

UC Berkeley

UC Berkeley Electronic Theses and Dissertations

Title

Computer-Aided Drug Discovery: Generative Molecular Design with Large Language Models and Structure-Based Benchmarking

Permalink

<https://escholarship.org/uc/item/8rh736g8>

ISBN

9798189271540

Author

Sun, Kunyang

Publication Date

2026-05-01

Peer reviewed|Thesis/dissertation

Computer-Aided Drug Discovery: Generative Molecular Design with Large Language
Models and Structure-Based Benchmarking

by

Kunyang Sun

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 Eric Neuscamman

Professor Liana Lareau

Spring 2026

Computer-Aided Drug Discovery: Generative Molecular Design with Large Language
Models and Structure-Based Benchmarking

Copyright 2026
by
Kunyang Sun

Abstract

Computer-Aided Drug Discovery: Generative Molecular Design with Large Language Models and Structure-Based Benchmarking

by

Kunyang Sun

Doctor of Philosophy in Chemistry

University of California, Berkeley

Professor Teresa Head-Gordon, Chair

The discovery of new small-molecule therapeutics remains a notoriously time-consuming and capital-intensive endeavor, constrained by the vastness of synthesizable chemical space and the limitations of traditional screening approaches[1–3]. Generative artificial intelligence has fundamentally reshaped this landscape by enabling the direct sampling of molecules with desired properties, yet persistent challenges in synthesizability, data quality, and translational validation continue to limit the practical impact of computational methods[4–7]. This dissertation addresses these challenges across three interconnected pillars: Large Language Model (LLM)-based generative molecular design, rigorous structure-based benchmarking, and prospective validation in competitive drug discovery settings.

In Chapters 2 and 3, I develop two LLM-driven generative frameworks that address core bottlenecks in molecular design. SynLlama [8], described in Chapter 2, tackles the synthesizability problem by fine-tuning an open-weight LLM to generate synthetic pathways that decompose arbitrary molecules into commercially available building blocks assembled via known and robust chemical reactions. By constraining generation to this synthesizable chemical space, every proposed candidate comes with an actionable synthesis route, bridging the gap between *in silico* design and practical wet-lab execution. In Chapter 3, I introduce LinkLlama, a complementary framework that reframes linker design as a conditional generation task to solve a persistent bottleneck in fragment-based drug discovery. By leveraging natural language conditioning rather than explicit geometric training objectives, LinkLlama produces chemically reasonable linkers with high spatial fidelity across diverse scales, from simple fragment joining and scaffold hopping to the demanding design of linkers for proteolysis targeting chimera (PROTAC). Together, these models form a versatile LLM-based generative suite where natural language prompts enable flexible multi-objective molecular design with natural language guidance.

In Chapters 4 and 5, I establish reliable structural foundations for training and evaluating

computational drug discovery methods. Chapter 4 introduces HiQBind [9], a comprehensive open-source workflow that systematically identifies and corrects structural errors in protein-ligand complexes. By integrating tools such as OpenMM [10] and PDBFixer [11] with native annotations from the PDB, the pipeline produces over 30,000 high-quality, non-covalent protein-ligand structures with accurately assigned bond orders, adjusted protonation states, and refined protein structures, directly resolving long-standing community concerns regarding the reproducibility of proprietary curation methods. In Chapter 5, I introduce KinConfBench [12], a conformation-aware benchmark that evaluates cofolding models based on their ability to capture the conformational landscape of human kinases. KinConfBench reveals a critical tension in current models that, while they often achieve strong pocket-centric metrics, they frequently generate structures that adopt incorrect regulatory conformations for a given ligand, exposing a pervasive issue of apo-state memorization. By establishing a protein-centric benchmark beyond pocket and ligand geometries, KinConfBench provides an explicit, clinically relevant standard for structure prediction tools. Together, both works lay the groundwork for the community to move forward with more robust and physically-correct model training and evaluations.

In Chapter 6, I describe my participation in the ASAP-Polaris-OpenADMET Antiviral Drug Discovery Challenge [13, 14], a prospective, blinded community challenge that tasks participants with predicting crystal binding poses for SARS-CoV-2 and MERS-CoV Main Protease ligands. Rather than deploying deep learning models, I investigate to what extent classical docking with AutoDock Vina [15] can be enhanced through fragment-guided sampling and scoring. This approach ranks 6th overall and 2nd among non-deep-learning methods, demonstrating that traditional docking paired with careful scoring refinement remains competitive with modern machine learning approaches. Furthermore, the results reveal that molecular docking retains a distinct advantage as a robust hypothesis generator due to its ability to freely sample diverse poses, whereas current cofolding models often exhibit constrained sampling capabilities, suggesting that the community can still greatly benefit from classical methods even as the field shifts toward learned structure prediction architectures.

Collectively, these efforts demonstrate that the advanced reasoning and sampling capabilities of Large Language Models, combined with rigorous structural benchmarking and prospective experimental validation, can produce practical, competitive tools for real-world drug discovery. By ensuring synthesizability at the point of generation, enforcing structural integrity and conformational correctness in evaluation datasets, and validating predictions in blinded community settings, this work establishes a coherent framework that bridges the critical gap between computational molecular design and successful translational outcomes.

Contents

Contents	i
1 Introduction	1
1.1 Generative Molecular Design with LLMs	2
1.2 Structure-Based Benchmarking	7
1.3 Translational Drug Discovery Workflows	9
2 SynLlama: Generating Synthesizable Molecules and Their Analogs with Large Language Models	13
2.1 Introduction	13
2.2 Methods	15
2.3 Results	22
2.4 Discussion and Conclusion	29
2.5 Supplementary Information	31
3 LinkLlama: Enabling Large Language Model for Chemically Reasonable Linker Design	50
3.1 Introduction	50
3.2 Methods	52
3.3 Results	57
3.4 Discussion and Conclusion	66
3.5 Supplementary Information	68
4 A Workflow to Create a High-Quality Protein–Ligand Binding Dataset for Training, Validation, and Prediction Tasks	82
4.1 Introduction	82
4.2 Methods	84
4.3 Results	88
4.4 Discussion and Conclusion	96
4.5 Supplementary Information	98

5	KinConfBench: A Curated Benchmark for Cofolding Models on Kinase Conformational States	112
5.1	Introduction	112
5.2	Methods	114
5.3	Results	114
5.4	Discussion and Conclusion	123
5.5	Supplementary Information	125
6	Polaris Challenge: Data-driven Priors to Improve Docking for Pose Prediction	134
6.1	Introduction	134
6.2	Methods	135
6.3	Results	140
6.4	Discussion and Conclusion	146
6.5	Supporting Information	148
7	Conclusion	153
	References	155

Acknowledgments

While history often attributes genius to the isolated outlier, the true engine of scientific discovery has always been the power of the collective mind. At Berkeley, I was fortunate to collaborate with many such great minds, both on and off campus, who ultimately shaped my research and my growth as a person. I have learned profound lessons working alongside exceptional collaborators, and I want to take a moment to pay tribute to everyone who helped and supported me along this invaluable journey.

First and foremost, I express my deepest gratitude to my research advisor and mentor, Prof. Teresa Head-Gordon, for her steadfast support throughout the past four years. From detailed code debugging to research ideation and collaborative paper editing, her guidance has been instrumental. Working with her taught me that it is perfectly fine to make mistakes and embrace trial and error in science. Having this freedom to explore ultimately shaped me into the researcher I am today: someone equipped with a wealth of ideas and, more importantly, the knowledge of how to bring them to fruition.

Beyond her direct mentorship, one of the greatest gifts Teresa has given me is the exceptional group of people she brings together in her lab. Being part of such a dynamic group with diverse research interests allowed me to expand my understanding of not only my own field, but other areas of physical chemistry as well. Specifically, I want to acknowledge past members Dr. Jie Li, Dr. Xingyi Guan, Dr. Oufan Zhang, Dr. Maria Tsanai, and my mentee Dorian Bagni, along with current members Yingze Wang, Joseph Cavanagh, Justin Purnomo, Giovanni Battista Alteri, Dr. Jordy Homing Lam, Stefano De Castro, and Scott Rankin, for their insightful discussions and support of my ideas within the Bio-ML subarea of the lab. I also deeply appreciated the stimulating conversations grounded with more physics I shared with past members Dr. Joseph Heindel, Dr. Selim Sami, and Dr. Wanlu Li, as well as current members Dr. Allen LaCour, Eric Chung-Yueh Yuan, Lukas Kim, Aalim Abdullah, Dr. Yunsheng Liu, Dr. Ruoqi Zhao, Dr. Sruthy Chandy, Joey Shin, and Jonathan Arnold. Research aside, I cherished my time in the lab at Berkeley thanks to you all. The fun hikes, board game nights, NBA watch parties, and casual life chats enriched my graduate school experience immensely.

Beyond my own lab, I was incredibly fortunate to find mentorship and camaraderie throughout the wider Berkeley community. I learned immensely from faculty, postdocs, and my graduate student peers through rigorous classes, collaborative projects, Chem kegs, softball games, and the shared trauma bonding inherent to a Ph.D. program. While it is impossible to name everyone who helped and supported me along the way, I am profoundly grateful to my qualification and thesis committees. The insightful suggestions and dedicated time from Prof. Eric Neuscamman, Prof. Liana Lareau, Prof. Eran Rabani, Prof. Ke Xu, and Prof. Jennifer Listgarten were instrumental in shaping this dissertation.

Beyond Berkeley, I was incredibly fortunate to collaborate with and receive mentorship from exceptional scientists who helped ground my passion for drug discovery in chemical reality, while simultaneously inspiring my work on LLM agents, cofolding, and generative models. I wish to express my sincere gratitude to all the researchers at the Midwest AVIDD

Center, specifically Prof. Reuben Harris for his outstanding leadership of the program, alongside Prof. Donghoon Chung, Prof. Thomas Bannister, Prof. Rommie Amaro, Dr. Ambuj Srivastava, and Alma C. Castañeda-Leautaud for our collaboration on Project 5, focusing on the design and testing of novel antivirals targeting the SARS-CoV-2 Nsp13 helicase. I also want to thank Prof. Damian Young and Dr. Kate Holloway for their insightful discussions and valuable suggestions regarding my research direction.

Parallel to that work, I am deeply grateful to have collaborated with Dr. Julie Forman-Kay and Zi Hao Liu on the study of intrinsically disordered proteins. I thank Dr. Jacob Sawyer and Bo Zhou for their help on SynLlama. Furthermore, I extend my thanks to Dr. Yuanqi Du, Dr. Tianyu Liu, Dr. Tony Shen, Botao Yu, Prof. Wengong Jin, and our other brilliant collaborators for our incredibly exciting work on the design of a scientific autonomous goal-evolving agent. A special shoutout goes to my dearest mentor, Dr. Mia A. Rosenfeld, as well as Dr. Matthew Welborn, Dr. Aleksander Durumeric, Dr. Michael Irvin, Qianzi Li, and Mac Kevin Braza. Thank you for being a vital part of the KinConfBench project and for fostering such a welcoming environment within the NeuralPLexer team at Iambic Therapeutics. Finally, I am deeply thankful to Dr. Hui Zhang, Dr. John Huang, Dr. Laurent Gomez, and all the other great scientists I met at Iambic for their insightful discussions and guidance during my internship.

Beyond the realm of research, this journey would have been entirely impossible without the unwavering love and support of the friends who have always believed in me. While the list is long, I want to extend my special thanks to Junyi Guo, Hancun Sun, Mingze Li, Yuxuan Fan, Dr. Rui Wang, Xiangyi Kong, Tongkun Liu, and Zhou Sun for their profound emotional support throughout my time here. I also wish to express my gratitude to Dr. Matthew Du, Prof. Joel Yuen-Zhou, and Prof. Michael Gilson for initially introducing me to this path of theoretical chemistry, and continuing their support throughout my graduate studies.

My deepest gratitude belongs to my parents, Ying Zhang and Changzhu Sun, for giving me the opportunity, courage, and support to study and build a career in a new country. And to our close family friends, Steven Sun, Sally Wang, Victor Sun, and Nadia Peng, thank you for treating me as one of your own and making the Bay Area truly feel like home. Finally, I want to thank my wife, Chang Liu, and our two cute family meow-mbers, Xiao Pang and Kwonka, for making my graduate school experience uniquely special. Chang, thank you for your unwavering support, your constant encouragement, and for always believing that I can accomplish great things.

Ultimately, this dissertation is proof that no great endeavor is achieved alone, as it really is the proud culmination of the brilliant collective mind and loving community I have been so fortunate to share this journey with. And thank you for reading it!

Chapter 1

Introduction

The process of discovering new small-molecule therapeutics is notoriously time-consuming and capital-intensive, typically requiring deep domain knowledge and years of iterative hit identification and lead optimization to arrive at a good drug candidate [1, 2]. Historically, researchers have relied on experimental small molecule high-throughput screening or DNA-encoded library screenings to search for initial hit molecules [16, 17]. With the rise of large molecular databases such as ChEMBL [18], PubChem [19], and ZINC [20], virtual screening has also become a viable alternative to save experimental cost by first filtering available molecules with heuristics like docking scores to prioritize computationally competent molecules for experimental testing [21]. However, even with the recent advent of ultra-large libraries such as the Enamine REAL library with billions of molecules [22], these small molecule collections are dwarfed by estimates for the total number of possible synthesizable small molecules, which range from 10^{24} to 10^{60} [3]. Screening all compounds with sufficiently sophisticated structure-based methodologies such as molecular docking remains intractable, and even simpler approaches are prone to ruling in false positives while missing potential leads [23, 24].

Generative AI has fundamentally reshaped the drug discovery landscape by offering a powerful alternative to traditional screening approaches [4–6, 25]. Rather than exploring fixed libraries, generative models learn the underlying “grammar” of chemical space [26, 27], allowing us to sample from a vast, continuous ocean of molecules. Early generative approaches, including variational autoencoders [27], recurrent neural networks, and reinforcement learning (RL) [28, 29], demonstrated that the distribution of generated molecules can be skewed towards candidates with specific properties such as target binding affinity. However, these methods commonly optimize against simple, low-fidelity proxy scores, such as Quantitative Estimate of Drug-likeness (QED), docking scores, and Synthetic Accessibility (SA) scores [15, 30, 31], which frequently leads to reward hacking, where the model exploits loopholes in the scoring function rather than learning meaningful chemistry [32, 33]. This behavior results in two well-documented failure modes: noisy guidance, in which unreliable proxy signals steer generative models towards suboptimal or impractical chemical domains [7, 34], and mode collapse, in which models become trapped in local regions of chemical

space rather than exploring diverse solutions.

Large Language Models (LLMs) offer a compelling paradigm shift for molecular design. Trained on massive corpora, LLMs are foundation models [35] that inherently possess broad chemical knowledge and can be efficiently adapted to downstream tasks through supervised fine-tuning (SFT) [36–40]. Unlike task-specific generative models for desired molecules, LLMs can be conditioned through natural language prompts, enabling direct specification of molecular properties and design constraints without retraining. Furthermore, fine-tuned LLMs can match or exceed the performance of chemical language models trained from scratch on chemical data alone, while requiring significantly less data [36]. These properties, such as natural language conditioning, data-efficient fine-tuning, and broad foundational knowledge, make LLMs uniquely suited for flexible, multi-objective molecular design that escapes the mode collapse problems from previous generations of generative models.

Yet even the most capable generative model is only as good as the data it learns from and the benchmarks it is evaluated against. The structural datasets that underlie scoring functions and model training contain well-documented noise, including incorrect bond geometries and data leakage between training and test splits [41–43]. Meanwhile, emerging structure prediction tools such as cofolding models [44–47] are rapidly being adopted for drug discovery, yet existing benchmarks remain predominantly ligand-centric and fail to assess whether these models capture the conformational dynamics that govern drug binding [43, 48]. Ultimately, computational predictions must be validated through physics-based simulations, such as molecular dynamics and free-energy calculations, as well as real experiments to demonstrate practical utility.

My thesis addresses these challenges across three interconnected pillars. First, I developed LLM-based generative approaches for molecular design, including SynLlama [8] and LinkLlama [49] as primary works, with contributions to SmileyLlama [36], TMCLlama [50], and SAGA[51]. Second, I established reliable structural benchmarks through HiQBind [9] and KinConfBench [12] as primary works, with contributions to Leak Proof PDBBind [41]. Third, I validated these tools in practical drug design campaigns, applying the iMiner workflow [52] in the Midwest AViDD collaboration [53], and demonstrating competitive performance in community challenges including the ASAP-Polaris-OpenADMET Antiviral Drug Discovery Challenge [13, 14] and the CACHE Challenge #7[54].

1.1 Generative Molecular Design with LLMs

Chemical Language Models (CLMs) trained on string representations such as SMILES [55] and SELFIES [56] have emerged as powerful tools for *de novo* molecular generation [4–6, 25]. Nearly all existing CLMs have been trained from scratch on large chemical databases using architectures ranging from variational autoencoders [57] to generative pre-trained transformers (GPTs) [58]. While effective at generating valid molecules, these models face persistent challenges when applied to drug discovery: they often lack practical guarantees of synthesiz-

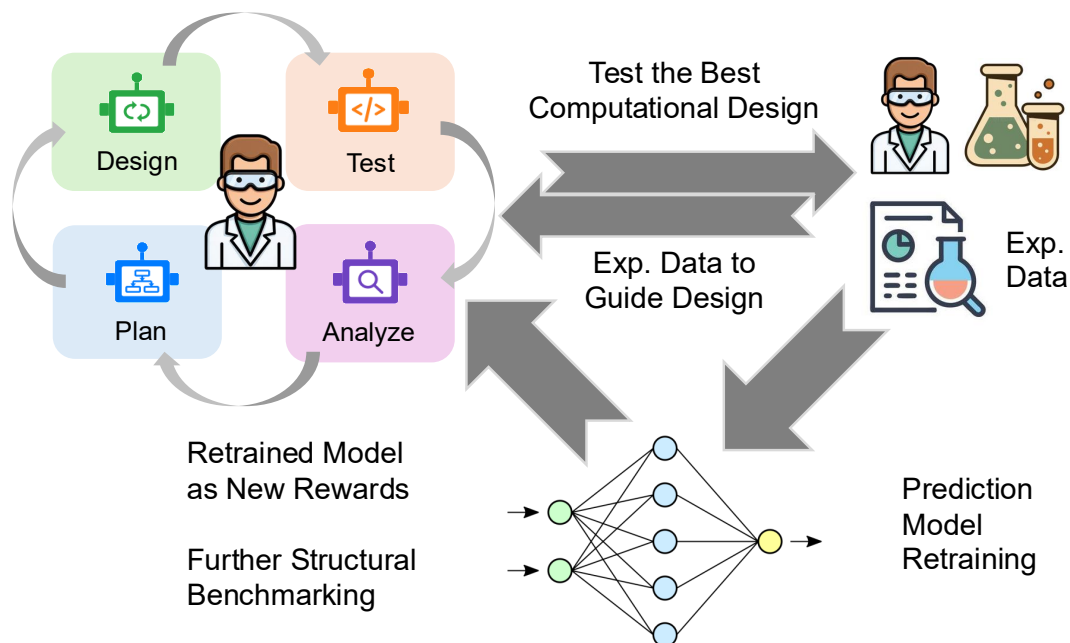


Figure 1.1: Synergy between LLM-based drug design, structural benchmarking, and prospective drug discovery.

ability [7, 34], struggle with multi-parameter optimization, and tend toward the aforementioned failure modes of reward hacking and mode collapse when steered by RL.

The advent of LLMs provides a fundamentally different starting point. Rather than training chemical models from scratch, LLMs such as the Meta Llama series [59] can be fine-tuned on chemical data through supervised fine-tuning (SFT), inheriting the representational capacity and broad knowledge of general-purpose language models [35]. This approach offers several advantages: natural language prompts allow direct specification of desired molecular properties without reward function engineering; SFT requires substantially less data than training from scratch; and the same fine-tuning paradigm generalizes across diverse chemical domains.

We first demonstrated this principle with SmileyLlama [36], which shows that fine-tuning an open-weight LLM produces a CLM that performs at or above the level of models trained solely on chemical SMILES data. By conditioning generation on natural language descriptions of molecular properties, such as molecular weight, number of hydrogen bond donors, and topological polar surface area, SmileyLlama effectively partitions chemical space, forcing the model to sample from desirable subspaces defined by specific property constraints. This natural language guidance also provides a remedy for mode collapse: because different prompts direct the model to sample from different regions of chemical space, the approach explores diverse solutions rather than greedily converging on a single optimum. The same SFT framework is further extended to transition metal complexes through TMCLlama [50],

demonstrating that the approach generalizes beyond drug-like small molecules to coordination chemistry and materials design. During this project, my contribution centers around evaluating the LLM outputs on standard chemistry benchmarks and exploring utilities for integrating this more powerful generative engine to our pre-existing iMiner workflow [52] that performs generation and validation at one step. By pre-conditioning a model to generate molecules within specific property constraints, a fine-tuned LLM like Smileyllama inherently becomes more efficient than the traditional paradigm of blind sampling followed by rigorous filtering. Rather than wasting computational resources generating and subsequently discarding irrelevant compounds, LLMs can be guided to explore only the most viable regions of chemical space from the outset. Recognizing the transformative potential of this capability, I shift my focus from benchmarking towards actively developing new, targeted applications for these generative techniques, which ultimately led to the development of SynLlama and LinkLlama.

Building on this foundation, SynLlama [8] addresses the synthesizability problem by fundamentally altering the generative search space itself with the power of an LLM in Chapter 2. Traditional approaches attempt to optimize synthesizability proxy scores on a molecular graph, but these scores are unreliable—they can assign high scores to impractical molecules and penalize complex yet synthesizable compounds that require multi-step pathways [32, 33]. Previous efforts to address this have either relied on molecular fragment assembly without explicit reaction pathways [60–63], or integrated computer-assisted synthetic planning (CASP) software with significant computational overhead [64, 65]. SynLlama takes a different approach where the LLM is fine-tuned to generate synthetic pathways for an arbitrary input molecule by effectively break it down into a synthesizable space formed by strictly assembling commercially available building blocks via known chemical reactions [64, 66]. By constraining generation to this “synthesizable chemical space,” every proposed candidate comes with an actionable synthesis pathway, which is extremely appealing to medicinal chemists who can directly examine, refine, and execute the proposed routes. Related models such as SynNet [67], ChemProjector [68], and Synformer [69] have also explored building-block-based generation, but the LLM backbone of SynLlama enables seamless integration with the natural language conditioning demonstrated in SmileyLlama.

Furthermore, leveraging data-efficient SFT transforms this general-purpose language model, Llama-3.2-1B, into a highly versatile generative engine for tasks related to chemical syntheses. Unlike models trained from scratch on strictly confined chemical libraries, this fine-tuning strategy allows SynLlama to internalize the underlying logic of organic synthesis rather than merely memorizing rigid reaction templates. This distinctly generative nature enables the model to extrapolate beyond its immediate training data, allowing it to propose novel synthetic routes utilizing unseen, commercially available building blocks. Because of this inherent adaptability, SynLlama is not limited to a single, narrow application. Rather, it can be seamlessly deployed across various critical stages of the drug discovery pipeline. Whether tasked with *de novo* synthesizable analog generation, complex synthesis planning, or localized hit expansion, the model serves as a flexible, expert-level computational assistant that ultimately bridges the gap between in silico design and practical wet-lab execution.

Although SmileyLlama and SynLlama fulfill many needs of structure-based drug design, including fragment growth, conditional generation, hit expansion, and R-group substitution on a central chemical scaffold, both models operate in a space with fixed chemical cores. To complete our tooling suite, we introduce LinkLlama [49], a specialized framework developed via LLM fine-tuning, designed to solve a persistent bottleneck in fragment-based drug discovery: the generation of chemically reasonable linkers. Traditional methods, such as 3D-coordinate diffusion models like DiffLinker [70] or 2D-graph growth models like DeLinker [71], often struggle to produce chemically reasonable molecules with favorable geometries. Common failure modes include the generation of molecules with strained fused rings, broken aromaticity, and other unstable ring systems or chemical motifs that do not typically appear in drug-like molecules. By reframing linker design as a conditional generation task given the fragment information and whole-molecule properties, LinkLlama inherently avoids these pitfalls, as the sampling of chemically reasonable linkers is incorporated into the input prompt for SFT. As a result, this approach yields a sizable leap in chemical quality, demonstrating a high percentage of chemically reasonable molecules compared to baseline models. Remarkably, LinkLlama maintains the spatial integrity of fragment hits across diverse scales, from simple scaffold hopping, fragment joining, to difficult design of PROTAC linkers, all without relying on complex, explicitly geometric training objectives.

Beyond its structural accuracy, a major advantage of LinkLlama is its capacity for “RL-free” conditional sampling. Unlike older generative models that depend on unstable and computationally expensive reinforcement learning cycles for multi-objective optimization, LinkLlama can be steered directly through natural language prompts, similar to SmileyLlama, to satisfy various design needs. This “alignment-by-design” approach allows medicinal chemists to easily tailor chemical property distributions without needing complex reward functions. Looking forward, LinkLlama is designed to function not in isolation, but as a core pillar of the broader “Chemical Llama Suite,” working synergistically with SmileyLlama for general molecular generation and SynLlama for retrosynthetic planning. Because the model communicates via standard SMILES and natural language, it is natively compatible with autonomous AI agents. This seamless integration paves the way for closed-loop drug discovery workflows where agentic systems can autonomously generate candidates, evaluate them through automated simulations, and continuously refine the model’s output using techniques like Direct Preference Optimization (DPO). Ultimately, this achieves an LLM-based equivalent of the commonly used REINVENT tool suite[6], while providing users with greater prompt-level control without expensive active learning or retarining.

Beyond individual molecular generation, SAGA (Scientific Autonomous Goal-evolving Agents) [51] represents a further evolution of the LLM-driven design philosophy by embedding generative models within an autonomous agentic framework. In SAGA, an Optimizer generates molecular candidates under current objectives, an Analyzer examines top hits using analysis tools and identifies patterns, a Planner proposes new reward terms and re-weights objectives, and an Implementer writes new scoring functions, automating the reward-tuning loop that scientists traditionally perform manually. The system operates under three modes of human involvement—co-pilot, semi-pilot, or autopilot—designed to augment, not replace,

human expertise. Although SAGA is a computational framework, its outputs have been experimentally validated across multiple domains. Specifically, 28 designed compounds targeting *E. coli* were synthesized, with 4 showing greater than 80% bacterial growth inhibition *in vitro* (the threshold conventionally used to identify a primary-screen hit). The resulting $\sim 14\%$ hit rate is roughly two orders of magnitude above the $\sim 0.1\%$ typical of high-throughput screening campaigns. In addition, the SAGA framework generalizes to *de novo* PD-L1 nanobody design, DNA enhancer optimization, inorganic materials discovery, and chemical process design.

In this project, my contribution centers primarily on the evaluation phase. Acting as the human computational chemist in the antibiotic discovery loop, I provide constructive feedback to guide the model’s exploration of diverse chemical spaces, interacting with SAGA by inputting specific analyses and strategic plans for the antibiotic portion of the study. One of the most surprising outcomes of the project is the system’s ability to perform sophisticated population-level analyses and assess crucial drug-related properties, such as metabolic stability, from a highly general prompt like, “Generate new antibiotics that are safe and synthesizable.” Throughout the project and the downstream analysis of the model’s performance, this experience suggests that autonomous scientific discovery is highly attainable when grounded in the thoughtful design of human-AI collaboration.

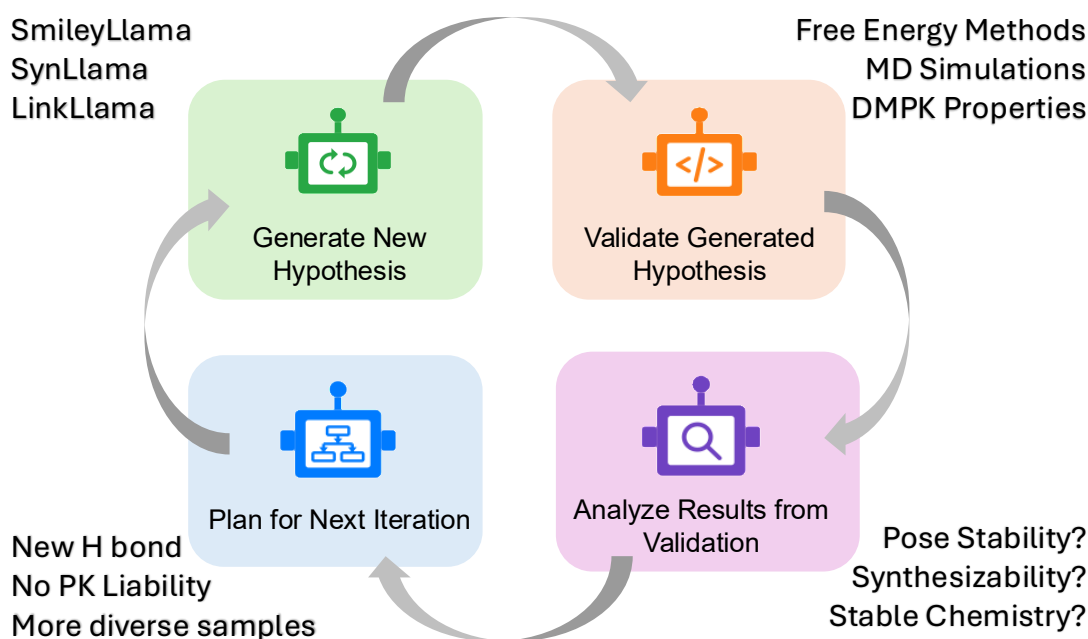


Figure 1.2: **Drug Discovery in the age of LLM.** An example workflow adapted from SAGA where agents can close the loops of design-test-analyze-plan cycle for drug discovery.

Beyond their individual applications, these projects also represent a fundamental paradigm shift toward autonomous research frameworks as illustrated in Figure 1.2. I envision an ap-

proaching evolution where systems like SAGA serve as high-level orchestrators, leveraging the reasoning capabilities of general-purpose LLMs to direct domain-specific models, such as SmileyLlama, SynLlama, and LinkLlama, via functional tool-calling. By seamlessly coordinating these specialized generative tasks with computationally rigorous validation software, these agentic systems can effectively close the design-test loop. Ultimately, these integrated LLM ecosystems will transcend their roles as static tools to become dynamic, collaborative partners, effectively serving as a "second brain" for human scientists to accelerate the frontier of drug discovery.

1.2 Structure-Based Benchmarking

Generative models and scoring functions are only as reliable as the structural data they are trained on and the benchmarks they are evaluated against. The PDBbind dataset [72], a curated set of approximately 19,500 biomolecular complex structures with experimentally measured binding affinities, serves as the primary training and evaluation resource for the majority of protein-ligand scoring function predictors, whether physics-based [15, 73, 74] or machine-learned [75–80]. Various benchmarks derived from PDBbind, such as the Comparative Assessment of Scoring Functions (CASF) series [81–84], are used to assess the scoring, ranking, docking, and screening power of these methods. However, two decades of reliance on PDBbind reveal significant structural and organizational issues that undermine model training and evaluation.

We first address the data leakage problem with Leak Proof PDBBind (LP-PDBBind) [41]. The commonly used splits of PDBBind include identical or highly similar proteins in both training and test sets, causing affinity prediction models to overfit to memorized protein families rather than learning generalizable binding patterns. By carefully re-splitting PDBBind to control for protein and ligand similarity, we construct a leak-free dataset and retrain four popular scoring functions: AutoDock Vina, RF-Score, InteractionGraphNet (IGN), and DeepDTA. Overall, the results show that performance drops significantly across these models on the leak-proof split, revealing that previously reported gains are partly attributable to data leakage rather than genuine predictive improvement. Furthermore, this work introduces BDB2020+, a new independent test set matching high-quality binding free energies from BindingDB [85] with co-crystallized complexes deposited in the PDB since 2020, providing a more stringent evaluation test set for future affinity models. In this work, my contribution centers on reimplementing and testing the DeepDTA method as the representative 1D model for affinity prediction. Interestingly, while DeepDTA is not as severely affected by the data leakage as the 3D-based methods, the leak-proof retraining still improves its generalizability on curated benchmarks, such as those for the SARS-CoV-2 Main Protease and human EGFR kinase.[41]

Beyond data leakage, the PDBbind dataset itself contains underlying structural noise that misrepresents physical interactions [41, 42], including incorrect ligand bond lengths and angles, steric clashes, and sub-optimal organization of protein-ligand classes. While the

biomolecular community relies heavily on databases like PDBbind[72], BioLiP2[86], BindingDB[85], and Binding MOAD[87], the curation of high-quality data requires rigorous, ongoing quality control. This challenge is compounded by the fact that PDBbind data curated after 2020 is commercialized into a closed system without published workflows, severely limiting accessibility and preventing fair methodological comparisons. To address this issue at its source, we introduce HiQBind-WF [9], a comprehensive, open-source optimization workflow that systematically identifies and cleans structural errors in protein-ligand systems. By deploying this pipeline, we compile the HiQBind dataset, leveraging sources like BioLiP2, BindingDB, and Binding MOAD to produce over 30,000 high-quality, non-covalent protein-ligand structures spanning more than 18,000 unique PDB entries. This new dataset successfully corrects common imperfections, offering the community a reliable resource with accurately assigned bond orders, adjusted protonation states, and highly refined protein structures.

Within this larger effort, my specific contribution centers on the protein processing pipeline, which is crucial because missing residues and atoms frequently cause breakages in molecular dynamics (MD) simulations and other downstream computational evaluations. By integrating open-source software such as OpenMM [10] and PDBFixer [11] with native annotations extracted directly from the PDB Data API and MMCIF file headers, the workflow retrieves rich structural context to accurately model missing segments. Importantly, this approach meticulously documents our restorative efforts, providing users with a transparent record of which structural elements are native to the original PDB and which are fixed by our workflow. Because the entire data processing pipeline is automated and open-source, it directly resolves long-standing community concerns regarding the reproducibility of proprietary curation methods. Ultimately, HiQBind provides a sustainable, transparent framework for continuously refining protein-ligand datasets, establishing a robust new benchmark resource that emerges as an accessible version of structural complexes with binding labels for the rigorous training and evaluation of future scoring functions.

Besides directly training and predicting binding affinities on fixed input structures, the rapid development of cofolding models, like AlphaFold3 [44], NeuralPLexer [47], Chai-1 [45], and Boltz [46], emerges as a powerful alternative for generating structural hypotheses for binding complexes and predicting their binding affinities, as demonstrated by Boltz-2 [88]. These advancements create an urgent need for benchmarks that go beyond basic structural recovery. Existing evaluations primarily assess ligand-centric metrics, such as protein pocket-aligned ligand RMSD and global fold confidence scores [48], but neglect the correctness of the protein dynamical hypothesis generated with the ligand placement. In practical drug discovery, only identifying the correct global fold is insufficient because successful lead optimization requires capturing the specific, induced-fit conformational states dictated by ligand binding.

To fill this gap, we introduce KinConfBench [12], which evaluates cofolding models based on their ability to capture the conformational landscape of human kinases. Kinases are ideal test systems given their high representation in the PDB, their importance as therapeutic targets [89–91], and their inherently dynamic nature. They function as molecular switches

that toggle between active and inactive states via subtle rearrangements of the activation loop (DFG motif) and the α C-helix [92, 93]. Critically, these conformational states dictate inhibitor binding kinetics and selectivity, distinguishing Type I from Type II inhibitors [94]. KinConfBench quantifies how well cofolding models recover these state-specific conformations, providing a rigorous assessment of whether structure prediction tools can truly drive rational drug design.

Specifically, our benchmarking of popular models reveals a critical tension between confident geometric poses and biophysically faithful conformational states. While models often achieve strong pocket-centric metrics, KinConfBench demonstrates that they frequently fail by generating structures that merely look bound while adopting the incorrect regulatory layout for the specific ligand. Furthermore, paired holo/apo analyses expose a pervasive issue of apo-state memorization such that current cofolding architectures often default to unbound states rather than generating the rare, induced-fit configurations required for holo states. By highlighting that models prioritize superficial geometric confidence over true functional state discrimination, KinConfBench establishes an explicit, clinically relevant standard that predicting the general architecture of a kinase is biologically insufficient if the specific, ligand-induced functional conformation is incorrect.

Together, these benchmarking efforts (eliminating data leakage via LP-PDBBind, cleaning structural noise through HiQBind, and introducing conformation-aware evaluation with KinConfBench) provide the reliable structural foundation that both generative models and scoring functions depend upon. By enforcing structural integrity, predictive generalizability, and functional correctness, these tools act as essential scientific guardrails against the data memorization and superficial geometric confidence that currently limit the field. Looking forward, as AI-driven drug discovery transitions from isolated generative tasks toward integrated, autonomous agentic workflows, the demand for high-fidelity data and stringent evaluation only intensifies. Ultimately, ensuring that next-generation cofolding models and autonomous systems navigate both chemical and conformational space with true biophysical awareness is what bridges the critical gap between computational design and successful clinical translation.

1.3 Translational Drug Discovery Workflows

The ultimate test of any computational method is whether it produces molecules that work in the real world. While Sections 1 and 2 describe the tools for molecular generation and the data foundations for evaluation, this section presents their application in iterative drug design workflows, collaborative drug discovery campaigns, and competitive community challenges.

The backbone of our drug design efforts for experimental projects is the iMiner workflow [52], an iterative generative-filtering framework that bridges *in silico* molecular design with physics-based validation. iMiner takes candidates from generative models and subjects them to successive rounds of filtering: first through molecular property and drug-likeness criteria [95–97], then through analysis of protein-ligand interaction profiles, and finally through

molecular dynamics (MD) simulations that assess binding stability under realistic physical conditions. For congeneric series of molecules, the expanded workflow now also incorporates binding free energy predictions to quantify relative affinities. This modular design makes iMiner adaptable to diverse targets and generative backends.

Within the iMiner framework, my specific contribution centers on training and fine-tuning models to handle distinct design choices, such as the retention of specific fragments and the geometric preservation of core functional groups. While these constraints can technically be managed through reward function engineering, practical experience reveals that traditional generative models suffer from severely limited sampling capabilities. For instance, in fragment-constrained tasks, fine-tuning forces the model to become hyper-specific to a single fragment. Tackling versatile drug discovery tasks would thus require training a separate model for every unique fragment. This lack of a generalizable framework highlights the inherent limitations of traditional generative models and directly motivates our exploration into the Large Language Model (LLM) approaches detailed in Section 1. By utilizing natural language prompts to flexibly guide conditional generation tasks, we successfully overcome these bottlenecks, replacing rigid fine-tuning with a highly adaptable design paradigm.

Despite the generative bottlenecks that motivated our pivot toward LLMs, the core filtering and validation pipeline of iMiner remains a highly robust engine for practical drug discovery. Putting this framework into practice, we deploy iMiner as part of the Midwest Antiviral Drug Discovery (AViDD) Center [53], one of several centers established to rapidly develop antivirals for emerging viral threats [98]. Working in close collaboration with medicinal chemists, we use this generative and physics-informed pipeline to design inhibitors targeting the SARS-CoV-2 NSP13 helicase. The workflow proposes both *de novo* molecules and strategic scaffold modifications based on high-throughput screening hits, which are subsequently synthesized and validated through experimental assays. Notably, the iMiner *de novo* molecular series achieved a 17.4% hit rate (2 orders of magnitude better compared to the average high-throughput screening hit rate of around 0.1%), with 4 compounds out of the 23 synthesized candidates showing an EC₅₀ value below 20 μ M without noticeable cytotoxicity, qualifying for primary hits for further optimization. Furthermore, the proposed molecules are structurally novel compared to known inhibitors for this protein, proving to us that generative design could serve as an efficient hypothesis generator where one need not test numerous compounds to get a structurally novel hit. This success demonstrates that our computational tools can produce viable candidates that survive the full translational pipeline, from initial *in silico* generation to expensive computational validations and rigorous experimental confirmation.

Beyond collaborative discovery campaigns, we validated our methods through prospective community challenges that provide rigorous, blinded assessments of computational approaches. In the ASAP-Polaris-OpenADMET Antiviral Drug Discovery Challenge [13], organized through the Polaris Hub [99] in collaboration with ASAP Discovery [100] and OpenADMET [101], participants were tasked with predicting crystal binding poses for 98 SARS-CoV-2 Mpro ligands and 97 MERS-CoV Mpro ligands using a training set of approximately 770 chronologically curated crystal structures [102]. This prospective design, with training

on historical data and testing on subsequently determined structures, mimics real-world hit-to-lead campaigns and provides a stringent test of generalizability. Rather than deploying deep learning models, we intentionally investigated to what extent classical docking with AutoDock Vina [15] could be enhanced through fragment-guided sampling and scoring. Our approach ranked 6th overall and 2nd among non-deep-learning methods, demonstrating that traditional docking paired with careful scoring refinement remains competitive with modern ML approaches and continues to play an important role in practical drug discovery workflows [14]. Furthermore, we demonstrate that molecular docking remains a robust hypothesis generator due to its ability to freely sample diverse poses, whereas modern cofolding models like Boltz and its fine-tuned variants often exhibit constrained sampling capabilities. This practical insight suggests that even as the field increasingly shifts toward cofolding architectures, the community can still greatly benefit from traditional docking methods, particularly in scenarios where crystal structures are scarce or prior protein-ligand information is highly limited.

In a separate prospective challenge, CACHE Challenge #7 [54], we deploy a methodological framework that combines SmileyLlama, SynLlama, and the iMiner workflow to generate *de novo* molecules that selectively target the protein PGK2 over its close analog PGK1 and avoids a specific scaffold. For this challenge, we are provided with initial crystal structures and DNA-encoded library (DEL) hits that demonstrate baseline selectivity for PGK2. To construct a comprehensive workflow that adheres to the CACHE requirement of proposing molecules strictly within the Enamine REAL space [22], we first utilize SmileyLlama to generate diverse *de novo* designs. These initial candidates are then filtered through the iMiner pipeline to isolate viable hits. Next, we project these molecules through SynLlama for hit expansion, generating structurally related compounds that retain strong predicted scores while maximizing synthetic accessibility. We then perform an explicit database query using a similarity cutoff to retrieve the closest commercially available matches directly from the Enamine catalog, as required by the organizers. On this final pool of synthesizable candidates, we apply our rigorous validation pipeline, calculating over 1,000 absolute binding free energy data points for this specific target to confidently select the top 150 molecules for our final submission. Ultimately, this workflow ranks 2nd overall in both the double-blind peer review and expert panel evaluations out of more than 40 participating groups. This end-to-end application demonstrates that the LLM tools developed across our individual methods integrate cohesively when deployed in a competitive drug design setting, and we eagerly await the impending experimental testing to confirm their practical success.

In summary, these practical applications validate that our computational methodologies transcend theoretical exercises to solve real-world drug discovery problems. By successfully deploying the iMiner pipeline to yield experimentally confirmed, structurally novel antivirals, we demonstrate the robustness of our iterative design and physics-based validation framework. Furthermore, our top-tier performance in competitive challenges underscores the versatility of our toolkit. We prove both the enduring value of classical docking as a diverse hypothesis generator when structural data is limited, and the immense capability of our integrated LLM ecosystems to navigate massive commercial chemical spaces for highly

selective, synthesizable de novo design. Ultimately, these real-world deployments confirm that by carefully balancing advanced generative AI with rigorous physical modeling, we can consistently produce highly competitive, experimentally viable candidates that stand at the forefront of modern computational chemistry.

Chapter 2

SynLlama: Generating Synthesizable Molecules and Their Analogs with Large Language Models

2.1 Introduction

Chemical space is enormous, built up via the exponential rise in functional group combinatorics that define an increasing diverse set of molecules. Traditional approaches that design synthetic pathways of unseen molecules under well-controlled laboratory conditions have been made possible by decades of research in synthetic chemistry, as well as mechanistic studies of key reaction steps and reaction classification[103, 104]. Using the wealth of data accumulated in libraries of chemical reactions, expert systems[105, 106] have been developed that deploy this knowledge to construct multi-step pathways to specified end-products. Such methods have become a key tool for the bench chemist, as illustrated by the Chematica software and its follow on commercial product now known as Synthia[106].

With recent advances in artificial intelligence and deep learning, generative models have begun to contribute to enumerating molecules at the stoichiometric scale. After training on databases containing various small molecules representations[72, 85, 107–109], string-based 1D generative models and structure-aware 3D *de novo* methods have paved the way for quick exploration of greater swaths of unseen chemical space[4–6, 36, 52, 110–122]. However, even with their exceptional generative capabilities, these models still face one major challenge: their proposed *de novo* molecules lack practical guarantees of synthesizability, which limits their utility in practice[7, 34, 123]. For generative approaches in drug and materials discovery to fulfill their potential, ensuring synthetic feasibility is essential to bridge the gap between *in silico* molecule design and the realistic applicability of computationally generated molecules.

¹Reproduced with permission from: SynLlama: Generating Synthesizable Molecules and Their Analogs with Large Language Models. K. Sun, D. Bagni, J. M. Cavanagh, Y. Wang, J. M. Sawyer, B. Zhou, A. Gritsevskiy, O. Zhang, and T. Head-Gordon. ACS Cent. Sci. 2025, 11, 11, 2108–2120.

In recognition of this dissonance, efforts have been made to address the problem of poor synthesizability of *de-novo*-generated molecules. One line of research focuses on integrating empirical or deep-learning scoring functions[30, 31, 124–128], such as the synthetic accessibility (SA) score [30] and DeepSA score[31], into the objective functions of learning algorithms. However, optimizing only the synthesizability score can still lead to the generation of unsynthesizable molecules because the scoring functions rely on identifying common fragments or reactive centers in molecules[32]. In addition, they often assign bad scores to complex yet synthesizable molecules that require multi-step synthetic pathways, causing generative models to miss viable candidates [33]. Others have proposed improving synthesizability by building molecules from common molecular fragments[60–63], but they still don’t guarantee synthesis as these methods do not explicitly consider the reaction pathways to build the molecular candidates. Another line of research integrates the explicit use of computer-assisted synthetic planning (CASP) software[64, 129] into the optimization[130]. However, the computational overhead is significant and the quality of the optimized molecules can be variable.[65].

Alternatively, proposing synthesizable molecular candidates using commercially available building blocks and commonly known organic reaction templates[64, 66] offers better synthetic tractability over simple molecule scoring. Importantly, this strategy is appealing to bench and medicinal chemists, since it offers actionable synthesis pathways for them to examine, refine, and execute. Some recent models in this direction apply rule-based synthesis and optimization on building blocks or entire synthetic pathways to generate novel molecules with desired chemical properties[106, 131–135]. Other models condition on input molecules to propose synthetic pathways using commercially available building blocks and well-validated reaction templates for either full construction of the target molecule or the generation of structurally similar analogs in a forward synthesis manner within the predefined chemical search space.[67–69] For example, SynNet[67] constructs synthetic trees via Markov Decision Processes (MDPs) and uses multilayer perceptrons to choose the next action space. More recent models such as ChemProjector[68] and Synformer[69] use transformers to decode for the next action space and have achieved good empirical performances for target and analog molecule reconstruction.

A compelling alternative is the use of Large Language Models (LLMs) due to their foundational nature and adaptability to downstream tasks. [35] LLMs inherently possess extensive chemical knowledge, and recent advancements have focused on extracting and applying this knowledge for predictive and optimization tasks using natural language guidance[37–40]. Furthermore, after fine-tuning, LLM models can perform as good or better than chemical language models trained solely on chemical representations, all while requiring less data.[36] The efficiency and unexpected performance gains from fine-tuning LLMs thus motivates us to explore their potential in more complex tasks, such as synthesis planning, which could pave the way for new chemical discoveries.

Herein, we present SynLlama, an LLM-based tool built on the open-source Llama-3.1-8B and Llama-3.2-1B foundation models [59] to deduce synthetic routes for target molecules or structurally related analogs. Specifically, the LLM component of SynLlama operates as

a constrained retrosynthesis module that breaks input molecules into building blocks (BBs) via well-validated (RXN) sequences, and the reconstruction module searches commercially available BBs based on LLM predictions and builds up molecules within a diverse yet synthesizable chemical space. As an illustration of utility, SynLlama demonstrates competitive performance in key tasks for drug discovery, including synthesis planning for target and analog molecules of pharmaceutical interest and expansion around existing molecular drug hits and leads. Moreover, because of its generative nature, the LLM component of SynLlama has the added ability to explore commercially available building blocks beyond the predefined synthetic space introduced during training - an ability that previous models lack. By integrating molecular design with synthetic feasibility, SynLlama represents a step forward in bridging computational chemistry with synthetic chemistry, providing chemists with actionable and experimentally accessible molecular candidates.

2.2 Methods

The SynLlama workflow, illustrated in Figure 2.1, is designed to generate synthesizable compounds within an expanded chemical space. When an input molecule passes through this workflow, it can either be fully reconstructed through valid synthetic pathways, or the workflow will produce a structurally similar yet synthesizable analog along with its synthesis route. To transform general-purpose LLMs, like the Llama 3 models [59], into expert models for synthetic pathways, we use three key components: 1) a reliable and diverse set of reaction data that covers a large synthesizable chemical space, 2) an efficient supervised fine-tuning (SFT) strategy to train a general-purpose LLM on these reaction data, and 3) a reconstruction algorithm that can convert the output of the fine-tuned LLM into valid synthesis routes, ensuring the proposed molecules lie within a commercially available chemical search space. These components are crucial for leveraging LLMs, which are known to perform well in diverse chemistry tasks [136, 137], to specialize in synthetic modeling.

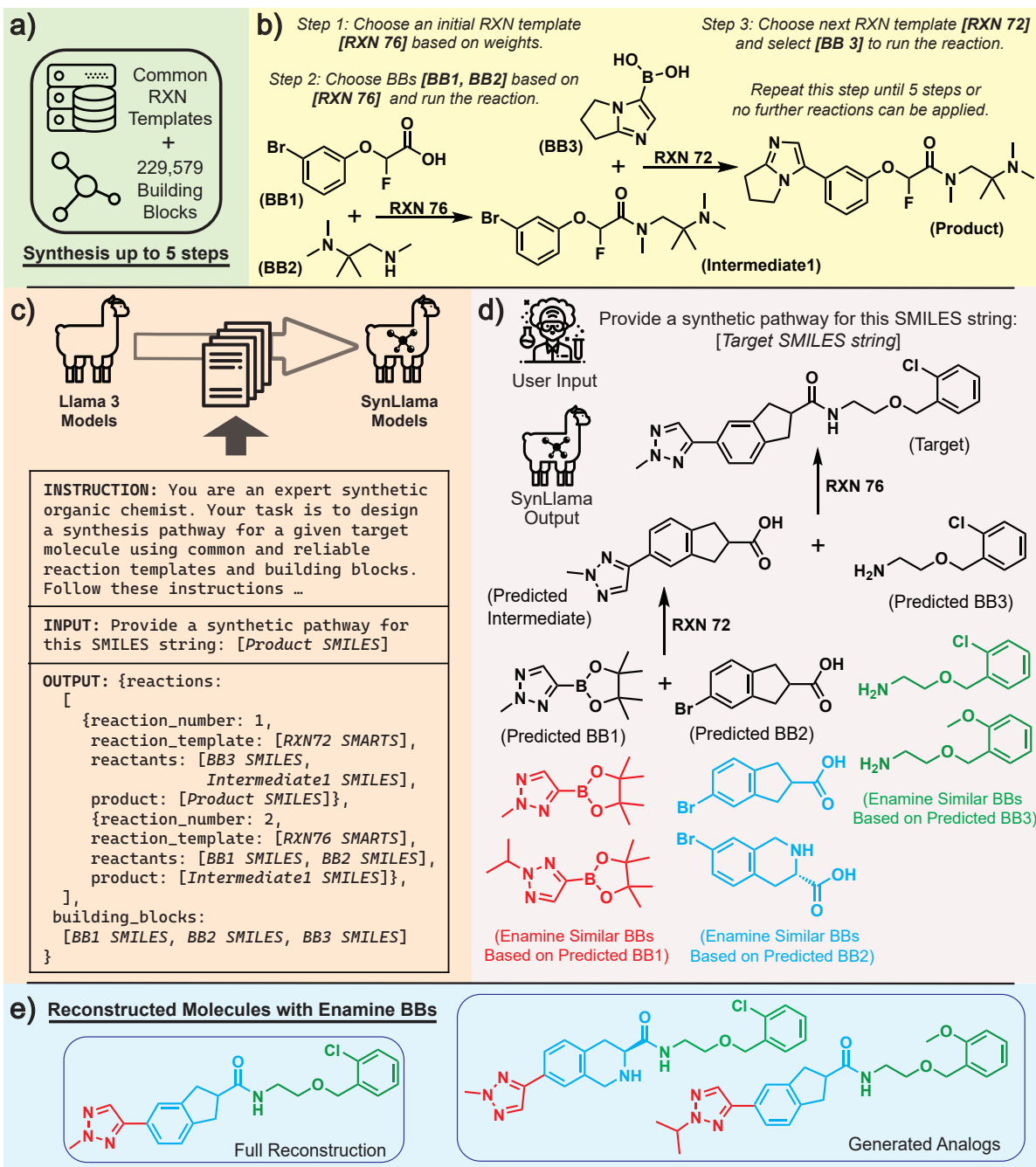


Figure 2.1: **Overview of the SynLlama workflow including data generation, supervised fine-tuning, inference, and reconstruction.** (a). The predefined synthesizable chemical space of reaction templates (RXN) and building blocks (BBs) that covers billions of molecules. (b). An example synthesis data and its generation process from the defined synthesizable chemical space to create training examples. Here, RXN 76 represents amide coupling and RXN 72 represents Suzuki coupling. (c). A schematic representation of supervised fine-tuning that converts Llama 3 models to SynLlama models, along with the instruction, input, and output for the example synthesis in (b). (d). SynLlama’s inference on an unseen test molecule. Black represents SynLlama’s raw retrosynthetic output consisting of RXN sequences and predicted BBs, while colored BBs indicate the top two most similar BBs to the predicted ones from the Enamine building block library. Here, RXN 76 represents amide coupling and RXN 72 represents Suzuki coupling. (e). Reconstructed molecules using the predicted reaction sequences and similar building blocks from the Enamine building block library. In this example, all predicted building blocks are present in the Enamine library, allowing for the complete reconstruction of the input molecule and the generation of close analogs.

Reaction Data for Training and Testing Sets

As illustrated in Figure 2.1(a), our defined chemical space for training consists of molecules that can be synthesized in at most five steps with Enamine building blocks[109] (BBs) and 2 sets of well-validated common organic reactions (RXNs). To define the training and testing BB data, we apply a time split whereby all Enamine BBs from the August 2024 release serve as the training BBs, and all new BBs from their February 2025 release that were not in the training set comprise the testing BBs. This procedure results in $\sim 230,000$ BBs for training and $\sim 13,000$ BBs for testing. Later in Section 3.1 we consider the target reconstruction for unseen molecules which are taken from the Enamine Diversity Set[109] and ChEMBL dataset[108].

We define two sets of reaction templates (RXN) which operate on the Enamine BBs. RXN 1 is formulated as a set of 91 reaction templates selected by Gao et al.[67] from the works of Hartenfeller et al.[138] and Button et al.[139]. RXN 2 is comprised of 115 reactions selected by Gao et al.[69] that contains reactions used to create the Enamine REAL space[109] plus some reactions from RXN 1. All reaction templates are accessible to both sets of BBs, thus defining the training and testing chemical spaces. As a result, there are $\sim 10^{30}$ molecules within this space that can be represented by a synthesis path that comprises a sequence of BBs and RXNs.

To enumerate molecules within this space, we use an iterative approach by selecting RXN templates and searching for compatible BBs. Specifically, as demonstrated in Figure 2.1(b), the selection of the initial RXN is guided by a probabilistic model based on the number of compatible BBs. Within these compatible BB reactants, the initial BBs are selected at random to form an intermediate via the selected RXN template. This intermediate is then used to match for subsequent RXNs and recruits additional BBs to expand the molecular

synthesis pathway until no further reactions are possible or the reaction reaches five steps.

After training on these representations, the resulting LLM will be able to build powerful connections by mapping input molecules to a sequence of BBs and RXNs that creates linear synthesis routes for the testing sets of molecules. Hence, we also construct test sets for the more difficult case of branch synthesis in which molecules have at least one or more reaction steps involving intermediates as reactants. To construct the branching synthesis test sets for both RXN 1 and 2 templates, we keep track of two synthesis trees at a time and check whether the intermediate molecules from both synthesis trees can react further with at least one reaction template from the corresponding RXN. We then filter for molecules that have at least one reaction step with reactants being intermediates to create the two test sets.

Supervised Fine-tuning and Inference from SynLlama

To create the SynLlama model, we need to establish data generation protocols for supervised fine-tuning (SFT) of the Llama 3 models as schematically shown in Figure 2.1(c). When generating reaction data in text format, we choose to represent the BBs and intermediates along the synthetic pathway using SMILES[140] strings, while RXNs are explicitly defined in the SMARTS[141] format. These structured chemical notations are designed to enhance SynLlama’s ability to systematically identify and deconstruct bonds according to RXN templates, effectively dismantling input molecules into building-block-sized fragments.

Since our goal is for SynLlama to learn to link molecules with their synthesis routes, our prompt-response pairs are structured according to retrosynthesis, as depicted in Figure 2.1(c) and shown in detail in Supplementary Figure S1. Such engineered prompts and responses allow the SynLlama model to learn to construct synthesis pathways for the input molecules by inferring sequences of BBs and RXNs, as well as the intermediate steps. While theoretically the model could predict BBs and RXNs without intermediates, we still include them in individual reaction steps in the hope of activating the inherent chemical knowledge in LLMs and enhancing their understanding of synthesis patterns. We have included some additional design choice analysis of forward synthesis versus retrosynthesis and whether to apply a drug-like product molecule filtering in the Supplementary Information.

We have considered both Llama-3.1-8B (8 Billion parameters) and Llama-3.2-1B (1 Billion parameters) for SFT using datasets of varying sizes. Specifically, Llama-3.1-8B is fine-tuned with datasets containing 100k and 500k synthesis routes, requiring 40 and 240 A-40 hours respectively. Llama-3.2-1B, on the other hand, is trained with datasets containing 500k and 2M synthesis entries, requiring approximately 60 and 240 hours respectively. Herein, we refer to the trained models as SynLlama-(parameter count)-(number of reactions trained) in the first part of Results. For example, SynLlama-1B-2M represents a model fine-tuned from Llama-3.2-1B with 2M synthesis routes. Further details of the SFT are provided in the Supplementary Information.

After training the SynLlama models, we apply the consistent prompt setup to perform inferences on molecules. For any given molecule, the SynLlama models predict reaction sequences in SMARTS format and generate SMILES strings for all the reactants, products,

and BBs for the reactions they predict. During inference time, the instruction to SynLlama remains the same, and SMILES strings in the input section are substituted with ones specified by the user. As depicted in Supplementary Figure S1, the responses of the SynLlama models follow the output structures enforced by the prepared training prompt-response pairs. To be more specific, the output response section consists of two parts: reactions and building blocks. In the ‘reactions’ component, the model sequentially deconstructs the target molecule by breaking bonds using provided reaction templates in a retrosynthetic manner. At each step, it predicts a reaction template, along with the reactants and product of the reaction, continuing until no further reactions are possible. Then, in the ‘building blocks’ section, the model compiles all building blocks, namely, reactants from each reaction that are not products in other reactions, identified from the ‘reaction’ section. A visual representation of the inference process is illustrated in black ink in Figure 2.1(d).

SynLlama Model Benchmarks

Since we are formulating the synthetic tasks using purely language-based modeling, where all reactions are expressed in SMARTS templates and molecules in SMILES strings, it is important to quantify the capacity of SynLlama for instruction following and comprehension of reaction chemistry. To assess SynLlama’s ability to follow instructions, we select three benchmarking criteria as shown in Table S2.3. The first is “Valid JSON,” which examines whether the output format is a parsable JSON following the fine-tuned templates that will be necessary for the downstream reconstruction algorithm. The second criterion is “Template Memorization,” which assesses the model’s ability to memorize the provided reaction templates that define our synthesizable chemical space. Lastly, we benchmark on “BB Selection,” which evaluates whether the “building blocks” section in the responses can accurately identify and select all the building blocks from the “reactions” section of the responses.

To assess SynLlama’s comprehension of reaction chemistry, we focus on individual reactions and summarize the three critical aspects as: (1) the percentage of “Valid SMILES” out of all SMILES strings in the responses, which is essential for assessing SynLlama’s learning outcome of string-based chemical representations in general, (2) the percentage of “Matched Reactants,” which calculates whether the generated reactants match the reactant templates specified in the predicted reactions, and (3) the percentage of “Good Products,” which assess if the predicted product can indeed be generated by applying the proposed reaction templates onto the reactants. Overall, these six benchmarks can collectively assess SynLlama’s capability to follow instructions and perform chemical reactions in string representations.

Dataset	Category	Model Config			
		8B-100k	8B-500k	1B-500k	1B-2M
Training Data	Valid JSON	96.20%	97.20%	96.60%	98.00%
	Template Mem.	99.95%	100.0%	100.0%	100.0%
	BB Selection	99.80%	100.0%	99.72%	99.96%
	Valid SMILES	94.74%	99.53%	95.17%	99.46%
	Matched Reactants	78.62%	96.42%	80.19%	97.64%
	Good Products	78.34%	96.58%	81.26%	98.58%
Testing Data	Valid JSON	91.00%	94.20%	88.70%	93.90%
	Template Mem.	99.91%	100.0%	99.97%	100.0%
	BB Selection	99.75%	99.96%	99.73%	99.98%
	Valid SMILES	94.37%	99.13%	87.66%	99.50%
	Matched Reactants	77.51%	94.83%	65.63%	96.90%
	Good Products	74.11%	94.16%	69.54%	96.39%
ChEMBL Data	Valid JSON	98.80%	99.00%	99.20%	99.00%
	Template Mem.	99.90%	99.82%	99.37%	99.82%
	BB Selection	99.57%	99.23%	99.50%	99.47%
	Valid SMILES	92.02 %	96.38%	95.86%	95.23%
	Matched Reactants	54.52%	69.25%	64.62%	70.93%
	Good Products	67.69%	85.03%	75.81%	87.02%

Table 2.1: **Benchmarks of SynLlama inferences on 1000 training, testing, and ChEMBL data.** We first select 1000 SMILES strings from the training examples, testing examples and the ChEMBL dataset, and then run inferences using SynLlama models trained on RXN 1. Benchmarking result for SynLlama models trained on RXN 2 can be found in Supplementary Table S1. The detailed descriptions of each benchmark can be found in the main text. Here, we run SynLlama inferences at $T = 0.1$ and $TopP = 0.1$ to generate reproducible benchmarking results (see Supplementary Table S2).

In Table S2.3, all four trained SynLlama models are evaluated on both in-distribution training data and out-of-distribution testing and ChEMBL [108] data to assess the benchmarks outlined above. In the instruction-following benchmarks, most models exhibit strong adherence (over 90 percent) to the fine-tuned response structure across all datasets. This impressive performance indicates that fine-tuning effectively retains the specified output structure when trained with over 100,000 samples. Furthermore, all four models successfully memorized the provided RXN templates and selected the building blocks (BBs) from all predicted reactants over 99 percent of the time. This capability further enhances the coupling effectiveness of the downstream reconstruction algorithm with the SynLlama raw output, as it only requires information about reaction sequences and predicted building blocks.

In the reaction chemistry benchmarking results, a clearer trend emerges: models, regardless of their parameter size, show improved comprehension of reaction chemistry in all three datasets as the amount of training data increases. Notably, most models maintain their performance from training to testing data, but exhibit a greater decline in “Matched Reactants” and “Good Products” performance when generalizing to the ChEMBL data. The reason behind this is that the testing data are generated in the same manner as the training data but with a different set of building blocks, while the ChEMBL data occupies a different chemical space, as previously noted by Gao et al. [67]. Despite the reductions in their performance for ChEMBL molecules, as shown in Supplementary Figure S2, SynLlama-8B-500k and SynLlama-1B-2M can still generate complete and valid syntheses over 50% of the time without any downstream processing. These results indicate that SynLlama’s raw results alone have potential utility for synthesis planning for unseen drug-like molecules.

When comparing SynLlama-8B-500k and SynLlama-1B-500k, we observe that the larger model demonstrates better performance when trained on the same amount of data. Although additional training data could further improve the 8B model based on the current trend, its higher computational cost makes this pursuit less practical. However, as the fine-tuning computational costs for SynLlama-8B-500k and SynLlama-1B-2M require approximately the same A40-GPU hours, and given the comparable benchmark performance between them, we decided to move forward with SynLlama-1B-2M, simplified as SynLlama, for the subsequent tasks due to its faster inference speed.

Reconstruction from Predicted Retrosynthesis

Using the predicted sequence of RXNs and BBs from SynLlama responses, we can synthesize the proposed target molecule or close analogs by applying the predicted reaction templates to the BBs in the inferred order, as shown in black ink in Figure 2.1(d). In some cases, the predicted BBs match known Enamine BBs, ensuring that the resulting molecules remain within an established chemical space for synthesis. However, due to SynLlama’s generative nature, some predicted BBs are novel while still providing valid synthesis pathways, and we only report new BBs that can be purchased from other suppliers identified by Molport [142]. Therefore, while SynLlama primarily produces molecules within the predefined chemical space using Enamine BBs, its output also offers an alternative strategy for molecule construction. We will revisit this point in the Results section.

When the input molecule cannot be fully reconstructed, we generate analogs by mapping the predicted BBs from SynLlama to known Enamine BBs, thereby sampling molecules from the well-defined Enamine chemical space. Under this scenario, we use nearest neighbor search algorithms with different molecular representations (SMILES and Morgan Fingerprints[143]) to sample Enamine neighboring BBs from the predicted BBs, as illustrated in colored inks in Figure 2.1(d). Since in SynLlama’s output, the RXN sequences are predicted concurrently with the BBs, our effective search space is constrained to Enamine BBs that can react through the specific RXN template. This smaller Enamine search space not only allows us to ensure the success rate of such forward syntheses but also allows us to effectively explore

segments of the input molecule. Further details of the nearest neighbor search algorithms are provided in the Supplementary Information.

When constructing full synthetic pathways for reactions with multiple possible products, we select the product that most closely matches the predicted product based on SMILES string similarity. As shown in Figure 2.1(e), the reconstruction algorithm iteratively builds synthesis routes, utilizing all predicted BBs and RXN sequences to reconstruct or generate variations of the original molecule from the synthesizable chemical space. This reconstruction algorithm enables the SynLlama model to function as a generator for synthesizable molecules along with their corresponding synthetic pathways.

2.3 Results

We examine SynLlama’s performance in synthesis planning of a diverse set of previously unseen compounds. We also explore the utility of the SynLlama workflow in real-world drug discovery applications, including its integration with generative algorithms to enhance the synthetic accessibility of proposed molecules while preserving their chemical properties and expanding the library of active compounds in the defined synthesizable space with as good or improved binding affinity metrics.

Synthesis Planning for Unseen Molecules

Having demonstrated that SynLlama models can reliably predict reaction sequences and building blocks in Table 1, we now use SynLlama to plan the synthesis of two groups of 1000 previously unseen molecules from the Enamine Diversity Set[109] and the publicly available ChEMBL database[108]. These datasets are specific to drug-like molecules, unlike the training data used for SynLlama; the drug-related property distribution of both sets of molecules against the training data is shown in Figure S3. In this validation, we test whether known synthesizable molecules can be reconstructed accurately from the baseline and SynLlama models as summarized in Table 2.

We first consider the standard reconstruction approach used by algorithms such as SynNet[67], ChemProjector[68], and Synformer[69] to create target molecules or analogs using BBs exclusively from the Enamine library. In this comparison, SynNet[67] and ChemProjector[68] serve as a baseline comparison for synthesizable chemical space coverage for RXN 1, whereas Synformer[69] is the baseline comparison on the expanded RXN 2 templates. As seen in the first column of Table 2.2 and Supplementary Table S3, when trained with their respective reaction sets and using only Enamine BBs, SynLlama outperforms all three methods while reducing the number of training data by 40-to 60-fold.

Although SynLlama yields higher percentages of successful ChEMBL reconstructions compared to SynNet and ChemProjector, and is on par with Synformer when only using Enamine BBs, there is a degradation of performance across all methods for ChEMBL compared to the Enamine Diversity set reconstructions. We attempted to improve upon the ChEMBL result by reformulating the training reaction data with extra filtering such that the product molecule distributions conform to a similar drug-like property distribution as

the ChEMBL set as seen in Supplementary Figure S4. We then performed supervised fine-tuning following the same procedure as described in Section 2, but now with this filtered set of training data, in hope of better generating synthetic pathways for molecules with drug-like properties. However, as shown in Supplementary Table S3, there is now a performance loss over both the Enamine Diversity and ChEMBL drug-like targets. This result suggests that training on more diverse product molecules ultimately benefits synthesis planning for the drug-like targets more than specializing the LLM further with more restrictive training data. In addition, the curated ChEMBL data appears unique and outside the synthesizable chemical space made up of Enamine BBs and RXN templates.

Dataset	Method	# of Recon. Mol.			Morgan Sim.
		Enamine BB	New BB	Total	
Enamine Diversity Set	SynNet	110	-	110	0.57
	ChemProjector	462	-	462	0.82
	Synformer	660	-	660	0.91
	SynLlama(RXN 1)	527	100	568	0.87
	SynLlama(RXN 2)	691	232	741	0.92
ChEMBL Data	SynNet	54	-	54	0.43
	ChemProjector	133	-	133	0.60
	Synformer	198	-	198	0.67
	SynLlama(RXN 1)	165	95	223	0.66
	SynLlama(RXN 2)	197	152	287	0.68

Table 2.2: **Comparison of synthesis planning performance among different methods.** Both the Enamine Diversity Set and ChEMBL Data are comprised of 1000 unseen molecules each. Details of each benchmark are described in the main text. The Morgan similarity scores include all analog molecules with successful synthesis pathways, as well as successfully reconstructed target molecules. The number of total training data and reaction set each method used for is: SynNet[67] (**200K**, RXN 1) ChemProjector[68] (**128M**, RXN 1), Synformer[69] (**85M**, RXN 2), and SynLlama (**2M**, RXN 1 & 2). Further details are provided in Supplementary Tables S4 and S5, and identification of purchasable New BBs with Molport[142] are described in the Supplementary Information.

However, unlike baseline methods that only generate molecules using Enamine BBs, SynLlama has the extra capacity of reconstruct target molecules with commercially available BBs beyond Enamine due to its generative capabilities, even without specific training for this purpose. As seen in Table 2.2 the ‘New BBs’, restricted to those purchasable through Molport[142], add possible synthetic pathways to reconstruct target molecules in all datasets and RXN templates. This also helps the ChEMBL data as well given that the molecules generated with the new BBs remain drug-like (Supplementary Figure S5). Since some target

molecules can be synthesized through multiple pathways, either using only Enamine BBs or with the addition of New BBs, the ‘Total’ column in Table 2.2 reflects the number of unique target molecules reconstructed with SynLlama. With these New BBs, SynLlama’s best reconstruction rates increase to 74.1% for the 1000 molecules in the Enamine Diversity Set, and encouragingly to 28.7% for the ChEMBL data. These results show that SynLlama learns reaction chemistry such that it can predict novel BBs to increase synthetic accessibility.

When the target molecule cannot be reconstructed, we assess the quality of the analog using a molecular similarity score between the target molecule and its most similar analog using Tanimoto similarity based on 4096-bit Morgan fingerprints[143]. In Table 2.2, the similarity metrics reported are average values of all generated molecules, including target molecules that are fully reconstructed (with a score of 1). Tables S6 and S7 also provide similarity metrics based on the 4096-bit Morgan fingerprints of Murcko scaffolds[144], and Gobbi 2D pharmacophore fingerprints[145], while including or excluding target molecules that are fully reconstructed. Overall, these results collectively show that SynLlama is highly capable of planning synthesis for related analog molecules with very good similarity, aided most by increased synthetic pathways using purchasable building blocks.

Finally, Supplementary Table S3 also considers reconstruction performances for the Enamine Diversity Set and ChEMBL Data based on forward synthesis as opposed to retrosynthesis, and for test molecules derived from tree-like synthesis pathways. It is evident that SynLlama performs better for retrosynthesis relative to forward synthesis that is used more successfully by Synformer, and retrosynthesis also performs better than the baseline methods for branching synthetic pathways.

Synthesizable Analog Search for *De Novo* Molecules

Previous research has shown that molecules proposed by generative models for binding to protein targets often face challenges in both reliability and practical synthesizability[7, 34, 123]. In particular, medicinal chemists are often reluctant to devote time and expensive resources to specialized synthesis of hit molecules due to the high false positive rates arising from the unreliability of docking scores in drug discovery. Instead finding closely related compounds that are constrained to Enamine BBs allows for an inexpensive purchase to verify hits before further refinements are deployed to gain lead drug molecules. In this section, we demonstrate SynLlama’s potential to bridge the gap between generative molecule design and practical synthesis for molecules that are optimized for drug-like applications such as binding assays.

Specifically, we consider two different generative methods for de novo drug molecules in Figure 2: a 1D-to-3D LSTM model, iMiner[52], which optimizes drug-likeness and AutoDock-vina[146] docking scores, and Pocket2Mol[118], a 3D graph transformer model which also optimizes AutoDock-vina[146] docking scores. Using iMiner we generated 500 molecules for the SARS-CoV-2 Main Protease (SARS2 MPro)[147], and used Pocket2Mol to generate 500 molecules each for the protein targets Thrombin[148] and TYK2[149, 150]. The first interesting observation is that only 1% of generated molecules from both iMiner and Pocket2Mol

can be synthetically reconstructed using SynLlama or ChemProjector, emphasizing that the generative models create difficult synthesis targets.

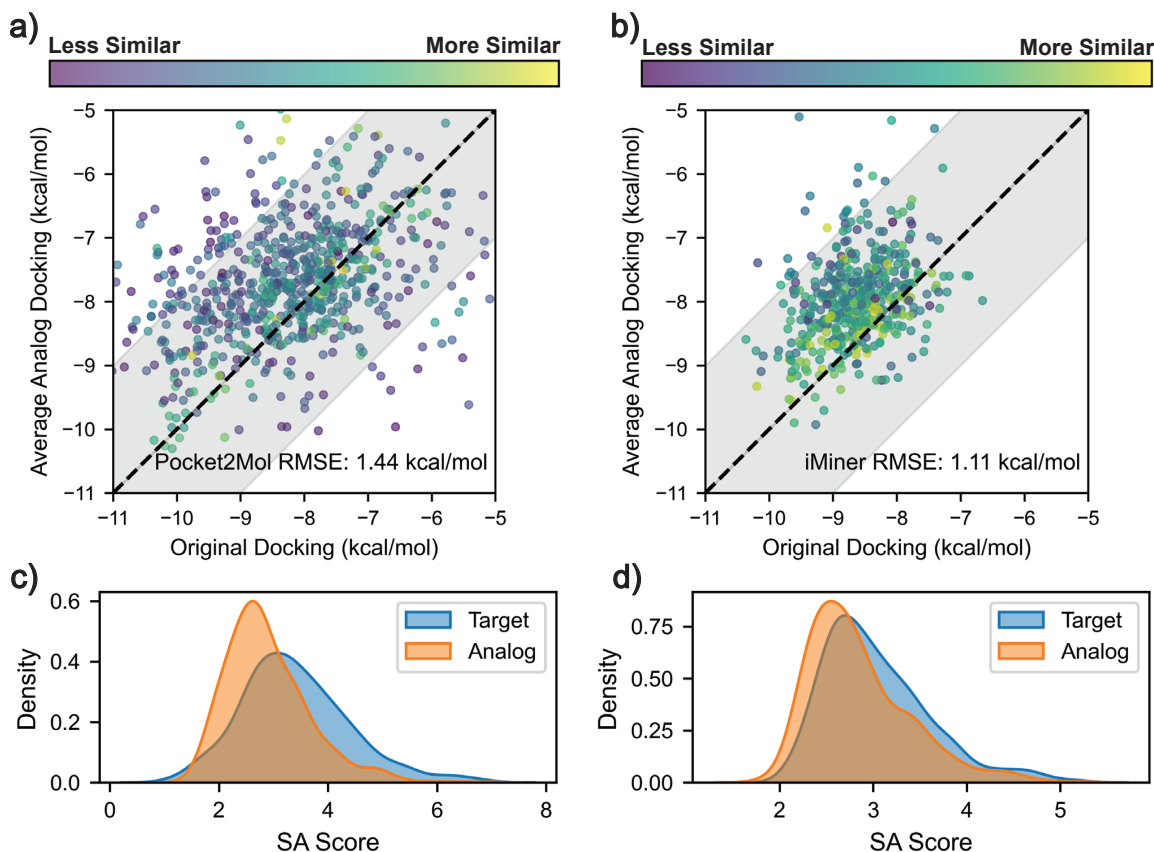


Figure 2.2: **SynLlama performance on generating synthesizable analogs for Pocket2Mol and iMiner proposed binders of SARS2 MPro[147], Thrombin[148], and TYK2[149, 150].** Correlation plot comparing docking scores of (a) Pocket2Mol and (b) iMiner generated molecules and the average Vina docking scores of ten most similar analogs from SynLlama trained with RXN 2. Each data point is color-coded by the average Morgan fingerprint similarity computed between the generated and analog molecules. The shaded area represents an energy uncertainty range of ± 2 kcal/mol for docking[146]. (c, d) Kernel density estimates of synthetic accessibility (SA) score distributions for target generated molecules and SynLlama-proposed analogs, with (c) showing the Pocket2Mol comparison and (d) showing the iMiner comparison.

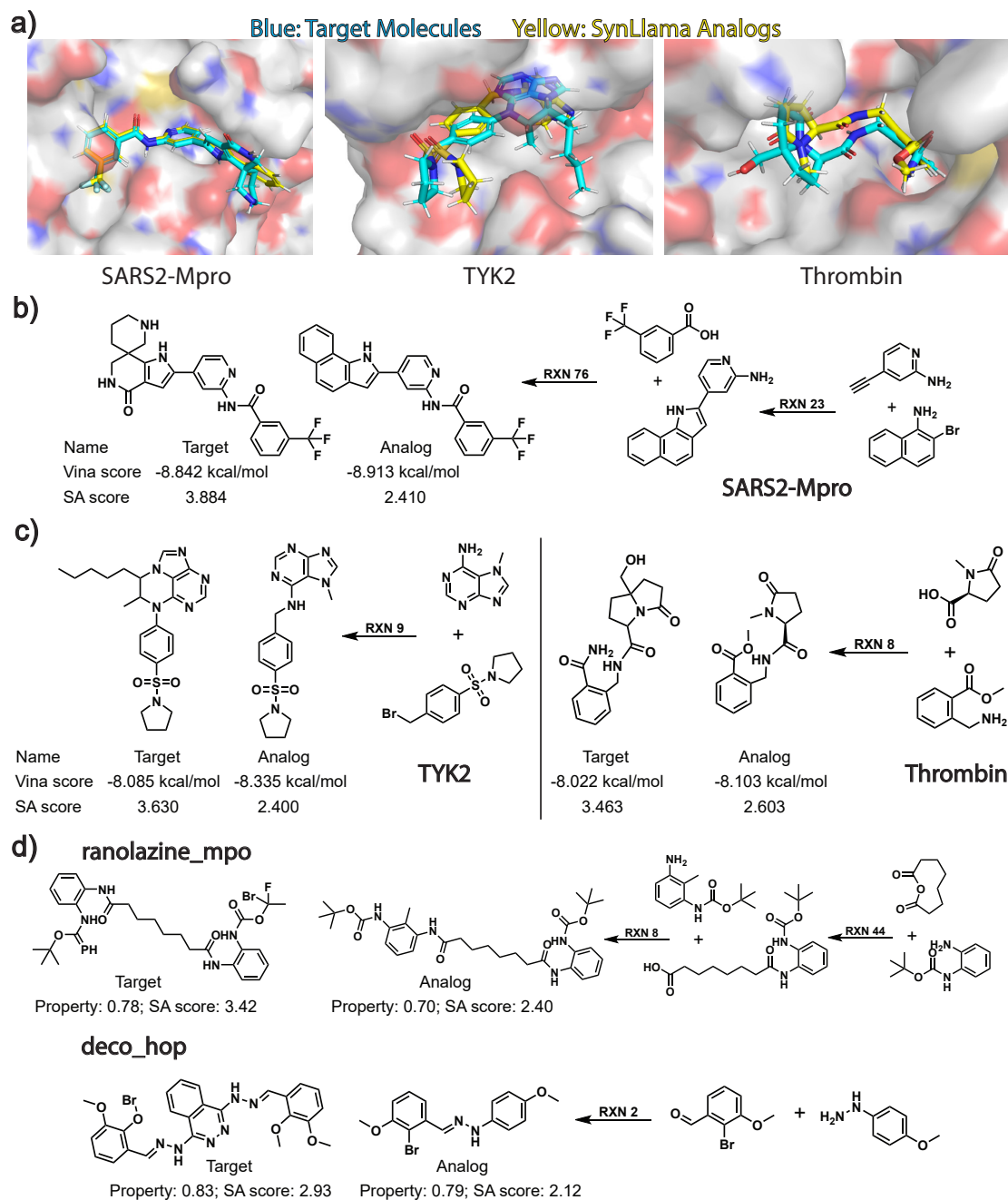


Figure 2.3: Examples of synthesizable analog generation for SARS2 MPro using iMiner and TYK2 and Thrombin with Pocket2Mol. (a) Docked pose visualization for all three protein targets. (b) Docking and SA scores for iMiner target and SynLlama analog for SARS2 MPro along with the predicted synthetic pathway. (c) Docking and SA scores for the Pocket2Mol targets and the SynLlama analogs for TYK2 and Thrombin along with their predicted synthetic pathways. (d) Two examples of ASKCOS-unsynthesizable target molecules (ranolazine_mpo and deco_hop) and their SynLlama – proposed analogs, showing property scores, SA scores, and predicted synthetic pathways.

Hence all 500 compounds for each protein target were then processed through SynLlama trained on RXN 2 to generate synthesizable analogs constrained to Enamine BBs. For all generated analogs we performed molecular docking with AutoDock-vina via using the same protocol as the original binders.

Figure 2.2(a,b) shows the RMSE of the docking scores between the target compounds and the generated analogs are 1.44 kcal/mol and 1.11 kcal/mol for Pocket2Mol and iMiner, respectively, both of which are within the acceptable range of inherent docking score errors reported by Trott et al[146]. Furthermore, as demonstrated in Figures 2.2(c,d), the SA score distribution of the SynLlama analogs generated for thrombin and TYK2 from Pocket2Mol show a notable improvement in synthetic accessibility at the expense of reduced similarity below the benchmarks in Table 2. SynLlama slightly improves SA for iMiner generated compounds for SARS2 MPro while maintaining a good similarity score on par with the benchmarks in Table 2. The fact that iMiner uses a drug-likeness score as part of its loss function[52] may explain its better performance, or perhaps SARS2 MPro is an easier protein target than Pocket2Mol’s thrombin and TYK2 protein targets.

The better synthesizable analogs have good retention of the binding mode of the original generated molecules, visually confirmed by the representative docking poses for target-analog molecular pairs shown in Figure 2.3(a) for the three proteins. Figure 2.3(b,c) provides a few examples of the synthesis pathways for the analogs of the molecules derived from the generative molecules.

The final comparison consists of molecules that were identified as unsynthesizable by Gao et al[7] via ASKCOS[66]. As also seen in Figure 2, the similarity scores of SynLlama analogs for this set are better than that observed for the Pocket2Mol generative model, and the generated analogs show a significant decrease in SA score, representing SynLlama’s effective strategy of improving synthetic accessibility via analog generation. In Supplementary Figure S8, we also highlight the few concrete examples of target-analog pairs where objective scores are maintained while target synthetic accessibility scores are substantially reduced. Overall, these collective results highlight SynLlama’s utility in effectively proposing synthesizable analogs for *de novo* molecules, thereby enhancing their synthetic accessibility without compromising their desired drug-related properties.

Local Hit Expansion for Binder Molecules

Because SynLlama breaks down the original target molecule for synthesis into building blocks, by nature this method allows diverse exploration around parts of the molecular scaffold rather than only on a whole target molecule. In a final task, we apply SynLlama to expand on hit molecules for three protein targets used in the previous section, SARS2 Mpro[151], Thrombin[148], and TYK2[149, 150] to discover synthesizable molecules that have better relative binding free energies (RBFs) confirmed by both experiments and accurate free energy perturbation (FEP) calculations.

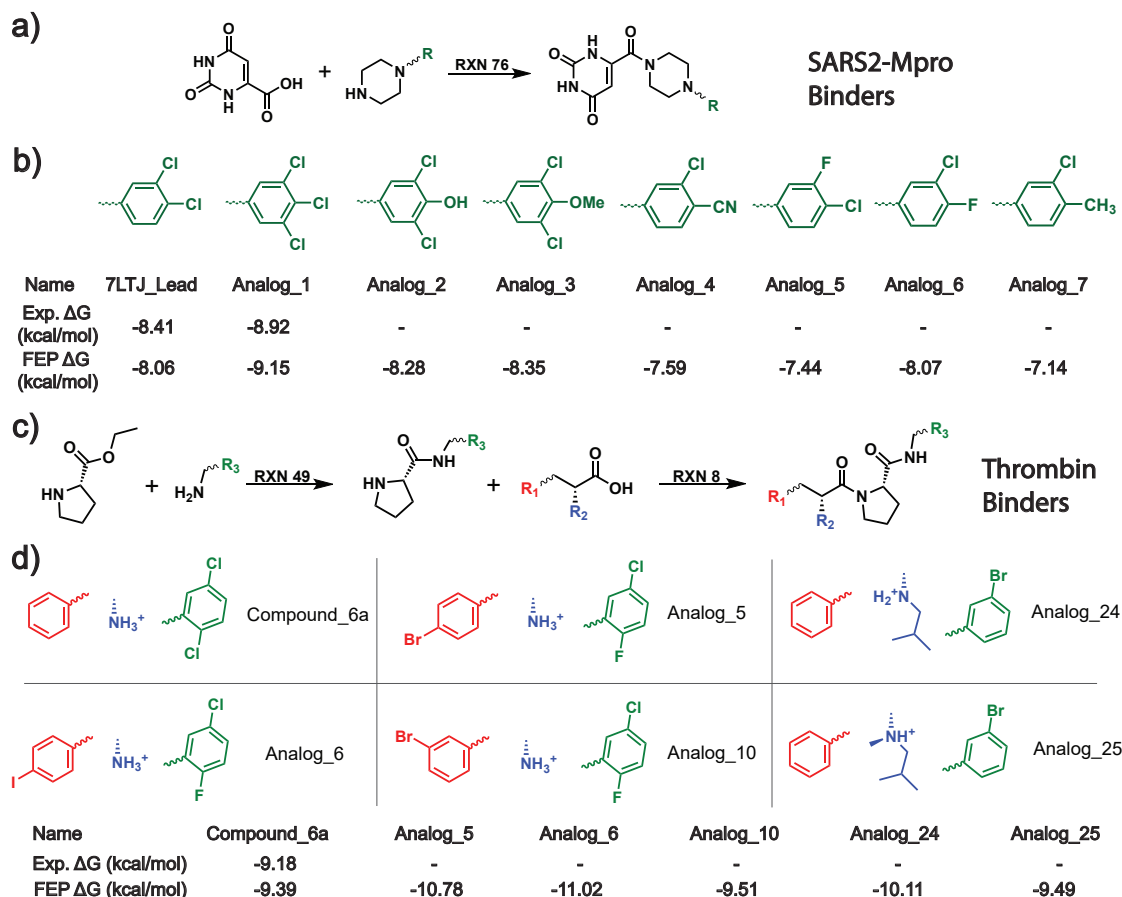


Figure 2.4: **Hit expansion of binders to SARS2 Mpro, Thrombin, and Tyk2 with SynLlama.** (a, c) SynLlama-predicted synthetic pathways that expand on the hit molecules for each protein target; substitution sites are labeled as R groups. (b, d) Binding free energies of the hit compounds and SynLlama-expanded analogs, with the proposed substitutions rendered in 2D using the same color scheme as the synthetic pathways. All potential binders either improve on the parent hit's FEP binding free energy or fall within the 1 kcal/mol uncertainty range.

As shown in Figure 2.4(a), the hit molecule for SARS2 Mpro (7LTJ_Lead) has a core scaffold of uracil and ortho-dichlorobenzene connected by a piperazine linker. Inspired by an experimental hit expansion campaign by Kneller et al. [152], we follow their practice to propose only functional group substitutions on the benzene ring while keeping the linker and uracil intact. In Figures 2.4(c) and 2.4(e), we use the best-performing molecules from the Schrodinger FEP benchmarking set[153] as our hits for Thrombin and TYK2. To expand on these hit molecules, we first identify the maximum common scaffolds among each group of molecules in the FEP benchmarking set and use the identified scaffolds to guide our selection of analog molecules. Specifically, we use SynLlama to generate 50 synthesizable analogs constrained to only Enamine BBs of the hit compound and filter for molecules that

retain the scaffold. In the end, we harvest a total of 8, 14, and 11 analog molecules that fulfill the criteria for SARS2 Mpro, Thrombin, and TYK2, respectively. The analogs are then placed in a pose configuration similar to the original hit molecule for downstream FEP calculations.

To verify the FEP results, we first choose ~ 10 synthesized and experimentally tested molecules to benchmark the accuracy of FEP for all three systems. In Supplementary Figure S9, all the calculated FEP values show a good correlation with the experimental ΔG converted from IC50, with an average RMSE of less than 1 kcal/mol. After this validation, we run FEP for the all proposed molecules to assess their binding affinities. As shown in Figures 2.4(b), 2.4(d), and 2.4(f), a significant portion of the proposed analogs (7 of 8, 5 of 14, and 6 of 11, respectively) showed potency compared to their parent hits for SARS2 Mpro, Thrombin, and TYK2. These results successfully demonstrate that SynLlama can propose diverse yet potent analogs when constrained on a molecular scaffold.

Moreover, the fact that all suggested analogs also come with predicted pathways using common reaction templates and purchasable BBs from Enamine suggests SynLlama’s practical use for drug discovery. Because of this composite capacity of optimizing both potency and synthetic accessibility for small molecules, SynLlama successfully rediscovered Analog_1, the most potent compound reported by Kneller et al. in their hit expansion campaign for SARS2 Mpro[152]. Furthermore, for the Thrombin case, SynLlama explores the R2 substitution site, a region previously unaddressed by earlier molecular series, and demonstrates the generation of more potent molecules via FEP. These results confirm that SynLlama effectively explores local chemistry with readily available building blocks, providing a direct and efficient path for medicinal chemists to accelerate hit expansion.

2.4 Discussion and Conclusion

Motivated by recent advances in LLMs for chemistry[36–39], we aim to leverage data-efficient supervised fine-tuning (SFT) to transform the general-purpose Meta Llama 3 into SynLlama, an LLM-based generator capable of proposing synthesizable molecules and deducing synthetic routes for target molecules or their close analogs. Throughout the study, we successfully show that SynLlama can effectively explore a custom-defined chemical search space composed of around 230,000 Enamine building blocks (BBs) and well-validated organic reactions (RXNs), after it has been fine-tuned on synthetic pathway data sampled from this specified chemical space. What’s more, despite utilizing nearly two orders of magnitude fewer synthetic pathways in training, SynLlama exhibits strong performance in key drug discovery tasks compared to existing models. Specifically, we have demonstrated that SynLlama can effectively aid in various stages of drug discovery that include synthesis planning, synthesizable analog generation for *de novo* molecules, and local hit expansion.

Because SynLlama is built on a general-purpose LLM instead of training from scratch[67–69], it offers a number of unique advantages and possibilities for further improvement. For example, when generating our fine-tuning data, we sampled from the predefined chemical

search space of 230K Enamine building blocks (BBs) and two sets of reaction templates (RXNs), but we did not embed these extensive requirements in the context window of the LLM. As a result, for our largest dataset with only 2 million synthetic pathways, the model only saw each BB a few dozen times, while each RXN template appeared hundreds of thousands of times. Consequently, while SynLlama efficiently memorizes the allowed RXNs, it only captures the distribution of Enamine BBs, which enables SynLlama to extrapolate to unseen yet purchasable building blocks outside of Enamine. This generative ability surpasses other existing methods such as ChemProjector and Synformer that can only explore a predefined building block search space, such as within the Enamine Diversity Set[109], and limits their ability to propose alternative synthesis pathway with novel building blocks.

In addition to its ability to extrapolate outside the training chemical space, the underlying Llama-3.2-1B used by SynLlama is relatively small and more predictive power would be expected if we train on larger LLMs with more data and compute power. However, we observe that a smaller LLM with fewer parameters can be turned into an expert model for complex tasks after SFT with relatively little data. This opens up opportunities to employ smaller expert models for various chemical tasks, benefiting from faster inference speeds, which can make these models more desirable. Moreover, optimal hyperparameters like temperature and top-p can vary between training and inference phases, depending on the downstream tasks. During inference, most valid raw outputs are generated under relatively low temperature and top-p settings. However, when the model is paired with reconstruction algorithms that require less strict adherence to reaction chemistry, higher temperature and top-p values can be used. This allows for a broader exploration of the Enamine chemical space, enabling the generation of more diverse and relevant analogs. This property is especially desirable in tasks that require extensive exploration, such as the hit expansion example we demonstrate. Another exciting direction is coupling SynLlama with another generative model, in which we have shown generates analogs while maintaining good docking scores and simultaneously shifting to better synthetic accessibility scores. This result suggests that SynLlama can serve effectively as a post-processor for other *de novo* generative models, ensuring the production of more synthesizable compounds with clear reaction pathways.

Among the numerous opportunities that LLMs bring to the field of drug discovery, their natural language capabilities and recent advancements in reasoning are the most exciting features that allow users without coding expertise to interact directly with the models, effectively bridging the gap between computational methods and experimental research. We envision that expert users can employ prompt engineering and fine-tuning data to incorporate more realistic factors than those explored here. For instance, medicinal chemists could fine-tune LLMs within this generalizable SFT framework with building blocks and reaction templates of their own choice. In addition, they can consider synthesis cost, reaction conditions, improved selectivity, and protection factors at specific reaction steps for more detailed and powerful synthesis planning. We see our work as an initial attempt to demonstrate the effectiveness of LLMs in real experimental research, encouraging further studies for better utilization of these models.

2.5 Supplementary Information

Additional Methodology Details

Supervised Fine Tuning protocol. After preparing the reaction data and prompt-response pairs from the training chemical space, we fine-tune Llama-3.1-8B (8 Billion parameters) and Llama-3.2-1B (1 Billion parameters) using the Axolotl package [59, 154] for 1 epoch. LLMs with more parameters require more resources to train and use, but they also typically perform better on a variety of tasks, which we consider in Results. For our SFT approach, we apply Low-Rank Adaptation (LoRA) with a rank of $r = 32$ and $\alpha = 16$ to the linear layers of the model.[155] We use FlashAttention-2[156], with the Adam optimizer[157], cross-entropy loss, and a cosine learning rate scheduler with a maximum learning rate of 2×10^{-5} .

Molecule Generation using Enamine BBs. When searching for the nearest neighbors of BBs, a natural choice is to perform a string-level similarity search based on SMILES strings, as this is the native format of SynLlama responses. For each RXN template, we systematically process all SMILES strings of its compatible building blocks that can participate in the reaction. First, we extract the full vocabulary of SMILES tokens and generate an n-gram representation by considering all possible consecutive token pairs (bigrams) and triplets (trigrams). Next, we identify the 1024 most frequently occurring n-grams across SMILES strings of all compatible BBs to form a representative token set for each individual RXN template. To facilitate efficient retrieval, we structure search trees based on the term frequency-inverse document frequency (TF-IDF) scores [158] of these n-grams, prioritizing highly informative substructures and accelerating inference. Consequently, when a new SMILES string of the predicted BB is introduced, it can be efficiently processed through the tree, yielding a list of the top K matching SMILES strings.

In addition, Gao et al.[67] investigated using Morgan fingerprints[143], a molecular representation capturing local chemical environment, to search for the nearest neighbors of BBs based on their Tanimoto similarity [159]. Similarly to that stated above, for each RXN template, we also build a separate search tree for all compatible Enamine BBs using 256-bit Morgan fingerprint representation with a searching radius of 2. Our empirical observations indicate that combining the top K molecules from both the SMILES and Morgan fingerprint methods offers better performance than relying on the top 2K molecules from a single method. However, since we are working with an LLM model, the generated SMILES strings still have a small chance of being invalid, which prevents us from calculating their Morgan fingerprints. Therefore, we employ the both combined TF-IDF and Morgan fingerprint search trees when dealing with valid molecules, and revert to only a SMILES-based search when the generated SMILES strings are invalid.

LLM Inference Hyperparameters for Various Tasks. A key advantage of SynLlama, and LLMs in general, is their sensitivity to variations in hyperparameters, such as temperature (T) and top-p ($TopP$), which can significantly impact the performance of reconstruction

and analog similarity. As shown in Supplementary Figure S2, SynLlama’s raw outputs exhibit enhanced reaction chemistry comprehension when inferences are run at lower $TopP$ and within a reasonable range of T for both test sets. This configuration allows SynLlama to explore purchasable building blocks outside the Enamine library while maintaining synthesis validity. Conversely, increasing T and $TopP$ generally reduces SynLlama’s ability to generate valid syntheses in its raw outputs. However, as Supplementary Figure S2 also illustrates, inferring with higher T and $TopP$ values than the optimal settings in raw outputs often leads to better overall average maximum similarity scores for reconstruction with Enamine BBs along. Nonetheless, excessively high settings can increase the failure rate.

Based on empirical observations, we recommend specific combinations of T and $TopP$ that effectively span a broad spectrum of tasks. These combinations optimize SynLlama’s performance by balancing exploration and precision during inference.

- **Frozen:** $T = 0.1, TopP = 0.1$, repeated once. This setting prioritizes deterministic generation, ensuring minimal variability and high reproducibility.
- **Low:** $T = 0.6, TopP = 0.5$, repeated multiple times. This configuration allows for limited exploration while maintaining a degree of precision.
- **Medium:** $T = 1.0, TopP = 0.7$, repeated multiple times. This setting balances exploration and diversity, generating outputs with moderate randomness.
- **High:** $T = 1.5, TopP = 0.9$, repeated multiple times. This configuration promotes high diversity and creativity in generation but may introduce more variability in results.

We define different sampling strategies based on these core settings:

- **Frugal Sampling:** A total of 4 inferences.
 - $T = 0.1, TopP = 0.1$, repeated one time.
 - $T = 0.6, TopP = 0.5$, repeated one time.
 - $T = 1.0, TopP = 0.7$, repeated one time.
 - $T = 1.5, TopP = 0.9$, repeated one time.
- **Greedy Sampling:** A total of 10 inferences.
 - $T = 0.1, TopP = 0.1$, repeated one time.
 - $T = 0.6, TopP = 0.5$, repeated two times.
 - $T = 1.0, TopP = 0.7$, repeated three times.
 - $T = 1.5, TopP = 0.9$, repeated four times.
- **Frozen Only:** A total of 1 inference.

- $T = 0.1, TopP = 0.1$, repeated one time.
- **Low Only:** A total of 5 inferences.
 - $T = 0.6, TopP = 0.5$, repeated five times.
- **Medium Only:** A total of 5 inferences.
 - $T = 1.0, TopP = 0.7$, repeated five times.
- **High Only:** A total of 5 inferences.
 - $T = 1.5, TopP = 0.9$, repeated five times.

Baseline Benchmarking Details. Since the Enamine BB catalog constantly updates new BBs and does not store historical data, we cannot access the exact training BBs used in training for the baseline methods ChemProjector[68] and Synformer[69]. The only training set data we had access to is described in Section 2.1 in the main document, and is $\sim 3\%$ (10k) more compared to that available to Synformer and Chemprojector (cutoff at October 2023). However, we highlight that 97% of our training data is identical to their previous work and the newly added building blocks (which is equivalent to a time split) show a similar distribution as the rest of the training BBs. We now have this comparison in Supplementary Figure S5. Therefore, we can fairly say that our training data are very similar to the baseline methods to which we compare. For a fair comparison at the inference stage, we provide ChemProjector and Synformer the same set of building blocks (cutoff at Feb. 2025) that SynLlama had access to during inference time. In Tables 2 and S3, we assess the performance of the trained baseline models with this more recent set of building blocks.

Checking Commercial Availability of Building Blocks via Molport. In Results, we used the Molport platform to check whether a predicted BB is commercially available or not. Initially, we compiled a list of building blocks for searching and used the ‘List Search’ tab in the Molport website (<https://www.molport.com/shop/swl-step-1>) to check their availability. Once the SMILES strings were entered into the search interface, we set the search criteria to a minimum acceptable quantity of 500 mg and match types restricted to ‘Exact’ and ‘Perfect’ to search in the database of ‘screening compounds’ and ‘building blocks.’ Once the search completed, we downloaded the excel file under the ‘Selected Items’ column from the List Search result tab (<https://www.molport.com/shop/swl-requests>), which contained both the commercially available compounds and information about the supplying vendors.

Calculation of SA Scores. We calculate SA scores for both the iMiner-proposed molecules and SynLlama-generated analogs using the oracle functions named ‘SA’ implemented in the TDC Commons package[160].

iMiner-Generated Molecules and Docking Procedures for Analogs. The iMiner algorithm[52], an 1D string-based LSTM model for SELFIES[161] string generation, was

employed in this study. The molecules generated by iMiner are optimized for 3D shape complementarity using a composite objective function comprising the AutoDock Vina[146] docking score, as well as a custom-defined druglikeness score[52].

For molecular docking tasks, we obtained the SARS-CoV-2 Mpro crystal structure (PDB ID: 7L11[147]) from the Protein Data Bank[162] and processed it with PDBFixer[163] to add missing hydrogens and remove heteroatoms. The docking grid was centered at the geometric center of the ligand (XF1) from the corresponding PDB file ($[x = -22, y = -4, z = -28]$) using a cubic box with 20 Å sides. Both proteins and ligands were converted to PDBQT format using Meeko (<https://github.com/forlilab/meeko>). Docking was performed with AutoDock Vina using an exhaustiveness parameter of 64, and the best pose for each ligand was recorded. This protocol was consistently applied during both iMiner training and analog docking assessments.

The custom drug-likeness score is a composite score that evaluates 13 key molecular properties derived from the ChEMBL database. These properties capture both basic structural features and nuanced physicochemical characteristics, including the fraction of sp^3 -hybridized carbons, the total number of heavy atoms, and the fraction of non-carbon atoms within these heavy atoms. Additionally, the score accounts for the counts of hydrogen bond donors and acceptors, the number of rotatable bonds, and the balance between aliphatic and aromatic rings, along with molecular weight. Complementing these are parameters such as the approximate log partition coefficient (alogP), polarizable surface area (PSA), the number of structural alerts, and the size of the largest ring present in the molecule. Each property contributes to the overall score through a weight that is inversely proportional to the entropy of its distribution in the ChEMBL database: properties with narrower and more informative distributions exert a stronger influence. By summing the log likelihoods of these properties with their respective weights, the score effectively biases the generative model to produce molecules that closely mimic the drug-like profiles observed in established therapeutics, ensuring that the exploration of chemical space remains focused on compounds with favorable bio-availability and efficacy profiles.

Pocket2Mol Generation. *De novo* generation with Pocket2Mol was performed for Thrombin and TYK2 targets using codes from the Pocket2Mol github repository[164]. Three default settings specified in `configs/sample_for_pdb.yml` were modified to generate at least 1000 molecules in one single run: `num_samples:1000`, `beam_size:500`, `max_steps:100`. The protein structure files were downloaded from the Schrödinger FEP benchmark github repository[153]. The pocket center was set to (-4.0, 26.5, -30.0) for TYK2 and (17.0, -12.5, 22.5) for Thrombin.

Unsynthesizable Molecules Identificaiton and Reward Calculation. To access the list of the unsynthesizable molecules, we query the first 50 top-scoring molecules that were identified as unsynthesizable by ASKCOS[129] for each property category listed in this csv (https://github.com/wenhao-gao/askcos_synthesizability/blob/master/results/goal_hard_cwo.csv). There are a total of 10 different individual rewards, including 7 multi-

property objectives (MPOs) centering around 7 different drug targets (Osimertinib, Fexofenadine, Ranolazine, Perindopril, Amlodipine, Sitagliptin, Zaleplon), Valsartan SMARTS, and 2 Hopping (Scaffold and deco). We used the TDC Commons package [160] to score both the original molecules and the generated analogs for their corresponding property category.

Free Energy Perturbation (FEP) Protocols. The relative binding free energies are calculated using GPU-accelerated AMBER22[165] (`pmemd.cuda.MPI`). AMBER14SB[166] and OpenFF-2.1.0[167] were used to parametrize the protein and the ligand, respectively. The SARS-CoV-2 Mpro protein structure (PDB code: 7LTJ) was downloaded from RCSB PDB and prepared with PDBFixer[11] to assign side-chain protonation states at pH=7.4 and add hydrogens. H163 was manually set to be its variant H1E (hydrogen added on N ϵ) to ensure the correct hydrogen bonding with the ligand. For TYK2 and Thrombin, their protein structures were downloaded from the github repository of Schrödinger benchmark dataset[153]. A submodule `app.Modeller` in OpenMM[10] was used to immerse the protein-ligand complexes and unbound ligands in a cubic water box with 15Å buffer size and add ions (Na $^{+}$, Cl $^{-}$) to neutralize the system and maintain 0.15M ionic strength. For perturbations involving charge changes, the alchemical water method[168] was used to eliminate the artifacts in PME simulation of system with net-charges.

We used 16 unevenly distributed lambdas (0.0, 0.174, 0.226, 0.265, 0.330, 0.383, 0.432, 0.477, 0.522, 0.568, 0.617, 0.670, 0.735, 0.774, 0.826, 1.0) to transform the initial state to the final state in the free energy. This lambda settings was designed to maximize the phase space overlap between adjacent states with the second-order smooth-step function introduced. The transformations were performed with the modified SSC(2) softcore potentials ($m = n = 2, \alpha_{LJ} = 0.5, \alpha_{Coul} = 1$)[169]. Kartograf[170] algorithm was used to determine the common core region (SC) and soft core region (SC) atoms.

Each lambda state was subjected to the following simulation protocol to equilibrate the system: (1) energy minimization without any constraints; (2) heating from 0 to 100 K at constant volume and temperature (NVT) ensemble over 20 ps, followed by MD at constant pressure and temperature (NPT) ensemble at 100 K for 20 ps; (3) heating to 200 K at NVT ensemble over 20 ps followed by another 20 ps at NPT ensemble at 200 K; (4) heating to 298.15 K at NVT ensemble over 20 ps followed by another 20 ps at NPT ensemble at 298.15 K; (5) another pre-production equilibrium run at NPT ensemble for 500 ps. During the equilibration steps 2-4, restraints ($5 \text{ kJ} \cdot \text{mol}^{-1} \cdot \text{\AA}^2$) were applied to heavy atoms on the solute. Finally, a 5-ns production run was performed for each lambda state with the ACES enhanced sampling method[171] and replica exchange was attempted every 0.5 ps. All the simulations employed 4 fs time step with the mass of solute hydrogens repartitioned to 3 amu[172]. MBAR algorithm implemented in `alchemlyb`[173] was used to estimate the free energy change between two states and yield $\Delta\Delta G$. Then, the maximum likelihood estimation (MLE) method[174] was used to calculate the absolute binding free energy (ΔG) of each ligand and the ΔG was shifted to make the average of calculated ΔG of the ligands equal to the average of their experimental ΔG :

$$\sum_i \Delta G_{\text{pred}}^{(i)} = \sum_i \Delta G_{\text{expt}}^{(i)} = \sum_i RT \ln \text{IC}_{50}^{(i)}$$

All the system preparation and analysis were performed with an in-house package named **easybfe** that automates the whole workflow and manage the calculations with high-performance computing platforms, and it will be described in a future publication in details.

Supporting Tables

Dataset	Category	SynLlama(RXN 1)	SynLlama(RXN 2)
Training Data	Valid JSON	98.00%	98.20%
	Template Mem.	100.0%	100.0%
	BB Selection	99.96%	99.96%
	Valid SMILES	99.46%	99.70%
	Matched Reactants	97.64%	97.95%
	Good Products	98.58%	97.97%
Testing Data	Valid JSON	93.90%	94.60%
	Template Mem.	100.0%	100.0%
	BB Selection	99.66%	100.0%
	Valid SMILES	99.50%	99.46%
	Matched Reactants	96.90%	97.25%
	Good Products	96.39%	96.19%
ChEMBL Data	Valid JSON	99.00%	99.00%
	Template Mem.	99.82%	100.0%
	BB Selection	99.47%	99.81%
	Valid SMILES	95.23%	97.33%
	Matched Reactants	70.93%	84.02%
	Good Products	87.02%	87.65%

Table S2.3: **Benchmarks of SynLlama inferences using SynLlama models trained with two sets of reaction templates.** Here, both models are fine-tuned on Llama-3.2-1B model with 2M reaction data generate using the same set of training building blocks. We select 1000 molecules for each model: training and testing data are generated using their corresponding reaction templates; ChEMBL data is the same set of 1000 molecules as described in the main text. All SynLlama inferences are run at $T = 0.1$ and $TopP = 0.1$.

Task	Sampling Method	K	N_{Syn}
LLM Benchmark	Frozen Only	5	25
Synthesis Planning	Greedy Sampling	5	25
Synthesizable Analog	High Only	10	50
Hit Expansion	High Only	20	100

Table S2.4: **Hyperparameters used for each task.** Under each task name we include the sampling method used for SynLlama inferences as defined in Additional Methodology Details. K represents the number of most similar SMILES string to take during the reconstruction algorithm. N_{Syn} represents the maximum number of synthesis routes to be tracked for each single SynLlama inference during the reconstruction algorithm.

Dataset	Method	# of Recon. Mol.			Morgan Sim.
		Enamine BB	New BB	Total	
Enamine Diversity	SynLlama	691	232	741	0.92
	SynLlama - druglike	574	100	595	0.86
	SynLlama - forward	529	50	546	0.85
ChEMBL Data	SynLlama	197	152	287	0.68
	SynLlama - druglike	124	107	197	0.61
	SynLlama - forward	132	58	168	0.59
Branching synthesis	ChemProjector	302	-	302	0.79
	SynLlama (RXN 1)	415	118	465	0.87
	Synformer†	39	-	39	0.61
	SynLlama (RXN 2)	358	101	408	0.84

† The released Synformer[69] model weights were fine-tuned extensively on smaller drug-like compounds, which caused it to fall short on synthesis planning for more complex molecules.

Table S2.5: **Reconstruction performances across various reaction data sets and prompt design choices.** SynLlama-druglike refers to fine-tuning on target molecule that falls within the same distribution as ChEMBL. More detailed analysis of the target molecule properties can be found in Figure S1. SynLlama-forward refers to fine-tune on prompt-response pairs structured as forward synthesis rather than retrosynthesis. SynLlama-branching refers to reconstructions based on tree-like synthesis with testing molecules generated similarly to the training data using testing building blocks and corresponding two sets of reaction templates. We compare to ChemProjector[68] (RXN 1) and Synformer[69] (RXN 2).

Dataset	% of BB in Enamine	# of Raw Reconstructed Mol.		
		Enamine BBs	New BBs	Total
Testing	75.85%	506	125	563
Enamine	73.51%	510	100	557
ChEMBL	48.07%	161	95	221

Table S2.6: Comparison of Enamine BB presence and reconstruction with purchasable BBs across datasets at greedy temperature and top-p combo when using 91 RXN templates (RXN 1).

Dataset	% of BB in Enamine	# of Raw Reconstructed Mol.		
		Enamine BBs	New BBs	Total
Testing	76.61%	465	114	520
Enamine	68.34%	647	232	711
ChEMBL	48.04%	179	152	280

Table S2.7: Comparison of Enamine BB presence and reconstruction with purchasable BBs across datasets at greedy temperature and top-p combo when using 115 RXN templates (RXN 2).

Dataset	Method	Morgan	Scaffold	Gobbi
Enamine Diversity	SynNet	0.57	0.57	0.52
	ChemProjector	0.82	0.85	0.83
	Synformer	0.91	0.92	0.89
	SynLlama (RXN 1)	0.87	0.88	0.85
	SynLlama (RXN 2)	0.92	0.94	0.92
ChEMBL Data	SynNet	0.43	0.20	0.27
	ChemProjector	0.60	0.59	0.56
	Synformer	0.67	0.72	0.72
	SynLlama (RXN 1)	0.66	0.67	0.63
	SynLlama (RXN 2)	0.68	0.69	0.66

Table S2.8: **Similarity metric comparisons over all successful reconstructions of target and analog molecules.** Similarity metrics using Morgan, Scaffold, and Gobbi similarity scores. SynNet[67] and Chemprojector[68] are trained using RXN 1, and Synformer[69] is trained using RXN 2. Scores are computed over successful reconstructions of target and analog molecules from Table 2.

Dataset	Method	Similarity		
		Morgan	Scaffold	Gobbi
Enamine Data	SynNet	0.51	0.51	0.45
	ChemProjector	0.67	0.72	0.69
	Synformer	0.74	0.76	0.69
	SynLlama(RXN 1)	0.69	0.72	0.65
	SynLlama(RXN 2)	0.69	0.75	0.70
ChEMBL Data	SynNet	0.39	0.38	0.22
	ChemProjector	0.54	0.52	0.49
	Synformer	0.59	0.65	0.65
	SynLlama(RXN 1)	0.56	0.57	0.51
	SynLlama(RXN 2)	0.54	0.56	0.52

Table S2.9: **Similarity metric comparisons over analog molecules when target molecules could not be constructed.** SynNet[67] and Chemprojector[68] are trained using RXN 1, and Synformer[69] is trained using RXN 2. Scores are computed over molecules from Table 2 that could not be fully reconstructed.

Supporting Figures

Instruction	You are an expert synthetic organic chemist. Your task is to design a synthesis pathway for a given target molecule using common and reliable reaction templates and building blocks. Follow these instructions:\n\n1. **Input the SMILES String:** Read in the SMILES string of the target molecule and identify common reaction templates that can be applied.\n\n2. **Decompose the Target Molecule:** Use the identified reaction templates to decompose the target molecule into different intermediates.\n\n3. **Check for Building Blocks:** For each intermediate:\n - Identify if it is a building block. If it is, wrap it in <bb> and </bb> tags and save it for later use.\n - If it is not a building block, apply additional reaction templates to further decompose it into building blocks.\n\n4. **Document Reactions:** For each reaction documented in the output, wrap the reaction template in <rxn> and </rxn> tags.\n\n5. **Repeat the Process:** Continue this process until all intermediates are decomposed into building blocks, and document each step clearly in a structured JSON format.
Input	Provide a synthetic pathway for this SMILES string: <chem>Cn1ncc(-c2ccc3c(c2)CC(C(=O)NCCOCc2ccccc2Cl)C3)n1</chem>
Output	<pre>"{'reactions': [{'reaction_number': 1, 'reaction_template': '<rxn>[Cl,OH,O-:3][C\$(C(=O)([CX4,c])),C\$([CH](=O)):2]=[O:4].[N\$([NH2,NH 3+1])([CX4,c]),N\$([NH])([CX4,c])([CX4,c]):6]>>[N+0:6]-[C:2]=[O:4]</rxn>' 'reactants': ['NCCOCc1ccccc1Cl', 'N1ncc(-c2ccc3c(c2)CC(C(=O)O)C3)n1'], 'product': 'N1ncc(-c2ccc3c(c2)CC(C(=O)NCCOCc2ccccc2Cl)C3)n1'}, {'reaction_number': 2, 'reaction_template': '<rxn>[Br:1][c\$(c(Br)),n\$(n(Br)),o\$(o(Br)),C\$([CH](Br)(=C)):2].[C\$(C(B) ([CX4])([CX4])([CX4])),C\$([CH](B)([CX4])([CX4])),C\$([CH2](B)([CX4])),C\$ ([CH2](B)),C\$(C(B)(=C)),c\$(c(B)),o\$(o(B)),n\$(n(B)):3][B\$(B([C,c,n,o])([OH,\$(OC)])([OH,\$(OC)])),B\$([B-1]([C,c,n,o])(N)([OH,\$(OC)])([OH,\$(OC)])) :4]>>[C,c,n,o:2][C,c,n,o:3]</rxn>' 'reactants': ['O=C(O)C1Cc2ccc(Br)cc2C1', 'N1ncc(B2OC(C)(C)C(C)(C)O2)n1'], 'product': 'N1ncc(-c2ccc3c(c2)CC(C(=O)O)C3)n1'}], 'building_blocks': ['<bb>NCCOCc1ccccc1Cl</bb>', '<bb>O=C(O)C1Cc2ccc(Br)cc2C1</bb>', '<bb>N1ncc(B2OC(C)(C)C(C)(C)O2)n1</bb>'] }"</pre>

Figure S2.5: Instruction, input, and output from the SynLlama model’s inference on example SMILES string from Fig.1d. During data generation, all instructions remain the same, and the input-output pairs are generated within the training synthesizable chemical space. We enforce the JSON format in the output for our post processing algorithms. The output JSON has two parts: reactions and building blocks. In ‘reactions’, a series of reaction steps are generated, where the product of the next reaction serves as the reactant for the previous one. In ‘building blocks’, BBs are selected from the ‘reaction’ section and compiled into a list.

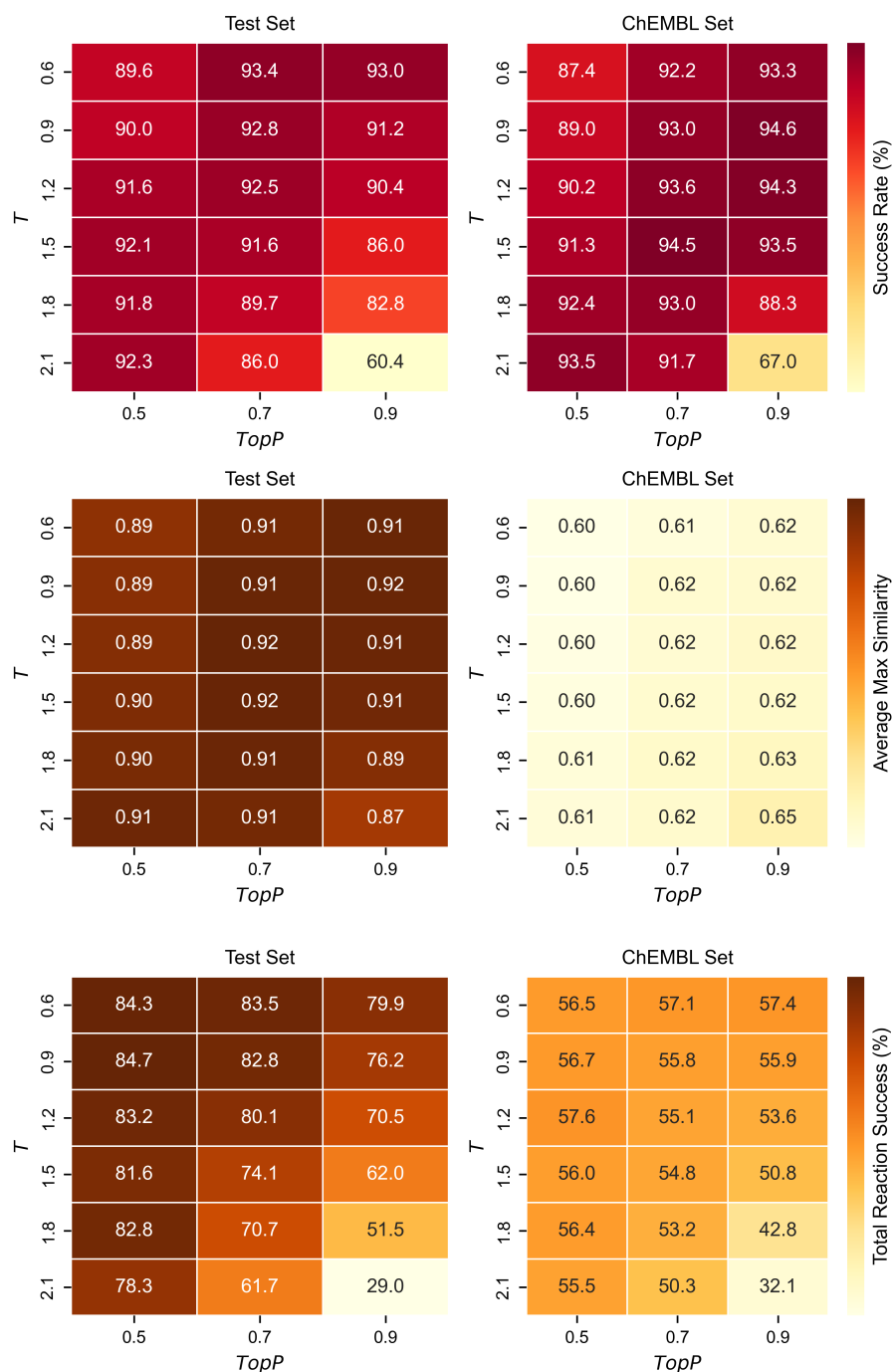


Figure S2.6: **Reconstruction algorithm and SynLlama raw output benchmarks for SynLlama inferences on the Testing and ChEMBL sets under various temperature and top-p combinations.** The first row represents the success rate of the Enamine reconstruction algorithm based on SynLlama inference outputs. The second row represents the average maximum Tanimoto similarity between the target and analogs generated via the reconstruction algorithm based on 4096-bit Morgan fingerprints. The last row represents the percentage of SynLlama raw outputs that can directly represent a retrosynthetic path for the input molecule without downstream processing with the reconstruction algorithm.

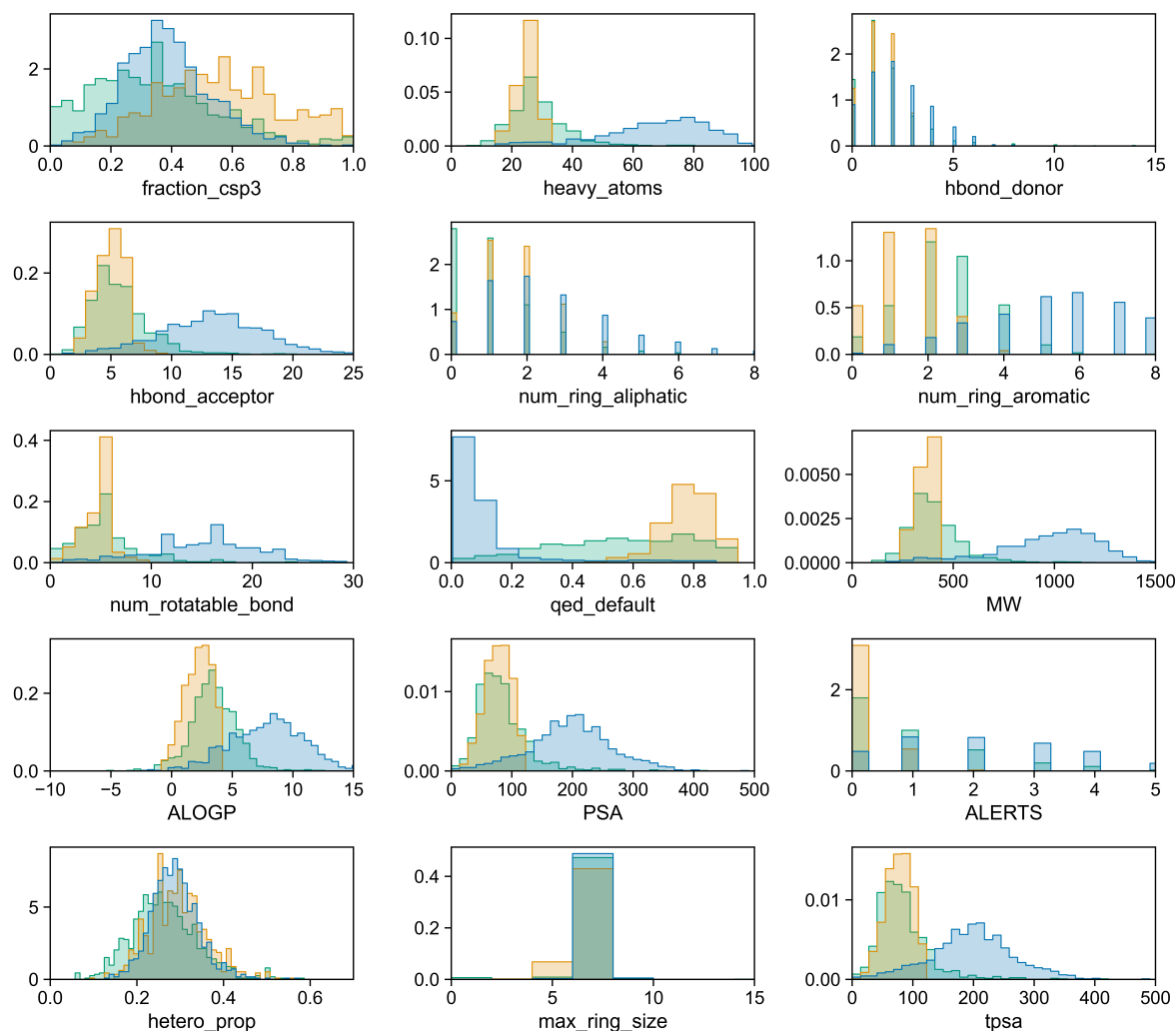


Figure S2.7: Drug-related property distributions between product molecules from normal training data (blue), Enamine Diversity Set (orange), and ChEMBL molecules (green).

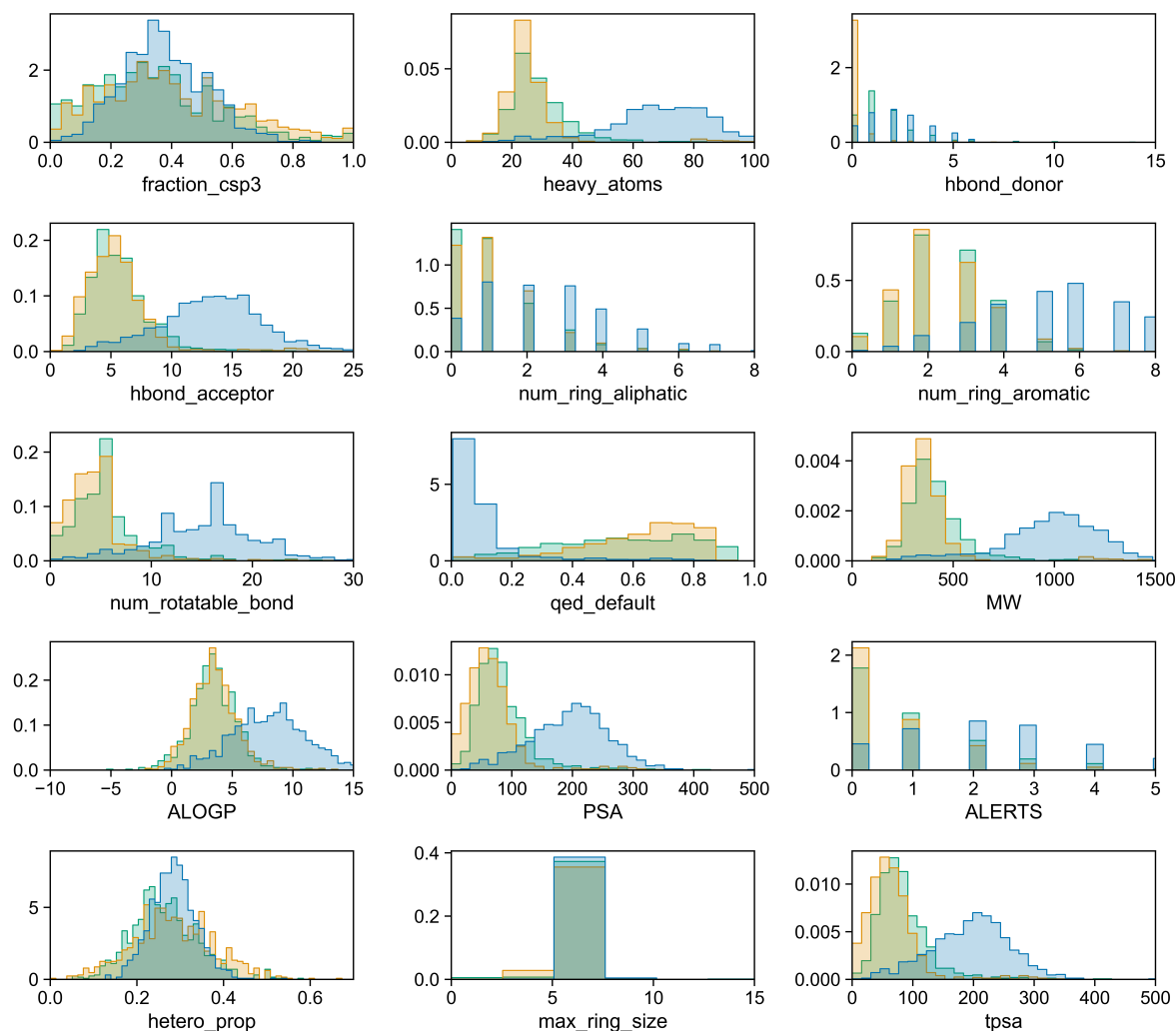


Figure S2.8: **Drug-related property distributions between product molecules from normal training data (blue), product molecules from training data constrained on druglike properties (orange), and ChEMBL molecules (green).** The generated product molecules under the constraint of druglike properties display similar distribution as ChEMBL molecules. The product molecules from normal training scheme occupies a very different chemical space with more larger molecules.

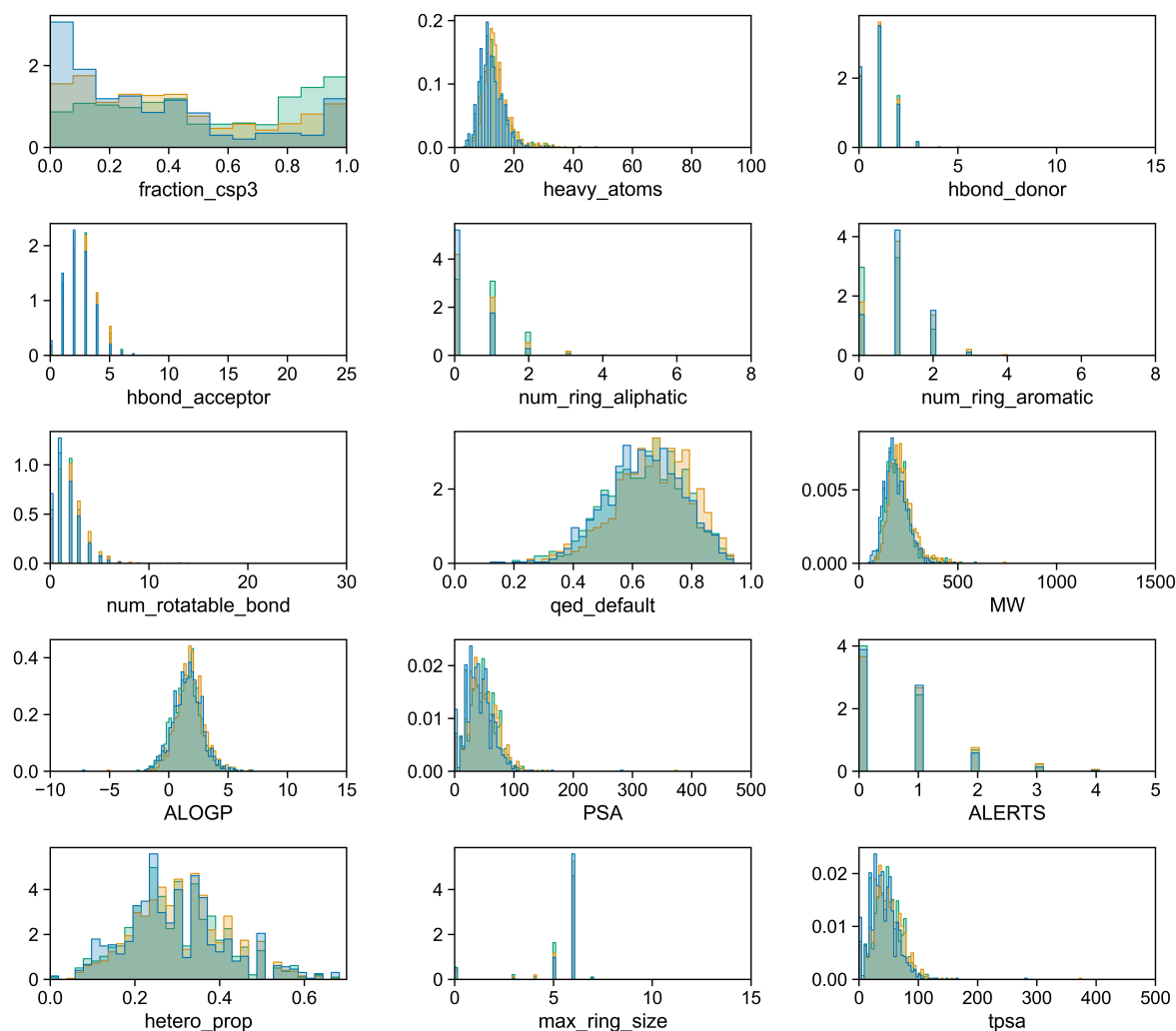


Figure S2.9: **Drug-related property distributions between training building blocks (blue), testing building blocks (orange), and Molport building blocks (green).** The three building blocks show very similar distribution in all categories.

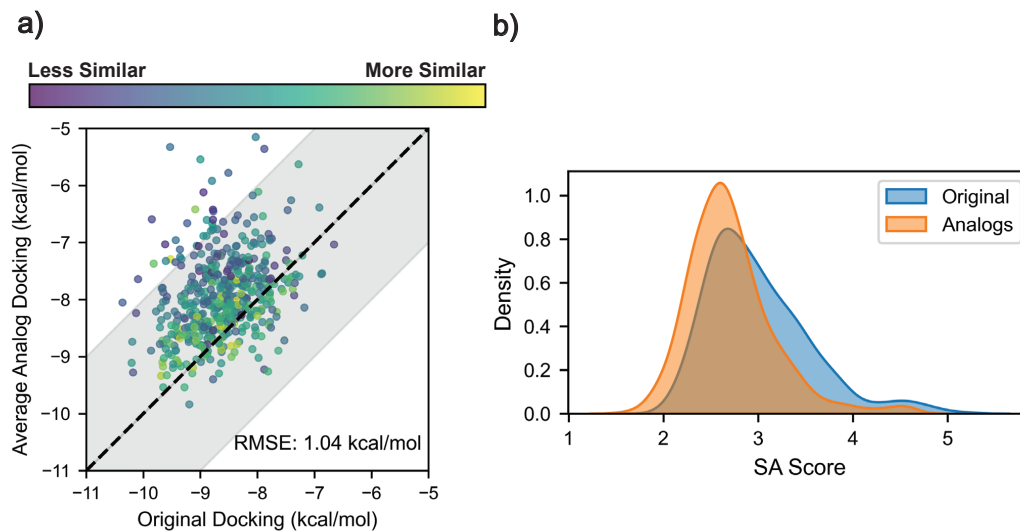


Figure S2.10: **Docking score and SA score distribution between 500 iMiner-generated molecules and proposed analogs from SynLlama model trained on RXN 1.** (a) Correlation plot comparing docking scores of 500 iMiner-generated molecules and the average docking scores of ten most similar analogs for each iMiner-generated molecule. Each data point is color-coded by the average Morgan fingerprint similarity computed between the iMiner target molecules and their corresponding analogs. The shaded area is the energy uncertainty range of $\pm 2 \text{ kcal/mol}$, which is typical for AutoDock Vina scores[146]. (b) SA score distribution of iMiner molecules and SynLlama-proposed analogs.

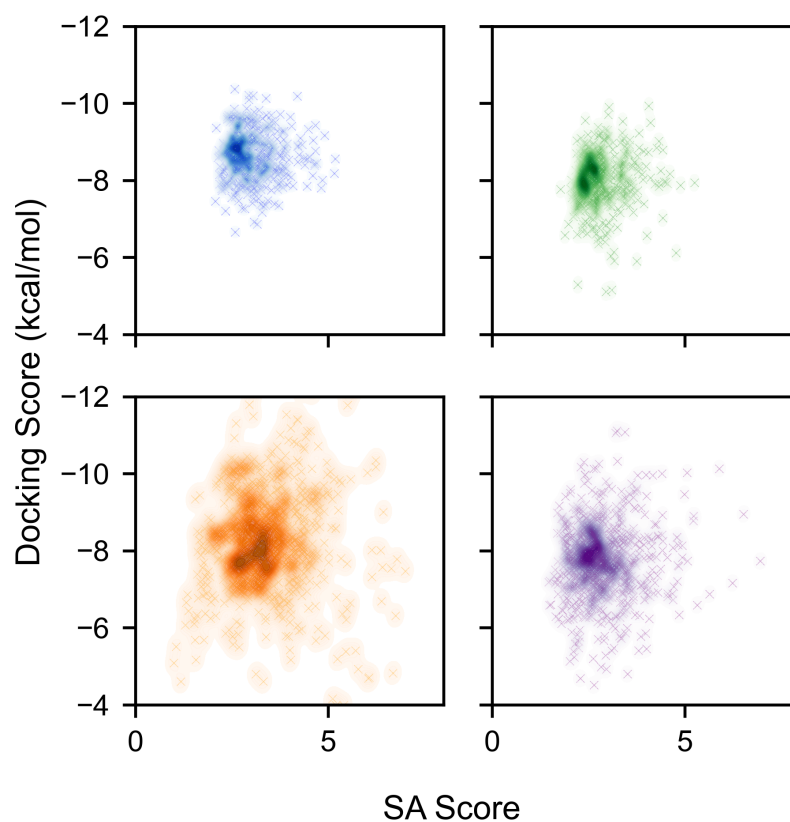


Figure S2.11: **Kernel density estimations of docking scores and SA score for target molecules and SynLlama-generated analogs.** Blue: iMiner targets. Green: iMiner analogs. Orange: Pocket2Mol targets. Purple: Pocket2Mol analogs.

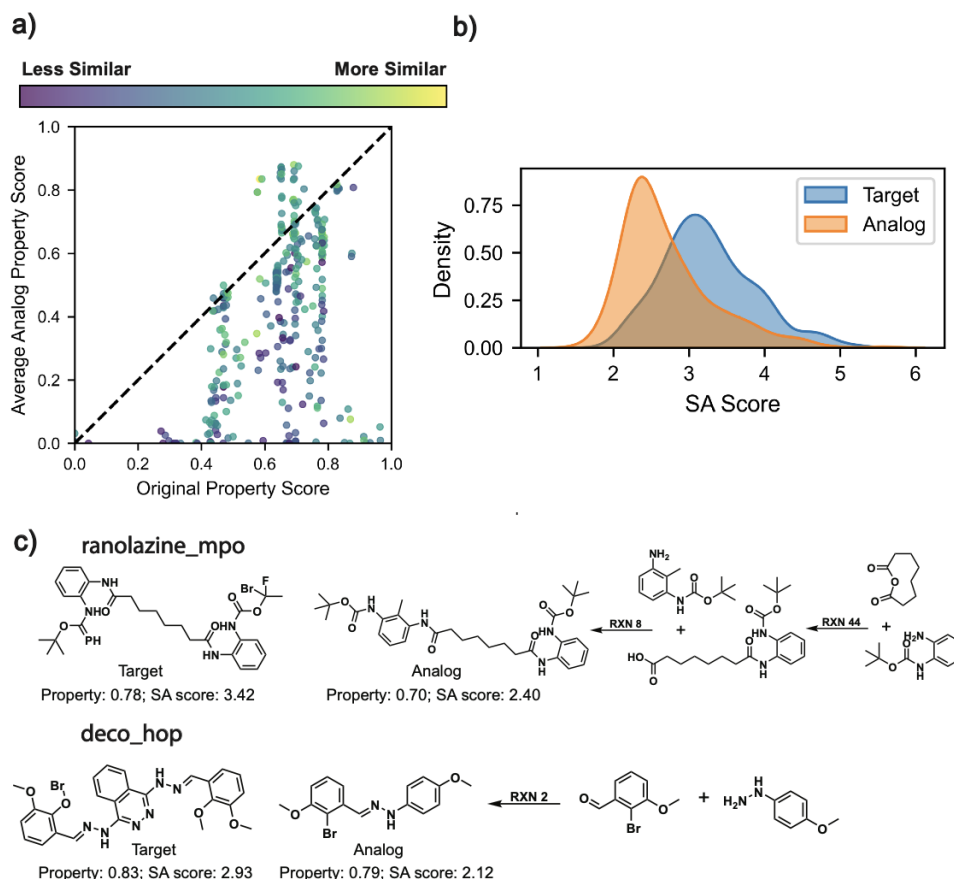


Figure S2.12: Oracle score and SA score distribution between 500 ASKCOS unsynthesizable molecules and proposed analogs from SynLlama model trained on RXN 2. (a) Correlation plot comparing property scores of 500 ASKCOS unsynthesizable molecules and the average docking scores of ten most similar analogs for each molecule. Each data point is color-coded by the average Morgan fingerprint similarity computed between the ASKCOS unsynthesizable molecules and their corresponding analogs. (b) SA score distribution of ASKCOS unsynthesizable molecules and SynLlama-proposed analogs. (c) Property and SA scores for example target-analog pairs along with the predicted analog synthetic pathways for two optimization targets: Ranolazine MPO and Deco Hop.

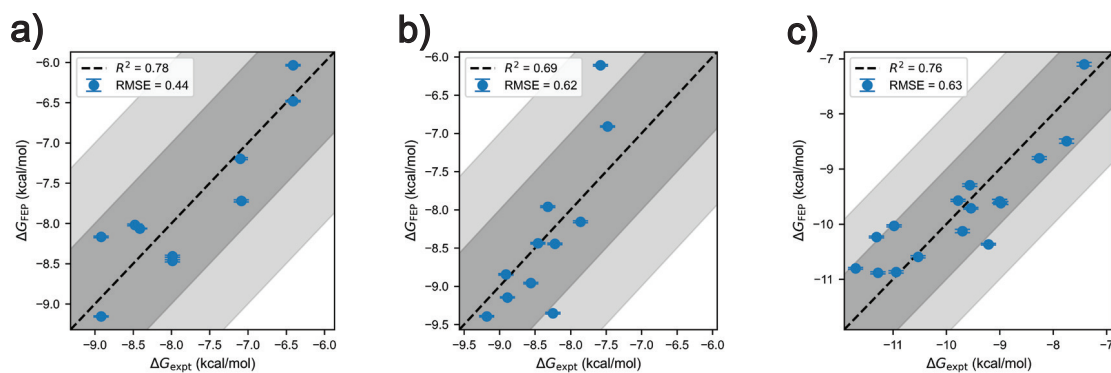


Figure S2.13: **FEP benchmarking on all three protein systems.** Correlation plots between ΔG extracted experimental IC50 values and ΔG calculated from FEP for (a) SARS-CoV-2 Mpro[152], (b) Thrombin[148], and (c) TYK2[149, 150]. The correlations across all three systems have RMSE < 1 kcal/mol, indicating the reliability of FEP calculations.

Chapter 3

LinkLlama: Enabling Large Language Model for Chemically Reasonable Linker Design

3.1 Introduction

Fragment-based drug discovery (FBDD) has established itself as a cornerstone of modern medicinal chemistry [175–180]. By identifying low-molecular-weight “hits” that bind to specific protein sub-pockets, FBDD provides an efficient starting point for exploring chemical space. However the central challenge in FBDD is the transition from individual fragments to lead-like molecules, a process that requires the design of optimal linker chemistry[181, 182]. Effective linker design is essential for connecting fragments from X-ray soaking experiments, performing scaffold hopping to replace problematic functional groups, and executing general hit expansion strategies[183]. However, the design of a linker is a complex multi-objective and constrained optimization task. Beyond simply connecting fragments, a chemically reasonable linker must ensure that the fragments maintain their original binding orientations. Furthermore, the resulting compound should exhibit low internal strain and possess drug-like properties while avoiding synthetically inaccessible motifs or undesired functional groups.

With the advancement of machine learning for drug discovery, several models have been developed to automate this process, broadly categorized into 2D and 3D approaches. 2D models such as DeLinker [71] and Link-INVENT [184] have demonstrated success in generating novel scaffolds. DeLinker utilizes an iterative atom-growth strategy, whereas Link-INVENT employs a reaction-based fragmentation and masked-encoder approach that can be integrated with reinforcement learning (RL) to optimize specific molecular properties. Conversely, 3D-aware models like DELETE [185] and DiffLinker [70] leverage Graph Neural

¹Reproduced with permission from: LinkLlama: Enabling Large Language Model for Chemically Reasonable Linker Design. K. Sun, Y. Wang, J. Purnomo, J. M. Cavanagh, G. B. Alteri, and T. Head-Gordon. BioRxiv, 2026.

Networks (GNNs) and diffusion models to propose linkers directly within a three-dimensional coordinate space, making them suitable methods for pocket-aware linker generation.

Despite these advancements, existing methods face significant limitations. The 2D models often require extensive RL or heavy post-hoc filtering to achieve the conditional sampling that satisfies medicinal chemistry standards. While 3D models are spatially aware, they frequently produce structures with geometric errors, such as unrealistic bond lengths or highly strained torsions, which may force fragments out of their bioactive conformations. Furthermore, standard evaluation metrics like the Quantitative Estimate of Drug-likeness (QED) [186] and Synthetic Accessibility (SA) scores [30] can be misleading. These metrics are often dominated by the properties of the initial fragments and fail to provide granular information regarding the quality of the generated linker itself. Furthermore, as pointed out by Walters[187], current models suffer greatly in generating molecules with stable drug-like chemistry, often rendering these models futile in real drug discovery deployment.

The recent success of Large Language Models (LLMs) in chemistry has opened new avenues for molecular and linker design to address these challenges. As foundation models [35], LLMs inherently possess broad chemical knowledge. Recent studies demonstrate that this representational capacity can be efficiently adapted to specific chemical tasks through supervised fine-tuning (SFT) [37–40, 50]. Several recent works exemplify this paradigm: SmileyLlama [36] showed that fine-tuning an open-weight LLM produces a Chemical Language Model (CLM) performing at or above the level of models trained exclusively on SMILES strings, utilizing natural language prompts to enable direct conditioning on molecular properties. Similarly, SynLlama [8] demonstrated that LLM fine-tuning can overcome synthesizability bottlenecks by constraining molecule generation to a manifold of commercially available building blocks and established chemical reactions. Ultimately, by coupling the deep internal chemical priors of an LLM with carefully curated fine-tuning datasets, complex chemical design problems can be effectively steered and resolved using natural language, typically with much less data and without RL retraining.

In this work, we present LinkLlama, a framework designed to bridge the gap between computational linker generation and practical medicinal chemistry. Our primary focus is on chemical reasoning that ensures the model consistently generating highly realistic molecules that chemists can readily act upon. We achieve this practical utility through an “alignment-by-design” approach that enables reinforcement learning (RL)-free conditional sampling. Rather than relying on computationally expensive RL cycles or complex reward-function engineering to optimize multiple objectives [28, 29, 184], users can dynamically steer LinkLlama via natural language prompts to satisfy precise structural and physicochemical constraints. This level of control is highlighted in our case analyses, where the model successfully navigates diverse design challenges, from standard scaffold hopping to satisfying the demanding architectural requirements of PROTAC linkers, while maintaining the vital spatial integrity of fragment hits. Ultimately, by combining robust chemical constraints with intuitive natural language guidance, LinkLlama emerges as a versatile and powerful tool for property-driven linker design.

3.2 Methods

Input Preparation

ChEMBL Data Processing. To obtain training data that covers a diverse, drug-like chemical space, we use ChEMBL36 [108] due to its breadth relative to earlier linker-generation work based on ZINC [71]. We apply a cleaning procedure of this database, which filters out molecules containing low-frequency atoms, exhibiting complex chemistry (such as low carbon fractions or complex ring systems), possessing excessive molecular weights, or being too small for fragmentation. Ultimately, approximately 93.0% of the ChEMBL36 molecules survive this culling, yielding a valid set of 2,665,082 molecules. The full cleaning steps are detailed in Chapter 5.5. Distributions comparing ChEMBL to other libraries (including PROTAC-relevant chemical space) are shown in Chapter S3.8.

Molecular Fragmentation. Once cleaned, the remaining ChEMBL molecules are fragmented into fragment-linker-fragment triplets following RDKit [188] matched molecular pair analysis. We perform exactly two cuts per molecule to generate these triplets, restricting the cuts strictly to acyclic single bonds incident to neutral sp^3 carbons. Because a single molecule can be fragmented in multiple ways, we apply specific criteria to filter the resulting decompositions. We retain cuts only if both the linker and fragments meet minimum heavy-atom counts, the linker possesses a topological path length of at least two between attachment points, and the linker is not larger than the smallest terminal fragment. Full details are provided in Chapter 5.5. From the cleaned ChEMBL36 set, this process yields 8,303,935 fragment-linker triplets with 256,135 unique linkers.

Assessment of Chemically Reasonable Molecules. We construct a rigorous analysis for both the designed linker and the complete molecule, utilizing two evaluation benchmarks from Walters and three additional "chemical reasonability" rules from iMiner [52]. We distinguish between properties specific to the isolated linker and those applied to the whole molecule. For the linker itself, we check for overly complex bridgehead ring structures or uncommon ring systems (defined as occurring fewer than 100 times in the entire ChEMBL36 database) [189, 190]. If a linker contains either motif, it is classified as "unreasonable." At the whole-molecule level, we apply a set of iMiner undesirable SMARTS [141] patterns that document common generative model errors, alongside PAINS[191] and Brenk filters[192] to identify and reject non-drug-like or problematic chemical motifs. Overall, a final molecule is classified as "reasonable" only if it passes all five metrics. Detailed implementations are provided in Chapter 5.5.

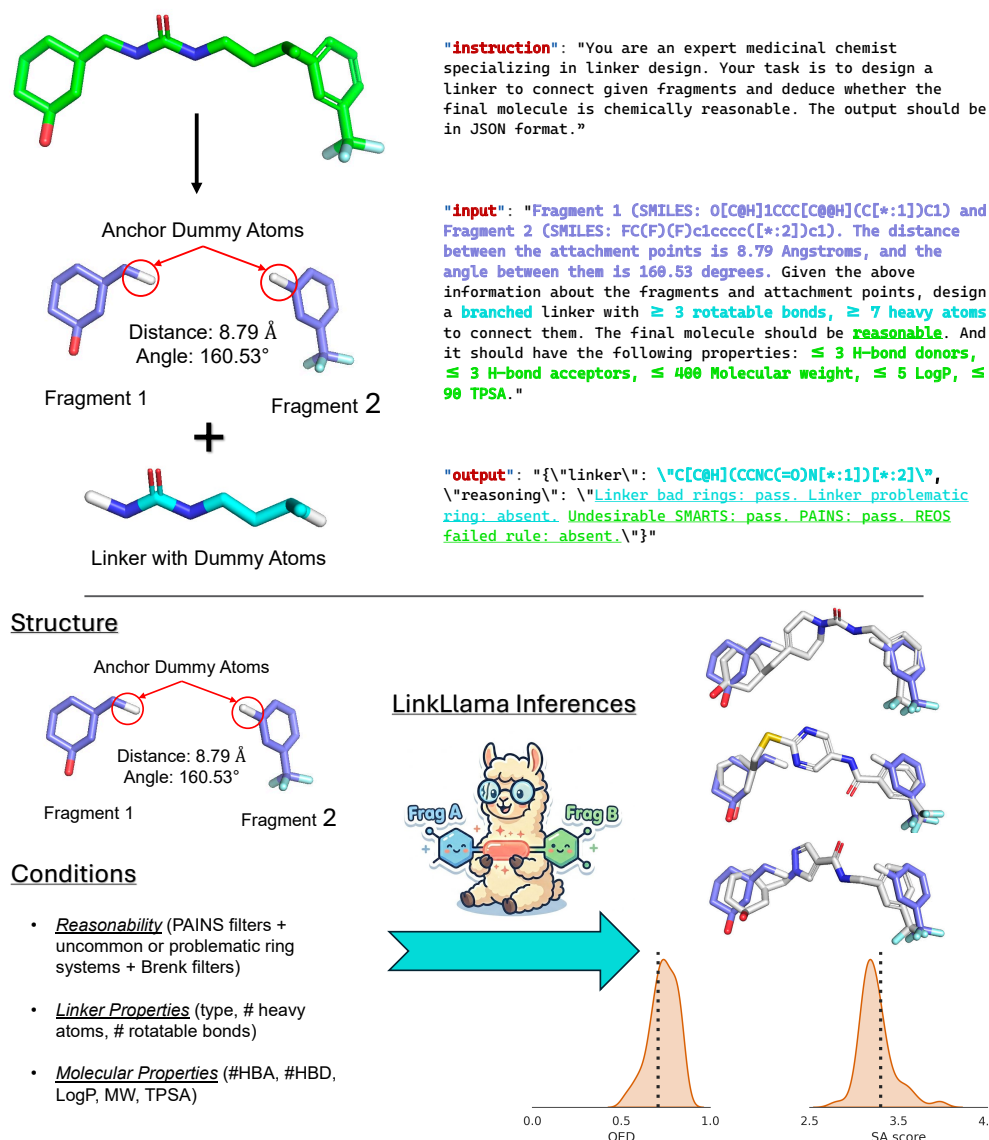


Figure 3.1: (Top). Schematic representation of the LinkLlama data curation pipeline, illustrating the transition from raw ChEMBL molecules to the filtered SFT corpus via fragmentation and chemical reasonability checks + the prompt construction and result processing. (Bottom). LinkLlama inference process with two parts of input specified in the SFT corpus along with 3 example generated molecules overlaid on crystal fragment poses and the distribution of QED and SA score for the 100 LinkLlama-generated whole molecules. The LinkLlama cartoon image is generated by Google Gemini.

Supervised Fine-tuning and Inference from LinkLlama

To create the LinkLlama model, we establish data generation protocols to format the cleaned input preparation of the ChEMBL dataset into a supervised fine-tuning (SFT) corpus. We frame the task as an instruction-following completion in Alpaca-style JSONL format. Each record provides a fixed instruction, a natural-language **input** that encodes the fragment pair and optional constraints, and a structured **output**. As illustrated in Chapter 3.1, the **input** always states the SMILES of both terminal fragments and their distance (in Å) and angle (in degrees). The distance is captured by the distance between the two anchor atoms, and the angle is calculated between the two unit vectors from each anchor atom pointing towards their connected dummy atoms. Additional linker-type and physicochemical band constraints (linker rotatable bonds, linker heavy atoms, H-bond donors/acceptors, molecular weight, $\log P$, TPSA) are included stochastically, each with an independent 50% chance during generation. The overall **reasonable/unreasonable** label is never dropped and is derived from the five reasonability checks as defined in the previous section. The complete template, all permissible input constraint strings, and the training randomization rules are documented in Chapter 5.5.

The model’s output is structured to predict not only the intervening linker but also to explicitly evaluate the chemical reasonability of the training structure. The expected response is a JSON object with the linker SMILES and a **reasoning** string that records the pass/fail status for the same chemical filters summarized above. By jointly modeling linker generation and this reasonability trace, LinkLlama learns to condition its generative space on medicinal chemistry heuristics.

LinkLlama is developed by fine-tuning the Meta Llama-3.2-1B-Instruct checkpoint on a subset of the prepared ChEMBL36 data filtered to minimize the over-representation of frequent linkers. This final subset contains 1,526,169 training examples. Fine-tuning is executed on a single node with $4 \times$ A100 GPUs in approximately 6.5 hours. Full training details are listed in Chapter 5.5.

During inference, prompts use the same sentence template, but the fragment geometry is always taken from the input data, while linker-type bands, molecule-property bands, and the requested **reasonable/unreasonable** condition are set explicitly in a YAML configuration. The model returns a JSON object with a linker SMILES and a reasoning string, where the linker SMILES string is then merged with the input fragments into the full molecule by aligning the anchor dummy atoms on the predicted linker and input fragments.

Hyperparameter Optimization and Model Selection

We first fine-tune a baseline model on two million randomly sampled fragment-linker pairs drawn from the full processed ChEMBL36 pool and assess its performance. In the initial pilot run, our primary focus is the validity, uniqueness, and novelty of the generated linkers produced by this model on the external ZINC-hard dataset used in the Results section. Supplementary linker-frequency diagnostics (Chapter S3.9) reveal a severe long tail in the

original ChEMBL data distribution, where a small set of simple yet common linkers, such as amide bonds, dominates the training counts. In practice, this checkpoint achieves high validity on the ZINC-hard dataset but relatively low uniqueness among valid generations and modest novelty with respect to the training linker vocabulary. As shown in Chapter 3.2, this trend confirms the potential memorization of the high-frequency head of the linker distribution rather than the exploration of diverse bridges.

To counteract this collapse in uniqueness, we curate two rebalanced training pools. First, the Cap50 corpus applies a strict per-linker frequency ceiling of 50 counts. Specifically, if a linker has more than 50 occurrences, we select 50 diverse samples based on their molecular properties and drop all other entries. This hard cap flattens the occurrence histogram and increases effective linker diversity at the cost of modest shifts toward heavier, more flexible linkers (Chapter S3.10), yielding the Diverse 1.5M split with 1,536,170 training samples. In addition, the Hybrid corpus employs a softer capping rule that preserves more of the natural rank-frequency ordering while still suppressing the most extreme repeats. In this approach, we apply the same diversity-based sampling for down-selection: linkers with fewer than 50 rows remain unchanged, those with between 50 and 500 rows are reduced by 10%, and those with more than 500 rows are hard-capped at 450 (i.e., $500 - 500 \times 10\%$). This procedure results in the **Diverse 2.9M** split containing 2,817,412 training samples. We train separate Llama-3.2-1B LoRA models on the Random 2M (unbalanced draw from the full pool), Diverse 1.5M (Cap50), and Diverse 2.9M (Hybrid) splits and compare them under identical evaluation protocols.

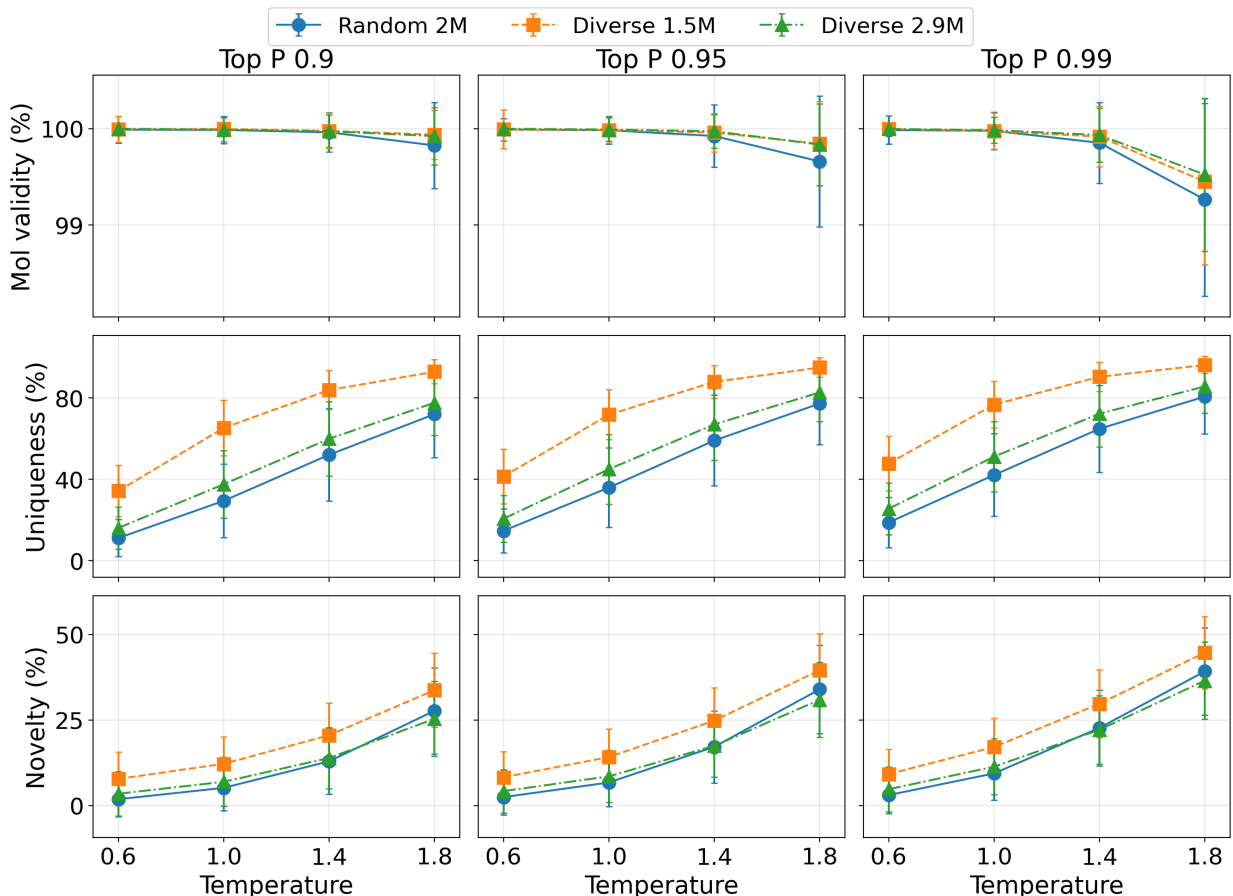


Figure 3.2: **Model hyperparameter searching performances of unconditional generation on the ZINC-hard dataset.** The legend denotes the three evaluated models trained on different data splits: **Random 2M** (the long-tailed, unbalanced pool), **Diverse 1.5M** (the strict Cap50 rebalanced pool), and **Diverse 2.9M** (the hybrid-capped pool). Individual data points represent temperature settings ($T \in \{0.6, 1.0, 1.4, 1.8\}$), and the columns correspond to Top-P sampling values (top- $p \in \{0.9, 0.95, 0.99\}$). The rows display mean molecule validity (%), uniqueness among valid samples (%), and the novelty of valid linkers compared to the training linker set (%), with shared y -axes within each row. Error bars indicate the dispersion across instances and aggregated runs.

In regards hyperparameters, we sweep temperature (T) and Top-P (top- p) as summarized in Chapter 3.2. The Diverse 1.5M (Cap50) checkpoint consistently offers the best compromise among validity, uniqueness, and novelty on this grid. As shown in Chapter S3.12, grids on another set we use in the Results section, the ZINC-random set, also uncovers the same overall trend. Therefore, we use this model with $T = 1.4$ and top- $p = 0.99$ everywhere below.

3.3 Results

Standard Benchmarking on ZINC

For the first set of generation tasks, we use the ZINC database [107], which is orthogonal to our ChEMBL training set. Although LinkLlama is trained on ChEMBL, ZINC offers a screening-library-like distribution distinct from the training corpus. Furthermore, it serves as the standard test evaluation setting in prior works, ensuring direct comparability of LinkLlama with established 2D and 3D baseline models. To perform a thorough evaluation, we construct a random 1k subset of fragment pairs to represent typical generation difficulty, alongside a hard 1k subset. In the hard 1k subset, the reference molecule is precomputed to pass all five chemistry checks (ensuring chemical reasonability), and the reference linker contains either more than three rotatable bonds or more than seven heavy atoms. This hard test stresses the models’ ability to design long or flexible bridges following the structural rules defined in the Methods section. After selecting the SMILES strings for each subset, we use RDKit to generate a conformer for each molecule and apply the MMFF94 force field to relax the energy. We then save the fragments into SDF files and calculate their structural input information for LinkLlama inferences.

During model inference, we evaluate the unconditional generation capabilities of LinkLlama to ensure a fair comparison. Under this constraint, all optional input keywords are omitted, and only the reasonability condition is enabled. Therefore, the input prompt is structured as follows:

```
FRAGMENT_INFO Given the above information about the fragments and attachment
points, design a linker to connect them. The final molecule should be reasonable.
```

Using this prompt, we compare LinkLlama against DeLinker (representing 2D graph generation baseline) and DiffLinker (representing 3D diffusion baseline). The input fragment SDF files are prepared for baseline method inferences.

To comprehensively assess generation quality, we measure a suite of standard metrics. *Validity* tracks the percentage of generated candidates per test instance that yield a well-formed product. Specifically, the structure must pass successfully in RDKit, carry compatible charges with the input fragments, and match the intended anchor connectivity. Among these valid outputs, *uniqueness* calculates the fraction of distinct canonicalized full-molecule SMILES, collapsing duplicate stereoisomer or tautomer representations. For LinkLlama, we additionally report *novelty*, which measures the fraction of valid, distinct linker graphs absent from the ChEMBL-derived training data inventory; novelty values for the baseline methods are extracted from Igashov et al. [70]. To evaluate general drug-likeness and synthesizability, we compute RDKit’s quantitative estimate of drug-likeness (*QED*, bounded in $[0, 1]$) and synthetic accessibility score (*SA*, where higher values indicate more difficult syntheses). Finally, *reasonability* represents the percentage of valid molecules that pass our composite medicinal-chemistry screen, which encompasses linker-centric ring rules alongside whole-molecule PAINS, Brenk, and related structural alerts.

Table 3.1: **Generative performance comparison on the ZINC random 1k and ZINC hard 1k benchmarks.** Validity, uniqueness, novelty, and reasonability are reported as percentages, while QED and SA are dimensionless. For the percentage columns, values represent the instance-aggregated mean $\pm$ the bootstrap standard deviation of that mean (using 2000 bootstrap resamples). Reasonability represents the average per-instance pass rate across all instances. QED and SA are reported as the instance-mean $\pm$ standard deviation across instances.

Split	Model	Validity	Uniqueness	Novelty	QED	SA	Reasonability
Random 1k	DeLinker	96.0 $\pm$ 0.2%	86.4 $\pm$ 0.1%	47.1%*	0.63 $\pm$ 0.11	3.39 $\pm$ 0.63	44.2 $\pm$ 0.8%
Random 1k	DiffLinker	89.8 $\pm$ 0.8%	66.1 $\pm$ 0.8%	47.7%*	0.65 $\pm$ 0.11	3.42 $\pm$ 0.74	25.1 $\pm$ 0.8%
Random 1k	LinkLlama	99.9 $\pm$ 0.0%	87.7 $\pm$ 0.3%	25.8 $\pm$ 0.3%	0.68 $\pm$ 0.13	3.30 $\pm$ 0.62	73.1 $\pm$ 1.1%
Hard 1k	DeLinker	89.7 $\pm$ 0.2%	88.3 $\pm$ 0.2%	47.1%*	0.64 $\pm$ 0.10	3.16 $\pm$ 0.58	43.4 $\pm$ 0.7%
Hard 1k	DiffLinker	95.6 $\pm$ 0.4%	64.1 $\pm$ 0.8%	47.7%*	0.63 $\pm$ 0.09	3.05 $\pm$ 0.65	31.0 $\pm$ 0.6%
Hard 1k	LinkLlama	99.9 $\pm$ 0.0%	90.4 $\pm$ 0.2%	29.7 $\pm$ 0.3%	0.68 $\pm$ 0.12	2.99 $\pm$ 0.55	87.4 $\pm$ 0.5%

* Data extracted from DiffLinker Table 1. Baseline models were trained on ZINC dataset which has less unique linkers.

Table 3.1 reports these benchmark metrics. On the random split, LinkLlama nearly saturates validity while maintaining competitive uniqueness, whereas DiffLinker shows the largest bootstrap uncertainty and the lowest validity and uniqueness among the three methods. Although the novelty score of LinkLlama appears to fall short compared to the baselines, novelty must be balanced with chemical realism, because the generated linker and overall chemistry should still be synthetically attainable and structurally sound. Therefore, the novelty score should be interpreted alongside the reasonability percentage. LinkLlama effectively doubles the reasonability rate relative to DiffLinker and substantially outperforms DeLinker, confirming that the model successfully conditions its generative space on chemistry that passes the five reasonability checks. QED and SA sit in a similar band across all baselines, with LinkLlama showing marginal improvements. On the hard split, the performance gap becomes even more pronounced. While DiffLinker’s validity rises relative to the random subset, its uniqueness remains the lowest. Conversely, LinkLlama’s uniqueness slightly exceeds DeLinker’s, its novelty increases, and its reasonability advantage widens dramatically, nearly doubling the success rates of the baseline methods. This observation indicates that in these harder, more drug-discovery-relevant cases, LinkLlama produces more sensible designs that medicinal chemists can readily utilize.

Beyond 2D graph validity, successful linker design requires physically realistic 3D geometries. Figure 3.3 summarizes 3D-aware strain and pose agreement on the ZINC datasets. Because DeLinker and LinkLlama generate SMILES strings as their final output, we assess their pose RMSD using the same generation procedure employed to create the ZINC input data: RDKit generates an initial conformer, which is subsequently relaxed via the MMFF94 force field. We quantify pose agreement using the mean and best fragment RMSD, which compare the heavy-atom positions of the generated ligand to the ZINC decoy reference after

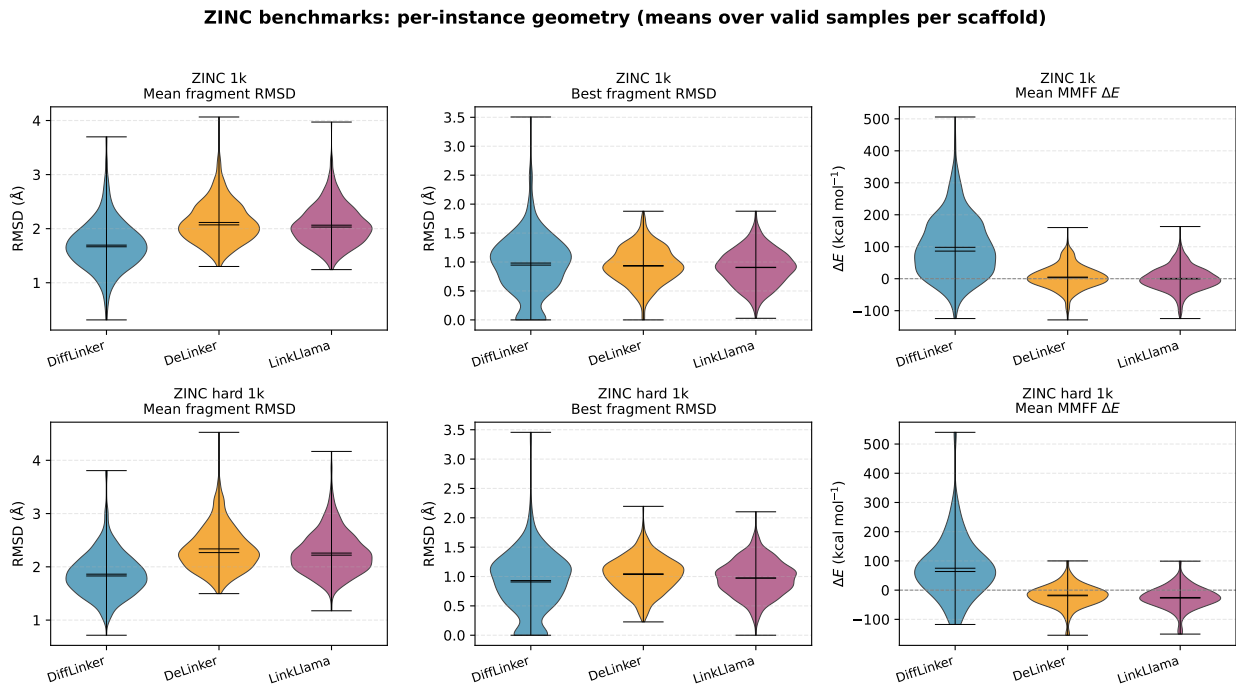


Figure 3.3: **Assessment of 3D geometric realism and internal strain for generated molecules.** The violin plots display the per-instance distributions for mean fragment RMSD, best fragment RMSD, and mean MMFF ΔE (kcal mol⁻¹) across the ZINC random (top row) and hard (bottom row) 1k datasets. The three columns correspond to these distinct geometry metrics. The distributions compare the generative outputs of DiffLinker, DeLinker, and LinkLlama, highlighting differences in pose agreement with the reference and the propensity of each method to generate high-strain, energetically unfavorable configurations.

alignment on the fixed fragments (lower values indicate a closer match to the reference pose). To evaluate physical realism, we calculate the mean MMFF ΔE after relaxing each pose. This metric is defined as the difference between the MMFF94 energy of the proposed full molecule (after a short force-field relaxation) and the reference ligand, serving as a proxy for internal strain where larger positive values indicate energetically unfavorable proposals. While the RMSD violin plots show that all methods achieve roughly overlapping pose agreement, their strain distributions differ significantly. DiffLinker’s ΔE distributions extend to much higher strain limits, suggesting frequent generation of distorted or sterically clashing geometries. In contrast, LinkLlama and DeLinker consistently produces low-strain structures, indicating that it naturally favors physically viable conformations even without explicit 3D spatial diffusion coordinates used by DiffLinker. Overall, on the ZINC test sets, LinkLlama generates diverse, unique candidates with competitive 1D and 3D assessment metrics, while yielding a significantly larger population of chemically reasonable structures.

Prompt-Conditioned Generation on ZINC Hard

Having evaluated unconditional generation, we design the next experiment to assess distribution control over the generated linkers. Specifically, we task the model with generating not only valid and reasonable linkers but also complete molecules whose physicochemical properties and linker topologies match specific constraints provided in the prompt. This setup utilizes the same instruction format established during supervised fine-tuning (detailed in the Methods and Supplementary Information). For this evaluation, we use the same ZINC hard 1k instances described in the previous section. For each instance, we generate and retain up to ten valid and unique SMILES strings per condition. We then compute physicochemical descriptors using the same RDKit-based pipeline applied throughout the benchmark suite. To rigorously quantify success, we define the success rate for each method to reflect the fraction of candidate molecules that simultaneously satisfy every active constraint specified for a given test.

We apply several distinct conditional constraints to test the model’s generative flexibility. *Reasonable* enforces the composite medicinal-chemistry filter used in previous sections, meaning a molecule is only counted if it passes all structural and strain rules, not merely if it is syntactically valid. The *Ro5* condition applies Lipinski-style drug-likeness limits to the fully assembled molecule: a molecular weight of at most 500 Da, a calculated $\log P$ of at most 5, no more than five hydrogen-bond donors, a maximum of ten hydrogen-bond acceptors, and a topological polar surface area not exceeding 140 Å². To constrain the linker topology itself, we use *Ring* and *Branched* labels, requiring the linker to contain at least one ring or to be classified as branched acyclic, respectively. Furthermore, we apply size limits using *rotB* and *heavy* labels, setting lower bounds that require more than two rotatable bonds and more than six heavy atoms in the linker alone to avoid trivial connections. To isolate the model’s instruction-following capabilities from its unconditional generative capabilities, we compare the prompt-conditioned LinkLlama outputs against an unconditional LinkLlama, as well as the baseline methods DeLinker and DiffLinker, scored with the exact same set of rules for each condition.

Chapter 3.2 summarizes the model success rate across various linker and molecular requirements. These results demonstrate LinkLlama’s strong adherence to user prompts, effectively shifting its generative distribution on demand. For the relatively mild Ro5 constraint, the prompted generation performs slightly above the unconditional baseline. This performance is expected, as the unconditional bulk distribution already heavily overlaps with standard drug-like property bands (Chapter S3.13). However, the performance gap widens significantly when introducing topology constraints. Prompts requiring specific topologies, such as ring or branched constraints, elevate the success rates well above the unconditional background distributions. When evaluating complex, joint constraints in the case of Ring + Ro5 + rotB/heavy + reasonable, LinkLlama retains a prompted success rate in the mid-40% range and meaningfully shift the distribution towards the desired distribution (Chapter S3.14). In stark contrast, the unconditional LinkLlama and baseline methods collapse to single digits or no success at all under these strict, overlapping criteria. Overall, these

Table 3.2: **Constraint satisfaction rates on the ZINC hard 1k dataset.** The table displays the percentage of generated molecules that simultaneously satisfy every active constraint for a given condition. All percentages represent the bootstrap mean $\pm$ standard deviation calculated over 2000 bootstrapping samples of the instances.

Condition	LinkLlama		Baselines	
	Conditional	Unconditional	DeLinker	DiffLinker
Ro5 + reasonable	81.2 $\pm$ 0.4	79.9 $\pm$ 0.4	21.8 $\pm$ 0.4	31.8 $\pm$ 0.6
Ring + reasonable	90.4 $\pm$ 0.3	63.6 $\pm$ 0.5	0.3 $\pm$ 0.1	9.5 $\pm$ 0.4
Ring + Ro5 + reasonable	80.1 $\pm$ 0.4	55.6 $\pm$ 0.5	0.3 $\pm$ 0.1	8.2 $\pm$ 0.3
Ring + rotB/heavy + reasonable	49.2 $\pm$ 0.6	9.0 $\pm$ 0.3	0.0 $\pm$ 0.0	1.1 $\pm$ 0.1
Ring + Ro5 + rotB/heavy + reasonable	43.1 $\pm$ 0.5	7.4 $\pm$ 0.3	0.0 $\pm$ 0.0	0.9 $\pm$ 0.1
Branched + reasonable	79.9 $\pm$ 0.4	22.5 $\pm$ 0.4	2.1 $\pm$ 0.2	24.5 $\pm$ 0.5
Branched + Ro5 + reasonable	76.7 $\pm$ 0.5	21.6 $\pm$ 0.4	2.1 $\pm$ 0.2	23.2 $\pm$ 0.5
Branched + Ro5 + rotB/heavy + reasonable	68.1 $\pm$ 0.5	15.8 $\pm$ 0.4	0.0 $\pm$ 0.0	6.2 $\pm$ 0.3
Branched + rotB/heavy + reasonable	70.6 $\pm$ 0.5	16.6 $\pm$ 0.4	0.0 $\pm$ 0.0	6.7 $\pm$ 0.3

results suggest that, under natural language guidance, LinkLlama can reliably navigate complex, multi-objective design spaces that remain inaccessible to unconditional generation or baseline methods.

Benchmarking on 3D-Based HiQBind Dataset

To further validate our model on a more challenging and structurally relevant benchmark, we utilize the HiQBind dataset [9], which consists of high-quality PDB chains paired with experimental crystal ligand poses. Because the structural fragments in this dataset form desirable interactions with target proteins, they serve as highly realistic starting points for fragment-based drug design. We evaluate the models on 1k random and 1k hard complexes, with the hard split enforcing the identical linker-difficulty rules defined previously (i.e., requiring linkers to contain more than three rotatable bonds or more than seven heavy atoms). It is important to note that we employ unconditional LinkLlama generation in this section to establish a rigorous baseline comparison. In practice, a user with a specific structural hypothesis could leverage our prompt-conditioning capabilities to combine various design properties, explicitly tailoring the generated linkers to the unique chemistry of the pocket.

As detailed in Chapter S3.4 and Figure 3.4, we observe generative performance and 3D structural trends consistent with our earlier ZINC benchmarking. Although overall QED scores shifts slightly lower across all methods on this harder dataset, LinkLlama maintains near-perfect validity and high uniqueness and demonstrates a stark advantage in structural reasonability, reaching 80.9% on the hard split compared to DiffLinker’s mid-20% mean pass rate. This divergence in chemical realism is heavily reflected in the 3D geometry evaluations.

HiQBind benchmarks: per-instance geometry (means over valid samples per scaffold) HiQBind 1k ΔE : instances with $|\Delta E| > 1000$ kcal mol⁻¹ omitted (viz).

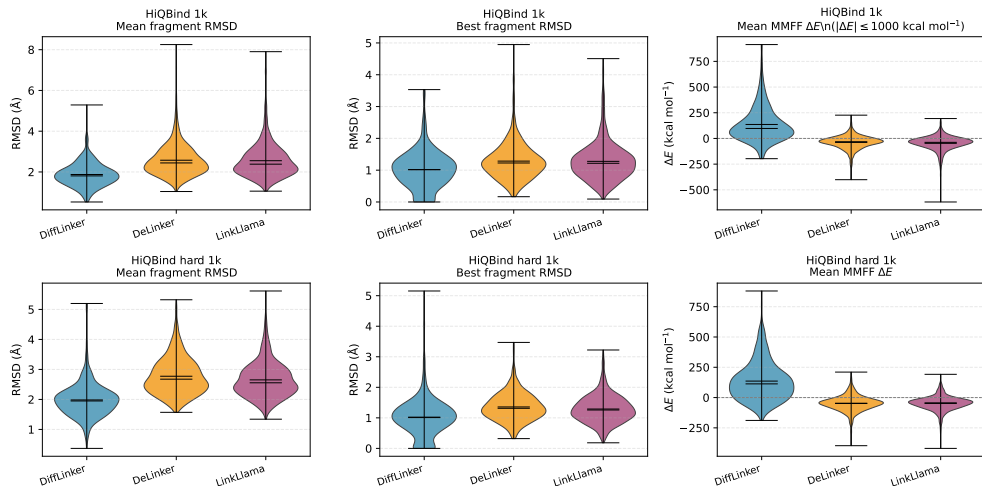


Figure 3.4: **Assessment of 3D geometric realism and internal strain on the HiQBind dataset.** The violin plots display per-instance distributions for mean fragment RMSD, best fragment RMSD, and mean MMFF ΔE (kcal/mol), following the layout of Figure 3.3. For the HiQBind random 1k plot specifically, the ΔE column filters out extreme outliers, retaining only instances with $|\Delta E| \leq 1000$ kcal/mol as an MMFF diagnostic cut-off. Both the HiQBind hard plots and all RMSD panels incorporate every valid generated instance.

Again, while DiffLinker achieves competitive pose agreement (mean RMSD), it does so at a severe energetic cost, exhibiting a heavy-tailed distribution of high internal strain (ΔE). Conversely, LinkLlama balances competitive RMSD with relaxed, low-strain geometries, confirming its ability to reliably produce physically viable and chemically reasonable ligand structures even within complex, protein-bound environments.

To evaluate functional binding potential, we use Uni-Dock [193] to dock valid proposed ligands into their respective protein pockets and calculate the per-instance mean docking score differences (Δ) relative to the experimental reference pose. Detailed procedures are include in Chapter 5.5. As shown in Figure 3.5, although most of the poses in all methods do not outperform the reference ligand, there still are some portions of all model’s distribution falls into the negative delta range. This observation indicates that the model routinely proposes novel, strong-binding linker conformations that match or improve upon the binding affinities of the native ligands within the actual pocket environment, suggesting that it can not only link the two fragments under such context, but also links the fragments in a physically meaningful way for generating good whole-molecule binders.

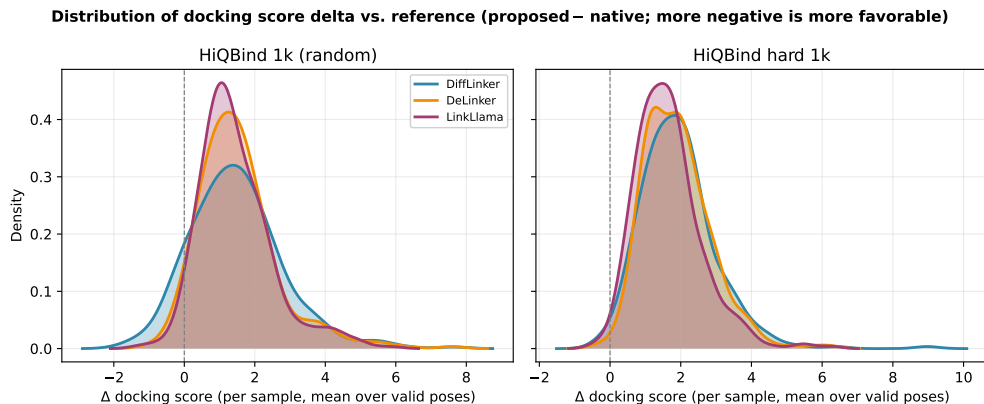


Figure 3.5: **Distribution of mean docking score differences relative to the reference ligands.** The plot displays kernel density estimates of the per-sample mean docking score delta (calculated as the proposed score minus the reference score) on valid instances for the HiQBind random and hard 1k splits.

Case Studies: Scaffold Hopping and PROTAC Design

To demonstrate the practical utility of *LinkLlama* in a realistic drug discovery scenario, we investigate a specific case from the HiQBind dataset of the mineralocorticoid receptor (MR) complexed with a potent antagonist (PDB ID: 6L88) [194]. The reference ligand exhibits a distinct binding mode that its terminal sulfonyl moiety forms a crucial hydrogen bond with Arg817, accessing a previously unexplored sub-pocket in the crystal structure. Simultaneously, its trifluoromethylbenzyl side chain occupies an adjacent small hydrophobic cavity. Recognizing an opportunity for scaffold hopping, we sought to replace the central core to discover novel chemical matter that maintains these critical anchoring interactions.

Although the reference molecule is highly optimized with a favorable docking score (-9.7 kcal/mol), we use *LinkLlama* to unconditionally sample 100 diverse linker candidates. These generated molecules are then rigorously filtered so that they have improved docking scores relative to the reference, low fragment RMSD ($\leq 2\text{\AA}$) to preserve the required 3D vectors of the terminal groups, and high chemical reasonability to pass all five checks. To rigorously evaluate binding stability, the top 10 designs are then subjected to 200 ns molecular dynamics (MD) simulations.

As shown in Figure 3.6, *LinkLlama* successfully generalizes to this complex task. As shown in Chapter 3.6, The model proposed novel, synthetically accessible heterocyclic cores, such as substituted isoxazoles and diazoles, that yields more favorable docking scores than the reference. Structural overlay confirms that the generated scaffolds perfectly maintain the trajectory of the functional terminal groups. Furthermore, the MD trajectories reveal that the generated ligands remain highly stable over the 200 ns duration, exhibiting mean ligand RMSDs (1.30–1.65 $\text{\AA}$) that are tighter than or comparable to the reference ligand (1.70 $\text{\AA}$).

This observation confirms that LinkLlama can reliably produce stable, synthetically realistic candidates that improve upon known potent inhibitors.

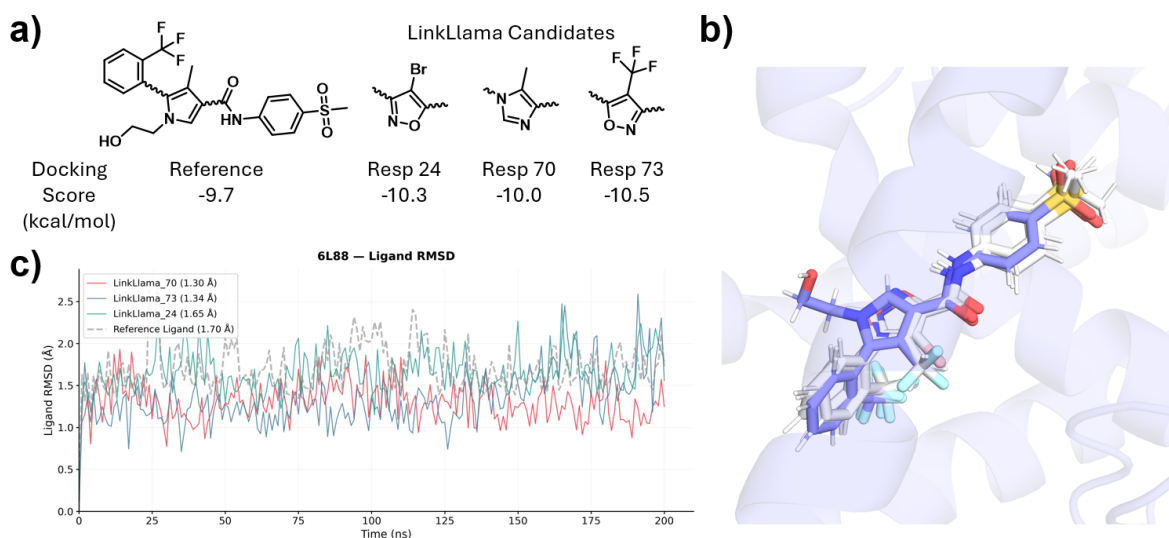


Figure 3.6: **Scaffold hopping case study on the mineralocorticoid receptor (PDB ID 6L88).** (a). 2D chemical structures of the reference ligand and three top-performing LinkLlama candidates (Resp 24, 70, and 73), illustrating the successful generation of diverse heterocyclic cores that improve upon the reference docking score. (b). 3D structural overlay of the docked poses of LinkLlama candidates on top of crystal ligand pose. (c). Ligand root-mean-square deviation (RMSD) over 200 ns molecular dynamics simulations.

We also demonstrate LinkLlama’s utility in proteolysis-targeting chimera (PROTAC) linker design. PROTACs are heterobifunctional molecules consisting of two ligands joined by a linker: one recruits a protein of interest (POI) and the other recruits an E3 ubiquitin ligase [195]. Unlike standard inhibitors, where efficacy is primarily driven by ligand binding affinity, PROTAC activity relies heavily on the linker’s ability to stabilize a cooperative ternary complex (POI–PROTAC–Ligase). Consequently, once the warhead and E3 anchoring ligands are identified, designing a linker with the optimal length, composition, and geometry becomes the critical bottleneck [195]. For this case study, we selected a well-characterized BRD4–VHL PROTAC from PROTAC-DB 3.0 [196] (PDB ID: 6SIS [197]).

In the original 6SIS structure, the PROTAC utilizes a complex macrocyclic architecture. As shown in Chapter 3.7(a), the VHL E3 ligand is shown in blue, and the BRD4 POI ligand is shown in green. We therefore task LinkLlama with generating linear linker alternatives to replace the specific bridging segment highlighted in red, aiming to simplify synthetic complexity while maintaining the vital ternary interaction. Given the expansive chemical space of PROTACs, we generate approximately 300 candidates using diverse property-constraining prompts to prioritize chemical reasonability (Chapter 5.5). Valid designs are docked into the ternary interface, and we select molecules that exhibited low fragment RMSDs compared

to the experimental crystal poses of the anchoring ligands. The top candidates are then evaluated using 200 ns MD simulations.

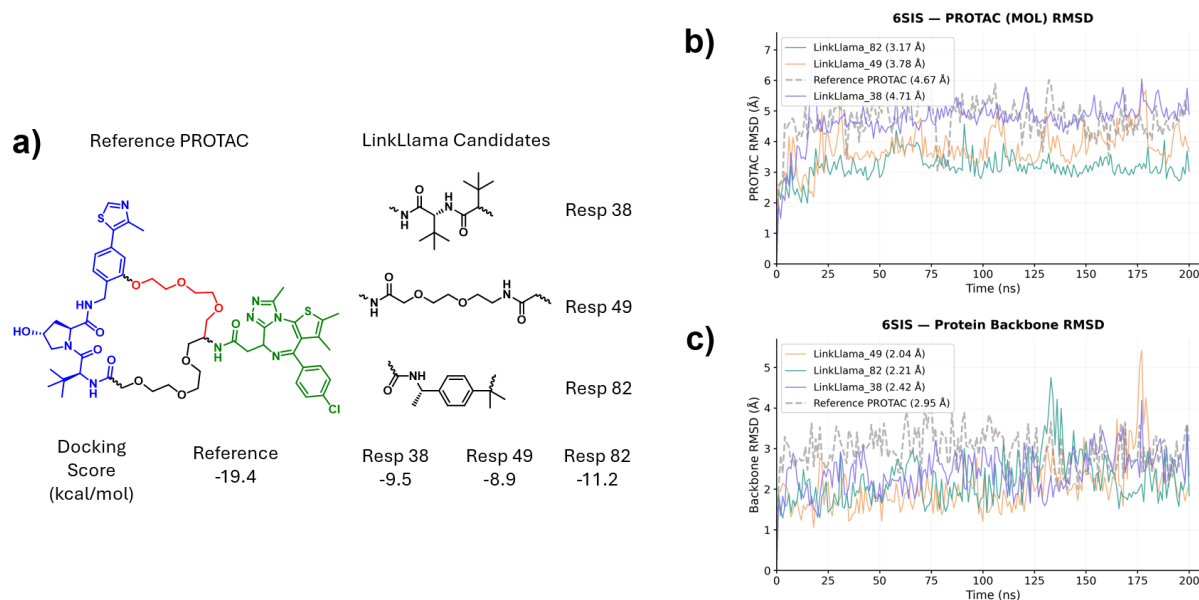


Figure 3.7: **PROTAC linker design for the BRD4-VHL ternary complex (PDB ID 6SIS).** (a). 2D chemical structures of the reference macrocyclic PROTAC (targeted linker region in red, VHL ligand in blue, and BRD4 ligand in green) alongside three linear LinkLlama candidates (Resp 38, 49, and 82) and their corresponding docking scores. (b). PROTAC ligand RMSD over 200 ns MD simulations. (c). Protein backbone RMSD over the simulation.

As shown in Figure 3.7(a), LinkLlama successfully generates diverse linear linkers, incorporating peptidic, PEG-based, and rigid aryl motifs. While these designs bypass the synthetic burden of macrocyclization, MD trajectories demonstrate excellent conformational stability. The generated PROTACs exhibit ligand RMSDs (3.17–4.71 Å) that are comparable to or significantly tighter than the reference macrocycle (4.67 Å). Crucially, the overall ternary complex stability is enhanced by these new linkers as the protein backbone RMSD for the generated candidates (2.04–2.42 Å) outperforms the reference complex (2.95 Å) throughout the trajectory. These results underscore the model’s capacity to navigate complex, long-range spatial constraints and propose viable PROTAC candidates. Overall, both cases demonstrate LinkLlama’s utility for linker proposal with various fragment starting points.

3.4 Discussion and Conclusion

In this study, we introduced LinkLlama, an LLM-based framework that reframes the fragment-linking problem as conditional SMILES string generation task. By leveraging supervised fine-tuning, LinkLlama successfully turns a general-purpose LLM into an expert model capable of navigating the accessible, drug-like chemical space that previous models often struggle to capture. By aligning the chemical reasonability in its training input prompts with the generated linkers, LinkLlama acts as a rational “slicer” of chemical space, successfully producing novel yet diverse linkers for various fragment inputs while strictly maintaining their chemical rationality. Our findings indicate a significant leap in chemical quality, achieving 70–80% chemically reasonable linker designs across various datasets, effectively doubling the performance of previous baseline methods. This capability suggests that medicinal chemists are far more likely to work with the raw hypotheses proposed by LinkLlama due to their superior chemical fidelity.

Another defining advantage of LinkLlama is its capacity for reinforcement learning-free conditional sampling via direct prompt manipulation. Generative models have historically relied on computationally expensive and often unstable RL cycles to achieve multi-objective optimization [28, 29, 184]. LinkLlama replaces this paradigm with an “alignment-by-design” approach. By steering the model through natural language prompts, medicinal chemists can seamlessly tailor chemical property distributions without engineering complex reward functions. This setup lowers the computational barrier for generating actionable leads while maintaining the spatial integrity of fragment hits across diverse modalities. Whether the task involves simple fragment joining, scaffold hopping, or navigating the demanding architectural constraints of PROTAC design, LinkLlama consistently outputs chemically reasonable designs that demonstrate strong computational validation.

Looking forward, the utility of frameworks like LinkLlama is further underscored by the rapidly shifting landscape of structural biology. With continuous advancements in high-throughput X-ray crystallography and cryo-EM like the OpenBind Initiative [198], experimental fragment hit data is becoming increasingly abundant. This wealth of structural data is driving a paradigm shift away from purely *de novo* whole-molecule generation toward a more grounded, iterative pipeline: fragment identification, experimental validation, generative linking, and subsequent validation. In line with the recent ASAP-Polaris-OpenADMET challenge [13] demonstrating that fragment anchors can effectively drive accurate pose predictions [14], LinkLlama capitalizes on these high-confidence spatial constraints. By reliably bridging experimentally validated fragments, it maximizes the value of structural data and ensures that generative design remains tethered to experimental reality.

In another aspect, LinkLlama is designed to function not in isolation, but as a foundational generative engine within a broader LLM-based chemical design ecosystem. Conceived as part of the “Chemical Llama Suite,” it works synergistically with models like SmileyLlama for general molecular generation and SynLlama for retrosynthetic planning. Because these models communicate natively via standard SMILES and natural language, they are perfectly positioned for integration with autonomous AI agents [37, 38, 51]. We envision a

“closed-loop” discovery workflow where an agentic system autonomously identifies promising candidates via LinkLlama, evaluates them through automated docking, FEP, or ADMET simulations, and continuously refines the model’s output using techniques such as Direct Preference Optimization (DPO). This architecture effectively achieves an LLM-based equivalent of the REINVENT tool suite [6], transforming the generative model from a static tool into a dynamic, evolving partner in the transition from isolated fragments to high-quality clinical leads.

3.5 Supplementary Information

Additional Methodology Details

Dataset Curation and Fragmentation Protocol. To ensure the quality and drug-like relevance of the parent pool prior to fragmentation, the initial dataset of 2,854,815 raw molecules sourced from ChEMBL36 undergoes a rigorous, multi-step filtering cascade. Molecules are systematically excluded based on standard physicochemical boundaries and structural viability using RDKit[188]. The sequential filtering protocol results in the removal of compounds based on the following criteria:

- **Validation and Salt Removal:** Molecules failing initial graph construction, failing salt-stripping protocols, or failing standard sanitization checks are removed (1 molecule).
- **Atom Type Restrictions:** Compounds containing non-standard elemental species or non-physiological charge states are excluded. Allowed atom types are strictly limited to the following symbol, valence, and formal charge combinations: C4(0), O2(0), N3(0), F1(0), S2(0), Cl1(0), S6(0), N4(+1), O1(-1), Br1(0), P5(0), and I1(0) (37,213 molecules removed).
- **Minimum Size Constraint:** Extremely small fragments lacking sufficient complexity for linker derivation, defined as having fewer than 6 heavy atoms, are removed (587 molecules).
- **Maximum Molecular Weight:** Exceptionally large macromolecules or polymers with a molecular weight of ≥ 1500 Da are discarded (23,957 molecules).
- **Ring System Constraints:** Molecules with excessively complex topologies, specifically those containing more than 10 total rings or more than 8 aromatic rings, are excluded (4,505 molecules removed).
- **Carbon Ratio:** Compounds with an unusually low fraction of carbon atoms (defined as a carbon-to-heavy-atom ratio of ≤ 0.5), which often indicate highly inorganic or non-drug-like frameworks, are removed (25,659 molecules).

Following this filtering cascade, the surviving structures are canonicalized and deduplicated. This yields a final robust pool of 2,665,082 unique, valid parent molecules (representing 93.4% of the initial input), which subsequently serve as the foundation for the fragmentation and supervised fine-tuning (SFT) corpus.

Computational Fragmentation and Structural Filtration. The generation of training triplets (fragment-linker-fragment) is executed using the Matched Molecular Pair Analysis (MMPA) fragmentation algorithm in RDKit. The procedure is designed to isolate

realistic, synthetically accessible linker units by strictly enforcing a two-cut mechanism per parent molecule.

Bonds to cleave are defined by the SMARTS[141] [#6+0;!\$(*=#,[!#6])]!@!=!#[*]. This specific query restricts fragmentation to acyclic single bonds attached to a neutral carbon atom. The pattern utilizes recursive SMARTS exclusions to prevent the disruption of complex substitution environments or conjugated systems.

Following MMPA decomposition, the resulting structural triplets are parsed into a central core (the linker) and terminal chains (the fragments). These isolated components are subjected to a rigorous filtration cascade. A fragmentation event is considered successful and added to the dataset only if all the following conditions are simultaneously satisfied:

- **Minimum Size Requirements:** The central linker unit must contain a minimum of 3 heavy atoms, and each separate terminal fragment must contain a minimum of 5 heavy atoms.
- **Attachment Point Validation:** The isolated linker subgraph must possess exactly two dummy atoms (atomic number equal to 0), confirming it represents a valid two-point connection bridge rather than a terminal appendage.
- **Topological Separation:** The shortest path between the two dummy atoms on the linker subgraph is calculated to ensure adequate spatial separation between the fragments. We calculate the topological distance as the total number of bonds in the shortest path minus two (accounting for the dummy bonds). The system requires a minimum internal path length of 2.
- **Mass Balance:** To prevent the generation of skewed decompositions where the linker dominates the molecular framework, the number of heavy atoms in the linker must be less than or equal to the heavy atom count of the smallest terminal fragment.

Molecules that fail to produce a valid double-cut matching the specified SMARTS pattern naturally fall back to an empty decomposition state and are subsequently removed from the training corpus, along with any decompositions that fail the topological or size-based filters.

The Chemical Reasonability Workflow. As mentioned in Main, a fragmentation case is reasonable if and only if all 5 chemical reasonability checks pass. We detail the five evaluation criteria below.

- **Bridgehead ring structure (Linker).** Fail if any linker heavy atom belongs to three or more rings in the smallest-set-of-rings decomposition, flagging a single center shared by too many ring systems.
- **Uncommon ring systems (Linker).** Fail if any ring system in the linker has an occurrence count below 100 in the full ChEMBL36 parent database used to build the training pool.

- Undesirable SMARTS (Molecule). Fail the full molecule matches to the curated list of undesirable SMARTS, including cyclopentadiene, cyclopentadiene ylidenes, aromaticity-breaking tautomers, antiaromatic system, unstable halogen-heteroatom bonds, unstable fused rings, allenic system, thiazyl linkages, and peroxide bonds.

```

- [C^2]1=[C^2]-[C^2]=[C^2]~[C;!d4]~[C;!^2;d2]1
- [C^2]1~[C^2]~[C^2]~[C^2]~[C;!^2;d2]~[N]1
- [#6^2]1~[#6^2]~[#6^3;!d4]~[#6^2]2~[#6^2]~
  [#6^2]~[#6^2]~[#6^2](~[*])~[#6^2]2~[#6^2]~1
- [#6]1(=[*])[#6]=[#6][#6]=[#6]1
- [#6]1=[#6][R{2-}]=[R{2-}]1
- [#6^2]1~[#6^2]~[#6^2]~[#6^2]~[#6^1]~[#6^1]~1
- [#7,#8,#16]-[#9,#17,#35,#53]
- [r3,r4]@[r5,r6]
- [*]=[#6,#7,#8]=[*]
- [#7,#16]=[#16]
- [#8]-[#8]

```

In addition to the patterns mentioned above, this check will fail if the generated pyrroles are not in one of the following correct forms.

```

- [N^2]1~[C,N;^2](=[*])~[C,N;^2]~[C,N;^2]~[C^3]1
- [N^2]1~[C,N;^2]~[C,N;^2](=[*])~[C,N;^2]~[C;^3]1
- [N^2]1~[C,N;^2]~[C,N;^2]~[C,N;^2](=[*])~[C;^3]1
- [C,N;^2](=[*])1~[N;^2]~[C,N;^2]~[C,N;^2]~[C;^3]1
- [C,N;^2]1~[N;^2]~[C,N;^2](=[*])~[C,N;^2]~[C;^3]1
- [C,N;^2]1~[N;^2]~[C,N;^2]~[C,N;^2](=[*])~[C;^3]1

```

- PAINS (Molecule). Fail if the full molecule matches to the standard PAINS structural-alert catalog.
- Brenk filters (Molecule). Fail if the full molecule matches to the Brenk medicinal-chemistry rule set.

Preparation of LLM Fine-Tuning Corpus. For all the training data of selection, we apply the same instruction prompt as follows.

You are an expert medicinal chemist specializing in linker design. Your task is to design a linker to connect given fragments and deduce whether the final molecule is chemically reasonable. The output should be in JSON format.

For each input-output training pairs, after collecting their fragment information, linker and molecular properties, and the reasonability checks, we structure these data into the following formats.

FRAGMENT_INFO Given the above information about the fragments and attachment points, design a LINKER_TYPE linker LINKER_PROPERTIES to connect them. The final molecule should be REASONABILITY.MOLECULE_PROPERTIES

The placeholders are substituted during training. FRAGMENT_INFO field is replaced with the fragment SMILES strings, and distance and angle between the two fragments. The REASONABILITY field is always either **reasonable** or **unreasonable** according to the overall reasonability check. All the other field in the input has a 50% chance of showing up in the training script to ensure our training robustness. The specific allowed input for these additional field is shown in Chapter S3.3.

The corresponding output line is a JSON object with keys **linker** (SMILES of the training linker) and **reasoning** (human-readable pass/fail summary for the five reasonability checks) as illustrated in Chapter 3.1.

Supervised Fine-Tuning. After preparing the aforementioned training data into the JSONL format, we fine-tune Llama-3.2-1B-Instruct model (1 Billion parameters) [59] using the Axolotl package [154] for 1 epoch. We apply Low-Rank Adaptation (LoRA) with a rank of $r = 32$ and $\alpha = 16$ to the linear layers of the model[155]. We use FlashAttention-2[156], with the Adam optimizer[157], cross-entropy loss, and a cosine learning rate scheduler with a maximum learning rate of 2×10^{-4} . The training time for our final selected model requires four A100 GPUs on one node for approximately 6.5 hours.

HiQBind/PROTAC docking procedures using Uni-Dock. For the HiQBind and PROTAC docking study, ligands were taken from the model-generated ensembles after the same validity screen used elsewhere in our linker benchmarks. For each protein target, we use the paired refined receptor PDB and co-crystallized reference ligand SDF as the center of the docking box. The box center is the midpoint of the ligand heavy-atom coordinate ranges, and each axis length is the larger of 24 Å and the ligand bounding span plus a 5 Å padding margin. The receptor is cleaned to coordinate records, converted to PDBQT with Meeko[199] using that same center and size, and treated as rigid during docking. Each accepted SMILES is embedded in 3D, prepared as PDBQT with Meeko, and docked with Uni-Dock in fast search mode against the rigid receptor within the defined box. The reported pose corresponds to the top-ranked output, with the docking score read from the PDBQT REMARK fields.

Molecular dynamics simulations via GROMACS[200]. The MD protocol used here follows the iMiner protocol[52]. It initializes with the top-scoring docked pose for each molecule, which is solvated and neutralized with Na^+ or Cl^- ions. The system is parameterized using the AMBER14SB [201], GAFF2 [202], and TIP3P force fields for the protein, ligand, and water, respectively. Following energy minimization, the complex undergoes successive 1-ns equilibrations in the NVT and NPT ensembles at 298.15 K and 1 bar. Then the final production is run for 200 ns for a full simulation and downstream analysis.

Prompts for generating PROTAC linker candidates. To sample diverse PROTAC linkers, we apply the following six prompt combinations and generate 50 samples for each prompt.

- **Unconditional:** reasonable.
- **Large MW / HBA:** reasonable; $\text{hbd} \leq 5$; $\text{hba} > 10$; $\text{mw} > 700$.
- **Lipophilic–polar:** reasonable; $\text{hbd} \leq 5$; $\text{hba} > 10$; $\text{mw} > 700$; $\log P > 6$; $\text{TPSA} > 200$.
- **Ring linker:** reasonable; ring-containing; $\text{hbd} \leq 5$; $\text{hba} > 10$; $\text{mw} > 700$.
- **Branched linker:** reasonable; branched; $\text{hbd} \leq 5$; $\text{hba} > 10$; $\text{mw} > 700$.
- **Chain flexible:** reasonable; chain; $\text{rotb} \geq 4$; heavy atoms ≥ 7 ; $\text{hbd} \leq 5$; $\text{hba} > 10$; $\text{mw} > 700$.

Supplementary Tables

Table S3.3: Configurable prompt fields and allowed input strings for linker and molecule property clauses. LogP uses the $\log P < 0$ branch only when the molecule’s computed $\log P$ is negative.

Field	Allowed literals
linker_type	chain, branched, ring-containing.
rotb_range	0, ≥ 1 , ≥ 2 , ≥ 3 , ≥ 4 .
heavy_atoms_range	≥ 3 , ≥ 4 , ≥ 5 , ≥ 6 , ≥ 7 , ≥ 10 .
hbd_range	> 5 , ≤ 5 , ≤ 3 , ≤ 2 , ≤ 1 .
hba_range	> 10 , ≤ 10 , ≤ 7 , ≤ 5 , ≤ 3 .
mw_range	> 700 , ≤ 700 , ≤ 600 , ≤ 500 , ≤ 400 , ≤ 300 .
logp_range	If $\log P < 0$: < 0 . Otherwise: > 6 , ≤ 6 , ≤ 5 , ≤ 4 , ≤ 3 .
tpsa_range	> 200 , ≤ 200 , ≤ 140 , ≤ 90 .
reasonability	reasonable or unreasonable.

Table S3.4: **Generative performance on the HiQBind random 1k and hard 1k benchmarks.** Conventions match Table 3.1: validity, uniqueness, novelty, and reasonability are reported as percentages, while QED and SA are dimensionless. Percentage columns display the instance-aggregated mean $\pm$ bootstrap standard deviation of that mean. QED and SA represent the instance-mean $\pm$ standard deviation across instances. Novelty values for the baseline methods are excluded due to differing training set derivations.

Split	Model	Validity	Uniqueness	Novelty	QED	SA	Reasonability
Random 1k	DeLinker	94.6 $\pm$ 0.4%	87.4 $\pm$ 0.2%	—	0.44 $\pm$ 0.18	3.84 $\pm$ 0.88	35.6 $\pm$ 1.1%
Random 1k	DiffLinker	87.2 $\pm$ 1.2%	64.2 $\pm$ 1.2%	—	0.53 $\pm$ 0.16	3.73 $\pm$ 0.98	18.3 $\pm$ 1.0%
Random 1k	LinkLlama	99.8 $\pm$ 0.0%	83.9 $\pm$ 0.4%	32.4 $\pm$ 0.5%	0.43 $\pm$ 0.20	3.83 $\pm$ 0.93	53.4 $\pm$ 1.5%
Hard 1k	DeLinker	92.6 $\pm$ 0.6%	88.2 $\pm$ 0.2%	—	0.44 $\pm$ 0.16	3.73 $\pm$ 0.75	44.0 $\pm$ 0.9%
Hard 1k	DiffLinker	88.5 $\pm$ 1.1%	60.9 $\pm$ 1.1%	—	0.52 $\pm$ 0.14	3.56 $\pm$ 0.87	25.7 $\pm$ 1.1%
Hard 1k	LinkLlama	99.8 $\pm$ 0.0%	84.7 $\pm$ 0.5%	40.1 $\pm$ 0.5%	0.44 $\pm$ 0.17	3.62 $\pm$ 0.75	80.9 $\pm$ 0.9%

Supplementary Figures

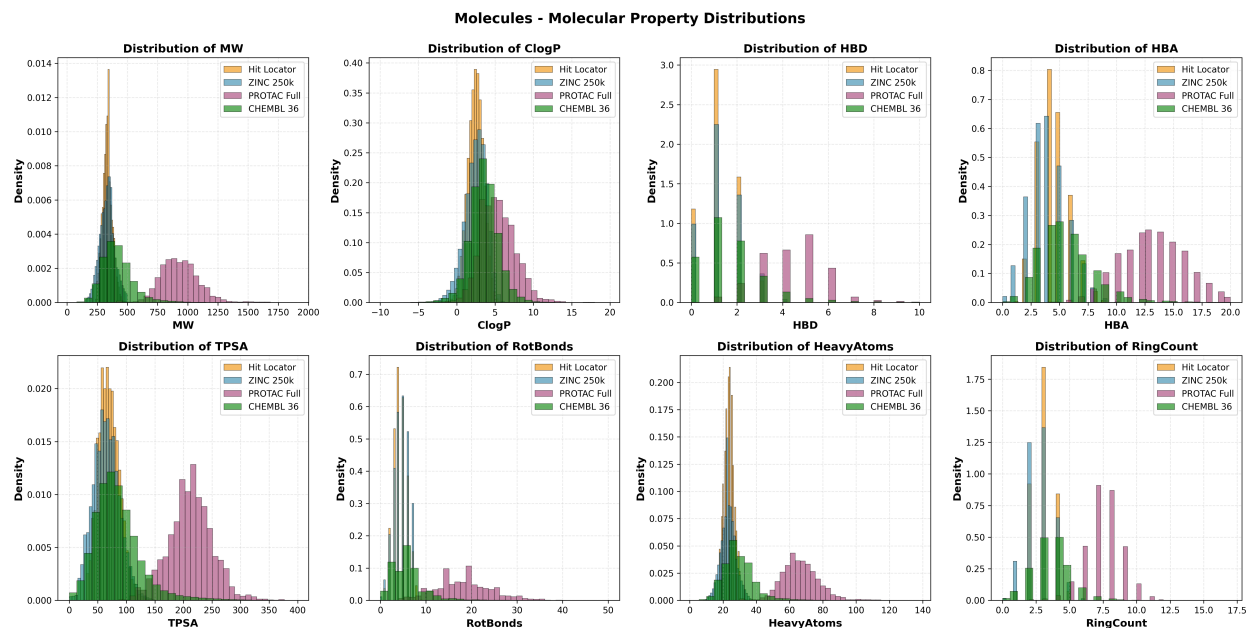


Figure S3.8: Molecular property distributions comparing the ChEMBL 36 dataset against ZINC 250k, Hit Locator, and a curated PROTAC dataset. Evaluated properties include Molecular Weight (MW), ClogP, Hydrogen Bond Donors (HBD), Hydrogen Bond Acceptors (HBA), Topological Polar Surface Area (TPSA), Rotatable Bonds (RotBonds), Heavy Atoms, and Ring Count. ChEMBL 36 exhibits a significantly broader coverage of chemical space, effectively bridging the gap between traditional small-molecule libraries (ZINC, Hit Locator) and larger chimeric molecules (PROTACs). This comprehensive diversity justifies its selection as the parent pool for LinkLlama’s training corpus.

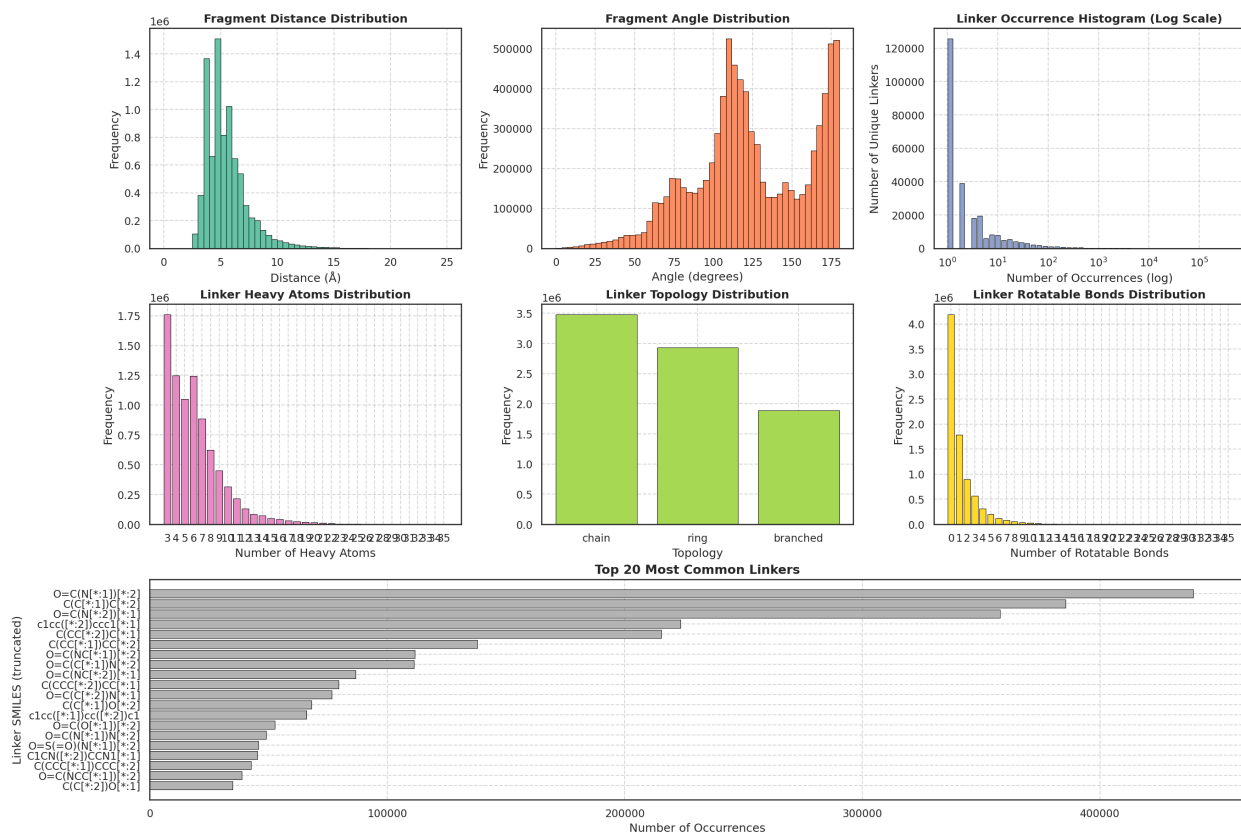


Figure S3.9: **Fragment and linker property distributions for the full ChEMBL36 fragmentation pool.** The top right panel (Linker Occurrence Histogram) and bottom panel (Top 20 Most Common Linkers) illustrate a severe long-tail distribution, where a small fraction of simple, highly prevalent linkers dominates the dataset. This extreme imbalance motivates the need for frequency-capping strategies during the curation of the supervised fine-tuning corpus to prevent model collapse toward common structural motifs.

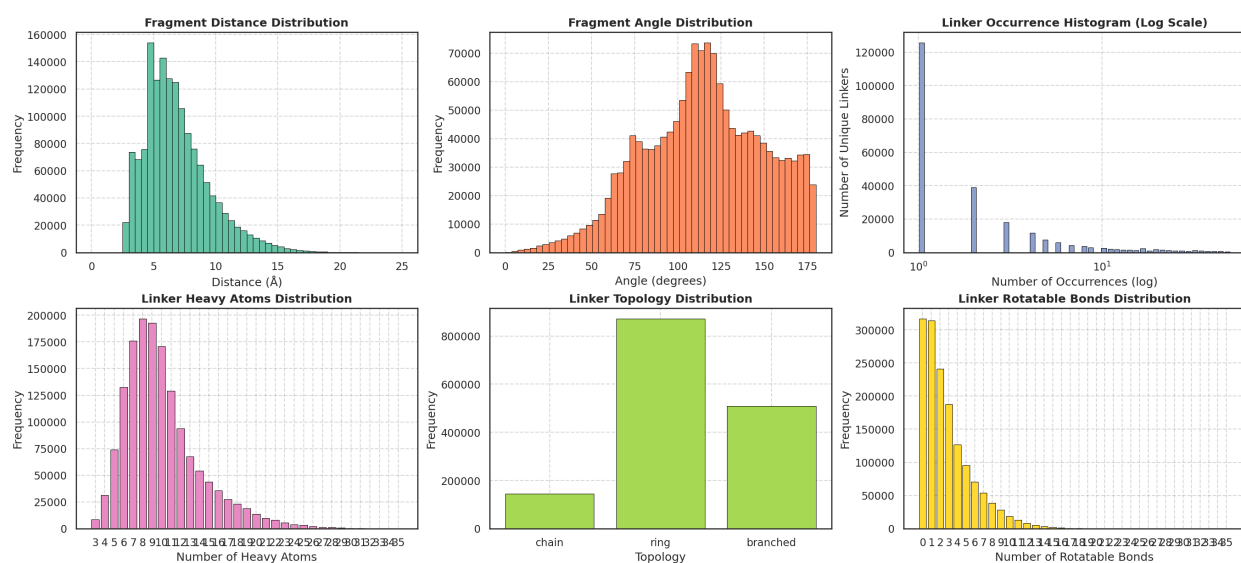


Figure S3.10: **Fragment and linker property distributions for the strictly balanced dataset (Cap50), where the maximum occurrence of any unique linker is restricted to a threshold of 50 examples.** While this approach effectively flattens the occurrence histogram and maximizes linker diversity, it slightly shifts the distributions of heavy atoms and rotatable bonds toward more complex topologies compared to the full parent pool.

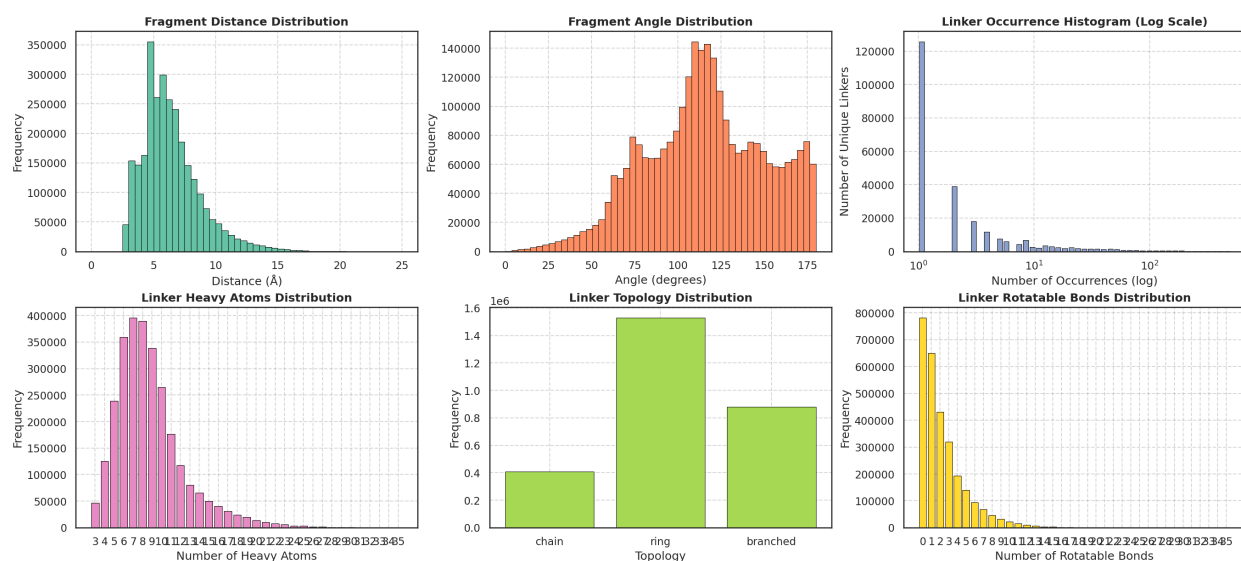


Figure S3.11: **Fragment and linker property distributions for the hybrid balanced dataset (Hybrid), where the maximum occurrence of any unique linker is 500, with 10% reduction for linkers with occurrences from 50 to 500.** This intermediate strategy applies a softened capping mechanism, preserving the natural frequency hierarchy of common synthetic linkages while aggressively truncating the extreme occurrences seen in the full dataset. This balance ensures the model learns a diverse linker vocabulary without losing the inductive bias toward chemically intuitive, frequently synthesized motifs.

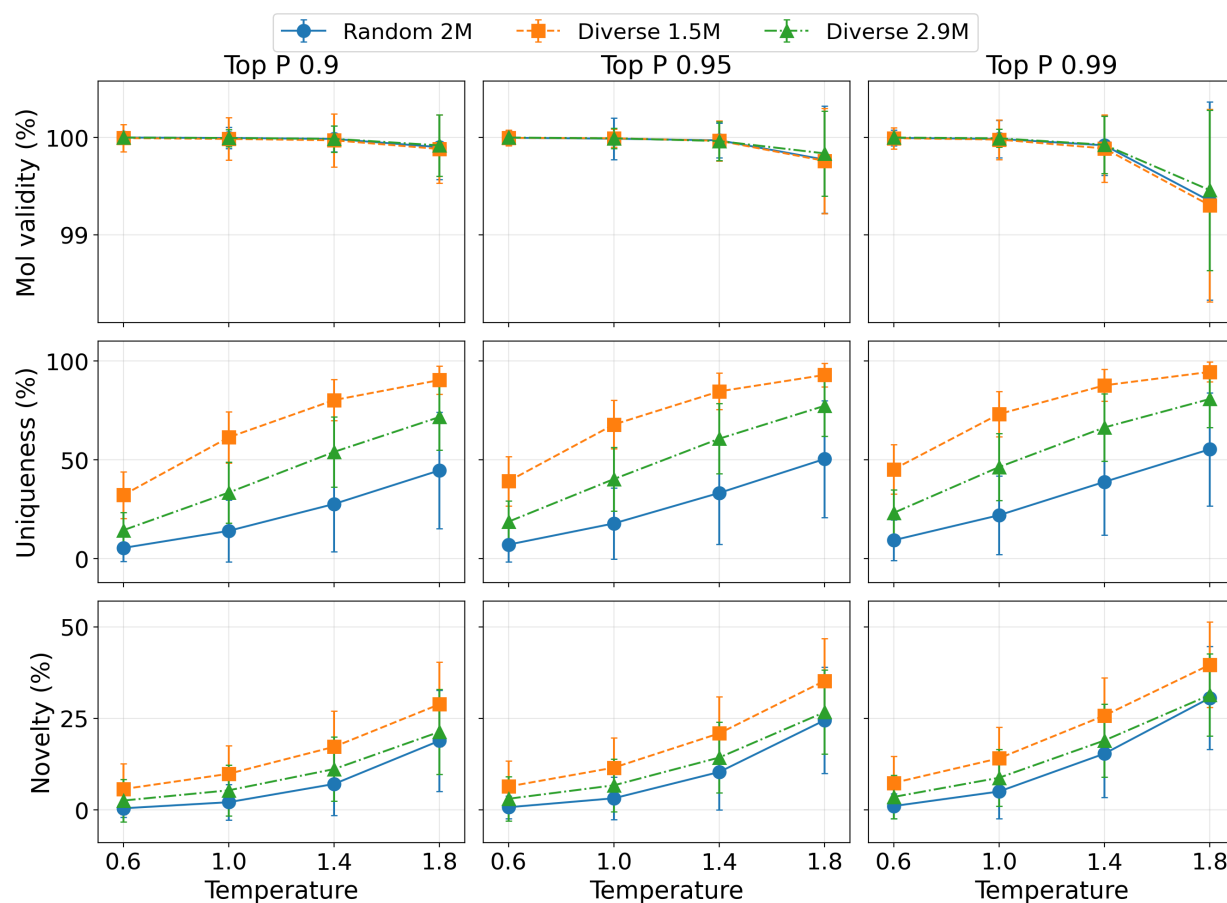


Figure S3.12: **Model hyperparameter searching performances of unconditional generation on the ZINC-random dataset.** The legend denotes the three evaluated models trained on different data splits: **Random 2M** (the long-tailed, unbalanced pool), **Diverse 1.5M** (the strict Cap50 rebalanced pool), and **Diverse 2.9M** (the hybrid-capped pool). Individual data points represent temperature settings ($T \in \{0.6, 1.0, 1.4, 1.8\}$), and the columns correspond to Top-P sampling values ($\text{top-}p \in \{0.9, 0.95, 0.99\}$). The rows display mean molecule validity (%), uniqueness among valid samples (%), and the novelty of valid linkers compared to the training linker set (%), with shared y -axes within each row. Error bars indicate the dispersion across instances and aggregated runs.

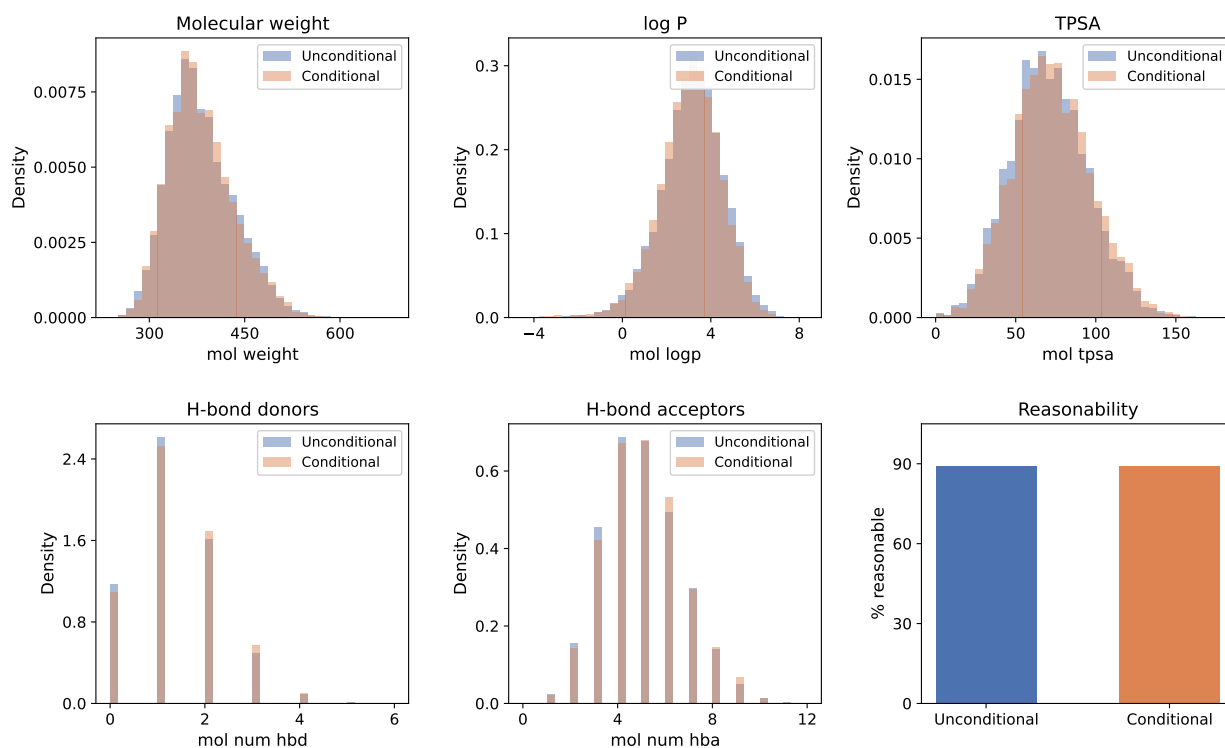


Figure S3.13: **ZINC hard 1k, Ro5-focused prompt: empirical marginals for physicochemical properties compared to the LinkLlama unconditional baseline.** The conditional generation (orange) has minimal changes compared to the unconditional baseline (blue) likely due to the fact that most training samples satisfy Ro5.

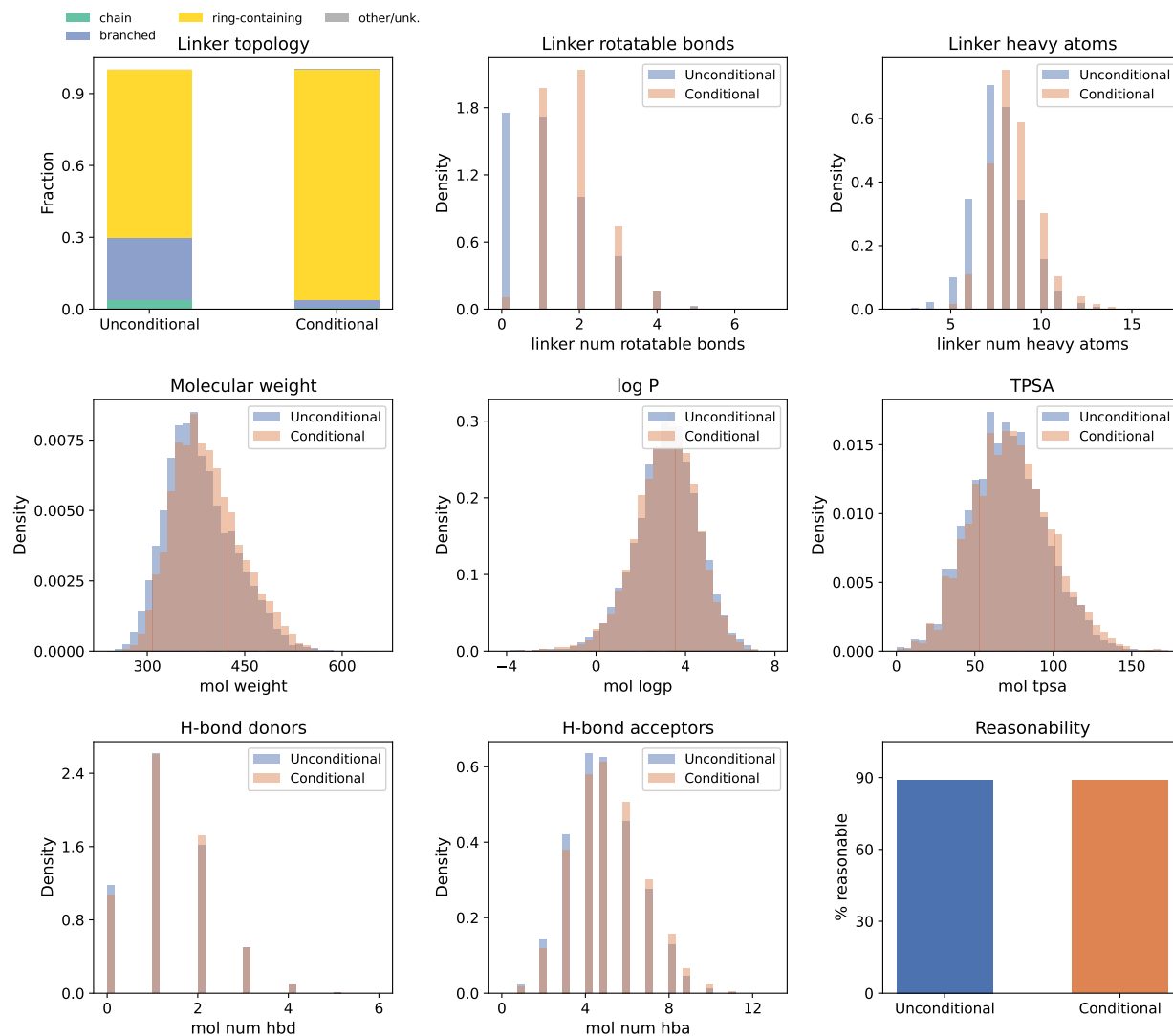


Figure S3.14: **ZINC hard 1k, joint prompt (ring-containing linker, Ro5 bands, rotatable-bond and heavy-atom floors): empirical marginals for properties and linker topology compared to the unconditional baseline.** The conditional generation (orange) demonstrates high prompt adherence across multiple simultaneous constraints, most notably forcing the linker topology to be nearly exclusively ring-containing, while shifting rotatable bonds (≥ 2) and heavy atoms (≥ 6) appropriately without violating Ro5 limits or degrading overall reasonability.

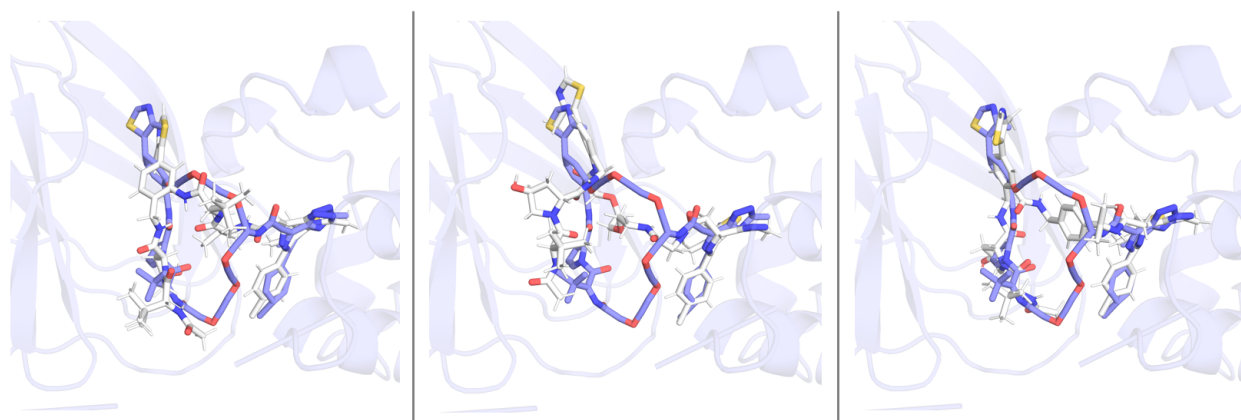


Figure S3.15: **3D structural overlays of the docked poses of LinkLlama-generated linear PROTAC candidates on the crystal poses of the BRD4-VHL PROTAC ternary complex (PDB ID 6SIS).** The reference macrocyclic PROTAC is shown in purple, while three top-performing generated linear candidates are depicted in white in the order of LinkLlama Resp 38, Resp 49, and Resp 82.

Chapter 4

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

‡Authors contributed equally. Kunyang Sun performed the protein annotation part of the HiQBind workflow as well as protein-related analysis.

4.1 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.[203] 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[15, 73, 74, 167, 204–210] or machine learning scoring functions.[75–80, 211, 212] The majority of protein-ligand SF predictors, whether physical or machine-learned, have been trained on the PDBbind dataset[72, 213–218] (<http://www.pdbbind-cn.org/>), specifically v2020, a curated set of $\sim 19,500$ 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 binding affinity data that is used for testing. Various benchmarks based on PDBbind, such as CASF (Comparative Assessment of Scoring Functions) series[81–84], CSAR (Community Structure Activity Resource) 2010[219] and

¹Reproduced 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.

PDBbind-blind-2013[220] 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[41, 42]. 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.[85, 221, 222] 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.[87, 223–225] 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.[86, 226] Other related datasets that focuses more on the geometries of proteins and ligands, including PLINDER[227] and DockGen[228], 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 $\sim 18,000$ unique PDB entries and $\sim 30,000$ 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.

4.2 Methods

The flowchart of HiQBind-WF is illustrated in Figure 6.1. We start by downloading the pdb and mmCIF formats directly from the RCSB PDB[229] 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.

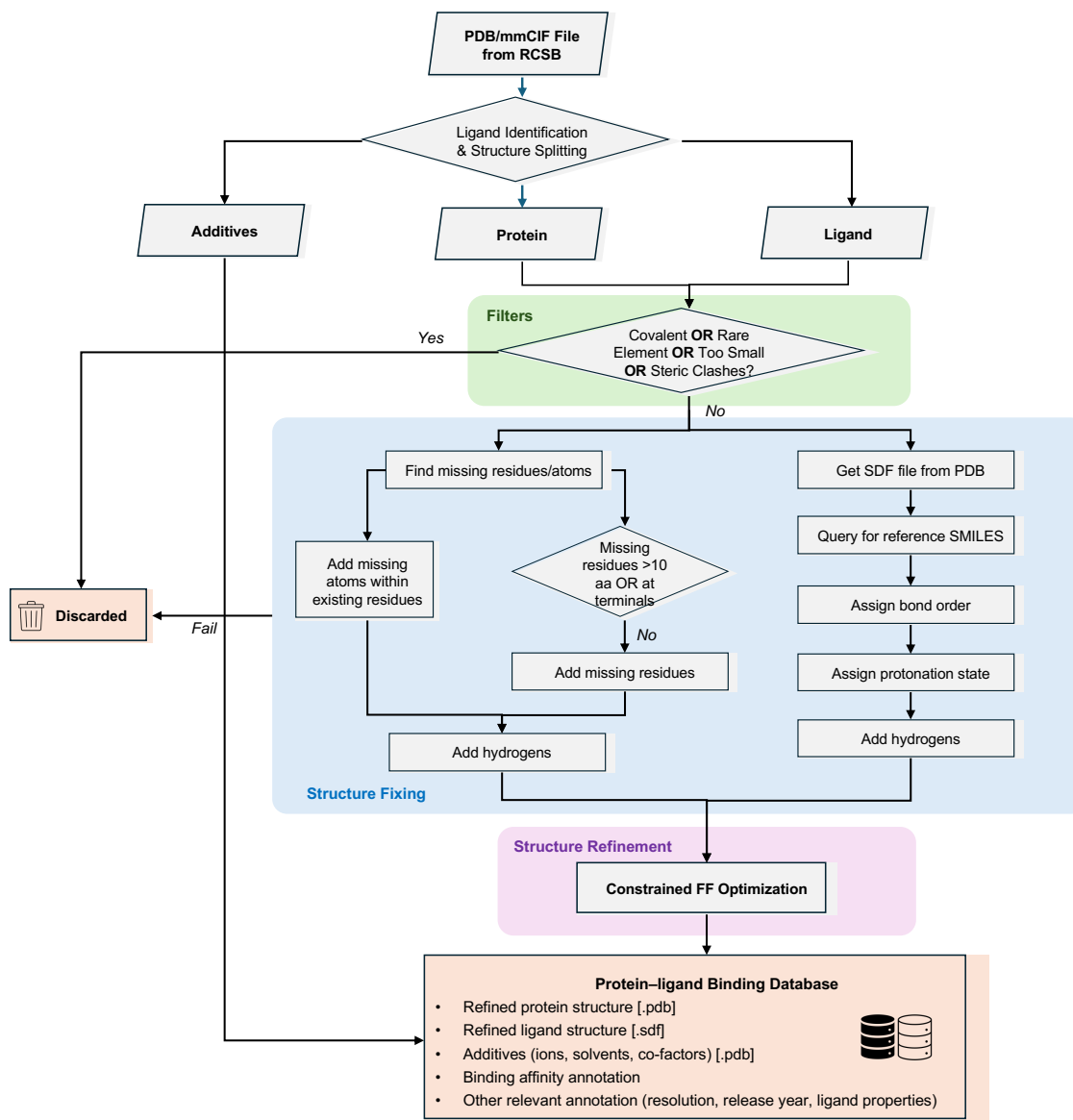


Figure 4.1: *Schematic representation of the semi-automated HiQBind-WF to refine protein-ligand binding structures.* HiQBind-WF downloads the pdb and mmCIF files from the RCSB PDB[229], followed by splitting each structure into three components —ligand, protein, and additives. A series of filters are 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[41]. 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[230].
- 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[11] (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)[231], domains that are not

expressed in the samples for crystallization, or his-tags introduced in the protein purification process[232]. 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 N δ or N ϵ , 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[233]. 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[234–236]. 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 piperazines 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[237] 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[229]) 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[188] (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[166] is used for the protein and OpenFF-2.1.0[167] together with Gasteiger charges[238] 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[10] 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 $\Delta G = RT \ln(K_{d/i})$. [239] However, a large portion of the data in the binding datasets is reported in terms of IC_{50} , which cannot be easily translated to ΔG s due to its dependence on other experimental conditions and inhibition mechanisms. [240] Furthermore, the IC_{50} values for the same protein-ligand complex can vary up to over order of magnitude in different assays[241], 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} .

4.3 Results

Figure 4.2 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)[229] 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 4.2a, which requires special methods to account for the covalent bond formulation[242, 243], and should remain distinct from protein-ligand complexes that are formed from non-covalent interactions only. Figure 4.2b illustrates an example from the PDBbind refined set, 5OUH[244], which is a noncovalent binder that exhibits a severe atomic clash with the protein. Similarly, for 3KMC[245] the chlorobenzene portion of the ligand is absent from the crystal structure as seen in Figure 4.2c. 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.

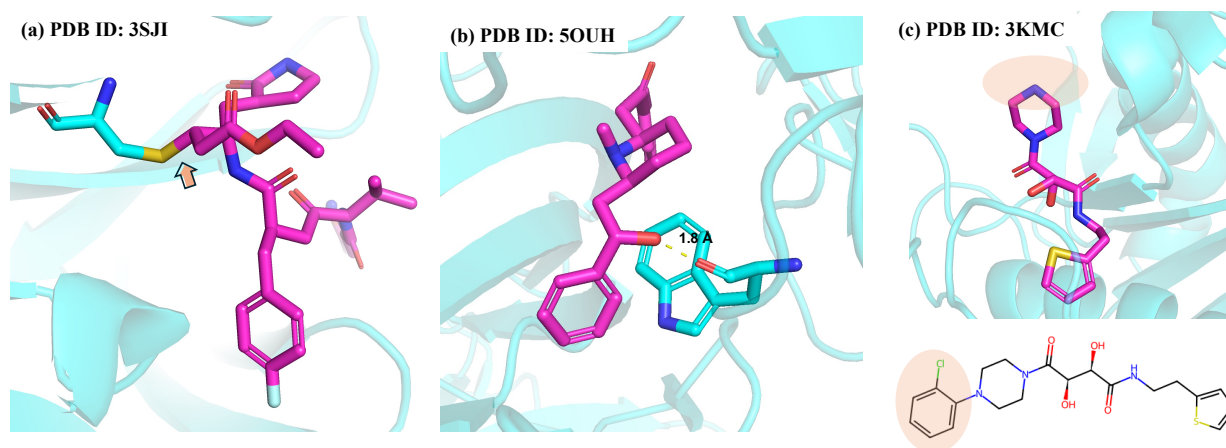


Figure 4.2: *Common structural imperfections in PDBbind dataset.* (a) Covalent binders. The ligand is covalently bound to cysteine with a Michael addition reaction. (b) Steric clashes with the distance between the clashing atoms being only 1.8 Å. (c) Missing atoms.

The HiQBind Workflow

We applied HiQBind-WF to refine the structures in the publicly available PDBbind v2020 dataset[218]. 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[218], 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 4.3. 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[246], the ligand of interest should be the cyclic peptide-like inhibitor, not the sulfonic acid buffer. In other cases, the PDBbind ligand structures are incomplete or the bonding is incorrect (Figure 4.3a). For example, the ligand in 1ALW[247] is missing an iodine atom and in 1DY4[248] the isopropyl is falsely reported as a cyclopropyl (Figure 4.3b,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[249] and 5ETT[250], 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 4.3d,e). In this case, the fixed ligand structures are more chemically feasible and also in line with the protonation states predicted by ChemAxon Marvin[236].

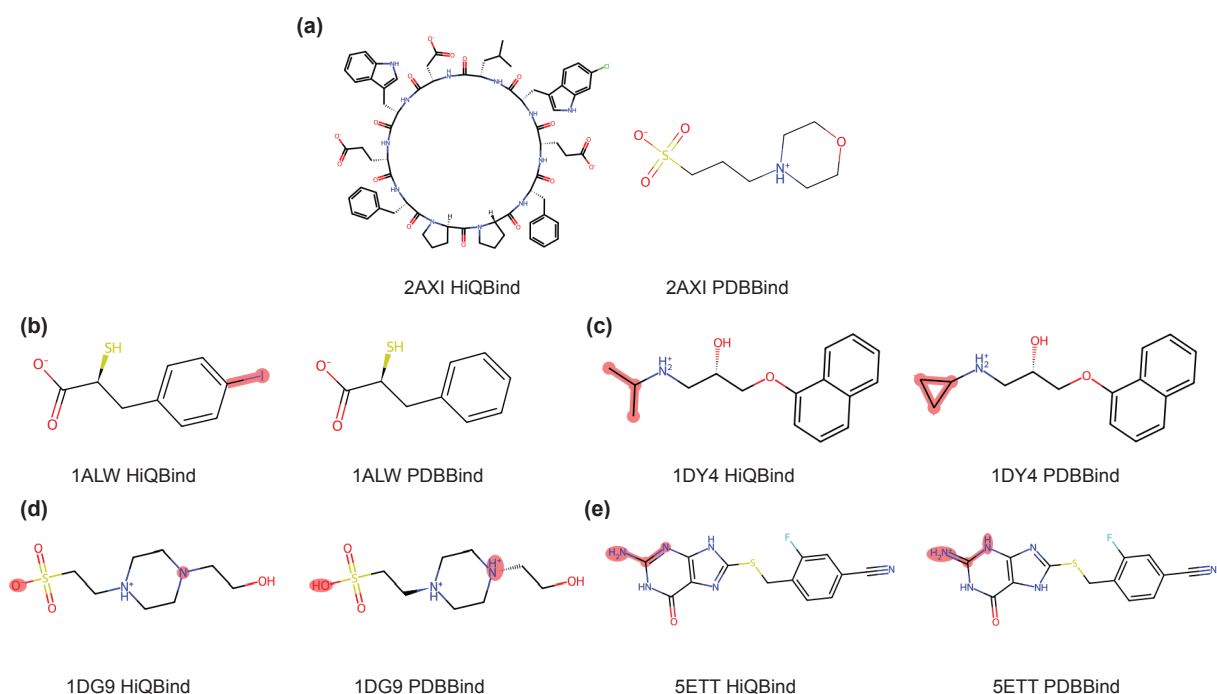


Figure 4.3: *Examples of corrected ligands by HiQBind-WF compared to the original PDBbind.* (a) Wrong ligand entity reported. (b) Missing atoms. (c) Wrong bond connectivity. (d-e) Undesired protonation and tautomer states. The mol2 format ligand files in PDBbind database were used for analysis.

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[188] (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 (oxygen 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[188] to process ligand structures.

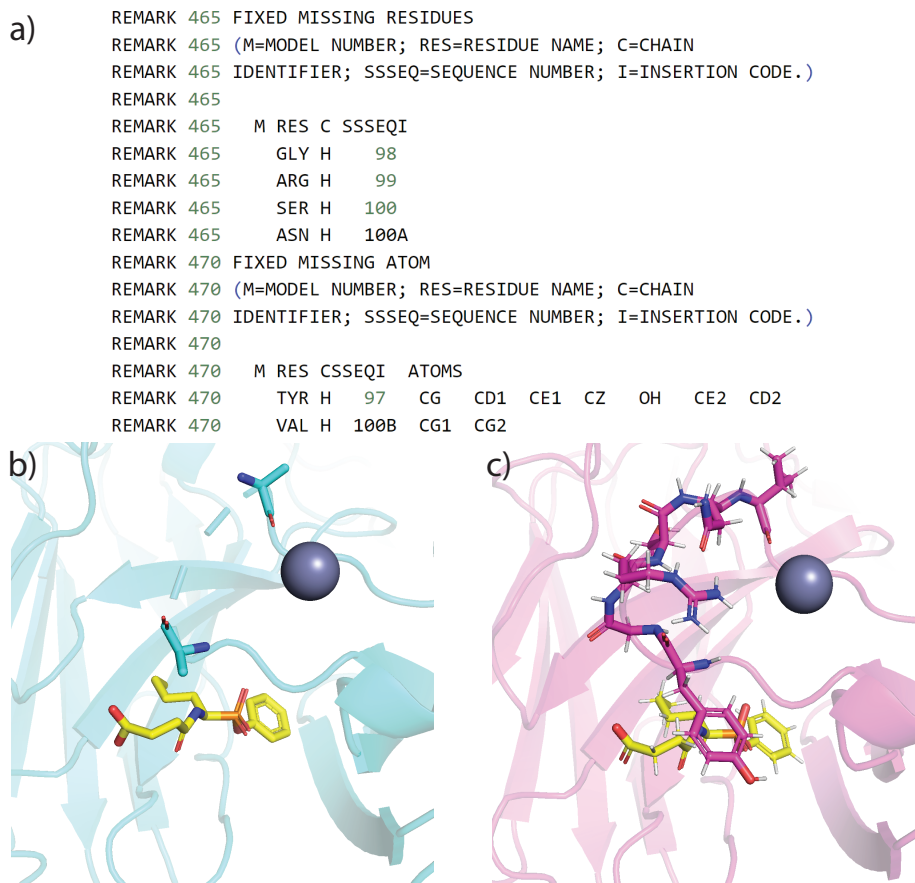


Figure 4.4: *Example of a refined protein structure derived from HiQBind-WF compared to the original PDBbind.* (a) The protein fixing metadata related to residue Y97 through V100B in the refined protein file. (b) Visualization of the original PDB entry with missing residues and atoms centralized at the 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 4.4 shows an example of protein 1A0Q[251] 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.

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[226] and Binding MOAD[224]. 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[224], BindingDB[85, 221], 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 4.5a); (2) the same ligand can bind to different protein pockets (Figure 4.5b); (3) there is more than one ligand with a reported binding affinity (Figure 4.5c); 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

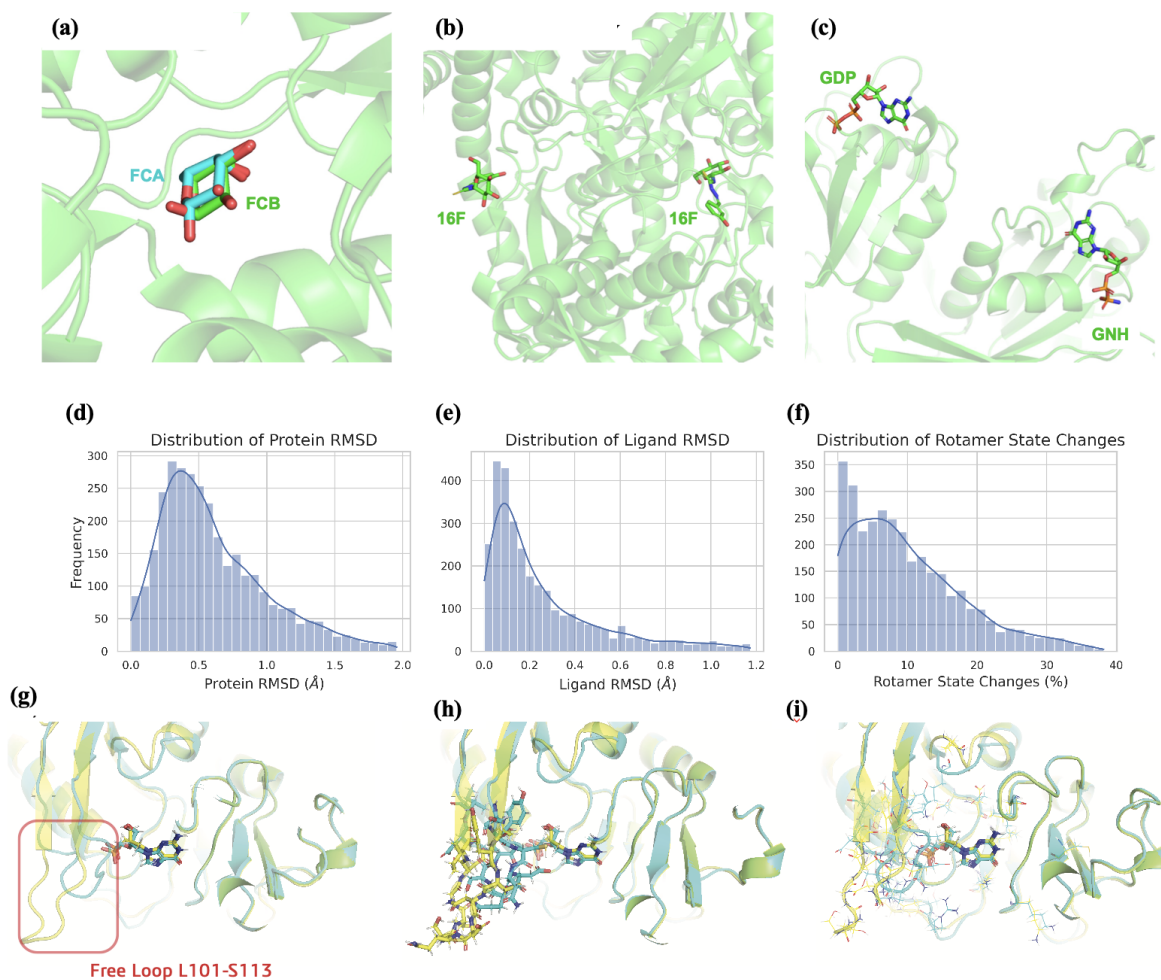


Figure 4.5: *Structural variations and more than one protein-ligand complexes in the same PDB entry.* (a) Two stereoisomers (FCA and FCB) contribute to the binding for 1ABF. (b) The same ligand (16F) binds to two different pockets in 3MRV(c) Two different ligands (GDP) and (GNH) bind to the protein in 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 [218]. 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 4.5(d,e)), there is a significant amount of side-chain rotamer state changes observed for different chains as shown in Figure 4.5(f). Here, following the common practice in the field[252], 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[253] in Figure 4.5(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 $4\text{\AA}$, 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 4.6 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[215], or split the data based on time[220, 227], sequence similarity[41], ECOD classifications[227, 228], or protein-ligand interaction profiles[227]. 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.

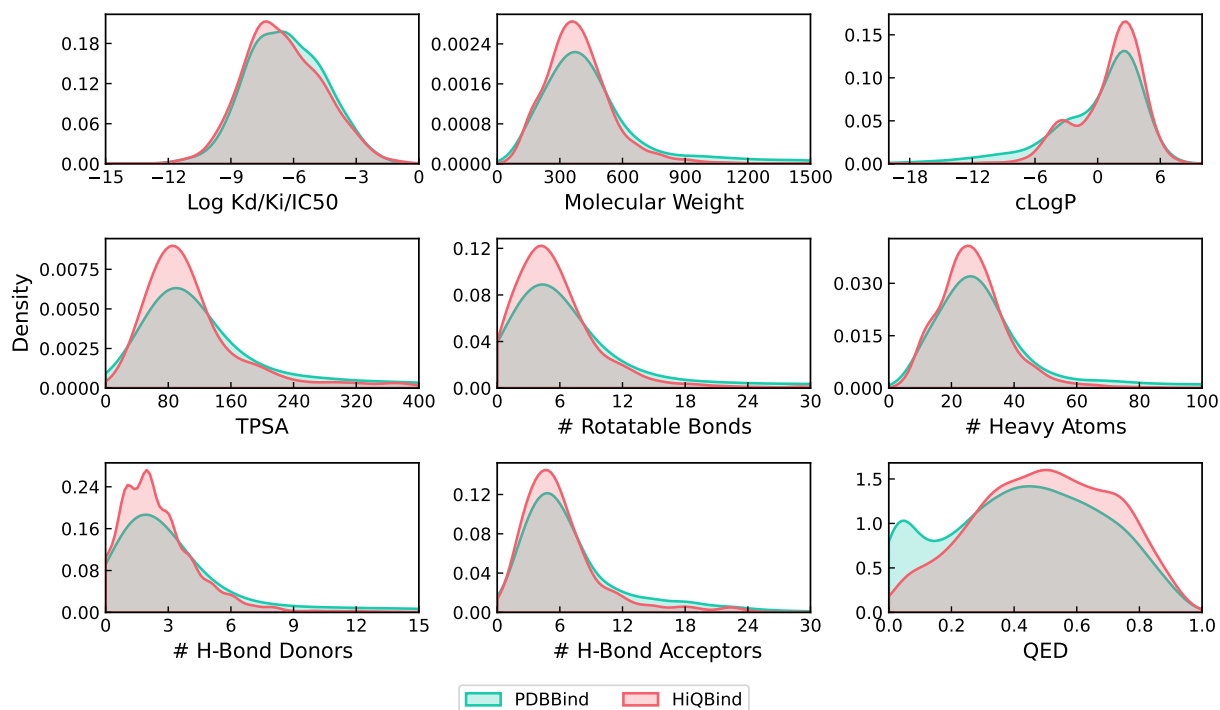


Figure 4.6: Comparison of the distribution of drug-related properties of ligands and their binding affinities between PDBbind and HiQBind. (a) Binding K_d , K_i , and IC_{50} values in log units. (b) molecular weight, (c) computed $\log P$ value, (d) topological polar surface area (TPSA), (e) the number 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.

4.4 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[72, 213–218], BioLiP2[86, 226], Binding MOAD[224], and related databases such as Binding DB[85, 221], Plinder[227], and Dockgen[228]. 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[226] and Binding MOAD[224] 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.

4.5 Supplementary Information

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 are small molecules
- `hiqbind_poly_metadata.csv`: Metadata for subset of HiQBind in which ligands are polymers
- `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 -xzf hiqbind.tar.gz` will yield two directories `raw_data_hiq_sm` and `raw_data_hiq_poly` corresponding 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 S4.1 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 S4.1: Overview of the files in HiQBind.

File	Description
*_ligand_refined.sdf	Refined ligand structures (hydrogen added, correct bond order, better tautomer states/protonation 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 metadata file in in Table S4.2.

Table S4.2: 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: "=", "i=", "j=" 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, separated 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 existing dataset, such as PDBbind v2020, here we provide some relevant analysis in Figure S4.7. 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-affinity annotation from.

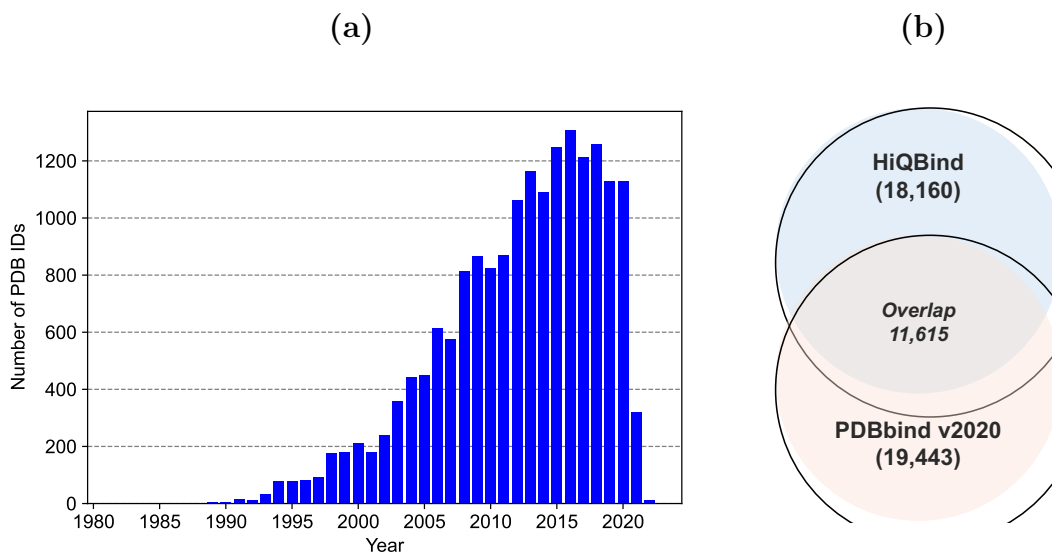


Figure S4.7: **(a)** The year distribution of PDB entries in HiQBind. HiQBind contains 1,463 PDB entries deposited after 1/1/2020. **(b)** HiQBind PDB entries overlapping with those in PDBbind v2020. Among 18,160 unique PDB IDs in HiQBind, 11,615 of them appear in PDBbind v2020.

Failed cases in processing HiQBind

Table S4.3: 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 successfully 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 structures.

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, 3uli, 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, 6ble, 6bl1, 6cli, 6cge, 6cha, 6di9, 6e6e, 6ezp, 6hhf, 6hhi, 6iuo, 6j99, 6jpp, 6jwl, 6jx0, 6jx4, 6jxt, 6k1s, 6lw2, 6lze, 6m0k, 6md1, 6mzw, 6o0e, 6o0f, 6oni, 6p69, 6p7i, 6pdz, 6pnm, 6pnn, 6pno, 6qw7, 6r4v, 6rnu, 6swm, 6t5y, 6tft, 6tpm, 6v6k, 6v6o, 6v9c, 6vh4, 6vim, 6why, 6wi0, 6wtj, 6wtt, 6wxz, 6xd3, 6xhm, 6xmz, 6xr3, 6y2f, 6yi8, 6yzt, 6zpi, 7b3e, 7bwd, 7c6u, 7c8u, 7cb7, 7cbt, 7dlm, 7dpp, 7jkv, 7jxl, 7jxp, 7jxw, 7k0e, 7k1h, 7lcl, 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, 3c wd, 3d04, 3fek, 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, 5gl, 5glz, 5gm0, 5h5o, 5h9j, 5hk9, 5hn7, 5hn9, 5hyr, 5i3b, 5is0, 5ivt, 5syf, 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, 2jip, 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, 4jff, 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, 5fqh, 5fqe, 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, 6hww, 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:

1lgs, 1a07, 1a2c, 1a5v, 1ad8, 1agm, 1apv, 1apw, 1aqc, 1at5, 1at6, 1axw, 1b11, 1b40, 1b5q, 1b6j, 1bdq, 1bfm, 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, 1lzf, 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, 1v1l, 1v16, 1v1m, 1vwl, 1vwn, 1wlg, 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, 1xxz, 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, 2ggv, 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, 2wgc, 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, 2yjg, 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, 3nkv, 3nq6, 3nw3, 3nxx,
 3ny3, 3o0w, 3o0x, 3o6m, 3oea, 3oeb, 3oka, 3okp, 3old, 3ole, 3olg, 3oli, 3oy8, 3p36, 3pj1, 3q8d, 3qg6,
 3qmo, 3qx3, 3qxx, 3qzs, 3r42, 3rjf, 3rme, 3rss, 3rul, 3run, 3ryb, 3sm1, 3sou, 3sow, 3sp5, 3t8v, 3tcg,
 3ti4, 3tib, 3tkz, 3tl0, 3tlh, 3u4s, 3u4w, 3u5l, 3ur0, 3uvm, 3uwn, 3v0w, 3vf9, 3voz, 3vp1, 3vp2, 3vtr,
 3vyk, 3waw, 3wch, 3wel, 3weo, 3wol, 3wom, 3won, 3wqd, 3wqv, 3zd3, 3zkk, 3zq9, 3zst, 3zvy, 3zw2,
 3zyb, 3zyh, 3zyr, 3zyx, 4a23, 4a3x, 4a45, 4a7a, 4a9u, 4aa1, 4ad2, 4ad3, 4agl, 4aif, 4ap0, 4aph, 4apr,
 4asl, 4aay, 4av0, 4av5, 4avi, 4avj, 4aze, 4b4q, 4b83, 4b9z, 4bdo, 4bgx, 4bh0, 4bj0, 4bnk, 4bqr, 4btl,
 4c02, 4c0r, 4clt, 4clw, 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,
 4l6j, 4l6k, 4l6l, 4l6m, 4l6n, 4l6o, 4l6t, 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, 4paw, 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, 5elb, 5eld, 5elo, 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, 5g6l, 5ggo, 5ggp, 5ghv, 5glu, 5glw, 5gm, 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, 5kef, 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, 5vr, 5vry, 5vrz, 5vs1, 5vs2, 5vs3, 5vv8, 5w3y, 5w7i, 5w7x, 5wei, 5we, 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, 6dq, 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, 6fnn, 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, 6ilu, 6i41, 6i5p, 6i5v, 6i5w, 6i68, 6i7a, 6ial, 6iam, 6idg, 6im4, 6inz, 6jad, 6jao, 6jba, 6jue, 6kgo, 6kgq, 6kxa, 6kxb, 6llf, 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, 6rfl, 6rk4, 6rt6, 6rt7, 6rti, 6ryo, 6sli, 6sey, 6sgf, 6slo, 6spw, 6spx, 6sq0, 6sw, 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, 8icu, 8icx, 8icy, 8kme, 9icd, 9icu

Note: 1gx4 failed only in small molecule set.

Failed cases in processing PDBBind

958 entries are identified with covalent binders:

1a09, 1a46, 1a5g, 1a6l, 1ad8, 1aht, 1amn, 1au0, 1au2, 1auj, 1avp, 1awf, 1awh, 1ayu, 1ayv, 1b0f, 1b5g, 1bgo, 1bio, 1bjr, 1bm, 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, 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,

Table S4.4: 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 ^c

^a In parenthesis is the number of protein-ligand structures.

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, 1vsu, 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, 2bd1, 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, 3zsl, 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, 4mnn, 4mnnv, 4mnnv,

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, 4qwv, 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, 5clx, 5clx, 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, 5gtq, 5gwa, 5gwz, 5h6v, 5hg5, 5hg7, 5hg8, 5hg9, 5hl9, 5h1b, 5h1d, 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, 5swb, 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, 6jpp, 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, 6rtu, 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, 5glp, 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, 6jjj, 6jki, 6jz0, 6k4r, 6kjf, 6myn, 6n55, 6n92, 6n94, 6n96, 6nt2, 6nzc, 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:

1lgs, 1a07, 1a0t, 1a2c, 1a37, 1a3e, 1abf, 1abt, 1af2, 1agm, 1apb, 1apv, 1apw, 1aqc, 1at5, 1at6, 1atl, 1aze, 1azx, 1b11, 1b2m, 1b40, 1b6j, 1bap, 1bd1, 1bdq, 1bm2, 1bm6, 1bsk, 1bt6, 1bux, 1bzh, 1c5o, 1c5p, 1c5z, 1cka, 1ckb, 1clu, 1cpi, 1cyn, 1czq, 1d4w, 1d6s, 1d8e, 1dkd, 1dmb, 1dva, 1dpx, 1e03, 1e5j, 1eb1, 1ec9, 1eef, 1ej4, 1eoj, 1eol, 1epq, 1eub, 1evh, 1ez9, 1f47, 1f4y, 1f5k, 1ff1, 1fh7, 1fh8, 1fh9, 1fhd, 1ffs, 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, 1vw1, 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, 2ggg, 2gh9, 2h13, 2h2d, 2h2e, 2h2g, 2h2h, 2h6k, 2h6q, 2h9m, 2h9n, 2h9p, 2hah, 2hdx, 2hj4, 2hjb, 2hkf, 2hnh, 2hrp, 2ig0, 2igv, 2igw, 2itk, 2iv9, 2ivz, 2j7w, 2j9n, 2jb5, 2jbu, 2jdk, 2jdl, 2jg8, 2jjb, 2jk9, 2jkr, 2jkt, 2jnj, 2jnw, 2jq9, 2jqk, 2k1q, 2k3w, 2kgi, 2knh, 2kup, 2loi, 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, 3qyv, 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, 3uyl, 3uvm, 3uvm, 3uvm, 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, 3wu, 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,
 4aay, 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,

4dec, 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, 4fc m, 4fe9, 4fem, 4fgx, 4fgy, 4fk7, 4fmn, 4fmo, 4fmq, 4fn5, 4ft2, 4fut, 4g0a, 4g5f, 4g68, 4g69, 4gah, 4gao, 4gj8, 4glr, 4glx, 4gne, 4gnf, 4ngn, 4gq6, 4gvc, 4gv d, 4gw1, 4gw5, 4gwi, 4gx l, 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, 4ll u, 4l58, 4l6t, 4lbl, 4lbo, 4lg6, 4lj h, 4lk6, 4lk7, 4lkd, 4lke, 4lkf, 4lkg, 4lkh, 4lkk, 4lkm, 4ln2, 4lnf, 4lno, 4lnp, 4lp6, 4lq3, 4lte, 4ml d, 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, 4oe e, 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, 4pv o, 4pxf, 4pz5, 4pz8, 4ql e, 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, 4tw t, 4tz m, 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, 4wc i, 4wey, 4wht, 4why, 4wj7, 4wko, 4wph, 4wq3, 4wrq, 4wv6, 4wy7, 4wym, 4x0z, 4x13, 4x14, 4xl n, 4xp, 4xlq, 4xlr, 4xls, 4x34, 4x3e, 4x3h, 4x3i, 4x3k, 4x3r, 4x3s, 4x6h, 4x6s, 4x8n, 4x8p, 4x9r, 4x9v, 4x9w, 4xc2, 4xe k, 4xgz, 4xh2, 4xqu, 4xtp, 4xx9, 4xxh, 4xyn, 4y32, 4y3b, 4y5i, 4yb5, 4yc8, 4ydf, 4ydn, 4yee, 4yef, 4yhp, 4yhz, 4yje, 4yj1, 4yk0, 4ykj, 4ykk, 4ym2, 4ynl, 4yoz, 4ysi, 4yw2, 4yy6, 4yyi, 4yy m, 4yy n, 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, 5bt, 5btv, 5c0m, 5c11, 5c13, 5clm, 5c6v, 5c7e, 5c7f, 5cbm, 5cfa, 5cil, 5cin, 5cqj, 5cqx, 5cr7, 5cs2, 5csz, 5cvd, 5cw8, 5cx i, 5d0j, 5dl u, 5d2a, 5d6y, 5d7e, 5dah, 5dif, 5dms, 5dtj, 5duw, 5dx b, 5dx e, 5dxg, 5e0l, 5e0m, 5elb, 5eld, 5elo, 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, 5gm v, 5gp7, 5gs4, 5gtr, 5gu4, 5gwy, 5gx6, 5gx7, 5hle, 5h5q, 5h5r, 5h5s, 5hct, 5hda, 5heb, 5hed, 5hes, 5hex, 5hey, 5hfl, 5hfb, 5hfc, 5hff, 5hhx, 5hjb, 5hjc, 5hjd, 5hkh, 5hlp, 5hog, 5hpm, 5hte, 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, 5kg n, 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, 5yjj,
 5ypo, 5ypp, 5ypw, 5yqw, 5yto, 5ytu, 5yv5, 5yvx, 5yy4, 5yy9, 5yyz, 5yzd, 5z89, 5z95, 5zbx, 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, 6civ, 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, 6hll, 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, 6qsx, 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 5

KinConfBench: A Curated Benchmark for Cofolding Models on Kinase Conformational States

5.1 Introduction

Highly accurate data-driven protein structure prediction models [254–259] have become powerful tools for structural biology. By combining evolutionary information with structure-based training using the Protein Data Bank (PDB) [260], these models have allowed researchers to generate plausible protein structural hypotheses from sequence, especially for proteins that are not structurally enabled by experiment. Inspired by this advancement, new directions have opened up to extend the framework beyond single-chain proteins, predicting protein complexes with other proteins[261, 262] or biologics [44–47, 88, 263, 264]. Consequently, a new class of “cofolding” models has emerged, designed to predict structures for biomolecular complexes including proteins, DNAs, RNAs, post-translational modifications, ions, small molecules, and other modalities critical for drug discovery. Models such as NeuralPLexer[47], AlphaFold3[44], Chai-1[45], Protenix[264], and Boltz 1 and 2[46, 88] have shown promise in accelerating structure-based drug discovery (SBDD) campaigns by generating structural hypotheses at a reasonable computational cost.

Despite their rapid adoption, data-driven cofolding models suffer from distinct generalization issues. For instance, Škrinjar et al. observe that these models exhibit significantly lower success rates and confidence scores when tasked with systems outside their training distribution [48]. Beyond these failures in low-data regimes, even predictions with high confidence scores can exhibit physically invalid steric clashes or distorted ligand geometries that fails the benchmark evaluations by the PoseBusters suite [43]. Moreover, recent benchmarking efforts have attempted to probe limitations in other regimes such as covalent modifications [265],

¹Reproduced with permission from: KinConfBench: A Curated Benchmark for Cofolding Models on Kinase Conformational States. K. Sun and T. Head-Gordon. BioRxiv, 2026, 10.64898/2026.04.07.716788.

PROTACs [266], and molecular glues [267]. Besides these emerging and less common modalities, even within the well-represented kinase complexes, Nittinger et al. reveal that cofolding models still struggle to place allosteric ligands into their allosteric pockets due to the overrepresentation of orthosteric ligands in the PDB training data [268]. In summary, these studies expose the weaknesses of the current cofolding models, collectively suggesting directions for improvements of the next-generation cofolding models.

However, while these benchmarks successfully identify domain-specific challenges, they largely overlook a more fundamental requirement of modern SBDD: the capacity for ligand-modulated conformational reasoning. As detailed above, current evaluations remain predominantly ligand-centric, prioritizing the geometric placement of the small molecule through mostly pocket-aligned ligand Root Mean Square Deviation (RMSD) and overall global structural confidence in the protein fold. While these metrics are excellent starting points to verify basic structural integrity, they treat the protein receptor as a static scaffold rather than a dynamic landscape [269]. In practical drug discovery, identifying the correct global fold is necessary but not sufficient; successful lead optimization requires capturing the specific induced-fit conformational states dictated by ligand binding. Without a focused assessment of this conformational response, there remains a critical gap in understanding whether cofolding models can truly drive rational drug design.

The human kinase family presents an ideal paradigm to address this gap, given its strong representation in the PDB, its importance in therapeutic discovery [89–91], and the inherently dynamic nature of kinase conformational landscape [270–272]. Functioning as molecular switches, kinases toggle between active and inactive states via subtle rearrangements of the activation loop (DFG motif) and the α C-helix [92, 93]. More importantly, the utility of a kinase structure in drug discovery is defined by these specific conformational states, which dictate inhibitor binding kinetics and selectivity that helps distinguish Type I from Type II inhibitors [94]. Consequently, it is imperative and timely to evaluate the capacity of cofolding models to provide biophysically valid structural hypotheses, specifically by quantifying how well they capture the kinase conformational landscape and generalize across unseen ligand-induced states for drug discovery.

In this work, we introduce KinConfBench, a curated benchmark of 2,225 high-quality human kinase chains representing a total of 1,420 unique kinase systems (43 unique apo targets and 1,377 unique holo complexes), designed to shift the evaluation focus from ligand RMSD to protein-ligand conformational fidelity. To rigorously quantify these states, we employ the structural nomenclature developed by Modi and Dunbrack [93], labeling kinase chains with categorical structural features to capture specific active and inactive conformations. Using this framework, we evaluate three widely used open-source cofolding models—Boltz-2[88], Chai-1[45], and Protenix[264]—on their capability to recover experimentally observed kinase states in the presence of bound small-molecule ligands. The resulting metrics show that high confidence in the active site geometry does not ensure high confidence in predicting the kinase conformational states, especially when ligands perturb the kinase away from common templates of the global fold. We further document a critical generalization gap on kinase holo-apo pairs. When challenged with unseen holo states after the training cutoff, all three

architectures increasingly predict the apo state, underscoring that the models are memorizing the apo-state baselines of the training set and poorly generalizing to true induced-fit outcomes. KinConfBench therefore foregrounds protein conformational correctness, not ligand and RMSD alone, as the quantities that ultimately need to be ranked and interpreted to make progress in SBDD campaigns.

5.2 Methods

When performing inferences for Boltz-2 [88], Chai-1 [45], and Protenix [264], we adhere to the official inference pipelines for each method to represent standard usage scenarios. Multiple Sequence Alignments (MSAs) are generated and cached with each tool’s default databases and pairing strategies. For Boltz-2, Chai-1, and Protenix, the ColabFold server [256] is used for MSA generation, while sequence pairing follows the defaults of the respective model repositories. Input sequences are restricted to the canonical kinase domain boundaries defined by UniProt, so that the benchmark stresses domain-level cofolding behavior.

Inferences are performed on NVIDIA A100 GPUs across the 1,420 unique configuration pairs (43 apo + 1,377 holo) defined during curation. Post-inference, structures are passed through the Dunbrack KinCoRe annotation pipeline; targets that failed annotation because of malformed outputs or missing motifs are excluded. To ensure robust sampling of the conformational landscape, we generate $N = 20$ cofolding samples per target for each inference run. Output structures are saved in mmCIF format to ensure compatibility with the KinCoRe annotation pipeline.

5.3 Results

The KinConfBench Conformational Benchmarks

To rigorously evaluate the capability of cofolding models to capture kinase conformational states, we curate KinConfBench, a high-quality benchmarking dataset derived from the RCSB PDB [260] as shown in Figure 5.1. To construct a comprehensive benchmarking dataset, we first identify 437 catalytically active human kinase domains and their corresponding UniProt [273] identifiers, following the classification established by Faezov and Dunbrack [274]. These UniProt IDs are then used to query the RCSB Protein Data Bank (PDB) [260] (snapshot 2025-08-01) to retrieve all associated structural entries. By prioritizing this UniProt-centric approach over broader sequence-similarity searches, we ensure the specific retrieval of human kinase domains while minimizing potential contamination from non-human homologs. This initial search yields a raw dataset of 7,763 unique PDB entries, each containing at least one relevant kinase chain.

To assign conformational labels to the selected kinase PDB entries, we process all chains from the identified entries using the KinCoRe software suite [275, 276]. KinCoRe categorizes kinase conformations by extracting and classifying key structural features into different

conformational groups. There are a total of eight categorical labels: a spatial label that captures the orientation of the DFG motif; a dihedral label that captures the torsion angle distribution of the X-DFG motif; a helix label that captures the orientation of the α C-helix; a salt bridge label that captures the orientation of a conserved kinase salt bridge; N-terminus and C-terminus activation loop labels that capture the orientations of activation loops for both terminus; a spine label that captures the orientation of the kinase regulatory spine; and finally, the activity label that assigns whether the kinase structure is in an active or inactive form based on the previous labels. More details on the specific labeling format are provided in Supplementary Information.

Following the labeling process, we apply a stringent filtering protocol to exclude structures with missing atomic coordinates in critical functional regions that have “None” annotations in the KinCoRe output. Specifically, kinase chains are discarded if they have “None” values for either the spatial label or the helix label due to missing residues. This rigorous screening for structural motif completeness ensures that our benchmark is based on high-fidelity structural data, resulting in a refined set of 7,638 kinase chains.

To ensure stringent data quality control, we excluded chains with a resolution worse than 4.5 Å or those determined by NMR spectroscopy, following the protocol established by NeuralPlexer [47]. Furthermore, for genes encoding both catalytically active and pseudokinase domains, such as members of the Janus Kinase (JAK) family, we use the reference sequences of the active kinase domains extracted from UniProt to selectively retain only the catalytic units while excluding pseudokinase domains. This final filtering stage yields a high-quality set of 6,703 kinase chains.

The filtered dataset is partitioned into three primary categories: mutants (1,915 chains), covalent complexes (428 chains), and wild-type (WT) noncovalent complexes (4,564 chains). For this study, we focus exclusively on the unphosphorylated subset of the WT Noncovalent entries comprising 2,900 chains to benchmark co-folding model performances on canonical kinase sequences. To refine this subset, we remove trivial redundancies, defined as entries sharing identical sequences (and ligand SMILES, where applicable) alongside identical KinCoRe conformational labels. Nevertheless, we preserve meaningful structural diversity by retaining any entry that exhibited at least one unique KinCoRe label relative to others of the same condition. This logic is applied both to apo chains of the same gene and to holo complexes featuring the same protein-ligand combination, ensuring that our benchmark accounts for the multi-state nature of kinase domains. By prioritizing these distinct conformational states, we collect the final 2,225 structurally unique kinase chains for cofolding model benchmarking. This entire curation process is illustrated in Figure 5.1(a).

To investigate the general distribution of the selected unique kinase chains, we supply further analysis based on their gene properties and KinCoRe labels. As shown in Figure 5.1(b), the dataset is primarily composed of single-ligand holo complexes ($\sim 80\%$), with apo structures and multi-ligand complexes each accounting for approximately 10%. This composition closely mirrors the broader distribution of kinases in the PDB, reflecting the data landscape encountered by co-folding models during training. Furthermore, while there appears to be a balanced global distribution between active and inactive states across the

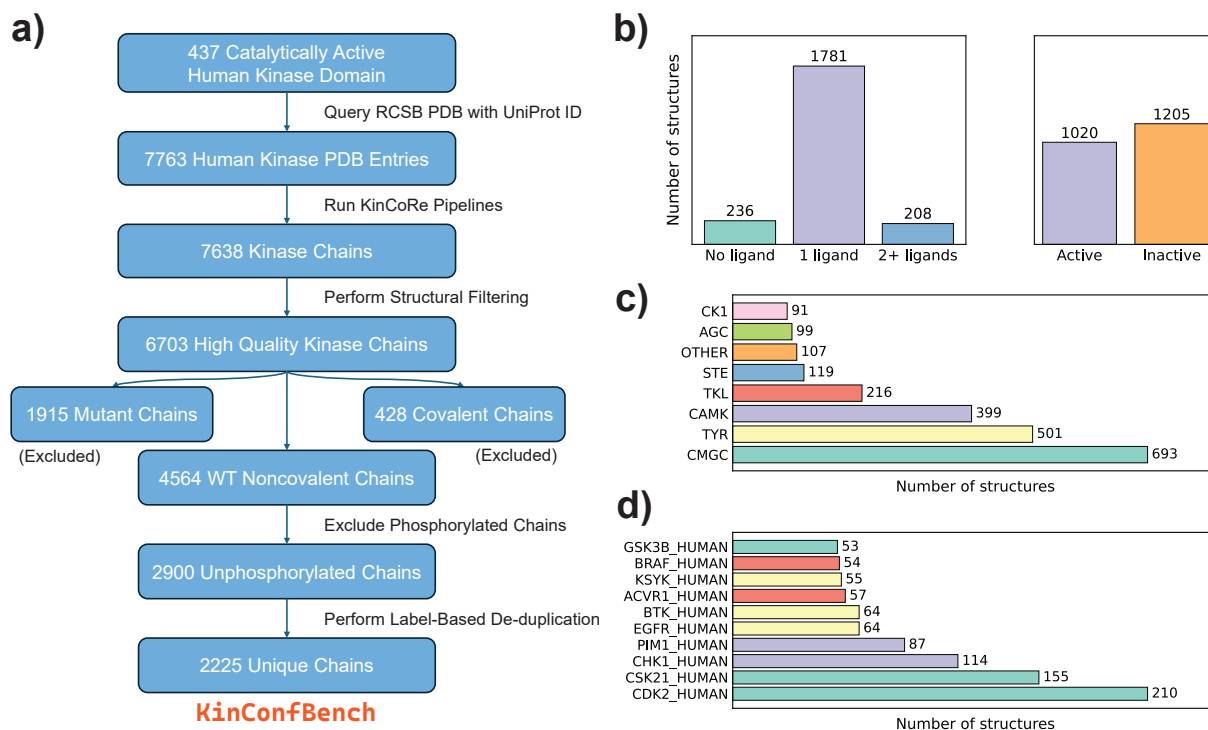


Figure 5.1: **Overview of the KinConfBench curation pipeline and its data distribution.** (a) Schematic of the curation workflow, detailing the sequence of filtering and labeling steps. (b) Distribution of ligand counts per kinase chain and their corresponding conformational activity. (c) Taxonomic distribution of kinase chains categorized by major kinase groups. Details for acronyms can be found in Supplementary Information. (d) Representation of the top 10 most frequent kinase genes in the dataset, colored by their kinase family.

entire dataset, it is worth noting that this ratio can vary quite significantly at an individual gene level.

In terms of the kinase family distribution, we take the Manning classification system [277] and find that CMGC, TYR, CAMK, and TKL groups dominate the dataset, as these families represent areas of intense therapeutic interest[89–91, 278]. Indeed, the ten most frequently occurring genes in our dataset all belong to these four groups (Figure 5.1(c) and (d)). Further distributional metrics, including ligand and KinCoRe label frequencies, are detailed in Supplementary Figure S1. We observe a high prevalence of DFG-in and BLAminus labels, alongside a majority of Type I ligands that aligns with KinCoRe server results as shown in Supplementary Table S1. This observation indicates a persistent bias in the PDB toward holo complexes where the ligands occupy the ATP-binding pocket in catalytically active states, as the DFG-in/BLAminus configuration is a structural prerequisite for kinase activity [93].

After filtering for experimental resolution, sequence completeness, and distinct functional domain labels, the final curated dataset of 2,225 kinase chains is comprised of 236 unique apo chains representing 43 unique gene targets, and 1,989 unique holo chains with 1,377 involving unique protein-ligand pairs, ensuring broad coverage of the human kinome.

Evaluation of Boltz-2, Chai-1, and Protenix against KinConfBench

Protein kinases are highly dynamic molecular machines that transition between multiple active and inactive isoforms to perform their function[270–272]. Molecular effectors such as small drug molecules introduce new interactions that can shift the thermodynamic equilibrium among these states. For instance, Type-II inhibitors specifically recognize and stabilize the inactive “DFG-out” conformation by binding into an adjacent, allosteric hydrophobic pocket [279, 280]. By establishing favorable new bonds within this newly exposed site, the ligand significantly lowers the potential energy minimum of the inactive state relative to the active one [281], and the kinase becomes thermodynamically trapped in this non-functional shape, effectively halting its catalytic activity. Therefore in the following sections we evaluate not only the ligand-centric geometric performance of the cofolding models, but establish the necessity of providing conformational selection metrics to enable drug selectivity for kinase targets.

From geometric filters to functional conformations Valid and conformationally diverse protein folded states is a fundamental prerequisite for the meaningful use of cofolding models. KinCoRe assigns each kinase protein structure to valid spatial and dihedral label combinations (36 theoretical, 12 observed in the PDB; Supplementary Table S1) [93]. By applying the KinCoRe labels to every prediction, we establish a “ground truth” based on the categorical orientation of the DFG motif and the α C-helix, providing a more functionally relevant metric than standard backbone RMSD. Mapping the inference results against the established kinase clusters reveals that all cofolding models generate protein structures falling exclusively within valid conformational categories as shown in Supplementary Table S2, no “forbidden” or physically unreasonable protein geometries appear, confirming that all tested cofolding models possess the baseline biophysical reasoning required to produce structurally plausible kinase conformations.

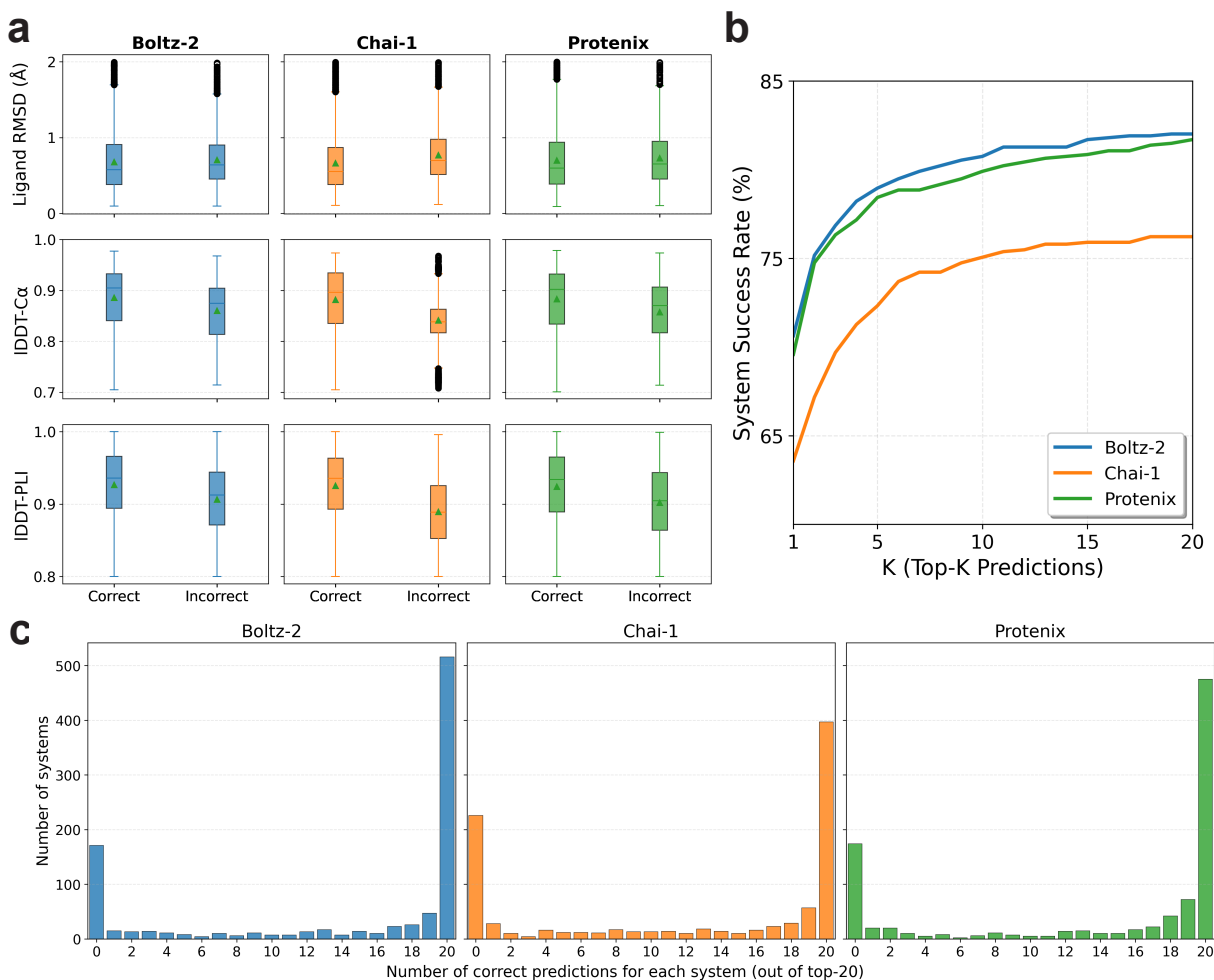


Figure 5.2: Correct and incorrect KinCoRe classification for all three cofolding model predictions passing geometric filters. (a) Cofolding model performance on geometric predictions ($\text{IDDT-C}\alpha \geq 0.7$, $\text{IDDT-PLI} \geq 0.8$, $\text{ligand RMSD} < 2.0 \text{ \AA}$) comparing samples with all eight KinCoRe labels being correct versus incorrect. To account for intrinsic structural flexibility where a single experimental entry may capture multiple conformational states, we consider a prediction successful if it matches any valid set of ground truth labels associated with that protein-ligand complex. (b) Success rate of each cofolding model versus a correct ranked prediction among the top- K by confidence. For each target, we evaluate an ensemble of $N = 20$ generated structures, ranked by their pLDDT confidence scores. (c) Histogram of the count of top-20 predictions that match all KinCoRe labels of their corresponding experimental structures. The x -axis is the number of correct predictions in the ranked ensemble (0-20), and the y -axis is how many of the 950 kinase systems fall in each bin. Panels share a common vertical scale to contrast how frequently each architecture saturates the ensemble with correct states versus how often it fails to identify a single exact match.

Beyond biophysical validity of the protein fold, a cofolding model must place the ligand correctly into the pocket, which we characterize using three structural metrics. Global fold quality, IDDT-C α , summarizes C α agreement of the predicted complex to the experimental structure; higher values indicate a closer global backbone match. Pocket-focused geometry, IDDT-PLI, emphasizes the agreement of the binding site and ligand-adjacent protein atoms compared to experimental values. Finally, ligand heavy-atom RMSD indicates agreement of the modeled ligand pose with the crystal pose. We define a geometrically successful prediction as one with global fold IDDT-C α ≥ 0.7 , binding pocket IDDT-PLI ≥ 0.8 , and ligand RMSD $< 2.0\text{\AA}$ following the definition of previous works [48, 255]. Out of 1420 protein-ligand kinase systems, geometric filtering resulted in only 950 kinase where all cofolding models produce at least one geometrically valid pose out of 20 samples for downstream analysis. The broad geometric distribution of these successful predictions for each cofolding model across kinase systems is detailed in Supplementary Figure S2.

However, a geometrically successful binding pose does not ensure that the cofolding model has captured the correct induced conformational state. Specifically, a prediction is only considered “correct” if all eight KinCoRe labels match those of the experimental structures. Figure 5.2(a) shows the relation of the three geometric metrics (IDDT-C α , IDDT-PLI, and ligand RMSD) to KinCoRe label correctness across all models. While correct conformational assignments skew better across the geometric metrics, the overlap with incorrect functional assignment remains large. The side-by-side comparison among the cofolding models highlights that mislabeled states can still achieve strong ligand-centric structural scores around the binding pocket, indicating that geometry alone is an incomplete proxy for robust cofolding prediction.

Figure 5.2(b) shows each co-folding models’ success versus K , defined as any correct functional state among the top- K by confidence ranking for each cofolding model. Both Boltz-2 and Protenix lead at Top-1 compared to Chai-1, while moving from Top-1 to Top-20 adds roughly 11-13 percentage points for each model (Boltz-2 +11.4; Chai-1 +12.6; Protenix +12.1), indicating that additional samples sometimes recover the correct conformational state even when the highest-confidence pose is wrong (see also Supplementary Table S3). However, Figure 5.2(c) shows that, despite the increase in accuracy, the co-folding model predictions are largely bimodal. This mostly all-right or all-wrong pattern suggests ensemble mode collapse rather than graded uncertainty within the twenty samples.

Supplementary Table S5.4 contains the 110 kinases where all three cofolding models fail to generate the correct conformation, providing an important benchmark for future cofolding model improvements. Figure 5.3 highlights MAP4K1/YK1 (PDB ID: 7M0M) [282] as a representative hard case where all cofolding model predictions fail to obtain an induced-fit to the correct conformational state. This observation shows that even when pocket metrics and ligand poses look favorable, Boltz-2, Chai-1, and Protenix often jointly settle into an alternate spatial and dihedral configuration that disagrees with the experimentally observed loop excursion, underscoring why ligand RMSD alone can mask protein-state errors.

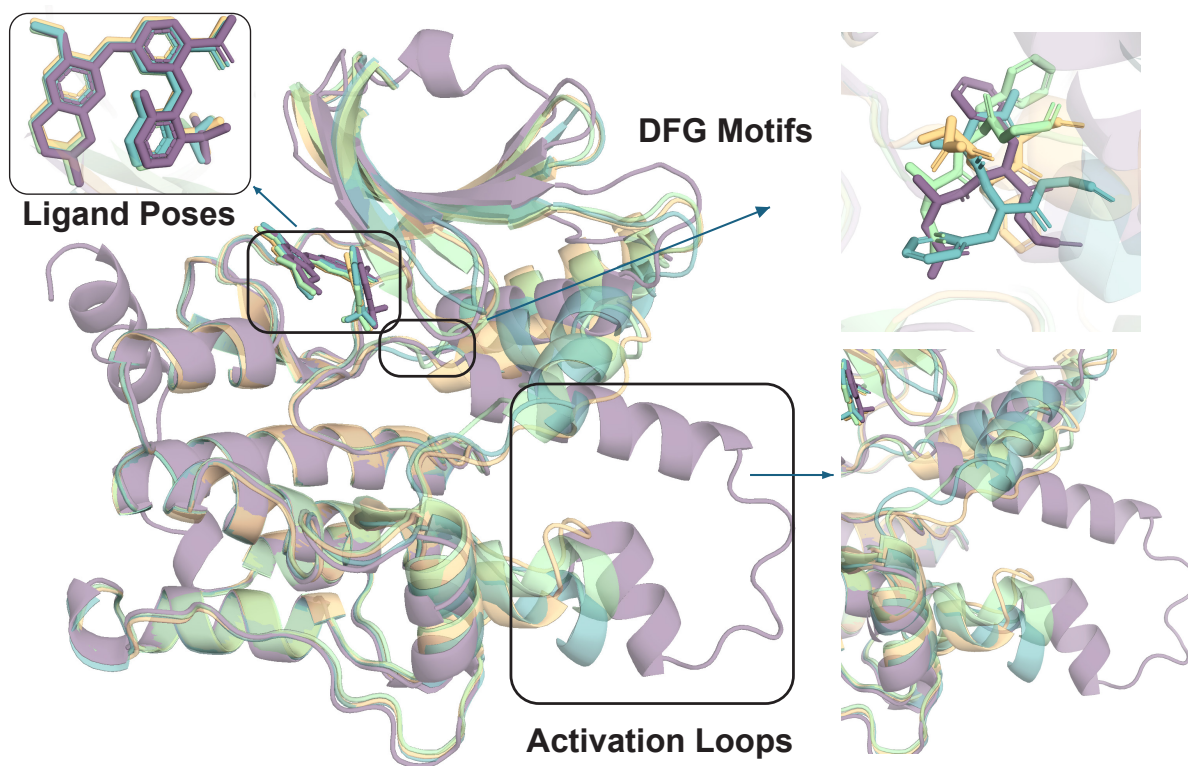


Figure 5.3: **Induced-fit failure modes for MAP4K1 (7M0M).** Superposition of the experimental Type-I diaminopyrimidine carboxamide complex (purple) with representative Boltz-2 (green), Chai-1 (orange), and Protenix (blue) predictions evaluated in KinConf-Bench. The experimental structure adopts DFG-in, BLAminus packing with an inward α C-helix and an extended, outward-biased activation loop. (**Top right**) DFG region; (**bottom right**) activation loop. Although ligand placements are perfect across models, the three evaluated cofolding predictions miss the cooperative loop sampled in the PDB, collapsing instead toward structure that is conformationally mislabeled relative to 7M0M.

Ensemble Diversity and Resolution of Complex Targets KinCoRe labels establish whether a model satisfies the mechanistic requirements of a kinase state such as satisfying the DFG and α C-helix motifs, but categorical success does not capture the nuances of “intra-basin exploration.” Hence, we extend our analysis beyond discrete accuracy to measure the structural dispersion within ensembles that are already deemed functionally correct. To quantify conformational diversity we use some of the same geometric descriptors that define the KinCoRe labels as well as some additional metrics, but on a continuous scale for distances and dihedral angles as described in Supplementary Information. For this benchmark we focus on the 509 systems where Boltz-2, Chai-1, and Protenix each have at least ten all-correct structures according to KinCoRe in the top-20, so every cofolding method contributes a comparably populated success set based on classification; Supplementary Figure

S3 summarizes how these targets distribute across the Top-K benchmark.

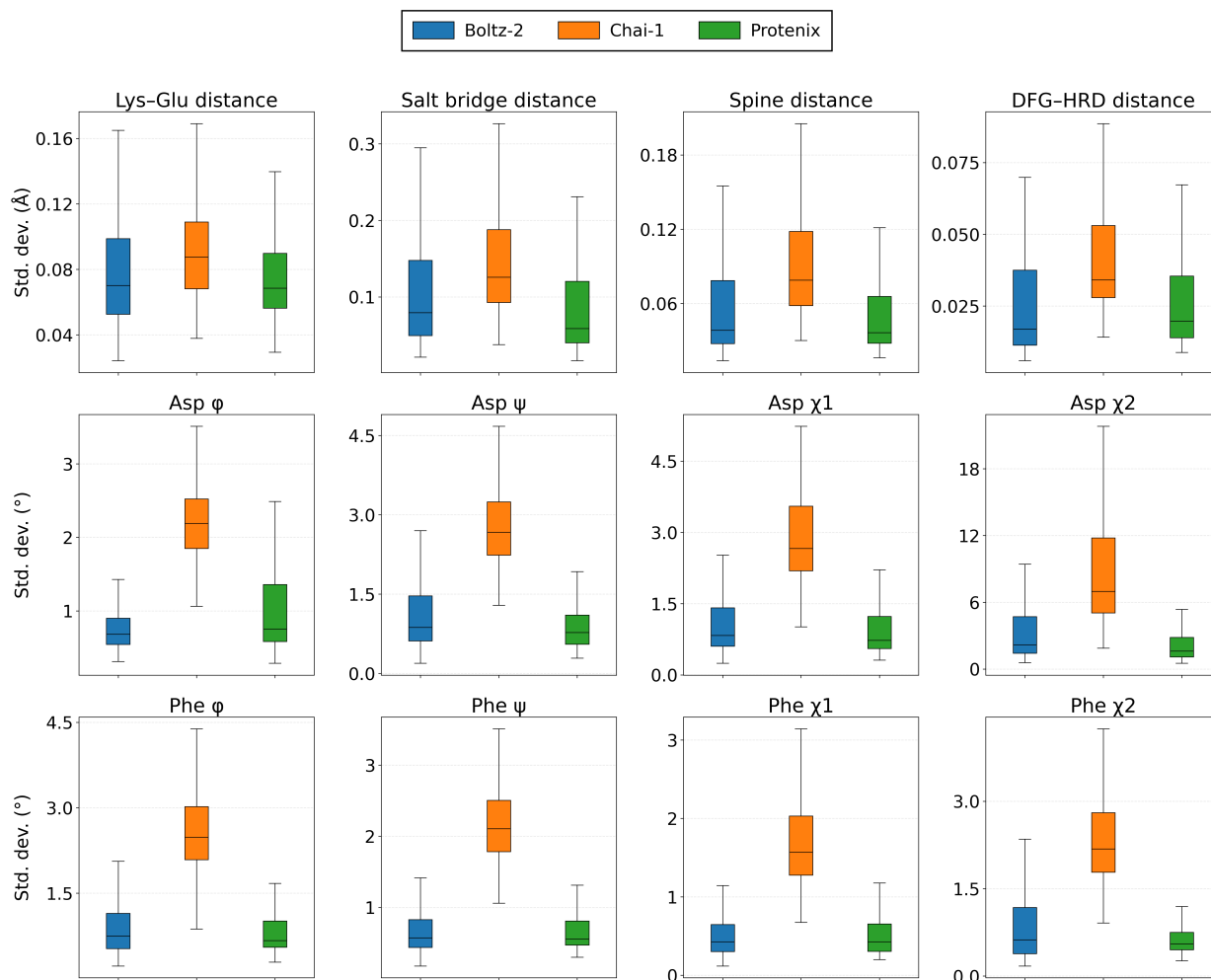


Figure 5.4: **Distance and torsion angle ensemble diversity among correct KinCoRe predictions across cofolding models.** For each system in the diversity subset from KinConBench, we retain only cofolding predictions that are all-label correct, then measure the distance and dihedral angle standard deviations across those correct poses. Overall, all cofolding models have little structural diversity. Panels compare Boltz-2 (blue), Chai-1 (orange), and Protenix (green).

Figure 5.4 shows the standard deviation in distance and torsion angle metrics around their average values across the 20 member ensembles for all three cofolding models. While Supplementary Figure S4 finds that all cofolding models capture the correct average distance and angle values compared to experiment, all models produce very small structural deviations of ~ 0.1 Å for the distance metrics and less than 5° average angle exploration except for Chai-

1’s performance on Asp χ_2 . To place these results into context, root mean square fluctuations of ~ 1 -2 Å and backbone or side chain dihedral angle ranges of 5° to 30° define the statistical fluctuations of stable protein conformations [252, 283], while significantly higher values would be needed to indicate transitions between distinct conformational states. Hence while the Chai-1 model shows greater explorations within the correct structural basin compared to Boltz-2 and Protenix, on an absolute scale the conformational diversity is still quite limited, in line with previous ligand-centric analysis of cofolding models[48].

Benchmarking Apo-to-Holo Conformational Shift In structure-based drug design, ligand binding represents a thermodynamic perturbation of the protein’s native conformational landscape. We conceptualize the apo (ligand-free) distribution as a cofolding model’s learned default energy landscape, whereas the introduced ligand acts as a physical modulator that distorts this baseline to stabilize a distinct holo conformation. If a cofolding model learns the underlying physics of molecular recognition, it successfully predicts this specific ligand-induced apo-to-holo conformational shift. Conversely, if a cofolding model relies heavily on memorization it will fail to generalize to an apo-to-holo conformational transition for novel ligand effectors, effectively “drifting” back to the unperturbed, default apo state.

To test this hypothesis, we perform a comparative analysis against the ground-truth holo and apo state pairs. We extract kinase systems from KinConfBench that possess resolved structures for both states, restricting our analysis to systems where the ligand induces a distinct structural change defined as a different set of KinCoRe functional labels compared to ligand-free structures. For each cofolding model, we define training and test sets based on a time split relevant to each model: a holo ligand combination counts as part of the test set if it first appears in the PDB only after that method’s training-data deposition cutoff: Boltz-2 (2023-06-01); Chai-1 (2021-01-12); Protenix (2021-09-30)). All apo-state structures are found in the training data for all cofolding models.

For each of the kinase system pairs, we generate an ensemble of 20 structures using the holo sequence-ligand as input for each cofolding model, and we classify their predictions according to:

- **Holo State:** At least one structure in the generated ensemble exactly matches the KinCoRe labels of the ground-truth holo reference and 0 matches the KinCoRe labels of the ground-truth apo reference. This metric quantifies the model’s ability to correctly predict the ligand-induced state.
- **Apo Drift:** At least one structure in the generated ensemble exactly matches the KinCoRe labels of the ground-truth apo reference and 0 matches the KinCoRe labels of the ground-truth holo reference. This metric quantifies the tendency of the model to revert to the memorized apo state despite the presence of the ligand.
- **Dual Match:** At least one structure of the ensemble matches the KinCoRe labels of the ground-truth holo reference, and another matches the ground-truth apo reference. This metric quantifies the possibility that the holo-state could be recovered with more sampling or an effective ranking, but otherwise is also evidence of apo-drift.

- **None:** No structure in the generated ensemble exactly matches the KinCoRe labels of either the ground-truth holo reference or the ground-truth apo reference.

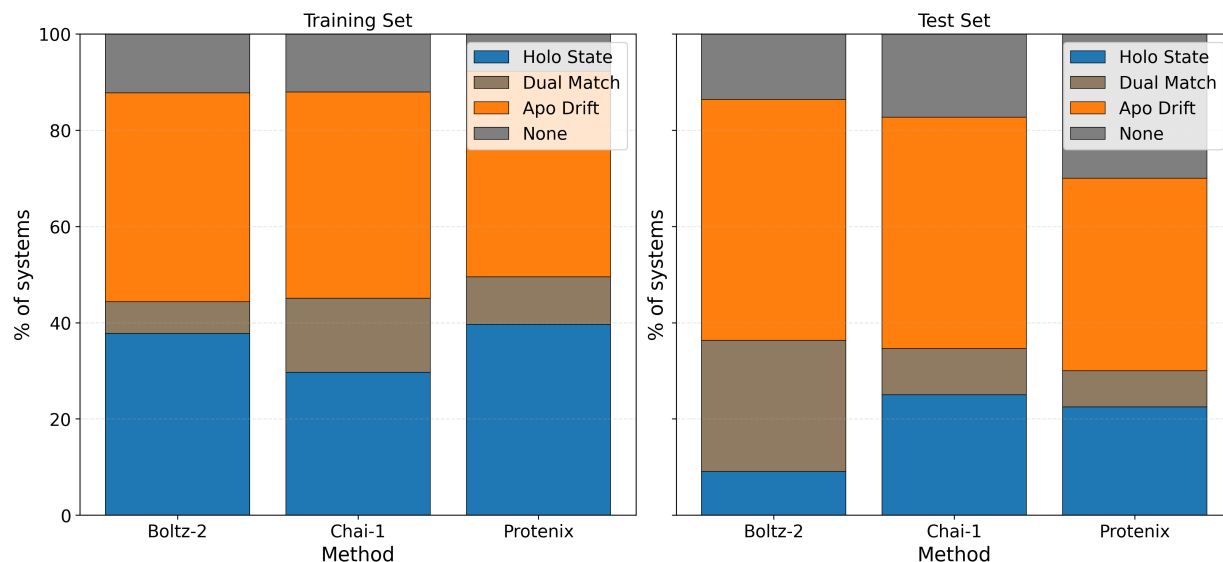


Figure 5.5: **Generalization and apo-drift across co-folding models.** Stacked fraction of systems in each holo/apo outcome category on holo→apo inference pairs for Boltz-2, Chai-1, and Protenix. Sample sizes per split (Train / Test systems): Boltz-2 212 / 22; Chai-1 182 / 52; Protenix 194 / 40. Categories (bottom to top in stack): Holo only; Dual Match; Apo only; None.

A robustly generalizing cofolding model should keep a large Holo-State fraction on both the training and testing data, whereas memorization shifts toward Apo-only or dual-match states predictions. As seen in Figure 5.5a, within the training set, the Holo-State prediction represents roughly 40% of the predictions for Boltz-2 and Protenix and around 30% for Chai-1, but all three models display nearly 40% Apo Drift within the training set. Moreover, when comparing to the test set distribution in Figure 5.5b against the training set, the stacked profiles further shift toward Apo-only and Dual-Match segments relative to the training systems—most clearly for Boltz-2 and Protenix—consistent with apo-drift when confronted with new holo states unsupported in the training statistics.

5.4 Discussion and Conclusion

Our benchmarking of three popular cofolding models highlights a tension between confident geometric poses and biophysically faithful kinase conformational states. Each architecture attains strong pocket-centric metrics on large portions of KinConfBench, yet functional KinCoRe agreement lags behind what raw IDDT or RMSD might suggest. Framing success

purely through the highest-confidence structure therefore obscures a clinically relevant failure mode: models can “look” bound while adopting the wrong regulatory layout for the ligand at hand.

Ensemble behavior compounds the issue for all cofolding models. When additional samples are drawn, Top-20 accuracy gains indicate that correct kinase conformational states sometimes appear deeper in the ranking, but per-target histograms remain dominated by all-or-nothing outcomes rather than graded uncertainty. None of the three cofolding models eliminates the need for external checks when structure-activity-relationships involving induced fit critical motions are required, as in MAP4K1/7M0M, where several models conspire on an alternate DFG-in basin. This likely arises because even when a cofolding model correctly predicts the KinCoRe classification, structural variety is seen to be well-below that expected for thermalized statistical fluctuations needed to create induced-fit configurations.

The holo/apo paired analysis sharpens this diagnosis. Memorization of apo kinase states is shared across Boltz-2, Chai-1, and Protenix rather than unique to one model. After the training cutoff (specific to each model), we find that the holo state prediction rates fall and apo drift rises for each architecture. Closing this gap likely demands both better generative breadth for rare induced-fit basins and confidence objectives that reward state discrimination, not only ligand-protein pocket aligned metrics.

At its core, drug discovery requires conformational state correctness rather than solely fold correctness or memorization of training instances. Predicting the general architecture of a kinase is biologically insufficient if the specific, ligand-induced functional conformation is wrong. KinConfBench is meant to keep that standard explicit as cofolding models continue to mature. Looking forward, the scope of kinase benchmarking must expand to encompass the intricate regulatory mechanisms that define this protein family in their real biological environment. Our current work focuses on the interaction between wild-type domains and non-covalent inhibitors, yet kinase signaling is tightly modulated by post-translational modifications (PTMs) and somatic mutations. Future benchmarks should rigorously assess whether cofolding models can predict how phosphorylation at specific regulatory sites or drug-resistant mutations alter the conformational ensemble. Furthermore, the rise of covalent inhibitors in oncology necessitates a focused evaluation of how cofolding models handle the shift in protein context when a covalent bond is formed, as covalent ligands typically bring more unforeseen perturbations than simple non-covalent docking.

Finally, the ultimate test of a cofolding model’s biophysical reasoning will be its ability to simulate competitive dynamics. In a physiological environment, kinases are rarely isolated with just a single ligand. For Type-I kinase inhibitors, they constantly exist in a competitive environment where inhibitors must displace ATP or compete with other endogenous substrates. Future generations of benchmarks should evaluate the ability of models to handle multi-ligand scenarios, probing whether the presence of competing binders drives the protein into distinct conformational equilibria. Moving beyond static snapshots to dynamic, competitive scenarios will be essential for evolving these tools from structural prediction engines into true computational assays for drug discovery.

5.5 Supplementary Information

Additional Methodology Details

Kinase Group Classifications. To categorize the structural and functional diversity of the kinome, we utilize the standard Manning classification system [277]. The primary groups are defined as follows.

- **AGC:** Proteins including PKA, PKG, and PKC families.
- **CAMK:** Calcium/calmodulin-dependent protein kinases.
- **CK1:** Casein kinase 1 and closely related isoforms.
- **CMGC:** Proteins including CDK, MAPK, GSK3, and CLK families.
- **STE:** Homologs of yeast Sterile 7, Sterile 11, and Sterile 20 kinases.
- **TYR:** Tyrosine kinases.
- **TKL:** Tyrosine kinase-like proteins that are serine-threonine kinases.
- **OTHER:** Other kinases that don't fall into above classification.

Details of KinCoRe Annotations. Introduced by Modi and Dunbrack[275], a kinase conformational label has the following eight annotations:

- **Spatial Label:** Describes the position and orientation of the DFG motif as DFGin, DFGout, DFGinter, or None when key residues are not correctly mapped.
- **Dihedral Label:** Captures the backbone and side chain dihedral angles of the X-DFG motif by using the Ramachandran region annotation (A, B, L, and E) for the X, D, and F residues and the DFG-Phe χ_1 rotamer (minus = -60° , plus = $+60^\circ$, and trans = 180°). These labels include BLAminus, BLBplus, ABAminus, BBAminus, BLBtrans, BLBminus, BLAplus, BABtrans, or None when key residues are not correctly mapped.
- **α C-helix Label:** Indicates the placement of the α C-helix as either Chelix-in, Chelix-out, or None when key residues are not correctly mapped.
- **Salt Bridge Label:** Denotes the formation of the conserved salt bridge as SaltBr-in, SaltBr-out, or None when key residues are not correctly mapped.
- **N-term. Act. Loop Label:** Classifies the N-terminal region of the activation loop as ActLoopNT-in, ActLoopNT-out, or None when key residues are not correctly mapped.
- **C-term. Act. Loop Label:** Classifies the C-terminal region of the activation loop as ActLoopCT-in, ActLoopCT-out, or None when key residues are not correctly mapped.

- **Spine Label:** Indicates the assembly state of the regulatory spine as Spine-in or Spine-out, or None when key residues are not correctly mapped.
- **Ligand Type:** Categorizes the bound ligand into Type 1, Type 1.5_Front, Type 1.5_Back, Type 2, Type 3, Allosteric, and No_ligand.

Details of Metrics for Diversity Analysis.

- **lys_glu_distance:** Tracks the separation between the conserved β 3-Lysine and α C-Glutamate, critical for evaluating α C-helix packing.
- **saltbridge_distance:** Monitors the formation and structural stability of the canonical salt bridge network.
- **spine_distance:** Measures the spatial arrangement of the hydrophobic residues comprising the regulatory spine, indicating its assembly state.
- **dfg_hrd_distance:** Evaluates N-terminal activation-loop coupling by measuring the spatial separation between the DFG and catalytic HRD motifs.
- **asp_phi:** Captures the ϕ backbone dihedral angle of the DFG Aspartate residue.
- **asp_psi:** Captures the ψ backbone dihedral angle of the DFG Aspartate residue.
- **asp_chi1:** Describes the primary side-chain torsion (χ_1) of the DFG Aspartate.
- **asp_chi2:** Describes the secondary side-chain torsion (χ_2) of the DFG Aspartate. This is treated with twofold (“ π ”) symmetry so that flipped carboxylate placements are not penalized as opposite phases on the circle.
- **phe_phi:** Captures the ϕ backbone dihedral angle of the DFG Phenylalanine residue.
- **phe_psi:** Captures the ψ backbone dihedral angle of the DFG Phenylalanine residue.
- **phe_chi1:** Describes the primary side-chain torsion (χ_1) of the DFG Phenylalanine.
- **phe_chi2:** Describes the secondary side-chain torsion (χ_2) of the DFG Phenylalanine. Like the Aspartate, this is treated with twofold (“ π ”) symmetry so that flipped aromatic substituent placements are not counted as opposite phases.

Supplementary Tables

Table S5.1: **Percentage of KinCoRe spatial and dihedral label combinations for human kinases.** The definitions of Spatial and Dihedral Labels are defined in Chapter 5.5. From the KinCoRe server PDB inventory (<https://dunbrack.fccc.edu/kincore/home>) and for KinConfBench (KCB) reference chains which focuses on a subset of high quality samples and filters out None labels during selection.

Spatial Label	Dihedral Label	All PDB	KinConfBench
DFGin	BLAminus	6806	1430
DFGin	BLAplus	316	27
DFGin	ABAminus	993	183
DFGin	BLBminus	433	96
DFGin	BLBplus	1008	238
DFGin	BLBtrans	255	102
DFGinter	BABtrans	26	5
DFGout	BBAminus	677	144
DFGin	None	836	0
DFGinter	None	224	0
DFGout	None	365	0
None	None	413	0

Table S5.2: **Distribution of KinCoRe spatial and dihedral labels for KinConfBench (KCB) and cofolding model predictions.** KCB percentages are derived from the 2,225-chain set described in the Results. Percentages for Boltz-2 ($n = 28,400$), Chai-1 ($n = 28,399$), and Protenix ($n = 28,400$) are calculated across 1,420 distinct kinase systems. * The small percentage of “None” labels results from residue mismatches during the KinCoRe labeling process.

Spatial	Dihedral	KCB (%)	Boltz-2 (%)	Chai-1 (%)	Protenix (%)
DFGin	BLAminus	64.3	65.8	77.5	65.1
DFGin	BLAplus	1.2	1.1	1.2	1.5
DFGin	ABAminus	8.2	3.9	4.8	3.3
DFGin	BLBminus	4.3	3.0	2.0	3.4
DFGin	BLBplus	10.7	8.0	8.5	8.3
DFGin	BLBtrans	4.6	11.6	0.1	11.7
DFGinter	BABtrans	0.2	0.0	0.0	0.0
DFGout	BBAminus	6.5	5.3	4.3	5.4
DFGin	None*	0.0	0.4	0.4	0.2
DFGinter	None*	0.0	0.8	0.6	0.1
DFGout	None*	0.0	0.1	0.5	0.1
None*	None*	0.0	0.0	0.1	1.0

Table S5.3: **Conformational (KinCoRe label) summary on 950 systems passing geometric filters.** Systems ≥ 1 correct: systems with at least one all-labels-correct prediction in top-20. Avg correct/system: mean KinCoRe-correct predictions per system (out of top-20).

Model	Systems ≥ 1 correct	Avg correct/system
Boltz-2	779 (82.0%)	14.2
Chai-1	724 (76.2%)	12.4
Protenix	776 (81.7%)	14.1

Table S5.4: 110 systems where all cofolding predictions fail to generate the correct conformation as the ground truth. Each entry is a KinConfBench system key (GENE.Name.Ligands).

System identifiers				
ABL1_HUMAN_66K	ABL1_HUMAN_STI_1N1	ACK1_HUMAN_1G0	ACK1_HUMAN_DBQ	ACK1_HUMAN_LWX
ACK1_HUMAN_R7P	ACK1_HUMAN_T74	ACK1_HUMAN_T95	ACK1_HUMAN_WTP	ALK_HUMAN_25J
ALK_HUMAN_6YL	ALK_HUMAN_AWJ	ALK_HUMAN_HKJ	ALK_HUMAN_J3Y	AURKA_HUMAN_9YQ
AURKA_HUMAN_X6D	BRAF_HUMAN_5I4	BRAF_HUMAN_FP3	BRAF_HUMAN_P06	BTB_HUMAN_YDA
CDK1_HUMAN_1QK	CDK2_HUMAN_02Z	CDK2_HUMAN_06Z	CDK2_HUMAN_09Z	CDK2_HUMAN_0S0
CDK2_HUMAN_106	CDK2_HUMAN_20K	CDK2_HUMAN_26Z	CDK2_HUMAN_A1A1H	CDK2_HUMAN_A1D6S
CDK2_HUMAN_A27	CDK2_HUMAN_DTQ	CDK2_HUMAN_ES4	CDK2_HUMAN_JWS	CDK2_HUMAN_LS1
CDK2_HUMAN_LS3	CDK2_HUMAN_LS4	CDK2_HUMAN_RON	CDK2_HUMAN_SU9	CDK2_HUMAN_WQ6
CDK2_HUMAN_X02	CDK2_HUMAN_X06	CDK2_HUMAN_X19	CDK2_HUMAN_X35	CDK2_HUMAN_X36
CDK2_HUMAN_X3A	CDK2_HUMAN_X40	CDK2_HUMAN_X42	CDK2_HUMAN_X43	CDK2_HUMAN_X44
CDK2_HUMAN_X62	CDK2_HUMAN_X6B	CDK2_HUMAN_Y8L	CDK2_HUMAN_Z19	CDK2_HUMAN_Z63
CDK2_HUMAN_Z71	CDK5_HUMAN_65L	CDK6_HUMAN_24V	CDK6_HUMAN_AP9	CDK6_HUMAN_LQQ
CDK7_HUMAN_I73	CDK7_HUMAN_WZ8	CDK8_HUMAN_C1I	CHK1_HUMAN_306	CHK1_HUMAN_373
CHK1_HUMAN_76A	CHK1_HUMAN_H0K	CHK1_HUMAN_YM6	CLK1_HUMAN_Q7K	CLK1_HUMAN_WAZ
DAPK1_HUMAN_BD4	DAPK1_HUMAN_LU2	DAPK1_HUMAN_PIT	DAPK1_HUMAN_STU	DAPK3_HUMAN_4RB
DCLK1_HUMAN_XBD	EPHA2_HUMAN_DXX	EPHA2_HUMAN_L66	EPHA2_HUMAN_QRD	EPHA2_HUMAN_QRR
EPHA2_HUMAN_WT3	IGF1R_HUMAN_PDR	JAK3_HUMAN_79T	KC1D_HUMAN_AUE	KCC2D_HUMAN_K88
KSYK_HUMAN_1B6	KSYK_HUMAN_4MG	KSYK_HUMAN_685	KSYK_HUMAN_X7G	M3K5_HUMAN_NJV
M3K5_HUMAN_STU	M4K1_HUMAN_2WI	M4K1_HUMAN_A1AP0	M4K1_HUMAN_YK1	MK01_HUMAN_2SH
MK01_HUMAN_33A	MK01_HUMAN_35X	MK01_HUMAN_362	MK01_HUMAN_5ID	MK01_HUMAN_F29
MK01_HUMAN_FRZ	MK07_HUMAN_R4L	MK14_HUMAN_GK1	SRPK1_HUMAN_RXZ	VRK2_HUMAN_7DZ
VRK2_HUMAN_KJD	WEE1_HUMAN_34W	WEE1_HUMAN_99J	WEE1_HUMAN_99M	WEE1_HUMAN_99V

Supplementary Figures

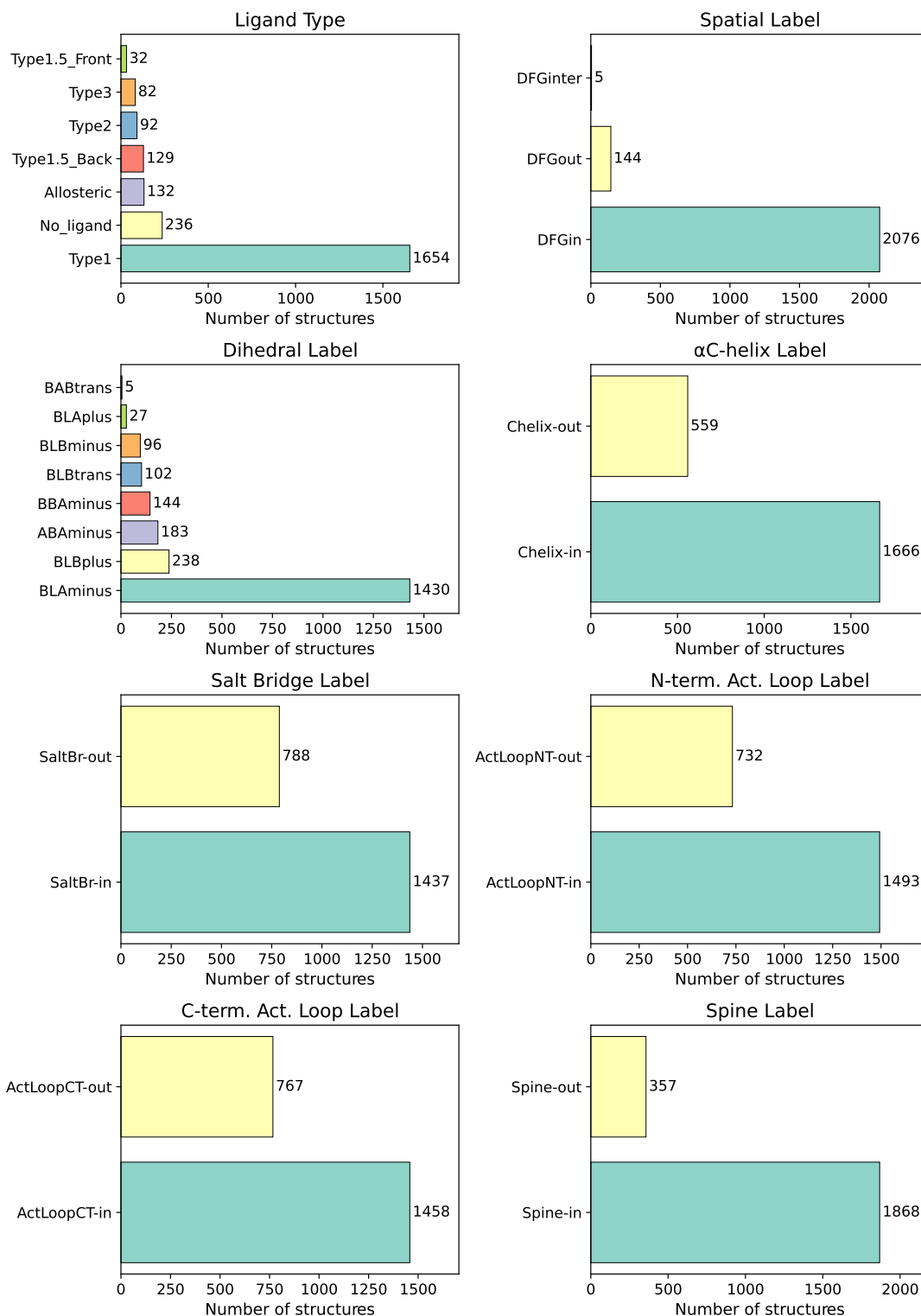


Figure S5.6: **KinConfBench curation and composition.** Summary of how the benchmark is built from PDB-derived kinase chains: counts by holo versus apo, single- versus multi-ligand complexes, redundancy filtering, and the distribution of entries across Manning kinase groups and ligand/chemotype categories. Together these panels document that the final benchmark is holo-heavy (reflecting the PDB) yet retains diverse genes and chemotypes, and that reported model statistics are computed on the curated non-redundant set rather than the raw PDB pull.

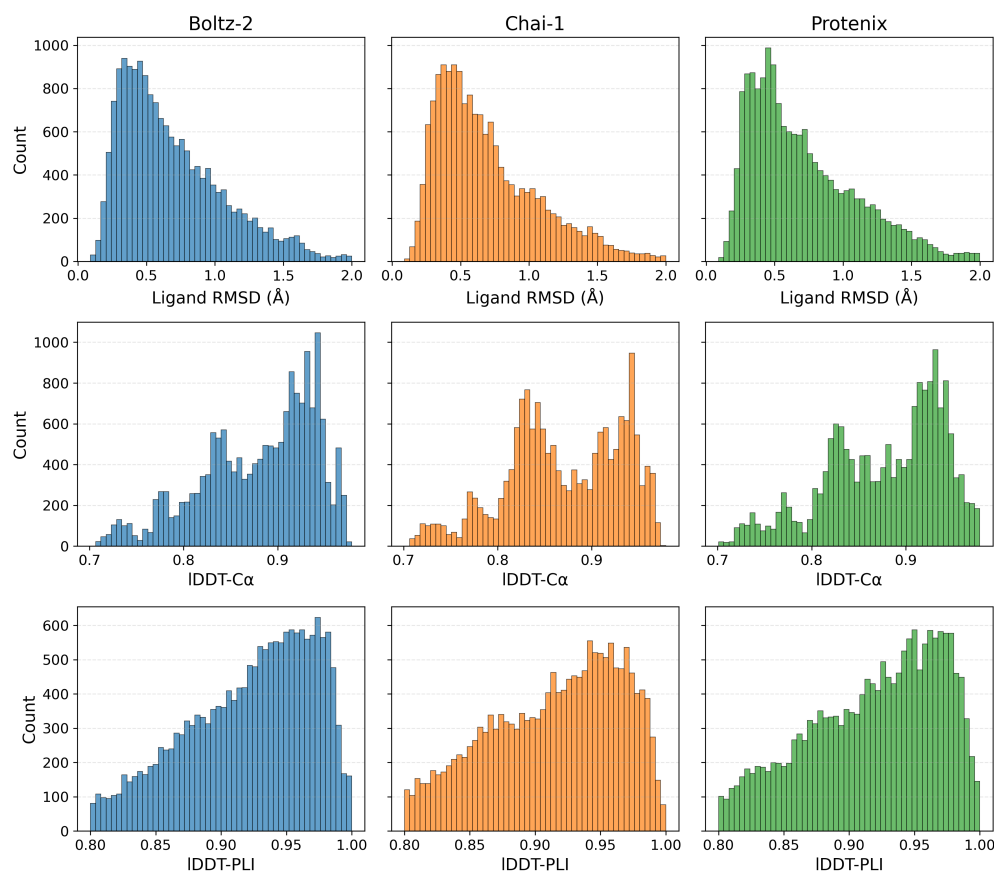


Figure S5.7: **Distribution of geometric metrics for cofolding predictions.** Distributions of ligand heavy-atom RMSD, IDDT-C α , and IDDT-PLI for ensemble predictions that pass the benchmark geometric filters (IDDT-C α $\geq$ 0.7, IDDT-PLI $\geq$ 0.8, ligand RMSD $<$ 2 Å). We have 950 protein-ligand kinase systems from Boltz-2 (Blue, 17,409 samples), Chai-1 (Orange, 16,471 samples), and Protenix (Green, 17,125 samples).

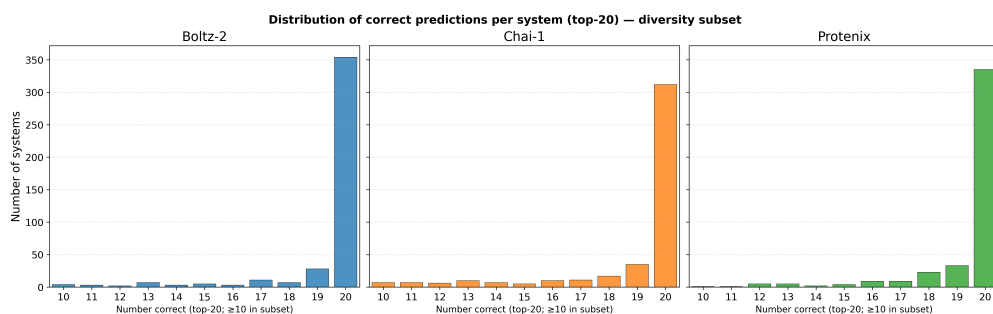


Figure S5.8: **Correct-prediction counts on the high-yield diversity subset.** Same quantity as in Figure 2(c), but restricted to 509 systems where Boltz-2, Chai-1, and Protenix each have at least ten all-labels-correct predictions in the top-20, the subset used for ensemble diversity analysis. The horizontal range is truncated to 10-20 so that differences in the upper tail (near-enumeration of the correct state) are visible. A single-row layout with a shared y -axis facilitates direct comparison of how densely each model covers the correct label manifold when it is readily accessible.

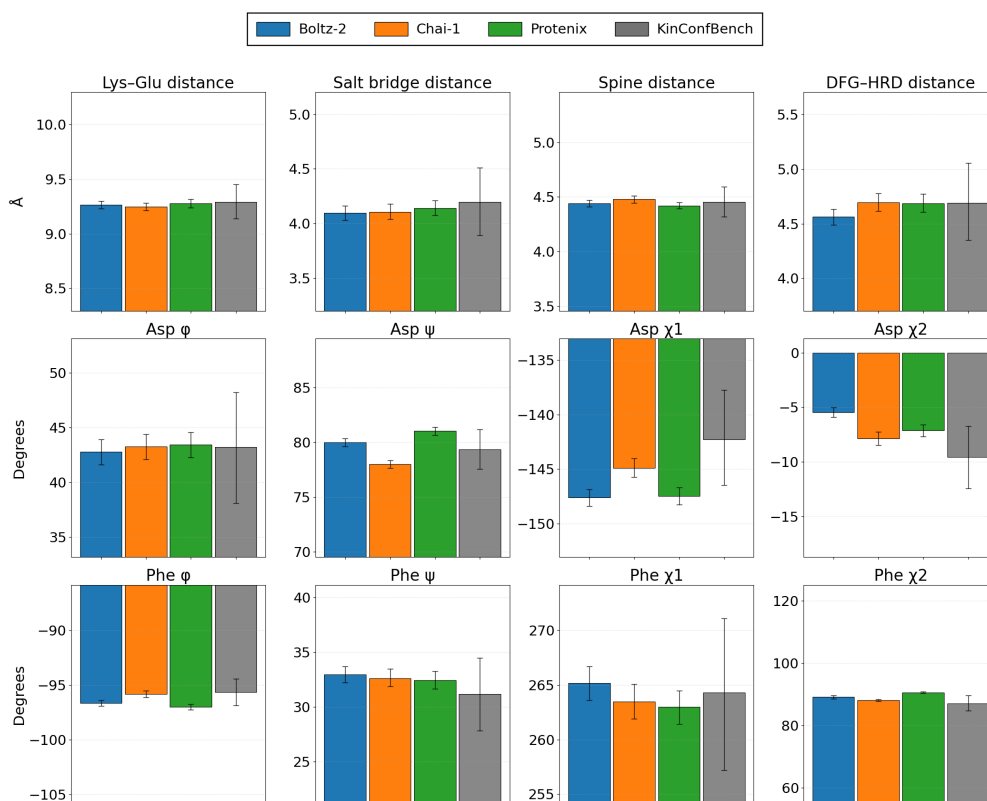


Figure S5.9: **Mean of key distance and angle metrics for cofolding predictions and KinConfBench.** The plot displays mean values for Boltz-2 (blue), Chai-1 (orange), and Protenix (green) predictions, with KinConfBench PDB reference values provided in gray for comparison. Error bars are bootstrap 95% confidence intervals for each pooled mean.

Chapter 6

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

‡Authors contributed equally. Kunyang Sun performed the fragment extraction and built the fragment-based Custom Score.

6.1 Introduction

The COVID-19 pandemic spurred an unprecedented global research effort aimed at discovering antiviral therapeutics. [284–287] Among the most intensively studied coronavirus targets is the main protease (Mpro), [288–291] 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[288] that stimulates multiple drug discovery projects, [292–306] with some eventually producing clinically approved inhibitors such as Nirmatrelvir (Paxlovid) [292] and Ensitrelvir. [293]

To accelerate antiviral discovery for SARS-CoV-2 Mpro, [102, 289, 307–310] a number of open-science initiatives[102, 286] produced rich experimental datasets. The COVID Moonshot consortium, [102, 286] for example, reported more than two thousand biochemical measurements and over eight hundred crystal structures for inhibitor complexes. Numerous Antiviral Drug Discovery (AViDD) centers[98] 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[13] organized through the Polaris Hub[99] in collaboration with ASAP Discovery[100] and OpenADMET[101]. Specifically, in the ligand pose prediction track, participants were provided

¹Reproduced 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, 21, 12058–12067

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 2Å heavy-atom RMSD) across all test ligands. [13]

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. [311] 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. [73, 74, 207, 312, 313] In particular, when a large amount of existing crystal structures are available, ligand-template-based docking [74, 314–316] can generate accurate results without the use of sophisticated algorithms. In parallel, thanks to the curation of high quality protein ligand structure datasets, [9, 41, 72, 227] recent developments in machine learning docking methods, [317] particularly protein–ligand co-folding models, [318–320] 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 [15] 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)[321] 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.

6.2 Methods

As depicted in Figure 6.1, 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) [322] conformers with clustered crystal complexes from either the organizers or the RCSB PDB. [229] As part of the standard drug discovery practice, [323] all ligands were then docked to these ensembles to produce a pool of candidate poses.

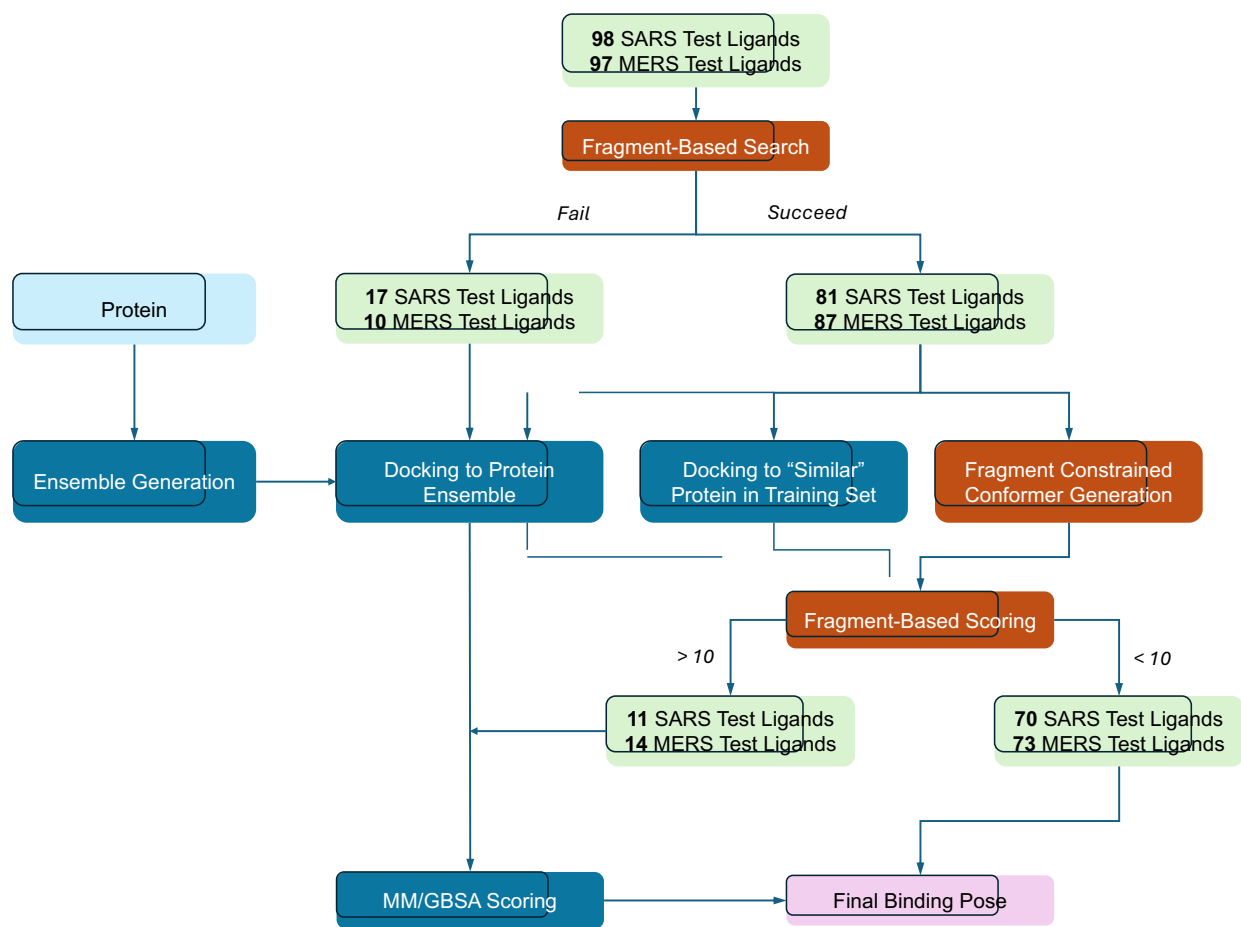


Figure 6.1: *Schematic representation of our multi-step workflow to generate the final ligand binding poses.* The fragment based search utilizes conserved chemical fragments of training ligands to help inform a set of structural and scoring priors for pose of the test ligand. Lack of train-test fragment 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.

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, [13, 306] 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, [322] 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.[322]

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. [324–327] 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. [328] 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 C α 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.

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.

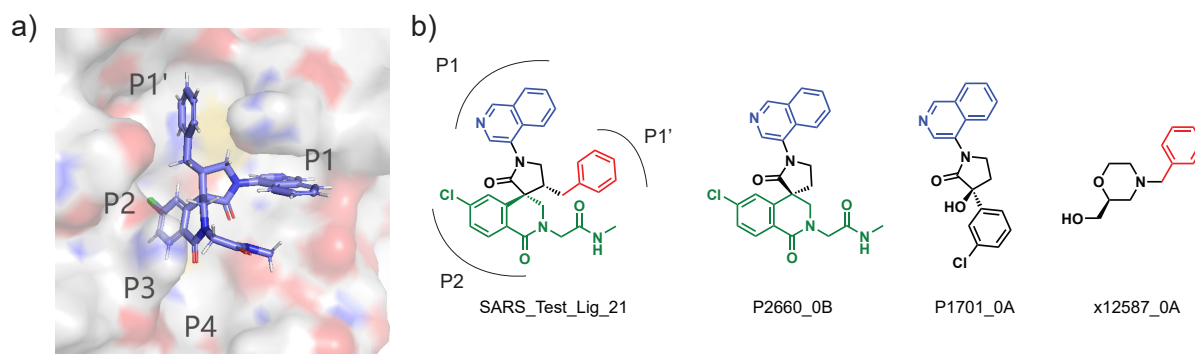


Figure 6.2: *Example of scaffold-aware ligand fragmentation.* (a) SARS_Test_Lig_21 crystal structure with binding pocket labels. (b) 3 example ligands with identical colored fragments as the test ligand SARS_Test_Lig_21 from the provided crystal structures.

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,[188] 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 6.2, 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, [327] 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 [188] 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.

Custom Fragment-Based Scoring. Traditional docking scores are largely empirical, parameterized to perform reasonably well across a wide range of targets and ligand chemotypes.[311] 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

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

where $\text{RMSD}_{\text{frag}}$ 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.

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. [329] Then, the GBSA score was calculated using `MMPBSA.py`[330] program in AmberTools with GB-OBC[331] implicit solvation model ($igb = 2$). Amber ff14SB force field[166] was used for protein and GAFF2[332] along with AM1-BCC charges[333, 334] was used for the ligand. The constrained geometry optimization was performed using OpenMM[10] 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.

6.3 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.

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 6.1), 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 6.3 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. [13] 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.

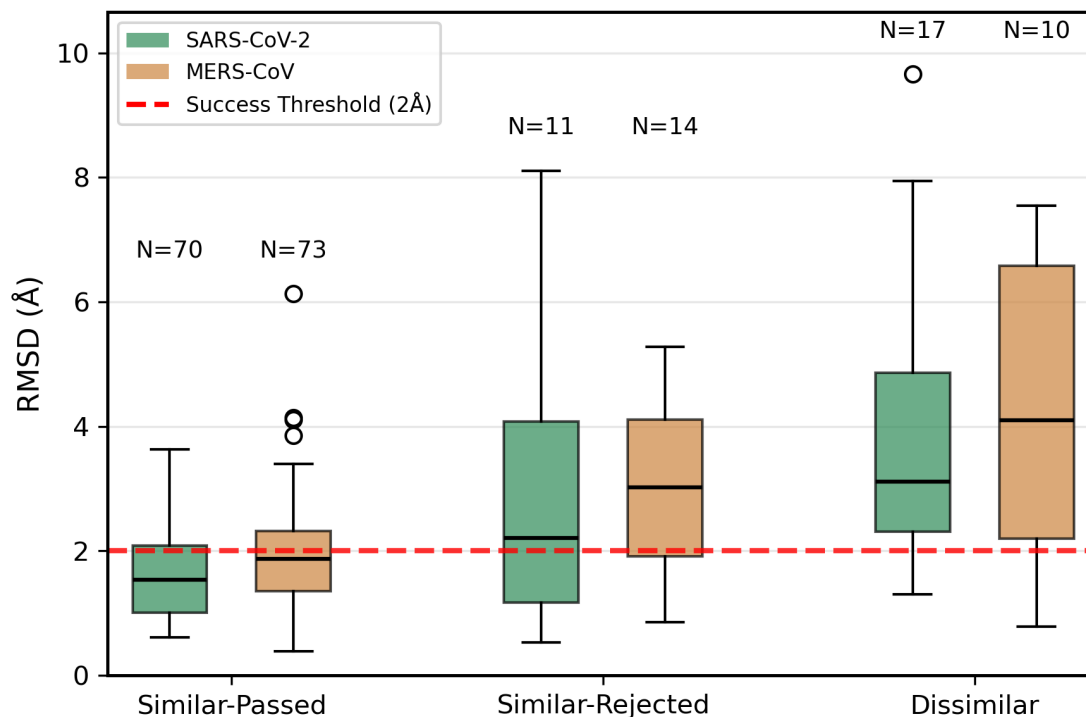


Figure 6.3: *Box plots of RMSD distribution for ligand pose prediction filtered into different categories of similarity to the training set for SARS-CoV-2 and MERS-CoV Mpro targets.* Similar-Passed consists of ligands that pass fragment similarity searches and custom score thresholds. Similar-Rejected 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.

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).

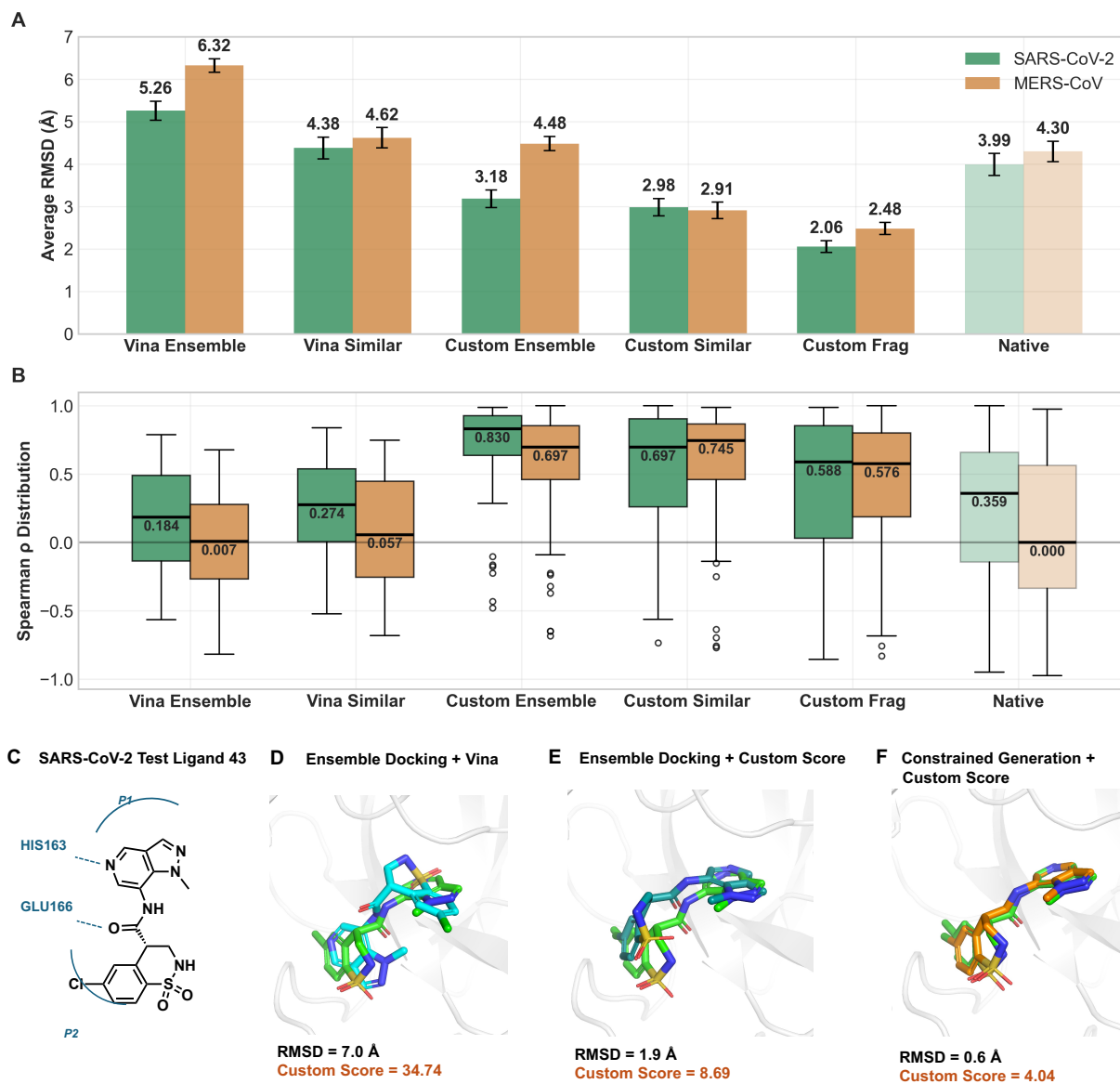


Figure 6.4: Comparison of pose prediction accuracy for ligands containing identical training fragments using different sampling and scoring approaches. (A) the average and standard deviation of best pose RMSD values determined by each method. (B) the distribution of score–RMSD correlation. For each testing 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 6.4A, 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 6.4B, 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 6.4, 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. 6.4C) 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 6.4D,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. 6.4F). 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.

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 (ligands 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 6.5, 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 6.5D). In this case, the top Vina Ensemble pose in Figure 6.5E 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 Å in 6.5F. 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 6.5G. 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.

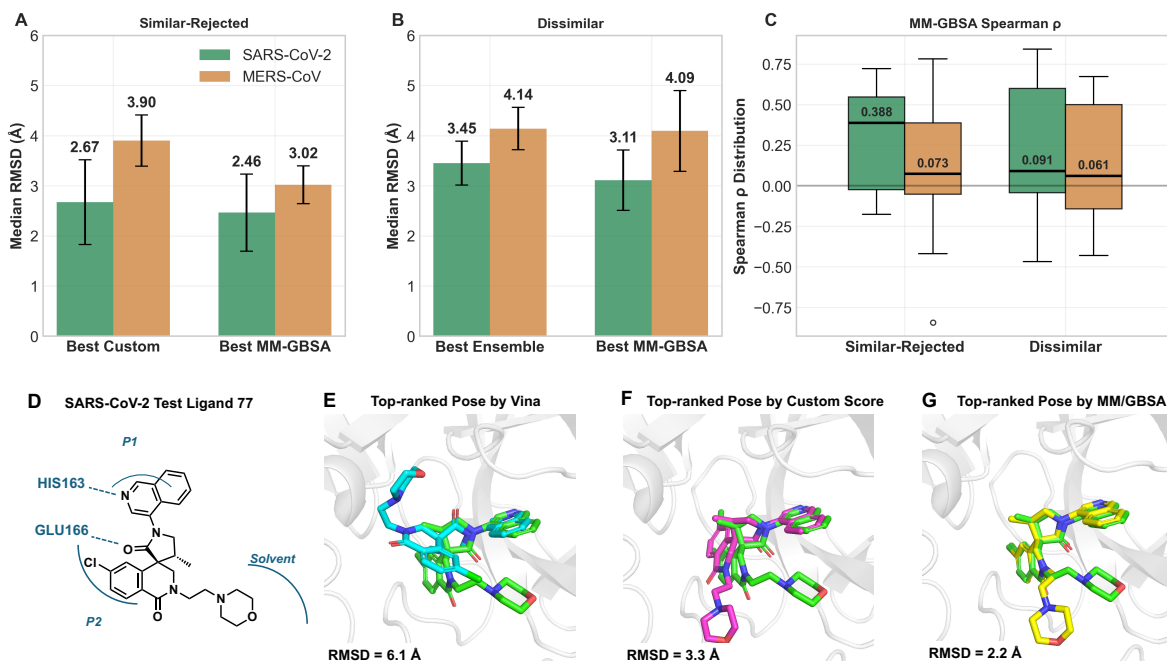


Figure 6.5: *Performance of MM/GBSA rescoring on challenging ligands without similar training fragments.* MM/GBSA provides modest improvements in (A,B) RMSD and (C) Spearman R , highlighting its role as a refinement step rather than a decisive predictor. (D) Illustration of MM/GBSA rescoring on SARS-CoV-2 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 an excellent tradeoff in computational efficiency and accuracy.

6.4 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 [335]. 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.

6.5 Supporting Information

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 $24 \text{ \AA}$ 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 $24 \text{ \AA}$ in each dimension.

Key Protein Residues. For protein alignment, we used the $C\alpha$ 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.

Supporting Figures

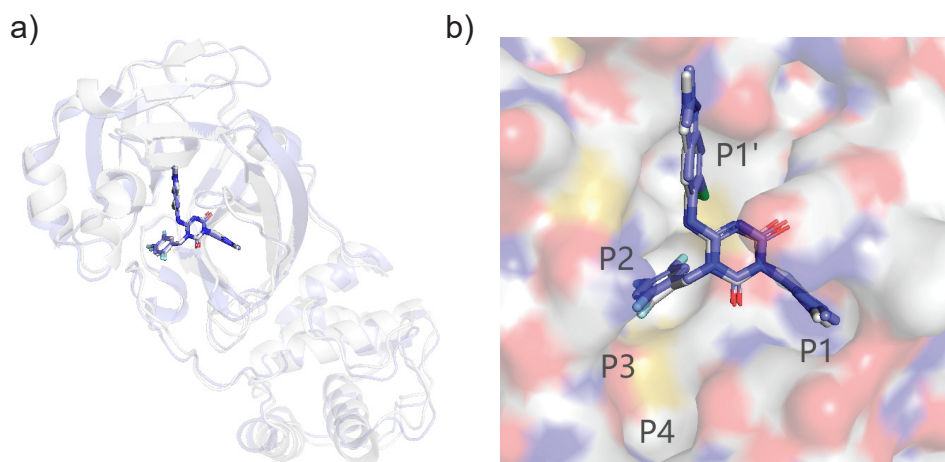


Figure S6.6: *Comparison of Ensitrelvir binding modes across SARS-CoV-2 and MERS-CoV Main proteases.* (a). Structure overlay between SARS-CoV-2 Mpro (PDB: 7VU6 [293]) and MERS-CoV Mpro (PDB: 8VUT[310]) in complex with Ensitrelvir. (b) Close-up visualization showing the highly homologous binding modes of Ensitrelvir across two proteins.

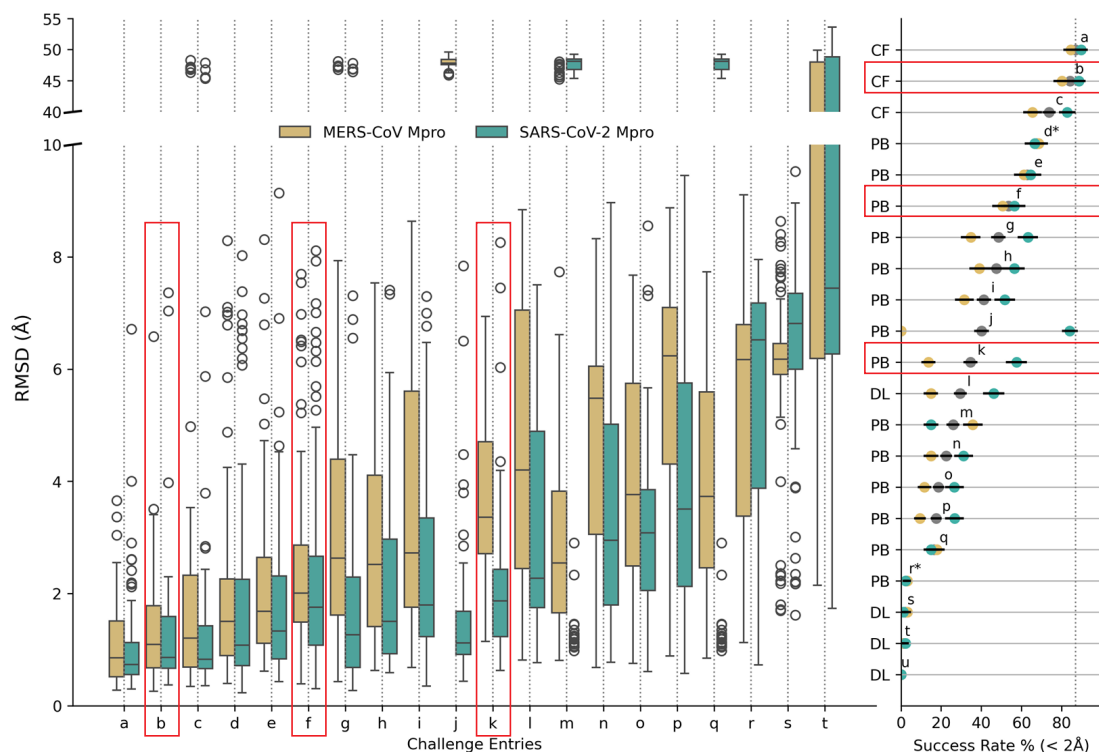


Figure S6.7: *Ligand pose challenge results from MacDermott-Opeskin et al. with boxes around key methods to highlight.*[13] Method b is a fine-tuned version of Boltz[320] using the provided training data; Method f is our submission; and Method k is a template-based docking method. Here, CF refers to 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. [13]

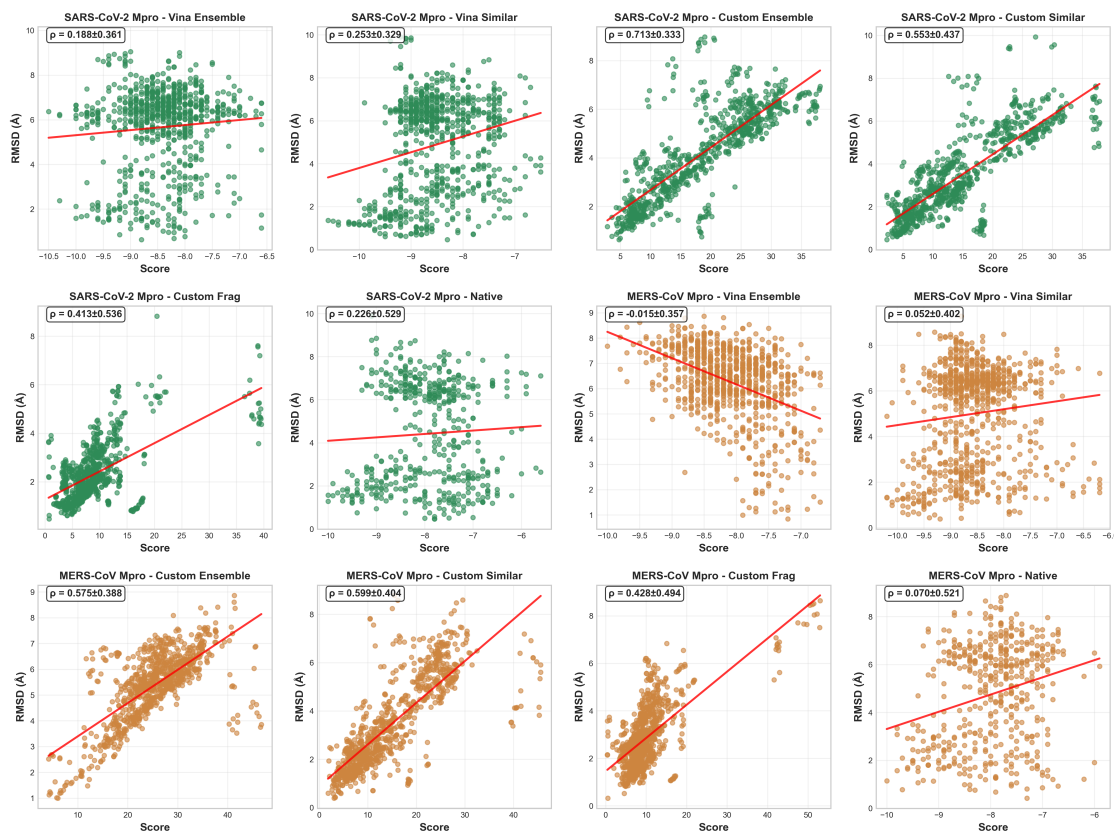


Figure S6.8: *Scatter plot of various scores versus RMSD between predicted and crystal poses.* For each target protein group, the top 10 poses ranked by each score along with their final RMSD are shown. Green points represent all 81 similar testing ligands for SARS-CoV-2 Mpro, and orange points represent all 87 similar testing ligands for MERS-CoV Mpro. The meaning of the ligand scoring category can be found in Main.

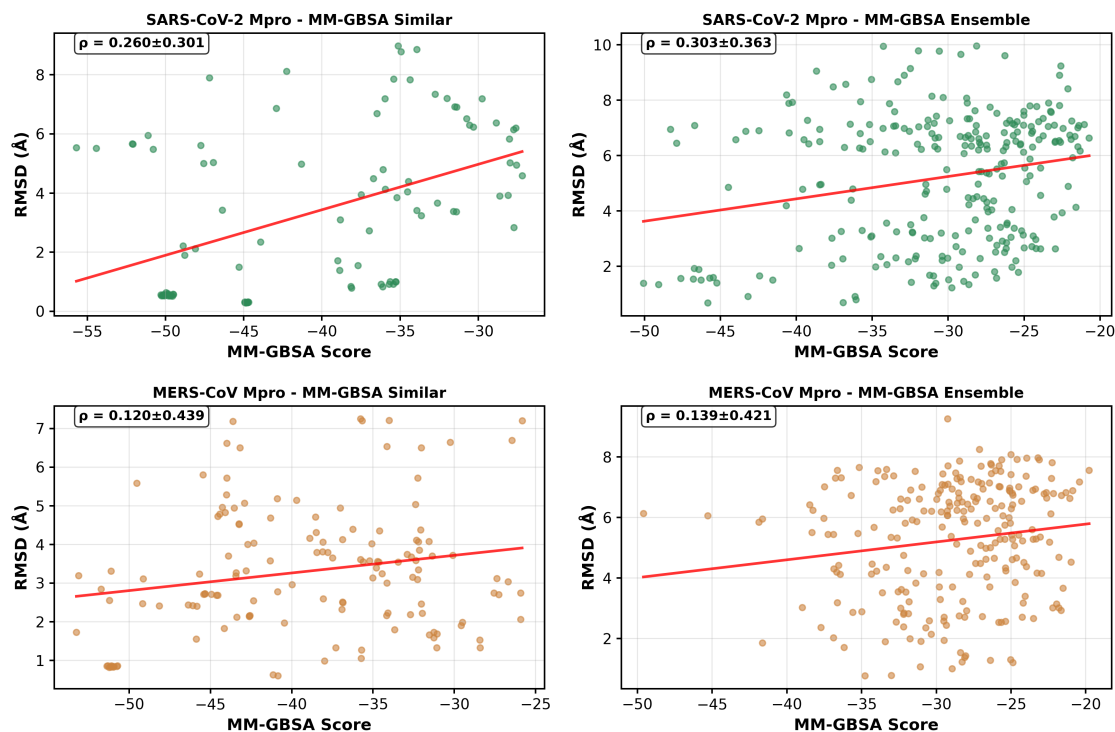


Figure S6.9: Scatter plot of MM/GBSA scores versus RMSD between final predicted and crystal poses. The top ten poses ranked by MM/GBSA energies are shown for each target protein group: (A) 14 MERS-CoV test ligands with identical fragments but rejected in earlier fragment-score filtering; (B) 11 SARS-CoV-2 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 S6.1: Summary of contest submission results by target and category. Values indicate the number of entries (N), the count of predictions with accuracy $< 2 \text{ \AA}$, and the corresponding percentage. Categories are described in main text.

Category	SARS-CoV-2			MERS-CoV		
	N	Count $< 2 \text{ \AA}$	% $< 2 \text{ \AA}$	N	Count $< 2 \text{ \AA}$	% $< 2 \text{ \AA}$
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 S6.2: Effect of MM/GBSA rescoring for the Similar-Passed ligand category.

Target	Condition	Count	RMSD ($\text{\AA}$)		$\leq 2 \text{ \AA}$	
			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 7

Conclusion

The fundamental challenge of modern drug discovery is not merely generating novel molecules from the practically infinite expanse of chemical space, but efficiently identifying the rare candidates that are synthesizable, biophysically active, and structurally sound. To address this problem, my thesis presents a comprehensive approach to modern computer-aided drug discovery, organized around two core pillars—generative molecular design with LLMs and structure-based benchmarking—unified by translational drug discovery workflows and experimental validation.

On the generative side, our works demonstrate that LLMs offer a paradigm-shifting alternative to traditional generative algorithms. SynLlama [8] and LinkLlama [49] show that language models can be fine-tuned to natively generate synthesizable molecules and molecular linkers with actionable synthesis pathways. Furthermore, my contributions to SmileyLlama [36] and TMCLlama [50] establish the broader utility of the "Chemical Llama Suite," demonstrating that natural language conditioning can intuitively guide multi-objective optimization without the computational overhead of reinforcement learning. Pushing this frontier further, my involvement in SAGA [51] embeds these generative models within autonomous agentic workflows that automate the reward-tuning loop traditionally performed by human scientists, highlighting a future where AI serves as an active, collaborative "second brain" for drug discovery campaigns and contribute to science in a supervised yet creative way.

However, intelligent generation is futile if grounded in flawed physical assumptions. On the benchmarking side, this thesis addresses the critical need for structural integrity and rigorous evaluation. Leak Proof PDBBind [41] reveals and corrects the pervasive data leakage that has historically inflated scoring function performance. HiQBind [9] systematically cleans structural noise and physical misrepresentations in protein-ligand datasets. Finally, KinConfBench [12] introduces conformation-aware evaluation of modern cofolding models for the kinase family, ensuring that structure prediction tools capture true functional states rather than prioritizing superficial geometric confidence. Together, these works provide the essential guardrails and reliable data that computational drug discovery depends upon.

Ultimately, the true validation of any computational methodology lies in its practical execution. These generative and benchmarking tools have been rigorously tested through

the iMiner workflow [52] in the Midwest AViDD campaign [53, 98] that yields experimentally confirmed, structurally novel inhibitors, and through competitive, blinded performance in the ASAP-Polaris-OpenADMET [13, 14] and CACHE [54] community challenges. These real-world deployments demonstrate that when machine learning tools are grounded in high-quality structural data, constrained by real-world chemical feasibility, and verified by physical simulation, they translate seamlessly from *in silico* theory to wet-lab reality. Looking forward, as the field transitions toward integrated, autonomous workflows, this thesis provides a robust foundation to ensure these computational engines consistently deliver novel, synthesizable, and effective therapeutics.

References

1. Amaro, R. E. & Mulholland, A. J. Multiscale methods in drug design bridge chemical and biological complexity in the search for cures. *Nature Rev. Chem.* **2**, 0148. ISSN: 2397-3358. <https://doi.org/10.1038/s41570-018-0148> (2018).
2. Batool, M., Ahmad, B. & Choi, S. A Structure-Based Drug Discovery Paradigm. *Inter. J. Mol. Sci.* **20**. ISSN: 1422-0067. <https://www.mdpi.com/1422-0067/20/11/2783> (2019).
3. Polishchuk, P. G., Madzhidov, T. I. & Varnek, A. Estimation of the size of drug-like chemical space based on GDB-17 data. *J. Comp.-Aid, Mol. Des.* **27**, 675–679 (2013).
4. Flam-Shepherd, D., Zhu, K. & Aspuru-Guzik, A. Language models can learn complex molecular distributions. *Nature Communications* **13**, 3293. ISSN: 2041-1723. <https://doi.org/10.1038/s41467-022-30839-x> (2022).
5. Skinnider, M., Stacey, R., Wishart, D. & Foster, L. Chemical language models enable navigation in sparsely populated chemical space. *Nature Machine Intelligence* **3**, 759–770 (2021).
6. Loeffler, H. H. *et al.* Reinvent 4: Modern AI-driven generative molecule design. en. *Journal of Cheminformatics* **16**, 20. ISSN: 1758-2946. <https://doi.org/10.1186/s13321-024-00812-5> (2026) (Feb. 2024).
7. Gao, W. & Coley, C. W. The Synthesizability of Molecules Proposed by Generative Models. en. *Journal of Chemical Information and Modeling* **60**, 5714–5723. ISSN: 1549-9596, 1549-960X. <https://pubs.acs.org/doi/10.1021/acs.jcim.0c00174> (2025) (Dec. 2020).
8. Sun, K. *et al.* SynLlama: Generating Synthesizable Molecules and Their Analogs with Large Language Models. *ACS Central Science* **11**, 2108–2120. ISSN: 2374-7943. <https://doi.org/10.1021/acscentsci.5c01285> (2025).
9. 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).

10. 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).
11. *PDBFixer* <https://github.com/openmm/pdbfixer>. accessed on Oct 29, 2024.
12. Sun, K. & Head-Gordon, T. *KinConfBench: A Curated Benchmark for Cofolding Models on Kinase Conformational States* en. Apr. 2026. <http://biorxiv.org/lookup/doi/10.64898/2026.04.07.716788> (2026).
13. MacDermott-Opeskin, H. *et al.* A Computational Community Blind Challenge on Pan-Coronavirus Drug Discovery Data. *Journal of Chemical Information and Modeling* **66**, 3129–3149. ISSN: 1549-9596, 1549-960X. <https://pubs.acs.org/doi/10.1021/acs.jcim.5c02106> (2026) (Mar. 23, 2026).
14. Sun, K., Wang, Y., Purnomo, J. & Head-Gordon, T. Polaris Challenge: Data-Driven Priors to Improve Docking for Pose Prediction. en. *Journal of Chemical Information and Modeling* **65**, 12058–12067. ISSN: 1549-9596, 1549-960X. <https://pubs.acs.org/doi/10.1021/acs.jcim.5c01976> (2026) (Nov. 2025).
15. 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).
16. Roy, R., Singh, S. K., Ahmad, N. & Misra, S. Challenges and advancements in high-throughput screening strategies for cancer therapeutics. *Global Translational Medicine* **3**, 2448. ISSN: 2811-0021. <https://accscience.com/journal/GTM/3/1/10.36922/gtm.2448> (2026) (Mar. 2024).
17. Neri, D. & Lerner, R. A. DNA-Encoded Chemical Libraries: A Selection System Based on Endowing Organic Compounds with Amplifiable Information. en. *Annual Review of Biochemistry* **87**, 479–502. ISSN: 0066-4154, 1545-4509. <https://www.annualreviews.org/doi/10.1146/annurev-biochem-062917-012550> (2026) (June 2018).
18. Gaulton, A. *et al.* ChEMBL: a large-scale bioactivity database for drug discovery. *Nuc. Acids Res.* **40**, D1100–D1107. ISSN: 0305-1048. <https://doi.org/10.1093/nar/gkr777> (2012).
19. Kim, S. *et al.* PubChem in 2021: new data content and improved web interfaces. *Nuc. Acids Res.* **49**, D1388–D1395. ISSN: 0305-1048. <https://doi.org/10.1093/nar/gkaa971> (2021).
20. Sterling, T. & Irwin, J. J. ZINC 15 – Ligand Discovery for Everyone. *J. Chem. Inform. Model.* **55**, 2324–2337. ISSN: 1549-9596. <https://doi.org/10.1021/acs.jcim.5b00559> (2015).

21. Neves, B. J. *et al.* QSAR-Based Virtual Screening: Advances and Applications in Drug Discovery. *Front. Pharma.* **9**. ISSN: 1663-9812. <https://www.frontiersin.org/articles/10.3389/fphar.2018.01275> (2018).
22. Grygorenko, O. O. *et al.* Generating Multibillion Chemical Space of Readily Accessible Screening Compounds. *iSci.* **23**. ISSN: 2589-0042. <https://doi.org/10.1016/j.isci.2020.101681> (2020).
23. Reulecke, I., Lange, G., Albrecht, J., Klein, R. & Rarey, M. Towards an integrated description of hydrogen bonding and dehydration: decreasing false positives in virtual screening with the HYDE scoring function. *ChemMedChem: Chemistry Enabling Drug Discovery* **3**, 885–897 (2008).
24. Duffy, B. C., Zhu, L., Decornez, H. & Kitchen, D. B. Early phase drug discovery: cheminformatics and computational techniques in identifying lead series. *Bioorg. Med. Chem.* **20**, 5324–5342 (2012).
25. Grisoni, F. Chemical language models for de novo drug design: Challenges and opportunities. *Current opinion in structural biology* **79**, 102527 (2023).
26. Sanchez-Lengeling, B. & Aspuru-Guzik, A. Inverse molecular design using machine learning: Generative models for matter engineering. *Science* **361**, 360–365 (2018).
27. Kusner, M. J., Paige, B. & Hernández-Lobato, J. M. *Grammar variational autoencoder* in *International Conference on Machine Learning* (2017), 1945–1954.
28. Olivecrona, M., Blaschke, T., Engkvist, O. & Chen, H. Molecular de-novo design through deep reinforcement learning. *J. Cheminform.* **9**, 1–14 (2017).
29. Popova, M., Isayev, O. & Tropsha, A. Deep reinforcement learning for de novo drug design. *Sci. Adv.* **4**, eaap7885 (2018).
30. Ertl, P. & Schuffenhauer, A. Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *Journal of Cheminformatics* **1**, 8. ISSN: 1758-2946. <https://jcheminf.biomedcentral.com/articles/10.1186/1758-2946-1-8> (2025) (2009).
31. Wang, S., Wang, L., Li, F. & Bai, F. DeepSA: a deep-learning driven predictor of compound synthesis accessibility. *Journal of Cheminformatics* **15**, 103. ISSN: 1758-2946. <https://doi.org/10.1186/s13321-023-00771-3> (2025) (Nov. 2023).
32. Bilodeau, C., Jin, W., Jaakkola, T., Barzilay, R. & Jensen, K. F. Generative models for molecular discovery: Recent advances and challenges. en. *WIREs Computational Molecular Science* **12**, e1608. ISSN: 1759-0876, 1759-0884. <https://wires.onlinelibrary.wiley.com/doi/10.1002/wcms.1608> (2025) (Sept. 2022).
33. Skoraczynski, G., Kitlas, M., Miasojedow, B. & Gambin, A. Critical assessment of synthetic accessibility scores in computer-assisted synthesis planning. *Journal of Cheminformatics* **15**, 6. ISSN: 1758-2946. <https://jcheminf.biomedcentral.com/articles/10.1186/s13321-023-00678-z> (2025) (2023).

34. Sumita, M., Yang, X., Ishihara, S., Tamura, R. & Tsuda, K. Hunting for Organic Molecules with Artificial Intelligence: Molecules Optimized for Desired Excitation Energies. en. *ACS Central Science* **4**, 1126–1133. ISSN: 2374-7943, 2374-7951. <https://pubs.acs.org/doi/10.1021/acscentsci.8b00213> (2025) (Sept. 2018).
35. Bommasani, R. *et al.* *On the Opportunities and Risks of Foundation Models* 2022. arXiv: 2108.07258 [cs.LG]. <https://arxiv.org/abs/2108.07258>.
36. Cavanagh, J. M. *et al.* *SmileyLlama: Modifying Large Language Models for Directed Chemical Space Exploration* 2024. arXiv: 2409.02231. <http://arxiv.org/abs/2409.02231> (2025).
37. Boiko, D. A., MacKnight, R., Kline, B. & Gomes, G. Autonomous chemical research with large language models. *Nature* **624**, 570–578. ISSN: 1476-4687. <https://www.nature.com/articles/s41586-023-06792-0> (2025) (Dec. 2023).
38. M. Bran, A. *et al.* Augmenting large language models with chemistry tools. *Nature Machine Intelligence* **6**, 525–535. ISSN: 2522-5839. <https://www.nature.com/articles/s42256-024-00832-8> (2025) (May 2024).
39. Yu, B., Baker, F. N., Chen, Z., Ning, X. & Sun, H. *LlaSMol: Advancing Large Language Models for Chemistry with a Large-Scale, Comprehensive, High-Quality Instruction Tuning Dataset* 2024. arXiv: 2402.09391 [cs.AI]. <https://arxiv.org/abs/2402.09391>.
40. Ramos, M. C., Collison, C. J. & White, A. D. A review of large language models and autonomous agents in chemistry. en. *Chemical Science*. ISSN: 2041-6539. <https://pubs.rsc.org/en/content/articlelanding/2025/sc/d4sc03921a> (2025) (Dec. 2024).
41. 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).
42. 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.
43. Buttenschoen, M., M. Morris, G. & M. Deane, C. PoseBusters: AI-based docking methods fail to generate physically valid poses or generalise to novel sequences. en. *Chemical Science* **15**, 3130–3139. <https://pubs.rsc.org/en/content/articlelanding/2024/sc/d3sc04185a> (2026) (2024).
44. 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).
45. Chai Discovery *et al.* *Chai-1: Decoding the molecular interactions of life* en. Oct. 2024. <http://biorxiv.org/lookup/doi/10.1101/2024.10.10.615955> (2025).

46. Wohlwend, J. *et al.* Boltz-1: Democratizing Biomolecular Interaction Modeling. *bioRxiv* (2024).
47. 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).
48. 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> (2026).
49. Sun, K. *et al.* LinkLlama: Enabling Large Language Model for Chemically Reasonable Linker Design en. Apr. 2026. <http://biorxiv.org/lookup/doi/10.64898/2026.04.15.718690> (2026).
50. Liu, Y. *et al.* Exploring Transition Metal Complexes with Large Language Models. *ChemRxiv* **2025**. <https://doi.org/10.26434/chemrxiv-2025-hm3zb>.
51. Du, Y. *et al.* Accelerating Scientific Discovery with Autonomous Goal-evolving Agents arXiv:2512.21782. Mar. 2026. <http://arxiv.org/abs/2512.21782> (2026).
52. Li, J. *et al.* Mining for Potent Inhibitors through Artificial Intelligence and Physics: A Unified Methodology for Ligand Based and Structure Based Drug Design. *Journal of Chemical Information and Modeling* **64**, 9082–9097. ISSN: 1549-9596. <https://doi.org/10.1021/acs.jcim.4c00634> (2024).
53. *Research Overview — Midwest Antiviral Drug Discovery (AViDD) Center* <https://midwestavidd.umn.edu/research/research-overview> (2026).
54. Ackloo, S. *et al.* CACHE (Critical Assessment of Computational Hit-finding Experiments): A public–private partnership benchmarking initiative to enable the development of computational methods for hit-finding. en. *Nature Reviews Chemistry* **6**, 287–295. ISSN: 2397-3358. <https://www.nature.com/articles/s41570-022-00363-z> (2026) (Apr. 2022).
55. Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **28**, 31–36 (1988).
56. Krenn, M., Häse, F., Nigam, A., Friederich, P. & Aspuru-Guzik, A. Self-referencing embedded strings (SELFIES): A 100% robust molecular string representation. *Machine Learning: Science and Technology* **1**, 045024. <https://doi.org/10.1088/2632-2153/aba947> (2020).
57. Gómez-Bombarelli, R. *et al.* Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. *ACS Central Science* **4**, 268–276. ISSN: 2374-7943. <https://doi.org/10.1021/acscentsci.7b00572> (2018).

58. Radford, A., Narasimhan, K., Salimans, T. & Sutskever, I. *Improving Language Understanding by Generative Pre-Training* 2018. https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf.
59. Dubey, A. *et al.* *The Llama 3 Herd of Models* Aug. 15, 2024. arXiv: 2407.21783[cs]. <http://arxiv.org/abs/2407.21783> (2024).
60. Podda, M., Bacciu, D. & Micheli, A. *A Deep Generative Model for Fragment-Based Molecule Generation* en. in *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics* (PMLR, June 2020), 2240–2250. <https://proceedings.mlr.press/v108/podda20a.html> (2025).
61. Yang, S., Hwang, D., Lee, S., Ryu, S. & Hwang, S. J. *Hit and Lead Discovery with Explorative RL and Fragment-based Molecule Generation* in *Advances in Neural Information Processing Systems* (eds Beygelzimer, A., Dauphin, Y., Liang, P. & Vaughan, J. W.) (2021). <https://openreview.net/forum?id=Msc9XKd-3bA>.
62. Chen, Z., Min, M. R., Parthasarathy, S. & Ning, X. A deep generative model for molecule optimization via one fragment modification. en. *Nature Machine Intelligence* **3**, 1040–1049. ISSN: 2522-5839. <https://www.nature.com/articles/s42256-021-00410-2> (2025) (Dec. 2021).
63. Seo, S., Lim, J. & Kim, W. Y. Molecular Generative Model via Retrosynthetically Prepared Chemical Building Block Assembly. en. *Advanced Science* **10**, 2206674. ISSN: 2198-3844, 2198-3844. <https://onlinelibrary.wiley.com/doi/10.1002/adv.202206674> (2025) (Mar. 2023).
64. Genheden, S. *et al.* AiZynthFinder: a fast, robust and flexible open-source software for retrosynthetic planning. *Journal of Cheminformatics* **12**, 70. ISSN: 1758-2946. <https://doi.org/10.1186/s13321-020-00472-1> (2025) (Nov. 2020).
65. Shen, Y. *et al.* Automation and computer-assisted planning for chemical synthesis. *Nature Reviews Methods Primers* **1**, 23. ISSN: 2662-8449. <https://doi.org/10.1038/s43586-021-00022-5> (2021).
66. Coley, C. W. *et al.* A robotic platform for flow synthesis of organic compounds informed by AI planning. *Science* **365**, eaax1566. ISSN: 0036-8075, 1095-9203. <https://www.science.org/doi/10.1126/science.aax1566> (2025) (Aug. 9, 2019).
67. Gao, W., Mercado, R. & Coley, C. W. *Amortized Tree Generation for Bottom-up Synthesis Planning and Synthesizable Molecular Design* arXiv:2110.06389. Mar. 2022. <http://arxiv.org/abs/2110.06389> (2025).
68. Luo, S. *et al.* *Projecting Molecules into Synthesizable Chemical Spaces* arXiv:2406.04628. June 2024. <http://arxiv.org/abs/2406.04628> (2025).
69. Gao, W., Luo, S. & Coley, C. W. *Generative Artificial Intelligence for Navigating Synthesizable Chemical Space* arXiv:2410.03494. Oct. 2024. <http://arxiv.org/abs/2410.03494> (2025).

70. Igashov, I. *et al.* Equivariant 3D-conditional diffusion model for molecular linker design. en. *Nature Machine Intelligence* **6**, 417–427. ISSN: 2522-5839. <https://www.nature.com/articles/s42256-024-00815-9> (2026) (Apr. 2024).
71. Imrie, F., Bradley, A. R., Van Der Schaar, M. & Deane, C. M. Deep Generative Models for 3D Linker Design. en. *Journal of Chemical Information and Modeling* **60**, 1983–1995. ISSN: 1549-9596, 1549-960X. <https://pubs.acs.org/doi/10.1021/acs.jcim.9b01120> (2026) (Apr. 2020).
72. 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).
73. Morris, G. M. *et al.* AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *J. Comp. Chem.* **30**, 2785–2791 (2009).
74. 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).
75. 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).
76. 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).
77. 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).
78. 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).
79. 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).
80. Lu, W. *et al.* Tankbind: Trigonometry-aware neural networks for drug-protein binding structure prediction. *bioRxiv*, 2022–06 (2022).
81. Su, M. *et al.* Comparative assessment of scoring functions: the CASF-2016 update. *J. Chem. Info. Model.* **59**, 895–913 (2018).
82. 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).

83. 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).
84. 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).
85. 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).
86. 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).
87. 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).
88. Passaro, S. *et al.* Boltz-2: Towards Accurate and Efficient Binding Affinity Prediction. *bioRxiv* (2025).
89. Wang, Y.-y., Zhao, R. & Zhe, H. The emerging role of CaMKII in cancer. *Oncotarget* **6**, 11725–11734. ISSN: 1949-2553. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4494900/> (2026) (Apr. 2015).
90. Zhang, M. *et al.* CDK inhibitors in cancer therapy, an overview of recent development. *American Journal of Cancer Research* **11**, 1913–1935. ISSN: 2156-6976. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8167670/> (2026) (May 2021).
91. Pottier, C. *et al.* Tyrosine Kinase Inhibitors in Cancer: Breakthrough and Challenges of Targeted Therapy. *Cancers* **12**, 731. ISSN: 2072-6694. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7140093/> (2026) (Mar. 2020).
92. Möbitz, H. The ABC of protein kinase conformations. eng. *Biochimica Et Biophysica Acta* **1854**, 1555–1566. ISSN: 0006-3002 (Oct. 2015).
93. Modi, V. & Dunbrack, R. L. Defining a new nomenclature for the structures of active and inactive kinases. en. *Proceedings of the National Academy of Sciences* **116**, 6818–6827. ISSN: 0027-8424, 1091-6490. <https://pnas.org/doi/full/10.1073/pnas.1814279116> (2025) (Apr. 2019).
94. Roskoski, R. Classification of small molecule protein kinase inhibitors based upon the structures of their drug-enzyme complexes. *Pharmacological Research* **103**, 26–48. ISSN: 1043-6618. <https://www.sciencedirect.com/science/article/pii/S1043661815301298> (2026) (Jan. 2016).

95. Lipinski, C. A. Drug-like properties and the causes of poor solubility and poor permeability. *J. Pharma. Tox. Meth.* **44**, 235–249. ISSN: 1056-8719. <http://www.sciencedirect.com/science/article/pii/S1056871900001076> (2000).
96. Lipinski, C. A., Lombardo, F., Dominy, B. W. & Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Delivery Rev.* **23**, 3–25 (1997).
97. Veber, D. F. *et al.* Molecular Properties That Influence the Oral Bioavailability of Drug Candidates. *Journal of Medicinal Chemistry* **45**, 2615–2623. ISSN: 0022-2623. <https://doi.org/10.1021/jm020017n> (2024) (June 1, 2002).
98. *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).
99. Polaris. *Polaris* en. <https://polarishub.io> (2025).
100. newpixcom. *AI-driven Structure-enabled Antiviral Platform (ASAP)* en-us. <https://asapdiscovery.org/> (2025).
101. OpenADMET. *OpenADMET* en-us. <https://openadmet.org/> (2025).
102. 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).
103. Corey, E. en. in *The Chemistry of Natural Products* 19–37 (Elsevier, 1967). ISBN: 9780080207414. <https://linkinghub.elsevier.com/retrieve/pii/B978008020741450004X> (2025).
104. Ihlenfeldt, W.-D. & Gasteiger, J. Computer-Assisted Planning of Organic Syntheses: The Second Generation of Programs. en. *Angewandte Chemie International Edition in English* **34**, 2613–2633. ISSN: 0570-0833. <https://onlinelibrary.wiley.com/doi/10.1002/anie.199526131> (2025) (Jan. 1996).
105. Corey, E. J., Long, A. K. & Rubenstein, S. D. Computer-Assisted Analysis in Organic Synthesis. en. *Science* **228**, 408–418. ISSN: 0036-8075, 1095-9203. <https://www.science.org/doi/10.1126/science.3838594> (2025) (Apr. 1985).
106. Kowalik, M. *et al.* Parallel Optimization of Synthetic Pathways within the Network of Organic Chemistry. *Angewandte Chemie International Edition* **51**, 7928–7932. ISSN: 1433-7851. <https://doi.org/10.1002/anie.201202209> (2012).
107. Irwin, J. J., Sterling, T., Mysinger, M. M., Bolstad, E. S. & Coleman, R. G. ZINC: A Free Tool to Discover Chemistry for Biology. *Journal of Chemical Information and Modeling* **52**, 1757–1768. ISSN: 1549-9596, 1549-960X. <https://pubs.acs.org/doi/10.1021/ci3001277> (2025) (2012).

108. Gaulton, A. *et al.* ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Research* **40**, D1100–D1107. ISSN: 0305-1048. <https://doi.org/10.1093/nar/gkr777> (2024) (D1 2012).
109. Enamine. *Building Block Catalogs* <https://enamine.net/>. Accessed: 2024-10-23.
110. Chithrananda, S., Grand, G. & Ramsundar, B. ChemBERTa: Large-Scale Self-Supervised Pretraining for Molecular Property Prediction. *CoRR* **abs/2010.09885**. <https://arxiv.org/abs/2010.09885> (2020).
111. Li, J. & Jiang, X. Mol-BERT: An Effective Molecular Representation with BERT for Molecular Property Prediction. *Wireless Communications and Mobile Computing* **2021** (ed Wang, Y.) 7181815. ISSN: 1530-8669, 1530-8677. <https://onlinelibrary.wiley.com/doi/10.1155/2021/7181815> (2025) (2021).
112. Bagal, V., Aggarwal, R., Vinod, P. K. & Priyakumar, U. MolGPT: Molecular Generation Using a Transformer-Decoder Model. *Journal of chemical information and modeling* **62**, 2064–2076 (2022).
113. Eckmann, P. *et al.* LIMO: Latent Inceptionism for Targeted Molecule Generation in *Proceedings of the 39th International Conference on Machine Learning* (eds Chaudhuri, K. *et al.*) **162** (PMLR, 17–23 Jul 2022), 5777–5792. <https://proceedings.mlr.press/v162/eckmann22a.html>.
114. Wang, Y., Zhao, H., Sciabola, S. & Wang, W. cMolGPT: A Conditional Generative Pre-Trained Transformer for Target-Specific De Novo Molecular Generation. *Molecules* **28**, 4430. ISSN: 1420-3049. <https://www.mdpi.com/1420-3049/28/11/4430> (2025) (2023).
115. Guan, J., Qian, W., Peng, X., *et al.* 3D equivariant diffusion for target-aware molecule generation and affinity prediction in *The Eleventh International Conference on Learning Representations* (Kigali, Rwanda, 2023).
116. Li, S. *et al.* LS-MolGen: Ligand-and-Structure Dual-Driven Deep Reinforcement Learning for Target-Specific Molecular Generation Improves Binding Affinity and Novelty. *Journal of Chemical Information and Modeling* (2023).
117. Luo, S., Guan, J., Jianzhu, M., *et al.* A 3D generative model for structure-based drug design in *Advances in Neural Information Processing Systems* **34** (2021), 6229–6239.
118. Peng, X. *et al.* Pocket2mol: Efficient molecular sampling based on 3d protein pockets in *International Conference on Machine Learning* (2022), 17644–17655.
119. Zhang, J. & Chen, H. De novo molecule design using molecular generative models constrained by ligand–protein interactions. *Journal of Chemical Information and Modeling* **62**, 3291–3306 (2022).
120. Qian, H., Lin, C., Zhao, D., *et al.* AlphaDrug: protein target specific de novo molecular generation. *PNAS Nexus* **1**, pgac227–238 (2022).

121. Schneuing, A. *et al.* Structure-based Drug Design with Equivariant Diffusion Models 2023. <https://openreview.net/forum?id=uKmuzIuVl8z>.
122. Zhang, O. *et al.* ResGen is a pocket-aware 3D molecular generation model based on parallel multiscale modelling. *Nature Machine Intelligence* **5**, 1020–1030 (2023).
123. Zhavoronkov, A. *et al.* Deep learning enables rapid identification of potent DDR1 kinase inhibitors. en. *Nature Biotechnology* **37**, 1038–1040. ISSN: 1087-0156, 1546-1696. <https://www.nature.com/articles/s41587-019-0224-x> (2025) (Sept. 2019).
124. Coley, C. W., Rogers, L., Green, W. H. & Jensen, K. F. SCScore: Synthetic Complexity Learned from a Reaction Corpus. en. *Journal of Chemical Information and Modeling* **58**, 252–261. ISSN: 1549-9596, 1549-960X. <https://pubs.acs.org/doi/10.1021/acs.jcim.7b00622> (2025) (Feb. 2018).
125. Voršilák, M., Kolář, M., Čmelo, I. & Svozil, D. SYBA: Bayesian estimation of synthetic accessibility of organic compounds. *Journal of Cheminformatics* **12**, 35. ISSN: 1758-2946. <https://doi.org/10.1186/s13321-020-00439-2> (2025) (May 2020).
126. Thakkar, A., Chadimová, V., Bjerrum, E. J., Engkvist, O. & Reymond, J.-L. Retrosynthetic accessibility score (RAscore) – rapid machine learned synthesizability classification from AI driven retrosynthetic planning. en. *Chemical Science* **12**, 3339–3349. ISSN: 2041-6539. <https://pubs.rsc.org/en/content/articlelanding/2021/sc/d0sc05401a> (2025) (Mar. 2021).
127. Kim, H., Lee, K., Kim, C., Lim, J. & Kim, W. Y. DFRscore: Deep Learning-Based Scoring of Synthetic Complexity with Drug-Focused Retrosynthetic Analysis for High-Throughput Virtual Screening. en. *Journal of Chemical Information and Modeling* **64**, 2432–2444. ISSN: 1549-9596, 1549-960X. <https://pubs.acs.org/doi/10.1021/acs.jcim.3c01134> (2025) (Apr. 2024).
128. Neeser, R. M., Correia, B. & Schwaller, P. FSscore: A Personalized Machine Learning-Based Synthetic Feasibility Score. en. *Chemistry–Methods* **4**, e202400024. ISSN: 2628-9725, 2628-9725. <https://chemistry-europe.onlinelibrary.wiley.com/doi/10.1002/cmt.202400024> (2025) (Nov. 2024).
129. Tu, Z. *et al.* ASKCOS: Open-Source, Data-Driven Synthesis Planning. en. *Accounts of Chemical Research* **58**, 1764–1775. ISSN: 0001-4842, 1520-4898. <https://pubs.acs.org/doi/10.1021/acs.accounts.5c00155> (2025) (June 2025).
130. Guo, J. & Schwaller, P. Directly optimizing for synthesizability in generative molecular design using retrosynthesis models. en. *Chemical Science* **16**, 6943–6956. ISSN: 2041-6539. <https://pubs.rsc.org/en/content/articlelanding/2025/sc/d5sc01476j> (2025) (Apr. 2025).

131. Li, P. *et al.* A deep learning approach for rational ligand generation with toxicity control via reactive building blocks. en. *Nature Computational Science* **4**, 851–864. ISSN: 2662-8457. <https://www.nature.com/articles/s43588-024-00718-0> (2025) (Nov. 2024).
132. Swanson, K. *et al.* Generative AI for designing and validating easily synthesizable and structurally novel antibiotics. *Nature Machine Intelligence* **6**, 338–353. ISSN: 2522-5839. <https://www.nature.com/articles/s42256-024-00809-7> (2025) (2024).
133. Cretu, M. *et al.* *SynFlowNet: Design of Diverse and Novel Molecules with Synthesis Constraints* in *The Thirteenth International Conference on Learning Representations* (2025). <https://openreview.net/forum?id=uvHmnaHyp1>.
134. Seo, S. *et al.* *Generative Flows on Synthetic Pathway for Drug Design* en. in (Oct. 2024). <https://openreview.net/forum?id=pB1XSj2y4X> (2025).
135. Wang, M. *et al.* ClickGen: Directed exploration of synthesizable chemical space via modular reactions and reinforcement learning. en. *Nature Communications* **15**, 10127. ISSN: 2041-1723. <https://www.nature.com/articles/s41467-024-54456-y> (2025) (Nov. 2024).
136. Hendrycks, D. *et al.* *Measuring Massive Multitask Language Understanding in International Conference on Learning Representations* (2021). <https://openreview.net/forum?id=d7KBJmI3GmQ>.
137. Wang, Y. *et al.* *MMLU-Pro: A More Robust and Challenging Multi-Task Language Understanding Benchmark* in *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track* (2024). <https://openreview.net/forum?id=y10DM6R2r3>.
138. Hartenfeller, M. *et al.* DOGS: Reaction-Driven de novo Design of Bioactive Compounds. en. *PLoS Computational Biology* **8** (ed Briggs, J. M.) e1002380. ISSN: 1553-7358. <https://dx.plos.org/10.1371/journal.pcbi.1002380> (2025) (Feb. 2012).
139. Button, A., Merk, D., Hiss, J. A. & Schneider, G. Automated de novo molecular design by hybrid machine intelligence and rule-driven chemical synthesis. en. *Nature Machine Intelligence* **1**, 307–315. ISSN: 2522-5839. <https://www.nature.com/articles/s42256-019-0067-7> (2025) (July 2019).
140. Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. en. *Journal of Chemical Information and Computer Sciences* **28**, 31–36. ISSN: 0095-2338, 1520-5142. <https://pubs.acs.org/doi/abs/10.1021/ci00057a005> (2025) (Feb. 1988).
141. Weininger, D. *Daylight Theory: SMARTS - A Language for Describing Molecular Patterns* <https://www.daylight.com/dayhtml/doc/theory/theory.smarts.html> (2025).
142. Molport. *List Search* <https://www.molport.com/shop/swl-step-1>. Accessed: 2024-12-23.

143. Morgan, H. L. The Generation of a Unique Machine Description for Chemical Structures—A Technique Developed at Chemical Abstracts Service. *Journal of Chemical Documentation* **5**, 107–113 (1965).
144. Bemis, G. W. & Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. *Journal of Medicinal Chemistry* **39**. PMID: 8709122, 2887–2893. eprint: <https://doi.org/10.1021/jm9602928>. <https://doi.org/10.1021/jm9602928> (1996).
145. Gobbi, A. & Poppinger, D. Genetic optimization of combinatorial libraries. *Biotechnology and Bioengineering* **61**, 47–54. <https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/abs/10.1002/%28SICI%291097-0290%28199824%2961%3A1%3C47%3A%3AAID-BIT9%3E3.0.CO%3B2-Z> (1998).
146. 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. ISSN: 0192-8651. <https://doi.org/10.1002/jcc.21334> (Jan. 2010).
147. 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 Cent. Sci.* **7**, 467–475 (2021).
148. Baum, B. *et al.* More than a simple lipophilic contact: a detailed thermodynamic analysis of nonbasic residues in the S1 pocket of thrombin. *Journal of molecular biology* **390**, 56–69 (2009).
149. Liang, J. *et al.* Lead identification of novel and selective TYK2 inhibitors. en. *European Journal of Medicinal Chemistry* **67**, 175–187. ISSN: 02235234. <https://linkinghub.elsevier.com/retrieve/pii/S0223523413002304> (2024) (Sept. 2013).
150. Liang, J. *et al.* Lead Optimization of a 4-Aminopyridine Benzamide Scaffold To Identify Potent, Selective, and Orally Bioavailable TYK2 Inhibitors. *Journal of Medicinal Chemistry* **56**. Publisher: American Chemical Society, 4521–4536. ISSN: 0022-2623. <https://doi.org/10.1021/jm400266t> (2024) (June 2013).
151. Clyde, A. *et al.* High-Throughput Virtual Screening and Validation of a SARS-CoV-2 Main Protease Noncovalent Inhibitor. *Journal of Chemical Information and Modeling* **62**. PMID: 34793155, 116–128. eprint: <https://doi.org/10.1021/acs.jcim.1c00851>. <https://doi.org/10.1021/acs.jcim.1c00851> (2022).
152. Kneller, D. W. *et al.* Structural, Electronic, and Electrostatic Determinants for Inhibitor Binding to Subsites S1 and S2 in SARS-CoV-2 Main Protease. *Journal of Medicinal Chemistry* **64**. PMID: 34705466, 17366–17383. eprint: <https://doi.org/10.1021/acs.jmedchem.1c01475>. <https://doi.org/10.1021/acs.jmedchem.1c01475> (2021).
153. *Schrodinger-FEP-Benchmark* https://github.com/schrodinger/public_binding_free_energy_benchmark. accessed on June 24, 2025.

154. Lian, W. *axolotl*. URL <https://github.com/axolotl-ai-cloud/axolotl/tree/main> <https://github.com/axolotl-ai-cloud/axolotl/tree/main> (2024).
155. Hu, E. J. *et al.* LoRA: Low-Rank Adaptation of Large Language Models in *International Conference on Learning Representations* (2022). <https://openreview.net/forum?id=nZeVKeeFYf9>.
156. Dao, T. *FlashAttention-2: Faster Attention with Better Parallelism and Work Partitioning* in *International Conference on Learning Representations (ICLR)* (2024).
157. Kingma, D. P. & Ba, J. Adam: A Method for Stochastic Optimization in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings* (eds Bengio, Y. & LeCun, Y.) (2015). <http://arxiv.org/abs/1412.6980>.
158. Spärck Jones, K. A Statistical Interpretation of Term Specificity and Its Application in Retrieval. *Journal of Documentation* **28**, 11–21 (1972).
159. Rogers, D. J. & Tanimoto, T. T. A Computer Program for Classifying Plants: The computer is programmed to simulate the taxonomic process of comparing each case with every other case. en. *Science* **132**, 1115–1118. ISSN: 0036-8075, 1095-9203. <https://www.science.org/doi/10.1126/science.132.3434.1115> (2025) (Oct. 1960).
160. Huang, K. *et al.* Therapeutics Data Commons: Machine Learning Datasets and Tasks for Drug Discovery and Development. *Proceedings of Neural Information Processing Systems, NeurIPS Datasets and Benchmarks* (2021).
161. Krenn, M., Häse, F., Nigam, A., Friederich, P. & Aspuru-Guzik, A. Self-referencing embedded strings (SELFIES): A 100% robust molecular string representation. *Mach. Learn.: Sci. Tech.* **1**, 045024. ISSN: 2632-2153 (2020).
162. Berman, H. *et al.* The Protein Data Bank. *Nucleic Acids Research* **28**, 235–242 (2000).
163. Eastman, P. *et al.* OpenMM 4: A Reusable, Extensible, Hardware Independent Library for High Performance Molecular Simulation. *Journal of Chemical Theory and Computation* **9**, 461–469 (2013).
164. *Pocket2Mol* <https://github.com/pengxingang/Pocket2Mol>. accessed on Jun 24, 2025.
165. Case, D. A. *et al.* AMBER22 San Francisco, CA, USA, 2022.
166. 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).
167. 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).

168. Chen, W. *et al.* Accurate calculation of relative binding free energies between ligands with different net charges. *Journal of chemical theory and computation* **14**, 6346–6358 (2018).
169. 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 (2023).
170. 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 (2024).
171. 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 (2023).
172. Hopkins, C. W., Le Grand, S., Walker, R. C. & Roitberg, A. E. Long-time-step molecular dynamics through hydrogen mass repartitioning. *Journal of chemical theory and computation* **11**, 1864–1874 (2015).
173. Wu, Z. *et al.* alchemlyb: the simple alchemistry library. *Journal of Open Source Software* **9**, 6934. <https://doi.org/10.21105/joss.06934> (2024).
174. Xu, H. Optimal measurement network of pairwise differences. *Journal of Chemical Information and Modeling* **59**, 4720–4728 (2019).
175. Rees, D. C., Congreve, Murray, C. W. & Carr, R. Fragment-based lead discovery. en. *Nature Reviews Drug Discovery* **3**, 660–672. ISSN: 1474-1784. <https://www.nature.com/articles/nrd1467> (2026) (Aug. 2004).
176. Erlanson, D. A., McDowell, R. S. & O'Brien, T. Fragment-Based Drug Discovery. en. *Journal of Medicinal Chemistry* **47**, 3463–3482. ISSN: 0022-2623, 1520-4804. <https://pubs.acs.org/doi/10.1021/jm040031v> (2026) (July 2004).
177. Congreve, M., Chessari, G., Tisi, D. & Woodhead, A. J. Recent Developments in Fragment-Based Drug Discovery. en. *Journal of Medicinal Chemistry* **51**, 3661–3680. ISSN: 0022-2623, 1520-4804. <https://pubs.acs.org/doi/10.1021/jm8000373> (2026) (July 2008).
178. Erlanson, D. A., Fesik, S. W., Hubbard, R. E., Jahnke, W. & Jhoti, H. Twenty years on: the impact of fragments on drug discovery. en. *Nature Reviews Drug Discovery* **15**, 605–619. ISSN: 1474-1784. <https://www.nature.com/articles/nrd.2016.109> (2026) (Sept. 2016).
179. Bian, Y. & Xie, X.-Q. (Computational Fragment-Based Drug Design: Current Trends, Strategies, and Applications. en. *The AAPS Journal* **20**, 59. ISSN: 1550-7416. <https://doi.org/10.1208/s12248-018-0216-7> (2026) (Apr. 2018).
180. Kirsch, P., Hartman, A. M., Hirsch, A. K. H. & Empting, M. Concepts and Core Principles of Fragment-Based Drug Design. en. *Molecules* **24**, 4309. ISSN: 1420-3049. <https://www.mdpi.com/1420-3049/24/23/4309> (2026) (Nov. 2019).

181. Bancet, A. *et al.* Fragment Linking Strategies for Structure-Based Drug Design. en. *Journal of Medicinal Chemistry* **63**, 11420–11435. ISSN: 0022-2623, 1520-4804. <https://pubs.acs.org/doi/10.1021/acs.jmedchem.0c00242> (2026) (Oct. 2020).
182. Bian, Y. & Xie, X.-Q. (Computational Fragment-Based Drug Design: Current Trends, Strategies, and Applications. en. *The AAPS Journal* **20**, 59. ISSN: 1550-7416. <https://doi.org/10.1208/s12248-018-0216-7> (2026) (Apr. 2018).
183. Grenier, D., Audebert, S., Preto, J., Guichou, J.-F. & Krimm, I. Linkers in fragment-based drug design: an overview of the literature. en. *Expert Opinion on Drug Discovery* **18**, 987–1009. ISSN: 1746-0441, 1746-045X. <https://www.tandfonline.com/doi/full/10.1080/17460441.2023.2234285> (2026) (Sept. 2023).
184. Guo, J. *et al.* Link-INVENT: generative linker design with reinforcement learning. en. *Digital Discovery* **2**, 392–408. ISSN: 2635-098X. <https://pubs.rsc.org/en/content/articlelanding/2023/dd/d2dd00115b> (2026) (Apr. 2023).
185. Chen, S. *et al.* Deep lead optimization enveloped in protein pocket and its application in designing potent and selective ligands targeting LTK protein. en. *Nature Machine Intelligence* **7**, 448–458. ISSN: 2522-5839. <https://www.nature.com/articles/s42256-025-00997-w> (2026) (Mar. 2025).
186. Bickerton, G. R., Paolini, G. V., Besnard, J., Muresan, S. & Hopkins, A. L. Quantifying the chemical beauty of drugs. *Nature Chem.* **4**, 90–98 (2012).
187. Walters, P. *Generative Molecular Design Isn't As Easy As People Make It Look* <https://practicalcheminformatics.blogspot.com/2024/05/generative-molecular-design-isnt-as.html> (2026).
188. *rdkit/rdkit* original-date: 2013-05-12T06:19:15Z. Apr. 2026. <https://github.com/rdkit/rdkit> (2026).
189. Walters, P. *Mining Ring Systems in Molecules for Fun and Profit* <https://practicalcheminformatics.blogspot.com/2022/12/identifying-ring-systems-in-molecules.html> (2026).
190. Yang, B., Xiang, C., Li, T., Xu, Y. & Li, J. Structure-Based Generation of 3D Small-Molecule Drugs: Are We There Yet? en. *Journal of Medicinal Chemistry* **68**, 22756–22768. ISSN: 0022-2623, 1520-4804. <https://pubs.acs.org/doi/10.1021/acs.jmedchem.5c01706> (2026) (Nov. 2025).
191. Baell, J. B. & Holloway, G. A. New Substructure Filters for Removal of Pan Assay Interference Compounds (PAINS) from Screening Libraries and for Their Exclusion in Bioassays. en. *Journal of Medicinal Chemistry* **53**, 2719–2740. ISSN: 0022-2623, 1520-4804. <https://pubs.acs.org/doi/10.1021/jm901137j> (2026) (Apr. 2010).
192. Brenk, R. *et al.* Lessons Learnt from Assembling Screening Libraries for Drug Discovery for Neglected Diseases. en. *ChemMedChem* **3**, 435–444. ISSN: 1860-7179, 1860-7187. <https://chemistry-europe.onlinelibrary.wiley.com/doi/10.1002/cmdc.200700139> (2026) (Mar. 2008).

193. Yu, Y. *et al.* Uni-Dock: GPU-Accelerated Docking Enables Ultralarge Virtual Screening. en. *Journal of Chemical Theory and Computation* **19**, 3336–3345. ISSN: 1549-9618, 1549-9626. <https://pubs.acs.org/doi/10.1021/acs.jctc.2c01145> (2026) (June 2023).
194. Takahashi, M. *et al.* Crystal structure of the mineralocorticoid receptor ligand-binding domain in complex with a potent and selective nonsteroidal blocker, esaxerenone (CS-3150). en. *FEBS Letters* **594**, 1615–1623. ISSN: 0014-5793, 1873-3468. <https://onlinelibrary.wiley.com/doi/10.1002/1873-3468.13746> (2026) (May 2020).
195. Békés, M., Langley, D. R. & Crews, C. M. PROTAC targeted protein degraders: the past is prologue. en. *Nature Reviews Drug Discovery* **21**, 181–200. ISSN: 1474-1784. <https://www.nature.com/articles/s41573-021-00371-6> (2026) (Mar. 2022).
196. Ge, J. *et al.* PROTAC-DB 3.0: an updated database of PROTACs with extended pharmacokinetic parameters. *Nucleic Acids Research* **53**, D1510–D1515. ISSN: 1362-4962. <http://dx.doi.org/10.1093/nar/gkae768> (Sept. 2024).
197. Testa, A., Hughes, S. J., Lucas, X., Wright, J. E. & Ciulli, A. Structure-Based Design of a Macrocyclic PROTAC. en. *Angewandte Chemie International Edition* **59**, 1727–1734. ISSN: 1433-7851, 1521-3773. <https://onlinelibrary.wiley.com/doi/10.1002/anie.201914396> (2026) (Jan. 2020).
198. Home — *openbind.uk* en-US. <https://openbind.uk/> (2026).
199. Santos-Martins, D. *et al.* Meeko: Molecule Parametrization and Software Interoperability for Docking and Beyond. en. *Journal of Chemical Information and Modeling* **65**, 13045–13050. ISSN: 1549-9596, 1549-960X. <https://pubs.acs.org/doi/10.1021/acs.jcim.5c02271> (2026) (Dec. 2025).
200. Abraham, M. J. *et al.* GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **1-2**, 19–25. ISSN: 2352-7110. <https://www.sciencedirect.com/science/article/pii/S2352711015000059> (2026) (Sept. 2015).
201. Maier, J. A. *et al.* ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. en. *Journal of Chemical Theory and Computation* **11**, 3696–3713. ISSN: 1549-9618, 1549-9626. <https://pubs.acs.org/doi/10.1021/acs.jctc.5b00255> (2026) (Aug. 2015).
202. 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**, 1157–1174. ISSN: 0192-8651, 1096-987X. <https://onlinelibrary.wiley.com/doi/10.1002/jcc.20035> (2026) (July 2004).
203. 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).

204. 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).
205. 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).
206. 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).
207. 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).
208. Muegge, I. PMF scoring revisited. *J. Med. Chem.* **49**, 5895–5902 (2006).
209. 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).
210. Dittrich, J., Schmidt, D., Pfleger, C. & Gohlke, H. Converging a knowledge-based scoring function: DrugScore2018. *J. Chem. Info. Model.* **59**, 509–521 (2018).
211. Somnath, V. R., Bunne, C. & Krause, A. Multi-scale representation learning on proteins. *Advances in Neural Information Processing Systems* **34**, 25244–25255 (2021).
212. 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).
213. 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).
214. 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).
215. 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).
216. 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).

- 217. Liu, Z. *et al.* PDB-wide collection of binding data: current status of the PDBbind database. *Bioinformatics* **31**, 405–412 (2015).
- 218. Liu, Z. *et al.* Forging the basis for developing protein–ligand interaction scoring functions. *Accounts of chemical research* **50**, 302–309 (2017).
- 219. 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).
- 220. 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).
- 221. 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).
- 222. Liu, T. *et al.* BindingDB in 2024: a FAIR knowledgebase of protein-small molecule binding data. *Nucleic acids research* **53**, D1633–D1644 (2025).
- 223. 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).
- 224. 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).
- 225. 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).
- 226. 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).
- 227. Durairaj, J. *et al.* PLINDER: The protein-ligand interactions dataset and evaluation resource. *bioRxiv*, 2024–07 (2024).
- 228. 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).
- 229. 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).
- 230. Guo, X.-K. & Zhang, Y. CovBinderInPDB: A Structure-Based Covalent Binder Database. *J. Chem. Info. Model.* **62**, 6057–6068 (2022).
- 231. 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>.

232. 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).
233. O’Boyle, N. M. *et al.* Open Babel: An open chemical toolbox. *Journal of cheminformatics* **3**, 1–14 (2011).
234. Luo, W. *et al.* Bridging Machine Learning and Thermodynamics for Accurate p K a Prediction. *JACS Au* **4**, 3451–3465 (2024).
235. 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).
236. *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.
237. 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).
238. 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).
239. Fermi, E. *Thermodynamics* (Courier Corporation, 2012).
240. MarÉchal, E. Measuring bioactivity: KI, IC50 and EC50. *Chemogenomics and Chemical Genetics: A User’s Introduction for Biologists, Chemists Informaticians*, 55–65 (2011).
241. Ross, G. A. *et al.* The maximal and current accuracy of rigorous protein-ligand binding free energy calculations. *Communications Chemistry* **6**, 222 (2023).
242. 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).
243. 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).
244. Delbart, F. *et al.* An allosteric binding site of the $\alpha 7$ 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).
245. 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).

246. 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. *Chem-BioChem* **7**, 515–526. ISSN: 1439-4227, 1439-7633. <https://chemistry-europe.onlinelibrary.wiley.com/doi/10.1002/cbic.200500452> (2024) (Mar. 2006).
247. 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/doi/10.1038/nsb0797-539> (2024) (July 1997).
248. 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).
249. 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).
250. 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).
251. 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).
252. 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).
253. 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).
254. Baek, M. *et al.* Accurate prediction of protein structures and interactions using a three-track neural network. en. *Science* **373**, 871–876. ISSN: 0036-8075, 1095-9203. <https://www.science.org/doi/10.1126/science.abj8754> (2025) (Aug. 2021).
255. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. en. *Nature* **596**, 583–589. ISSN: 1476-4687. <https://www.nature.com/articles/s41586-021-03819-2> (2025) (Aug. 2021).

256. Mirdita, M. *et al.* ColabFold: making protein folding accessible to all. en. *Nature Methods* **19**, 679–682. ISSN: 1548-7105. <https://www.nature.com/articles/s41592-022-01488-1> (2025) (June 2022).
257. Lin, Z. *et al.* Evolutionary-scale prediction of atomic-level protein structure with a language model. en. *Science* **379**, 1123–1130. ISSN: 0036-8075, 1095-9203. <https://www.science.org/doi/10.1126/science.ade2574> (2025) (Mar. 2023).
258. Fang, X. *et al.* A method for multiple-sequence-alignment-free protein structure prediction using a protein language model. en. *Nature Machine Intelligence* **5**, 1087–1096. ISSN: 2522-5839. <https://www.nature.com/articles/s42256-023-00721-6> (2025) (Oct. 2023).
259. Ahdriz, G. *et al.* OpenFold: retraining AlphaFold2 yields new insights into its learning mechanisms and capacity for generalization. en. *Nature Methods* **21**, 1514–1524. ISSN: 1548-7105. <https://www.nature.com/articles/s41592-024-02272-z> (2025) (Aug. 2024).
260. Berman, H. M. The Protein Data Bank. *Nucleic Acids Research* **28**, 235–242. ISSN: 13624962. <https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/28.1.235> (2026) (Jan. 2000).
261. Evans, R. *et al.* Protein complex prediction with AlphaFold-Multimer en. Oct. 2021. <http://biorxiv.org/lookup/doi/10.1101/2021.10.04.463034> (2025).
262. Gao, M., Nakajima An, D., Parks, J. M. & Skolnick, J. AF2Complex predicts direct physical interactions in multimeric proteins with deep learning. en. *Nature Communications* **13**, 1744. ISSN: 2041-1723. <https://www.nature.com/articles/s41467-022-29394-2> (2025) (Apr. 2022).
263. Krishna, R. *et al.* Generalized biomolecular modeling and design with RoseTTAFold All-Atom. en. *Science* **384**, eadl2528. ISSN: 0036-8075, 1095-9203. <https://www.science.org/doi/10.1126/science.adl2528> (2025) (Apr. 2024).
264. ByteDance AML AI4Science Team *et al.* Protenix - Advancing Structure Prediction Through a Comprehensive AlphaFold3 Reproduction en. Jan. 2025. <http://biorxiv.org/lookup/doi/10.1101/2025.01.08.631967> (2025).
265. Zhang, T.-h. *et al.* Benchmarking co-folding methods to predict the structures of covalent protein–ligand complexes. en. *Acta Pharmacologica Sinica*, 1–11. ISSN: 1745-7254. <https://www.nature.com/articles/s41401-025-01721-5> (2026) (Jan. 2026).
266. Dunlop, N., Erazo, F., Jalalypour, F. & Mercado, R. Predicting PROTAC-mediated ternary complexes with AlphaFold3 and Boltz-1. en. *Digital Discovery* **4**, 3782–3809. <https://pubs.rsc.org/en/content/articlelanding/2025/dd/d5dd00300h> (2026) (2025).

267. Liao, Y., Zhu, J., Xie, J., Lai, L. & Pei, J. Benchmarking Cofolding Methods for Molecular Glue Ternary Structure Prediction. en. *Journal of Chemical Information and Modeling* **65**, 11136–11148. ISSN: 1549-9596, 1549-960X. <https://pubs.acs.org/doi/10.1021/acs.jcim.5c01860> (2026) (Oct. 2025).
268. Nittinger, E., Yoluk, O., Tibo, A., Olanders, G. & Tyrchan, C. Co-folding, the future of docking – prediction of allosteric and orthosteric ligands. *Artificial Intelligence in the Life Sciences* **8**, 100136. ISSN: 2667-3185. <https://www.sciencedirect.com/science/article/pii/S2667318525000121> (2026) (Dec. 2025).
269. Wei, H. & McCammon, J. A. Structure and dynamics in drug discovery. en. *npj Drug Discovery* **1**, 1. ISSN: 3005-1452. <https://www.nature.com/articles/s44386-024-00001-2> (2026) (Nov. 2024).
270. McClendon, C. L., Kornev, A. P., Gilson, M. K. & Taylor, S. S. Dynamic architecture of a protein kinase. eng. *Proceedings of the National Academy of Sciences of the United States of America* **111**, E4623–4631. ISSN: 1091-6490 (Oct. 2014).
271. Taylor, S. S. & Kornev, A. P. Protein kinases: evolution of dynamic regulatory proteins. eng. *Trends in Biochemical Sciences* **36**, 65–77. ISSN: 0968-0004 (Feb. 2011).
272. Taylor, S. S. *et al.* From structure to the dynamic regulation of a molecular switch: A journey over 3 decades. eng. *The Journal of Biological Chemistry* **296**, 100746. ISSN: 1083-351X (2021).
273. The UniProt Consortium *et al.* UniProt: the Universal Protein Knowledgebase in 2025. en. *Nucleic Acids Research* **53**, D609–D617. ISSN: 0305-1048, 1362-4962. <https://academic.oup.com/nar/article/53/D1/D609/7902999> (2026) (Jan. 2025).
274. Faezov, B. & Dunbrack, R. L. *AlphaFold2 models of the active form of all 437 catalytically competent human protein kinase domains* en. July 2023. <http://biorxiv.org/lookup/doi/10.1101/2023.07.21.550125> (2025).
275. Modi, V. & Dunbrack, R. L. Kincore: a web resource for structural classification of protein kinases and their inhibitors. en. *Nucleic Acids Research* **50**, D654–D664. ISSN: 0305-1048, 1362-4962. <https://academic.oup.com/nar/article/50/D1/D654/6395339> (2025) (Jan. 2022).
276. Faezov, B. & Dunbrack, R. L. *AlphaFold2 models of the active form of all 437 catalytically competent human protein kinase domains* en. July 2023. <http://biorxiv.org/lookup/doi/10.1101/2023.07.21.550125> (2025).
277. Manning, G., Whyte, D. B., Martinez, R., Hunter, T. & Sudarsanam, S. The Protein Kinase Complement of the Human Genome. en. *Science* **298**, 1912–1934. ISSN: 0036-8075, 1095-9203. <https://www.science.org/doi/10.1126/science.1075762> (2026) (Dec. 2002).

278. Śmiech, M., Leszczyński, P., Kono, H., Wardell, C. & Taniguchi, H. Emerging BRAF Mutations in Cancer Progression and Their Possible Effects on Transcriptional Networks. *Genes* **11**, 1342. ISSN: 2073-4425. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7697059/> (2026) (Nov. 2020).
279. Kufareva, I. & Abagyan, R. Type-II Kinase Inhibitor Docking, Screening, and Profiling Using Modified Structures of Active Kinase States. en. *Journal of Medicinal Chemistry* **51**, 7921–7932. ISSN: 0022-2623, 1520-4804. <https://pubs.acs.org/doi/10.1021/jm8010299> (2026) (Dec. 2008).
280. Vijayan, R. S. K. *et al.* Conformational Analysis of the DFG-Out Kinase Motif and Biochemical Profiling of Structurally Validated Type II Inhibitors. en. *Journal of Medicinal Chemistry* **58**, 466–479. ISSN: 0022-2623, 1520-4804. <https://pubs.acs.org/doi/10.1021/jm501603h> (2026) (Jan. 2015).
281. Zhao, Z. *et al.* Exploration of Type II Binding Mode: A Privileged Approach for Kinase Inhibitor Focused Drug Discovery? en. *ACS Chemical Biology* **9**, 1230–1241. ISSN: 1554-8929, 1554-8937. <https://pubs.acs.org/doi/10.1021/cb500129t> (2026) (June 2014).
282. Lesburg, C. *HPK1 IN COMPLEX WITH COMPOUND 1: 7m0m* en. Apr. 2021. https://www.wwpdb.org/pdb?id=pdb_00007m0m (2026).
283. Kalaivani, R., Narwani, T. J., De Brevern, A. G. & Srinivasan, N. Long-range molecular dynamics show that inactive forms of Protein Kinase A are more dynamic than active forms. en. *Protein Science* **28**, 543–560. ISSN: 0961-8368, 1469-896X. <https://onlinelibrary.wiley.com/doi/10.1002/pro.3556> (2026) (Mar. 2019).
284. 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).
285. 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).
286. 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).
287. 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).
288. 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).

289. 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).
290. 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).
291. 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).
292. 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).
293. 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).
294. Lockbaum, G. J. *et al.* Crystal structure of SARS-CoV-2 main protease in complex with non-covalent inhibitor ML188. *Viruses* **13**, 174 (2021).
295. 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).
296. Günther, S. *et al.* X-ray screening identifies active site and allosteric inhibitors of SARS-CoV-2 MPro. *Science* **372**, 642–646 (2021).
297. Drayman, N. *et al.* Masitinib is a broad coronavirus 3CL inhibitor that blocks replication of SARS-CoV-2. *Science* **373**, 931–936 (2021).
298. Kitamura, N. *et al.* Expedited approach toward rational design of noncovalent SARS-CoV-2 MPro inhibitors. *J. Med. Chem.* **65**, 2848–2865 (2021).
299. 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).
300. Glaser, J. *et al.* Hit expansion of a noncovalent SARS-CoV-2 main protease inhibitor. *ACS Pharm. Trans. Sci.* **5**, 255–265 (2022).
301. Deshmukh, M. G. *et al.* Structure-guided design of perampanel-derived pharmacophore targeting SARS-CoV-2 main protease. *Structure* **29**, 823–833 (2021).
302. Hou, N. *et al.* Development of highly potent noncovalent inhibitors of SARS-CoV-2 3CLpro. *ACS central science* **9**, 217–227 (2023).

303. 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).
304. 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).
305. 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).
306. 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).
307. 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).
308. 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).
309. Moghadas, 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).
310. 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).
311. 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).
312. 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).
313. 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).

314. 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).
315. 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).
316. 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).
317. 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).
318. 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).
319. 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).
320. Wohllwend, J. *et al.* Boltz-1 Democratizing Biomolecular Interaction Modeling en. Nov. 2024. <http://biorxiv.org/lookup/doi/10.1101/2024.11.19.624167> (2025).
321. 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).
322. 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).
323. 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).
324. 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).

325. 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).
326. 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).
327. 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. *npj Viruses* **2**, 1–20. ISSN: 2948-1767. <https://www.nature.com/articles/s44298-024-00028-2> (2025) (June 2024).
328. 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.
329. 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).
330. 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).
331. 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).
332. 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).
333. 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).
334. 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).
335. 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).
336. Brown, N., Fiscato, M., Segler, M. H. & Vaucher, A. C. GuacaMol: benchmarking models for de novo molecular design. *J. Chem. Inform. Model.* **59**, 1096–1108 (2019).

337. Yuan, E. C. -. *et al.* *Foundation Models for Atomistic Simulation of Chemistry and Materials* 2025. arXiv: 2503.10538 [physics.chem-ph]. <https://arxiv.org/abs/2503.10538>.
338. Li, Y., Zhang, L. & Liu, Z. Multi-objective de novo drug design with conditional graph generative model. *Journal of Cheminformatics* **10**, 33. ISSN: 1758-2946. <https://jcheminf.biomedcentral.com/articles/10.1186/s13321-018-0287-6> (2025) (Dec. 2018).
339. Levin, I., Fortunato, M. E., Tan, K. L. & Coley, C. W. Computer-aided evaluation and exploration of chemical spaces constrained by reaction pathways. *AIChE Journal* **69**, e18234. ISSN: 0001-1541, 1547-5905. <https://aiche.onlinelibrary.wiley.com/doi/10.1002/aic.18234> (2025) (Dec. 2023).
340. Gómez-Bombarelli, R. *et al.* Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. *ACS central science* **4**, 268–276. ISSN: 2374-7943 (2018).
341. Paul, S. M. *et al.* How to improve R&D productivity: the pharmaceutical industry's grand challenge. *Nature Reviews Drug Discovery* **9**, 203–214. ISSN: 1474-1776, 1474-1784. <https://www.nature.com/articles/nrd3078> (2025) (2010).
342. Nicolaou, K. C. Advancing the Drug Discovery and Development Process. *Angewandte Chemie* **126**, 9280–9292. ISSN: 0044-8249, 1521-3757. <https://onlinelibrary.wiley.com/doi/10.1002/ange.201404761> (2025) (2014).
343. Shaw, J. W., Horrace, W. C. & Vogel, R. J. The Determinants of Life Expectancy: An Analysis of the OECD Health Data. *Southern Economic Journal* **71**, 768–783. ISSN: 0038-4038, 2325-8012. <https://onlinelibrary.wiley.com/doi/10.1002/j.2325-8012.2005.tb00675.x> (2025) (2005).
344. Lichtenberg, F. R. Sources of U.S. longevity increase, 1960–2001. *The Quarterly Review of Economics and Finance* **44**, 369–389. ISSN: 10629769. <https://linkinghub.elsevier.com/retrieve/pii/S1062976904000377> (2025) (2004).
345. Lichtenberg, F. R. The effect of new drug approvals on HIV mortality in the US, 1987–1998. *Economics & Human Biology* **1**, 259–266. ISSN: 1570677X. <https://linkinghub.elsevier.com/retrieve/pii/S1570677X0200031X> (2025) (2003).
346. Weisfeldt, M. L. & Zieman, S. J. Advances In The Prevention And Treatment Of Cardiovascular Disease. *Health Affairs* **26**, 25–37. ISSN: 0278-2715, 1544-5208. <http://www.healthaffairs.org/doi/10.1377/hlthaff.26.1.25> (2025) (2007).
347. Grootendorst, P., Piérard, E. & Shim, M. Life-expectancy gains from pharmaceutical drugs: a critical appraisal of the literature. *Expert Review of Pharmacoeconomics & Outcomes Research* **9**, 353–364. ISSN: 1473-7167, 1744-8379. <http://www.tandfonline.com/doi/full/10.1586/erp.09.35> (2025) (2009).

348. Buxbaum, J. D., Chernew, M. E., Fendrick, A. M. & Cutler, D. M. Contributions Of Public Health, Pharmaceuticals, And Other Medical Care To US Life Expectancy Changes, 1990-2015: Study examines the conditions most responsible for changing US life expectancy and how public health, pharmaceuticals, other medical care, and other factors may have contributed to the changes. *Health Affairs* **39**, 1546–1556. ISSN: 0278-2715, 1544-5208. <http://www.healthaffairs.org/doi/10.1377/hlthaff.2020.00284> (2025) (2020).
349. Sadybekov, A. A. *et al.* Synthon-based ligand discovery in virtual libraries of over 11 billion compounds. *Nature* **601**, 452–459. ISSN: 1476-4687. <https://doi.org/10.1038/s41586-021-04220-9> (2022).
350. Gimeno, A. *et al.* The Light and Dark Sides of Virtual Screening: What Is There to Know? *Int J Mol Sci* **20**. ISSN: 1422-0067 (2019).
351. Scior, T. *et al.* Recognizing Pitfalls in Virtual Screening: A Critical Review. *Journal of Chemical Information and Modeling* **52**, 867–881. ISSN: 1549-9596. <https://doi.org/10.1021/ci200528d> (2012).
352. Liu, F. *et al.* The impact of library size and scale of testing on virtual screening. *Nature Chemical Biology*. ISSN: 1552-4469. <https://doi.org/10.1038/s41589-024-01797-w> (2025).
353. Reymond, J.-L. The Chemical Space Project. *Accounts of Chemical Research* **48**, 722–730. ISSN: 0001-4842. <https://doi.org/10.1021/ar500432k> (2015).
354. Gironda-Martínez, A., Donckele, E. J., Samain, F. & Neri, D. DNA-Encoded Chemical Libraries: A Comprehensive Review with Successful Stories and Future Challenges. *ACS Pharmacology & Translational Science* **4**, 1265–1279. <https://doi.org/10.1021/acsptsci.1c00118> (2021).
355. Grygorenko, O. O. Enamine Ltd.: The Science and Business of Organic Chemistry and Beyond. *European Journal of Organic Chemistry* **2021**, 6474–6477. ISSN: 1434-193X. <https://doi.org/10.1002/ejoc.202101210> (2021).
356. Satz, A. L. *et al.* DNA-encoded chemical libraries. *Nature Reviews Methods Primers* **2**, 3. ISSN: 2662-8449. <https://doi.org/10.1038/s43586-021-00084-5> (2022).
357. Lyu, J. *et al.* Ultra-large library docking for discovering new chemotypes. *Nature* **566**, 224–229. ISSN: 1476-4687. <https://doi.org/10.1038/s41586-019-0917-9> (2019).
358. Shultz, M. D. Two Decades under the Influence of the Rule of Five and the Changing Properties of Approved Oral Drugs. *Journal of Medicinal Chemistry* **62**. PMID: 30212196, 1701–1714. eprint: <https://doi.org/10.1021/acs.jmedchem.8b00686>. <https://doi.org/10.1021/acs.jmedchem.8b00686> (2019).

359. Lipinski, C. A., Lombardo, F., Dominy, B. W. & Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced Drug Delivery Reviews* **23**. In Vitro Models for Selection of Development Candidates, 3–25. ISSN: 0169-409X. <https://www.sciencedirect.com/science/article/pii/S0169409X96004231> (1997).
360. Coley, C. W., Green, W. H. & Jensen, K. F. Machine Learning in Computer-Aided Synthesis Planning. en. *Accounts of Chemical Research* **51**, 1281–1289. ISSN: 0001-4842, 1520-4898. <https://pubs.acs.org/doi/10.1021/acs.accounts.8b00087> (2025) (May 2018).
361. *USPTO dataset* http://figshare.com/articles/dataset/Chemical_reactions_from_US_patents_1976-Sep2016_/5104873. accessed on July, 2025.
362. *Reaxys* <https://www.reaxys.com/>. accessed on July, 2025.
363. *Open Reaction Database* <https://open-reaction-database.org/>. accessed on July, 2025.
364. *IBM RXN for Chemistry* <https://rxn.app.accelerate.science/rxn>. accessed on July, 2025.
365. Lin, K., Xu, Y., Pei, J. & Lai, L. Automatic retrosynthetic route planning using template-free models. en. *Chemical Science* **11**, 3355–3364. ISSN: 2041-6539. <https://pubs.rsc.org/en/content/articlelanding/2020/sc/c9sc03666k> (2025) (Mar. 2020).
366. Liu, Z. *et al.* Forging the Basis for Developing Protein–Ligand Interaction Scoring Functions. *Accounts of Chemical Research* **50**, 302–309. ISSN: 0001-4842. <https://doi.org/10.1021/acs.accounts.6b00491> (2017).
367. 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. ISSN: 1868-1743. <https://doi.org/10.1002/minf.201400132> (2015).
368. 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].
369. 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. <https://doi.org/10.1021/jz500737m> (2014).
370. Haghighatlari, M. *et al.* Learning to Make Chemical Predictions: The Interplay of Feature Representation, Data, and Machine Learning Methods. *Chem* **6**, 1527–1542. <https://doi.org/10.1016/j.chempr.2020.05.014> (2020).

- 371. Needleman, S. B. & Wunsch, C. D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Bio.* **48**, 443–453 (1970).
- 372. Bickerton, G. R., Paolini, G. V., Besnard, J., Muresan, S. & Hopkins, A. L. Quantifying the chemical beauty of drugs. *Nature chemistry* **4**, 90–98 (2012).
- 373. Gohlke, H., Hendlich, M. & Klebe, G. Knowledge-based scoring function to predict protein-ligand interactions. *J. Mol. Bio.* **295**, 337–356 (2000).
- 374. Mooij, W. T. & Verdonk, M. L. General and targeted statistical potentials for protein–ligand interactions. *Proteins: Struct., Funct., and Bioinform.* **61**, 272–287 (2005).
- 375. Shen, C. *et al.* Accuracy or novelty: what can we gain from target-specific machine-learning-based scoring functions in virtual screening? *Briefings in Bioinformatics* **22**, bbaa410 (2021).
- 376. Li, Y. & Yang, J. Structural and sequence similarity makes a significant impact on machine-learning-based scoring functions for protein–ligand interactions. *J. Chem. Info. Model.* **57**, 1007–1012 (2017).
- 377. Li, H. *et al.* The impact of protein structure and sequence similarity on the accuracy of machine-learning scoring functions for binding affinity prediction. *Biomolecules* **8**, 12 (2018).
- 378. Pushpakom, S. *et al.* Drug repurposing: progress, challenges and recommendations. *Nature reviews Drug discovery* **18**, 41–58 (2019).
- 379. Zervosen, A., Sauvage, E., Frère, J.-M., Charlier, P. & Luxen, A. Development of new drugs for an old target—the penicillin binding proteins. *Mol.* **17**, 12478–12505 (2012).
- 380. Newman, A. H., Ku, T., Jordan, C. J., Bonifazi, A. & Xi, Z.-X. New drugs, old targets: tweaking the dopamine system to treat psychostimulant use disorders. *Annual review of pharmacology and toxicology* **61**, 609–628 (2021).
- 381. Zhang, L. *et al.* Crystal structure of SARS-CoV-2 main protease provides a basis for design of improved α -ketoamide inhibitors. *Science* **368**, 409–412 (2020).
- 382. Herbst, R. S. Review of epidermal growth factor receptor biology. *International Journal of Radiation Oncology* Biology* Physics* **59**, S21–S26 (2004).
- 383. Gao, K. *et al.* Perspectives on SARS-CoV-2 main protease inhibitors. *J. Med. Chem.* **64**, 16922–16955 (2021).
- 384. Nelder, J. A. & Mead, R. A simplex method for function minimization. *The computer journal* **7**, 308–313 (1965).
- 385. Rogers, D. & Hahn, M. Extended-connectivity fingerprints. *J. Chem. Info. Model.* **50**, 742–754 (2010).
- 386. Dice, L. R. Measures of the amount of ecologic association between species. *Ecology* **26**, 297–302 (1945).

387. Unoh, Y. *et al.* Discovery of S-217622, a noncovalent oral SARS-CoV-2 3CL protease inhibitor clinical candidate for treating COVID-19. *J. Med. Chem.* **65**, 6499–6512 (2022).
388. Yun, C.-H. *et al.* Structures of lung cancer-derived EGFR mutants and inhibitor complexes: mechanism of activation and insights into differential inhibitor sensitivity. *Cancer cell* **11**, 217–227 (2007).
389. Zhu, S.-J. *et al.* Structural insights into drug development strategy targeting EGFR T790M/C797S. *Oncotarget* **9**, 13652–13665. <https://api.semanticscholar.org/CorpusID:4090455> (2018).
390. Yan, X.-E. *et al.* Structural basis of mutant-selectivity and drug-resistance related to CO-1686. *Oncotarget* **8**, 53508 (2017).
391. Peng, Y.-H. *et al.* Protein kinase inhibitor design by targeting the Asp-Phe-Gly (DFG) motif: the role of the DFG motif in the design of epidermal growth factor receptor inhibitors. *J. Med. Chem.* **56**, 3889–3903 (2013).
392. Hanan, E. J. *et al.* 4-Aminoindazolyl-dihydrofuro [3, 4-d] pyrimidines as non-covalent inhibitors of mutant epidermal growth factor receptor tyrosine kinase. *Bioorg. Med. Chem. Lett.* **26**, 534–539 (2016).
393. Sogabe, S. *et al.* Structure-based approach for discovery of pyrrolo [3, 2-d] pyrimidine-based EGFR T790M/L858R mutant inhibitors. *ACS Med. Chem. Lett.* **4**, 201–205 (2013).
394. Stamos, J., Sliwkowski, M. X. & Eigenbrot, C. Structure of the epidermal growth factor receptor kinase domain alone and in complex with a 4-anilinoquinazoline inhibitor. *J. Bio. Chem.* **277**, 46265–46272 (2002).
395. Bryan, M. C. *et al.* Pyridones as highly selective, noncovalent inhibitors of T790M double mutants of EGFR. *ACS Med. Chem. Lett.* **7**, 100–104 (2016).
396. Peng, Y.-H. *et al.* Protein kinase inhibitor design by targeting the Asp-Phe-Gly (DFG) motif: the role of the DFG motif in the design of epidermal growth factor receptor inhibitors. *J. Med. Chem.* **56**, 3889–3903 (2013).
397. Aertgeerts, K. *et al.* Structural analysis of the mechanism of inhibition and allosteric activation of the kinase domain of HER2 protein. *J. Bio. Chem.* **286**, 18756–18765 (2011).
398. Planken, S. *et al.* Discovery of n-((3 r, 4 r)-4-fluoro-1-(6-((3-methoxy-1-methyl-1 h-pyrazol-4-yl) amino)-9-methyl-9 h-purin-2-yl) pyrrolidine-3-yl) acrylamide (pf-06747775) through structure-based drug design: A high affinity irreversible inhibitor targeting oncogenic egfr mutants with selectivity over wild-type egfr. *J. Med. Chem.* **60**, 3002–3019 (2017).
399. Yun, C.-H. *et al.* The T790M mutation in EGFR kinase causes drug resistance by increasing the affinity for ATP. *Proc. Natl. Acad. Sci.* **105**, 2070–2075 (2008).

- 400. Xu, G. *et al.* Discovery of novel 4-amino-6-arylaminopyrimidine-5-carbaldehyde oximes as dual inhibitors of EGFR and ErbB-2 protein tyrosine kinases. *Bioorg. Med. Chem. Lett.* **18**, 3495–3499 (2008).
- 401. Yosaatmadja, Y. *et al.* Binding mode of breakthrough inhibitor AZD9291 to EGF receptor revealed. *J. Struct. Bio.* **192**, 539–544 (2015).
- 402. Wood, E. R. *et al.* A unique structure for epidermal growth factor receptor bound to GW572016 (Lapatinib) relationships among protein conformation, inhibitor off-rate, and receptor activity in tumor cells. *Cancer research* **64**, 6652–6659 (2004).
- 403. Jacobs, J. *et al.* Discovery, synthesis, and structure-based optimization of a series of N-(tert-butyl)-2-(N-arylamido)-2-(pyridin-3-yl) acetamides (ML188) as potent noncovalent small molecule inhibitors of the severe acute respiratory syndrome coronavirus (SARS-CoV) 3CL protease. *J. Med. Chem.* **56**, 534–546 (2013).
- 404. 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).
- 405. 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).
- 406. 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.
- 407. Li, H. *et al.* Classical scoring functions for docking are unable to exploit large volumes of structural and interaction data. *Bioinformatics* **35**, 3989–3995 (2019).
- 408. Wang, D. D., Xie, H. & Yan, H. Proteo-chemometrics interaction fingerprints of protein–ligand complexes predict binding affinity. *Bioinformatics* **37**, 2570–2579 (2021).
- 409. 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).
- 410. Francoeur, P. G. *et al.* Three-dimensional convolutional neural networks and a cross-docked data set for structure-based drug design. *Journal of chemical information and modeling* **60**, 4200–4215 (2020).
- 411. Paszke, A. *et al.* in *Advances in Neural Information Processing Systems 32* 8024–8035 (Curran Associates, Inc., 2019). <http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf>.
- 412. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **12**, 2825–2830 (2011).
- 413. Pham, T. N. H. *et al.* Improving ligand-ranking of AutoDock Vina by changing the empirical parameters. *Journal of Computational Chemistry* **43**, 160–169 (2022).

- 414. Li, J., Chen, Q., Weng, J., Wu, J. & Xu, X. *Canonized then Minimized RMSD for Three-Dimensional Structures* 2024. arXiv: 2405.00339 [physics.chem-ph].
- 415. Hastie, T., Tibshirani, R., Friedman, J. H. & Friedman, J. H. *The elements of statistical learning: data mining, inference, and prediction* (Springer, 2009).
- 416. Kullback, S. & Leibler, R. A. On information and sufficiency. *The annals of mathematical statistics* **22**, 79–86 (1951).
- 417. Cutler, A. & Zhao, G. PERT-perfect random tree ensembles. *Computing Science and Statistics* **33**, 90–4 (2001).
- 418. Belkin, M., Hsu, D. J. & Mitra, P. Overfitting or perfect fitting? risk bounds for classification and regression rules that interpolate. *Advances in neural information processing systems* **31** (2018).
- 419. Gorantla, R., Kubincova, A., Weiße, A. Y. & Mey, A. S. From Proteins to Ligands: Decoding Deep Learning Methods for Binding Affinity Prediction. *Journal of Chemical Information and Modeling* (2023).
- 420. Volkov, M. *et al.* On the frustration to predict binding affinities from protein–ligand structures with deep neural networks. *Journal of medicinal chemistry* **65**, 7946–7958 (2022).
- 421. 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).
- 422. Cock, P. J. *et al.* Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422 (2009).
- 423. Bajusz, D., Rácz, A. & Héberger, K. Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *Journal of cheminformatics* **7**, 1–13 (2015).
- 424. Castilho, M. S., Postigo, M. P., Pereira, H. M., Oliva, G. & Andricopulo, A. D. Structural basis for selective inhibition of purine nucleoside phosphorylase from *Schistosoma mansoni*: Kinetic and structural studies. en. *Bioorganic & Medicinal Chemistry* **18**, 1421–1427. ISSN: 09680896. <https://linkinghub.elsevier.com/retrieve/pii/S096808961000043X> (2024) (Feb. 2010).
- 425. Wang, Y. *et al.* *HiQBind Dataset* Nov. 2024. <https://doi.org/10.6084/m9.figshare.27430305>.
- 426. Xiao, Y. *et al.* Phosphorylation releases constraints to domain motion in ERK2. en. *Proceedings of the National Academy of Sciences* **111**, 2506–2511. ISSN: 0027-8424, 1091-6490. <https://pnas.org/doi/full/10.1073/pnas.1318899111> (2026) (Feb. 2014).

427. Li, Y. & Nam, K. Dynamic, structural and thermodynamic basis of insulin-like growth factor 1 kinase allostery mediated by activation loop phosphorylation. en. *Chemical Science* **8**, 3453–3464. ISSN: 2041-6520, 2041-6539. <https://xlink.rsc.org/?DOI=C7SC00055C> (2026) (2017).
428. Ruff, E. F. *et al.* A dynamic mechanism for allosteric activation of Aurora kinase A by activation loop phosphorylation. en. *eLife* **7**, e32766. ISSN: 2050-084X. <https://elifesciences.org/articles/32766> (2026) (Feb. 2018).
429. Sultan, M. M., Denny, R. A., Unwalla, R., Lovering, F. & Pande, V. S. Millisecond dynamics of BTK reveal kinome-wide conformational plasticity within the apo kinase domain. en. *Scientific Reports* **7**, 15604. ISSN: 2045-2322. <https://www.nature.com/articles/s41598-017-10697-0> (2026) (Nov. 2017).
430. Pucheta-Martínez, E. *et al.* An Allosteric Cross-Talk Between the Activation Loop and the ATP Binding Site Regulates the Activation of Src Kinase. en. *Scientific Reports* **6**, 24235. ISSN: 2045-2322. <https://www.nature.com/articles/srep24235> (2026) (Apr. 2016).
431. Gough, N. R. & Kalodimos, C. G. Exploring the conformational landscape of protein kinases. en. *Current Opinion in Structural Biology* **88**, 102890. ISSN: 0959440X. <https://linkinghub.elsevier.com/retrieve/pii/S0959440X24001179> (2026) (Oct. 2024).
432. Steichen, J. M. *et al.* Global Consequences of Activation Loop Phosphorylation on Protein Kinase A. en. *Journal of Biological Chemistry* **285**, 3825–3832. ISSN: 00219258. <https://linkinghub.elsevier.com/retrieve/pii/S0021925820810539> (2026) (Feb. 2010).
433. Xie, T., Saleh, T., Rossi, P. & Kalodimos, C. G. Conformational states dynamically populated by a kinase determine its function. en. *Science* **370**, eabc2754. ISSN: 0036-8075, 1095-9203. <https://www.science.org/doi/10.1126/science.abc2754> (2025) (Oct. 2020).
434. Palhano Zanela, T. M., Woudenberg, A., Romero Bello, K. G. & Underbakke, E. S. Activation loop phosphorylation tunes conformational dynamics underlying Pyk2 tyrosine kinase activation. en. *Structure* **31**, 447–454.e5. ISSN: 09692126. <https://linkinghub.elsevier.com/retrieve/pii/S0969212623000370> (2026) (Apr. 2023).
435. Masters, M. R., Mahmoud, A. H. & Lill, M. A. Investigating whether deep learning models for co-folding learn the physics of protein-ligand interactions. en. *Nature Communications* **16**, 8854. ISSN: 2041-1723. <https://www.nature.com/articles/s41467-025-63947-5> (2026) (Oct. 2025).
436. Vara, B. A. *et al.* Discovery of Diaminopyrimidine Carboxamide HPK1 Inhibitors as Preclinical Immunotherapy Tool Compounds. en. *ACS Medicinal Chemistry Letters* **12**, 653–661. ISSN: 1948-5875, 1948-5875. <https://pubs.acs.org/doi/10.1021/acsmchemlett.1c00096> (2026) (Apr. 2021).

- 437. Pillaiyar, T., Manickam, M., Namasivayam, V., Hayashi, Y. & Jung, S.-H. An Overview of Severe Acute Respiratory Syndrome–Coronavirus (SARS-CoV) 3CL Protease Inhibitors: Peptidomimetics and Small Molecule Chemotherapy. en. *Journal of Medicinal Chemistry* **59**, 6595–6628. ISSN: 0022-2623, 1520-4804. <https://pubs.acs.org/doi/10.1021/acs.jmedchem.5b01461> (2025) (July 2016).
- 438. De Wit, E., Van Doremalen, N., Falzarano, D. & Munster, V. J. SARS and MERS: recent insights into emerging coronaviruses. en. *Nature Reviews Microbiology* **14**, 523–534. ISSN: 1740-1526, 1740-1534. <https://www.nature.com/articles/nrmicro.2016.81> (2025) (Aug. 2016).
- 439. Hou, T., Wang, J., Li, Y. & Wang, W. Assessing the performance of the molecular mechanics/Poisson Boltzmann surface area and molecular mechanics/generalized Born surface area methods. II. The accuracy of ranking poses generated from docking. *Journal of computational chemistry* **32**, 866–877 (2011).
- 440. Friesner, R. A. *et al.* Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy. en. *Journal of Medicinal Chemistry* **47**, 1739–1749. ISSN: 0022-2623, 1520-4804. <https://pubs.acs.org/doi/10.1021/jm0306430> (2025) (Mar. 2004).