UC Berkeley
UC Berkeley Electronic Theses and Dissertations

Title

Advancing Computational Drug Discovery through Data Curation, Binding Free Energy Calculations, and Force Field Development

Permalink

https://escholarship.org/uc/item/9tt9q1n5

ISBN

9798173788986

Author

Wang, Yingze

Publication Date

2026-08-10

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

Advancing Computational Drug Discovery through Data Curation, Binding Free Energy Calculations, and Force Field Development

by

Yingze Wang

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Chemistry

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Teresa Head-Gordon, Chair
Professor Martin Head-Gordon
Professor Alexis Bell

Summer 2026

Advancing Computational Drug Discovery through Data Curation, Binding Free Energy Calculations, and Force Field Development

Copyright 2026

by

Yingze Wang

Abstract

Advancing Computational Drug Discovery through Data Curation, Binding Free Energy Calculations, and Force Field Development

by

Yingze Wang

Doctor of Philosophy in Chemistry

University of California, Berkeley

Professor Teresa Head-Gordon, Chair

The success rate of drug discovery has historically been limited by iterative trial and error across an immense chemical space, making the development of an effective drug extremely lengthy, costly, and risky. Structure-based methods offer a promising route toward rational drug design by using three-dimensional receptor–ligand structures and physics-based modeling to generate and prioritize molecular hypotheses. Their accuracy and success, however, remain limited by three challenges: the quality of the structural data used to construct and evaluate models, the development of efficient and high-fidelity physical interaction models, and the practicality of estimating protein–ligand binding free energies at the accuracy and throughput required for prospective applications. This dissertation correspondingly addresses these challenges through high-quality structural data curation (Chapter 2), development of next-generation advanced force fields (Chapter 3–5), and binding free-energy workflows with prospective applications (Chapter 6–7).

Chapter 2 describes HiQBind, an automated, open-source, and comprehensive workflow that prepares high-quality noncovalent protein–ligand complex structures with binding affinity annotations. This workflow combines stringent filters that control primary data quality, two fixing modules that correct ligand bond orders and protonation states and add missing protein atoms, and a refinement module that ensures final structural integrity. This pipeline was applied to produce a new dataset of more than 30,000 protein–ligand complexes with reliable structural and binding data from open sources, providing a useful resource for training and validating next-generation structure-based drug-discovery models. More broadly, this work provides the community with a transparent, reproducible, and freely accessible resource that bridges the gap between noisy public databases and proprietary curated datasets.

Chapters 3–5 advance the development of next-generation many-body polarizable force fields through reference data, software infrastructure, and model design. In Chapter 3, I con-

struct a protein–ligand interaction-energy dataset by extracting representative motifs from the HiQBind database developed in Chapter 2 and applying quantum-mechanical energy decomposition analysis. The resulting 14,905 noncovalent interaction-energy records across six types, together with their decomposition into electrostatics, Pauli repulsion, dispersion, polarization, and charge transfer, enable quantitative analysis of the energetic origins of protein–ligand interactions and provide physically interpretable targets for model development and validation. In Chapter 4, I develop TorchFF, a high-performance, GPU-accelerated differentiable force-field library by introducing customized CUDA operators within PyTorch to accelerate computational routines in molecular mechanics, such as neighbor list, multipolar electrostatics, and Ewald and particle mesh Ewald (PME) to treat long-ranged electrostatics. This design retains the automatic differentiation feature of PyTorch while significantly improving the performance, providing efficient and reusable infrastructure for gradient-based optimization and molecular dynamics with advanced force fields containing complicated physics terms. In Chapter 5, I improved the short-range intramolecular interactions in the AMOEBA polarizable force field with a simplified treatment. The component-specific scaling convention of AMOEBA is replaced by a unified scheme excluding all 1–4 nonbonded interactions, which are evaluated with a more comprehensive valence potential containing bond, angle, torsion and coupling terms. Along with a streamlined parameterization protocol, this model better reproduces quantum-mechanical potential energy surfaces of model dipeptides while achieving reasonable solution-phase conformational behavior, making it a viable alternative for modeling protein-backbone energetics.

Chapters 6–7 demonstrate how binding free energies across different levels of computational rigor and cost can accelerate real-world drug discovery campaigns. In Chapter 6, I present EasyBFE, an open-source package that automates alchemical protein–ligand binding free-energy calculations within a modular, user- and developer-friendly framework. I apply it to SARS-CoV-2 Mpro lead optimization and PGK2 hit identification, showing how relative and absolute binding free-energy calculations can support candidate prioritization at different stages of drug development. Chapter 7 introduces my participation in ASAP–Polaris–OpenADMET Challenge for predicting binding poses of unreleased SARS-CoV-2 and MERS-CoV Mpro ligands. An approach that re-ranked the outcomes of Autodock Vina with fragment-based scoring and MM/GBSA was developed and it ranked 6-th over all the $\sim$ 100 participants. The results revealed the sampling power of molecular docking and showed how to efficiently benefit from it with knowledge derived from prior data or rapid end-state binding free energy estimations grounded in more rigorous physics compared to empirical scoring functions.

In summary, this dissertation presents my endeavors across several aspects of structure-based drug discovery including data quality improvement, force field development, and binding free energy calculations. To generate practical impacts, computational models must be trained with trustworthy data, grounded in rigorous physics, organized in accessible automated

workflows, and evaluated in real-world drug discovery campaigns. Integrating these elements provides a systematic and reproducible route for turning *in silico* hypotheses into actionable drug-discovery outcomes.

Contents

1 Introduction **1**
1.1 Overview . . . 1
1.2 High-Quality Protein–Ligand Data Curation . . . 3
1.3 Advanced Force Field Development . . . 5
1.4 Binding Free Energy Workflows and Applications . . . 8

2 A Workflow to Create a High-Quality Protein–Ligand Binding Dataset for Training, Validation, and Prediction Tasks **11**
2.1 Introduction . . . 11
2.2 Methods . . . 13
2.3 Results . . . 19
2.3.1 The HiQBind Workflow . . . 20
2.3.2 Creation of the HiQBind Dataset . . . 23
2.4 Discussion and Conclusion . . . 27
2.5 Supporting Information . . . 29
2.5.1 Data Usage . . . 29
2.5.2 Failed cases in processing HiQBind . . . 33
2.5.3 Failed cases in processing PDBBind . . . 36

3 Energetics of Non-covalent Interactions of Protein–Ligand Complexes for Drug Discovery **44**
3.1 Introduction . . . 44
3.2 Methods . . . 46
3.2.1 Construction of Protein-Ligand Interacting Fragments . . . 46
3.2.2 Protein-Ligand Interacting Fragments by NCI Type . . . 48
3.2.3 Energy calculations and Energy Decomposition Analysis . . . 50
3.2.4 Benchmarking Force Fields and MLIPs . . . 50
3.3 Results . . . 51
3.3.1 Electrostatically dominated NCIs . . . 52
3.3.2 Pi-stacking NCIs are dominated by dispersion forces . . . 56

3.3.3 Weakly interacting NCIs . . . 58
3.3.4 Benchmarking Force Fields and MLIPs . . . 62
3.3.5 Trends Beyond Isolated NCI Fragment Dimers . . . 64
3.4 Conclusions . . . 66
3.5 Supporting Information . . . 68
3.5.1 Supporting Tables . . . 68
3.5.2 Supporting Figures . . . 75

4 TorchFF: A High-Performance GPU-Accelerated Differentiable Force Field Library 87
4.1 Introduction . . . 87
4.2 Methodology . . . 89
4.2.1 TorchFF Code Architecture . . . 89
4.2.2 Classical Potentials . . . 90
4.2.3 Ewald and Particle Mesh Ewald . . . 92
4.2.4 Neighbor List . . . 94
4.2.5 Completely Multipolar Model (CMM) . . . 96
4.3 Results . . . 99
4.3.1 Performance . . . 99
4.4 Conclusions and Outlook . . . 108
4.5 Supporting Information . . . 109
4.5.1 Derivation of Multipolar Ewald equations . . . 109
4.5.2 Derivation of Multipolar PME equations . . . 112
4.5.3 Correctness Validation of TorchFF Implementation . . . 116

5 Improved and Simplified Treatment of Short-Range Intramolecular Interactions in Polarizable Force Fields 118
5.1 Introduction . . . 118
5.2 Theory and Methods . . . 120
5.2.1 New Unified Nonbonded Scaling Scheme in AMOEBA . . . 120
5.2.2 Valence Potential . . . 123
5.2.3 Parameterization Procedure . . . 125
5.2.4 Replica-Exchange Molecular Dynamics for NMR Observables . . . 128
5.3 Results . . . 129
5.3.1 Validation of the Permanent-Dipole Refit . . . 129
5.3.2 Reproduction of Quantum-Mechanical Ramachandran Surfaces . . . 131
5.3.3 Energy and Force Accuracy on Validation Conformers . . . 136
5.3.4 Condensed-Phase Simulations and NMR Scalar Couplings . . . 138
5.4 Conclusions and Outlook . . . 141
5.5 Supporting Information . . . 143

5.6 Supplementary Tables 149

6 EasyBFE: An Open-source Package to Automate Protein-Ligand Binding Free Energy Calculations 151
6.1 Introduction 151
6.2 Theory 153
6.2.1 Software Design Principles 153
6.2.2 Alchemical Transformation and Free Energy Estimators 154
6.2.3 Molecular Dynamics 156
6.2.4 Relative Binding Free Energy 158
6.2.5 Absolute Binding Free Energy 162
6.3 Application 165
6.3.1 SARS-CoV-2 Main-Protease Lead Optimization 165
6.3.2 CACHE Challenge 7: De Novo Selective PGK2 Hit Design 168
6.4 Conclusion and Outlook 171

7 Polaris Challenge: Data-Driven Priors to Improve Docking for Pose Prediction 173
7.1 Introduction 173
7.2 Methods 175
7.2.1 Protein Ensemble Generation and Docking 176
7.2.2 Fragment Identification and Pose Generation 177
7.2.3 Custom Fragment-Based Scoring 178
7.2.4 Fallback MM/GBSA Scoring 179
7.3 Results 179
7.3.1 Overall Performance 180
7.3.2 Performance of Similar Test Ligands 181
7.3.3 Performance of MM/GBSA Rescoring 183
7.4 Discussion and Conclusion 186
7.5 Supporting Information 188
7.5.1 Additional Methodology Details 188
7.5.2 Supporting Figures 189

8 Conclusion 194

Acknowledgments

First of all, I am profoundly grateful to my PhD advisor, Prof. Teresa Head-Gordon for her continuous guidance, encouragement and support. She possesses an extraordinarily broad breadth of scientific interest alongside a remarkably rigorous attitude toward science, which has greatly influenced the way I think and conduct research. She also created a friendly, warm and intellectually stimulating environment in the lab that I really enjoyed, which also motivated me to think critically and grow as a scientist.

I also want to extend my appreciation to my excellent colleagues in the THG Lab. Specifically, thanks to Dr. Jie Li, Dr. Xingyi Guan, Dr. Oufan Zhang, Dr. Kunyang Sun, Dorian Bagni, Joseph Cavanagh, and Justin Purnomo for the wonderful teamwork on the drug discovery projects and insightful discussions across a wide range topics regarding machine learning and durg discovery. I also want to acknowledge Dr. Joseph Heindel, Dr. Selim Sami, Lukas Kim, Aalim Abdullah, and my mentee Joey Shin. It was my pleasure to work closely with these brilliant minds and appreciate their dedication to pushing the force field development forward.

My academic growth has also been generously supported by faculty members at UC Berkeley. I am grateful to my qualifying examination committee members—Prof. Martin Head-Gordon, Prof. Alexis Bell, Prof. John Groves, and Prof. Ronald Cohen—for their insightful feedback, constructive guidance to my research. I benefited greatly from Prof. Martin Head-Gordon's Quantum Mechanics course and Prof. David Limmer's Statistical Mechanics course that have consolidated my theoretical foundation in this field. Without a doubt, these were the best classes I have ever taken in those subjects. I would also like to express deep and special thanks to Martin for organizing numerous wonderful parties alongside Teresa that brought our labs closer together. Finally, I am especially grateful to Prof. Jim Demmel and Prof. Aydın Buluç for their course Applications of Parallel Computers. The high-performance computing principles I learned in their class provided the crucial technical foundation that directly inspired and enabled my work on TorchFF.

My research was significantly enriched by external collaborators across various institutions. I would like to thank Prof. Reuben Harris and Prof. Peter I. Dosa for leading the MidWest AVIDD Center. My gratitude goes to Prof. Rommie Amaro, Prof. Donghoon Chung, Prof. Thomas Bannister, Prof. Daniel A. Harki, Prof. Dahai Luo and Prof. Srinivas Chamakuri and their respective research teams for our close collaboration on antiviral drug discovery

projects, specifically Project 5 (SARS-CoV-2 nsp13) and Project 3 (SARS-CoV-2/ZIKV main protease). I also extend my thanks to Prof. Yang Zhang (NYU) and Prof. Heather A. Carlson (Odyssey Therapeutics Inc.) for their valuable perspective on HiQBind. I am also grateful to Prof. Zhongyue Yang and Dr. Qianzhen Shao from Vanderbilt University for their collaboration in exploring the application of EasyBFE in enzyme engineering (though not presented in this dissertation). Additionally, I would like to thank Rohith Srinivaas Mohanakrishnan for his contributions to the development of TorchFF.

Beyond the research, I am deeply thankful to my friends, whose emotional help, care, and presence kept me grounded and supported through the challenges of graduate school. Special thanks to Kunyang Sun and his wife Chang Liu for the weekend hangouts, game nights, and emotional support. I am also grateful to Dr. Yuanqi Du and Yanze Wang for countless online conversations spanning deep research discussions, career advice, and casual talks that provided much-needed comfort and perspective.

Finally, I owe everything to my family. To my beloved parents, Xijun Wang and Xiuyu Yang, thank you for your lifelong devotion, endless sacrifices, and unconditional love. You built the foundation for everything I am today, always believing in me, comforting me, and supporting my dreams every step of the way. I am also deeply grateful to my parents-in-law, Zhanming Yu and Hui Jin, as well as my other family members, for their constant warmth, kindness, and encouragement throughout my life.

Most importantly, I offer my deepest love and gratitude to my wife, Jin Yu. Marrying you in the during my PhD journey is the most precious milestone of my life. You have always been my emotional anchor, keeping our life wonderfully organized, vibrant, and full of color amidst the demanding pace of graduate school. Your patience, warmth, and belief in me have made every challenge manageable and every achievement meaningful. I could not have accomplished any of this without you by my side.

Chapter 1

Introduction

1.1 Overview

The discovery of new drugs is central to advancing public health and human well-being, but the process is extremely lengthy, costly, and risky. This difficulty is reflected in the exceptionally high attrition rate of drug candidates, in which only a small fraction of which ultimately receive regulatory approval [1]. A successful drug molecule must maintain sufficient binding potency for its therapeutic target at low concentrations while satisfying other requirements, including safety and appropriate physicochemical and pharmacokinetic properties. To achieve this goal, researchers commonly use high-throughput screening to identify hit molecules and then perform iterative lead optimization to balance these properties. Because wet-lab experiments, including compound synthesis and subsequent *in vitro* and *in vivo* testing, remain expensive and provide limited information at each step in each design cycle, there has been a long-standing demand for rational computational design methods that effectively leverage mechanistic evidence to prioritize the most informative and promising molecular hypotheses before committing wet-lab resources.

Structure-based drug discovery (SBDD) offers one route toward this objective. By leveraging three-dimensional protein–ligand structures, receptor–ligand interactions can be examined at atomistic resolution, providing a physical basis for understanding how chemical modifications modulate binding [2]. Two core challenges in SBDD are the efficient determination of high-fidelity protein–ligand complex structures and the accurate prediction of binding affinities. Experimental advances in X-ray crystallography, NMR spectroscopy, cryo-electron microscopy, and high-throughput surface plasmon resonance have expanded the structural and binding data available through resources such as the Protein Data Bank (PDB) and BindingDB [3–5]. In parallel, *in silico* methods have matured rapidly along both physics-based and data-driven directions over the past decade.

Physics-based methods, as represented by docking and binding free energy calculations, provide atomistic and physically interpretable models of drug-target interaction with different

levels of computational cost and rigor. Molecular docking can rapidly generate binding poses through conformational sampling and rank them with empirical or physics-inspired scoring functions, making virtual screening of large libraries feasible [2, 6, 7]. Free-energy methods, including end-point MM/PB(GB)SA and free-energy perturbation (FEP) based on alchemical molecular dynamics, provide more physically rigorous estimates of binding affinity but are more computationally intensive [8–10]. These methods are therefore particularly suitable for lead optimization, in which potency can be sensitive to minor chemical modifications of a ligand [10]. Data-driven approaches instead use machine-learning models and neural networks to learn structural and interaction patterns from experimental data. Representative examples include the rapidly growing class of "co-folding" models, such as AlphaFold and Boltz, that directly predict protein–ligand complex structures and can also provide confidence or affinity-related outputs [11, 12]. However, these SBDD methods remain limited in accuracy, efficiency, and applicability by several challenges for which this dissertation provides corresponding solutions.

First, both data-driven approaches and physics-based approaches with empirically fitted parameters heavily depend on the quality of the structural and binding data used to construct, train, and evaluate them. Public resources such as the PDB and BindingDB contain invaluable primary data, but their entries are not uniformly ready for molecular modeling and can propagate structural errors, inconsistent annotations, and dataset biases into downstream models [3, 4, 13, 14]. Manual curation can improve quality but does not scale readily with the continued growth of these resources. Therefore, in Chapter 2, I present HiQBind, an automated workflow that filters, repairs, and refines protein–ligand complexes and associates them with binding-affinity annotations to reproducibly construct a high-quality dataset for model development and evaluation [15].

Second, the accuracy and transferability of physics-based predictors are limited by the potential energy surfaces used to describe biomolecular interactions. Conventional force fields achieve computational efficiency through simplified functional forms and approximations, whereas more physically complete models require suitable reference data, robust parameterization strategies, and efficient software infrastructure [16–18]. Chapters 3–5 address these needs through three complementary contributions: a quantum-mechanical protein–ligand interaction dataset with energies decomposed into physically meaningful components, a high-performance GPU-accelerated differentiable force-field library, and an improved treatment of short-range intramolecular interactions in the AMOEBA polarizable force field [17, 19, 20].

Third, even the most rigorous and accurate binding free-energy calculations have limited practical impact unless they can be prepared, executed, and analyzed with the ease and throughput required for prospective real-world drug discovery. Chapter 6 presents EasyBFE, an open-source package that automates alchemical binding free-energy workflows and applies them to lead optimization and hit prioritization in two real-world drug discovery campaigns [21]. As a complementary free-energy approach, Chapter 7 describes how end-point MM/GBSA can be used together with fragment-based scoring based on prior data to increase

the accuracy of traditional docking in a prospective binding pose-prediction challenge [6, 22, 23].

1.2 High-Quality Protein–Ligand Data Curation

The Protein Data Bank (PDB) and derivative resources that connect experimental structures to functional or biophysical assay annotations—including PDBbind, BioLiP2, Binding MOAD, and BindingDB—provide much of the empirical foundation for contemporary structure-based drug discovery [3–5, 24–26]. PDBbind, in particular, has been used extensively and has supported the development of physics-based/machine-learning molecular docking and scoring functions [27–34]. Widely used benchmarks to evaluate the performance of these models, such as the Comparative Assessment of Scoring Functions series [35–40], were also built on top of this dataset. However, across more than two decades of PDBbind releases and updates, structural defects within individual complexes and organizational overlap between training and test splits have persisted and accumulated as the dataset has expanded [13, 15]. These issues propagate into downstream development and evaluation tasks, result in less accurate or generalizable models, and challenge the validity of the derived benchmarks [14].

The first deficiency is data leakage, which we addressed in Leak-Proof PDBBind (LP-PDBBind) [13]. The conventional division of PDBbind into the general, refined, and core sets focused more on the structural quality and led to highly similar or even identical proteins or ligands across different subsets. Other splitting schemes, including random or chronological splitting, prevented leakage to some extent but never fully removed it by design. Models relying on such splits for training and testing would experience overfitting problems by memorizing prevalent protein sequences or ligand chemotypes rather than actual patterns related to binding [13, 41–43]. In this work, we reorganized PDBbind v2020 into leak-free training, validation, and test sets by jointly controlling protein-sequence, ligand-fingerprint, and three-dimensional interaction similarity [13, 44]. We showed that the reported performance gains of such models were indeed inflated by the data leakage issue because we observed a consistent accuracy decrease after retraining four methodologically distinct scoring functions—AutoDock Vina, RF-Score, IGN, and DeepDTA— with our new splits [27, 31, 32, 45]. Additionally, we prepared three external protein–ligand datasets—BDB2020+, Mpro, and EGFR—with the same leak-proof control to provide the community with more stringent tests that are closer to real-world discovery scenarios. BDB2020+ was constructed from PDB structures deposited after 2020 with high-quality binding free-energy data available in BindingDB. The latter two datasets were curated from the literature with a specific focus on the SARS-CoV-2 main protease and epidermal growth factor receptor, respectively. My specific contribution to this work was to collect the literature and formulate the Mpro dataset.

The second issue is that the structural records themselves are not uniformly reliable. Owing to historical limitations in structure deposition and preparation, two broad classes of artifacts have accumulated in the PDB: (1) incorrect or undocumented ligand bond orders,

tautomeric forms, and protonation states; and (2) missing atoms or residues and steric clashes caused by atoms placed at physically implausible positions because of uncertainties in the experimental density. Although PDBbind corrects some aspects of ligand chemistry, it does not systematically resolve this issue and completely ignores the second class of defects, allowing problematic structures to persist in the "general" set and, in some cases, propagate into the nominally higher-quality "refined" set. Because these artifacts alter formal charges, hydrogen-bonding networks, interfacial contacts, and short-range repulsion, they can compromise both SBDD models trained on these data and stand-alone benchmarks derived from them. A further limitation is that PDBbind relies on a largely manual and nontransparent curation process, which is difficult to sustain as the PDB continues to expand. Moreover, data curated after PDBbind v2020 are no longer freely accessible, while the corresponding preparation workflow has not been publicly documented. These restrictions limit the reproducibility, accessibility, and continuous maintenance of the dataset.

To address these limitations, I developed HiQBind, an automated and open-source workflow for preparing high-quality noncovalent protein–ligand complexes with binding-affinity annotations. By orchestrating open-source tools including RDKit, PDBFixer, and the RCSB PDB API, HiQBind integrates data filtering, ligand and protein fixing, and complex structure refinement within a reproducible pipeline [15, 46, 47]. For each PDB entry, the workflow first excludes systems outside its intended scope or containing severe defects that cannot be reconciled automatically. Such cases include covalent binders, non-drug-like ligands, steric clashes, and incomplete ligand structures. It then corrects ligand bond orders and aromaticity using reference chemical information from the Chemical Component Dictionary, assigns protonation states appropriate for physiological pH, and adds hydrogen atoms. Missing protein residues, heavy atoms, and hydrogens are repaired with PDBFixer when they can be modeled reliably. Finally, the reconstructed complex undergoes constrained force-field minimization, allowing added atoms to relax while preserving the coordinates of experimentally resolved heavy atoms [48–50]. We applied this workflow to prepare both an optimized version of PDBbind v2020 and an external dataset containing 32,275 protein–ligand complexes drawn from 18,160 unique PDB entries and associated with binding-affinity annotation data from BioLiP2 and Binding MOAD [15, 25, 26]. These new, accessible datasets establish a robust benchmark resource for the rigorous training and evaluation of future scoring functions and other SBDD models.

From a longer-term perspective, SBDD is shifting increasingly toward AI-driven approaches that extend beyond static structure and affinity prediction to fully autonomous discovery engines coordinating generative design, computational evaluation, experimental feedback, and iterative decision-making across the entire design–make–test–analyze (DMTA) cycle. In this context, the volume of data is expected to grow rapidly, while the data themselves will become increasingly heterogeneous, noisy, and difficult to curate, which amplifies the significance of high-fidelity data and rigorous validation benchmarks. Therefore, efforts such as LP-PDBBind and HiQBind provide accessible, reproducible, and sustainable frameworks

for strengthening the reliability of the structural foundation that future computational SBDD pipelines will hinge on.

1.3 Advanced Force Field Development

Molecular simulation has been viewed as one of the most powerful tools in SBDD by providing dynamical atomistic details and capturing protein backbone and side-chain reorganization, induced-fit processes, and solvation or ionic effects. It is also the foundation of the "gold-standard" binding-affinity prediction method, i.e., free-energy perturbation, because it provides the correct thermodynamic ensembles and energies. However, such a calculation is only as reliable as the potential energy surface (PES) that drives these simulations. Quantum mechanics (QM) provides the most accurate description by directly solving the first-principles electronic Schrödinger equation that defines this surface, but it is computationally prohibitive for the length and time scales required in routine protein–ligand applications [51]. Machine-learning interatomic potentials (MLIPs), which learn QM energies and forces from data, offer a promising route toward narrowing this accuracy–cost gap [52–54], yet remain too resource-intensive. Furthermore, their accuracy and transferability depend strongly on the reference QM level and the chemical diversity, configurational coverage, and charge- and spin-state coverage of the training data. Extensive benchmarks are consequently needed to validate their generalization to the specific system of interest before application [55, 56]. Molecular-mechanics force fields therefore remain the most common models for biomolecular simulation and binding free-energy calculations due to their efficiency and established ability to reproduce thermodynamic observables.

Conventional force fields, represented by AMBER, CHARMM, OPLS, and OpenFF [16, 57–60], adopt simple functional forms, including harmonic bond and angle terms, periodic torsion terms, Coulomb interactions with fixed atomic charges, and Lennard–Jones potentials to describe bonded and nonbonded interactions. This simplicity enables these FFs to attain the exceptional speed and memory efficiency needed for extensive sampling of large complex systems, but restricts their accuracy and transferability [61, 62]. For example, even modern free-energy protocols commonly retain errors on the order of $1 \sim 2$ kcal/mol with well-parametrized fixed-charge models, and the performance varies substantially across targets and perturbations [63, 64]. A principled path forward is to enrich the FF with physics absent from the fixed-charge, pairwise-additive approximation, including multipolar electrostatics, many-body polarization, charge transfer, and short-range charge penetration [62, 65–67]. These advanced polarizable FFs, however, introduce more strongly coupled components and therefore require more detailed fitting targets, more systematic parameterization procedures, and software that can evaluate both the model and its derivatives efficiently. The three contributions developed in Chapters 3–5 address these requirements accordingly.

At the data level, a central problem is that fitting only total QM interaction energies permits errors in different physical components to cancel, so a model may reproduce an

aggregate value for the wrong reason and fail when the balance of interactions changes. Many QM interaction-energy datasets exist, but none of them satisfies the requirements for developing advanced force fields for protein–ligand interaction modeling: adequate chemical and conformational coverage with high-level quantum-mechanical energies decomposed into physically meaningful components. In Chapter 3, I address this limitation by extracting representative local interaction motifs from the HiQBind databases and computing their QM interaction energies at ωB97X-V/def2-TZVPD with absolutely localized molecular orbital energy decomposition analysis (ALMO-EDA). The decomposition separates permanent electrostatics, Pauli repulsion, dispersion, polarization, and charge transfer, thereby converting each complex from a single fitting target into a physically interpretable profile [19, 68]. The dataset contains $\sim 15,000$ protein–ligand fragment dimers spanning the major noncovalent interaction (NCI) classes encountered at protein–ligand interfaces, while a complementary small dataset was also generated with the same procedure, consisting of ~ 100 full ligand–pocket structures from the OMOL dataset beyond the fragments. The relative proportions of the energy components across different NCI types were analyzed, providing insights into their energetic origins. The data were then used to benchmark three types of models: one fixed-charge model (GAFF2), one polarizable force field (AMOEBA), and four MLIPs (MACE-OFF, MACE-OMOL, UMA, and AIMNet2). The resulting trends show that dispersion-dominated interactions can be represented reasonably well by mature MM models, whereas polarizable force fields systematically improve the description of interactions with stronger electrostatics and polarization. Modern MLIPs can approach QM accuracy when their training data adequately represent the relevant chemistry. Additionally, I envision that this dataset will serve as a useful resource for future EDA-guided force-field development and validation.

Another challenge is the efficiency of implementation and validation. Development of advanced force fields requires exploring the endless space of functional forms to describe different aspects of physics and deriving the best parameter set for any given analytical formula. However, the mathematically complex and nonadditive nature of these potentials makes it an arduous task to manually derive, maintain, and implement the analytical gradients with respect to coordinates or parameters, which impedes the rapid prototyping of the model with dynamics or optimization of the parameters with gradient-based approaches [18, 20]. Automatic differentiation offers a natural solution [20, 69–72], but the speed performance becomes an obstacle, because frameworks such as PyTorch were specially designed and optimized for deep learning, not molecular mechanics. The latter is dominated by algorithms that fail to map as efficiently onto generic matrix operations, which are characteristic of deep learning and are the primary focus of existing differentiable programming packages.

In Chapter 4, I develop TorchFF to bridge this gap by developing backend-optimized C++/CUDA operators and registering them in PyTorch, thereby preserving the automatic-differentiation computation graph and a composable high-level interface. TorchFF provides reusable primitives that are commonly used in molecular mechanics, including a variety of bonded and nonbonded potentials, long-range multipolar Ewald and particle-mesh Ewald

electrostatics, and neighbor-list construction, so that new physical models can be assembled from these optimized pieces without reimplementing them from scratch. Numerical validation confirms the correctness of this implementation by reproducing absolute values of energies and forces, as well as MD-based properties including radial distribution functions (RDFs) and densities. Benchmarking shows a consistent and significant performance advantage for the optimized kernels over the native PyTorch implementation, particularly for larger or more complex workloads, for which TorchFF can be two orders of magnitude faster. I also apply TorchFF to implement the Completely Multipolar Model (CMM), an exceptionally sophisticated many-body potential that features ubiquitous coupling between bonded and nonbonded terms [73]. TorchFF enables condensed-phase simulations with CMM to be performed efficiently enough to validate it against thermodynamic properties, which cannot be achieved with native PyTorch. This demonstrates that TorchFF serves as infrastructure that turns differentiability from a convenient prototyping capability into a practical foundation for advanced force-field development.

My third contribution to force-field development is to improve the functional forms in the existing AMOEBA polarizable model, in which a particularly persistent ambiguity arises for short-range intramolecular interactions, especially between atom pairs separated by three covalent bonds. In this regime, electron-density overlap, partial covalency, charge penetration, and exchange repulsion blur the nominal boundary between bonded and nonbonded physics. Conventional treatment involves scaling the nonbonded interactions and fitting torsional parameters to compensate for the residual error [74–77]. This convention couples torsional parameters to an arbitrary nonbonded scaling rule and does not guarantee an accurate potential energy surface. The problem is more pronounced in AMOEBA, whose permanent multipoles, direct induction, polarization energy, and van der Waals interactions follow distinct short-range scaling conventions, complicating its implementation and introducing additional computational overhead [78, 79].

In Chapter 5, I replace these component-specific conventions with a unified partition: nonbonded interactions are excluded through the 1–4 level and included from the 1–5 level onward, while the excluded short-range intramolecular region is described by an improved valence potential with Morse bonds, cosine angles, periodic torsions, and their coupling terms [80, 81]. After refitting the permanent atomic dipoles against QM electrostatic potentials to compensate for the change in direct field scaling that produces induced dipoles, the newly introduced valence potential is then parameterized with Q-Force, which provides an automated pipeline to generate parameters by fitting against reference QM energies, forces, and Hessians. I applied this force field construction to three model dipeptides, showing that it improves agreement with QM potential energy surfaces, especially the forces, as validated by a 2D ϕ/ψ relaxed scan and an independent set of 600 conformers. The new model also supports stable condensed-phase dynamics and preserves reasonable solution-phase conformational behavior. These results indicate that a richer bonded representation can replace ad hoc short-range scaling without sacrificing the advantages of a polarizable multipolar nonbonded

model, thereby providing a clearer and more systematic route to future development of models with more advanced nonbonded treatment beyond AMOEBA functional forms.

From a broader perspective, as more physical effects are explicitly accounted for by modern force fields, their development has become an integrated scientific and engineering problem. On the scientific side, progress requires determining which physical effects should be represented and through what functional forms, designing fitting targets that meaningfully constrain those terms, and developing parameterization and validation strategies that produce accurate and transferable models. On the engineering side, these components must be integrated within a reproducible, extensible, and automated framework, so that model development can proceed efficiently and remain traceable and revisable as new data, physical insights, and functional forms emerge. This integration will become increasingly important as future force fields combine ML with MM, for example, by using neural networks for parameter assignment or range-separated architectures in which MM and ML, respectively, describe transferable long-range and complex short-range physics. Collectively, through reference data generation, infrastructure for efficient and continuous development and validation, and improved functional-form design, the efforts described in Chapters 3–5 support this perspective by providing scientific insights and practical tools for future force-field development in this direction.

1.4 Binding Free Energy Workflows and Applications

A central goal of structure-based drug discovery is the accurate prediction of protein–ligand binding affinity. Empirical or machine-learned scoring functions (SFs) that are widely used in molecular docking provide very fast estimates, making them useful in large-scale screening for hit identification. However, their simplified treatment of molecular interactions and limited conformational sampling restrict their accuracy, making them less effective in lead optimization, which requires the model to capture affinity changes caused by minor chemical modifications [2, 6, 7]. Instead, obtaining binding free energies from physics-based simulations and statistical estimators provides a more numerically accurate and theoretically rigorous approach [63, 64, 82]. The most representative method is free-energy perturbation (FEP), in which an alchemical pathway is used to gradually transform one thermodynamic state into another. The free energy is then estimated from samples along the alchemical trajectory using approaches including thermodynamic integration (TI), the Bennett acceptance ratio (BAR), or its multistate variant, MBAR [9].

However, the practical application of alchemical binding free-energy calculations remains constrained not only by their computational cost but also by the complexity of the workflow. Reliable calculations require coordinated steps spanning system preparation and parameterization, definition of the alchemical pathway, construction of thermodynamic cycles and simulation protocols, simulation execution, and statistical analysis with appropriate diagnostics [9, 83]. Each stage involves numerous technical or scientific decisions and often

requires substantial manual intervention to translate inputs and outputs across scattered software tools. Consequently, these calculations remain largely manageable only by domain experts and are difficult to scale to the throughput required for prospective drug-discovery campaigns.

To reduce this technical barrier, Chapter 6 presents EasyBFE, an open-source package for alchemical protein–ligand binding free-energy workflows [21]. By using AMBER as the MD engine [84] and integrating open-source tools, including components from the alchemlyb [85] and OpenFE [86] communities, EasyBFE automates the end-to-end workflow for both relative and absolute binding free-energy calculations by implementing and connecting the required components described above. The software is not merely a collection of wrapper scripts, but is carefully designed to support users with different levels of domain expertise. Entry-level users can launch routine calculations with predefined settings using a few Python or Linux shell commands, while more experienced users can tailor the calculation by adjusting choices in a configuration file, such as the ligand force field, atom-mapping method, alchemical schedule, and simulation time to achieve optimal performance for specific projects. Developers can also benefit from its modular architecture when introducing new algorithms for ligand parameterization, atom mapping, or Boresch-restraint selection [87]. Because EasyBFE currently relies on AMBER for production simulations, it can readily benefit from continued improvements to the engine, including new enhanced-sampling methods or better GPU support. In addition to its core functionality, EasyBFE provides supporting utilities such as protein-structure preparation and constrained docking against a reference, both of which are important for preparing reliable input structures. EasyBFE was validated with established benchmarks by reaching accuracy comparable to that of other open-source MD engines or industry counterparts.

I demonstrate the practical significance of this tool in two prospective studies of real-world drug-discovery campaigns that were at the lead-optimization and hit-identification stages, respectively. First, in the SARS-CoV-2 main-protease project in which I participated as part of the Midwest Antiviral Drug Discovery (AViDD) Center [88], I deployed the RBFE workflow to optimize a hit molecule series that was previously identified by DEL [89]. I first conducted a retrospective study to confirm that EasyBFE was reliable for this specific protein and chemical series with an RMSE below 1 kcal/mol compared with experimental readouts. Then, I proposed several chemical modifications for the next synthesis round that were predicted by RBFE to enhance binding. Experimental assays then confirmed the predicted improvement, and the optimized compound showed an approximately fivefold lower IC_{50} than its parent.

The other program is CACHE Challenge #7, which asked the participants to find orthosteric inhibitors of phosphoglycerate kinase 2 (PGK2) that were selective over the closely related PGK1 enzyme [90]. In this project, I first applied an LLM-based generation workflow to obtain a pool of promising candidates: SmileyLlama [91] to produce diverse *de novo* candidates and SynLlama [92] to project the promising ones into synthesizable analogs from the Enamine REAL space. After filters were applied to screen for solubility, drug-likeness,

promiscuity, and stability of the PGK2-selective interactions, I applied the ABFE workflow as the final selection stage because the remaining molecules were too chemically diverse to enable reliable RBFE network construction. Remarkably, EasyBFE's automation made it possible to complete calculations for 1,200 compounds on NERSC within six weeks. The top 200 compounds were submitted for experimental validation, of which a substantial fraction were predicted to bind comparably to or more favorably than a known binder. The results indicate that the combined generation and free-energy workflow can produce a promising prospective set in real-world drug discovery campaigns with little prior knowledge.

Chapter 7 examines the use of a complementary end-state binding free-energy estimation method with lower cost and less physical rigor in a competition hosted by the ASAP–Polaris–OpenADMET community to predict binding poses of unreleased SARS-CoV-2 and MERS-CoV main protease ligands [22]. We developed a workflow to rerank candidate poses generated by AutoDock Vina [6] using a fragment-based score that encapsulates prior knowledge from the provided training data, with MM/GBSA [8, 93] as a fallback in lower-information regimes. This strategy ranked 6th overall and 2nd among non-deep-learning methods, demonstrating that traditional docking remains competitive with modern ML approaches if it can be improved with prior knowledge or more rigorous physics.

Collectively, Chapters 6 and 7 show that binding free-energy methods, with different levels of physical rigor and computational cost, can generate practical impact by prioritizing molecular hypotheses in the drug discovery design–make–test–analyze (DMTA) cycle if they are implemented in a user-friendly manner, executed automatically, and applied strategically.

Chapter 2

A Workflow to Create a High-Quality Protein–Ligand Binding Dataset for Training, Validation, and Prediction Tasks

‡Yingze Wang and Kunyang Sun contributed equally. Yingze Wang designed and implemented the HiQBind workflow, including data cleaning, structure fixing, and ligand fixing, and analyzed the ligand-fixing results.

2.1 Introduction

Scoring functions (SFs) are crucial in computer aided drug discovery, utilized for selecting the most probable ligand geometry and its binding pose with a protein that best correlates or predicts their free energy of binding.[36] There are a plethora of SFs being developed and widely used by computational and medicinal chemists, and they can be broadly categorized into either classical scoring functions[27–30, 49, 94–99] or machine learning scoring functions.[31–34, 45, 100–102] The majority of protein-ligand SF predictors, whether physical or machine-learned, have been trained on the PDBbind dataset[24, 35, 103–107] (http://www.pdbbind-cn.org/), specifically v2020, a curated set of ∼19,500 biomolecular complex structures and their experimentally measured binding affinities. PDBbind is further organized into a "general" data subset that is often adopted by SFs for training, and separate "refined" and "core" datasets which contain protein-ligand complexes with the best structural quality and most reliable

[1]Reproduced with permission from: A workflow to create a high-quality protein–ligand binding dataset for training, validation, and prediction tasks. Y. Wang‡, K. Sun‡, J. Li, X. Guan, O. Zhang, D. Bagni, Y. Zhang, H. A. Carlson, and T. Head-Gordon. *Digital Discovery* 2025, 4, 1209–1220.

binding affinity data that is used for testing. Various benchmarks based on PDBbind, such as CASF (Comparative Assessment of Scoring Functions) series[37–40], CSAR (Community Structure Activity Resource) 2010[108] and PDBbind-blind-2013[109] have been proposed to assess the scoring power, ranking power, docking power and screening power of various SFs.

PDBbind has been an invaluable resource to the biomolecular community during its two-decade development, but a significant portion of the PDBbind dataset contains structural errors, statistical anomalies, and a sub-optimal organization of protein-ligand classes that can limit SF training and validation[13, 110]. These inconsistencies undermines the purpose of the refined set, which is intended to serve as a high-quality benchmark for evaluation of scoring functions and docking methods. Another concern in regards PDBbind is that the data processing procedure is neither open-sourced nor automated, potentially relying on individual groups needing to introduce their own manual intervention that may lead to inconsistencies. Furthermore, the PDBbind data curation process became more problematic in 2021 when PDBbind ceased to be freely available for data curated after 2020, which limits access and hinders the development and validation of new scoring functions (and other additional uses).

Fortunately, other curation efforts have created alternative protein-ligand structural and/or binding datasets that have increased the size and comprehensiveness of available data for drug discovery efforts. BindingDB is a database containing 2.9 million binding measurements spanning 1.3 million compounds for thousands of protein targets, which are curated from the literature and patents.[4, 5, 111] Binding MOAD is a curated database of 41,409 protein–ligand structural complexes, with binding affinity data available for 15,223 (37%) of them; Binding MOAD's curation involved extracting high-quality structures from the PDB and finding associated binding data from publications with the aid of an NLP-based annotation tool.[26, 112–114] BioLiP is a large database of over 900,000 biologically-relevant protein-ligand interactions curated from the PDB, and enriched with various functional annotations, including Enzyme Commission numbers, Gene Ontology terms, catalytic sites, and binding affinities from Binding MOAD, BindingDB, as well as manual surveys.[25, 115] Other related datasets that focuses more on the geometries of proteins and ligands, including PLINDER[116] and DockGen[117], contain an expanded set of protein-ligand structural complexes but do not have annotations of binding affinity data. However, in general, these curation efforts have largely focused on increasing the size and comprehensiveness of protein-ligand data, rather than increasing the quality and reliability of the data themselves. Therefore, there is a pressing need for an open-source and systematic workflow to prepare protein-ligand binding datasets with well-defined binding affinity annotations and higher-quality structures in order to foster greater reproducibility, transparency, and accessibility.

In this work, we introduce HiQBind-WF, a workflow of algorithms for data cleaning and structural preparation that creates a curated dataset of high-quality, non-covalent protein-ligand complex structures with binding affinity annotations. This workflow contains several modules: (1) a curating procedure that rejects ligands covalently bonded to proteins, ligands with rarely-occurring elements, and structures containing severe steric clashes; (2)

a ligand-fixing module to ensure the correctness of the ligand structure including correct bond order and reasonable protonation states; (3) a protein-fixing module to extract and, when necessary, add missing atoms to all chains involved in the protein-ligand binding; (4) a structure refinement module to simultaneously add hydrogens to both proteins and ligands in their complex state, as opposed to the current practice in PDBbind that completes the hydrogen chemistry for protein and ligand independently. The motivation for adding this hydrogen growth module is that although many SFs only take heavy atoms into consideration, future physics-based SFs could potentially benefit from explicit hydrogens to better model intermolecular interactions such as hydrogen bonding.

We utilized this workflow to optimize PDBbind v2020 and compared the processed structures. Analysis of the structural differences between the same PDB entry demonstrated that HiQBind-WF is able to correct for various observed structural imperfections. Further, to illustrate the applicability of the HiQBind-WF, we created HiQBind, a new dataset with high-quality protein-ligand binding structures and affinities by processing PDB entries included in BioLiP2 and Binding MOAD associated with binding affinities drawn from BindingDB. The HiQBind dataset includes >18,000 unique PDB entries and >30,000 protein-ligand complex structures. We also confirmed that HiQBind shares similar properties with existing datasets like PDBbind, demonstrating its feasibility to be used for developing and validating SFs and other structure-based drug-design tools. The HiQBind-WF and HiQBind dataset are provided open-source to foster transparency and sustainability as new data appears, in order to maintain this important resource for the biology and drug discovery communities.

2.2 Methods

The flowchart of HiQBind-WF is illustrated in Figure 2.1. We start by downloading the pdb and mmcif formats directly from the RCSB PDB[3] for supplied entries. The pdb files are used for structure preparation and the headers in mmcif files are used to extract useful metadata, such as resolution, deposit date and sequence information. For each PDB entry, we split the structure into three components: ligand, protein and additives, and curate these categories as follows.

We define three classes of ligand(s) for any given protein-ligand complex structure: (1) Any residue will be identified as a ligand if its name matches the Chemical Component Dictionary (CCD) code deposited in given reference datasets (PDBbind, BioLiP or Binding MOAD). Ligands identified in this manner are referred to as ”small molecules”. (2) Otherwise, chains in the original PDB file that are less than 20 residues but more than one residue as ligands will be selected as ligands. For example, PDBbind entries that contain patterns such as “*-mer,” or symbols like ”-”, ”&” or ”+”, MOAD entries with more than one CCD in the `name` column of its csv-formatted dataset, and BioLiP entries with `Ligand CCD` column to be ”peptide”, ”dna” or ”rna”. These ligands are typically polypeptides, oligosaccharides, or oligonucleotides, collectively referred to as ”polymers”. (3) For each identified ligand, we

label any biopolymer chains within 10 Å as the associated protein structure. Then, for each protein structure, we labeled residues specified by the ”HETATM” record in the pdb file within 4 Å as additives, which includes ions, solvents, and co-factors. The additives are saved in pdb format and directly deposited in the database, and the protein and ligand structure are ready to proceed to the next workflow steps.

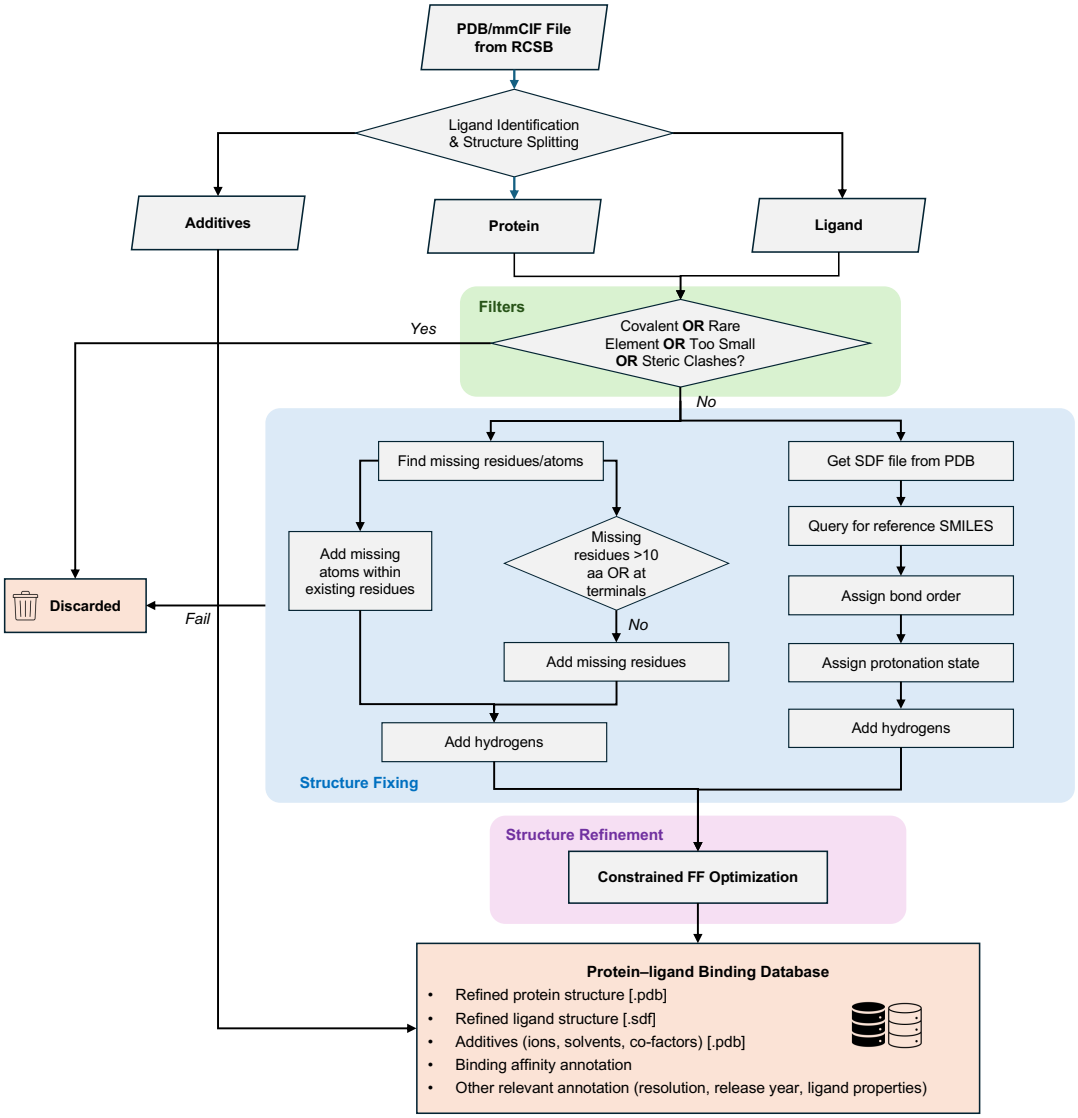

Figure 2.1: *Schematic representation of the semi-automated HiQBind-WF to refine protein-ligand binding structures.* HiQBind-WF downloads the pdb and mmCIF files from the RCSB PDB[3], followed by splitting each structure into three components —ligand, protein, and additives. A series of filters are then applied to remove covalent binders, ligands with rare elements, very small ligands, and complexes exhibiting steric clashes. Subsequently, the protein structure is fixed by adding missing atoms and residues (ProteinFixer) and the ligand structure is fixed by correcting bond orders, protonation states, and aromaticity (LigandFixer). Finally, the fixed protein and ligand structures are recombined and subjected to a constrained energy minimization to resolve potential unreasonable structures and to refine hydrogen positions.

After the splitting for the ligand categories and their associated protein-ligand complexes, we define an additional set of downselect filters, including some borrowed from the processing protocols of LP-PDBbind[13]. The purpose of these filters are to exclude protein-ligand complex structures that specifically can interfere with training of SFs, with eliminations that would meet any of the following criteria:

- Covalent binder filter: Excludes ligands covalently bound to the protein, as indicated by the "CONECT" record in the pdb file. When covalently-bound ligands are identified, they are eliminated. This is because covalent binding inherently is different from non-covalent binding which does not involve bond breaking, and thus requires special treatment in any SFs. Those covalent binder entries are included in the SI which may be helpful as a separate curation of the data or may be accessible from CovBinderInPDB[118].

- Rare element filter: Excludes ligands containing elements other than H, C, N, O, F, P, S, Cl, Br, I. For example, Te or Se are infrequently encountered, and their inclusion can make it challenging for SFs to learn key binding features giving data sparsity for these ligands. These ligand entries are also included in the SI which may be helpful as a separate curation of the data.

- Small ligand filter: Excludes ligands containing less than 4 heavy atoms, which includes small inorganic binders like $O_2, NH_3, CO_2, NO_2, N_3^-$, which are beyond the scope of common protein-ligand binding studies.

- Steric clashes filter: Excludes structures with protein-ligand heavy atom pairs closer than 2 angstroms. Such steric clashes often arise from electron density uncertainties or inaccurate structural reconstruction from electron densities and are not physically feasible non-covalent interactions. Including such structures in SF development could be detrimental, for example leading to an underestimation of the repulsion energy in physics-based SFs. Additionally, the steric clash filter helps to exclude covalent ligands if the covalent bond is not properly represented in the "CONECT" record.

For protein-ligand complexes that pass these filters, two structure-fixing modules are implemented separately for proteins and ligands. In the ProteinFixer module, we first use the sequence information from the mmcif file header to detect missing atoms and residues. Then, for missing residues or missing atoms within an existing residue, PDBFixer[46] (version 1.9) is used to add them, except when the missing residues are longer than 10 amino acids or are located at the sequence terminals. Adding missing atoms to protein structure is essential near binding sites because incomplete structures can compromise accurate modeling of binding interactions, and any molecular dynamics or alchemical binding free energy calculation also require complete structures to ensure the correct structural ensemble are sampled during simulations. However, long missing segments or missing terminus residues in crystal structure

are often attributable to intrinsically disordered regions (IDR)[119], domains that are not expressed in the samples for crystallization, or his-tags introduced in the protein purification process[120]. If far enough removed from the ligand binding site(s), we regard it safe to skip modeling these residues explicitly. Hence, we leave these regions in their original form and in themselves do not define a criterion for being discarded in the final dataset. The final step of the protein-fixing module is to add hydrogen atoms at pH=7.4 with PDBFixer. At this pH, the protonation state assignment of titratable side-chains obeys the following rules: all lysine (LYS) and arginine (ARG) are positively charged and glutamic acid (GLU) and aspartic acid (ASP) are deprotonated. Histidine remains in a neutral form and whether the HID or HIE variant (the hydrogen is added to Nδ or $N\epsilon$, respectively) is selected will be based on which one forms a better hydrogen bond, which is the default behavior of PDBFixer.

In the LigandFixer module, we first obtain an sdf file for each ligand instance either by downloading from the RCSB PDB (if possible) or converting from the native pdb format with OpenBabel[121]. Since explicit atom connections may not be present in the pdb format, the bond orders in this converted sdf file are typically inferred from local atomic geometries and the resulting structure is herein referred to as "inferred structure". Then, a reference SMILES is obtained, which is used to correct bond orders and aromaticity specifications that could sometimes be mislabeled in the inferred structure. The bond order assignment protocol is implemented as follows: if the inferred and reference structure are isomorphic, a one-to-one atom mapping will be generated by structure matching and then bond orders, atom hybridization and aromatic specifications will be assigned according to the reference. Otherwise, the bond order assignment will come to a failure point, which means that the inferred structure does not share the same number of atoms or bond connectivity as the reference, indicating that there are missing atoms or distorted geometries in the crystal structure. Therefore, such structures will be excluded.

After a correct structure is obtained, protonation states are assigned to the ligand. We acknowledge that it is a non-trivial task to correctly determine protonation states for titratable groups within a ligand at a given pH and many algorithms that use empirical rules, QM/MM calculations or machine learning have been reported[122–124]. However, since our workflow is designed for high-throughput processing, we improve the efficiency using a simple set of predefined rules to determine the protonation states by relevant matching functional groups in SMARTS patterns. Acids, nitro groups, thiophenols, azides, and N-oxides are deprotonated. Aliphatic amines and guanidines/imines are protonated, while anilines are not protonated. There are other special considerations that should also be accounted for: amines will not be protonated if the nitrogen is directly bonded with atoms other than H or C. Only one nitrogen atom on diamines and piperaizines will be protonated to avoid two positive charged groups close to each other, which is not favorable at normal biologically-relevant pH. Enols with the motif O=C-C=C-OH are deprotonated. The protonation state assignment is implemented by modifying the default behavior of the `dimophite_dl` package[125] which can be found in the Github repository.

One thing that should be noted here is the source of the reference SMILES string. If the ligand is a small molecule with a CCD code or is a polymer with a BIRD (Biologically Interesting Molecule Reference Dictionary[3]) code, we will query RCSB PDB for its reference SMILES. If the ligand is a polymer consisting only of alpha-amino acids, we will assume it is a simple non-cyclic peptide and generate a SMILES string based on its sequence information and amide-bond formation rules. Apparently, for the latter case, any mismatch between the inferred and reference structure does not mean the inferred structure is wrong - the ligand may just be a cyclic peptide or contain disulfide bonds. However, such structures will also be excluded as we are unable to verify its correctness automatically at this stage. For such cases, human inspection will be inevitable and it's beyond the scope of the workflow. In addition, we found that some of the SMILES strings deposited in RCSB PDB are incorrect such that all the bonds are labeled as single bonds. Most of these errors were caught by a geometric check for $sp^3/sp^2/sp$ carbons. For these cases, we manually corrected the SMILES according to the original literature and use the corrected one to do the ligand fixing. The list of manually corrected SMILES can be found in the public Github repository. The bond-order assignment, protonation state assignment, and added hydrogens in the ligand-fixing module are all performed with RDKit[47] (version 2024.03.4).

The last part of the HiQBind-WF structure preparation is a refinement module in which the fixed ligand structure and protein structure are combined, followed by a constrained energy minimization with a well-established force field. AMBER14SB[48] is used for the protein and OpenFF-2.1.0[49] together with Gasteiger charges[126] are used for the ligand. Coordinate constraints are applied to all atoms that are experimentally resolved, which means only positions of hydrogens (both on the ligand and protein) and atoms added by PDBFixer in the protein-fixing module are allowed to be optimized. We found this physically-based structural optimization is useful to resolve any remaining steric clashes between added atoms introduced by treating the protein and ligand structure separately in the previous structure fixing modules. Additionally, the hydrogen-bonding network between the ligand and protein is also optimized in this process. The constrained energy minimization was performed with OpenMM 8.1.1[50] by setting masses of all constrained atoms to zero.

The binding affinity in terms of ΔG is directly related to the dissociation coefficient K_d or K_i through the standard relationship $\Delta G = RT \ln(K_{d/i})$.[127] However, a large portion of the data in the binding datasets is reported in terms of IC_{50}, which cannot be easily translated to ΔGs due to its dependence on other experimental conditions and inhibition mechanisms.[128] Furthermore, the IC_{50} values for the same protein-ligand complex can vary up to over order of magnitude in different assays[129], and some deposited binding data are not reported as exact values but just ranges. Therefore, the binding affinity data is reorganized into a machine-readable format (csv) with comments as to the form of the experimental binding free energy data: K_d, K_i, IC_{50} and EC_{50}.

2.3 Results

Figure 2.2 illustrates the common problems that arise in the training and tests sets for protein-ligand interactions and associated binding assays when developing a scoring function using various curated databases. Some of the structural imperfections are inherited from the original RCSB Protein Data Bank (PDB)[3] dataset, such as missing hydrogen atoms and/or incomplete residues due to uncertainties in the modeled electron densities, whereas some errors originate from the preparation of ligand structures that results in incorrect bond order, protonation state, tautomer state and aromaticity specifications. Some entries are covalent binders such as shown in Figure 2.2a, which requires special methods to account for the covalent bond formulation[130, 131], and should remain distinct from protein-ligand complexes that are formed from non-covalent interactions only. Figure 2.2b illustrates an example from the PDBbind refined set, 5OUH[132], which is a noncovalent binder that exhibits a severe atomic clash with the protein. Similarly, for 3KMC[133] the chlorobenzene portion of the ligand is absent from the crystal structure as seen in Figure 2.2c. This again highlights the need for the community to have a free open-source tool to curate high-quality protein-ligand structures in a reproducible way.

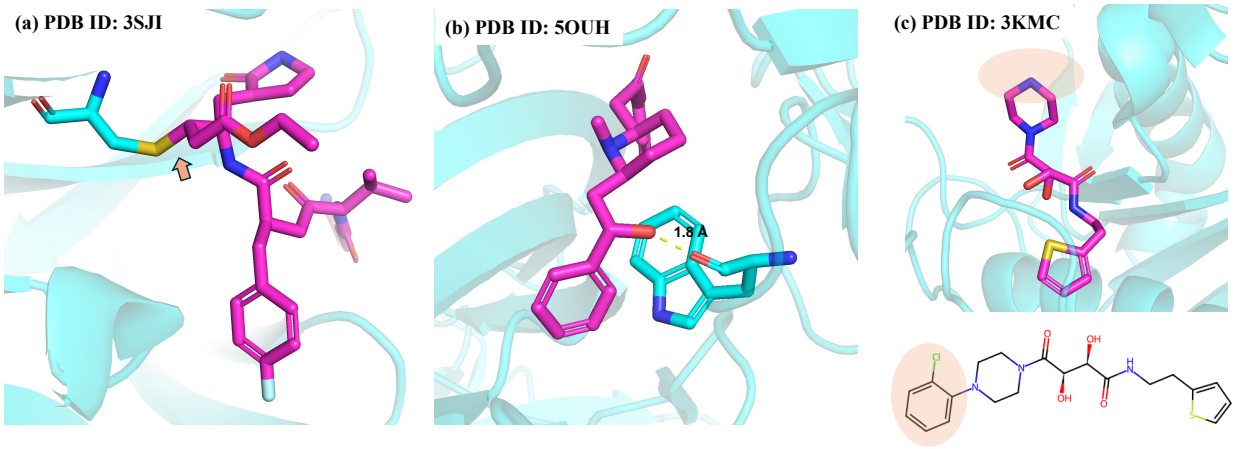

Figure 2.2: *Common structural imperfections in PDBbind dataset.* (a) Covalent binders. The ligand is covalently bound to cysteine with a Michael addition reaction. (b) Steric clashes with the distance between the clashing atoms being only 1.8 Å. (c) Missing atoms.

2.3.1 The HiQBind Workflow

We applied HiQBind-WF to refine the structures in the publicly available PDBbind v2020 dataset[35]. While we are not able to publish this optimized PDBbind dataset because the user's agreement of PDBbind prohibits any distribution of any derivative dataset, we can report some general statistics of the workflow and provide examples of structural fixes of the protein and ligand data. However, users can reproduce an optimized PDBbind data set using HiQBind-WF following step-by-step instructions in our Github repository.

Of the original 19,443 unique PDB entries for proteins with ligands in PDBbind v2020, 1,330 entries were discarded by the filters and 2,452 entries were discarded because they were unable to pass the structure fixing and refinement modules. Our final optimized PDBbind dataset contains 15,661 unique PDB IDs and 27,757 protein-ligand complexes structures. In addition, considering that the original PDBbind general set was further filtered to create a "refined" and "core" set based on structure quality, binding data quality, and redundancy reduction[35], the optimized PDBbind data yields totals of 4,969 and 279 entries in the refined and core sets, respectively. Finally, the associated binding affinity data is reorganized into a machine-readable format (csv) with comments as to the form of the experimental binding free energy data: K_d, K_i, IC_{50} and EC_{50}.

Example of refined ligand structures using HiQBind-WF. Here, we also provide examples of the fixed ligand structures obtained from HiQBind-WF and compared with the deposited ligands in the original PDBbind dataset as provided in Figure 2.3. In some cases, we find that some of the ligands in PDBbind are different from what was actually reported in the literature from which they were derived. For 2AXI[134], the ligand of interest should be the cycflic peptide-like inhibitor, not the sulfonic acid buffer. In other cases, the PDBbind ligand structures are incomplete or the bonding is incorrect (Figure 2.3a). For example, the ligand in 1ALW[135] is missing an iodine atom and in 1DY4[136] the isopropyl is falsely reported as a cyclopropyl (Figure 2.3b,c). This type of problem may arise for historical reasons, i.e. some structures in PDBbind were derived from older version of RCSB PDB that contained these incorrect structures. We also find that HiQBind-WF yields ligands with better protonation/tautomer states. Two examples are 1DG9[137] and 5ETT[138], for which PDBbind shows that the former case contains a neutral sulfonic acid and a divalent piperazine cation motif while the latter case falsely makes a guanosine-like compound positively charged (Figure 2.3d,e). In this case, the fixed ligand structures are more chemically feasible and also in line with the protonation states predicted by ChemAxon Marvin[124].

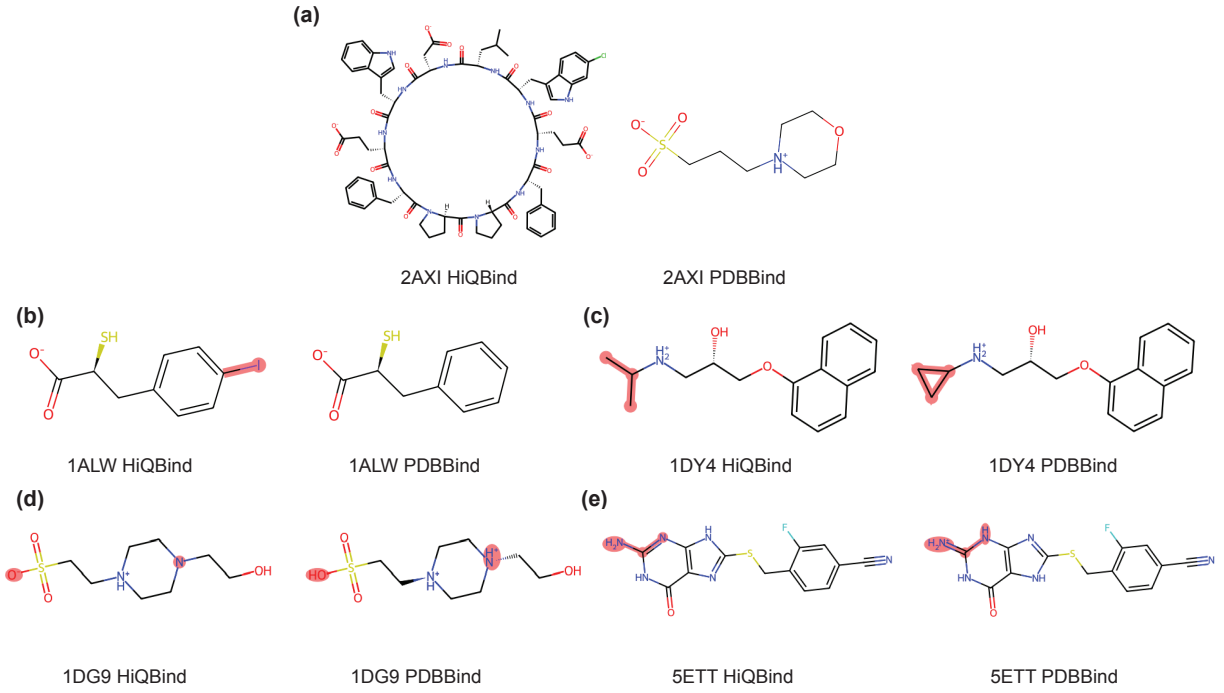

Figure 2.3: *Examples of corrected ligands by HiQBind-WF compared to the original PDBbind.* (a) Wrong ligand entity reported. (b) Missing atoms. (c) Wrong bond connectivity. (d-e) Undesired protonation and tautomer states. The mol2 format ligand files in PDBbind database were used for analysis.

HiQBind-WF also fixes a small but practical problem in PDBbind. PDBbind provides two file formats for the ligand structure, mol2 and sdf. However, among all 19,443 entries, 45 mol2 files and 3175 sdf files cannot be processed by RDKit[47] (version 2024.03.4), a widely-used open-source cheminformatics tool. This may be due to the fact that these files are prepared by some other software and their sdf specifications are not compatible with RDKit. Examples are undesired aromaticity specification (oxygens tagged as aromatic to represent equivalent atoms in $RSO_3^-, RCOO^-, RPO_3^{2-}$) or formal charge specification (nitrogen with 4 explicit valence tagged to be neutral). HiQBind-WF naturally addresses this technical problem because it uses RDKit[47] to process ligand structures.

a)

```
REMARK 465 FIXED MISSING RESIDUES
REMARK 465 (M=MODEL NUMBER; RES=RESIDUE NAME; C=CHAIN
REMARK 465 IDENTIFIER; SSSEQ=SEQUENCE NUMBER; I=INSERTION CODE.)
REMARK 465
REMARK 465   M RES C SSSEQI
REMARK 465     GLY H    98
REMARK 465     ARG H    99
REMARK 465     SER H   100
REMARK 465     ASN H   100A
REMARK 470 FIXED MISSING ATOM
REMARK 470 (M=MODEL NUMBER; RES=RESIDUE NAME; C=CHAIN
REMARK 470 IDENTIFIER; SSSEQ=SEQUENCE NUMBER; I=INSERTION CODE.)
REMARK 470
REMARK 470   M RES CSSEQI  ATOMS
REMARK 470     TYR H   97   CG   CD1  CE1  CZ   OH   CE2  CD2
REMARK 470     VAL H  100B  CG1  CG2
```

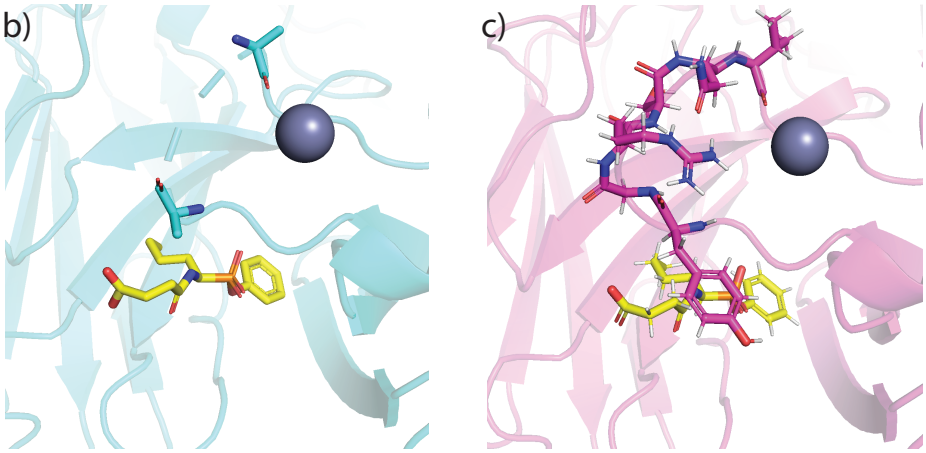

Figure 2.4: *Example of a refined protein structure derived from HiQBind-WF compared to the original PDBbind.* (a) The protein fixing metadata related to residue Y97 through V100B in the refined protein file. (b) Visualization of the original PDB entry with missing residues and atoms centralized at the region Y97-V100B close to the binding site. (c) Visualization of the refined protein structure after the protein-fixing module.

Example of refined protein structures using HiQBind-WF. With the protein-fixing module, users interested in training 3D-based SFs and capturing local protein-ligand interactions would benefit from a more complete protein and binding site representation. To demonstrate our protein-fixing module, Figure 2.4 shows an example of protein 1A0Q[139] that has both missing atoms and missing residues around the binding site. Here, the protein-fixing module first identified those missing data and fixed them based on the sequence information provided in the mmcif file header and the predefined residue templates. The reason behind using information from the mmcif header rather than the pdb "SEQRES" field is that in

some of the deposited structures, missing residues are also omitted in the "SEQRES" field. As a result, an unphysical peptide bond will be placed between the start and the end of a short sequence of the missing residues, which will cause problems in training SFs. The metadata from this fixing call is stored in the refined pdb file in case users want to label the original crystal residues and repair residues differently.

2.3.2 Creation of the HiQBind Dataset

In order to further demonstrate the utility of HiQBind-WF, we have created a new dataset of high-quality, non-covalent protein-ligand complex structures and their associated binding affinity values. To prepare the HiQBind dataset, we used two biologically relevant protein-ligand datasets as a starting point: BioLiP2[25] and Binding MOAD[26]. We downloaded the txt-formatted BioLiP database from its official website and csv-formatted dataset MOAD from its Github repository and entries with at least one reported binding affinity (K_i, K_d, IC_{50} or EC_{50}) data were selected. BioLiP2 itself provides a sizable collection of protein-ligand entries deposited in RCSB PDB and enriched with multiple annotations, including binding affinity data from various sources including Binding MOAD[26], BindingDB[4, 5], and manual annotation. Although BioLiP2 encompasses much of Binding MOAD, we still found additional entries from Binding MOAD that we also include in our new dataset. Over this entire merged set seven PDB entries (2BXG, 2I30, 3T74, 3T8G, 4H57, 6TMN and 7JWN) were discarded because their binding affinities are invalid (with $K_i > 10^3$ M), and all entries as part of the publicly available PDBbind v2020 dataset are not included. In total, 20,349 unique PDB entries with reported binding affinities were obtained.

We then applied HiQBind-WF to process all these starting entries, yielding 18,160 unique PDB entries. For the 2,189 entries that failed to pass HiQBind-WF, 761 of them are discarded by the filters and the remaining 1,428 entries are due to the failure of structure fixing and refinement modules (Table S4). A large portion of the discarded datapoints are "polymers" for which it is hard to verify their structural correctness because of the difficulty in obtaining a reliable reference SMILES string which is more suitable to small molecules. Almost certainly, human inspection and expertise will rescue some of the discarded data, but the design goal here is to automate the corrections with a high throughput procedure as much as possible.

At the same time we retain 32,275 protein-ligand complex structures. The reasons behind the increase in the amount of data compared to PDBbind is that we have included multiple protein-ligand complexes from the same RCSB PDB entry because: (1) multiple conformers or stereoisomers can contribute to the binding (Figure 2.5a); (2) the same ligand can bind to different protein pockets (Figure 2.5b); (3) there is more than one ligand with a reported binding affinity (Figure 2.5c); 131 PDB entries are of this reason; and (4) for structures containing homo-multimers, structural fluctuations between chains that share sequence identity are non-negligible.

A moderate amount of PDB entries contain multiple records of the same ligand of interest

in the deposited structures. The reason lies in the fact that proteins can form various quaternary structures using copies of the same chain, and ligands as binders can interact with the macromolecule at the tertiary or quaternary level. For example, when a ligand

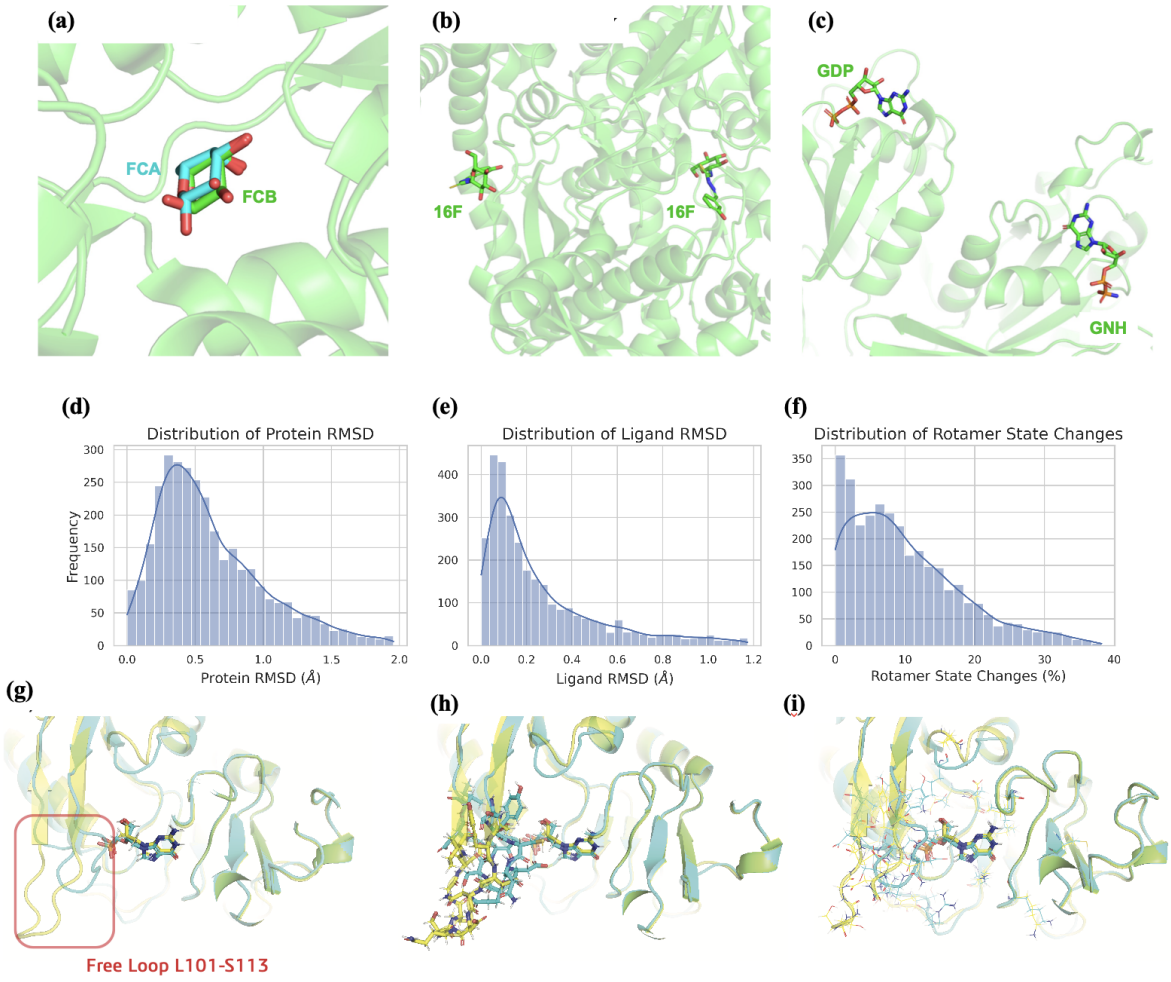

Figure 2.5: *Structural variations and more than one protein-ligand complexes in the same PDB entry.* (a) Two stereoisomers (FCA and FCB) contribute to the binding for 1ABF. (b) The same ligand (16F) binds to two different pockets in 3MRV(c) Two different ligands (GDP) and (GNH) bind to the protein in different pockets for 1A4R. (d) Distribution of protein RMSD. (e) Distribution of ligand RMSD. (f) Percentage of rotamer state changes for residues around the ligand binding sites. Bottom row: visualization of changes in rotamer states between chain A (blue) and chain B (yellow) of PDB ID: 3GEP. (g) Structure overlay between two chains and their respective ligands. (h) Structural differences around the free loop regions between two chains visualized as sticks. (i) Rotamer comparisons of all 29 residues that change their states across chains.

binds to a specific chain of a homodimer protein, two PDB entries are present. PDBbind [35].

usually keeps only one randomly-selected sample of the interacting protein-ligand complex. However, since different chains in PDB are resolved separately using their electron density maps, there are still some level of non-negligible structural variations among different copies, making them valuable data sources for training SFs.

Although the overall protein and ligand RMSD distributions between identical chains of the same entry do not show a great difference (Figure 2.5(d,e)), there is a significant amount of side-chain rotamer state changes observed for different chains as shown in Figure 2.5(f). Here, following the common practice in the field[140], we used the angle cutoff of 60° to any of the four side-chain torsion angles to define a switch in the rotamer states. To illustrate, the protein chains A and B for PDB entry 3GEP[141] in Figure 2.5(g-h) shows that 29 out of 57 residues near the binding site have a change in their side-chain rotameric states, including 12 residues in the free loop area (L101, S103-I113). In particular, the distance between the side chain of D107 and the ligand in chain A (blue) is smaller than 4Å, compared to chain B where the free loop is further from the ligand. Therefore, including multiple records of protein-ligand interactions with the same PDB entry can be informative and beneficial.

To characterize and validate the HiQBind dataset, Figure 2.6 provides the distributions of binding affinities and drug-like properties compared to the PDBbind dataset. It is seen that the new HiQBind dataset shares a very close set of distributions of drug-like properties with PDBbind, especially for the binding affinities in which both datasets cover a large window of approximately 10 log units. We also noticed that HiQBind dataset is a bit more druglike (QED score) with smaller ligands having fewer rotatable bonds and better cLogP/hydrogen-bonding properties. This demonstrates the feasibility of the new HiQBind dataset as a useful resource for future SFs development, benchmarks and other structure-based drug design studies. It is also important to note that, since no standardized method exists for further filtering and data splitting, it is up to the users decide how to perform these additional operations. For example, one might filter out NMR structures and entries with IC_{50} or EC_{50} values, as done in the PDBbind refined set[105], or split the data based on time[109, 116], sequence similarity[13], ECOD classifications[116, 117], or protein–ligand interaction profiles[116]. In the Supporting Information, we also provided analysis upon the time distribution of PDB entries in HiQBind as well as its overlapping with PDBbind v2020 (Figure S1). This shall benefit users who want to do time-based splitting or create independent test set for those models that have already been trained on PDBbind v2020.

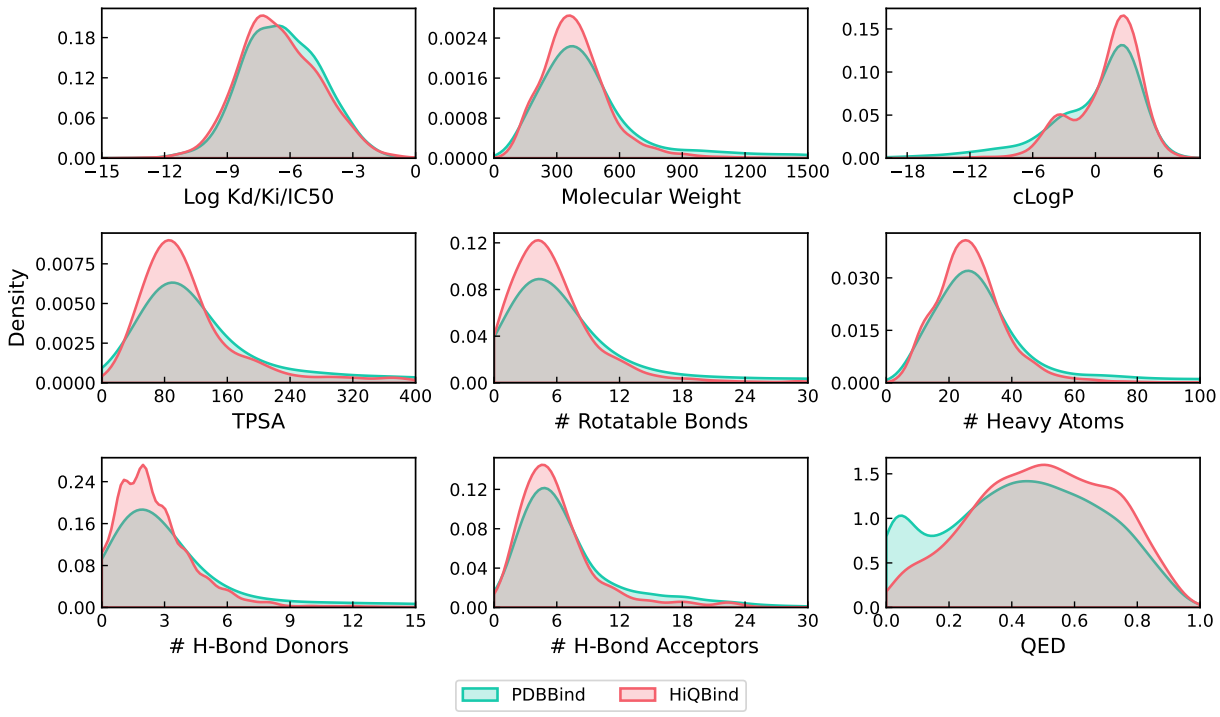

Figure 2.6: *Comparison of the distribution of drug-related properties of ligands and their binding affinities between PDBbind and HiQBind.* (a) Binding K_d, K_i, and IC_{50} values in log units. (b) molecular weight, (c) computed $\log P$ value, (d) topological polar surface area (TPSA), (e) the number of rotatable bonds, (f) the number of heavy atoms, (g) the number of hydrogen-bond donor atoms, (h)the number of hydrogen-bond acceptor atoms, and (i) quantitative estimation of drug-likeness (QED) values.

2.4 Discussion and Conclusion

Many physics-based and machine-learned scoring functions used to predict protein-ligand binding affinities rely on powerful databases such as PDBbind[24, 35, 103–107], BioLiP2[25, 115], Binding MOAD[26], and related databases such as Binding DB[4, 5], Plinder[116], and Dockgen[117]. While central to the biomolecular and drug discovery communities, all data curation efforts require ongoing quality-control efforts. In fact, the latest PDBbind version hosted on the PDBbind+ website have performed some corrections, but PDBbind data curated after v2021 has been commercialized such that it is only accessible to paid users, and there is no published literature describing their workflow. Hence, we are unable to make

a fair comparison between the quality of their generated data compared to that presented here. Therefore, we believe there is a need for the community to have a free open-source tool to curate high-quality protein-ligand structures in a reproducible way.

We have developed an optimization workflow, HiQBind-WF that aims to improve the structural integrity in a semi-automated way and produce high-quality structures with binding affinity annotations. We compared PDBbind v2020 to the structures processed with HiQBind-WF. Differences between the complexes highlighted the strength of our workflow in assigning correct bond orders, protonation states, and protein structure refinements. We also used this workflow to prepare HiQBind, a newly compiled dataset based on BioLiP2[25] and Binding MOAD[26] that offers high-quality, non-covalent protein-ligand complexes with binding-affinity data. HiQBind provides more than 30,000 protein-ligand structures spanning over 18,000 unique PDB entries and is feasible to be deployed in the development of scoring functions or force fields or related activities.

As an open source effort, we believe that HiQBind-WF provides a sustainable framework for continuously updating and refining protein-ligand binding datasets for drug discovery, by meeting scientific goals of ensuring transparency and reproducibility. We also envision that structure-based drug design studies can benefit from the new HiQBind data that has no overlap with PDBbind, and thus reporting evaluation metrics upon this new dataset that will become a common practice as part of future computational modeling efforts.

2.5 Supporting Information

2.5.1 Data Usage

The HiQBind dataset can be found in the following Figshare repository:

- https://doi.org/10.6084/m9.figshare.27430305

which contains the following files:

- `hiqbind.tar.gz`: Protein-Ligand complexes structures in HiQBind
- `hiqbind_metadata.csv`: Metadata for HiQBind dataset
- `hiqbind_sm_metadata.csv`: Metadata for subset of HiQBind in which ligands all small molecules
- `hiqbind_poly_metadata.csv`: Metadata for subset of HiQBind in which ligands are polyemers
- `README.md`: Description of the structural dataset and columns in the metadata csv file

The metadata files are also provided in the Github repository:

- https://github.com/THGLab/HiQBind/blob/main/figshare

Unzipping the data tarball with command `tar -xzvf hiqbind.tar.gz` will yield two directories `raw_data_hiq_sm` and `raw_data_hiq_poly` corresponding to to the "small molecule" and "polymer" subset of HiQBind, respectively. In each directory, you will see a file structure like this:

```
-- 1a69/
   |-- 1a69_FMB_A_240/
       |-- 1a69_FMB_A_240_ligand.pdb
       |-- 1a69_FMB_A_240_protein.pdb
       |-- 1a69_FMB_A_240_protein_hetatm.pdb
       |-- 1a69_FMB_A_240_hetatm.pdb
       |-- 1a69_FMB_A_240_ligand_refined.sdf
       |-- 1a69_FMB_A_240_protein_refined.pdb
   |-- 1a4m_FMB_B_240/
   |-- 1a4m_FMB_C_240/
-- 1a85/
```

An overview of the file contents is in Table S2.1 and here is a description of the naming conventions of the files:

- `1a69`: 4-letter PDB ID
- `FMB`: Name of the ligand. If the ligand is a polymer, it will be format like "ACE-DIP", where "ACE" is the name of the first residue and "DIP" is the name of the last residue.
- `A`: Ligand chain ID.
- `240`: Ligand residue number. If the ligand is a polymer, it will be format like "1-3", where "1" is the residue number of the first residue and "3" is the number of the last residue. Note the residue number may contain insertion code, or be a negative integer or zero.

Table S2.1: Overview of the files in HiQBind.

File	**Description**
*_ligand_refined.sdf	Refined ligand structures (hydrogen added, correct bond order, better tautomer states/protonataion states) with HiQBind-WF workflow
*_protein_refined.pdb	Refined protein structures (hydrogen added, missing atoms/residues added) with HiQBind-WF workflow
*_ligand.pdb	Ligand structure extracted from the original PDB (not processed)
*_protein.pdb	Protein structure extracted from the original PDB (not processed). A protein is defined as chains within 10 angstrom of the ligand structure.
*_hetatm.pdb	Additives' structure extracted from the original PDB (not processed)
*_protein_hetatm.pdb	Protein structure with additives (solvents, ions) extracted from the original PDB (not processed). Additives are specified with "HETATM" atoms that are within 4 angstroms of the protein chains.

A description of all fields in the csv-formatted metatdata file in in Table S2.2.

Table S2.2: Overview of the HiQBind metadata fields.

Field	Type	Description
PDBID	string	Four-letter PDB code
Resolution	string / float	Resolution of the structure or "NMR"
Year	int	Initial release year in RCSB PDB database
Ligand Name	string	Name of the ligand.[a]
Ligand Chain	string	Chain ID of the ligand.
Ligand Residue Number	string	Residue number of the ligand.[b]
Binding Affinity Measurement	string	Type of binding affinity assay: "kd", "ki", "ic50" or "ec50".[c]
Binding Affinity Sign	string	Sign of the binding affinity measurement: "=", ">=", "<=" or "∼".
Binding Affinity Value	float	Value of the binding affinity
Binding Affinity Unit	string	Unit of the binding affinity: "fM", "pM", "nM", "uM", "mM" and "M".
Log Binding Affinity	float	Binding affinity in log unit.
Binding Affinity Source	string	Source of the binding affinity annotations: "BindingMOAD", "BindingDB" or "BioLiP"
Binding Affinity Annotation	string	The annotation in the original source.
Protein UniProtID	string	UniProtID of the proteins, seperated by a comma if the ligand bound to more than one chain.
Protein UniProtName	string	Name of the proteins, separated by a comma if the ligand bound to more than one chain.
Ligand SMILES	string	SMILES of the ligand.
Ligand MW	float	Molecular weight of the ligand.
Ligand LogP	float	LogP value of the ligand computed by RDKit.
Ligand TPSA	float	TPSA value of the ligand computed by RDKit.
Ligand NumRotBond	int	Number of rotatable bonds in the ligand.
Ligand NumHeavyAtoms	int	Number of heavy atoms in the ligand.
Ligand NumHDon	int	Number of hydrogen bond donors in the ligand.
Ligand NumHAcc	int	Number of hydrogen bond acceptors in the ligand.
Ligand QED	float	QED value of the ligand computed by RDKit.

[a] If the ligand is a polymer, its name will be format like "ACE-DIP", where "ACE" is the name of the first residue and "DIP" is the name of the last residue.

[b] If the ligand is a polymer, its residue number will be format like "1-3", where "1" is the number of the first residue and "3" is the number of the last one. Note the residue number may contain insertion code, or be a negative integer or zero.

[c] In some sources where binding data is labeled as K_a or K_b, they are converted to K_d using $K_a = 1/K_d$.

Since users may want to create their own data splits based on time or create subset that does not overlap with exisiting dataset, such as PDBbind v2020, here we provide some relevant analysis in Figure S2.3. All entries in PDBbind v2020 are before 1/1/2020, and HiQBind contains 1,463 PDB entries deposited after this date. Overall, 11,615 PDB entries out of HiQBind's total 18,160 entries are in PDBbind v2020. This overlapping is due to the fact that many entries in also appear in Binding MOAD and BindingDB, which are datasets that HiQBind queries binding-affiniy annotation from.

Figure S2.1: *

Figure S2.2: *

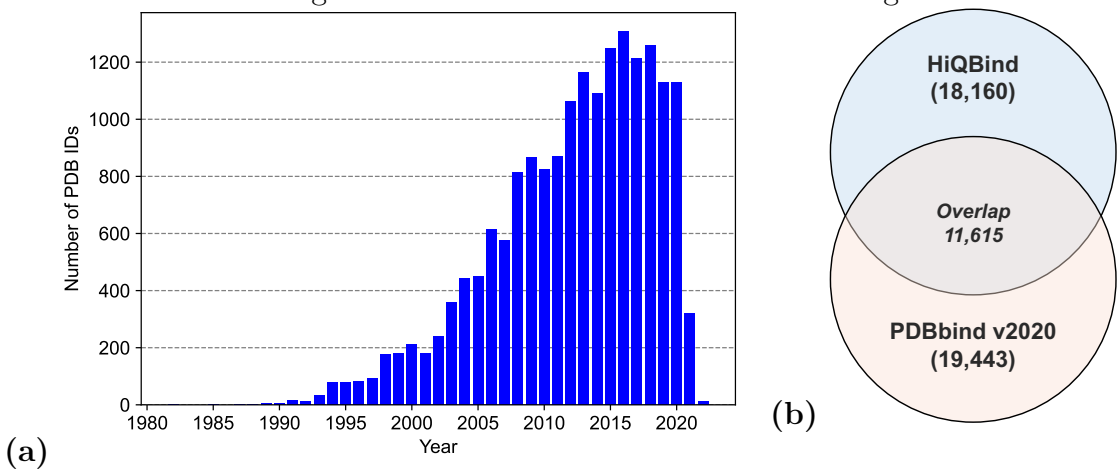

Figure S2.3: **(a)** The year distribution of PDB entries in HiQBind. HiQBind contains 1,463 PDB entries deposited after 1/1/2020. **(b)** HiQBind PDB entries overlapping with those in PDBbind v2020. Among 18,160 unique PDB IDs in HiQBind, 11,615 of them appear in PDBbind v2020.

2.5.2 Failed cases in processing HiQBind

Table S2.3: Overview of the number of entries during filter and structure fixing process of HiQBind.

	Number of unique PDB IDs		
	Small Molecule	Polymer	Total
Before processing	19105	1250	20349[a]
Successfully processed[d]	17725 (31572)	438 (703)	18160[b] (32275)
Failed	1380	812	2189[c]
Covalent binders	262	16	278
Steric clashes	105	7	112
Contain uncommon elements	355	6	361
Small ligands	10	0	10
Fail to fix structures	638	783	1428[c]

[a] 6 PDB entries are included in both small molecule set and polymer set: 1ga8, 1gwv, 1gx4, 1o7o, 1o9f, 4d4u.
[b] 3 PDB entries are sucessfully processed in both small molecule set and polymer set: 1gwv, 1o7o, 1o9f.
[c] 1ga8 and 4d4u are failed in both small molecule set and polymer set; 1gx4 is successful in small molecule set but failed in polymer set.
[d] In parenthesis is the number of protein-ligand structrues.

278 entries are identified with covalent binders:

1au0, 1ayu, 1ayv, 1ayw, 1b0f, 1b4e, 1b6a, 1c3b, 1dru, 1drw, 1erm, 1eta, 1etb, 1fsw, 1fsy, 1ga9, 1gfw, 1h2y, 1hbj, 1i72, 1iau, 1kds, 1kdw, 1ke0, 1ke3, 1lhc, 1lhd, 1ms0, 1mt5, 1my8, 1nl6, 1nlj, 1o2t, 1o41, 1o45, 1o4d, 1o4e, 1o4i, 1o5f, 1o5w, 1pi4, 1pi5, 1q6k, 1qfs, 1qhr, 1qm5, 1re1, 1rhm, 1rhr, 1rhu, 1rwm, 1rww, 1s2y, 1s3b, 1s3e, 1snk, 1sri, 1tu6, 1u9v, 1u9w, 1u9x, 1uks, 1w31, 1wof, 1yt7, 2abe, 2alv, 2amd, 2aux, 2auz, 2bxr, 2bxs, 2c72, 2c73, 2c75, 2c76, 2clx, 2dc6, 2dc8, 2dc9, 2dca, 2dcc, 2dcd, 2fj0, 2fs9, 2fzi, 2g5p, 2g5t, 2g63, 2gsu, 2hpp, 2hpq, 2hwp, 2j5f, 2jai, 2jdh, 2jdk, 2op9, 2qky, 2ql5, 2ql7, 2ql9, 2qlb, 2qlq, 2qq7, 2v6n, 2wn2, 2y4a, 2y55, 2yld, 2zu4, 2zu5, 2zz3, 2zz4, 2zz6, 3afk, 3b0r, 3bls, 3dwq, 3dz5, 3fkv, 3g3d, 3g3m, 3gzn, 3hj0, 3i06, 3i4a, 3ika, 3jwe, 3k7f, 3k83, 3k84, 3kjq, 3kw9, 3kwb, 3kwz, 3lok, 3m3c, 3m3e, 3o1g, 3ovx, 3qug, 3rdh, 3sgu, 3sw6, 3u1i, 3w2t, 3zim, 3zs0, 3zs1, 3zvt, 3zvw, 4an0, 4b16, 4bxn, 4d0l, 4dmx, 4e3n, 4hbp, 4hjs, 4hpc, 4i24, 4ll0, 4lqm, 4lrm, 4lv0, 4oyq, 4rus, 4twy, 4wkt, 4x21, 4yqm, 4ywb, 4z16, 4zzm, 4zzo, 5cyi, 5f2e, 5fct, 5feq, 5gmp, 5i3k, 5j87, 5j9y, 5j9z, 5lf3, 5lf7, 5mxq, 5u4f, 5weu, 5xdk, 5xdl, 5xxd, 5xxg, 5y9t, 5yu9, 5za2, 6alj, 6avi, 6ax1, 6b1e, 6bl1, 6c1i, 6cge, 6cha, 6di9, 6e6e, 6ezp, 6hhf, 6hhi, 6iuo, 6j99, 6jpj, 6jwl, 6jx0, 6jx4, 6jxt, 6k1s, 6lw2, 6lze, 6m0k, 6md1, 6mzw, 6o0e, 6o0f, 6oni, 6p69, 6p7i, 6pdz, 6pnm, 6pnn, 6pno, 6qw7, 6r4v, 6rnu, 6swm, 6t5y, 6tfp, 6tpm, 6v6k, 6v6o, 6v9c, 6vh4, 6vim, 6why, 6wi0, 6wtj, 6wtt, 6wxz, 6xd3, 6xhm, 6xmk, 6xr3, 6y2f, 6yi8, 6yzr, 6zpi, 7b3e, 7bwd, 7c6u, 7c8u, 7cb7, 7cbt, 7d1m, 7dpp, 7jkv, 7jxl, 7jxp, 7jxw, 7k0e, 7k1h, 7lel, 7ltn, 7mlf, 7orf, 7rbz, 7rc0, 7tob

112 entries are identified to exhibit steric clashes:

1bb0, 1ca8, 1e55, 1fbp, 1h00, 1qga, 1rdy, 1vyr, 1w6p, 1ykl, 1ykp, 1z8a, 2c01, 2ime, 2j9m, 2jjk, 2r24, 2r9m, 2vj1, 2w68, 2w92, 2wnj, 2xuc, 2zyk, 3a23, 3cwd, 3d04, 3fck, 3kfy, 3lb4, 3mdj, 3n8y, 3pch, 3qce, 3qcf, 3rho, 3rv7, 3uo9, 3w4k, 3wzu, 3zrc, 4bii, 4d0m, 4e26, 4ele, 4fh7, 4fm5, 4g5y, 4gs9, 4he9, 4ic0, 4jlm, 4otj, 4ucf, 4v04, 4xi3, 4yx9, 4zx6, 4zzw, 5afk, 5dx3, 5dxb, 5ewa, 5fb7, 5fry, 5glt, 5glz, 5gm0, 5h5o, 5hjg, 5hk9, 5hn7, 5hn9, 5hyr, 5i3b, 5is0, 5ivt, 5sye, 5syf, 5t6p, 5vn1, 5wbf, 5xg4, 5z75, 6a6k, 6ah4, 6ah5, 6c7g, 6ch5, 6co5, 6cup, 6eum, 6gno, 6jz0, 6kiu, 6kiv, 6kix, 6kiz, 6n92, 6n94, 6n96, 6ncf, 6ohu, 6oo8, 6ppl, 6pww, 6qdf, 6ula, 6whv, 6wv1, 7c7h, 7cr4

361 entries are identified contain ligand with uncommon elements:

1cp6, 1cws, 1d3v, 1dd7, 1dzj, 1el7, 1el8, 1esz, 1hq5, 1hyv, 1hyz, 1k2v, 1lvk, 1mu9, 1nop, 1rff, 1rfi, 1rg1, 1rg2, 1rgt, 1rgu, 1rh0, 1suo, 1svk, 1tqv, 1tqw, 1v97, 1w0g, 1wva, 1xuf, 1xuj, 1y3g, 1ztz, 2aeb, 2ato, 2bdm, 2bt3, 2c2s, 2c2t, 2cfd, 2ci0, 2cib, 2d0t, 2dy5, 2fb3, 2fdu, 2fdv, 2fdw, 2fou, 2fov, 2foy, 2gj8, 2h7r, 2ij7, 2jjp, 2jld, 2nz5, 2oi4, 2oro, 2orp, 2p8o, 2pll, 2rar, 2rav, 2rb5, 2rbk, 2v0m, 2v96, 2v97, 2w09, 2w0b, 2wfg, 2wh8, 2whf, 2wpu, 2wuz, 2wv2, 2wx2, 2xfh, 2yak, 2ydm, 3b6h, 3b99, 3ben, 3bwf, 3e6k, 3e6v, 3e9b, 3fxz, 3fy0, 3g5n, 3gph, 3i7z, 3i80, 3i8w, 3ixb, 3ixg, 3jt4, 3jt9, 3jus, 3khm, 3koh, 3ksw, 3lc4, 3mdt, 3mke, 3mmr, 3mnu, 3nxu, 3o0j, 3ozu, 3ozv, 3ozw, 3qoa, 3r9c, 3ruk, 3swz, 3t3q, 3t3r, 3t3s, 3t3z, 3tmz, 3vjs, 3vjt, 3wax, 3way, 3zg2, 3zg3, 4c0c, 4c4r, 4c4s, 4coh, 4d2y, 4d2z, 4d30, 4d31, 4d32, 4d33, 4d34, 4d35, 4d36, 4d37, 4d38, 4d39, 4d3a, 4d3b, 4d75, 4d78, 4d7d, 4dtw, 4dtz, 4du2, 4dub, 4due, 4duf, 4ehr, 4ejh, 4eji, 4enh, 4fdh, 4fil, 4g0c, 4g44, 4g45, 4h4d, 4h4e, 4hpd, 4hww, 4hxq, 4hze, 4i06, 4i4g, 4i60, 4ie2, 4ie3, 4iu0, 4iu4, 4ixu, 4ixv, 4j14, 4jfv, 4jfw, 4jjf, 4jjg, 4k9t, 4k9v, 4k9w, 4k9x, 4mdg, 4mdl, 4mdm, 4nkv, 4ob0, 4pf7, 4pk5, 4pk6, 4q3q, 4q3r, 4q3s, 4u72, 4u74, 4uch, 4uhi, 4uhl, 4uqh, 4uzi, 4v3u, 4v3v, 4v3w, 4v3x, 4v3y, 4v3z, 4wmz, 4xrz, 4xud, 4z0l, 4zdy, 4zdz, 4ze0, 4ze1, 4ze2, 4ze3, 4zee, 4zgx, 5agp, 5brv, 5c8i, 5cg5, 5cg6, 5cjh, 5e58, 5e7v, 5ek2, 5ek3, 5ek4, 5etw, 5fom, 5fqb, 5fqc, 5fsa, 5hs1, 5ibe, 5ibf, 5ibg, 5irq, 5irv, 5jqt, 5jws, 5jwu, 5jww, 5k1l, 5k7k, 5l2r, 5l92, 5l94, 5mon, 5moo, 5tz1, 5u48, 5u4a, 5u4c, 5u4e, 5vc0, 5vce, 5veu, 5wmu, 5wn8, 6a16, 6a17, 6abk, 6bbs, 6bc9, 6bcc, 6c6n, 6c8x, 6ceh, 6chi, 6cir, 6ciz, 6cr2, 6d1l, 6d1m, 6da3, 6da5, 6daa, 6dab, 6dac, 6dag, 6daj, 6e40, 6e41, 6f0a, 6ffk, 6frj, 6gcy, 6h1l, 6h1t, 6h3q, 6hwz, 6hx5, 6i96, 6j8r, 6jn6, 6kof, 6kps, 6kvq, 6kw7, 6m7x, 6ma6, 6ma7, 6msn, 6mso, 6o3i, 6pgx, 6pht, 6pu7, 6q2t, 6q37, 6q39, 6qfu, 6qfv, 6qfw, 6qfx, 6qou, 6rpn, 6rvf, 6rvk, 6rvl, 6rw1, 6sp7, 6t7u, 6t9z, 6u30, 6u7o, 6ung, 6unh, 6uni, 6unk, 6unl, 6v7c, 6v7d, 6v7e, 6v7f, 6vbn, 6wpl, 6wr0, 6wr1, 6xz8, 6xz9, 6yzq, 6yzs, 6yzt, 6yzu, 6yzv, 6yzw, 6yzx, 6zjc, 7e0p, 7e7f, 7m8v, 9nse

10 entries are identified to contain small ligands:

1lg6, 1ray, 1ugb, 1ugf, 1zdq, 2ca2, 2hds, 4e5q, 4qef, 6pdv

1428 entries failed the structure fixing:

11gs, 1a07, 1a2c, 1a5v, 1ad8, 1agm, 1apv, 1apw, 1aqc, 1at5, 1at6, 1axw, 1b11, 1b40, 1b5q, 1b6j, 1bdq, 1bfn, 1bm2, 1bzh, 1c0p, 1c5o, 1c5p, 1c5z, 1cka, 1ckb, 1ckp, 1clu, 1d6s, 1deh, 1dmb, 1e4w, 1e5j, 1eb1, 1ec9, 1eef, 1epq, 1evh, 1eyn, 1ez9, 1f5k, 1f86, 1fh7, 1fh8, 1fh9, 1fhd, 1fls, 1fm1, 1fwu, 1fwv, 1g3c, 1g42, 1g9r, 1ga8, 1gah, 1gai, 1gbn, 1gni, 1gnm, 1gnn, 1gno, 1gny, 1gse, 1gu3, 1gui, 1gvu, 1gvx, 1gwl, 1gwq, 1gwr, 1gww, 1h24, 1h2t, 1h2u, 1h9g, 1hew, 1hkk, 1hpb, 1hps, 1htb, 1htg, 1hut, 1i30, 1i7c, 1i7m, 1igj, 1is0, 1it6, 1j1a, 1j84, 1j8v, 1jd5, 1jfh, 1jlx, 1jn2, 1ju9, 1jvp, 1k1y, 1k7t, 1k7u, 1kjl, 1kjr, 1kl3, 1kl5, 1l6m, 1l7z, 1lax, 1lek, 1lf9, 1lkk, 1lqe, 1lr8, 1lzb, 1lze, 1lzg, 1m7d, 1m7i, 1md2, 1mf4, 1mfa, 1mfd, 1mfl, 1mhw, 1n92, 1nm5, 1nzy, 1oau, 1obx, 1oby, 1ocq, 1of4, 1ogg, 1oh3, 1oh4,

1ols, 1olu, 1olx, 1oxn, 1oxq, 1p2g, 1ph0, 1phw, 1pig, 1pl0, 1pmh, 1ppi, 1pzi, 1pzk, 1q0y, 1q4k, 1qaw, 1qiw, 1qj6, 1qj7, 1qja, 1qjb, 1qlu, 1qsc, 1ra1, 1ra9, 1rbm, 1rbz, 1rc0, 1rcv, 1rd9, 1rdj, 1rdl, 1rdm, 1rdp, 1rf2, 1rr8, 1rrj, 1rsu, 1rx9, 1s7y, 1sem, 1seu, 1shd, 1skx, 1sld, 1sle, 1slg, 1sln, 1ssq, 1str, 1sts, 1syn, 1t7d, 1t8i, 1tb4, 1tc1, 1tf8, 1tlc, 1tps, 1tsd, 1tw6, 1u5b, 1ua7, 1ugx, 1ugy, 1ukq, 1ukt, 1uld, 1ule, 1ulg, 1umw, 1ur8, 1ur9, 1utc, 1ux7, 1uxa, 1uxb, 1uz8, 1uzn, 1v0k, 1v0l, 1v11, 1v16, 1v1m, 1vwl, 1vwn, 1w1g, 1w2h, 1w3k, 1w3l, 1w3y, 1w8f, 1w8h, 1w9d, 1w9e, 1w9o, 1w9w, 1wdn, 1wdr, 1wgc, 1wu1, 1x08, 1x09, 1x7w, 1x7x, 1x7y, 1x7z, 1x80, 1xff, 1xkx, 1xl1, 1xnk, 1xt3, 1xt8, 1xzx, 1y3n, 1y3p, 1y3y, 1y79, 1y7l, 1yf4, 1yhm, 1yp9, 1yxd, 1zfp, 1zhs, 1zkk, 1zky, 2a3y, 2ak5, 2aoc, 2aod, 2aoh, 2aou, 2ar6, 2arb, 2ay7, 2az8, 2az9, 2azb, 2azc, 2bcd, 2bel, 2bgr, 2bmz, 2c05, 2c7u, 2cb8, 2cdp, 2chn, 2cjw, 2cvq, 2dt3, 2e95, 2eax, 2er9, 2euk, 2eum, 2evl, 2ff4, 2fgu, 2fgv, 2foq, 2fos, 2gfa, 2gfj, 2ggu, 2gh9, 2h13, 2h6k, 2h6n, 2h6q, 2hah, 2hj4, 2hjb, 2hl3, 2ho2, 2i74, 2ifw, 2ii3, 2imf, 2iqd, 2itk, 2j1t, 2j6o, 2jg8, 2jjb, 2mip, 2msb, 2ntb, 2nwl, 2nwx, 2nxd, 2nxl, 2nxm, 2o9v, 2oei, 2ok0, 2oor, 2ooz, 2orv, 2p4t, 2peh, 2q5a, 2q7q, 2q8y, 2qmj, 2qt5, 2qta, 2qtz, 2r0h, 2r1w, 2r1x, 2r1y, 2r23, 2r3y, 2rca, 2ri9, 2rkm, 2rkn, 2tpi, 2uue, 2uvh, 2uvi, 2uvj, 2uyq, 2v7d, 2v83, 2v86, 2v88, 2v8l, 2vco, 2vj0, 2vl1, 2vnf, 2vnk, 2vpe, 2vqz, 2vsl, 2vt5, 2vuz, 2vzl, 2w1u, 2w3o, 2w5k, 2w5l, 2w7y, 2wa8, 2wcg, 2wd3, 2wdb, 2wev, 2wgc, 2whp, 2wk2, 2wly, 2wm0, 2wn3, 2wra, 2wt0, 2wt1, 2wt2, 2wyf, 2wyn, 2x4t, 2xd3, 2xg3, 2xg9, 2xog, 2xoi, 2xom, 2xon, 2xqq, 2xwe, 2y4s, 2y64, 2y6g, 2y6l, 2yfz, 2yg0, 2yjq, 2ylg, 2yln, 2ypp, 2z8d, 2z8e, 2z8f, 2zcr, 2zcs, 2zex, 2zey, 2zga, 2zhf, 2zhk, 2zhl, 2zhp, 2zm3, 2zno, 2zq0, 2zy1, 3a06, 3ach, 3aci, 3al3, 3avf, 3ayc, 3ayd, 3b4y, 3b9a, 3bbt, 3bho, 3bib, 3biv, 3boo, 3bpc, 3bwr, 3bzi, 3c0z, 3c3d, 3c9e, 3cbs, 3d3x, 3d6o, 3d9t, 3db3, 3dcq, 3diw, 3drg, 3dz2, 3dz4, 3dz6, 3e7a, 3ebb, 3er3, 3eys, 3f3a, 3f5j, 3f5k, 3f5o, 3f5p, 3f70, 3f81, 3f9w, 3f9y, 3fi2, 3fn0, 3g0e, 3g0f, 3g3r, 3g42, 3gjf, 3gl6, 3gss, 3gxz, 3h9f, 3hkn, 3hlo, 3hzk, 3hzy, 3i6c, 3iet, 3if7, 3iiq, 3iit, 3iqg, 3iqh, 3iqi, 3iqj, 3iqv, 3isd, 3ivv, 3jvk, 3jyr, 3jzj, 3k00, 3k45, 3k8d, 3ka2, 3kce, 3kmc, 3kyf, 3kyg, 3lbl, 3lex, 3lgs, 3ll2, 3lpl, 3lq2, 3lq4, 3lqi, 3lqj, 3lsj, 3lug, 3luh, 3lw1, 3m3o, 3m3r, 3mbp, 3muk, 3mxc, 3mxy, 3n0i, 3n7y, 3n8m, 3n8w, 3nf3, 3nkx, 3nq6, 3nw3, 3nxr, 3ny3, 3o0w, 3o0x, 3o6m, 3oea, 3oeb, 3oka, 3okp, 3old, 3ole, 3olg, 3oli, 3oy8, 3p36, 3pj1, 3q8d, 3qg6, 3qmo, 3qx3, 3qxv, 3qzs, 3r42, 3rjf, 3rme, 3rss, 3rul, 3run, 3ryb, 3sm1, 3sou, 3sow, 3sp5, 3t8v, 3tcg, 3ti4, 3tib, 3tkz, 3tl0, 3tlh, 3u4s, 3u4w, 3u51, 3ur0, 3uvm, 3uwn, 3v0w, 3vf9, 3voz, 3vp1, 3vp2, 3vtr, 3vyk, 3waw, 3wch, 3wel, 3weo, 3wol, 3wom, 3won, 3wqd, 3wqv, 3zdv, 3zkk, 3zq9, 3zst, 3zvy, 3zw2, 3zyb, 3zyh, 3zyr, 3zyx, 4a23, 4a3x, 4a45, 4a7a, 4a9u, 4aa1, 4ad2, 4ad3, 4agl, 4aif, 4ap0, 4aph, 4apr, 4asl, 4auy, 4av0, 4av5, 4avi, 4avj, 4aze, 4b4q, 4b83, 4b9z, 4bdo, 4bgx, 4bh0, 4bj0, 4bnk, 4bqr, 4btl, 4c02, 4c0r, 4c1t, 4c1u, 4c5w, 4c9w, 4cd4, 4cd5, 4cd6, 4cd8, 4cdr, 4cov, 4cow, 4coy, 4coz, 4cqm, 4cub, 4d3w, 4d44, 4d4c, 4d4d, 4d4u, 4dhl, 4dro, 4drq, 4e6c, 4egi, 4ep2, 4eqj, 4er4, 4exh, 4ezq, 4ezt, 4ezy, 4fch, 4fem, 4fgx, 4fk7, 4fut, 4g0u, 4g0w, 4g5f, 4g68, 4gah, 4gk9, 4gxk, 4gzf, 4gzw, 4h39, 4h3b, 4hcz, 4hgw, 4hoe, 4hp0, 4hpi, 4hva, 4i23, 4iaw, 4iax, 4iea, 4ifi, 4igk, 4ii9, 4ikn, 4iwd, 4j48, 4jaw, 4je8, 4jih, 4jir, 4jof, 4jog, 4joh, 4joj, 4jok, 4jor, 4jvn, 4k3l, 4k3m, 4k6y, 4k75, 4k76, 4kup, 4l6t, 4lbj, 4lbk, 4lbl, 4lbm, 4lbn, 4lbo, 4lbt, 4ljh, 4lk7, 4lkd, 4lke, 4lkf, 4lkk, 4lkm, 4lmb, 4lnp, 4lo6, 4m7j, 4m9h, 4mbp, 4mdn, 4mdr, 4mh5, 4mo4, 4mr3, 4mr5, 4mx5, 4n84, 4ngp, 4ngs, 4ngt, 4nmo, 4nmp, 4nmq, 4nmr, 4nms, 4nmt, 4nmv, 4nrt, 4nxq, 4ny3, 4o6l, 4o6w, 4oee, 4oef, 4oeg, 4ojz, 4onc, 4ou3, 4owl, 4oxy, 4oyr, 4p0a, 4p7r, 4p8v, 4p8x, 4p9v, 4p9z, 4pa0, 4pfw, 4pfy, 4phv, 4pk3, 4ptc, 4q4s, 4qhp, 4ql1, 4qlk, 4qll, 4qme, 4qvb, 4qyn, 4r5e, 4r5i, 4r6o, 4r6p, 4r6q, 4r6r, 4r6t, 4r73, 4ra5, 4rdd, 4rh5, 4rmh, 4rx0, 4rxh, 4tqm, 4tro, 4u7w, 4ua8, 4uac, 4ufb, 4up4, 4utv, 4uyw, 4v27, 4w9f, 4w9n, 4wa2, 4wey, 4wq3, 4x1r, 4x3h, 4x3r, 4x3s, 4x9r, 4x9v, 4x9w, 4xbl, 4xbn, 4xbq, 4xoq, 4xur, 4xx9, 4y32, 4y5i, 4yee, 4yef, 4yk0, 4ykj, 4ykk, 4ylz, 4ym0, 4ym2, 4ynl, 4yvv, 4yw2, 4yw8,

4yyt, 4yz5, 4yzc, 4z0u, 4z1n, 4zfz, 4zh7, 4zhc, 4zhl, 4zhm, 4zo7, 4zs9, 4zwy, 4zze, 5a3o, 5a4w, 5ab0, 5ab9, 5aci, 5acw, 5acx, 5ajb, 5aom, 5apr, 5ayf, 5azg, 5b2d, 5bmm, 5bo9, 5boo, 5bpc, 5bs8, 5bta, 5btd, 5btf, 5btr, 5btv, 5bxp, 5bxr, 5bxs, 5bxt, 5c0m, 5c1m, 5cpm, 5cqj, 5cs2, 5cvd, 5cxi, 5d6y, 5d8u, 5d9j, 5dms, 5dxg, 5e1b, 5e1d, 5e1o, 5e4t, 5e50, 5e70, 5e76, 5e7g, 5e8f, 5eel, 5eie, 5elf, 5elq, 5elz, 5em9, 5ema, 5ey9, 5f08, 5f3c, 5f3e, 5f3g, 5f3i, 5f4n, 5f5k, 5f7v, 5f90, 5fa5, 5fa6, 5fiv, 5fjx, 5fjz, 5flc, 5fpp, 5fra, 5fre, 5fu2, 5fu3, 5fvs, 5g4c, 5g5z, 5g61, 5ggo, 5ggp, 5ghv, 5glu, 5glw, 5gmv, 5gu4, 5gwk, 5gx6, 5gx7, 5gza, 5gzk, 5hct, 5hda, 5heb, 5hed, 5hes, 5hey, 5hf1, 5hfb, 5hfc, 5hff, 5hla, 5hlb, 5hld, 5hlp, 5htb, 5htc, 5hza, 5hzb, 5i2f, 5i75, 5icv, 5ifu, 5igm, 5iok, 5izf, 5j41, 5j5y, 5j8u, 5jf2, 5jf3, 5jf4, 5jf5, 5jf7, 5jf8, 5jiz, 5jq7, 5jvi, 5kcf, 5kcw, 5kd9, 5kle, 5klf, 5knj, 5kox, 5l41, 5l4l, 5l79, 5l7f, 5lb7, 5ldq, 5lkc, 5ls7, 5lub, 5lyr, 5m17, 5m1z, 5m28, 5m5d, 5m77, 5may, 5mb0, 5mb1, 5mk9, 5mka, 5mks, 5mm9, 5mnh, 5mrd, 5mtt, 5mtu, 5mwj, 5mxg, 5mxr, 5n16, 5n99, 5n9c, 5nf7, 5nfa, 5npr, 5nps, 5o22, 5o4z, 5o58, 5o7u, 5o7v, 5o7w, 5oca, 5ofl, 5ofx, 5omw, 5on2, 5on3, 5onh, 5osx, 5osy, 5q0l, 5swb, 5sz2, 5t4j, 5t54, 5t55, 5t5j, 5t5l, 5t5p, 5t78, 5t7i, 5t7t, 5t8r, 5tdb, 5tdc, 5tdd, 5tef, 5tha, 5tlm, 5tln, 5tlo, 5tlu, 5tp0, 5tpc, 5ttf, 5ttg, 5ttw, 5tyi, 5tzo, 5u06, 5u0f, 5u1q, 5udk, 5uf1, 5uf4, 5ufc, 5ur1, 5v61, 5v62, 5vdk, 5vnf, 5vrw, 5vrx, 5vry, 5vrz, 5vs1, 5vs2, 5vs3, 5vv8, 5w3y, 5w7i, 5w7x, 5wei, 5wje, 5x2a, 5xo2, 5xof, 5y5u, 5yba, 5yd3, 5yd4, 5yd5, 5ygf, 5yqw, 5ysd, 5yse, 5ysf, 5yto, 5ytu, 5yx2, 5yy4, 5zbz, 5zrf, 6a56, 6a80, 6a8n, 6apr, 6apu, 6asz, 6at0, 6ayh, 6b2c, 6b3p, 6b67, 6bn0, 6brr, 6c19, 6c5h, 6c5j, 6c5k, 6c9n, 6c9p, 6c9r, 6cb5, 6cbz, 6ccu, 6cd8, 6cdg, 6cff, 6cy7, 6d6t, 6dfr, 6dmf, 6dqn, 6dqs, 6dqv, 6dqz, 6dr0, 6drc, 6e3d, 6ecz, 6eeh, 6ek3, 6eor, 6eq1, 6eqv, 6eqw, 6eqx, 6f4p, 6f4r, 6f4t, 6f57, 6f5u, 6f6i, 6f6n, 6f6s, 6f8n, 6fau, 6fav, 6faw, 6fbw, 6fby, 6ffl, 6fft, 6fhu, 6fi4, 6fi5, 6fiv, 6fjj, 6flg, 6fmn, 6fn9, 6fpu, 6fqr, 6fu1, 6fuv, 6g15, 6g9b, 6g9i, 6gd0, 6gdo, 6gex, 6gtx, 6gty, 6gtz, 6gu0, 6gw1, 6gx6, 6h0h, 6h4o, 6h4p, 6h4r, 6h4s, 6h4u, 6h4v, 6h4w, 6h4x, 6h4y, 6h50, 6h51, 6h52, 6h5w, 6hcu, 6hhp, 6hi4, 6hi6, 6hi8, 6hm0, 6hmt, 6how, 6hpg, 6hqx, 6hro, 6hs4, 6i0l, 6i1u, 6i41, 6i5p, 6i5v, 6i5w, 6i68, 6i7a, 6ial, 6iam, 6idg, 6im4, 6inz, 6jad, 6jao, 6jba, 6jue, 6kgo, 6kgq, 6kxa, 6kxb, 6l1f, 6lcf, 6ld3, 6ld4, 6ld5, 6lf2, 6lfj, 6lj2, 6lk4, 6lra, 6lrd, 6m6p, 6m8p, 6mle, 6mlo, 6n19, 6n93, 6n97, 6nlk, 6nll, 6nln, 6nxz, 6ny0, 6o5i, 6o6n, 6orv, 6p02, 6pa7, 6phx, 6pre, 6prg, 6pwv, 6q9w, 6qdx, 6qpl, 6qsw, 6qsx, 6qux, 6r18, 6r3m, 6r6v, 6r6w, 6rav, 6rd2, 6rfh, 6rk4, 6rt6, 6rt7, 6rti, 6ryo, 6s1i, 6sey, 6sgf, 6slo, 6spw, 6spx, 6sq0, 6swn, 6swo, 6swq, 6swx, 6t1d, 6t7z, 6tg4, 6th7, 6tkm, 6tmp, 6tmq, 6tmz, 6tn0, 6tn2, 6tpk, 6tpx, 6tpy, 6tpz, 6tv4, 6twu, 6twx, 6u4q, 6u4t, 6u5m, 6u67, 6u8p, 6u8v, 6u8w, 6u8x, 6u90, 6u91, 6uc3, 6udc, 6v2e, 6vhz, 6vuf, 6vw7, 6wf5, 6wj7, 6wm1, 6wo2, 6wp5, 6wy7, 6x23, 6x3v, 6x9c, 6x9i, 6x9j, 6xwd, 6y2g, 6y58, 6yb4, 6yqn, 6yqo, 6yqr, 6yqs, 6yt1, 6yt2, 6z4q, 6z4s, 6z84, 6zb0, 6zb1, 6zb2, 6zb3, 6zfm, 6zin, 6zpd, 7a5m, 7at8, 7bvt, 7c67, 7c68, 7c69, 7c6g, 7c6i, 7c6k, 7c6m, 7c6t, 7c6w, 7c70, 7jjc, 7k0l, 7k0q, 7k5e, 7k5f, 7k5g, 7k5h, 7kkp, 7kme, 7lan, 7luf, 7n7x, 7otm, 7oty, 7pw4, 7pw5, 7pw6, 7pw7, 8icj, 8ico, 8icw, 8icx, 8icy, 8kme, 9icd, 9icu

Note: 1gx4 failed only in small molecule set.

2.5.3 Failed cases in processing PDBBind

958 entries are identified with covalent binders:

1a09, 1a46, 1a5g, 1a61, 1ad8, 1aht, 1amn, 1au0, 1au2, 1auj, 1avp, 1awf, 1awh, 1ayu, 1ayv, 1b0f, 1b5g, 1bgo, 1bio, 1bjr, 1bmq, 1bwn, 1c3b, 1doj, 1e34, 1e37, 1eas, 1eat, 1ekb, 1ero, 1erq, 1exw, 1f1j, 1f7b, 1f9e, 1fsw, 1fsy, 1ft4, 1g37, 1ga9, 1gbt, 1gfw, 1ggd, 1gzg, 1gzv, 1h0w, 1h1b, 1h8y, 1hbj, 1i72,

Table S2.4: Overview of the number of entries during filter and structure fixing process of PDBBind.

	Number of unique PDB IDs		
	Small Molecule	Polymer	Total
Before processing	16738	2705	19443
Successfully processed[a]	14915 (26626)	746 (1131)	15661 (27757)
Failed	1823	1959	3782
Covalent binders	857	101	958
Steric clashes	128	37	165
Contain uncommon elements	197	9	206
Small ligands	1	0	1
Fail to fix structures	640	1812	2452

[a] In parenthesis is the number of protein-ligand structrues.

1i8j, 1iau, 1iem, 1iew, 1inc, 1jlx, 1k2i, 1kds, 1kdw, 1ke0, 1ke3, 1l6s, 1l6y, 1lhc, 1lhd, 1lhe, 1lhf, 1lhg, 1llb, 1me3, 1me4, 1mem, 1mns, 1mpl, 1ms0, 1ms6, 1mwt, 1mxo, 1my8, 1nc6, 1njt, 1nju, 1nkm, 1nl6, 1nlj, 1nms, 1no9, 1npz, 1nqc, 1ny0, 1nyy, 1o2t, 1o41, 1o43, 1o45, 1o4a, 1o4d, 1o4e, 1o4i, 1o4k, 1o5f, 1ong, 1onh, 1p01, 1p02, 1p03, 1p04, 1p05, 1p06, 1p10, 1pau, 1pi4, 1pi5, 1q6k, 1qcp, 1qfs, 1qhr, 1qj1, 1qj6, 1qj7, 1qtn, 1qwu, 1qx1, 1re1, 1rhj, 1rhk, 1rhm, 1rhq, 1rhr, 1rhu, 1rtl, 1rww, 1rwx, 1rxp, 1snk, 1sre, 1sri, 1tbz, 1tlo, 1tmb, 1tu6, 1tyn, 1u9v, 1u9w, 1u9x, 1ukt, 1uod, 1vgc, 1vsn, 1w10, 1w12, 1w14, 1w31, 1x6u, 1y19, 1yk7, 1ylv, 1yly, 1ym1, 1yms, 1yt7, 1z6f, 1zom, 1zpb, 1zpc, 2a4g, 2a4q, 2ajb, 2ajd, 2ajl, 2alv, 2asu, 2aux, 2auz, 2bdl, 2bz5, 2clx, 2eep, 2f9u, 2f9v, 2fda, 2fj0, 2fm2, 2fs8, 2fs9, 2ftd, 2fxr, 2g5p, 2g5t, 2g63, 2g83, 2gbf, 2gbg, 2gph, 2gsu, 2gvf, 2h5d, 2h5i, 2h5j, 2h65, 2ha0, 2hds, 2hob, 2hwo, 2hwp, 2i03, 2i72, 2jal, 2jbv, 2jdh, 2k0x, 2lpr, 2mlm, 2nqg, 2nqi, 2o7e, 2o7v, 2o9a, 2obo, 2oc0, 2oc1, 2oc7, 2op3, 2op9, 2oz2, 2q3z, 2q80, 2q9m, 2q9n, 2qaf, 2qcn, 2qky, 2ql5, 2ql7, 2ql9, 2qlb, 2qlf, 2qlj, 2qlq, 2qnz, 2qq7, 2qve, 2r4b, 2r6n, 2rcx, 2rjr, 2rjs, 2uzj, 2v6n, 2vgc, 2wap, 2wgi, 2wig, 2wij, 2wik, 2wj1, 2wj2, 2woq, 2wzx, 2wzz, 2xcn, 2xdm, 2xdw, 2xe4, 2xk1, 2xlc, 2xln, 2xni, 2xow, 2xu1, 2xu3, 2xu4, 2xu5, 2xzc, 2y2h, 2y2i, 2y2j, 2y2k, 2y2n, 2y2p, 2y4a, 2y55, 2y59, 2yj2, 2yj8, 2yj9, 2yjb, 2yjc, 2z3z, 2zu3, 2zu4, 2zu5, 2zz6, 3a73, 3b1t, 3b1u, 3b9s, 3bar, 3bh3, 3bjm, 3bls, 3blt, 3blu, 3bm6, 3bm8, 3bwk, 3c9e, 3d4f, 3d62, 3dz5, 3e90, 3ewu, 3ex3, 3ex6, 3eyd, 3fkv, 3fmq, 3fmr, 3fnm, 3g3d, 3g3m, 3gjs, 3gpj, 3gzn, 3h0e, 3hd3, 3hha, 3hj0, 3hwn, 3i06, 3i4a, 3ibc, 3ika, 3iut, 3k7f, 3k83, 3k84, 3kjf, 3kjn, 3kjq, 3kqa, 3kw9, 3kwb, 3kwz, 3lj7, 3lle, 3lok, 3lox, 3lpr, 3lxs, 3m3c, 3m3e, 3mbz, 3mkf, 3mxr, 3mxs, 3n4c, 3n5e, 3ns7, 3nzi, 3o1g, 3o6t, 3o86, 3o87, 3o88, 3of8, 3oj8, 3oli, 3opp, 3opr, 3ovx, 3oyp, 3p8e, 3pa8, 3pcb, 3pcf, 3pch, 3pr0, 3qkv, 3qsd, 3rdh, 3rjm, 3s1y, 3s22, 3s3q, 3s3r, 3sji, 3sjo, 3sn8, 3sna, 3snb, 3snc, 3snd, 3sv6, 3sv7, 3sv8, 3svv, 3sz9, 3szb, 3t9t, 3tdz, 3tjm, 3u1i, 3ufa, 3ur9, 3v4j, 3v4x, 3v6r, 3v6s, 3vb4, 3vb5, 3vb6, 3vb7, 3vgc, 3w2p, 3w2q, 3w2t, 3wnr, 3wns, 3wnt, 3zcz, 3zeb, 3zim, 3zmh, 3zmi, 3zmj, 3zot, 3zs0, 3zs1, 3zvt, 3zvw, 4amx, 4amy, 4amz, 4an0, 4an1, 4axm, 4bcb, 4bcc, 4bs5, 4bsq, 4bxn, 4ccd, 4d8e, 4d8i, 4dcd, 4dkt, 4dmy, 4e3i, 4e3j, 4e3k, 4e3l, 4e3m,

4e3n, 4e3o, 4ede, 4eej, 4efg, 4ehm, 4ejf, 4est, 4exz, 4f49, 4fgt, 4fzc, 4fzg, 4gd6, 4ght, 4gk7, 4gkc, 4gs6, 4hbp, 4hcu, 4hcv, 4hnp, 4hrc, 4hrd, 4i7c, 4i7d, 4i9o, 4i9r, 4i9s, 4imq, 4imz, 4inh, 4inr, 4int, 4inu, 4ivk, 4j5p, 4j70, 4jg6, 4jg7, 4jg8, 4jj7, 4jj8, 4jje, 4jmx, 4jr0, 4kqo, 4kw6, 4l0l, 4len, 4lv1, 4lv2, 4lv3, 4lys, 4m1j, 4m8t, 4mao, 4mbf, 4mnv, 4mvn, 4mz4, 4mzo, 4mzs, 4nk3, 4nnn, 4nnw, 4no1, 4no6, 4no8, 4no9, 4o7d, 4ob2, 4ool, 4oon, 4osf, 4pid, 4piq, 4pis, 4pji, 4pkb, 4pl3, 4pl4, 4pl5, 4pnc, 4q1s, 4q2k, 4qbb, 4qkx, 4qps, 4qq5, 4qqc, 4qvl, 4qvm, 4qvn, 4qvp, 4qvq, 4qvv, 4qvw, 4qvy, 4qw0, 4qw1, 4qw3, 4qw4, 4qw5, 4qw6, 4qw7, 4qwf, 4qwg, 4qwi, 4qwj, 4qwk, 4qwl, 4qwr, 4qws, 4qwu, 4qwx, 4qxj, 4qz0, 4qz1, 4qz2, 4qz3, 4qz4, 4qz5, 4qz6, 4qz7, 4qzw, 4qzx, 4r02, 4r17, 4r18, 4r3b, 4r6v, 4rsp, 4ruu, 4s2i, 4tkn, 4tky, 4twy, 4u0g, 4u0x, 4uuq, 4vgc, 4wbg, 4wks, 4wkt, 4wku, 4wkv, 4wm9, 4wmc, 4wsj, 4wsk, 4wx4, 4wx6, 4wx7, 4wyy, 4wz4, 4wz5, 4x0u, 4x21, 4x68, 4x69, 4x6j, 4xbb, 4xbd, 4xcu, 4xjr, 4xuz, 4yas, 4yec, 4yhf, 4yqm, 4yqu, 4yqv, 4yrs, 4yrt, 4yv8, 4z16, 4zro, 5acb, 5ahj, 5c1x, 5c1y, 5c20, 5c91, 5cls, 5cyi, 5d11, 5d6e, 5d6f, 5d9p, 5dg6, 5dgj, 5dp4, 5dp5, 5dp6, 5dp7, 5dp8, 5dp9, 5dpa, 5e0g, 5e0h, 5e0j, 5e7r, 5eb2, 5ee8, 5eec, 5est, 5f02, 5f90, 5fa7, 5fao, 5fap, 5faq, 5fas, 5fat, 5foo, 5fq9, 5g0q, 5gmp, 5gnk, 5gso, 5gty, 5gwa, 5gwz, 5h6v, 5hg5, 5hg7, 5hg8, 5hg9, 5hl9, 5hlb, 5hld, 5i23, 5i24, 5inh, 5j5d, 5j7p, 5j7s, 5j87, 5j8i, 5j8x, 5j9y, 5j9z, 5jh6, 5jk3, 5kre, 5kyk, 5l6h, 5l6i, 5l6j, 5l6o, 5l6p, 5lc0, 5lcj, 5lck, 5lpr, 5mae, 5maj, 5mjb, 5mqy, 5mxq, 5ne1, 5ne3, 5ngf, 5npb, 5nud, 5nwz, 5om9, 5orl, 5swh, 5sys, 5t66, 5t6f, 5t6g, 5tdi, 5teh, 5tg1, 5tg2, 5tg4, 5tg5, 5tg6, 5tg7, 5tig, 5toz, 5tts, 5ttu, 5ttv, 5tyj, 5tyk, 5tyl, 5tyn, 5tyo, 5typ, 5u4f, 5u4g, 5ug8, 5ug9, 5ugc, 5v4q, 5v88, 5vnd, 5vqe, 5vqv, 5vqx, 5vqy, 5vqz, 5w12, 5w13, 5w14, 5wac, 5wad, 5wae, 5waf, 5wag, 5wdl, 5wej, 5wfj, 5x02, 5x5g, 5x79, 5xhr, 5xyz, 5yof, 5yu9, 5za2, 5zde, 5zdg, 5zwf, 6a87, 6af9, 6afa, 6afc, 6afd, 6afe, 6aff, 6afg, 6afh, 6afi, 6afj, 6afl, 6alz, 6ary, 6ax1, 6b0v, 6b0y, 6b1e, 6b1f, 6b1h, 6b1j, 6b1o, 6b1w, 6b1x, 6b1y, 6b41, 6b95, 6bib, 6bic, 6bid, 6bkx, 6bl1, 6bl2, 6bq0, 6cha, 6cn8, 6cqt, 6cqz, 6czu, 6d3g, 6d8e, 6da4, 6db4, 6dge, 6dud, 6e5b, 6e5g, 6e5s, 6e7m, 6ert, 6euv, 6eyz, 6f34, 6f6r, 6fdq, 6fdu, 6ffn, 6ffs, 6fv1, 6fv2, 6g7f, 6g8m, 6g8n, 6g9f, 6g9s, 6gch, 6gcr, 6gop, 6gxy, 6gzy, 6h0u, 6hgy, 6hhg, 6hhh, 6hhi, 6hhj, 6htc, 6htd, 6htp, 6htr, 6hub, 6huc, 6huq, 6huu, 6huv, 6hv4, 6hv5, 6hv7, 6hva, 6hvr, 6hvs, 6hvt, 6hvu, 6hvv, 6hvw, 6i0x, 6ib0, 6ib2, 6ic5, 6ic6, 6iuo, 6iyv, 6iyw, 6j6m, 6jpj, 6k1s, 6lpr, 6m8w, 6m8y, 6m9c, 6m9d, 6m9f, 6mhb, 6mhc, 6mhd, 6mhm, 6mkq, 6mny, 6mu1, 6mzw, 6n4t, 6n9p, 6n9t, 6nd3, 6nng, 6nnr, 6nvg, 6nvh, 6nvi, 6nvj, 6nvl, 6o8i, 6oim, 6ovz, 6p8x, 6p8y, 6p8z, 6pgo, 6pgp, 6pnm, 6pnn, 6pno, 6q35, 6q5b, 6qft, 6qg4, 6qg7, 6qho, 6qhr, 6qmu, 6qw7, 6qw8, 6qw9, 6qwa, 6qwb, 6r4v, 6rjp, 6rmm, 6rn6, 6rn7, 6rn9, 6rne, 6rni, 6rnu, 6rrm, 6rtn, 6s1s, 6s9w, 6s9x, 6skb, 6skd, 6un1, 6un3, 7gch, 7lpr, 8lpr, 9lpr

165 entries are identified to exhibit steric clashes:

1ba8, 1bb0, 1ca8, 1e55, 1gj5, 1gvk, 1h9l, 1ibc, 1jbd, 1jrs, 1nu8, 1oxg, 1pop, 1xxh, 1ykp, 1yyy, 1zzz, 2a4r, 2jjk, 2l75, 2p59, 2pre, 2psx, 2q6f, 2qpj, 2qqs, 2r9m, 2vcb, 2vj1, 2w68, 2w92, 2wnj, 2ww0, 2xuc, 3aav, 3atp, 3b3x, 3bcn, 3d04, 3dkj, 3e0p, 3e16, 3fck, 3fv7, 3fyz, 3fzc, 3gjq, 3gjt, 3gpe, 3gxy, 3kl8, 3lce, 3lpg, 3mo0, 3o0u, 3qce, 3qcf, 3qzq, 3rsb, 3rv7, 3sgv, 3t2c, 3tyq, 3uo9, 3wdc, 3wdd, 3wde, 3wzu, 3zmq, 3zrc, 4a2a, 4e26, 4ele, 4f3h, 4g5y, 4ga3, 4gs9, 4h38, 4h3a, 4he9, 4jlm, 4kiu, 4li5, 4lil, 4loj, 4p00, 4ps0, 4ps1, 4qlq, 4qls, 4qlt, 4qlu, 4qlv, 4v04, 4w5j, 4whs, 4ym4, 4yv2, 4yx9, 4z7f, 4z7q, 4zx6, 5afk, 5dpw, 5dx3, 5ewa, 5f5b, 5fb7, 5g1p, 5gsw, 5h5o, 5hk9, 5hn7, 5hn9, 5htb, 5hvp, 5ivt, 5izk, 5jer, 5n6s, 5ngb, 5npf, 5nqe, 5ouh, 5t6p, 5v5o, 5vfn, 5wbf, 5xg4, 5xw6, 5yjo, 5ykp, 5yr4, 5yr5,

5yr6, 5yun, 5z1d, 6a6k, 6a73, 6b22, 6byz, 6c7g, 6ccx, 6dqb, 6eqs, 6eum, 6fmp, 6hmy, 6ijl, 6ikm, 6jij, 6jki, 6jz0, 6k4r, 6kjf, 6myn, 6n55, 6n92, 6n94, 6n96, 6nt2, 6nzg, 6ohu, 6pk7, 6pvs

206 entries are identified to contain ligand with uncommon elements:

1cp6, 1d3v, 1dzj, 1esz, 1hq5, 1hyv, 1hyz, 1k2v, 1lkx, 1lvk, 1nxy, 1nym, 1pq3, 1wva, 1y3g, 2aeb, 2cfd, 2cfg, 2fou, 2fov, 2foy, 2jld, 2p8o, 2pll, 2v0c, 2v96, 2wf5, 2wfg, 2wq4, 2yak, 2ydm, 2z97, 3bwf, 3c7n, 3csl, 3cst, 3e6k, 3e6v, 3e81, 3e9b, 3fxz, 3fy0, 3ixg, 3m1s, 3mg0, 3mke, 3mmr, 3mnu, 3o0j, 3p3h, 3p3j, 3p44, 3p55, 3pup, 3q4c, 3qlb, 3qo9, 3rj7, 3ro0, 3sjt, 3skk, 3sl0, 3sl1, 3vjs, 3vjt, 3w8o, 3wax, 3way, 3wc5, 3whw, 3zp9, 4aw8, 4daw, 4dcx, 4dcy, 4ehr, 4fea, 4fil, 4g0c, 4h4d, 4h4e, 4hmq, 4hww, 4hxq, 4hze, 4i06, 4i60, 4ido, 4ie2, 4ie3, 4iu0, 4iu4, 4ixu, 4ixv, 4jfv, 4jfw, 4jhq, 4jjf, 4jjg, 4k6t, 4kai, 4kb7, 4kbi, 4kii, 4l6q, 4ob0, 4ob1, 4q3q, 4q3r, 4q3s, 4rlp, 4u5t, 4xkc, 4z46, 5agi, 5agj, 5agr, 5ags, 5agt, 5dhf, 5fjw, 5fom, 5fqb, 5fqc, 5fsb, 5hj9, 5hja, 5hki, 5hlm, 5ll7, 5m29, 5m2q, 5m34, 5m3b, 5mgi, 5mnx, 5mny, 5mo2, 5mon, 5moo, 5mos, 5nxx, 5nxy, 5od5, 5tgy, 5u48, 5u4a, 5u4c, 5u4e, 5ujo, 5vrl, 5wys, 5zeq, 6abk, 6ajz, 6b30, 6bbs, 6bc9, 6c8x, 6ceh, 6d1l, 6d1m, 6e4v, 6frj, 6h3q, 6hwz, 6hx5, 6i0k, 6i0p, 6i96, 6i97, 6ibs, 6ibv, 6j8q, 6j8r, 6jn3, 6jn4, 6jn5, 6jn6, 6msn, 6mso, 6naf, 6nti, 6ntj, 6pgx, 6pht, 6q2y, 6q30, 6q37, 6q39, 6q92, 6qaf, 6qfu, 6qfv, 6qfw, 6qfx, 6rmf, 6rsa, 6rvf, 6rvk, 6rvl, 6rw1, 6skc, 6u7o, 6u7p, 6uhu

1 entry are identified to contain small ligands:

6eu6

2452 entries failed the structure fixing:

11gs, 1a07, 1a0t, 1a2c, 1a37, 1a3e, 1abf, 1abt, 1af2, 1agm, 1apb, 1apv, 1apw, 1aqc, 1at5, 1at6, 1atl, 1aze, 1azx, 1b11, 1b2m, 1b40, 1b6j, 1bap, 1bdl, 1bdq, 1bm2, 1bm6, 1bsk, 1bt6, 1bux, 1bzh, 1c5o, 1c5p, 1c5z, 1cka, 1ckb, 1clu, 1cpi, 1cyn, 1czq, 1d4w, 1d6s, 1d8e, 1dkd, 1dmb, 1dva, 1dxp, 1e03, 1e5j, 1eb1, 1ec9, 1eef, 1ej4, 1eoj, 1eol, 1epq, 1eub, 1evh, 1ez9, 1f47, 1f4y, 1f5k, 1ff1, 1fh7, 1fh8, 1fh9, 1fhd, 1fls, 1fwu, 1fwv, 1g42, 1g6g, 1g9r, 1ga8, 1gag, 1gah, 1gai, 1gmy, 1gni, 1gnj, 1gnm, 1gnn, 1gno, 1gny, 1gu3, 1gui, 1gvu, 1gvx, 1gwm, 1gwq, 1gwr, 1gwv, 1gzc, 1h00, 1h07, 1h24, 1h25, 1h26, 1h27, 1h28, 1h2t, 1h2u, 1h5v, 1h6e, 1hc9, 1hgt, 1hkj, 1hkk, 1hkm, 1hps, 1htg, 1i3z, 1i6v, 1i7c, 1i7m, 1i8h, 1i8i, 1idg, 1igj, 1iht, 1ikt, 1ilq, 1iq1, 1is0, 1it6, 1iwq, 1j19, 1j1a, 1j4q, 1jd5, 1jd6, 1jfh, 1jh1, 1jm4, 1jmq, 1jn2, 1jp5, 1jpl, 1juq, 1jvp, 1k1y, 1k9q, 1kat, 1kc5, 1kcs, 1kjr, 1kl3, 1kl5, 1kna, 1kne, 1l6m, 1lek, 1lf8, 1lf9, 1lkk, 1ll4, 1lqe, 1lt5, 1lxh, 1m7d, 1m7i, 1mf4, 1mfa, 1mfd, 1mhw, 1mpa, 1mv0, 1n3w, 1n4m, 1n5z, 1n7m, 1nde, 1ngw, 1nlo, 1nlp, 1nlt, 1ny2, 1o9k, 1oau, 1obx, 1ocn, 1ocq, 1od8, 1oeb, 1ogg, 1oh4, 1oj5, 1ok7, 1oko, 1ols, 1olu, 1olx, 1om9, 1orw, 1osg, 1osv, 1ov3, 1ow6, 1ow7, 1ow8, 1oxn, 1oxq, 1oy7, 1ozv, 1p28, 1p2g, 1p4u, 1pcg, 1pdq, 1ph0, 1pig, 1pl0, 1pmx, 1ppi, 1pum, 1pxh, 1py1, 1pyw, 1pzi, 1q4k, 1qaw, 1qi0, 1qiw, 1qja, 1qjb, 1qm5, 1qsc, 1r17, 1r2b, 1r6z, 1rdj, 1rdl, 1rdn, 1rgj, 1s9v, 1shd, 1sje, 1sld, 1sle, 1slg, 1sln, 1sm3, 1sps, 1ssq, 1str, 1sts, 1szm, 1t29, 1t2v, 1t37, 1t79, 1t7d, 1t7f, 1t7r, 1tc1, 1tet, 1ths, 1tl9, 1tps, 1ttv, 1tyr, 1u8t, 1uef, 1ugx, 1ugy, 1uh1, 1uj0, 1ujj, 1ujk, 1ukh, 1ule, 1ulg, 1umw, 1upk, 1ur9, 1urc, 1urg, 1utc, 1uti, 1uvu, 1ux7, 1uxa, 1uxb, 1uz8, 1v0k, 1v0l, 1v0m, 1v0n, 1v11, 1v16, 1v1m, 1vj6, 1vr1, 1vwl, 1vwn, 1w1g, 1w2h, 1w3k, 1w3l, 1w70, 1w80, 1wdn, 1wdq, 1wdr, 1ws5, 1wu1, 1x11, 1x8s, 1x9d, 1xb7, 1xff, 1xhm, 1xn2, 1xn3, 1xt3, 1xt8, 1y3a, 1y3n, 1y3p, 1y3y, 1ybg, 1ybo, 1yhm, 1ymx, 1yp9, 1yvh, 1ywi, 1yxd, 1yy6, 1z3t, 1z3v, 1zfp, 1zkk, 1zky, 1zub, 2a25, 2aez, 2aof, 2aoh, 2aoi, 2aoj, 2aou, 2aq9, 2auc, 2ay7, 2az8, 2az9, 2azb, 2azc, 2azm,

2b1q, 2b1r, 2b2v, 2b7f, 2bba, 2bcd, 2bgn, 2bgr, 2bmz, 2br8, 2byp, 2c1n, 2c9t, 2ce9, 2cht, 2ci9, 2cia, 2co0, 2d1x, 2d2v, 2df6, 2dwx, 2e7l, 2e95, 2e98, 2e9a, 2e9c, 2eh8, 2emt, 2er0, 2er9, 2euk, 2eum, 2evl, 2ez5, 2f5t, 2f6j, 2fci, 2fgu, 2fgv, 2flu, 2fr8, 2frd, 2fsa, 2fts, 2fuu, 2fx9, 2fys, 2g6q, 2gfa, 2ggu, 2gh9, 2h13, 2h2d, 2h2e, 2h2g, 2h2h, 2h6k, 2h6q, 2h9m, 2h9n, 2h9p, 2hah, 2hdx, 2hj4, 2hjb, 2hkf, 2hmh, 2hrp, 2ig0, 2igv, 2igw, 2itk, 2iv9, 2ivz, 2j7w, 2j9n, 2jb5, 2jbu, 2jdk, 2jdl, 2jg8, 2jjb, 2jk9, 2jkr, 2jkt, 2jmj, 2jnw, 2jq9, 2jqk, 2k1q, 2k3w, 2kgi, 2knh, 2kup, 2l0i, 2l65, 2l7u, 2l8j, 2lcs, 2lct, 2liq, 2llq, 2lp8, 2m0o, 2m3o, 2mg5, 2mip, 2mkr, 2mov, 2mow, 2mpa, 2mwy, 2n3k, 2n7b, 2nn8, 2nwl, 2nwn, 2nxd, 2nxl, 2nxm, 2o9k, 2o9r, 2o9v, 2odd, 2oei, 2oi9, 2ooz, 2peh, 2pem, 2pl9, 2pmc, 2pnx, 2pv3, 2q7q, 2q7y, 2q8y, 2qbx, 2qic, 2qki, 2qmj, 2qt5, 2qta, 2qtr, 2qv7, 2qwe, 2r02, 2r03, 2r05, 2r0h, 2r0y, 2r1w, 2r1x, 2r1y, 2r23, 2r2b, 2r3c, 2r3y, 2r5b, 2r7g, 2rfy, 2ri9, 2rkm, 2rkn, 2rok, 2rol, 2rvn, 2srt, 2tpi, 2uw0, 2uyq, 2uz6, 2v7d, 2v83, 2v85, 2v86, 2v87, 2v88, 2vl1, 2vnf, 2vpe, 2vpg, 2vr3, 2vsl, 2vwf, 2vxj, 2w0p, 2w0z, 2w10, 2w16, 2w2u, 2w3o, 2w47, 2w6c, 2w6t, 2w6u, 2w73, 2w76, 2w77, 2w78, 2w7y, 2w9r, 2wa8, 2wcg, 2wd3, 2whp, 2wk2, 2wly, 2wlz, 2wm0, 2wp1, 2wyf, 2wyn, 2wzf, 2x2i, 2x3t, 2x4t, 2x4z, 2x52, 2x6w, 2x6x, 2x6y, 2x85, 2xaf, 2xag, 2xah, 2xaj, 2xaq, 2xas, 2xcs, 2xct, 2xg3, 2xg9, 2xhs, 2xl2, 2xl3, 2xn6, 2xn7, 2xog, 2xoi, 2xqq, 2xrw, 2xs0, 2xs8, 2xwd, 2xwe, 2xxn, 2xzq, 2y06, 2y07, 2y1n, 2y36, 2y4m, 2y4s, 2y6s, 2y8i, 2y8o, 2y9g, 2y9q, 2ydt, 2yhw, 2yjq, 2ylc, 2yln, 2ymt, 2ynr, 2yns, 2ypp, 2yq6, 2z5o, 2z5s, 2z5t, 2zcr, 2zcs, 2zg3, 2zga, 2zgm, 2zm3, 2zpk, 2zq0, 2zy1, 2zym, 2zyn, 3afk, 3al3, 3alt, 3ap4, 3ap7, 3ary, 3arz, 3as3, 3ask, 3asl, 3au6, 3avf, 3avg, 3avh, 3ax5, 3aya, 3ayc, 3ayd, 3b3s, 3b95, 3bbb, 3bbt, 3bg8, 3bho, 3bim, 3bpc, 3btr, 3bu6, 3bu8, 3bum, 3bun, 3buo, 3buw, 3bux, 3bzi, 3c0z, 3c1n, 3c6w, 3c94, 3cbs, 3cfs, 3cfv, 3ck7, 3ck8, 3ckb, 3coj, 3cs8, 3d1e, 3d1f, 3d3x, 3d45, 3d6o, 3d9k, 3d9l, 3d9m, 3d9n, 3d9o, 3d9p, 3dab, 3dcq, 3diw, 3dla, 3dnj, 3dow, 3dpo, 3drf, 3drg, 3dri, 3ds1, 3ds9, 3dvp, 3dz2, 3dz4, 3dz6, 3e7a, 3e8u, 3ebb, 3eg6, 3ehn, 3eht, 3emh, 3eqs, 3er3, 3ery, 3eu7, 3evc, 3evd, 3evf, 3eyf, 3eys, 3eyu, 3f3a, 3f5j, 3f5k, 3f5l, 3f5p, 3f69, 3f70, 3f81, 3f9w, 3f9y, 3fbr, 3fdm, 3fdt, 3fi2, 3fn0, 3fqa, 3fuc, 3fv8, 3g0e, 3g0f, 3g2s, 3g2t, 3g2u, 3g2v, 3g2w, 3g3r, 3g42, 3g5v, 3g5y, 3g7l, 3gds, 3ggw, 3ghe, 3gl6, 3gsm, 3gss, 3gv6, 3gxz, 3h52, 3h6z, 3h91, 3h9f, 3hkn, 3hkt, 3hlo, 3hqh, 3hs8, 3hs9, 3hzk, 3hzv, 3hzy, 3i02, 3i5r, 3i6c, 3i8t, 3i90, 3i91, 3iet, 3if7, 3ifl, 3ifo, 3ifp, 3iit, 3iiw, 3iiy, 3ij0, 3ij1, 3ijy, 3ikc, 3iqg, 3iqh, 3iqi, 3iqj, 3iqq, 3iss, 3isw, 3iux, 3ivq, 3ivv, 3iw7, 3jpx, 3juq, 3jvk, 3jyr, 3jzg, 3jzh, 3jzj, 3k00, 3k26, 3k27, 3k48, 3k8d, 3ka2, 3kmc, 3krd, 3ktr, 3kyf, 3kyg, 3kze, 3l3q, 3l3x, 3l3z, 3l6x, 3lbl, 3lgl, 3lgs, 3lk1, 3lnj, 3lnz, 3lpl, 3lq2, 3lq4, 3lqi, 3lqj, 3luo, 3m3o, 3m3r, 3m53, 3m54, 3m55, 3m56, 3m57, 3m58, 3m59, 3m5a, 3mbp, 3me9, 3mea, 3met, 3meu, 3ml4, 3mp1, 3mp6, 3muk, 3mxc, 3mxy, 3n5u, 3nf3, 3nfk, 3nfl, 3nii, 3nij, 3nil, 3nin, 3nkx, 3nsn, 3nti, 3nw3, 3ny3, 3o0e, 3o1d, 3o1e, 3o6l, 3o6m, 3ob0, 3ob1, 3ob2, 3odi, 3odl, 3ogx, 3oka, 3okp, 3old, 3ole, 3olg, 3omc, 3omg, 3oq5, 3oy8, 3oyw, 3p4f, 3pdh, 3pfp, 3pgu, 3pj1, 3pkn, 3plu, 3pma, 3poa, 3pp7, 3pqz, 3psl, 3puj, 3puk, 3pxe, 3q5u, 3q6s, 3q8d, 3qfy, 3qfz, 3qg6, 3ql9, 3qlc, 3qmk, 3qn7, 3qnj, 3qo2, 3qs4, 3qxd, 3qxv, 3qzt, 3qzv, 3r42, 3r93, 3rbq, 3rdv, 3rg2, 3rl7, 3rl8, 3rme, 3rqe, 3rqf, 3rqg, 3rtx, 3rul, 3rum, 3run, 3rv6, 3rv8, 3rz9, 3rzi, 3s7f, 3shb, 3shv, 3sm1, 3so6, 3sou, 3sov, 3sow, 3stj, 3sw9, 3sxu, 3szm, 3t5i, 3t6r, 3t7g, 3t83, 3t8v, 3tcg, 3tdu, 3tf6, 3tf7, 3tg5, 3th0, 3ti4, 3tib, 3tiw, 3tkz, 3tl0, 3tlh, 3tpx, 3tsz, 3twr, 3tws, 3twu, 3twv, 3tww, 3twx, 3tzd, 3u3f, 3u78, 3ual, 3uat, 3ud7, 3ud8, 3ud9, 3uda, 3ued, 3uef, 3ueo, 3ui2, 3uig, 3uih, 3uii, 3uij, 3uik, 3upk, 3ur0, 3uri, 3uvk, 3uvl, 3uvm, 3uvn, 3uvo, 3uvu, 3uvw, 3uvx, 3uw9, 3uwl, 3ux0, 3uxg, 3uyr, 3uzd, 3v2o, 3v30, 3v3b, 3v43, 3v4t, 3v7d,

3va4, 3vf9, 3vfj, 3voz, 3vp1, 3vp2, 3vp3, 3vp4, 3vtr, 3vzg, 3w37, 3waw, 3wcb, 3wch, 3wdz, 3wp0, 3wp1, 3wqv, 3wqw, 3wsy, 3wut, 3wuu, 3wuv, 3zdv, 3zev, 3zha, 3zhf, 3zi8, 3zjt, 3zju, 3zjv, 3zke, 3zkf, 3zlv, 3zmp, 3zmt, 3zmu, 3zmv, 3zmz, 3zn0, 3zn1, 3zq9, 3zqi, 3zst, 3zvy, 3zyb, 3zyh, 3zyr, 4a0j, 4a1w, 4a23, 4a4c, 4a50, 4a7j, 4a9t, 4a9u, 4aa1, 4aa2, 4abi, 4abj, 4ad2, 4ad3, 4agl, 4aif, 4aom, 4ap0, 4aph, 4apr, 4auy, 4av0, 4av5, 4avi, 4avj, 4ay6, 4ayp, 4aze, 4b4n, 4b4q, 4b60, 4b83, 4b8o, 4b8p, 4b8y, 4b9h, 4b9w, 4b9z, 4ba3, 4bea, 4bg6, 4bgx, 4blb, 4bpi, 4bpj, 4btl, 4bv2, 4bxu, 4c0r, 4c16, 4c1t, 4c1u, 4c1w, 4c4n, 4c5w, 4c9w, 4cc2, 4cc7, 4cd4, 4cd5, 4cd6, 4cd8, 4cdr, 4ch2, 4ch8, 4ciz, 4cpq, 4cps, 4cpx, 4csy, 4cy1, 4czs, 4d1d, 4d4d, 4de7, 4dhl, 4djs, 4dma, 4dow, 4dro, 4ds1, 4dx9, 4e35, 4e3b, 4e6c, 4e81, 4e9c, 4e9d, 4edu, 4egi, 4elb, 4elg, 4elh, 4ep2, 4eqf, 4eqj, 4er4, 4erq, 4ery, 4erz, 4es0, 4esg, 4ewr, 4exh, 4ezo, 4ezq, 4ezt, 4ezy, 4f14, 4f20, 4fbx, 4fcm, 4fe9, 4fem, 4fgx, 4fgy, 4fk7, 4fmn, 4fmo, 4fmq, 4fn5, 4ft2, 4fut, 4g0a, 4g5f, 4g68, 4g69, 4gah, 4gao, 4gj8, 4glr, 4glx, 4gne, 4gnf, 4gng, 4gq6, 4gvc, 4gvd, 4gw1, 4gw5, 4gwi, 4gxl, 4gy5, 4gye, 4gzf, 4gzw, 4gzx, 4h36, 4h39, 4h3b, 4h3q, 4hcz, 4hfz, 4hgc, 4hp0, 4hpi, 4hpy, 4hs6, 4hs8, 4htp, 4hva, 4hy9, 4hyb, 4i2z, 4i31, 4i32, 4i33, 4i67, 4i7b, 4iaw, 4iax, 4ib5, 4ifi, 4igk, 4igq, 4ii9, 4ikn, 4ipn, 4is6, 4iur, 4iut, 4iuu, 4iuv, 4iwd, 4j09, 4j24, 4j26, 4j2c, 4j3u, 4j48, 4j4v, 4j73, 4j77, 4j7i, 4j84, 4j8g, 4j8r, 4j8s, 4jc1, 4jck, 4je8, 4jfx, 4jfz, 4jg0, 4jg1, 4jiz, 4jjq, 4jmg, 4jmh, 4jof, 4jog, 4joh, 4joj, 4jok, 4k0o, 4k0u, 4k3l, 4k3m, 4k63, 4k64, 4k66, 4k67, 4k6u, 4k6v, 4k6w, 4k6y, 4k72, 4k75, 4k76, 4kc1, 4kc2, 4kc4, 4kmd, 4kn7, 4kom, 4kon, 4ktu, 4kup, 4kvm, 4kx8, 4l1u, 4l58, 4l6t, 4lbl, 4lbo, 4lg6, 4ljh, 4lk6, 4lk7, 4lkd, 4lke, 4lkf, 4lkg, 4lkh, 4lkk, 4lkm, 4ln2, 4lnf, 4lno, 4lnp, 4lp6, 4lq3, 4lte, 4m1d, 4m7j, 4mbp, 4mdn, 4mdr, 4mg5, 4mo4, 4mr3, 4mr5, 4mrd, 4mx5, 4mz5, 4mz6, 4mzf, 4mzh, 4mzj, 4mzk, 4mzl, 4n3w, 4n6g, 4n7g, 4n7h, 4n7j, 4n7y, 4n84, 4nb3, 4ngn, 4ngp, 4ngq, 4ngs, 4ngt, 4nku, 4nl1, 4nmo, 4nmp, 4nmq, 4nmr, 4nms, 4nmt, 4nmv, 4nrk, 4nrl, 4nrt, 4nuf, 4nw2, 4nxq, 4ny3, 4o0r, 4o36, 4o3t, 4o3u, 4o42, 4o45, 4o4y, 4o62, 4o6w, 4oak, 4odk, 4odl, 4odm, 4odn, 4odp, 4odq, 4oee, 4oef, 4oeg, 4oel, 4oem, 4ofl, 4onf, 4oru, 4orx, 4ory, 4ou3, 4ouj, 4ov5, 4oyk, 4oz1, 4p0a, 4p0b, 4p0n, 4p4s, 4pft, 4pfu, 4pgc, 4phv, 4pl6, 4pli, 4pn1, 4pnw, 4po7, 4pry, 4psx, 4ptc, 4pvo, 4pxf, 4pz5, 4pz8, 4q1e, 4q4s, 4q6f, 4qaa, 4qc1, 4qf7, 4qfl, 4qfn, 4qfo, 4qfp, 4qh7, 4qh8, 4qhp, 4ql1, 4qlk, 4qll, 4qme, 4qq4, 4qqi, 4qsk, 4qxt, 4qy8, 4r1e, 4r3s, 4r6t, 4ra1, 4ra5, 4rh5, 4rhu, 4ris, 4rme, 4rqi, 4rqz, 4rrv, 4rxh, 4rxz, 4tk1, 4tk2, 4tk3, 4tk4, 4tmp, 4tnw, 4tt2, 4tw8, 4twt, 4tzm, 4tzn, 4tzq, 4u0a, 4u0b, 4u0c, 4u0d, 4u2w, 4u68, 4u6x, 4u7t, 4u90, 4ua8, 4uac, 4ud7, 4ue1, 4um9, 4umn, 4utn, 4utr, 4utv, 4utx, 4uu5, 4uu7, 4uu8, 4uua, 4uub, 4uw1, 4ux9, 4uxj, 4v1f, 4v27, 4w4z, 4w50, 4w5a, 4w9f, 4w9n, 4wci, 4wey, 4wht, 4why, 4wj7, 4wko, 4wph, 4wq3, 4wrq, 4wv6, 4wy7, 4wym, 4x0z, 4x13, 4x14, 4x1n, 4x1p, 4x1q, 4x1r, 4x1s, 4x34, 4x3e, 4x3h, 4x3i, 4x3k, 4x3r, 4x3s, 4x6h, 4x6s, 4x8n, 4x8p, 4x9r, 4x9v, 4x9w, 4xc2, 4xek, 4xgz, 4xh2, 4xqu, 4xtp, 4xx9, 4xxh, 4xyn, 4y32, 4y3b, 4y5i, 4yb5, 4yc8, 4ydf, 4ydn, 4yee, 4yef, 4yhp, 4yhz, 4yje, 4yjl, 4yk0, 4ykj, 4ykk, 4ym2, 4ynl, 4yoz, 4ysi, 4yw2, 4yy6, 4yyi, 4yym, 4yyn, 4yyt, 4yz5, 4yzc, 4z0d, 4z0e, 4z0f, 4z0u, 4z1n, 4z2o, 4z2p, 4z68, 4z7i, 4z83, 4z88, 4z89, 4z8m, 4zdu, 4zeb, 4zhl, 4zhm, 4znx, 4zs9, 4zwy, 5a0e, 5a2i, 5a2j, 5a2k, 5a3h, 5a3o, 5ab0, 5ab1, 5ab9, 5abp, 5acw, 5acx, 5ajc, 5ajo, 5ajp, 5aom, 5apr, 5awt, 5awu, 5ayf, 5azg, 5b2d, 5b4w, 5b56, 5b6g, 5bjt, 5bmm, 5btr, 5btv, 5c0m, 5c11, 5c13, 5c1m, 5c6v, 5c7e, 5c7f, 5cbm, 5cfa, 5cil, 5cin, 5cqj, 5cqx, 5cr7, 5cs2, 5csz, 5cvd, 5cw8, 5cxi, 5d0j, 5d1u, 5d2a, 5d6y, 5d7e, 5dah, 5dif, 5dms, 5dtj, 5duw, 5dxb, 5dxe, 5dxg, 5e0l, 5e0m, 5e1b, 5e1d, 5e1o, 5e2v, 5e2w, 5e4w, 5e8f, 5eay, 5eel, 5eeq, 5eie, 5ekg, 5elf, 5elq, 5em9, 5ema, 5emb, 5eoc, 5eok, 5epp, 5esq, 5eta,

5etf, 5etu, 5euk, 5ewz, 5ey8, 5ey9, 5eyz, 5ez0, 5f08, 5f2u, 5f3c, 5f3e, 5f3g, 5f3i, 5f4n, 5f67, 5f88, 5fb0, 5fb1, 5ff6, 5fh6, 5fiv, 5fjx, 5fkj, 5fos, 5fpi, 5fpp, 5fyq, 5g5z, 5g60, 5g61, 5g6u, 5gg4, 5ggo, 5ggp, 5ghv, 5glu, 5gmi, 5gmj, 5gmv, 5gp7, 5gs4, 5gtr, 5gu4, 5gwy, 5gx6, 5gx7, 5h1e, 5h5q, 5h5r, 5h5s, 5hct, 5hda, 5heb, 5hed, 5hes, 5hex, 5hey, 5hf1, 5hfb, 5hfc, 5hff, 5hhx, 5hjb, 5hjc, 5hjd, 5hkh, 5hlp, 5hog, 5hpm, 5htc, 5huw, 5huy, 5hyq, 5hyr, 5i25, 5i2f, 5i2i, 5i8c, 5iaw, 5ick, 5icv, 5icx, 5icy, 5icz, 5id0, 5id1, 5ifu, 5igm, 5igq, 5ijj, 5ijp, 5iok, 5iop, 5ir1, 5itf, 5iv2, 5ivn, 5ivz, 5ix1, 5ixt, 5iy4, 5iyv, 5iz6, 5izf, 5izj, 5j19, 5j31, 5j3v, 5j41, 5j5x, 5j7j, 5j8u, 5j9k, 5jek, 5jeo, 5jf2, 5jf3, 5jf4, 5jf5, 5jf7, 5jf8, 5jin, 5jiy, 5jjm, 5jlz, 5jm4, 5jop, 5jq7, 5jqb, 5jr2, 5jvi, 5jy0, 5k5c, 5k6s, 5kez, 5kgn, 5klr, 5klt, 5knj, 5kqd, 5ksu, 5ksv, 5kzp, 5l0c, 5l0h, 5l3f, 5l3g, 5l7f, 5l7k, 5lax, 5lb7, 5lbq, 5lgp, 5lgq, 5lgr, 5lgs, 5lrk, 5lsh, 5lso, 5lu2, 5lub, 5lvx, 5ly1, 5ly2, 5ly3, 5lyr, 5lzh, 5m17, 5m1z, 5m28, 5m5d, 5m63, 5m77, 5mav, 5may, 5mb1, 5mby, 5mgx, 5mk1, 5mk3, 5mk9, 5mka, 5mks, 5mlo, 5mlw, 5mm9, 5mng, 5mnh, 5mo0, 5moq, 5mrd, 5mtw, 5mwj, 5mxo, 5mxr, 5myk, 5myo, 5myx, 5n16, 5n31, 5n7b, 5n7g, 5n7x, 5n8e, 5n8j, 5n8t, 5n8w, 5n99, 5n9n, 5nfa, 5nin, 5njx, 5nne, 5npr, 5nps, 5nw8, 5nwk, 5nx2, 5nxq, 5o22, 5o45, 5o4y, 5o4z, 5o58, 5o5m, 5ocj, 5ods, 5ofx, 5ogl, 5ok6, 5osy, 5ot3, 5oua, 5ous, 5oxk, 5oxl, 5oxm, 5oxn, 5oy3, 5oyd, 5q0l, 5sve, 5svi, 5svx, 5svy, 5svz, 5swf, 5sz2, 5szb, 5szc, 5t1i, 5t1k, 5t1l, 5t1m, 5t31, 5t52, 5t54, 5t6j, 5t78, 5t7s, 5t8r, 5t90, 5tdb, 5tdr, 5tdw, 5tef, 5teg, 5th2, 5th7, 5tha, 5tkj, 5tkk, 5tln, 5tp0, 5tpb, 5tpc, 5tq1, 5tqs, 5ttf, 5ttg, 5ttw, 5twg, 5twh, 5tyi, 5tzo, 5u06, 5u0f, 5u1q, 5u2j, 5u66, 5u6k, 5ufc, 5uff, 5umz, 5un1, 5unj, 5ur1, 5uw5, 5uwi, 5uwj, 5uwp, 5v1d, 5v1y, 5v2p, 5v2q, 5v3r, 5v4b, 5v6y, 5va9, 5vb9, 5vdk, 5vk0, 5vkm, 5vlh, 5vlk, 5vll, 5vlp, 5vnb, 5vqi, 5vtb, 5vzu, 5vzy, 5w0l, 5w0q, 5w38, 5w4e, 5w5s, 5w5u, 5w6i, 5w6r, 5w6t, 5w6u, 5w7i, 5w7j, 5w7x, 5w94, 5wa1, 5wa4, 5wbk, 5wbl, 5wei, 5wg8, 5wgd, 5wgq, 5wir, 5wkl, 5wkm, 5wle, 5wqd, 5wtt, 5wxf, 5wxg, 5wxh, 5wxo, 5wxp, 5wyr, 5x72, 5xgl, 5xhs, 5xhz, 5xo2, 5xof, 5xs8, 5xup, 5xvw, 5xwr, 5xxf, 5xxk, 5xyf, 5y1u, 5y20, 5y21, 5y53, 5y59, 5y5u, 5y5w, 5y6k, 5y7w, 5y97, 5yba, 5yc1, 5yc2, 5yc3, 5yc4, 5yco, 5ygd, 5ygf, 5yjy, 5ypo, 5ypp, 5ypw, 5yqw, 5yto, 5ytu, 5yv5, 5yvx, 5yy4, 5yy9, 5yyz, 5yzd, 5z89, 5z95, 5zbz, 5zia, 5zjy, 5zjz, 5zk5, 5zk7, 5zk9, 5zml, 5znp, 5znr, 5zoo, 5zop, 5zuj, 6a30, 6a5e, 6a6w, 6a80, 6a8g, 6a8n, 6a9c, 6a9o, 6abp, 6aox, 6apr, 6apu, 6ar2, 6asz, 6at0, 6au5, 6ax4, 6axj, 6axk, 6axp, 6ayh, 6ayn, 6azk, 6azl, 6b27, 6b2c, 6b5m, 6b5o, 6b5r, 6b5t, 6b67, 6bcr, 6bcy, 6bd1, 6bgg, 6bhd, 6bhe, 6bhh, 6bhi, 6bij, 6bil, 6bin, 6bir, 6biv, 6bix, 6biy, 6biz, 6bj2, 6bmi, 6bnt, 6buu, 6bvb, 6bvh, 6bw3, 6bw4, 6byk, 6c4u, 6c5h, 6c5j, 6c5k, 6c9n, 6c9p, 6c9r, 6cb5, 6cct, 6ccu, 6cd8, 6cdg, 6cdm, 6cdo, 6cdp, 6cer, 6cf6, 6cgt, 6cjv, 6co4, 6d07, 6d08, 6d1u, 6d3x, 6d3y, 6d3z, 6d40, 6d4o, 6d6t, 6df1, 6df2, 6do5, 6drt, 6dub, 6e49, 6e5x, 6e8k, 6e8m, 6ecz, 6eeh, 6egw, 6eiz, 6ek3, 6em6, 6em7, 6ema, 6eo0, 6epy, 6epz, 6eq1, 6eqv, 6eqw, 6eqx, 6er3, 6eru, 6esa, 6evp, 6eww, 6ex0, 6ezi, 6f08, 6f09, 6f55, 6f5m, 6f5u, 6f6d, 6f6i, 6f6n, 6f6s, 6f7t, 6f8g, 6fam, 6fau, 6fav, 6faw, 6fbw, 6fby, 6fc6, 6fel, 6fhu, 6fi4, 6fi5, 6fiv, 6fkp, 6fkq, 6fky, 6fkz, 6flg, 6fmn, 6fn9, 6fpu, 6fsd, 6fse, 6fu1, 6fvn, 6fx1, 6fzf, 6fzj, 6fzp, 6g0q, 6g15, 6g2n, 6g47, 6g6x, 6g84, 6g85, 6g86, 6g8i, 6g8j, 6g8k, 6g8l, 6g8p, 6g8q, 6g9b, 6g9i, 6gfx, 6ggb, 6gjj, 6gw1, 6gwe, 6gxe, 6gzl, 6h0b, 6h41, 6h4o, 6h4p, 6h4r, 6h4s, 6h4u, 6h4v, 6h4w, 6h4x, 6h4y, 6h50, 6h51, 6h52, 6h5w, 6h7b, 6h8c, 6h96, 6h9v, 6hck, 6hcu, 6hhp, 6hks, 6hlb, 6hld, 6hle, 6hm4, 6hmg, 6hmt, 6hoi, 6hol, 6hpg, 6hro, 6hs4, 6hv2, 6hy7, 6hza, 6hzb, 6hzc, 6hzd, 6hzx, 6i41, 6i4x, 6i5j, 6i5n, 6i5p, 6i68, 6i7a, 6iae, 6iam, 6idg, 6iiw, 6im4, 6inz, 6iqg, 6iso, 6j9w, 6j9y, 6jad, 6jag, 6jam, 6jan, 6jao, 6jap, 6jax, 6jb0, 6jb4, 6jbb, 6jjm, 6jjn, 6jjz, 6k2n, 6k5r, 6k5t, 6kdi, 6kmj, 6md6,

6me1, 6mil, 6mim, 6min, 6miq, 6mle, 6mlo, 6mm5, 6mnf, 6mqc, 6mqe, 6mqm, 6msy, 6mtv, 6mu3, 6mub, 6n19, 6n3e, 6n3f, 6n5x, 6n7q, 6n87, 6n93, 6nao, 6ncp, 6njz, 6nk0, 6nk1, 6nkp, 6nsx, 6nxz, 6ny0, 6o21, 6o3w, 6o3x, 6o3y, 6o7g, 6oie, 6om2, 6om4, 6oxl, 6p3w, 6p7p, 6p7q, 6pek, 6peu, 6phx, 6pi7, 6pit, 6prg, 6pxc, 6q38, 6q4q, 6q9t, 6q9w, 6qc0, 6qcg, 6qdx, 6qk8, 6qpl, 6qs1, 6qsz, 6qtm, 6qto, 6qtq, 6qtr, 6qts, 6qtw, 6qtx, 6qzr, 6r0x, 6r8i, 6rhe, 6rk4, 6rml, 6rr0, 6s07, 6sen, 6sq0, 6tyz, 6u5m, 6uyx, 6uyy, 6uyz, 6v1c, 7abp, 7kme, 8abp, 9abp, 9icd

Chapter 3

Energetics of Non-covalent Interactions of Protein–Ligand Complexes for Drug Discovery

‡Yingze Wang and Dong Jun Shin contributed equally. Yingze Wang developed the ligand-fragmentation workflow, analyzed the EDA results, performed the auxiliary OMol full-pocket analysis, and benchmarked existing force fields.

3.1 Introduction

Accurate prediction of protein-ligand binding affinities has been the central task of computer-aided drug design (CADD), requiring models for noncovalent interactions (NCIs) to accelerate the chemical design of drug candidates for protein targets. There is an admirable history of large community efforts on describing NCIs using two-body force fields (FFs), and persistent efforts to improve their accuracy in binding affinities is ongoing.[142, 143] But their limited functional forms can lead to poor pose predictions and/or protein-ligand binding energy interactions that incur average errors of $1.5 \sim 2$ kcal/mol[144, 145] in practical drug discovery campaigns.[146, 147] Hence, many-body FFs such as CHARMM-Drude[143] and AMOEBA[17] have taken hold in the biomolecular simulation community for proteins and nucleic acids[148], but are still relatively underdeveloped compared to fixed charge models. In part this is a limitation in the functional form for the many-body terms, but just as importantly the creation of enough diverse ab initio data for the piecewise description of NCIs to parameterize both types of FF models.

We have advocated for new approaches to FFs by using energy decomposition analysis

[1]Reproduced with permission from: Energetics of Non-covalent Interactions of Protein–Ligand Complexes for Drug Discovery. Y. Wang‡, D. J. Shin‡, M. Head-Gordon, and T. Head-Gordon. *J. Chem. Inf. Model.* 2025, 65, 21, 12058–12067.

(EDA) of the QM energy and forces for non-bonded interactions.[19, 149, 150] The EDA approach is itself a chemical model that decomposes the energy (and forces) into the following terms

$$E_{\text{int}} = E_{\text{elec}} + E_{\text{Pauli}} + E_{\text{disp}} + E_{\text{pol}} + E_{\text{CT}} \tag{S2.1}$$

where E_{elec} and E_{Pauli} are predominately 2-body terms describing permanent electrostatics and Pauli repulsion, while E_{disp}, E_{pol}, and E_{CT} correspond to the many contributions from dispersion, polarization, and charge transfer, respectively.[68] We have used the EDA approach to create new water force fields of high accuracy such as MB-UCB[151, 152] and the Completely Multipolar Model (CMM)[153, 154]. We find good agreement with the Absolutely localized molecular orbital (ALMO)-EDA breakdown term-by-term for water dimers and trimers that leads to excellent reproduction of large water cluster benchmarks and spectroscopic observables. We note that the EDA approach has also been used to guide protein force field development for polarizable models.[155]

Several datasets of quantum-mechanical (QM) noncovalent interaction (NCI) energies have been reported,[156–160], including some that also use an EDA scheme. For instance, Donchev et al. compiled DES370K[157], comprising ∼370K dimer geometries with CCSD(T)-level interaction energies and SAPT0 energy decomposition for generic organic systems. However, its chemical diversity is relatively limited, as the dataset comprises only approximate 3,700 unique molecular pairs and is largely restricted to small fragments. Spronk et al. introduced Splinter[156], which offers substantially greater conformational and chemical coverage (1.7M geometries across >9,000 distinct dimers); however, this space is generated from only 332 manually selected monomers for protein and drug molecule fragments. Moreover, the reference level used in Splinter (SAPT0) is generally less reliable than higher-level DFT benchmarks and can incur sizable errors, with deviations that may exceed 10 kcal/mol in challenging cases. As a result, Splinter is well suited for large-scale pretraining, but is less appropriate for rigorous validation. Details of these data sets are provided in Supplementary Table S1.

In this work we have formulated a large and diverse NCI dataset for protein-ligand chemistry aiming to simultaneously achieve the "sweet spot" of (i) high-quality, experimentally derived structures, (ii) broad chemical diversity with real-world protein-ligand interactions, (iii) high-accuracy electronic-structure reference levels, and (iv) physically meaningful energy decomposition for each NCI type using ALMO-EDA[68, 161]. We first created a high-quality protein-ligand dataset derived from the HiQBind[15] workflow operating on the PDBBind v2020[162] and BioLiP2[163, 164] data to create experimentally relevant protein-ligand complexes. We then used Protein Ligand Interaction Profiler (PLIP)[165–167] to break down the complex structure into 6 types of NCIs: hydrophobic, hydrogen bonding, cation-π, $\pi - \pi$ stacking, salt bridges, and halogen bonds. Hence each full protein-ligand structure is fragmented to create dimer substructures between a protein fragment and a ligand fragment, resulting in a total of 14,905 dimer complexes and organized into data subsets for each NCI type. For each NCI type we further create a meta data set by performing an energy decomposition analysis (EDA)[19, 149] using the absolutely localized molecular

orbital (ALMO)-EDA[19] to decompose the high-level DFT energy (ωB97X-V/def2-TZVPD) into the terms described in Eq. (1). We believe that this type of meta data is better suited to force field development or for machine learning training of interaction potentials than simply the amalgamated data of a single energy value. When presented as ternary diagrams that visualize and separate Pauli, permanent electrostatics, and dispersion from polarization and charge transfer, we find that salt bridges, pi-stacking, and hydrophobic NCIs should be well-captured by fixed charge models, whereas pi-cation, halogen bonding, and hydrogen bonding would ideally require formulation of many-body polarization and charge transfer FF terms. Finally we assess the errors in reproducing energies in all NCI categories for several popular physical force fields and machine learned interaction potentials.

3.2 Methods

3.2.1 Construction of Protein-Ligand Interacting Fragments

We first generated a high-quality protein-ligand complex structure dataset by processing PDBBind v2020[162] and BioLiP2[163, 164] with the HiQBind workflow[15]. This workflow first employed a set of filters to remove structures with steric clashes (heavy atoms < 2Å), rare elements (elements other than H, C, N, O, F, S, P, Cl, Br, I), inorganic binders (number of heavy atoms smaller than 4) and covalent binders. Then, two structure fixing modules are applied to both the ligand and the protein to assign correct bond order and protonation states under physiological pH as well as to add atoms that are missing from experiment due to insufficient electron density. For details of the HiQBind workflow, please refer to ref [15].

Next, we analyzed the types of protein-ligand interactions using Protein Ligand Interaction Profiler (PLIP)[165–167]. PLIP annotated the complex structure with 8 types of interactions: hydrophobic, hydrogen bond, cation-π, $\pi - \pi$ stacking, salt bridge and halogen bond, metal bond and water bridges. This project focuses on the first six of them and does not consider the latter two NCIs. In the same protein-ligand complex, interactions under the same NCI type were merged if the marked interacting atoms overlap (mainly for fused aromatic rings). This prevented generating duplicate fragment dimers. At the next step we devise an algorithm to fragment the protein-ligand complex into protein-ligand dimers that capture a particular NCI type as shown in Figure 3.1. Our algorithm fragments the protein-ligand complex based on functional groups (FGs) inspired by the work of Ji et al.[168] The fragmentation should follow the rationale that it keeps the minimal FG motifs that are most relevant to protein-ligand binding. Additionally, it should also include any neighboring FGs that have a potential electronic structure effect on the interacting center. For each ligand, we extracted FGs with the following definition:

- Rings: For each N-member ring ($N \leq 8$ and defined by the smallest set of smallest rings convention) in a molecule, we define it as an FG if the ring is aromatic or $N \leq 4$ or containing less than 3 sp^3 carbons.

- Non-cyclic FGs: For the non-cyclic part of a molecule, we first mark all functional atoms, including heteroatoms, carbons in non-aromatic double or triple bonds, and sp^3 carbons connected to two or more hetero atoms (oxygens, nitrogens, sulfurs and halogens). Then, marked functional atoms that are connected with chemical bonds are merged as FGs.
- sp^3 carbons: Each carbon atom that does not belong to the first two types of FGs, i.e. a sp^3 carbon, is also defined as an FG to achieve a complete split of the molecule.

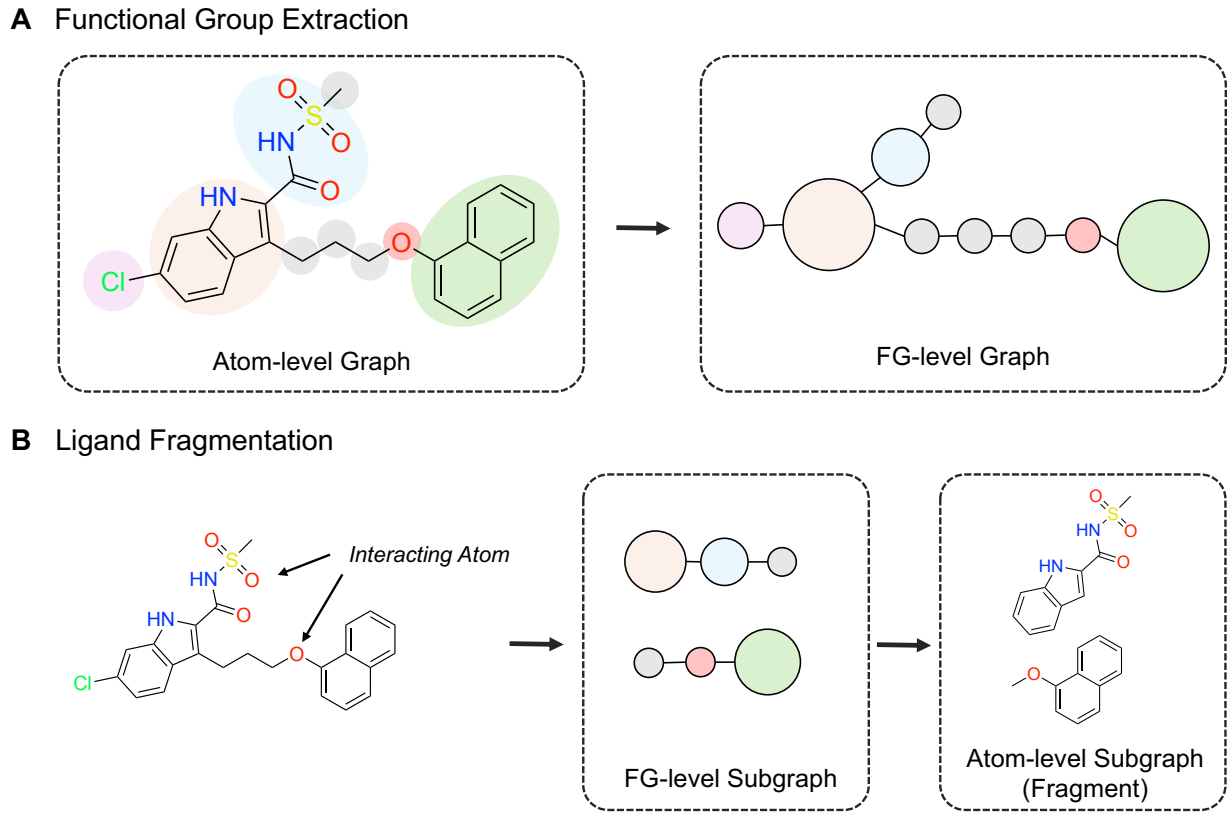

Figure 3.1: *An overview of the ligand fragmentation algorithm.* (A) Functional-group (FG) extraction: the ligand is represented as an atom-level graph and coarse-grained into an FG-level graph by identifying ring and non-cyclic functional motifs and residual sp^3 carbons as nodes. (B) Fragmentation: for each PLIP-annotated interacting heavy atom, the corresponding FG and its neighboring FGs are selected to form an FG-level subgraph, which is then mapped back to an atom-level substructure.

For each heavy atom of interest, the algorithm extracts the FG consisting of it plus its neighboring FGs, and each terminal open valance is capped with a methyl group if it

was originally connected with a carbon, otherwise capped with a hydrogen. For the protein fragmentation, if the interacting atom marked by PLIP is a backbone amide O or N (hydrogen bond donor/acceptor or halogen bond acceptor), N-methylacetamide (NMA) will be used as representative structures or N,N-dimethylacetamide (DMA) if the residue is a proline. If the interacting atom marked by PLIP locates on the side chain, relevant side chain motif will be selected as fragment (Figure S7). For each NCI type, after the all protein ligand fragment dimers are constructed, we calculated the SOAP descriptor using the ASAP package[169] and clustered them using K-Means algorithm implemented in scikit-learn when the number of generated fragment dimers are too large.

To assess the sensitivity of the ALMO-EDA results to the fragmentation scheme, we performed a validation study on randomly selected examples across all NCI classes by comparing the dimer fragments against minimally extended models retaining the full residue (ACE/NME-capped) and full ligand. As summarized in Supporting Information Tables S6–S7, the absolute interaction energies become somewhat more favorable in the extended models (because more interacting atoms are included), but the relative balance and qualitative trends among the ALMO-EDA components remain unchanged. These results support that our fragmentation scheme retains the essential local electronic structure features needed to capture the NCI profile of the protein–ligand environment.

3.2.2 Protein-Ligand Interacting Fragments by NCI Type

While small gas-phase protein–ligand fragment dimers cannot fully reproduce the electrostatic, hydrogen-bonding, and dielectric environment of a complete and intact protein active site, the scope of this work is to provide controlled, high-quality fragment-level training and validation data for next-generation force fields and MLIPs. In this setting, the fragment dimers are the leading order term in the many-body expansion, and the fragments are designed to isolate individual NCIs and investigate their intrinsic physical energetic origins using ALMO-EDA, which also provides quantitative guidance for how these NCIs can be studied and tuned.

Hydrogen-bonding. Among all the hydrogen-bond (H-bond) interactions identified by PLIP, we generated a set of 5,000 protein–ligand fragment dimers using the fragmentation and clustering protocol described above. ALMO-EDA calculations were performed on all dimers, of which 4,977 completed successfully. The failed cases were primarily due to unsuccessful SCF convergence under the default Q-Chem criteria. Another 394 dimers exhibited positive total interaction energies because the interacting atoms are positioned too close to one another, which is an artifact of extracting geometries from experimental structures, where atomic positions are mostly refined against X-ray electron densities rather than quantum-mechanical energetics, occasionally yielding repulsive local configurations. Among the remaining 4,583 attractive dimers, we also found 281 of them are dimers with oppositely charged monomers, and hence they are treated and discussed separately as salt bridges. This resulted in a total of 4302 hydrogen-bonded dimers used for EDA analysis in the next section.

Salt bridges. A total of 12,678 salt bridge interactions were identified by PLIP and 8,364

remained after merging duplications and removing metal ion-meditated interactions. 2000 representative salt bridge dimers are yielded by the clustering and underwent ALMO-EDA calculation. 1976 dimers successfully converged and 5 of them are discarded because of net repulsion. These structures contained ligand fragments that were manually verified not to represent typical salt bridges: three were be falsely identified as ions by PLIP, and two correspond to zwitterionic species that contain both cationic and anionic centers but carry an overall charge with the same sign as the participating protein fragment. Additionally, 281 hydrogen bond dimers with oppositely charged fragments were reclassified as salt bridges due to their similarity in nature. Hence a total of 2252 dimers were classified as salt bridges for EDA analysis.

Halogen bonding. Our procedure identified 1064 halogen bonding with pattern $\mathrm{C-X\cdots Y}$, with chlorine being the most prevalent participant (501 dimers), followed by fluorine (376 dimers), bromine (110 dimers) and iodine (77 dimers).

Pi Stacking. 8,663 π-stacking interactions are identified by PLIP in total and 7,101 distinct protein ligand fragment dimers were generated by the fragmentation protocol. We also discard dimers involving charges, as such structures are dominated by electrostatics and exhibit π-cation like behavior. This gave rise to 6,578 dimers, of which 1000 clusters underwent ALMO-EDA calculation. However, a post analysis of these 1000 dimers revealed that 22 of them are with one non-aromatic fragment, due to the fact that PLIP falsely flags certain geometrically planar ring systems like imidazolidine-2,4-dione (PDB ID: 5OH7) as aromatic. Among the remaining 978 structures, 12 of them exhibit net repulsion due to compressed geometry and were discarded, yielding 966 structures for the following analysis.

Pi Cation. A total of 2280 pi-cation interactions were identified by PLIP. Double protonated histidines were then discarded since they should be mostly neutral at physiological pH, let alone in the presence of a cationic ligand. Similarly, we also removed non-aromatic and non-cationic ligand fragments that were falsely flagged as π-cation interaction by PLIP. These were due to a technical fact that PLIP internally uses OpenBabel to guess the ligand bond order as well as charge state, occasionally leading to a false interpretation but HiQBind applied a strict criteria to assign correct bond orders. Finally, 1520 dimers were generated and underwent ALMO-EDA calculations. A total of 1505 π-cation dimers exhibiting net attraction were then analyzed.

Hydrophobic. PLIP identified 86,267 hydrophobic contacts across the dataset. After the fragmentation, we also discarded fragments with a net charge and dimers that involve in other NCI types, since the hydrophobic interactions may locally co-exist with other NCI types and fragmentation including neighboring motifs would potentially bring them in and mess up the energy profile. Finally, this additional filter yielded 78,758 dimers and 5000 dimers were selected with the clustering protocol described above, of which 184 dimers exhibited positive total interaction energies and were thus eliminated, resulting in 4816 hydrophobic dimers used for the EDA analysis.

3.2.3 Energy calculations and Energy Decomposition Analysis

All energies calculations were performed on all dimers with the Q-Chem[170] software package (version 6.0) at the ωB97X-V/def2-TZVPD level of theory, a range-separated hybrid Generalized Gradient Approximation (hybrid-GGA) functional with VV10 nonlocal correlation[171]. The exchange-correlation potential was evaluated on a fine Lebedev quadrature using 99 radial and 590 angular grid points (99, 590) while a smaller (50, 195) grid was used for evaluating the VV10 functional. This level of theory is a good balance between accuracy and efficiency, and has been successfully applied to the development of force fields[151–154, 172] and studies of intermolecular interactions[173, 174]. This is supported by a recent protein–ligand benchmark[174] in which ωB97X-V achieves a mean absolute error of 0.35 kcal/mol relative to DLPNO-CCSD(T) across a representative collection of $\sim$ 200 protein–pocket+ligand complexes. This dataset contains substantially larger structures than the fragment dimers considered here, suggesting that the expected error in our smaller, more localized systems is likely even lower than 0.35 kcal/mol. Therefore we believe the level of DFT theory used in this work is reliable reference without introducing systematic bias across NCI types.

We used the energy decomposition analysis (EDA) based on absolutely localized molecular-orbitals (ALMO) scheme for Eq. (1).[68, 161, 175, 176] ΔE_{elec}(Electrostatics) describes permanent electrostatics via a classical Coulombic interaction of the total charge distributions of the interacting fragments. ΔE_{Pauli} (Pauli repulsion) arises from the Pauli exclusion principle of electrons and captures the energetic penalty of abiding by wave function antisymmetry between fragments. ΔE_{disp} (Dispersion) is an attractive interaction due to correlated fluctuations of electrons. ΔE_{pol} (Polarization) describes energy lowering due to the distortion of the electron density localized on each molecule in the presence of other molecules. ΔE_{ct} (Charge transfer) is the energy lowering associated with orbital mixing across different fragments. Absolute energies of each term are stable with basis set size as ALMO-EDA is designed to approach a useful and well-defined complete basis set (CBS) limit.[68, 161, 175, 176]

Energy components of the decomposed energies from ALMO-EDA were analyzed on ternary plots, [177] similarly to previous work[173]. The position on the ternary plot represents the ratio of each ALMO-EDA component X

$$\sigma_X = \frac{\Delta E_X}{|\Delta E_{\text{disp}}| + |\Delta E_{\text{frz}}| + |\Delta E_{\text{orb}}|} \tag{S2.2}$$

where $\Delta E_{\text{frz}} = \Delta E_{\text{Pauli}} + \Delta E_{\text{elec}}$ is defined across two vertices to distinguish stabilizing and destabilizing interaction energies whereas ΔE_{disp} and $\Delta E_{\text{orb}} = \Delta E_{\text{pol}} + \Delta E_{\text{CT}}$ are strictly stabilizing.

3.2.4 Benchmarking Force Fields and MLIPs

The GAFF2 parameters were assigned with `acpype` package[178] with both AM1-BCC charge method[179] and RESP charges derived under ωB97X-V/def2-TZVPD using Q-Chem. The

GAFF2 calculation were performed using OpenMM. AMOEBA parameters were assigned using Poltype2[180], where atomic multipoles are calculated with Psi4 and GDMA under ωB97X-V/def2-TZVPD level of theory. AMOEBA calculations were performed using Tinker v8.11. MACE-OFF-23(M) and MACE-OMOL calculations were performed using the `MACE` package v0.3.14[181]. UMA(S) model is downloaded from huggingface[182] (name uma-s-1.1) and the calculation is performed using the `fairchem` package v2.13.10[183] with `task_name` set to `"omol"`. The AIMNet2 model and calculations were performed using the `aimnet2` package[184].

3.3 Results

Table 3.1 presents the average as well as median interaction energy by NCI type, roughly organized by row into the strongest to weakest of protein-ligand interactions. As these averages are over many different subclasses within each NCI category, below we consider each specific type of NCI and their EDA components in more detail. This is useful for understanding how to tie certain atom types to EDA components in force field development or to featurize input representations in creating machine learning force fields.

Table 3.1: *Non-covalent interactions energies by type and their ALMO-EDA components.* Average and (median) energies in kJ/mol. NCI categories are ordered from the strongest to the weakest interactions.

Energy (kJ/mol)	Salt-Bridge	H-Bond charge	H-bond neutral	Pi-Cation
Total	−527.16 (−444.28)	−76.43 (−73.14)	−23.85 (−23.41)	−41.13 (−38.67)
Pauli	78.30 (66.02)	54.60 (42.62)	40.93 (36.12)	20.25 (16.70)
Electrostatics	−507.24 (−432.55)	−70.43 (−68.06)	−32.97 (−31.57)	−21.90 (−19.63)
Dispersion	−19.20 (−18.67)	−17.52 (−16.90)	−16.69 (−17.52)	−19.14 (−18.77)
Polarization	−57.53 (−54.54)	−30.84 (−27.58)	−7.82 (−6.95)	−16.70 (−14.47)
Charge Transfer	−21.49 (−18.85)	−12.24 (−9.19)	−7.29 (−5.99)	−3.64 (−2.96)
Number of data	2252	1872	2430	1505
Energy (kJ/mol)	Pi-stack Parallel	Pi-stack T-shaped	Hydrophobic	
Total	−22.54 (−21.91)	−13.97 (−12.62)	−11.05 (−9.38)	
Pauli	28.61 (27.28)	20.32 (17.62)	19.48 (16.52)	
Electrostatics	−10.17 (−9.55)	−9.07 (−7.74)	−7.18 (−5.71)	
Dispersion	−35.23 (−34.41)	−20.29 (−19.45)	−19.60 (−17.24)	
Polarization	−2.55 (−2.34)	−2.05 (−1.61)	−1.37 (−1.03)	
Charge Transfer	−3.19 (−3.01)	−2.87 (−2.53)	−2.38 (−2.06)	
Number of data	527	439	4816	
Energy (kJ/mol)	Halogen-F	Halogen-Cl	Halogen-Br	Halogen-I
Total	−5.95 (−3.12)	−5.80 (−4.74)	−10.01 (−8.15)	−12.72 (−12.40)
Pauli	7.85 (4.79)	9.83 (7.22)	13.56 (11.12)	19.80 (12.64)
Electrostatics	−3.03 (−0.17)	−3.63 (−2.18)	−8.09 (−6.42)	−13.09 (−9.92)
Dispersion	−7.76 (−6.86)	−8.75 (−8.05)	−9.79 (−9.52)	−12.65 (−11.55)
Polarization	−1.92 (−0.72)	−1.86 (−1.18)	−3.59 (−2.25)	−3.78 (−2.75)
Charge Transfer	−1.10 (−0.65)	−1.39 (−1.03)	−2.10 (−1.73)	−2.99 (−1.94)
Number of data	376	501	110	77

3.3.1 Electrostatically dominated NCIs

Salt bridge energetics are the strongest NCIs and are dominated by favorable permanent electrostatics as evident in Table 3.1. Separating the dimers by the location of the positive charge, the average interaction energy is lower by -173 kJ/mol for the 1651 dimers with a positively charged protein fragment with a negatively charged ligand, as opposed to the 601 dimers where the positive charge is carried instead by the ligand as shown in Table 3.2 and Figure 3.2 (also see Supplementary Figure S1a). We also investigate the difference between different anion types, classifying the total dataset into carboxylate, phosphate and sulfate based on the element that the negatively charged center atom is connected to. As shown in Table 3.2 and Supplementary Figure S1b, phosphate groups show nearly twice the average stabilization energy from permanent electrostatics, polarization, and charge transfer interactions in magnitude, in part due to chains of phosphates on the ligand fragments which increase the total negative charge. Overall protein-ligand salt-bridge NCIs should in principle

be well described by fixed charge force fields.

The 4,302 hydrogen-bond dimers are comprised of 2,430 neutral clusters whose average interaction energy of –23.85 kJ/mol is not particularly exceptional, whereas the 1872 charged fragments due to ligand interactions with amino acid residues such as lysine, aspartic and glutamic acid, etc. increase the average total interaction energy to –76.43 kJ/mol (Table 3.1).

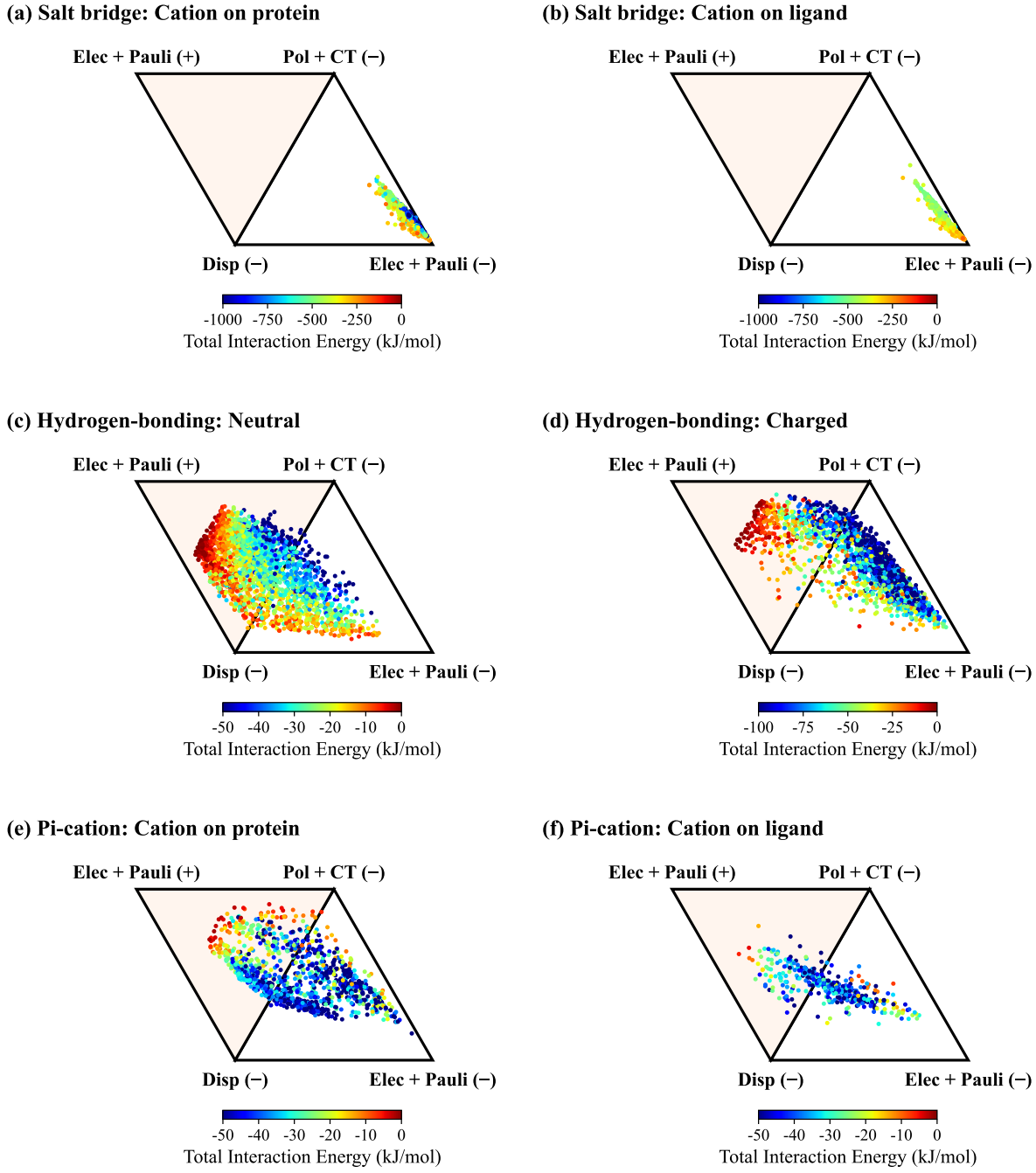

Figure 3.2: *Ternary plots of protein-ligand interactions for strongly interacting NCIs.* (top) Salt Bridges, where position of the cation charge is found on (a) the protein and (b) the ligand. (middle) hydrogen bond interactions for (c) neutral and (d) charged hydrogen bonded complexes. (bottom) π-cation interactions where the cation is on the (e) protein or (f) ligand.

Hence hydrogen-bonding intermolecular energies span a broad range of interaction strengths, from –254.7 kJ/mol to –0.01 kJ/mol as seen in the violin plots in Supplementary Figure S2. However for either neutral or charged hydrogen-bonded complexes, polarization plays a significant role beyond just permanent electrostatics and dispersion as seen in Figure 3.2(c,d). This highlights the necessity to explicitly model the polarization and charge transfer in molecular mechanics force fields.

Table 3.2: *Salt-Bridge interaction energies and their ALMO-EDA components.* Average and (median) energies in kJ/mol. Protein and Ligand refer to the fragment which contains the cation. We also consider the energetics of three types of oxyanions on the ligand fragment.

	All	Protein	Ligand
Total	−527.16 (−444.28)	−573.63 (−462.20)	−399.50 (−406.97)
Pauli	78.30 (66.02)	80.27 (67.74)	72.91 (60.04)
Electrostatics	−507.24 (−432.55)	−550.81 (−453.64)	−387.56 (−387.69)
Dispersion	−19.20 (−18.67)	−19.57 (−18.65)	−18.17 (−18.71)
Polarization	−57.53 (−54.54)	−61.46 (−56.07)	−46.73 (−46.29)
Charge Transfer	−21.49 (−18.85)	−22.05 (−19.39)	−19.95 (−15.67)
Number of data	2252	1651	601
	Carboxylate	Phosphate	Sulfate
Total	−404.28 (−407.02)	−827.65 (−785.66)	−360.32 (−360.27)
Pauli	69.94 (57.34)	101.28 (87.75)	46.31 (36.00)
Electrostatics	−391.33 (−390.12)	−791.89 (−754.80)	−340.30 (−334.31)
Dispersion	−17.42 (−17.26)	−23.57 (−22.44)	−16.59 (−15.42)
Polarization	−46.55 (−47.35)	−84.91 (−83.14)	−38.28 (−39.33)
Charge Transfer	−18.92 (−16.26)	−28.57 (−27.59)	−11.45 (−8.93)
Number of data	1509	662	81

Table 3.1 also summarizes the interaction energies and their ALMO-EDA components for the 1505 π-cation protein-ligand dimers. Pi-cation complexes are found to be predominately stabilized by permanent electrostatics, and while still strongly supported by dispersion forces, polarization and charge transfer energies play a much larger role (especially when the cation is on the protein), thus making it more difficult for fixed charge models to adequately capture this type of NCI as seen in Figure 3.2(e,f) and Supplementary Figure S3. Table 3.3 also considers the protein π-fragment residues TYR, PHE, TRP, HIS interacting with a cationic ligand fragment.[185] It appears that ligand design could benefit from the stronger electrostatics and polarization components of TRP, followed by TYR and PHE, making them more desirable targets compared to HIS which has significantly weaker π-cation interactions.

Table 3.3: *π-cation interaction energies of aromatic residues with a ligand cation fragment and their ALMO-EDA components.* Average and (median) energies in kJ/mol for TYR, PHE, TRP and HIS.

	TYR	PHE	TRP	HIS
Total	−35.29 (−36.25)	−34.44 (−34.78)	−46.94 (−48.74)	−22.52 (−19.68)
Pauli	20.82 (19.62)	24.51 (20.90)	25.74 (23.85)	16.06 (8.65)
Electrostatics	−21.57 (−21.99)	−22.91 (−21.89)	−30.20 (−30.62)	−13.63 (−9.80)
Dispersion	−18.67 (−18.96)	−19.72 (−19.03)	−22.90 (−22.26)	−12.70 (−12.72)
Polarization	−12.17 (−12.30)	−12.18 (−11.81)	−14.91 (−14.55)	−9.00 (−7.03)
Charge Transfer	−3.70 (−3.59)	−4.14 (−3.93)	−4.67 (−4.17)	−3.25 (−1.73)
Number of data	147	61	171	38

3.3.2 Pi-stacking NCIs are dominated by dispersion forces

Table 3.1 summarizes the interaction energies and their ALMO-EDA components for 966 $\pi - \pi$ protein-ligand dimers for parallel and T-shaped geometries. These NCIs are ubiquitous in proteins as the strength of their energetics are comparable to hydrogen-bonding while also being directional through the orientation of their π electron systems, but have not hitherto been carefully analyzed for protein-ligand complexes.

Until recently, the Hunter-Sanders (HS) model had long been the principle paradigm for interpreting the structure of many protein $\pi - \pi$ complexes. Under perfect co-facial stacking arrangements which maximize dispersion forces, the HS model assumes that the repulsive electrostatic quadrupole moments of the interacting pi systems will instead favor offset stacking to retain most of the attractive dispersion while reducing quadrupolar repulsions. In previous work we used ALMO-EDA to support an alternative interpretative model by Carter-Fenk and Herbert (CFH) that off-stack geometries in proteins arise from the competition between Pauli repulsion and dispersion, and furthermore typically show favorable electrostatic interactions[173].

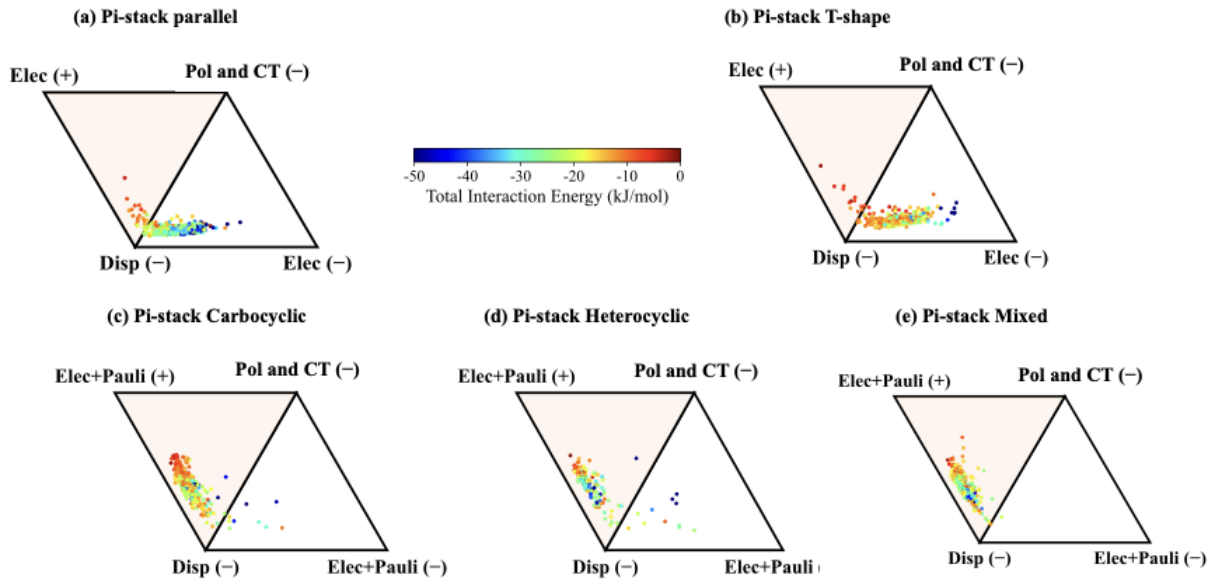

Figure 3.3: *Ternary plots of protein-ligand interactions for $\pi-$stacking NCIs.* We first consider electrostatics, dispersion, polarization and charge transfer for (a) parallel and (b) T-shaped π-stacking complexes. (c-e) The introduction of Pauli shows that Pauli repulsion weakens π-stacking in general, but introduction of heteroatoms (i.e. non-carbon) on (d) both or (e) one of the aromatic rings strengthens π-stacking energetics.

Figure 3.3(a,b) show that pi-stacking in protein-ligand systems is a dispersion dominated NCI supported by favorable permanent electrostatics, with parallel stacking gaining $\sim$ 8 kJ/mol more stability on average through geometries that better maximize dispersion compared to T-shaped geometries. The ternary diagrams shown in Figures 3.3(c-e) confirm that the inclusion of Pauli is the source of repulsive interactions in π-stacking as predicted by the CFH model. Introducing heteroatoms into the aromatic rings is a common approach to modulate the $\pi - \pi$ interaction strength in drug or material discovery, as shown by the number of examples found from the PDB and tabulated in Table 3.4. Less than half of of the identified $\pi - \pi$ interaction happens between aromatic rings composed exclusively of carbon atoms ("carbocyclic" in Fig. 3.3c), while introducing heteroatoms into the ring systems stabilizes the $\pi - \pi$ interaction by $\sim$5 kJ/mol on average, mostly driven by stronger electrostatics of the introduces permanent dipoles. Taken together, these results indicate that $\pi - \pi$ interactions for protein-ligand interactions are governed by a balance between geometry-dependent dispersion and chemistry-dependent electrostatic and induction effects, with heteroatom incorporation providing an additional, non-negligible stabilization mechanism beyond dispersion alone.

Table 3.4: Pi–Stack interaction energies and their ALMO-EDA components in kJ/mol. Average (mean) energies out of parentheses and median in parentheses.

	All	Parallel	T-Shaped
Total	−18.64 (−17.31)	−22.54 (−21.91)	−13.97 (−12.62)
Pauli	24.84 (22.60)	28.61 (27.28)	20.32 (17.62)
Electrostatics	−9.67 (−8.79)	−10.17 (−9.55)	−9.07 (−7.74)
Dispersion	−28.44 (−27.26)	−35.23 (−34.41)	−20.29 (−19.45)
Polarization	−2.33 (−2.03)	−2.55 (−2.34)	−2.05 (−1.61)
Charge Transfer	−3.05 (−2.80)	−3.19 (−3.01)	−2.87 (−2.53)
Number of data	966	527	439
	Heterocyclic	Carbocyclic	Mixed
Total	−21.13 (−20.37)	−16.00 (−14.93)	−20.50 (−20.00)
Pauli	28.42 (27.88)	22.68 (19.95)	25.68 (24.18)
Electrostatics	−11.65 (−10.74)	−8.82 (−7.65)	−9.74 (−9.24)
Dispersion	−31.33 (−29.60)	−24.88 (−22.72)	−31.16 (−31.22)
Polarization	−3.18 (−2.92)	−2.07 (−1.75)	−2.22 (−2.00)
Charge Transfer	−3.39 (−3.20)	−2.91 (−2.64)	−3.05 (−2.88)
Number of data	168	423	375

3.3.3 Weakly interacting NCIs

The ALMO-EDA components presented in Table 3.1 shows that hydrophobic interactions are primarily driven by dispersion interactions, but moderate energy scales can be reached when aided with some electrostatic stabilization in the protein-ligand complex (Figure 3.4 and Supplementary Figures S5). The proportions of each ALMO-EDA component for hydrophobic interactions is very similar to the components of pi-stacking interactions, but the lack of aromaticity decreases the magnitude of the interaction by $\sim 60\%$.

Halogen bonding has realized accumulated importance as an NCI in drug discovery, illustrated by their use in the rational design of inhibitors targeting cathespin L[186] and HIV-1 reverse transcriptase[187]. The halogen bond is described as the attraction between an electrophilic region associated with a halogen atom (X) and a nucleophilic motif (Lewis base). As summarized in Table 3.1, the total interaction energy becomes more favorable down the halogen group, from weakly stabilizing F-bonds and Cl-bonds ($\sim$ 6 kJ/mol) to increasing strength involving I-bonds ($\sim$13 kJ/mol). As seen in Figure 3.4 and Supplementary Fig. S6, fluorine and chlorine halogen bonding is dominated by dispersion, which accounts for more than half of the attractive interaction. For bromine and iodine, the permanent electrostatics become larger, consistent with the fact that heavier atoms are less electron-negative and more polarizable covalently, resulting in a more flexible electron cloud and thus development of a stronger σ-hole under the affect of electron withdrawing group. Dispersion still remains a major stabilizing component for these two heavier halogens with $\sim 40\%$ contribution. The absolute strength of polarization and charge-transfer terms also grow very modestly with

halogen size, and this indicates that the experimentally derived halogen bonds are weak.

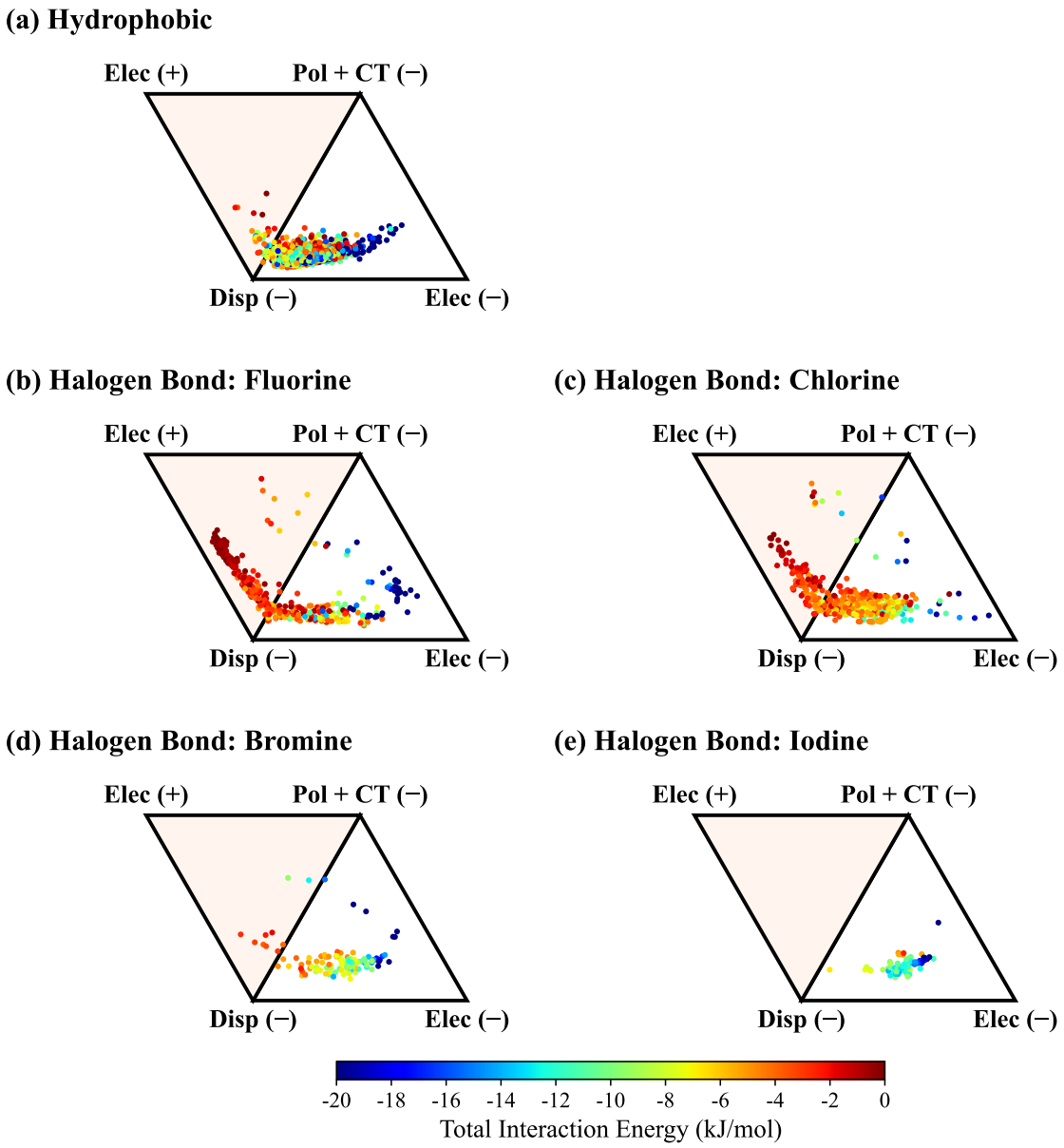

Figure 3.4: *Ternary plots of weakly interacting protein-ligand NCIs.* All weak interactions are dominated by dispersion as represented by (a) hydrophobic interactions and for halogen bonding involving (b) fluorine, (c) chlorine, dc) bromine, and (e) iodine.

We also noticed that the protein-ligand structures from the PDB database are not necessarily the optimal geometry of a halogen bond. As shown in Figure S10, angular distributions of $\angle(\mathrm{C}-\mathrm{X}\cdots\mathrm{Y})$ show significant deviations from the ideal 180°. Therefore, we also performed rigid potential energy surface (PES) scan between the halobenzene (PhX) and N-methylacetamide (NMA) to investigate more favorable geometries. NMA was used as the model compound because analysis of our dataset indicates that the most prevalent halogen bond acceptor is the protein backbone carbonyl oxygen and NMA serves as a minimal but chemically representative model of this motif. The scan was performed along the $r(\mathrm{O}-\mathrm{X})$ and $\angle(\mathrm{C}=\mathrm{O}\cdots\mathrm{X})$ for both T-shaped and co-planar geometries with the $\angle(\mathrm{C}-\mathrm{X}\cdots\mathrm{O})$ fixed at 180°. In the T-shaped geometry, the benzene ring and the amide are perpendicular to each other while in the co-planar geometry, they are on the same plane.

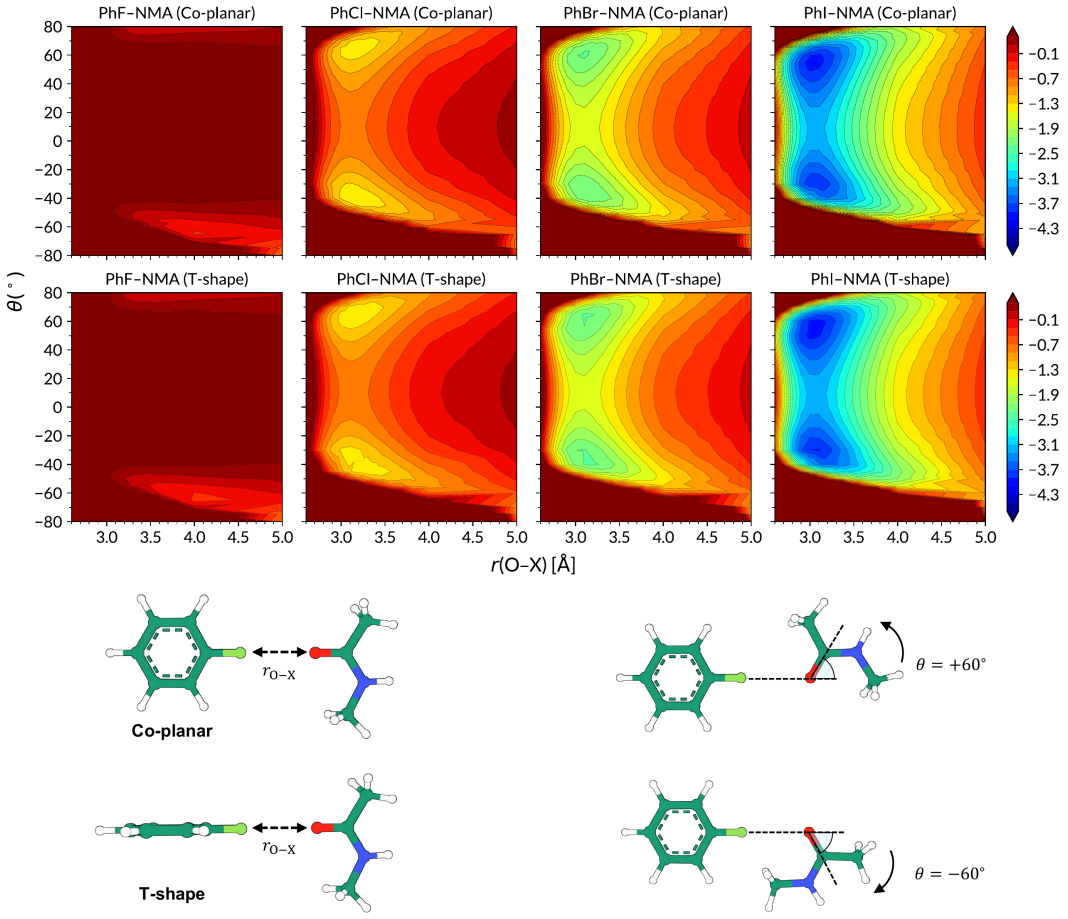

Figure 3.5: *Potential energy scan of an NMA fragment interacting with halogenated benzenes in both co-planar and T-shaped orientation.* Fluorine does not induce a classic sigma-hole type halogen bonds unlike chlorine, bromine, and iodine, but stabilization of halogen bonding still requires ideal geometries. The color is based on the total interaction energy in kcal/mol.

As shown in Figure 3.5, there are local minima on the PES except for PhF-NMA, showing that fluorine does not form a typical halogen bond by definition for any geometry. We still retained F-containing interactions in our halogen-bond analysis to maintain consistency with the PLIP-based annotation pipeline used to define NCI types across the dataset. In terms of the energy components, we found that while dispersion dominates by ∼5 kJ/mol over electrostatics in PhCl-NMA, and almost equally contributes in PhBr-NMA, electrostatics stabilizes the PhI-NMA complex by ∼10 kJ/mol than dispersion. These support the σ hole interpretation of halogen bonding under protein-ligand context.

3.3.4 Benchmarking Force Fields and MLIPs

Table 3.5 benchmarks two molecular mechanics (MM) force fields, including a fixed-charge model GAFF2[188] with both AM1-BCC[179] and ωB97X-V RESP charges, a polarizable force field AMOEBA[189], and four modern MLIPs, including MACE-OFF-23(M)[190], MACE-OMOL[181], UMA(S)[191] and AIMNet2[192], against the high quality DFT reference energies across six NCI types, revealing clear performance stratification tied to the underlying physical character of each interaction.

GAFF2 with ωB97X-V RESP charges performs slightly worse than AM1-BCC[179] charges for these NCI types. This may be because AM1-BCC was designed to reproduce HF/6-31G* electrostatic potentials, a level of theory which is known to overestimate molecular dipole moments, thereby partially implicitly accounting for some polarization effects.[179] In contrast, ωB97X-V RESP charges may provide a more accurate description of permanent electrostatics, but by removing this empirical error-cancellation they also make the absence of explicit polarization in GAFF2 more evident. Among MM models, AMOEBA consistently outperforms GAFF2 with either charge model, reflecting the importance of explicit polarization and higher-order electrostatics via atomic multipoles. Replacing AM1-BCC with RESP charges produces only modest and non-systematic changes across NCI classes and therefore does not alter this comparison, indicating that the performance gap is primarily associated with the force-field model rather than charge assignment alone. This advantage is most pronounced for electrostatically dominated interactions, including salt bridges, hydrogen bonds and π-cation interactions, where many-body polarization and charge-transfer effects are substantial.

In contrast, for dispersion-dominated NCIs such as $\pi-\pi$ stacking and hydrophobic contacts, fixed-charge force fields perform comparatively well, with GAFF2 errors approaching those of MLIPs, consistent with the secondary role of many-body effects in these interactions. The good performance of GAFF and AMOEBA on these dispersion dominated NCIs can also be explained by the fact that these models have been parameterized and extensively validated against experimental properties (density, heat of vaporization) of organic liquids[188, 189]. Even so, halogen-bonding interaction errors are high, and should be improvable under the fixed charge force field model since it is largely a dispersion dominant interaction.

AIMNet2 shows intermediate accuracy and is on par with the performance of the AMOEBA MM force field, which may simply reflect differences in the level of DFT used to train the MLIP. In contrast, MACE-OFF-23 performs more poorly than AMOEBA in all NCI categories. For strongly interacting NCIs dominated by electrostatic, most notably π-cation interactions, salt bridges and hydrogen bonds in charged environments, this deficiency can be traced to the absence of explicit charge-state modeling and hence is corrected in MACE-OMOL where charge states are embedded and properly represented in the OMol25 training dataset[193]. A detailed breakdown reveals that while MACE-OFF-23 achieves a modest MAE of 1.83 kJ/mol for neutral hydrogen bonds, the error increases dramatically to over 20 kJ/mol for charged hydrogen-bonded dimers (see Table 3.5).

Across MLIPs, UMA and MACE-OMOL exhibit the strongest and most uniform agreement

with the DFT benchmark, ranking best across nearly all NCI classes. But this has to be calibrated by the fact that both are trained on the large and chemically diverse OMOL25 dataset[193], that contains protein-ligand configurations derived from BioLiP and augmented by molecular dynamics sampling, while our benchmark is derived from HiQBind structures that are refined versions of BioLiP and PDBBind, and which uses the same level of DFT that we use here. As a result, there is likely meaningful overlap in the underlying chemical and structural space between the OMOL25 training set and the present benchmark, and therefore this evaluation should not be regarded as a strictly out-of-distribution test for OMOL-trained models. Our objective here is instead to assess whether current MLIPs can reproduce high-level QM interaction energies for realistic protein–ligand NCIs, even in the in-distribution regime relevant to practical applications. Overall, the ranking in Table 3.5 highlights that accurate treatment of polarization, charge transfer, and charge states—either through physically motivated force fields or sufficiently rich ML training data—is essential for predictive modeling of protein–ligand interactions beyond dispersion-dominated regimes.

Table 3.5: *Benchmark performance of molecular mechanics force fields and machine-learning interaction potentials against DFT reference interaction energies across six protein–ligand NCI types.* Mean absolute errors (MAE) in kJ/mol; values in parentheses report the normalized error defined as MAE divided by the absolute value of the average DFT total interaction energy for each NCI category
. Additional comparisons in NCI sub-categories are provided in Supplementary Table S2-S4.

Model	Salt-Bridge	H-Bond charge	H-Bond neutral	Pi-cation
GAFF2/BCC	43.81 (8.31%)	21.25 (27.80%)	4.17 (17.47%)	9.64 (23.44%)
GAFF2/RESP	47.01 (8.92%)	25.02 (32.74%)	4.24 (17.77%)	12.89 (31.33%)
AMOEBA	19.23 (3.65%)	10.33 (13.51%)	3.62 (15.17%)	3.45 (8.39%)
MACE-OFF-23 (M)	367.26 (69.67%)	20.57 (26.91%)	1.83 (7.65%)	20.01 (48.66%)
AIMNET2	9.99 (1.89%)	4.76 (6.22%)	2.97 (12.46%)	4.94 (12.02%)
MACE-OMOL	4.62 (0.88%)	2.20 (2.88%)	1.09 (4.55%)	1.57 (3.81%)
UMA (S)	3.80 (0.72%)	1.12 (1.46%)	0.36 (1.52%)	1.49 (3.63%)
Number of data	2252	1872	2430	1505
Model	Pi-stack Parallel	Pi-stack T-shaped	Hydrophobic	
GAFF2/BCC	3.57 (15.86%)	2.35 (16.84%)	1.56 (14.09%)	
GAFF2/RESP	2.17 (9.65%)	2.60 (18.64%)	1.43 (12.95%)	
AMOEBA	1.86 (8.23%)	2.13 (15.27%)	1.13 (10.21%)	
MACE-OFF-23 (M)	2.63 (11.66%)	2.10 (15.03%)	1.60 (14.46%)	
AIMNET2	6.16 (27.36%)	3.80 (27.20%)	2.74 (24.77%)	
MACE-OMOL	1.21 (5.38%)	1.16 (8.32%)	1.06 (9.56%)	
UMA (S)	0.76 (3.37%)	0.42 (2.98%)	0.40 (3.60%)	
Number of data	527	439	4816	
Model	Halogen-F	Halogen-Cl	Halogen-Br	Halogen-I
GAFF2/BCC	2.76 (46.30%)	5.23 (90.03%)	11.21 (112.05%)	18.39 (144.66%)
GAFF2/RESP	2.22 (37.32%)	4.96 (85.48%)	10.19 (101.83%)	14.47 (113.81%)
AMOEBA	1.44 (24.12%)	1.10 (18.88%)	1.65 (16.45%)	3.34 (26.27%)
MACE-OFF-23 (M)	3.11 (52.17%)	3.82 (65.78%)	5.69 (56.87%)	9.61 (75.54%)
AIMNET2	2.69 (45.12%)	2.85 (49.15%)	3.45 (34.50%)	5.45 (42.88%)
MACE-OMOL	1.01 (17.04%)	1.10 (19.00%)	0.85 (8.45%)	1.43 (11.26%)
UMA (S)	0.31 (5.16%)	0.32 (5.44%)	0.39 (3.89%)	0.39 (3.04%)
Number of data	376	501	110	77

3.3.5 Trends Beyond Isolated NCI Fragment Dimers

To assess how a more extended protein environment influences the ALMO-EDA analysis beyond the fragment-dimer approximation, we constructed an auxiliary dataset from the OMOL25 evaluation dataset consisting of full ligand–protein pocket models. We retained systems of moderate size (100–150 atoms) containing only H, C, N, O, F, P, S, Cl, Br, and I, consistent with the chemical scope of the main HiQBind-derived benchmark, and selected 111 representative examples spanning different charge states. In contrast to the main dataset, this auxiliary dataset is not designed to isolate individual NCI classes, because each structure

contains the full ligand together with multiple surrounding residues and therefore multiple NCI classes are involved in one single structure. This small dataset instead provides an investigation of the ALMO-EDA profile in a more extended setting that includes additional local many-body effects and environmental screening.

We categorize this set into 16 binding conformations without hydrogen bonds, 65 with hydrogen bonds, and 30 salt-bridge-like charged complexes, and results are summarized in Table S8 and Figures S11–S12. The binding conformations without hydrogen bonds are the weakest and are dominated primarily by dispersion. Hydrogen-bonded complexes exhibit total interaction energies that are approximately twofold stronger, with polarization and charge transfer playing a more prominent role, compensating the additional destabilization caused by the increased Pauli repulsion. Salt-bridge complexes, involving oppositely charged species, display the strongest interactions, with the enhanced stabilization arising primarily from much stronger permanent electrostatics together with increased polarization. These results are fully consistent with the physical picture obtained from the fragment-based analysis.

Table 3.6: *Benchmark performance of molecular mechanics force fields and machine-learning interaction potentials against DFT reference interaction energies for the auxiliary OMOL full ligand–protein pocket dataset.* Mean absolute errors (MAE) in kJ/mol; values in parentheses report the normalized error defined as MAE divided by the magnitude of the average total DFT interaction energy.

	ALL	No H-Bond	H-Bond	Salt Bridges
AMBER14SB/GAFF2/BCC	10.63 (5.19%)	2.56 (4.83%)	7.65 (6.64%)	21.28 (4.43%)
AMOEBA	14.17 (6.92%)	5.64 (10.63%)	12.48 (10.83%)	22.27 (4.64%)
MACE-OFF-23 (M)	80.26 (39.18%)	10.10 (19.03%)	17.59 (15.27%)	253.45 (52.82%)
AIMNET2	17.27 (8.43%)	12.59 (23.72%)	20.87 (18.11%)	11.98 (2.50%)
MACE-OMOL	6.35 (3.10%)	2.91 (5.47%)	2.89 (2.51%)	15.70 (3.27%)
UMA (S)	2.96 (1.44%)	0.57 (1.08%)	1.84 (1.60%)	6.65 (1.39%)

We also benchmarked this full ligand–protein pocket dataset against force fields and MLIPs, as summarized in Table 6 and Figure S13. To better mimic practical protein–ligand modeling, the fixed-charge baseline uses AMBER14SB[194] for the protein and GAFF2/AM1-BCC for the ligand, while the AMOEBA model uses predefined AMOEBABIO18 parameters[195] for the protein and Poltype-generated AMOEBA parameters for the ligand. With the larger system size and the presence of multiple simultaneous NCIs, the overall ranking reveals new insight into the fragment-dimer benchmark results. Interestingly, the AMBER14SB/GAFF2 fixed-charge model performs competitively on this full-pocket dataset, with errors comparable to or lower than AMOEBA and AIMNET2 across the three subsets of weak to strong protein-ligand energetics. This likely reflects the more extensive empirical parametrization used by mature fixed-charge FFs that are tuned and validated on large fragments, whole proteins, and experimental condensed-phase properties such as densities and heats of vaporization.

Hence the fixed charge models are largely practical but "unprincipled", i.e. in the sense that they have not performed adequately based on NCIs but are largely better at recapitulaing amalgamated energies. Although AMOEBA protein parameters were derived from MP2-level calculations[196], which may introduce some additional errors relative to the ωB97X-V/def2-TZVPD benchmark, the polarizable model overall has a principled understanding of NCIs that is extensible to collective many-body effects of multiple NCIs, but just not at the level of accuracy of MLIPs such as UMA and MACE-OMOL. Although UMA and MACE-OMOL have been trained on similar full protein-ligand complexes with the same level of DFT as the full protein-ligand benchmark, the two-body NCIs are not as representative in their training set, indicating that they too understand the underlying physics of molecular interactions. Finally, MACE-OFF-23 performs substantially worse, especially for salt-bridge-like complexes, again highlighting the importance of charge-state coverage and more careful treatment of training data in general in which the AIMNet2, UMA, and MACE-OMOL models have better accomplished.

3.4 Conclusions

In this work, we present a large, systematically curated dataset of protein–ligand noncovalent interaction energies derived from high-quality experimental structures and analyzed using ALMO energy decomposition analysis. By resolving interaction energies into physically meaningful components—electrostatics, Pauli repulsion, dispersion, polarization, and charge transfer—we provide a quantitative framework for understanding how different classes of NCIs arise and how they should be modeled. Our analysis reveals that while dispersion dominates hydrophobic and π-stacking interactions, polarization and charge transfer play a decisive role in hydrogen bonds, π-cation interactions, halogen bonds, and salt bridges, particularly in charged environments.

Benchmarking classical force fields and modern MLIPs against DFT highlights both progress and remaining challenges. Fixed-charge models remain competitive for dispersion-driven interactions but exhibit systematic deficiencies for strongly electrostatic motifs. Polarizable force fields such as AMOEBA substantially reduce these errors, while state-of-the-art MLIPs—especially UMA and MACE-OMOL—achieve near-quantitative agreement with DFT across all interaction types when trained on sufficiently diverse datasets. Conversely, the failure of models lacking charge-state coverage underscores the importance of explicitly representing charged and many-body effects in training data. However, we clarify that this benchmark is based on gas-phase single-point DFT calculations of protein–ligand fragment dimers, and the reported energies therefore do not fully represent "binding free energy", which is a thermodynamic property that should be estimated by ensemble of configurations in the presence of full biological environment, including extended hydrogen bond network and many-body effects, solvation effects, conformational reorganization etc.

In conclusion, the dataset and benchmark presented here establish a physically grounded

reference for next-generation force field development and MLIP training and validation, because accurate single-point interaction energies and their physical decomposition are the foundation on which reliable and interpretable potential energy surfaces, molecular simulations, and ultimately free-energy calculations are built. By linking interaction accuracy directly to underlying EDA components, our results provide clear guidance for designing transferable, predictive models of protein–ligand interactions for drug discovery and molecular design.

3.5 Supporting Information

3.5.1 Supporting Tables

Table S3.1: Overview of NCI datasets.

Name	No. of geometries	No. of distinct dimers	Level of EDA theory
Splinter[156]	1.7M	>9,000	SAPT0/aug-cc-pV(D+d)Z
DES370K[157]	370K	$\sim 3,700$	SAPT0/aug-cc-PVTZ
Schriber21[158]	30K	$\sim 1,000$	SAPT2+(3)δMP2/aug-cc-pVTZ
SAPT10K[159]	9,982	944	SAPT2+(3)(CCD)/aug-cc-pVTZ
NENCI-2021[160]	7,763	141	SAPT2+/aug-cc-pVDZ
Ours	14,905	9,883	ALMO-EDA/ωB97X-V/def2-TZVPD

Table S3.2: *Benchmark performance of molecular mechanics force fields and machine-learning interaction potentials against DFT reference interaction energies for pi-cation.* Mean absolute errors (MAE) in kJ/mol; values in parentheses report the normalized error defined as MAE divided by the magnitude of the total DFT interaction energy.

	TYR	PHE	TRP	HIS
GAFF2/BCC	5.44 (15.40%)	6.92 (20.08%)	5.22 (11.12%)	8.33 (37.00%)
GAFF2/RESP	7.37 (20.87%)	7.36 (21.37%)	8.55 (18.22%)	6.54 (29.02%)
AMOEBA	1.75 (4.97%)	1.83 (5.31%)	1.91 (4.06%)	2.93 (12.99%)
MACE-OFF-23 (M)	18.47 (52.33%)	19.51 (56.64%)	29.76 (63.40%)	12.16 (53.97%)
AIMNET2	5.54 (15.70%)	6.55 (19.03%)	5.94 (12.65%)	3.44 (15.28%)
MACE-OMOL	0.86 (2.44%)	1.07 (3.10%)	0.84 (1.79%)	1.15 (5.09%)
UMA (S)	0.97 (2.74%)	0.91 (2.64%)	1.42 (3.02%)	1.17 (5.18%)
Number of data	147	61	171	38

Table S3.3: *Benchmark performance of molecular mechanics force fields and machine-learning interaction potentials against DFT reference interaction energies for pi-stacking.* Mean absolute errors (MAE) in kJ/mol; values in parentheses report the normalized error defined as MAE divided by the magnitude of the total DFT interaction energy.

Model	All	Parallel	T-shaped
GAFF2/BCC	3.02 (16.19%)	3.57 (15.86%)	2.35 (16.84%)
GAFF2/RESP	2.37 (12.71%)	2.17 (9.65%)	2.60 (18.64%)
AMOEBA	1.98 (10.63%)	1.86 (8.23%)	2.13 (15.27%)
MACE-OFF-23 (M)	2.39 (12.81%)	2.63 (11.66%)	2.10 (15.03%)
AIMNET2	5.09 (27.30%)	6.16 (27.36%)	3.80 (27.20%)
MACE-OMOL	1.19 (6.38%)	1.21 (5.38%)	1.16 (8.32%)
UMA (S)	0.60 (3.24%)	0.76 (3.37%)	0.42 (2.98%)
Number of data	966	527	439
Model	Heterocyclic	Carbocyclic	Mixed
GAFF2/BCC	3.55 (16.80%)	2.71 (16.90%)	3.13 (15.29%)
GAFF2/RESP	2.77 (13.12%)	2.49 (15.58%)	2.05 (10.00%)
AMOEBA	2.27 (10.72%)	1.97 (12.32%)	1.87 (9.10%)
MACE-OFF-23 (M)	3.11 (14.70%)	2.12 (13.25%)	2.37 (11.55%)
AIMNET2	5.05 (23.91%)	4.29 (26.81%)	6.01 (29.30%)
MACE-OMOL	1.26 (5.97%)	1.10 (6.88%)	1.26 (6.13%)
UMA (S)	0.85 (4.04%)	0.49 (3.08%)	0.62 (3.01%)
Number of data	168	423	375

Table S3.4: *Benchmark performance of molecular mechanics force fields and machine-learning interaction potentials against DFT reference interaction energies for salt-bridge.* Mean absolute errors (MAE) in kJ/mol; values in parentheses report the normalized error defined as MAE divided by the magnitude of the total DFT interaction energy.

Model	All	Protein	Ligand
GAFF2/BCC	43.81 (8.31%)	47.40 (8.26%)	33.97 (8.50%)
GAFF2/RESP	47.01 (8.92%)	50.53 (8.81%)	37.32 (9.34%)
AMOEBA	19.23 (3.65%)	19.74 (3.44%)	17.83 (4.46%)
MACE-OFF-23 (M)	367.26 (69.67%)	405.16 (70.63%)	263.15 (65.87%)
AIMNET2	9.99 (1.89%)	11.65 (2.03%)	5.43 (1.36%)
MACE-OMOL	4.62 (0.88%)	5.74 (1.00%)	1.55 (0.39%)
UMA (S)	3.80 (0.72%)	4.64 (0.81%)	1.49 (0.37%)
Number of data	2252	1651	601
Model	Carboxylate	Phosphate	Sulfate
GAFF2/BCC	34.59 (8.55%)	66.27 (8.01%)	32.20 (8.94%)
GAFF2/RESP	37.48 (9.27%)	70.73 (8.55%)	30.54 (8.47%)
AMOEBA	17.26 (4.27%)	24.34 (2.94%)	14.12 (3.92%)
MACE-OFF-23 (M)	265.54 (65.68%)	614.78 (74.28%)	239.25 (66.40%)
AIMNET2	5.81 (1.44%)	20.09 (2.43%)	5.14 (1.43%)
MACE-OMOL	2.13 (0.53%)	10.66 (1.29%)	1.72 (0.48%)
UMA (S)	1.57 (0.39%)	9.18 (1.11%)	1.26 (0.35%)
Number of data	1509	662	81

Table S3.5: Overview of the Metadata fields.

Field	Type	Description
PDBID	string	Four-letter PDB code
subdir	string	HiQBind sub-category `pdbbind_opt_sm`, `pdbbind_opt_poly`, or `biolip2_opt`
smiles0	string	Ligand fragment SMILES string
charge0	int	Charge of the ligand fragment
natoms0	int	Number of atoms in ligand fragment
smiles1	string	Protein fragment SMILES string.
charge1	int	Charge of the protein fragment.
natoms1	int	Number of atoms in protein fragment
elements	list of string	List of atom symbols of the dimer, separate by whitespace. Protein fragment concatenated after ligand fragment.
xyz	list of float	List of Cartesian coordinates of the dimer, separate by whitespace. Protein fragment concatenated after ligand fragment.
ELEC	float	Electrostatics component of ALMO-EDA from orthogonal frozen decomposition[a]
PAULI	float	Pauli repulsion component of ALMO-EDA from orthogonal frozen decomposition[a]
DISP	float	Dispersion component of ALMO-EDA[c]
CLS_ELEC	float	Electrostatics component of ALMO-EDA from classical frozen decomposition[a,b]
MOD_PAULI	float	Pauli repulsion component of ALMO-EDA from classical frozen decomposition[a,b]
FROZEN	float	Sum of ELEC, PAULI, and DISP
POLARIZATION	float	Polarization component of ALMO-EDA
CHARGE_TRANSFER	float	Charge transfer component of ALMO-EDA
TOTAL	float	Total interaction Energy of dimer system

[a] The difference between classical and orthogonal decomposition is described in Ref [68] as well as the QChem manual section 12.6.4.

[b] In the main text and supporting figures, all the permanent electrostatics ("ELEC" in the figures) and Pauli repulsion ("PAULI" in the figures) actually refers to the classical decomposition, i.e. CLS_ELEC and MOD_PAULI in this table.

[c] Dispersion is always under orthogonal decomposition by default except for structures with iodine because QChem is not able to perform orthogonal decomposition for iodine.

Table S3.6: *ALMO-EDA energies fragment dimers and extended models (full ligand + full protein residue) for electrostatically dominated protein–ligand NCIs.* Values are average ALMO-EDA energies in kJ/mol, with medians in parentheses. Portions are ALMO-EDA components normalized by the average total interaction energy.

Energy (kJ/mol)	Salt-Bridge	Salt-Bridge-full	H-Bond	H-Bond-full
Total	−369.38 (−367.82)	−366.86 (−387.03)	−44.28 (−35.83)	−59.65 (−49.28)
Pauli	35.97 (35.01)	53.96 (54.22)	50.61 (51.47)	74.52 (77.60)
Electrostatics	−346.88 (−346.44)	−335.74 (−349.11)	−52.10 (−49.26)	−63.50 (−59.90)
Dispersion	−13.04 (−12.26)	−23.33 (−18.74)	−17.68 (−18.26)	−36.65 (−32.69)
Polarization	−34.89 (−39.88)	−48.47 (−51.32)	−14.88 (−11.43)	−21.09 (−17.13)
Charge Transfer	−10.52 (−11.81)	−13.28 (−13.43)	−10.22 (−10.78)	−12.92 (−14.45)
Portions	Salt-Bridge	Salt-Bridge-full	H-Bond	H-Bond-full
Pauli	−0.10 (−0.10)	−0.15 (−0.14)	−1.14 (−1.44)	−1.25 (−1.57)
Electrostatics	0.94 (0.94)	0.92 (0.90)	1.18 (1.37)	1.06 (1.22)
Dispersion	0.04 (0.03)	0.06 (0.05)	0.40 (0.51)	0.61 (0.66)
Polarization	0.09 (0.11)	0.13 (0.13)	0.34 (0.32)	0.35 (0.35)
Charge Transfer	0.03 (0.03)	0.04 (0.03)	0.23 (0.30)	0.22 (0.29)
Number of data	14	14	15	15
Energy (kJ/mol)	Pi-cation	Pi-cation-full		
Total	−32.76 (−33.94)	−47.69 (−55.41)		
Pauli	15.61 (17.42)	44.26 (35.42)		
Electrostatics	−15.23 (−13.81)	−25.85 (−20.60)		
Dispersion	−16.22 (−18.38)	−39.58 (−42.46)		
Polarization	−13.92 (−14.65)	−20.65 (−19.07)		
Charge Transfer	−3.00 (−3.10)	−5.87 (−6.01)		
Portions	Pi-cation	Pi-cation-full		
Pauli	−0.48 (−0.51)	−0.93 (−0.64)		
Electrostatics	0.46 (0.41)	0.54 (0.37)		
Dispersion	0.50 (0.54)	0.83 (0.77)		
Polarization	0.43 (0.43)	0.43 (0.34)		
Charge Transfer	0.09 (0.09)	0.12 (0.11)		
Number of data	9	9		

Table S3.7: *ALMO-EDA energies of fragment dimers and minimally extended models (full ligand + full protein residue) for weakly interacting protein–ligand NCIs.* Values are average ALMO-EDA energies in kJ/mol, with medians in parentheses. Portions are ALMO-EDA components normalized by the average total interaction energy.

Energy (kJ/mol)	Pi-stack	Pi-stack-full	Hydrophobic	Hydrophobic-full
Total	−16.45 (−14.47)	−33.98 (−28.76)	−11.41 (−10.85)	−18.10 (−19.36)
Pauli	22.89 (20.98)	46.20 (34.21)	17.99 (17.32)	21.09 (19.67)
Electrostatics	−10.53 (−9.27)	−25.68 (−13.42)	−6.08 (−5.69)	−8.41 (−6.92)
Dispersion	−22.49 (−20.11)	−39.57 (−35.83)	−20.11 (−20.48)	−25.70 (−22.96)
Polarization	−2.81 (−1.93)	−7.63 (−3.77)	−0.98 (−1.00)	−2.32 (−2.05)
Charge Transfer	−3.51 (−2.82)	−7.31 (−4.14)	−2.23 (−2.30)	−2.77 (−2.69)
Portions	Pi-stack	Pi-stack-full	Hydrophobic	Hydrophobic-full
Pauli	−1.39 (−1.45)	−1.36 (−1.19)	−1.58 (−1.60)	−1.17 (−1.02)
Electrostatics	0.64 (0.64)	0.76 (0.47)	0.53 (0.52)	0.46 (0.36)
Dispersion	1.37 (1.39)	1.16 (1.25)	1.76 (1.89)	1.42 (1.19)
Polarization	0.17 (0.13)	0.22 (0.13)	0.09 (0.09)	0.13 (0.11)
Charge Transfer	0.21 (0.20)	0.22 (0.14)	0.20 (0.21)	0.15 (0.14)
Number of data	15	15	15	15
Energy (kJ/mol)	Halogen	Halogen-full		
Total	−8.34 (−8.03)	−16.33 (−15.13)		
Pauli	13.09 (8.86)	30.17 (21.13)		
Electrostatics	−6.74 (−4.06)	−13.99 (−10.58)		
Dispersion	−10.38 (−10.18)	−23.07 (−22.04)		
Polarization	−2.74 (−1.63)	−5.58 (−3.01)		
Charge Transfer	−1.57 (−1.29)	−3.86 (−2.87)		
Portions	Halogen	Halogen-full		
Pauli	−1.57 (−1.10)	−1.85 (−1.40)		
Electrostatics	0.81 (0.50)	0.86 (0.70)		
Dispersion	1.25 (1.27)	1.41 (1.46)		
Polarization	0.33 (0.20)	0.34 (0.20)		
Charge Transfer	0.19 (0.16)	0.24 (0.19)		
Number of data	16	16		

Table S3.8: *Total interaction energies and their ALMO-EDA components of the auxiliary OMOL full ligand–protein pocket dataset.* Average and (median) energies in kJ/mol.

Energy (kJ/mol)	ALL	No H-Bond	H-Bond	Salt Bridges
Total	−204.82 (−120.51)	−53.08 (−55.07)	−115.23 (−111.97)	−479.88 (−494.18)
Pauli	144.92 (132.20)	53.31 (56.24)	156.81 (146.50)	168.03 (138.33)
Electrostatics	−188.09 (−110.77)	−27.50 (−27.07)	−114.64 (−99.36)	−432.88 (−441.81)
Dispersion	−93.50 (−92.32)	−64.14 (−66.64)	−99.35 (−100.16)	−96.48 (−92.26)
Polarization	−43.21 (−32.59)	−7.45 (−7.34)	−32.02 (−28.31)	−86.54 (−82.94)
Charge Transfer	−24.94 (−23.49)	−7.29 (−6.94)	−26.03 (−23.49)	−32.01 (−26.95)
Number of data	111	16	65	30

3.5.2 Supporting Figures

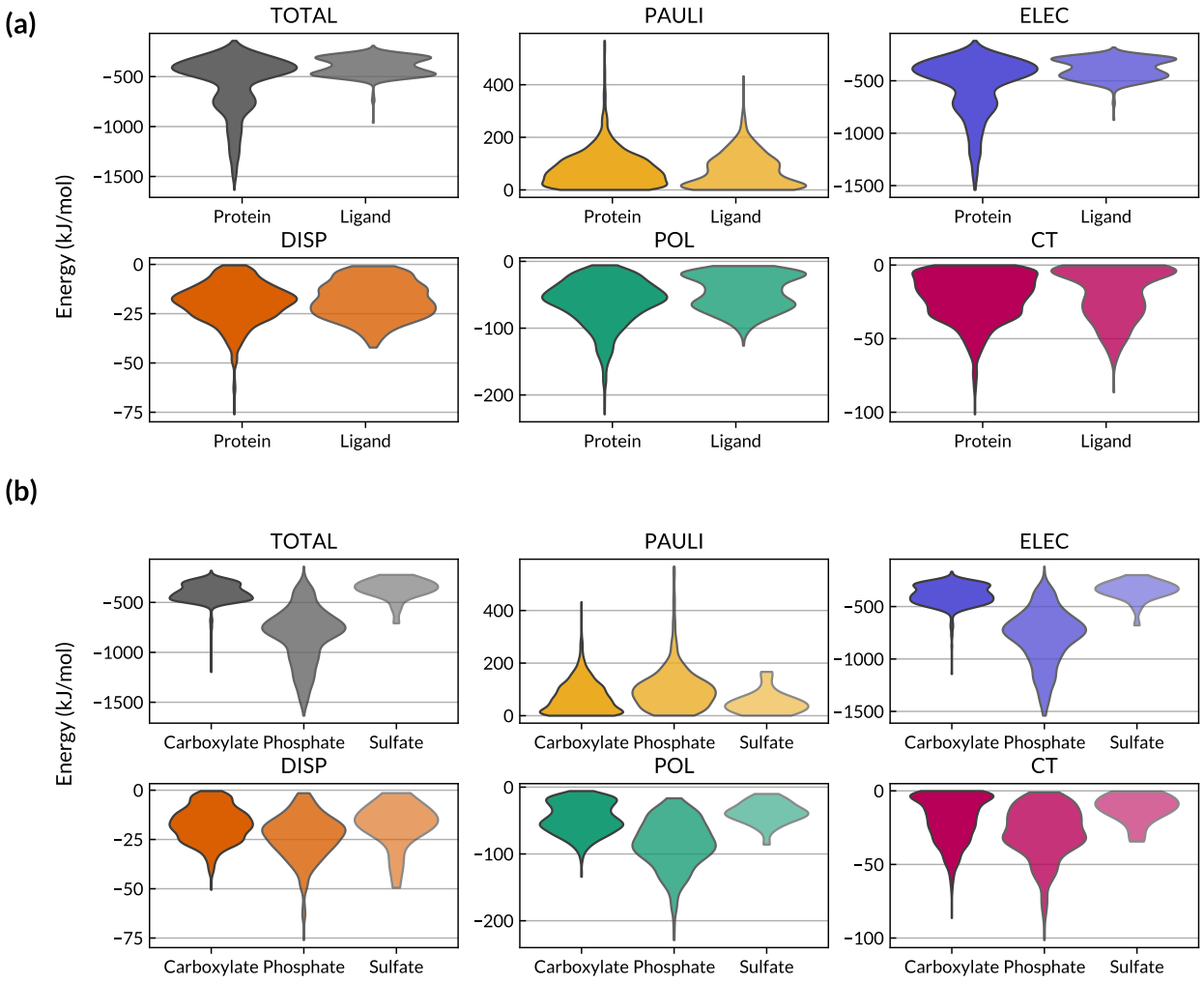

Figure S3.1: Distribution of total energy and ALMO-EDA energy components for protein-ligand salt bridges. Violin plots show (a) lower energetics when the cation is centered on the protein versus the ligand, and (b) among three types of negatively charged functional group, the most favorable interactions found from phosphate groups.

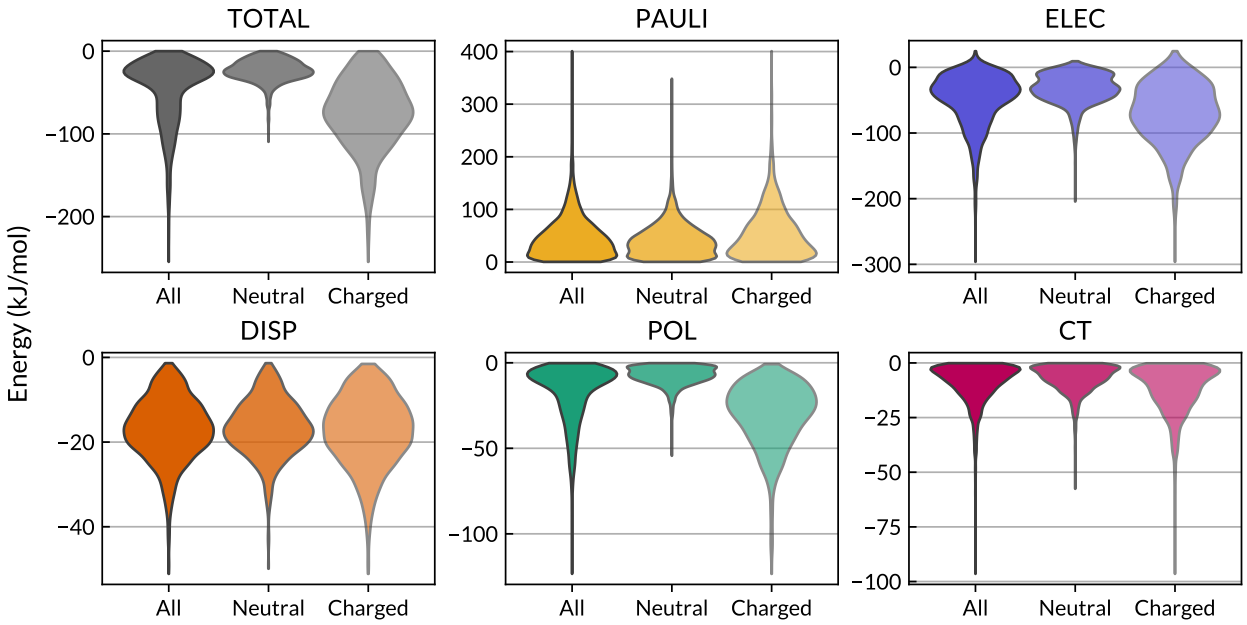

Figure S3.2: Distribution of total energy and ALMO-EDA energy components for protein-ligand hydrogen bonds. Violin plots show comparison of neutral and charged hydrogen bonded complexes.

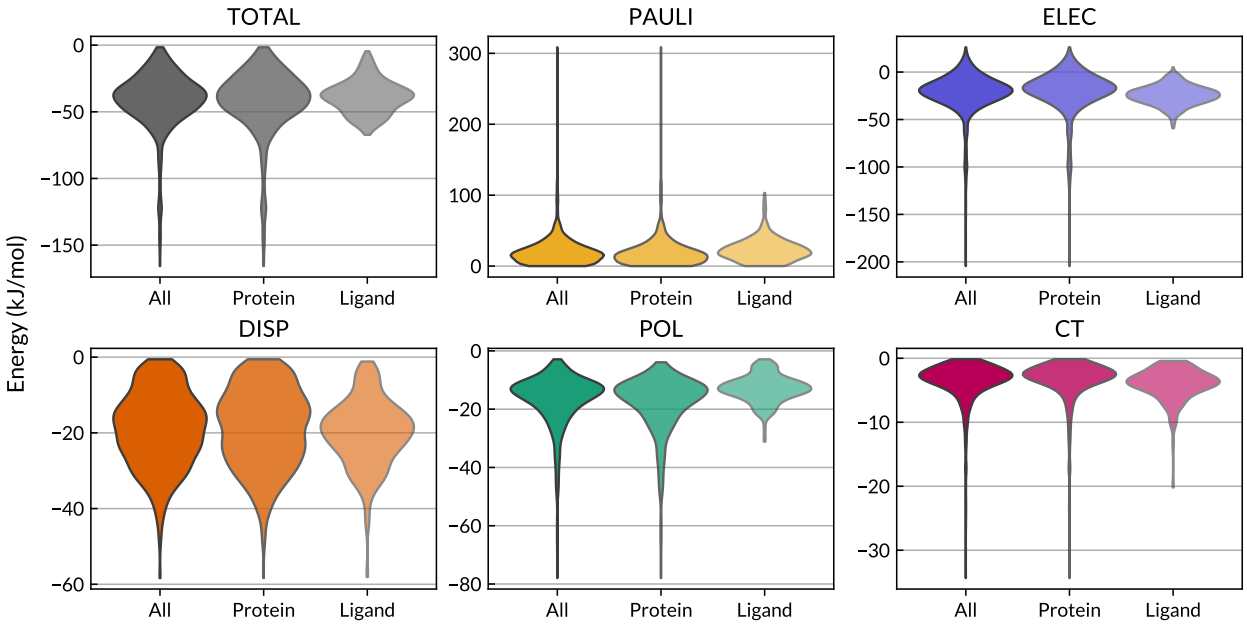

Figure S3.3: Distribution of total energy and ALMO-EDA energy components for protein-ligand π-cation interactions. Violin plots show lower energetics when the cation is centered on the protein gives a larger energetic range supported by polarization and charge transfer.

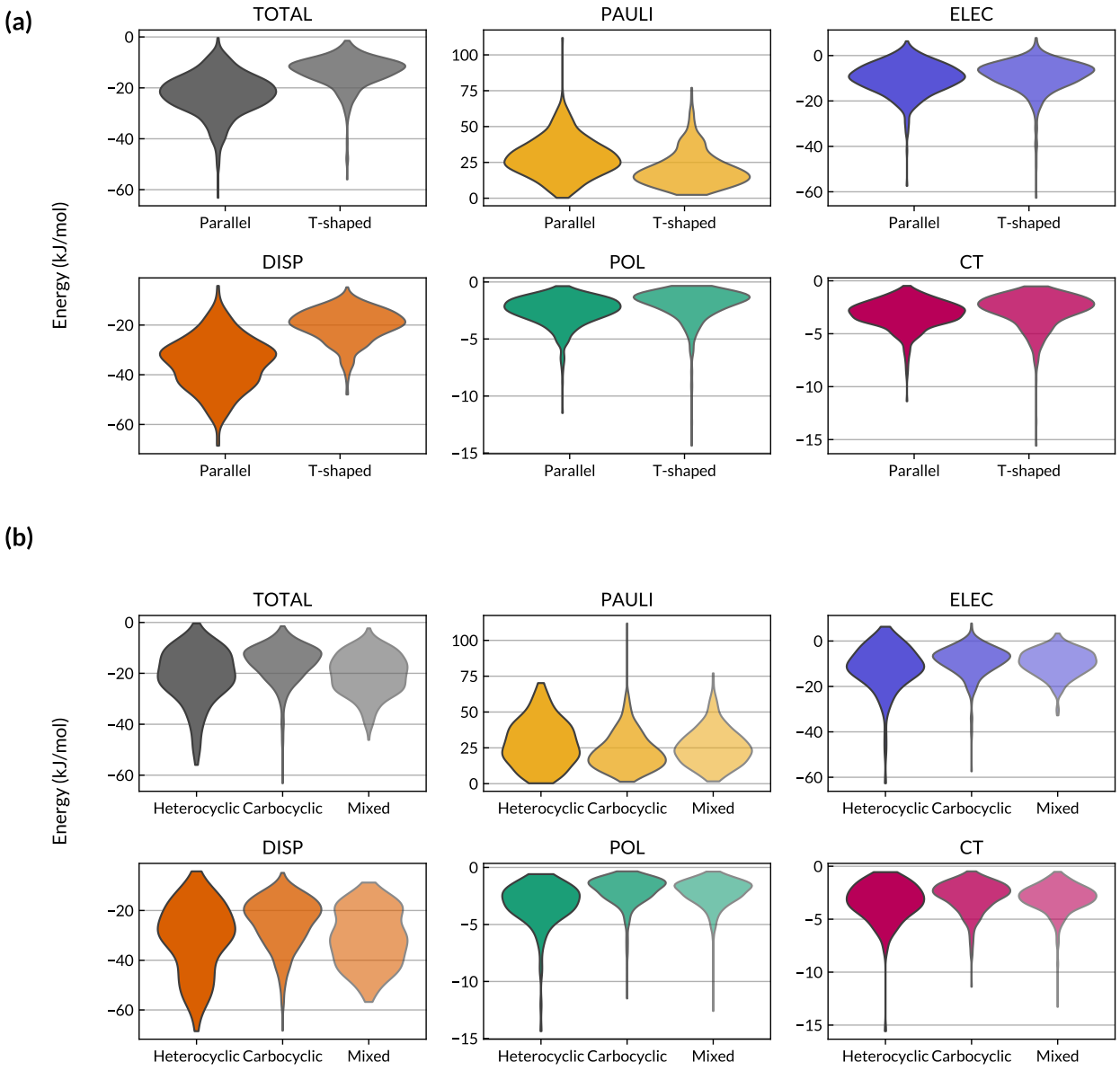

Figure S3.4: Distribution of total energy and ALMO-EDA energy components for protein-ligand π-stacking interactions. Violin plots show (a) lower energetics for stacked versus T-shaped geometries due to greater surface area for dispersion stabilization, and (b) greater stabilization found by introducing heteroatoms into the carbocyclic ring.

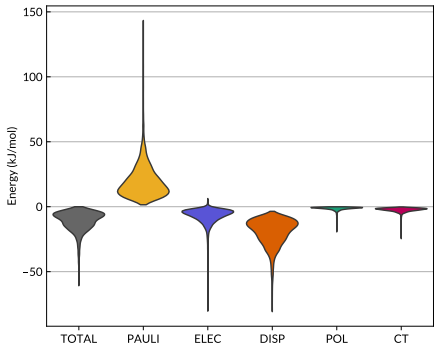

Figure S3.5: Distribution of total energy and ALMO-EDA energy components for protein-ligand hydrophobic interactions. Violin plots show (a) lower energetics for stacked versus T-shaped geometries due to greater surface area for dispersion stabilization, and (b) greater stabilization found by introducing heteroatoms into the carbocyclic ring.

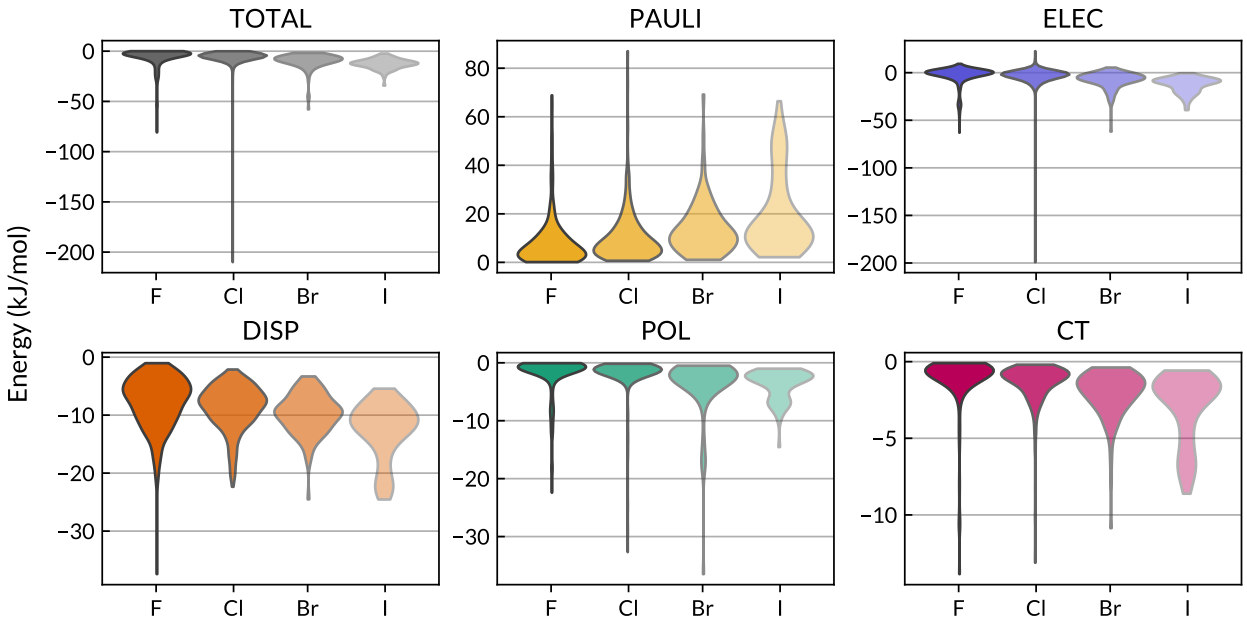

Figure S3.6: Distribution of total energy and ALMO-EDA energy components for protein-ligand π-stacking interactions. Violin plots show (a) lower energetics for stacked versus T-shaped geometries due to greater surface area for dispersion stabilization, and (b) greater stabilization found by introducing heteroatoms into the carbocyclic ring.

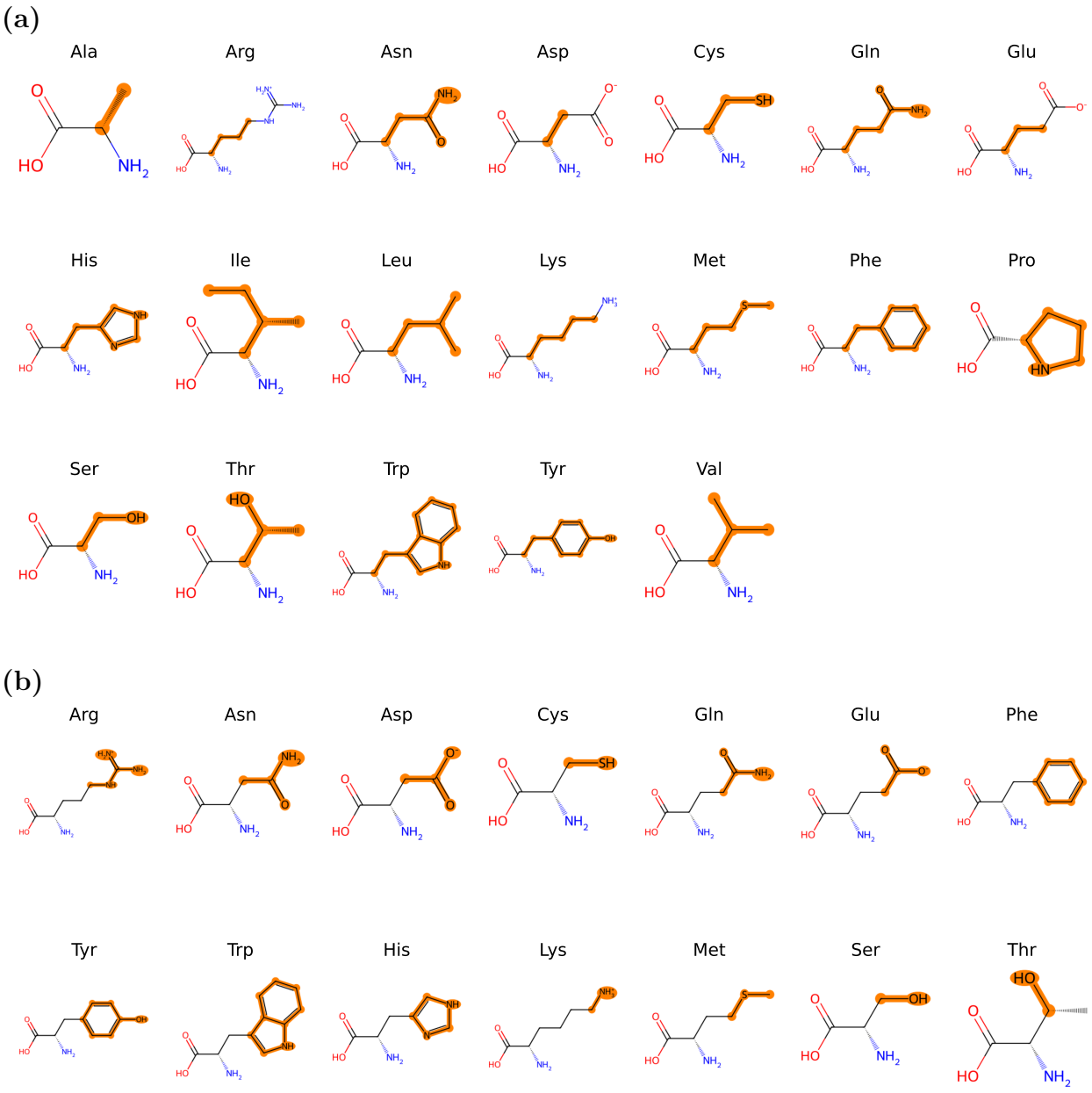

Figure S3.7: Protein fragmentation scheme when the interacting atom is located on a side chain forming (a) hydrophobic interactions or (b) other NCI types. Highlighted atoms are selected as the fragment, and open valences are capped with hydrogens.

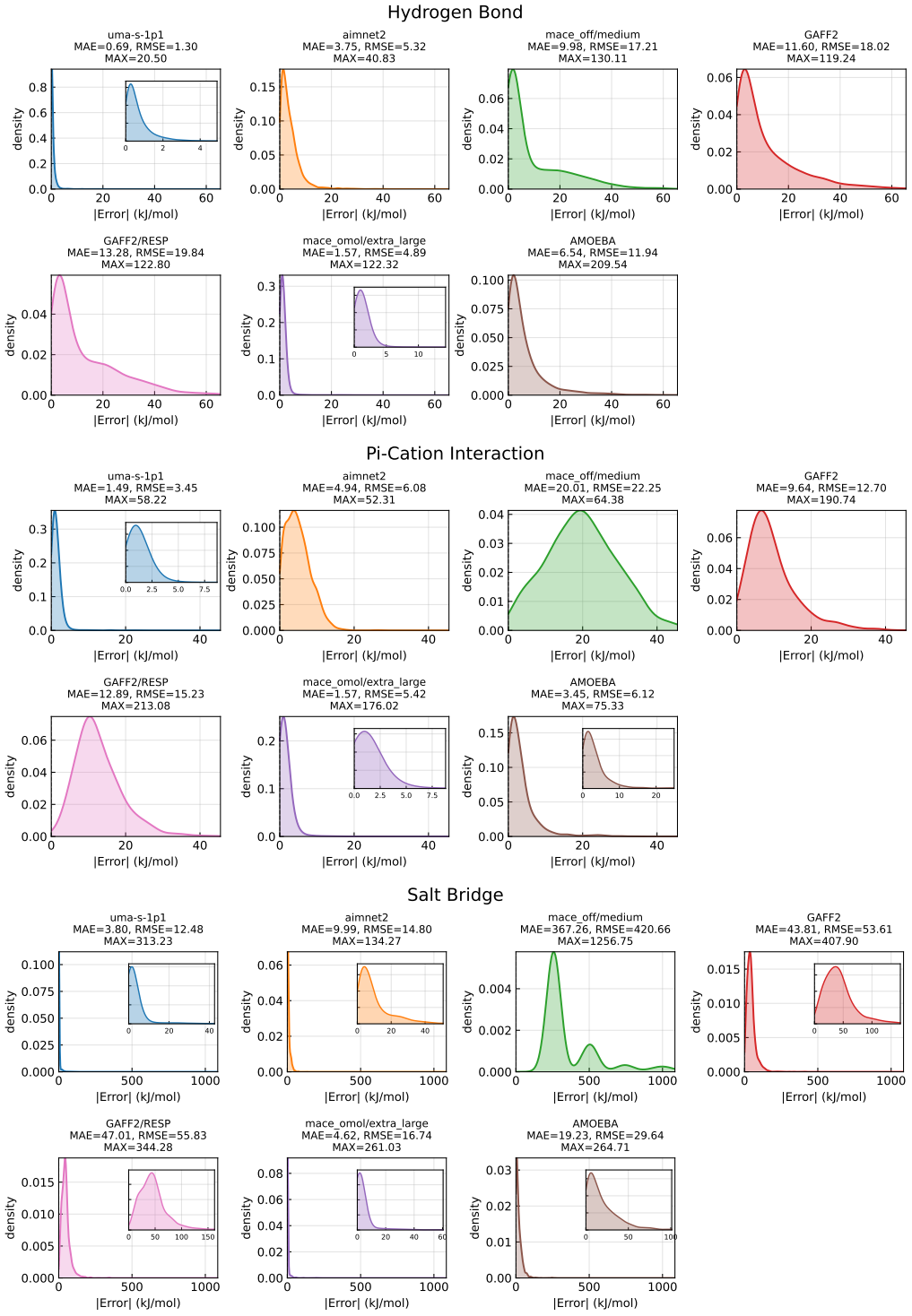

Figure S3.8: Absolute error distributions of total interaction energies for electrostatically dominated NCIs: hydrogen bond, pi-cation and salt bridges

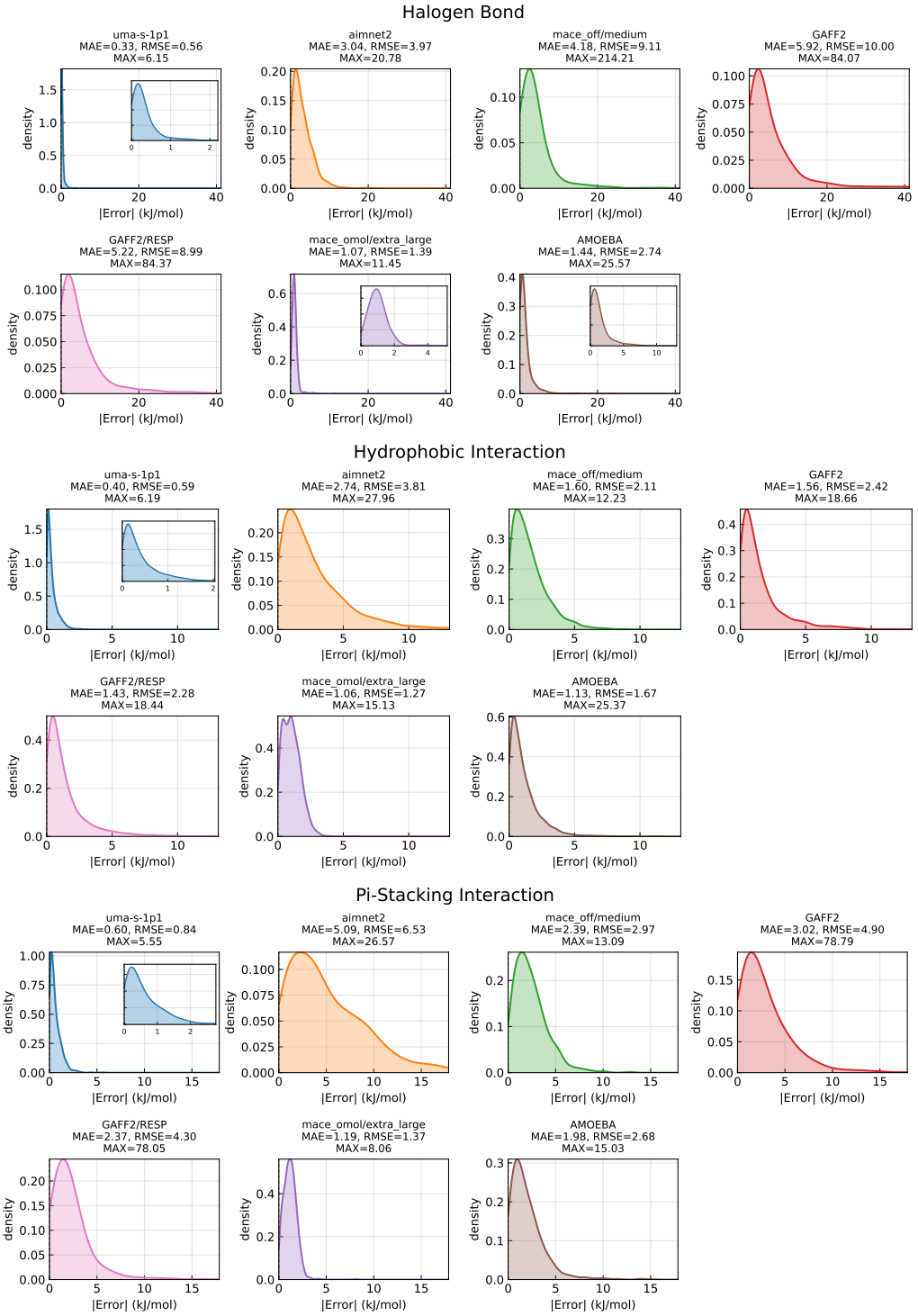

Figure S3.9: Absolute error distributions of total interaction energies for weakly interacting NCIs: halogen bond, hydrophobic and pi-stacking.

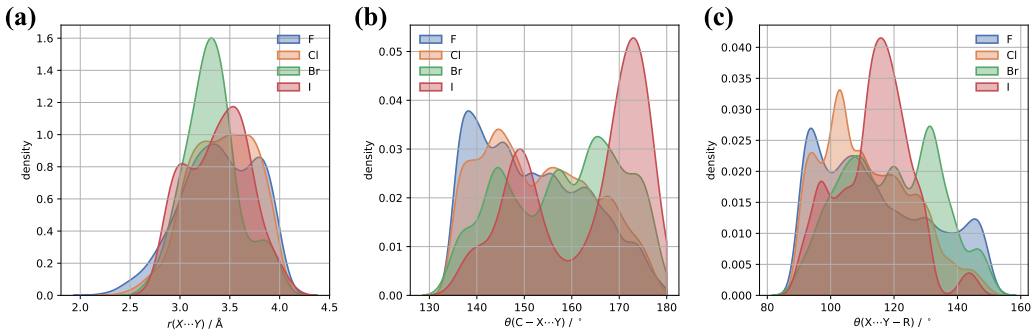

Figure S3.10: *Distribution of the distances and angles of halogen bonds separated by halogen atom types.* (a) Distance between halogen X and electron-pair donor Y, (b) Angle of halogen acceptor $\angle(\mathrm{C}-\mathrm{X}\cdots\mathrm{Y})$, and (c) Angle of Donor atom $\angle(\mathrm{X}\cdots\mathrm{Y}-\mathrm{R})$.

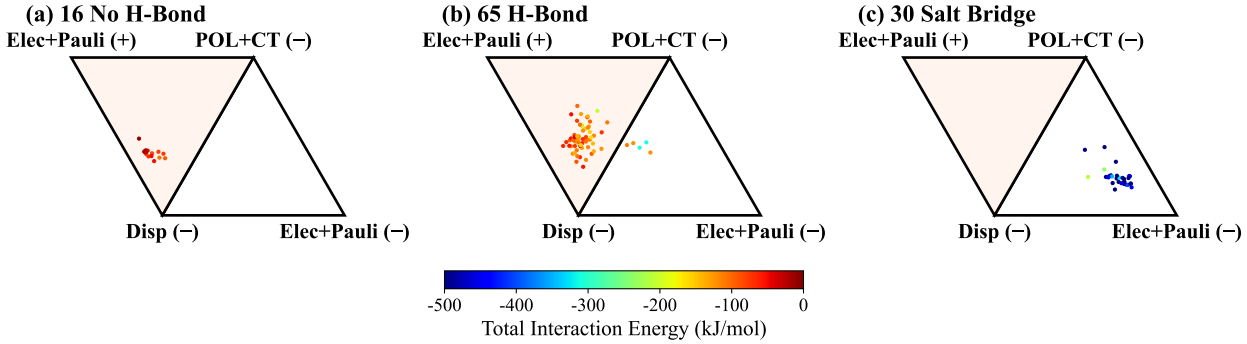

Figure S3.11: *Ternary plots of the ALMO-EDA energy components of the auxiliary OMOL full protein-pocket ligand dataset.*

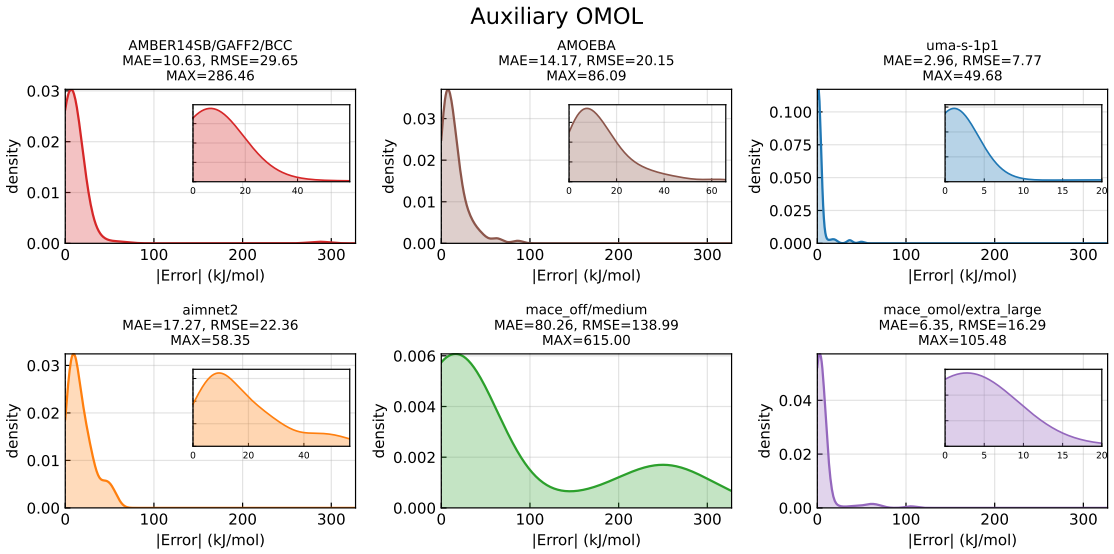

Figure S3.13: Absolute error distributions of total interaction energies for the auxiliary OMOL dataset.

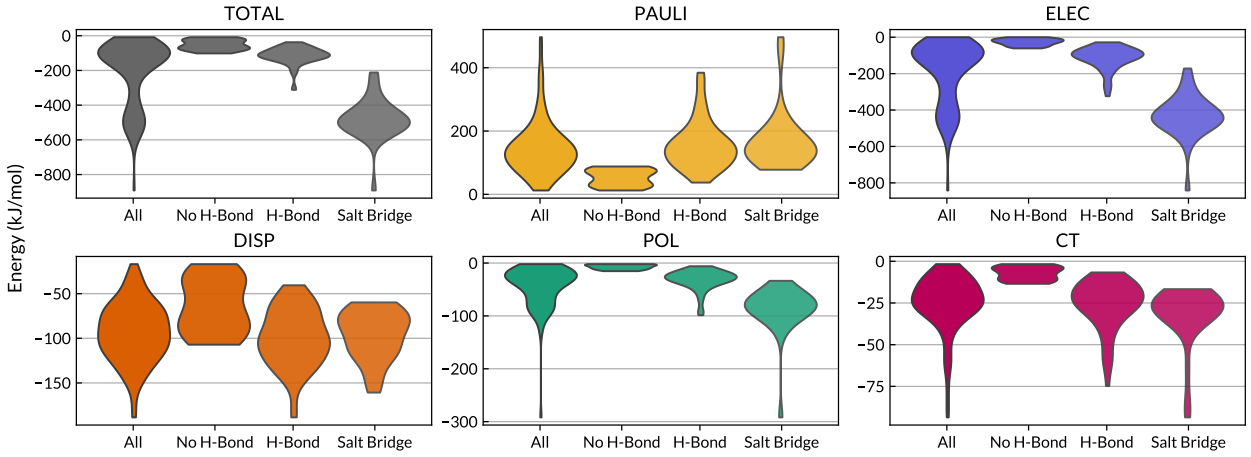

Figure S3.12: *Distribution of total interaction energies and their ALMO-EDA components of the auxiliary OMOL full protein-pocket ligand dataset.*

Chapter 4

TorchFF: A High-Performance GPU-Accelerated Differentiable Force Field Library

4.1 Introduction

Molecular dynamics (MD) simulation serves as a cornerstone of modern computational chemistry, offering an atomistic-resolution perspective on the thermodynamic and kinetic behavior of complex molecular systems. The predictive fidelity of these simulations is fundamentally governed by the accuracy of the underlying force field (FF), which defines the potential energy surface (PES). Conventional force fields, such as AMBER[57], CHARMM[58, 59], and OPLS-AA[60, 197], rely on a pairwise-additive framework to describe intermolecular interactions via point charges and Lennard-Jones potentials.[16] While these models achieve remarkable computational efficiency and scalability through highly optimized code and hardware, their simplified mathematical forms often neglect critical physics including polarization, charge transfer and charge penetration.[62, 65, 66] Consequently, conventional FFs frequently lack transferability and must be meticulously tailored for specific chemical environments through a top-down parameterization process and fitting parameters to reproduce macroscopic experimental data.

To address these limitations, a new generation of advanced FFs have emerged, utilizing functional forms that extend beyond the classical pairwise-additive paradigm. These developments can be broadly categorized into data-driven and physics-driven models. Data-driven approaches leverage various machine learning (ML) architectures trained on extensive *ab initio* datasets[52–55], while physics-driven models analytically incorporate sophisticated terms such as bonded term couplings[198–200], multipolar interactions, explicit polarization[201,

[1]Reproduced with permission from: TorchFF: A High-Performance GPU-Accelerated Differentiable Force Field Library. Y. Wang, A. S. Abdullah, R. S. Mohanakrishnan, and T. Head-Gordon. ChemRxiv, 2026

202], and many-body charge transfer.[73, 203] Recent efforts have also focused on hybrid ML/physics-based frameworks, employing neural networks to describe short-range, localized interactions while retaining analytical forms (e.g., Ewald summation) for long-range electrostatics[204–206]. However, the implementation of these complex, non-additive interactions is an arduous task, as the requirement to manually derive and code analytical forces often impedes the rapid prototyping and validation of new models.

These challenges have been unified as the requirement to differentiate the potential energy with respect to both atomic coordinates and force field parameters[18, 20]. This task is naturally suited for automatic differentiation (AD), the foundational technology of modern machine learning frameworks such as PyTorch[69], JAX[70], and TensorFlow[207]. By leveraging AD, developers can focus on the underlying physical equations while the framework handles the generation of exact gradients. Several differentiable force field platforms have recently introduced, including TorchMD[71, 208], JAX-MD[72, 209] DMFF[20], and the recently developed DIMOS framework[210], which provides a fast, modular and end-to-end differentiable simulation environment in PyTorch that not only provides interfaces to conventional FFs and ML potentials but also various integrators with thermostat/barostat support. All of these efforts demonstrate the promise of this paradigm. At the same time, many efforts have been devoted to improve the performance of the AD frameworks through compiler- and runtime-based optimization strategies, including just-in-time (JIT) compilation (XLA-backed compiler in JAX[70], Triton in PyTorch[211]) to automatically "translate" high-level Python operations into low-level C++/CUDA code with kernel fusion, and CUDA graphs for reducing CPU-side launch overhead through the "capture-replay" model.

Nevertheless, differentiable molecular mechanics continues to face substantial performance bottlenecks because its dominant workloads differ fundamentally from those of deep learning. Rather than being dominated by large, high-dimensional matrix operations, molecular mechanics is characterized by numerous small-kernel computations and specialized algorithmic patterns, such as neighbor-list construction and Ewald/PME summation, that are not naturally expressed through generic tensor operations. Consequently, compiler-based optimizations alone may not produce the most efficient execution paths for these workloads, especially when performance improvement gains from fine-grained control over thread scheduling, warp-level communication, and memory management. In addition, generic reverse-mode AD, which mechanically applies the chain rule, does not automatically exploit opportunities for algorithm-specific reuse of intermediate quantities in gradient calculation, such as structure factors in Ewald methods, and can therefore introduce avoidable overhead in the backward pass.

In this respect, we view that the current state of differentiable molecular mechanics resembles the historical evolution of deep learning itself, where major advances in practical performance were enabled not only by high-level compilers but also by specialized kernels and domain-specific libraries. For instance, the highly tuned convolution routines in cuDNN, and more recent FlashAttention[212–214] for Transformers. Molecular mechanics likewise requires

a dedicated computational backend if differentiable force fields are to achieve production-level efficiency. To address this need, we introduce TorchFF, a high-performance library that extends PyTorch with a suite of customized operators tailored to the algorithmic demands of molecular mechanics. TorchFF implements performance-critical components—including bonded and nonbonded interactions, multipolar Ewald/PME, and neighbor-list construction—as reusable backend-optimized primitives, while preserving a Python-level interface for flexible model development. In this work, we show that this design yields substantial performance gains over native PyTorch implementations with high-level JIT optimizations, and we further demonstrate how TorchFF enables the efficient construction and evaluation of an advanced force-field model for water in a condensed-phase simulation within a fully differentiable workflow.

4.2 Methodology

4.2.1 TorchFF Code Architecture

As illustrated in Figure 4.1, similar to many ML libraries such as PyG[215] and FlashAttention[216], TorchFF adopts a layered design that extends the functionality of the core PyTorch framework while preserving the usability and flexibility of the PyTorch programming model. In the lower C++ layer, computationally intensive molecular mechanics routines, including the energy (forward) and forces (backward) calculations, are implemented in customized CUDA kernels and are integrated with torch's automatic differentiation machinery and registered as "operators" through libtorch APIs. Once registered, these operators can be invoked through the `torch.ops` interface from Python and further wrapped as higher-level PyTorch modules, allowing them to be composed naturally with standard PyTorch components. Thus, users can easily leverage them and build their own FFs and apply them to MD simulation software like atomic simulation environment (ASE)[217] or OpenMM[218].

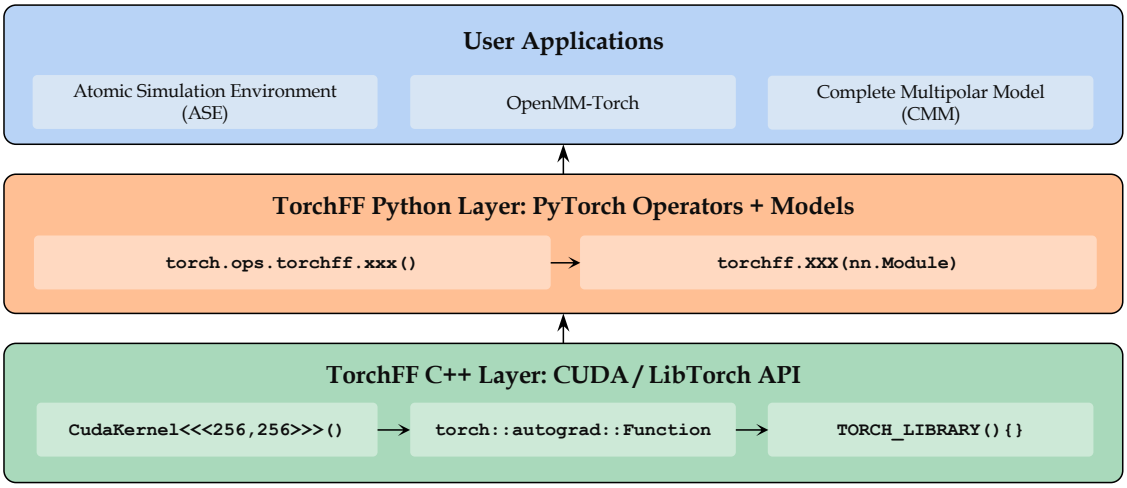

Figure 4.1: *Architecture of* ***TorchFF***. TorchFF implements CUDA-optimized primitives for critical components of many force fields including bonded and nonbonded interactions, multipolar Ewald/PME, and neighbor-list construction, while preserving a Python-level interface for flexible model development.

4.2.2 Classical Potentials

TorchFF supports a set of analytical potential forms widely used in molecular mechanics, spanning both bonded and nonbonded interactions. Although machine-learned interaction potentials (MLIPs) have become increasingly important, the physically motivated functional forms used in molecular mechanics remain essential because they provide interpretable energy decompositions[65], robust extrapolation behavior[56], and substantially lower computational cost. Furthermore, they are useful not only as standalone force-field models, but also as components of hybrid ML/physics frameworks and Δ–learning approaches[204–206].

The bonded interactions implemented in TorchFF include multiple bond and angle potentials (Eq. 1-6), including harmonic forms, anharmonic forms with higher-order polynomial corrections, the Morse bond potential, and the harmonic cosine angle potential.[198, 199] TorchFF also provides two classes of torsional potentials: periodic torsions for flexible dihedrals associated with rotatable bonds (Eq. 8), and harmonic improper torsions (Eq. 9) commonly used to enforce planarity around a central atom. Together, these terms describe the local geometric structure and intramolecular flexibility of molecular systems.

$$U_{\text{bond}} = \frac{1}{2}k_b(b - b_0)^2 \tag{S3.1}$$

$$U_{\text{bond}} = k_b(b - b_0)^2[1 - 2.55(b - b_0) + 3.793125(b - b_0)^2] \tag{S3.2}$$

$$U_{\text{bond}} = D[1 - \exp(-\sqrt{k_b/2D}(b - b_0)]^2 \tag{S3.3}$$

$$U_{\text{angle}} = \frac{1}{2}k_\theta(\theta - \theta_0)^2 \tag{S3.4}$$

$$\begin{aligned} U_{\text{angle}} = k_\theta(\theta - \theta_0)^2[1 - 0.014(\theta - \theta_0) + 5.6 \times 10^{-5}(\theta - \theta_0)^2 \\ - 7.0 \times 10^{7}(\theta - \theta_0)^3 + 2.2 \times 10^{-8}(\theta - \theta_0)^4] \end{aligned} \tag{S3.5}$$

$$U_{\text{angle}} = \frac{1}{2}k_\theta(\cos\theta - \cos\theta_0)^2 \tag{S3.6}$$

$$U_{\text{torsion}} = k_n\,(1 + \cos(n\phi - \phi_0)) \tag{S3.7}$$

$$U_{\text{torsion}} = \frac{1}{2}k_\phi(\phi - \phi_0)^2 \tag{S3.8}$$

For nonbonded interactions, TorchFF supports two van der Waals potentials that treat Pauli repulsion and dispersion in one formulation. The 12-6 Lennard-Jones potential parameterized against condensed-phase properties[197] is the default choice of almost all fixed-charge force fields (Eq. 9).[57–60] The buffered 14-7 van der Waals form (Eq. 10) is well-known as a component of AMOEBA[201] and MB-UCB[219] polarizable models, by introducing a "softer" repulsion region.

$$U_{6-12} = 4\varepsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right] \tag{S3.9}$$

$$U_{7-14} = \varepsilon_{ij}\left(\frac{1.07}{\rho_{ij} + 0.07}\right)^{7}\left(\frac{1.12}{\rho_{ij}^{7} + 0.12} - 2.0\right), \quad \rho_{ij} = r_{ij}/\sigma_{ij} \tag{S3.10}$$

TorchFF also supports a Slater-like potential (Eq. 11) and a pure Tang-Tonnies dispersion model (Eq. 12)[220], as they are increasingly adopted by modern QM-inspired force fields for their more physically grounded description of short-range overlap and long-range dispersion interactions.[73, 202, 203] In addition, TorchFF supports permanent multipolar electrostatics (Eq. 13) up to quadrupole moments[51], enabling a more accurate representation of anisotropic charge distributions than point-charge models and providing a foundation for advanced polarizable force fields[221] and models such as MB-Pol[222].

$$U_{\text{Slater}} = A_{ij}\left[\frac{1}{3}(B_{ij}r_{ij})^2 + B_{ij}r_{ij} + 1\right]\exp(-B_{ij}r_{ij}) \tag{S3.11}$$

$$U_{\text{disp}} = \sum_{i<j} -f_{6,ij}^{\text{TT}}(r_{ij})\frac{C_{6,ij}}{r_{ij}^6}, \quad f_{n,ij}^{\text{TT}}(r) = 1 - e^{-b_{ij}r}\sum_{k=0}^{n}\frac{(b_{ij}r)^k}{k!} \tag{S3.12}$$

$$U_{\text{multipole}} = \mathbf{M}_j^T\mathbf{T}_{ij}\mathbf{M}_i, \quad \mathbf{T}_{ij} = \begin{bmatrix} 1 & \nabla & \nabla^2 \\ \nabla & \nabla^2 & \nabla^3 \\ \nabla^2 & \nabla^3 & \nabla^4 \end{bmatrix}\left(\frac{1}{r_{ij}}\right) \tag{S3.13}$$

4.2.3 Ewald and Particle Mesh Ewald

Accurate and efficient calculation of long-range electrostatics interactions is indispensable in molecular simulation under periodic boundary conditions, because the Coulomb potential decays slowly as $1/r$, making a direct summation over all periodic images conditionally convergent and computationally prohibitive. The Ewald summation method[223, 224], developed in the last century, resolved this difficulty by splitting the interaction into short-range and long-range components that are evaluated in real and reciprocal space respectively. Recent work[206] has shown that augmenting ML models with Ewald formulations can eliminate artifacts that are caused by the locality assumption of MLIPs, thus improving the transferability and accuracy in describing target physics that is dominated by long-range interactions.

TorchFF provides high-performance differentiable implementations of the Ewald summation through the `torchff.Ewald` module and is briefly described here. Under three-dimensional periodic boundary conditions, the electrostatic energy of a unit cell with lattice matrix $\mathbf{A} = [\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3]$ can be written compactly as

$$U = \frac{1}{2}{\sum_{i,j,\boldsymbol{\ell}}}' \hat{L}_i\hat{L}_j\left(\frac{1}{|\mathbf{r}_i - \mathbf{r}_j + \mathbf{A}\boldsymbol{\ell}|}\right), \tag{S3.14}$$

$$\hat{L}_i = q_i + \boldsymbol{\mu}_i\cdot\nabla_i + \frac{1}{3}\boldsymbol{\Theta}_i\colon\nabla_i\nabla_i, \tag{S3.15}$$

where the prime indicates omission of the $i = j$, $\boldsymbol{\ell} = \mathbf{0}$ term. In the Ewald framework, this energy is decomposed into real-space, reciprocal-space, and self interaction contributions,

$$U = U_{\text{real}} + U_{\text{rec}} + U_{\text{self}}. \tag{S3.16}$$

The real-space contribution is

$$U_{\text{real}} = \frac{1}{2}{\sum_{i,j,\boldsymbol{\ell}}}' \hat{L}_i\hat{L}_j\frac{\operatorname{erfc}(\alpha|\mathbf{r}_j - \mathbf{r}_i + \mathbf{A}\boldsymbol{\ell}|)}{|\mathbf{r}_j - \mathbf{r}_i + \mathbf{A}\boldsymbol{\ell}|}, \tag{S3.17}$$

where α is the Ewald splitting parameter. Because the complementary-error-function damping renders this term short-ranged, it can be evaluated efficiently by direct summation within a real-space cutoff. Accordingly, in TorchFF the real-space part is handled by the multipolar real-space module embodied in Eq. (13), which provides implementation of properly damped interaction tensors for higher moments.

To express the reciprocal-space contribution, we define the reciprocal lattice as $\mathbf{A}^* = (\mathbf{A}^{-1})^T$ and introduce $\mathbf{m} = \mathbf{A}^*\mathbf{k}$ with $\mathbf{k} \in \mathbb{Z}^3$. The reciprocal-space Ewald energy is then

$$U_{\text{rec}} = \frac{1}{2\pi V} \sum_{\mathbf{m}\neq 0} \frac{\exp(-\pi^2\mathbf{m}^2/\alpha^2)}{\mathbf{m}^2} |S(\mathbf{m})|^2, \tag{S3.18}$$

where $V = \det(\mathbf{A})$ and the multipolar structure factor is

$$S(\mathbf{m}) = \sum_j \hat{L}_j \exp(2\pi i\, \mathbf{m}\cdot\mathbf{r}_j)\,. \tag{S3.19}$$

Owing to the Gaussian damping in Eq. (18), this reciprocal-space sum also decays rapidly and can therefore be truncated in practice. Finally, the remaining term removes the artificial self-interaction introduced by the reciprocal-space formulation. For permanent multipoles up to quadrupoles, it is given by

$$U_{\text{self}} = -\frac{\alpha}{\sqrt{\pi}} \sum_i \left(q_i^2 + \frac{2\alpha^2}{3}\, \boldsymbol{\mu}_i\cdot\boldsymbol{\mu}_i + \frac{8\alpha^4}{45}\, \boldsymbol{\Theta}_i : \boldsymbol{\Theta}_i \right). \tag{S3.20}$$

TorchFF exploits several methods to accelerate the evaluation of U_{rec} and $S(\mathbf{m})$ in the rate limiting steps of the reciprocal sum. For instance, only a nonredundant half-space of reciprocal vectors is explicitly enumerated, with the remaining half accounted for by symmetry. The structure factor $S(\mathbf{m})$, reciprocal vectors, and the Gaussian prefactors are cached to the torch autograd context after being calculated in the forward pass, in order to reuse them the backward pass. TorchFF also leverages a hierarchical memory and synchronization strategy to accelerate the computation of U_{rec} and $S(\mathbf{m})$: each thread block is assigned to one reciprocal vector, and atomic contributions are cooperatively reduced using warp-level and block-level synchronization with intermediate accumulation in block-level shared memory before the final write to global memory.

Although Ewald summation is formally exact, its computational cost becomes prohibitive for large systems as it scales at $O(N^{3/2} - N^2)$.[225] Particle Mesh Ewald (PME), and in particular its smooth variant[226], was further developed to overcome this limitation by replacing the explicit reciprocal-space summation with a mesh-based representation evaluated using fast Fourier transforms (FFTs), reducing the complexity to $O(N \log N)$.[227] Therefore, PME has become the standard treatment to long-range multipolar electrostatics in condensed-phase simulations for biomolecular and material systems in modern MD packages. In addition,

over the past two decades, PME methods have been generalized beyond point charges to higher electric moments, supporting modern advanced FFs with multipolar interaction and explicit polarization.[228–231] The importance of PME has also extended into machine-learning interatomic potentials (MLIPs). Recent work[206] has shown that augmenting ML models with Ewald formulations can eliminate artifacts that are caused by the locality assumption of MLIPs, thus improving the transferability and accuracy in describing target physics that is dominated by long-range interactions. In this sense, Ewald and PME are no longer only classical electrostatics algorithms, but also reusable computational primitives for the development of next-generation FFs and MLIPs.

Hence TorchFF also provides high-performance differentiable implementations of particle mesh Ewald (PME) for permanent multipoles up to quadrupoles through the `torchff.PME` module. More specifically, the complex exponential entering $S(\mathbf{m})$ is interpolated onto nearby grid points using B-splines, and the reciprocal-space convolution is then evaluated through FFTs. The resulting PME reciprocal-space energy can be written as

$$U_{\text{rec}} = \frac{1}{2V} \sum_{\mathbf{n}} \Phi(\mathbf{n})\, Q^R(\mathbf{n}), \tag{S3.21}$$

where $\Phi(\mathbf{n}) = \mathcal{F}^{-1}[G(\mathbf{k})\mathcal{F}[Q^R(\mathbf{n})]]$, $\mathcal{F}$ is the Fourier transform and $Q^R(\mathbf{n})$ denote the mesh multipole array:

$$Q^R(\mathbf{n}) = \sum_{i=1}^{N} \sum_{\ell_1,\ell_2,\ell_3} \hat{L}_i \left[\prod_{\beta=1}^{3} M_p(u_{\beta i} - n_\beta - \ell_\beta N_\beta) \right], \quad 0 \le n_\beta < N_\beta \tag{S3.22}$$

For brevity, only the essential expressions are given here for Ewald and PME. The complete derivations of the multipolar Ewald and PME formulas (including the forward and backward calculation), together with algorithmic details for their TorchFF implementations, are provided in the Supporting Information. Another important feature of TorchFF is that the Ewald and PME operators are not limited to returning energies. In the forward pass, they can also return the electrostatic potential and electric field, with full backward support for all returned quantities. This functionality is particularly useful in advanced force fields and MLIPs, where the potential or electric field may couple directly to other terms or serve as physically meaningful descriptors that should be differentiable.

4.2.4 Neighbor List

Neighbor lists identify atom pairs within a defined distance cutoff and are essential for the efficient evaluation of interactions in both classical force fields and machine-learned interatomic potentials (MLIPs). In classical models, they accelerate pairwise short-ranged nonbonded force calculations, while in MLIPs they define the local atomic environments over which message passing or local descriptors are constructed. By avoiding the $O(N^2)$ cost of

naive all-pairs enumeration, neighbor lists reduce the computational scaling to approximately linear with system size and are therefore a fundamental component of high-performance molecular simulation frameworks. TorchFF provides a neighbor list construction engine (`torchff.nblist.NeighborList`) that includes both a naive $O(N^2)$ brute-force algorithm and a cell-list method. The brute-force approach is primarily intended for small systems, whereas the cell-list algorithm achieves $O(N)$ scaling and is therefore more suitable for large-scale molecular dynamics simulations.

In the brute-force implementation, all atom pairs are examined and pairs within a user-defined cutoff distance are recorded. The computation is parallelized on the GPU by assigning each thread to evaluate a single atom pair, resulting in $N(N-1)/2$ parallel distance checks. For each pair, the thread reads the coordinates of the two atoms, computes the interatomic distance with respect to periodic boundary conditions, and determines whether the pair satisfies the cutoff. To efficiently accumulate detected neighbor pairs, intermediate results are first stored in block-level shared memory before being flushed to global memory. Since shared memory provides significantly lower latency than global memory, buffering the results in shared memory reduces the frequency of global atomic updates, thereby alleviating write contention among threads.

In the cell-list implementation, the neighbor-list construction proceeds in three main stages: (1) spatial binning and clustering, (2) identification of interacting clusters, and (3) construction of atomic neighbor pairs. In the first stage, atoms are spatially organized into clusters of 32 atoms to improve memory locality and parallel efficiency. The choice of 32 atoms per cluster matches the warp size of modern NVIDIA GPUs, enabling efficient use of warp-level synchronization primitives and avoiding costly block-level or global synchronization. This is achieved by partitioning the periodic simulation box into small spatial cells and assigning each atom to a cell according to its fractional coordinates. Atoms are then sorted by their cell indices so that atoms located in nearby regions of space are grouped together. The cell size is chosen as

$$l_{\text{cell}} = \sqrt[3]{\frac{32}{\rho}}, \tag{S3.23}$$

where ρ is the particle number density, such that the expected number of atoms in each cell is close to the cluster size.

In the second stage, possible interactions between clusters are determined using a bounding-sphere criterion. The bounding box and radius of each cluster are computed cooperatively by threads within a warp using warp-level reduction operations, which avoids explicit global synchronization and improves performance. For a pair of clusters i and j, the clusters are considered potentially interacting if

$$\|\mathbf{c}_i - \mathbf{c}_j\| \leq r_i + r_j + r_c, \tag{S3.24}$$

where r_c is the interaction cutoff distance, $\mathbf{c}_i$ is the center of cluster i, and r_i is the radius of

the bounding sphere enclosing the cluster. These quantities are defined as

$$\mathbf{c}_i = \mathbf{r}_{\text{anchor}} + \frac{\max_{a\in i}\{\mathbf{r}'_a\} + \min_{a\in i}\{\mathbf{r}'_a\}}{2}, \tag{S3.25}$$

$$r_i = \left\| \frac{\max_{a\in i}\{\mathbf{r}'_a\} - \min_{a\in i}\{\mathbf{r}'_a\}}{2} \right\|, \tag{S3.26}$$

where $\mathbf{r}'_a$ denotes the position of atom a relative to an anchor atom within cluster i, and $\|\cdot\|$ represents the Euclidean norm. Clusters that violate Eq. (S3.24) are sufficiently separated and cannot contain interacting atom pairs, allowing them to be discarded at this stage.

In the third and final stage, for each interacting cluster pair identified in the previous step, atomic neighbor pairs are evaluated. Each GPU warp (32 threads) processes a cluster pair. Each thread loads the coordinates of one atom from each cluster, and warp-level data exchange is used to systematically evaluate all $32 \times 32 = 1024$ possible atom pairs within the cluster pair. Pairs whose interatomic distance falls within the cutoff are recorded as neighbors.

The overall algorithmic complexity of the cell-list method can be approximated as

$$O(N_c^2 + mN) \tag{S3.27}$$

where $N_c = N/32$ is the number of clusters and $m \sim \frac{4}{3}\pi\rho r_c^3$ represents the average number of neighbors per atom within the cutoff radius r_c. The first term corresponds to the traversal of cluster pairs, whereas the second term arises from the evaluation of atom pairs within interacting clusters. In practice, the condition $mN \gg N_c^2$ typically holds, and the overall computational cost scales approximately linearly with the system size N, as will be demonstrated in the results section.

TorchFF neighbor list construction also supports exclusion criteria commonly used in classical force fields. Atom pairs within a specified topological distance (e.g., those connected through bonds, angles, or torsions) are excluded from non-bonded neighbor list construction because their interactions are handled by bonded terms[198]. Users can provide the excluded pairs as a sparse matrix in a user-friendly coordinate (COO) format, which is then transformed into a computation-friendly compressed sparse row (CSR) format internally and accessed by each GPU thread during pair evaluation.

4.2.5 Completely Multipolar Model (CMM)

The Completely Multipolar Model (CMM)[73, 203] is a physics-based many-body force-field framework designed to reproduce the individual components of the ALMO energy decomposition analysis (ALMO-EDA)[66]. In ALMO-EDA, the total quantum mechanical intermolecular interaction energy is decomposed into physically meaningful contributions,

$$E_{\text{int}} = E_{\text{Pauli}} + E_{\text{elec}} + E_{\text{disp}} + E_{\text{pol}} + E_{\text{CT}} \tag{S3.28}$$

corresponding to Pauli repulsion, permanent electrostatics, dispersion, polarization, and charge transfer. The CMM force field constructs a classical potential that mirrors this decomposition:

$$U_{\mathrm{CMM}} = U_{\mathrm{elec}} + U_{\mathrm{disp}} + U_{\mathrm{Pauli}} + U_{\mathrm{pol}} + U_{\mathrm{exch,pol}} + U_{\mathrm{CT}}^{\mathrm{direct}} + U_{\mathrm{CT}}^{\mathrm{indirect}} + U_{\mathrm{1b}} \tag{S3.29}$$

A key feature of CMM is that its nonbonded terms are expressed through multipolar interaction formulism up to the quardupoles with damped Coulomb interaction tensors:

$$U_{\mathrm{elec}} = \sum_{i<j} \left[\frac{Z_i Z_j}{r_{ij}} + Z_i \mathbf{T}_{ij}^{\mathrm{damp}} \mathbf{M}_j + \mathbf{M}_i \mathbf{T}_{ij}^{\mathrm{damp}} Z_j + \mathbf{M}_i \mathbf{T}_{ij}^{\mathrm{overlap}} \mathbf{M}_j \right] \tag{S3.30}$$

$$U_{\mathrm{Pauli}} = \sum_{i<j} \mathbf{M}_i^{\mathrm{Pauli}} \, \mathbf{T}_{ij}^{\mathrm{Pauli}} \, \mathbf{M}_j^{\mathrm{Pauli}} \tag{S3.31}$$

$$U_{\mathrm{Exch-pol}} = \sum_{i<j} \mathbf{M}_i^{\mathrm{Exch-pol}} \, \mathbf{T}_{ij}^{\mathrm{Exch-pol}} \, \mathbf{M}_j^{\mathrm{Exch-pol}} \tag{S3.32}$$

$$U_{\mathrm{CT}}^{\mathrm{direct}} = \sum_{i<j} \mathbf{M}_i^{\mathrm{CT,fwd}} \mathbf{T}_{ij}^{\mathrm{CT}} \mathbf{M}_i^{\mathrm{CT,bwd}} + \mathbf{M}_i^{\mathrm{CT,bwd}} \mathbf{T}_{ij}^{\mathrm{CT}} \mathbf{M}_i^{\mathrm{CT,fwd}} \tag{S3.33}$$

$$\mathbf{T}_{ij} = \begin{bmatrix} 1 & \nabla & \nabla^2 \\ \nabla & \nabla^2 & \nabla^3 \\ \nabla^2 & \nabla^3 & \nabla^4 \end{bmatrix} \left(\frac{f(r_{ij}, b_{ij})}{r_{ij}} \right) \tag{S3.34}$$

$$U_{\mathrm{disp}} = \sum_{i<j} -f_{6,ij}^{\mathrm{TT}}(r_{ij}) \frac{C_{6,ij}}{r_{ij}^6} \tag{S3.35}$$

where $f(r_{ij})$ terms are damping function featuring an exponential decay term modulated by specific polynomials. This unified formulation ensures the correct asymptotic behavior at long range while maintaining physically meaningful interactions in the short-range regime where electron densities overlap. Each interaction type possesses unique asymptotic behaviors and can be controlled via specific choices of formulating $f(\cdot)$ and the value of damping factor b. Furthermore, the permanent electrostatics V_{elec} takes charge penetration (CP) contribution into account through a "core-shell" model where Z_i is the core charge. The dispersion in CMM is given by the Tang-Toennies model, although the CMM formalism allows for many-body dispersion. CMM also defines a direct charge transfer model that adopts a "forward-backward" approach to capture the intrinsic bidirectional electron transfer process.

Many-body coupling in CMM is manifested in polarization, described through intramolec-

ular charge fluctuations and induced dipoles:

$$U_{\text{pol}} = \frac{1}{2}\sum_i \eta_i(\delta q_i)^2 + \sum_i \delta q_i V_i + \sum_A \lambda_A(\sum_{i\in A}\delta q_i - \Delta Q_A) + \frac{1}{2}\sum_i \boldsymbol{\mu}_i^{\mathrm{T}}\boldsymbol{\alpha}_i\boldsymbol{\mu}_i - \sum_i \boldsymbol{\mu}_i\cdot\boldsymbol{E}_i + \sum_{i<j}\boldsymbol{M}_i^{\text{ind}}\boldsymbol{T}_{ij}^{\text{pol}}\boldsymbol{M}_j^{\text{ind}} \tag{S3.36}$$

where η represents the atomic hardness, $\boldsymbol{\alpha}$ is the 3×3 anisotropic polarizability tensor, V_i and $\boldsymbol{E}_i$ denote the electric potential and field generated by CP-corrected permanent multipoles. $\boldsymbol{M}^{\text{ind}} = [\delta q, \boldsymbol{\mu}]$ denotes the induced multipole vector. ΔQ_A is the total charge change of water molecule A. In CMM, the charge rearrangements are only allowed within a molecule but not between molecules, which is enforced by the method of Lagrange multipliers. This constraint is also coherent with the definition of polarization energy in ALMO-EDA. The fluctuating charges $\{\delta q_i\}$ and induced dipoles $\{\boldsymbol{\mu}_i\}$ are obtained via variational method, i.e. minimizing equation (S3.36) by solving a linear system:

$$\begin{bmatrix} \mathbf{T}_{qq}^{\text{pol}} & \mathbf{T}_{q\mu}^{\text{pol}} & \mathbf{1} \\ \mathbf{T}_{\mu q}^{\text{pol}} & \mathbf{T}_{\mu\mu}^{\text{pol}} & \mathbf{0} \\ \mathbf{1}^{\dagger} & \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \delta\mathbf{q} \\ \boldsymbol{\mu} \\ \boldsymbol{\lambda} \end{bmatrix} = \begin{bmatrix} -\mathbf{V} \\ \mathbf{E} \\ \mathbf{\Delta Q} \end{bmatrix} \tag{S3.37}$$

Many-body charge transfer is contained in the indirect charge transfer energy through the coupling with polarization. The transferred charge due to charge transfer is taken proportional to the isotropic component of the direct CT interaction energy and thus the net charge change of atom i is defined as:

$$\Delta Q_i^{\text{CT}} = \frac{1}{\epsilon_{ij}}(V_{\text{iso},j\to i}^{\text{CT}} - V_{\text{iso},i\to j}^{\text{CT}}) \tag{S3.38}$$

which modifies the charge constraints entering the polarization equations. The resulting many-body contribution is therefore:

$$U_{\text{CT}}^{\text{indirect}} = U_{\text{pol}}(\boldsymbol{\Delta Q}^{\text{CT}}) - U_{\text{pol}}(\mathbf{0}) \tag{S3.39}$$

Besides the nonbonded interactions, CMM defines the one-body potential for water using a Morse bond potential, a cosine angle potential, a bond-bond coupling term and a bond-angle coupling term:

$$U_{\text{1b}} = U_{\text{bond}} + U_{\text{angle}} + U_{\text{bb}} + U_{\text{ba}} \tag{S3.40}$$

$$U_{\text{bond}} = D\left[1 - \exp\left(-\sqrt{\frac{k_{\text{b}}}{2D}}(R - R_{\text{eq}})^2\right)\right] \tag{S3.41}$$

$$U_{\text{angle}} = \frac{k_a}{2}(\cos\theta - \cos\theta_0)^2 \tag{S3.42}$$

$$U_{\text{bb}} = k_{\text{bb}}(R_1 - R_{1,\text{eq}})(R_2 - R_{2,\text{eq}}) \tag{S3.43}$$

$$U_{\text{ba}} = k_{\text{ba}}(R - R_{\text{eq}})(\cos\theta - \cos\theta_0) \tag{S3.44}$$

CMM also constructs coupling between the one-body potential[199] with the non-bonded interactions by (1) making the bond length equilibrium R_{eq} and k_b dependent on permanent electric field and the amount of transferred charge; (2) geometry-dependent charge flux:

$$R_{\text{eq}}(E_{\text{OH}}, \Delta Q_{\text{H}}^{\text{CT}}) = R_{\text{eq}}^0 + \frac{\mu^{(1)} E_{\text{OH}}}{k_b^0 - \mu^{(2)} E_{\text{OH}}} + k_{\text{ct}}^{(1)} (\Delta Q_{\text{H}}^{\text{CT}})^2 \tag{S3.45}$$

$$k_b(E_{\text{OH}}, \Delta Q_{\text{H}}^{\text{CT}}) = k_b^0 - 3k_b^0 \alpha \Delta R_{\text{eq}} - \mu^{(2)} E_{\text{OH}} + k_{\text{ct}}^{(2)} (\Delta Q_{\text{H}}^{\text{CT}})^2 \tag{S3.46}$$

$$dq_{\text{H1}} = j_b(R_{\text{OH1}} - R_0) + j_{bb}(R_{\text{OH2}} - R_0) + j_\theta(\cos\theta - \cos\theta_0) \tag{S3.47}$$

$$dq_{\text{H2}} = j_b(R_{\text{OH2}} - R_0) + j_{bb}(R_{\text{OH1}} - R_0) + j_\theta(\cos\theta - \cos\theta_0) \tag{S3.48}$$

$$dq_{\text{O}} = -(dq_{\text{H1}} + dq_{\text{H2}}) \tag{S3.49}$$

to ensure that dipole surface is well described. Further details of this advanced FF for water simulations is described in the original CMM paper.[73]

4.3 Results

All the simulation and benchmarks reported in this work were conducted on the NERSC Perlmutter supercomputing platform, using a single NVIDIA A100-SXM4-40GB GPU on a GPU node with one 64-core AMD EPYC 7763 CPU, with software based on Python 3.12, PyTorch 2.5.1, and TorchFF compiled using CUDA 12.9.

4.3.1 Performance

Classical Molecular Mechanics Potentials. We benchmarked the performance of TorchFF across the classical potential terms described in the Methods section and compared it against a reference native PyTorch implementation. All benchmarks were conducted using a consistent and carefully controlled protocol for both implementations. For each system size and potential, we performed 10 warm-up runs followed by 1000 production runs, and report the mean wall-clock time over the production runs. Each run consists of a full forward and backward pass, i.e., the evaluation of energies and coordinate gradients, thereby reflecting realistic molecular simulation workloads. To maximize performance, CUDA Graphs were employed to reduce kernel launch overhead, and the native PyTorch implementation was further optimized using `torch.compile` to enable just-in-time (JIT) compilation for operator fusion. Throughout the following sections, "native PyTorch" refers to implementations of the potential terms using only standard PyTorch tensor operations, without customized C++/CUDA kernels. This setup is designed to approximate the upper bound of performance achievable within the standard PyTorch framework.

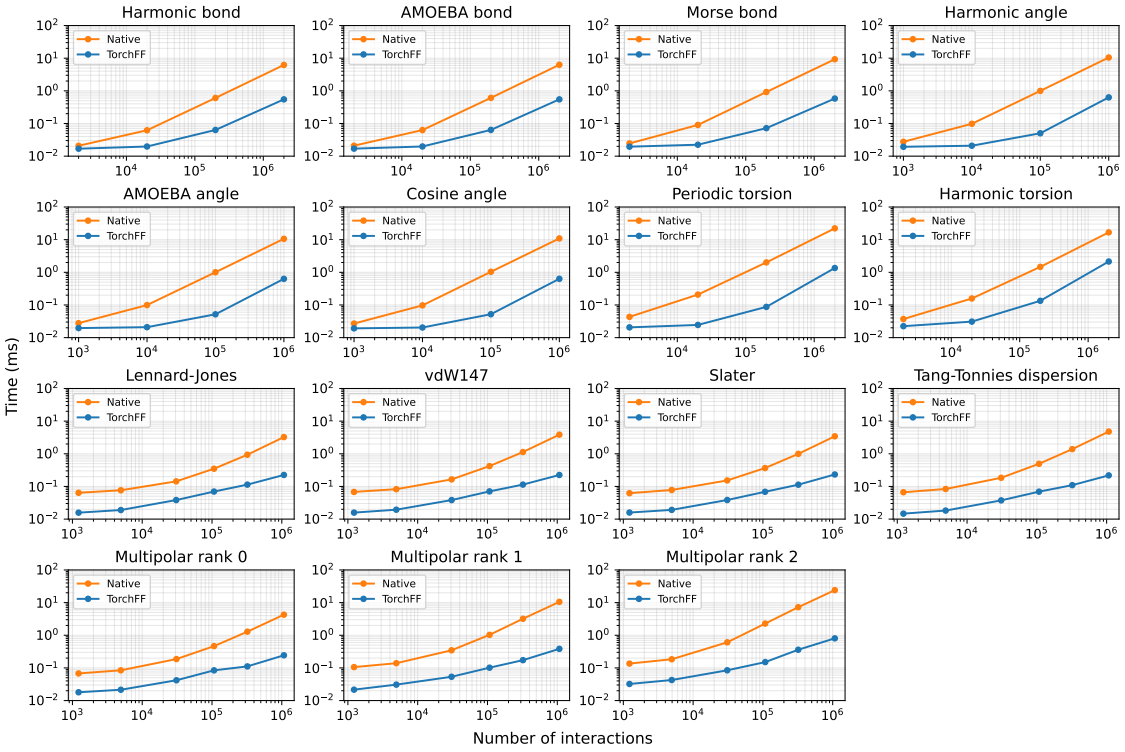

Figure 4.2: *Performance comparison between TorchFF and native PyTorch implementations for common bonded and non-bonded terms of classical molecular mechanics potentials.* The x-axis ("Number of interactions") denotes the total number of evaluated terms for each potential, i.e., the number of bonds, angles, torsions, or nonbonded interacting pairs included in the calculation.

The results comparing TorchFF against native PyTorch are shown in Figure 4.2. A comparison with the native PyTorch implementation is intended primarily as a simple reference for validating the numerical correctness of TorchFF rather than as a highly optimized baseline. Across all interaction types, TorchFF consistently outperforms the optimized native implementation, with speedups that increase systematically with system size, reflecting the benefits of kernel fusion and improved memory efficiency for computationally intensive, pairwise-dominated workloads. For bonded interactions (bonds, angles, and torsions), TorchFF achieves modest improvements of $1.5 \sim 3\times$ at small to medium system sizes ($10^3 \sim 10^4$ interactions), while the advantage grows to approximately $10 \sim 20\times$ for larger systems ($10^5 \sim 10^6$ interactions). Nonbonded interactions exhibit more pronounced performance gains, consistently achieving approximately $4 \sim 15\times$ acceleration across system sizes. This is because the nonbonded interactions involve more operations when accounting for periodic boundary conditions. The speedup becomes even more significant

as the arithmetic complexity increases, as demonstrated by multipolar interactions that include dipoles and quadrupoles (rank 1 and 2), where TorchFF reaches up to approximately $20 \sim 30\times$ acceleration at $10^5 \sim 10^6$ interacting pairs. These results indicate that, despite extensive optimization of the native PyTorch baseline, a substantial performance gap remains, highlighting the importance of backend-specialized CUDA kernels for achieving high-performance differentiable molecular mechanics.

Ewald and PME Electrostatics. We benchmarked the TorchFF implementations of Ewald summation and Particle Mesh Ewald (PME) across various system sizes and multipolar ranks, with results summarized in Figure 4.3. The benchmarks utilized systems containing 10^2 to 10^5 particles randomly distributed in a cubic box with a volume of 10 Å^3 per atom, approximating the density of liquid water. The Ewald separation parameter α and the reciprocal space cutoff $K_{\max}$ and mesh spacing for PME were selected to maintain a relative force accuracy of 10^{-6} using the approach implemented in OpenMM[218, 232]. We evaluated three multipolar ranks: Rank 0 (monopoles), Rank 1 (dipoles), and Rank 2 (quadrupoles). Reported performance metrics represent the mean execution time of 200 independent force evaluations after 20 warmup iterations.

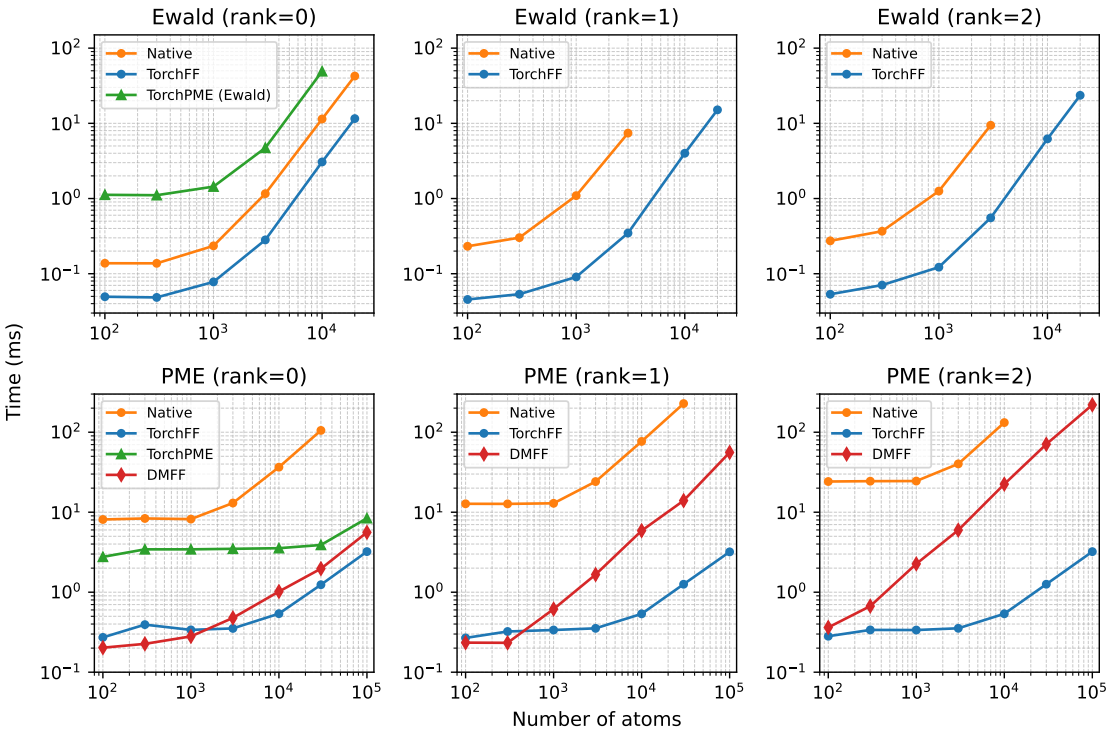

Figure 4.3: *TorchFF performance compared to native PyTorch, Torch-PME[233] and DMFF[20] Ewald and PME implementations.* Shown are nonbonded force-evaluation performance for Ewald (top) and PME (bottom) across cubic water boxes containing 100 to 10K TIP3P water molecules, and for multipolar ranks 0 (monopoles), Rank 1 (dipoles), and Rank 2 (quadrupoles). Note Torch-PME is only implemented for monopoles whereas DMFF is only implemented for PME.

The performance of the TorchFF customized kernels demonstrates a clear trade-off between Ewald and PME as a function of system size. For small systems ($N \leq 10^3$), Ewald summation remains more efficient than PME across all ranks due to the lower algorithmic overhead of the direct reciprocal sum compared to the 3D FFT and grid interpolation steps required for PME. For instance, at $N = 1000$ and Rank 2, the Ewald implementation (0.22 ms) is approximately 2× faster than PME (0.46 ms). However, PME exhibits superior $O(N \log N)$ scaling, overtaking the $O(N^2)$ Ewald summation between 10^3 and 10^4 atoms. At $N = 10^4$, PME is $5 \sim 9\times$ faster than Ewald for different ranks. Notably, in the small-system regime, the execution times for both methods remain nearly constant, indicating that performance is limited by kernel launch overhead and fixed memory orchestration rather than the numerical calculation itself.

Not surprisingly, because the native PyTorch implementation relies on generic tensor

operations and invokes many small GPU kernels, it is substantially worse than TorchFF performance, especially for the PME where algorithmic overhead is large. For instance, the TorchFF implementation is two order of magnitude faster ($\sim 180\times$) for PME at Rank 2 and $N = 10^4$, demonstrating the substantial gains enabled by low-level optimization. In addition, the native code contains many tensor broadcasts that generates large intermediate tensors, which easily triggers out-of-memory errors at larger system sizes.

We further compared TorchFF against two more performance optimized frameworks that provide differentiable Ewald and/or PME implementations: `torch-pme`[233] and Differentiable Molecular Force Field (DMFF)[20]. `torch-pme` is an optimized PyTorch-based implementation, but it is currently limited to monopolar electrostatics (Rank 0), whereas DMFF is a JAX-based framework that supports multipolar PME up to Rank 2 through XLA-based JIT compilation. Results show that `torch-pme` provides a faster PME implementation but slower Ewald than the native PyTorch implementation, while TorchFF is nearly $\sim 5 - 30\times$ faster across both Ewald and PME. DMFF is reasonably optimized for charge-only models , but its cost rises sharply with increasing multipolar rank. At Rank 2 and $N = 3 \times 10^4$, TorchFF requires only 1.41 ms, whereas DMFF takes 71.13 ms, corresponding to a speedup of more than 50-fold. This trend suggests that, although compiler-based approaches can accelerate high-level tensor expressions, they may still fail to produce machine code as efficient as hand-written CUDA kernels for complex multipolar PME workloads. By contrast, TorchFF maintains comparatively stable PME performance across ranks through kernel fusion, reduced memory traffic, and explicit control over GPU execution, making it a more suitable high-performance backend for advanced force field development.

Neighbor List. To evaluate the performance of the neighbor list construction engines, we benchmarked the TorchFF "brute-force" and "cell-list" implementations across a range of system sizes, with results summarized in Figure 4.4. While the brute-force kernel exhibits a slight performance advantage for very small systems ($\leq 10^3$ atoms) due to minimal setup overhead, the TorchFF cell-list algorithm demonstrates superior scaling, overtaking the brute-force approach at approximately 10^4 atoms. Beyond this crossover point, the performance gap widens rapidly: the cell-list implementation is $6.4\times$ faster at 3×10^4 atoms, $19.3\times$ faster at 10^5 atoms, and $127\times$ faster at 10^6 atoms. In contrast, the native PyTorch brute-force implementation (orange line, Figure 4.4) fails beyond 10^4 atoms due to out-of-memory (OOM) errors. This failure is a direct result of the $O(N^2)$ memory requirement of standard PyTorch broadcasting, which generates massive intermediate tensors for distance computing under PBC. Such scalability constraints underscore the necessity of implementing customized CUDA operators for neighbor list construction, as standard high-level library functions are inherently unable to manage the memory overhead required for large-scale atomistic systems.

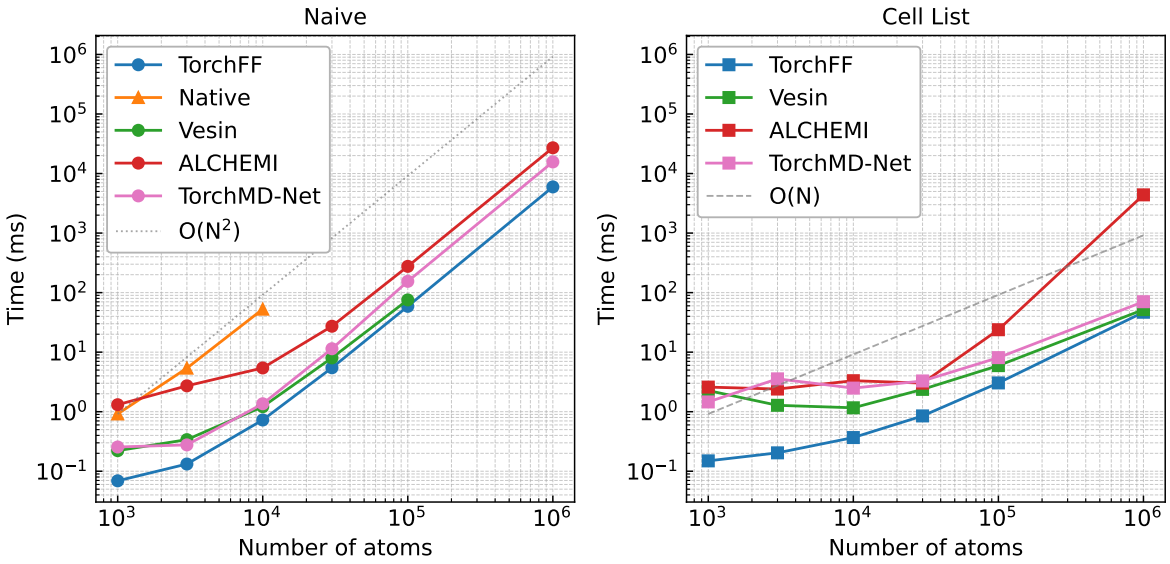

Figure 4.4: *Neighbor list evaluation performance of TorchFF compared to native PyTorch, Vesin[234], TorchMD-Net[208], and ALCHEMI[235] implementations.* The benchmarks utilized systems containing 10^3 to 10^6 particles randomly distributed in a cubic box with a density of 100 particles/nm^3, approximating liquid water under ambient conditions. A cutoff radius of $r_c = 0.8$ nm was employed. Reported performance represent the mean time of 10 independent runs after 10 warmup runs.

For comparison, we also compare against three other packages offering similar functionality—Vesin[234], TorchMD-Net[208], and ALCHEMI[235]—all of which primarily target the development and deployment of MLIPs. Vesin and TorchMD-Net adopt a similar development philosophy to TorchFF, utilizing customized CUDA operators exposed through a PyTorch API. ALCHEMI, developed by NVIDIA, is implemented on top of NVIDIA Warp, a framework that JIT-compiles Python kernels to CUDA. While this offers flexibility, the JIT process acts as an intermediate "translator" that may not produce the optimal machine code achievable through hand-written C++/CUDA primitives. The results show that TorchFF consistently exhibits the fastest performance and most favorable scaling. In the brute-force case, the performance gain of TorchFF can be attributed to the use of a writing buffer in block-level shared memory, which reduces access contention to global memory. Furthermore, TorchFF allows users to pre-allocate the neighbor list size based on estimated density (e.g., $500 \times N$ slots), avoiding the memory-prohibitive full allocation ($N(N-1)/2$) in other implementations such as Vesin.

Most notably, the TorchFF cell-list implementation is significantly faster (2–4×) than its counterparts in the small-to-medium regime. TorchFF reaches 0.368 ms at 10^4 atoms and

3.04 ms at 10^5 atoms, compared to 1.16/5.94 ms for Vesin and 2.49/8.09 ms for TorchMD-Net. This favorable performance is consistent with the TorchFF cell-list design, which focuses on reducing the search space while minimizing algorithmic overhead. As discussed previously (equation S3.27), although the theoretical leading term of the algorithm's complexity is N^2, the massive parallelism of modern GPUs efficiently handles the N_c^2 cluster-pair search, allowing the performance to scale near-linearly.

End-to-end Simulation of FFs with Increasing Complexity. We further evaluated TorchFF in end-to-end condensed-phase molecular dynamics simulations for three representative water models: the fixed-charge TIP3P model[197], the polarizable AMOEBA model[201], and the advanced many-body CMM model[73]. For the former two water models we compared their performance against OpenMM, a highly optimized production MD engine, with results summarized in Table 4.1 for different system sizes. The correctness of TorchFF's implementation is also validated by direct comparisons with OpenMM using the TIP3P and AMOEBA water models. The potential energies and atomic forces show close agreement within numerical errors across periodic water boxes of different sizes, and the equilibrium densities and radial distribution functions obtained from TorchFF- and OpenMM-based simulations are also consistent. Detailed results are provided in the Supporting Information. In all cases, OpenMM still outperforms TorchFF, with the gap being more pronounced for the simpler TIP3P model, where OpenMM is approximately 6–9× faster, whereas for the more complex AMOEBA model the difference is reduced to about 3–4× speedup. This trend is consistent with the design goal of TorchFF: as the force field becomes more sophisticated and the force evaluation contains a larger fraction of computation, the relative overhead of a differentiable PyTorch-based workflow becomes less dominant. At the same time, TorchFF substantially outperforms the native PyTorch implementations for both TIP3P and AMOEBA, typically by about one order of magnitude, demonstrating that customized CUDA kernels are even more critical when measured within a full end-to-end MD workflow rather than in isolated force evaluations. Furthermore, we note that this comparison is not strictly one-to-one. OpenMM is a tightly integrated MD engine with additional workflow-level optimizations beyond force evaluation itself, including Verlet-style neighbor-list buffering and infrequent list rebuilds, highly optimized coupling between neighbor-list construction and nonbonded kernels, and a lower PME interpolation order than used here (order 4 in OpenMM versus order 6 in TorchFF). By contrast, the present TorchFF simulations were executed through the `openmm-torch` interface except for the native PyTorch implementation for AMOEBA which uses `torch-sim` package[236] because we found it faster than `openmm-torch` in this case. This couples the an extra MD integrator to a `torch.nn.Module` and therefore introduces extra communication overhead between the MD driver and the force computation. This suggests that there remains significant room for further improvement through MD engine-level optimizations tailored specifically to differentiable molecular dynamics workflows.

Table 4.1: *Comparison of end-to-end simulation performance for water models with increasing complexity reported in ns/day.* All simulations were carried out in the NVE ensemble for 10,000 steps with a 1 fs time step. The non-bonded interaction cutoff was 0.8 nm for TIP3P and AMOEBA and 0.9 nm for CMM. Mutual polarization was turned off in AMOEBA to avoid any algorithmic differences regarding SCF-solver, making the comparison more reflecting the energy/force calculation kernels.

Number of water	Force Field	Native PyTorch (PME)	OpenMM (PME)	TorchFF (PME)
300	TIP3P	15.38	977.34	138.43
1000	TIP3P	11.60	603.34	104.78
3000	TIP3P	3.43	605.46	64.75
300	AMOEBA	1.36	168.27	44.49
1000	AMOEBA	0.99	99.95	23.50
3000	AMOEBA	0.40	51.20	14.96
Number of water	Force Field	Native PyTorch (Ewald)	TorchFF (Ewald)	TorchFF (PME)
300	CMM	0.95	5.21	4.16
1000	CMM	0.27	2.83	2.86
3000	CMM	N/A	0.55	0.67

Here we further demonstrate the utility of TorchFF by implementing it for the more complex CMM model many-body force field and applying it to condensed-phase MD simulations. Although current molecular mechanics and molecular dynamic packages are designed to be as modular as possible by providing low-level APIs for users to customize the potential, the novel functional forms of the CMM potential present significant implementation challenges since no existing MD software supports all the terms. The ubiquitous intricate coupling between terms renders the manual derivation and implementation of analytical forces extremely laborious and error-prone. While the original CMM implementation was developed in Julia and limited to non-periodic gas-phase clusters, we utilized functionalities in TorchFF, including the bonded kernels, multipolar interaction kernels and Ewald/PME to construct an implementation (hereafter "PyCMM") within the TorchFF framework. PyCMM integrates the long-range contributions to both permanent electrostatics and polarization interactions through Ewald or PME summation. The fluctuating charges and induced dipoles in Eq. 37 are obtained iteratively using a preconditioned conjugate-gradient (PCG) solver. The response operator ("polarization matrix") was explicitly constructed and applied to the trial induced multipoles during each PCG iteration. For the periodic box containing 216 water molecules used in our tests, the PCG solver typically converges within 5 iterations. Crucially, the automatic differentiation (AD) capabilities of PyTorch are leveraged to compute forces directly from the potential energy.

We interfaced PyCMM with the Atomic Simulation Environment[217] (ASE) as the MD driver. Figure 4.5 illustrates the energy conservation when simulating the CMM water

model with PyCMM. The system exhibits excellent stability, with a negligible drift rate of 1.1×10^{-3} kJ/mol/ps over a 1-ns trajectory (polarization solver tolerance of 10^{-7} a.u.) in line with other reported simulations. These results verify the numerical correctness of the implementation and the smoothness of the CMM potential surface, particularly the customized CUDA operators in TorchFF, where the backward operations were manually implemented. Further results on water properties will be reported in a future publication.

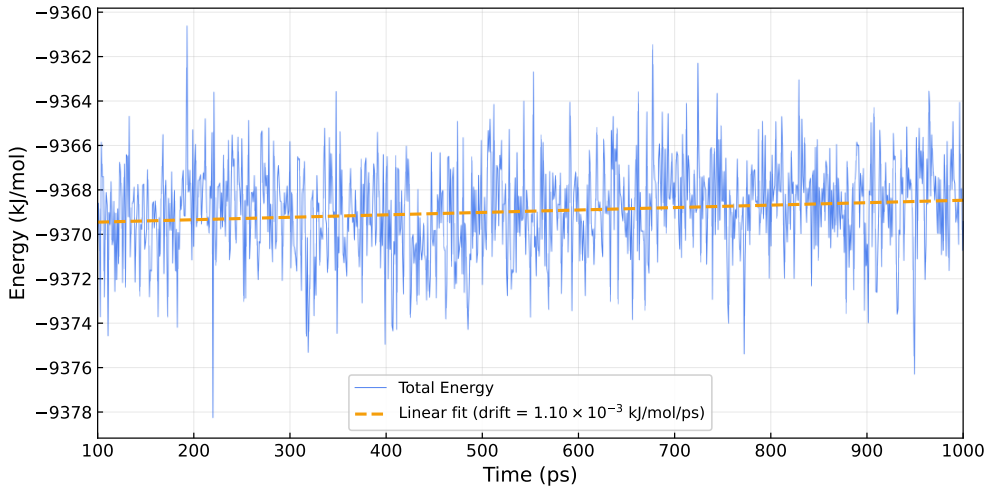

Figure 4.5: *PyCMM simulation of the CMM water model.* Shown is energy conservation in the NVE simulation of a cubic box containing 216 water molecules with a 1-fs timestep to demonstrate correctness of the implementation.

We also compared the speed performance of PyCMM with and without TorchFF customized CUDA operators and results are shown in Table 4.1. For 300 and 1000 water molecules, TorchFF achieves remarkable speedups of ∼5.5× and ∼10×, respectively, relative to the native PyTorch implementation. In the small-system regime, PME is slightly less efficient than direct Ewald (e.g., 4.16 vs 5.21 ns/day at 300 waters), reflecting the additional algorithmic overhead associated with mesh construction and FFTs. As system size increases, this overhead is amortized, and PME becomes competitive or slightly more favorable (e.g., 0.67 vs 0.55 ns/day at 3000 waters). Notably, the native PyTorch implementation fails at 3000 waters due to GPU out-of-memory limitations, whereas TorchFF remains tractable. These results represent a proof-of-concept demonstration, as only selected kernels were replaced by TorchFF operators while the overall simulation workflow remains largely unoptimized. Therefore, further performance gains of CMM are expected with deeper integration and end-to-end optimization.

4.4 Conclusions and Outlook

In this work, we presented TorchFF, a high-performance differentiable force-field library that extends PyTorch with customized C++/CUDA operators for molecular mechanics. TorchFF was designed to address a central limitation of existing automatic-differentiation frameworks for atomistic simulation: although these environments provide great flexibility for model development, they often incur substantial overhead when applied to the many-atom, small-kernel workloads that dominate force-field calculations. By implementing performance-critical routines—including bonded interactions, multipolar electrostatics, PME, and neighbor-list construction—as backend-optimized primitives, TorchFF combines the programmability of the PyTorch ecosystem with execution characteristics that are much closer to those of specialized compiled MD engines. The benchmark results show that this strategy yields substantial gains in efficiency and scalability for popular fixed charge and polarizable models.

Beyond performance, TorchFF is intended as an enabling infrastructure for the development of next-generation force fields where rapid model development are essential but highly complex. Because its operators remain fully compatible with automatic differentiation, the library supports both molecular simulation and gradient-based force-field optimization within a unified software environment. We further demonstrated this capability by implementing the Completely Multipolar Model and applying it to condensed-phase water simulations, showing that TorchFF can support sophisticated many-body, physics-based potentials while retaining numerical stability and favorable computational performance. Future development will focus on expanding the set of supported functional forms and further narrowing the performance gap with production MD packages. More broadly, since TorchFF provides a reusable computational backend for hybrid ML/physics models in which analytical long-range interactions and differentiable short-range models must coexist efficiently, we will also work with the community to increase interoperability with emerging differentiable simulation engines[236, 237] and training workflows[20] by replacing selected performance-critical low-level implementations in existing packages with TorchFF operators. This would allow users to benefit from the performance improvements of TorchFF without substantially modifying their existing workflows or adopting a new software stack.

4.5 Supporting Information

4.5.1 Derivation of Multipolar Ewald equations

For a system containing point charges, Cartesian dipoles, and quadrupoles, the reciprocal-space energy defined by Ewald summation is

$$U_{\text{rec}} = \frac{1}{2}\sum_{i,j,\boldsymbol{\ell}} \hat{L}_i \hat{L}_j \frac{\operatorname{erf}(\alpha|\mathbf{r}_i - \mathbf{r}_j + \mathbf{A}\boldsymbol{\ell}|)}{|\mathbf{r}_i - \mathbf{r}_j + \mathbf{A}\boldsymbol{\ell}|}, \tag{S4.1}$$

$$\hat{L}_i = q_i + \boldsymbol{\mu}_i \cdot \nabla_i + \frac{1}{3}\boldsymbol{\Theta}_i : \nabla_i \nabla_i. \tag{S4.2}$$

If we define the generalized potential

$$V(\mathbf{r}) = \sum_{j,\boldsymbol{\ell}} \hat{L}_j \frac{\operatorname{erf}(\alpha|\mathbf{r} - \mathbf{r}_j + \mathbf{A}\boldsymbol{\ell}|)}{|\mathbf{r} - \mathbf{r}_j + \mathbf{A}\boldsymbol{\ell}|}, \tag{S4.3}$$

then, since $V(\mathbf{r})$ is periodic and satisfies $V(\mathbf{r} + \mathbf{A}\boldsymbol{\ell}) = V(\mathbf{r})$, we can expand it as a Fourier series:

$$V(\mathbf{r}) = \frac{1}{V}\sum_{\mathbf{m}} \hat{V}(\mathbf{m}) \exp(2\pi i\, \mathbf{m}\cdot\mathbf{r}), \quad \mathbf{m} = \mathbf{A}^{-1}\mathbf{k}, \quad V = \det(\mathbf{A}), \tag{S4.4}$$

$$\hat{V}(\mathbf{m}) = \int_A V(\mathbf{r}) \exp(-2\pi i\, \mathbf{m}\cdot\mathbf{r})\, d\mathbf{r} \tag{S4.5}$$

$$= \sum_{j,\boldsymbol{\ell}} \int_A \hat{L}_j \frac{\operatorname{erf}(\alpha|\mathbf{r} - \mathbf{r}_j + \mathbf{A}\boldsymbol{\ell}|)}{|\mathbf{r} - \mathbf{r}_j + \mathbf{A}\boldsymbol{\ell}|} \exp(-2\pi i\, \mathbf{m}\cdot\mathbf{r})\, d\mathbf{r} \tag{S4.6}$$

$$= \sum_{j} \hat{L}_j \int_{\mathbb{R}^3} \frac{\operatorname{erf}(\alpha|\mathbf{r} - \mathbf{r}_j|)}{|\mathbf{r} - \mathbf{r}_j|} \exp(-2\pi i\, \mathbf{m}\cdot\mathbf{r})\, d\mathbf{r} \tag{S4.7}$$

$$= \sum_{j=1}^{N} \hat{L}_j \exp(-2\pi i\, \mathbf{m}\cdot\mathbf{r}_j) \int_{\mathbb{R}^3} \frac{\operatorname{erf}(\alpha|\mathbf{r}|)}{|\mathbf{r}|} \exp(-2\pi i\, \mathbf{m}\cdot\mathbf{r})\, d\mathbf{r} \tag{S4.8}$$

$$= \frac{1}{2\pi}\frac{\exp(-\pi^2\mathbf{m}^2/\alpha^2)}{\mathbf{m}^2} \sum_{j} \hat{L}_j \exp(-2\pi i\, \mathbf{m}\cdot\mathbf{r}_j). \tag{S4.9}$$

Then the reciprocal-space energy becomes

$$\begin{aligned} U_{\text{rec}} &= \frac{1}{2}\sum_i \hat{L}_i V(\mathbf{r}_i) \\ &= \frac{1}{2\pi V}\sum_{\mathbf{m}} \frac{\exp(-\pi^2\mathbf{m}^2/\alpha^2)}{\mathbf{m}^2} \sum_{i,j} \hat{L}_i\hat{L}_j \exp(2\pi i\,\mathbf{m}\cdot(\mathbf{r}_i-\mathbf{r}_j)) \\ &= \frac{1}{2\pi V}\sum_{\mathbf{m}} \frac{\exp(-\pi^2\mathbf{m}^2/\alpha^2)}{\mathbf{m}^2} |S(\mathbf{m})|^2. \end{aligned} \tag{S4.10}$$

where

$$S(\mathbf{m}) = \sum_j \hat{L}_j \exp(2\pi i\,\mathbf{m}\cdot\mathbf{r}_j) \tag{S4.11}$$

$$= \sum_j L_j(\mathbf{m}) \exp(2\pi i\,\mathbf{m}\cdot\mathbf{r}_j) \tag{S4.12}$$

$$L_j(\mathbf{m}) = q_j + 2\pi i\,\boldsymbol{\mu}_j\cdot\mathbf{m} - \frac{4\pi^2}{3}\mathbf{m}^T\boldsymbol{\Theta}_j\mathbf{m} \tag{S4.13}$$

The gradients of the energy with respect to the multipoles and coordinates can then be written as

$$\frac{\partial U_{\text{rec}}}{\partial q_j} = \text{Re}\left[\frac{1}{\pi V}\sum_{\mathbf{m}\neq 0}\frac{\exp(-\pi^2\mathbf{m}^2/\alpha^2)}{\mathbf{m}^2} e^{-2\pi i\,\mathbf{m}\cdot\mathbf{r}_j} S(\mathbf{m})\right] \tag{S4.14}$$

$$\frac{\partial U_{\text{rec}}}{\partial \boldsymbol{\mu}_j} = \text{Re}\left[-\frac{2i}{V}\sum_{\mathbf{m}\neq 0}\mathbf{m}\frac{\exp(-\pi^2\mathbf{m}^2/\alpha^2)}{\mathbf{m}^2}\, e^{-2\pi i\,\mathbf{m}\cdot\mathbf{r}_j}\, S(\mathbf{m})\right] \tag{S4.15}$$

$$\frac{\partial U_{\text{rec}}}{\partial \boldsymbol{\Theta}_j} = \text{Re}\left[\frac{4\pi}{3V}\sum_{\mathbf{m}\neq 0}\mathbf{m}\mathbf{m}^T\frac{\exp(-\pi^2\mathbf{m}^2/\alpha^2)}{\mathbf{m}^2}\, e^{-2\pi i\,\mathbf{m}\cdot\mathbf{r}_j}\, S(\mathbf{m})\right] \tag{S4.16}$$

$$\frac{\partial U_{\text{rec}}}{\partial \mathbf{r}_j} = \text{Re}\left[-\frac{2i}{V}\sum_{\mathbf{m}\neq 0}\mathbf{m}\frac{\exp(-\pi^2\mathbf{m}^2/\alpha^2)}{\mathbf{m}^2}\, e^{-2\pi i\,\mathbf{m}\cdot\mathbf{r}_j}\, S(\mathbf{m})L_j^{\dagger}(\mathbf{m})\right] \tag{S4.17}$$

Require: Atomic coordinates $\{\mathbf{r}_j\}$, multipoles $\{q_j, \boldsymbol{\mu}_j, \boldsymbol{\Theta}_j\}$, box matrix $\mathbf{A}$, reciprocal cutoff K, and Ewald parameter α
Ensure: Reciprocal-space energy U_{rec} and its derivatives with respect to multipoles and coordinates

Forward Step I: Prepare reciprocal-space vectors
for each thread **do** ▷ TorchFF kernel
    Determine $\mathbf{m}$ and compute Gaussian prefactor in Eq. (S4.10) from thread ID
    Write them to global memory
end for

Forward Step II: Evaluate $S(\mathbf{m})$ and U_{rec}
for each thread block **do** ▷ TorchFF kernel
    Load $\mathbf{m}$ and prefactor
    for each thread in the block **do**
        Loop over atoms j in a strided manner
        Read $\mathbf{r}_j$ and atomic multipoles
        Compute and accumulate the contribution to $S(\mathbf{m})$ with Eq.(S4.12)
    end for
    Reduce $S(\mathbf{m})$ within the block using warp-level sync and block-level shared memory.
    Write $S(\mathbf{m})$ to global memory
    Accumulate the contribution to U_{rec} using Eq. (S4.10)
end for

Forward Step III: Save intermediate quantities
Save coordinates, multipoles, reciprocal vectors, structure factors, and output buffers
return U_{rec}

Backward: Accumulate derivatives
for each thread block **do** ▷ TorchFF kernel
    Read $\mathbf{r}_i$ and atomic multipoles
    Initialize block-local partial sums for multipole and coordinate derivatives
    for each thread in the block **do**
        Loop over reciprocal vectors $\mathbf{m}$ in a strided manner
        Read $\mathbf{m}$ and $S(\mathbf{m})$
        Compute and accumulate contributions to gradients using Eqs.(S4.14)–(S4.17)
    end for
    Reduce gradients within the block using warp-level sync and block-level shared memory.
    Write gradients to global memory
end for

Algorithm 1: Reciprocal-space multipolar Ewald in TorchFF

4.5.2 Derivation of Multipolar PME equations

Prefactors introduced by the B-spline. The sixth-order B-spline is defined as

$$M_6(x) = \frac{1}{5!}\sum_{k=0}^{6}(-1)^k \binom{6}{k} (x-k)_+^5, \tag{S4.18}$$

$$(x-k)_+^5 = \begin{cases} (x-k)^5, & x > k, \\ 0, & x \le k. \end{cases} \tag{S4.19}$$

Given the numbers of grid points N_1, N_2, and N_3, we first denote the scaled fractional coordinates of atom i by $\mathbf{u}_i$, i.e.,

$$u_{\beta i} = N_\beta \mathbf{a}_\beta^* \cdot \mathbf{r}_i. \tag{S4.20}$$

The exponential factor in Eq. (S4.11) can be approximated by

$$\exp\left(2\pi i \frac{k_\beta}{N_\beta} u_\beta\right) \approx b\left(\frac{k_\beta}{N_\beta}\right) \sum_{n'_\beta \in \mathbb{Z}} M_p(u_\beta - n'_\beta) \exp\left(2\pi i \frac{k_\beta}{N_\beta} n'_\beta\right). \tag{S4.21}$$

The coefficient $b(k_\beta/N_\beta)$ is determined by requiring the two sides of the above equation to be equal at $u_\beta = p$:

$$\exp\left(2\pi i \frac{k_\beta}{N_\beta} p\right) = b\left(\frac{k_\beta}{N_\beta}\right) \sum_{n'_\beta \in \mathbb{Z}} M_p(p - n'_\beta) \exp\left(2\pi i \frac{k_\beta}{N_\beta} n'_\beta\right). \tag{S4.22}$$

Since $M_p(x)$ is nonzero only on $0 < x < p$ (i.e., it has compact support) and is symmetric, $M_p(x) = M_p(p-x)$, the above sum involves only $p-1$ terms:

$$\exp\left(2\pi i \frac{k_\beta}{N_\beta} p\right) = b\left(\frac{k_\beta}{N_\beta}\right) \sum_{n'_\beta=1}^{p-1} M_p(n'_\beta) \exp\left(2\pi i \frac{k_\beta}{N_\beta} n'_\beta\right), \tag{S4.23}$$

$$\exp\left(2\pi i \frac{k_\beta}{N_\beta} (p-1)\right) = b\left(\frac{k_\beta}{N_\beta}\right) \sum_{n'_\beta=0}^{p-2} M_p(n'_\beta + 1) \exp\left(2\pi i \frac{k_\beta}{N_\beta} n'_\beta\right), \tag{S4.24}$$

$$b\left(\frac{k_\beta}{N_\beta}\right) = \exp\left(2\pi i \frac{k_\beta}{N_\beta}(p-1)\right) \left[\sum_{n'_\beta=0}^{p-2} M_p(n'_\beta + 1) \exp\left(2\pi i \frac{k_\beta}{N_\beta} n'_\beta\right)\right]^{-1}. \tag{S4.25}$$

For simplicity, we define the prefactor introduced by the B-spline as

$$b(\mathbf{k}) = b\left(\frac{k_1}{N_1}\right) b\left(\frac{k_2}{N_2}\right) b\left(\frac{k_3}{N_3}\right). \tag{S4.26}$$

Derivation of the PME reciprocal energy. Substituting Eqs. (S4.21) and (S4.26) into Eq. (S4.11), we obtain

$$S(\mathbf{m}) = b(\mathbf{k}) \sum_j \sum_{n_1',n_2',n_3'} \hat{L}_j[M_p(u_{1j} - n_1') \times M_p(u_{2j} - n_2') \times M_p(u_{3j} - n_3')] \times \exp\left[2\pi i \left(\frac{k_1 n_1'}{N_1} + \frac{k_2 n_2'}{N_2} + \frac{k_3 n_3'}{N_3}\right)\right]. \tag{S4.27}$$

Because $n_\beta' \in \mathbb{Z}$, we can reorganize the summation index as

$$\sum_{n_\beta'} f(n_\beta') = \sum_{n_\beta=0}^{N_\beta - 1} \sum_{\ell_\beta} f(n_\beta + \ell_\beta N_\beta). \tag{S4.28}$$

Therefore,

$$\begin{aligned} S(\mathbf{m}) = b(\mathbf{k}) \sum_j \sum_{\ell_1,\ell_2,\ell_3} \sum_{n_1=0}^{N_1-1} \sum_{n_2=0}^{N_2-1} \sum_{n_3=0}^{N_3-1} \hat{L}_j[M_p(u_{1j} - n_1 - \ell_1 N_1) \\ \times M_p(u_{2j} - n_2 - \ell_2 N_2) \times M_p(u_{3j} - n_3 - \ell_3 N_3)] \\ \times \exp\left[2\pi i \left(\frac{k_1 n_1}{N_1} + \frac{k_2 n_2}{N_2} + \frac{k_3 n_3}{N_3} + k_1\ell_1 + k_2\ell_2 + k_3\ell_3\right)\right]. \end{aligned} \tag{S4.29}$$

If we define the generalized multipolar array $Q^R(\mathbf{n})$ on each grid point by

$$Q^R(\mathbf{n}) = \sum_j \sum_{\ell_1,\ell_2,\ell_3} \hat{L}_j \left[\prod_{\beta=1}^{3} M_p(u_{\beta j} - n_\beta - \ell_\beta N_\beta)\right], \tag{S4.30}$$

then, using the periodicity of the exponential, we obtain

$$S(\mathbf{m}) = b\,(\mathbf{k}) \sum_{n_1=0}^{N_1-1} \sum_{n_2=0}^{N_2-1} \sum_{n_3=0}^{N_3-1} Q^R(\mathbf{n}) \exp\left[2\pi i \left(\frac{k_1 n_1}{N_1} + \frac{k_2 n_2}{N_2} + \frac{k_3 n_3}{N_3}\right)\right]. \tag{S4.31}$$

Note that $S(\mathbf{m})$ is the discrete Fourier transform of $Q^R(\mathbf{n})$ scaled by the prefactor $b(\mathbf{k})$, i.e.,

$$\mathcal{F}[Q^R(\mathbf{n})] \equiv Q^F(\mathbf{k}) = \sum_{n_1=0}^{N_1-1} \sum_{n_2=0}^{N_2-1} \sum_{n_3=0}^{N_3-1} Q^R(\mathbf{n}) \exp\left[2\pi i \left(\frac{k_1 n_1}{N_1} + \frac{k_2 n_2}{N_2} + \frac{k_3 n_3}{N_3}\right)\right], \tag{S4.32}$$

so that

$$S(\mathbf{m}) = b(\mathbf{k}) Q^F(\mathbf{k}). \tag{S4.33}$$

We denote the Gaussian prefactor from Ewald summation together with the prefactor from B-spline interpolation by $G^F(\mathbf{k})$:

$$G^F(\mathbf{k}) = \frac{1}{\pi V}\frac{\exp(-\pi^2\mathbf{m}^2/\alpha^2)}{\mathbf{m}^2}|b(\mathbf{k})|^2. \tag{S4.34}$$

The total reciprocal-space energy in PME is then

$$\begin{aligned} U_{\text{rec}} &= \frac{1}{2\pi V}\sum_{\mathbf{m}}\frac{\exp(-\pi^2\mathbf{m}^2/\alpha^2)}{\mathbf{m}^2}|S(\mathbf{m})|^2 \\ &= \frac{1}{2}\sum_{\mathbf{k}}|Q^F(\mathbf{k})|^2 G^F(\mathbf{k}). \end{aligned} \tag{S4.35}$$

In practice, the PME energy is actually evaluated on the real-space grid:

$$U_{\text{rec}} = \frac{1}{2}\sum_{\mathbf{n}}\Phi(\mathbf{n})Q^R(\mathbf{n}), \tag{S4.36}$$

where

$$\Phi(\mathbf{n}) = \mathcal{F}^{-1}[G^F(\mathbf{k})\mathcal{F}[Q^R(\mathbf{n})]]. \tag{S4.37}$$

Derivation of the gradients.

$$\frac{\partial U_{\text{rec}}}{\partial X} = \sum_{\mathbf{n}}\Phi(\mathbf{n})\frac{\partial Q^R(\mathbf{n})}{\partial X}. \tag{S4.38}$$

Therefore,

$$\frac{\partial U_{\text{rec}}}{\partial q_j} = \sum_{\mathbf{n}}\Phi(\mathbf{n})\sum_{\ell_1,\ell_2,\ell_3}\left[\prod_{\beta=1}^{3}M_p(u_{\beta j}-n_\beta-\ell_\beta N_\beta)\right] \tag{S4.39}$$

$$\frac{\partial U_{\text{rec}}}{\partial \boldsymbol{\mu}_j} = \sum_{\mathbf{n}}\Phi(\mathbf{n})\sum_{\ell_1,\ell_2,\ell_3}\nabla_j\left[\prod_{\beta=1}^{3}M_p(u_{\beta j}-n_\beta-\ell_\beta N_\beta)\right] \tag{S4.40}$$

$$\frac{\partial U_{\text{rec}}}{\partial \boldsymbol{\Theta}_j} = \sum_{\mathbf{n}}\Phi(\mathbf{n})\sum_{\ell_1,\ell_2,\ell_3}\frac{1}{3}\nabla_j^2\left[\prod_{\beta=1}^{3}M_p(u_{\beta j}-n_\beta-\ell_\beta N_\beta)\right] \tag{S4.41}$$

$$\frac{\partial U_{\text{rec}}}{\partial \mathbf{r}_j} = \sum_{\mathbf{n}}\Phi(\mathbf{n})\sum_{\ell_1,\ell_2,\ell_3}\left(q_j\nabla_j + \boldsymbol{\mu}_j^{\mathrm{T}}\nabla_j\nabla_j + \frac{1}{3}\boldsymbol{\Theta}_j\nabla_j^3\right)\left[\prod_{\beta=1}^{3}M_p(u_{\beta j}-n_\beta-\ell_\beta N_\beta)\right] \tag{S4.42}$$

Require: Atomic coordinates $\{\mathbf{r}_j\}$, multipoles $\{q_j, \boldsymbol{\mu}_j, \boldsymbol{\Theta}_j\}$, box matrix $\mathbf{A}$, grid dimensions (K_1, K_2, K_3), Ewald parameter α, B-spline order p, and precomputed one-dimensional modulus factors $b(k_1/N_1)$, $b(k_2/N_2)$, and $b(k_3/N_3)$
Ensure: Reciprocal-space energy U_{rec} and derivatives with respect to multipoles and coordinates

Forward Step I: Spread multipoles onto the real-space grid
Initialize the real-space grid multipolar array $Q^R(\mathbf{n})$ to zero
for each thread **do** ▷ TorchFF kernel
  Read box and the coordinate and multipoles of atom i
  Compute the scaled fractional coordinate $\mathbf{u}_i$
  for each neighboring grid point (p^3 points in total) **do**
    Compute interpolated multipoles with Eq.(S4.30)
    Accumulate onto $Q^R(\mathbf{n})$ using `atomicAdd`
  end for
end for

Forward Step II: Transform to reciprocal space and apply convolution
Compute the forward FFT: $Q^F(\mathbf{k}) \leftarrow \mathcal{F}[Q^R(\mathbf{n})]$ ▷ `torch::rfftn`
for each thread **do** ▷ TorchFF kernel
  Determine the indices k_1, k_2, and k_3 from the thread ID
  Read $Q^F(\mathbf{k})$ and the precomputed modulus factors $b(k_1/N_1)$, $b(k_2/N_2)$, and $b(k_3/N_3)$
  Compute the reciprocal-space coefficient $G^F(\mathbf{k})$ using Eq. (S4.34)
  Update the reciprocal-space potential as $\Phi^F(\mathbf{k}) \leftarrow G^F(\mathbf{k})Q^F(\mathbf{k})$
end for

Forward Step III: Transform back and evaluate the energy
Compute the inverse FFT: $\Phi(\mathbf{n}) \leftarrow \mathcal{F}^{-1}[\Phi^F(\mathbf{k})]$ ▷ `torch::irfftn`
Compute the reciprocal-space energy: $U_{\text{rec}} \leftarrow \frac{1}{2}\sum_{\mathbf{n}} \Phi(\mathbf{n})Q^R(\mathbf{n})$ ▷ `torch::sum`

Forward Step IV: Save intermediate quantities for the backward pass
Save $\mathbf{r}_j$, q_j, $\boldsymbol{\mu}_j$, $\boldsymbol{\Theta}_j$, and $\Phi(\mathbf{n})$ to autograd context
return U_{rec}

Backward
for each thread **do** ▷ TorchFF Kernel
  Read box the coordinate and multipoles of atom i
  Compute the scaled fractional coordinate $\mathbf{u}_i$
  Initialize gradients
  for each neighboring grid point (p^3 points in total) **do**
    Evaluate the B-spline interpolation weights $M_p(\cdot)$
    Compute gradient Eqs. (S4.39)–(S4.42) and accumulate them
  end for
  Write result back to global memory
end for
return Derivatives of U_{rec}

Algorithm 2: Reciprocal-space multipolar PME in TorchFF

4.5.3 Correctness Validation of TorchFF Implementation

As shown in Table S4.1, the potential energies and atomic forces calculated by TorchFF were compared directly against the corresponding OpenMM results for periodic boxes containing 300, 1000, and 3000 water molecules with TIP3P and AMOEBA (without mutual polarization). The nonbonded interaction cutoff was set to 0.8 nm. PME is used for long-range electrostatics with an Ewald parameter $\alpha = 3.65$ nm^{-1} and grid size $K = 128$. This relatively large grid size was selected to minimize numerical differences arising from the different B-spline interpolation orders used by the two implementations: TorchFF uses sixth-order B-splines, whereas OpenMM uses fifth-order B-splines. We evaluated energy and force errors as follows:

$$\text{Rel. } \Delta E \ \% = \frac{|E(\text{OpenMM}) - E(\text{TorchFF})|}{|E(\text{OpenMM})|} \times 100 \tag{S4.43}$$

$$F_{\text{RMSE}} = \sqrt{\frac{1}{3N} \sum_{i=1}^{N} \sum_{\alpha \in x,y,z} [F_{i,\alpha}(\text{OpenMM}) - F_{i,\alpha}(\text{TorchFF})]^2} \tag{S4.44}$$

TorchFF closely reproduces the energies and forces calculated by OpenMM for both force fields and across all tested system sizes. The relative deviations in the total potential energy remain below (2×10^{-5}%) and the force RMSE values range from 10^{-4} to 10^{-3} kJ/mol/nm, demonstrating close numerical agreement between the TorchFF and OpenMM implementations. All the calculations are performed with double precision.

Table S4.1: Direct comparison of the potential energies and atomic forces calculated using OpenMM and TorchFF for periodic water boxes of different sizes with the TIP3P and AMOEBA force fields. The unit of energy is kJ/mol and the unit of force is kJ/mol/nm.

# Waters	Force Field	E(OpenMM)	E(TorchFF)	Rel. ΔE %	F_{RMSE}
300	TIP3P	-15047.0132	-15047.0159	1.75×10^{-5}	1.48×10^{-4}
300	AMOEBA	-11779.1261	-11779.1271	8.45×10^{-6}	3.26×10^{-4}
1000	TIP3P	-51159.1689	-51159.1781	1.81×10^{-5}	1.56×10^{-4}
1000	AMOEBA	-39778.5374	-39778.5413	9.72×10^{-6}	5.68×10^{-4}
3000	TIP3P	-154727.4695	-154727.4985	1.87×10^{-5}	3.00×10^{-4}
3000	AMOEBA	-120925.2234	-120925.2342	8.90×10^{-6}	1.09×10^{-3}

We additionally performed molecular dynamics simulations of liquid water using TorchFF and OpenMM with the TIP3P and AMOEBA force fields. Each system was first energy-minimized, followed by 500 ps of equilibration in the NVT ensemble and 500 ps of equilibration in the NPT ensemble. A final 5 ns NPT production simulation was then performed and the snapshot was saved every 5 ps. The time step was 1 fs. Langevin dynamics was used to maintain the temperature, and pressure was controlled using a Monte Carlo barostat.

The TorchFF simulations were driven through the `openmm-torch` interface. The radial distribution functions were calculated from the production trajectories using MDAnalysis and are shown in Figure S4.1. The RDFs obtained from TorchFF and OpenMM also closely overlap, demonstrating the consistency between TorchFF and OpenMM implementations.

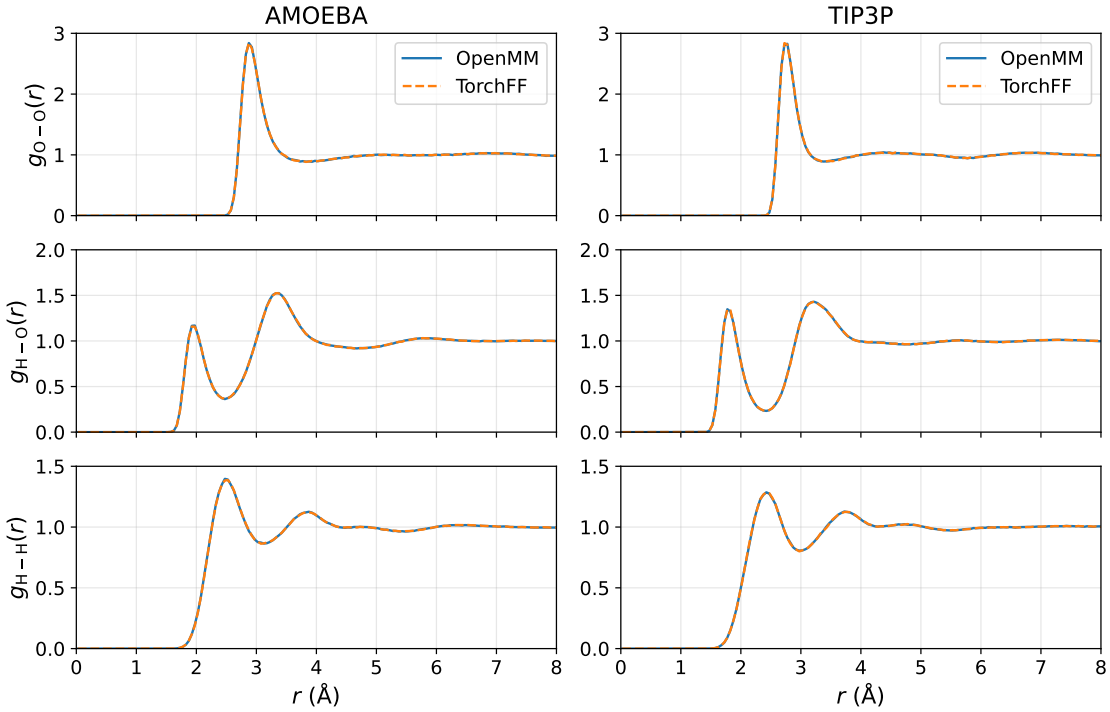

Figure S4.1: Comparison of the radial distribution functions (RDFs) obtained from OpenMM and TorchFF simulations of liquid water using the AMOEBA and TIP3P force fields.

The equilibrium densities obtained from the production trajectories are reported in Table S4.2. The uncertainties were estimated as the standard deviations of the mean densities calculated from 20 trajectory blocks. For both force fields, the TorchFF and OpenMM densities agree within the estimated statistical uncertainty.

Table S4.2: Comparison of the equilibrium densities (g/mL) of liquid water obtained from OpenMM and TorchFF simulations using the AMOEBA and TIP3P force fields.

Force Field	OpenMM	TorchFF
AMOEBA	0.986 ± 0.002	0.987 ± 0.003
TIP3P	0.997 ± 0.002	0.998 ± 0.002

Chapter 5

Improved and Simplified Treatment of Short-Range Intramolecular Interactions in Polarizable Force Fields

5.1 Introduction

Accurate and efficient representation of molecular potential energy surfaces (PESs) is one of the central topics in computational chemistry, with direct applications ranging from materials science to drug discovery. Despite the emerging machine-learning interatomic potentials[238–241] (MLIPs) that are fully data driven, molecular-mechanics force fields (MMFFs), which adopt carefully parameterized analytical potentials, continue to occupy an essential role in molecular dynamics (MD) for real-world applications with their remarkable efficiency and interpretability. In these physical models, the potential energy is conventionally partitioned into bonded and nonbonded contributions.[242] Nonbonded terms account for intermolecular and nonlocal intramolecular interactions, including electrostatics, Pauli repulsion, dispersion, and many-body effects such as polarization and charge transfer in more advanced models. Bonded terms capture local intramolecular distortions through bond-stretching, angle-bending, and torsional bond rotations.

Although this decomposition is physically straightforward and computationally convenient, the boundary between bonded and nonbonded physics is ambiguous, particularly for 1–4 interactions, i.e., interactions between atoms separated by three covalent bonds, which are critical to describe conformational energetics and directly affect the predictive power for macroscopic phenomena such as protein folding[243] or binding free energies[244]. At such distances, the overlap of electron densities gives rise to partial covalency[245, 246] and short-range effects such as charge penetration[247, 248] and exchange repulsion, which cannot be well described by the nonbonded model. As a consequence, most FFs adopt a complicated treatment for 1–4 interactions by first arbitrarily scaling the nonbonded

interactions to attenuate their errors and relying on torsional parameters to compensate for residual inaccuracies.[249] For example, in fixed-charge models like AMBER[74, 243, 249–251], OPLS[76, 252], CHARMM[75, 253], and OpenFF[77, 254], the Coulomb and Lennard-Jones interactions between 1–4 atom pairs are multiplied by an arbitrary factor such as 1/2 or 5/6, and truncated Fourier-series torsion terms are introduced to bridge the gap between MM and QM energy profiles. Besides the accuracy problem, this treatment has several other drawbacks. It introduces a fundamental interdependence between the bonded and nonbonded models, complicating parametrization and potentially reducing transferability. Moreover, agreement with a one-dimensional torsional energy profile does not guarantee accurate forces, relaxed geometries, or the correct multidimensional PES, which are essential for reliable simulations.

These difficulties persist, and in some respects become more pronounced, in many-body force fields.[255, 256] AMOEBA provides a substantially richer description of molecular interactions than fixed-charge models through multipolar interactions, explicit many-body polarization, and an improved 14–7 van der Waals potential[78, 257, 258]. Nevertheless, its short-range intramolecular treatment still relies on empirical scaling conventions. Moreover, the permanent multipole interaction, van der Waals interaction, and polarization are governed by separate scaling rules that are not even solely based on topological distances[79]. This further introduces uncertainty and difficulty in parameterization, as well as additional implementation and computational complexity, thus impeding its wider application.

Two complementary strategies can be considered for improving the treatment of short-range intramolecular interactions. The first is to describe 1–4 interactions entirely through a nonbonded model with complete physics. More advanced force fields like HIPPO[259], AMOEBA+[260], MB-UCB[261, 262], or CMM[245, 263] are promising via this route because they incorporate explicit treatments of related effects like charge penetration and may reduce the magnitude of empirical torsional corrections. In practice, however, balancing many strongly coupled interaction components is difficult, and the resulting models can also become challenging to parametrize. In addition, residual errors in any individual component can still propagate, making a torsional correction necessary to recover an accurate PES.

The second strategy is the bonded-only treatment of the short-range 1–4 contributions. In our previous work[81], we explored this idea and developed such an "expanded" bonded model containing bond, angle, and torsional potentials together with their coupling terms. We adapted the Q-Force[80] parametrization framework and developed a workflow to derive the parameters in an automated and systematic manner for fixed charge force fields by fitting against QM reference energies, forces, and Hessians. For small molecules without discernible nonbonded contacts beyond the 1–4 level, this construction reproduced the reference QM PES with near-QM accuracy[81].

In the present work, we extend this bonded-only description of short-range intramolecular interactions to embed in larger systems with 1-5 and beyond interactions. Using the AMOEBA polarizable force field, we replace the original component-specific intramolecular scaling

conventions with a unified scheme in which permanent electrostatics, direct induction, polarization energy, and van der Waals interactions are excluded through the 1–4 level and fully included, i.e. unscaled, beginning at the 1–5 level. The local intramolecular PES is then described by the Q-Force bonded and bonded-coupling model, while the permanent atomic multipoles are refitted under the revised intramolecular polarization convention. Capped alanine, glycine, and proline dipeptides are used as representative models of the protein backbone torsional space captured with a Ramachandran plot.

The resulting force field provides substantially improved agreement with QM gas-phase and implicit-solvent Ramachandran surfaces, as well as lower energy and force errors for independent validation conformations. It also supports stable condensed-phase molecular dynamics and retains conformational behavior and solution-phase observables comparable to those obtained with the established AMOEBA model with its complicated scaling schemes for 1-4 interactions. These results show that an advanced bonded representation can be combined naturally with a polarizable multipolar nonbonded model, providing a more accurate and transparent alternative to conventional short-range scaling schemes and a foundation for the development of future molecular-mechanics force fields.

5.2 Theory and Methods

5.2.1 New Unified Nonbonded Scaling Scheme in AMOEBA

The nonbonded energy in the AMOEBA polarizable force field is composed of permanent electrostatics, many-body polarization, and van der Waals interactions[78]:

$$V_{\text{nonbonded}} = V_{\text{perm}} + V_{\text{pol}} + V_{\text{vdW}}. \tag{S4.1}$$

The permanent electrostatic interaction is represented by atom-centered permanent multipoles through quadrupole order:

$$V_{\text{perm}} = \frac{1}{2} \sum_{i \neq j} s_{ij}^{m} \mathbf{M}_i^{\mathrm{T}} \mathbf{T}_{ij} \mathbf{M}_j, \tag{S4.2}$$

where $\mathbf{M}_i = (q_i, \boldsymbol{\mu}_i, \boldsymbol{\Theta}_i)$ is the multipole vector and $\mathbf{T}_{ij}$ is the multipole interaction tensor generated from spatial derivatives of the Coulomb kernel $1/r_{ij}$. The coefficient s_{ij}^{m}, often referred to as the "m-scale", is the factor that scales the permanent multipolar interactions between atom pairs with short topological separations. Specifically, in the current version of AMOEBA, permanent electrostatics between 1–2 and 1–3 pairs are fully excluded, scaled by 0.4 for 1–4 pairs, by 0.8 for 1–5 pairs, and fully included thereafter.

As the most significant improvement to fixed-charge force fields, AMOEBA explicitly models polarization effects through induced atomic dipoles, which serve as a response to the

electric field generated by the permanent multipoles and by all other induced dipoles,

$$\boldsymbol{\mu}_i^{\text{ind}} = \alpha_i \left[\mathbf{E}_i^d + \sum_{j \neq i} \mathbf{T}_{ij}^{\text{damp}} \boldsymbol{\mu}_j^{\text{ind}} \right] \tag{S4.3}$$

where α_i is the atomic polarizability and $\mathbf{T}_{ij}^{\text{damp}}$ is the Thole-damped interaction tensor. $\mathbf{E}_i^d$ is the electric field on atom i that produces the induced dipoles, i.e.

$$\mathbf{E}_i^d = \sum_{j \neq i} s_{ij}^d \mathbf{T}_{ij}^{\text{damp}} \mathbf{M}_j, \tag{S4.4}$$

The polarization energy is then defined as

$$V_{\text{pol}} = -\frac{1}{2} \sum_i \boldsymbol{\mu}_i^{\text{ind}} \cdot \mathbf{E}_i^p \tag{S4.5}$$

where $\mathbf{E}_i^p$ is the field actually used in the polarization energy calculation and it differs from $\mathbf{E}_i^d$ only by the scaling settings:

$$\mathbf{E}_i^p = \sum_{j \neq i} s_{ij}^p \mathbf{T}_{ij}^{\text{damp}} \mathbf{M}_j \tag{S4.6}$$

The scaling factors applied to topologically proximal atom pairs during the calculation of the electric field s_{ij}^d and polarization energy s_{ij}^p are often referred to as the "d-scale" and "p-scale", respectively. Unlike the m-scales, these two polarization-related scales are not determined solely by the number of bonds separating the atom pairs. Instead, a concept called "polarization group" is introduced and it is typically a functional group with limited conformational flexibility[79]. Permanent multipoles only polarize atoms in other groups, i.e., the d-scale is 0.0 for atom pairs in the same polarization group and is 1.0 for atom pairs in distinct groups, regardless of their topological distance. The p-scale, however, excludes 1–2 and 1–3 interactions but fully includes 1–4 and more distant interactions, except that intra-group 1–4 interactions receive a factor of 0.5.

The van der Waals interaction is described by the buffered 14–7 potential in AMOEBA:

$$V_{\text{vdW}} = \sum_{i<j} s_{ij}^v \epsilon_{ij} \left(\frac{1+\delta}{\rho_{ij}+\delta} \right)^7 \left(\frac{1+\gamma}{\rho_{ij}^7+\gamma} - 2 \right), \quad \rho_{ij} = \frac{r_{ij}}{R_{ij}^0} \tag{S4.7}$$

Here, R_{ij}^0 and ϵ_{ij} are the combined equilibrium separation and well depth, respectively, and the AMOEBA buffering constants are $\delta = 0.07$ and $\gamma = 0.12$. The buffering softens the rapidly increasing repulsive wall at short distances while retaining the attractive long-range interaction. The coefficient s_{ij}^v denotes the vdW scaling factor, which fully excludes 1–2 and 1–3 pairs but fully includes 1–4 and more distant pairs.

As summarized above, the nonbonded scaling scheme in AMOEBA is apparently more intricate than those commonly used in fixed-charge force fields. In our view, this complexity partly reflects the historical and incremental development of the force field over the past two decades. AMOEBA was originally developed for water and was later extended to molecules with greater flexibility. This extension required a set of conformer-independent parameters, i.e., intrinsic permanent atomic multipoles (PAMs) and vdW parameters, to reproduce both conformer-dependent electrostatic properties, such as molecular dipoles and electrostatic potentials, and conformational energetics. These different objectives, however, are controlled through separate scaling treatments in the current framework: the *d*-scales affect the conformer-dependent electrostatic properties via intramolecular polarization, whereas the *m*-, *p*-, and vdW scaling factors influence the energetic profiles. Consequently, although AMOEBA incorporates more physical details than fixed-charge models, its present scaling framework remains, to some extent, an empirically accumulated set of interaction-specific rules rather than a scheme uniquely determined by the underlying physics. This empirical structure can heavily entangle the parameterization of bonded and nonbonded terms, while also increasing implementation and computational complexity.

Therefore, in the present work, we simplify the original complicated scaling scheme in AMOEBA with a single unified topological rule applied consistently to permanent electrostatics, polarization, direct induction, and vdW interactions:

$$s_{ij}^{m} = s_{ij}^{p} = s_{ij}^{d} = s_{ij}^{v} = \begin{cases} 0, & n_{ij} \leq 3, \\ 1, & n_{ij} \geq 4. \end{cases} \tag{S4.8}$$

where n_{ij} is the number of covalent bonds along the shortest path between atoms i and j. Thus, all nonbonded interactions between 1–2, 1–3, and 1–4 atom pairs are excluded, whereas all 1–5 and more distant intramolecular interactions are included without any scaling. Table 5.1 summarizes the differences between the conventional and new scaling schemes for AMOEBA nonbonded interactions.

We believe that this unified scheme offers several advantages. First, it establishes a clearer separation between local and nonlocal physics. At typical 1–4 separations, overlap between electronic densities gives rise to charge penetration, exchange repulsion, and other short-range effects that are not explicitly represented by the current AMOEBA nonbonded functional forms. Applying constant scaling factors can only mimic, rather than recover, the missing physics and consequently introduces additional compensation between the nonbonded interactions and the bonded torsional parameters. By contrast, removing all 1–4 nonbonded interactions more effectively decouples the parameterization of the bonded and nonbonded models. The remaining 1–5 and longer-range interactions generally occur at separations where conventional nonbonded functional forms provide a physically reasonable description. As demonstrated in the following sections, this revised scheme allows the advanced bonded model to be optimized directly against intramolecular quantum-mechanical energies and forces, resulting in improved accuracy.

Second, applying a consistent scaling treatment across all nonbonded components reduces implementation complexity and improves computational efficiency. One important consequence of the original inconsistent *d*- and *p*-scaling treatment is the need to maintain two distinct sets of induced dipoles to ensure correct evaluation of polarization energies and forces[258], because the electric field used to generate the induced dipoles differs from that used in the polarization energy evaluation. This duplication increases both memory usage and computational cost and complicates the implementation of efficient polarization solvers, particularly in condensed-phase simulations under periodic boundary conditions, where the induced dipoles must be determined iteratively through a self-consistent-field (SCF) procedure. Unifying the *d*- and *p*-scaling treatments eliminates this requirement. Moreover, because the new scheme employs a binary exclusion rule, separate 1–4 and 1–5 neighbor lists are no longer needed for applying interaction-specific scaling factors. The group-based *d*-scaling treatment used to construct the electric field for direct induction is also eliminated, making the model more systematic, reproducible, and less dependent on chemical-perception rules.

Table 5.1: Comparison of the conventional AMOEBA scaling conventions and the unified scaling scheme used in the present force field.

Interaction component	Scheme	1–2	1–3	1–4	1–5	≥ 1–6
Permanent multipoles, s^m	Original AMOEBA	0.0	0.0	0.4	0.8	1.0
	This work	0.0	0.0	0.0	1.0	1.0
Polarization energy, s^p	Original AMOEBA	0.0	0.0	1.0[a]	1.0	1.0
	This work	0.0	0.0	0.0	1.0	1.0
Direct induction, s^d	Original AMOEBA	Polarization-group dependent[b]				
	This work	0.0	0.0	0.0	1.0	1.0
Buffered 14–7 vdW, s^v	Original AMOEBA	0.0	0.0	1.0	1.0	1.0
	This work	0.0	0.0	0.0	1.0	1.0

[a] The conventional intragroup 1–4 polarization factor can be 0.5.

[b] In the conventional group-based construction, the direct permanent field is excluded within the same polarization group and is generally unscaled between distinct polarization groups. Consequently, no unique *d*-scale can be assigned from the atom-pair 1–*n* separation alone.

5.2.2 Valence Potential

In our previous work[81, 264], we demonstrated that with proper selection and fitting of bond, angle, torsion, and their coupling terms, we could reproduce the QM PES of small molecules with very high accuracy. This novel bonded-only framework for 1–4 interactions

was also compatible with fixed-charge models, including OPLS, CHARMM, and AMBER, that describe nonbonded interactions beyond 1–4 atom pairs, as validated by reproducing the QM Ramachandran surface of alanine dipeptide. Here, we retain the functional forms of AMOEBA electrostatic, polarization, and van der Waals terms, whereas the bonded energy is evaluated using the following functional form:

$$\begin{aligned} V_{\text{bonded}} = V_{\text{bond}} + V_{\text{angle}} + V_{\text{torsion}} + V_{\text{bond-bond}} + V_{\text{angle-angle}} \\ + V_{\text{bond-angle}} + V_{\text{torsion-bond}} + V_{\text{torsion-angle}} \\ + V_{\text{torsion-angle-angle}} + V_{\text{improper}} \end{aligned} \tag{S4.9}$$

Bond stretching, angle bending and torsional terms. In our model, bond stretching and angle bending are described using a Morse potential and a cosine-based potential, respectively:

$$V_{\text{bond}}(b) = D_e \left[1 - \exp\left(-\alpha(b - b_0)\right)\right]^2, \quad \alpha = \sqrt{K_b/2D_e} \tag{S4.10}$$

$$V_{\text{angle}}(\theta) = K_\theta \left[\cos(\theta) - \cos(\theta_0)\right]^2 \tag{S4.11}$$

These forms recover harmonic behavior near the equilibrium geometry while naturally introducing the physically appropriate anharmonicity at larger extensions, without requiring additional higher-order polynomial terms used in the original AMOEBA bond and angle potential.

For proper flexible torsions, our model adopts the common truncated Fourier series:

$$V_{\text{torsion}}(\phi) = \sum_{n=1}^{4} K_n \left[1 + \cos(n\phi - \phi_n)\right] \tag{S4.12}$$

where ϕ is the dihedral angle, n is the periodicity, K_n is the force constant and ϕ_n is the phase factor associated with the n-th Fourier component. ϕ_n is 0 for odd n values and π for even n values, which is the common choice to respect relevant symmetries. The use of periodicities $n = 1$–4 provides sufficient flexibility to describe the positions and relative energies of diverse torsional minima and barriers.

For trigonal centers that are expected to remain planar, an improper dihedral potential is included to penalize out-of-plane distortions,

$$V_{\text{improper}}(\phi_{\text{imp}}) = K_{\text{imp}} \phi_{\text{imp}}^2 \tag{S4.13}$$

where ϕ_{imp} measures the displacement from the planar reference geometry and K_{imp} is the corresponding improper force constant. For a planar center bonded to three neighboring atoms, all three possible improper-dihedral combinations are included.

Bond, angle, and torsion coupling terms. Correlations among neighboring internal coordinates are represented explicitly through bond-bond, angle-angle, bond-angle, torsion-bond, torsion-angle and torsion-angle-angle coupling terms. These terms were developed and

implemented in force fields such as MM3 [265, 266] and MMFF94 [267, 268], CFF93 [269, 270].

Simultaneous distortions of two bonds that share a common atom are described by

$$V_{\text{bond-bond}}(b, b') = K_{bb'}(b - b_0)(b' - b'_0) \tag{S4.14}$$

and the angle-angle coupling term

$$V_{\text{angle-angle}}(\theta, \theta') = K_{\theta\theta'} \left[\cos(\theta) - \cos(\theta_0)\right] \left[\cos(\theta') - \cos(\theta'_0)\right] \tag{S4.15}$$

captures the correlation between two adjacent angles that share at least two common atoms. The interaction between bond-angle pairs that share two common atoms is expressed by

$$V_{\text{bond-angle}}(b, \theta) = K_{b\theta}(b - b_0) \left[\cos(\theta) - \cos(\theta_0)\right] \tag{S4.16}$$

and the following torsional-related coupling terms are included to account for cross-interactions between dihedral-bond, dihedral-angle and dihedral-angle-angle:

$$V_{\text{torsion-bond}}(b, \phi) = \sum_{n=1}^{4} K_{n,b\phi}(b - b_0) \left[1 + \cos(n\phi - \phi_n)\right] \tag{S4.17}$$

$$V_{\text{torsion-angle}}(\theta, \phi) = \sum_{n=1}^{4} K_{n,\theta\phi} \left(\theta - \theta_0\right) \left[1 + \cos(n\phi - \phi_n)\right] \tag{S4.18}$$

$$V_{\text{torsion-angle-angle}}(\theta, \theta', \phi) = K_{\theta\theta'\phi}(\theta - \theta_0)(\theta' - \theta'_0)[1 + \cos(n\phi - \phi_n)] \tag{S4.19}$$

The torsion-bond term applies to torsion-bond pairs sharing one common atom, and the torsion-angle terms apply to torsion-angle pairs sharing two common atoms. The higher-order torsion-angle-angle term only applies to adjacent angles that form a torsion, i.e., for a torsion defined by atoms A–B–C–D, this term couples the A–B–C and B–C–D angles to the A–B–C–D dihedral and therefore captures energetic contributions arising from their simultaneous deformation that cannot be represented by independent torsion-angle terms alone.

5.2.3 Parameterization Procedure

Replacing the conventional AMOEBA intramolecular scaling scheme with the unified rule in Equation (S4.8) requires reparameterization of two components of the force field. First, because the direct-induction scaling that governs the induced-dipole response is altered, the permanent atomic multipoles must be refit to preserve the quantum-mechanical electrostatic properties of the model compounds. Second, the short-range intramolecular energy through the 1–4 level, now assigned entirely to the bonded model described in Section 5.2.2, must be

parametrized directly against quantum-mechanical reference data. We address these tasks sequentially by first optimizing the electrostatics under the unified scaling scheme and then fitting the bonded and bonded-coupling terms with the nonbonded model fixed.

In the original AMOEBA protein force field, permanent atomic multipoles were obtained by distributed multipole analysis and subsequently optimized against quantum-mechanical electrostatic potentials evaluated on grids surrounding model dipeptides[257]. Multiple conformers were fitted simultaneously so that the resulting conformer-independent parameters, together with the explicit intramolecular polarization, were able to describe the conformational dependence of the electrostatic potential. We follow the same electrostatic-fitting protocol and settings while applying the revised scaling scheme. Capped ACE–X–NME dipeptides, with X = Gly, Ala, or Pro, serve as the model compounds. For each compound, several representative conformers are fitted simultaneously to MP2/aug-cc-pVTZ electrostatic potentials sampled on four layers of grid points spaced by 0.35 Å, beginning 1.0 Å from the van der Waals surface; an additional conformer is reserved for validation. In order to introduce minimal changes to the original parameter set, we reoptimize only the atomic dipoles, while the atomic charges and quadrupoles are retained at their original values. This targeted fit allows the dipoles to compensate for the modified induction scheme while preserving the overall charge distribution and higher-order moments. We assess the refit using the root-mean-square error (RMSE, Equation (S4.20)) and the relative RMSE (RRMSE, Equation (S4.21)) between the model and QM electrostatic potentials over both the fitting and validation sets, comparing the original AMOEBA parameters with the new scaling scheme before and after dipole reoptimization. Electrostatic-potential grids and atomic multipole fits are generated with the POTENTIAL program distributed with Tinker[271], and the QM electrostatic potential at each grid point is evaluated with Psi4[272].

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\mathrm{ESP}_i^{\mathrm{QM}} - \mathrm{ESP}_i^{\mathrm{MM}})^2} \tag{S4.20}$$

$$\mathrm{RRMSE} = \frac{\mathrm{RMSE}}{N^{-1}\sum_{i=1}^{N}|\mathrm{ESP}_i^{\mathrm{QM}}|} \times 100\%. \tag{S4.21}$$

The bonded potential in Section 5.2.2 is parametrized using the Q-Force framework[80, 81], which provides an automated and reproducible route to generate force field parameters from the reference QM data. We follow the same protocol as in our previous work to generate the fitting data, which is briefly outlined here:

- First, for each input molecule, a full geometry optimization at the QM level is performed, followed by calculation of the Hessian for the optimized structure.
- Then, multiple molecular dynamics simulations are performed at 500 K with GFN2-xTB[273], starting from conformers generated by CREST[274]. A total of 400 snapshots

are sampled equally across the trajectories, and their energies and atomic forces are evaluated at the reference QM level. Half of the snapshots are included in the fit, and the remaining half are reserved for validation.

- Relaxed 1D and 2D torsion scans are performed to provide additional coverage of the torsional potential energy surface, especially the high-energy areas that are rarely accessed by the MD. One-dimensional scans are performed around noncyclic bonds in increments of 15°, and 2D scans are performed for the backbone ϕ/ψ surface on a two-dimensional grid with 15° increments (i.e., a 24×24 grid) along both coordinates. TorsionDrive[275] at the GFN2-xTB level is first used to generate the initial geometries, after which each grid point is reoptimized at the reference QM level with the scanned dihedral angle(s) constrained.

- In addition, the QM 2D ϕ/ψ scan is also performed with the SMD[276] continuum-water model to account for solvation effects on the conformational energetics. These implicit-solvent reference data are included as additional fitting targets. A similar strategy was also adopted in the development of CMAP terms for the AMBER force field[243].

- All reference QM calculations for this stage are performed with Q-Chem[277] at the ωB97X-V/def2-TZVPD level[278], which balances cost and accuracy and has been widely adopted in molecular mechanics studies[245, 263, 279].

After the reference data generation, the force constants for all the terms in the bonded potential are optimized using lasso regression to minimize the differences between QM and MM Hessians, energies, and forces. The fitting targets are the QM data with the AMOEBA nonbonded contribution, evaluated with the new scaling scheme and reoptimized atomic dipoles, subtracted. For QM–SMD data, the Generalized-Kirkwood (GK) model[280] is used for MM subtraction. To maintain consistent amide parameters across the capped dipeptides, N-methylacetamide (NMA) is parametrized first. Bonded parameters for the chemically equivalent terms associated with both amide linkages of each ACE–X–NME dipeptide are then transferred from the NMA fit and held fixed. The subsequent residue-specific fits therefore optimize only the remaining bonded and bonded-coupling terms.

All the MM calculations are performed with OpenMM 8[281]. Because OpenMM hard-codes the AMOEBA scaling scheme and does not support modifying it through force field definitions or APIs, we patch the source code to implement the new scaling scheme. The switch from group-based d-scaling to topological-distance-based d-scaling is enabled by assigning each atom to its own polarization group.

5.2.4 Replica-Exchange Molecular Dynamics for NMR Observables

To validate the reparameterized force field in the condensed phase, backbone scalar ${}^3J(\mathrm{H_N}, \mathrm{H}_\alpha)$-couplings are computed from explicit-solvent simulations and compared against experimental NMR data[282] for the alanine dipeptide and glycine dipeptide (proline is excluded because it lacks $\mathrm{H_N}$). Since these couplings are sensitive to the population of the accessible backbone conformers, we use temperature replica-exchange molecular dynamics (T-REMD) to obtain an extensive sampling of the Ramachandran backbone space.

All molecular dynamics simulations are performed with the patched OpenMM using GPU acceleration. Each model dipeptide is solvated in a periodic box containing 997 explicit water molecules. The solute is described using the new AMOEBA scaling scheme together with the new bonded potential, whereas the original AMOEBA parameters, including the bond terms, angle terms with higher-order polynomials, Urey-Bradley term, and nonbonded parameters, remain unchanged for water. Long-range permanent electrostatics are evaluated with particle-mesh Ewald (PME) using a 0.7 nm real-space cutoff and an Ewald error tolerance of 7.5×10^{-4}; the vdW interactions use a 0.9 nm cutoff, and the induced dipoles are converged to a tolerance of 10^{-5} D at every step. No bond constraints or hydrogen-mass repartitioning are applied. The multiple-time-step (RESPA) Langevin integrator with a friction coefficient of 1 ps^{-1} is adopted as the thermostat, with the slow nonbonded forces evaluated every 2 fs and the fast bonded forces evaluated every 1 fs.

36 replicas with geometrically spaced temperatures from 300 to 620 K are used in the T-REMD, and the trajectory at 300 K is used for subsequent analysis. The simulation consists of five stages: (1) energy minimization at 300 K; (2) equilibration in the NVT ensemble at 300 K; (3) equilibration in the NPT ensemble at 300 K and 1 atm, maintained with a Monte Carlo barostat; (4) NVT equilibration of all 36 replicas using the coordinates and velocities from the previous step as starting points; and (5) production in the NVT ensemble with temperature replica exchange. The equilibration time for steps (2)–(4) is 0.25 ns, and production is run for 30 ns per replica, giving an aggregate sampling of 1.08 μs across the 36 replicas. Exchanges between replicas are attempted every 1 ps using the standard Metropolis criterion based on the replica potential energies, with velocities rescaled by $\sqrt{T_i/T_j}$ upon an accepted swap to maintain the target temperatures. Snapshots are saved every 1 ps for subsequent analysis.

The potential of mean force (PMF) is evaluated as $F(\phi, \psi) = -k_\mathrm{B}T \ln P(\phi, \psi)$, where $P(\phi, \psi)$ is the torsion distribution from the trajectory evaluated on a periodic 36×36 grid. For each saved configuration, the backbone scalar coupling is computed from the Karplus relation[283]:

$$ {}^3J(\mathrm{H_N}, \mathrm{H}_\alpha) = A\cos^2(\phi + \Delta\phi) + B\cos(\phi + \Delta\phi) + C \tag{S4.22}$$

We use $A = 7.09$ Hz, $B = -1.42$ Hz, $C = 1.55$ Hz, and $\Delta\phi = -60°$, an empirical parameterization derived from protein NMR data.[282] This set is selected because it is widely used and was also adopted in the ff19SB and AMOEBA protein force-field validation studies.[243,

257] Other literature parameterizations can yield different back-calculated couplings; their coefficients, origins, and effects on the present results are discussed in the Supporting Information (Supplementary Tables S5.1 and S5.2). The ensemble averages are compared with the experimental values of 6.06 Hz for alanine and 5.85 Hz for glycine.[284] Following previous work,[243, 257, 285, 286] the discrepancy between simulated and experimental NMR data is quantified as

$$\chi^2 = \left(\frac{J_{\text{sim}} - J_{\text{exp}}}{\sigma} \right)^2 , \tag{S4.23}$$

where J_{sim} is calculated using equation (S4.22) for each configuration in the sampled trajectory at 300 K and then averaged. σ is the systematic error of the Karplus model and is set to 0.91 for this specific coupling type[243, 285, 286].

5.3 Results

5.3.1 Validation of the Permanent-Dipole Refit

Table 5.2 reports the RMSE and relative RMSE of the model electrostatic potential relative to the MP2/aug-cc-pVTZ reference for each fitting and validation conformer. The assignment of conformers to the fitting and validation sets follows that used in the original development of the AMOEBA parameters, thereby preserving a consistent basis for evaluating the effect of the new scaling scheme. Applying the unified intramolecular scaling scheme while retaining the original permanent multipoles substantially degrades the molecular electrostatic potential. Across all conformers, the RMSE before refitting ranges from 1.66 to 3.72 kcal/mol/e, with relative errors reaching approximately 75%. Reoptimizing only the atomic dipoles markedly improves agreement with the reference potential. For the fitting structures, the RMSE is reduced to 0.42–0.53 kcal/mol/e, demonstrating that the dipole refit is necessary to compensate for the change in intramolecular induction scaling.

After the dipole refit, the electrostatic potentials obtained with the new scaling scheme are generally comparable in accuracy to those of the Original AMOEBA model. For the fitting structures, the resulting RMSEs are typically only 0.1–0.2 kcal/mol/e larger than the corresponding Original AMOEBA values. The remaining discrepancy could potentially be reduced by additionally reoptimizing the atomic quadrupoles. Such an optimization is not pursued here because the objective of this work is to introduce minimal changes to the existing force field and to assess the feasibility of the unified scaling scheme, rather than to perform an exhaustive refit of all permanent multipole parameters.

The validation results show that the refitted dipoles retain reasonable transferability to conformations not included in the optimization. For alanine dipeptide, the validation error is close to the range observed for the fitting conformers. The somewhat larger error for glycine dipeptide is consistent with the behavior of the Original AMOEBA parameters, which also exhibit a higher error for this validation structure. The proline validation conformer shows

Table 5.2: Electrostatic-potential RMSEs for the dipeptide conformers used in the permanent-dipole refit validation. The three model variants correspond to the Original AMOEBA parameters, the unified intramolecular scaling scheme applied without refitting the permanent multipoles, and the unified scaling scheme after reoptimizing the atomic dipoles. RMSEs are evaluated relative to the MP2/aug-cc-pVTZ reference electrostatic potential and are reported in kcal/mol/e; values in parentheses denote the corresponding relative RMSEs.

Split	Peptide	ϕ/ψ (degree)	Original AMOEBA	New scaling before refit	New scaling after refit
Validation	Ala	(−137.9, 22.8)	0.44 (6.35%)	2.67 (38.23%)	0.59 (8.51%)
	Gly	(−90.0, 180.0)	0.67 (10.36%)	2.19 (33.82%)	0.88 (13.64%)
	Pro	(−85.7, 74.0)	0.38 (7.66%)	3.68 (74.94%)	1.02 (20.76%)
Fit	Ala	(63.4, 34.8)	0.34 (4.30%)	2.75 (35.01%)	0.51 (6.50%)
	Ala	(−166.2, −37.2)	0.34 (4.71%)	2.50 (34.33%)	0.50 (6.84%)
	Ala	(−158.4, 161.3)	0.38 (6.94%)	2.07 (37.98%)	0.53 (9.64%)
	Ala	(73.9, −64.0)	0.37 (6.80%)	2.26 (41.85%)	0.47 (8.69%)
	Ala	(−82.9, 77.9)	0.29 (5.90%)	2.46 (49.47%)	0.50 (10.13%)
	Gly	(−83.0, 74.7)	0.30 (5.42%)	1.66 (29.61%)	0.42 (7.50%)
	Gly	(172.3, 177.0)	0.37 (6.17%)	2.08 (34.70%)	0.45 (7.50%)
	Pro	(39.9, 55.0)	0.27 (3.98%)	3.72 (54.34%)	0.51 (7.43%)

the largest residual error after refitting. Because only one proline conformer was included in the original fitting set, this result suggests that a more conformationally diverse training set may be required to obtain fully transferable permanent dipoles for proline under the new scaling scheme.

5.3.2 Reproduction of Quantum-Mechanical Ramachandran Surfaces

The backbone ϕ/ψ 2D conformational energy surfaces calculated with the original AMOEBA and the present model are presented and compared with the QM reference in both the gas phase and in implicit solvent. Table 5.3 reports the corresponding mean absolute error (MAE) and root-mean-square error (RMSE) relative to the QM surface. For each grid point, the molecular-mechanics energy is obtained by MM relaxation with ϕ and ψ restrained to the target angles, mirroring the constrained optimization used to generate the QM reference surface. The QM surface is referenced to its global minimum, whereas the MM surface is referenced to the MM energy at the ϕ/ψ grid point corresponding to the reference QM minimum. This is a more stringent validation because the fitting is performed with only QM-relaxed geometries, and overfitting to QM geometry energies may lead to substantial force errors that drive relaxation away from the reference basins or cause unstable or nonconvergent minimization.

Table 5.3: Mean absolute error (MAE) and root-mean-square error (RMSE), in kcal/mol, of the relaxed molecular-mechanics ϕ/ψ Ramachandran surfaces relative to the quantum-mechanical reference, in the gas phase and in implicit solvent.

		Gas		Implicit solvent	
Peptide	Model	MAE	RMSE	MAE	RMSE
Ala	Original AMOEBA	0.91	1.15	1.00	1.22
	This work	0.70	0.90	0.66	0.87
Gly	Original AMOEBA	1.24	1.47	1.47	1.81
	This work	0.85	1.12	0.65	0.80
Pro	Original AMOEBA	4.29	5.67	4.15	5.94
	This work	1.14	1.53	1.18	1.55

The present model accurately reproduces the QM surfaces for all three model dipeptides in both gas-phase (Figure 5.1) and implicit-water environments (Figure 5.2) by identifying the correct basin regions, demonstrating the model's ability to describe the local curvature and overall landscape of the QM surfaces. Overall, for all three model dipeptides, there is

an improvement compared to the Original AMOEBA in both environments. For alanine dipeptide, the MAE is slightly reduced by 0.21 kcal/mol in gas phase versus 0.34 kcal/mol under implicit solvent effects. Under the implicit solvent environment, the new model better balances the relative energy difference between the ppII and β regions and recovers the metastable $C7_{eq}$ basin between the ppII and α_R regions. For glycine dipeptide, the new model recovers the low-energy profile in the fully extended region near $(180°, 180°)$ in both environments, where the Original AMOEBA yields relatively higher energies around these regions. The largest improvement occurs for proline, where the Original AMOEBA gas-phase surface differs qualitatively in shape from the QM reference. The new model achieves an MAE of around 1 kcal/mol compared to more than 4 kcal/mol with the original AMOEBA. A major contribution arises from the high-energy regions with distorted geometries where AMOEBA energy errors can be substantially large, showing the ability of the bond/angle/torsion coupling terms to extend the predictive power to larger conformational spaces. Nevertheless, we should also acknowledge that the torsional parameters used in the current AMOEBA were initially fit against a different QM level of theory (RI-TRIM MP2/CBS vs ωB97X-V/def2-TZVPD) and subsequently manually refined against PDB-derived ϕ/ψ distributions.[257]

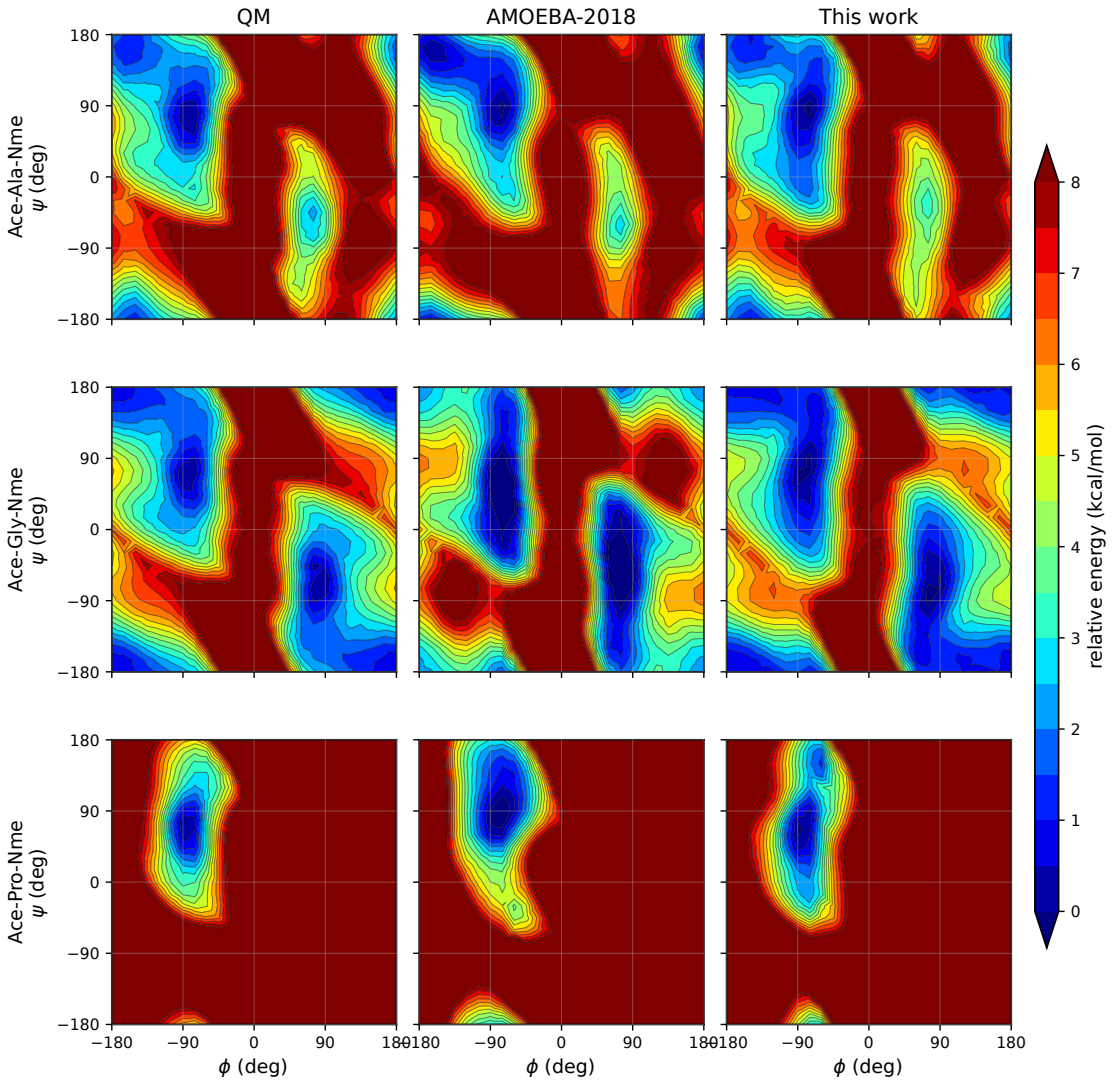

Figure 5.1: Gas-phase relaxed ϕ/ψ Ramachandran surfaces (relative energy in kcal/mol) for the alanine, glycine, and proline dipeptides (rows). Columns compare the quantum-mechanical reference (left) with the Original AMOEBA model (center) and the present unified-scaling model (right). The "AMOEBA-2018" label in the figure denotes the Original AMOEBA model. Each molecular-mechanics grid point is constrained-relaxed to the quantum-mechanical ϕ/ψ target.

Including SMD reference data as per Figure 5.2 is physically important because solvation effects strongly affect protein conformational preferences.[287] This is particularly relevant here because the unified scheme removes all 1–4 nonbonded interactions. Therefore, solvent effects

on short-range conformational energetics are less effectively modeled through nonbonded interactions and must be somewhat compensated by the bonded and coupling terms. A similar philosophy was also proposed in the development of CMAP terms to compensate for the lack of explicit polarization treatment in fixed-charge FFs.[243] The ablation study confirms the role of QM–SMD fitting data, as presented in Supplementary Table S5.3, Figure S5.1 and Figure S5.2.

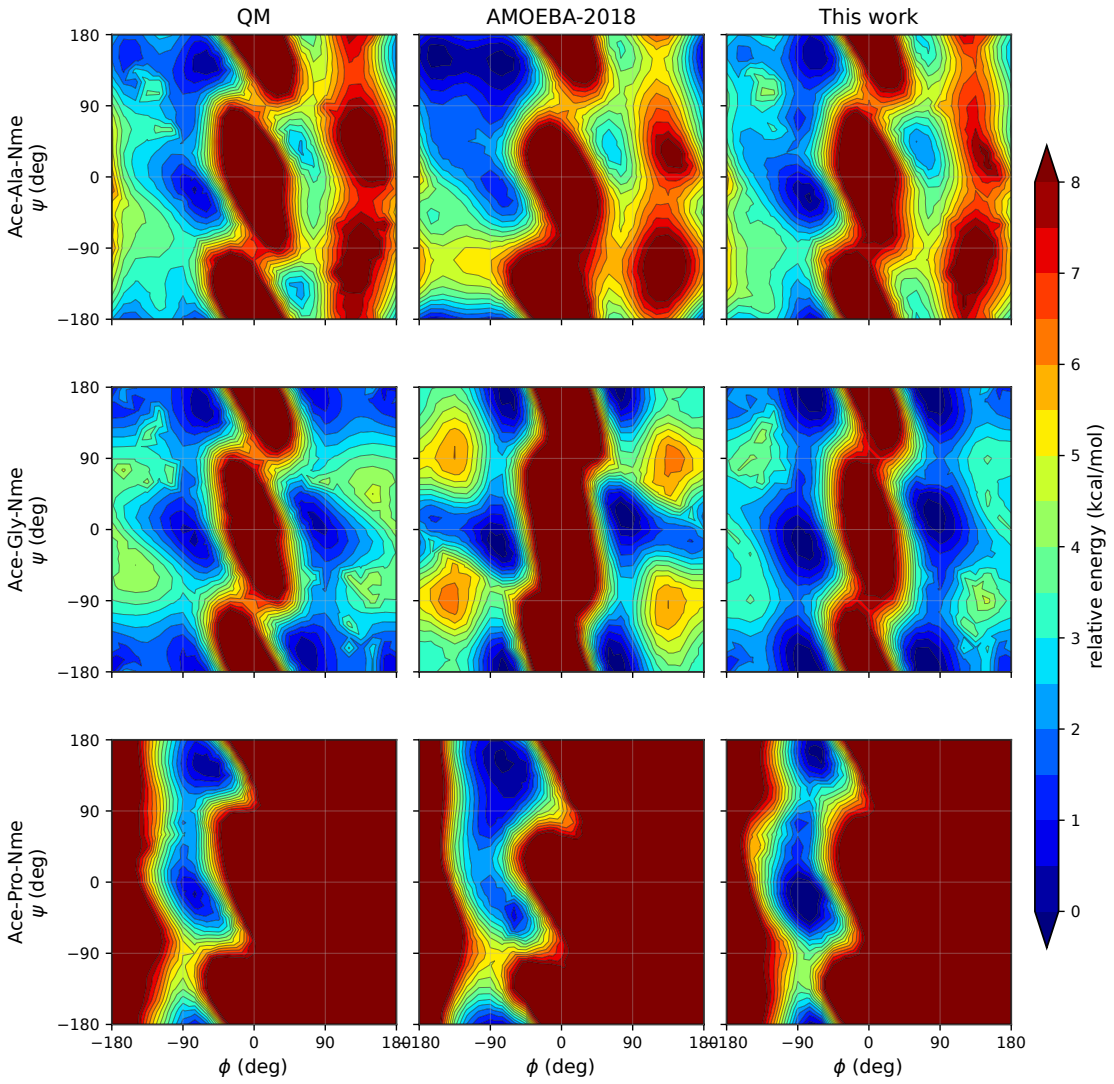

Figure 5.2: Implicit-solvent relaxed ϕ/ψ Ramachandran surfaces (relative energy in kcal/mol), organized as in Figure 5.1. The quantum-mechanical reference uses the SMD continuum-water model and the molecular-mechanics surfaces use the AMOEBA Generalized-Kirkwood implicit solvent.

We find that including SMD targets lowers the implicit-solvent MAE by 0.44, 0.75, and 0.86 kcal/mol for alanine, glyine, and proline dipeptides, respectively, with a negligible effect on the quality of the gas-phase surfaces. For alanine dipeptide, SMD training raises the otherwise overstabilized α_L region, which aligns with previous studies showing that the solvent reaction field changes the energetics of aligned peptide dipoles and can stabilize α_R

conformations.[287] For glycine dipeptide, it contracts overly broad low-energy regions while retaining the low-energy corner around the fully extended geometry. For proline dipeptide, it better reproduces the accessible region with solvation effects, despite some remaining error in the relative depths of its two minima in the ppII and α_R regions.

5.3.3 Energy and Force Accuracy on Validation Conformers

As described in Section 5.2.3, we evaluate transferability of the model using an independent set of 600 conformers excluded from the Q-Force fit. The set contains 200 MD-sampled structures for each dipeptide, all re-evaluated at the reference QM level. For each peptide, the relative QM and MM energies are computed with respect to the same structure: the conformer with the lowest QM energy. Table 5.4 reports both the combined and peptide-specific MAE and RMSE values. The corresponding peptide-specific plots are provided in the Supplementary Figures S5.3 to S5.5

Table 5.4: Relative-energy and Cartesian force-component errors for the held-out validation conformers. Each peptide contributes 200 conformers, and the combined set contains 600 conformers. Within each peptide, all energy series are referenced to the conformer with the lowest quantum-mechanical energy. MAE and RMSE values are reported in kcal/mol for energy and kcal/mol/Å for force.

Set	Model	E MAE	E RMSE	F MAE	F RMSE
Combined	Original AMOEBA	3.02	3.80	10.51	13.23
	This work	1.53	1.92	3.16	5.32
Ala	Original AMOEBA	2.67	3.32	10.39	12.92
	This work	1.89	2.21	3.18	5.38
Gly	Original AMOEBA	3.41	4.30	11.13	14.08
	This work	0.86	1.10	3.59	6.05
Pro	Original AMOEBA	2.99	3.72	10.15	12.84
	This work	1.84	2.23	2.82	4.64

Figure 5.3(a,b) shows the energy and force-component correlations and absolute-error distributions across the 600 conformers. The present model reduces the relative-energy MAE from 3.02 to 1.53 kcal/mol and, more strikingly, the force-component MAE from 10.51 to 3.16 kcal/mol/Å compared with the original AMOEBA. The more concentrated error distributions near zero (Figure 5.3c,d) further show that the improvement is systematic across the validation set rather than arising from only a few conformers. The different sensitivities of the two observables help explain why the force improvement is more remarkable than the

energy improvement. Forces are local derivatives of the PES and therefore respond sharply to distortions along stiff covalent coordinates, which can generate large force components without contributing proportionally to the relative energy of the entire conformer. In our model, the Morse-bond and cosine-angle terms improve the description of bond stretches and angle bends, while extensive coupling terms between bond, angle, and torsion coordinates capture the correlated motions of these local coordinates. In the normal-mode picture, these cross terms supply off-diagonal curvature between internal coordinates and thereby help reproduce the mixed bond, angle, and torsional displacements of molecular vibrations. [80, 81]

Relative conformational energies, by contrast, accumulate both local and nonlocal contributions in a more even manner across the molecule and thus are largely influenced by torsional and nonbonded interactions. Interactions between 1–5 and more distant atom pairs retain the AMOEBA nonbonded functional forms and, apart from the refitted permanent dipoles, the existing nonbonded parameters. Residual discrepancies in these contributions therefore remain in the conformational energies and cannot be removed independently through the bonded fit. Furthermore, the same bonded parameters are optimized against both gas-phase and SMD reference data. This joint objective intentionally balances the two environments rather than minimizing the gas-phase energy alone, as also reflected by the modest gas/solution tradeoff in the SMD ablation study (Supplementary Table S5.3). Together, these factors explain why the force MAE is reduced by more than the energy MAE.

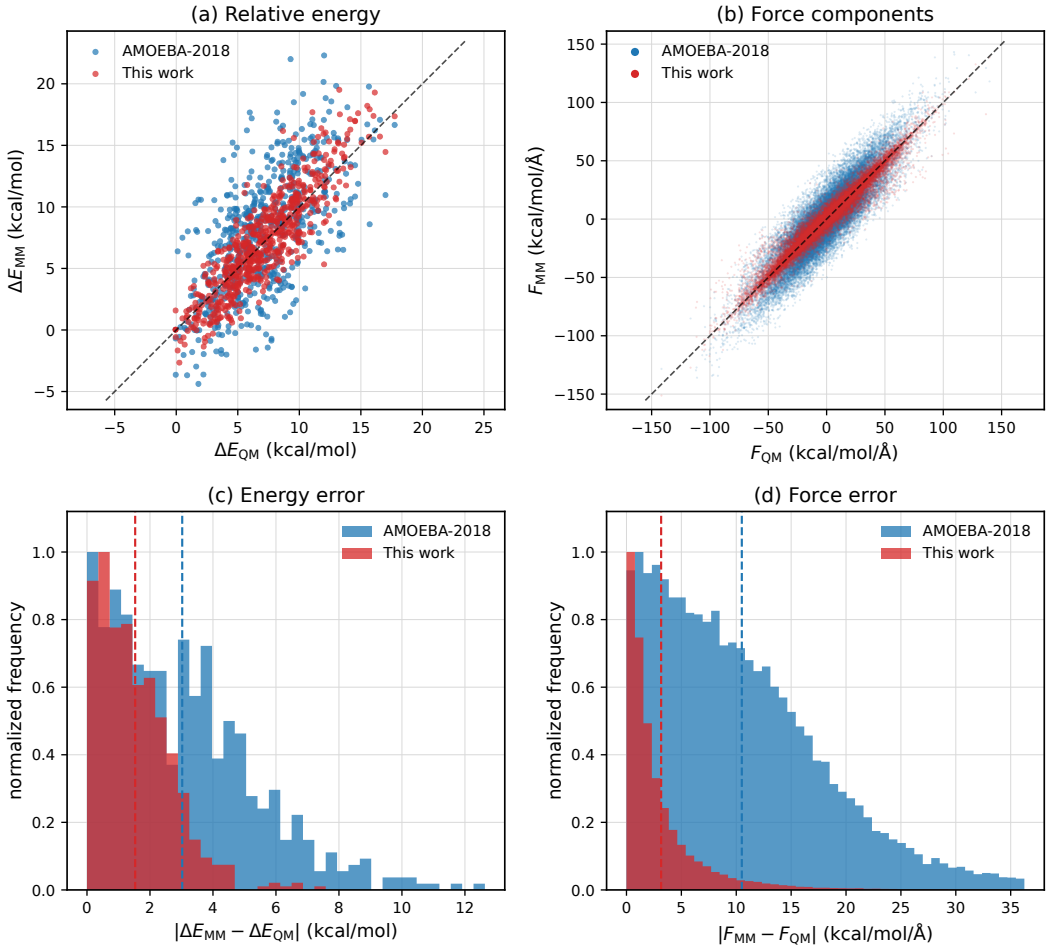

Figure 5.3: Energy and force accuracy on the 600 held-out validation conformers (alanine, glycine, and proline pooled). (a) Relative-energy correlation and (b) force-component correlation between molecular mechanics and quantum mechanics; (c,d) peak-normalized distributions of the absolute relative-energy and force-component errors, with dashed lines marking each model's MAE. The "AMOEBA-2018" labels in the figure denote the Original AMOEBA model.

5.3.4 Condensed-Phase Simulations and NMR Scalar Couplings

We perform T-REMD simulations for the alanine and glycine dipeptides with explicit water using both the original AMOEBA and our model, and compare the computed scalar couplings between the two backbone hydrogens, ${}^3J(\mathrm{H_N}, \mathrm{H}_\alpha)$, with the experimental data reported by Avbelj et al.[284] This not only validates that our model is compatible with the current

AMOEBA water model and able to yield stable simulations in liquid water (even at high temperature), but also serves as a probe of the torsional parameters' ability to describe conformational behavior in a solution environment[243, 250, 257].

Table 5.5: Backbone ${}^3J(\mathrm{H_N}, \mathrm{H}_\alpha)$ scalar couplings (Hz) from the 300 K ensemble and per-peptide χ^2 values against experiment. All simulated values use the Karplus coefficients specified in Section 5.2.4.

		Original AMOEBA		This work	
Peptide	J_{exp}	J_{sim}	χ^2	J_{sim}	χ^2
Alanine	6.06	6.30	0.07	6.43	0.16
Glycine	5.85	6.31	0.26	6.49	0.50

As summarized in Table 5.5, we find that our new model reasonably well reproduces the experimental NMR couplings for these two model dipeptides, with absolute errors of 0.37 and 0.64 Hz, respectively. This conclusion is also supported by the fact that the χ^2 values are smaller than 1, indicating that the deviation from experimental values is smaller than the intrinsic uncertainty of the Karplus model with this specific coefficient set (0.91 Hz). For the same reason, even though the results with the original AMOEBA model are numerically closer to the experiment, the deviation ($\sim$ 0.1 Hz) is too small to distinguish the two models. As a complement to this analysis, we study the sensitivity of simulated J-couplings to the choice of Karplus coefficients by re-evaluating the scalar couplings with the same trajectories but with other coefficient sets from different sources. The results are summarized in Table S5.1 and S5.2. We find that the seven different sets yield a variation of approximately 1 Hz, which is substantially larger than the differences between the two force fields, and some of them even favor our model over the original AMOEBA. We also find that our model always yields higher absolute J-coupling values than the original AMOEBA, which can be explained by our model sampling more structures with slightly more negative ϕ values, as indicated by the sampled ϕ distribution in Supplementary Figure S5.6, which is shifted toward $\phi = -120°$; this region lies on the steepest portion of the Karplus curve. Therefore, the NMR scalar coupling data do not provide sufficient resolution to rank the two force fields, but indicate that the present model gives solution-phase conformational behavior comparable to the original AMOEBA.

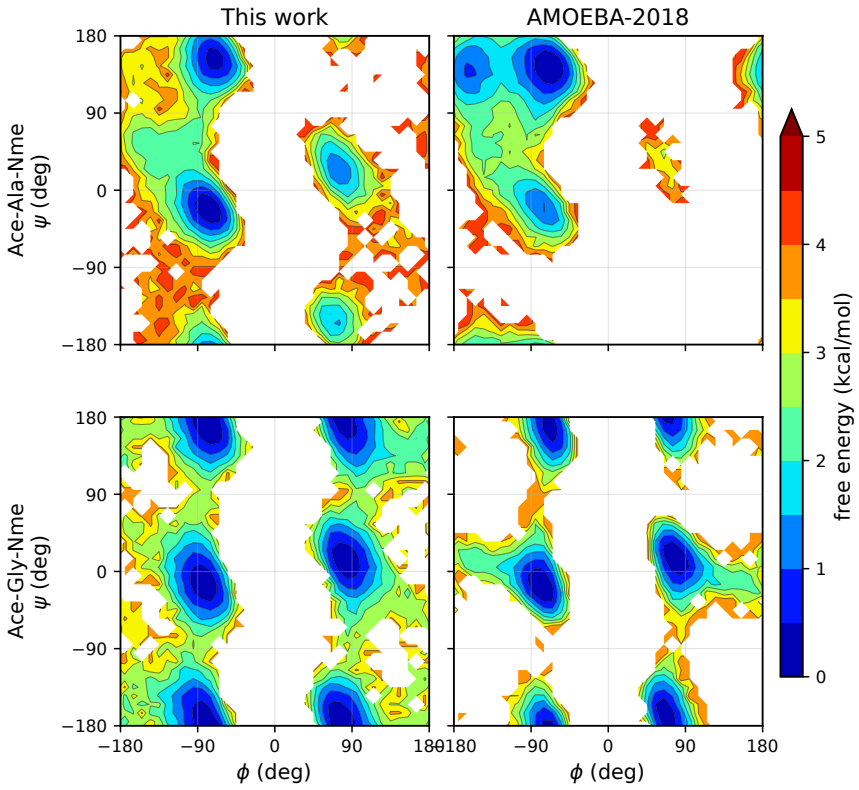

Figure 5.4: Condensed-phase two-dimensional ϕ/ψ potential of mean force (kcal/mol) at 300 K from temperature replica-exchange molecular dynamics, for the present model (left) and Original AMOEBA (right), for the alanine and glycine dipeptides (rows). The "AMOEBA-2018" label in the figure denotes Original AMOEBA.

Since ${}^3J(\mathrm{H_N}, \mathrm{H}_\alpha)$ depends only on the ϕ dihedral, we also plot and compare the 2D PMF on the ϕ/ψ space using the T-REMD trajectory at 300 K for both force fields (Figure 5.4). For alanine, our model successfully identifies the two major basins around the ppII and α_R regions, as well as the two shallow α_L and $\mathrm{C7_{ax}}$ regions. However, our model predicts the depths of the ppII and α_L basins equally, as opposed to the original AMOEBA which favors the ppII region. Interestingly, our model does not reproduce the β basin at the upper-left corner of the Ramachandran plot, which is visible in the original AMOEBA simulation as well as in the relaxed scan with the implicit-solvent model. For the achiral glycine dipeptide, our force field identifies the six symmetric basins but samples more extended structures near $(180°, 180°)$ than the original AMOEBA, which can be explained by the 2D MM relaxed-scan result in section 5.3.2, where our model reproduces that region better than the original AMOEBA does relative to QM. The differences between the two FFs are understandable because after being

fitted to the QM energies, the original AMOEBA's torsional parameters underwent several other iterations of manual adjustment to balance the 2D torsion distributions against PDB statistics.[257] We do not do this in our work because the goal of this research is to propose an accurate force field model with a reproducible fitting procedure, so that the same procedure can be easily generalized to other molecules while leaving enough room for potential changes to the model. Moreover, because PDB structures are biased toward folded proteins, if the statistics are not properly collected to remove such bias, empirical fitting against those data will erroneously reflect this bias and degrade the resulting FF's performance on, for example, intrinsically disordered proteins (IDPs).

The primary purpose of the explicit-solvent simulations is not to outperform Original AMOEBA, but to test whether the reparametrized solute model remains stable and compatible with the established AMOEBA water model. The simulations of both alanine and glycine complete 30 ns in each of the 36 replicas, including the highest temperature of 620 K. Completion of this high-temperature stress test supports the stability of the new bonded potential and unified scaling scheme in condensed-phase simulations.

5.4 Conclusions and Outlook

The conventional treatment of short-range intramolecular interactions (mainly interactions between 1–4 pairs) in molecular mechanics force fields combines bonded torsional potentials with scaled nonbonded terms. This construction makes the torsional parameters dependent on the somewhat arbitrarily chosen nonbonded scaling convention and obscures the boundary between local and nonlocal physics. The problem is amplified in advanced polarizable force fields like AMOEBA, where permanent electrostatics, the fields used to determine induced dipoles, polarization energy, and van der Waals interactions follow different scaling rules. The resulting component-specific treatment leads to inaccuracies in reproducing the potential energies and underlying forces, and complicates both parameterization and implementation.

We previously proposed a bonded-only 1–4 treatment framework to mitigate such limitations, and in this work, we extend it to a polarizable force field. Specifically, we replace the component-specific nonbonded scaling rules in the current AMOEBA with a simple and unified topology-based exclusion rule, in which all nonbonded interactions through 1–4 are removed and interactions from 1–5 onward are retained in full. The 1–4 interactions are modeled entirely by the bonded-only model containing Morse bonds, cosine angles, periodic and harmonic torsions, as well as their coupling terms. After refitting the permanent atomic dipoles to compensate for the changes in the direct-induction scaling, we fit the bonded terms against QM reference data in an automated manner using the Q-Force toolkit. We apply this framework to model dipeptides representing protein backbones and show that it reproduces QM electrostatic properties with accuracy similar to that of the well-developed original AMOEBA parameters, but substantially reduces the errors in energy and force predictions. We also demonstrate the utility of this model in condensed-phase simulations

with explicit water, showing that the model produces smooth energy surfaces and NMR couplings comparable to the original AMOEBA model. These results establish the simplified treatment as a viable alternative for protein-backbone 1–4 interactions without empirical adjustment to experimental conformational populations.

Future development should extend this framework beyond the present protein-backbone test to a broader range of chemical environments. Transferable parameter assignment could be built from discrete, chemically interpretable atom and interaction types, or from machine-learned chemical perception that maps molecular environments directly to force-field parameters, as illustrated by data-driven models such as Espaloma and ByteFF[288–290] and powered by differentiable force field development toolkits. The same separation of local bonded and longer-range nonbonded physics can also be examined with more elaborate nonbonded interaction models, for example, the Completely Multipolar Model[245, 263], to provide a physically complete molecular mechanical model with near-QM accuracy.

5.5 Supporting Information

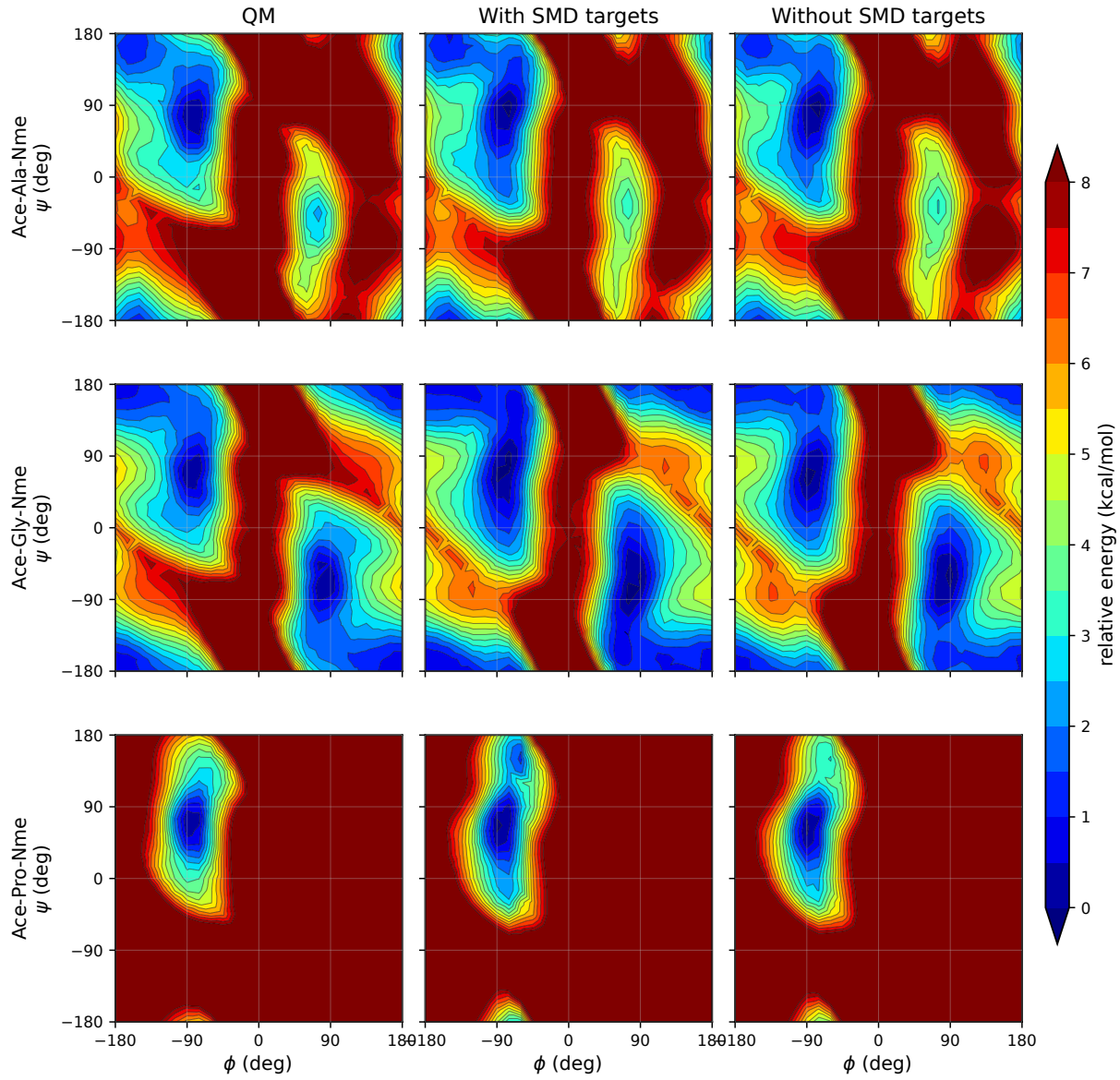

Figure S5.1: Gas-phase relaxed ϕ/ψ Ramachandran surfaces (relative energy in kcal/mol) for the alanine, glycine, and proline dipeptides (rows). Columns show the quantum-mechanical reference, the fit including SMD reference data, and the paired fit excluding SMD reference data.

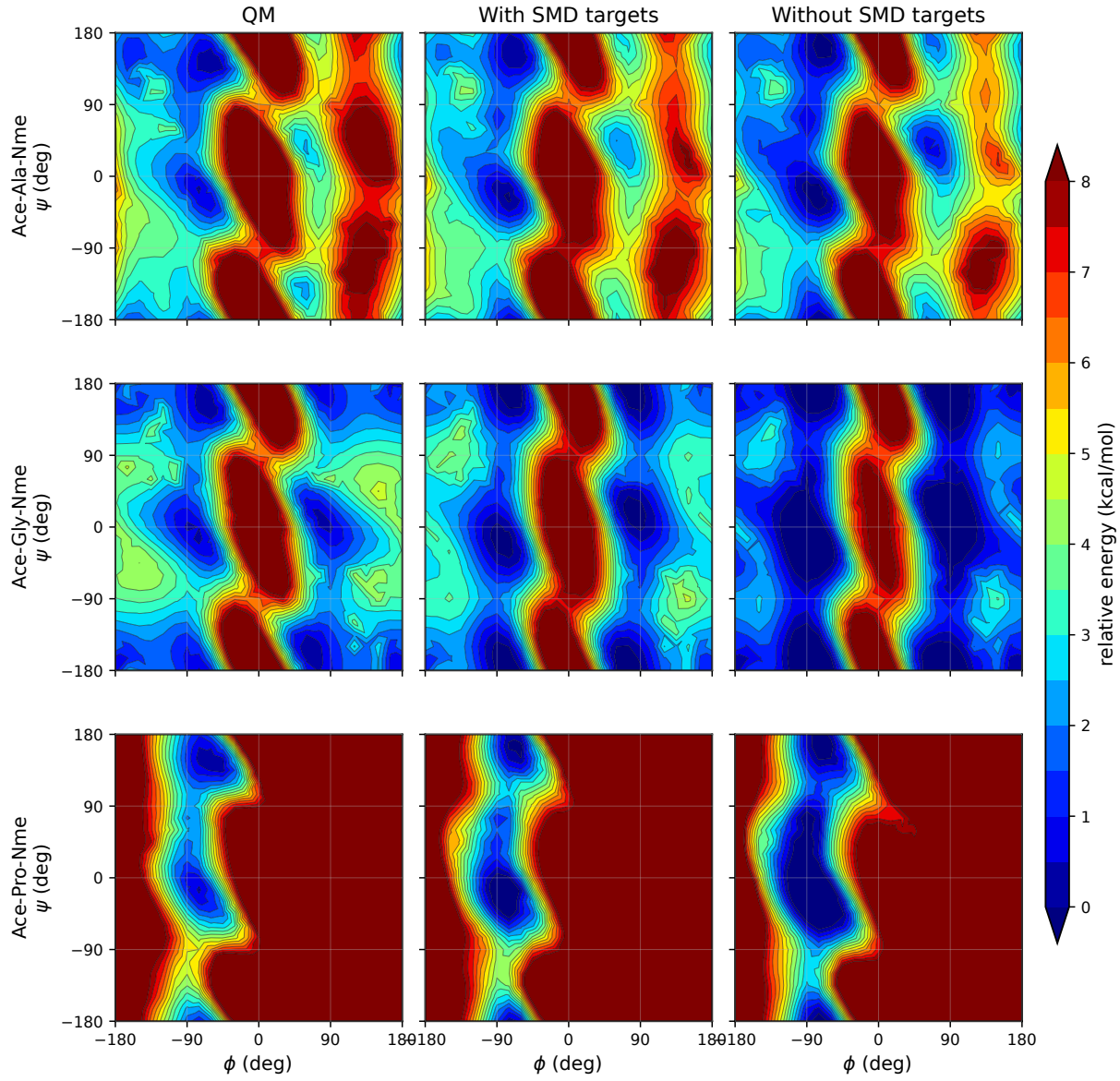

Figure S5.2: Implicit-solvent relaxed ϕ/ψ Ramachandran surfaces (relative energy in kcal/mol), organized as in Figure S5.1. The quantum-mechanical reference uses SMD continuum water and the molecular-mechanics surfaces use the AMOEBA Generalized-Kirkwood model.

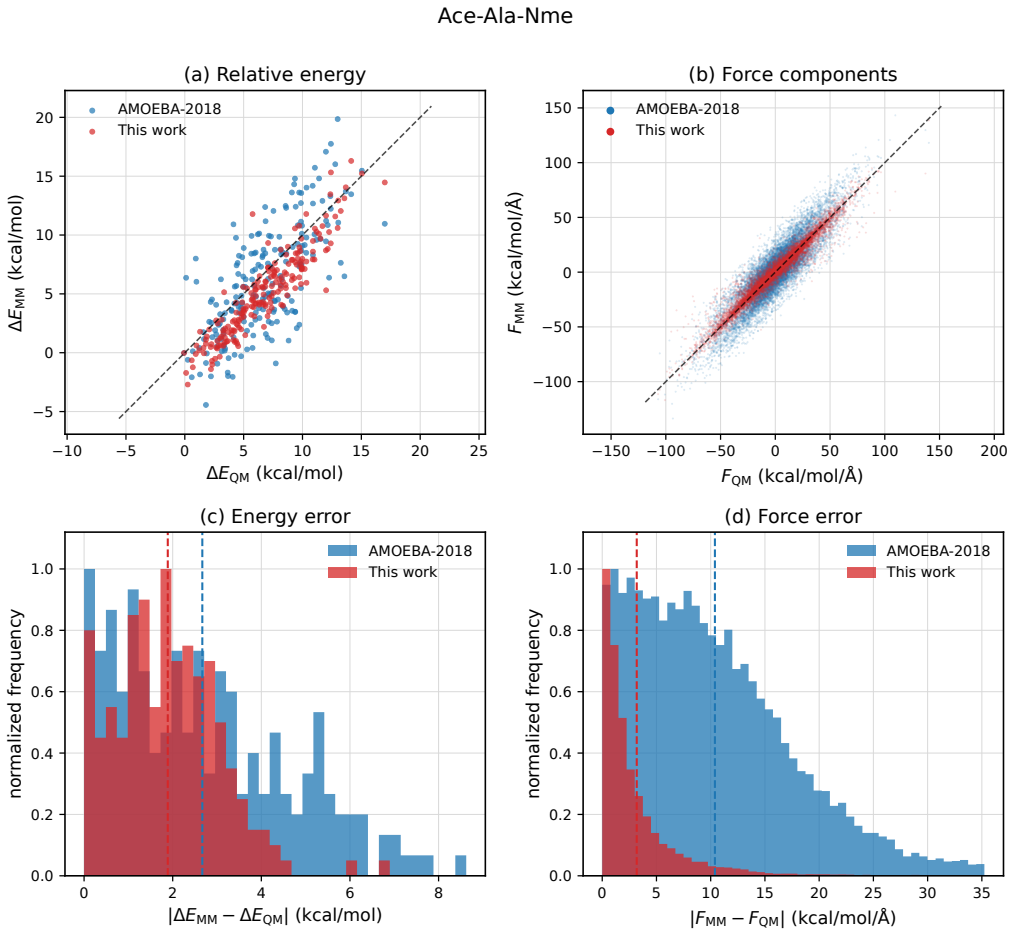

Figure S5.3: Energy and force accuracy on the 200 held-out alanine dipeptide validation conformers. Panels as in the main-text validation figure: (a) relative-energy correlation, (b) force-component correlation, and (c,d) peak-normalized absolute-error distributions with dashed MAE lines. The "AMOEBA-2018" labels in the figure denote the Original AMOEBA model.

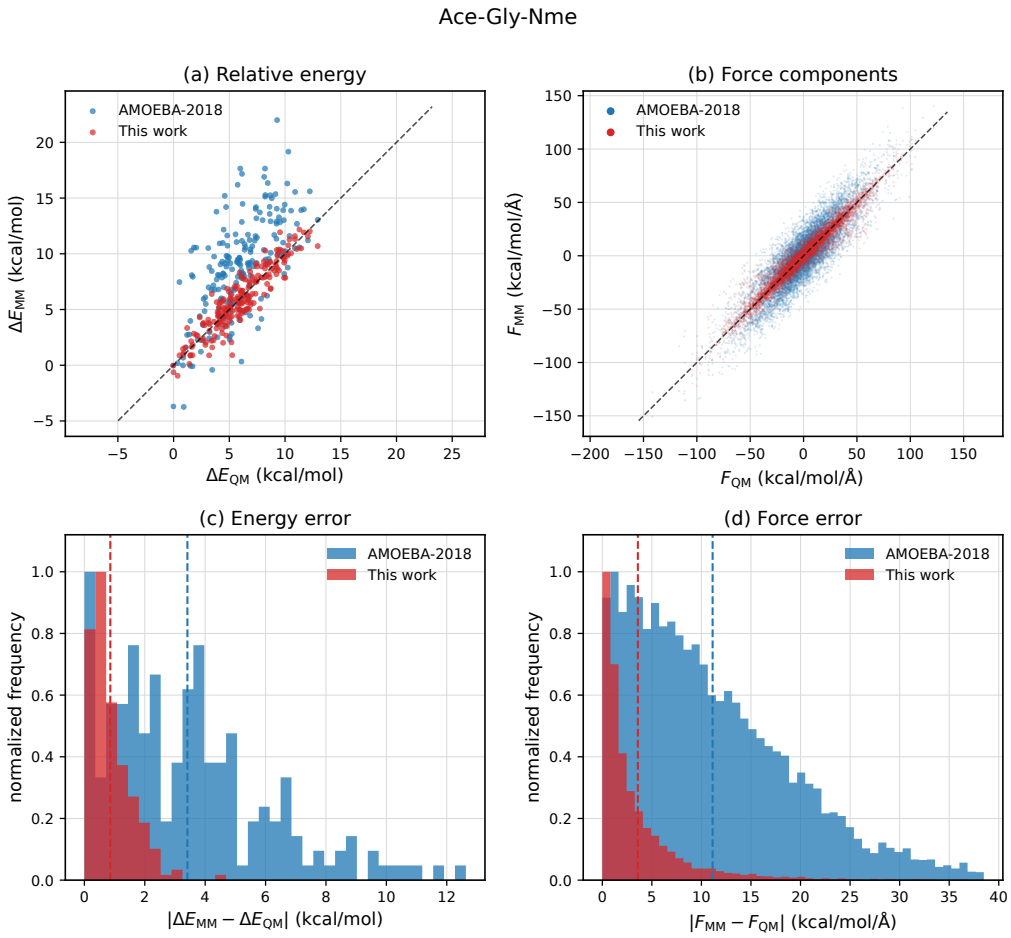

Figure S5.4: Energy and force accuracy on the 200 held-out glycine dipeptide validation conformers, organized as in Figure S5.3. The "AMOEBA-2018" labels in the figure denote the Original AMOEBA model.

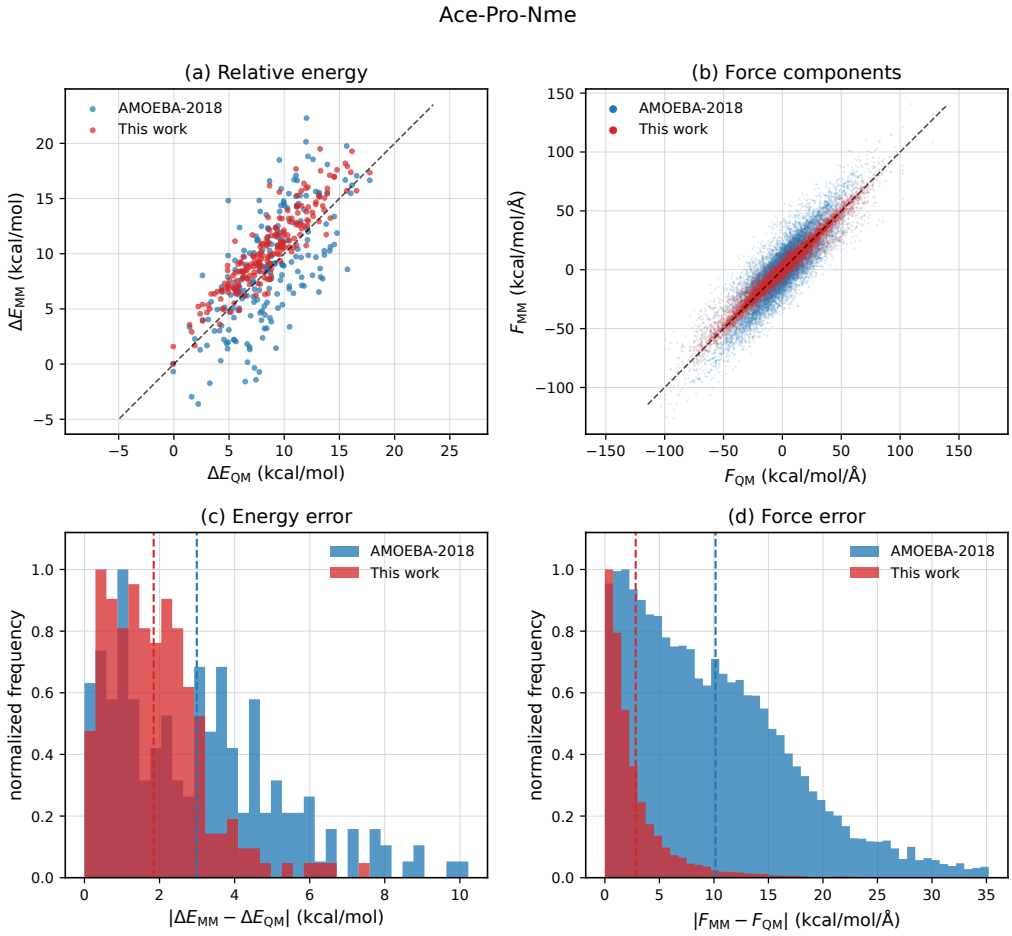

Figure S5.5: Energy and force accuracy on the 200 held-out proline dipeptide validation conformers, organized as in Figure S5.3. The "AMOEBA-2018" labels in the figure denote the Original AMOEBA model.

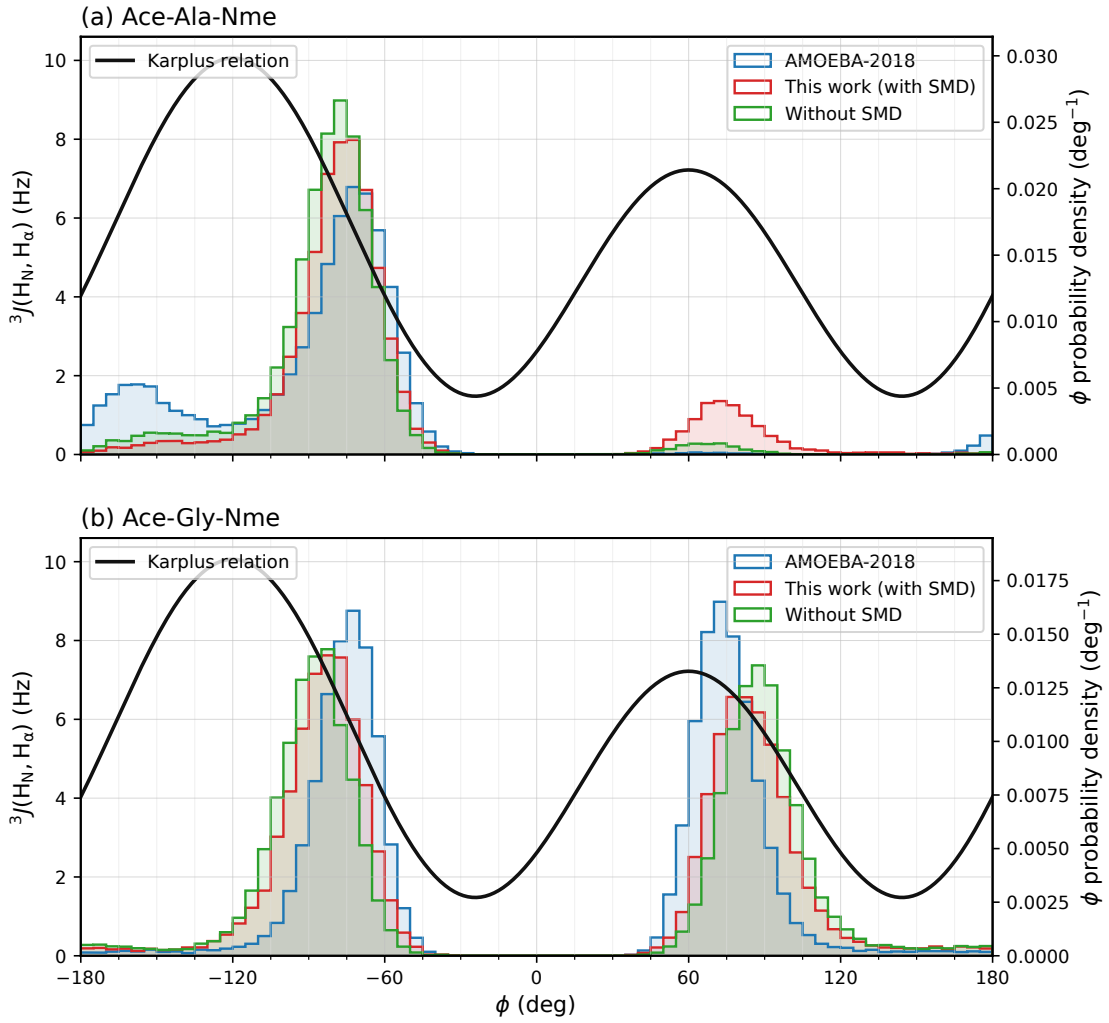

Figure S5.6: The Karplus curve used for the main-text analysis overlaid with the 300 K ϕ probability densities for alanine (a) and glycine (b). Distributions are shown for Original AMOEBA, the present fit including SMD reference data, and the paired fit excluding SMD reference data. The "AMOEBA-2018" labels in the figure denote Original AMOEBA.

5.6 Supplementary Tables

Table S5.1: Literature coefficient sets used to evaluate the backbone $^3J(\mathrm{H_N}, \mathrm{H}_\alpha)$ Karplus relation. The coefficients A, B, and C are in Hz and $\Delta\phi$ is in degrees.

Set	A	B	C	$\Delta\phi$	Reference	Description
1	7.09	-1.42	1.55	-60.00	[282]	Ubiquitin couplings fitted to solution-derived ϕ angles
2	9.44	-1.53	-0.07	-60.00	[291]	DFT fit for optimized Ace–Ala–NMe conformers
3	9.14	-2.28	-0.29	-64.51	[291]	DFT fit for optimized Ala–Ala–NH_2 conformers
4	7.90	-1.05	0.65	-60.00	[292]	Self-consistent fit to flavodoxin NMR couplings
5	9.50	-1.40	0.30	-60.00	[293]	Zero-motion extrapolation from empirical protein calibrations
6	-4.02	-0.54	6.80	0.10	[294]	B972/EPR-III fit to scanned glycine conformers
7	-4.60	-0.69	7.63	0.80	[294]	B3LYP/EPR-III fit to scanned glycine conformers

Table S5.2: Ensemble-averaged backbone $^3J(\mathrm{H_N}, \mathrm{H}_\alpha)$ scalar couplings (Hz) from the 300 K replica for Original AMOEBA and the paired bonded fits trained with and without SMD reference data. Values are the running averages after 30,000 frames (30 ns). Experimental values are 6.06 Hz for alanine and 5.85 Hz for glycine.[284]

Peptide	Model	Set 1	Set 2	Set 3	Set 4	Set 5	Set 6	Set 7
Ala ($J_{\mathrm{exp}} = 6.06$)	Original AMOEBA	6.30	6.00	6.55	5.57	6.31	5.68	6.33
	With SMD targets	6.43	6.25	6.83	5.82	6.59	6.18	6.89
	Without SMD targets	6.79	6.65	7.38	6.11	6.97	6.21	6.92
Gly ($J_{\mathrm{exp}} = 5.85$)	Original AMOEBA	6.31	6.32	6.27	6.02	6.74	6.20	6.93
	With SMD targets	6.49	6.51	6.58	6.15	6.92	6.37	7.13
	Without SMD targets	6.57	6.60	6.68	6.23	7.01	6.45	7.22

Table S5.3: Mean absolute error (MAE) and root-mean-square error (RMSE), in kcal/mol, of the relaxed alanine, glycine, and proline ϕ/ψ surfaces for bonded fits with and without the SMD surfaces included as fitting targets.

		Gas		Implicit solvent	
Peptide	Fit	MAE	RMSE	MAE	RMSE
Ala	With SMD targets	0.70	0.90	0.66	0.87
	Without SMD targets	0.70	0.92	1.10	1.40
Gly	With SMD targets	0.85	1.12	0.65	0.80
	Without SMD targets	0.62	0.86	1.40	1.61
Pro	With SMD targets	1.14	1.53	1.18	1.55
	Without SMD targets	1.01	1.36	2.04	2.44

Chapter 6

EasyBFE: An Open-source Package to Automate Protein-Ligand Binding Free Energy Calculations

6.1 Introduction

Accurate prediction of protein–ligand binding affinity is a central objective of structure-based drug discovery, guiding critical decisions regarding which molecular designs should advance to synthesis and experimental evaluation. While molecular docking and end-state methods can evaluate large numbers of candidates at a relatively low computational cost, their predictive accuracy is largely limited by approximations in conformational sampling, underlying physical models, and solvation effects. Such inaccuracies restrict their utility during lead optimization, a stage where it is essential to reliably resolve the small energetic differences that separate chemically similar compounds [8, 295]. In contrast, alchemical binding free energy calculations provide a more rigorous alternative by evaluating free energy differences between molecular states using explicit statistical-mechanical ensembles. As a result, these methods can achieve predictive accuracies approaching experimental error (typically 1–2 kcal/mol). When rigorously validated for a specific system of interest, they can effectively function as a "computational assay" that reliably drives decision-make-test-analyze (DMTK) cycle [10, 296].

Despite their utility, the practical application of alchemical binding free-energy calculations is hindered by the complexity of such workflows. Reliable execution demands meticulously coordinated steps: system preparation and parameterization, alchemical pathway design, simulation protocol construction, simulation execution, and rigorous statistical analysis [9, 83]. Furthermore, each stage requires critical scientific decisions and substantial manual intervention to bridge scattered software tools. For instance, in relative binding free energy (RBFE) calculations, which determine the affinity difference between two ligands via alchemical

perturbation from one to the other, accuracy is largely affected on optimal atom mapping and perturbation graph construction [9, 297, 298]. Absolute binding free energy (ABFE) calculations, which directly predict binding affinity of given ligand, involve even more intricate alchemical processes, and the selection of atoms forming the geometric restraints is challenging [87, 299, 300]. As a result, these techniques remain largely restricted to domain experts and are difficult to automate and scale for high-throughput, prospective drug discovery.

To address these technical barriers, an extensive software ecosystem has emerged, encompassing commercial platforms [10], open-source end-to-end workflows [301–306], and specialized utility packages [297, 298]. However, despite these advancements, current open-source tools suffer from two fundamental architectural limitations. First, absolute binding free energy (ABFE) and relative binding free energy (RBFE) calculations are largely fragmented. Most open-source packages restrict themselves exclusively to either RBFE [301–303] or ABFE [304–306], forcing researchers to navigate disjointed software environments, inconsistent APIs, and distinct file formats when switching between different drug-discovery tasks. Second, many existing software tools operate as monolithic assemblies of ad-hoc scripts tied to specific simulation protocols, rendering them fragile, difficult to maintain, and non-extensible to novel force fields, atom-mapping algorithms, or sampling strategies. OpenFE [86] represents a notable step forward in addressing these software engineering challenges through a modern, modular, and extensible architecture. However, it is primarily tailored to the OpenMM engine, leaving a critical void for workflows relying on AMBER—a widely deployed engine featuring mature GPU acceleration, advanced soft-core potentials, and enhanced sampling algorithms [84, 307].

To bridge this gap, I developed **EasyBFE**, an open-source, modular, and extensible software package tailored for AMBER-based alchemical binding free energy calculations [83]. EasyBFE unifies both RBFE and ABFE workflows within a single, consistent object-oriented architecture, allowing the choice of ABFE or RBFE to follow the underlying chemical question rather than software limitations. Rather than introducing a new molecular dynamics engine, EasyBFE serves as an end-to-end orchestrator that seamlessly connects receptor preparation, ligand parameterization, atom mapping, perturbation network design, AMBER alchemical system construction, distributed HPC execution, and statistical analysis through a common set of abstractions and APIs. Through standardized plugin interfaces, EasyBFE functions both as a robust, automated pipeline for routine CADD applications and as an extensible development environment for benchmarking new force fields, mapping algorithms, restraint strategies, and simulation protocols.

The remainder of this paper is organized as follows. First we will detail the software design principles, basic alchemical free energy theory, alongside the specific implementation details for both RBFE and ABFE workflows in EasyBFE. Then we will demonstrate the practical utility of EasyBFE across two real-world drug discovery campaigns: lead optimization targeting the SARS-CoV-2 main protease (M^{pro}) using RBFE, and hit identification targeting PGK2 in CACHE Challenge #7 using ABFE. The final section will conclude with a summary of

EasyBFE's contributions and future perspectives on integrating the package into autonomous, agentic AI-driven discovery workflows.

6.2 Theory

6.2.1 Software Design Principles

An ideal computational tool for alchemical binding free-energy calculations must serve as a unified workspace. While there have been a lot of efforts to automate these calculations, existing solutions typically suffer from critical drawbacks: they often focus narrowly on isolated components (such as atom mapping or perturbation graph construction), function as simple assemblies of scripts that lack long-term extensibility and sustainability, or restrict themselves to exclusively supporting either ABFE or RBFE workflows. Consequently, researchers have historically been forced to navigate disjointed packages and inconsistent APIs to execute calculation for specific needs [86, 301, 304]. A truly effective platform must eliminate this friction by consolidating these methods under a single interface while accommodating a diverse, collaborative user base. For experimentalists, junior computational chemists, or AI researchers seeking rigorous physical validation, the software must function as a robust, automated end-to-end engine that reliably generates predictions with validated defaults, requiring minimal (idealy no) human intervention. Senior CADD experts, however, require granular control to tune simulation parameters for complex targets, while method developers need a plug-and-play environment to test novel force fields or mapping algorithms without hacking into a hard-coded pipeline. Underpinning all these use cases is the absolute necessity for the software to scale flawlessly to meet the need of real-world discovery campaigns, which leads to the following principles of software design and engineering:

- **Modular architecture and separation of concerns.** Unlike a monolithic assembly of ad-hoc scripts commonly used to string together specific free-energy calculations, EasyBFE is fundamentally built upon a modular architecture. The core operations of the simulation pipeline, such as ligand parameterization, explicit solvent system setup, MD input generation and exceution, and statistical analysis, are decoupled into strictly independent components. Rather than implementing RBEF or ABFE workflows as entirely separate codebases, these macroscopic thermodynamic assays are designed as higher-level orchestrators that utilize the exact same underlying foundation. This separation of concerns prevents code duplication and ensures strict physical consistency across different assays. Furthermore, this foundation makes the software inherently robust, as improvements to a core module immediately and uniformly benefit all supported free-energy protocols.
- **Extensibility through object-oriented interfaces.** While modularity dictates the high-level organization of the workflow, EasyBFE achieves algorithmic extensibility

through a standardized plugin architecture utilizing Object-Oriented Programming (OOP) principles. Operations that inherently require diverse methodological choices, such as ligand force fields, constructing alchemical perturbation graphs, finding atom mapping, or selecting Boresch restraint atoms, are decoupled from the main execution logic. Instead, they are defined by standard interfaces (abstract base classes). Specific algorithms simply inherit from these interfaces and implement the required methods. This enables rapid integration of new techniques as method developers can implement and test a novel atom-mapping or restraint-selection algorithm simply by writing a single compliant class, without the need to decipher, hook into, or modify the broader automation workflow.

- **Progressive disclosure for diverse expertise levels.** EasyBFE is designed to accommodate researchers across varying levels of computational expertise. For scientists with entry-level experience in free-energy methods, standard calculations can be executed out-of-the-box using robust, physically validated default settings via a straightforward command-line interface (CLI). Meanwhile, EasyBFE remains fully controllable via a YAML configuration file or the Python API, allowing professional users to tune the settings, such as simulation protocols, λ scheduling, perturbation graph, to optimize the accuracy and reliability for specific systems of interest.

- **Automation with staged architecture.** Although EasyBFE is capable of executing an end-to-end automated workflow that takes a raw chemical structure and directly outputs a final free-energy number, it deliberately avoids functioning as an opaque "black box". The pipeline is carefully staged into discrete functional steps with clear command-line boundaries: ligand parameterization (`easybfe ligand pargen`), system setup and MD input generation (`easybfe abfe/rbfe setup`), MD execution (`sbatch`), and statistical result analysis (`easybfe abfe/rbfe analyze`). This staged architecture provides researchers with critical flexibility. Scientifically, it creates explicit review boundaries where experts can pause the workflow to inspect critical intermediate decisions, such as atom mapping or Boresch restraint selections, are desirable before committing expensive computational resources to the MD execution phase. Technically, this division allows the execution stage to be highly tailored to diverse computing environments; for instance, the independent thermodynamic legs generated by the setup command can be seamlessly distributed across multiple machines in a high-performance computing cluster if sufficient bandwidth is available.

6.2.2 Alchemical Transformation and Free Energy Estimators

Rather than their absolute values, the free-energy difference between specific initial and final states (denoted by "0" and "1") that define a chemical or physical process is generally of interest. In theory, the Gibbs free energy difference is directly related to the ratio of

configurational partition functions characterized by the potential energy U at temperature T [9, 308]:

$$\Delta G_{0\to1} = G_1 - G_0 = -\beta^{-1} \ln\left(\frac{Z_1}{Z_0}\right), \tag{6.1}$$

$$Z_i = \int \exp\left[-\beta U_i(\mathbf{x})\right] \mathrm{d}\mathbf{x}, \qquad \beta = (k_\mathrm{B}T)^{-1}, \tag{6.2}$$

Direct evaluation of this ratio, however, is generally difficult because the important regions of configuration space may differ substantially between the two states. Therefore, intermediate thermodynamic states, characterized by a parametric potential $U(\mathbf{x};\lambda)$, are introduced to ensure sufficient phase overlap. The two physical end states satisfy $U(\mathbf{x};0) = U_0(\mathbf{x})$ and $U(\mathbf{x};1) = U_1(\mathbf{x})$, whereas intermediate values of λ define nonphysical, or so-called "alchemical", states. For example, in terms of ligand binding, the two end states are the ligand fully interacts or completely does not interact with the environment and the intermediate states correspond to some partially interacting states, which are not physical. Once an alchemical path has been specified, the free-energy difference can be estimated from derivatives along the path or from reweighting the samples from different states.

Thermodynamic integration (TI) follows the first strategy by expressing the free-energy difference as a one-dimensional integral of the thermodynamic derivative [295, 308]:

$$\Delta G_{0\to1} = \int_0^1 \frac{\partial G}{\partial \lambda} d\lambda = \int_0^1 \left\langle \frac{\partial U(\mathbf{x};\lambda)}{\partial \lambda} \right\rangle_\lambda \mathrm{d}\lambda \tag{6.3}$$

where $\langle\cdot\rangle_\lambda$ denotes an equilibrium average at fixed λ. In practice, simulations are performed at a finite and discrete set of λ windows, so the integral must be evaluated by numerical quadrature. Consequently, TI requires both adequate sampling of the derivative within each window and sufficient λ resolution wherever the integrand is strongly curved.

Reweighting-based estimators provide a complementary strategy by evaluating the same configurations under neighboring potential-energy functions. The Zwanzig formula is the earliest approach in this regard [295]:

$$\Delta G = -\beta^{-1} \ln\langle \exp[-\beta(U_1(\mathbf{x}) - U_0(\mathbf{x}))]\rangle_0 \tag{6.4}$$

This one-sided perturbation estimate uses configurations from only one state and can therefore depend strongly on rare configurations when overlap is incomplete. In contrast, the Bennett acceptance ratio (BAR) method improves the estimate by using information more efficiently from both states [309], and the free energy difference is obtained by solving the following equation self-consistently:

$$\left\langle \frac{1}{1+\exp[\beta(U_1(\mathbf{x}) - U_0(\mathbf{x}) - \Delta G)]} \right\rangle_0 = \left\langle \frac{1}{1+\exp[\beta(U_0(\mathbf{x}) - U_1(\mathbf{x}) + \Delta G)]} \right\rangle_1 \tag{6.5}$$

Then, considering K states along the alchemical path (including two end states), the total free energy difference can be expressed by the sum of differences between neighboring states:

$$\Delta G_{0\to 1} = \sum_{i=1}^{K-1} \Delta G_{i\to i+1} \equiv \sum_{i=1}^{K-1} G_{i+1} - G_i \tag{6.6}$$

If the alchemical path is well designed, the overlap between adjacent intermediate states will be large enough so that each $\Delta G_{i\to i+1}$ can be reliably estimated by Equation (6.5).

The multistate Bennett acceptance ratio (MBAR) further generalizes the BAR method by applying the same reweighting principle to all sampled states simultaneously [310]. Specifically, for K states, let N_k be the number of uncorrelated samples collected at λ_k, let $u_k(\mathbf{x}) = \beta U(\mathbf{x};\lambda_k)$ be the reduced potential, and let $\{\mathbf{x}_n\}_{n=1}^{N}$ denote the pooled samples, where $N = \sum_{k=1}^{K} N_k$. MBAR then determines the dimensionless free energies $f_i = \beta G_i$ from the self-consistent equations and applies Equation (6.6):

$$f_i = -\ln \sum_{n=1}^{N} \frac{\exp[-u_i(\mathbf{x}_n)]}{\sum_{k=1}^{K} N_k \exp[f_k - u_k(\mathbf{x}_n)]}. \tag{6.7}$$

$$\Delta G_{0\to 1} = \sum_{i=1}^{K-1} \beta^{-1}(f_{i+1} - f_i) \tag{6.8}$$

By incorporating information from all sampled states, MBAR is asymptotically unbiased and statistically efficient under its assumptions [310]. Moreover, neither BAR nor MBAR relies on the potential derivative $\partial U/\partial\lambda$, making these methods more applicable when enhanced sampling biases are introduced along the alchemical path. In such cases, the intermediate-state potentials can be substantially altered, making analytical calculation of $\partial U/\partial\lambda$ cumbersome and even impractical. For these reasons, MBAR is chosen as the default free energy estimator in EasyBFE.

Beyond reporting a single free-energy estimate, EasyBFE leverages the `alchemlyb` package [85] to implement a complete analysis pipeline comprising equilibration detection and decorrelation [310], MBAR estimation, convergence analysis, block-averaging analysis, and phase-space overlap analysis. Together, these diagnostics enable users to assess statistical uncertainty and reliability of the final free-energy estimate and provide guidance to optimize the input settings [9].

6.2.3 Molecular Dynamics

Protein preparation. Preparing simulation-ready protein structures from raw experimental data is a critical prerequisite for molecular dynamics simulations, but the process has traditionally been labor-intensive and technically demanding, particularly for non-expert

users, because it often requires manually coordinating multiple programs or writing custom scripts. To address this bottleneck, EasyBFE provides an end-to-end automated preparation pipeline that minimizes manual intervention. Users can specify the polymer chains to retain, which is useful when the crystallized structure is a multimer, choose whether to preserve crystallographic waters, and remove irrelevant heterogens such as crystallization additives. The selected structure is then processed with PDBFixer to reconstruct missing atoms and residues, cap unresolved termini where appropriate, and assign protonation states at a user-specified pH while accounting for nearby metal coordination [311]. Finally, a constrained energy minimization relaxes the coordinates of newly added atoms while keeping the experimentally resolved structure fixed. The complete workflow is available through both the Python API for workflow integration and the `easybfe protein fix` command-line interface (CLI) for execution with one single command.

Ligand parameterization. The ligand force field fundamentally dictates the accuracy of binding free-energy calculations. Because the field is actively evolving and no single force field is universally superior across all chemical spaces, researchers often rely on retrospective studies to identify the optimal parameters for their specific target systems. Recognizing this critical need for flexibility, EasyBFE leverages the abovementioned object-oriented plugin architecture to decouple ligand parameterization from the broader execution workflow. Through a standardized interface, the platform natively supports major force field families, including GAFF and OpenFF [312, 313], with configurable partial-charge assignment methods Gasteiger, AM1-BCC [314, 315], or RESP. Furthermore, this extensible design allows users to seamlessly register custom parameterization algorithms (for instance, to integrate closed-source force fields) simply by implementing the standard interface, without needing to modify the core codebase. This parameterization protocol can be invoked programmatically via the Python API or executed as reproducible batch jobs using the `easybfe ligand pargen` command-line interface. Regardless of the chosen backend algorithm, the unified output is always a dedicated directory containing the force field parameters in AMBER `prmtop`, GROMACS `top`, and OpenMM XML formats, guaranteeing a seamless handoff for downstream system construction.

Molecular Dynamics. The current version of EasyBFE uses AMBER as the alchemical molecular dynamics engine. AMBER is available at no cost for academic use, provides mature GPU-accelerated free-energy functionality, and benefits from sustained development by an active community [83, 316]. EasyBFE supports releases after AMBER22 to use some advanced features, including the smoothstep soft-core potentials and ACES enhanced sampling [84, 307]. In addition to regularizing short-range interactions and avoiding endpoint singularities, smoothstep soft-core potentials provide a smoother alchemical pathway near the end states, thereby improving numerical stability and reducing estimator variance. ACES further accelerates sampling by scaling selected short-range interactions along the λ-path to lower both potential and kinetic barriers, while propagating enhanced samples to physical states through Hamiltonian replica exchange [317]. Together, these methods improve the

overall reliability and accuracy of alchemical free-energy calculations [318].

Instead of the `tleap` program distributed with `ambertools`, EasyBFE uses the OpenMM toolchain [232] to construct simulation system by combining protein and ligand structure, assigning force field parameters, defining water box, and adding solvent/counter ions. ParmEd then converts the OpenMM topology, system, and coordinates into the AMBER `prmtop` and `inpcrd` formats used for simulation. This choice is made because OpenMM provides a more programmable Python API, offering more flexibility for debugging, unittesting or configuration (for example, deploying non-AMBER force fields). Whereas `tleap` requires extensive intermediate-file manipulation which is not elegant nor sustainable.

6.2.4 Relative Binding Free Energy

The primary goal of relative binding free energy calculation (RBFE) is to evaluate the binding free energy difference of two different ligands (A and B) to the same target, which is particularly useful in lead optimization to answer whether a specific chemical modification improves binding affinity [9, 295]. By using the thermodynamic cycle as illustrated in Figure 6.1, the free energy differences in the two physical binding process equals to the differences between two alchemical process [295]:

$$\begin{aligned}\Delta\Delta G_{\text{bind}}^{\text{A}\to\text{B}} &= \Delta G_{\text{bind}}(\text{B}) - \Delta G_{\text{bind}}(\text{A}) \\ &= \Delta G_{\text{complex}}^{\text{A}\to\text{B}} - \Delta G_{\text{solvent}}^{\text{A}\to\text{B}}.\end{aligned} \tag{6.9}$$

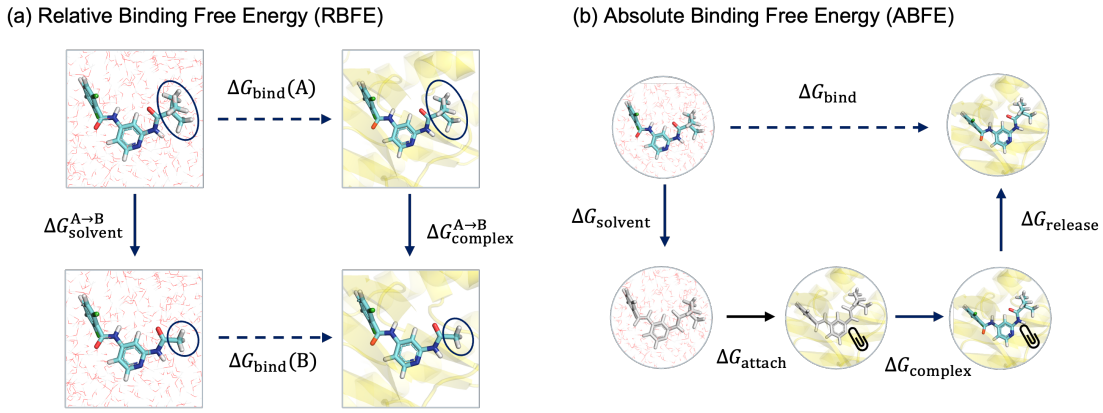

Figure 6.1: Thermodynamic cycles implemented for (a) relative and (b) absolute binding free-energy calculations. In the RBFE cycle, ligand A is transformed into ligand B in bulk solvent and in the solvated protein–ligand complex. In the ABFE cycle, the ligand is decoupled from solvent, attached to the binding site with restraints, coupled back with the protein, and released from the restraints. Dashed arrows denote the physical binding processes, and solid arrows denote the alchemical legs.

Instead of performing two absolute binding free energies (ABFEs) that simulate the entire process from the "unbound" state to "bound" state, which is relatively hard to implement and converge as discussed in the following section, RBFE converts the problem by directly connecting two physical ligands through a relatively short alchemical path and evaluating the differences caused by this small perturbation under two environments. When the two ligands share a common binding mode and a large chemical core, most of interactions with the surrounding environment remain similar along the alchemical path. RBFE benefits from this similarity in two ways. First, the resulting perturbation is typically shorter and smoother, providing better configurational overlap between neighboring states and reducing the sampling effort and number of intermediate λ windows required. Second, systematic errors in the force field and simulation protocol can partially cancel between the two thermodynamic legs and between the similar end states [9, 308].

Topology construction and Atom mapping. A key component in RBFE is how to setup the topology that represents the alchemically modified system. Common approaches include dual, single or hybrid topology [9, 308]. Dual topology represents the changing molecules separately by two non-interacting set of atoms, each with independent coordinates. Although it is straightforward and technically feasible for dissimilar molecules, the dual topology requires more λ states and extensive sampling, as it is essentially the same as two simultaneous ABFE calculations with complete ligand creation and annihilation occurring oppositely. Single or hybrid topology approaches address this problem by identifying "common-core" and "soft-core" regions and treating them differently. The common core is a shared substructure of the two ligands in which the atoms retain the same set of coordinates and topological structure, but undergo changes in force field parameters. The remaining part is often referred to as "soft core" atoms, because they are not mapped to physical atoms in the other state and rely on soft-core potentials to reliably simulate vanishing or appearing from noninteracting "dummy state".

This construction, however, sacrifices some flexibility because the two ligands must be sufficiently similar to define a reasonable common core. The distinction between single and hybrid topology is primarily implementation-specific, depending on whether the shared coordinates of the common-core atoms are encoded directly in the topology or enforced dynamically through constraints during the simulation.

An additional challenge introduced by the single or hybrid topology approach is how to find the atoms that define the common core region and their one-to-one mapping. A reliable map should not only maximize the number of common-core atoms, but also apply constraints to avoid nonoptimal or even incorrect end-state descriptions caused by bond, hybridization or ring size changes. Automatic methods generally follow either a two-dimensional graph-based or a three-dimensional geometry-based strategy. LOMAP [297] represents the former approach: it selects the promising maximum-common-structure (MCS) candidate by a elaborately designed score that accounts for chemical rules including ring size, sterochemistry, hybridization etc. Kartograf instead maps atoms directly from pre-aligned 3D ligand structures and then applies

chemical filters to avoid problematic correspondences [298]. Each of the approach has its own advantage. Three-dimensional methods avoid the expensive MCS searches used in 2D methods and can resolve ambiguities associated with symmetry and stereochemistry, but 2D methods does not require high-quality starting poses and reliable pre-alignment, which may be difficult to obtain in practice. Furthermore, for some edge-cases, automatic mapping results can remain suboptimal and may require expert adjustment [86].

Through the plugin mechanism described above, EasyBFE integrates two commonly used algorithms, LOMAP and Kartograf, and also supports defining a common core via a SMILES or SMARTS pattern. Users can also specify the mapping directly using atom-index pairs, which is useful for manually adjusting automated results. Once the mapping is established, EasyBFE reorders the ligand atoms to fulfill AMBER's requirement that the order of common-core atoms in both ligands must be identical. It also prepares the mask selection string to define the soft-core atoms. Additionally, EasyBFE forcibly aligns the common-core coordinates to ensure they are exactly identical, satisfying another key geometric requirement of AMBER.

Perturbation Graph. Lead optimization often requires ranking a series of molecules on a common affinity scale, which can be achieved by reconstructing absolute binding affinity values from pairwise RBFE estimates. For this purpose, a *perturbation graph* is constructed, in which each node represents a ligand and each edge represents an alchemical transformation between two ligands. The graph must be connected so that every ligand can be related to every other ligand through one or more transformations. A fully connected graph, however, is neither necessary nor generally desirable, because some ligand pairs may be too dissimilar for a reliable RBFE calculation. Instead, the graph should preferentially connect chemically similar ligands while retaining sufficient redundancy. In particular, the inclusion of cycles enables *cycle-closure analysis*: because free energy is a state function, the sum of the calculated free-energy changes around any closed cycle should be zero within statistical uncertainty. Deviations from cycle closure therefore provide a useful diagnostic for identifying inconsistent or poorly converged transformations and can also be used to improve the global estimation of absolute binding free energy values [10, 297].

EasyBFE supports several perturbation-graph construction algorithms suited to different circumstances. A star graph, i.e. the single-edge RBFE calculations, connects every ligand to one reference ligand and is useful during early-stage hit expansion, when chemical space must be explored at relatively high throughput. A bi-star or wheel graph introduces additional edges to enable cycle-closure analysis at minimal extra cost. The minimum spanning tree (MST) selects the minimum $N - 1$ transformations required to construct a connected graph while maximizing the overall pairwise similarity. The minimal redundant tree is constructed by running the MST procedure multiple times, thereby introducing redundancy and ensuring that each node is connected by at least two edges. This design improves robustness because the network can remain connected even if some subsequent simulations fail. The LOMAP algorithm further balances ligand similarity, path length, and cycle coverage. As with atom-

mapping algorithms, however, no single graph-construction strategy is universally optimal. EasyBFE therefore supports user-defined networks, allowing users to construct graphs based on chemical intuition or import them from other software.

Finally, EasyBFE leverages the maximum-likelihood estimation algorithm to obtain ligand-level estimates of the absolute binding free energies from the RBFE network. Assuming that the statistical errors of the pairwise estimates are independent and normally distributed, the algorithm determines the absolute binding energies by maximizing the following likelihood[319]:

$$\mathcal{L}(\{\Delta G_i\}) = \prod_{(i,j)\in E} \frac{1}{\sqrt{2\pi\sigma_{ij}^2}} \exp\left\{-\frac{1}{2\sigma_{ij}^2}\left[\Delta\Delta G_{i\to j} - (\Delta G_j - \Delta G_i)\right]^2\right\}. \tag{6.10}$$

Here, E is the set of edges in the perturbation graph, $\Delta\Delta G_{i\to j}$ is the calculated relative binding free energy from ligand i to ligand j, σ_{ij} is its statistical uncertainty, and ΔG_i is the estimated absolute binding free energy of ligand i. Because the likelihood depends only on pairwise differences, these absolute estimates are defined only up to a common additive constant. When experimental affinities are available, the estimates are therefore commonly shifted so that their mean matches the experimental mean in order to ensure a fair comparison against experimental data.

Constrained Ligand Docking. RBFE calculations are highly sensitive to the quality of input structures, necessitating that paired ligands are aligned to reflect a shared binding mode. Furthermore, single- or hybrid-topology methods require the input coordinates of the common-core atoms to be strictly identical between the two ligands. EasyBFE addresses this requirement through a robust constrained docking workflow that orchestrates conditioned conformer generation followed by targeted geometry optimization.

Starting from a reference ligand with a known binding pose, EasyBFE first identifies a common substructure between the reference and the target ligand, either algorithmically via LOMAP or through direct user specification. Next, it leverages RDKit to generate a 3D conformer for the target ligand, constraining the coordinates of the mapped substructure to match those of the reference [320]. Because RDKit's conformer embedding is not protein-aware, the newly generated peripheral atoms may clash with the receptor pocket. To resolve these steric clashes, the structure undergoes a local optimization using AutoDock Vina [6]. However, since the Vina optimization can inadvertently shift the coordinates of the previously constrained core atoms, EasyBFE provides an optional, subsequent force-field optimization in the presence of the rigid protein. This final step forcibly restores and locks the common core to its exact reference coordinates while preserving a physical binding mode. Ultimately, this protocol guarantees that the generated ligand pose possesses identical coordinates for the shared core, strictly satisfying the geometric requirements of AMBER's hybrid-topology RBFE framework. This procedure is available both programmatically and via the command-line interface as `easybfe ligand cdock`.

Execution. As discussed above, RBFE calculations in EasyBFE are executed through a staged architecture. Specifically, users specify the paths to the parameterized ligands and the simulation settings within a YAML configuration file. The `easybfe rbfe setup` command then validates this input, generates the alchemical perturbation graph, and prepares the coordinate systems and executable scripts required to run the transformations for all perturbation edges in both the complex and solvent environments. At this intermediate stage, users can explicitly review the generated atom mappings to ensure physical correctness prior to simulation. The prepared scripts can then be submitted to a high-performance computing (HPC) scheduler for parallel invocation. Once the simulations conclude, users can collect the results using the `easybfe rbfe analyze` command. This step computes the free-energy estimate for each perturbation edge, accompanied convergence and phase-space overlap diagnostics, and applies the maximum-likelihood estimation (MLE) model to calculate the "absolute" affinities across the entire ligand network.

6.2.5 Absolute Binding Free Energy

EasyBFE follows a standard double decoupling scheme to calcuate ABFE as illustrated in Figure 6.1b [9, 87]. In this approach, the physical binding process of a ligand to its target receptor is decomposed into four alchemical steps: (1) Annihilating the "free" ligand from its physical environment by turning of its interaction with surrounding water; (2) Linking this "dummy" ligand to the target protein through Boresch-style harmonic geometric restraints; (3) Turning on the interaction between the ligand and its environment in "bound" state, including protein, water and counter ions with the restraints existing; (4) Releasing the Boresch restraints. For step (1), (2) and (4), the free energy change is evaluated alchemically by altering the corresponding potential energy term (Lennard–Jones, Coulomb with soft-core treatment or restraints). The free energy associated with attaching the restraints in step 2, whereas, is calculated analytically as described in the following. The total binding free energy is thus the sum of free energy changes of these four steps:

$$\begin{aligned}\Delta G &= G_{\mathrm{PL}} - G_{\mathrm{P}} - G_{\mathrm{L}} \\ &= -\Delta G_{\mathrm{solvent}} + \Delta G_{\mathrm{attach}} + \Delta G_{\mathrm{complex}} + \Delta G_{\mathrm{release}}\end{aligned} \tag{6.11}$$

The minus sign before the $\Delta G_{\mathrm{solvent}}$ is to follow the conventional definition of solvation free energy $\Delta G_{\mathrm{solvent}} = G_L^{\mathrm{water}} - G_L^{\mathrm{gas}}$.

Boresch restraints and the selection algorithm

As illustrated in Figure 6.2a, the dummy ligand is restrained to its binding pocket by applying harmonic restraint potential to six degrees of freedom formulated by three protein atoms $(\mathrm{P}_1, \mathrm{P}_2, \mathrm{P}_3)$ and three ligand atoms $(\mathrm{L}_1, \mathrm{L}_2, \mathrm{L}_3)$, including one bond $r(\mathrm{P}_1 - \mathrm{L}_1)$, two angles $\theta_1(\mathrm{P}_2 - \mathrm{P}_1 - \mathrm{L}_1)$ and $\theta_2(\mathrm{P}_1 - \mathrm{L}_1 - \mathrm{L}_2)$ and three dihedrals $\phi_1(\mathrm{P}_3 - \mathrm{P}_2 - \mathrm{P}_1 - \mathrm{L}_1)$,

$\phi_2(\mathrm{P}_1-\mathrm{L}_1-\mathrm{L}_2-\mathrm{L}_3)$, and $\phi_3(\mathrm{P}_2-\mathrm{P}_1-\mathrm{L}_1-\mathrm{L}_2)$. Without this special treatment introduced by Boresch et al. [87], the ligand will diffuse freely throughout the simulation box in the fully decoupled state. Since protein system is much more heterogeneous than water, this freely sampled configurations share little phase-space overlap with the physically bound state where the ligand should remain stable in the pocket. As a result, the corresponding alchemical free energy calculation will be much less reliable.

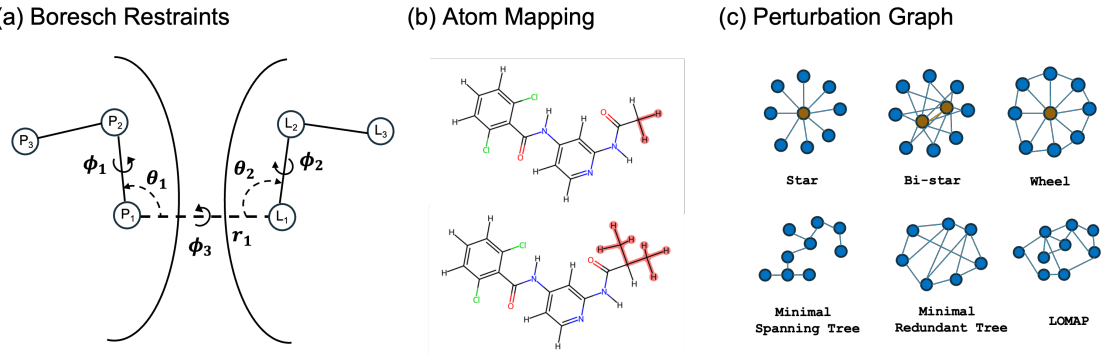

Figure 6.2: Illustration of scientific components in RBFE and ABFE. (a) The Boresch restraint scheme defines one distance, two angles, and three dihedrals from three protein anchors (P_1–P_3) and three ligand anchors (L_1–L_3). (b) Atom mapping identifies the common atoms between two ligands while the atoms that differ between the end states are highlighted in red. (c) Perturbation-network options include star, bi-star, wheel, minimum spanning tree, minimally redundant tree, and LOMAP designs. Blue nodes represent ligands, edges represent alchemical transformations, and brown nodes identify reference ligands.

Because the restraints are applied at the dummy state where the protein and ligand are not interacting, the free energy associated with applying the Boresch-style restraints can be derived analytically [87, 321]:

$$\Delta G_{\mathrm{attach}} = -\beta^{-1} \ln \frac{Z_r Z_{\theta_1} Z_{\theta_2} Z_{\phi_1} Z_{\phi_2} Z_{\phi_3}}{8\pi^2 V^\circ} \tag{6.12}$$

where

$$Z_r = \frac{r_0}{2\beta k_r} \exp(-\beta k_r r_0^2) + \frac{\sqrt{\pi}}{4\beta k_r \sqrt{\beta k_r}} \left(1 + 2\beta k_r r_0^2\right) \left[1 + \mathrm{erf}\left(\sqrt{\beta k_r}\, r_0\right)\right] \tag{6.13}$$

$$Z_{\theta_j} = \sqrt{\frac{\pi}{\beta k_{\theta_j}}} \exp\left(-\frac{1}{4\beta k_{\theta_j}}\right) \sin\theta_{j,0} \tag{6.14}$$

$$Z_{\phi_i} = \sqrt{\frac{\pi}{\beta k_{\phi_i}}} \mathrm{erf}\left(\pi\sqrt{\beta k_{\phi_i}}\right). \tag{6.15}$$

Here, $V^\circ = 1660$ Å^3 is the standard-state volume at 1 mol/L. By default, EasyBFE uses force constants $k_r = 10$ kcal/mol/Å^2 and $k_{\theta_j} = k_{\phi_i} = 10$ kcal/mol/rad^2; r_0, $\theta_{j,0}$, and $\phi_{i,0}$ are the corresponding equilibrium values.

Although the exact ΔG_{attach} is independent of the choice of anchor atoms in the limit of complete sampling, in practice they have to be selected algorithmically so that the restrained bonds, angles or dihedrals do not fluctuate drastically around equilibrium values. This enables faster convergence and avoids ill-defined dihedral values when the angle approaches 0° or 180°.

EasyBFE offers three approaches to select the Boresch restraint atoms. First of all, following previous work [300], EasyBFE first performs a short plain MD simulation (1 ns by default) of the protein–ligand complex, and identifies hydrogen bonds and salt bridges between the ligand and protein with occupancies above 50%. For each interaction, the interacting ligand atom, if contains two heavy-atom neighbors, is assigned as a candidate anchor atom L_1. Otherwise, the nearest atom satisfying this criterion will be selected. For example, a terminal carboxylate oxygen is mapped to the central carbon atom. Similarly, the C_α atom of the residue (denoted by i) containing the interacting atom on the protein will be selected as P_1 candidate. All valid six-atom combinations are then enumerated - L_2 is a heavy-atom neighbor of L_1, and L_3 is a heavy-atom neighbor of L_2. The protein anchors P_2 and P_3 are chosen from the neighboring backbone atoms (N_i, C_{i-1}) or (C_i, N_{i+1}). Candidates with a mean value of either θ_1 or θ_2 outside the range 45°–135° are discarded. Among the remaining candidates, those with L_1 closest to the ligand center of mass are prioritized, and the final anchor set is selected as the one with the lowest score:

$$S_{\text{rest}} = \frac{\sigma_r \, \sigma_{\theta_1} \, \sigma_{\theta_2} \, \sigma_{\phi_1} \, \sigma_{\phi_2} \, \sigma_{\phi_3}}{\left(\sin \overline{\theta}_1 \sin \overline{\theta}_2\right)^2}, \tag{6.16}$$

where σ is the standard deviation of the geometric variable. This MD-based selection is the most robust of the three options because it evaluates the stability of the anchors and minimizes the risk of encountering ill-defined dihedral constraints. EasyBFE also provides a faster alternative for Boresch restraint selection. It applies the same geometric selection procedure as the MD-based approach but evaluates only the initial structure, without performing MD or using MD-derived scores. Finally, EasyBFE supports users to define custom anchor atoms directly by specifying their atomic indices.

Execution. All ABFE-related functionalities are organized under the `easybfe abfe` command suite. EasyBFE provides both a staged workflow and an integrated pipeline for ABFE. As discussed previously, in the staged mode, ligand parameterization, system setup, simulation execution, and data analysis remain distinct operations. This modularity allows the three thermodynamic legs of an ABFE calculation to be independently distributed across high-performance computing resources and enables expert inspection of the prepared systems prior to job submission.

Conversely, the integrated pipeline accepts a prepared protein, a ligand, and a configuration file, consolidating all stages via a single `easybfe abfe pipeline` command. This mode

sequentially executes the three alchemical transformations directly on the host machine where the command is invoked. Additionally, the integrated mode features an early-stopping mechanism that permits users to define a preliminary affinity threshold. If the pre-production simulation phase yields an estimated ABFE value less favorable (more positive) than this predefined threshold, the pipeline automatically omits the final, computationally expensive production stage. This is useful to save resources in high-throughput screening where candidates may be notoriously bad. However, the specific threshold value must be carefully calibrated for the target system of interest to minimize the risk of false negatives.

6.3 Application

6.3.1 SARS-CoV-2 Main-Protease Lead Optimization

The SARS-CoV-2 main protease (M^{pro}) campaign demonstrated that EasyBFE can be applied to guide decision-making in lead optimization within a congeneric medicinal-chemistry series. Conducted as part of Project 3 ("Pandemic Virus Protease Inhibitors") under the Midwest AViDD Center, the project targeted the discovery and optimization of non-covalent, non-peptidic inhibitors of M^{pro} [322]. Initial hit discovery via DNA-encoded library (DEL) screening identified a benzimidazole series [323], from which MWAC-1871 emerged as a promising early hit with a biochemical IC_{50} of 0.45 μM. Subsequent chemical optimization produced MWAC-2748, which exhibited enhanced potency with a biochemical IC_{50} of 0.098 μM and a cell-based EC_{50} of 0.58 μM. Compound MWAC-2748 was subsequently selected as the optimization starting point for prospective RBFE calculations. The structural progression from MWAC-1871 to MWAC-2748 and the two synthesized prospective analogues is summarized in Figure 6.3.

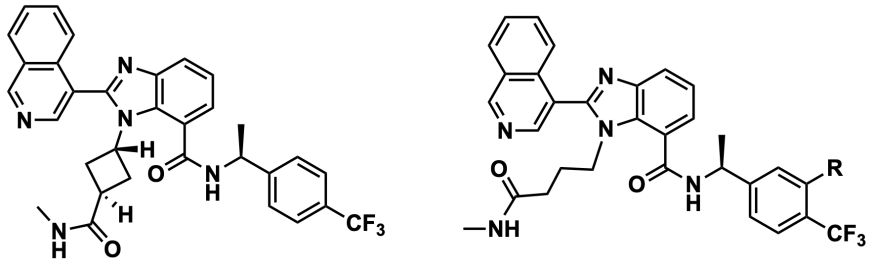

Figure 6.3: Chemical structures of MWAC-1871 and the MWAC-2748 analogue series evaluated in the SARS-CoV-2 M^{pro} lead-optimization campaign. MWAC-2748, MWAC-3290, and MWAC-3707 correspond to R = H, Cl, and CN, respectively.

Before prospective deployment, the RBFE protocol was retrospectively evaluated across 23 compounds synthesized and tested during the hit-to-lead stage (Figure 6.4). It should be noted that the experimental ΔG values in this retrospective benchmark were converted from biochemical IC_{50} measurements rather than obtained from direct biophysical binding assays, serving as an operational proxy to facilitate comparison with computational predictions. The protocol achieved a coefficient of determination (R^2) of 0.42 and a root-mean-square error (RMSE) of 0.94 kcal/mol relative to these experimental free-energy estimates. The majority of predictions fell within 1 kcal/mol of experimental values, with seven compounds showing errors between 1 and 2 kcal/mol, and only one outlier exceeding 2 kcal/mol. These retrospective findings confirmed that the workflow provided a reliable relative ranking to guide prioritization, while establishing that experimental validation remained essential.

Motivated by earlier structural observations involving naphthyl substitutions, four prospective functional groups (−F, −Cl, −CN, and $-CH_3$) were evaluated at the *ortho*-position of the 4-(trifluoromethyl)phenyl ring in MWAC-2748. These modifications yielded calculated $\Delta\Delta G$ values of −0.75, −1.34, −1.25, and −0.61 kcal/mol, respectively, indicating that substitutions at this position were generally predicted to favor binding affinity. The two top-ranked candidates were synthesized and experimentally evaluated. MWAC-3290 (R = Cl) demonstrated a nearly 2-fold improvement in biochemical potency ($IC_{50} = 0.057$ μM) and a nearly 5-fold enhancement in cellular activity ($EC_{50} = 0.12$ μM) relative to MWAC-2748. MWAC-3707 (R = CN) maintained the biochemical potency ($IC_{50} = 0.097$ μM) while yielding a modest 1.5-fold improvement in cellular antiviral activity ($EC_{50} = 0.38$ μM). Converting these biochemical IC_{50} values to experimental relative free energies yielded $\Delta\Delta G$ estimates of −0.32 kcal/mol for MWAC-3290 and 0.00 kcal/mol for MWAC-3707. Although RBFE was

somewhat optimistic regarding the absolute energetic gains of these two perturbations, it correctly predicted the relative ranking between them, identifying the chloro substitution as superior to the cyano group. These results showed RBFE, if validated by experimental data, can be applied to reliably prioritize structural hypotheses during lead optimization.

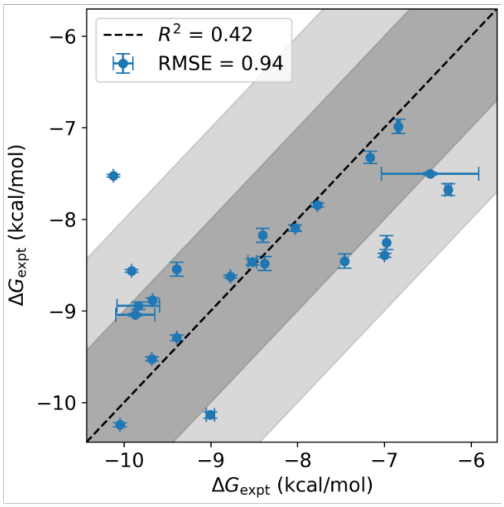

Figure 6.4: Retrospective comparison of calculated and experimental binding free energies for the SARS-CoV-2 Mpro inhibitor series studied within Project 3 of the Midwest AViDD Center. The dashed line denotes equality between calculation and experiment, whereas the dark and light shaded regions indicate deviations within 1 kcal/mol and 2 kcal/mol, respectively.

Table 6.1: Prospective substitutions evaluated relative to MWAC-2748 in the SARS-CoV-2 Mpro campaign. Free-energy differences are in kcal/mol, and activity measurements are in μM. NT denotes not tested.

Compound ID	R-group	IC_{50} (μM)	EC_{50} (μM)	$\Delta\Delta G_{RBFE}$ kcal/mol	$\Delta\Delta G_{expt}$ kcal/mol
MWAC-2748	H	0.098	0.58	0.00	0.00
N/A	F	NT	NT	-0.75	NT
MWAC-3290	Cl	0.057	0.12	-1.34	-0.32
MWAC-3707	CN	0.097	0.38	-1.25	0.00
N/A	CH_3	NT	NT	-0.61	NT

6.3.2 CACHE Challenge 7: De Novo Selective PGK2 Hit Design

For CACHE Challenge 7, we set out to identify orthosteric inhibitors of phosphoglycerate kinase 2 (PGK2) selective over its close homologue, PGK1 [324]. To explore a broad chemical space beyond fixed-size libraries, we initiated an LLM-based generative pipeline for *de novo* design. We first deployed SmileyLlama—a reinforcement-learning molecular generator pre-trained on approximately four million Enamine screening compounds—to propose ∼300,000 *de novo* candidate structures guided by multi-objective rewards incorporating Vina docking scores, predicted solubility, and quantitative estimate of drug-likeness (QED). To ensure chemical accessibility, we leveraged SynLlama, an LLM-based retrosynthetic planning tool that maps *de novo* designed molecules onto synthesizable chemical space. By combining SynLlama with commercial chemical-space searches, we identified approximately 100,000 synthesizable analogs with exact Enamine REAL identifiers from the initial 300,000 candidates. Because these LLM-generated candidates exhibited high chemical diversity without a shared core scaffold, we selected independent absolute binding free energy (ABFE) calculations for final physics-based prioritization rather than constructing long or weakly connected relative free energy (RBFE) networks across the entire set.

Crucially, alignment of the binding pockets revealed that Tyr242 in PGK2 is substituted by a phenylalanine residue in PGK1, whereas all other active-site residues are completely identical between the two isoforms. We therefore adopted the formation of a specific hydrogen bond with Tyr242 as a core structural hypothesis for isoform selectivity—a mechanism supported by the structures of two selective reference inhibitors identified from a DNA-encoded library (DEL) screen provided in the challenge. Prior to free-energy evaluation, we subjected the candidate pool to sequential coarse-grained primary filters: thresholds of Vina score < -9.0, UCBBind > 7.0, molecular weight < 500 g/mol, and predicted $\log S > -4.5$, alongside the geometric hydrogen-bond requirement with Tyr242. Candidates meeting these criteria underwent 1-ns explicit-solvent molecular dynamics (MD) simulations to verify the conformational stability of the proposed Tyr242 interaction.

Because a substantial number of candidates survived the MD filter, executing physics-based ABFE calculations across the entire remaining pool was computationally prohibitive. To address this, we implemented an active learning strategy to efficiently navigate the candidate space. We initially ran EasyBFE calculations on a seed batch selected based on favorable Vina docking scores and trained a machine-learning surrogate model (ML-ABFE) using molecular and interaction fingerprints. Given the surrogate model's modest predictive accuracy ($R^2 \approx 0.3$–0.4), we utilized it strictly as a coarse negative filter to discard clearly non-binding candidates (predicted ABFE > -6.0 kcal/mol) prior to launching full ABFE evaluations. Across three iterative rounds of active-learning filtering and physics-based scoring, we completed a total of 1,276 EasyBFE calculations with production run 2.5 ns for all alchemical legs.

The calculated binding free energies across these 1,276 completed records spanned a broad range from -14.45 to 4.69 kcal/mol. This substantial spread, despite all candidates having

passed the initial stringent Vina docking filter, highlights that Vina docking scores are an unreliable ranking method for affinity prioritization, particularly given that numerous candidates exhibited poor binding affinities and several even yielded positive ABFE values. In accordance with the CACHE Challenge requirements, we selected the top-200 ranked compounds for final submission, corresponding to a score cutoff of -7.89 kcal/mol. This shortlisted candidate pool proved highly promising when benchmarked against the DEL-derived selective reference binder (calculated ABFE = -11.56 kcal/mol): 16 of our candidate compounds surpassed this reference estimate. Furthermore, considering a standard computational uncertainty margin of 2.0 kcal/mol, 80 compounds achieved calculated binding free energies more favorable than -9.56 kcal/mol. As an example, we present the two-dimensional chemical structures of the five highest-ranking candidates by ABFE (ranging from -14.45 to -12.51 kcal/mol) in Figure 6.6.

In summary, this campaign demonstrates that EasyBFE is sufficiently robust and fully automated to execute campaign-scale, high-throughput free-energy calculations with direct practical impact. Deployed on NERSC supercomputing resources, our end-to-end pipeline—spanning *de novo* generation to final ABFE rescoring—was completed in just six weeks. We are currently awaiting experimental feedback from the CACHE Challenge and look forward to advancing these predictions through further experimental validation.

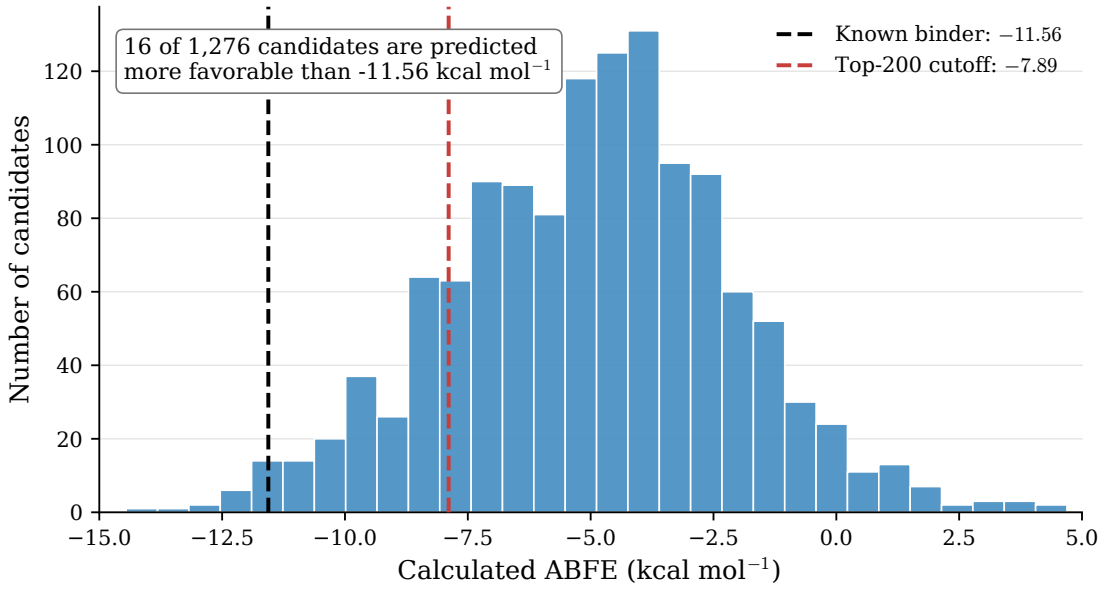

Figure 6.5: Distribution of calculated PGK2 absolute binding free energies for 1,276 LLM-generated candidates that reached the high-throughput EasyBFE stage of the CACHE Challenge 7 workflow. The black dashed line marks the known-binder reference at −11.56 kcal/mol, and the red dashed line marks the −7.89 kcal/mol cutoff for the 200 most favorable records.

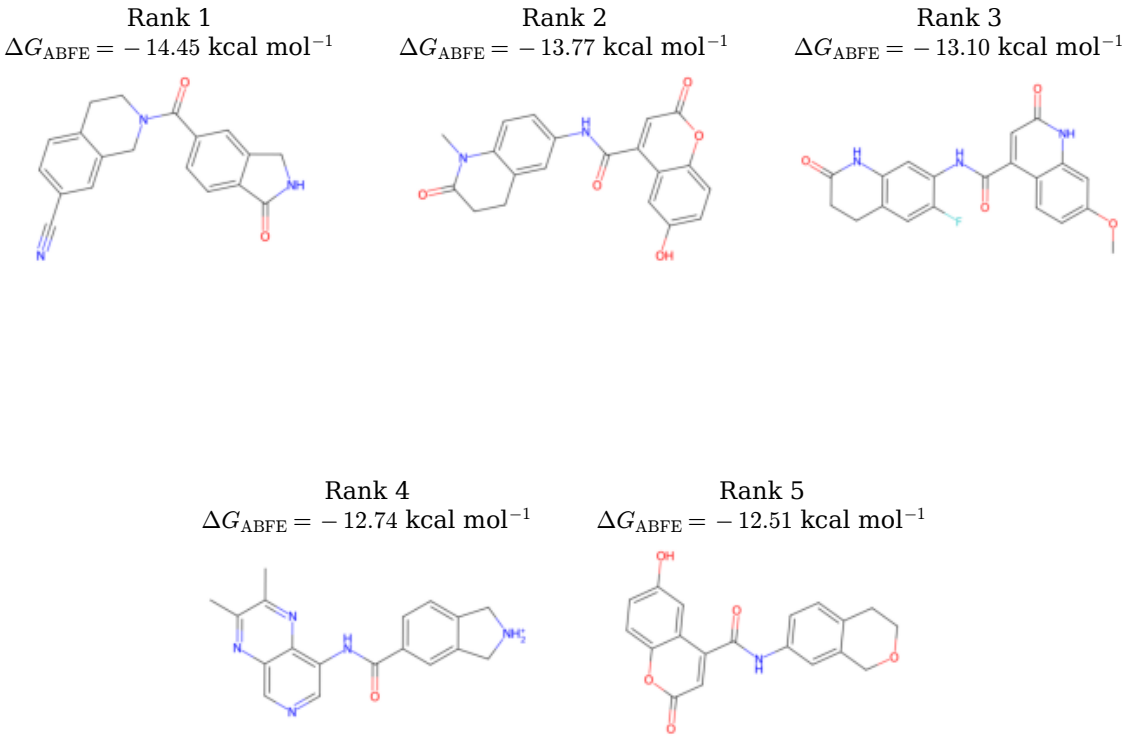

Figure 6.6: Two-dimensional structures of the five LLM-generated CACHE Challenge 7 candidates with the most favorable calculated PGK2 absolute binding free energies.

6.4 Conclusion and Outlook

In conclusion, this work introduced EasyBFE, an open-source, automated, end-to-end software package designed to streamline protein-ligand binding free energy calculations using the AMBER molecular dynamics engine. Despite the recognized theoretical accuracy of alchemical free energy methods, their routine application in drug discovery has historically been hindered by the fragmented nature of available tools and the extensive manual expertise required to manipulate complex workflows. EasyBFE addresses these critical and systematic bottlenecks by providing a unified platform that seamlessly integrates both relative (RBFE) and absolute (ABFE) binding free energy calculations within a single, cohesive architecture.

EasyBFE is built upon robust design principles tailored to meet the practical needs of users across varying levels of expertise, while ensuring the software remains sustainable and extensible for future development. Its modular architecture cleanly separates discrete stages of the simulation pipeline—including ligand parameterization, system construction,

MD execution, and free energy analysis. Accessible via programmable APIs and intuitive command-line interfaces, the entire workflow remains highly configurable and fully auditable. Furthermore, EasyBFE features an object-oriented plugin system that enables the seamless implementation of diverse algorithms through a uniform interface. This covers essential functions such as ligand parameterization, atom mapping, and perturbation graph generation, empowering developers to effortlessly integrate new algorithms as the field advances.

The practical utility of EasyBFE was demonstrated through two actual drug development programs. First, during a lead optimization campaign for the SARS-CoV-2 main protease, a retrospective study was conducted to compare calculated RBFE values against in-house experimental data for a series of compounds originating from a DNA-encoded library (DEL) screen. The predictive accuracy was highly satisfactory, achieving an RMSE of less than 1 kcal/mol alongside a descent correlation. A subsequent prospective study successfully utilized this workflow to identify a structural modification that improved the experimental IC_{50} by 5-fold. The second application involved CACHE Challenge #7, which aimed to discover selective PGK2 inhibitors. In this campaign, the absolute binding free energy (ABFE) workflow in EasyBFE was leveraged as the final physics-based scoring layer to rank LLM-generated candidates. Over a six-week period, more than 1,200 ABFE calculations were executed, yielding a shortlist of the top 200 compounds with promising binding affinities. Notably, 16 of these candidates exhibited more favorable calculated ABFE values than the established reference binder.

Future development will focus on expanding the package's functionality to support broader force field coverage, enhanced sampling techniques, and the integration of recent AMBER advances. Looking forward, these continuous improvements will position EasyBFE as a highly reliable, physics-based scoring engine uniquely suited for integration into autonomous, agentic AI workflows in modern drug discovery.

Chapter 7

Polaris Challenge: Data-Driven Priors to Improve Docking for Pose Prediction

‡Kunyang Sun and Yingze Wang contributed equally. Yingze Wang performed the MM/GBSA calculations and the case-studies.

7.1 Introduction

The COVID-19 pandemic spurred an unprecedented global research effort aimed at discovering antiviral therapeutics. [325–328] Among the most intensively studied coronavirus targets is the main protease (Mpro), [329–332] a cysteine protease responsible for cleaving viral polyproteins into functional enzymes. Because Mpro is essential for viral replication and lacks close human homologs, it is an attractive drug target[329] that stimulates multiple drug discovery projects, [333–347] with some eventually producing clinically approved inhibitors such as Nirmatrelvir (Paxlovid) [333] and Ensitrelvir. [334]

To accelerate antiviral discovery for SARS-CoV-2 Mpro, [330, 348–352] a number of open-science initiatives[327, 349] produced rich experimental datasets. The COVID Moonshot consortium, [327, 349] for example, reported more than two thousand biochemical measurements and over eight hundred crystal structures for inhibitor complexes. Numerous Antiviral Drug Discovery (AViDD) centers[88] were established in late 2021 to rapidly expand anti-virals for a range of viruses and targets that were often structurally enabled. These resources laid the groundwork for the Antiviral Drug Discovery Challenge[22] organized through the Polaris Hub[353] in collaboration with ASAP Discovery[354] and OpenADMET[355]. Specifically, in

[1]Reproduced with permission from: Polaris Challenge: Data-Driven Priors to Improve Docking for Pose Prediction. K. Sun‡, Y. Wang‡, J. Purnomo, and T. Head-Gordon. *J. Chem. Inf. Model.* 2025, 65, 12058–12067.

the ligand pose prediction track, participants were provided with a chronologically curated training set of approximately 770 SARS-CoV-2 Mpro crystal structures and tasked with predicting the crystal binding poses of 98 new SARS-CoV-2 Mpro ligands and 97 MERS-CoV Mpro ligands. This prospective design, training on data available up to a fixed time point and testing on subsequent ligands, mimics real-world hit-to-lead campaigns and provides a stringent assessment of a given methods' generalizability. During evaluations, predictions were scored using the success rate (fraction of poses within 2Åheavy-atom RMSD) across all test ligands. [22]

Ligand pose prediction is inherently a two-component problem: candidate conformations must be generated and then evaluated using scoring functions to identify the most probable binding modes. [2] Traditional docking software follows this paradigm, and significant advances in both sampling strategies and scoring methodologies have led to improved screening efficiency and greater scoring accuracy. [356–360] In particular, when a large amount of existing crystal structures are available, ligand-template-based docking [358, 361–363] can generate accurate results without the use of sophisticated algorithms. In parallel, thanks to the curation of high quality protein ligand structure datasets, [364–367] recent developments in machine learning docking methods, [368] particularly protein–ligand co-folding models, [11, 12, 369] have shown promise in predicting binding poses within a single inference step. Nevertheless, there remains considerable value in benchmarking and refining classical docking approaches, both to establish robust baselines and to locate the performance boundaries of traditional methods. In this work, we intentionally forgo deep learning models and instead investigate to what extent widely adopted tools like AutoDock Vina [6] can be enhanced through fragment-guided sampling and scoring. By treating the Polaris challenge as a prospective benchmark for classical docking, we aim to systematically characterize the conditions under which traditional methods succeed, where they fail, and how they can be effectively incorporated into a typical drug discovery workflow.

Here, we present a docking and rescoring workflow developed for the challenge. The process begins with ensemble docking using GPU accelerated Vina (Vina-GPU)[370] to generate initial poses, followed by a scaffold-aware ligand fragmentation step to distinguish test ligands with training-set analogs from those without. For ligands with similar fragments, we leverage fragment alignment from related training complexes to guide pose generation similar to ligand template-based docking and apply a custom fragment-based scoring function for rescoring and selection. Ligands lacking such similarity, or failing to meet fragment-scoring thresholds, are advanced instead through a physics-based MM/GBSA refinement of generated poses. By combining fragment-informed priors with physics-based optimization, this workflow is designed to exploit the complementary strengths of data-driven and structure-based approaches to predict ligand poses of target-specific proteins.

7.2 Methods

As depicted in Figure 7.1, our ligand pose prediction strategy follows a multi-step, two-branch workflow that integrates ensemble docking, fragment-based structural similarity detection, and physics-based MM/GBSA rescoring. For each of the two task targets, we first generated a diverse ensemble of protein conformations by combining clustered Monte Carlo Side Chain Ensemble (MCSCE) [371] conformers with clustered crystal complexes from either the organizers or the RCSB PDB. [372] As part of the standard drug discovery practice, [373] all ligands were then docked to these ensembles to produce a pool of candidate poses.

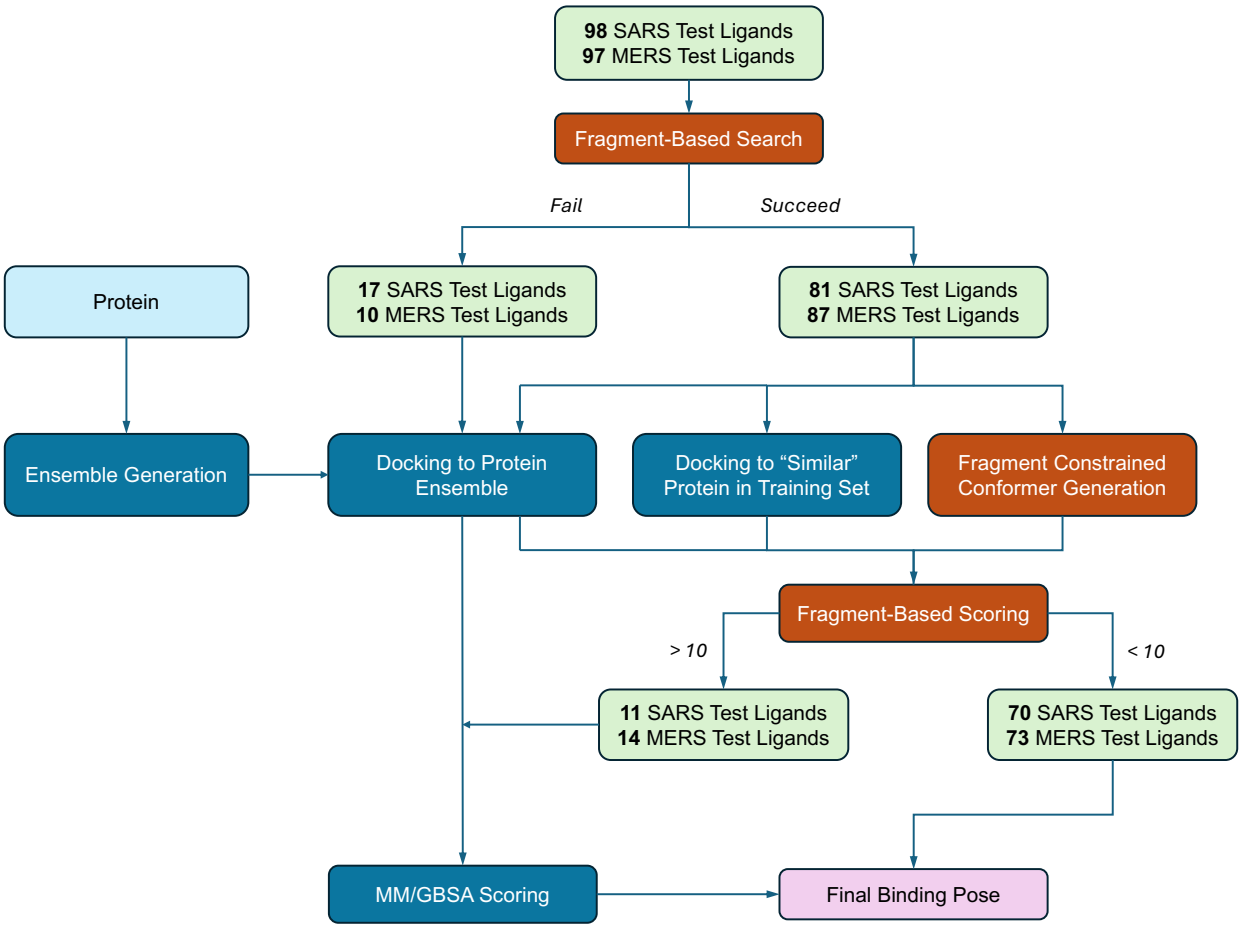

Figure 7.1: *Schematic representation of our multi-step workflow to generate the final ligand binding poses.* The fragment based search utilizes conserved chemical fragments of training ligands to help inform a set of structural and scoring priors for pose of the test ligand. Lack of train-test fragment similarity generates a ligand pose based on standard docking and free energy scoring approaches.

Ligands sharing conserved chemical fragments with those in the training set were routed through a fragment-guided branch, where additional poses were generated by fragment-constrained conformer sampling and docking into structurally similar training-set proteins. These poses were ranked using a custom fragment-based score that quantified structural agreement with known fragment conformations. Ligands without shared fragments in the training set, or whose fragment-based scores failed to meet a quality threshold, were instead processed through the alternative refinement branch, where MM/GBSA calculations were applied to the ensemble-docked poses to select the most favorable binding mode. This workflow leveraged available training data while applying physically rigorous refinement in data-poor regimes, ensuring a balance between data-driven and physics-driven guidance.

7.2.1 Protein Ensemble Generation and Docking

Since previous studies have shown that the backbone of SARS-CoV-2 Mpro and MERS-CoV Mpro remains largely rigid upon ligand binding, [22, 347] we focused on sampling diverse side-chain conformations while keeping the backbone fixed. To this end, we used the Monte Carlo Side Chain Ensemble (MCSCE) method, [371] which proposes rotamers from backbone-dependent libraries, incrementally builds all-atom side chains with hydrogens, and accepts or rejects configurations based on force-field energies. This procedure efficiently generates physically reasonable ensembles with fixed backbone atom positions. The resulting side chain ensembles capture side-chain rearrangements that are important for induced fit, while not requiring running more expensive molecular dynamics simulations for longer timescales. The MCSCE method has been validated extensively against X-ray crystallography and NMR J-coupling data.[371]

For each reference protein, we generated 2,000 MCSCE conformers and clustered them by side chain atom features over binding site residues identified from literature. [374–377] Specifically, since all MCSCE conformers share an identical backbone, we represented each conformer solely by the 3D coordinates of its binding site side chain atoms. These coordinates were concatenated into a feature vector for each structure, which were then clustered using DBSCAN. [378] To further expand diversity, we incorporated 770 SARS-CoV-2 Mpro structures from the challenge training set and 44 MERS-CoV Mpro structures curated from the RCSB PDB. These conformers were first aligned and then clustered on their side chain atom positions of binding site residues following the same procedure. The final protein ensemble consisted of nine representative SARS-CoV-2 Mpro structures (four from MCSCE and five from the training crystal structure set) and four representative MERS-CoV Mpro structures (one from MCSCE and three from the PDB crystal structures). All representative structures were aligned to their respective reference proteins using Cα atoms and served as docking targets for the test ligands, using Vina-GPU for pose generation. Details about the docking box information are included in the Supplementary Information. For each docking run, we retained the top four poses for SARS-CoV-2 and the top nine for MERS-CoV,

ensuring a comparable total number of candidate poses between the two targets. Each ligand docking run on average takes around one second to finish on a single A-40 GPU.

7.2.2 Fragment Identification and Pose Generation

To better leverage structural information from the provided training set, we incorporated a fragment-based search strategy alongside conventional docking. This approach is motivated by the observation that many test ligands share conserved chemical substructures, including core scaffolds or functional moieties, with ligands in the training set. Since these fragments often mediate key interactions with pocket residues, recognizing them enables us to bias pose generation toward geometries consistent with experimentally validated binding modes, which will improve the accuracy of binding pose predictions beyond docking alone.

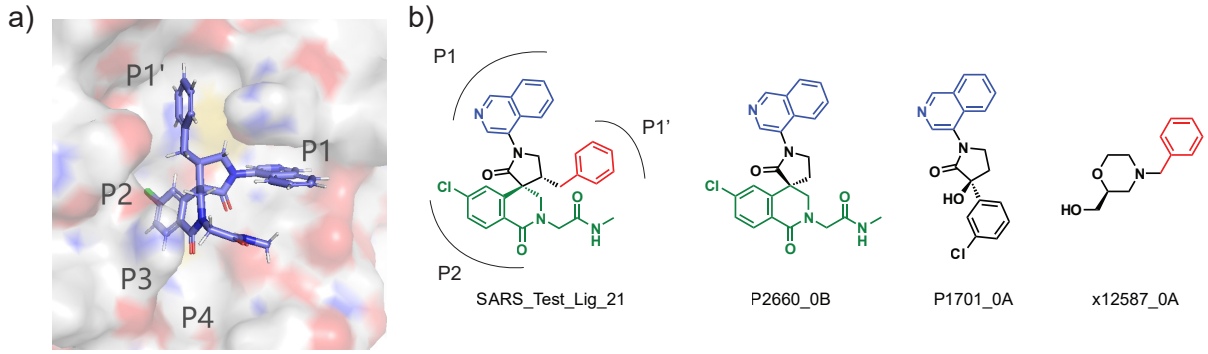

Figure 7.2: *Example of scaffold-aware ligand fragmentation.* (a) `SARS_Test_Lig_21` crystal structure with binding pocket labels. (b) 3 example ligands with identical colored fragments as the test ligand `SARS_Test_Lig_21` from the provided crystal structures.

To create the fragments, we implemented a scaffold-aware ligand fragmentation protocol using the training set ligands as references. The protocol begins by identifying all ring systems in the molecule using RDKit,[320] where rings sharing two or more atoms are treated as a single fused system. Each fragment is then defined by the ring system plus its attached atoms, extending outwards until one bond away from any other ring system. As illustrated in Figure 7.2, the target ligand is partitioned into four fragments denoted by different colors. By design, this protocol keeps individual ring systems intact, preserving the local chemical context essential for protein interactions and effectively accommodating the trefoil-like structures of SARS-CoV-2 Mpro binders. Following this protocol, we identified 81 of 98 SARS-CoV-2 test ligands that shared at least one identical fragment with a training-set ligand. Given the high conservation and structural similarity between the SARS-CoV-2 Mpro and MERS-CoV Mpro binding pockets, [377] MERS-CoV test ligands were also compared against the SARS-CoV-2 training set, yielding 87 of 97 ligands eligible for downstream fragment-based processing.

For ligands with fragment matches, we sampled additional poses beyond the ensemble docking results using two strategies: (1) re-docking the test ligand into every training protein whose ligand shared a similar fragment, following the same Vina-GPU protocol, and (2) generating 10 conformers per protein in RDKit [320] with the matched fragment fixed at its crystallographic position while allowing the remainder of the ligand to flexibly fit. This procedure yielded up to ∼300 poses for high-similarity ligands, compared with 36 poses from ensemble docking for ligands without fragment matches. Owing to the high binding mode homology between SARS-CoV-2 and MERS-CoV as shown in Supplementary Figure S1, the same workflow was applied cross-target for MERS-CoV ligands using SARS-CoV-2 proteins, further increasing their number of candidate poses despite having no MERS-CoV Mpro structures in the training set. The expanded sampling framework remains computationally efficient. On average, docking against 8 training proteins takes approximately 8 seconds per ligand on a single A40 GPU, while the RDKit conformer generation step took less than one second on a CPU.

7.2.3 Custom Fragment-Based Scoring

Traditional docking scores are largely empirical, parameterized to perform reasonably well across a wide range of targets and ligand chemotypes.[2] While effective for general-purpose virtual screening, they are not necessarily optimal for specialized tasks, as will be shown in the next section. Moreover, docking scores are calculated independently for each candidate pose and cannot directly incorporate prior knowledge from a training dataset. These limitations motivated us to develop a new pose-ranking metric that explicitly leverages information from the provided experimentally resolved protein-ligand complexes in the training set.

In our formulation, candidate poses with fragment matches were ranked using a geometry-based metric defined as

$$S_{\text{frag}} = \sum_{\text{frag}} \text{RMSD}_{\text{frag}} \times \frac{N_{\text{frag}}}{N_{\text{ligand}}}, \tag{6.1}$$

where RMSDfrag is the heavy-atom RMSD between the fragment in the candidate pose and the corresponding fragment from the training set, N_{frag} is the number of heavy atoms in the matched fragment, and N_{ligand} is the total number of heavy atoms in the ligand. The weighting factor $N_{\text{frag}}/N_{\text{ligand}}$ ensures that larger fragments, which capture more of the ligand geometry, contribute proportionally more to the final score than smaller ones. Summing over all matched fragments yields a single measure of structural agreement between a candidate pose and experimentally validated binding conformations. Herein, we refer to this metric as the Custom Score.

The underlying intuition behind the design of the Custom Score parallels templated ligand docking, where chemically similar ligands are expected to adopt comparable binding geometries to preserve favorable protein–ligand interactions, particularly when the protein

structure is unchanged. Lower scores indicate closer alignment of candidate fragments with training crystal structures, reflecting higher structural fidelity and a greater likelihood of predicting correct binding poses. In contrast, higher scores signify larger geometric deviations and reduced retention of key interactions. To calibrate a score cutoff for filtering unreliable matches, we created a validation set by holding out 70 random ligands from the training data and used the other 700 ligands to build the fragment library. Our analysis of the hold-out set showed a linear correlation between our Custom Score and ligand RMSD, where a score of 10 corresponds to an RMSD of 2Å. Therefore, for ligands with at least one pose below the custom score threshold of 10, we chose its lowest-scoring pose as the prediction.

7.2.4 Fallback MM/GBSA Scoring

For ligands lacking structurally similar fragments in the training set, or when the fragment-based score is too high (exceeding 10) to be reliable, we employed MM/GBSA, a more rigorous physics-based approach, to rank the poses. We chose to apply MM/GBSA selectively because our analysis of the 70-ligand validation set showed that MM/GBSA rescoring actually worsened the average RMSD for poses already considered reliable by our Custom Score.

To perform MM/GBSA rescoring, each candidate pose first underwent a geometry optimization in complex with the protein, in which the positions of all protein heavy atoms were constrained. This optimization served to relieve steric clashes while preserving the overall binding geometry, a procedure shown in prior studies to improve agreement with experimental binding affinities. [23] Then, the GBSA score was calculated using `MMPBSA.py`[93] program in AmberTools with GB-OBC[379] implicit solvation model ($igb = 2$). Amber ff14SB force field[380] was used for protein and GAFF2[312] along with AM1-BCC charges[314, 315] was used for the ligand. The constrained geometry optimization was performed using OpenMM[381] by assigning zero mass to frozen atoms. The pose with the most favorable MM/GBSA score (i.e., the lowest predicted binding free energy) was selected as the final model, providing a robust, physics-grounded fallback when similarity-based guidance proved insufficient.

7.3 Results

In this section, we present a retrospective analysis of binding poses extracted from different stages of our workflow, evaluating how effective each stage would be if considered in isolation. This stepwise assessment allows us to identify the contributions of prior knowledge from training data and physics-based refinement to overall prediction accuracy. By examining the incremental improvements across stages, we highlight why the combined workflow leads to more reliable ligand pose prediction compared to a baseline docking protocol.

7.3.1 Overall Performance

As highlighted in Supplementary Figure S2, our workflow ranked 6th overall in the contest, exceeding the performance of purely ligand template-based methods but trailing behind methods that fine-tune cofolding models. To systematically evaluate the components of our workflow shown in Figure 7.1), we categorized our contest submissions for the SARS-CoV-2 and MERS-CoV Mpro targets into three groups based on their origin and filtering outcome. The largest group, Similar-Passed, consisted of ligands identified through similarity searches that successfully passed our Custom Score threshold. The second group, Similar-Rejected, included ligands that were initially identified by similarity but were subsequently filtered out due to poor performances in Custom Scores. The final category, Dissimilar, comprised ligands that lacked known structural analogues and were processed using ensemble-based solely physics methods. This tiered framework allows for a direct assessment of how each stage of our methodology contributed to the final prediction quality.

The results shown in Figure 7.3 and Supplementary Table S1 demonstrate a clear performance hierarchy across these categories, validating our prioritization strategy. The Similar-Passed group achieved the highest success rates, with 51 out of 70 SARS-CoV-2 predictions and 44 out of 73 MERS-CoV predictions having an RMSD < 2 Å. In contrast, the accuracy was substantially lower for the Similar-Rejected group (4/11 for SARS-CoV-2 and 4/14 for MERS-CoV). The Dissimilar group, representing the lowest-confidence predictions, had the lowest success rates, with 3/17 for SARS-CoV-2 and 3/10 for MERS-CoV. Overall, our workflow yielded greater accuracy for the SARS-CoV-2 target (59.00 ± 5.03 %) compared to the MERS-CoV target (52.49 ± 5.11 %), as determined by bootstrapping with 1000 samples. [22] We attribute this difference to the more extensive training data available for SARS-CoV-2. The consistently better performance of the Similar-Passed category across both targets underscores the efficacy of using similarity searches to leverage existing structural knowledge. Collectively, these results establish that our combined approach of similarity-based selection and rigorous scoring provides a robust method for distinguishing high-confidence predictions from less reliable ones, significantly improving pose prediction accuracy.

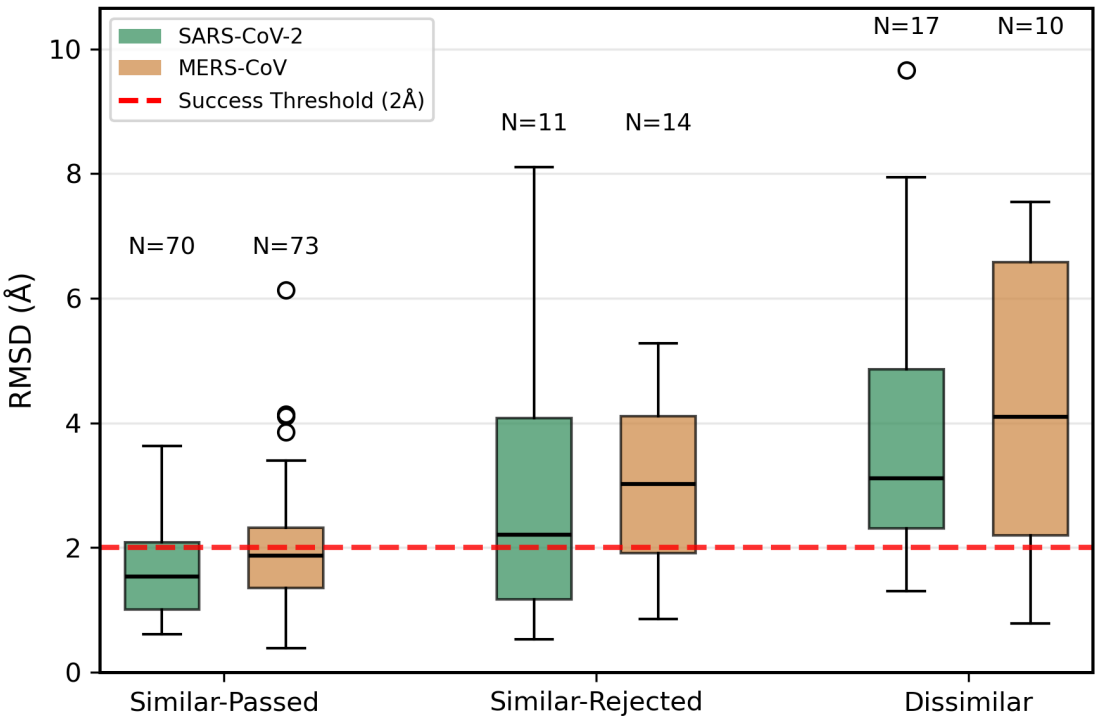

Figure 7.3: *Box plots of RMSD distribution for ligand pose prediction filtered into different categories of similarity to the training set for SARS-CoV-2 and MERS-CoV Mpro targets.* Similar-Passed consists of ligands that pass fragment similarity searches and custom score thresholds. Similar-Rejected consists of ligands that pass fragment similarity searches but fail on the custom score threshold. Dissimilar requires a pose prediction using standard docking and MM/GBSA scoring.

7.3.2 Performance of Similar Test Ligands

Next we consider the test ligands identified by our scaffold-aware fragmentation method with at least one fragment identical to that in the training set. This analysis allowed us to systematically evaluate six distinct approaches: Vina Ensemble (ensemble docking with Vina Score), Vina Similar (docking to the most similar training protein with Vina Score), Custom Ensemble (ensemble docking with Custom Score), Custom Similar (similar protein docking with Custom Score), Custom Frag (fragment-constrained generation with Custom Score), and our baseline Native (retrospective docking to the test protein structure with Vina Score).

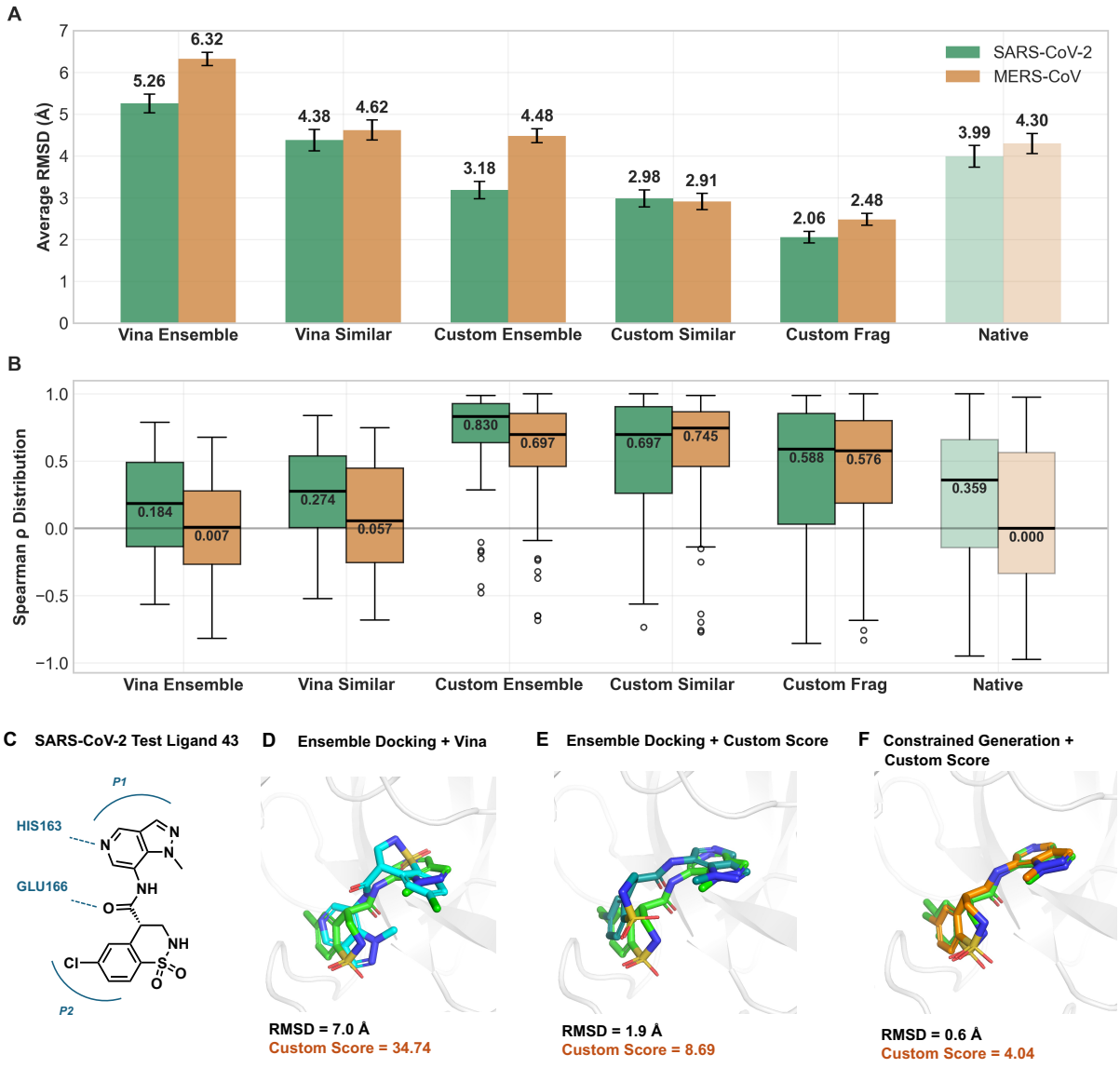

Figure 7.4: *Comparison of pose prediction accuracy for ligands containing identical training fragments using different sampling and scoring approaches.* (A) the average and standard deviation of best pose RMSD values determined by each method. (B) the distribution of score–RMSD correlation. For each testing ligand, top 10 poses selected by each method were used to calculate the Spearman R between score and RMSD. A total of N=81 ligands for SARS-CoV-2 and N=87 for MERS-CoV. (C) 2D ligand structure and binding mode of SARS-CoV-2 test ligand 43. Crystal ligand poses (green) aligned with best poses from ensemble docking ranked by (D) Vina Score (cyan), (E) Custom Score (teal), and (F) best pose from fragment-constrained conformer generation ranked by Custom Score (orange). Further details are provided in Supplementary Figure S3.

As shown in Figure 7.4A, ensemble docking with a Vina Score yielded inferior performance compared to Native docking for best pose selection. Docking to similar training proteins rather than pre-generated ensembles significantly improved RMSD values, although it still falls short of the Native baseline. Hence sampling of similar protein structures and ligands provides clear advantages over ensemble-based approaches, effectively better capturing the conformational specificity required for accurate pose prediction. As seen in Figure 7.4B, even Vina Score and using native testing proteins yielded only weak correlations between Vina Score and RMSD, indicating that the Vina Score is not well suited for reliably identifying the best poses from a large candidate pool. To address this, we applied our fragment-based Custom Score to rescore poses from both Ensemble and Similar docking. As shown in Figure 7.4, rescoring with the Custom Score consistently reduced RMSD values and improved score–RMSD correlations, confirming that fragment-derived information from the training set was successfully incorporated.

The example of test ligand 43 bound to SARS-Cov-2 MPro (Fig. 7.4C) further highlights this point: while the best Vina-scored pose was entirely misoriented, rescoring with the Custom Score identified a near-native conformation, validating the ranking power of fragment-based scoring (Figure 7.4D,E). Beyond rescoring, we also tested fragment-constrained conformer generation combined with the Custom Score. This approach achieved the best RMSD values overall. As illustrated for test ligand 43, fragment-constrained generation produced a pose within 0.6 ÅRMSD of the crystal structure, faithfully recovering all key interactions, including the pyrazolopyridine-HIS163 hydrogen bond, the chlorobenzene hydrophobic contacts, and the linker amide-GLU166 hydrogen bond that was absent in the rescored pose (Fig. 7.4F). The benefit of this approach is particularly evident for flexible ligands, where docking alone often fails to sample correct conformations. By constraining conformer generation with fragment priors, the sampling space is reduced and the likelihood of recovering the correct pose increases.

Finally, the same performance hierarchy was observed for MERS-CoV ligands, despite all upsampling and scoring being derived from SARS-CoV-2 training fragments and protein structures. Although MERS-CoV results were generally lower in absolute terms, the relative improvement trends remained consistent, with Custom Frag achieving the best RMSD, Custom Score improving upon Vina Score, and Similar docking outperforming Ensemble docking. This transferability provides strong evidence that fragment-based approaches can leverage conserved binding site architectures to improve pose prediction across related coronavirus proteases.

7.3.3 Performance of MM/GBSA Rescoring

Our final analysis focuses on ligands in the low-data or low-confidence regime, which represent the most challenging prediction scenarios. These were categorized as either Similar-Rejected (ligands with fragment matches that failed our Custom Score threshold) or Dissimilar (lig-

ands lacking any matching fragments in the training set). For these cases, we applied MM/GBSA rescoring as a physics-based refinement stage to assess whether more rigorous energy calculations could improve pose selection.

As shown in Figure 7.5, MM/GBSA rescoring led to modest improvements across these difficult cases. It slightly reduced average RMSDs and yielded small but consistent gains in Spearman R for the best poses sampled from different methods. The refinement is best illustrated by SARS-CoV-2 test ligand 77 (Figure 7.5D). In this case, the top Vina Ensemble pose in Figure 7.5E was poorly aligned (RMSD $>$ 6 Å), deviating from the ground truth ligand pose. With the help of our fragment-based Custom Score, the γ-lactam scaffold was placed to form desirable interaction with GLU166 and reduced the RMSD to 3.3 Åin 7.5F. Furthermore, MM/GBSA helped the ligand recovered all key moieties except the solvent-exposed morpholine, yield a pose with 2.2 Å RMSD to the crystal structure as demonstrated in 7.5G. Overall, MM/GBSA functions as a useful optimizer that can refine poses and improve rankings to some degree, but it does not fundamentally change the outcome for cases where the initial sampling fails. Its value therefore depends critically on the quality of the starting ensemble generated by fragment-guided sampling.

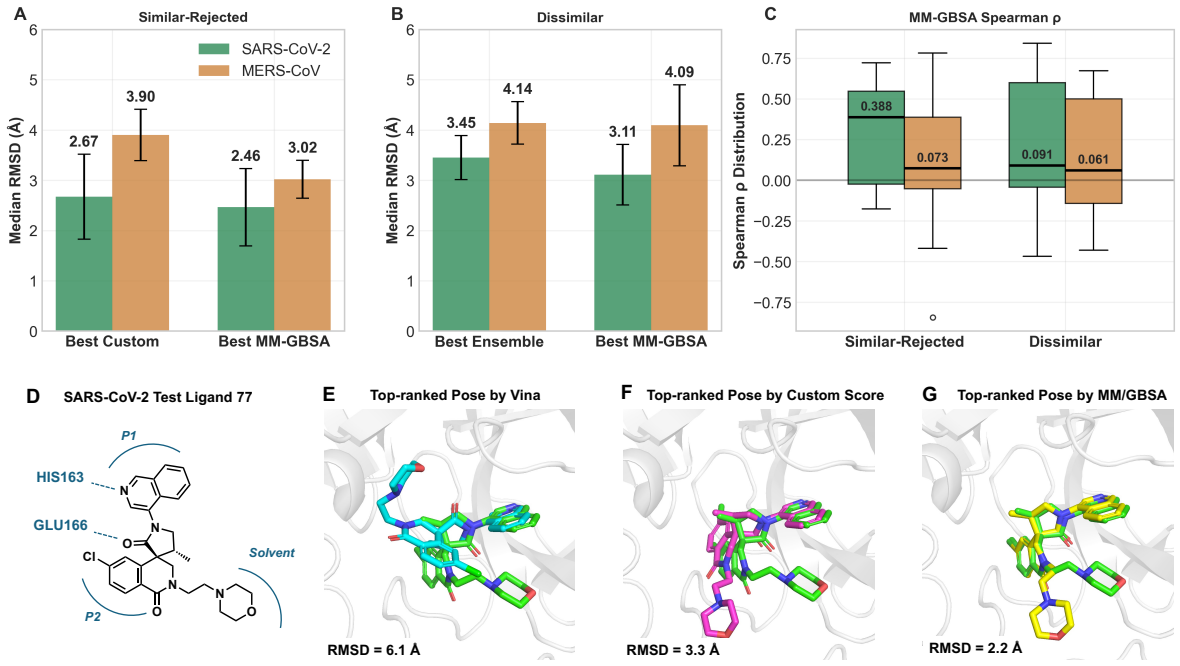

Figure 7.5: *Performance of MM/GBSA rescoring on challenging ligands without similar training fragments.* MM/GBSA provides modest improvements in (A,B) RMSD and (C) Spearman R, highlighting its role as a refinement step rather than a decisive predictor. (D) Illustration of MM/GBSA rescoring on SARS-CoV-2 test ligand 77. Best pose selected by (E) Vina score (cyan), (F) fragment-based score (magenta), and (G) MM/GBSA score (yellow), superposed with the crystal structure (green). More details are provided in Supplementary Figure S4.

To investigate the practical utility of MM/GBSA further, we performed a retrospective analysis by applying MM/GBSA rescoring to the high-confidence Similar-Passed category. As shown in Supplementary Table S2, for SARS-CoV-2, where the fragment priors were derived from abundant homologous crystal structures, MM/GBSA rescoring worsened the average RMSD of these already accurate poses. For MERS-CoV, for which no specific training data was available, MM/GBSA performed only slightly better in pose accuracy compared to our data-driven prior approach. This observation suggests that fragment-based scoring is superior when extensive, homologous training data exists, and furthermore performs comparable to a physics-based approach like MM/GBSA for related but a less-characterized target as in this case for MERS-CoV. Given that MM/GBSA is far more expensive, and performs less well on SARS-CoV-2, this analysis further supports that fragment-guided sampling and scoring offers and excellent tradeoff in computational efficiency and accuracy.

7.4 Discussion and Conclusion

The ligand pose prediction problem is inherently two-step: candidate poses must first be sampled, and then accurately scored and ranked to obtain the best pose. In this work, we present a docking-based workflow that achieves over 50% success in predicting ligand poses within 2 Å RMSD for both SARS-CoV-2 Mpro and MERS-CoV Mpro. While Vina docking and scoring alone performed poorly, their combination with fragment-derived priors and robust rescoring proved effective, highlighting the importance of incorporating prior knowledge to guide the selection stage.

Within this framework, fragment-informed scoring and fragment-constrained sampling provided the largest performance gains, steering pose selection toward geometries consistent with experimental observations. Instead of focusing on maximum common substructure in template-based docking, our fragment-constrained sampling method yields a more diverse set of candidate pose while retaining key protein-ligand interactions, leading to more effective prediction when coupled with a scoring function. Interestingly, even though our fragment information was derived exclusively from SARS-CoV-2, it transferred effectively to MERS-CoV, suggesting that conserved pockets across related proteins can provide valuable priors for pose prediction. While for ligands without suitable fragment matches, the more expensive MM/GBSA rescoring served as a moderately effective fallback, often refining initially suboptimal poses into more native-like conformations, although its value depends critically on the quality of the starting ensemble. As shown in our retrospective analysis, the physics-based refinement worsened the results for SARS-CoV-2 and only provided small improvements for MERS-CoV from the fragment transfers. Together, these findings emphasize that physics-based docking engines can generate sufficient diversity at the sampling stage, but their success depends on tailored scoring strategies that integrate domain knowledge.

We also observed an orthogonality between sampling and scoring. The score–RMSD correlation was strongest when sampling with Vina and scoring with the fragment-inspired Custom Score, even though fragment-constrained generation achieved lower RMSD values overall. By contrast, combining fragment-driven sampling with fragment-informed scoring yielded weaker correlations. This suggests that mixing sampling and scoring strategies, rather than tightly coupling them, may improve discrimination among candidate poses. An interesting direction for future study would be to score fragment-constrained poses with Vina or apply Custom Score to constrained docking results to further differentiate pose quality. These findings highlight not only the utility of fragments in scoring but also the broader advantages of docking-based sampling for generating diverse candidate ensembles.

Compared with end-to-end co-folding models, physics-based docking retains additional advantages. AutoDock Vina generated a wide range of plausible poses without any learned scoring augmentation, preserving structural diversity that may be lost in co-folding models [382]. For new proteins not represented in the co-folding training data, or for proteins with multiple binding sites, this pose diversity can benefit downstream rescoring, hybrid methods,

or expert evaluation.

These findings suggest that physics-based docking remains a valuable tool for ligand pose prediction, particularly for targets with limited training data, multiple pockets, or less-conserved binding environments. By integrating fragment-derived priors when available, and falling back on physics-based rescoring when not, such workflows can balance diversity in sampling with selectivity in pose ranking. Moving forward, hybrid approaches that combine the sampling power of docking with the discriminatory strength of learned or knowledge-based scoring functions may offer a robust and generalizable route to high-accuracy pose prediction.

7.5 Supporting Information

7.5.1 Additional Methodology Details

Vina Config for Docking

For SARS-CoV-2 Mpro, the docking box was defined to have a center at $(-8.46,\ -1.10,\ -20.79)$ with a box size of 24 Å in each dimension. For MERS-CoV Mpro, the docking box was defined to have a center at $(8.09,\ -0.03,\ 21.95)$ with a box size of 24 Å in each dimension.

Key Protein Residues

For protein alignment, we used the C α atoms of residues known to be critical for ligand binding. In SARS-CoV-2 Mpro, the alignment was based on residues 41, 140–145, 163–166, 172, and 187–189. In MERS-CoV Mpro, the corresponding set included residues 41, 143–148, 166–169, 175, and 190–192.

PDB IDs for MERS-CoV Mpro

For clustering of MERS-CoV Mpro crystal structures, we used the following 44 PDB entries: 4YLU, 5WKJ, 5WKK, 5WKL, 5WKM, 6VGY, 6VGZ, 6VH0, 6VH1, 6VH2, 6VH3, 7ENE, 7T3Y, 7T3Z, 7T40, 7T41, 7TQ7, 7TQ8, 7VTC, 7WQJ, 7XRY, 8CZT, 8CZU, 8CZV, 8DGY, 8E6B, 8E6C, 8E6D, 8E6E, 8HUT, 8IG6, 8J34, 8YKL, 9ATA, 9ATD, 9ATE, 9ATDF, 9ATG, 9ATH, 9ATI, 9ATJ, 9ATS, 9ATT, 9BOO.

7.5.2 Supporting Figures

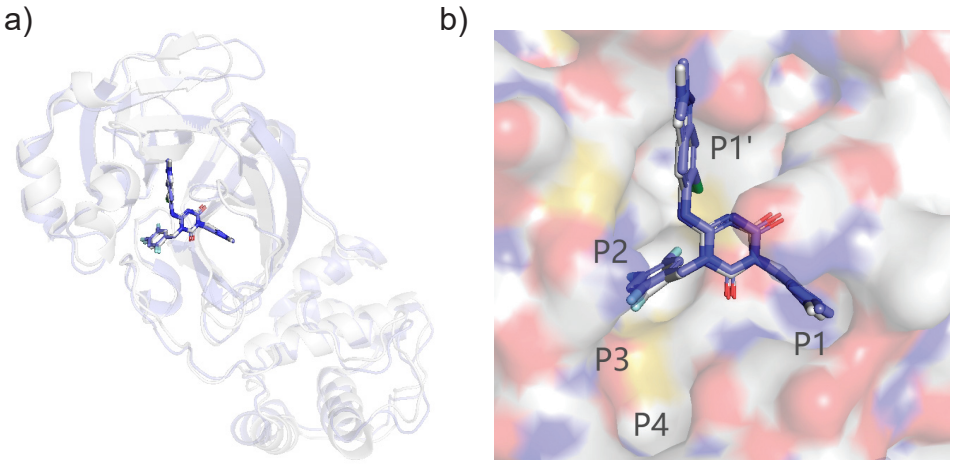

Figure S7.1: *Comparison of Ensitrelvir binding modes across SARS-CoV-2 and MERS-CoV Main proteases.* (a). Structure overlay between SARS-CoV-2 Mpro (PDB: 7VU6 [334]) and MERS-CoV Mpro (PDB: 8VUT[352]) in complex with Ensitrelvir. (b) Close-up visualization showing the highly homologous binding modes of Ensitrelvir across two proteins.

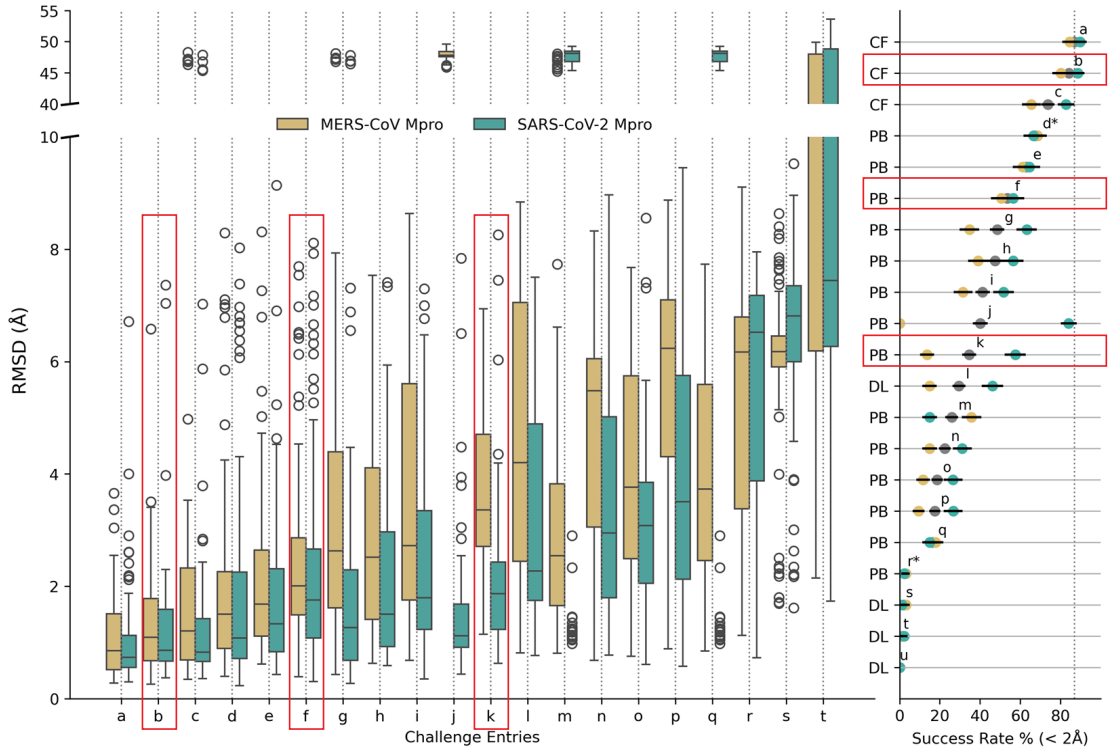

Figure S7.2: *Ligand pose challenge results from MacDermott-Opeskin et al. with boxes around key methods to highlight.[22]* Method b is a fine-tuned version of Boltz[12] using the provided training data; Method f is our submission; and Method k is a template-based docking method. Here, CF refers to co-folding models; PB refers to physics-based methods; and DL represents deep learning methods. More details about each individual submission can be found in the challenge summary paper. [22]

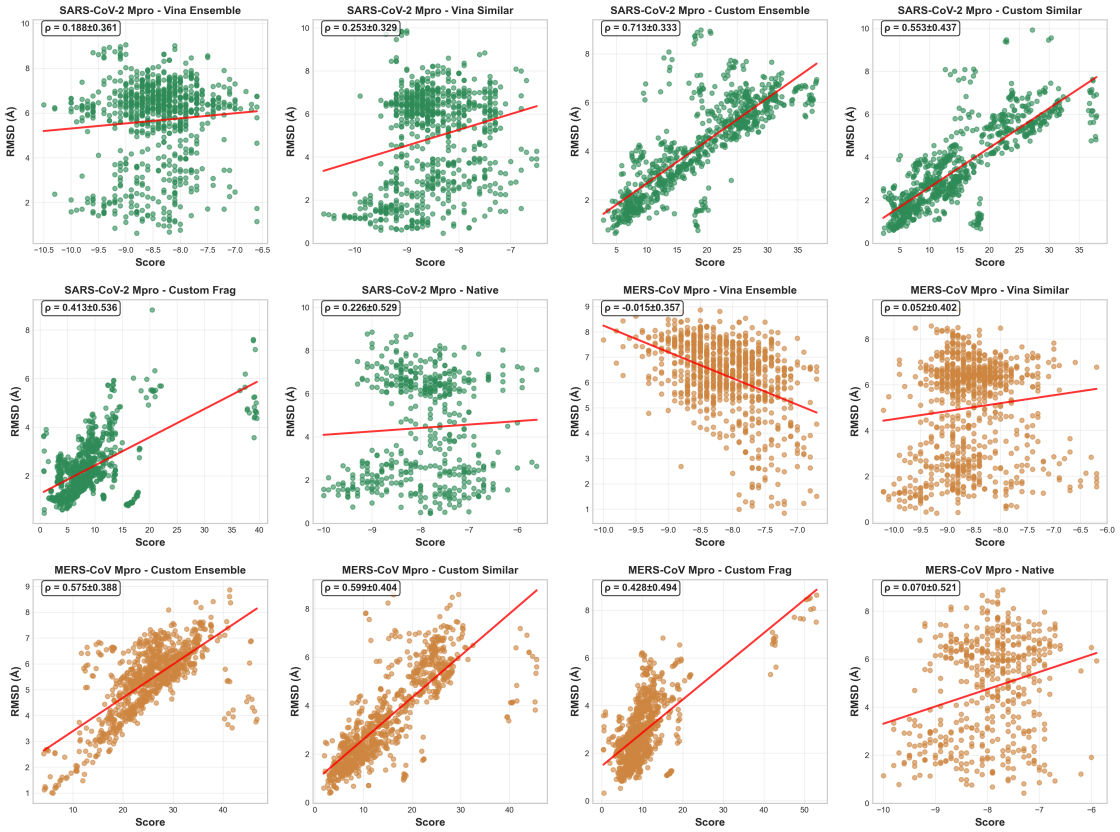

Figure S7.3: *Scatter plot of various scores versus RMSD between predicted and crystal poses.* For each target protein group, the top 10 poses ranked by each score along with their final RMSD are shown. Green points represent all 81 similar testing ligands for SARS-CoV-2 Mpro, and orange points represent all 87 similar testing ligands for MERS-CoV Mpro. The meaning of the ligand scoring category can be found in Main.

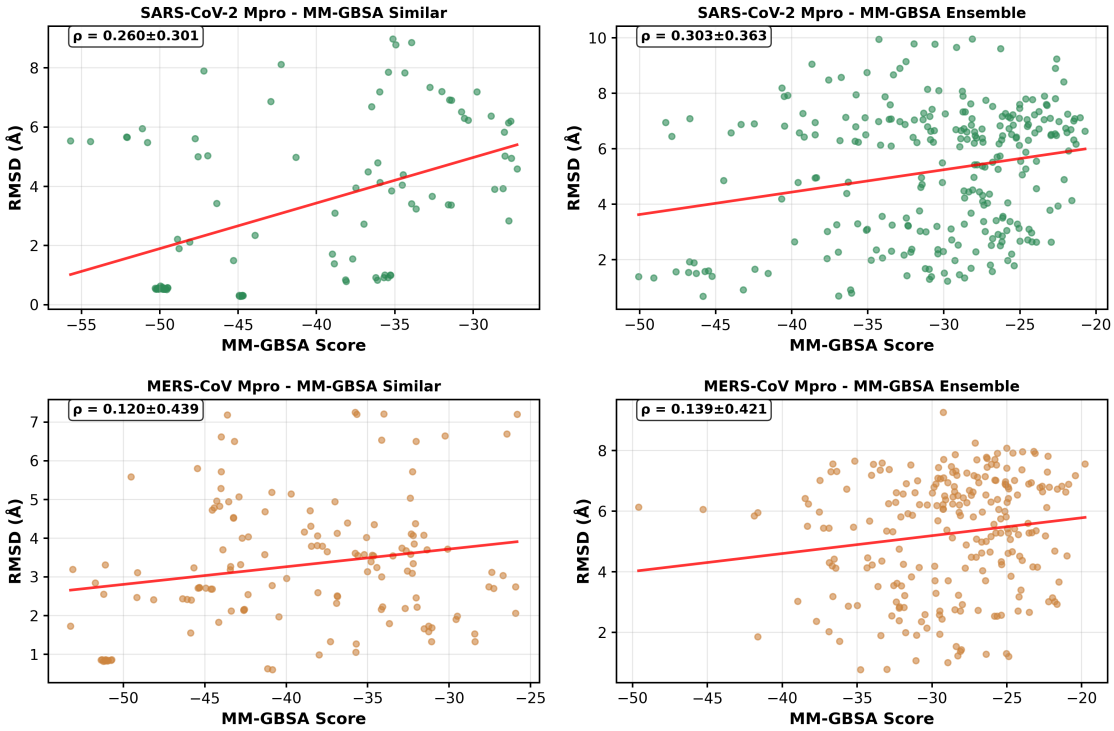

Figure S7.4: *Scatter plot of MM/GBSA scores versus RMSD between final predicted and crystal poses.* The top ten poses ranked by MM/GBSA energies are shown for each target protein group: (A) 14 MERS-CoV test ligands with identical fragments but rejected in earlier fragment-score filtering; (B) 11 SARS-CoV-2 test ligands with identical fragments but rejected; (C) 10 MERS-CoV ligands without identical fragments as training ligands; and (D) 17 SARS-CoV-2 ligands without identical fragments.

Supporting Tables

Table S7.1: Summary of contest submission results by target and category. Values indicate the number of entries (N), the count of predictions with accuracy $< 2\,\text{Å}$, and the corresponding percentage. Categories are described in main text.

Category	**SARS-CoV-2**			**MERS-CoV**		
	N	Count $< 2\,\text{Å}$	% $< 2\,\text{Å}$	N	Count $< 2\,\text{Å}$	% $< 2\,\text{Å}$
Similar–Passed	70	51	72.9%	73	44	60.3%
Similar–Rejected	11	4	36.4%	14	4	28.6%
Dissimilar	17	3	17.6%	10	3	30.0%
Overall	98	58	59.2%	97	51	52.6%

Table S7.2: Effect of MM/GBSA rescoring for the Similar-Passed ligand category.

Target	**Condition**	**Count**	**RMSD (Å)**		**<2Å**	
			Avg.	Std.	Count	%
SARS-CoV-2	Without MM/GBSA	70	1.602	0.690	51	72.9
	With MM/GBSA	70	1.774	0.860	44	62.9
MERS-CoV	Without MM/GBSA	73	1.939	0.941	44	60.3
	With MM/GBSA	73	1.876	0.977	46	63.0

Chapter 8

Conclusion

The discovery of novel therapeutics is a complex endeavor that increasingly relies on computational structure-based drug discovery (SBDD) to prioritize promising molecular candidates. However, the efficacy of SBDD pipelines has historically been constrained by the quality of foundational data, the accuracy of physics-based model that describes molecular interactions, and the practical accessibility of rigorous free-energy calculation workflows. This dissertation addresses these critical bottlenecks by presenting new methodologies and open-source tools spanning data curation, force-field development, and automated binding-affinity predictions.

First, this dissertation addressed the critical data quality problem that frequently compromises structure-based scoring models. By introducing HiQBind, together with contributions to LP-PDBBind, my work proposed a framework to curate high-fidelity and unbiased protein-ligand complexes datasets by fixing common structural artifacts in open-source dataset while systematically eliminating data leakage. More importantly, the resulting workflow establishes a scalable and sustainable continuous data infrastructure, shifting the paradigm of dataset construction from labor-intensive manual curation to fully programmatic, automated pipelines.

Second, to bridge the gap between computational efficiency and quantum-mechanical accuracy, this dissertation advanced the development of polarizable force fields through a coupled approach of data generation, software infrastructure, and functional-form optimization. The creation of a protein–ligand EDA dataset provided physically meaningful reference targets for parameterizing force fields with more complex physics. TorchFF addressed the critical performance bottleneck in differentiable force field implementations by delivering optimized differentiable operators specialized for molecular mechanics, which transforms differentiable programming from a prototyping concept into a practical foundation for advanced force field validation and optimization. Furthermore, by restructuring short-range intramolecular interactions within the AMOEBA framework using an improved valence potential and a regularized exclusion scheme, this work demonstrates how a richer and well-parameterized bonded representation can replace ad hoc scaling rules and yield a more accurate potential

energy surface.

Third, this dissertation introduced EasyBFE to address the steep operational barriers that have historically hindered the real-world application of rigorous binding free-energy calculations. By providing an accessible, end-to-end automated platform, EasyBFE bridges the gap between physical rigor and practical execution, and its practical utility was validated in prospective drug discovery campaigns, including the lead optimization of SARS-CoV-2 main protease inhibitors and selective PGK2 inhibitor *de novo* design. Additionally, this research explored how a lower-cost end-state free-energy method, MM/GBSA, can be strategically combined with prior data to improve binding pose predictions for SARS-CoV-2 and MERS main protease ligands in a blinded challenge .

Looking forward, the landscape of structure-based drug discovery is rapidly shifting toward specialized AI models and autonomous AI agentic workflows that drive the design–make–test–analyze (DMTA) cycle. In this AI-driven paradigm, the foundational contributions of this dissertation become even more indispensable. This is driven by a dual need: the rising demand for high-quality data to reliably train and validate these models, and the necessity for automated, high-accuracy computational pipelines that AI agents can seamlessly integrate with. By delivering automated data curation pipelines, more accurate physics models, and automated free-energy engines, this dissertation provides the essential, high-precision building blocks required to empower future AI agents and accelerate rational drug design.

References

1. Waring, M. J. *et al.* An analysis of the attrition of drug candidates from four major pharmaceutical companies. *Nature Reviews Drug Discovery* **14,** 475–486. ISSN: 1474-1776. https://www.nature.com/articles/nrd4609 (2015).
2. Pinzi, L. & Rastelli, G. Molecular Docking: Shifting Paradigms in Drug Discovery. en. *International Journal of Molecular Sciences* **20,** 4331. ISSN: 1422-0067. https://www.mdpi.com/1422-0067/20/18/4331 (2025) (Sept. 2019).
3. Berman, H. M. *et al.* The Protein Data Bank. *Nucleic Acids Research* **28,** 235–242. ISSN: 0305-1048. https://doi.org/10.1093/nar/28.1.235 (2000).
4. Liu, T., Lin, Y., Wen, X., Jorissen, R. N. & Gilson, M. K. BindingDB: a web-accessible database of experimentally determined protein–ligand binding affinities. *Nucleic acids research* **35,** D198–D201 (2007).
5. Gilson, M. K. *et al.* BindingDB in 2015: A public database for medicinal chemistry, computational chemistry and systems pharmacology. *Nucleic Acids Research* **44,** D1045–D1053. ISSN: 0305-1048. https://doi.org/10.1093/nar/gkv1072 (2016).
6. Trott, O. & Olson, A. J. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J. Comp. Chem.* **31,** 455–461 (2010).
7. Huang, S.-Y., Grinter, S. Z. & Zou, X. Scoring functions and their evaluation methods for protein–ligand docking: recent advances and future directions. *Phys. Chem. Chem. Phys.* **12,** 12899–12908 (2010).
8. Wang, E. *et al.* End-Point Binding Free Energy Calculation with MM/PBSA and MM/GBSA: Strategies and Applications in Drug Design. *Chemical Reviews* **119,** 9478–9508. ISSN: 0009-2665. https://doi.org/10.1021/acs.chemrev.9b00055 (2026) (Aug. 2019).
9. Mey, A. S. J. S. *et al.* Best Practices for Alchemical Free Energy Calculations [Article v1.0]. en. *Living Journal of Computational Molecular Science* **2,** 18378–18378. ISSN: 2575-6524. https://livecomsjournal.org/index.php/livecoms/article/view/v2i1e18378 (2026) (Dec. 2020).

10. Wang, L. *et al.* Accurate and Reliable Prediction of Relative Ligand Binding Potency in Prospective Drug Discovery by Way of a Modern Free-Energy Calculation Protocol and Force Field. *Journal of the American Chemical Society* **137,** 2695–2703. ISSN: 0002-7863. https://doi.org/10.1021/ja512751q (2024) (Feb. 2015).

11. Abramson, J. *et al.* Accurate structure prediction of biomolecular interactions with AlphaFold 3. en. *Nature* **630,** 493–500. ISSN: 1476-4687. https://www.nature.com/articles/s41586-024-07487-w (2025) (June 2024).

12. Wohlwend, J. *et al. Boltz-1 Democratizing Biomolecular Interaction Modeling* en. Nov. 2024. http://biorxiv.org/lookup/doi/10.1101/2024.11.19.624167 (2025).

13. Li, J. *et al.* Leak Proof PDBBind: A Reorganized Dataset of Protein-Ligand Complexes for More Generalizable Binding Affinity Prediction. eng. *ArXiv,* arXiv:2308.09639v2. ISSN: 2331-8422 (May 2024).

14. Buttenschoen, M., Morris, G. M. & Deane, C. M. *PoseBusters: AI-based docking methods fail to generate physically valid poses or generalise to novel sequences* 2023. arXiv: 2308.05777 [q-bio.QM].

15. Wang, Y. *et al.* A workflow to create a high-quality protein–ligand binding dataset for training, validation, and prediction tasks. *Digital Discovery* **4,** 1209–1220 (2025).

16. Nerenberg, P. S. & Head-Gordon, T. New developments in force fields for biomolecular simulations. *Curr Opin Struct Biol* **49,** 129–138. ISSN: 1879-033X (Electronic) 0959-440X (Linking). https://www.ncbi.nlm.nih.gov/pubmed/29477047 (2018).

17. Ponder, J. W. *et al.* Current Status of the AMOEBA Polarizable Force Field. *J Phys Chem B* **114,** 2549–64. ISSN: 1520-6106 (Print) 1520-5207 (Electronic) (2010).

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

19. Mao, Y. *et al.* From intermolecular interaction energies and observable shifts to component contributions and back again: A tale of variational energy decomposition analysis. *Annu. Rev. Phys. Chem.* **72,** 641–666 (2021).

20. Wang, X. *et al.* Dmff: an open-source automatic differentiable platform for molecular force field development and molecular dynamics simulation. *J. Comp. Theo. Comput.* **19,** 5897–5909 (2023).

21. Wang, Y. *EasyBFE: An Open-Source Package for Protein–Ligand Binding Free-Energy Calculations* 2026. https://github.com/Ericwang6/easybfe (2026).

22. MacDermott-Opeskin, H. *et al. A Computational Community Blind Challenge on Pan-Coronavirus Drug Discovery Data* en. July 2025. https://chemrxiv.org/engage/chemrxiv/article-details/6878bef4fc5f0acb52a813f5 (2025).

23. Yang, M. *et al.* Uni-GBSA: an open-source and web-based automatic workflow to perform MM/GB (PB) SA calculations for virtual screening. *Briefings in Bioinformatics* **24,** bbad218 (2023).

24. Wang, R., Fang, X., Lu, Y. & Wang, S. The PDBbind Database: Collection of Binding Affinities for Protein-Ligand Complexes with Known Three-Dimensional Structures. en. *Journal of Medicinal Chemistry* **47,** 2977–2980. ISSN: 0022-2623, 1520-4804. `https://pubs.acs.org/doi/10.1021/jm0305801` (2024) (June 2004).

25. Zhang, C., Zhang, X., Freddolino, P. L. & Zhang, Y. BioLiP2: an updated structure database for biologically relevant ligand–protein interactions. *Nucleic Acids Research* **52,** D404–D412. ISSN: 0305-1048. eprint: `https://academic.oup.com/nar/article-pdf/52/D1/D404/55039597/gkad630.pdf`. `https://doi.org/10.1093/nar/gkad630` (July 2023).

26. Wagle, S. *et al.* Sunsetting binding MOAD with its last data update and the addition of 3D-ligand polypharmacology tools. *Scientific Reports* **13,** 3008 (2023).

27. Trott, O. & Olson, A. J. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J. Comp. Chem.* **31,** 455–461 (2010).

28. Morris, G. M. *et al.* AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *J. Comp. Chem.* **30,** 2785–2791 (2009).

29. Friesner, R. A. *et al.* Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *J. Med. Chem.* **47,** 1739–1749 (2004).

30. Friesner, R. A. *et al.* Extra precision glide: Docking and scoring incorporating a model of hydrophobic enclosure for protein- ligand complexes. *J. Med. Chem.* **49,** 6177–6196 (2006).

31. Ballester, P. J. & Mitchell, J. B. A machine learning approach to predicting protein–ligand binding affinity with applications to molecular docking. *Bioinformatics* **26,** 1169–1175 (2010).

32. Jiang, D. *et al.* Interactiongraphnet: A novel and efficient deep graph representation learning framework for accurate protein–ligand interaction predictions. *J. Med. Chem.* **64,** 18209–18232 (2021).

33. Moon, S., Zhung, W., Yang, S., Lim, J. & Kim, W. Y. PIGNet: a physics-informed deep learning model toward generalized drug–target interaction predictions. *Chemical Science* **13,** 3661–3673 (2022).

34. Shen, C. *et al.* Boosting protein–ligand binding pose prediction and virtual screening based on residue–atom distance likelihood potential and graph transformer. *J. Med. Chem.* **65,** 10691–10706 (2022).

35. Liu, Z. *et al.* Forging the basis for developing protein–ligand interaction scoring functions. *Accounts of chemical research* **50,** 302–309 (2017).

36. Huang, S.-Y., Grinter, S. Z. & Zou, X. Scoring functions and their evaluation methods for protein–ligand docking: recent advances and future directions. *Phys. Chem. Chem. Phys.* **12,** 12899–12908 (2010).

37. Cheng, T., Li, X., Li, Y., Liu, Z. & Wang, R. Comparative assessment of scoring functions on a diverse test set. *J. Chem. Info. Model.* **49,** 1079–1093 (2009).

38. Li, Y. *et al.* Comparative assessment of scoring functions on an updated benchmark: 1. Compilation of the test set. *Journal of chemical information and modeling* **54,** 1700–1716 (2014).

39. Li, Y., Han, L., Liu, Z. & Wang, R. Comparative assessment of scoring functions on an updated benchmark: 2. Evaluation methods and general results. *Journal of chemical information and modeling* **54,** 1717–1736 (2014).

40. Su, M. *et al.* Comparative assessment of scoring functions: the CASF-2016 update. *J. Chem. Info. Model.* **59,** 895–913 (2018).

41. Kramer, C. & Gedeck, P. Leave-cluster-out cross-validation is appropriate for scoring functions derived from diverse protein data sets. *Journal of chemical information and modeling* **50,** 1961–1969 (2010).

42. Ballester, P. J. & Mitchell, J. B. *Comments on "leave-cluster-out cross-validation is appropriate for scoring functions derived from diverse protein data sets": Significance for the validation of scoring functions* 2011.

43. Zhu, H., Yang, J. & Huang, N. Assessment of the generalization abilities of machine-learning scoring functions for structure-based virtual screening. *Journal of Chemical Information and Modeling* **62,** 5485–5502 (2022).

44. Wang, D. D., Xie, H. & Yan, H. Proteo-chemometrics interaction fingerprints of protein–ligand complexes predict binding affinity. *Bioinformatics* **37,** 2570–2579 (2021).

45. Ozturk, H., Ozgur, A. & Ozkirimli, E. DeepDTA: deep drug–target binding affinity prediction. *Bioinformatics* **34,** i821–i829. ISSN: 1367-4803. `https://doi.org/10.1093/bioinformatics/bty593` (2018).

46. *PDBFixer* `https://github.com/openmm/pdbfixer`. accessed on Oct 29, 2024.

47. Landrum, G. *et al.* RDKit: A software suite for cheminformatics, computational chemistry, and predictive modeling. *Greg Landrum* **8** (2013).

48. Maier, J. A. *et al.* ff14SB: improving the accuracy of protein side chain and backbone parameters from ff99SB. *Journal of chemical theory and computation* **11,** 3696–3713 (2015).

49. Boothroyd, S. *et al.* Development and Benchmarking of Open Force Field 2.0.0: The Sage Small Molecule Force Field. *J. Chem. Theo. Comp.* **19,** 3251–3275. ISSN: 1549-9618. https://doi.org/10.1021/acs.jctc.3c00039 (2023).

50. Eastman, P. *et al.* OpenMM 8: Molecular Dynamics Simulation with Machine Learning Potentials. *The Journal of Physical Chemistry B* **128.** PMID: 38154096, 109–116. eprint: https://doi.org/10.1021/acs.jpcb.3c06662. https://doi.org/10.1021/acs.jpcb.3c06662 (2024).

51. Albaugh, A. *et al.* Advanced Potential Energy Surfaces for Molecular Simulation. *J Phys Chem B* **120,** 9811–32. ISSN: 1520-6106 (Print) 1520-5207 (Electronic) (2016).

52. Kovács, D. P. *et al.* Mace-off: Short-range transferable machine learning force fields for organic molecules. *J. Am. Chem. Soc.* **147,** 17598–17611 (2025).

53. Anstine, D. M., Zubatyuk, R. & Isayev, O. AIMNet2: a neural network potential to meet your neutral, charged, organic, and elemental-organic needs. *Chem. Sci.* **16,** 10228–10244 (2025).

54. Wood, B. M. *et al. UMA: A Family of Universal Models for Atoms* 2025. arXiv: 2506.23971 [cs.LG]. https://arxiv.org/abs/2506.23971.

55. Levine, D. S. *et al.* The open molecules 2025 (omol25) dataset, evaluations, and models. *arXiv preprint arXiv:2505.08762* (2025).

56. Yuan, E. C.-Y. & Head-Gordon, T. Teachers that teach the irrelevant: Pre-training machine learned interaction potentials with classical force fields for robust molecular dynamics simulations. *arXiv preprint arXiv:2509.14205* (2025).

57. Wang, J., Wolf, R. M., Caldwell, J. W., Kollman, P. A. & Case, D. A. Development and testing of a general amber force field. *J. Comp. Chem.* **25,** 1157–1174 (2004).

58. MacKerell, A. D. *et al.* All-Atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B* **102,** 3586–3616 (1998).

59. Vanommeslaeghe, K. *et al.* CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields. *J. Comp. Chem.* **31,** 671–690 (2010).

60. Jorgensen, W. L. & Schyman, P. Treatment of halogen bonding in the OPLS-AA force field: application to potent anti-HIV agents. *J. Comp. Theo. Comput.* **8,** 3895–3901 (2012).

61. Boothroyd, S. *et al.* Development and Benchmarking of Open Force Field 2.0.0: The Sage Small Molecule Force Field. *J. Comp. Theo. Comput.* **19,** 3251–3275. ISSN: 1549-9618. https://doi.org/10.1021/acs.jctc.3c00039 (2023).

62. Demerdash, O., Wang, L.-P. & Head-Gordon, T. Advanced models for water simulations. *WIREs Comp. Mol. Sci.* **8,** e1355. ISSN: 1759-0876. https://doi.org/10.1002/wcms.1355 (2018).

63. Ross, G. A. *et al.* The maximal and current accuracy of rigorous protein-ligand binding free energy calculations. *Communications Chemistry* **6,** 222 (2023).

64. Baumann, H. M. *et al.* Large-scale collaborative assessment of binding free energy calculations for drug discovery using OpenFE. *ChemRxiv* **2025.** https://doi.org/10.26434/chemrxiv-2025-7sthd.

65. Demerdash, O., Mao, Y., Liu, T., Head-Gordon, M. & Head-Gordon, T. Assessing many-body contributions to intermolecular interactions of the AMOEBA force field using energy decomposition analysis of electronic structure calculations. *J. Chem. Phys.* **147,** 161721. ISSN: 1089-7690 (Electronic). https://www.ncbi.nlm.nih.gov/pubmed/29096520 (2017).

66. Mao, Y. *et al.* From intermolecular interaction energies and observable shifts to component contributions and back again: A tale of variational energy decomposition analysis. *Annu. Rev. Phys. Chem.* **72,** 641–666 (2021).

67. Ponder, J. W. *et al.* Current Status of the AMOEBA Polarizable Force Field. *J Phys Chem B* **114,** 2549–64. ISSN: 1520-6106 (Print) 1520-5207 (Electronic) (2010).

68. Horn, P. R., Mao, Y. & Head-Gordon, M. Defining the contributions of permanent electrostatics, Pauli repulsion, and dispersion in density functional theory calculations of intermolecular interaction energies. *The Journal of chemical physics* **144** (2016).

69. Paszke, A. *et al.* in *Adv Neural Inf Process Syst 32* 8024–8035 (Curran Associates, Inc., 2019).

70. Bradbury, J. *et al. JAX: composable transformations of Python+NumPy programs* version 0.3.13. 2018. http://github.com.

71. Doerr, S. *et al.* TorchMD: A deep learning framework for molecular simulations. *J. Comp. Theo. Comput.* **17,** 2355–2363 (2021).

72. Schoenholz, S. & Cubuk, E. D. Jax md: a framework for differentiable physics. *Adv Neural Inf Process Syst* **33,** 11428–11441 (2020).

73. Heindel, J. P., Sami, S. & Head-Gordon, T. Completely multipolar model as a general framework for many-body interactions as illustrated for water. *J. Comp. Theo. Comput.* **20,** 8594–8608 (2024).

74. Wang, J., Wolf, R. M., Caldwell, J. W., Kollman, P. A. & Case, D. A. Development and testing of a general amber force field. en. *Journal of Computational Chemistry* **25.** _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/jcc.20035, 1157–1174. ISSN: 1096-987X. https://onlinelibrary.wiley.com/doi/abs/10.1002/jcc.20035 (2026) (2004).

75. MacKerell, A. D. J. *et al.* All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins. *J. Phys. Chem. B* **102,** 3586–3616. ISSN: 1520-6106. https://doi.org/10.1021/jp973084f (2026) (Apr. 1998).

76. Robertson, M. J., Tirado-Rives, J. & Jorgensen, W. L. Improved Peptide and Protein Torsional Energetics with the OPLS-AA Force Field. *J. Chem. Theory Comput.* **11,** 3499–3509. ISSN: 1549-9618. https://doi.org/10.1021/acs.jctc.5b00356 (2026) (July 2015).

77. Qiu, Y. *et al.* Development and benchmarking of open force field v1. 0.0—the parsley small-molecule force field. *Journal of chemical theory and computation* **17,** 6262–6280 (2021).

78. Ponder, J. W. *et al.* Current Status of the AMOEBA Polarizable Force Field. *J. Phys. Chem. B* **114,** 2549–2564. ISSN: 1520-6106. https://doi.org/10.1021/jp910674d (2025) (Mar. 2010).

79. Ren, P. & Ponder, J. W. Consistent treatment of inter- and intramolecular polarization in molecular mechanics calculations. en. *Journal of Computational Chemistry* **23.** _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/jcc.10127, 1497–1506. ISSN: 1096-987X. https://onlinelibrary.wiley.com/doi/abs/10.1002/jcc.10127 (2026) (2002).

80. Sami, S., Menger, M. F., Faraji, S., Broer, R. & Havenith, R. W. A. Q-Force: Quantum Mechanically Augmented Molecular Force Fields. *J. Chem. Theory Comput.* **17,** 4946–4960. ISSN: 1549-9618. https://doi.org/10.1021/acs.jctc.1c00195 (2026) (Aug. 2021).

81. Abdullah, A. S., Wang, Y., Menger, M. F., Sami, S. & Head-Gordon, T. Improved Treatment of 1–4 Interactions in Force Fields for Molecular Dynamics Simulations. en-US. *J. Chem. Theory Comput.* **21,** 11669–11678. ISSN: 1549-9618. https://doi.org/10.1021/acs.jctc.5c01458 (2026) (Nov. 2025).

82. Wang, L. *et al.* Accurate and reliable prediction of relative ligand binding potency in prospective drug discovery by way of a modern free-energy calculation protocol and force field. *J. Am. Chem. Soc.* **137,** 2695–2703 (2015).

83. Lee, T.-S. *et al.* Alchemical Binding Free Energy Calculations in AMBER20: Advances and Best Practices for Drug Discovery. *Journal of Chemical Information and Modeling* **60,** 5595–5623. ISSN: 1549-9596. https://doi.org/10.1021/acs.jcim.0c00613 (2024) (Nov. 2020).

84. Tsai, H.-C. *et al.* AMBER Free Energy Tools: A New Framework for the Design of Optimized Alchemical Transformation Pathways. *Journal of Chemical Theory and Computation* **19,** 640–658. ISSN: 1549-9618. https://doi.org/10.1021/acs.jctc.2c00725 (2024) (Jan. 2023).

85. Wu, Z. *et al.* alchemlyb: the simple alchemistry library. en. *Journal of Open Source Software* **9,** 6934. ISSN: 2475-9066. https://joss.theoj.org/papers/10.21105/joss.06934 (2026) (Sept. 2024).

86. Baumann, H. M. *et al.* Large-Scale Collaborative Assessment of Binding Free Energy Calculations for Drug Discovery Using OpenFE. *Journal of Chemical Information and Modeling* **66,** 6429–6452. ISSN: 1549-9596. https://doi.org/10.1021/acs.jcim.6c00089 (2026) (June 2026).

87. Boresch, S., Tettinger, F., Leitgeb, M. & Karplus, M. Absolute Binding Free Energies: A Quantitative Approach for Their Calculation. *The Journal of Physical Chemistry B* **107,** 9535–9551. ISSN: 1520-6106. https://doi.org/10.1021/jp0217839 (2026) (Sept. 2003).

88. *Antiviral Drug Discovery (AViDD) Centers for Pathogens of Pandemic Concern | NIAID: National Institute of Allergy and Infectious Diseases* en. May 2022. https://www.niaid.nih.gov/research/antiviral-drug-discovery-centers-pathogens-pandemic-concern (2025).

89. Jimmidi, R. *et al.* DNA-encoded chemical libraries yield non-covalent and non-peptidic SARS-CoV-2 main protease inhibitors. *Communications Chemistry* **6,** 164. https://doi.org/10.1038/s42004-023-00961-y (2023).

90. CACHE. *Finding Selective Inhibitors of Phosphoglycerate Kinase 2: CACHE Challenge #7* 2025. https://cache-challenge.org/challenges/finding-selective-inhibitors-of-phosphoglycerate-kinase-2 (2026).

91. Cavanagh, J. M. *et al. SmileyLlama: Modifying Large Language Models for Directed Chemical Space Exploration* arXiv:2409.02231. 2024. https://doi.org/10.48550/arXiv.2409.02231.

92. Sun, K. *et al.* SynLlama: Generating Synthesizable Molecules and Their Analogs with Large Language Models. *ACS Central Science* **11,** 2108–2120. https://doi.org/10.1021/acscentsci.5c01285 (2025).

93. Miller III, B. R. *et al.* MMPBSA. py: an efficient program for end-state free energy calculations. *Journal of chemical theory and computation* **8,** 3314–3321 (2012).

94. Qiu, Y. *et al.* Development and Benchmarking of Open Force Field v1.0.0-the Parsley Small-Molecule Force Field. *J Chem Theory Comput* **17,** 6262–6280. ISSN: 1549-9618 (Print) 1549-9618 (2021).

95. Wang, R., Lai, L. & Wang, S. Further development and validation of empirical scoring functions for structure-based binding affinity prediction. *Journal of computer-aided molecular design* **16,** 11–26 (2002).

96. Jones, G., Willett, P., Glen, R. C., Leach, A. R. & Taylor, R. Development and validation of a genetic algorithm for flexible docking. *J. Mol. Bio.* **267,** 727–748 (1997).

97. Muegge, I. PMF scoring revisited. *J. Med. Chem.* **49,** 5895–5902 (2006).

98. Huang, N., Kalyanaraman, C., Bernacki, K. & Jacobson, M. P. Molecular mechanics methods for predicting protein–ligand binding. *Phys. Chem. Chem. Phys.* **8,** 5166–5177 (2006).

99. Dittrich, J., Schmidt, D., Pfleger, C. & Gohlke, H. Converging a knowledge-based scoring function: DrugScore2018. *J. Chem. Info. Model.* **59,** 509–521 (2018).

100. Somnath, V. R., Bunne, C. & Krause, A. Multi-scale representation learning on proteins. *Advances in Neural Information Processing Systems* **34,** 25244–25255 (2021).

101. Lu, W. *et al.* Tankbind: Trigonometry-aware neural networks for drug-protein binding structure prediction. *bioRxiv,* 2022–06 (2022).

102. Yang, Z., Zhong, W., Lv, Q., Dong, T. & Yu-Chian Chen, C. Geometric interaction graph neural network for predicting protein–ligand binding affinities from 3d structures (gign). *The journal of physical chemistry letters* **14,** 2020–2033 (2023).

103. Wang, R., Fang, X., Lu, Y., Yang, C.-Y. & Wang, S. The PDBbind Database: Methodologies and Updates. en. *Journal of Medicinal Chemistry* **48,** 4111–4119. ISSN: 0022-2623, 1520-4804. `https://pubs.acs.org/doi/10.1021/jm048957q` (2024) (June 2005).

104. Cheng, T., Li, X., Li, Y., Liu, Z. & Wang, R. Comparative Assessment of Scoring Functions on a Diverse Test Set. en. *Journal of Chemical Information and Modeling* **49,** 1079–1093. ISSN: 1549-9596, 1549-960X. `https://pubs.acs.org/doi/10.1021/ci9000053` (2024) (Apr. 2009).

105. Li, Y. *et al.* Comparative Assessment of Scoring Functions on an Updated Benchmark: 1. Compilation of the Test Set. en. *Journal of Chemical Information and Modeling* **54,** 1700–1716. ISSN: 1549-9596, 1549-960X. `https://pubs.acs.org/doi/10.1021/ci500080q` (2024) (June 2014).

106. Li, Y., Han, L., Liu, Z. & Wang, R. Comparative Assessment of Scoring Functions on an Updated Benchmark: 2. Evaluation Methods and General Results. en. *Journal of Chemical Information and Modeling* **54,** 1717–1736. ISSN: 1549-9596, 1549-960X. `https://pubs.acs.org/doi/10.1021/ci500081m` (2024) (June 2014).

107. Liu, Z. *et al.* PDB-wide collection of binding data: current status of the PDBbind database. *Bioinformatics* **31,** 405–412 (2015).

108. Dunbar Jr, J. B. *et al.* CSAR benchmark exercise of 2010: selection of the protein–ligand complexes. *Journal of chemical information and modeling* **51,** 2036–2046 (2011).

109. Li, H., Leung, K.-S., Wong, M.-H. & Ballester, P. J. Improving AutoDock Vina using random forest: the growing accuracy of binding affinity prediction by the effective exploitation of larger data sets. *Molecular informatics* **34,** 115–126 (2015).

110. Stärk, H., Ganea, O., Pattanaik, L., Barzilay, R. & Jaakkola, T. *Equibind: Geometric deep learning for drug binding structure prediction* in *International conference on machine learning* (2022), 20503–20521.

111. Liu, T. *et al.* BindingDB in 2024: a FAIR knowledgebase of protein-small molecule binding data. *Nucleic acids research* **53,** D1633–D1644 (2025).

112. Hu, L., Benson, M. L., Smith, R. D., Lerner, M. G. & Carlson, H. A. Binding MOAD (mother of all databases). *Proteins: Structure, Function, and Bioinformatics* **60,** 333–340 (2005).

113. Ahmed, A., Smith, R. D., Clark, J. J., Dunbar Jr, J. B. & Carlson, H. A. Recent improvements to Binding MOAD: a resource for protein–ligand binding affinities and structures. *Nucleic acids research* **43,** D465–D469 (2015).

114. Benson, M. L. *et al.* Binding MOAD, a high-quality protein-ligand database. eng. *Nucleic Acids Research* **36,** D674–678. ISSN: 1362-4962 (Jan. 2008).

115. Yang, J., Roy, A. & Zhang, Y. BioLiP: a semi-manually curated database for biologically relevant ligand-protein interactions. eng. *Nucleic Acids Research* **41,** D1096–1103. ISSN: 1362-4962 (Jan. 2013).

116. Durairaj, J. *et al.* PLINDER: The protein-ligand interactions dataset and evaluation resource. *bioRxiv,* 2024–07 (2024).

117. Corso, G., Deng, A., Polizzi, N., Barzilay, R. & Jaakkola, T. *The Discovery of Binding Modes Requires Rethinking Docking Generalization* in *NeurIPS 2023 Generative AI and Biology (GenBio) Workshop* (2023).

118. Guo, X.-K. & Zhang, Y. CovBinderInPDB: A Structure-Based Covalent Binder Database. *J. Chem. Info. Model.* **62,** 6057–6068 (2022).

119. Liu, Z. H., Tsanai, M., Zhang, O., Forman-Kay, J. & Head-Gordon, T. *Computational Methods to Investigate Intrinsically Disordered Proteins and their Complexes* 2024. arXiv: 2409.02240 [physics.bio-ph]. https://arxiv.org/abs/2409.02240.

120. Gall, T. L., Romero, P. R., Cortese, M. S., Uversky, V. N. & Dunker, A. K. Intrinsic disorder in the protein data bank. *Journal of Biomolecular structure and dynamics* **24,** 325–341 (2007).

121. O'Boyle, N. M. *et al.* Open Babel: An open chemical toolbox. *Journal of cheminformatics* **3,** 1–14 (2011).

122. Luo, W. *et al.* Bridging Machine Learning and Thermodynamics for Accurate p K a Prediction. *JACS Au* **4,** 3451–3465 (2024).

123. Johnston, R. C. *et al.* Epik: p K a and Protonation State Prediction through Machine Learning. *Journal of chemical theory and computation* **19,** 2380–2388 (2023).

124. *Prediction of dissociation constant using microconstants.* https://docs.chemaxon.com/display/docs/attachments/attachments_1814016_1_Prediction_of_dissociation_constant_using_microconstants.pdf. accessed on Oct 31, 2024.

125. Ropp, P. J., Kaminsky, J. C., Yablonski, S. & Durrant, J. D. Dimorphite-DL: an open-source program for enumerating the ionization states of drug-like small molecules. *Journal of Cheminformatics* **11,** 1–8 (2019).

126. Gasteiger, J. & Marsili, M. A new model for calculating atomic charges in molecules. *Tetrahedron Letters* **19,** 3181–3184. https://api.semanticscholar.org/CorpusID:97016098 (1978).

127. Fermi, E. *Thermodynamics* (Courier Corporation, 2012).

128. MarÉchal, E. Measuring bioactivity: KI, IC50 and EC50. *Chemogenomics and Chemical Genetics: A User's Introduction for Biologists, Chemists Informaticians,* 55–65 (2011).

129. Ross, G. A. *et al.* The maximal and current accuracy of rigorous protein-ligand binding free energy calculations. *Communications Chemistry* **6,** 222 (2023).

130. Zhu, K. *et al.* Docking covalent inhibitors: a parameter free approach to pose prediction and scoring. *Journal of chemical information and modeling* **54,** 1932–1940 (2014).

131. Bianco, G., Forli, S., Goodsell, D. S. & Olson, A. J. Covalent docking using autodock: Two-point attractor and flexible side chain methods. *Protein Science* **25,** 295–301 (2016).

132. Delbart, F. *et al.* An allosteric binding site of the α7 nicotinic acetylcholine receptor revealed in a humanized acetylcholine-binding protein. en. *Journal of Biological Chemistry* **293,** 2534–2545. ISSN: 00219258. https://linkinghub.elsevier.com/retrieve/pii/S0021925820317142 (2024) (Feb. 2018).

133. Rosner, K. E. *et al.* The discovery of novel tartrate-based TNF-α converting enzyme (TACE) inhibitors. en. *Bioorganic & Medicinal Chemistry Letters* **20,** 1189–1193. ISSN: 0960894X. https://linkinghub.elsevier.com/retrieve/pii/S0960894X09017053 (2024) (Feb. 2010).

134. Fasan, R. *et al.* Structure–Activity Studies in a Family of β-Hairpin Protein Epitope Mimetic Inhibitors of the p53–HDM2 Protein–Protein Interaction. en. *ChemBioChem* **7,** 515–526. ISSN: 1439-4227, 1439-7633. https://chemistry-europe.onlinelibrary.wiley.com/doi/10.1002/cbic.200500452 (2024) (Mar. 2006).

135. Lin, G.-d. *et al.* Crystal structure of calcium bound domain VI of calpain at 1.9 Å resolution and its role in enzyme assembly, regulation, and inhibitor binding. *Nature Structural Biology* **4,** 539–547. ISSN: 1072-8368. https://www.nature.com/doifinder/10.1038/nsb0797-539 (2024) (July 1997).

136. Ståhlberg, J. *et al.* Structural basis for enantiomer binding and separation of a common β-blocker: crystal structure of cellobiohydrolase Cel7A with bound (S)-propranolol at 1.9 Å resolution. en. *Journal of Molecular Biology* **305,** 79–93. ISSN: 00222836. https://linkinghub.elsevier.com/retrieve/pii/S0022283600942372 (2024) (Jan. 2001).

137. Zhang, M., Zhou, M., Van Etten, R. L. & Stauffacher, C. V. Crystal Structure of Bovine Low Molecular Weight Phosphotyrosyl Phosphatase Complexed with the Transition State Analog Vanadate . en. *Biochemistry* **36,** 15–23. ISSN: 0006-2960, 1520-4995. https://pubs.acs.org/doi/10.1021/bi961804n (2024) (Jan. 1997).

138. Dennis, M. L. *et al.* Structural Basis for the Selective Binding of Inhibitors to 6-Hydroxymethyl-7, 8-dihydropterin Pyrophosphokinase from Staphylococcus aureus and Escherichia coli. *Journal of Medicinal Chemistry* **59,** 5248–5263 (2016).

139. Buchbinder, J. L., Stephenson, R. C., Scanlan, T. S. & Fletterick, R. J. A comparison of the crystallographic structures of two catalytic antibodies with esterase activity11Edited by I. A. Wilson. *Journal of Molecular Biology* **282,** 1033–1041. ISSN: 0022-2836. https://www.sciencedirect.com/science/article/pii/S0022283698920253 (1998).

140. Shapovalov, M. V. & Dunbrack, R. L. A smoothed backbone-dependent rotamer library for proteins derived from adaptive kernel density estimates and regressions. *Structure* **19,** 844–858 (2011).

141. Keough, D. T. *et al.* Inhibition of Hypoxanthine-Guanine Phosphoribosyltransferase by Acyclic Nucleoside Phosphonates: A New Class of Antimalarial Therapeutics. *Journal of Medicinal Chemistry* **52.** PMID: 19527031, 4391–4399. eprint: https://doi.org/10.1021/jm900267n. https://doi.org/10.1021/jm900267n (2009).

142. Boothroyd, S. *et al.* Development and Benchmarking of Open Force Field 2.0.0: The Sage Small Molecule Force Field. *Journal of Chemical Theory and Computation* **19,** 3251–3275. ISSN: 1549-9618. https://doi.org/10.1021/acs.jctc.3c00039 (2023).

143. Lemkul, J. A., Huang, J., Roux, B. & MacKerell Alexander D., J. An Empirical Polarizable Force Field Based on the Classical Drude Oscillator Model: Development History and Recent Applications. *Chem. Rev.* **116,** 4983–5013 (2016).

144. Henriksen, N. M. & Gilson, M. K. Evaluating Force Field Performance in Thermodynamic Calculations of Cyclodextrin Host–Guest Binding: Water Models, Partial Charges, and Host Force Field Parameters. *J. Chem. Theo. Comput.* **13,** 4253–4269. ISSN: 1549-9618. https://doi.org/10.1021/acs.jctc.7b00359 (2017).

145. Chen, C., Yu, B., Yousefi, R., Iwahara, J. & Pettitt, B. M. Assessment of the Components of the Electrostatic Potential of Proteins in Solution: Comparing Experiment and Theory. *J. Phys. Chem. B* **126,** 4543–4554. ISSN: 1520-6106. https://doi.org/10.1021/acs.jpcb.2c01611 (2022).

146. León, A. M., Andersen, L. & Hub, J. S. BindFlow: a free, user-friendly pipeline for absolute binding free energy calculations using free energy perturbation or MM (PB/GB) SA. *Journal of Chemical Theory and Computation* (2025).

147. Wang, L. *et al.* Accurate and reliable prediction of relative ligand binding potency in prospective drug discovery by way of a modern free-energy calculation protocol and force field. *Journal of the American Chemical Society* **137,** 2695–2703 (2015).

148. Albaugh, A. *et al.* Advanced Potential Energy Surfaces for Molecular Simulation. *J Phys Chem B* **120,** 9811–32. ISSN: 1520-6106 (Print) 1520-5207 (Electronic) (2016).

149. Szalewicz, K. Symmetry-adapted perturbation theory of intermolecular forces. *WIRES: Comput. Mol. Sci.* **2,** 254–272 (2012).

150. Demerdash, O., Mao, Y., Liu, T., Head-Gordon, M. & Head-Gordon, T. Assessing many-body contributions to intermolecular interactions of the AMOEBA force field using energy decomposition analysis of electronic structure calculations. *J. Chem. Phys.* **147,** 161721. ISSN: 1089-7690 (Electronic). `https://www.ncbi.nlm.nih.gov/pubmed/29096520` (2017).

151. Das, A. K. *et al.* Development of an advanced force field for water using variational energy decomposition analysis. *Journal of chemical theory and computation* **15,** 5001–5013 (2019).

152. Das, A. K., Liu, M. & Head-Gordon, T. Development of a many-body force field for aqueous alkali metal and halogen ions: an energy decomposition analysis guided approach. *Journal of Chemical Theory and Computation* **18,** 953–967 (2022).

153. Heindel, J. P., Sami, S. & Head-Gordon, T. Completely multipolar model as a general framework for many-body interactions as illustrated for water. *Journal of Chemical Theory and Computation* **20,** 8594–8608 (2024).

154. Heindel, J. P., Kim, L., Head-Gordon, M. & Head-Gordon, T. Completely Multipolar Model for Many-Body Water–Ion and Ion–Ion Interactions. *The Journal of Physical Chemistry Letters* **16,** 975–984 (2025).

155. Liu, C., Piquemal, J.-P. & Ren, P. AMOEBA+ classical potential for modeling molecular interactions. *Journal of chemical theory and computation* **15,** 4122–4139 (2019).

156. Spronk, S. A., Glick, Z. L., Metcalf, D. P., Sherrill, C. D. & Cheney, D. L. A quantum chemical interaction energy dataset for accurately modeling protein-ligand interactions. *Scientific Data* **10,** 619 (2023).

157. Donchev, A. G. *et al.* Quantum chemical benchmark databases of gold-standard dimer interaction energies. *Scientific data* **8,** 55 (2021).

158. Schriber, J. B. *et al.* CLIFF: A component-based, machine-learned, intermolecular force field. *The Journal of Chemical Physics* **154** (2021).

159. Villot, C. & Lao, K. U. Ab initio dispersion potentials based on physics-based functional forms with machine learning. *The Journal of Chemical Physics* **160** (2024).

160. Sparrow, Z. M., Ernst, B. G., Joo, P. T., Lao, K. U. & DiStasio, R. A. NENCI-2021. I. A large benchmark database of non-equilibrium non-covalent interactions emphasizing close intermolecular contacts. *The Journal of Chemical Physics* **155** (2021).

161. Horn, P. R. & Head-Gordon, M. Polarization contributions to intermolecular interactions revisited with fragment electric-field response functions. *The Journal of Chemical Physics* **143,** 114111. ISSN: 0021-9606. https://doi.org/10.1063/1.4930534 (2015).

162. Su, M. *et al.* Comparative assessment of scoring functions: the CASF-2016 update. *J. Chem. Info. Model.* **59,** 895–913 (2018).

163. Yang, J., Roy, A. & Zhang, Y. BioLiP: a semi-manually curated database for biologically relevant ligand-protein interactions. eng. *Nucleic Acids Research* **41,** D1096–1103. ISSN: 1362-4962 (Jan. 2013).

164. Zhang, C., Zhang, X., Freddolino, P. L. & Zhang, Y. BioLiP2: an updated structure database for biologically relevant ligand–protein interactions. *Nucleic Acids Research* **52,** D404–D412. ISSN: 0305-1048. eprint: https://academic.oup.com/nar/article-pdf/52/D1/D404/55039597/gkad630.pdf. https://doi.org/10.1093/nar/gkad630 (July 2023).

165. Salentin, S., Schreiber, S., Haupt, V. J., Adasme, M. F. & Schroeder, M. PLIP: fully automated protein–ligand interaction profiler. *Nucleic acids research* **43,** W443–W447 (2015).

166. Adasme, M. F. *et al.* PLIP 2021: expanding the scope of the protein–ligand interaction profiler to DNA and RNA. *Nucleic acids research* **49,** W530–W534 (2021).

167. Schake, P., Bolz, S. N., Linnemann, K. & Schroeder, M. PLIP 2025: introducing protein–protein interactions to the protein–ligand interaction profiler. *Nucleic Acids Research,* gkaf361 (2025).

168. Ji, Z., Shi, R., Lu, J., Li, F. & Yang, Y. ReLMole: molecular representation learning based on two-level graph similarities. *Journal of Chemical Information and Modeling* **62,** 5361–5372 (2022).

169. Cheng, B. *et al.* Mapping materials and molecules. *Accounts of Chemical Research* **53,** 1981–1991 (2020).

170. Epifanovsky, E. *et al.* Software for the frontiers of quantum chemistry: An overview of developments in the Q-Chem 5 package. *The Journal of chemical physics* **155** (2021).

171. Vydrov, O. A. & Van Voorhis, T. Nonlocal van der Waals density functional: The simpler the better. *The Journal of chemical physics* **133** (2010).

172. Abdullah, A. S., Wang, Y., Menger, M. F., Sami, S. & Head-Gordon, T. Improved Treatment of 1–4 Interactions in Force Fields for Molecular Dynamics Simulations. *Journal of Chemical Theory and Computation* (2025).

173. Carter-Fenk, K. *et al.* The energetic origins of pi–pi contacts in proteins. *Journal of the American Chemical Society* **145,** 24836–24851 (2023).

174. Puleva, M. *et al.* Extending quantum-mechanical benchmark accuracy to biological ligand-pocket interactions. *Nature Communications* **16,** 8583 (2025).

175. Mao, Y., Ge, Q., Horn, P. R. & Head-Gordon, M. On the Computational Characterization of Charge-Transfer Effects in Noncovalently Bound Molecular Complexes. *Journal of Chemical Theory and Computation* **14,** 2401–2417. ISSN: 1549-9618. https://doi.org/10.1021/acs.jctc.7b01256 (2018).

176. Aldossary, A., Shen, H., Wang, Z. & Head-Gordon, M. Uncoupled fragment electric-field response functions: An accelerated model for the polarization energy in energy decomposition analysis of intermolecular interactions. *Chemical Physics Letters* **862,** 141825. ISSN: 0009-2614. https://www.sciencedirect.com/science/article/pii/S000926142400767X (2025).

177. Burns, L. A. *et al.* The BioFragment Database (BFDb): An open-data platform for computational chemistry analysis of noncovalent interactions. *The Journal of chemical physics* **147** (2017).

178. Sousa da Silva, A. W. & Vranken, W. F. ACPYPE-Antechamber python parser interface. *BMC research notes* **5,** 367 (2012).

179. Jakalian, A., Jack, D. B. & Bayly, C. I. Fast, efficient generation of high-quality atomic charges. AM1-BCC model: II. Parameterization and validation. *Journal of Computational Chemistry* **23,** 1623–1641. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/jcc.10128. https://onlinelibrary.wiley.com/doi/abs/10.1002/jcc.10128 (2002).

180. Walker, B., Liu, C., Wait, E. & Ren, P. Automation of AMOEBA polarizable force field for small molecules: Poltype 2. *Journal of computational chemistry* **43,** 1530–1542 (2022).

181. *MACE Package release 0.3.14* https://github.com/ACEsuit/mace/releases/tag/v0.3.14. accessed on Jan 26, 2026.

182. *UMA Models* https://huggingface.co/facebook/UMA. accessed on Jan 26, 2026.

183. *fairchem* https://github.com/facebookresearch/fairchem. accessed on Jan 26, 2026.

184. *AIMNet Code* https://github.com/isayevlab/aimnetcentral. accessed on Jan 26, 2026.

185. Mecozzi, S., West, A. P. & Dougherty, D. A. Cation-pi interactions in aromatics of biological and medicinal interest: electrostatic potential surfaces as a useful qualitative guide. *Proceedings of the National Academy of Sciences* **93,** 10566–10571 (1996).

186. Hardegger, L. A. *et al.* Systematic investigation of halogen bonding in protein–ligand interactions. *Angewandte Chemie International Edition* **50,** 314–318 (2011).

187. Jorgensen, W. L. & Schyman, P. Treatment of halogen bonding in the OPLS-AA force field: application to potent anti-HIV agents. *Journal of chemical theory and computation* **8,** 3895–3901 (2012).

188. Wang, J., Wolf, R. M., Caldwell, J. W., Kollman, P. A. & Case, D. A. Development and testing of a general amber force field. *Journal of computational chemistry* **25,** 1157–1174 (2004).

189. Ponder, J. W. *et al.* Current status of the AMOEBA polarizable force field. *The journal of physical chemistry B* **114,** 2549–2564 (2010).

190. Kovács, D. P. *et al.* Mace-off: Short-range transferable machine learning force fields for organic molecules. *Journal of the American Chemical Society* **147,** 17598–17611 (2025).

191. Wood, B. M. *et al.* UMA: A Family of Universal Models for Atoms. *arXiv preprint arXiv:2506.23971* (2025).

192. Anstine, D. M., Zubatyuk, R. & Isayev, O. AIMNet2: a neural network potential to meet your neutral, charged, organic, and elemental-organic needs. *Chemical Science* **16,** 10228–10244 (2025).

193. Levine, D. S. *et al.* The open molecules 2025 (omol25) dataset, evaluations, and models. *arXiv preprint arXiv:2505.08762* (2025).

194. Maier, J. A. *et al.* ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *Journal of Chemical Theory and Computation* **11,** 3696–3713. ISSN: 1549-9618. https://doi.org/10.1021/acs.jctc.5b00255 (2015).

195. Liu, C., Qi, R., Wang, Q., Piquemal, J. P. & Ren, P. Capturing Many-Body Interactions with Classical Dipole Induction Models. *Journal of Chemical Theory and Computation* **13,** 2751–2761. ISSN: 1549-9618. https://doi.org/10.1021/acs.jctc.7b00225 (2017).

196. Shi, Y. *et al.* Polarizable atomic multipole-based AMOEBA force field for proteins. *Journal of chemical theory and computation* **9,** 4046–4063 (2013).

197. Jorgensen, W. L. & Tirado-Rives, J. The OPLS [optimized potentials for liquid simulations] potential functions for proteins, energy minimizations for crystals of cyclic peptides and crambin. *J. Am. Chem. Soc.* **110,** 1657–1666 (1988).

198. Abdullah, A. S., Wang, Y., Menger, M. F., Sami, S. & Head-Gordon, T. Improved Treatment of 1–4 Interactions in Force Fields for Molecular Dynamics Simulations. *J. Comp. Theo. Comput.* (2025).

199. Sami, S., LaCour, R. A., Heindel, J. P. & Head-Gordon, T. Simple and Accurate One-Body Energy and Dipole Moment Surfaces for Water and Beyond. *J. Phys. Chem. Lett.* **15,** 6712–6721 (2024).

200. Zhang, C. *et al.* AMOEBA polarizable atomic multipole force field for nucleic acids. *J. Comp. Theo. Comput.* **14,** 2084–2108 (2018).

201. Ponder, J. W. *et al.* Current status of the AMOEBA polarizable force field. *J. Phys. Chem. B* **114,** 2549–2564 (2010).

202. Rackers, J. A., Silva, R. R., Wang, Z. & Ponder, J. W. Polarizable Water Potential Derived from a Model Electron Density. *J. Chem. Theo. Comput.* **17,** 7056–7084 (2021).

203. Heindel, J. P., Kim, L., Head-Gordon, M. & Head-Gordon, T. Completely Multipolar Model for Many-Body Water–Ion and Ion–Ion Interactions. *J. Phys. Chem. Lett.* **16,** 975–984 (2025).

204. Chen, J. & Yu, K. Phyneo: a neural-network-enhanced physics-driven force field development workflow for bulk organic molecule and polymer simulations. *J. Comp. Theo. Comput.* **20,** 253–265 (2023).

205. Chen, J. *et al.* A Hybrid Physics-Driven Neural Network Force Field for Liquid Electrolytes. *J. Comp. Theo. Comput.* **22,** 3011–3022 (6 2026).

206. Cheng, B. Latent Ewald summation for machine learning of long-range interactions. *npj Comp. Mat.* **11,** 80 (2025).

207. Martín Abadi *et al. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems* Software available from tensorflow.org. 2015. `https://www.tensorflow.org/`.

208. Pelaez, R. P. *et al.* Torchmd-net 2.0: Fast neural network potentials for molecular simulations. *J. Comp. Theo. Comput.* **20,** 4076–4087 (2024).

209. Schoenholz, S. S. & Cubuk, E. D. JAX, MD A framework for differentiable physics. *J. Stat. Mech. Theory Exp.* **2021,** 124016 (2021).

210. Christiansen, H. *et al.* Fast, modular, and differentiable framework for machine learning-enhanced molecular simulations. *The Journal of Chemical Physics* **163** (2025).

211. Tillet, P., Kung, H.-T. & Cox, D. *Triton: an intermediate language and compiler for tiled neural network computations* in *Proceedings of the 3rd ACM SIGPLAN International Workshop on Machine Learning and Programming Languages* (2019), 10–19.

212. Dao, T., Fu, D., Ermon, S., Rudra, A. & Ré, C. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Adv Neural Inf Process Syst* **35,** 16344–16359 (2022).

213. Dao, T. Flashattention-2: Faster attention with better parallelism and work partitioning. *arXiv preprint arXiv:2307.08691* (2023).

214. Shah, J. *et al.* Flashattention-3: Fast and accurate attention with asynchrony and low-precision. *Adv Neural Inf Process Syst* **37,** 68658–68685 (2024).

215. *PyG* `https://github.com/pyg-team/pytorch_geometric`. accessed on Apr 10, 2026.

216. *FlashAttention* `https://github.com/dao-ailab/flash-attention`. accessed on Apr 10, 2026.

217. Hjorth Larsen, A. *et al.* The atomic simulation environment—a Python library for working with atoms. *Journal of Physics: Condensed Matter* **29,** 273002 (2017).

218. Eastman, P. *et al.* OpenMM 8: molecular dynamics simulation with machine learning potentials. *J. Phys. Chem. B* **128,** 109–116 (2023).

219. Das, A. K. *et al.* Development of an advanced force field for water using variational energy decomposition analysis. *J. Comp. Theo. Comput.* **15,** 5001–5013 (2019).

220. Tang, K. & Toennies, J. P. An improved simple model for the van der Waals potential based on universal damping functions for the dispersion coefficients. *J. Chem. Phys.* **80,** 3726–3741 (1984).

221. Demerdash, O., Yap, E. H. & Head-Gordon, T. Advanced potential energy surfaces for condensed phase simulation. *Annu Rev Phys Chem* **65,** 149–74. ISSN: 1545-1593 (Electronic) 0066-426X (Linking). `https://www.ncbi.nlm.nih.gov/pubmed/24328448` (2014).

222. Zhu, X., Riera, M., Bull-Vulpe, E. F. & Paesani, F. MB-pol (2023): Sub-chemical accuracy for water simulations from the gas to the liquid phase. *J. Comp. Theo. Comput.* **19,** 3551–3566 (2023).

223. Ewald, P. P. Evaluation of optical and electrostatic lattice potentials. *Ann. Phys.* **64,** 253–268 (1921).

224. De Leeuw, S. W., Perram, J. W. & Smith, E. R. Simulation of electrostatic systems in periodic boundary conditions. I. Lattice sums and dielectric constants. *Proc. R. Soc. Lond. A Math. Phys. Sci.* **373,** 27–56 (1980).

225. Perram, J. W., Petersen, H. G. & de Leeuw, S. W. An Algorithm for the Simulation of Condensed Matter Which Grows as the 3/2 Power of the Number of Particles. *Mol. Phys.* **65,** 875–893 (4 1988).

226. Essmann, U. *et al.* A smooth particle mesh Ewald method. *J. Chem. Phys.* **103,** 8577–8593 (1995).

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

228. Aguado, A. & Madden, P. A. Ewald summation of electrostatic multipole interactions up to the quadrupolar level. *J. Chem. Phys.* **119,** 7471–7483 (2003).

229. Sagui, C., Pedersen, L. G. & Darden, T. A. Towards an accurate representation of electrostatics in classical force fields: Efficient implementation of multipolar interactions in biomolecular simulations. *J. Chem. Phys.* **120,** 73–87 (2004).

230. Giese, T. J., Panteva, M. T., Chen, H. & York, D. M. Multipolar Ewald methods, 1: Theory, accuracy, and performance. *J. Comp. Theo. Comput.* **11,** 436–450 (2015).

231. Symons, B. C. & Popelier, P. L. Flexible multipole moments in smooth particle mesh Ewald. *J. Chem. Phys.* **156** (2022).

232. Eastman, P. *et al.* OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. *PLOS Comp. Bio.* **13,** e1005659 (2017).

233. Loche, P. *et al.* Fast and flexible long-range models for atomistic machine learning. *J. Chem. Phys.* **162** (2025).

234. Bigi, F. *et al.* Metatensor and metatomic: foundational libraries for interoperable atomistic machine learning. *J. Chem. Phys.* **164** (2026).

235. *NVIDIA ALCHEMI Toolkit-Ops* `https://github.com/NVIDIA/nvalchemi-toolkit-ops`. accessed on Apr 10, 2026, v0.2.0.

236. *Torch-Sim* `https://github.com/TorchSim/torch-sim`. accessed on Apr 10, 2026, v0.2.0.

237. Gupta, A., Tadmor, E. B. & Martiniani, S. *KUSP: Python server for deploying ML interatomic potentials* in *AI for Accelerated Materials Design-Vienna 2024* (2024).

238. Kovács, D. P. *et al.* Mace-off: Short-range transferable machine learning force fields for organic molecules. *Journal of the American Chemical Society* **147,** 17598–17611 (2025).

239. Wood, B. M. *et al.* UMA: A Family of Universal Models for Atoms. *arXiv preprint arXiv:2506.23971* (2025).

240. Anstine, D. M., Zubatyuk, R. & Isayev, O. AIMNet2: a neural network potential to meet your neutral, charged, organic, and elemental-organic needs. *Chemical Science* **16,** 10228–10244 (2025).

241. Haghighatlari, M. *et al.* NewtonNet: a Newtonian message passing network for deep learning of interatomic potentials and forces. *Digital Discovery* **1,** 333–343 (2022).

242. Nerenberg, P. S. & Head-Gordon, T. New developments in force fields for biomolecular simulations. *Current Opinion in Structural Biology* **49,** 129–138. ISSN: 0959-440X. `https://www.sciencedirect.com/science/article/pii/S0959440X17301604` (2018).

243. Tian, C. *et al.* ff19SB: Amino-Acid-Specific Protein Backbone Parameters Trained against Quantum Mechanics Energy Surfaces in Solution. en-US. *J. Chem. Theory Comput.* **16,** 528–552. ISSN: 1549-9618. https://doi.org/10.1021/acs.jctc.9b00591 (2025) (Jan. 2020).

244. Chang, J. *et al.* Efficient and Precise Force Field Optimization for Biomolecules Using DPA-3. *Journal of Chemical Information and Modeling* (2026).

245. Heindel, J. P., Kim, L., Head-Gordon, M. & Head-Gordon, T. Completely Multipolar Model for Many-Body Water–Ion and Ion–Ion Interactions. *J. Phys. Chem. Lett.* **16,** 975–984. https://doi.org/10.1021/acs.jpclett.4c02940 (2026) (Jan. 2025).

246. Kim, L. & Head-Gordon, T. Near equivalence of polarizability and bond order flux metrics for describing covalent bond rearrangements. *Physical Chemistry Chemical Physics* **26,** 27459–27465. ISSN: 1463-9076. https://doi.org/10.1039/d4cp03337j (2024).

247. Rackers, J. A. *et al.* An optimized charge penetration model for use with the AMOEBA force field. en. *Phys. Chem. Chem. Phys.* **19,** 276–291. ISSN: 1463-9084. https://pubs.rsc.org/en/content/articlelanding/2017/cp/c6cp06017j (2025) (Dec. 2016).

248. Freitag, M., Gordon, M., Jensen, J. & Stevens, W. Evaluation of Charge Penetration Between Distributed Multipolar Expansions. en (2000-05-01 00:05:00 2000).

249. Qiu, Y., Smith, D., Stern, C., Feng, M. & Wang, L.-P. Driving Torsion Scans with Wavefront Propagation. *ChemRxiv* **2020.** eprint: https://chemrxiv.org/doi/pdf/10.26434/chemrxiv.12044673.v1. https://chemrxiv.org/doi/abs/10.26434/chemrxiv.12044673.v1 (2020).

250. Maier, J. A. *et al.* ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J. Chem. Theory Comput.* **11,** 3696–3713. ISSN: 1549-9618. https://doi.org/10.1021/acs.jctc.5b00255 (2026) (Aug. 2015).

251. Cornell, W. D. *et al.* A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *J. Am. Chem. Soc.* **117,** 5179–5197. ISSN: 0002-7863. https://doi.org/10.1021/ja00124a002 (2026) (May 1995).

252. Jorgensen, W. L. & Tirado-Rives, J. The OPLS [optimized potentials for liquid simulations] potential functions for proteins, energy minimizations for crystals of cyclic peptides and crambin. *J. Am. Chem. Soc.* **110,** 1657–1666. ISSN: 0002-7863. https://doi.org/10.1021/ja00214a001 (2026) (Mar. 1988).

253. Vanommeslaeghe, K. *et al.* CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields. en. *Journal of Computational Chemistry* **31.** _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/jcc.21367, 671–690. ISSN: 1096-987X. https://onlinelibrary.wiley.com/doi/abs/10.1002/jcc.21367 (2026) (2010).

254. Boothroyd, S. *et al.* Development and benchmarking of open force field 2.0. 0: the sage small molecule force field. *Journal of chemical theory and computation* **19,** 3251–3275 (2023).

255. Demerdash, O., Wang, L.-P. & Head-Gordon, T. Advanced models for water simulations. *WIREs Computational Molecular Science* **8,** e1355. eprint: https://wires.onlinelibrary.wiley.com/doi/pdf/10.1002/wcms.1355. https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/wcms.1355 (2018).

256. Demerdash, O., Yap, E.-H. & Head-Gordon, T. Advanced Potential Energy Surfaces for Condensed Phase Simulation. *Annual Review of Physical Chemistry* **65,** 149–174. ISSN: 1545-1593. https://www.annualreviews.org/content/journals/10.1146/annurev-physchem-040412-110040 (2014).

257. Shi, Y. *et al.* Polarizable Atomic Multipole-Based AMOEBA Force Field for Proteins. en-US. *J. Chem. Theory Comput.* **9,** 4046–4063. ISSN: 1549-9618. https://doi.org/10.1021/ct4003702 (2026) (Sept. 2013).

258. Ren, P., Wu, C. & Ponder, J. W. Polarizable Atomic Multipole-Based Molecular Mechanics for Organic Molecules. *J. Chem. Theory Comput.* **7,** 3143–3161. ISSN: 1549-9618. https://doi.org/10.1021/ct200304d (2025) (Oct. 2011).

259. Rackers, J. A., Silva, R. R., Wang, Z. & Ponder, J. W. Polarizable Water Potential Derived from a Model Electron Density. en-US. *J. Chem. Theory Comput.* **17,** 7056–7084. ISSN: 1549-9618. https://doi.org/10.1021/acs.jctc.1c00628 (2025) (Nov. 2021).

260. Liu, C., Piquemal, J.-P. & Ren, P. AMOEBA+ Classical Potential for Modeling Molecular Interactions. *J. Chem. Theory Comput.* **15,** 4122–4139. ISSN: 1549-9618. https://doi.org/10.1021/acs.jctc.9b00261 (2025) (July 2019).

261. Das, A. K., Liu, M. & Head-Gordon, T. Development of a Many-Body Force Field for Aqueous Alkali Metal and Halogen Ions: An Energy Decomposition Analysis Guided Approach. *J. Chem. Theory Comput.* **18,** 953–967. ISSN: 1549-9618. https://doi.org/10.1021/acs.jctc.1c00537 (2026) (Feb. 2022).

262. Das, A. K. *et al.* Development of an Advanced Force Field for Water Using Variational Energy Decomposition Analysis. *J. Chem. Theory Comput.* **15,** 5001–5013. ISSN: 1549-9618. https://doi.org/10.1021/acs.jctc.9b00478 (2025) (Sept. 2019).

263. Heindel, J. P., Sami, S. & Head-Gordon, T. Completely Multipolar Model as a General Framework for Many-Body Interactions as Illustrated for Water. *J. Chem. Theory Comput.* **20,** 8594–8608. ISSN: 1549-9618. https://doi.org/10.1021/acs.jctc.4c00812 (2025) (Oct. 2024).

264. Sami, S., LaCour, R. A., Heindel, J. P. & Head-Gordon, T. Simple and Accurate One-Body Energy and Dipole Moment Surfaces for Water and Beyond. *J. Phys. Chem. Lett.* **15,** 6712–6721. https://doi.org/10.1021/acs.jpclett.4c00587 (2025) (July 2024).

265. Allinger, N. L., Yuh, Y. H. & Lii, J. H. Molecular mechanics. The MM3 force field for hydrocarbons. 1. *J. Am. Chem. Soc.* **111,** 8551–8566 (1989).

266. Lii, J. H. & Allinger, N. L. Molecular mechanics. The MM3 force field for hydrocarbons. 2. Vibrational frequencies and thermodynamics. *J. Am. Chem. Soc.* **111,** 8566–8575 (1989).

267. Halgren, T. A. Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94. *J. Comput. Chem.* **17,** 490–519. https://onlinelibrary.wiley.com/doi/abs/10.1002/%28SICI%291096-987X%28199604%2917%3A5/6%3C490%3A%3AAID-JCC1%3E3.0.CO%3B2-P (1996).

268. Halgren, T. A. MMFF VII. Characterization of MMFF94, MMFF94s, and other widely available force fields for conformational energies and for intermolecular-interaction energies and geometries. *J. Comput. Chem.* **20,** 730–748 (1999).

269. Hwang, M. J., Stockfisch, T. P. & Hagler, A. T. Derivation of Class II Force Fields. 2. Derivation and Characterization of a Class II Force Field, CFF93, for the Alkyl Functional Group and Alkane Molecules. *J. Am. Chem. Soc.* **116,** 2515–2525 (1994).

270. Maple, J. R. *et al.* Derivation of class II force fields. I. Methodology and quantum force field for the alkyl functional group and alkane molecules. *J. Comput. Chem.* **15,** 162–182. https://onlinelibrary.wiley.com/doi/abs/10.1002/jcc.540150207 (1994).

271. Rackers, J. A. *et al.* Tinker 8: software tools for molecular design. *Journal of chemical theory and computation* **14,** 5273–5289 (2018).

272. Smith, D. G. *et al.* PSI4 1.4: Open-source software for high-throughput quantum chemistry. *The Journal of chemical physics* **152** (2020).

273. Bannwarth, C., Ehlert, S. & Grimme, S. GFN2-xTB—An Accurate and Broadly Parametrized Self-Consistent Tight-Binding Quantum Chemical Method with Multipole Electrostatics and Density-Dependent Dispersion Contributions. *J. Chem. Theory Comput.* **15,** 1652–1671. ISSN: 1549-9618. https://doi.org/10.1021/acs.jctc.8b01176 (2026) (Mar. 2019).

274. Pracht, P. *et al.* CREST—A program for the exploration of low-energy molecular chemical space. *J. Chem. Phys.* **160,** 114110. ISSN: 0021-9606. https://doi.org/10.1063/5.0197592 (2026) (Mar. 2024).

275. Qiu, Y. *et al.* Driving torsion scans with wavefront propagation. *J. Chem. Phys.* **152,** 244116. ISSN: 0021-9606. https://doi.org/10.1063/5.0009232 (2026) (June 2020).

276. Marenich, A. V., Cramer, C. J. & Truhlar, D. G. Universal solvation model based on solute electron density and on a continuum model of the solvent defined by the bulk dielectric constant and atomic surface tensions. *The Journal of Physical Chemistry B* **113,** 6378–6396 (2009).

277. Epifanovsky, E. *et al.* Software for the frontiers of quantum chemistry: An overview of developments in the Q-Chem 5 package. *J. Chem. Phys.* **155,** 084801. ISSN: 0021-9606. https://doi.org/10.1063/5.0055522 (2026) (Aug. 2021).

278. Mardirossian, N. & Head-Gordon, M. *omega*B97X-V: A 10-parameter, range-separated hybrid, generalized gradient approximation density functional with nonlocal correlation, designed by a survival-of-the-fittest strategy. *Phys. Chem. Chem. Phys.* **16,** 9904–9924. ISSN: 1463-9076. https://doi.org/10.1039/c3cp54374a (2026) (June 2014).

279. Wang, Y., Shin, D. J., Head-Gordon, M. & Head-Gordon, T. Energetics of Noncovalent Interactions of Protein–Ligand Complexes for Drug Discovery. *J. Chem. Inf. Model.* ISSN: 1549-9596. https://doi.org/10.1021/acs.jcim.6c01291 (2026) (June 2026).

280. Corrigan, R. A. *et al.* A generalized Kirkwood implicit solvent for the polarizable AMOEBA protein model. *The Journal of chemical physics* **159** (2023).

281. Eastman, P. *et al.* OpenMM 8: Molecular Dynamics Simulation with Machine Learning Potentials. *J. Phys. Chem. B* **128,** 109–116. ISSN: 1520-6106. https://doi.org/10.1021/acs.jpcb.3c06662 (2025) (Jan. 2024).

282. Hu, J.-S. & Bax, A. Determination of ϕ and χ_1 Angles in Proteins from ^{13}C–^{13}C Three-Bond J Couplings Measured by Three-Dimensional Heteronuclear NMR. How Planar Is the Peptide Bond? *J. Am. Chem. Soc.* **119,** 6360–6368. ISSN: 0002-7863. https://doi.org/10.1021/ja970067v (2026) (July 1997).

283. Karplus, M. Contact electron-spin coupling of nuclear magnetic moments. *The Journal of chemical physics* **30,** 11–15 (1959).

284. Avbelj, F., Golic Grdadolnik, S., Grdadolnik, J. & Baldwin, R. L. Intrinsic backbone preferences are fully present in blocked amino acids. *Proc. Natl. Acad. Sci. U.S.A.* **103,** 1272–1277. https://doi.org/10.1073/pnas.0510420103 (2006).

285. Best, R. B., Buchete, N.-V. & Hummer, G. Are Current Molecular Dynamics Force Fields too Helical? English. *Biophysical Journal* **95,** L07–L09. ISSN: 0006-3495, 1542-0086. https://www.cell.com/biophysj/abstract/S0006-3495(08)70277-7 (2026) (July 2008).

286. Wickstrom, L., Okur, A. & Simmerling, C. Evaluating the performance of the ff99SB force field based on NMR scalar coupling data. *Biophysical journal* **97,** 853–856 (2009).

287. Shang, H. S. & Head-Gordon, T. Stabilization of Helixes in Glycine and Alanine Dipeptides in a Reaction Field Model of Solvent. *J. Am. Chem. Soc.* **116,** 1528–1532. ISSN: 0002-7863. https://doi.org/10.1021/ja00083a042 (2026) (Feb. 1994).

288. Wang, Y. *et al.* End-to-end differentiable construction of molecular mechanics force fields. *Chem. Sci.* **13,** 12016–12033. ISSN: 2041-6520. https://doi.org/10.1039/d2sc02739a (2026) (Nov. 2022).

289. Zheng, T. *et al.* Bridging quantum mechanics to liquid properties via a universal organic force field. en. *Nat Commun.* ISSN: 2041-1723. https://www.nature.com/articles/s41467-026-73566-3 (2026) (May 2026).

290. Zheng, T. *et al.* Data-driven parametrization of molecular mechanics force fields for expansive chemical space coverage. *Chem. Sci.* **16,** 2730–2740. ISSN: 2041-6520. https://doi.org/10.1039/d4sc06640e (2026) (Feb. 2025).

291. Case, D. A., Scheurer, C. & Brüschweiler, R. Static and Dynamic Effects on Vicinal Scalar J Couplings in Proteins and Peptides: A MD/DFT Analysis. *J. Am. Chem. Soc.* **122,** 10390–10397. ISSN: 0002-7863. https://doi.org/10.1021/ja001798p (2026) (Oct. 2000).

292. Schmidt, J. M., Blümel, M., Löhr, F. & Rüterjans, H. Self-consistent ^3J coupling analysis for the joint calibration of Karplus coefficients and evaluation of torsion angles. *Journal of Biomolecular NMR* **14,** 1–12. https://doi.org/10.1023/A:1008345303942 (1999).

293. Brüschweiler, R. & Case, D. Adding harmonic motion to the Karplus relation for spin-spin coupling. *Journal of the American Chemical Society* **116,** 11199–11200. https://doi.org/10.1021/ja00103a062 (1994).

294. Aliev, A. E. & Courtier-Murias, D. Experimental Verification of Force Fields for Molecular Dynamics Simulations Using Gly-Pro-Gly-Gly. *J. Phys. Chem. B* **114,** 12358–12375. ISSN: 1520-6106. https://doi.org/10.1021/jp101581h (2026) (Sept. 2010).

295. Qian, R., Xue, J., Xu, Y. & Huang, J. Alchemical Transformations and Beyond: Recent Advances and Real-World Applications of Free Energy Calculations in Drug Discovery. *Journal of Chemical Information and Modeling* **64,** 7214–7237. ISSN: 1549-9596. https://doi.org/10.1021/acs.jcim.4c01024 (2024) (Oct. 2024).

296. Ross, G. A. *et al.* The maximal and current accuracy of rigorous protein-ligand binding free energy calculations. en. *Communications Chemistry* **6,** 1–12. ISSN: 2399-3669. https://www.nature.com/articles/s42004-023-01019-9 (2025) (Oct. 2023).

297. Liu, S. *et al.* Lead optimization mapper: automating free energy calculations for lead optimization. en. *Journal of Computer-Aided Molecular Design* **27,** 755–770. ISSN: 1573-4951. https://doi.org/10.1007/s10822-013-9678-y (2024) (Sept. 2013).

298. Ries, B. *et al.* Kartograf: A Geometrically Accurate Atom Mapper for Hybrid-Topology Relative Free Energy Calculations. *Journal of Chemical Theory and Computation* **20,** 1862–1877. ISSN: 1549-9618. https://doi.org/10.1021/acs.jctc.3c01206 (2024) (Mar. 2024).

299. Alibay, I., Magarkar, A., Seeliger, D. & Biggin, P. C. Evaluating the use of absolute binding free energy in the fragment optimisation process. en. *Communications Chemistry* **5,** 105. ISSN: 2399-3669. https://www.nature.com/articles/s42004-022-00721-4 (2026) (Sept. 2022).

300. Wu, Z., Konig, G., Boresch, S. & Cossins, B. P. Optimizing Absolute Binding Free Energy Calculations for Production Usage. *Journal of Chemical Theory and Computation* **21,** 8330–8340. ISSN: 1549-9618. https://doi.org/10.1021/acs.jctc.5c00861 (2026) (Sept. 2025).

301. Carvalho Martins, L., Cino, E. A. & Ferreira, R. S. PyAutoFEP: An Automated Free Energy Perturbation Workflow for GROMACS Integrating Enhanced Sampling Methods. *Journal of Chemical Theory and Computation* **17,** 4262–4273. ISSN: 1549-9618. https://doi.org/10.1021/acs.jctc.1c00194 (2026) (July 2021).

302. Alencar Araripe, D. *et al.* Doing More with Less: Accurate and Scalable Ligand Free Energy Calculations by Focusing on the Binding Site. *Journal of Chemical Information and Modeling* **66,** 3164–3172. ISSN: 1549-9596. https://doi.org/10.1021/acs.jcim.5c02932 (2026) (Mar. 2026).

303. Liu, R. *et al.* ALCHEMD: Bridging Accessibility and Accuracy in Automated Relative Binding Free Energy Workflows. *Journal of Chemical Theory and Computation* **22,** 750–762. ISSN: 1549-9618. https://doi.org/10.1021/acs.jctc.5c01857 (2026) (Jan. 2026).

304. Heinzelmann, G., Huggins, D. J. & Gilson, M. K. BAT2: an Open-Source Tool for Flexible, Automated, and Low Cost Absolute Binding Free Energy Calculations. *Journal of Chemical Theory and Computation* **20,** 6518–6530. ISSN: 1549-9618. https://doi.org/10.1021/acs.jctc.4c00205 (2026) (Aug. 2024).

305. Li, P., Pu, T. & Mei, Y. FEP-SPell-ABFE: An Open-Source Automated Alchemical Absolute Binding Free-Energy Calculation Workflow for Drug Discovery. *Journal of Chemical Information and Modeling* **65,** 2711–2721. ISSN: 1549-9596. https://doi.org/10.1021/acs.jcim.4c01986 (2026) (Mar. 2025).

306. León, A. M., Andersen, L. & Hub, J. S. BindFlow: A Free, User-Friendly Pipeline for Absolute Binding Free Energy Calculations Using Free Energy Perturbation or MM(PB/GB)SA. *Journal of Chemical Theory and Computation* **22,** 1198–1213. ISSN: 1549-9618. https://doi.org/10.1021/acs.jctc.5c02026 (2026) (Jan. 2026).

307. Lee, T.-S. *et al.* Improved Alchemical Free Energy Calculations with Optimized Smoothstep Softcore Potentials. *Journal of Chemical Theory and Computation* **16,** 5512–5525. ISSN: 1549-9618. https://doi.org/10.1021/acs.jctc.0c00237 (2024) (Sept. 2020).

308. York, D. M. Modern Alchemical Free Energy Methods for Drug Discovery Explained. *ACS Physical Chemistry Au* **3,** 478–491. https://doi.org/10.1021/acsphyschemau.3c00033 (2026) (Nov. 2023).

309. Bennett, C. H. Efficient estimation of free energy differences from Monte Carlo data. *Journal of Computational Physics* **22,** 245–268. ISSN: 0021-9991. https://www.sciencedirect.com/science/article/pii/0021999176900784 (2026) (Oct. 1976).

310. Shirts, M. R. & Chodera, J. D. Statistically optimal analysis of samples from multiple equilibrium states. en. *The Journal of Chemical Physics* **129,** 124105. ISSN: 0021-9606, 1089-7690. https://pubs.aip.org/jcp/article/129/12/124105/957527/Statistically-optimal-analysis-of-samples-from (2026) (Sept. 2008).

311. *PDBFixer* https://github.com/openmm/pdbfixer. accessed on Oct 29, 2024.

312. Wang, J., Wolf, R. M., Caldwell, J. W., Kollman, P. A. & Case, D. A. Development and testing of a general amber force field. *Journal of computational chemistry* **25,** 1157–1174 (2004).

313. Qiu, Y. *et al.* Development and Benchmarking of Open Force Field v1.0.0-the Parsley Small-Molecule Force Field. *J Chem Theory Comput* **17,** 6262–6280. ISSN: 1549-9618 (Print) 1549-9618 (2021).

314. Jakalian, A., Bush, B. L., Jack, D. B. & Bayly, C. I. Fast, efficient generation of high-quality atomic charges. AM1-BCC model: I. Method. *Journal of computational chemistry* **21,** 132–146 (2000).

315. Jakalian, A., Jack, D. B. & Bayly, C. I. Fast, efficient generation of high-quality atomic charges. AM1-BCC model: II. Parameterization and validation. *Journal of computational chemistry* **23,** 1623–1641 (2002).

316. Amber Molecular Dynamics. *Download Amber MD* https://ambermd.org/GetAmber.php (2026).

317. Lee, T.-S., Tsai, H.-C., Ganguly, A. & York, D. M. ACES: Optimized Alchemically Enhanced Sampling. *Journal of Chemical Theory and Computation* **19,** 472–487. ISSN: 1549-9618. https://doi.org/10.1021/acs.jctc.2c00697 (2024) (Jan. 2023).

318. Tsai, H.-C. *et al.* Improvements in Precision of Relative Binding Free Energy Calculations Afforded by the Alchemical Enhanced Sampling (ACES) Approach. *Journal of Chemical Information and Modeling* **64,** 7046–7055. ISSN: 1549-9596. https://doi.org/10.1021/acs.jcim.4c00464 (2024) (Sept. 2024).

319. Xu, H. Optimal Measurement Network of Pairwise Differences. *Journal of Chemical Information and Modeling* **59,** 4720–4728. ISSN: 1549-9596. https://doi.org/10.1021/acs.jcim.9b00528 (2024) (Nov. 2019).

320. Landrum, G. *et al.* RDKit: A software suite for cheminformatics, computational chemistry, and predictive modeling. *Greg Landrum* **8** (2013).

321. Chen, W. *et al.* Enhancing Hit Discovery in Virtual Screening through Absolute Protein–Ligand Binding Free-Energy Calculations. *Journal of Chemical Information and Modeling* **63,** 3171–3185. ISSN: 1549-9596. https://doi.org/10.1021/acs.jcim.3c00013 (2026) (May 2023).

322. Midwest Antiviral Drug Discovery (AViDD) Center. *Project 3: Pandemic Virus Protease Inhibitors* https://midwestavidd.umn.edu/research/project-3-pandemic-virus-protease-inhibitors (2026).

323. Jimmidi, R. *et al.* DNA-encoded chemical libraries yield non-covalent and non-peptidic SARS-CoV-2 main protease inhibitors. *Communications Chemistry* **6,** 164. ISSN: 2399-3669. https://doi.org/10.1038/s42004-023-00961-y (2026) (Aug. 2023).

324. CACHE. *Application for Challenge #7: Finding Selective Inhibitors of Phosphoglycerate Kinase 2* https://cache-challenge.org/application/2501 (2026).

325. Phillips, N. The coronavirus is here to stay — here's what that means. en. *Nature* **590,** 382–384. ISSN: 0028-0836, 1476-4687. https://www.nature.com/articles/d41586-021-00396-2 (2025) (Feb. 2021).

326. Von Delft, F. *et al.* A white-knuckle ride of open COVID drug discovery. en. *Nature* **594,** 330–332. ISSN: 0028-0836, 1476-4687. https://www.nature.com/articles/d41586-021-01571-1 (2025) (June 2021).

327. Chodera, J., Lee, A. A., London, N. & Von Delft, F. Crowdsourcing drug discovery for pandemics. en. *Nature Chemistry* **12,** 581–581. ISSN: 1755-4330, 1755-4349. https://www.nature.com/articles/s41557-020-0496-2 (2025) (July 2020).

328. Zhao, T. *et al.* Warnings regarding the potential coronavirus disease 2019 (COVID-19) transmission risk: Vaccination is not enough. en. *Infection Control & Hospital Epidemiology* **43,** 679–680. ISSN: 0899-823X, 1559-6834. https://www.cambridge.org/core/product/identifier/S0899823X21000635/type/journal_article (2025) (May 2022).

329. Ullrich, S. & Nitsche, C. The SARS-CoV-2 main protease as drug target. en. *Bioorganic & Medicinal Chemistry Letters* **30,** 127377. ISSN: 0960894X. https://linkinghub.elsevier.com/retrieve/pii/S0960894X20304881 (2025) (Sept. 2020).

330. Douangamath, A. *et al.* Crystallographic and electrophilic fragment screening of the SARS-CoV-2 main protease. en. *Nature Communications* **11,** 5047. ISSN: 2041-1723. https://www.nature.com/articles/s41467-020-18709-w (2025) (Oct. 2020).

331. Zhang, L. *et al.* Crystal structure of SARS-CoV-2 main protease provides a basis for design of improved α-ketoamide inhibitors. en. *Science* **368,** 409–412. ISSN: 0036-8075, 1095-9203. https://www.science.org/doi/10.1126/science.abb3405 (2025) (Apr. 2020).

332. Jin, Z. *et al.* Structure of Mpro from SARS-CoV-2 and discovery of its inhibitors. en. *Nature* **582,** 289–293. ISSN: 1476-4687. https://www.nature.com/articles/s41586-020-2223-y (2025) (June 2020).

333. Owen, D. R. *et al.* An oral SARS-CoV-2 M^{pro} inhibitor clinical candidate for the treatment of COVID-19. en. *Science* **374,** 1586–1593. ISSN: 0036-8075, 1095-9203. https://www.science.org/doi/10.1126/science.abl4784 (2025) (Dec. 2021).

334. Unoh, Y. *et al.* Discovery of S-217622, a Noncovalent Oral SARS-CoV-2 3CL Protease Inhibitor Clinical Candidate for Treating COVID-19. en. *Journal of Medicinal Chemistry* **65,** 6499–6512. ISSN: 0022-2623, 1520-4804. https://pubs.acs.org/doi/10.1021/acs.jmedchem.2c00117 (2025) (May 2022).

335. Lockbaum, G. J. *et al.* Crystal structure of SARS-CoV-2 main protease in complex with non-covalent inhibitor ML188. *Viruses* **13,** 174 (2021).

336. Su, H.-X. *et al.* Anti-SARS-CoV-2 activities in vitro of Shuanghuanglian preparations and bioactive ingredients. *Acta Pharm. Sin.* **41,** 1167–1177 (2020).

337. Günther, S. *et al.* X-ray screening identifies active site and allosteric inhibitors of SARS-CoV-2 MPro. *Science* **372,** 642–646 (2021).

338. Drayman, N. *et al.* Masitinib is a broad coronavirus 3CL inhibitor that blocks replication of SARS-CoV-2. *Science* **373,** 931–936 (2021).

339. Kitamura, N. *et al.* Expedited approach toward rational design of noncovalent SARS-CoV-2 MPro inhibitors. *J. Med. Chem.* **65,** 2848–2865 (2021).

340. Zhang, C.-H. *et al.* Potent noncovalent inhibitors of the main protease of SARS-CoV-2 from molecular sculpting of the drug perampanel guided by free energy perturbation calculations. *ACS central science* **7,** 467–475 (2021).

341. Glaser, J. *et al.* Hit expansion of a noncovalent SARS-CoV-2 main protease inhibitor. *ACS Pharm. Trans. Sci.* **5,** 255–265 (2022).

342. Deshmukh, M. G. *et al.* Structure-guided design of perampanel-derived pharmacophore targeting SARS-CoV-2 main protease. *Structure* **29,** 823–833 (2021).

343. Hou, N. *et al.* Development of highly potent noncovalent inhibitors of SARS-CoV-2 3CLpro. *ACS central science* **9,** 217–227 (2023).

344. Kneller, D. W. *et al.* Structural, electronic, and electrostatic determinants for inhibitor binding to subsites S1 and S2 in SARS-CoV-2 main protease. *J. Med. Chem.* **64,** 17366–17383 (2021).

345. Gao, S. *et al.* Discovery and crystallographic studies of trisubstituted piperazine derivatives as non-covalent SARS-CoV-2 MPro inhibitors with high target specificity and low toxicity. *J. Med. Chem.* **65,** 13343–13364 (2022).

346. Han, S. H. *et al.* Structure-based optimization of ML300-derived, noncovalent inhibitors targeting the severe acute respiratory syndrome coronavirus 3CL protease (SARS-CoV-2 3CLpro). *J. Med. Chem.* **65,** 2880–2904 (2021).

347. Li, J. *et al.* Mining for Potent Inhibitors through Artificial Intelligence and Physics: A Unified Methodology for Ligand Based and Structure Based Drug Design. en. *Journal of Chemical Information and Modeling* **64,** 9082–9097. ISSN: 1549-9596, 1549-960X. https://pubs.acs.org/doi/10.1021/acs.jcim.4c00634 (2025) (Dec. 2024).

348. Lin, M. *et al.* Molecular mechanism of ensitrelvir inhibiting SARS-CoV-2 main protease and its variants. en. *Communications Biology* **6,** 1–7. ISSN: 2399-3642. https://www.nature.com/articles/s42003-023-05071-y (2025) (July 2023).

349. Boby, M. L. *et al.* Open science discovery of potent noncovalent SARS-CoV-2 main protease inhibitors. en. *Science* **382,** eabo7201. ISSN: 0036-8075, 1095-9203. https://www.science.org/doi/10.1126/science.abo7201 (2025) (Nov. 2023).

350. Heilmann, E. *et al.* SARS-CoV-2 3CL^{pro} mutations selected in a VSV-based system confer resistance to nirmatrelvir, ensitrelvir, and GC376. en. *Science Translational Medicine* **15,** eabq7360. ISSN: 1946-6234, 1946-6242. https://www.science.org/doi/10.1126/scitranslmed.abq7360 (2025) (Jan. 2023).

351. Moghadasi, S. A. *et al.* Transmissible SARS-CoV-2 variants with resistance to clinical protease inhibitors. en. *Science Advances* **9,** eade8778. ISSN: 2375-2548. https://www.science.org/doi/10.1126/sciadv.ade8778 (2025) (Mar. 2023).

352. Lin, C. *et al.* Structural basis for the inhibition of coronaviral main proteases by ensitrelvir. en. *Structure* **31,** 1016–1024.e3. ISSN: 09692126. https://linkinghub.elsevier.com/retrieve/pii/S0969212623002071 (2025) (Sept. 2023).

353. Polaris. *Polaris* en. https://polarishub.io (2025).

354. newpixcom. *AI-driven Structure-enabled Antiviral Platform (ASAP)* en-us. https://asapdiscovery.org/ (2025).

355. OpenADMET. *OpenADMET* en-us. https://openadmet.org/ (2025).

356. Eberhardt, J., Santos-Martins, D., Tillack, A. F. & Forli, S. AutoDock Vina 1.2.0: New Docking Methods, Expanded Force Field, and Python Bindings. *J. Chem. Info. Model.* **61,** 3891–3898. ISSN: 1549-9596. https://doi.org/10.1021/acs.jcim.1c00203 (2021).

357. Morris, G. M. *et al.* AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *J. Comp. Chem.* **30,** 2785–2791 (2009).

358. Friesner, R. A. *et al.* Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *J. Med. Chem.* **47,** 1739–1749 (2004).

359. Jones, G., Willett, P., Glen, R. C., Leach, A. R. & Taylor, R. Development and validation of a genetic algorithm for flexible docking. *J. Mol. Bio.* **267,** 727–748 (1997).

360. Lang, P. T. *et al.* DOCK 6: Combining techniques to model RNA–small molecule complexes. en. *RNA* **15,** 1219–1230. ISSN: 1355-8382, 1469-9001. http://rnajournal.cshlp.org/lookup/doi/10.1261/rna.1563609 (2025) (June 2009).

361. Alekseenko, A. *et al.* ClusPro LigTBM: Automated Template-based Small Molecule Docking. *Journal of Molecular Biology. Computation Resources for Molecular Biology* **432,** 3404–3410. ISSN: 0022-2836. https://www.sciencedirect.com/science/article/pii/S0022283619307119 (2025) (May 2020).

362. Yang, X., Liu, Y., Gan, J., Xiao, Z.-X. & Cao, Y. FitDock: protein–ligand docking by template fitting. en. *Briefings in Bioinformatics* **23,** bbac087. ISSN: 1467-5463, 1477-4054. https://academic.oup.com/bib/article/doi/10.1093/bib/bbac087/6548375 (2025) (May 2022).

363. Pang, M. *et al.* CoDock-Ligand: combined template-based docking and CNN-based scoring in ligand binding prediction. *BMC Bioinformatics* **24,** 444. ISSN: 1471-2105. https://doi.org/10.1186/s12859-023-05571-y (2025) (Nov. 2023).

364. Wang, R., Fang, X., Lu, Y. & Wang, S. The PDBbind Database: Collection of Binding Affinities for Protein-Ligand Complexes with Known Three-Dimensional Structures. en. *Journal of Medicinal Chemistry* **47,** 2977–2980. ISSN: 0022-2623, 1520-4804. https://pubs.acs.org/doi/10.1021/jm0305801 (2024) (June 2004).

365. Li, J. *et al.* Leak Proof PDBBind: A Reorganized Dataset of Protein-Ligand Complexes for More Generalizable Binding Affinity Prediction. eng. *ArXiv,* arXiv:2308.09639v2. ISSN: 2331-8422 (May 2024).

366. Durairaj, J. *et al.* PLINDER: The protein-ligand interactions dataset and evaluation resource. *bioRxiv,* 2024–07 (2024).

367. Wang, Y. *et al.* A workflow to create a high-quality protein–ligand binding dataset for training, validation, and prediction tasks. en. *Digital Discovery* **4,** 1209–1220. https://pubs.rsc.org/en/content/articlelanding/2025/dd/d4dd00357h (2025) (2025).

368. Yu, Y., Lu, S., Gao, Z., Zheng, H. & Ke, G. Do deep learning models really outperform traditional approaches in molecular docking? *arXiv preprint arXiv:2302.07134* (2023).

369. Qiao, Z., Nie, W., Vahdat, A., Miller, T. F. & Anandkumar, A. State-specific protein–ligand complex structure prediction with a multiscale deep generative model. en. *Nature Machine Intelligence* **6,** 195–208. ISSN: 2522-5839. https://www.nature.com/articles/s42256-024-00792-z (2025) (Feb. 2024).

370. Ding, J. *et al.* Vina-GPU 2.0: Further Accelerating AutoDock Vina and Its Derivatives with Graphics Processing Units. en. *Journal of Chemical Information and Modeling* **63,** 1982–1998. ISSN: 1549-9596, 1549-960X. https://pubs.acs.org/doi/10.1021/acs.jcim.2c01504 (2025) (Apr. 2023).

371. Bhowmick, A. & Head-Gordon, T. A Monte Carlo Method for Generating Side Chain Structural Ensembles. *Structure* **23,** 44–55. ISSN: 0969-2126. https://www.sciencedirect.com/science/article/pii/S0969212614003566 (2025) (Jan. 2015).

372. Berman, H. M. *et al.* The Protein Data Bank. *Nucleic Acids Research* **28,** 235–242. ISSN: 0305-1048. https://doi.org/10.1093/nar/28.1.235 (2000).

373. Amaro, R. E. *et al.* Ensemble Docking in Drug Discovery. en. *Biophysical Journal* **114,** 2271–2278. ISSN: 00063495. https://linkinghub.elsevier.com/retrieve/pii/S0006349518303242 (2025) (May 2018).

374. Nguyen, D. D., Gao, K., Chen, J., Wang, R. & Wei, G.-W. Unveiling the molecular mechanism of SARS-CoV-2 main protease inhibition from 137 crystal structures using algebraic topology and deep learning. en. *Chemical Science* **11,** 12036–12046. ISSN: 2041-6539. https://pubs.rsc.org/en/content/articlelanding/2020/sc/d0sc04641h (2025) (Nov. 2020).

375. Flynn, J. M. *et al.* Comprehensive fitness landscape of SARS-CoV-2 Mpro reveals insights into viral resistance mechanisms. *eLife* **11** (eds Ogbunugafor, C. B., Barkai, N., Fraser, J. S. & Weinreich, D.) e77433. ISSN: 2050-084X. https://doi.org/10.7554/eLife.77433 (2025) (June 2022).

376. Mengist, H. M., Dilnessa, T. & Jin, T. Structural Basis of Potential Inhibitors Targeting SARS-CoV-2 Main Protease. English. *Frontiers in Chemistry* **9.** ISSN: 2296-2646. https://www.frontiersin.org/journals/chemistry/articles/10.3389/fchem.2021.622898/full (2025) (Mar. 2021).

377. Krismer, L. *et al.* Study of key residues in MERS-CoV and SARS-CoV-2 main proteases for resistance against clinically applied inhibitors nirmatrelvir and ensitrelvir. en. *npj Viruses* **2,** 1–20. ISSN: 2948-1767. https://www.nature.com/articles/s44298-024-00028-2 (2025) (June 2024).

378. Ester, M., Kriegel, H.-P., Sander, J. & Xu, X. *A density-based algorithm for discovering clusters in large spatial databases with noise* in *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD'96)* (1996), 226–231.

379. Onufriev, A., Bashford, D. & Case, D. A. Exploring protein native states and large-scale conformational changes with a modified generalized born model. *Proteins: Structure, Function, and Bioinformatics* **55,** 383–394 (2004).

380. Maier, J. A. *et al.* ff14SB: improving the accuracy of protein side chain and backbone parameters from ff99SB. *Journal of chemical theory and computation* **11,** 3696–3713 (2015).

381. Eastman, P. *et al.* OpenMM 8: Molecular Dynamics Simulation with Machine Learning Potentials. *The Journal of Physical Chemistry B* **128.** PMID: 38154096, 109–116. eprint: https://doi.org/10.1021/acs.jpcb.3c06662. https://doi.org/10.1021/acs.jpcb.3c06662 (2024).

382. Škrinjar, P., Eberhardt, J., Tauriello, G., Schwede, T. & Durairaj, J. *Have protein-ligand cofolding methods moved beyond memorisation?* en. Feb. 2025. http://biorxiv.org/lookup/doi/10.1101/2025.02.03.636309 (2025).