Simulating Protein–Ligand Binding with Neural Network Potentials

by

© **Shae-Lynn Lahey**

A thesis submitted to the School of Graduate Studies in partial fulfillment of the requirements for the degree of Master of Science.

Department of Chemistry

Memorial University

August 2021

St. John's, Newfoundland and Labrador, Canada

Abstract

Computational methods have been developed to predict the structures and energetics of protein-ligands complexes. However these methods are limited by the accuracy and transferability of the molecular mechanical (MM) models used to calculate the potential energy. Neural network potentials (NNPs) eliminate the need for parameterization and avoid many of the limiting assumptions of MM models. We evaluated the accuracy of ANI-type NNP models for predicting the potential energy surface of biaryl torsions. The ANI-2X and ANI-1ccX NNPs were found to be more accurate and reliable than popular molecular mechanical models. We then developed a new method where the NNP is used to describe the intramolecular terms of a ligand while a conventional MM model is used to describe the environment. This method was found to be effective for predicting the binding pose of ligands bound to proteins and could be used to calculate the conformational component of the binding energy. We also show that these methods can be used to refine low-resolution cryo-EM structures of protein-ligand complexes.

I dedicate this work to my parents for making me believe I could do anything.

Lay summary

Most drugs work by binding to a protein in our bodies. This binding blocks or activates a protein, which results in the desired effect of the drug. A drug molecule must have a very specific shape and interactions to bind to a specific protein. In order to develop more effective or new drugs, we need to understand which shapes and interactions result in the strongest binding. Drugs that bind strongly have a lower, more negative binding energy. If we can minimize this energy we can maximize the effect of the drug; however, experimentally observing the structure of a drug bound in a protein and its binding energy is not always possible and requires sophisticated instruments. It is often more advantageous to model the protein–ligand complex using computer simulations. This is commonly done using a method called molecular dynamics (MD), which mimics the natural movement of the system by calculating the forces on the atoms of the drugs at every moment in time. The calculation of these forces have generally used molecular mechanical models. These methods can be effective, but can lack transferability: if optimal parameters have not been defined, the methods may not describe a given drug molecule accurately. Neural Network Potentials (NNP) are a new method that can calculate forces accurately while also having good transferability. NNP's have been trained to calculate forces on a new molecule based on the energies of similar molecules. The "machine learning" approach is based on the same type of method that Netflix uses to suggests new movies for you to watch based on what you have already seen. In this thesis, I have developed a new method to simulate the binding of drug molecules to proteins using the Accurate NeurAl engINe for Molecular Energies (ANAKIN-ME, or ANI for short) NNP. First I test their accuracy in calculating rotational energies of drug fragments against popular molecular mechanical models. Then I test their accuracy for predicting the bound conformations of drug molecules in comparison to the pose measured experimentally using X-ray crystallography. Lastly I use the ANI NNP to help better refine results

in cryo-electron microscopy (cryo-EM), a technique that flash freezes proteins and blasts them with a beam of electrons to capture an image of the protein structure. Cryo-EM is used to determine structures of proteins that could not be determined using X-ray crystallography, but the structures determined with this method are often lower resolution.

For each of these tests, the NNP consistently performed as well or better than conventional molecular mechanical models. Perhaps more importantly, they eliminate the need for parameterization for specific molecules, which makes computational drug discovery workflows simpler. Because of these advantages, NNP's could be the new way forward for protein–drug modelling.

Acknowledgements

I would like to acknowledge Dr. Chris Rowley, my supervisor, for taking me on as an undergraduate student and convincing me to do a graduate degree. Thank you for all your knowledge, help and support over the years. I would also like to acknowledge the Rowley group for being great co-workers and friends and helping me with my research. Thank you Dr. Christina Bottaro for becoming my co-supervisor. Thanks to the School of Graduate Studies at Memorial University for a graduate fellowship, and Dr Liqin Chen for a scholarship. Thank you to NSERC for funding through the CG-SM scholarship. Finally thank you to Compute Canada for the allocation for resources needed to do this research.

Statement of contribution

Nguyen Thien Phuc Tu performed the OpenMM simulations for Chapter 2.

Christopher Rowley ran the CGenFF, GAFF, and OpenFF scans in Chapter 2, as well as made the bootstrap analysis Figures (2.2) and the ozanimod figure (Figure 2.10). He calculated the PMF's of erlotinib in Chapter 3 (Figure 3.2) and created the erlotinib torsional energy figure (Figure 3.3).

I performed all other calculations and created all other figures not explicitly cited.

Table of contents

List of tables

List of figures

List of abbreviations

GAFF	Generalized AMBER Force Field
OpenFF	Open Force Field
CGenFF	CHARMM General Force Field
OPLS	Optimized Potentials for Liquid Simulations
NNP	Neural Network Potential
RMSD	Root Mean Squared Deviation
MADB	Mean Absolute Deviation of Barrier Height
MM	Molecular Mechanics
MD	Molecular Dynamics
EM	Electron Microscopy
2D	Two-Dimensional
3D	Three-Dimensional
MDFF	Molecular Dynamics Flexible Fitting
QM	Quantum Mechanics
ML	Machine Learning
NN	Neural Network
NMS	Normal Mode Sampling
DFT	Density Functional Theory
ED	Electron Density
SMIRNOFF	SMIRKS Native Open Force Field
RESP	Restrained Electrostatic Potential
HF	Hartree–Fock
PES	Potential Energy Surface
PMF	Potential of Mean Force
ABF	Adaptive Biasing Force
WHAM	Weighted Histogram Analysis Method
MSE	Mean Signed Error
EGFR	Epidermal Growth Factor Receptor
JAK1	Janus Kinase 1
DYRK1A	Dual specificity Tyrosine Phosphorylation Regulated Kinase 1A
XRD	X-Ray Diffraction

Chapter 1

Introduction

1.1 The Study of Protein–Ligand Binding

Proteins are an important class of macromolecules that are involved in cellular processes, including structural, chemical, mechanical, and cell signaling roles. They control these processes by forming intermolecular complexes with other cellular components, such as other proteins, nucleic acids, lipids, carbohydrates, or molecular species. Many proteins will bind to a specific chemical species (a.k.a., the ligand), which is often an organic molecule. This includes protein enzymes that catalyze the chemical transformation of a molecular substrate and protein receptors that can isomerize between two functional states when a molecule is bound to them. Many drug molecules act by binding to the protein targets in similar ways to their native substrates, where they block the activity of an enzyme on its substrate or bind to a receptor in a way that modulates its conformational state.

The identification of small molecules that affect the activity of a targeted protein by binding to them is a popular strategy to identify new drugs for the treatment of diseases. Drugs typically bind to a pocket in the protein where its natural substrate binds. The amino acids that form this binding pocket define the shape that a drug molecule must hold to fit in the pocket and the types of intermolecular interactions that can be formed between the protein and the drug [1]. Protein–ligand binding must also be strong enough to compensate for the interactions between the ligand and the surrounding solution that are lost when the ligand binds. Small-molecule

ligands of biological molecules can hold a range of geometries, both in solution and in their bound state. The strain and reduced flexibility of bound drugs can partially counter the intermolecular interactions that drive protein–ligand binding. All of these competing factors make it hard to predict whether or not a molecule or drug would be a good drug target for a protein.

The protein–ligand binding will only occur if the transfer of the ligand from solution into the binding site is spontaneous. In other words, if the intermolecular driving forces of binding outweigh all the other forces that oppose binding then protein-ligand binding will occur. The spontaneity of a reaction is measured by Gibbs energy (ΔG). Gibbs energy is the sum of the enthalpy (ΔH) minus the entropy (ΔS) at a given temperature (T).

$$\Delta G = \Delta G - T\Delta S \tag{1.1}$$

In order to test whether or not a drug is effectively binding to a protein, a range of experimental and computational methods have been developed. For the purposes of this thesis only X-ray crystallography and cryo-EM experimental techniques for studying protein–drug interactions are described, although there are many more techniques that are used such as Nuclear Magnetic Resonance Imaging (NMR), Laue X-ray diffraction, small angle X-ray scattering, binding assays and isothermal titration calorimetry. An extensive review of these techniques was done by Xing et al. [1]

1.1.1 X-ray Crystallography and Cryo-EM

X-ray crystallography is one of the most popular methods for determining the structure of a protein and has been in wide use for several decades. The first instance of X-ray diffraction of protein crystals was reported in the early 1930s [2]. Thirty years later the structure of myoglobin was captured by Kendrew *et al.*[3] marking the first crystal structure of a protein captured by X-ray crystallography. The procedure for determining a crystal structure of a protein is complex. A protein sample is purified, crystallized and then exposed to an X-ray beam to yield a three-dimensional structure. The X-ray beam is diffracted by the crystal in very specific directions. By measuring the angle, intensity and pattern of the diffracted beams, a three-dimensional electron

density map can be generated [2]. For example, the intensity of the spots is used to determine structure factors, and the pattern of the diffraction spots are examined to give information about the crystal packing symmetry and size of the unit cell that lead to an electron density map. Various methods are then applied to improve the quality of the map so that a molecular structure can be built using the protein sequence. The resulting structure is then refined to fit the map more accurately and represent the thermodynamically favored conformation [4].

X-ray crystallographic determination of protein structure has significant limitations. This technique relies on the growth of the protein into a crystal. A reliable source of protein must be available, and a purification/concentration protocol that will yield high quality, homogeneous, soluble material is needed [5]. There are also many solution conditions, such as pH, buffer, protein concentration, temperature, possible inclusion of additives, and choice of precipitant that all have to be optimized correctly in order for the protein to crystallize in a form suitable for X-ray crystallography [4]. A protein may also crystallize into a non-biological structure which is not of interest. A minimum crystal size of 0.1 mm is needed for analysis as well [4]. The quality of an XRD structure depends on factors like the intensity of the incident beam of X-rays, the degree of disorder in the crystal cell, and the degree to which the crystal diffracts the X-rays. The resolution of the protein structure is defined in terms of the shortest distance between points on the diffraction pattern, with a resolution of 1–2 Å corresponding to atomistic structures, while structures derived from patterns with a lower resolution (e.g., $>$ 2 Å) may only be reliable for the coarser features of the secondary and tertiary structure of a protein.

Cryogenic electron microscopy (cryo-EM) is a rapidly emerging method for protein structure determination. Cryo-EM has been used since the 1980's when Dubochet *et al.*[6] used this technique to study the structure of viruses. The advantage to cryo-EM is that it does not need the proteins to be crystallized. Instead, the proteins are placed on a sample grid and flash frozen by plunging the sample in liquid ethane and trapped between a thin film of amorphous ice. Two dimensional images of the sample are generated by bombarding it with beams of electrons. These 2D structures are then combined computationally to give a 3D structure. Because this technique avoids the crystal packing effects and solution conditions required to grow protein crystals, cryo-EM structures can be more accurate representations of a protein's native or biological state. Proteins can be observed in multiple conformations in their native environment,

providing a better insight to the behaviour of these molecules [1].

A drawback to cryo-EM is that the inherent signal to noise ratio that accompanied the low-power electron beam made resolutions of early cryo-EM structures poor, typically equivalent to an XRD structure with a resolution of 10 Å. In recent years, this technique has been improved to the point that structures routinely have a solution equivalent to a 3–4 Å resolution XRD structure, with more and more structures emerging with resolutions in the 2–3 Å range. Kato *et al.* [7] set a record with their structure of apoferritin at 1.54 Å resolution, and in 2020, Yip *et al.*[8] were able to capture the structure of a stable iron-storing protein called ferritin at a resolution of approximately 1.2 Å. These studies proved that cryo-EM techniques are able to achieve X-ray crystallographic levels of resolution of protein structures.

Cryo-electron microscopy has allowed the determination of many protein structures that could not be determined by methods like X-ray crystallography; however, it has seen more limited use in the determination of protein–ligand structures. Even with the recent improvements in the resolution of cryo-EM techniques, the structural resolution of bound ligands in cryo-EM structures is often too low to provide a definitive pose of the bound molecule or its intermolecular interactions. Molecular modeling of the ligand, with or without reference to experimental XRD or cryo-EM density of the protein, offers a concrete solution to resolving the structures of drug–protein complexes beyond what can be achieved using cryo-EM and X-ray crystallography alone.

1.2 Conventional Force Fields and Molecular Dynamics

Experimental methods for studying protein–ligand complexes can be time-consuming, expensive, and laborious. Modern drug design usually involves a study of hundreds to thousands of molecules to find the best molecule which is not realistic to approach with low-throughput XRD and cryo-EM experiments. Computer modeling has huge potential to predict the structures of protein–ligand complexes and to calculate the affinity of a drug to a protein. Computer modeling can also be used to predict the affinity of a drug to the protein, which cryo-EM and XRD cannot.

Molecular dynamics (MD) is one of the most popular methods for simulating

protein–ligand complexes. MD mimics the natural movement and interactions of the atoms of a biomolecular system. In these simulations, atoms are represented as single point masses and their trajectories are simulated according to Newton's equations of motion. MD can be used to sample equilibrium ensembles of configurations, which can then be analyzed to determine thermodynamic properties like the binding affinity.

MD simulations of protein–ligand binding require accurate methods to calculate the intramolecular interactions of these ligands to predict the relative stability of these conformations. Although high-level *ab initio* methods (e.g., MP2 or CCSD) provide accurate predictions of the stability of molecular geometry, simulations of protein–ligand binding commonly require the evaluation of millions or billions of configurations, making these methods impractical to be used directly. Instead, empirical molecular mechanical (MM) models are used to provide a simplified model for the potential energy and forces of the system.

In MM, a set of simple mathematical functions are used to approximate the intermolecular and intramolecular interactions of the system [9]. The potential energy ($\mathcal{V}$) is divided into two components: bonded and nonbonded terms.

$$\mathcal{V}_{\text{total}}(\mathbf{r}) = \mathcal{V}_{\text{bonded}}(\mathbf{r}) + \mathcal{V}_{\text{non-bonded}}(\mathbf{r}) \tag{1.2}$$

The bonded term describes interactions of atoms linked by covalent bonds,

$$\begin{aligned}\mathcal{V}_{\text{bonded}}(\mathbf{r}) = \sum_{\text{bonds}} \frac{1}{2} k_{\text{bond}} \left(r - r_{eq}\right)^2 + \sum_{\text{angles}} \frac{1}{2} k_{\text{angle}} \left(\theta - \theta_{eq}\right)^2 \\ + \sum_{\text{torsions}} \sum_{\text{i}} k_{torsion} \cos\left(n_i \left(\phi_i - \delta\right)\right),\end{aligned} \tag{1.3}$$

where k_{bond} is the spring constant for the bond stretch, r_{eq} is the equilibrium bond length, k_{angle} is the spring constant for the angle bend, θ_{eq} is the equilibrium bond angle, $\|_{torsion}$ is the barrier for torsional rotation, n_i is the rotational multiplicity, and δ is the phase shift.

The non-bonded term describes Pauli repulsive, van der Waals, and electrostatic interactions for pairs of atoms that are not bonded with each other. The Pauli and van der Waals terms are combined into the Lennard-Jones potential while the electrostatic

interactions are calculated as the sum of Coulombic interactions. The total non-bonded interaction potential is,

$$\mathcal{V}_{\text{non-bonded}}(\mathrm{r}) = (\mathbf{r}) = \sum_i \sum_j \frac{q_i q_j}{4\pi\epsilon r_{ij}} + 4\varepsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right] \tag{1.4}$$

where q is the partial charge of an atom, σ is the Lennard-Jones radius, and ε is the Lennard-Jones well-depth.

Assigning appropriate parameters for all the bonded and nonbonded terms is one of the primary challenges in developing and applying MM models. Partial atomic charges are assigned to each atom using empirical or quantum-mechanically derived methods which calculates the distribution of charges within a molecule [10], but the other parameters are typically assigned based on their homology to molecules with similar chemical structures. Each atom is assigned an "atom type" based on its element and chemical environment [11]. Appropriate bond stretch, angle bend, and dihedral rotation parameters must then be defined for all permutations of atom types present in a molecule.

The definition of optimal parameters for molecular mechanical models of protein–ligand binding remains a major challenge in computational chemistry. The electrostatic [12, 13], repulsive, and dispersion[14, 15] interaction terms have been developed actively; however, accurate representation of intramolecular potential energy of the ligand is particularly challenging and no complete, general solution has been developed. Force fields for drug-like compounds are particularly difficult to develop because of the enormous variety of chemical motifs, which often feature complex chemical effects like conjugation, hyperconjugation, and aromaticity. This is compounded by the enormous variety of chemical motifs that are possible in the chemical drug space, where each could require a distinct set of parameters.

1.3 Neural Networks and the ANI Neural Network Potential

Neural networks (NN) developed through machine learning (ML) are an emerging and powerful tool that could serve as an alternative to MM models. ML uses computer

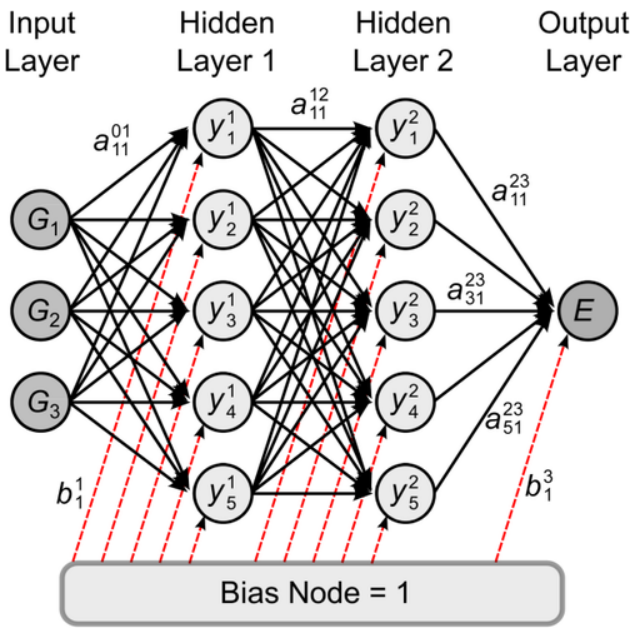

Figure 1.1: A general NN architecture with input coordinates G, weights a, nodes y, biasing weights b, and atomic energy E (This figure was reproduced from Behler *et al.*, "Generalized neural-network representation of high-dimensional potential-energy surfaces"[17]. Published by Physical Review Letters with permission).

algorithms to generate a model based on sample data to make predictions without being explicitly programmed to do so [16]. In this instance, the ML algorithm is trained to predict the potential energy of a molecular configuration using its coordinates as the input, which is known as a Neural Network Potential (NNP).

The NNs used in these models are highly flexible, non-linear functions with optimizable parameters called weights. The NN contains multiple hidden layers consisting of nodes that relate the input layer to the output layer through the weights that are adjusted to the desired level of theory [17]. The NN's can also contain a biasing weight that is used to offset the activation function, which is a function used to predict output based on input.

The weights are initially chosen randomly, but are adjusted to minimize a cost function. In this instance, the cost function is the deviation between the potential energy of a set of molecular configurations predicted by the NN and those calculated using a QM method. The weights are updated through the computation of analytic derivatives of a cost function until the weights that result in the lowest deviation are obtained. The data set used to optimize the weights of a NN is called a training set.

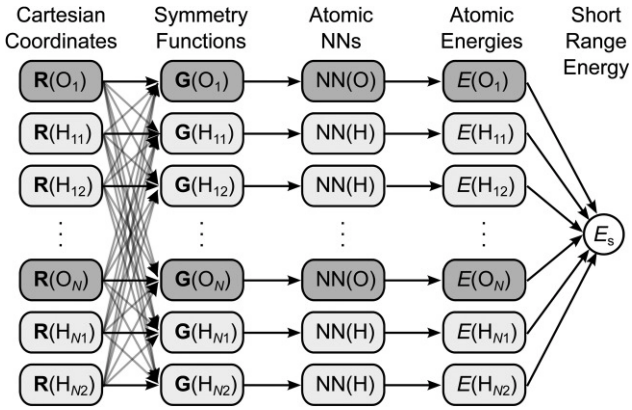

Figure 1.2: Higher dimensional atomic NN for water (This figure was reproduced from Behler et al., "Generalized neural-network representation of high-dimensional potential-energy surfaces"[17]. Published by Physical Review Letters with permission).

For a given set of coordinates, the output of the NN is given by

$$E_i = f_a^2 \left[w_{01}^2 + \sum_{j=1}^{2} w_j^2 f_a^1 \left(w_{0j}^1 + \sum_{u=1}^{2} w_{uj}^1 G_i^u \right) \right] \tag{1.5}$$

where w_{ij}^k is the weight parameter connecting node j in layer k with node i in layer $k-1$, w_{0j}^k is a bias weight that is used as an adjustable offset for the activation functions f_a^k, and G_i^u is a set of symmetry function values [17].

A naive implementation of an NNP would take the Cartesian coordinates of the atoms as the inputs of the NNP directly and output a total energy, as seen in Figure 1.1. There are several disadvantages to this type of NN. Since all the weights are different, the order in which the atoms are entered into the NN matters. For example, in a water molecule, the two hydrogens are chemically equivalent but would have different NN's for this method. Interchanging the same type of atoms will lead to a different total energy. Also, the size of the network is fixed so the NN will only work for systems of the same size. To overcome these issues, Behler and Parrinello developed a new type of NN, seen in Figure 1.2, where the output of the network is the total energy evaluated based the sum of atomic energies [17].

This type of NN takes Cartesian coordinates of atoms, R, and transforms them into symmetry functions, G. Symmetry functions describe the energetically-relevant local environment of each atom. These symmetry functions resolve the interchangeability

and consistency issues, which make them more suitable as the inputs of the NN [17]. Each atom then has its own standard NN that has the same weights for equivalent atoms. The atomic energies for each atomic NN is then summed to calculate the total energy E_s. This type of NNP can be trained and used on a variety of systems of varying size.

There are three key steps to training a neural network. First, a data set is selected. The accuracy of empirical potential is dependent on the amount, quality, and types of interactions included in the data used to train the model. The intended use of the NNP also factors into the choice of a dataset (i.e., protein–ligand binding and protein folding). Once a dataset is chosen, single-point energies for various molecular conformations either at equilibrium and/or non-equilibrium are calculated by the chosen ab initio theory (i.e., DFT or CCSD(T)). Time and practicality must be considered when choosing the theory. Using a higher level of theory to generate the training data could theoretically yield a more accurate NNP, but this has to be weighted against the decrease in extent of the training data if the computational cost to generate results from a higher level method are significantly larger [18].

The Accurate NeurAl networK engINe for Molecular Energies (ANAKIN-ME, or ANI for short) were the first effective general-purpose NNPs for organic molecules built using the NNP architecture developed by Behler and Parrinello [16, 19, 20]. To ensure the NNPs could be used to simulate arbitrary energies of drug molecules, it was trained on an exhaustive set of molecules. The GDB-11 data set includes all molecules composed of elements C, N, O, F, and H with up to 11 non-hydrogen atoms. These compounds are filtered by rules to exclude unstable molecules and those are not accessible synthetically [21]. This data set was used as the basis for the training set used to develop the ANI-1X NNP [16].

Although the GDB-11 data set includes a comprehensive set of molecules, they are exclusively at minimum-energy geometries. To generate a comprehensive set of molecular coordinates to serve as a training set, the developers of the ANI-1 NNP used Normal Mode Sampling (NMS) to generate non-equilibrium structures. The vibrational frequencies of each molecule in the GDB-11 data set that contained up to eight non-hydrogen atoms with elements C, N, O, and H were calculated and linear combinations of Cartesian displacements of these vibrational modes were used to generate structures displaced from the potential energy minimum. Generating the

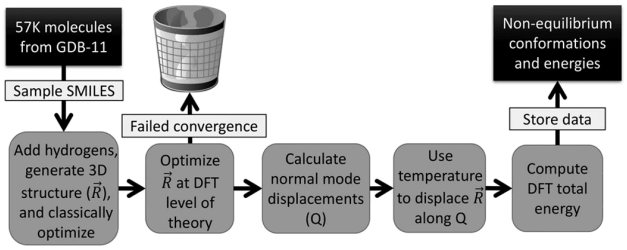

Figure 1.3: Normal mode sampling (This figure was reproduced from Smith *et al.*, "ANI-1: an extensible neural network potential with DFT accuracy at force field computational cost"[16]. Published by The Royal Society of Chemistry under the CC-BY).

training set using NMS ensures that the NNP will be able to describe accurately the configurations that could emerge spontaneously in a molecular dynamics simulation [22].

The first step to NMS is to compute a set of normal coordinates, N_f. Normal coordinates (c_N) are a linear combination of the Cartesian coordinates, c_i, that describes the coupled motion of all the atoms that comprise a molecule.

$$c_N = \sum_{i}^{N_f} c_i \tag{1.6}$$

Next, force constants are obtained by computing the quantum mechanical Hessian at the minimum energy geometry. Then, a set of pseudo-random numbers are obtained, and a displacement for each normal mode coordinate is calculated by assuming a harmonic potential.

$$R_i = \pm\sqrt{\frac{3c_i N_a k_b T}{K_i}} \tag{1.7}$$

where N_a is the number of atoms in an energy minimized molecule, k_b is Boltzmann's constant, T is temperature, and K_i is a pseudo-random number. This displacement is used to scale each normal mode coordinate and then scaled potential energy is calculated using the displaced coordinates. Figure 1.3 is a visual representation of the steps described above.

Once the potential energy surfaces have been calculated, a neural network is trained to reproduce the calculated results. This is done by optimizing the NN using

a cost derivative. The cost derivative depends on the network architecture and the symmetry parameters (i.e., radial and angular shifting parameters, cutoff distances). The program randomly samples structures from the training set in a mini-batch of 1024 molecules, calculating the cost derivative with respect to each weight. The parameters and weights are adjusted such that the cost derivative is as small as possible and the output of the NN is as accurate as possible to the QM level of theory used to generate the training data [22].

Using the higher dimensional NN's and NMS the ANI-1X dataset was created to train and validate the ANI-1X NNP. ANI was created by Smith *et al.* specifically for use in modelling drug molecules and protein ligands [16]. For each structure in the ANI-1X data set, the energy of the structure was calculated using density functional theory (DFT) at the ωB97X/6-31G* level. In total, ANI-1X dataset contains molecules up to eight non-hydrogen atoms that include only the elements C, H, N and O. In total, this training set includes 20 million conformations for approximately 60 thousand molecules.

The ANI-1ccX NNP was developed next using transfer learning (Figure 1.4), where the inner layers of the ANI-1x model were transferred while the input and output layers were trained to CCSD(T)*/CBS data calculated for a subset of the ANI-1X dataset [22]. Transfer learning allows a NN to be trained using a far more limited set of data because only the weights of the input and output layers need to be optimized, and the more extensive ANI-1X dataset provides the data needed to train the inner layers of the NN.

The ANI-2X potential is the most recent ANI NNP. This NNP was trained to reproduce density functional theory (DFT) calculated energies (ωB97X/6-31G*) of 5 million molecular geometries for compounds containing elements C, N, O, H, F, S, and Cl, which overcame the limitation of the ANI-1X and ANI-1ccX NNPs that were limited to the elements C, N, O, and H [20].

These models have shown remarkable transferability; they provide accurate predictions of molecules that are not present in their training sets. Further, they avoid the standard force field approximations where intramolecular interactions are cast into harmonic, cosine, Coulombic, and Lennard-Jones potentials. Replacing MM models with NNP models, specifically ANI NNP's, in simulations of protein-ligand interactions could provide ab initio accuracy at a similar computational cost to molecular

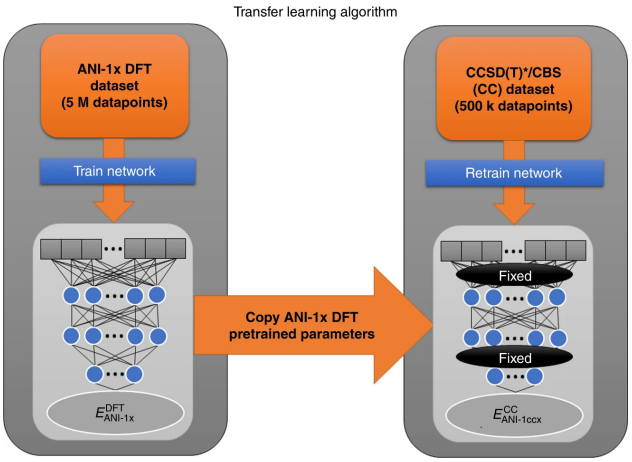

Figure 1.4: Training of ANI-1ccx using transfer learning from ANI-1x NNP (This figure was reproduced from Smith *et al.*, "Approaching coupled cluster accuracy with a general-purpose neural network potential through transfer learning" [22], published by Nature Communications under the CC-BY.)

mechanical models while avoiding the parameterization of individual ligands.

1.3.1 Multiscale Models

Although NNPs could theoretically describe every component of a biochemical system, the current ANI family of NNPs were only trained for neutral, organic molecular structures. Full biochemical systems include water, ions, and proteins, and the intermolecular interactions of these components are essential parts of this description, although ANI NNPs were not developed for modeling intermolecular interactions.

Multiscale methods are an established approach in computational modeling for describing one component of the system using one method, while the rest of the system is described using another method [23]. In chemistry, the most widely used multiscale method is Quantum Mechanical / Molecular Mechanical (QM/MM) models, where a small but important part of the system is described using a QM model, while the balance of the system is described using an MM model [24, 25, 26, 27, 28].

QM/MM methods have been used to describe protein–ligand complexes, where the ligand is described using a QM method while the rest of the system is described with a conventional MM model. This avoids the issue of defining parameters for the intramolecular terms of the ligand, but the high computational cost of an accurate QM

method makes it difficult to perform MD simulations of protein–ligand with sufficient time scales to calculate statistically-meaningful quantities.

In this thesis, I develop a new multiscale computational method, termed NNP/MM. In these models, a component of the system is described using an NNP, while the rest of the system is described using MM. As the NNP provides comparable accuracy at a much lower computational cost than a QM method, this model could provide a systematic way to perform MD simulations of protein–ligand complexes without the need for an MM force field for the intramolecular terms of the ligand.

In this framework, the total potential energy of the system ($\mathcal{V}$) is defined as a sum of the the potential energy of the MM region ($\mathcal{V}_{MM}$), the potential energy of the NNP region ($\mathcal{V}_{NNP}$) and the interaction between these two regions ($\mathcal{V}_{NNP/MM}$):

$$\mathcal{V}(\mathbf{r}) = \mathcal{V}_{MM}\left(\mathbf{r}_{MM}\right) + \mathcal{V}_{NNP}\left(\mathbf{r}_{NNP}\right) + \mathcal{V}_{NNP/MM}\left(\mathbf{r}\right) \tag{1.8}$$

The interactions between the atoms represented using the NNP and atoms represented using MM are calculated using Lennard-Jones and Coulombic potentials (Eqn. 1.9).

$$\mathcal{V}_{NNP/MM}\left(\mathbf{r}\right) = \sum_{i}^{MM}\sum_{j}^{NNP}\frac{q_i q_j}{4\pi\epsilon r_{ij}} + 4\varepsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right] \tag{1.9}$$

where q is the molecular mechanical partial charge of an atom, σ is the Lennard-Jones radius, and ε is the Lennard-Jones well-depth.

1.3.2 Molecular Dynamics Flexible Fitting

Another application of NNP/MM used in this thesis is to describe protein–ligand complexes in Molecular Dynamics Flexible Fitting (MDFF) refinement of cryo-EM structures. MDFF works by supplementing an MD force field ($\mathcal{V}_{force-field}$) with an electrostatic-like potential derived from the cryo-EM density ($\mathcal{V}_{EM}$). The $\mathcal{V}_{force-field}$ can be split into three terms describing the potential energies of the protein ($\mathcal{V}_{protein}$), the ligand ($\mathcal{V}_{ligand}$) and the protein-ligand interactions ($\mathcal{V}_{protein-ligand}$). $\mathcal{V}_{EM}$ biases MD simulations towards structures that are consistent with the cryo-EM electron density maps. Structural models are refined against the EM density by determining

atomic positions that minimize the weighted sum of $\mathcal{V}_{force-field}$ and $\mathcal{V}_{EM}$.

MM models are commonly used as the potential energy functions for $\mathcal{V}_{ligand}$ and $\mathcal{V}_{protein-ligand}$. This complicates the workflow because the parameters for the MM potential of the ligand must be defined by the user and can introduce an additional source of error into the MDFF simulation if the parameters are not optimal. A truly general strategy for resolving the structures of cryo-EM protein-ligand complexes requires an accurate method for calculating $\mathcal{V}_{ligand}$ for all possible ligands.

This QM/MM model for a protein–ligand complex can be also combined with the Molecular Dynamics Flexible Fitting (MDFF) method to refine cryo-EM structures of protein–ligands complexes. Although QM/MM-MDFF can provide well-resolved cryo-EM structures with accuracy, their computational cost is much greater and it is difficult to perform extended MDFF MD simulations with these methods.

Neural network potential (NNP)/MM has already served as a more effective ligand model in molecular dynamics flexible fitting (MDFF) refinement of cryo-EM structures [29]. This method is similar to QM/MM, where the ligand is defined by an NNP and the rest of the system is defined using conventional MM. In MDFF simulations of protein-ligand complexes, NNPs can be used to represent the ligand embedded within a MM model for the protein. The protein-ligand interactions are calculated by pairwise additive Lennard-Jones and electrostatic potentials, as in a mechanically-embedded QM/MM model. The use of neural networks to define ligands in complex systems could provide the answer to finding a model that is flexible, transferable and can be applied to a large variety of systems without the need for parameterization.

1.4 Thesis Outline

In Chapter 2 of this thesis, I explore whether NNPs are effective for calculating challenging potential energy surfaces of drug-like molecules. A testing set of 88 biaryl drug fragments was used to compare the accuracy of ANI-1ccX and ANI-2X NNP's to four conventional force fields. The results of each method is compared with high-level ab initio data for validation. We present the first application of NNP/MM, where I use the NNP to simulate the intramolecular terms of the ligand inside a aqueous solvent described using a MM method, which were then used to the potential of mean force and calculate isomerization rates using enhanced-sampling MD methods.

In Chapter 3, I take the NNP's a step further by using the ANI-1ccX NNP to represent the intramolecular terms of protein-bound ligands when embedded into an MM model for the protein and solvent [30]. This NNP/MM model is investigated to determine if it is a competitive model to the conventional CHARMM General Force Field (CGenFF) MM model. Electron density (ED) maps from X-ray crystallography data are used as a measure of accuracy, wherein the ANI-1ccX NNP ligand pose and the CGenFF ligand pose is superimposed onto the ED map. We also compare the conformational Gibbs energy of binding for these ligands, calculated using conventional MM and NNP/MM.

In the final chapter of this thesis, Chapter 4, an NNP/MM-MDFF is employed on several protein-ligand systems that have low resolution cryo-EM structures in the hopes to get a better refinement of the structure and ligand positions. The ANI-2X NNP is used to represent the intramolecular terms of the ligand, as in Chapter 3, and MDFF is used to bias the simulation to match the cryo-EM data. We demonstrate this method on a set of published cryo-EM protein–ligand structures.

Chapter 2

Using Neural Network Potentials to Model Torsional Potential Energy Surfaces of Biaryl Drug Fragments

The content of this chapter has been published in the *Journal of Chemical Information and Modelling*: Lahey, S., Phuc, T-N., Rowley, C.N. Benchmarking Force Field and the ANI Neural Network Potentials for the Torsional Potential Energy Surface of Biaryl Drug Fragments. Published December 2020.

2.1 Introduction

Many natural products and drug molecules contain biaryl motifs and rotation around the bonds connecting them that introduces a torsional degree of freedom in these molecules. The potential energy surface for the rotation around these bonds varies due to conjugation, steric interactions, intramolecular hydrogen bonding, and electron repulsion. These effects determine the equilibrium conformations held by the molecules and the rates of conformational isomerization. Accurate computational models for these torsional potential energy surfaces is essential for modeling conformational dynamics and protein–ligand binding. Typically conventional MD force

fields are used to model these energy surfaces.

Torsional potentials are generally defined as a sum of cosine functions with a variety of periods, offsets, and amplitudes,

$$\mathcal{V}_{\text{torsion}}(\theta) = \sum_i k_i \left(1 + \cos\left(n_i\theta + \phi_i\right)\right) \tag{2.1}$$

Additionally, intramolecular non-bonded interactions can significantly affect torsional potential energy surfaces. The treatment of these interactions varies with the force field used and the simulation options must be carefully chosen to match the selected force field. Intramolecular interactions between atoms separated by three bonds (i.e., 1,4 nonbonded interactions) can have a large effect on the torsional potential energy surface. Some force fields are designed to be used with the 1,4 Coulombic and Lennard-Jones interactions between the atoms reduced. There is significant variety in how these interactions are calculated: in the Generalized AMBER Force Field (GAFF), 1,4 electrostatic interactions are scaled by a factor of 0.833, while 1,4 Lennard-Jones interactions are scaled by a factor of 0.5 [31]. The OpenFF standard allows these factors to be specified for given atomic pairs, but default to the same value as GAFF [32]. 1,4 scaling factors are not used at all with CGenFF [33, 34]. The OPLS force field employs a scaling factor of 0.5 on both the Coulombic and Lennard-Jones 1,4 interactions but also uses the geometric combination rule for Lennard-Jones radii parameters while the other force fields use the arithmetic mean [35]. Intramolecular non-bonded interactions between atoms beyond the 1,4 interactions also significantly affect torsional potential energy surfaces due to steric or electrostatic interactions [36].

Innovations have been introduced to simplify this process. New methods have been developed to determine parameters for the possible permutations of atom types automatically. Where a force field lacks a specific term, this parameter can be fit to an ab initio potential energy surface. The SMIRKS Native Open Force Field (SMIRNOFF) format assigns force field parameters by searching for chemical substructures with SMIRKS format.[37] These methods still require the definition of extensive sets of parameters and follow most of the standard approximations inherent to conventional force fields.

NNPs could provide a solution to this problem. They can provide the same accuracy of QM calculations but at a fraction of the computational cost. Although

Devereux *et al.* reported that the ANI-2X NNP was effective for some torsional profiles, it is not clear if they will be as effective as the standard MM models for the type of torsional profiles present in drug-like molecules (esp. biaryl torsions). Jorgensen and coworkers published a test set of biaryls present in drug molecules and drug candidates, which provides an extensive, diverse, and relevant test set for assessing the accuracy of computational methods for predicting biaryl torsional potential energy surfaces [38]. This will help establish whether ANI potentials are viable as replacement for MM models.

In this chapter, we compare the performance of four MM models (GAFF, CGenFF, OPLS, and OpenFF) and two NNPs (ANI-2X and ANI-1ccX) for the prediction of biaryl torsional potential energy surfaces calculated using high-level ab initio method (CCSD(T1*)/CBS).

2.2 Computational Methods

2.2.1 Test Set

The test set of biaryls used here is largely the same as the biaryl torsional test set developed by Dahlgren *et al.*[38] This test set was generated by extracting biaryl fragments from drug and drug-like molecules. Because some of the methods used here are not designed for use with charged compounds (e.g., the NNPs and the CGenFF parameterization server), all compounds were modeled in their neutral protonation state and the one charged compound in this test set (1-phenylpyridazin-1-ium) was excluded in our study. Also, 5-phenyl-1,2,4-oxadiazole, a fragment of the drug ozanimod, was added to our test set. The structures of the molecules in the test set and their associated numbering are illustrated in Figure 2.1. The structures, topology, and parameter files of this test set are available on our GitHub repository [39].

2.2.2 Molecular Mechanical Parameterization and Calculations

The CGenFF, GAFF, and OPLS potential energy surfaces were calculated using relaxed scans with CHARMM (i.e., the torsional degree of freedom was restrained and

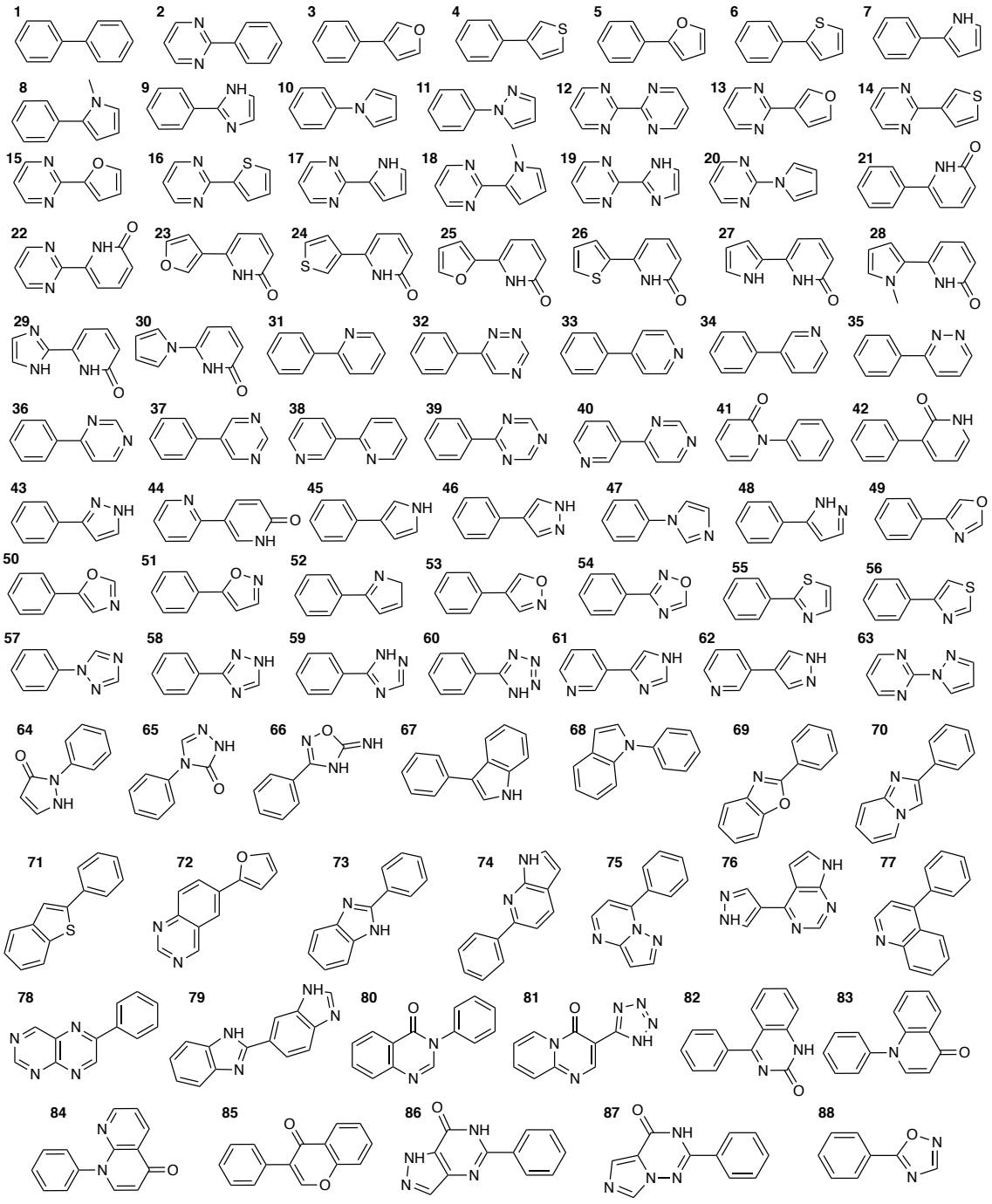

Figure 2.1: Test set of biaryl fragments.

the remaining degrees of freedom were energy-minimized). These calculations were performed using CHARMM 41b2 [40].

CGenFF calculations employed the CGenFF force field version 2.2.0 [33]. The charge, topology, and parameters were assigned using the CGenFF parameterization server. 1,4 non-bonded interactions calculated using the CGenFF force field were not scaled [12].

The GAFF parameters were assigned using AmberTools [31]. Atomic charges in the GAFF model were calculated using the Restrained Electrostatic Potential method (RESP) using Hartree–Fock (HF) with the 6-31G* basis set as the target quantum mechanical (QM) electrostatic potential [41]. 1,4 nonbonded interactions in the GAFF force field calculations were scaled by a factor of 0.833.

The OPLS [42] topology and parameters were generated using the LigParGen server [43]. The CM1A-LBCC charge model was used [44]. 1,4 nonbonded interactions in the OPLS force field were scaled by a factor of 0.5. The geometric combination rule was used for the σ Lennard-Jones parameters.

The OpenFF charges and parameters were generated using the SMIRNOFF Open Force Field version 1.1.1 with the code name "Parsley" [32]. The relaxed potential scans were performed in OpenMM package version 7.4.1 [45] with external harmonic restraint on the interest dihedral angle. The harmonic strength was set to ensure the torsional angle variance was less than 0.02°.

2.2.3 ANI Potential Energy Surfaces

The potential energy surfaces for the ANI-2X [20] and ANI-1ccX [22] NNPs were calculated using TorchANI [46] interfaced with the External feature of Gaussian 09[47] through a python script. This script is available for download from our GitHub repository [48]. Each surface was calculated from complete scans in the forward and reverse directions for rotation around this torsion, where the minimum of the energies from each scan was used to construct the PES.

2.2.4 QM Potential Energy Surfaces

The QM potential energy surfaces were calculated using a relaxed scan of the biaryl torsional degree of freedom using the RIMP2/def2-TZVP level of theory [49, 50, 51]. ORCA 4.2.1[52] was used to calculate the single point potential energy of these configurations using the composite CCSD(T)*/CBS method described by Smith *et al.*[22] The iterative triples method, DLNPO-CCSD(T1) [53], was used because the DLNPO-CCSD(T) torsional potential energy surfaces of some molecules were discontinuous due to differences in the conventional triples correction energy.

2.2.5 NNP/MM MD Simulations

An NNP/MM simulation was performed of ozanimod (5-[3-[(1S)-1-(2-hydroxyethylamino)-2,3-dihydro-1H-inden-4-yl]-1,2,4-oxadiazol-5-yl]-2-propan-2-yloxybenzonitrile) in an explicit aqueous solvent. In this method, the intramolecular potential energy of the ligand ($\mathcal{V}_{NNP}(\mathbf{r}_{NNP})$) is calculated using the ANI-1ccX NNP, while the potential energy of the solvent is represented using a conventional MM model ($\mathcal{V}_{MM}(\mathbf{r}_{MM})$). The potential energy of the system is the sum of these two components and an additional term corresponding to the interaction of the NNP and MM atoms ($\mathcal{V}_{NNP/MM}$) (Eqn. 2.2).

$$\mathcal{V}(\mathbf{r}) = \mathcal{V}_{MM}(\mathbf{r}_{MM}) + \mathcal{V}_{NNP}(\mathbf{r}_{NNP}) + \mathcal{V}_{NNP/MM}(\mathbf{r}) \tag{2.2}$$

The interactions between the atoms represented using the NNP and atoms represented using MM are calculated using Lennard-Jones and Coulombic potentials (Eqn. 1.8).

In these simulations, the solvent–solute Lennard-Jones parameters (σ and ϵ)[31] were generated using the Lorentz–Berthelot combination rules using the GAFF parameters for the solute and the TIP3P-FB parameters for the water molecules. The partial atomic charges (q) of the solute were calculated using the RESP method [41].

The simulations were performed using NAMD 2.13 [54] interfaced with TorchANI [46] using the NAMD-ANI interface script [48]. The total number of water molecules was 2158. The dimensions of the periodic simulation cell were 48.7 Å $\times$ 38.5Å $\times$ 34.8 Å. The ozanimod molecule was represented using the ANI-1ccX NNP and the solvent

was represented using the TIP3P-FB water model [55]. The GAFF Lennard-Jones parameters were used for the solute–solvent non-bonded interactions [31]. Intermolecular electrostatic interactions were calculated based on RESP charges assigned to the atoms of the ozanimod [41]. In the NNP/MM framework, the intramolecular interactions of the ligand are calculated using only the NNP [30]. A 2 fs timestep was used. Bonds containing hydrogen atoms were constrained using the SHAKE algorithm. In these simulations, the temperature was coupled to a 298.15 K bath using a Lowe–Anderson thermostat [56].

Calculation of the potential of mean force (PMF) for the rotation of the N-C-C-C biaryl torsion was performed using the adaptive biasing force (ABF) [57, 58] and umbrella sampling [59, 60]. Umbrella sampling simulations were performed on the N-C-C-C biaryl torsional coordinate with a harmonic bias potential with a force constant of 0.25 kcal/mol/degree^2 with windows at 5° spacings. Each window was simulated for 2 ns with a 200 ps equilibration. The PMF was constructed from the umbrella sampling time series using the Weighted Histogram Analysis Method (WHAM) [61, 62, 63]. In these simulations, temperature was regulated using a Langevin thermostat with a bath temperature of 298.15 K and a damping coefficient of 1 ps^{-1}.

2.3 Results and Discussion

2.3.1 Overall Performance

Only general trends and some notable examples are discussed here, although the plots of potential energy surfaces of all 88 torsional rotations are included in the supporting information of Lahey *et al.*[64] We define two metrics for the overall performance of each method. The root-mean-squared-deviation (RMSD) of a method for a given torsion is calculated as the root-mean-square of the potential energy calculated using a given method ($\mathcal{V}_{i,\text{method}}$) relative to the CCSD(T1)*/CBS//MP2/def2-TZVP reference value ($\mathcal{V}_{i,CCSD}$), calculated at 5° increments between 0° and 360°,

$$RMSD = \sqrt{\frac{1}{n}\sum_{j}^{n}\frac{1}{N_{bins}}\sum_{i}\left(\mathcal{V}_{i,\text{method}} - \mathcal{V}_{i,CCSD}\right)^2} \tag{2.3}$$

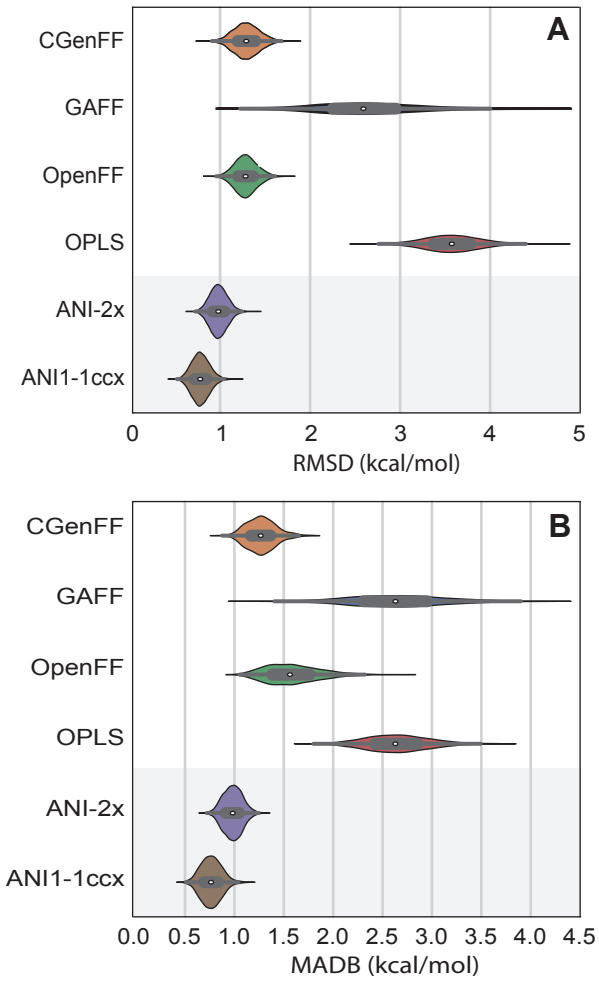

Figure 2.2: A: RMSD of the PES (Eq. 2.3) for each method. B: Mean absolute deviation of the rotational barrier height of each method. The CCSD(T1*)/CBS profiles are used as the reference. Distributions are calculated using bootstrap analysis. NNP methods (ANI-2X and ANI-1ccX) are shaded.

where N_{bins} is the number of points calculated on the potential energy surface ($N_{bins} = 72$) and n is the number of PESs in the test set ($n = 88$).

Our second metric is the mean absolute deviation of the torsion rotational barrier height (MADB) for each method, relative to the CCSD(T1)*/CBS//MP2/def2-TZVP barrier height (Eqn. 2.4).

$$MADB = \frac{1}{n} \sum_{j}^{n} |\Delta\mathcal{V}^{\ddagger}_{CCSD} - \Delta\mathcal{V}^{\ddagger}| \tag{2.4}$$

where $\Delta\mathcal{V}^{\ddagger}$ is the barrier height of torsional rotation, defined as the difference between the minimum and maximum energy point on the PES.

This analysis is sensitive to the composition of the test set. The arbitrary inclusion of some of the potential energy surfaces that a given method performs poorly for could shift the results significantly. To estimate the uncertainty of these rankings due to the composition of the test set, we have used bootstrap error analysis, where an alternative set of 88 compounds were chosen randomly-with-replacement from the total set. This process was repeated 10,000 times and these sets were used to calculate a distribution for each metric.

The bootstrap analysis of the averages of these metrics for the test set are presented in Figure 2.2. Only profiles that were available for all methods were included in this analysis, so the sulfur-containing compounds (**4**, **6**, **14**, **16**, **24**, **26**, **55**, **56**, and **71**) were not included because the ANI-1ccX NNP is not defined for this element. The results for the sets including sulfur-containing compounds generally follow the same trends, with ANI-2X performing as well or better than the best force-field methods. These results are included in Supporting Information. The difference of means for each pair of methods was calculated to quantify the difference in performance and are presented in Table 2.1. The standard error of these differences were uniformly < 0.01 kcal/mol.

For both metrics, the ANI-2X and ANI-1ccX methods outperform all four force fields. The ANI-2X and ANI-1ccX NNPs have a similar level of performance on this test set, although the ANI-1ccX barrier heights are more accurate than the ANI-2X by 0.2 kcal/mol on average. Notably, the ANI-1ccX MADB is only 0.7 kcal/mol, indicating that the goal of "sub-kcal" accuracy has already been achieved for these

Table 2.1: Difference of means for the RMSD (top) and MADB (bottom) of the test set (excluding sulfur-containing compounds). Positive values indicate that the first method (column) agrees more closely with the CCSD(T1*) reference values than the second method (row). Standard errors are omitted because they are all negligible (< 0.01 kcal/mol).

	GAFF	OpenFF	OPLS	ANI-2X	ANI-1ccX
CGenFF	-0.44	0.07	-1.21	0.31	0.32
GAFF		0.50	-0.78	0.75	0.76
OpenFF			-1.28	0.25	0.26
OPLS				1.52	1.53
ANI-2X					0.01

	GAFF	OpenFF	OPLS	ANI-2X	ANI-1ccX
CGenFF	1.33	0.00	-2.29	0.31	0.51
GAFF		1.34	-0.96	1.64	1.84
OpenFF			-2.29	0.31	0.50
OPLS				2.60	2.80
ANI-2X					0.20

PESs.

Of the force field methods, CGenFF and OpenFF have similar levels of accuracy and are both significantly more accurate than the others. The mean signed error (MSE) in the barrier height is positive for all the models except CGenFF, which tends to underestimate barriers by 0.2 kcal/mol. Notably, GAFF has an MSE of 1.2 kcal/mol and OPLS has an MSE of 2.5 kcal/mol, indicating a significant tendency to overestimate torsional barriers. The distributions generated by bootstrap error analysis are narrow for ANI-2X, ANI-1ccX, CGenFF, and OpenFF indicating a fairly consistent level of performance across the test set. The distributions for OPLS and GAFF are much broader, indicating the performance on some compounds is more varied.

The relative accuracy of these methods can also be quantified by ranking which method provides the PES with the lowest RMSD or the most accurate barrier. These rankings are presented in Figure 2.3 (top). By these measures, the ANI-1ccX NNP is most accurate. It should be noted that the ANI-2X method predicts barrier heights

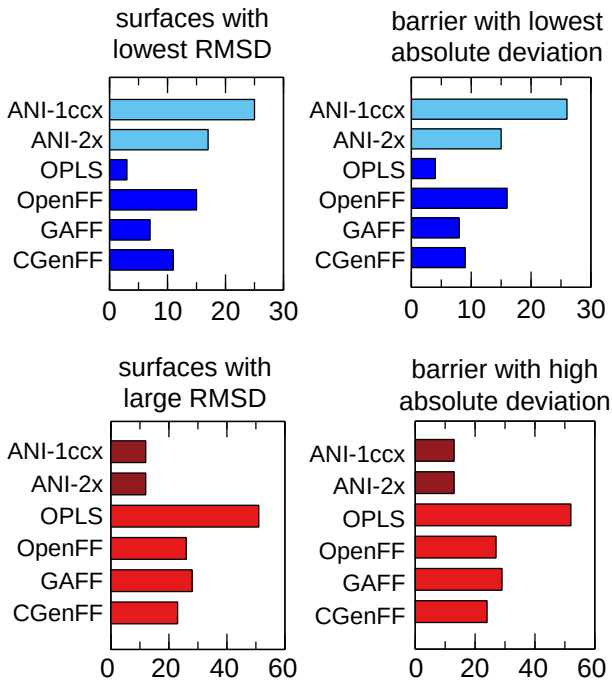

Figure 2.3: Top left: Number of torsional PESs where a given method has the lowest RMSD. Top right: Number of torsional PESs where a given method has the lowest barrier height deviation. Light blue color indicates methods is a NNP and dark blue indicates method is a conventional force field model. Bottom left: Number of torsions where a method gives a high RMSD (i.e., on average, more than 1 kcal/mol in error at each point on the PES). Bottom right: Number of torsional PESs where a method predicts the barrier height inaccurately (i.e., more than 2 kcal/mol in error). To allow comparison of the ANI-1ccX method, S-containing compounds were not included ($n = 79$).

with similar results to ANI-1ccX, so it has nearly the same accuracy. Although ANI-1ccX is ranked higher because it is generally incrementally more accurate than ANI-2X.

Finally, we can also assess the methods according to the number of torsional PESs where a method performs poorly. These rankings are presented in Figure 2.3 (bottom). For the RMSD criteria, this is defined as a mean squared deviation (MSD) per point on the surface that is greater than 1 kcal/mol and for the barrier height a poorly performing is defined as one where the predicted barrier height is in error by 2 kcal/mol or more. Based on these metrics for poor performance, the ANI-2X and ANI-1ccX NNP models are also superior compared to the force field models, with the ANI-1ccX methods demonstrating poor performance for only 10 PESs and 10 barriers lower than the best force field (CGenFF). Among the force fields, the CGenFF performs poorly for the fewest torsions, followed by GAFF, OpenFF, and OPLS.

This highlights a major advantage of the NNP methods - the number of cases where they perform poorly is small. The strategy of training these potentials to reproduce molecular energies in general rather than specific interactions results in methods that are robust for PESs outside their training sets. It should be noted that none of the biaryl compounds in this test set were part of the ANI-2X or ANI-1ccX training sets, so the success of these methods show that they are remarkably robust and provide accurate predictions for molecules and bonding motifs that they were not explicitly trained to describe. This success for molecules outside of their training set is a significant advantage for high-throughput screening of protein–ligand binding, where the validation and possible reparameterization of a force field is too time-consuming.

2.3.2 CGenFF

Overall, CGenFF performs as well or better than the other force fields; however, there are some instances where the barriers are predicted to be much lower than the CCSD(T1)*. These examples are shown in Figure 2.4. This is apparent in N-rich heterocyclics, like **20** and **39**, suggesting that the parameters for C-CA-CA-NA and N-CA-NA-CA dihedrals have a maximum barrier height that is too small. For example, the OPLS force field predicts a torsional barrier of **20** more accurately (9 kcal/mol),

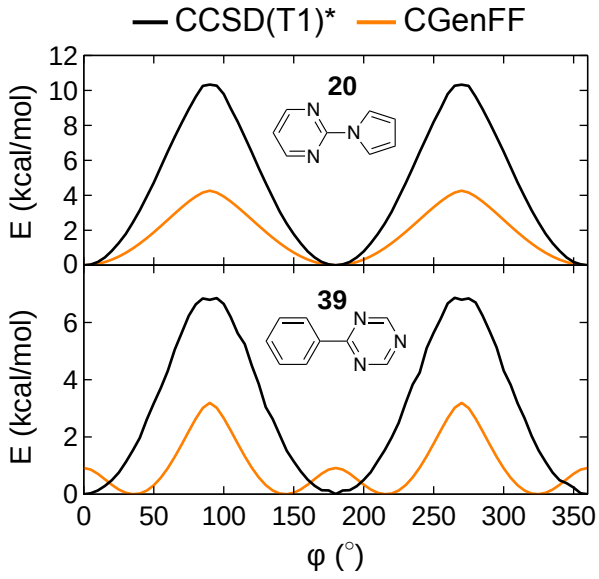

Figure 2.4: The torsional potential energy surfaces of **20** and **39** are examples where the CGenFF force field is an inaccurate model.

in part because it uses a C-N-C-N dihedral potential with a maximum of 3.6 kcal/mol instead of the barrier maximum of 1.8 kcal/mol used by CGenFF. Adjustment of a handful of biaryl dihedral terms would improve the accuracy of CGenFF even further.

2.3.3 OpenFF

OpenFF is the "newest" of the force fields evaluated here and is designed to avoid duplicate or unnecessary parameters. As a result, there are far fewer parameters in the current version of OpenFF compared to the other force fields (i.e., 342 parameters in OpenFF vs more than 6000 parameters for CGenFF). Nevertheless, based on the RMSD and barrier height metrics, it generally performs better than GAFF and OPLS for this test set and performs as well or better than CGenFF. Because relatively few "specific" torsional parameters are currently defined, much of its success is derived from the general biaryl potentials, 1,4 Lennard-Jones and electrostatic parameters. This strategy is less successful for torsions containing aromatic nitrogen atoms in the ortho or ipso positions, such as **10** and **12**. The torsional PESs of these N-containing

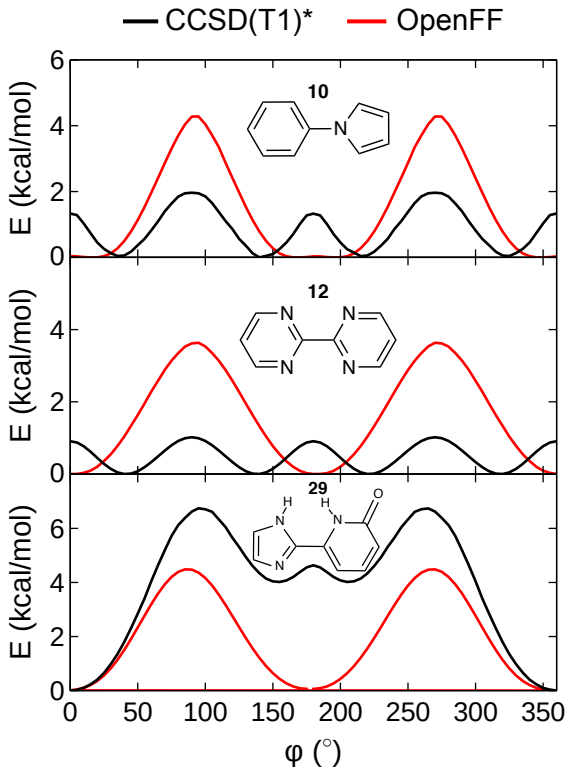

Figure 2.5: The torsional potential energy surfaces of **10**, **12**, and **29** are examples where the OpenFF force field is an inaccurate model.

aromatics are influenced by complex hyper conjugative and electron-repulsive interactions, which require explicit parameterization for the force field to describe quantitatively. These examples are shown in Figure 2.5. The OpenFF model defines a systemic process for improving its description of torsional interactions through fitting to QM surfaces, so subsequent revisions are likely to show even better performance for these surfaces.

2.3.4 GAFF

An instance where the GAFF force field significantly deviates from the reference PES is where the aryl linkage is through the nitrogen of a pyrrole group. Repulsion between the pyrrole non-bonded pair and the π system of the benzene ring destabilizes planar conformations, but lone-pair–CH repulsion occurs when the rings are perpendicular, so the minimum energy conformations occur at 30° deviations from planarity.

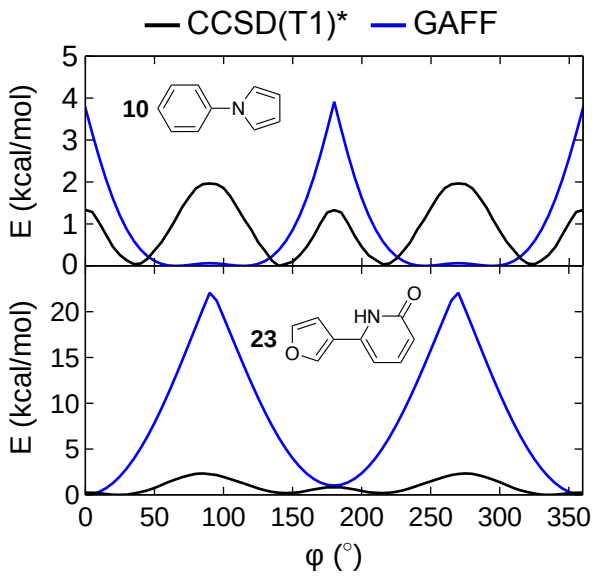

Figure 2.6: The torsional potential energy surfaces of **10** and **23** are examples where the GAFF force field is an inaccurate model.

In contrast to this, the GAFF force field predicts a broad minimum corresponding to conformations where the rings are non-planar. The CA–CA–NA–CA torsional parameter is the immediate cause of this issue. Examples of GAFF's poor performance are seen in Figure 2.6.

The GAFF force field significantly overestimates the barrier for rotations where there is an amide NH group in the ortho position of one of the rings. For example, in **23**, GAFF predicts a barrier of 22 kcal/mol, while it is only 2 kcal/mol with CCSD(T1)*. This issue is present in **24**, **25**, **26**, **27**, **28**, and **29**.

2.3.5 OPLS

The OPLS model performs relatively poorly on this test set. In many cases, this is due to a significant overestimation of the rotational barrier. Examples are shown in Figure 2.7. The mean signed error for the OPLS barrier is 2.5 kcal/mol, indicating that the tendency to overestimate torsional barriers is systematic in this force field. This is evident in the PES for **7** and **12**, where the barrier is overestimated by a factor of 2 and 6, respectively. In other cases, the topology file generated by the LigParGen server includes torsions that result in asymmetric potential energy surfaces on torsions

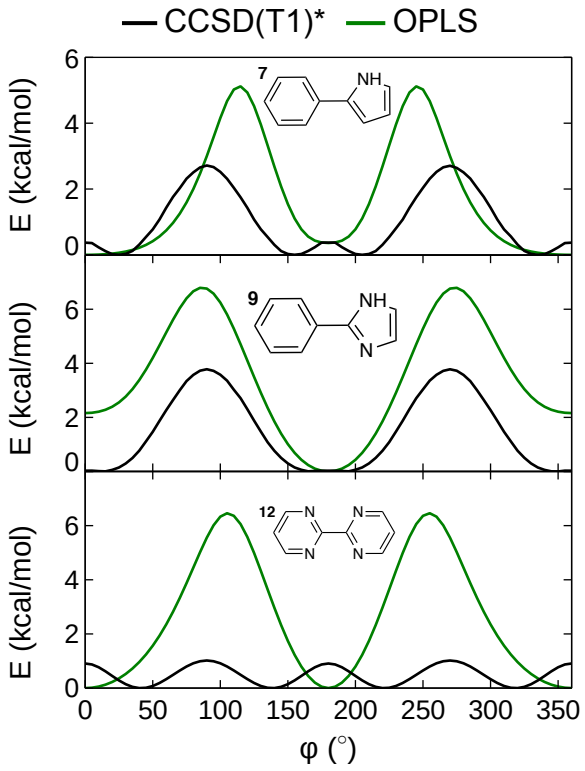

Figure 2.7: The torsional potential energy surfaces of **7**, **9**, **12** are examples where the OPLS force field is an inaccurate model.

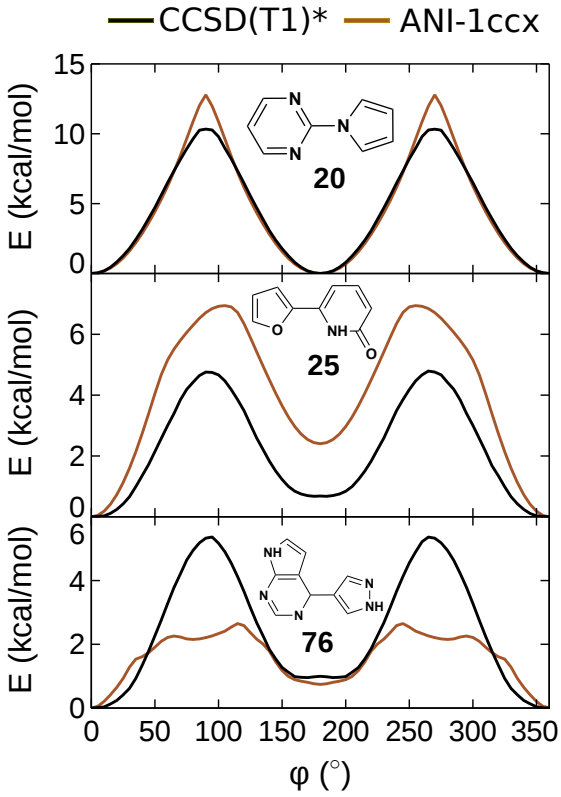

Figure 2.8: The torsional potential energy surfaces of **20**, **25**, and **76** are examples where the ANI-1ccX NNP is an inaccurate model.

that should be symmetric. The PES of **9** is an example of this effect. Dahlgren *et al.* showed that new parameters could improve the performance of the OPLS force field specifically for biaryl torsions [38], so including these parameters in the models generated by LigParGen could immediately improve the accuracy of this model.

2.3.6 ANI

The ANI-2X and ANI-1ccX NNPs generally outperform the MM models and rarely fail to provide a reasonably accurate surface. The ANI-1ccX NNP gives incrementally greater accuracy than ANI-2X, although it only supports a smaller set of ligands because only the C, N, O, and H elements are defined for it. Examples of where ANI-1ccx performed poorly are shown in Figure 2.8 The relative stability of the cis and trans conformations of **25** is overestimated by the ANI-1ccX potential and the barrier to rotation is significantly overestimated. The PES of **76** is generally irregular

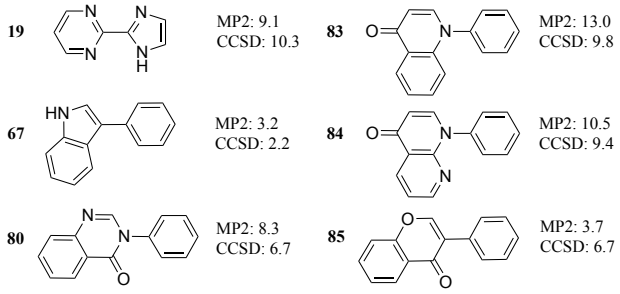

Figure 2.9: Test set structures where the MP2 barrier height differs by 1 kcal/mol or higher from the CCSD(T1)* surface.

and the barrier is significantly underestimated. The lack of explicit electrostatic terms limits the accuracy of these NNPs when the relative stability of two conformations depends on a long-range polar interaction.

2.3.7 MP2 Failures

Torsional potentials of force fields are often parameterized to reproduce MP2 potential energy surfaces because MP2 is a computationally tractable ab initio method with analytical gradients. Having calculated the energies at both the CCSD(T*)/CBS//MP2/def2-TZVP and MP2/def2-TZVP levels, we can test if these levels of theory provide the same level of accuracy for these torsional profiles. The MP2/def2-TZVP PESs are generally in very good agreement with the CCSD(T1)*/CBS PESs, with a RMSD of 0.1 kcal/mol and a MADB of 0.3 kcal/mol.

Although the agreement between MP2 and CCSD(T1)* is generally close, of the 88 torsions in the test set, the MP2 and CCSD(T1)* barrier heights differed by more than 1 kcal/mol in six instances, as seen in Figure 2.9. In four instances (**83**,**80**,**84**, and **85**), one of the aryls in the rotation was bicyclic, so the transition state occurs when the structure is planar and a steric interaction arises between the ortho hydrogen of the phenyl ring and an atom in the ortho position of the other ring. The CCSD barrier is lower than the MP2 barrier in 3 out of 4 of these examples, suggesting that MP2 overestimates the strength of this type of steric repulsion. The other examples are nitrogen-containing heteroaromatics (**19**, **67**). The MP2 barrier is lower than the CCSD(T1)* barrier in one of these examples, suggesting that there is a small tendency for MP2 to underestimate the electronic repulsion associated with π lone pairs. The

origin of the deviation in **19** and **67** is less evident, although both are N-containing heteroaromatics.

This raises questions about the common practice of using MP2 potential energy surfaces as the target data for fitting force field torsional potential energy surfaces. Although MP2 is in good agreement with CCSD(T1)* surfaces for most of the molecules, the MP2 and CCSD(T1)* barrier heights differed by 1 kcal/mol or higher for 6 of the torsions. This suggests that parameterizing a force field to reproduce the torsional surfaces of an NNP trained for CCSD(T) data could be advantageous because it is not necessary to calculate CCSD(T1)* potential energy surfaces for each torsion in the ligand.

The ANI-1ccX NNP is incrementally more accurate than the ANI-2X NNP. The ANI-1ccX NNP was developed through a transfer learning approach, where input and output layers of the ANI-1X NNP were retrained using a subset of CCSD*(T1)/CBS data, while the ANI-2X potential was trained to DFT (ωB97X/6-31G*) data exclusively.

2.3.8 Molecular Dynamics of Torsional Rotations

The large barriers to torsional rotations present in some biaryl compounds can result in large activation energies for conformation isomerization. Consequently, in molecular dynamics simulations of drug compounds containing biaryls, the timescales associated with rotation around the biaryl bond can be much longer than other, more facile, types of conformational isomerization. The atomic forces of the ANI NNP can be calculated analytically using the auto-differentiation (autograd) feature of the Torch library, so geometry optimizations and MD simulations can performed simply and efficiently by calling this differentiation routine in the TorchANI library [46].

To demonstrate that these methods are practical for use in molecular dynamics simulations for real drug molecules, we have performed simulations of ozanimod in an explicit aqueous solution. Ozanimod is a drug for the treatment of multiple sclerosis that binds the sphingosine 1-phosphate receptor [65]. The rotation around the bond connecting the isopropyl benzonitrile and the diazafuran has a significant barrier height because of the conjugation between the rings and low steric repulsion in planar conformations. The time series of this angle over the course of a 10 ns simulation

shows that this molecule undergoes conformational isomerization through rotation around this torsional degree of freedom on the multi-nanosecond timescale. This can be seen in Figure 2.10. These simulations are tractable using conventional Graphical Processing Unit hardware; these NNP/MM MD simulations required 0.34 days/ns on an Intel(R) Core(TM) i7-8700 with an NVIDIA Titan Xp GPU. This performance will likely improve when NNPs are implemented directly into simulation codes and when GPU-accelerated implementations of symmetry function calculations are completed.

To investigate this isomerization further, the potential of mean force (PMF) of this degree of freedom was calculated using umbrella sampling and adaptive biasing force[57, 58] molecular dynamics simulations. Both methods provide similar PMFs (Figure 2.10) and can be used with the TorchANI NNP/MM interface with NAMD without modification.

Using the umbrella sampling PMF, the rate constant of isomerization (k_{KS}) was calculated using Kramers–Smoluchowski transition state theory [66, 67].

$$k_{KS} = D_{TS} \frac{\sqrt{|W''(q_{\text{minimum}}) \cdot W''(q_{TS})|}}{2\pi k_B T} e^{-\Delta W^{\ddagger}/k_B T} \tag{2.5}$$

where D_{TS} is the diffusion coefficient at the transition state, W'' is the second derivative of the PMF, q_{TS} is the position on the reaction coordinate where the maximum of the PMF occurs, q_{minimum} is the position on the reaction coordinate where the minimum of the PMF occurs, and $\Delta W^{\ddagger}$ is the barrier height of the PMF.

In these simulations, the TIP3P-FB water model was used, which predicts a viscosity close to the experimental value [55], allowing more accurate predictions of diffusion rates in aqueous solutions[10] and more accurate predictions of the solvent friction on the reaction coordinate.

D_{TS} was calculated using the generalized Langevin approach [68, 69, 70], where a strong harmonic potential was used to restrain the simulation to the transition state and the diffusion coefficient was determined by the rate of relaxation of the position autocorrelation function of the time series of the reaction coordinate (q).

$$D_{TS} = \frac{\text{var}(q)^2}{\int_0^{\infty} \langle q(0) \cdot q(\tau) \rangle \mathrm{d}\tau} \tag{2.6}$$

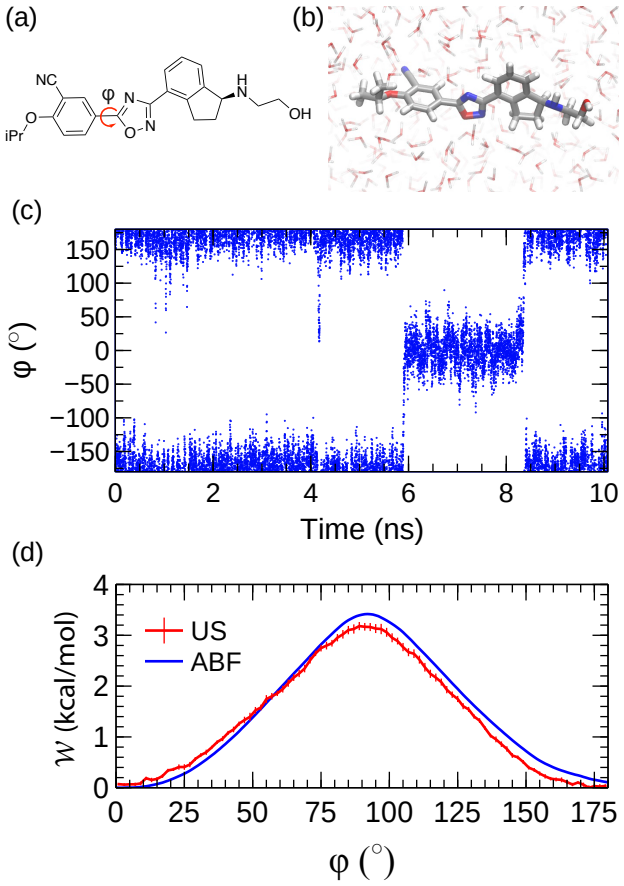

Figure 2.10: (a) Structure of ozanimod with ϕ torsional angle indicated. (b) Representative structure of ozanimod in solution (c) Time series of ϕ angle from 10 ns MD simulation of ozanimod in aqueous solution (d) ANI-1ccX/TIP3P-FB potential of mean force for rotation around ϕ dihedral.

Table 2.2: Rate theory prediction of isomerization of ozanimod

	ANI-1ccX	CGenFF
D_{TS} (rad^2/ns)	67.5	102.4
$\Delta W^{\ddagger}$ (kcal/mol)	3.2	4.7
rate (ns^{-1})	4.3×10^{-1}	6.4×10^{-2}

This theory predicted a rate of isomerization of 4.30×10^{-1} ns^{-1}, which is generally consistent with the slow rates of isomerization observed in the NNP/MM MD simulation (Table 2.2). These simulations demonstrate that the ANI NNPs can be used immediately to describe the dynamics of slow degrees of freedom of arbitrary drug-like organic molecules using existing simulation methods. In comparison, the rate predicted by the most reliable MM model, CGenFF, is roughly 7 times slower. This is predominantly due to the larger activation energy (3.2 kcal/mol for ANI-1ccX/MM vs 4.7 kcal/mol for CGenFF), although the diffusion coefficient of the system along the reaction coordinate at the transition state also differs.

2.4 Conclusions

Force field and NNP methods were evaluated for their ability to predict the potential energy surfaces of biaryl torsions found in drug and drug-like molecules ($n = 88$). As these torsions are important features for the structure and dynamics of these molecules, efficient but accurate computational models of these terms are essential for accurate protein–ligand binding simulations. In comparison to high-level ab initio reference data, the ANI-1ccX NNP was the most accurate method and generally predicted barrier heights within 1 kcal/mol, although this method only supports the elements C, N, O, and H. The ANI-2X NNP had a comparable level of accuracy and can be used with element C, N, O, H, S, F, and Cl. Significantly, these NNPs provided accurate models in most of the cases and provided poor descriptions in relatively few cases. The robustness and reliability of these NNPs without specific parameterization is particularly useful for simplifying modeling workflows, although the NNPs examined here currently are not appropriate for simulations of charged compounds, which limits their applicability somewhat.

The force field methods were less accurate, although there were significant differences in the accuracy of the force fields. The CGenFF was most accurate, followed by the OpenFF, GAFF, and OPLS. The OpenFF model is notable because it performed relatively well despite having been parameterized with relatively little data and including relatively few parameters. Although MP2 potential energy surfaces were generally in good agreement with the CCSD(T1)* reference values, there were significant differences in 6 instances, suggesting force fields and NNPs should be parameterized to reproduce CCSD(T1)* data for optimal and comprehensive accuracy.

The NNP/MM method was used to simulate the conformational isomerization of the biaryl-containing drug molecule ozanimod. Multi-nanosecond molecular dynamics simulations in an explicit aqueous solvent were performed, as well as umbrella sampling and adaptive biasing force enhanced sampling techniques. These free energy methods can be used in NNP/MM simulations through the NAMD–TorchANI interface, which makes a diverse set of simulation methods available without modification and allows for facile construction. This provides a method for computationally-efficient but highly-accurate models for the intramolecular potential energy surfaces of ligands within biomolecular simulations without relying on a parameterized force field.

Chapter 3

Simulating Protein–Ligand Binding with Neural Network Potentials

The content of this chapter has been published in Chemical Science: Lahey, S., Rowley, C.N., Simulating Protein–Ligand Binding with Neural Network Potentials. Published Jan 2020.

3.1 Introduction

Molecular simulation of the binding of small molecules to proteins has provided computational prediction and rationalization of the affinity and selectivity of drugs with their targets. These simulations rely on molecular mechanical (MM) force fields to describe the intra and intermolecular interactions of the solvent, protein, and ligand. These "force fields" are constructed from simple mathematical functions that approximate the potential energy surface of the protein–ligand complex. A force field requires the definition of a large set of parameters, which are typically chosen to yield the closest agreement with empirical or quantum chemical data.

As shown in Chapter 2, NNPs perform remarkably well at modelling torsional potentials of drug like molecules versus conventional MM models. Taking it one step further, here, we present a strategy to simulate protein–ligand complexes using a machine-learned NNP to represent the intramolecular interactions of the ligand. This model is embedded inside a conventional MM force field for the protein and solvent,

so established models for these components can be used without modification. We call this method NNP/MM, as it functions the same as Quantum Mechanical / Molecular Mechanical (QM/MM) models do, but with the NNP used in place of the QM method. This method is tested for its ability to predict the poses of protein-bound drugs in comparison to electron density distributions determined by X-ray crystallography. The Gibbs energies for restraining the ligands to their bound conformations are calculated using NNP/MM and compared to the CGenFF force field.

The drug erlotinib is used as a standout example for the differences between ANI and CGenFF. Erlotinib is an inhibitor of the epidermal growth factor receptor (EGFR) tyrosine kinase, and is used in the treatment of non small cell lung cancer, pancreatic cancer, and several other types of cancer [71]. The PMF for erlotinib is calculated using ANI and CGenFF and compared to CCSD(T) data.

3.2 Computational Methods

3.2.1 Theory

In this method, the potential energy of the whole system is defined as the sum of the potential energy of the subsystem described by the NNP (i.e., the intramolecular interactions of the ligand) ($\mathcal{V}_{NNP}$), the potential energy of the environment around the ligand ($\mathcal{V}_{MM}$), and the interactions between the ligand and its environment ($\mathcal{V}_{NNP/MM}$). (Eqn. 3.1).

$$\mathcal{V}(\mathbf{r}) = \mathcal{V}_{MM}\left(\mathbf{r}_{MM}\right) + \mathcal{V}_{NNP}\left(\mathbf{r}_{NNP}\right) + \mathcal{V}_{NNP/MM}\left(\mathbf{r}\right) \tag{3.1}$$

The MM region is represented using a conventional MM force field, so $\mathcal{V}_{MM}$ is calculated in the normal fashion for an additive force field. For non-covalent protein–ligand binding, the $\mathcal{V}_{(NNP/MM)}$ term is the conventional MM non-bonded interactions between the protein and the ligand, which is simply the sum of Lennard-Jones and pairwise Coulombic interactions between the NNP atoms and MM atoms (Eqn. 3.2).

$$\mathcal{V}_{NNP/MM}\left(\mathbf{r}\right) = \sum_{i}^{MM}\sum_{j}^{NNP}\frac{q_i q_j}{4\pi\epsilon r_{ij}} + 4\varepsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right] \tag{3.2}$$

This functions similarly to mechanically-embedded QM/MM models [72], where the NNP serves as the "QM" model embedded within the MM system. This method can be employed in many established simulation codes without modification because they can be implemented using existing QM/MM features, which allows the energy and forces of a critical subsection of the system to be calculated using an external method.

The immediate advantage of this method is that highly-accurate intramolecular forces can be calculated for ligands without parameterization and without modifications to current molecular simulation codes. A limitation of this approach is that the protein–ligand interactions are still calculated by the CGenFF/CHARMM electrostatic and Lennard-Jones terms. The development of efficient NNPs that are capable of describing the entire system could provide more accurate and non-empirical protein-ligand binding energies.

There have been several reports where QM/MM simulations were used to model protein–ligand complexes [73, 74, 75]. The drawbacks of these QM methods is that typically they use semi-empirical quantum mechanics in order to calculate the energy and forces of the ligand sufficiently quickly to perform sufficiently long MD simulations. These methods generally are less accurate than the ANI NNPs for the calculations of the relative conformational stability of ligand conformations and the computational cost is generally greater. One advantage of QM/MM methods over the NNP/MM method used here is that the electron density of the ligand can be polarized by the protein and solvent (i.e., through electrostatic-embedding QM/MM[72]). This is not possible for the NNPs used here because these methods do make any calculation of the electron density of the ligand, so they are effectively mechanically embedded.

3.2.2 Technical Details

All molecular dynamics (MD) simulations were performed using NAMD 2.13 [76]. The ligand intramolecular energies and forces were calculated using the ANI-1ccX [22] NNP implemented in the TorchANI package [46]. The programs were interfaced through the general-purpose external-force functionality of the NAMD QM/MM code [77]. The CHARMM36m force field[78] was used to represent the protein and the mTIP3P model [79, 80] was used to represent the water molecules. Sample input files and our scripts can be downloaded from our online repository [48] and will be included

in future distributions of NAMD. The CGenFF [12] Lennard-Jones and electrostatic parameters were used to calculate the non-bonded ligand–protein interactions (i.e., $\mathcal{V}_{(NNP/MM)}$). Non-bonded interactions were calculated using a 12 Å cutoff, although lattice-summation methods are also available in the QM/MM NAMD interface.

The calculation of the erlotinib potential energy surface was performed using ORCA 4.2.1 [52]. Optimizations with constraints on the amine torsional angle were performed using the resolution of identity 2nd-order Møller–Plesset theory (RI-MP2) with the def2-TZVP basis set [51]. Single point energy evaluations were performed at these optimized structures using Domain-based Local Pair Natural Orbital - Coupled Cluster Singles and Doubles with perturbative triples[81] with the def2-TZVP basis set (DLPNO-CCSD(T)/def2-TZVP//RI-MP2/def2-TZVP) to generate the QM potential energy surface.

3.2.3 Test Set

To evaluate the ability of the ANI-1ccX potential to predict the pose of a bound ligand, we developed a test set of protein–ligand complexes. We selected a structurally-diverse set of complexes where a high-resolution crystallographic structure of the protein–ligand complex was available, including several where the ligand is in a conformationally-strained pose. The ANI-1ccX NNP is only defined for carbon, nitrogen, hydrogen, and oxygen, so only ligands composed of these elements were selected. The full details of the structures are available in Appendix A.

3.2.4 Simulations of Ligand Binding Poses

The NNP/MM ligand binding poses were generated by MD simulations of the protein–ligand complexes. The crystallographic structure (including crystallographic water molecules) was placed in a periodic unit cell of liquid water. The protonation states of the protein and ligand were assigned using H++ 3.2 [82] and by examining the intermolecular interactions of titratable residues in the crystallographic structure. A 5 ns equilibration MD simulation using the CGenFF force field for the ligand was performed where all non-hydrogen atoms of the ligand and protein were restrained to their crystallographic positions. The equilibrated structures were used as the initial structures of 2 ns NNP/MM MD simulations of the complexes. In these simulations,

the C_α atom of the protein backbone were restrained to their crystallographic positions using harmonic potentials (k_c = 10 kcal/mol Å^{-2}). These simulations were performed with a thermostat temperature set to correspond to the temperature the crystallographic structure was collected for (e.g., 100 K). The ligand electron density was obtained from the crystallographic electron density map, selecting all points within 2 Å of the ligand atoms in the PDB structure. An isosurface value of 0.5 was used in the renderings.

3.2.5 Calculation of Conformational Gibbs Energy

Confine-and-release alchemical free energy perturbation is a popular technique for calculating absolute protein–binding energies [83, 84, 85, 86]. In these methods, the total binding energy is divided into a set of Gibbs energies for each step in a path where the ligand is constrained to its bound conformation and is then decoupled from its environment. The component corresponding to the reversible work required to constrain the ligand to its bound conformation is defined as ΔG_{cons}. Physically, this energy corresponds to the reduction of conformational freedom and isomerization to a higher energy conformation that occurs when a ligand binds to a protein. In confine-and-release absolute binding energy calculation schemes, this is the only term where the intramolecular interactions of the ligand are significant. Accordingly, it is only necessary to use the NNP/MM method when calculating this term; the remaining terms can be calculated using conventional force fields. Notably, this step does not include any alchemical transformation, so performing the calculation with NNP/MM does not present any special challenges.

This term can be calculated by defining the root-mean-square deviation (RMSD) of the ligand relative to its bound conformation (ζ) and then calculating the Gibbs energy required to impose a harmonic restraint on the RMSD ($\frac{1}{2}k_c\zeta^2$) so that the ligand is restricted to hold its bound conformation. This procedure is performed for the ligand in solution and in the site to obtain Gibbs energies for restricting the conformation of the ligand in each of these states. The difference of these energies provides the conformational or "strain" component of the absolute binding energy, (ΔG_{cons}).

Using umbrella sampling, the potential of mean force (PMF) can be calculated as a function of the RMSD. Integration of this PMF biased by the harmonic restraining

function provides the $\Delta G_{cons,site/solvent}$ (Eqn. 3.3).

$$e^{-\Delta G_{cons,site/solvent}/k_BT} = \frac{\int e^{-\left[w(\zeta,site/solvent)+\frac{1}{2}k_c\zeta^2\right]/k_BT}d\zeta}{\int e^{-w(\zeta,site/solvent)/k_BT}d\zeta} \tag{3.3}$$

where k_c is a harmonic potential to restrain the conformation of the ligand at the reference structure. In this work, a value of $k_c = 10$ kcal/mol Å^{-2} was used.

These PMFs are calculated from an umbrella sampling simulation where the windows were separated by 0.5 Å and a harmonic biasing potential with a spring constant of 50 kcal/mol Å $^{-2}$ was used. Each window was sampled by performing a 1 ns equilibration simulation followed by a 4 ns sampling simulation. The PMF was constructed from the umbrella sampling simulations using Weighted Histogram Analysis Method (WHAM) with statistical uncertainties of the profiles estimated by bootstrap analysis [61, 62, 63].

These calculations are performed for the ligand bound to the protein and in solution to yield ($\Delta G_{cons,site}$) and $\Delta G_{cons,solvent}$, respectively. The difference of these two energies provides ΔG_{cons} (Eqn. 3.4).

$$\Delta G_{cons} = \Delta G_{cons,site} - \Delta G_{cons,solvent} \tag{3.4}$$

3.3 Results and Discussion

3.3.1 Prediction of Ligand Poses

Figure 3.1 shows the ligand poses generated from the ANI/MD simulations overlaid with the crystallographic electron density maps of the ligand. Generally, the NNP/MM ligand pose overlaps well with the crystallographic density. The positions of the ligand phenyl rings in the thrombin complex (3DA9) and the biotin carboxylase complex (2W6N) are the most significant deviation. The NNP/MM model still relies on conventional MM parameters for the protein–ligand and water–ligand interactions, so these deviations may not be related to the NNP component of the model.

One notable success of the NNP/MM potential is in predicting the binding pose of erlotinib to the epidermal growth factor receptor (EGFR). The core scaffold of this

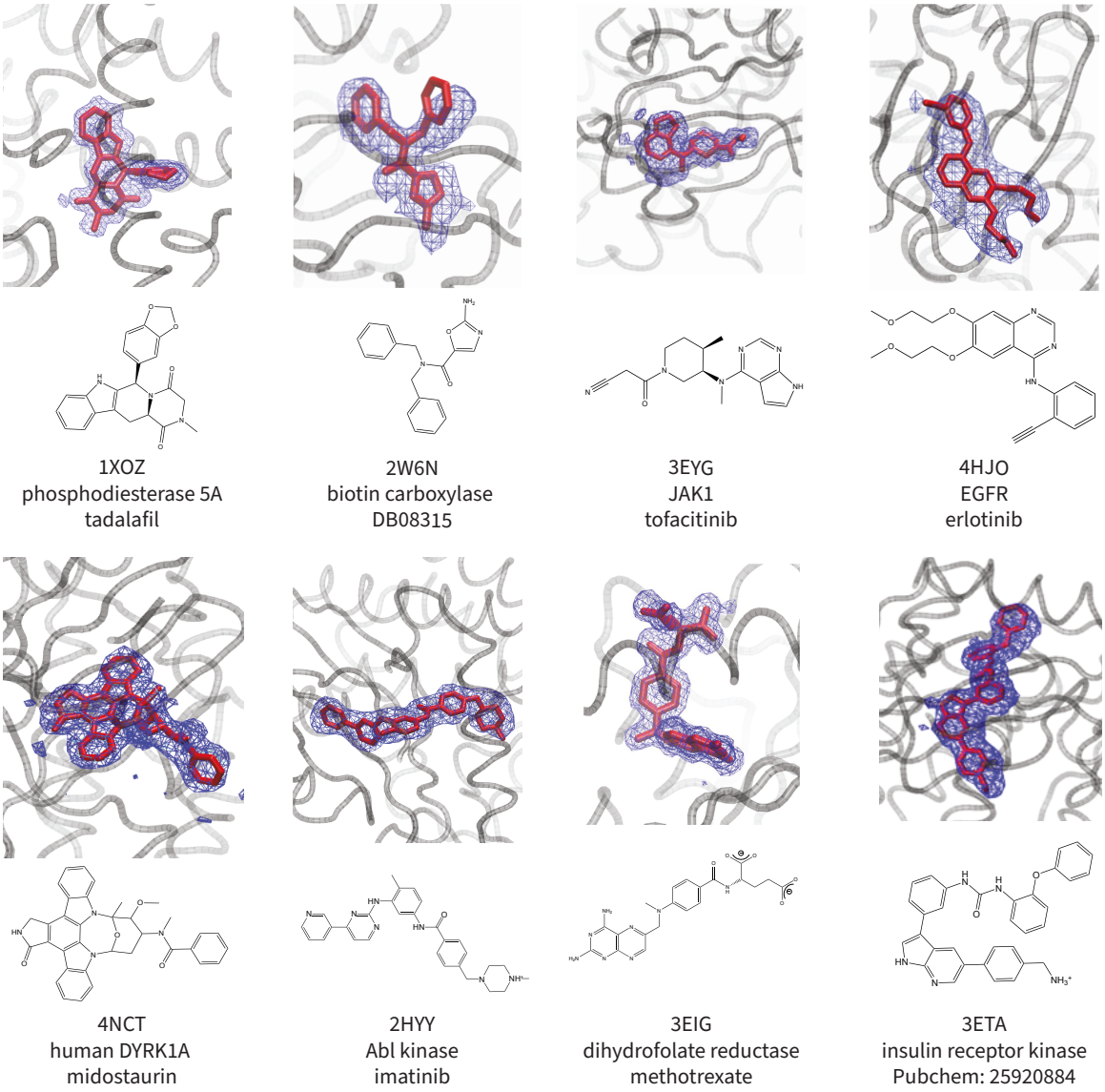

Figure 3.1: Calculated poses of ligands (red) in protein binding sites. The crystallographic electron density of the ligands are shown in blue. The PDB ID, protein name, and ligand name are included beneath the image.

drug is composed of amine-linked ethynyl-phenyl and quinazoline rings. Crystallographic structures of the protein-bound complex show the quinazoline ring bound in the adenosine-binding site while the ethynyl-phenyl group binds in a pocket formed by the T702, T830, and K721 residues. The binding pose predicted by CGenFF is inconsistent with the XRD data, in which the two rings form a more acute angle relative to each other ($\phi_1 = 63 \pm 1°, \phi_2 = 4 \pm 1°$). The simulation using the NNP/MM model is more consistent with the crystallographic data, whereby ($\phi_1 = 44 \pm 1°, \phi_2 = 4 \pm 1°$).

Surprisingly, the poses predicted for the ligands that contain charged functional groups (2HYY, 3ETA, and 3EIG) are reasonable even though the ANI-1ccX potential was not designed to describe charged species and none of the molecules this NNP was trained for were charged.

3.3.2 Conformational Free Energies

The conformational strain of the ligand that occurs in protein–ligand binding arises from the need for the ligand to adopt the conformation it holds in its bound form. The bound conformation may be more strained than the lowest energy conformation it can hold in solution. Further, some ligands can adopt multiple conformations in solution, so limiting the conformational space of the ligand to the bound conformation is endergonic. For example, Roux and coworkers' calculations of the binding affinity of imatinib to Abl kinase predicted that while the net interaction energy of binding was -27.7 kcal/mol, the conformational energy countered this by 11.3 kcal/mol [87]. The conformational energies for the test set of ligands were estimated by calculating the PMF ($w(z)$) for the deviation from the bound pose using umbrella-sampling MD simulations with both the CGenFF and NNP/MM models. ΔG_{cons} was calculated from these PMFs using Eqn. 3.3. These energies are collected in Table 3.1. The PMFs for all complexes are presented in Appendix A.

Amongst the neutral ligands, the NNP/MM conformational energies are generally similar in magnitude to the CGenFF strain energies. This indicates that the ANI-1ccX model can achieve similar results to the CGenFF model despite the lack of any explicit parameterization for these molecules. The conformational energies of 4HJO (erlotinib bound to EGFR) show the largest difference, with the NNP/MM strain energy being 4.7 kcal/mol smaller than the CGenFF strain energy. The high strain predicted by the CGenFF model is due to the amine functional group of erlotinib

Table 3.1: Conformational Gibbs energy of binding for protein–ligand complexes calculated using the MM(CGenFF) and NNP/MM methods. All energies are in kcal/mol.

PDB ID	$\Delta G_{cons,CGENFF}$	$\Delta G_{cons,NNP/MM}$	charge
1XOZ	0.4 ± 0.0	0.5 ± 0.0	0
2W6N	4.7 ± 0.1	5.2 ± 0.1	0
3EYG	1.9 ± 0.1	1.0 ± 0.2	0
4HJO	13.0 ± 0.1	8.3 ± 0.1	0
4NCT	3.4 ± 0.1	2.3 ± 0.1	0
2HYY	8.1 ± 0.1	326.9 ± 0.1	1
3EIG	11.1 ± 0.0	37.7 ± 0.0	-2
3ETA	5.6 ± 0.1	15.2 ± 0.2	1

holding a pyramidal geometry in the solution simulations, creating a large energetic penalty to force the drug into its bound conformation. In the NNP/MM simulation of erlotinib in solution, the amine group remains close to a co-planar geometry with respect to the quinazoline ring, with a moderate skew in the dihedral angle between the phenyl group and the amine.

The ligands that contain charged functional groups (2HYY, 3ETA, and 3EIG) have anomalously high conformational energies. This issue originates from the use of the ANI-1ccX NNP, which was only trained on neutral molecules. This NNP predicts reasonable geometries of the ammonium and carboxylate groups in these molecules, but these ionic functional groups form spurious intramolecular contacts in the solution NNP/MM MD simulations. For example, the ligand of 3EIG adopts a conformation where the carboxylates groups are in close contact, rather than repelling each other like they should (see Appendix A). This results in the stabilization of regions of the PMF corresponding to large structural deviations from the bound pose. As the NNP(ANI-1ccX) model was not designed for the description of charged molecules like this, it is unsuitable for calculating their conformational energies.

Extensive MD simulations are needed to calculate ΔG_{cons} by calculating the PMF of the RMSD, but these simulations were completed at a modest computational cost because of efficient implementations of the ANI model for execution on graphical processing units. For example, the NNP/MM MD simulations of imatinib (69 atoms) executed at a rate of 3.4 ns/day on a single Titan Xp NVIDIA GPU. Even faster performance is anticipated after the planned integration of NNPs directly into NAMD and other molecular simulation codes.

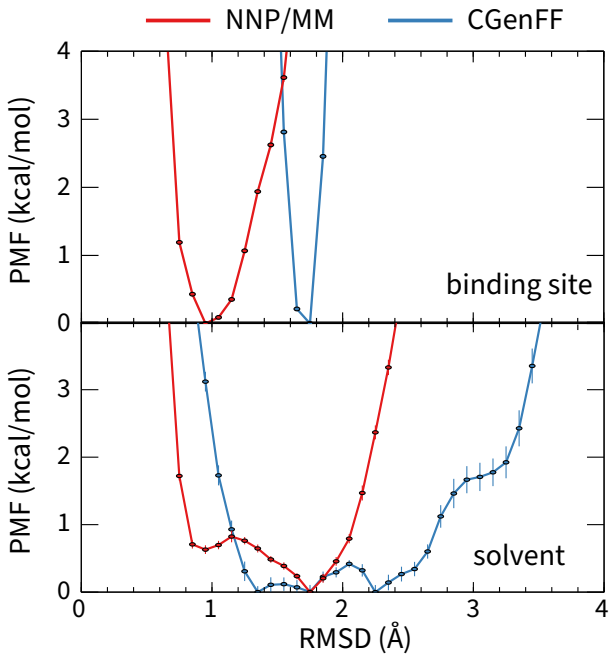

Figure 3.2: The potential of mean force for the deviation of the structure of erlotinib from its bound conformation when it is bound to EGFR (top, PDB ID: 4HJO) and when it is in solution (bottom) calculated using the hybrid NNP/MM and pure MM (CGenFF) methods.

Empirical force fields are parameterized in an internally consistent manner, so it is possible that the MM parameters used to describe the non-bonded interactions between the ligand and its surroundings will not be optimal for the NN/MM term. In particular, the balance between the MM ligand–water, ligand–protein, and the NNP ligand intramolecular dispersion interactions will not necessarily be consistent [14, 15]. This issue has been addressed in some QM/MM models by defining new parameters for the QM–MM Lennard-Jones terms [88, 89]. Nevertheless, the common practice has been to parameterize the intramolecular terms of ligands to gas phase potential energy surfaces, so the ANI-1ccX should be a suitable replacement for these terms. This effect should also lead to a systematic difference in the conformational energies of the ligands, but the CGenFF and NNP/MM conformational energies are close in magnitude for 1XOZ, 2W6N, 3EYG, and 4NCT.

3.3.3 Torsional Potential Energy Surface of Erlotinib

The large difference in the ANI-1ccX and CGenFF conformational energies of 4HJO (erlotinib bound to EGFR) originate from the ligand adopting conformations in solution that are drastically different than the bound conformation when the CGenFF model is used, while the NNP/MM model predicts similar conformations in both the binding site and solution. This is evident in the CGenFF PMF of the ligand's conformation relative to its bound pose in Figure 3.2, which is considerably broader than the NNP/MM PMF and is higher energy in the crystallographic pose (RMSD=0 Å).

The geometry of the erlotinib amine linker and its aromatic substituents deviates sharply from the bound pose in the CGenFF solution structure (Figure 3.3 (b)); the amine is partially pyramidalized and the aromatic substituents are skew to each other. In contrast, in the NNP/MM simulation, the amine predominantly remains in a planar geometry, conjugated with the quinazoline and phenyl rings.

The potential energy surface corresponding to rotations around the amine torsion angles of erlotinib is presented in Figure 3.3 (c). The minima on the CGenFF surface corresponds to structures where the amine is significantly pyramidal and the substituent phenyl and quinazoline rings adopt angles that reduce steric repulsion between them. The ANI-1ccX surface is consistent with the DLPNO-CCSD(T) surface, where there is a broad global minimum centered around ($\phi_1 = 0°, \phi_2 = 0°$) and the amine nitrogen holds a planar arrangement with the aromatic groups.

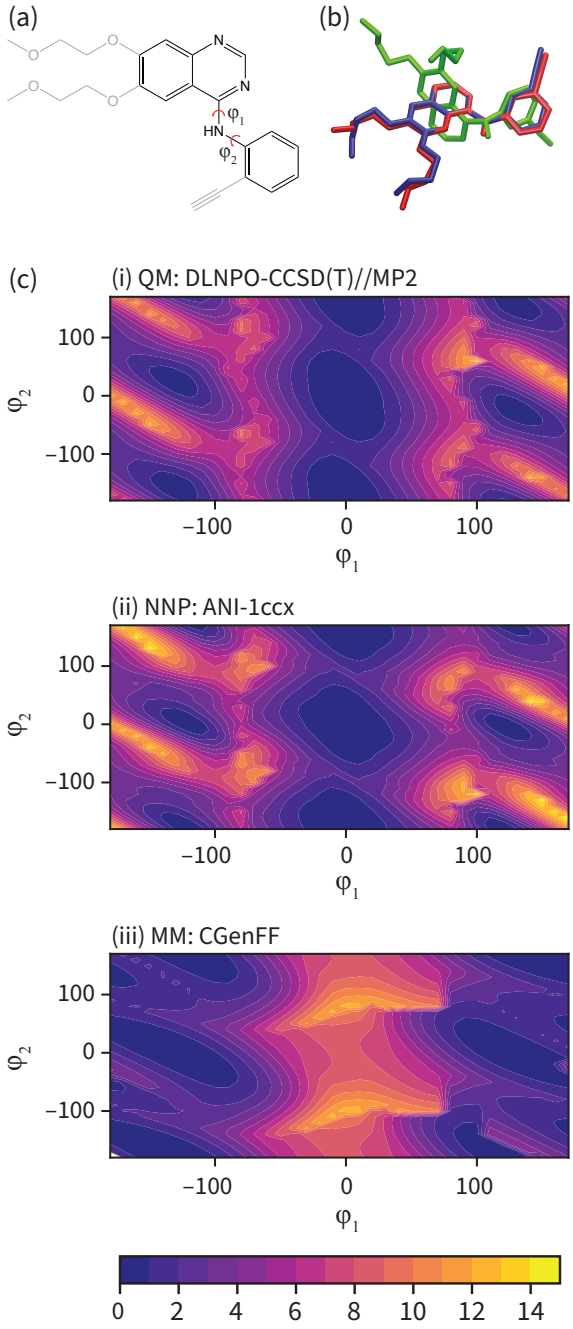

Figure 3.3: (a) The fragment of erlotinib used to calculate the potential energy surface. Truncated groups are shown in grey. (b) Representative solution conformations of erlotinib for the CGenFF MM model (green) and NNP/MM model (red) overlaid with the ligand pose from the 4HJO crystal structure (c) The relaxed potential energy surfaces for rotation around the erlotinib fragment amine bonds calculated using (i) DLPNO-CCSD/def2-TZVP//MP2/def2-TZVP (ii) NNP(ANI-1ccX) and (iii) the CGenFF MM model. Energies are in kcal/mol.

The failure of the CGenFF force field stems from the lack of a distinct atom type for amines conjugated with aromatic rings. While it would be possible to adjust the parameters of the CGenFF force field to improve its description of the arylamine potential energy surfaces, this introduces a new fitting stage and requires computationally demanding QM calculations to provide the target data. Generally, it is not immediately apparent where a general-purpose force field will fail. By using NNPs to calculate these interactions, these issues are avoided entirely because energy surfaces with near-CCSD(T) accuracy can be generated efficiently and without the need to parameterize the intramolecular potential energy surface explicitly.

3.4 Conclusions

NNPs provide accurate representations of the intramolecular interactions of drug molecules in molecular simulations of protein–ligand binding. These simulations take advantage of established MM models of the protein and solution, while eliminating the need to develop a force field for the intramolecular interactions of each ligand. By employing a NNP that has already been trained on a broad set of molecular species, the fundamental intramolecular interactions that give rise to the molecular energy surface are captured without the need to parameterize a force field. This representation is also free of the harmonic/torsional/improper scheme used in conventional force fields. This allows the simulations to be deployed immediately, without the development of parameters for each new chemical moiety.

These methods can be incorporated directly into existing confine-and-release methods to calculate the absolute binding energy because these methods include a step where the ligand's conformation is constrained to its bound pose. In several cases, the conformational energies calculated using the NNP(ANI-1ccX)/MM model were similar to those predicted by the popular general-purpose CGenFF force field, although chemically-significant differences (i.e., $>$ 1 kcal/mol) were found in several instances.

Chapter 4

The Refinement of Cryo-EM structures using Neural Network Potentials

4.1 Introduction

The three-dimensional structure of ligands bound to their protein targets allows the molecular effect of ligand binding to be understood structurally and for drug-protein interactions to be tuned. X-ray crystallography has been used to determine protein structures beginning with Kendrew *et al.*'s determination of the structure of myoglobin in 1960 [3]. Between this first structure and the year 2021, over 178,229 crystallographic structures have been deposited into the protein data bank archive of protein structures. X-ray crystallography has some limitations however, such as the need for proteins to crystallize, and the complicated nature of the method.

Cryo-EM is an emerging technique of interest competing with XRD because of its lack of need for crystallization. This provides access to structures that are unable to be studied by XRD. However, cryo-EM is still a limited method due to its low resolution imaging, which is often unable to provide a definitive pose of a bound molecule or its intermolecular interactions.

Molecular modeling of the ligand, guided by the density of the protein, offers a

concrete solution to resolving the drug-protein problems in cryo-EM. Notably, Molecular Dynamics Flexible Fitting (MDFF) provides a straightforward strategy where an all-atom molecular mechanics-based model is generated of the protein-ligand complex. An external biasing potential is imposed that favours structures where the atoms lie on top of regions where the cryo-EM-derived electron density is highest. As a consequence, over the course of a molecular dynamics simulation, the atoms of the protein-ligand complex will dynamically move to structures consistent with the cryo-EM data. The potential energy of the force field and ligand ensures that these structures will also be chemically reasonable in terms of bond lengths, torsional angles, intermolecular interactions, etc.

MDFF was first used successfully to refine the cryo-EM structure of the ribosome, after its development by Trabuco *et al.* [90]. The developer group has since went on to solve structural models of photosynthetic proteins [91, 92], myosin [93], chaperonins [94], bacterial chemosensory array [95], and virus capsids [96, 97], including the first all-atom structure of the HIV capsid [97]. Other groups have also found success using MDFF to model structures such as the actin-myosin interface [98] and the HIV-1 virus [99, 100].

Although this strategy has proven effective for modeling proteins, representing the intramolecular terms of the ligand remains an issue. This process can require a tedious parameterization of the ligand force field based on experimental or quantum chemical data. Several force field construction protocols have been developed to capture the ligand interactions [29]. The quantum mechanical (QM)/MM interface of NAMD allows partitioning of a system into quantum mechanical and molecular mechanical levels of description. The ligand is described quantum chemically, while the protein and the solvent are probed classically. The energies from the protein are computed using an MM model, such as CHARMM. The ligand energies can be calculated with an external QM software. QM/MM can be combined with MDFF to resolve cryo-EM structures of protein-ligand complexes, preventing the ligand geometry from deviating towards unphysical structures. Although QM/MM-MDFF can provide well-resolved cryo-EM structures, QM/MM prohibitively time-consuming the nanosecond length MD simulations needed for standard MDFF fitting. A truly general strategy for resolving the structures of cryo-EM protein–ligand complexes requires an accurate method for calculating the potential energy for all possible ligands at computational cost that allows routine nanosecond length MD simulations.

Neural network potentials (NNPs) have recently emerged as an alternative that has the parameter-free accuracy of QM models, but the efficiency of MM models. These methods have recently been employed to model the intramolecular terms of ligands that are in solution or protein-ligand complexes. This would allow protein–ligand complexes to be refined using MDFF with a highly accurate representation of the intramolecular ligand terms, without requiring parameterization.

In this chapter, we test MDFF using a NNP/MM representation of six published cryo-EM structures of protein–ligand complexes. NNP/MM and conventional MM are compared to the published structures.

4.2 Methods

4.2.1 Selection of the Test Set

Structures were identified where the ligand contains only the elements supported by the ANI-2X potential (C, N, O, H, S, F, and Cl) and do not contain charged functional groups. Proteins with large unmodeled regions were also excluded. These structures identified are summarized in Table 4.1. The reported resolution of these structures ranges between 1.9 to 3.84 Å.

4.2.2 Computational Methods

MDFF Simulations were performed using NAMD 2.14 interfaced to TorchANI using our NNP/MM interfacing scripts. Protein segments less than 50 amino acids in length that were missing from the protein were completed using SWISS-MODEL homology modeling [101]. Larger unresolved regions were not included in the model.

All proteins were represented using the CHARMM36m force field. The intermolecular interactions between the protein and the ligand in the NNP simulations use the CGenFF atomic charges and Lennard-Jones parameters. The parameter for strength of the MDFF coupling between the MDFF and the ligand was set to 1.0 (GSCALE = 1.0). The simulations were performed for 0.5 ns where the timestep was 2 fs. A Langevin thermostat with a friction coefficient of 5 ps^{-1} was applied. Restraints were imposed to preserve the secondary structure, chirality, and peptide-bond cis/trans

geometry.

4.3 Results and Discussion

The structures generated using MDFF are overlaid with the structures reported in the PDB in Figure 4.1.The molecular structure of the ligands calculated using the MDFF/CGenFF simulations, the MDFF/NNP simulations, and structure deposited in the PDB are presented in Figure 4.2. These simulations were broadly successful in refining structural models of cryo-EM of protein-ligand complexes. In all cases, the protein secondary and tertiary structures were the same for the published cryo-EM derived structure and the MDFF refined structures. In most cases, the structure of the ligand calculated using MDFF was similar to the PDB model, although in some cases, there were significant differences.

For the structures with PDB IDs 6OT0,7L1V and 6X3X, the MDFF ligand structures aligned with the PDB structures with minimal differences. These minimal differences can arise from slight differences in bond angles and lengths between the three methods. All three ligand structures fit within the cryo-EM density, so there is no indication that one method is superior to the others in these cases. These examples show that MDFF/NNP can perform just as accurately as traditional MDFF/CGenFF, while avoiding the need for parameterizing the intramolecular terms of a molecular mechanical force field.

In the structures of (4-oxo-5-phenyl-3,4-dihydrothieno[2,3-d]pyrimidin-2-yl)methyl-3-(3-oxo-2,3-dihydro-4H-1,4-benzoxazin-4-yl)propanoate bound to TRPV5 (PDBID: 6PBE), the MDFF/NNP and the PDB structures are in good agreement, with the exception of the conformation of the ester. Figure 4.3 shows the structure of (4-oxo-5-phenyl-3,4-dihydrothieno[2,3-d]pyrimidin-2-yl)methyl 3-(3-oxo-2,3-dihydro-4H-1,4-benzoxazin-4-yl)propanoate. In the MDFF-NNP/MM structure, the ester conformation is trans, whereas in the PDB the conformation is cis. Both conformations still fit within the cryo-EM density and it is uncertain if one conformation is preferred over the other because the rest of the structure does not appear to be affected by this difference. The experimental cryo-EM electron density in this region is low, so it is not immediately apparent which conformation is observed experimentally. It is worth noting that MDFF/CGenFF also preferred a cis conformation. If the actual ligand

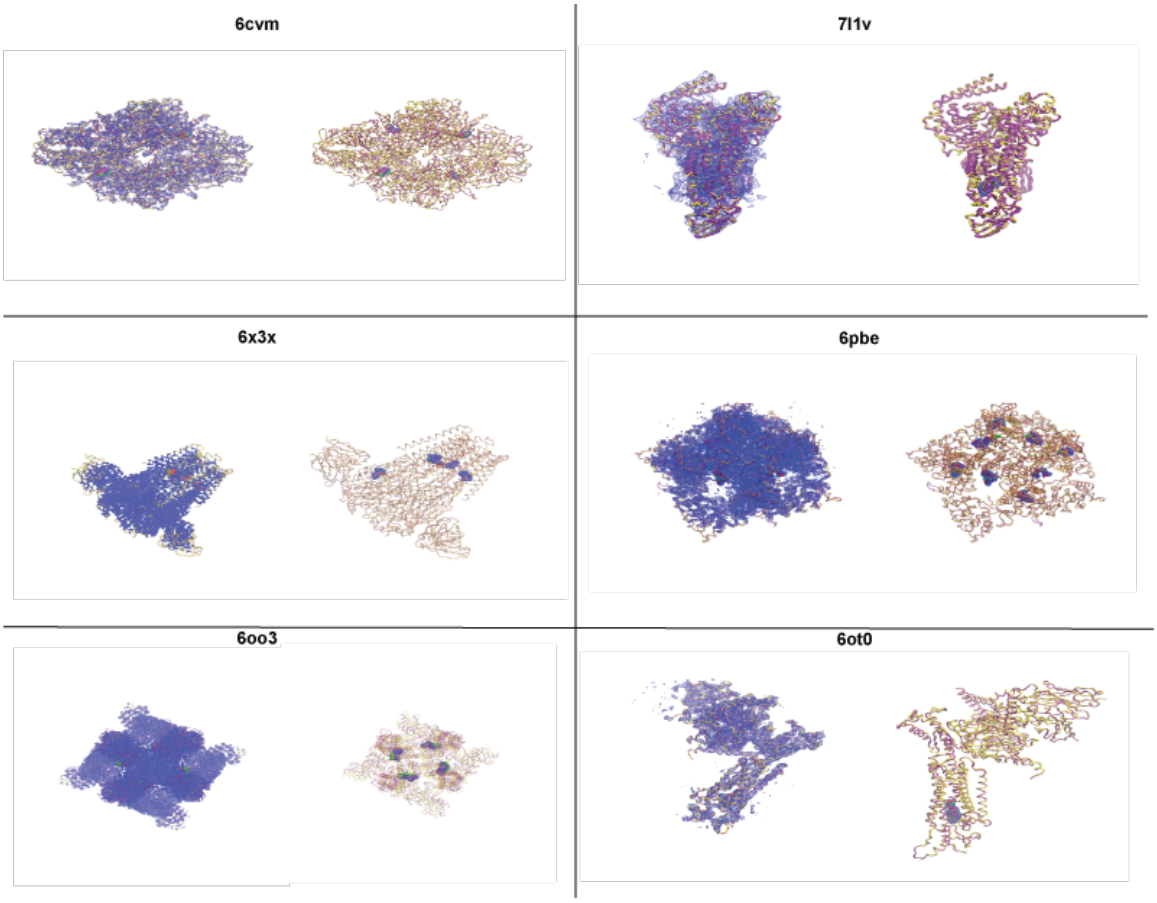

Figure 4.1: Structures of the protein-ligand complex generated with MDFF/NNP overlaid with the PDB structure and cryo-EM density. MDFF/NNP protein is in purple, and MDFF/NNP ligand is in green. PDB protein structure is in yellow, and PDB ligand structure is in red. Cryo-EM density is in blue.

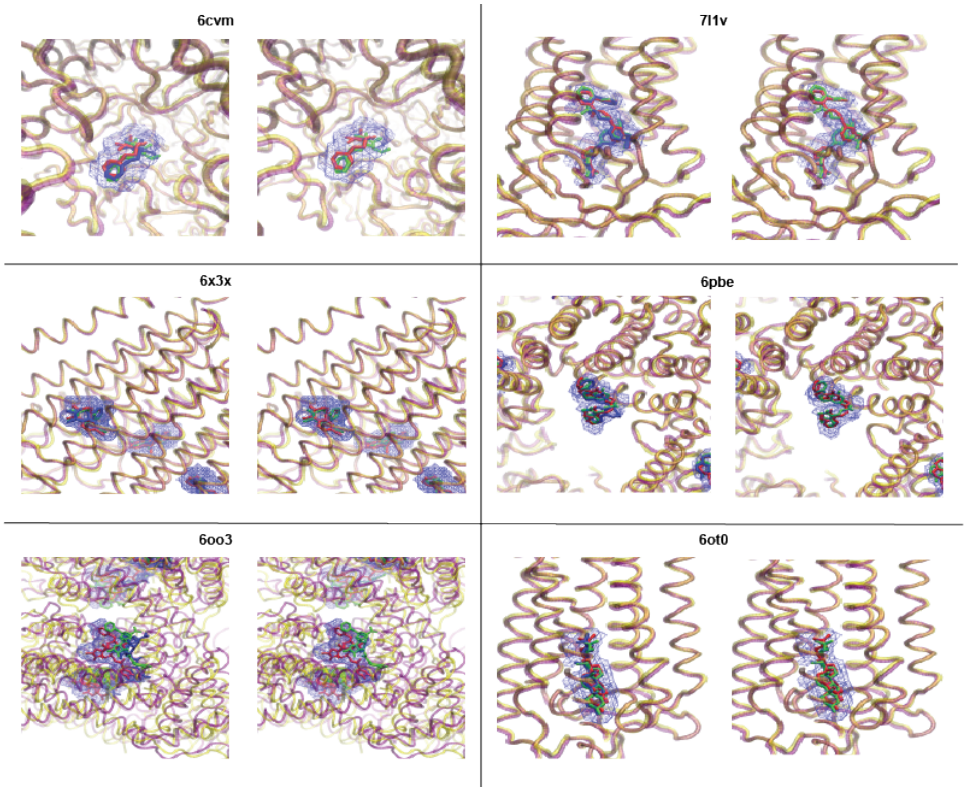

Figure 4.2: Comparison of MDFF protein–ligand complex structures, PDB structure, and cryo-EM density. Left figure shows the MDFF-NNP/MM structure of the ligand overlaid with the PDB and MDFF-CGenFF structures and cryo-EM density. Right figure shows only the MDFF-NNP/MM structure of the ligand overlaid with the PDB and cryo-EM density. The MDFF-NNP/MM protein is in purple, and MDFF-NNP/MM ligand is in green. PDB protein is in yellow, and PDB ligand is in red. MDFF-CGenFF ligand is in blue. Cryo-EM density is in blue.

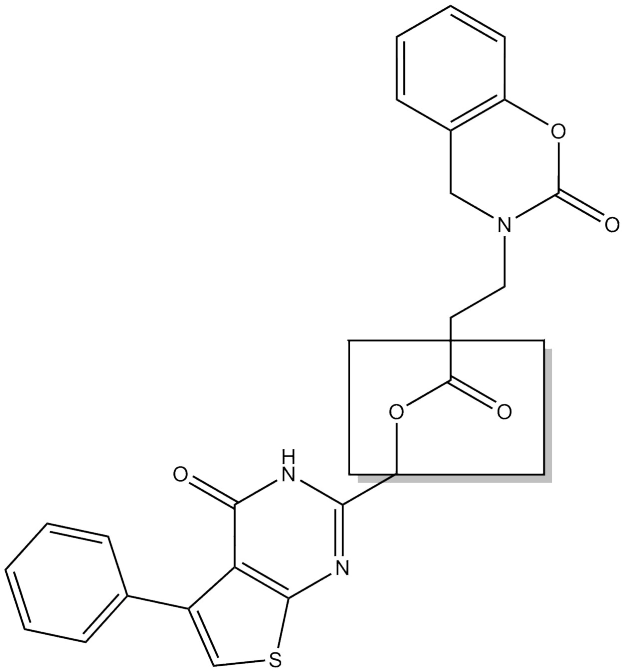

Figure 4.3: Structure of (4-oxo-5-phenyl-3,4-dihydrothieno[2,3-d]pyrimidin-2-yl)methyl 3-(3-oxo-2,3-dihydro-4H-1,4-benzoxazin-4-yl)propanoate.

conformation is trans, this would give the MDFF/NNP method an advantage over MDFF/CGenFF because it would predict a more accurate structure.

The ligand of β-galactosidase bound to 2-phenylethyl 1-thio-β-D-galactopyranoside (PETG) (PDB ID: 6CVM) has a methyl alcohol group attached to an oxane ring, as seen in Figure 4.4. In the MDFF/NNP derived structure, this functional group is below the plane, but it is above the plane in the PDB structure. The methyl alcohol group being below the plane causes the alcohol to slightly stick out of the density region. MDFF/CGenFF also has the methyl alcohol group below the plane causing it to stick out of the density slightly. Again, because the resolution is low in cryo-EM it is difficult to tell which conformation is correct, however in this case it seems as though the PDB structure is a more accurate representation of the real structure as the ligand is entirely within the density region. In this case, differences in bond angles could be the cause of this deviation in the MDFF/NNP structure.

Resiniferatoxin bound to TRPV2/RTx (PDB ID: 6OO3) shows one of the largest differences between MDFF-refined structures and the PDB structure (Figure 4.5. The PDB structure has several unusual bond lengths in the tetracyclic cage-like orthoester

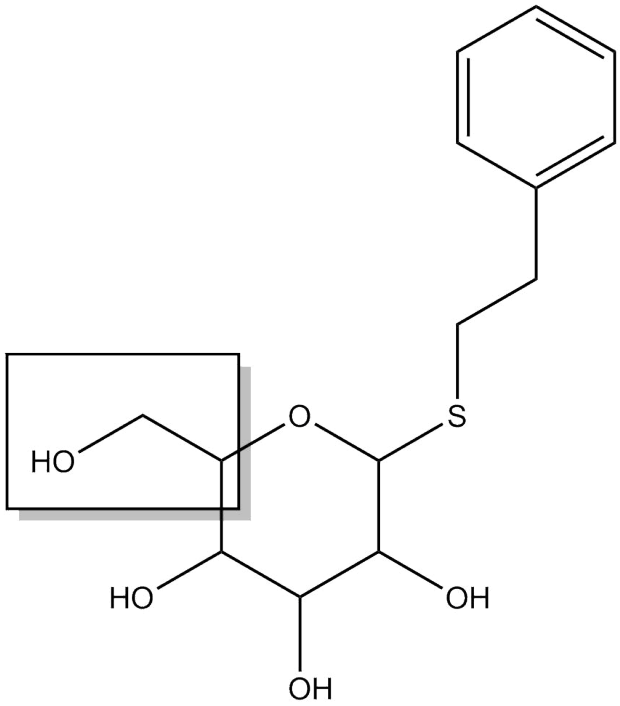

Figure 4.4: Structure of 2-phenylethyl 1-thio-β-D-galactopyranoside.

motif, with elongated C–O bonds. In both the CGenFF and NNP derived structures, these bond lengths are reduced to more normal values near 1.3 Å. In the MDFF models, the benzyl group of the ligand rotates to occupy an unassigned concentration of electron density in a hydrophobic pocket near Val633. This binding mode is shown in Figure 4.6.

4.4 Conclusion

The MDFF protocol was shown to be effective in refining published cryo-EM structures of protein–ligand complexes. This method can be used when a molecular mechanical force field is used to represent both the ligand and the protein, which we demonstrated with the structures generated using CGenFF/CHARMM36m model. We have also shown that an NNP can be used to represent the intramolecular terms of the ligand, which we demonstrated with the structures generated using the ANI-2X/CHARMM36m model.

This type of NNP/MM model has several advantages in MDFF refinement of

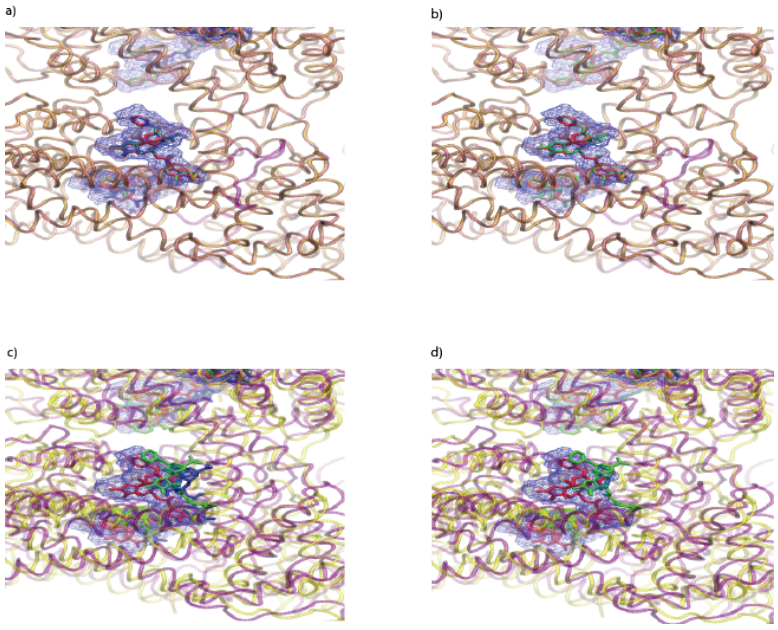

Figure 4.5: Close up view of Resiniferatoxin. MDFF/NNP is in green, PDB is in red and MDFF/CGenFF is in blue. A) is at the start of the simulation with MDFF/NNP, PDB and MDFF/CGenFF. B) is at the start of the simulation with no MDFF/CGenFF. C) is halfway through the simulation with MDFF/NNP, PDB, and MDFF/CGenFF. D) is halfway through the simulation with no MDFF/CGenFF.

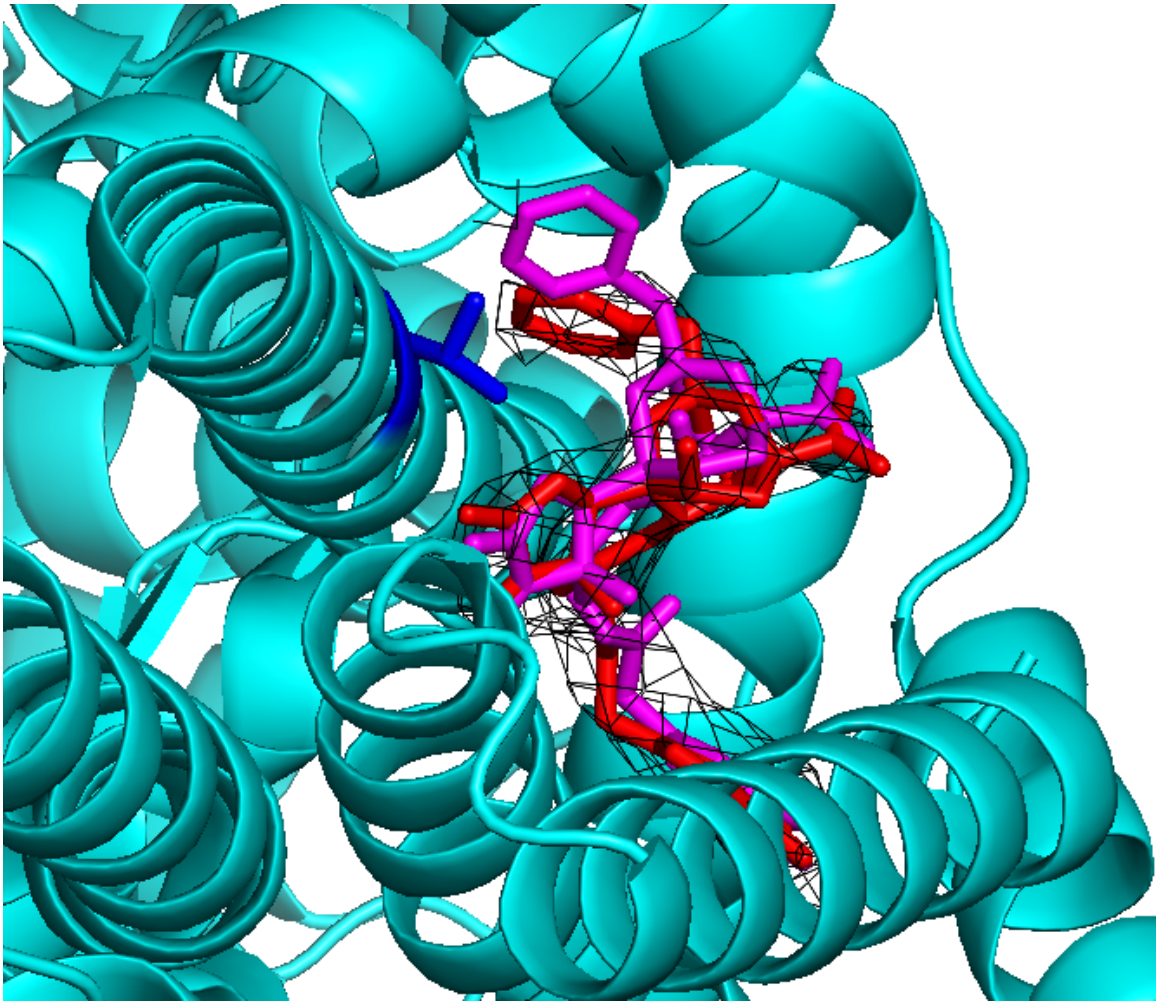

Figure 4.6: Resiniferatoxin bound to the TRPV2/RTx channel from the MDFF NNP/MM structures (red) and the PDB structure (magenta) with the cryo-EM electron density map of the ligand in black wireframe. In the NNP structure, the benzyl group in the uppermost section of the ligand occupies a region with a significant unassigned concentration of electron density, forming an hydrophobic contact with Val633 (blue).

cryo-EM data because it provides a more accurate potential energy function for the intramolecular terms of the ligand that avoids the need to parameterize a molecular mechanical force field for each ligand. These models were found to be effective for a broad range of chemical motifs, including complex natural product compounds and drug molecules. There were a couple examples where MDFF/NNP had a different conformation for a particular molecule or group of molecules than the PDB. It was unclear due to low resolution of cryo-EM data which conformation is accurate to the actual ligand structure, as well as if these conformation differences had an effect on binding. The MDFF structures of resiniferatoxin bound to TRPV2 showed the most significant deviations from the PDB structure. As the simulation progressed, both MDFF-NNP and MDFF-CGenFF deviated away from the cryo-EM density. We suspect this is because of major differences in bond lengths between the above methods and the PDB. The MDFF also showed that the benzyl group in resiniferatoxin rotates to occupy an unassigned concentration of electron density in a hydrophobic pocket near a valine residue. More tests and data is needed to better understand these types of examples and to conclusively demonstrate if MDFF-NNP/MM is generally superior to conventional MDFF-NNP/MM simulations, although the calculations presented in Chapters 2 and 3 suggest that ANI-NNPs are generally more accurate and reliable for simulating ligand structure and dynamics than standard MM models.

EMD	PDB ID	Protein	Ligand	Resolution (Å)
20190	6OT0	G-protein coupled receptor	Oxysterol	3.84
20291	6PBE	Transient receptor potential vanilloid 5 (TRPV5)	(4-oxo-5-phenyl-3,4-dihydrothieno [2,3-d]pyrimidin-2-yl)methyl 3-(3-oxo-2,3-dihydro-4H-1, 4-benzoxazin-4-yl)propanoate	3.78
7770	6CVM	β-galactosidase	2-phenylethyl 1-thio-β-D-galactopyranoside	1.90
23119	7L1V	Orexin Receptor 2	4'-methoxy-N,N-dimethyl-3'-[3-(2-[2-(2H-1,2,3-triazol-2-yl) benzene-1-carbonyl]aminoethyl)phenyl] [1,1'-biphenyl]-3-carboxamide	3.00
20143	6OO3	Transient receptor potential vanilloid 2 (TRPV2) channel	resiniferatoxin	2.90
22036	6X3X	Human γ-aminobutyric acid (GABAA) receptor	diazepam	2.92

Table 4.1: Information on the protein-ligand complexes used in the test set, including EMD and PDB IDs, protein and ligand names, and cryo-EM resolution in Å.

Chapter 5

Conclusions and Future Work

5.1 Conclusions

Understanding the structure and thermodynamics of protein-ligand binding is an important part of drug development. Experimental techniques for determining the structure of protein-ligand complexes, such as X-ray crystallography and cryo-EM, can be costly and unreliable. Molecular simulation can provide computational models that enhance or even replace these experimental structures, but these methods require accurate computational methods to predict the potential energy of a molecule configuration. Simple molecular mechanical models are commonly used for this purpose, but their accuracy is inconsistent and they require extensive parameterization. Neural Network Potentials are an attractive alternative to these MM models because they have comparable computational cost to MM but could have greater transferability and accuracy.

In this thesis, the ANI family of NNPs were used to represent the intramolecular terms of small-molecule drugs. In each chapter, the accuracy and transferability of these NNPs was tested in comparison to other common methods used to model protein-ligand binding.

5.1.1 Conclusions from Using Neural Network Potentials to Model Torsional Potential Energy Surfaces of Biaryl Drug Fragments

In Chapter 2, torsional energies of 88 biaryl drug fragments are calculated using ANI-1ccx, ANI-2X, CGenFF, OPLS, OpenFF, and GAFF and compared to CCSD(T). The tests of accuracy performed on these molecules showed that ANI NNP's could perform as well or better than conventional force fields in predicting energies of drug fragments. The strategy of training these NNP's to reproduce molecular energies in general rather than specific interactions results in methods that are robust for PES's outside their training sets. It should be noted that none of these biaryl compounds in this test set were part of the ANI-2x or ANI-1ccx training sets, so the success of these methods show that they are remarkably robust and provide accurate predictions for molecules and bonding motifs that they were not explicitly trained to describe.

As a practical example of using NNP's to calculate torsional energies in real systems, we performed simulations of the drug ozanimod in an explicit aqueous solution. Multi-nanosecond molecular dynamics simulations in an explicit aqueous solvent were performed, as well as umbrella sampling and adaptive biasing force enhanced sampling techniques. These free energy methods can be used in NNP/MM simulations through the NAMD–TorchANI interface, which makes a diverse set of simulation methods available without modification and allows for facile construction. This provides a method for computationally-efficient but highly-accurate models for the intramolecular potential energy surfaces of ligands within biomolecular simulations without relying on a parameterized force field.

This chapter demonstrated that NNP's provide a model for the intramolecular interactions of drug-molecule fragments that are as accurate or more accurate than conventional molecular mechanical models. The next logical step was to test their accuracy in full protein–ligand simulations.

5.1.2 Conclusions from Simulating Protein–Ligand Binding with Neural Network Potentials

In Chapter 3, eight protein-ligand complexes were simulated using NNP/MM to test the accuracy of NNP's compared to the conventional CHARMM Generational Force Field (CGenFF) in predicting ligand binding poses. High resolution X-ray crystallographic electron density maps had been published for all these structures, which were used as reference to assess the accuracy of the predicted binding pose. The ANI ligand pose, and CGenFF ligand pose were overlaid with the electron density map taken from X-ray crystallography. The Gibbs energy of conformational change upon binding was also calculated for each ligand. Amongst the neutral ligands, the NNP/MM conformational energies are generally similar in magnitude to the CGenFF strain energies. This indicates that the ANI-1ccX model can achieve similar results to the CGenFF model despite the lack of any explicit parameterization for these molecules. The ligands that contain charged functional groups have anomalously high conformational energies. This issue originates from the use of the ANI-1ccX NNP, which was only trained on neutral molecules. Overall, the NNP/MM method worked as well or better than CGenFF and matched crystallography data accurately.

The large difference between the conformational energy of the anti-cancer drug erlotinib calculated using the CGenFF model and the NNP/MM model resulted from a large change in conformation in solution in comparison to the bound state predicted by the CGenFF model. This was largely due to a pyramidalization of the aryl amine group. We found that this was a spurious effect in the CGenFF force field because it does not have a distinct atom type for planar aryl amines and thereby predicts that pyramidal configurations would be most stable. Calculation of the 2D potential energy surface for rotation around the aryl amine bonds showed that CGenFF predicts this surface incorrectly, while the ANI-1ccX model is in good agreement with the high-level ab initio surface (CCSD(T)). This highlights that NNPs can be more reliable than molecular mechanical force fields, which can provide inconsistent results for compounds outside the set used to parameterize them.

The torsional potential energy of erlotinib was calculated both in solution and in the binding site. ANI showed that the conformation did not deviate largely from the binding pose when in solution, whereas CGenFF showed a drastic deviation. The deviation came from the amine linker and its aromatic substituents. The potential

of mean forces were calculated and compared to CCSD(T) theory for the torsional energy of the amine linker. CGenFF predicted a non-planar conformation, whereas ANI and CCSD(T) predicted a planar conformation. This error arises from CGenFF not having an atom type for an amine linked to two aromatic rings. Therefore CGenFF gets the conformational energy wrong.

The work in this chapter has shown again that the ANI NNPs can predict accurate energies for molecules outside of its test set. These NNP's can be used in large scale simulations and retain the accuracy and efficiency of conventional MM models, while increasing transferability. Also, because NNP's do not need to be parameterized, errors in atom types and parameterization of force fields, like in the case of erlotinib, are not present. This is significant as it is not always known that a force field will fail until the simulation is complete. Using NNP's eliminates this problem and therefore saves time.

We proved that NNP's could be used accurately in large scale simulations. Lastly, in this work, we used ANI to help refine cryo-EM protein-ligand structures, specifically the ligand binding pose in these structures.

5.1.3 Conclusions from The Refinement of Cryo-EM structures using Neural Network Potentials

In Chapter 4, six protein–ligand complexes are used to test ANI-NNP's ability to work with molecular fitting methods to better refine cryo-EM structures. Overall, this testset was a success with the majority of structures aligning well with structures from PDB data and the cryo-EM density. There were some cases in which MDFF/NNP held a different conformation for some groups in the ligand structure than the PDB. This did not seem to affect the fitting of the ligands into the cryo-EM density and because cryo-EM data resolution is low it is impossible to tell in this study which conformation is accurate. This conformational difference could be important in binding, and more calculations and tests are needed.

The example of resiniferatoxin bound to TRPV2/RTx was the most interesting. In the MDFF simulations, the ligand benzyl group adopted a different conformation in comparison to the PDB structure, forming a hydrophobic constant with a valine residue. Although this structure is consistent with the cryo-EM electron density, this

binding mode was not identified in the structure reported in the PDB. Also, the structure in the PDB had abnormal bond lengths for several C–O bonds whereas MDFF/NNP and MDFF/CGenFF refined those bond lengths to normal values.

All structural examples in the test set with the exception of resiniferatoxin showed that NNP's are useful in refining cryo-EM structures as they are just as accurate as MDFF/CGenFF methods and match PDB data well. There may be an advantage to using NNP's in the cases where there are multiple possible binding poses because NNP's have been demonstrated to describe ligand structures more accurately and reliably than force fields in some instances. The resiniferatoxin provides a clear example of where structures determined using conventional cryo-EM structure fitting software can be inaccurate.

This chapter is a continuation of the first two chapters in which it is proven once again that the ANI NNP's can be used in the place of conventional MM models, and can be used on a broad range of protein–ligand systems. Another theme that is apparent is that using ANI NNP's can help identify flaws in original methods that otherwise might have gone unnoticed.

5.2 Future Work

General purpose NNPs that are capable of describing protein-ligand systems are in the early stages of their development. The first ANI-type general-purpose NNP, ANI-1, was published by Smith *et at* [16] in 2017. As a result, there are some limitations in these models that must be resolved before they can be applied more widely.

Currently, the ANI-type NNPs published to date can only describe a limited number of elements. The most extensive ANI potential is ANI-2X, which can describe molecules containing the elements C, N, O, F, Cl, S and H. Drug molecules that contain elements such as P, Br, and B cannot be described using these models. The development of NNPs that are capable of describing a broader set of elements will allow a fuller set of drug molecules to be modeled.

Another significant issue is the description of long-range interactions with ANI

NNPs. The ANI-2X NNP has no interatomic interactions beyond 5 Å, so intramolecular dispersion and Coulombic interactions outside this range are not described correctly. NNPs that support charge–charge interactions have been proposed and these methods may eventually resolve these limitations [102].

A related limitation of our NNP/MM embedding scheme is that the NNP region does not experience induced polarization by the MM region. Induced polarization has been found to be significant in some chemical systems [103, 104, 105, 106, 107, 108, 109], so neglecting these effects could limit the accuracy of the model. Gastegger *et al.* have proposed a strategy for describing induced polarization within an NNP, which could eventually resolve this limitation [110].

Another major issue with these ANI-type NNPs is that they cannot be used to describe charged molecules or even molecules with charged functional groups. ANI also does not contain any charged molecules in its training set and therefore cannot be used with charged molecules. This was evident in Chapter 3, where the conformational energies of drugs containing charged functional groups were not physically realistic. The development of NNPs trained to describe charged functional groups and incorporation of effects for long-range electrostatic interactions would allow a broader set of ligands to be described.

Lastly, the NNP/MM models we use rely on conventional Lennard-Jones potentials to describe the dispersion and Pauli repulsion interactions between the NNP and MM regions. These require the definition of pairwise Lennard-Jones well-depth and atomic radii parameters. We have used standard force field combination rules with tabulated parameters, although QM analysis has shown that these parameters may overestimate the strength of C_6 dispersion interactions [15, 14]. More complex non-bonded interactions have been proposed, which use a more realistic exponential function to describe Pauli repulsive interactions and include C_8 dispersion [111]. The use of these improved potentials may further improve the accuracy of NNP/MM simulations.

One of the major reasons for applying NNPs to protein–ligand interactions is that they are as efficient and easy to use as the conventional MM models. With regards to computational cost, the ANI NNP's scale linearly and have performance that is comparable to MM force fields, so formally, NNP/MM simulations could have similar performance to pure MM simulations. For practical reasons, our NNP/MM code integrated into NAMD through its QM/MM interface, which is used to call

the TorchANI python-based ANI-NNP code. In turn, TorchANI calls the PyTorch library to calculate the NN's. If ANI NNP's were directly implemented into NAMD, the computational efficiency would likely improve considerably and it would make NNP/MM more accessible for other researchers who may be familiar with conventional molecular mechanical models but not NNP's.

Our NNP/MM method was also promising for the refinement of cryo-EM structures when combined with MDFF. Although we showed this method could refine a structure where there was already an atomistic model available, this method would be more widely applicable if it could be used to determine the ligand pose without an existing atomistic model. It would also be necessary to prove that the accuracy of high resolution structures can be achieved using low-resolution data. This work is underway by using high resolution XRD maps that are artificially coarsened to the resolution typical of cryo-EM experiments, then used as the inputs of NNP/MM MMDF simulation to see if the original structures determined using the high resolution XRD data can be recovered.

Throughout this thesis, ANI-type NNPs were shown to be effective models for predicting the structures, conformational energies, and dynamics of drug-like molecules. The modest computational cost of these models in comparison to QM methods allows for them to be used in molecular dynamics simulations. Although these methods are not yet mature enough to perform a realistic simulation of a complete protein-ligand complex, we have found our NNP/MM method is an effective way to take advantage of ANI-NNP's accurate description of intramolecular interactions of small organic molecules, while the solvent, ions, and protein can be described using the mature MM force fields for these components (e.g., CHARMM). These simulations can be extended further by using enhanced sampling methods, like umbrella sampling, ABF, and MDFF.

The main theme of this thesis is that NNP's provide a powerful new way to model protein–ligand complexes. They are generally more accurate and reliable than popular MM models, without the need for parameterization, which makes them very transferable from system to system. Although they are more computationally intensive than MM models, this cost is tractable on modern computing facilities. Expanding the role of computer modeling in drug development will require methods that are more accurate but are also computationally efficient. Uniquely, NNP's have an efficiency

comparable to MM models but have similar accuracy to the QM models they are trained to reproduce. They provide a promising path forward in the evolution of molecular simulation methods.

Appendix A

Supporting Information for Chapter 3

Table A.1: Table of crystallographic data. Ligands that did not have a common drug name are listed by their Drugbank ID or their compound ID number(CID).

PDB ID	Protein	EC	Ligand	Resolution (Å)	Temp (K)
1XOZ [112]	phosphodi-esterase 5A	3.1.4.17	tadalafil	1.37	93
3EYG [113]	JAK1	2.7.10.2	tofacitinib	1.9	100
2W6N [114]	biotin carboxylase	6.3.4.14	DB08315	1.87	100
4HJO [115]	EGFR	2.7.10.1	erlotinib	2.21	110
4NCT [116]	Human DYRK1A	2.7.12.1	midostaurin	2.6	100
2HYY [117]	Abl kinase	2.7.10.2	imatinib	2.4	100
3EIG [118]	dihydrofolate reductase	1.5.1.3	methotrexate	1.7	93
3ETA [119]	IGF-1R	2.7.10.1	CID 45272927	2.6	93

Table A.2: Table of RMSD of the calculated structures of the ligands relative to the PDB structure.

PDB ID	RMSD (Å)	
	CGenFF	NNP/MM
1XOZ	0.19	0.13
3EYG	0.24	0.50
2W6N	0.54	0.57
4HJO	0.79	0.59
4NCT	1.1	0.60
2HYY	0.39	0.42
3EIG	0.37	0.28
3ETA	0.26	0.35

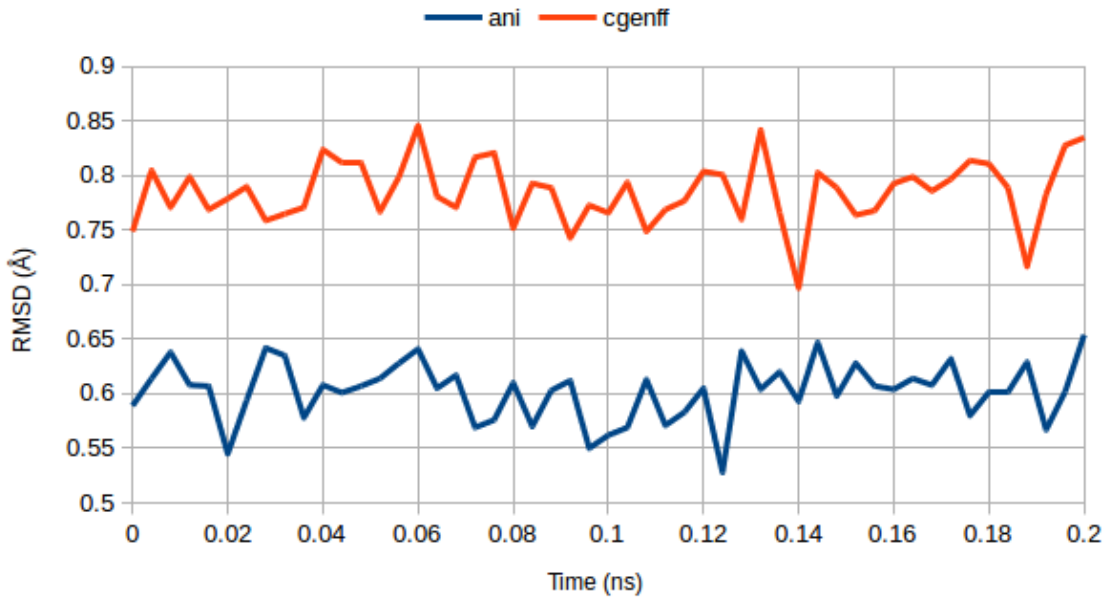

Figure A.1: Trajectory of the RMSD of the calculated structure of 4HJO vs the PDB structure vs time.

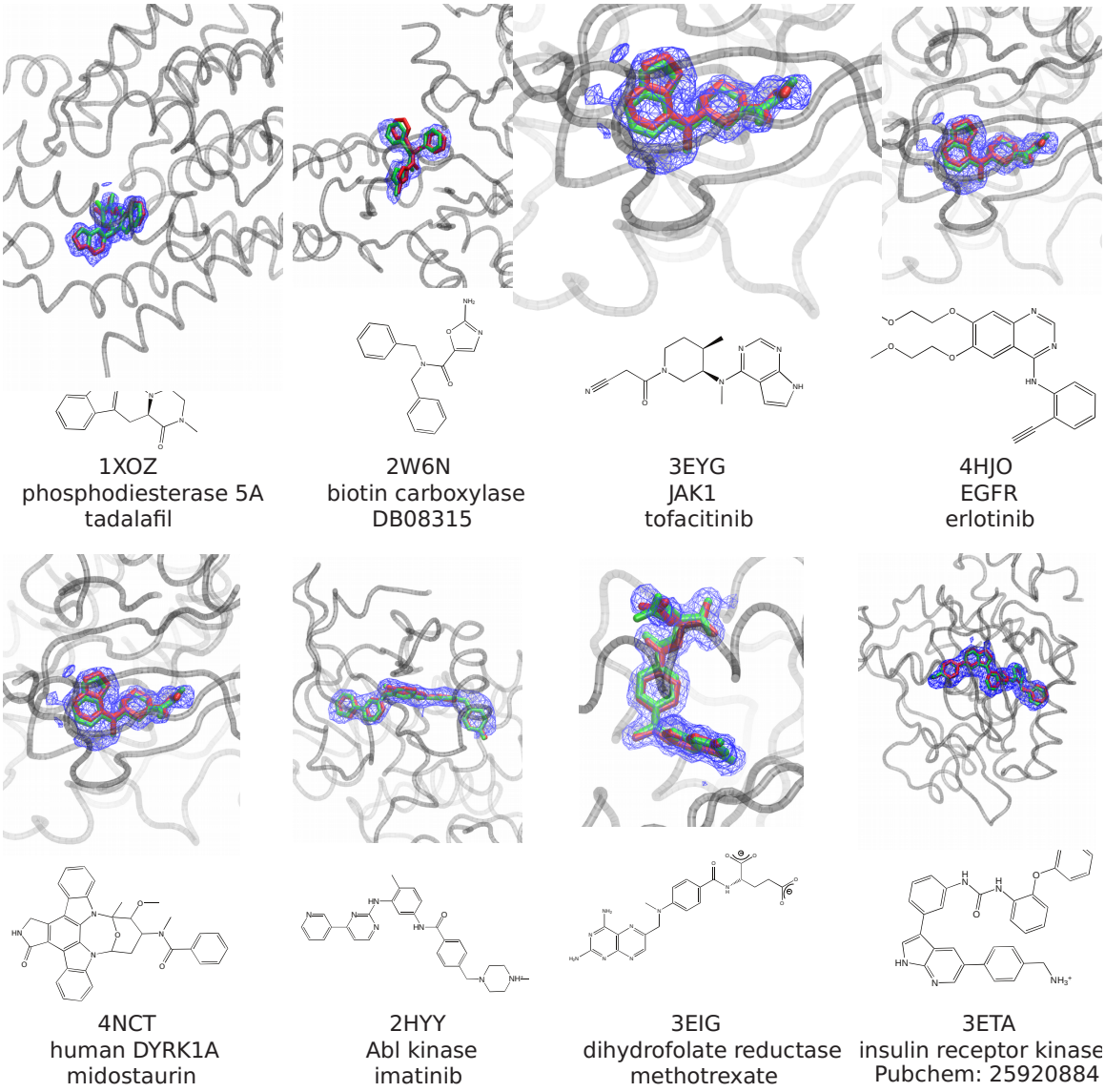

Figure A.2: Calculated poses of ligands. CGenFF is in green and ANI is in red. The crystallographic electron density of the ligands are shown in blue. The PDB ID, protein name, and ligand name are included beneath the image.

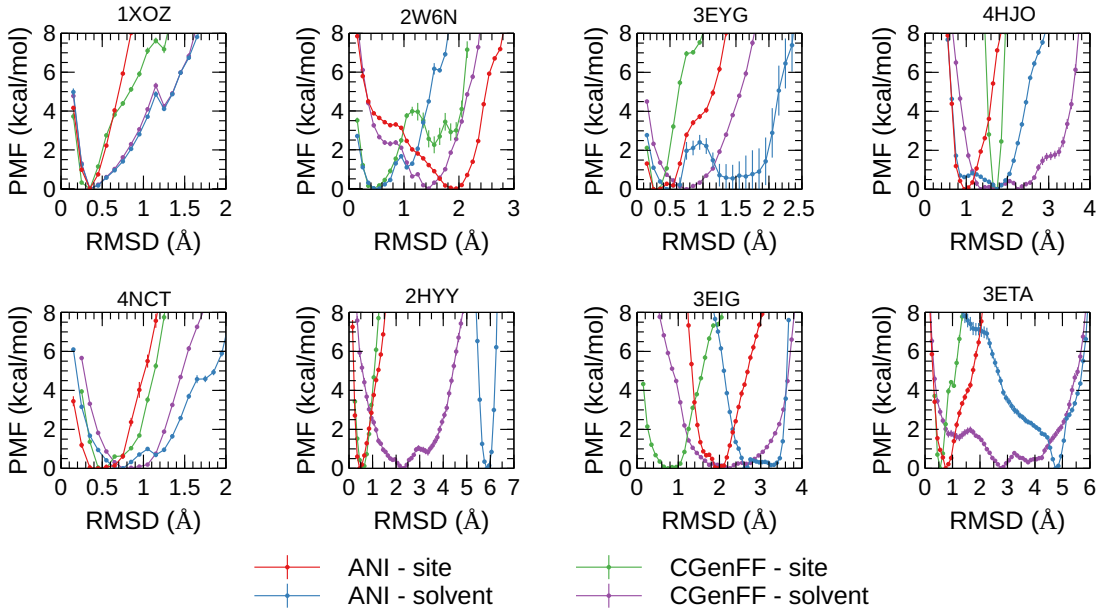

Figure A.3: The potential of mean force for the deviation of the structure of a ligand from its bound conformation when it is bound to its protein target.

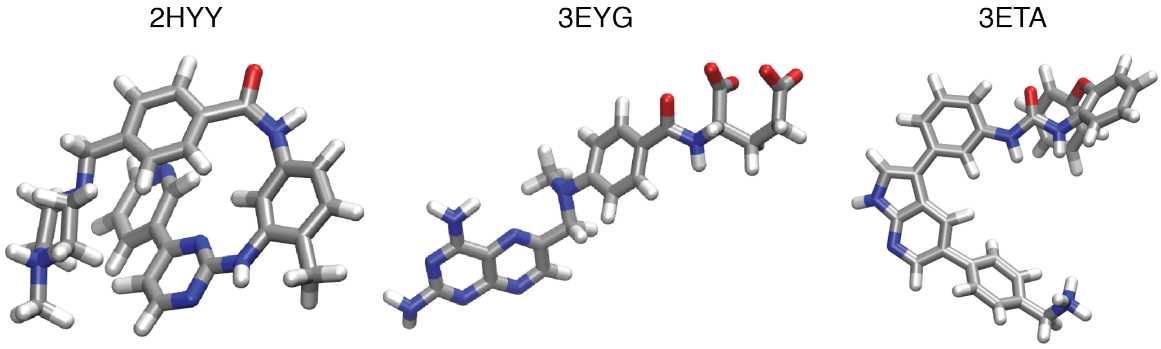

Figure A.4: The representative structures of the ionic ligands from the lowest energy of the PMF for the NNP/MM simulations of the ligands in explicit water (the solvent is not shown for clarity). The charged functional groups of the ligands form spurious intramolecular contacts

A.0.1 NNP Technical Details

The ANI-1ccX model was used in all instances where NNP calculations. The TorchANI implementation was used [46]. No additional intramolecular force terms for the ligand were included. In this NNP, there are 16 radial elements and the radial cutoff is 5.2 Å. The complete details at described in the Supplementary Materials of Smith *et al.* [22]

The parameters of the NNP used here are defined in the TorchANI parameter file:

```
TM = 1
Rcr = 5.2000e+00
Rca = 3.5000e+00
EtaR = [1.6000000e+01]
ShfR = [9.0000000e-01,1.1687500e+00,1.4375000e+00,1.7062500e+00,
1.9750000e+00,2.2437500e+00,2.5125000e+00,2.7812500e+00,
3.0500000e+00,3.3187500e+00,3.5875000e+00,3.8562500e+00,
4.1250000e+00,4.3937500e+00,4.6625000e+00,4.9312500e+00]
Zeta = [3.2000000e+01]
ShfZ = [1.9634954e-01,5.8904862e-01,9.8174770e-01,1.3744468e+00,
1.7671459e+00,2.1598449e+00,2.5525440e+00,2.9452431e+00]
EtaA = [8.0000000e+00]
ShfA = [9.0000000e-01,1.5500000e+00,2.2000000e+00,2.8500000e+00]
Atyp = [H,C,N,O]
```

A.0.2 NAMD Input File Section for NNP/MM Simulation

```
qmforces on
qmParamPDB qmmm.pdb
qmSoftware custom
qmexecpath client.py
qmBaseDir /dev/shm/
QMColumn occ
qmChargeMode none
qmElecEmbed off
```

A.0.3 Sample ORCA RI-MP2 Input File

```
! RI-MP2 RIJCOSX def2-TZVP def2-TZVP/C def2/J TIGHTSCF Opt PAL8
% maxcore 15000

* xyzfile 0 1 pes.xyz
%geom
Constraints
{D 2 3 4 5 0.0 C }
{D 3 4 5 6 0.0 C }
end
end
```

A.0.4 Sample ORCA DLPNO-CCSD(T) Input File

```
! DLPNO-CCSD(T) def2-TZVP def2-TZVP/C TIGHTSCF
% pal nprocs 8 end
% maxcore 15000

* xyzfile 0 1 pes.xyz
end
end
```

Bibliography

[1] X. Du, Y. Li, Y.-L. Xia, S.-M. Ai, J. Liang, P. Sang, X.-L. Ji, and S.-Q. Liu. Insights into protein–ligand interactions: Mechanisms, models, and methods. *International Journal of Molecular Sciences*, 17(2), 2016.

[2] L. Maveyraud and L. Mourey. Protein X-ray crystallography and drug discovery. *Molecules*, 25(5), 2020.

[3] J. C. Kendrew, R. E. Dickerson, B. E. Strandberg, R. G. Hart, D. R. Davies, D. C. Phillips, and V. C. Shore. Structure of myoglobin: A three-dimensional Fourier synthesis at 2 å resolution. *Nature*, 185(4711):422–427, 1960.

[4] M. S. Smyth and J. H. Martin. X-ray crystallography. *Mol. Pathol.*, 53(1):8–14, 2000.

[5] M. W. Parker. Protein structure from X-ray diffraction. *J. Biol. Phys.*, 29(4):341–362, 2003.

[6] J. Dubochet, M. Adrian, J. J. Chang, J. C. Homo, J. Lepault, A. W. McDowall, and P. Schultz. Cryo-electron microscopy of vitrified specimens. *Q. Rev. Biophys.*, 21(2):129–228, 1988.

[7] T. Kato, F. Makino, T. Nakane, N. Terahara, T. Kaneko, Y. Shimizu, S. Motoki, I. Ishikawa, K. Yonekura, and K. Namba. CryoTEM with a cold field emission gun that moves structural biology into a new stage. *Microsc. Microanal.*, 25(S2):998–999, 2019.

[8] K. M. Yip, N. Fischer, E. Paknia, A. Chari, and H. Stark. Atomic-resolution protein structure determination by cryo-EM. *Nature*, 587(7832):157–161, 2020.

[9] O. G. Kenno Vanommeslaeghe and A. D. Mackerell Jr. Molecular mechanics. *Curr. Pharm. Des.*, 20(20):3281–3292, 2014.

[10] F. Sajadi and C. Rowley. Simulations of lipid bilayers using the CHARMM36 force field with the TIP3P-FB and TIP4P-FB water models. *PeerJ*, 6, 2018.

[11] J. Wang, W. Wang, P. A. Kollman, and D. A. Case. Automatic atom type and bond type perception in molecular mechanical calculations. *J. Mol. Graph. Model.*, 25(2):247–260, 2006.

[12] K. Vanommeslaeghe, E. P. Raman, and A. D. MacKerell. Automation of the CHARMM general force field (CGenFF) ii: Assignment of bonded parameters and partial atomic charges. *J. Chem. Inf. Model.*, 52(12):3155–3168, 2012.

[13] M. Riquelme, A. Lara, D. L. Mobley, T. Verstraelen, A. R. Matamala, and E. Vöhringer-Martinez. Hydration free energies in the freesolv database calculated with polarized iterative hirshfeld charges. *J. Chem. Inf. Model.*, 58(9):1779–1797, 2018.

[14] 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(12):6146–6157, 2017.

[15] E. T. Walters, M. Mohebifar, E. R. Johnson, and C. N. Rowley. Evaluating the london dispersion coefficients of protein force fields using the exchange-hole dipole moment model. *J. Phys. Chem. B*, 122(26):6690–6701, 2018.

[16] 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.

[17] J. Behler and M. Parrinello. Generalized neural-network representation of high-dimensional potential-energy surfaces. *Phys. Rev. Lett.*, 98(14):146401, 2007.

[18] J. Behler. First principles neural network potentials for reactive simulations of large molecular and condensed systems. *Angew. Chem. Int. Ed.*, 56(42):12828–12840, 2017.

[19] 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(4):3192–3203, 2017.

[20] C. Devereux, J. S. Smith, K. K. Davis, 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.*, 0(0):null, 0.

[21] T. Fink and J.-L. Reymond. Virtual exploration of the chemical universe up to 11 atoms of C, N, O, F: Assembly of 26.4 million structures (110.9 million stereoisomers) and analysis for new ring systems, stereochemistry, physicochemical properties, compound classes, and drug discovery. *J. Chem. Inf. Model.*, 47(2):342–353, 2007.

[22] J. S. Smith, B. T. Nebgen, R. Zubatyuk, N. Lubbers, C. Devereux, K. Barros, S. Tretiak, O. Isayev, and A. E. Roitberg. Approaching coupled cluster accuracy with a general-purpose neural network potential through transfer learning. *Nat. Comm.*, 10(1):2903, 2019.

[23] M. F. Horstemeyer. *Multiscale Modeling: A Review*, pages 87–135. Springer Netherlands, Dordrecht, 2010.

[24] H. M. Senn and W. Thiel. Qm/mm methods for biomolecular systems. *Angew. Chem. Int. Ed.*, 48(7):1198–1229, 2009.

[25] C. N. Rowley and T. K. Woo. Counteranion effects on the zirconocene polymerization catalyst olefin complex from qm/mm molecular dynamics simulations. *Organometallics*, 30(8):2071–2074, 2011.

[26] S. Riahi, B. Roux, and C. N. Rowley. Qm/mm molecular dynamics simulations of the hydration of mg(ii) and zn(ii) ions. *Can. J. Chem.*, 91(7):552–558, 2013.

[27] S. Riahi and C. N. Rowley. The charmm–turbomole interface for efficient and accurate qm/mm molecular dynamics, free energies, and excited state properties. *J. Comput. Chem.*, 35(28):2076–2086, 2014.

[28] E. Awoonor-Williams and C. N. Rowley. The hydration structure of methylthiolate from qm/mm molecular dynamics. *J. Chem. Phys*, 149(4):045103, 2018.

[29] J. W. Vant, S.-L. J. Lahey, K. Jana, M. Shekhar, D. Sarkar, B. H. Munk, U. Kleinekathöfer, S. Mittal, C. Rowley, and A. Singharoy. Flexible fitting of small molecules into electron microscopy maps using molecular dynamics simulations with neural network potentials. *J. Chem. Inf. Model.*, 60(5):2591–2604, 2020.

[30] S.-L. J. Lahey and C. N. Rowley. Simulating protein–ligand binding with neural network potentials. *Chem. Sci.*, 11:2362–2368, 2020.

[31] J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman, and D. A. Case. Development and testing of a general amber force field. *J. Comput. Chem.*, 25(9):1157–1174, 2004.

[32] Y. Qiu, D. Smith, S. Boothroyd, H. Jang, J. Wagner, C. Bannan, T. Gokey, V. Lim, C. Stern, A. Rizzi, X. Lucas, B. Tjanaka, M. Shirts, M. Gilson, J. Chodera, C. Bayly, D. Mobley, and L.-P. Wang. Development and benchmarking of open force field v1.0.0, the parsley small molecule force field. *ChemRxiv*, 2020.

[33] K. Vanommeslaeghe, E. Hatcher, C. Acharya, S. Kundu, S. Zhong, J. Shim, E. Darian, O. Guvench, P. Lopes, I. Vorobyov, and A. D. Mackerell Jr. CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields. *J. Comput. Chem.*, 31(4):671–690, 2010.

[34] K. Vanommeslaeghe, E. Hatcher, C. Acharya, S. Kundu, S. Zhong, J. Shim, E. Darian, O. Guvench, P. Lopes, I. Vorobyov, and A. D. Mackerell Jr. CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields. *J. Comput. Chem.*, 31(4):671–690, 2010.

[35] E. Harder, W. Damm, J. Maple, C. Wu, M. Reboul, J. Y. Xiang, L. Wang, D. Lupyan, M. K. Dahlgren, J. L. Knight, and et al. OPLS3: A force field providing broad coverage of drug-like small molecules and proteins. *J. Chem. Theory Comput.*, 12(1):281–296, 2016.

[36] S. M. Gutiérrez Sanfeliciano and J. M. Schaus. Rapid assessment of conformational preferences in biaryl and aryl carbonyl fragments. *PLOS ONE*, 13(3):1–21, 2018.

[37] D. L. Mobley, C. C. Bannan, A. Rizzi, C. I. Bayly, J. D. Chodera, V. T. Lim, N. M. Lim, K. A. Beauchamp, D. R. Slochower, M. R. Shirts, M. K. Gilson, and P. K. Eastman. Escaping atom types in force fields using direct chemical perception. *J. Chem. Theory Comput.*, 14(11):6076–6092, 2018.

[38] M. K. Dahlgren, P. Schyman, J. Tirado-Rives, and W. L. Jorgensen. Characterization of biaryl torsional energetics and its treatment in OPLS all-atom force fields. *J. Chem. Inf. Model.*, 53(5):1191–1199, 2013.

[39] C. Rowley. Repository of biaryl test set structures and topology files. `https://github.com/RowleyGroup/torsionbenchmark`, 2019.

[40] B. R. Brooks, C. L. Brooks III, A. D. Mackerell Jr., L. Nilsson, R. J. Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels, S. Boresch, A. Caflisch, L. Caves, Q. Cui, A. R. Dinner, M. Feig, S. Fischer, J. Gao, M. Hodoscek, W. Im, K. Kuczera, T. Lazaridis, J. Ma, V. Ovchinnikov, E. Paci, R. W. Pastor, C. B. Post, J. Z. Pu, M. Schaefer, B. Tidor, R. M. Venable, H. L. Woodcock, X. Wu, W. Yang, D. M. York, and M. Karplus. CHARMM: The biomolecular simulation program. *J. Comput. Chem.*, 30(10):1545–1614, 2009.

[41] C. I. Bayly, P. Cieplak, W. Cornell, and P. A. Kollman. A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: the RESP model. *J. Phys. Chem.*, 97(40):10269–10280, 1993.

[42] W. L. Jorgensen and J. Tirado-Rives. Potential energy functions for atomic-level simulations of water and organic and biomolecular systems. *Proc. Natl. Acad. Sci. U.S.A.*, 102(19):6665–6670, 2005.

[43] L. S. Dodda, I. Cabeza de Vaca, J. Tirado-Rives, and W. L. Jorgensen. LigParGen web server: an automatic OPLS-AA parameter generator for organic ligands. *Nucleic Acids Res.*, 45(W1):W331–W336, 2017.

[44] L. S. Dodda, J. Z. Vilseck, J. Tirado-Rives, and W. L. Jorgensen. 1.14*CM1A-LBCC: Localized bond-charge corrected cm1a charges for condensed-phase simulations. *J. Phys. Chem. B*, 121(15):3864–3870, 2017.

[45] P. Eastman, J. Swails, J. D. Chodera, R. T. McGibbon, Y. Zhao, K. A. Beauchamp, L.-P. Wang, A. C. Simmonett, M. P. Harrigan, C. D. Stern, R. P. Wiewiora, B. R. Brooks, and V. S. Pande. OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. *PLOS Computational Biology*, 13(7):1–17, 2017.

[46] 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(7):3408–3415, 2020.

[47] M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, B. Mennucci, G. A. Petersson, H. Nakatsuji, M. Caricato, X. Li, H. P. Hratchian, A. F. Izmaylov, J. Bloino, G. Zheng, J. L. Sonnenberg, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, J. A. Montgomery, Jr., J. E. Peralta, F. Ogliaro, M. Bearpark, J. J. Heyd, E. Brothers, K. N. Kudin, V. N. Staroverov, R. Kobayashi, J. Normand, K. Raghavachari, A. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, N. Rega, J. M. Millam, M. Klene, J. E. Knox, J. B. Cross, V. Bakken, C. Adamo, J. Jaramillo, R. Gomperts, R. E. Stratmann, O. Yazyev, A. J. Austin, R. Cammi, C. Pomelli, J. W. Ochterski, R. L. Martin, K. Morokuma, V. G. Zakrzewski, G. A. Voth, P. Salvador, J. J. Dannenberg, S. Dapprich, A. D. Daniels, Ö. Farkas, J. B. Foresman, J. V. Ortiz, J. Cioslowski, and D. J. Fox. Gaussian 09 Revision D.01. Gaussian Inc. Wallingford CT 2009.

[48] C. Rowley. NAMD–TorchANI interface scripts. `https://github.com/RowleyGroup/NNP-MM`, 2019.

[49] C. Møller and M. S. Plesset. Note on an approximation treatment for many-electron systems. *Phys. Rev.*, 46:618–622, 1934.

[50] C. Hättig. Optimization of auxiliary basis sets for ri-mp2 and ri-cc2 calculations: Core–valence and quintuple-ζ basis sets for h to ar and QZVPP basis sets for li to kr. *Phys. Chem. Chem. Phys.*, 7:59–66, 2005.

[51] F. Weigend and R. Ahlrichs. Balanced basis sets of split valence, triple zeta

valence and quadruple zeta valence quality for h to rn: Design and assessment of accuracy. *Phys. Chem. Chem. Phys.*, 7:3297–3305, 2005.

[52] F. Neese. The orca program system. *Wiley Interdiscip. Rev. Comput. Mol. Sci*, 2(1):73–78, 2012.

[53] Y. Guo, C. Riplinger, U. Becker, D. G. Liakos, Y. Minenkov, L. Cavallo, and F. Neese. Communication: An improved linear scaling perturbative triples correction for the domain based local pair-natural orbital based singles and doubles coupled cluster method [DLPNO-CCSD(T)]. *J. Chem. Phys.*, 148(1):011101, 2018.

[54] J. C. Phillips, D. a. J. Hardy, J. D. C. Maia, J. E. Stone, J. a. V. Ribeiro, R. C. Bernardi, R. Buch, G. Fiorin, J. H'enin, W. Jiang, R. McGreevy, M. C. R. Melo, B. K. Radak, R. D. Skeel, A. Singharoy, Y. Wang, B. Roux, A. Aksimentiev, Z. Luthey-Schulten, L. V. Kalé, K. Schulten, C. Chipot, and E. Tajkhorshid. Scalable molecular dynamics on CPU and GPU architectures with NAMD. *J. Chem. Phys.*, 153(4):044130, 2020.

[55] L.-P. Wang, T. J. Martinez, and V. S. Pande. Building force fields: An automatic, systematic, and reproducible approach. *J. Phys. Chem. Lett.*, 5(11):1885–1891, 2014.

[56] E. A. Koopman and C. P. Lowe. Advantages of a lowe–andersen thermostat in molecular dynamics simulations. *J. Chem. Phys.*, 124(20):204103, 2006.

[57] E. Darve, D. Rodríguez-Gómez, and A. Pohorille. Adaptive biasing force method for scalar and vector free energy calculations. *J. Chem. Phys.*, 128(14):144120, 2008.

[58] J. Hénin, G. Fiorin, C. Chipot, and M. L. Klein. Exploring multidimensional free energy landscapes using time-dependent biases on collective variables. *J. Chem. Theory Comput.*, 6(1):35–47, 2010.

[59] G. M. Torrie and J. P. Valleau. Nonphysical sampling distributions in monte carlo free-energy estimation: Umbrella sampling. *J. Comput. Phys.*, 23:187–199, 1977.

[60] K. Johannes. Umbrella sampling. *Wiley Interdiscip. Rev. Comput. Mol. Sci.*, 1(6):932–942, 2011.

[61] S. Kumar, J. M. Rosenberg, D. Bouzida, R. H. Swendsen, and P. A. Kollman. The weighted histogram analysis method for free-energy calculations on biomolecules. i. the method. *J. Comput. Chem.*, 13(8):1011–1021, 1992.

[62] B. Roux. The calculation of the potential of mean force using computer simulations. *Comput. Phys. Commun.*, 91(1–3):275–282, 1995.

[63] A. Grossfield. WHAM: the weighted histogram analysis method, version 2.0.6, http://membrane.urmc.rochester.edu/content/wham, 2018.

[64] S.-L. J. Lahey, T. N. Thien Phuc, and C. N. Rowley. Benchmarking force field and the ani neural network potentials for the torsional potential energy surface of biaryl drug fragments. *J. Chem. Inf. Model.*, 60(12):6258–6268, 2020.

[65] J. A. Cohen, D. L. Arnold, G. Comi, A. Bar-Or, S. Gujrathi, J. P. Hartung, M. Cravets, A. Olson, P. A. Frohna, and K. W. Selmaj. Safety and efficacy of the selective sphingosine 1-phosphate receptor modulator ozanimod in relapsing multiple sclerosis (radiance): a randomised, placebo-controlled, phase 2 trial. *Lancet Neurol.*, 15(4):373 – 381, 2016.

[66] H. Kramers. Brownian motion in a field of force and the diffusion model of chemical reactions. *Physica*, 7(4):284 – 304, 1940.

[67] T. B. Woolf and B. Roux. Conformational flexibility of o-phosphorylcholine and o-phosphorylethanolamine: A molecular dynamics study of solvation effects. *J. Am. Chem. Soc.*, 116(13):5916–5926, 1994.

[68] J. E. Straub, M. Borkovec, and B. J. Berne. Calculation of dynamic friction on intramolecular degrees of freedom. *J. Phys. Chem.*, 91(19):4995–4998, 1987.

[69] G. Hummer. Position-dependent diffusion coefficients and free energies from bayesian analysis of equilibrium and replica molecular dynamics simulations. *New J. Phys.*, 7:34, 2005.

[70] K. Gaalswyk, E. Awoonor-Williams, and C. N. Rowley. Generalized langevin methods for calculating transmembrane diffusivity. *J. Chem. Theory Comput.*, 12(11):5609–5619, 2016.

[71] J. Dowell, J. D. Minna, and P. Kirkpatrick. Erlotinib hydrochloride. *Nat. Rev. Drug Discov.*, 4(1):13–14, 2005.

[72] D. Bakowies and W. Thiel. Hybrid models for combined quantum mechanical and molecular mechanical approaches. *J. Phys. Chem.*, 100(25):10580–10594, 1996.

[73] M. P. Gleeson and D. Gleeson. QM/MM calculations in drug discovery: A useful method for studying binding phenomena? *J. Chem. Inf. Model.*, 49(3):670–677, 2009.

[74] Z. Fu, X. Li, and K. M. Merz Jr. Accurate assessment of the strain energy in a protein-bound drug using QM/MM X-ray refinement and converged quantum chemistry. *J. Comput. Chem.*, 32(12):2587–2597, 2011.

[75] K. D. Dubey and R. P. Ojha. Binding free energy calculation with QM/MM hybrid methods for Abl-Kinase inhibitor. *J. Biol. Phys.*, 37(1):69–78, 2011.

[76] J. C. Phillips, R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R. D. Skeel, L. Kalé, and K. Schulten. Scalable molecular dynamics with NAMD. *J. Comput. Chem.*, 26(16):1781–1802, 2005.

[77] M. C. R. Melo, R. C. Bernardi, T. Rudack, M. Scheurer, C. Riplinger, J. C. Phillips, J. D. C. Maia, G. B. Rocha, J. V. Ribeiro, J. E. Stone, and et al. NAMD goes quantum: an integrative suite for hybrid simulations. *Nat. Methods*, 15(5):351–354, 2018.

[78] J. Huang, S. Rauscher, G. Nawrocki, T. Ran, M. Feig, B. L. de Groot, H. Grubmüller, and J. MacKerell, Alexander D. CHARMM36m: an improved force field for folded and intrinsically disordered proteins. *Nat. Methods*, 14(1):71–73, 2017.

[79] 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(2):926–935, 1983.

[80] E. Neria, S. Fischer, and M. Karplus. Simulation of activation free energies in molecular systems. *J. Chem. Phys.*, 105(5):1902–1921, 1996.

[81] Y. Guo, C. Riplinger, U. Becker, D. G. Liakos, Y. Minenkov, L. Cavallo, and F. Neese. Communication: An improved linear scaling perturbative triples correction for the domain based local pair-natural orbital based singles and doubles coupled cluster method [dlpno-ccsd(t)]. *J. Chem. Phys.*, 148(1):011101, 2018.

[82] R. Anandakrishnan, B. Aguilar, and A. V. Onufriev. H++ 3.0: automating pK prediction and the preparation of biomolecular structures for atomistic molecular modeling and simulations. *Nucleic Acids Res.*, 40(W1):W537–W541, 2012.

[83] J. Wang, Y. Deng, and B. Roux. Absolute binding free energy calculations using molecular dynamics simulations with restraining potentials. *Biophys. J.*, 91(8):2798–2814, 2006.

[84] D. L. Mobley, J. D. Chodera, and K. A. Dill. Confine-and-release method: Obtaining correct binding free energies in the presence of protein conformational change. *J. Chem. Theory Comput.*, 3(4):1231–1235, 2007.

[85] D. L. Mobley and K. A. Dill. Binding of small-molecule ligands to proteins: "what you see" is not always "what you get". *Structure*, 17(4):489–498, 2009.

[86] J. C. Gumbart, B. Roux, and C. Chipot. Efficient determination of protein–protein standard binding free energies from first principles. *J. Chem. Theory Comput.*, 9(8):3789–3798, 2013.

[87] Y.-L. Lin, Y. Meng, W. Jiang, and B. Roux. Explaining why gleevec is a specific and potent inhibitor of abl kinase. *Proc. Natl. Acad. Sci. U.S.A.*, 110(5):1664–1669, 2013.

[88] P. M. Zimmerman, M. Head-Gordon, and A. T. Bell. Selection and validation of charge and lennard-jones parameters for QM/MM simulations of hydrocarbon interactions with zeolites. *J. Chem. Theory Comput.*, 7(6):1695–1703, 2011.

[89] C. N. Rowley and B. Roux. The solvation structure of na^+ and k^+ in liquid water determined from high level ab initio molecular dynamics simulations. *J. Chem. Theory Comput.*, 8(10):3526–3535, 2012.

[90] L. G. Trabuco, E. Villa, E. Schreiner, C. B. Harrison, and K. Schulten. Molecular dynamics flexible fitting: a practical guide to combine cryo-electron microscopy and X-ray crystallography. *Methods*, 49(2):174–180, 2009.

[91] J. Hsin, J. Gumbart, L. G. Trabuco, E. Villa, P. Qian, C. N. Hunter, and K. Schulten. Protein-induced membrane curvature investigated through molecular dynamics flexible fitting. *Biophys. J.*, 97(1):321–329, 2009.

[92] M. Sener, J. Hsin, L. G. Trabuco, E. Villa, P. Qian, C. N. Hunter, and K. Schulten. Structural model and excitonic properties of the dimeric RC-LH1-PufX complex from rhodobacter sphaeroides. *Chem. Phys.*, 357(1-3):188–197, 2009.

[93] H. Kim, J. Hsin, Y. Liu, P. R. Selvin, and K. Schulten. Formation of salt bridges mediates internal dimerization of myosin VI medial tail domain. *Structure*, 18(11):1443–1449, 2010.

[94] K. Zhang, L. Wang, Y. Liu, K.-Y. Chan, X. Pang, K. Schulten, Z. Dong, and F. Sun. Flexible interwoven termini determine the thermal stability of thermosomes. *Protein Cell*, 4(6):432–444, 2013.

[95] C. K. Cassidy, B. A. Himes, F. J. Alvarez, J. Ma, G. Zhao, J. R. Perilla, K. Schulten, and P. Zhang. CryoEM and computer simulations reveal a novel kinase conformational switch in bacterial chemotaxis signaling. *Elife*, 4, 2015.

[96] X. Wang, F. Xu, J. Liu, B. Gao, Y. Liu, Y. Zhai, J. Ma, K. Zhang, T. S. Baker, K. Schulten, D. Zheng, H. Pang, and F. Sun. Atomic model of rabbit hemorrhagic disease virus by cryo-electron microscopy and crystallography. *PLoS Pathog.*, 9(1):e1003132, 2013.

[97] G. Zhao, J. R. Perilla, E. L. Yufenyuy, X. Meng, B. Chen, J. Ning, J. Ahn, A. M. Gronenborn, K. Schulten, C. Aiken, and P. Zhang. Mature HIV-1 capsid structure by cryo-electron microscopy and all-atom molecular dynamics. *Nature*, 497(7451):643–646, 2013.

[98] M. Lorenz and K. C. Holmes. The actin-myosin interface. *Proc. Natl. Acad. Sci. U. S. A.*, 107(28):12529–12534, 2010.

[99] T. A. M. Bharat, L. R. Castillo Menendez, W. J. H. Hagen, V. Lux, S. Igonet, M. Schorb, F. K. M. Schur, H.-G. Kräusslich, and J. A. G. Briggs. Cryo-electron microscopy of tubular arrays of HIV-1 gag resolves structures essential for immature virus assembly. *Proc. Natl. Acad. Sci. U. S. A.*, 111(22):8233–8238, 2014.

[100] F. K. M. Schur, W. J. H. Hagen, M. Rumlová, T. Ruml, B. Müller, H.-G. Kräusslich, and J. A. G. Briggs. Structure of the immature HIV-1 capsid in intact virus particles at 8.8 å resolution. *Nature*, 517(7535):505–508, 2015.

[101] A. Waterhouse, M. Bertoni, S. Bienert, G. Studer, G. Tauriello, R. Gumienny, F. T. Heer, T. A. de Beer, C. Rempfer, L. Bordoli, R. Lepore, and T. Schwede. SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Res.*, 46(W1):W296–W303, 05 2018.

[102] T. W. Ko, J. A. Finkler, S. Goedecker, and J. Behler. General-Purpose machine learning potentials capturing nonlocal charge transfer. *Acc. Chem. Res.*, 54(4):808–817, 2021.

[103] C. N. Rowley and B. Roux. The solvation structure of Na+ and K+ in liquid water determined from high level ab initio molecular dynamics simulations. *J. Chem. Theory Comput.*, 8(10):3526–3535, 2012.

[104] S. Riahi and C. N. Rowley. A Drude polarizable model for liquid hydrogen sulfide. *J. Phys. Chem. B*, 117(17):5222–5229, 2013.

[105] S. Riahi and C. N. Rowley. Solvation of hydrogen sulfide in liquid water and at the water–vapor interface using a polarizable force field. *J. Phys. Chem. B*, 118(5):1373–1380, 2014.

[106] S. Riahi and C. N. Rowley. Why can hydrogen sulfide permeate cell membranes? *J Am. Chem. Soc.*, 2014-10.

[107] A. N. S. Adluri, J. N. Murphy, T. Tozer, and C. N. Rowley. Polarizable force field with a σ-hole for liquid and aqueous bromomethane. *J. Phys. Chem. B*, 119(42):13422–13432, 2015.

[108] A. J. Hazel, E. T. Walters, C. N. Rowley, and J. C. Gumbart. Folding free energy landscapes of β-sheets with non-polarizable and polarizable CHARMM force fields. *J. Chem. Phys.*, 149(7):072317, 2018.

[109] V. S. Inakollu, D. P. Geerke, C. N. Rowley, and H. Yu. Polarisable force fields: what do they add in biomolecular simulations? *Curr. Opin. Struct. Biol.*, 61:182–190, 2020.

[110] M. Gastegger, K. T. Schütt, and K.-R. Müller. Machine learning of solvent effects on molecular spectra and reactions, 2020.

[111] M. Mohebifar and C. N. Rowley. An efficient and accurate model for water with an improved non-bonded potential. *J. Chem. Phys.*, 153(13):134105, 2020.

[112] G. L. Card, B. P. England, Y. Suzuki, D. Fong, B. Powell, B. Lee, C. Luu, M. Tabrizizad, S. Gillette, P. N. Ibrahim, D. R. Artis, G. Bollag, M. V. Milburn, S.-H. Kim, J. Schlessinger, and K. Y. Zhang. Structural basis for the activity of drugs that inhibit phosphodiesterases. *Structure*, 12(12):2233–2247, 2004.

[113] N. K. Williams, R. S. Bamert, O. Patel, C. Wang, P. M. Walden, A. F. Wilks, E. Fantino, J. Rossjohn, and I. S. Lucet. Dissecting specificity in the janus kinases: The structures of JAK-specific inhibitors complexed to the JAK1 and JAK2 protein tyrosine kinase domains. *J. Mol. Biol.*, 387(1):219–232, 2009.

[114] I. Mochalkin, J. R. Miller, L. Narasimhan, V. Thanabal, P. Erdman, P. B. Cox, J. V. N. V. Prasad, S. Lightle, M. D. Huband, and C. K. Stover. Discovery of antibacterial biotin carboxylase inhibitors by virtual screening and fragment-based approaches. *ACS Chemical Biology*, 4(6):473–483, 2009.

[115] J. H. Park, Y. Liu, M. A. Lemmon, and R. Radhakrishnan. Erlotinib binds both inactive and active conformations of the EGFR tyrosine kinase domain. *Biochemical Journal*, 448(3):417–423, 2012.

[116] M. Alexeeva, E. Åberg, R. A. Engh, and U. Rothweiler. The structure of a dual-specificity tyrosine phosphorylation-regulated kinase 1A–PKC412 complex reveals disulfide-bridge formation with the anomalous catalytic loop HRD(HCD) cysteine. *Acta Crystallographica Section D*, 71(5):1207–1215, 2015.

[117] S. W. Cowan-Jacob, G. Fendrich, A. Floersheimer, P. Furet, J. Liebetanz, G. Rummel, P. Rheinberger, M. Centeleghe, D. Fabbro, and P. W. Manley. Structural biology contributions to the discovery of drugs to treat chronic myelogenous leukaemia. *Acta Crystallographica Section D Biological Crystallography*, 63(1):80–93, 2006.

[118] J. P. Volpato, B. J. Yachnin, J. Blanchet, V. Guerrero, L. Poulin, E. Fossati, A. M. Berghuis, and J. N. Pelletier. Multiple conformers in active site of human

dihydrofolate reductase F31R/Q35E double mutant suggest structural basis for methotrexate resistance. *J. Biol. Chem.*, 284(30):20079–20089, 2009.

[119] S. Patnaik, K. L. Stevens, R. Gerding, F. Deanda, J. B. Shotwell, J. Tang, T. Hamajima, H. Nakamura, M. A. Leesnitzer, A. M. Hassell, L. M. Shewchuck, R. Kumar, H. Lei, and S. D. Chamberlain. Discovery of 3, 5-disubstituted-1h-pyrrolo[2, 3-b]pyridines as potent inhibitors of the insulin-like growth factor-1 receptor (IGF-1r) tyrosine kinase. *Bioorganic & Medicinal Chemistry Letters*, 19(11):3136–3140, 2009.