# Modern semiempirical electronic structure methods and machine learning potentials for drug discovery: conformers, tautomers and protonation states

Jinzhe Zeng (曾晋哲),[1] Yujun Tao (陶玉君),[1] Timothy J. Giese,[1] and Darrin M. York[1]

*Laboratory for Biomolecular Simulation Research, Institute for Quantitative Biomedicine and Department of Chemistry and Chemical Biology, Rutgers University, Piscataway, NJ 08854, USA*

(*Electronic mail: Darrin.York@rutgers.edu)

## ABSTRACT

Modern semiempirical electronic structure methods have considerable promise in drug discovery as universal "force fields" that can reliably model biological and drug-like molecules, including alternative tautomers and protonation states. Herein, we compare the performance of several NDDO-based semiempirical (MNDO/d, AM1, PM6, PM6-D3H4X, PM7, and ODM2), density-functional tight-binding based (DFTB3, DFTB/ChIMES, GFN1-xTB, and GFN2-xTB) models with pure machine learning potentials (ANI-1x and ANI-2x) and hybrid quantum mechanical/machine learning potentials (AIQM1 and QD$\pi$) for a wide range of data computed at a consistent $\omega$B97X/6-31G* level of theory (as in the ANI-1x database). This data includes conformational energies, intermolecular interactions, tautomers, and protonation states. Additional comparisons are made to a set of natural and synthetic nucleic acids from the artificially expanded genetic information system (AEGIS) that have important implications in the design of new biotechnology and therapeutics. Finally, we examine acid/base chemistry relevant for RNA cleavage reactions catalyzed by small nucleolytic ribozymes, DNAzymes and ribonucleases. Overall, the hybrid quantum mechanical/machine learning potentials appear to be the most robust for these datasets, and the recently developed QD$\pi$ model performs exceptionally well, having especially high accuracy for tautomers and protonation states relevant to drug discovery.

## I. INTRODUCTION

Alchemical free energy (AFE) simulations[1] are widely used for the prediction of ligand-protein binding energies in drug discovery. These predictions are used to prioritize compounds for costly synthesis and testing in the lead optimization cycle.[2] The predictive capability of these methods relies critically on the accuracy of the force fields that are used.[3] For well-studied biological systems such as proteins[4–6], and common solvents such as water[7–11] and monovalent ions[12–15], several molecular mechanical (MM) force fields[16,17] have been developed and undergone extensive validation and revision based on comparison with a wide range of experiments. These force fields have evolved to become increasingly robust and reliable in long-time molecular dynamics simulations, despite the simplicity of their functional forms. On the other hand, the "general" molecular mechanical force fields needed to model

drug-like molecules that may not have ever been synthesized before, are generally much less reliable. Moreover, conventional MM force fields are not "universal" in the sense that they use a pre-defined covalent bonding topology, and are thus limited in their ability to model alternative tautomers and protonation states. This is important as 30% of the compounds in vendor databases and 21% drug databases have potential tautomers[18,19]; further it has been estimated that up to 95% of drug molecules contain ionizable groups[18] (~75% weak bases and ~20% weak acids[20,21]).

Modern semiempirical quantum mechanical (QM) electronic structure methods[22,23] provide an attractive alternative to the general MM force fields for drug discovery. The reason is that, unlike a typical protein that may contain several thousands of atoms, approximately 79% of drugs are between 10-40 non-hydrogen atoms and the vast majority are less than 100 non-hydrogen atoms.[24] This is of the size range where semiempirical QM methods are able to be used in combined quantum mechanical/molecular mechanical (QM/MM) simulations that include explicit MM representation of the entire protein and surrounding solvent bath under periodic boundary conditions.[25–28] Highly efficient (including parallel and GPU-accelerated) implementations of semiempirical molecular orbital[29] and density-functional tight-binding[30] have been made and are available for molecular dynamics simulations. More importantly, in the context of AFE simulations, these QM/MM potentials can be efficiently integrated into thermodynamic cycles using an indirect (or sometimes referred as "book-ending" or "reference potential") approaches[31–35] that apply an end-state MM→QM free energy correction to a high-precision MM AFE simulation.

One potential caveat is the high level of accuracy required by drug discovery applications that seek to distinguish binding free energies to a resolution of below $k_BT$ (0.59 kcal/mol at 300K)[36–38]. This is extremely challenging for even the most advanced modern semiempirical QM methods. One path forward that appears promising is to use machine-learning potentials (MLPs) either as stand-alone alternative models[39–44], or else to augment existing semiempirical QM methods.[45–51] We will refer to the former class as "pure MLPs" and the latter class as "QM/Δ-MLPs". MLPs have emerged as powerful tools to enable fast and accurate chemical models within the scope of their training[39,41–44]. Many such models have emerged for different applications[52–67], although few, if any, have been used to their full potential in rigorous AFE simulations. Application of these models in drug discovery AFE simulations is challenging because they must: 1) make robust predictions for molecules

within the relevant medicinal chemistry space that may have never been synthesized or characterized[68], 2) model a wide range of intra- and intermolecular interactions, including relative conformational energies, hydrogen bonding[69], $\pi$ stacking[70,71], London dispersion[72] and mixed interactions, 3) quantitatively handle different tautomers[18,19,73] and protonation states[21]. Currently, the ANI[63,74–76] class of models, and particularly the second generation ANI-2x[76] have received widespread attention. A limitation of these models is that they were built for neutral molecules, and their functional forms do not explicitly account for total molecular charge nor spin state. Consequently, they are not able to reliably predict energetics for changing protonation states. This is a serious limitation, as it has been estimated that up to 95% of drug molecules contain ionizable groups.[18] Related to this, some of the pure MLPs did not initially treat long-ranged electrostatic interactions, although there have been efforts to remedy this.[66] Alternatively, there have been several recent efforts to develop new QM/$\Delta$-MLPs[45–51,77], the most relevant in the current context being AIQM1[46] that is based as the novel ODMx class of semiempirical models[78] and has recently been demonstrated to be robust for transition state optimizations[79].

Very recently we introduced a first-generation QM/$\Delta$-MLP for drug discovery.[77] The Quantum Deep-learning Potential Interaction (QD$\pi$) model uses a fast, robust 3rd-order self-consistent density-functional tight binding (DFTB3/3OB) model[80,81] that is corrected to high-level accuracy through an MLP correction ($\Delta$-MLP) based on our range-corrected deep-learning potential (DPRc)[47,48] as part of DeePMD-kit[82] interfaced with AMBER[83]. The underlying DFTB3 model is able to capture long-range electrostatic interactions, as well as changes in charge, protonation, and spin state. The intramolecular and short- to mid-range intermolecular interactions are made quantitatively accurate by training the DPRc model to correct the total energy and forces to match those of high-level *ab initio* methods.

In the present work, we compare the performance of several modern semiempirical QM, QM/$\Delta$-MLP and pure MLP models against consistent reference data derived from databases relevant for drug discovery. Of particular focus of the present work is in characterizing the ability of different potentials to accurately model intermolecular interactions, tautomers and protonation states. Toward that end, we consider the a dataset of natural and synthetic nucleic acids from the artificially expanded genetic information system (AEGIS)[84–87] that is being used for a wide range of biotechnology applications[88]. The system uses 12 different nucleobases in its genetic code that include 4 canonical nucleobases found in DNA

(adenine, cytosine, guanine and thymine) in addition to 8 synthetic nucleobases. These serve as good test systems as they contain complex covalent bonding and exhibit a rich set of tautomer forms, hydrogen bonded complexes, and alternative protonation states. The remainder of the manuscript is organized as follows. The Methods section describes the computational details pertaining to the various semiempirical QM (MNDO/d[89], AM1[90], PM6[91], PM6-D3H4X[92,93], PM7[94], ODM2[78], DFTB3[95], GFN1-xTB[96], GFN2-xTB[97], and DFTB/ChIMES[98]), MLP (ANI-1x[74] and ANI-2x[76]) and QM/Δ-MLP (AIQM1[46] and most recently QDπ[77]) models, as well as the key modified databases (DBs) used as reference data at the ωB97X/6-31G*[77,99] level. The Results and Discussion section presents and analyzes data for a set of 10 broad-spectrum databases for intermolecular interactions, tautomers, and protonation states, and 2D conformational energy profiles. Further application is made to examine the performance of modern semiempirical QM, MLP and QM/Δ-MLPs against the AEGIS dataset[85,86]. Finally, the paper provides contextual examples of acid/base chemistry relevant for RNA cleavage reactions catalyzed by small nucleolytic ribozymes and ribonucleases[100].

## II. METHODS

### A. Models compared in the current work

*a. Density-functional reference data.* ωB97X/6-31G*[99] was performed using Gaussian 16[101]. Reference energy and forces (including geometry optimizations, where needed) were performed at a consistent ωB97X/6-31G*[99] level of theory.

*b. NDDO-based semiempirical models.* Semiempirical quantum mechanical (QM) models based on the neglect of diatomic differential overlap (NDDO) approximation enable the number of electron repulsion integrals to be drastically reduced and the single-particle density matrix to be decomposed into effective atom-centered atomic orbital products (and their resulting electrostatics represented as multipoles).[102] The NDDO approximation also eliminates the need to explicitly enforce orthogonalization of the molecular orbitals that normally would be achieved by having an overlap matrix in the generalized eigenvalue equation. Consequently, this may lead to poor modeling of conformational energies and their barriers if left uncorrected. Much work has been made to introduce orthogonalization corrections

into the theoretical framework which has resulted in the OMX class of methods[103–106]. In the current work, the following NDDO-based methods are considered: MNDO/d[89], AM1[90], and PM6[91] which were evaluated with the AMBER 20[107] SQM module[108]; and the ODM2[78] method which was evaluated using the MNDO program[109] kindly provided by Dr. Axel Koslowski, and PM6-D3H4X[92,93], PM7[94] that was performed using the MOPAC software[110]. PM6-D3H4X and PM7 correct PM6 using classical potentials and are often claimed to be the most suitable methodology for drug design among NDDO-based semiempirical models.[111,112]

*c. DFTB-based semiempirical models.* Density-functional tight binding methods offer an intriguing alternative to the NDDO-based semiempirical models. DFTB methods use an expansion of the energy[113] about a sum of neutral atoms densities together with a two-center integral approximation to enable a framework for highly efficient calculations (speed very comparable with NDDO-based methods). Unlike the NDDO-based methods, DFTB methods keep the overlap matrix in the generalized eigenvalue equation, and thus explicitly deal with orbital orthogonalization. However, this complicates the decomposition of the density matrix which now contains 2-center products. Various density-matrix partition schemes can be used to map the density onto atomic centers such that an atom-centered (typically monopolar) representation can be made for the second-order electrostatic term in the expansion. The DFTB-based methods considered here include: DFTB3[95] (3OB parameters[114]) that was performed using the AMBER 20[107] SQM module[108,115]; and GFN1-xTB[96], GFN2-xTB[97], DFTB/ChIMES[98] (3OB parameters[114] and ChIMES parameters[116] kindly provided by Dr. Cong Huy Pham) models evaluated with the DFTB+ software[30].

Compared to DFTB3 and GFN1-xTB, GFN2-xTB represents the first broadly parametrized tight-binding method, primarily designed for the fast calculation of structures and noncovalent interaction energies, to include electrostatic and exchange-correlation Hamiltonian terms up to second order in the multipole expansion.[97] In this way, the model takes into account anisotropic second order density fluctuation effects via short-range damped interactions of cumulative atomic multipole moments. DFTB/ChIMES[116], on the other hand, leverages the relative simplicity of linear regression machine learning in the recently developed Chebyshev Interaction Model for Efficient Simulation (ChIMES) method.[117] Validation tests of DFTB/ChIMES demonstrate the model exhibits both transferability and extensibility, and enable physical and chemical predictions with up to coupled-cluster accuracy.[116]

It should be noted that the use of machine learning methods to enhance DFTB models in

one form or other is not new. Notable works along these lines, in addition to DFTB/ChIMES, include, but are not limited to, the ML-Hamiltonian approach of Yaron and co-workers[118], the development of many-body potentials from deep tensor neural networks[119,120], Gaussian process regression[121], and unsupervised machine learning[122].

*d. Machine learning potentials (MLPs).* The pure machine learning potentials considered in this work produce energies and atomic forces of a molecule given the positions and elements. These potentials are quite fast compared with semiempirical QM models, and they have more favorable scaling properties. However, some initial pure MLPs were built for neutral molecules in singlet ground states, so they do not reliably model changes in charge state that occur with the addition or loss of electrons and/or protons. The latter of which is important for drug molecules that contain ionizable sites. The pure MLPs considered here include: ANI-1x[74] and ANI-2x[76] models performed using the TorchANI software[123]. Both ANI-1x and ANI-2x models use the ANI descriptor[63] with a cutoff radius of 6 Å and were trained against $\omega$B97X/6-31G* with the active learning cycles. The training data of ANI-1x only include energies, and the training data of ANI-2x include both energies and forces.

*e. Combined semiempirical quantum mechanical and machine-learning potentials (QM/$\Delta$-MLPs).* An attractive alternative to either semiempirical QM or pure MLPs is to combine the strengths of both into a combined QM/$\Delta$-MLP. In this way, it builds off of a fast and robust semiempirical QM that inherently can accommodate changes in electronic charge and spin states while using MLPs to greatly enhance the accuracy across a broad spectrum of chemical environments. The QM/$\Delta$-MLPs considered here include: the QD$\pi$[77] model, which is based on DFTB3/3OB[95,107,108,114,115] and the deep-learning potential available in DeePMD-kit[82,83], and the AIQM1@DFT*[46] model based on a ODM2[78,109] model (which includes the D4 dispersion correction[124]) and a trained neural network correction using TorchANI[123]. The MLP component of QD$\pi$ uses the DeepPot-SE descriptor[61] with a cutoff radius of 6 Å and was trained against $\omega$B97X/6-31G* energies and forces for 241 M steps; the MLP part of AIQM1@DFT* uses the ANI descriptor[63] with a cutoff radius of 6 Å and was trained against $\omega$B97X/def2-TZVPP energies and forces for 1000 epochs.[46]

All geometry optimizations using semiempirical QM, MLP or QM/$\Delta$-MLP models were performed using the Limited-memory Broyden-Fletcher-Goldfarb-Shanno (LBFGS) algorithm[125] in the ASE[126] package. Relaxed 2D torsion profiles were made using the same way described in Ref. 77.

### B. Databases and reference data used in the current work

The reference data used in the current work includes the modified ANI-1x[74,77,127], the modified COMP5[74,77,128–130], S66×8[74,131,132], HB375×10[77,133], TautoBase (TB)[77,134,135], amino acids (AA) and nucleic acid (NA)[77,136], PA26 and TAUT15[77,137], RegioSQM20[77,138], and artificially expanded genetic information system (AEGIS)[77,84–86]. All reference data was computed (or re-computed[77]) at the $\omega$B97X/6-31G* level of theory (consistent with the most extensive ANI-1x and COMP5 databases).

Among all reference data, ANI-1x (or the modified version) dataset was used to parametrize DFTB/ChIMES, ANI-1x, ANI-2x, QD$\pi$, and AIQM1; S66×8 was used to parametrize PM6-D3H4X, ANI-2x and QD$\pi$; and TB, AA, NA, PA, and AEGIS were used for the training of QD$\pi$.

## III. RESULTS AND DISCUSSION

The focus of the current article is on comparing modern semiempirical electronic structure methods and machine learning potentials with respect to their ability to accurately model conformers, tautomers and protonation states of biological and drug-like molecules. These methods have potential impact for drug discovery owing to their efficiency and robustness.

### A. Comparison of broad-spectrum databases

Important properties for consideration include: relative conformational energies, a wide range of intermolecular interactions, as well as tautomeric and protonation state relative energies. The QD$\pi$ model was trained with the same reference theory level as ANI-2x[76] ($\omega$B97X/6-31G*) and considered a number of DBs that encompass conformational energies (ANI-1x, COMP5), intermolecular interactions (S66×8, HB375×10) and tautomer (Tauto-Base, Taut15) and protonation state relative energies (AA, NA, PA26, RegioSQM20) that are described in detail elsewhere[77]. A comparison of 11 semiempirical quantum and machine learning models are compared against 10 databases in Table I.

*Conformational energy datasets.* With respect to the diverse conformational energy datasets (ANI-1x[74,127] and COMP5[74,77,128–130]), the mean absolute errors (maEs) in the forces are smallest for the MLP and $\Delta$-MLP potentials (QD$\pi$, ANI-2x, AIQM1 and ANI-1x),

**Table I:** Mean absolute errors for different data sets used for training and testing of the QDπ model.[a]

| | ANI-1x | | S66 | TB | AA | NA | PA | COMP5 | | HB | T15 | SQM |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | E | **F** | ΔE | ΔE | ΔE | ΔE | ΔE | E | **F** | ΔE | ΔE | ΔE |
| QDπ | 0.83 | 1.16 | 0.13 | 0.82 | 0.09 | 0.17 | 0.39 | 1.48 | 1.14 | 0.44 | 0.70 | 2.53 |
| AIQM1 | ... | 3.10 | 0.57 | 2.07 | 7.30 | 4.71 | 5.06 | ... | 2.59 | 0.71 | 1.37 | 2.75 |
| ANI-1x | 1.48 | 4.48 | 1.41 | 1.73 | 86.95 | 52.68 | 43.02 | 1.96 | 3.72 | 1.25 | 1.63 | 16.85 |
| ANI-2x | 1.07 | 2.11 | 0.37 | 1.76 | 70.52 | 52.48 | 23.80 | 1.67 | 1.86 | 1.40 | 1.00 | 13.64 |
| GFN2-xTB | ... | 5.81 | 0.74 | 5.68 | 5.77 | 8.45 | 7.35 | ... | 4.33 | 0.85 | 2.84 | 4.12 |
| GFN1-xTB | ... | 4.69 | 0.77 | 5.23 | 5.00 | 11.73 | 4.43 | ... | 3.68 | 0.87 | 5.32 | 4.10 |
| DFTB3 | ... | 7.58 | 1.14 | 5.45 | 8.63 | 10.85 | 12.54 | ... | 5.46 | 1.17 | 3.65 | 4.59 |
| DFTB/ChIMES | ... | 4.82 | 1.72 | 5.04 | 9.47 | 9.70 | 12.87 | ... | 4.14 | 1.36 | 3.00 | 6.70 |
| ODM2 | ... | 12.80 | 1.24 | 3.37 | 9.13 | 5.26 | 6.04 | ... | 9.97 | 1.29 | 3.64 | 3.99 |
| PM6 | ... | 12.96 | 1.19 | 4.90 | 11.23 | 11.03 | 17.84 | ... | 9.33 | 1.24 | 5.60 | 5.30 |
| PM6-D3H4X | ... | 13.60 | 0.63 | 5.44 | 9.67 | 11.72 | 7.78 | ... | 10.27 | 0.84 | 6.16 | 6.61 |
| PM7 | ... | 11.98 | 0.84 | 4.34 | 7.24 | 10.72 | 10.11 | ... | 8.54 | 1.00 | 3.74 | 5.93 |
| AM1 | ... | 14.95 | 2.17 | 5.01 | 4.43 | 7.32 | 13.51 | ... | 12.13 | 2.57 | 3.99 | 4.13 |
| MNDO/d | ... | 15.14 | 6.67 | 9.69 | 11.71 | 11.29 | 13.07 | ... | 11.52 | 9.36 | 7.78 | 5.18 |

[a]Mean absolute errors in the energy (E, kcal/mol), forces (**F**, kcal/mol/Å) and ΔE for ANI-1x[74,127], S66×8 (S66)[74,131,132], TautoBase (TB)[134,135], amino acid and nucleic acid proton affinities (AA and NA)[136], PA26 (PA)[137], COMP5[74,128–130], HB375×10 (HB)[133], Taut15 (T15)[137] and RegioSQM20 (SQM)[138] databases. Datasets on the right were not part of the QDπ training.

and the QDπ model performs the overall best (maE values 1.16 and 1.14 kcal/mol/Å for ANI-1x and COMP5 datasets, respectively). This is likely due to the fact that the ANI-1x dataset was an integral part of the training of these models. In general, the DFTB models (GFN1-xTB, GFN2-xTB and DFTB3/3OB) have lower force errors with respect to the reference ωB97X/6-31G* values (maE values range from 4.69-7.58 and 3.68-5.46 kcal/mol/Å for ANI-1x and COMP5, respectively), whereas the NDDO-based methods have consider-

ably larger errors (maE values range from 11.98-15.14 and 8.54-12.13 kcal/mol/Å for ANI-1x and COMP5, respectively), with PM7 performing the best of the NDDO methods.

*Intermolecular interaction datasets.* With respect to intermolecular interaction DBs (S66×8[74,131,132] and HB375×10[133]), several models have ΔE values below 1 kcal/mol on average (QDπ, AIQM1, GFN1-xTB, GFN2-xTB, PM6-D3H4X and PM7), with QDπ and AIQM1 having exceptional agreement with the reference data: QDπ has maE values of 0.13 and 0.44 kcal/mol, and AIQM1 has maE values of 0.57 and 0.71 kcal/mol for S66×8 and HB375×10, respectively. The ANI-2x model has excellent maE values for S66×8 (maE 0.37 kcal/mol) but does not perform quite as well for the HB375×10 DB (maE 1.40 kcal/mol). The DFTB3, DFTB/ChIMES, ODM2 and PM6 methods perform similarly with ΔE maE values that range from 1.14-1.72 (S66×8) and 1.17-1.36 (HB375×10) kcal/mol for these DBs. The MNDO/d method has the largest ΔE errors (6.67-9.36 kcal/mol), stemming from known limitations in the core-core interactions that particularly affect hydrogen bonding, and that the empirical modified core-core repulsions in AM1 were designed in part to partially alleviate (AM1 maE values range from 2.17-2.57 kcal/mol).

*Tautomer datasets.* With respect to the tautomer databases, TautoBase[134,135] (TB) and Taut15[137] (T15), only the QDπ model achieves ΔE errors less than 1 kcal/mol (maE values 0.82 and 0.70 kcal/mol for TB and T15, respectively). The AIQM1, and ANI models perform admirably with errors generally below 2 kcal/mol (maE values range from 1.73-2.07 and 1.00-1.37 kcal/mol for TB and T15, respectively). The remainder of the DFTB-based methods have maE values in excess of 5 kcal/mol for TB, and similar values for AM1, PM6 and PM6-D3H4X methods. The ODM2 method makes notable improvement with reduced errors relative to the other NDDO-based methods (maE values of 3.37 and 3.64 kcal/mol for TB and T15, respectively). The MNDO/d method overall performs the worst with maE values for TB and T15 exceeding 9 kcal/mol.

*Relative protonation datasets.* The relative protonation datasets include amino and nucleic acid models compounds[136] (AA and NA) as well as more general proton affinity (PA26[137]) datasets and a subset of the RegioSQM20[138] (SQM) database containing C, H, O, and N elements. The latter involves many relative protonation energies not related to ionizable sites in biological or drug-like molecules, and hence may be of less relevance for drug discovery. For the AA, NA and PA26 datasets, the QDπ model stands alone with respect to having very high accuracy in relative deprotonation energies (maE values range from 0.09-

0.39 kcal/mol). The next best models are AIQM1 (maE 4.71-7.30 kcal/mol). The other semiempirical QM models exhibit much larger ranges: GFN2-xTB (5.77-8.45 kcal/mol), GFN1-xTB (5.00-11.73 kcal/mol), DFTB3 (8.63-12.54 kcal/mol), DFTB/ChIMES (9.47-12.87), ODM2 (5.26-9.13 kcal/mol), PM6 (11.03-17.84 kcal/mol), PM6-D3H4X (7.78-11.72), PM7 (7.24-10.72), AM1 (4.43-13.51 kcal/mol) and MNDO/d (11.29-13.07 kcal/mol). With respect to the SQM dataset, again the QD$\pi$ and AIQM1 models perform best (maE values of 2.53 and 2.75 kcal/mol, respectively), and the remaining semiempirical QM models perform similarly with maE values that range from 3.99-6.70 kcal/mol. The pure MLP models (ANI-1x and ANI-2x) break down with respect to their ability to predict relative protonation/deprotonation energies as these potentials were designed for neutral molecules.

Overall, the QD$\pi$ model performs exceptionally well across all datasets. The AIQM1 model is also impressive in this regard, with the exception of the protonation/deprotonation energies where AIQM1 have larger errors for the AA, NA and PA datasets. Clearly the QM/$\Delta$-MLP form, using DFTB3 or ODM2 as a QM base model, considerably enhances the accuracy across all datasets listed in Table I. The pure MLP models, and particularly ANI-2x, generally perform better than the semiempirical QM models, with the exception of protonation/deprotonation energies where the model gives very larger errors. Of the semiempirical QM models, the DFTB-based methods have smaller force errors than the NDDO-based models. The GFN1-xTB, GFN2-xTB, PM6-D3H4X and PM7 models perform well for intermolecular interactions, slightly better than the DFTB3, DFTB/ChIMES and ODM2 models. All of the semiempirical QM models are fairly comparable in modeling tautomer energy differences (with the exception of MNDO/d that is less accurate), with ODM2 performing best over a broad range of data. For protonation/deprotonation energies, however, there is no clear trend with the semiempirical QM potentials – they all deviate from the reference data with $\Delta$E maE values exceeding 8 kcal/mol for at least one of the datasets (AA, NA, PA or SQM).

In the remainder of the manuscript, we focus comparison to the most recent modern semiempirical QM (DFTB3, DFTB/ChIMES, GFN2-xTB, ODM2, PM6-D3H4X, PM7), MLP (ANI-2x) and QM/$\Delta$-MLP (QD$\pi$ and AIQM1) models.

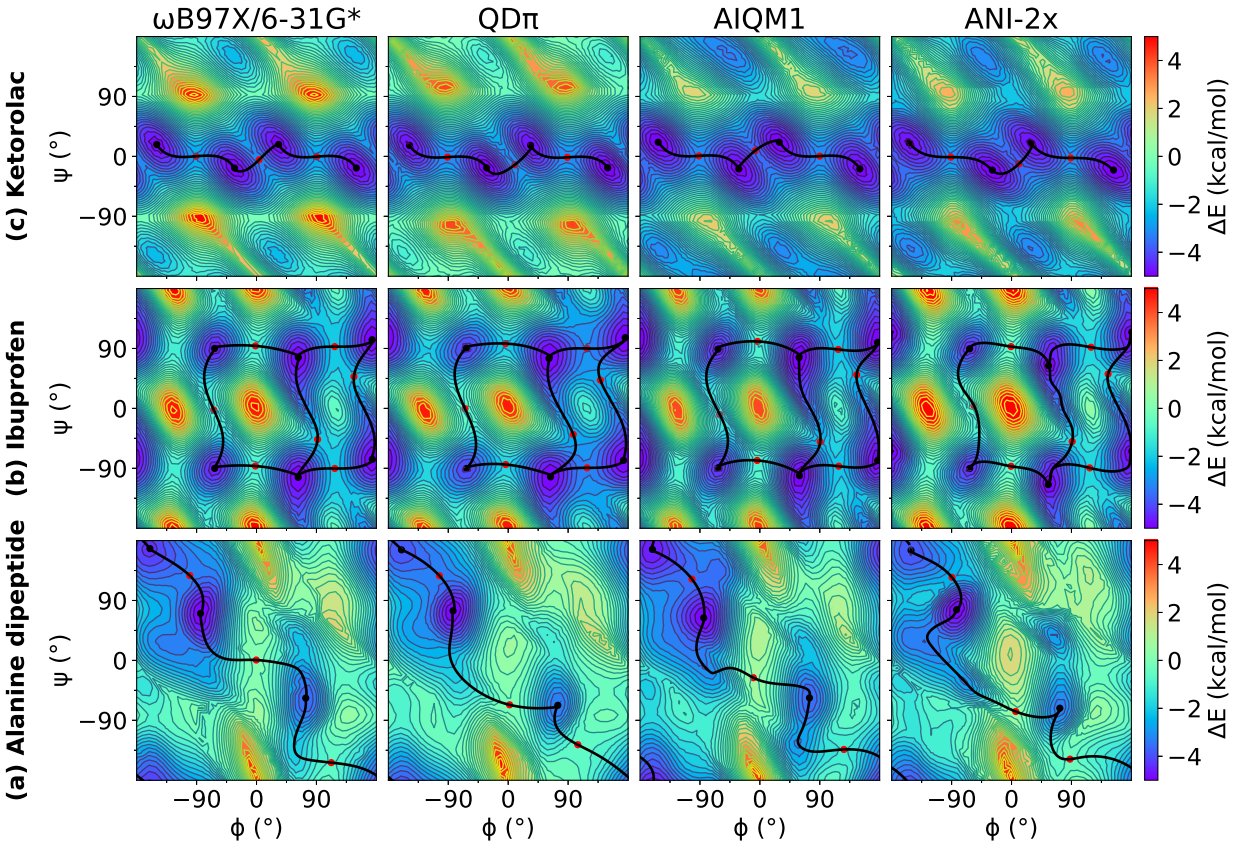


**Figure 1:** Relaxed 2D torsion profiles for (a) alanine dipeptide; (b) ibuprofen; (c) ketorolac. Each molecule was computed using ωB97X/6-31G*, QDπ, AIQM1, and ANI-2x, respectively. The reference level of theory is ωB97X/6-31G*. The color bars represent of the potential energy (with respect to the minimum energy) in kcal/mol.

## B. Comparison of 2D conformation energy profiles

We examined relaxed 2D torsion profiles for three systems: the alanine dipeptide, and the drug molecules ibuprofen and ketorolac illustrated in Figures 1, 2, and 3. These figures compare 2D torsion profiles at the ωB97X/6-31G* reference level with the QM/Δ-MLP/pure MLP models QDπ, AIQM1 and ANI-2x (Fig. 1), DFTB-based GFN2-xTB, DFTB3 and DFTB/ChIMES (Fig. 2), and NDDO-based ODM2, PM3-D3H4X and PM7 (Fig. 3) models. The relative energy values for the stationary points are provided in Table S1 of the Supporting Information. All of the models qualitatively predict the correct trends. A modest exception occurs with PM6-D3H4X and PM7 that do not predict a pronounced minimum in the $\beta$ region (~180/180) of the $\phi/\psi$ map (Fig. 3). Overall, the QDπ and AIQM1 mod-

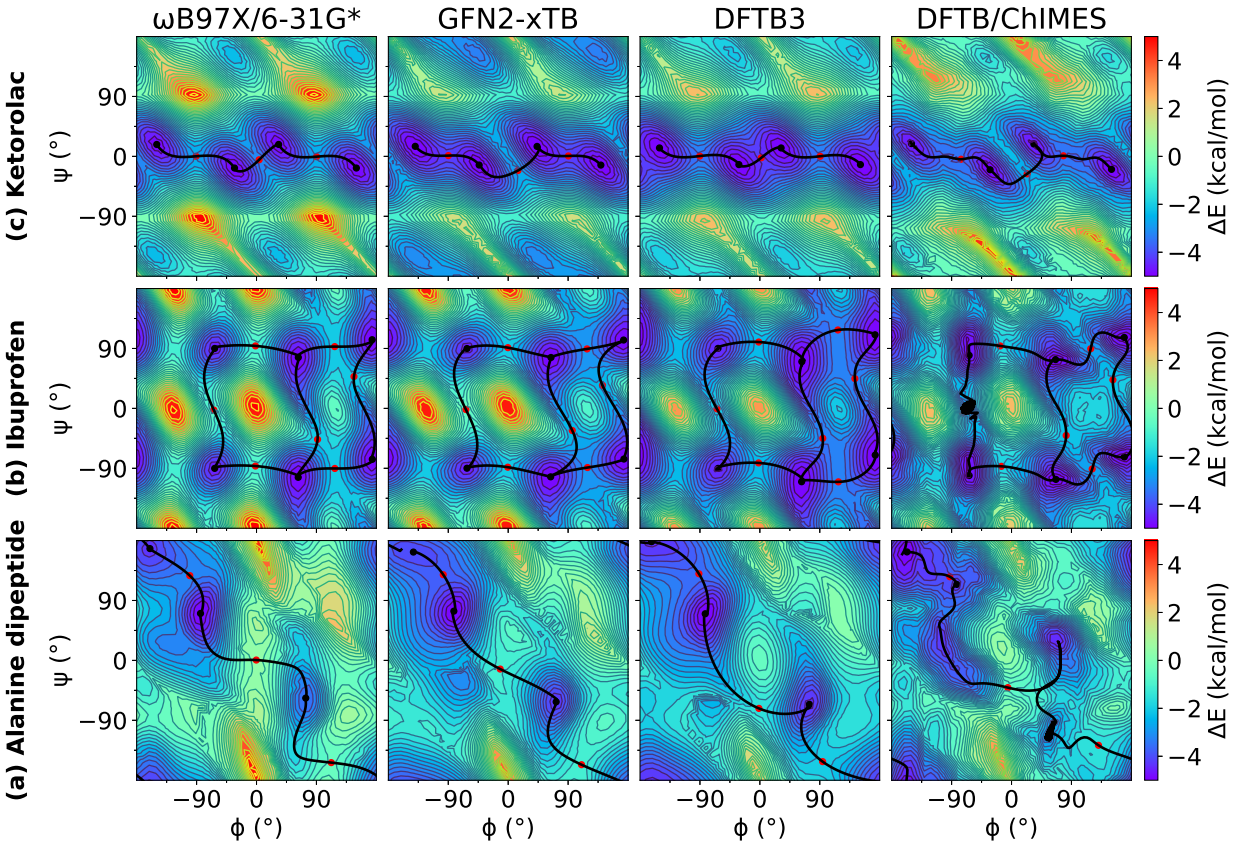


**Figure 2:** Relaxed 2D torsion profiles for (a) alanine dipeptide; (b) ibuprofen; (c) ketorolac. Each molecule was computed using $\omega$B97X/6-31G*, GFN2-xTB, DFTB3, and DFTB/ChIMES, respectively. The reference level of theory is $\omega$B97X/6-31G*. The color bars represent of the potential energy (with respect to the minimum energy) in kcal/mol.

els have the closest agreement with $\omega$B97X/6-31G*, with ANI-2x only slightly worse. The GFN2-xTB, DFTB3 and ODM2 semiempirical QM models tend to systematically underestimate the conformational barriers between minima (Table S1 in Supporting Information). The largest errors that occur for the QD$\pi$ model are for the transition states in the ibuprofen example, that like the semiempirical QM models, are systematically underestimated.

## C. Comparison of hydrogen bond complex energies for natural and artificial nucleic acids

The natural and modified nucleic acids exhibit a wide range of canonical and non-canonical hydrogen bonded base pairs, including some that involve non-standard tautomer

| | | QM/Δ-MLP or MLP | | | DFTB | | | NDDO | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | ωB97X | QDπ | AIQM1 | ANI-2x | GFN2 | DFTB3 | ChIMES | ODM2 | D3H4X | PM7 |
| Complex | ΔE | Err | Err | Err | Err | Err | Err | Err | Err | Err |
| C –●– G | −32.90 | 0.16 | 7.75 | 9.84 | 3.66 | 10.98 | 0.08 | 10.35 | 4.87 | 1.92 |
| T –●– A | −18.22 | −0.14 | 6.71 | 3.65 | 2.15 | 9.33 | 2.19 | 7.51 | 2.90 | 0.88 |
| U –●– A | −18.36 | 0.17 | 7.39 | 3.39 | 2.12 | 9.35 | 2.22 | 7.45 | 2.94 | 0.96 |
| S –●– B | −37.40 | 0.03 | 10.01 | 7.67 | 3.99 | 11.34 | −0.64 | 10.18 | 6.62 | 3.80 |
| V –●– J | −34.52 | 0.00 | 9.70 | 8.04 | 2.43 | 9.32 | −2.73 | 8.57 | 5.25 | 1.83 |
| K –●– X | −22.46 | −0.08 | 7.06 | 6.34 | 2.13 | 9.56 | 1.99 | 9.14 | 1.69 | −1.24 |
| Z –●– P | −33.11 | −0.05 | 8.13 | 10.77 | 3.65 | 10.14 | −0.76 | 10.93 | 5.50 | 1.83 |
| B –●– G | −32.50 | −0.31 | 8.20 | 8.38 | 3.72 | 10.21 | −0.69 | 10.13 | 5.28 | 1.28 |
| B –●– P | −33.68 | 0.13 | 7.45 | 10.02 | 4.56 | 11.02 | −0.56 | 11.61 | 6.47 | 1.77 |
| B –■– G | −22.46 | 0.09 | 8.76 | 9.08 | 3.56 | 10.54 | 3.50 | 9.95 | 4.24 | −0.24 |
| G –●– T* | −33.99 | 0.03 | 7.28 | 10.79 | 4.59 | 10.11 | −3.34 | 9.56 | 5.51 | 3.22 |
| G* –●– T | −23.00 | 0.39 | 7.82 | 7.14 | 3.03 | 9.49 | −0.04 | 8.55 | 0.41 | −0.67 |
| B* –●– G | −25.59 | −0.11 | 8.99 | 13.00 | 3.14 | 10.75 | −2.45 | 8.60 | 1.71 | −3.49 |
| T –●– B* | −22.92 | −0.03 | 5.80 | 6.18 | 3.04 | 9.44 | 0.24 | 8.88 | 0.43 | −0.08 |
| K+ –●– X− | −144.48 | 0.10 | 12.53 | 79.73 | 14.30 | 18.59 | 8.26 | 15.79 | 15.52 | 15.29 |
| Z− –●– G | −43.33 | −0.07 | 12.36 | 11.58 | 6.69 | 15.35 | 3.42 | 12.16 | 4.25 | 1.90 |
| C –●– P+ | −47.17 | 0.04 | 6.77 | 24.45 | 3.38 | 13.41 | 0.87 | 9.41 | 7.35 | 4.82 |
| maE | ... | 0.11 | 8.46 | 14.17 | 4.25 | 11.22 | 1.99 | 10.08 | 4.87 | 2.77 |
| rmsE | ... | 0.15 | 8.66 | 22.51 | 5.09 | 11.49 | 2.83 | 10.26 | 5.97 | 4.44 |

**Table II:** Hydrogen bond complex energies from ωB97X/6-31G* and model errors (kcal/mol) for artificially expanded genetic information system (AEGIS) base pair dataset[77,85,86] with Leontis and Westhof symbols used for classification of nucleic acid base pairs[139–141], including complexes that involve alternative tautomers and protonation states.[a]

[a]Models and datasets are described in the Methods section. An illustration of each of the complexes is provided in Figure 4. Complexes include: Adenine (**A**), cytosine (**C**), guanine (**G**), thymine (**T**), uracil (**U**), isoguanine (**B**), isocytosine (**S**), 6-amino-5-nitropyridin-2-one (**Z**), 2-aminoimidazo[1,2a][1,3,5]triazin-4(1H)-one (**P**), imidazo[1,2-a]-1,3,5-triazine-2(8H)-4(3H)-dione (**X**), 2,4-diaminopyrimidine (**K**), 4-aminoimidazo[1,2-a][1,3,5]triazin-2(8H)-one (**J**), 6-amino-3-methylpyridin-2(1H)-one (**V**)[86,142]. The "*" symbol refers to tautomeric form, and "+" and "-" symbols refer to the positive and negative charge.

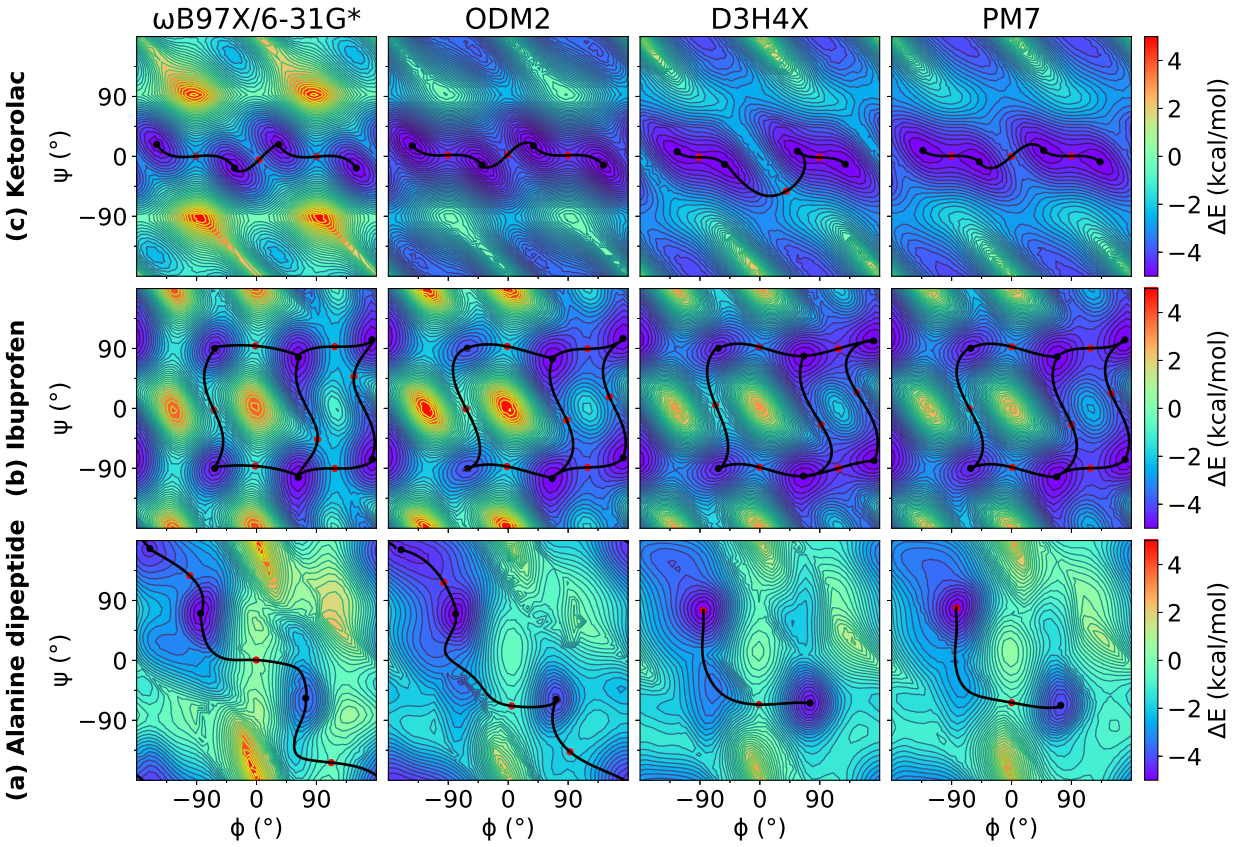


**Figure 3:** Relaxed 2D torsion profiles for (a) alanine dipeptide; (b) ibuprofen; (c) ketorolac. Each molecule was computed using ωB97X/6-31G*, ODM2, PM6-D3H4X, PM7, respectively. The reference level of theory is ωB97X/6-31G*. The color bars represent of the potential energy (with respect to the minimum energy) in kcal/mol.

forms and protonation states. The base pairs considered in the AEGIS dataset[77,85,86] are illustrated in Figure 4. This dataset represents a rich set of hydrogen bonding interactions between endocyclic and exocyclic amine, carbonyl and hydroxyl functional groups. The results are listed in Table II, and the neutral base pairs are illustrated in Figure 5. Overall, the QDπ model gives excellent agreement with the ωB97X/6-31G* reference level over the entire set with ΔE maE of 0.11 kcal/mol and maximum error of 0.39 kcal/mol for G*-T. The DFTB/ChIMES model has next lowest error (maE 1.99 kcal/mol), followed by PM7 (maE 2.77 kcal/mol), and GFN2-xTB (maE 4.25 kcal/mol) and PM6-D3H4X (maE 4.87 kcal/mol). The remainder of the models have maE values in excess of 8 kcal/mol. The ANI-2x model has a large maE value (14.17 kcal/mol), but the errors are dominated by base pairs involving ionized nucleobases that range from 11.58-79.73 kcal/mol, whereas the range

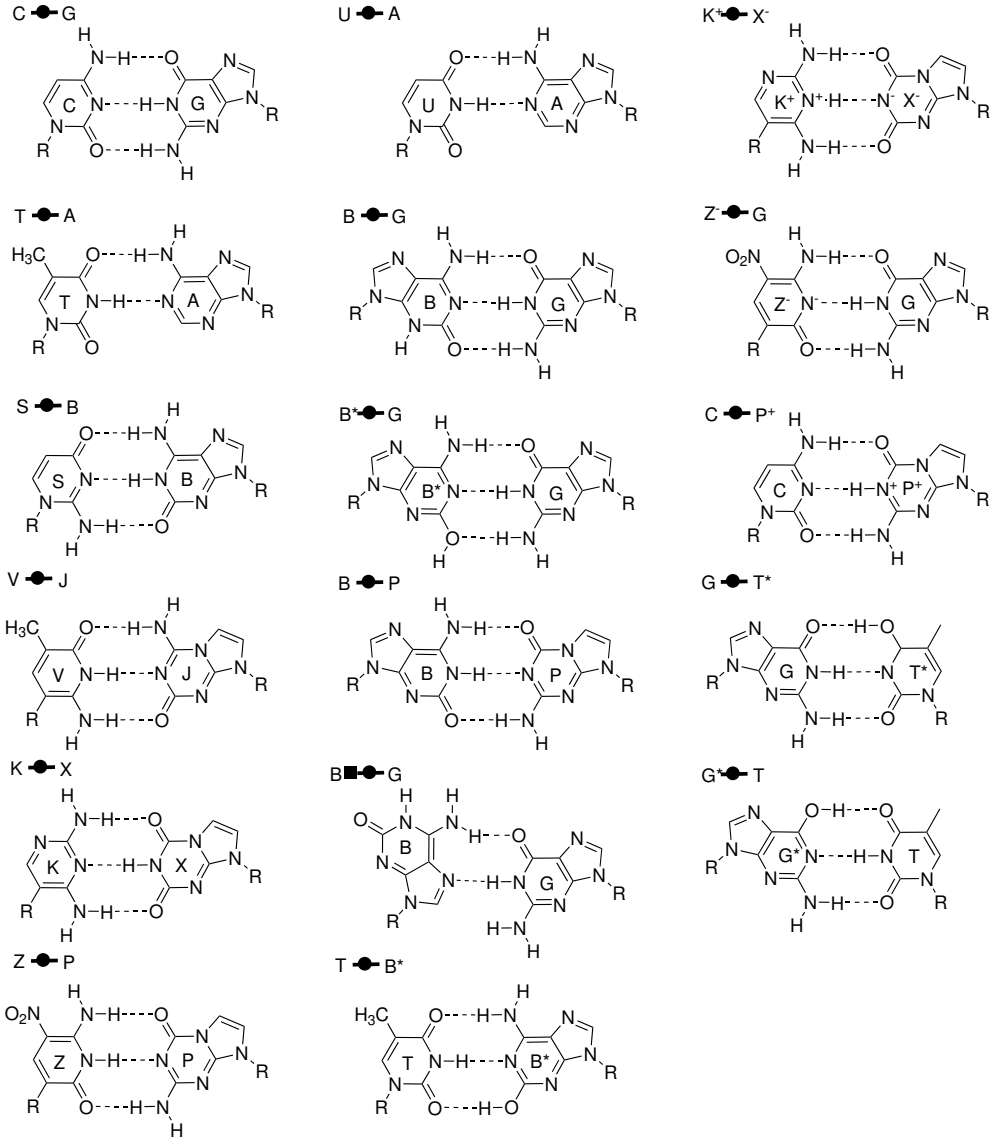


**Figure 4:** Structures for artificially expanded genetic information system (AEGIS) base pair dataset[77,85,86] with Leontis and Westhof symbols used for classification of nucleic acid base pairs[139–141].

of errors for neutral base pairs in much smaller (3.39-13.00 kcal/mol, maE of the neutral subset 9.72 kcal/mol).

Examination of the correlation of hydrogen complex energies for neutral nucleobases reveals that QD$\pi$ has the highest correlation ($R^2$ value 0.999), followed by DFTB/ChIMES, AIQM1 and ODM2 with $R^2$ values of 0.99. Whereas DFTB/ChIMES is well aligned with the reference data, the ODM2 and related AIQM1 models have values systematically shifted to lower $\Delta$E values. Both PM7 and PM6-D3H4X models show impressive correlation ($R^2$

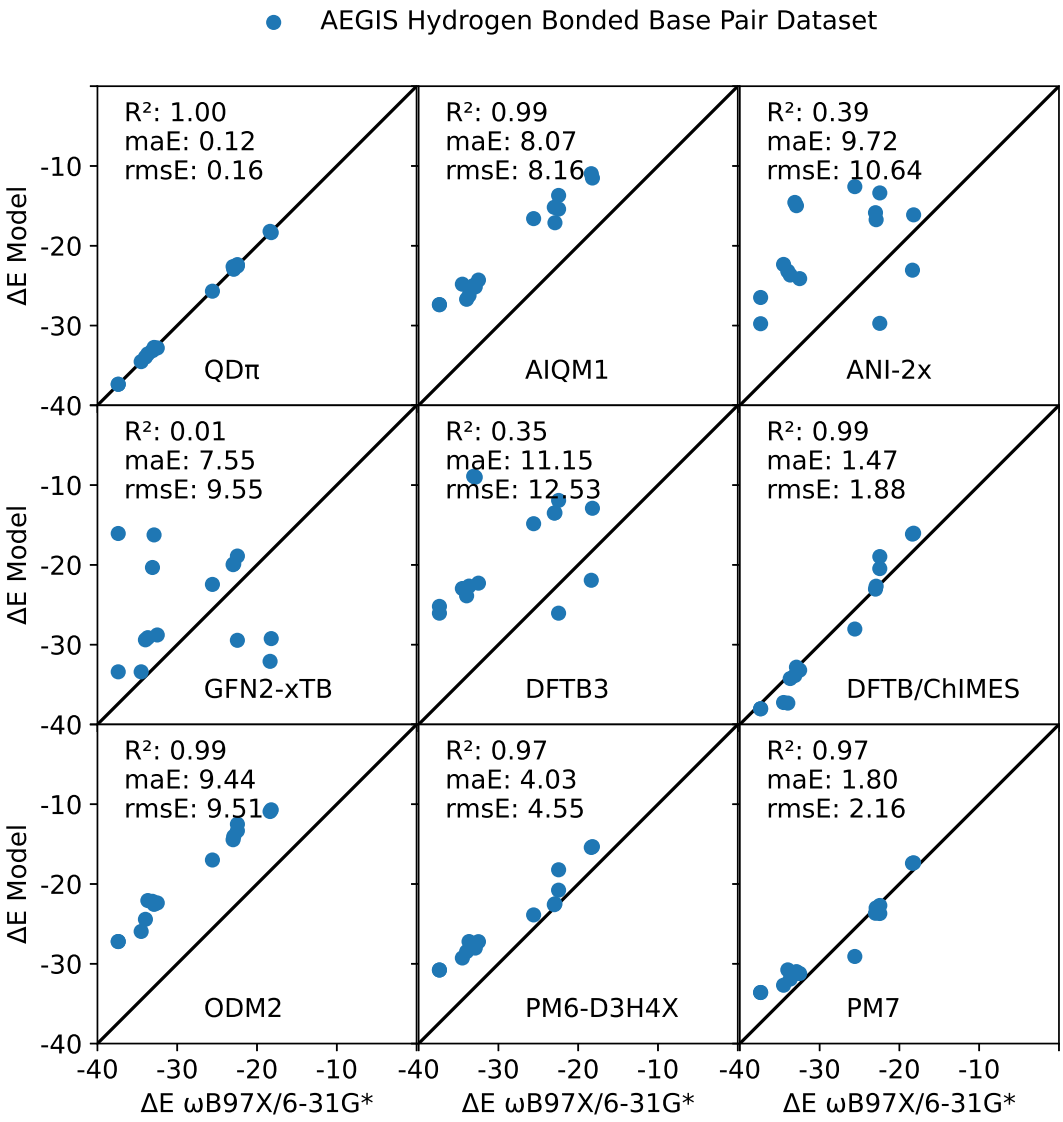


**Figure 5:** Relation between hydrogen bond complex energies calculated by $\omega$B97X/6-31G* and QD$\pi$, AIQM1, ANI-2x, GFN2-xTB, DFTB3, DFTB/ChIMES, ODM2, PM6-D3H4X, and PM7 for the artificially expanded genetic information system (AEGIS) base pair dataset[86], including complexes that involve alternative tautomers and protonation states. Illustrations of each of the complexes is provided in Figure 4. The three base pairs that involve ionized nucleobases are excluded from the regression as they have much larger binding energy values that would artificially skew the correlation.

values of 0.97) and low maE values (1.80 and 4.03 kcal/mol, respectively) for complexes of these neutral nucleobases.

### D. Comparison of tautomer energies for natural and artificial nucleic acids

The artificially expanded genetic information system (AEGIS) dataset also exhibits a rich set of tautomeric forms that have been studied extensively with computational methods.[77,85–87] These tautomeric pairs are illustrated in Figure 6, and their $\Delta$E values are listed in Table III and illustrated in Figure 7. Overall, both QD$\pi$ and AIQM1 give excellent agreement with the $\omega$B97X/6-31G* reference values, with $\Delta$E maE values of 0.71 and 0.77 kcal/mol, respectively, and high correlation ($R^2$ value 0.99). The ANI-2x is the next most accurate, but with errors roughly twice as large (maE 1.41 kcal/mol) and ($R^2$ 0.97). The DFTB/ChIMES and GFN2-xTB models have considerably higher errors (maE 2.20 and 3.16 kcal/mol, respectively) but maintains high correlation with the reference values ($R^2$ value 0.97), whereas DFTB3 and ODM2 have larger errors (maE 5.25 and 4.69 kcal/mol, respectively) and lower correlation ($R^2$ values of 0.61 and 0.85, respectively). The largest errors occur for PM7 and PM6-D3H4X (maE 5.70 and 7.98 kcal/mol, respectively).

It has been estimated that 30% of the compounds in vendor databases and 21% drug databases have potential tautomers.[18,19] For drug discovery applications, it is thus vitally important to be able to model alternative tautomer forms, discern which forms are relevant for ligand-protein binding, and if binding induces a change in tautomer state, to quantitatively determine the tautomerization energy contribution to binding with sub-kcal/mol accuracy. In some cases, the semiempirical QM models incorrectly predict the lowest energy tautomer (1 case for GFN2-xTB, 2 cases for DFTB3, and 9 cases each for ODM2, PM6-D3H4X and PM7). For the models compared here, only QD$\pi$ and AIQM1 are able to achieve the requisite accuracy for quantitative prediction of ligand-protein binding applications.

### E. Comparison of protonation energies for common general acids and bases

Modeling protonation states is important for drug discovery applications as it has been estimated that up to 95% of drug molecules contain ionizable groups[18] (~75% weak bases and ~20% weak acids[20,21]), and protonation states can sometimes change upon ligand binding.

| | | QM/Δ-MLP or MLP | | | DFTB | | | NDDO | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | ωB97X | QDπ | AIQM1 | ANI-2x | GFN2 | DFTB3 | ChIMES | ODM2 | D3H4X | PM7 |
| Tautomer pair | ΔE | Err | Err | Err | Err | Err | Err | Err | Err | Err |
| 1b - 1a | 2.43 | −0.36 | −1.07 | −1.35 | −2.75* | −6.25* | −2.38 | 3.48 | 6.60 | 2.67 |
| 1c - 1b | 17.39 | 0.36 | 0.40 | 0.83 | −0.72 | 2.68 | 2.58 | −6.65 | −11.01 | −7.46 |
| 2b - 2a | −5.41 | −1.12 | 0.13 | −0.29 | 2.81 | −4.39 | −1.02 | 6.31* | 12.62* | 8.85* |
| 2c - 2b | 4.95 | 1.22 | −0.74 | 2.30 | −2.61 | 4.74 | 1.54 | −7.50* | −14.46* | −10.85* |
| 3b - 3a | −6.32 | −0.28 | −0.16 | 0.89 | 3.53 | −3.90 | −0.77 | 6.39* | 12.95* | 9.24* |
| 3c - 3b | 4.05 | −0.15 | −0.15 | 3.05 | −2.48 | 4.44 | 1.02 | −7.07* | −13.56* | −9.97* |
| 4b - 4a | −6.81 | 0.69 | 0.37 | 0.58 | 3.60 | −4.06 | −0.94 | 7.06* | 13.49* | 10.01* |
| 4c - 4b | 2.76 | 0.06 | −0.82 | 1.14 | −1.90 | 5.47 | 1.98 | −6.48* | −12.55* | −8.93* |
| 5b - 5a | −6.12 | −0.41 | 0.66 | −0.28 | 3.07 | −4.60 | −1.43 | 7.34* | 13.36* | 9.71* |
| 5c - 5b | 3.23 | −0.57 | −0.57 | 0.49 | −1.92 | 5.97 | 2.56 | −6.59* | −13.03* | −9.41* |
| 6b - 6a | 12.24 | −0.01 | −0.20 | −2.56 | −1.85 | −3.22 | −3.63 | 3.76 | −4.94 | −0.92 |
| 6c - 6b | 20.12 | −0.10 | −0.35 | 2.63 | −5.80 | −3.10 | −0.26 | −4.60 | −8.47 | −8.48 |
| 7b - 7a | 20.15 | 1.40 | −0.63 | −1.39 | −4.87 | −5.71 | −5.32 | 5.63 | −0.67 | −0.43 |
| 7c - 7b | −19.17 | −0.96 | 0.94 | −0.22 | 4.79 | 6.93 | 1.64 | −0.25 | −0.01 | 2.86 |
| 8b - 8a | 21.32 | 0.84 | −1.79 | −0.08 | −5.43 | −10.25 | −2.12 | 1.57 | 3.71 | 1.24 |
| 8c - 8b | −6.16 | 0.31 | −0.43 | 1.10 | −1.52 | 0.18 | 1.28 | −1.34 | −3.27 | −3.05 |
| 9b - 9a | 5.42 | 0.77 | −0.89 | −0.45 | −5.39 | −7.41* | −4.94 | 4.64 | −3.31 | −3.53 |
| 9c - 9b | −10.02 | −0.74 | 0.40 | 0.29 | 3.20 | 6.10 | 1.29 | −0.76 | −1.38 | 1.91 |
| 10b - 10a | 7.95 | 2.11 | −0.57 | −2.50 | −1.83 | 0.77 | −3.26 | 3.28 | −3.84 | −1.36 |
| 10c - 10b | 22.20 | −2.92 | −2.92 | −5.08 | −6.40 | −15.21 | 0.78 | −9.14 | −11.36 | −9.26 |
| 10d - 10c | 4.01 | 0.04 | 1.54 | 2.86 | −1.71 | 5.13 | −3.27 | 9.90 | 12.36 | 7.71 |
| 11b - 11a | −0.86 | −0.63 | −0.19 | −0.36 | −0.47 | −4.04 | −0.58 | 3.42* | 7.69* | 4.68* |
| 11c - 11b | 24.35 | −0.84 | −0.90 | −4.13 | −4.19 | −5.43 | −6.39 | 1.67 | −8.47 | −6.13 |
| 12b - 12a | 22.00 | 0.41 | −1.80 | 0.14 | −5.51 | −10.37 | −2.16 | 1.34 | 3.73 | 1.31 |
| 12c - 12b | −7.79 | 0.54 | 0.54 | 0.29 | −0.62 | 0.87 | 1.78 | −0.98 | −2.70 | −2.59 |
| maE | ... | 0.71 | 0.77 | 1.41 | 3.16 | 5.25 | 2.20 | 4.69 | 7.98 | 5.70 |
| rmsE | ... | 0.97 | 1.00 | 1.94 | 3.59 | 6.12 | 2.67 | 5.42 | 9.28 | 6.72 |

**Table III:** Tautomerization energies from ωB97X/6-31G* and model errors (kcal/mol) for the artificially expanded genetic information system (AEGIS) tautomer dataset.[a]

[a]Models and datasets are described in the Methods section. Illustrations of each of the tautomerization reactions is provided in Figure 7. Errors corresponding to wrong prediction of more stable tautomer are indicated by an asterisk (*).

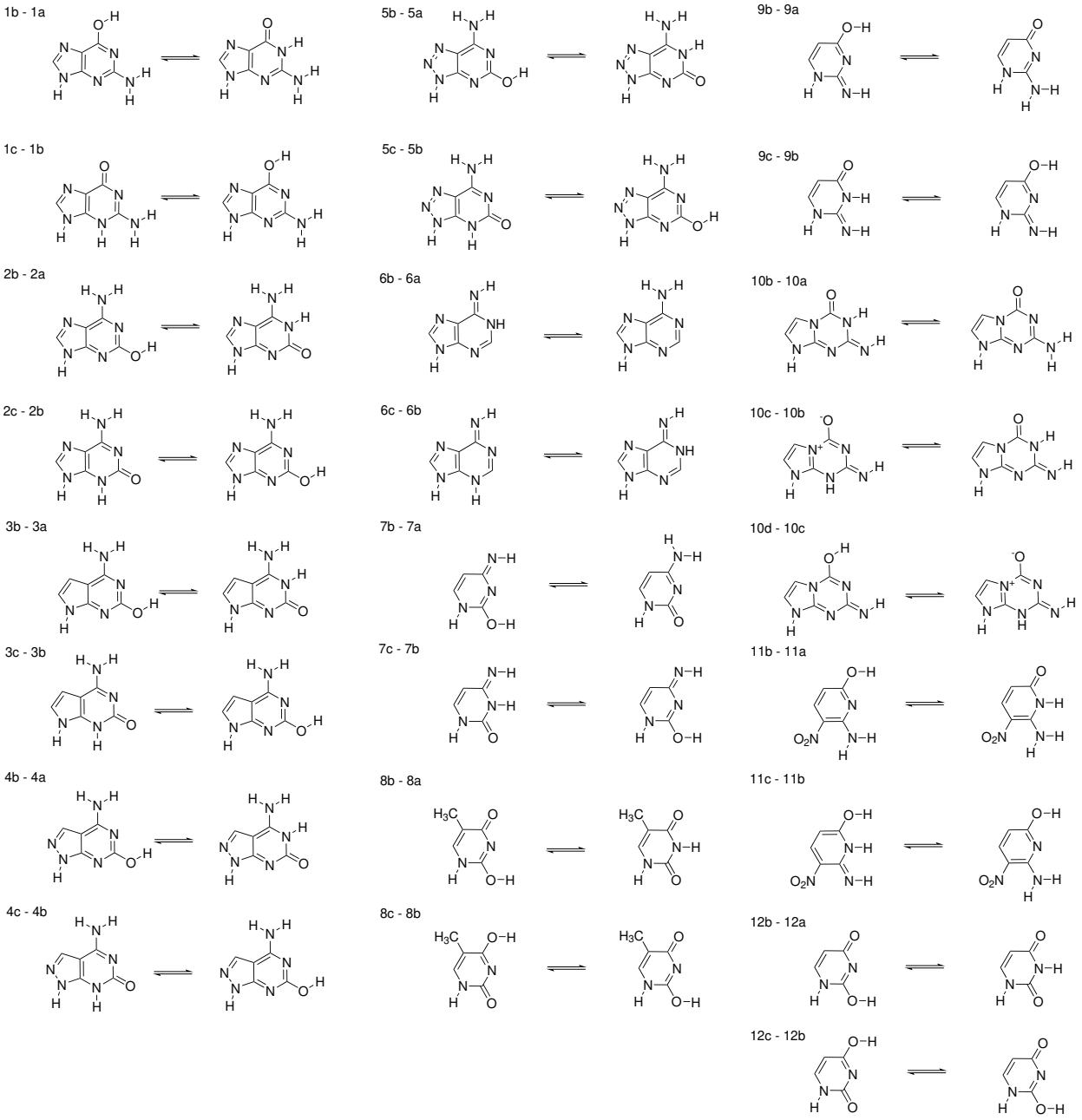


**Figure 6:** Structures for artificially expanded genetic information system (AEGIS) tautomer dataset.[77,85] Guanine derivatives(1-5, 2: nucleobase code B), codes 6: A, 7: C, 8: T, 9: S, 10: P, 11: Z.

Hence, quantitatively accurate modeling of protonation/deprotonation events at these ionizable sites is critical to obtain high predictive capability. As an illustrative set of examples, we examine simple model systems that mimic the acid/base chemistry associated with RNA cleavage reactions catalyzed by small nucleolytic RNA enzymes (ribozymes) and protein

| | | QM/Δ-MLP or MLP | | | DFTB | | | NDDO | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | ωB97X | QDπ | AIQM1 | ANI-2x | GFN2 | DFTB3 | ChIMES | ODM2 | D3H4X | PM7 |
| Protonation pair | ΔE | Err | Err | Err | Err | Err | Err | Err | Err | Err |
| [Lys:$NH_2$,iPrOH] | 167.76 | 0.00 | −0.64 | −115.04 | 0.04 | 6.11 | −6.87 | −15.24 | −13.15 | −10.15 |
| [His:$N_\epsilon$,iPrOH] | 158.33 | 0.08 | −9.22 | −126.62 | −7.02 | −11.33 | −17.47 | −18.74 | −12.96 | −6.19 |
| [$EtO^-$,His:$N_\epsilon H^+$] | −160.25 | −0.02 | 12.82 | 137.70 | 9.66 | 11.71 | 18.09 | 21.92 | 12.24 | 5.36 |
| [G:$N_1{}^-$,iPrOH] | 43.06 | −1.11 | −1.15 | −28.62 | −2.69 | −8.63 | −11.17 | −10.84 | 0.89 | 5.45 |
| [$EtO^-$,A:$N_1H^+$] | −165.06 | 1.25 | 12.94 | 137.24 | 10.02 | 15.21 | 23.15 | 20.74 | 8.07 | 1.25 |
| [$EtO^-$,A:$N_3H^+$] | −190.89 | 1.21 | 12.88 | 143.42 | 11.40 | 16.00 | 24.17 | 19.43 | 17.97 | 8.39 |
| [$EtO^-$,C:$N_3H^+$] | −160.33 | 0.89 | 12.78 | 145.20 | 6.58 | 4.66 | 16.93 | 20.19 | 7.03 | 2.22 |
| [G:$N_1{}^-$,A:$N_1H^+$] | −120.07 | 0.08 | 8.19 | 97.55 | 4.69 | 6.20 | 11.36 | 6.73 | 9.69 | 7.53 |
| [G:$N_1{}^-$,A:$N_3H^+$] | −145.91 | 0.04 | 8.12 | 103.73 | 6.07 | 6.99 | 12.38 | 5.41 | 19.58 | 14.67 |
| [G:$N_1{}^-$,C:$N_3H^+$] | −115.34 | −0.27 | 8.03 | 105.50 | 1.25 | −4.35 | 5.14 | 6.17 | 8.64 | 8.51 |
| maE | ... | 0.50 | 8.68 | 114.06 | 5.94 | 9.12 | 14.67 | 14.54 | 11.02 | 6.97 |
| rmsE | ... | 0.72 | 9.72 | 118.72 | 6.96 | 9.96 | 15.87 | 15.84 | 12.17 | 7.88 |

**Table IV:** Selected relative protonation/deprotonation energies from ωB97X/6-31G* and model error (kcal/mol) relevant to acid/base catalysis in RNA cleavage reactions.[a]

[a]Models and datasets are described in the Methods section. Protonation pairs are written in the general form as follows. [B,A]: B + A → $BH^+$ + $A^-$, or [$B^-$,$AH^+$]: $B^-$ + $AH^+$ → BH + A. Here B/$BH^+$ and $B^-$/BH are base/conjugate acids pairs and A/$A^-$ and $AH^+$/A are acid/conjugate base pairs. These are model systems for general acid and base catalysis in RNA cleavage reactions by small nucleolytic ribozymes and ribonucleases[100]. Molecules indicated are: isopropanol (iPrOH), ethoxide ($EtO^-$), neutral lysine (Lys:$NH_2$), neutral histidine (His:$N_\epsilon$), protonated histidine (His:$N_\epsilon H^+$), deprotonated guanine at the N1 position (G:$N_1{}^-$), protonated adenine at the N1 and N3 positions (A:$N_1H^+$ and A:$N_3H^+$) and protonated cytosine at the N3 position (C:$N_3H^+$).

enzymes (ribonucleases)[100]. In these reactions, the 2'OH of an RNA nucleotide, modeled by the secondary alcohol isopropanol (iPrOH), becomes activated (deprotonated) by a general base that in ribozymes is often an ionized (deprotonated) guanine residue (G:$N_1{}^-$), and in RNase A[143–145] is generally believed to be a histidine (His:$N_\epsilon$) although it has been specu-

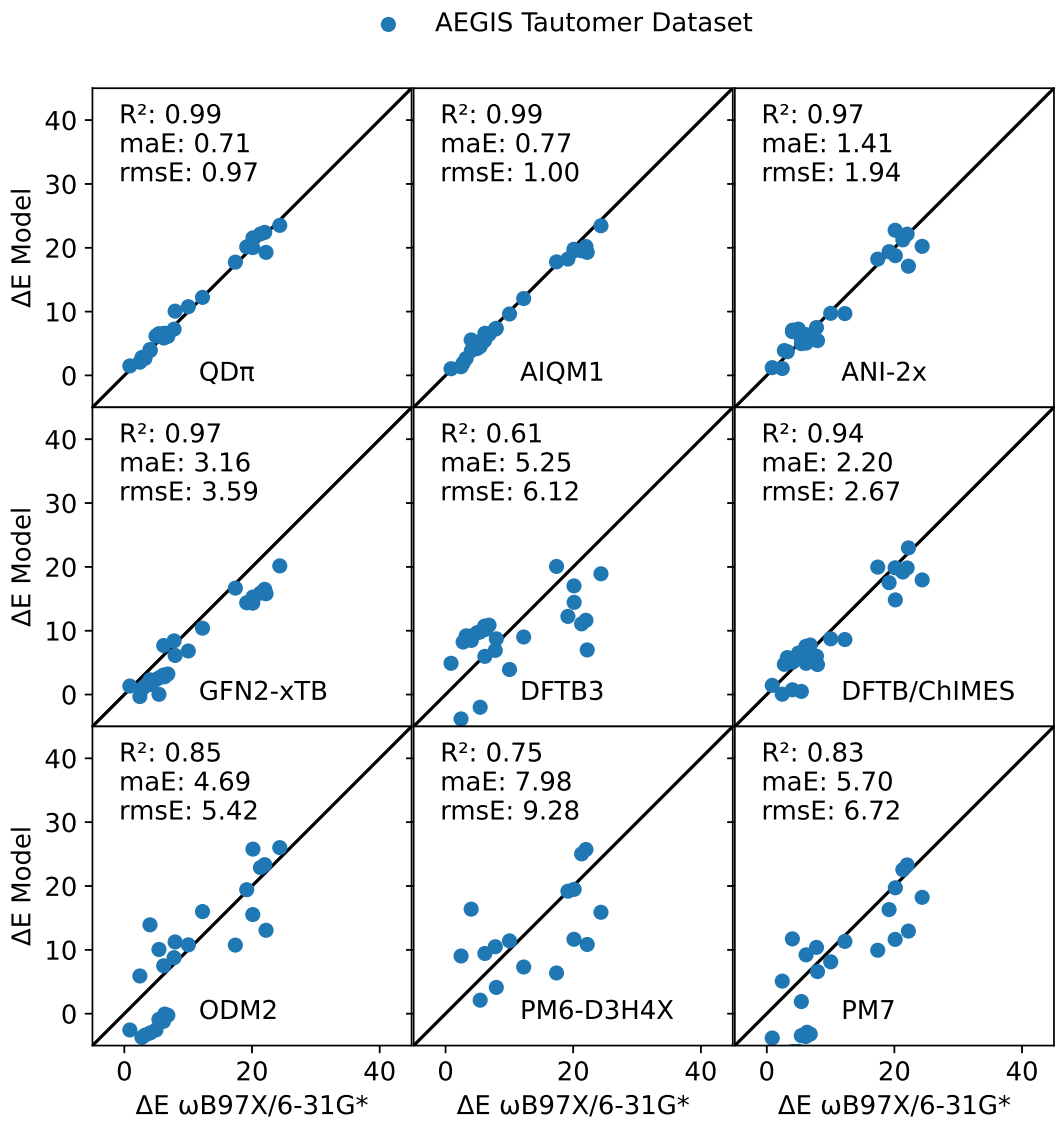


**Figure 7:** Relation between tautomerization energies calculated by ωB97X/6-31G* and QDπ, AIQM1, ANI-2x, GFN2-xTB, DFTB3, DFTB/ChIMES, ODM2, PM6-D3H4X, and PM7 for the artificially expanded genetic information system (AEGIS) tautomer dataset. An illustration of each of the complexes is provided in Figure 6. In the regression plot shown, the sign convention (direction of the tautomer reaction) is chosen such that the reference ΔE value is positive (this is done to circumvent "spreading out" of the data and artificially inflating the correlation).

lated that a neutral lysine (Lys:$NH_2$) might also be capable. The activated nucleophile then attacks the scissile phosphate, passing through a pentavalent transition state, followed by departure of the 5'O leaving group (modeled by the primary alcoxide ethoxide ($EtO^-$) with the assistance of a general acid that in ribozymes can be either a protonated adenine at the N1 or N3 positions (A:$N_1H^+$ and A:$N_3H^+$, respectively) or an ionized (protonated) cytosine (C:$N_3H^+$), and in RNase A is a protonated histidine (His:$N_\epsilon H^+$).

Table IV lists relative protonation/deprotonation reactions that model general acid/base events in RNA cleavage reactions[100]. Overall, QD$\pi$ performs extremely well, with $\Delta$E maE of 0.50 kcal/mol. Of the semiempirical QM methods, GFN2-xTB is the least inaccurate (maE 5.94 kcal/mol) followed by PM7 (6.97 kcal/mol), with other models notably higher (maE values range from 9.12 to 14.67 kcal/mol). As mentioned earlier, the ANI-2x model was not designed to handle ions; it produces errors on the order of 100 kcal/mol. The AIQM1 model is greatly improved with respect to ANI-2x and ODM2 (the base QM model). The QD$\pi$ $\Delta$E maE value is dominated by large positive errors involving the ethoxide and protonated nucleobases (0.89-1.25 kcal/mol). The ethoxide anion is a primary alkoxide that is only marginally stable in the gas phase, and thus especially challenging. The QD$\pi$ model is by far the most accurate for protonation/deprotonation energies. It is a promising candidate for use in drug discovery applications.

## IV. CONCLUSION

We have compared the performance of several NDDO-based (MNDO/d, AM1, PM6, PM6-D3H4X, PM7, and ODM2) and density-functional tight-binding based (DFTB3, DFTB/ChIMES, GFN1-xTB, and GFN2-xTB) semiempirical models with pure machine learning potentials (ANI-1x and ANI-2x) and hybrid quantum mechanical/machine learning potentials (AIQM1 and QD$\pi$). We examine broad datasets computed at a consistent $\omega$B97X/6-31G* level of theory that includes conformational energies, intermolecular interactions, tautomers, and protonation states. The methods were further compared against the AEGIS dataset and acid/base chemistry relevant for RNA cleavage reactions catalyzed by small nucleolytic ribozymes and ribonucleases. Overall, the recently developed QD$\pi$ model performs exceptionally well across all datasets, having especially high accuracy for tautomers and protonation states relevant to drug discovery. The AIQM1 model also has

impressive performance for many cases, including tautomerization energies. All other methods examined have various strengths and weaknesses, but none have the broad range of quantitative accuracy of the QD$\pi$ model for the data examined. Taken together, this suggests that QM/$\Delta$-MLPs such as QD$\pi$ and AIQM1 have considerable promise as universal force fields for drug discovery applications.

## SUPPLEMENTARY MATERIAL

See supplementary material for relative energies for the minima and transition state of the alanine dipeptide, ibuprofen, and ketorolac.

## ACKNOWLEDGMENTS

The authors are grateful for the financial support provided by the National Institutes of Health (No. GM107485 to DMY). 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 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).

## DATA AVAILABILITY

QD$\pi$-v1.0 is openly available in our GitLab repository at https://gitlab.com/RutgersLBSR/qdpi, which was previously released[77]. The data that support the findings of this study are available from the corresponding author upon reasonable request.

## REFERENCES

[1]T.-S. Lee, B. K. Allen, T. J. Giese, Z. Guo, P. Li, C. Lin, T. D. M. Jr., D. A. Pearlman, B. K. Radak, Y. Tao, H.-C. Tsai, H. Xu, W. Sherman, and D. M. York, "Alchemical Binding Free Energy Calculations in AMBER20: Advances and Best Practices for Drug Discovery," J. Chem. Inf. Model. **60**, 5595–5623 (2020).

[2]W. L. Jorgensen, "Efficient drug lead discovery and optimization," Acc. Chem. Res. **42**, 724–733 (2009).

[3]D. J. Cole, J. T. Horton, L. Nelson, and V. Kurdekar, "The future of force fields in computer-aided drug design," Future Med. Chem. **11**, 2359–2363 (2019).

[4]A. D. MacKerell, Jr., "Empirical Force Fields for Biological Macromolecules: Overview and Issues," J. Comput. Chem. **25**, 1584–1604 (2004).

[5]K. Lindorff-Larsen, P. Maragakis, S. Piana, M. P. Eastwood, R. O. Dror, and D. E. Shaw, "Systematic validation of protein force fields against experimental data," PLoS One **7**, 32131 (2012).

[6]S. Piana, J. L. Klepeis, and D. E. Shaw, "Assessing the accuracy of physical models used in protein-folding simulations: quantitative evidence from long molecular dynamics simulations," Curr. Opin. Struct. Biol. **24**, 98–105 (2014).

[7]W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein, "Comparison of simple potential functions for simulating liquid water," J. Chem. Phys. **79**, 926–935 (1983).

[8]H. W. Horn, W. C. Swope, J. W. Pitera, J. D. Madura, T. J. Dick, G. L. Hura, and T. Head-Gordon, "Development of an improved four-site water model for biomolecular simulations: TIP4P-Ew," J. Chem. Phys. **120**, 9665–9678 (2004).

[9]Y. Wu, H. L. Tepper, and G. A. Voth, "Flexible simple point-charge water model with improved liquid-state properties," J. Chem. Phys. **124**, 024503 (2006).

[10]S. Izadi, R. Anandakrishnan, and A. V. Onufriev, "Building Water Models: A Different Approach," J. Phys. Chem. Lett. **5**, 3863–3871 (2014).

[11]S. Izadi and A. V. Onufriev, "Accuracy limit of rigid 3-point water models," J. Chem. Phys. **145**, 074501–074510 (2016).

[12]I. S. Joung and T. E. Cheatham III, "Molecular dynamics simulations of the dynamic and energetic properties of alkali and halide ions using water-model-specific ion parameters," J. Phys. Chem. B **113**, 13279–13290 (2009).

[13]P. Li, B. P. Roberts, D. K. Chakravorty, and K. M. Merz, Jr., "Rational design of Particle Mesh Ewald compatible Lennard-Jones parameters for +2 metal cations in explicit solvent," J. Chem. Theory Comput. **9**, 2733–2748 (2013).

[14]P. Li and K. M. Merz, Jr., "Taking into account the ion-induced dipole interaction in the nonbonded model of ions," J. Chem. Theory Comput. **10**, 289–297 (2014).

[15]P. Li and K. M. Merz, "Metal Ion Modeling Using Classical Mechanics," Chem. Rev. **117**, 1564–1686 (2017).

[16]P. E. M. Lopes, O. Guvench, and A. D. MacKerell, Jr., "Current status of protein force fields for molecular dynamics simulations," Methods Mol. Biol. **1215**, 47–71 (2015).

[17]C. Tian, K. Kasavajhala, K. A. A. Belfon, L. Raguette, H. Huang, A. N. Migues, J. Bickel, Y. Wang, J. Pincay, Q. Wu, and C. Simmerling, "ff19SB: Amino-Acid-Specific Protein Backbone Parameters Trained against Quantum Mechanics Energy Surfaces in Solution," J. Chem. Theory Comput. **16**, 528–552 (2020).

[18]Y. C. Martin, "Experimental and pK prediction aspects of tautomerism of drug-like molecules," Drug Discov. Today. Technol. **27**, 59–64 (2018).

[19]F. Milletti and A. Vulpetti, "Tautomer Preference in PDB Complexes and its Impact on Structure-Based Drug Discovery," J. Chem. Inf. Model. **50**, 1062–1074 (2010).

[20]J. I. Wells, *Pharmaceutical preformulation : the physicochemical properties of drug substances* (Chichester:E.Horwood, U.K., 1988).

[21]C. D. Navo and G. Jiménez-Osés, "Computer Prediction of p$K_a$ Values in Small Molecules and Proteins," ACS Med. Chem. Lett. **12**, 1624–1628 (2021).

[22]W. Thiel, "Semiempirical quantum–chemical methods," WIREs Comput. Mol. Sci. **4**, 145–157 (2014).

[23]K. Kříž and J. Řezáč, "Benchmarking of Semiempirical Quantum-Mechanical Methods on Systems Relevant to Computer-Aided Drug Design," J. Chem. Inf. Model. **60**, 1453–1460 (2020).

[24]V. Khanna and S. Ranganathan, "Physicochemical property space distribution among human metabolites, drugs and toxins," BMC Bioinformatics **10**, S10 (2009).

[25]T. Darden, D. York, and L. Pedersen, "Particle mesh Ewald: An N log(N) method for Ewald sums in large systems," J. Chem. Phys. **98**, 10089–10092 (1993).

[26]K. Nam, J. Gao, and D. M. York, "An efficient linear-scaling Ewald method for long-range electrostatic interactions in combined QM/MM calculations," J. Chem. Theory Comput. **1**, 2–13 (2005).

[27]T. J. Giese, M. T. Panteva, H. Chen, and D. M. York, "Multipolar Ewald methods, 2: Applications using a quantum mechanical force field," J. Chem. Theory Comput. **11**, 451–461 (2015).

[28]T. J. Giese and D. M. York, "Ambient-Potential Composite Ewald Method for ab Initio Quantum Mechanical/Molecular Mechanical Molecular Dynamics Simulation," J. Chem. Theory Comput. **12**, 2611–2632 (2016).

[29]J. T. Margraf, M. Hennemann, and T. Clark, "EMPIRE: a highly parallel semiempirical molecular orbital program: 3: Born-Oppenheimer molecular dynamics," J. Mol. Model. **26**, 43 (2020).

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

[31]T. J. Giese and D. M. York, "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. **15**, 5543–5562 (2019).

[32]F. L. Kearns, P. S. Hudson, H. L. Woodcock, and S. Boresch, "Computing converged free energy differences between levels of theory via nonequilibrium work methods: Challenges and opportunities," J. Comput. Chem. **38**, 1376–1388 (2017).

[33]S. Boresch and H. L. Woodcock, "Convergence of single-step free energy perturbation," Mol. Phys. **115**, 1200–1213 (2017).

[34]P. S. Hudson, F. Aviat, R. Meana-Pañeda, L. Warrensford, B. C. Pollard, S. Prasad, M. R. Jones, H. L. Woodcock, and B. R. Brooks, "Obtaining QM/MM binding free energies in the SAMPL8 drugs of abuse challenge: indirect approaches," J. Comput.-Aided Mol. Des. **36**, 263–277 (2022).

[35]A. Schöller, F. Kearns, H. L. Woodcock, and S. Boresch, "Optimizing the Calculation of Free Energy Differences in Nonequilibrium Work SQM/MM Switching Simulations," J. Phys. Chem. B **126**, 2798–2811 (2022).

[36]G. Arya, "Models for recovering the energy landscape of conformational transitions from single-molecule pulling experiments," Molecular Simulation **42**, 1102–1115 (2016).

[37]A. N. Naganathan, U. Doshi, and V. Muñoz, "Protein Folding Kinetics: Barrier Effects in Chemical and Thermal Denaturation Experiments," J. Am. Chem. Soc. **129**, 5673–82 (2007).

[38]J. Basran, S. Patel, M. J. Sutcliffe, and N. S. Scrutton, "Importance of Barrier Shape in Enzyme-catalyzed Reactions," J. Biol. Chem. **276**, 6234–42 (2001).

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

[40]K. T. Butler, D. W. Davies, H. Cartwright, O. Isayev, and A. Walsh, "Machine learning for molecular and materials science," Nature **559**, 547–555 (2018).

[41]F. Noé, A. Tkatchenko, K.-R. Müller, and C. Clementi, "Machine Learning for Molecular Simulation," Annu. Rev. Phys. Chem. **71**, 361–390 (2020).

[42]M. Pinheiro Jr, F. Ge, N. Ferré, P. O. Dral, and M. Barbatti, "Choosing the right molecular machine learning potential," Chem. Sci. **12**, 14396–14413 (2021).

[43]S. Manzhos and T. Carrington Jr, "Neural Network Potential Energy Surfaces for Small Molecules and Reactions," Chem. Rev. **121**, 10187–10217 (2021).

[44]J. Zeng, L. Cao, and T. Zhu, "Neural network potentials," in *Quantum Chemistry in the Age of Machine Learning*, edited by P. O. Dral (Elsevier, 2022) Chap. 12, pp. 279–294.

[45]X. Pan, J. Yang, R. Van, E. Epifanovsky, J. Ho, J. Huang, J. Pu, Y. Mei, K. Nam, and Y. Shao, "Machine-Learning-Assisted Free Energy Simulation of Solution-Phase and Enzyme Reactions," J. Chem. Theory Comput. **17**, 5745–5758 (2021).

[46]P. Zheng, R. Zubatyuk, W. Wu, O. Isayev, and P. O. Dral, "Artificial intelligence-enhanced quantum chemical method with broad applicability," Nat. Commun. **12**, 7022 (2021).

[47]J. Zeng, T. J. Giese, Ş. Ekesan, and D. M. York, "Development of Range-Corrected Deep Learning Potentials for Fast, Accurate Quantum Mechanical/Molecular Mechanical Simulations of Chemical Reactions in Solution," J. Chem. Theory Comput. **17**, 6993–7009 (2021).

[48]T. J. Giese, J. Zeng, Ş. Ekesan, and D. M. York, "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. **18**, 4304–4317 (2022).

[49]C. L. Gómez-Flores, D. Maag, M. Kansari, V.-Q. Vuong, S. Irle, F. Gräter, T. Kubař, and M. Elstner, "Accurate Free Energies for Complex Condensed-Phase Reactions Us-

ing an Artificial Neural Network Corrected DFTB/MM Methodology," J. Chem. Theory Comput. **18**, 1213–1226 (2022).

[50]J. Böser, T. Kubař, M. Elstner, and D. Maag, "Reduction pathway of glutaredoxin 1 investigated with QM/MM molecular dynamics using a neural network correction," J. Chem. Phys. **157**, 154104 (2022).

[51]P. O. Dral, T. Zubatiuk, and B.-X. Xue, "Learning from multiple quantum chemical methods: Δ-learning, transfer learning, co-kriging, and beyond," in *Quantum Chemistry in the Age of Machine Learning*, edited by P. O. Dral (Elsevier, 2022) Chap. 21, pp. 491–507.

[52]J. Behler and M. Parrinello, "Generalized Neural-Network Representation of High-Dimensional Potential-Energy Surfaces," Phys. Rev. Lett. **98**, 146401–146404 (2007).

[53]A. P. Bartók, M. C. Payne, R. Kondor, and G. Csányi, "Gaussian Approximation Potentials: The Accuracy of Quantum Mechanics, without the Electrons," Phys. Rev. Lett. **104**, 136403 (2010).

[54]J. Behler, "Atom-centered Symmetry Functions for Constructing High-dimensional Neural Network Potentials," J. Chem. Phys. **134**, 074106 (2011).

[55]M. Gastegger, L. Schwiedrzik, M. Bittermann, F. Berzsenyi, and P. Marquetand, "wACSF—Weighted atom-centered symmetry functions as descriptors in machine learning potentials," J. Chem. Phys. **148**, 241709 (2018).

[56]S. Chmiela, A. Tkatchenko, H. E. Sauceda, I. Poltavsky, K. T. Schütt, and K.-R. Müller, "Machine learning of accurate energy-conserving molecular force fields," Sci. Adv. **3**, 1603015 (2017).

[57]K. T. Schütt, F. Arbabzadah, S. Chmiela, K. R. Müller, and A. Tkatchenko, "Quantum-chemical insights from deep tensor neural networks," Nat. Commun. **8**, 13890 (2017).

[58]K. Schütt, H. Sauceda, P. Kindermans, A. Tkatchenko, and K. Müller, "SchNet - A Deep Learning Architecture for Molecules and Materials," J. Chem. Phys. **148**, 241722 (2018).

[59]X. Chen, M. S. Jørgensen, J. Li, and B. Hammer, "Atomic Energies from a Convolutional Neural Network," J. Chem. Theory Comput. **14**, 3933–3942 (2018).

[60]L. Zhang, J. Han, H. Wang, R. Car, and W. E, "Deep potential molecular dynamics: a scalable model with the accuracy of quantum mechanics," Phys. Rev. Lett. **120**, 143001 (2018).

[61]L. Zhang, J. Han, H. Wang, W. Saidi, R. Car, and W. E, "End-to-end symmetry preserving inter-atomic potential energy model for finite and extended systems," in *Advances in Neural Information Processing Systems 31*, edited by S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (Curran Associates, Inc., 2018) pp. 4436–4446.

[62]Y. Zhang, C. Hu, and B. Jiang, "Embedded Atom Neural Network Potentials: Efficient and Accurate Machine Learning with a Physically Inspired Representation," J. Phys. Chem. Lett. **10**, 4962–4967 (2019).

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

[64]O. Unke and M. Meuwly, "PhysNet: A Neural Network for Predicting Energies, Forces, Dipole Moments, and Partial Charges," J. Chem. Theory Comput. **15**, 3678–3693 (2019).

[65]Z. L. Glick, D. P. Metcalf, A. Koutsoukas, S. A. Spronk, D. L. Cheney, and C. D. Sherrill, "AP-Net: An atomic-pairwise neural network for smooth and transferable interaction potentials," J. Chem. Phys. **153**, 044112 (2020).

[66]T. Zubatiuk and O. Isayev, "Development of Multimodal Machine Learning Potentials: Toward a Physics-Aware Artificial Intelligence," Acc. Chem. Res. **54**, 1575–1585 (2021).

[67]E. R. Khajehpasha, J. A. Finkler, T. D. Kühne, and S. A. Ghasemi, "CENT2: Improved charge equilibration via neural network technique," Phys. Rev. B **105**, 144106 (2022).

[68]Y. Hirano, N. Okimoto, S. Fujita, and M. Taiji, "Molecular Dynamics Study of Conformational Changes of Tankyrase 2 Binding Subsites upon Ligand Binding," ACS Omega **6**, 17609–17620 (2021).

[69]P. W. Kenny, "Hydrogen-Bond Donors in Drug Design," J. Med. Chem. **65**, 14261–14275 (2022).

[70]H. Yuki, Y. Tanaka, M. Hata, H. Ishikawa, S. Neya, and T. Hoshino, "Implementation of $\pi$-$\pi$ Interactions in Molecular Dynamics Simulation," J. Comput. Chem. **28**, 1091–1099 (2007).

[71]T. Chen, M. Li, and J. Liu, "$\pi$–$\pi$ Stacking Interaction: A Nondestructive and Facile Means in Material Engineering for Bioapplications," Cryst. Growth Des. **18**, 2765–2783 (2018).

[72]M. Mohebifar, E. R. Johnson, and C. N. Rowley, "Evaluating Force-Field London Dispersion Coefficients Using the Exchange-Hole Dipole Moment Model," J. Chem. Theory Comput. **13**, 6146–6157 (2017).

[73]L. I. Vazquez-Salazar, E. D. Boittier, O. T. Unke, and M. Meuwly, "Impact of the Characteristics of Quantum Chemical Databases on Machine Learning Prediction of Tautomerization Energies," J. Chem. Theory Comput. **17**, 4769–4785 (2021).

[74]J. S. Smith, B. Nebgen, N. Lubbers, O. Isayev, and A. E. Roitberg, "Less is more: Sampling chemical space with active learning," J. Chem. Phys. **148**, 241733–241743 (2018).

[75]J. Smith, B. Nebgen, R. Zubatyuk, N. Lubbers, C. Devereux, K. Barros, S. Tretiak, O. Isayev, and A. Roitberg, "Approaching Coupled Cluster Accuracy with a General-purpose Neural Network Potential Through Transfer Learning," Nat. Commun. **10**, 2903 (2019).

[76]C. Devereux, J. S. Smith, K. K. Huddleston, K. Barros, R. Zubatyuk, O. Isayev, and A. E. Roitberg, "Extending the Applicability of the ANI Deep Learning Molecular Potential to Sulfur and Halogens," J. Chem. Theory Comput. **16**, 4192–4202 (2020).

[77]J. Z. amd Yujun Tao, T. J. Giese, and D. M. York, "QD$\pi$: A Quantum Deep Potential Interaction Model for Drug Discovery," J. Chem. Theory Comput. (2023), 10.1021/acs.jctc.2c01172.

[78]P. O. Dral, X. Wu, and W. Thiel, "Semiempirical Quantum-Chemical Methods with Orthogonalization and Dispersion Corrections," J. Chem. Theory Comput. **15**, 1743–1760 (2019).

[79]Y. Chen, Y. Ou, P. Zheng, Y. Huang, F. Ge, and P. O. Dral, "Benchmark of General-Purpose Machine Learning-Based Quantum Mechanical Method AIQM1 on Reaction Barrier Heights," J. Chem. Phys. (2023), 10.1063/5.0137101.

[80]Y. Yang, H. Yu, D. M. York, Q. Cui, and M. Elstner, "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 **111**, 10861–10873 (2007).

[81]M. Gaus, X. Lu, M. Elstner, and Q. Cui, "Parameterization of DFTB3/3OB for Sulfur and Phosphorus for Chemical and Biological Applications," J. Chem. Theory Comput. **10**, 1518–1537 (2014).

[82]H. Wang, L. Zhang, J. Han, and W. E, "DeePMD-kit: A deep learning package for many-body potential energy representation and molecular dynamics," Comput. Phys. Commun. **228**, 178–184 (2018).

[83]W. Liang, J. Zeng, D. M. York, L. Zhang, and H. Wang, "Learning deepmd-kit: A guide to building deep potential models," in *A Practical Guide to Recent Advances in Multiscale Modeling and Simulation of Biomolecules*, edited by Y. Wang and R. Zhou (AIP Publishing, 2023) Chap. Chapter 6, pp. 6–1–6–20.

[84]Z. Yang, D. Hutter, P. Sheng, A. M. Sismour, and S. A. Benner, "Artificially expanded genetic information system: a new base pair with an alternative hydrogen bonding pattern," Nucleic Acids Res. **34**, 6095–6101 (2006).

[85]L. Eberlein, F. R. Beierlein, N. J. R. van Eikema Hommes, A. Radadiya, J. Heil, S. A. Benner, T. Clark, S. M. Kast, and N. G. J. Richards, "Tautomeric Equilibria of Nucleobases in the Hachimoji Expanded Genetic Alphabet," J. Chem. Theory Comput. **16**, 2766–2777 (2020).

[86]E. Biondi and S. A. Benner, "Artificially Expanded Genetic Information Systems for New Aptamer Technologies," Biomedicines **6**, 53 (2018).

[87]B. Behera, P. Das, and N. R. Jena, "Accurate Base Pair Energies of Artificially Expanded Genetic Information Systems (AEGIS): Clues for Their Mutagenic Characteristics," J. Phys. Chem. B **123**, 6728–6739 (2019).

[88]C. A. Jerome, S. Hoshika, K. M. Bradley, S. A. Benner, and E. Biondi, "In vitro evolution of ribonucleases from expanded genetic alphabets," Proc. Natl. Acad. Sci. USA **119**, e2208261119 (2022).

[89]M. J. S. Dewar and W. Thiel, "A semiempirical model for the two-Center repulsion integrals in the NDDO approximation," Theor. Chim. Acta **46**, 89–104 (1977).

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

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

[92]J. Řezáč and P. Hobza, "Advanced Corrections of Hydrogen Bonding and Dispersion for Semiempirical Quantum Mechanical Methods," J. Chem. Theory Comput. **8**, 141–151

(2012).

[93]J. Řezáč and P. Hobza, "A halogen-bonding correction for the semiempirical PM6 method," Chem. Phys. Lett. **506**, 286–289 (2011).

[94]J. J. P. Stewart, "Optimization of parameters for semiempirical methods VI: more modifications to the NDDO approximations and re-optimization of parameters," J. Mol. Model. **19**, 1–32 (2013).

[95]M. Gaus, Q. Cui, and M. Elstner, "DFTB3: Extension of the seld-consistent-charge density-functional tight-binding method (SCC-DFTB)," J. Chem. Theory Comput. **7**, 931–948 (2011).

[96]S. Grimme, C. Bannwarth, and P. Shushkov, "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. **13**, 1989–2009 (2017).

[97]C. Bannwarth, S. Ehlert, and S. Grimme, "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. **15**, 1652–1671 (2019).

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

[99]J.-D. Chai and M. Head-Gordon, "Systematic optimization of long-range corrected hybrid density functionals," J. Chem. Phys. **128**, 084106 (2008).

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

[101]M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, G. A. Petersson, H. Nakatsuji, X. Li, M. Caricato, A. V. Marenich, J. Bloino, B. G. Janesko, R. Gomperts, B. Mennucci, H. P. Hratchian, J. V. Ortiz, A. F. Izmaylov, J. L. Sonnenberg, D. Williams-Young, F. Ding, F. Lipparini, F. Egidi, J. Goings, B. Peng, A. Petrone, T. Henderson, D. Ranasinghe, V. G. Zakrzewski, J. Gao, N. Rega, G. Zheng, W. Liang, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa,

M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, K. Throssell, J. A. Montgomery, Jr., J. E. Peralta, F. Ogliaro, M. J. Bearpark, J. J. Heyd, E. N. Brothers, K. N. Kudin, V. N. Staroverov, T. A. Keith, R. Kobayashi, J. Normand, K. Raghavachari, A. P. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, J. M. Millam, M. Klene, C. Adamo, R. Cammi, J. W. Ochterski, R. L. Martin, K. Morokuma, O. Farkas, J. B. Foresman, and D. J. Fox, "Gaussian˜16 Revision A.03," (2016), gaussian Inc. Wallingford CT.

[102]D. L. Beveridge, *Approximate Molecular Orbital Theory of Nuclear and Electron Magnetic Resonance Parameters* (Springer, Boston, 1977).

[103]T. Tuttle and W. Thiel, "OMx-D: semiempirical methods with orthogonalization and dispersion corrections. Implementation and biochemical application," Phys. Chem. Chem. Phys. **10**, 2159–2166 (2008).

[104]D. Tuna, Y. Lu, A. Koslowski, and W. Thiel, "Semiempirical Quantum-Chemical Orthogonalization-Corrected Methods: Benchmarks of Electronically Excited States," J. Chem. Theory Comput. **12**, 4400–4422 (2016), pMID: 27380455, http://dx.doi.org/10.1021/acs.jctc.6b00403.

[105]P. O. Dral, X. Wu, L. Spörkel, A. Koslowski, W. Weber, R. Steiger, M. Scholten, and W. Thiel, "Semiempirical Quantum-Chemical Orthogonalization-Corrected Methods: Theory, Implementation, and Parameters," J. Chem. Theory Comput. **12**, 1082–1096 (2016), pMID: 26771204, http://dx.doi.org/10.1021/acs.jctc.5b01046.

[106]P. O. Dral, X. Wu, L. Spörkel, A. Koslowski, and W. Thiel, "Semiempirical Quantum-Chemical Orthogonalization-Corrected Methods: Benchmarks for Ground-State Properties," J. Chem. Theory Comput. **12**, 1097–1120 (2016), pMID: 26771261, http://dx.doi.org/10.1021/acs.jctc.5b01047.

[107]D. A. Case, K. Belfon, I. Y. Ben-Shalom, S. R. Brozell, D. S. Cerutti, T. E. Cheatham III, V. W. D. Cruzeiro, T. A. Darden, R. E. Duke, G. Giambasu, , M. K. Gilson, H. Gohlke, A. W. Goetz, R. Harris, S. Izadi, S. A. Izmailov, K. Kasavajhala, K. Kovalenko, R. Krasny, T. Kurtzman, T. Lee, S. Le-Grand, P. Li, C. Lin, J. Liu, T. Luchko, R. Luo, V. Man, K. Merz, Y. Miao, O. Mikhailovskii, G. Monard, , H. Nguyen, A. Onufriev, F. Pan, S. Pantano, R. Qi, D. R. Roe, A. Roitberg, C. Sagui, S. Schott-Verdugo, J. Shen, C. L. Simmerling, N. Skrynnikov, J. Smith, J. Swails, R. C. Walker, J. Wang, R. M. Wilson, R. M. Wolf, X. Wu, Y. Xiong, Y. Xue, D. M. York, and P. A. Kollman, *AMBER 20*,

University of California, San Francisco, San Francisco, CA (2020).

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

[109]M. Thiel, "Mndo," (2022), max-Planck-Institut für Kohlenforschung, Mülheim an der Ruhr.

[110]J. J. P. Stewart, "MOPAC: A semiempirical molecular orbital program," J. Comput.-Aided Mol. Des. **4**, 1–105 (1990).

[111]J. Hostaš, J. Řezáč, and P. Hobza, "On the performance of the semiempirical quantum mechanical PM6 and PM7 methods for noncovalent interactions," Chem. Phys. Lett. **568**, 161–166 (2013).

[112]A. V. Sulimov, D. C. Kutov, E. V. Katkova, I. S. Ilin, and V. B. Sulimov, "New generation of docking programs: Supercomputer validation of force fields and quantum-chemical methods for docking," J. Mol. Graph. Model. **78**, 139–147 (2017).

[113]T. J. Giese and D. M. York, "Density-functional expansion methods: grand challenges," Theor. Chem. Acc. **131**, 1145 (2012).

[114]M. Gaus, A. Goez, and M. Elstner, "Parametrization and Benchmark of DFTB3 for Organic Molecules," J. Chem. Theory Comput. **9**, 338–354 (2013).

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

[116]C. H. Pham, R. K. Lindsey, L. E. Fried, and N. Goldman, "High-Accuracy Semiempirical Quantum Models Based on a Minimal Training Set," J. Phys. Chem. Lett. **13**, 2934–2942 (2022).

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

[118]H. Li, C. Collins, M. Tanha, G. J. Gordon, and D. J. Yaron, "A Density Functional Tight Binding Layer for Deep Learning of Chemical Hamiltonians," J. Chem. Theory Comput. **14**, 5764–5776 (2018).

[119]J. Zhu, V. Q. Vuong, B. G. Sumpter, and S. Irle, "Artificial neural network correction for density-functional tight- binding molecular dynamics simulations," MRS Communications **9**, 867–873 (2019).

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

[121]C. Panosetti, A. Engelmann, L. Nemec, K. Reuter, and J. T. Margraf, "Learning to Use the Force: Fitting Repulsive Potentials in Density- Functional Tight-Binding with Gaussian Process Regression," J. Chem. Theory Comput. **16**, 2181–2191 (2020).

[122]J. J. Kranz, M. Kubillus, R. Ramakrishnan, O. A. von Lilienfeld, and M. Elstner, "Generalized Density-Functional Tight-Binding Repulsive Potentials from Unsupervised Machine Learning," J. Chem. Theory Comput. **14**, 2341–2352 (2018).

[123]X. Gao, F. Ramezanghorbani, O. Isayev, J. S. Smith, and A. E. Roitberg, "TorchANI: A Free and Open Source PyTorch-Based Deep Learning Implementation of the ANI Neural Network Potentials," J. Chem. Inf. Model. **60**, 3408–3415 (2020).

[124]E. Caldeweyher, C. Bannwarth, and S. Grimme, "Extension of the D3 dispersion coefficient model," J. Chem. Phys. **147**, 034112 (2017).

[125]D. C. Liu and J. Nocedal, "On the limited memory BFGS method for large scale optimization," Mathematical Programming **45**, 503–528 (1989).

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

[127]J. S. Smith, R. Zubatyuk, B. Nebgen, N. Lubbers, K. Barros, A. E. Roitberg, O. Isayev, and S. Tretiak, "The ANI-1ccx and ANI-1x data sets, coupled-cluster and density functional theory properties for molecules," Sci Data **7**, 134 (2020).

[128]T. Fink, H. Bruggesser, and J.-L. Reymond, "Virtual Exploration of the Small-Molecule Chemical Universe below 160 Daltons," Angew. Chem. Int. Ed. **44**, 1504–8 (2005).

[129]L. C. Blum and J.-L. Reymond, "970 Million Druglike Small Molecules for Virtual Screening in the Chemical Universe Database GDB-13," J. Am. Chem. Soc. **131**, 8732–8733 (2009).

[130]V. Law, C. Knox, Y. Djoumbou, T. Jewison, A. C. Guo, Y. Liu, A. Maciejewski, D. Arndt, M. Wilson, V. Neveu, A. Tang, G. Gabriel, C. Ly, S. Adamjee, Z. T. Dame, B. Han, Y. Zhou, and D. S. Wishart, "DrugBank 4.0: shedding new light on drug metabolism," Nucleic Acids Res. **42**, 1091–1097 (2014).

[131]L. Goerigk, H. Kruse, and S. Grimme, "Benchmarking Density Functional Methods against the S66 and S66x8 Datasets for Non-Covalent Interactions," Chem. Phys. Chem. **12**, 3421–33 (2011).

[132]B. Brauer, M. K. Kesharwani, S. Kozuch, and J. M. L. Martin, "The S66x8 benchmark for noncovalent interactions revisited: explicitly correlated ab initio methods and density functional theory," Phys. Chem. Chem. Phys. **18**, 20905–25 (2016).

[133]J. Řezáč, "Non-Covalent Interactions Atlas Benchmark Data Sets: Hydrogen Bonding," J. Chem. Theory Comput. **16**, 2355–2368 (2020).

[134]O. Wahl and T. Sander, "Tautobase: An Open Tautomer Database," J. Chem. Inf. Model. **60**, 1085–1089 (2020).

[135]M. Wieder, J. Fass, and J. D. Chodera, "Fitting quantum machine learning potentials to experimental free energy data: predicting tautomer ratios in solution," Chem. Sci. **12**, 11364–11381 (2021).

[136]A. Moser, K. Range, and D. M. York, "Accurate Proton Affinity and Gas-Phase Basicity Values for Molecules Important in Biocatalysis," J. Phys. Chem. B **114**, 13911–13921 (2010).

[137]L. Goerigk, A. Hansen, C. Bauer, S. Ehrlich, A. Najibi, and S. Grimme, "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. **19**, 32184–32215 (2017).

[138]N. Ree, A. H. Göller, and J. H. Jensen, "RegioSQM20: improved prediction of the regioselectivity of electrophilic aromatic substitutions," J. Cheminform. **13**, 10 (2021).

[139]N. B. Leontis and E. Westhof, "Geometric nomenclature and classification of RNA base pairs," RNA **7**, 499–512 (2001).

[140]N. B. Leontis, J. Stombaugh, and E. Westhof, "The non-watson-crick base pairs and their associated isostericity matrices," Nucleic Acids Res. **30**, 3497–3531 (2002).

[141]N. B. Leontis and E. Westhof, "Analysis of RNA motifs," Curr. Opin. Struct. Biol. **13**, 300–308 (2003).

[142]I. Singh, M.-J. Kim, R. W. Molt, S. Hoshika, S. A. Benner, and M. M. Georgiadis, "Structure and Biophysics for a Six Letter DNA Alphabet that Includes Imidazo[1,2-a]-1,3,5-triazine-2(8H)-4(3H)-dione (X) and 2,4-Diaminopyrimidine (K)," ACS Synth. Biol. **6**, 2118–2129 (2017).

[143]R. T. Raines, "Ribonuclease A," Chem. Rev. **98**, 1045–1066 (1998).

[144]H. Gu, S. Zhang, K.-Y. Wong, B. K. Radak, T. Dissanayake, D. L. Kellerman, Q. Dai, M. Miyagi, V. E. Anderson, D. M. York, J. A. Piccirilli, and M. E. Harris, "Experimental and computational analysis of the transition state for ribonuclease A-catalyzed RNA 2'-*O*-transphosphorylation," Proc. Natl. Acad. Sci. USA **110**, 13002–13007 (2013).

[145]M. E. Harris, J. A. Piccirilli, and D. M. York, "Integration of kinetic isotope effect analyses to elucidate ribonuclease mechanism," Biochim. Biophys. Acta **1854**, 1801–1808 (2015).