JCTC Journal of Chemical Theory and Computation

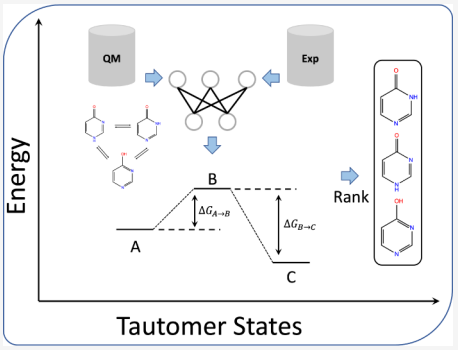
pubs.acs.org/JCTC Article

Fast and Accurate Prediction of Tautomer Ratios in Aqueous Solution via a Siamese Neural Network

Xiaolin Pan, Xudong Zhang, Song Xia, and Yingkai Zhang*

Cite This: J. Chem. Theory Comput. 2025, 21, 3132−3141 Read Online

ABSTRACT: Tautomerization plays a critical role in chemical and biological processes, influencing molecular stability, reactivity, biological activity, and ADME-Tox properties. Many drug-like molecules exist in multiple tautomeric states in aqueous solution, complicating the study of protein−ligand interactions. Rapid and accurate prediction of tautomer ratios and identification of predominant species are therefore crucial in computational drug discovery. In this study, we introduce sPhysNet-Taut, a deep learning model fine-tuned on experimental data using a Siamese neural network architecture. This model directly predicts tautomer ratios in aqueous solution based on MMFF94-optimized molecular geometries. On experimental test sets, sPhysNet-Taut achieves state-of-the-art performance with root-mean-square error (RMSE) of 1.9 kcal/mol on the 100-tautomers set and 1.0 kcal/mol on the SAMPL2 challenge, outperforming all other methods. It also provides superior ranking power for tautomer pairs on multiple test sets. Our results demonstrate that fine-tuning on experimental data significantly enhances model performance compared to training from scratch. This work not only offers a valuable deep learning model for predicting tautomer ratios but also presents a protocol for modeling pairwise data. To promote usability, we have developed an accessible tool that predicts stable tautomeric states in aqueous solution by enumerating all possible tautomeric states and ranking them using our model. The source code and web server are freely accessible at https://github.com/xiaolinpan/sPhysNet-Taut and https://yzhang.hpc.nyu.edu/tautomer.



INTRODUCTION

Tautomeric equilibrium plays a crucial role in various chemical and biological processes, affecting molecular stability, reactivity, and biological activity. In drug discovery, studying tautomeric equilibrium is essential because many drug-like compounds exhibit heterocyclic structures with potential tautomeric transformations.[1,2] Notably, approximately 26% of approved drugs exist in different tautomeric states.[3] Tautomeric interconversions mainly involve three types: prototropic tautomerism, ring−chain tautomerism, and valence tautomerism. Prototropic tautomerism is the most common type in drug molecules, involving bond reformation and proton transfer. This transformation interchanges pharmacophore types with hydrogen bond donors becoming acceptors and vice versa, which alters the interaction between proteins and ligands.[1,3] Some studies have discussed the effects of tautomerization on structure-based and ligand-based screening methods, where high-energy tautomer states may form different interactions that lead to an increase in false positives and unnecessary computational costs.[4−9] Correctly assigning tautomer states in protein−ligand complexes is also crucial for molecular dynamic simulations and protein−ligand binding free energy calculations.[10,11] Several experimental techniques are available to determine the tautomer ratio in various solutions, including NMR, UV−visible spectroscopy, IR, and fluorescence spectroscopy.[12−14] However, the small free energy differences between tautomeric states and their rapid interconversion make experimental measurements challenging. Therefore, rapid and accurate prediction of tautomer ratios and favorable tautomeric states in aqueous solution is essential for computational drug discovery.

Traditional computational methods are generally based on empirical rules or quantum mechanical (QM) calculations. Empirical rules-based methods[4,15−19] rely on rules derived from experimental and computed data to determine the tautomeric preference, considering factors like the number of aromatic rings and double bonds. However, empirical-based scoring methods only rank tautomeric states without providing energy information. Quantum mechanical calculations-based methods, which combine QM method and an implicit solvent model in a thermodynamic cycle, can accurately calculate energy differences between tautomer states. While these methods offer remarkable performance, their substantial computational requirements limit high-throughput applica-

Received: January 10, 2025
Revised: March 9, 2025
Accepted: March 11, 2025
Published: March 17, 2025

https://doi.org/10.1021/acs.jctc.5c00041
J. Chem. Theory Comput. 2025, 21, 3132−3141

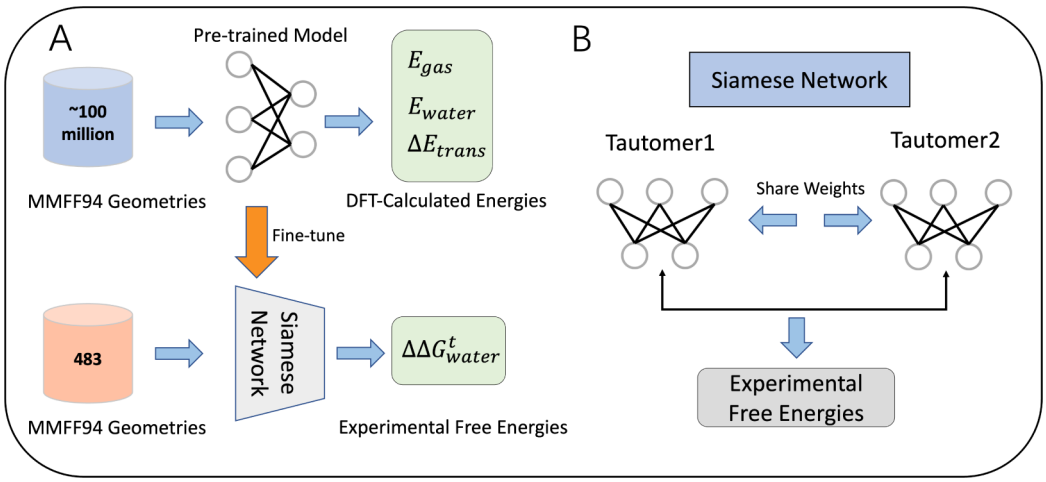


Figure 1. A strategy to develop a deep learning model for predicting tautomer ratios in aqueous solution with limited experimental data, based only on the structures of two tautomeric states. (A) The pretrained model was first trained on data from B3LYP/6−31G*/SMD calculations and subsequently fine-tuned with experimental data. This approach enables the prediction of relative free energies between tautomer pairs using a Siamese neural network. (B) The Siamese neural network architecture. The inputs are two tautomeric states with 3D geometries, and the model predicts their free energy difference. Basic models are applied to each input, sharing their weights.

tions. The tautomer ratios can be converted into free energies by ΔG = -RTln *K*, allowing for accurate calculation of it through free energy calculation methods. In the SAMPL2 blind challenge, designed to evaluate computational methods for predicting hydration free energies and tautomer ratios in aqueous solution, the top four submissions used quantum mechanical methods with implicit solvent models, and they achieved root-mean-square error (RMSE) deviations ranging from 1.9 to 3.4 kcal/mol.[20−23] Additionally, Wieder et al.[24] achieved comparable results with an RMSE of 2.2 kcal/mol in a retrospective study using B3LYP/aug-cc-pVTZ//B3LYP/6−31G(d)/SMD.

Although quantum mechanical based methods have achieved some success in predicting tautomer ratios, their accuracy is limited, and they require significant computational resources. Recently, deep learning strategies have made progress in predicting electronic energies, solvation energies, atomic forces, and various molecular properties. One approach is to develop deep potentials to improve the accuracy of molecular simulations by learning density functional theory (DFT) calculated energies, forces, and partial charges.[25−36] This type of method optimizes molecular geometries iteratively by itself and then calculates the equilibrium electronic energy. It also requires significant computational time for processing large chemical library, although it is faster than DFT methods. Another approach is to learn the electronic energy and solvation energy using force field-optimized geometries.[37−39] This method saves a lot of computational resources and time by avoiding the need to obtain high-quality geometries while still achieving good accuracy. Advances in artificial intelligence offer new avenues for developing methods that can predict tautomer ratios in aqueous solutions quickly and accurately. Wieder et al.[24] also utilized experimental data to fine-tune the ANI-1ccx deep potential model for predicting tautomer ratios in aqueous solution, including solvent effects by employing a relative alchemical free energy calculation protocol. Their optimized ANI-1ccx model achieved an RMSE of 2.8 [2.2, 3.2] kcal/mol with the alchemical free energy calculation, which is better than the native ANI-1ccx model's performance with an RMSE of 6.7 [5.7, 7.7] kcal/mol. Ji et al.[40] developed a tool for predicting favorable tautomeric states, which includes a deep learning-based scoring method for tautomer ranking. This scoring method combines the ANI-2x deep potential with a deep learning-based solvation model trained on DFT-calculated data, which achieved a similar performance to wB97*X*/6−31G*//M062*X*/6−31G*/SMD on the experimental data set, with an RMSE of 3.15 kcal/mol. All these methods are somewhat time-consuming or have limited performance. The method developed by Wieder et al. needs to run MD and alchemical free energy calculations; the method developed by Ji et al. requires optimizing molecular geometries using the ANI-2x deep potential model and its performance cannot surpass that of the DFT methods it was trained on. To our best knowledge, no deep learning approaches have developed to directly predict experimental tautomer ratios based on the input of two tautomer structures. Meanwhile, due to the limited size of experimental data, to train a robust and accurate deep learning model from scratch can be challenging.

In this study, we introduce sPhysNet-Taut, a model fine-tuned from the pretrained neural network using experimental data. This model takes MMFF94-optimized conformations as input and directly predicts the relative energy between tautomer pairs. We constructed the Frag20-Taut database using DFT calculations, which includes electronic energy data for both gas and aqueous phases. This data set contains approximately one million well-selected molecules. Initially, we trained the sPhysNet-MT model on the Frag20-Taut data set to predict calculated energies. Subsequently, we designed a Siamese neural network[14,41] based on sPhysNet-MT and fine-tuned it using experimental data collected from Tautobase[42] (as shown in Figure 1). After training and fine-tuning on the Frag20-Taut and experimental data sets using 5-fold cross-validation, our model significantly outperforms other methods, achieving an RMSE of 1.9 kcal/mol on the 100-tautomers set and 1.0 kcal/mol on the SAMPL2 challenge. On an external test set, our model achieved a 77% success rate in ranking the favorable tautomer as the preferred species. Additionally, we developed a user-friendly web server and command line tool for tautomer enumeration and ranking in aqueous solution based on the sPhysNet-Taut model. The web server is available

https://doi.org/10.1021/acs.jctc.5c00041
J. Chem. Theory Comput. 2025, 21, 3132−3141

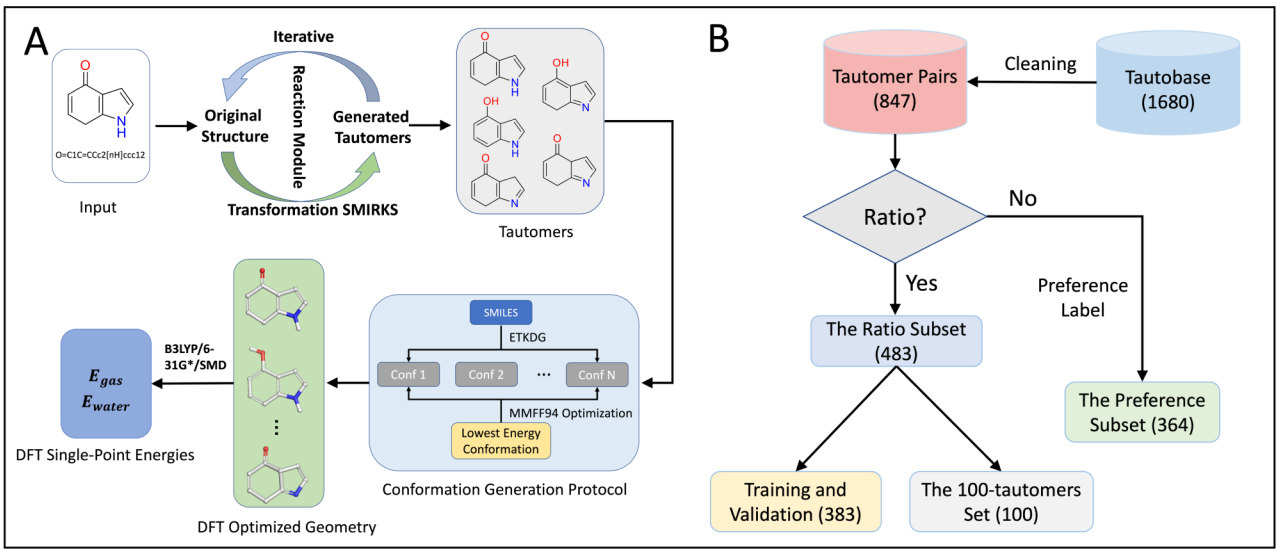


Figure 2. (A) Workflow for preparing Frag20-Taut data set. This process includes four steps: Enumerating all possible tautomeric states using transformation rules, optimizing the conformation of each tautomer using the MMFF94 force field and selecting the lowest-energy conformation, performing geometry optimization using B3LYP/6−31G*/SMD, and calculating single-point energies. (B) The workflow used to compile experimental data from the Tautobase database. Two data sets were extracted: the ratio subset with experimentally measured log K values, and the preference subset with experimental preferred states.

at https://yzhang.hpc.nyu.edu/tautomer, and the source code is accessible at https://github.com/xiaolinpan/sPhysNet-Taut.

MATERIAL AND METHODS

Data Set. *Data Set for Calculated Energetics: Frag20-Taut.* To train sPhysNet-MT for accurately predicting the relative energies between different tautomeric states in aqueous solution, we constructed the Frag20-Taut data set comprising a large set of generated tautomer pairs. In our previous work, we developed the Frag20 data set to model molecular electronic energies and transfer energies. The molecules within Frag20 data set were meticulously selected using a well-designed protocol to ensure sufficient diversity while limiting the maximum heavy atom count to 20, as detailed in our earlier publications.[37,39] We adapted the SMIRKS strings of the 54 prototropic tautomeric transformation rules summarized by Dhaked et al. for compatibility with the RDKit reaction module.[40,43] These transformation rules were applied to generate tautomer pairs for molecules within the Frag20 data set. If the 54 prototropic tautomeric rules did not yield tautomeric states, we employed the tautomer generation module in RDKit. We randomly selected 62,688 molecules with potential tautomeric transformation from Frag20, generating a total of 313,630 tautomeric molecules. We retained only the molecules whose SMILES strings did not change after DFT optimization, resulting in a total of 250,822 molecules. The Frag20-Taut combines Frag20 and the generated tautomeric molecules, containing a total of 929,738 molecules with calculated energies. Figure 2 illustrates the preparation workflow for Frag20-Taut. The first step involves generating all possible tautomeric states for each molecule and discarding those without alternative tautomeric states. The second step is to generate conformations for each tautomer structure using the ETKDG method[44,45] and optimize them with the MMFF94 force field.[46−50] For each tautomer, we generated 300 initial conformations and optimized them using MMFF94, retaining the lowest-energy optimized conformation as the input geometry for DFT calculations. The third step involves optimizing the geometric structures and calculating the electronic energies E_{gas} and E_{water} using the DFT method at the B3LYP/6−31G* level with the universal solvation model (SMD) in Gaussian 16 software.[51] The transfer energies are determined by calculating the difference between the single-point electronic energies in the gas phase and the water phase, defined as follows: $\Delta E_{transfer} = E_{water} - E_{gas}$.

Experimental Tautomer Data Set: Tautobase. Wahl et al. have published an open-source tautomer database, Tautobase, containing experimentally measured and estimated tautomer ratios for 1680 entries in various solvents, primarily water.[42] In this study, our goal is to develop a model for predicting tautomer ratios to accelerate drug discovery, focusing on water as the solvent since the tautomeric state within high abundance preferentially binds to receptor. To obtain a clean experimental data set for fine-tuning our model, we filtered records in Tautobase using the following criteria: (1) only records measured in water were retained; (2) the record must have a logK value; (3) all molecules must be neutral; (4) the elements in the molecule must be limited to C, H, O, N, S, P, F, Cl, and Br; (5) the tautomerism must be a prototropic transformation. After applying these rules and excluding molecules from the SAMPL2 challenge, we obtained 483 records and named it as the ratio subset. Additional records without logK values in water were used as an external test set to evaluate the ranking power of our model, defined by its ability to prioritize the most favorable tautomer as the top-ranked species. The external test set contains a total of 364 records and named it as the preference subset. We then use the same protocol as in the Frag20-Taut preparation to generate the lowest-energy MMFF94-optimized conformation for each tautomer from the SMILES string. The logK values were converted to free energy by $\Delta G = = RT\ln K$.

Deep Learning Models. *Siamese Neural Network.* Siamese neural networks[14,41] are commonly used to process two inputs and determine the distance between them, such as in object tracking and matching tasks. The goal of model

https://doi.org/10.1021/acs.jctc.5c00041
J. Chem. Theory Comput. 2025, 21, 3132−3141

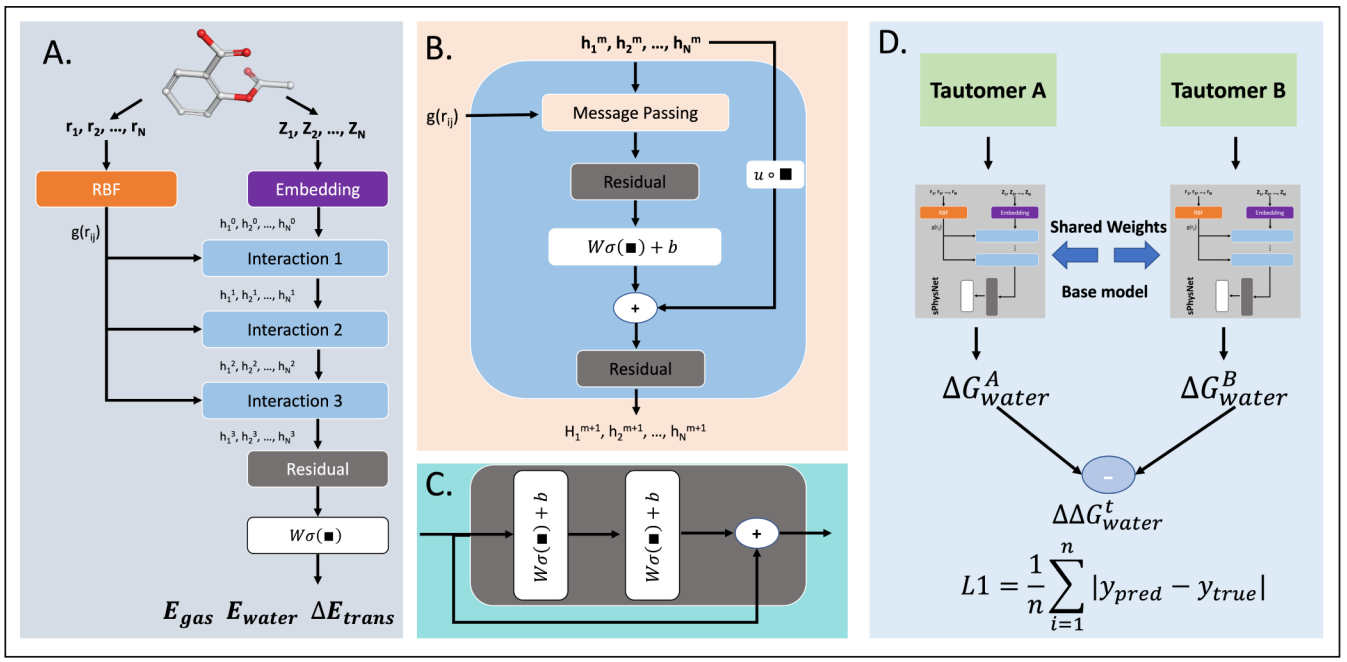


Figure 3. Overview of the sPhysNet-Taut model based on sPhysNet[37,39] and PhysNet.[34] (A) The architecture of the sPhyNet-MT model. It consists of a radial basis function (RBF) layer, an embedding layer, interaction layers, and residual layers. The RBF layer encodes distances between atom pairs into vectors, while the embedding layer encodes element types into vectors. This information is then processed through the interaction layer and residual layer to predict the final targets. (B) The interaction layer, which includes message passing, residual layers, and gate layers. (C) The residual layer, used within the interaction layer and before the model output, helps refine the predictions. (D) The architecture of the Siamese neural network. The core of the Siamese neural network is the sPhysNet-MT model with shared weights across both branches. The output is calculated as the difference between the predicted energies in water for the two mirrored models.

optimization is to minimize the distance between the predicted vectors for inputs of the same category and maximize it for those of different categories, as shown in eq 1, where $f(x)$ represents any differentiable function. These neural networks are also suitable for modeling differences between two relative states to predict pairwise properties, such as protein−ligand relative binding affinities[52] and molecular property predictions.[53] In this work, we designed a Siamese neural network based on the pretrained sPhysNet-MT model and fine-tuned its parameters using experimental data. The model architecture is detailed in Figure 3D. Two input tautomeric states are processed by the shared base model to predict their electronic energies in water. The relative energies, $\Delta\Delta G^{t}_{solv}$, are calculated as the difference between the two predicted energies, as shown in eq 2. These two base models share their weights, ensuring consistent processing of both inputs.

$$G((x_1, y_1), (x_2, y_2)) = \begin{cases} \min(\|f(x_1) - f(x_2)\|) & y_1 = y_2 \\ \max(\|f(x_1) - f(x_2)\|) & y_1 \neq y_2 \end{cases} \quad (1)$$

$$\Delta\Delta G^{t_A \rightarrow t_B}_{solv} = f(Taut_B) - f(Taut_A) \quad (2)$$

sPhysNet-MT. To accurately predict the DFT calculated electronic energies in both gas and aqueous phases, we developed a multitask deep learning model called sPhysNet-MT, using MMFF94-optimized conformations as input structures. sPhysNet-MT is a modification of PhysNet,[34] featuring fewer parameters, faster computation speed, and is implemented using PyTorch. Figure 3A−C illustrate the architecture of sPhyNet-MT, which includes an RBF layer, an element-type embedding layer, interaction modules, and residual modules. Initially, the N atoms in a molecule go through an embedding layer and are embedded into vectors h_1^0, h_2^0, ..., h_N^0 (node embeddings). The distances between each atom pair are expanded with RBFs into vectors {g(rij) | i, j ∈ {1,2,···, N}, rij <10Å} (edge embeddings). The node embeddings are then updated by three interaction modules with message passing, residual layers, and gate layers. The node embeddings of the final layer go through the output layer to predict targeted properties. For each input molecule, sPhysNet-MT provides three predicted energies by summing the predicted atomic energies: electronic energies both in the gas phase and the aqueous phase, and transfer energies between these phases. The details of each module are described in Text S1.

Model Training. *Pretraining on Frag20-Taut.* Since the experimental data on the relative free energy between tautomer pairs is extremely limited, it is difficult to train a robust and accurate deep learning model from scratch. Therefore, as described earlier, we aim to pretrain our model based on DFT-calculated data to obtain good initial model parameters, ensuring fast convergence and improved accuracy when fine-tuning the model based on experimental data, while also integrating the robustness of the DFT method. The Frag20-Taut data set comprises three target energies: E_{gas}, E_{water}, and ΔE_{trans}. To train the multitask model on these properties, we employ a loss function that optimizes the model weights by summing the L1 losses for each of the three energies. The loss function is defined as follows:

https://doi.org/10.1021/acs.jctc.5c00041
J. Chem. Theory Comput. 2025, 21, 3132−3141

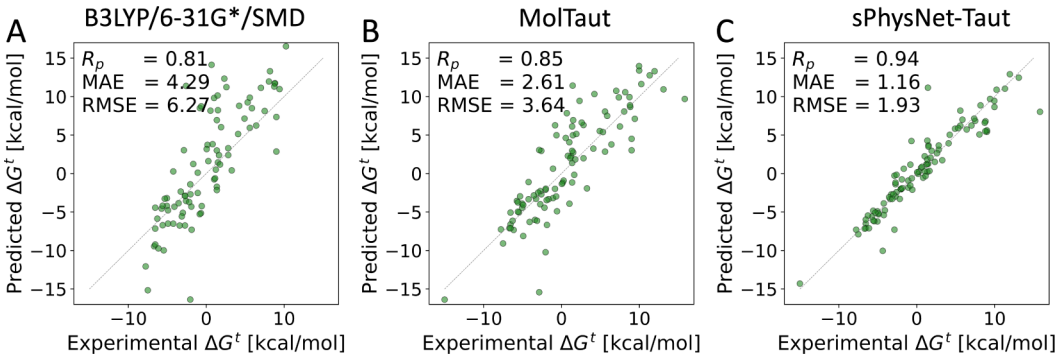


Figure 4. Comparative performance of three methods for predicting the experimental relative free energy between tautomer pairs in aqueous solution: (A) B3LYP/6−31G*/SMD: Uses the B3LYP functional with the 6−31G* basis set the SMD solvation model. (B) MolTaut: Uses ANI-2x for predicting electronic energy in the gas phase and the MolSolv model for solvation energy prediction, (C) sPhysNet-Taut: Fine-tuned on experimental data using MMFF94-optimized geometries.

$$L = \frac{1}{N}\sum_{i}^{N} |\widehat{E^i_{\text{gas}}} - E^i_{\text{gas}}| + \frac{1}{N}\sum_{i}^{N} |\widehat{E^i_{\text{water}}} - E^i_{\text{water}}| + \frac{1}{N}\sum_{i}^{N} |\widehat{\Delta E^i_{\text{trans}}} - \Delta E^i_{\text{trans}}| \quad (3)$$

We randomly split the Frag20-Taut data set into a training set, validation set, and test set, containing 893,954, 17,942, and 17,844 samples, respectively. Training a model based on DFT-optimized geometries is often impractical for high-throughput energy calculations; therefore, we used MMFF94-optimized geometries as input structures for the pretrained model. For comparison, we also trained a model using DFT-optimized geometries to assess the performance differences attributable to input geometries. sPhysNet-MT was implemented using PyTorch, with the message-passing module and data loader module based on PyTorch Geometric (PyG).[54] We used the EmaAmsGrad optimizer with a learning rate of 0.001, a β parameter set from 0.0 to 0.99, epsilon at 1e-8, and no weight decay. During training, we saved the model weights that achieved the best performance (RMSE) on the validation set over 500 epochs, utilizing an NVIDIA A100 80GB GPU.

Fine-Tuning on Experimental Data. Our objective with the experimental data set was to fine-tune the predictions of tautomer pairs in aqueous solution using MMFF94-optimized geometries within our designed Siamese neural network. We utilized 5-fold cross-validation to train and validate the neural network. The experimental data set was split into a training set and a validation set in an 8:2 ratio, resulting in 306 tautomer pairs for training and 77 for validation. The test set comprised a fixed set of 100 tautomer pairs, referred to as the "100-tautomers set". For gas-phase energies, we used DFT-calculated values as learning targets. We employed the EmaAmsGrad optimizer in our fine-tuning process over 2,000 epochs with a learning rate of 0.0001, saving the model weights that exhibited the best performance on the validation set. An L1 loss function was used to optimize the model weights, with α and β set to 0.8 and 0.2, respectively. The loss function is defined as follows:

$$L = \alpha\frac{1}{N}\sum_{i}^{N} |\widehat{\Delta G^t_{\text{water}}} - \Delta G^t_{\text{water}}| + \beta\frac{1}{N}\sum_{i}^{N} |\widehat{\Delta G^t_{\text{gas}}} - \Delta G^t_{\text{gas}}| \quad (4)$$

Workflow for Predicting Favorable Tautomers. In this work, we have developed a rapid workflow to predict favorable tautomeric states in aqueous solution. Initially, we enumerate all possible tautomeric states using transformation rules in the RDKit reaction module, covering 54 types of prototropic tautomerism summarized by Dhaked et al.[43] For each input molecule, we apply these transformation rules iteratively up to five times using a cycle generation protocol. We generate molecular conformations using ETKDG[44,45] and optimize each one using MMFF94[46−50] within RDKit package, retaining the lowest-energy conformation as the input for our energy prediction model. Next, we predict the relative energy of each tautomer pair and rank the tautomer states based on their predicted relative energies in water. Finally, we identify all low-energy tautomeric states based on an energy cutoff (default is 2.76 kcal/mol, corresponding to 1% tautomer ratio). To evaluate the runtime of our workflow, we selected 483 molecules from the ratio subset as a molecular pool to generate tautomeric states and conformations, which we then ranked using our model. All calculations were running on a single CPU core. Our workflow can process an average of 14.4 molecules per minute. In comparison, MolTaut can only process an average of 1.8 molecules per minute, and the DFT method can process only 0.005 molecules per minute. Our workflow is about 2880 times faster than the DFT method and about 8 times faster than MolTaut.

RESULTS AND DISCUSSION

Performance in Predicting Relative Free Energies between Tautomer Pairs. In a previous study focused on favorable tautomer prediction, Ji et al.[40] developed a scoring method called MolTaut for tautomer ranking, which combines the ANI-2x deep potential with a deep learning-based solvation model (MolSolv) trained on DFT-calculated data. However, the performance of MolTaut was limited by the inherent constraints of DFT methods. To enhance accuracy, we aimed to integrate both calculated and experimental data. As previously mentioned, we fine-tuned a pretrained model using experimental data within a Siamese neural network, resulting in the sPhysNet-Taut model. A detailed description of the performance for the pretrained model is provided in the Text S2.

To comparatively evaluate the performance of our model, we compared the performance of B3LYP/6−31G*/SMD, MolTaut, and sPhysNet-Taut model based on MMFF94

https://doi.org/10.1021/acs.jctc.5c00041
J. Chem. Theory Comput. 2025, 21, 3132−3141

optimized geometries. As illustrated in Figure 4, sPhysNet-Taut predicts the relative free energies on the **100-tautomers set** with a MAE of 1.16 kcal/mol and an RMSE of 1.93 kcal/mol. This performance significantly outperforms both the MolTaut scoring method (MAE = 2.61 kcal/mol, RMSE = 3.64 kcal/mol) and B3LYP/6−31G*/SMD (MAE = 4.29 kcal/mol, RMSE = 6.27 kcal/mol), achieving state-of-the-art results. Additionally, we analyzed the distribution of prediction errors across different ranges of heavy atom counts in the test set (Figure S1). For molecules with fewer than 15 heavy atoms, we observed that the prediction error increased with molecular size. In contrast, for molecules with more than 15 heavy atoms, the prediction error did not further increase, which may be attributed to data set imbalance. Overall, these findings suggest that molecular size plays a significant role in influencing the performance of the model.

SAMPL2,[55] a blind computational challenge, provides a list of tautomer pairs with experimentally measured ratio values, making it an excellent data set for evaluating various models. This data set is split into two subsets: an obscure set containing 8 tautomer pairs and an explanatory set with 12. We selected tautomer pairs that contain only neutral molecules with reliable error estimations. These molecules were excluded from our training set to ensure unbiased evaluation. We assessed the performance of our model against four methods that excelled in the SAMPL2 challenge (refs.[20−23]), one method distinguished for its performance in a retrospective study (ref,24 and the previously proposed AI-based method, MolTaut. Further details on these methods are available inText S3. According to the results shown in Tables 1 and S2, our

Table 1. Performance of This Work, Wieder et al.'s Method, and Several Submissions of SAMPL2 Challenge

	RMSE (kcal/mol)		
name	the obscure set	the explanatory set	ref
sPhysNet-Taut	1.3	0.8	this work
MolTaut[40]	1.3	3.9	40
Wieder et al.[24]	1.3	2.5	24
Klamt et al.[21]	1.4	3.6	
Ribeiro et al.[22]	1.5	2.9	
Kast et al.[20]	2.8	0.8	
Soteras et al.[23]	1.3	3.8	

sPhysNet-Taut model achieved a total RMSE of 1.0 kcal/mol with MMFF94-optimized geometries, greatly outperforming other quantum mechanics-based methods and MolTaut. Notably, whereas most methods showed less robust performance on the explanatory set compared to the obscure set (except for ref,20 our method demonstrated the opposite trend, excelling on the explanatory data set with an RMSE of 0.8 kcal/mol using MMFF94-optimized geometries. Furthermore, as illustrated in Figure S2, most tautomeric transformations in the SAMPL2 test set involve 1,3 aromatic heteroatom H-shifts, which are common and crucial in drug-like heterocyclic compounds. This result suggests that sPhysNet-Taut is capable of accurately predicting tautomer ratios for this type of tautomerization.

Building on the finding of Zhang et al.,[38] who demonstrated that fine-tuning significantly enhances model performance in learning aqueous free solvation energies on small experimental data sets, we investigated the effect of training set size on learning tautomer ratios to further illustrate the advantages of fine-tuning over training models from scratch. We created 10 different training sets, ranging from 75 to 300 tautomer pairs, and used them to both fine-tune sPhysNet-Taut and train it from scratch using MMFF94-optimized geometries with 5-fold cross-validation. All models were tested on the 100-tautomers set, and their performance is depicted in Figure 5. Notably,

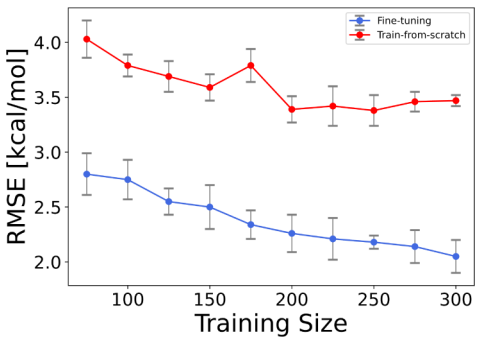


Figure 5. Performance of sPhysNet-Taut on the 100-tautomers test set across different training set sizes. This compares the model's performance when fine-tuned versus trained from scratch over varying training set sizes. The error bars indicate standard deviations, while the central points represent the mean values.

fine-tuning with just 75 tautomer pairs achieved a commendable RMSE of 2.80 kcal/mol, surpassing the performance of models trained from scratch. Furthermore, while the performance of models trained from scratch plateaued as the data set size increased, the fine-tuned models continued to improve with larger data sets. These results affirm that fine-tuning is an effective strategy for training deep learning models on small experimental data sets, offering a promising approach for model development with limited data.

Ranking Power for Predicting Favorable Tautomers. To evaluate the ranking ability of our model is also important for its application in drug discovery. We further assessed the ranking ability of sPhysNet-Taut on the two experimental test sets, as presented in Figure 6. On the 100-tautomers set, sPhysNet-Taut accurately ranked 94% of tautomer pairs with MMFF94-optimized geometries, outperforming both the DFT method and MolTaut. To evaluate our model on a larger data set, we selected an external set containing 364 tautomer pairs without logK value from Tautobase, as described in the data set section. On this external test set, sPhysNet-Taut correctly ranked 77% of pairs, again outperforming the DFT method and MolTaut. While the ranking success rate decreased from 94% to 77%, sPhysNet-Taut consistently demonstrated superior performance compared to the other methods.

Application of sPhysNet-Taut on the PDBbind Database. Tautomerization significantly impacts structure-based and ligand-based screening methods, as different tautomeric states have different pharmacophores, 3D shapes, electrostatic surfaces, and conformations. These variations can lead to different interactions, potentially increasing false positives and computational costs. Generally, the tautomeric state with the lowest energy often dominates, while higher-energy tautomeric states is less abundant in solution and may incur energetic penalties in binding free energy. In this study, we used the PDBbind v2020 refined set[56−58] to assess the effects of reassigning tautomeric states using sPhysNet-Taut on the hydrogen bonding. To simplify our analysis, we focused solely on neutral molecules in the refined set, identifying 2,086

https://doi.org/10.1021/acs.jctc.5c00041
J. Chem. Theory Comput. 2025, 21, 3132−3141

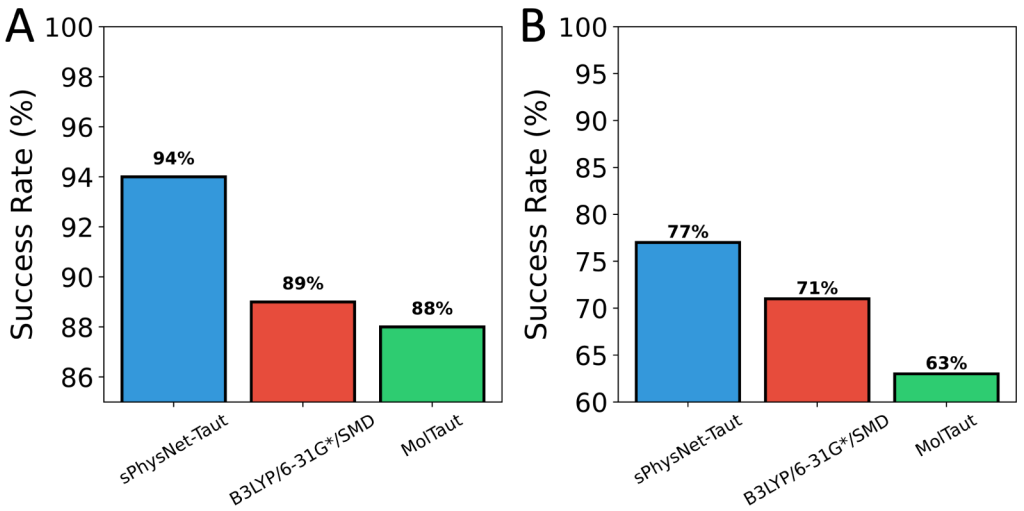


Figure 6. Success rate of ranking tautomer pairs on (A) the 100-tautomers set and (B) the external test set.

structures with neutral ligands out of the 5,316 available complex crystal structures available. We excluded crystal structures where the ligands differed from those reported in the original publications. We compared the tautomeric states in the original structures to the lowest-energy tautomeric states predicted by sPhysNet-Taut. We found that 70 original tautomeric states of ligands were more than 2.76 kcal/mol above the predicted lowest-energy tautomeric state, indicating that this state exists in aqueous solution at less than 1%. These findings and the details of the ligands are provided in Table S4. Before analyzing hydrogen bonds, we prepared the structures by removing water molecules, adjusting protein hydrogens, and minimizing all structures with a 0.5 Å RMSD constraint using the Schrödinger protein preparation wizard,[59] keeping ligand structures unchanged. The number of hydrogen bonds between proteins and ligands was calculated using the Schrödinger Python API.

Figure 7 illustrates the changes in the number of hydrogen bonds between original structures and reassigned tautomeric

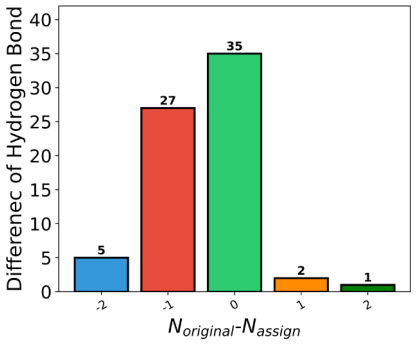


Figure 7. Differences in the number of hydrogen bonds were observed between the original structures in PDBbind and the reassigned tautomer structures generated by sPhysNet-Taut.

states in the PDBbind refined set. Our results show that reassigning to the predicted lowest-energy tautomer increased the number of hydrogen bonds in 32 complex structures, as detailed in Figure S3A. In 35 structures, no changes in hydrogen bonds were observed because the proton shifts occurred at the solvent interface or between nonpolar atoms (Figure S3B), and interactions with bridging water molecules were excluded from the analysis. In three crystal structures, the number of hydrogen bonds decreased following tautomer reassignment. For instance, in the crystal structure of PDB ID 2VK2 (Figure S3C), the predicted conversion of a hydroxyl group to a carbonyl group led to the loss of two hydrogen bonds with ARG17 and GLU13. Additionally, for two other structures (PDB ID: 4G90 and 5OHA), our predictions—though missing one hydrogen bond—are consistent with the original publications but differ from the entries in the PDBbind database, as shown in Figure S3D,E. These findings suggest that reassigning tautomeric states is essential for modeling the protein−ligand interactions.

CONCLUSIONS

Many organic small molecules exist multiple tautomeric states, but typically only one or a few are dominant, which complicates ligand preparation in chemical libraries. Although QM based methods can predict tautomer ratios with high accuracy across various solvents, their substantial computational requirements limit their use in high-throughput applications. To address this challenge, we developed a deep learning model based on a Siamese neural network to predict tautomer ratios in aqueous solution using MMFF94-optimized conformations. We enhanced the pretrained model by fine-tuning it with experimental data to improve its predictions of molecular internal energies and solvent effects. Our pretrained model leverages multitask learning to predict three types of energies: electronic energies in gas and water phases, and transfer energies between these phases, achieving chemical accuracy in these predictions. On the experimental data set, our fine-tuned model achieves state-of-the-art performance, with an RMSE of 1.93 kcal/mol on the test set and 1.16 kcal/mol in the SAMPL2 challenge. It also provides superior ranking power for tautomer pairs on several test sets. This study not only provides a valuable deep learning model for predicting tautomer ratios directly from force field-optimized geometries but also offers a framework for modeling pairwise data. Additionally, we developed a user-friendly tool to estimate all possible tautomeric states using transformation rules and rank them using our sPhysNet-Taut model to predict favorable tautomeric states in aqueous solution. We believe this model will be a handy tool in computational drug discovery by reducing tautomeric conflicts in large chemical libraries,

https://doi.org/10.1021/acs.jctc.5c00041
J. Chem. Theory Comput. 2025, 21, 3132−3141

eliminating unstable tautomeric states during virtual screening, and evaluating the effect of substitution on tautomeric equilibrium during lead optimization.

■ ASSOCIATED CONTENT

Data Availability Statement

All data sets and source codes used in this work are available. The Frag20-Taut data set can be accessed at 10.5281/zenodo.13870370, the experimental data used to fine-tune our model is extracted from Tautobase (https://github.com/WahlOya/Tautobase) and the ratio subset and the preference subset) for model training and testing are located at https://github.com/xiaolinpan/sPhysNet-Taut/exp_data. All source codes for data set preparation, model training, and model usage are at https://github.com/xiaolinpan/sPhysNet-Taut. The Web server is located at https://yzhang.hpc.nyu.edu/tautomer.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jctc.5c00041.

Text S1: Detailed description of sPhysNet-MT model; Text S2: Detailed description of the performance for the pretrained model; Text S3: Detailed description of the methods used in the SAMPL2 challenge; Figure S1: Error analysis for different ranges of heavy atom counts; Figure S2: Visualization of molecular structures in the SAMPL2 challenge; Figure S3: Comparative analysis of hydrogen bond interactions between original structures and their reassigned tautomeric states in the PDBbind database; Table S1: Test Performance on the Frag20-Taut calculated test set, which contains 17,844 molecules and 5,844 tautomer pairs; Table S2: Prediction details for each method in the SAMPL2 challenge; Table S3: Data for 70 Ligands from the PDBbind v2020 refined set with tautomeric differences (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Yingkai Zhang − *Department of Chemistry, New York University, New York, New York 10003, United States; Simons Center for Computational Physical Chemistry at New York University, New York, New York 10003, United States; NYU-ECNU Center for Computational Chemistry at NYU Shanghai, Shanghai 200062, China;* orcid.org/0000-0002-4984-3354; Email: yingkai.zhang@nyu.edu

Authors

Xiaolin Pan − *Department of Chemistry, New York University, New York, New York 10003, United States;* orcid.org/0000-0002-9465-3971

Xudong Zhang − *Department of Chemistry, New York University, New York, New York 10003, United States;* orcid.org/0009-0002-8631-3632

Song Xia − *Department of Chemistry, New York University, New York, New York 10003, United States;* orcid.org/0000-0002-6077-1718

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jctc.5c00041

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was supported by the U.S. National Institutes of Health (R35-GM127040). We thank NYU-ITS and NYUAD for providing computational resources.

■ REFERENCES

(1) Milletti, F.; Vulpetti, A. Tautomer Preference in PDB Complexes and its Impact on Structure-Based Drug Discovery. *J. Chem. Inf. Model.* **2010**, *50* (6), 1062−1074.

(2) Bharatam, P. V.; Valanju, O. R.; Wani, A. A.; Dhaked, D. K. Importance of tautomerism in drugs. *Drug Discovery Today* **2023**, *28* (4), 103494.

(3) Martin, Y. C. Let's not forget tautomers. *J. Comput.-Aided Mol. Des.* **2009**, *23* (10), 693−704.

(4) Oellien, F.; Cramer, J.; Beyer, C.; Ihlenfeldt, W.-D.; Selzer, P. M. The Impact of Tautomer Forms on Pharmacophore-Based Virtual Screening. *J. Chem. Inf. Model.* **2006**, *46* (6), 2342−2354.

(5) Kalliokoski, T.; Salo, H. S.; Lahtela-Kakkonen, M.; Poso, A. The Effect of Ligand-Based Tautomer and Protomer Prediction on Structure-Based Virtual Screening. *J. Chem. Inf. Model.* **2009**, *49* (12), 2742−2748.

(6) ten Brink, T.; Exner, T. E. Influence of Protonation, Tautomeric, and Stereoisomeric States on Protein−Ligand Docking Results. *J. Chem. Inf. Model.* **2009**, *49* (6), 1535−1546.

(7) Dhaked, D. K.; Nicklaus, M. C. Tautomeric Conflicts in Forty Small-Molecule Databases. *J. Chem. Inf. Model.* **2024**, *64*, 7409.

(8) Zhang, Y.; Zhang, Z.; Ke, D.; Pan, X.; Wang, X.; Xiao, X.; Ji, C. FragGrow: A Web Server for Structure-Based Drug Design by Fragment Growing within Constraints. *J. Chem. Inf. Model.* **2024**, *64* (10), 3970−3976.

(9) Shan, J.; Pan, X.; Wang, X.; Xiao, X.; Ji, C. FragRep: A Web Server for Structure-Based Drug Design by Fragment Replacement. *J. Chem. Inf. Model.* **2020**, *60* (12), 5900−5906.

(10) Hu, Y.; Sherborne, B.; Lee, T.-S.; Case, D. A.; York, D. M.; Guo, Z. The importance of protonation and tautomerization in relative binding affinity prediction: a comparison of AMBER TI and Schrödinger FEP. *J. Comput.-Aided Mol. Des.* **2016**, *30* (7), 533−539.

(11) Champion, C.; Hünenberger, P. H.; Riniker, S. Multistate Method to Efficiently Account for Tautomerism and Protonation in Alchemical Free-Energy Calculations. *J. Chem. Theory Comput.* **2024**, *20* (10), 4350−4362.

(12) Martin, Y. C. Experimental and pKa prediction aspects of tautomerism of drug-like molecules. *Drug Discovery Today: Technol.* **2018**, *27*, 59−64.

(13) Blakeley, M. Neutron crystallography aids in drug design. *IucrJ* **2016**, *3* (5), 296−297.

(14) Bromley, J.; Guyon, I.; LeCun, Y.; Säckinger, E.; Shah, R. Signature verification using a "Siamese" time delay neural network. In *In Proceedings of the 6th International Conference on Neural Information Processing Systems*; NeurIPS: Denver, Colorado, 1993.

(15) 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* (1), 68−75.

(16) Cruciani, G.; Milletti, F.; Storchi, L.; Sforna, G.; Goracci, L. In silico pKa Prediction and ADME Profiling. *Chem. Biodiversity* **2009**, *6* (11), 1812−1821.

(17) Urbaczek, S.; Kolodzik, A.; Rarey, M. The Valence State Combination Model: A Generic Framework for Handling Tautomers and Protonation States. *J. Chem. Inf. Model.* **2014**, *54* (3), 756−766.

(18) Bietz, S.; Urbaczek, S.; Schulz, B.; Rarey, M. Protoss: a holistic approach to predict tautomers and protonation states in protein-ligand complexes. *J. Cheminf.* **2014**, *6* (1), 12.

(19) Greenwood, J. R.; Calkins, D.; Sullivan, A. P.; Shelley, J. C. Towards the comprehensive, rapid, and accurate prediction of the favorable tautomeric states of drug-like molecules in aqueous solution. *J. Comput.-Aided Mol. Des.* **2010**, *24* (6), 591−604.

(20) Kast, S. M.; Heil, J.; Güssregen, S.; Schmidt, K. F. Prediction of tautomer ratios by embedded-cluster integral equation theory. *J. Comput.-Aided Mol. Des.* **2010**, *24* (4), 343−353.
(21) Klamt, A.; Diedenhofen, M. Some conclusions regarding the predictions of tautomeric equilibria in solution based on the SAMPL2 challenge. *J. Comput.-Aided Mol. Des.* **2010**, *24* (6), 621−625.
(22) Ribeiro, R. F.; Marenich, A. V.; Cramer, C. J.; Truhlar, D. G. Prediction of SAMPL2 aqueous solvation free energies and tautomeric ratios using the SM8, SM8AD, and SMD solvation models. *J. Comput.-Aided Mol. Des.* **2010**, *24* (4), 317−333.
(23) Soteras, I.; Orozco, M.; Luque, F. J. Performance of the IEF-MST solvation continuum model in the SAMPL2 blind test prediction of hydration and tautomerization free energies. *J. Comput.-Aided Mol. Des.* **2010**, *24* (4), 281−291.
(24) 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* (34), 11364−11381.
(25) Anstine, D.; Zubatyuk, R.; Isayev, O. AIMNet2: a neural network potential to meet your neutral, charged, organic, and elemental-organic needs, *ChemRxiv*, **2024**, .
(26) Zeng, J.; Tao, Y.; Giese, T. J.; York, D. M. QDπ: A Quantum Deep Potential Interaction Model for Drug Discovery. *J. Chem. Theory Comput.* **2023**, *19* (4), 1261−1275.
(27) Qiao, Z.; Welborn, M.; Anandkumar, A.; Manby, F. R.; Miller, T. F., III. OrbNet: Deep learning for quantum chemistry using symmetry-adapted atomic-orbital features. *J. Chem. Phys.* **2020**, *153* (12), 124111.
(28) 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* (7), 3408−3415.
(29) 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* (7), 4192−4202.
(30) Zubatyuk, R.; Smith, J. S.; Leszczynski, J.; Isayev, O. Accurate and transferable multitask prediction of chemical properties with an atoms-in-molecules neural network. *Sci. Adv.* **2019**, *5* (8), No. eaav6490.
(31) Smith, J. S.; Nebgen, B. T.; Zubatyuk, R.; Lubbers, N.; Devereux, C.; Barros, K.; Tretiak, S.; Isayev, O.; Roitberg, A. E. Approaching coupled cluster accuracy with a general-purpose neural network potential through transfer learning. *Nat. Commun.* **2019**, *10* (1), 2903.
(32) 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* (4), 3192−3203.
(33) 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., III OrbNet Denali: A machine learning potential for biological and organic chemistry with semi-empirical cost and DFT accuracy. *J. Chem. Phys.* **2021**, *155* (20), 204103.
(34) Unke, O. T.; Meuwly, M. PhysNet: A Neural Network for Predicting Energies, Forces, Dipole Moments, and Partial Charges. *J. Chem. Theory Comput.* **2019**, *15* (6), 3678−3693.
(35) 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* (24), 241733.
(36) Liu, Z.; Zubatiuk, T.; Roitberg, A.; Isayev, O. Auto3D: Automatic Generation of the Low-Energy 3D Structures with ANI Neural Network Potentials. *J. Chem. Inf. Model.* **2022**, *62* (22), 5373−5382.
(37) Xia, S.; Zhang, D.; Zhang, Y. Multitask Deep Ensemble Prediction of Molecular Energetics in Solution: From Quantum Mechanics to Experimental Properties. *J. Chem. Theory Comput.* **2023**, *19* (2), 659−668.
(38) Zhang, D.; Xia, S.; Zhang, Y. Accurate Prediction of Aqueous Free Solvation Energies Using 3D Atomic Feature-Based Graph Neural Network with Transfer Learning. *J. Chem. Inf. Model.* **2022**, *62* (8), 1840−1848.
(39) Lu, J.; Xia, S.; Lu, J.; Zhang, Y. Dataset Construction to Explore Chemical Space with 3D Geometry and Deep Learning. *J. Chem. Inf. Model.* **2021**, *61* (3), 1095−1104.
(40) Pan, X.; Zhao, F.; Zhang, Y.; Wang, X.; Xiao, X.; Zhang, J. Z. H.; Ji, C. MolTaut: A Tool for the Rapid Generation of Favorable Tautomer in Aqueous Solution. *J. Chem. Inf. Model.* **2023**, *63* (7), 1833−1840.
(41) Chopra, S.; Hadsell, R.; LeCun, Y.Learning a similarity metric discriminatively, with application to face verification*In 2005 Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*IEEE2005*1*539−546
(42) Wahl, O.; Sander, T. Tautobase: An Open Tautomer Database. *J. Chem. Inf. Model.* **2020**, *60* (3), 1085−1089.
(43) Dhaked, D. K.; Ihlenfeldt, W.-D.; Patel, H.; Delannée, V.; Nicklaus, M. C. Toward a Comprehensive Treatment of Tautomerism in Chemoinformatics Including in InChI V2. *J. Chem. Inf. Model.* **2020**, *60* (3), 1253−1275.
(44) Riniker, S.; Landrum, G. A. Better Informed Distance Geometry: Using What We Know To Improve Conformation Generation. *J. Chem. Inf. Model.* **2015**, *55* (12), 2562−2574.
(45) Wang, S.; Witek, J.; Landrum, G. A.; Riniker, S. Improving Conformer Generation for Small Rings and Macrocycles Based on Distance Geometry and Experimental Torsional-Angle Preferences. *J. Chem. Inf. Model.* **2020**, *60* (4), 2044−2058.
(46) Halgren, T. A. Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94. *J. Comput. Chem.* **1996**, *17* (5−6), 490−519.
(47) Halgren, T. A. Merck molecular force field. II. MMFF94 van der Waals and electrostatic parameters for intermolecular interactions. *J. Comput. Chem.* **1996**, *17* (5−6), 520−552.
(48) Halgren, T. A. Merck molecular force field. III. Molecular geometries and vibrational frequencies for MMFF94. *J. Comput. Chem.* **1996**, *17* (5−6), 553−586.
(49) Halgren, T. A. Merck molecular force field. V. Extension of MMFF94 using experimental data, additional computational data, and empirical rules. *J. Comput. Chem.* **1996**, *17* (5−6), 616−641.
(50) Halgren, T. A.; Nachbar, R. B. Merck molecular force field. IV. conformational energies and geometries for MMFF94. *J. Comput. Chem.* **1996**, *17* (5−6), 587−615.
(51) Marenich, A. V.; Cramer, C. J.; Truhlar, D. G. Universal Solvation Model Based on Solute Electron Density and on a Continuum Model of the Solvent Defined by the Bulk Dielectric Constant and Atomic Surface Tensions. *J. Phys. Chem. B* **2009**, *113* (18), 6378−6396.
(52) McNutt, A. T.; Koes, D. R. Improving ΔΔG Predictions with a Multitask Convolutional Siamese Network. *J. Chem. Inf. Model.* **2022**, *62* (8), 1819−1829.
(53) Tynes, M.; Gao, W.; Burrill, D. J.; Batista, E. R.; Perez, D.; Yang, P.; Lubbers, N. Pairwise Difference Regression: A Machine Learning Meta-algorithm for Improved Prediction and Uncertainty Quantification in Chemical Search. *J. Chem. Inf. Model.* **2021**, *61* (8), 3846−3857.
(54) Fey, M.; Lenssen, J. E. Fast graph representation learning with PyTorch Geometric. *arXiv*, **2019**, .
(55) Geballe, M. T.; Skillman, A. G.; Nicholls, A.; Guthrie, J. P.; Taylor, P. J. The SAMPL2 blind prediction challenge: introduction and overview. *J. Comput.-Aided Mol. Des.* **2010**, *24* (4), 259−279.
(56) Su, M.; Yang, Q.; Du, Y.; Feng, G.; Liu, Z.; Li, Y.; Wang, R. Comparative Assessment of Scoring Functions: The CASF-2016 Update. *J. Chem. Inf. Model.* **2019**, *59* (2), 895−913.
(57) Li, Y.; Su, M.; Liu, Z.; Li, J.; Liu, J.; Han, L.; Wang, R. Assessing protein−ligand interaction scoring functions with the CASF-2013 benchmark. *Nat. Protoc.* **2018**, *13* (4), 666−680.
(58) Liu, Z.; Su, M.; Han, L.; Liu, J.; Yang, Q.; Li, Y.; Wang, R. Forging the Basis for Developing Protein−Ligand Interaction Scoring Functions. *Acc. Chem. Res.* **2017**, *50* (2), 302−309.

https://doi.org/10.1021/acs.jctc.5c00041
J. Chem. Theory Comput. 2025, 21, 3132−3141

(59) Madhavi Sastry, G.; Adzhigirey, M.; Day, T.; Annabhimoju, R.; Sherman, W. Protein and ligand preparation: parameters, protocols, and influence on virtual screening enrichments. *J. Comput.-Aided Mol. Des.* **2013**, *27* (3), 221−234.

https://doi.org/10.1021/acs.jctc.5c00041
J. Chem. Theory Comput. 2025, 21, 3132−3141