulation of Solution-Phase and Enzyme Reactions. *J. Chem. Theory Comput.* **2021**, *17*, 5745−5758.
(36) Smith, J. S.; Nebgen, B.; Lubbers, N.; Isayev, O.; Roitberg, A. E. Less is more: Sampling chemical space with active learning. *J. Chem. Phys.* **2018**, *148*, 241733−241743.
(37) Smith, J. S.; Zubatyuk, R.; Nebgen, B.; Lubbers, N.; Barros, K.; Roitberg, A. E.; Isayev, O.; Tretiak, S. The ANI-1ccx and ANI-1x data sets, coupled-cluster and density functional theory properties for molecules. *Sci. Data* **2020**, *7*, 134.
(38) Rai, B. K.; Sresht, V.; Yang, Q.; Unwalla, R.; Tu, M.; Mathiowetz, A. M.; Bakken, G. A. TorsionNet: A Deep Neural Network to Rapidly Predict Small-Molecule Torsional Energy Profiles with the Accuracy of Quantum Mechanics. *J. Chem. Inf. Model.* **2022**, *62*, 785−800.
(39) Řezáč, J. Non-Covalent Interactions Atlas Benchmark Data Sets: Hydrogen Bonding. *J. Chem. Theory Comput.* **2020**, *16*, 2355−2368.
(40) Goerigk, L.; Kruse, H.; Grimme, S. Benchmarking Density Functional Methods against the S66 and S66×8 Datasets for Non-Covalent Interactions. *ChemPhysChem* **2011**, *12*, 3421−33.
(41) Brauer, B.; Kesharwani, M. K.; Kozuch, S.; Martin, J. M. L. The S66×8 benchmark for noncovalent interactions revisited: explicitly correlated ab initio methods and density functional theory. *Phys. Chem. Chem. Phys.* **2016**, *18*, 20905−25.
(42) Goerigk, L.; Hansen, A.; Bauer, C.; Ehrlich, S.; Najibi, A.; Grimme, S. A look at the density functional theory zoo with the advanced GMTKN55 database for general main group thermochemistry, kinetics and noncovalent interactions. *Phys. Chem. Chem. Phys.* **2017**, *19*, 32184−32215.
(43) Moser, A.; Range, K.; York, D. M. Accurate Proton Affinity and Gas-Phase Basicity Values for Molecules Important in Biocatalysis. *J. Phys. Chem. B* **2010**, *114*, 13911−13921.
(44) Smith, J.; Nebgen, B.; Zubatyuk, R.; Lubbers, N.; Devereux, C.; Barros, K.; Tretiak, S.; Isayev, O.; Roitberg, A. Approaching Coupled Cluster Accuracy with a General-purpose Neural Network Potential Through Transfer Learning. *Nat. Commun.* **2019**, *10*, 2903.
(45) Devereux, C.; Smith, J. S.; Huddleston, K. K.; Barros, K.; Zubatyuk, R.; Isayev, O.; Roitberg, A. E. Extending the Applicability of the ANI Deep Learning Molecular Potential to Sulfur and Halogens. *J. Chem. Theory Comput.* **2020**, *16*, 4192−4202.
(46) Martin, Y. C. Experimental and pK prediction aspects of tautomerism of drug-like molecules. *Drug Discovery Today. Technol.* **2018**, *27*, 59−64.
(47) Yang, Y.; Yu, H.; York, D. M.; Cui, Q.; Elstner, M. Extension of the Self-Consistent-Charge Density-Functional Tight-Binding Method: Third-Order Expansion of the Density Functional Theory Total Energy and Introduction of a Modified Effective Coulomb Interaction. *J. Phys. Chem. A* **2007**, *111*, 10861−10873.
(48) 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.
(49) Zeng, J.; Giese, T. J.; Ekesan, Ş.; York, D. M. Development of Range-Corrected Deep Learning Potentials for Fast, Accurate Quantum Mechanical/Molecular Mechanical Simulations of Chemical Reactions in Solution. *J. Chem. Theory Comput.* **2021**, *17*, 6993−7009.
(50) Giese, T. J.; Zeng, J.; Ekesan, Ş.; York, D. M. Combined QM/MM, Machine Learning Path Integral Approach to Compute Free Energy Profiles and Kinetic Isotope Effects in RNA Cleavage Reactions. *J. Chem. Theory Comput.* **2022**, *18*, 4304−4317.
(51) Zheng, P.; Zubatyuk, R.; Wu, W.; Isayev, O.; Dral, P. O. Artificial intelligence-enhanced quantum chemical method with broad applicability. *Nat. Commun.* **2021**, *12*, 7022.
(52) Gómez-Flores, C. L.; Maag, D.; Kansari, M.; Vuong, V.-Q.; Irle, S.; Gräter, F.; Kubař, T.; Elstner, M. Accurate Free Energies for Complex Condensed-Phase Reactions Using an Artificial Neural

Network Corrected DFTB/MM Methodology. *J. Chem. Theory Comput.* **2022**, *18*, 1213−1226.
(53) Böser, J.; Kubař, T.; Elstner, M.; Maag, D. Reduction pathway of glutaredoxin 1 investigated with QM/MM molecular dynamics using a neural network correction. *J. Chem. Phys.* **2022**, *157*, 154104.
(54) Dral, P. O.; Zubatiuk, T.; Xue, B.-X. In *Quantum Chemistry in the Age of Machine Learning*; Dral, P. O., Ed.; Elsevier: 2022; Chapter 21, pp 491−507, DOI: 10.1016/B978-0-323-90049-2.00012-3.
(55) Darden, T.; York, D.; Pedersen, L. Particle mesh Ewald: An N log(N) method for Ewald sums in large systems. *J. Chem. Phys.* **1993**, *98*, 10089−10092.
(56) Nam, K.; Gao, J.; York, D. M. An efficient linear-scaling Ewald method for long-range electrostatic interactions in combined QM/MM calculations. *J. Chem. Theory Comput.* **2005**, *1*, 2−13.
(57) Giese, T. J.; Panteva, M. T.; Chen, H.; York, D. M. Multipolar Ewald methods, 1: Theory, accuracy, and performance. *J. Chem. Theory Comput.* **2015**, *11*, 436−450.
(58) Giese, T. J.; Panteva, M. T.; Chen, H.; York, D. M. Multipolar Ewald methods, 2: Applications using a quantum mechanical force field. *J. Chem. Theory Comput.* **2015**, *11*, 451−461.
(59) Giese, T. J.; York, D. M. Ambient-Potential Composite Ewald Method for ab Initio Quantum Mechanical/Molecular Mechanical Molecular Dynamics Simulation. *J. Chem. Theory Comput.* **2016**, *12*, 2611−2632.
(60) Eberlein, L.; Beierlein, F. R.; van Eikema Hommes, N. J. R.; Radadiya, A.; Heil, J.; Benner, S. A.; Clark, T.; Kast, S. M.; Richards, N. G. J. Tautomeric Equilibria of Nucleobases in the Hachimoji Expanded Genetic Alphabet. *J. Chem. Theory Comput.* **2020**, *16*, 2766−2777.
(61) Wahl, O.; Sander, T. Tautobase: An Open Tautomer Database. *J. Chem. Inf. Model.* **2020**, *60*, 1085−1089.
(62) Wieder, M.; Fass, J.; Chodera, J. D. Fitting quantum machine learning potentials to experimental free energy data: predicting tautomer ratios in solution. *Chem. Sci.* **2021**, *12*, 11364−11381.
(63) Ree, N.; Göller, A. H.; Jensen, J. H. RegioSQM20: improved prediction of the regioselectivity of electrophilic aromatic substitutions. *J. Cheminform.* **2021**, *13*, 10.
(64) Fink, T.; Bruggesser, H.; Reymond, J.-L. Virtual Exploration of the Small-Molecule Chemical Universe below 160 Da. *Angew. Chem., Int. Ed.* **2005**, *44*, 1504−8.
(65) Blum, L. C.; Reymond, J.-L. 970 Million Druglike Small Molecules for Virtual Screening in the Chemical Universe Database GDB-13. *J. Am. Chem. Soc.* **2009**, *131*, 8732−8733.
(66) Law, V.; Knox, C.; Djoumbou, Y.; Jewison, T.; Guo, A. C.; Liu, Y.; Maciejewski, A.; Arndt, D.; Wilson, M.; Neveu, V.; Tang, A.; Gabriel, G.; Ly, C.; Adamjee, S.; Dame, Z. T.; Han, B.; Zhou, Y.; Wishart, D. S. DrugBank 4.0: shedding new light on drug metabolism. *Nucleic Acids Res.* **2014**, *42*, D1091.
(67) Giese, T. J.; York, D. M. Development of a Robust Indirect Approach for MM → QM Free Energy Calculations That Combines Force-Matched Reference Potential and Bennett's Acceptance Ratio Methods. *J. Chem. Theory Comput.* **2019**, *15*, 5543−5562.
(68) König, G.; Mei, Y.; Pickard 4th, F. C.; Simmonett, A. C.; Miller, B. T.; Herbert, J. M.; Woodcock, H. L.; Brooks, B. R.; Shao, Y. Computation of Hydration Free Energies Using the Multiple Environment Single System Quantum Mechanical/Molecular Mechanical Method. *J. Chem. Theory Comput.* **2016**, *12*, 332−44.
(69) Kearns, F. L.; Hudson, P. S.; Boresch, S.; Woodcock, H. L. Methods for Efficiently and Accurately Computing Quantum Mechanical Free Energies for Enzyme Catalysis. *Methods Enzymol* **2016**, *577*, 75−104.
(70) Schöller, A.; Kearns, F.; Woodcock, H. L.; Boresch, S. Optimizing the Calculation of Free Energy Differences in Non-equilibrium Work SQM/MM Switching Simulations. *J. Phys. Chem. B* **2022**, *126*, 2798−2811.
(71) Wang, H.; Zhang, L.; Han, J.; E, W. DeePMD-kit: A deep learning package for many-body potential energy representation and molecular dynamics. *Comput. Phys. Commun.* **2018**, *228*, 178−184.
(72) Liang, W.; Zeng, J.; York, D. M.; Zhang, L.; Wang, H. In *A Practical Guide to Recent Advances in Multiscale Modelling and Simulation for Biomolecules*; Wang, Y., Ed.; AIP Publishing: 2023; Chapter Learning DeePMD-kit: A guide to building Deep Potential models.
(73) Chai, J.-D.; Head-Gordon, M. Systematic optimization of long-range corrected hybrid density functionals. *J. Chem. Phys.* **2008**, *128*, 084106.
(74) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Petersson, G. A.; Nakatsuji, H.; Li, X.; Caricato, M.; Marenich, A. V.; Bloino, J.; Janesko, B. G.; Gomperts, R.; Mennucci, B.; Hratchian, H. P.; Ortiz, J. V.; Izmaylov, A. F.; Sonnenberg, J. L.; Williams-Young, D.; Ding, F.; Lipparini, F.; Egidi, F.; Goings, J.; Peng, B.; Petrone, A.; Henderson, T.; Ranasinghe, D.; Zakrzewski, V. G.; Gao, J.; Rega, N.; Zheng, G.; Liang, W.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Throssell, K.; Montgomery, J. A., Jr.; Peralta, J. E.; Ogliaro, F.; Bearpark, M. J.; Heyd, J. J.; Brothers, E. N.; Kudin, K. N.; Staroverov, V. N.; Keith, T. A.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A. P.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Millam, J. M.; Klene, M.; Adamo, C.; Cammi, R.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Farkas, O.; Foresman, J. B.; Fox, D. J.*Gaussian 16*, Revision A.03; Gaussian Inc.: Wallingford, CT, 2016.
(75) Gaus, M.; Goez, A.; Elstner, M. Parametrization and Benchmark of DFTB3 for Organic Molecules. *J. Chem. Theory Comput.* **2013**, *9*, 338−354.
(76) Ramakrishnan, R.; Dral, P. O.; Rupp, M.; von Lilienfeld, O. A. Quantum chemistry structures and properties of 134 kilo molecules. *Sci. Data* **2014**, *1*, 140022.
(77) Rupp, M.; Tkatchenko, A.; Müller, K.-R.; von Lilienfeld, O. Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning. *Phys. Rev. Lett.* **2012**, *108*, 058301.
(78) Smith, J. S.; Isayev, O.; Roitberg, A. E. ANI-1, A data set of 20 million calculated off-equilibrium conformations for organic molecules. *Sci. Data* **2017**, *4*, 170193−170200.
(79) 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.
(80) Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31−36.
(81) Gao, X.; Ramezanghorbani, F.; Isayev, O.; Smith, J. S.; Roitberg, A. E. TorchANI: A Free and Open Source PyTorch-Based Deep Learning Implementation of the ANI Neural Network Potentials. *J. Chem. Inf. Model.* **2020**, *60*, 3408−3415.
(82) Biondi, E.; Benner, S. A. Artificially Expanded Genetic Information Systems for New Aptamer Technologies. *Biomedicines* **2018**, *6*, 53.
(83) O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. Open Babel: An open chemical toolbox. *J. Cheminform* **2011**, *3*, 33.
(84) RDKit: Open-source cheminformatics. http://www.rdkit.org (accessed 2023-01-19).
(85) Zeng, J.; Cao, L.; Xu, M.; Zhu, T.; Zhang, J. Z. H. Complex reaction processes in combustion unraveled by neural network-based molecular dynamics simulation. *Nat. Commun.* **2020**, *11*, 5713.
(86) Wen, T.; Zhang, L.; Wang, H.; E, W.; Srolovitz, D. J. Deep potentials for materials science. *Mater. Futures* **2022**, *1*, 022601.
(87) Lu, D.; Jiang, W.; Chen, Y.; Zhang, L.; Jia, W.; Wang, H.; Chen, M. DP Compress: A Model Compression Scheme for Generating Efficient Deep Potential Models. *J. Chem. Theory Comput.* **2022**, *18*, 5559−5567.
(88) Goodfellow, I.; Bengio, Y.; Courville, A.*Deep Learning*; MIT Press: 2016. http://www.deeplearningbook.org (accessed 2023-01-19).
(89) Case, D. A.; Belfon, K.; Ben-Shalom, I. Y.; Brozell, S. R.; Cerutti, D. S.; Cheatham, T. E., III; Cruzeiro, V. W. D.; Darden, T. A.;

Duke, R. E.; Giambasu, G.; Gilson, M. K.; Gohlke, H.; Goetz, A. W.; Harris, R.; Izadi, S.; Izmailov, S. A.; Kasavajhala, K.; Kovalenko, K.; Krasny, R.; Kurtzman, T.; Lee, T.; Le-Grand, S.; Li, P.; Lin, C.; Liu, J.; Luchko, T.; Luo, R.; Man, V.; Merz, K.; Miao, Y.; Mikhailovskii, O.; Monard, G.; Nguyen, H.; Onufriev, A.; Pan, F.; Pantano, S.; Qi, R.; Roe, D. R.; Roitberg, A.; Sagui, C.; Schott-Verdugo, S.; Shen, J.; Simmerling, C. L.; Skrynnikov, N.; Smith, J.; Swails, J.; Walker, R. C.; Wang, J.; Wilson, R. M.; Wolf, R. M.; Wu, X.; Xiong, Y.; Xue, Y.; York, D. M.; Kollman, P. A.*AMBER 20*; University of California: San Francisco: San Francisco, CA, 2020.

(90) Walker, R. C.; Crowley, M. F.; Case, D. A. The implementation of a fast and accurate QM/MM potential method in Amber. *J. Comput. Chem.* **2008**, *29*, 1019−1031.

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

(92) Seabra, G.; Walker, R. C.; Elstner, M.; Case, D. A.; Roitberg, A. E. Implementation of the SCC-DFTB method for hybrid QM/MM simulations within the Amber molecular dynamics package. *J. Phys. Chem. A* **2007**, *111*, 5655−5664.

(93) Dewar, M. J. S.; Thiel, W. A semiempirical model for the two-Center repulsion integrals in the NDDO approximation. *Theor. Chim. Acta* **1977**, *46*, 89−104.

(94) Dewar, M. J. S.; Zoebisch, E.; Healy, E. F.; Stewart, J. J. P. Development and use of quantum mechanical molecular models. 76. AM1: a new general purpose quantum mechanical molecular model. *J. Am. Chem. Soc.* **1985**, *107*, 3902−3909.

(95) Stewart, J. J. P. Optimization of parameters for semiempirical methods V: Modification of NDDO approximations and application to 70 elements. *J. Mol. Model.* **2007**, *13*, 1173−1213.

(96) Hourahine, B.; Aradi, B.; Blum, V.; Bonafe, F.; Buccheri, A.; Camacho, C.; Cevallos, C.; Deshaye, M. Y.; Dumitrica, T.; Dominguez, A.; Ehlert, S.; Elstner, M.; van der Heide, T.; Hermann, J.; Irle, S.; Kranz, J. J.; Kohler, C.; Kowalczyk, T.; Kubar, T.; Lee, I. S.; Lutsker, V.; Maurer, R. J.; Min, S. K.; Mitchell, I.; Negre, C.; Niehaus, T. A.; Niklasson, A. M. N.; Page, A. J.; Pecchia, A.; Penazzi, G.; Persson, M. P.; Rezac, J.; Sanchez, C. G.; Sternberg, M.; Stohr, M.; Stuckenberg, F.; Tkatchenko, A.; z. Yu, V. W.; Frauenheim, T. DFTB+, a software package for efficient approximate density functional theory based atomistic simulations. *J. Chem. Phys.* **2020**, *152*, 124101.

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

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

(99) Hjorth Larsen, A.; Jørgen Mortensen, J.; Blomqvist, J.; Castelli, I. E.; Christensen, R.; Dułak, M.; Friis, J.; Groves, M. N.; Hammer, B.; Hargus, C.; Hermes, E. D.; Jennings, P. C.; Bjerre Jensen, P.; Kermode, J.; Kitchin, J. R.; Leonhard Kolsbjerg, E.; Kubal, J.; Kaasbjerg, K.; Lysgaard, S.; Bergmann Maronsson, J.; Maxson, T.; Olsen, T.; Pastewka, L.; Peterson, A.; Rostgaard, C.; Schiøtz, J.; Schütt, O.; Strange, M.; Thygesen, K. S.; Vegge, T.; Vilhelmsen, L.; Walter, M.; Zeng, Z.; Jacobsen, K. W. The atomic simulation environment—a Python library for working with atoms. *J. Phys.: Condens. Matter* **2017**, *29*, 273002.

(100) Liu, D. C.; Nocedal, J. On the limited memory BFGS method for large scale optimization. *Mathematical Programming* **1989**, *45*, 503−528.

(101) Cournia, Z.; Chipot, C.; Roux, B.; York, D. M.; Sherman, W. In *Free Energy Methods in Drug Discovery—Introduction*; Armacost, K. A., Thompson, D. C., Eds.; ACS Symposium Series; 2021; Vol. *1397*, pp 1−38, DOI: 10.1021/bk-2021-1397.ch001.

(102) Weber, W.; Thiel, W. Orthogonalization corrections for semiempirical methods. *Theor. Chem. Acc.* **2000**, *103*, 495−506.

(103) Huang, M.; Giese, T. J.; Lee, T.-S.; York, D. M. Improvement of DNA and RNA Sugar Pucker Profiles from Semiempirical Quantum Methods. *J. Chem. Theory Comput.* **2014**, *10*, 1538−1545.

(104) Kuechler, E. R.; Giese, T. J.; York, D. M. Charge-dependent many-body exchange and dispersion interactions in combined QM/MM simulations. *J. Chem. Phys.* **2015**, *143*, 234111.

(105) Perreault, D. M.; Anslyn, E. V. Unifying the Current Data on the Mechanism of Cleavage-Transesterification of RNA. *Angew. Chem., Int. Ed.* **1997**, *36*, 432−450.

(106) Emilsson, G. M.; Nakamura, S.; Roth, A.; Breaker, R. R. Ribozyme speed limits. *RNA* **2003**, *9*, 907−918.

(107) Bevilacqua, P. C.; Harris, M. E.; Piccirilli, J. A.; Gaines, C.; Ganguly, A.; Kostenbader, K.; Ekesan, Ş.; York, D. M. An Ontology for Facilitating Discussion of Catalytic Strategies of RNA-Cleaving Enzymes. *ACS Chem. Biol.* **2019**, *14*, 1068−1076.

(108) Anslyn, E.; Breslow, R. On the mechanism of catalysis by ribonuclease: cleavage and isomerization of the dinucleotide UpU catalyzed by imidazole buffers. *J. Am. Chem. Soc.* **1989**, *111*, 4473−4482.

(109) Raines, R. T. Ribonuclease A. *Chem. Rev.* **1998**, *98*, 1045−1066.

(110) Gu, H.; Zhang, S.; Wong, K.-Y.; Radak, B. K.; Dissanayake, T.; Kellerman, D. L.; Dai, Q.; Miyagi, M.; Anderson, V. E.; York, D. M.; Piccirilli, J. A.; Harris, M. E. Experimental and computational analysis of the transition state for ribonuclease A-catalyzed RNA 2′-*O*-transphosphorylation. *Proc. Natl. Acad. Sci. U.S.A.* **2013**, *110*, 13002−13007.

(111) Lilley, D. M. J. Classification of the nucleolytic ribozymes based upon catalytic mechanism. *F1000 Res.* **2019**, *8*, 1462.

(112) Gaines, C. S.; Piccirilli, J. A.; York, D. M. The L-platform/L-scaffold framework: a blueprint for RNA-cleaving nucleic acid enzyme design. *RNA* **2020**, *26*, 111−125.

(113) Breaker, R. R.; Joyce, G. F. A DNA enzyme that cleaves RNA. *Chem. Biol.* **1994**, *1*, 223−229.

(114) Liu, H.; Yu, X.; Chen, Y.; Zhang, J.; Wu, B.; Zheng, L.; Haruehanroengra, P.; Wang, R.; Li, S.; Lin, J.; Li, J.; Sheng, J.; Huang, Z.; Ma, J.; Gan, J. Crystal Structure of an RNA-Cleaving DNAzyme. *Nat. Commun.* **2017**, *8*, 2006−2015.

(115) Ekesan, Ş.; York, D. M. Dynamical ensemble of the active state and transition state mimic for the RNA-cleaving 8−17 DNAzyme in solution. *Nucleic Acids Res.* **2019**, *47*, 10282−10295.

(116) Borggräfe, J.; Victor, J.; Rosenbach, H.; Viegas, A.; Gertzen, C. G. W.; Wuebben, C.; Kovacs, H.; Gopalswamy, M.; Riesner, D.; Steger, G.; Schiemann, O.; Gohlke, H.; Span, I.; Etzkorn, M. Time-resolved structural analysis of an RNA-cleaving DNA catalyst. *Nature* **2022**, *601*, 144−149.

(117) Formoso, E.; Matxain, J. M.; Lopez, X.; York, D. M. Molecular dynamics simulation of bovine pancreatic ribonuclease A-CpA and transition state-like complexes. *J. Phys. Chem. B* **2010**, *114*, 7371−7382.

(118) Brandstetter, H.; Grams, F.; Glitz, D.; Lang, A.; Huber, R.; Bode, W.; Krell, H. W.; Engh, R. A. The 1.8-Å Crystal Structure of a Matrix Metalloproteinase 8-Barbiturate Inhibitor Complex Reveals a Previously Unobserved Mechanism for Collagenase Substrate Recognition. *J. Biol. Chem.* **2001**, *276*, 17405−12.

(119) Natesan, S.; Subramaniam, R.; Bergeron, C.; Balaz, S. Binding Affinity Prediction for Ligands and Receptors Forming Tautomers and Ionization Species: Inhibition of Mitogen-Activated Protein Kinase-Activated Protein Kinase 2 (MK2). *J. Med. Chem.* **2012**, *55*, 2035−47.

(120) Soliva, R.; Gelpí, J. L.; Almansa, C.; Virgili, M.; Orozco, M. Dissection of the Recognition Properties of p38 MAP Kinase. Determination of the Binding Mode of a New Pyridinyl−Heterocycle Inhibitor Family. *J. Med. Chem.* **2007**, *50*, 283−93.

(121) Christensen, A. S.; Sirumalla, S. K.; Qiao, Z.; O'Connor, M. B.; Smith, D. G. A.; Ding, F.; Bygrave, P. J.; Anandkumar, A.; Welborn, M.; Manby, F. R.; Miller, T. F., 3rd OrbNet Denali: A

machine learning potential for biological and organic chemistry with semi-empirical cost and DFT accuracy. *J. Chem. Phys.* **2021**, *155*, 204103.
(122) Zubatyuk, R.; Smith, J. S.; Nebgen, B. T.; Tretiak, S.; Isayev, O. Teaching a neural network to attach and detach electrons from molecules. *Nat. Commun.* **2021**, *12*, 4870.
(123) Isert, C.; Atz, K.; Jiménez-Luna, J.; Schneider, G. QMugs, quantum mechanical properties of drug-like molecules. *Sci. Data* **2022**, *9*, 273.
(124) Eastman, P.; Behara, P. K.; Dotson, D. L.; Galvelis, R.; Herr, J. E.; Horton, J. T.; Mao, Y.; Chodera, J. D.; Pritchard, B. P.; Wang, Y.; De Fabritiis, G.; Markland, T. E.SPICE, A Dataset of Drug-like Molecules and Peptides for Training Machine Learning Potentials. 2022, *arXiv:2209.10702*. https://arxiv.org/abs/2209.10702 (accessed 2023-01-19).
(125) Khrabrov, K.; Shenbin, I.; Ryabov, A.; Tsypin, A.; Telepov, A.; Alekseev, A.; Grishin, A.; Strashnov, P.; Zhilyaev, P.; Nikolenko, S.; Kadurin, A. nablaDFT: Large-Scale Conformational Energy and Hamiltonian Prediction benchmark and dataset. *Phys. Chem. Chem. Phys.* **2022**, *24*, 25853−25863.
(126) Tian, C.; Kasavajhala, K.; Belfon, K. A. A.; Raguette, L.; Huang, H.; Migues, A. N.; Bickel, J.; Wang, Y.; Pincay, J.; Wu, Q.; Simmerling, C. ff19SB: Amino-Acid-Specific Protein Backbone Parameters Trained against Quantum Mechanics Energy Surfaces in Solution. *J. Chem. Theory Comput.* **2020**, *16*, 528−552.
(127) Pérez, A.; Marchán, I.; Svozil, D.; Sponer, J.; Cheatham, T. E., III; Laughton, C. A.; Orozco, M. Refinement of the AMBER force field for nucleic acids: Improving the description of α/γ conformers. *Biophys. J.* **2007**, *92*, 3817−3829.
(128) Zgarbová, M.; Otyepka, M.; Šponer, J.; Mládek, A.; Banáš, P.; Cheatham, T. E., III; Jurečka, P. Refinement of the Cornell et al. nucleic acids force field based on reference quantum chemical calculations of glycosidic torsion profiles. *J. Chem. Theory Comput.* **2011**, *7*, 2886−2902.
(129) Bergonzo, C.; Cheatham, T. E., III Improved Force Field Parameters Lead to a Better Description of RNA Structure. *J. Chem. Theory Comput.* **2015**, *11*, 3969−3972.
(130) Izadi, S.; Anandakrishnan, R.; Onufriev, A. V. Building Water Models: A Different Approach. *J. Phys. Chem. Lett.* **2014**, *5*, 3863−3871.
(131) Izadi, S.; Onufriev, A. V. Accuracy limit of rigid 3-point water models. *J. Chem. Phys.* **2016**, *145*, 074501−074510.
(132) Li, P.; Roberts, B. P.; Chakravorty, D. K.; Merz, K. M., Jr. Rational design of Particle Mesh Ewald compatible Lennard-Jones parameters for + 2 metal cations in explicit solvent. *J. Chem. Theory Comput.* **2013**, *9*, 2733−2748.
(133) Li, P.; Merz, K. M., Jr. Taking into account the ion-induced dipole interaction in the nonbonded model of ions. *J. Chem. Theory Comput.* **2014**, *10*, 289−297.
(134) Li, P.; Merz, K. M. Metal Ion Modeling Using Classical Mechanics. *Chem. Rev.* **2017**, *117*, 1564−1686.