

Deep learning-based protein–ligand docking

Martin Buttenschön

The Queen's College, University of Oxford



A thesis submitted for the degree of
Doctor of Philosophy
in the Department of Statistics

Michaelmas term 2025



Abstract

This thesis investigates deep learning (DL) approaches for docking small molecules into protein receptors, a central challenge in structure-based drug discovery. While DL has revolutionised protein structure prediction, DL-based docking methods have yet to achieve a comparable breakthrough. My work demonstrates how better evaluation, data, and modelling assumptions might close this gap.

First, I introduced PoseBusters, a test suite and an accompanying benchmark, both developed by me, that define chemical consistency and physical plausibility checks; protein-ligand complexes and structures that pass all tests are said to be “PB-valid”. Applying PoseBusters to DL-based and classical docking methods showed that, once physical and stereochemical validity was required, current DL methods did not outperform classical docking, and that a simple molecular mechanics energy minimisation step could correct failures, indicating that important docking-relevant physics has not been learnt by DL models.

I also showed that commonly used time-based train–test splits overestimate generalisation capabilities; stratification by receptor sequence, pocket similarity, ligand chemistry, and interaction patterns would give a more realistic view of model performance. Using PoseBusters to analyse widely used training and test sets revealed that the training data often contained systematic structural and stereochemical problems sharing some failure modes with the models, whereas curated benchmarks were substantially cleaner. However, training data issues alone could not explain the observed error rates.

Next, I mined the Protein Data Bank for pairs of similar ligands bound to the same protein pocket to assess the hypothesis that the structure of a smaller molecule can be used to guide the prediction of the structure of a larger elaborated molecule. The data show that binding-mode conservation statistically increases as the chemical similarity of the ligands increases; this relationship was particularly strong for elaborations, confirming the experimental basis for both fragment-based and drug discovery and providing substructure constraints.

Last, I present GuideDock, which augments the diffusion-based docking model DiffDock-L with physics- and knowledge-based guidance. By incorporating force-field terms alongside user-specified pocket, substructure, and interaction constraints, GuideDock enables pocket-specific and constrained docking while improving physical plausibility and accuracy under substructure constraints.

These studies provide a comprehensive assessment of DL-based protein–ligand docking, highlight the central role of robust metrics and carefully curated data, and show how diffusion guidance can inject physical priors and user knowledge into generative docking models, thus informing the development of a new class of protein–ligand docking and co-folding methods.

Meiner Mutter und meinem Vater.

Acknowledgements

I am deeply grateful to my academic supervisors, Charlotte Deane, Garrett Morris, and Michael Bronstein, and to my industrial supervisor, Giorgio Tamo, for their guidance, encouragement, and support throughout my DPhil. I especially thank Garrett and Charlotte for welcoming me into the fields of computational structural biology and cheminformatics, and Giorgio for sharing insights into real-world drug discovery. These disciplines sit at the intersection of many of the subjects I have long loved: mathematics, physics, chemistry, and biology. As a result, I find the pursuit of drug discovery and the scientific questions surrounding how life works at the molecular level to be profoundly rewarding. I am grateful that this degree has enabled me to become a research scientist able to contribute to society, whether through developing medicines that improve human health or advancing our scientific understanding of life.

I am also grateful to my transfer of status examiners, Paul Brennan and Michael Bronstein; my confirmation of status examiners, Yee Whye Teh and Jakob Foerster; and my viva examiners, Fernanda Duarte and Mohammed AlQuraishi, for their time, insight, and constructive engagement with this work. I thoroughly enjoyed the scientific discussions we shared and am thankful for their encouragement and perspective throughout these milestones.

I thank my colleagues at the Oxford Protein Informatics Group and the Department of Statistics for creating such a welcoming, social, and enjoyable environment. From punting and cycling pub trips to shared formals and summer garden parties, they made Oxford a wonderful place to work and live. In particular, I thank Tom, Mark, Fergus, Brennan, Leo, Tobias, Steph, Fabian, Dan, Carlos, Oliver, Anna, Matteo, Rubén, Andrew, Lars, Yael, Saifuddin, Henry, and Nic for countless helpful conversations, much good humour, and for making the work even more enjoyable.

I gratefully acknowledge The Queen's College, the Department of Statistics, and the University of Oxford for providing such a stimulating and supportive environment in which to carry out this research. I also acknowledge the generous financial support of The Queen's College, Bristol Myers Squibb, and the McKinsey & Company education leave programme.

Beyond my research, I have greatly enjoyed life in the City of Dreaming Spires. I am especially grateful to The Queen's College for welcoming me into its friendly community and providing so many opportunities to make lasting friendships. I particularly thank the boat club; although my rowing career was distinguished chiefly by winning spoons, the camaraderie of the team more than compensated for any lack of athletic achievement. I also have fond memories of the time spent on the tennis courts and croquet lawns, at chess tournaments, and the Town and Gown races. Oxford will remain a treasured memory: late summer evenings watching garden plays, dancing at college balls, and the tranquillity of the beautiful gardens and parks at all times of day. My years here have been

immensely fulfilling, and this thesis reflects what has been a truly wonderful chapter of my life.

I thank my friends Michael, Gabriel, Krzysztof, Felix, Johannes, Luis, James, and Wojtek: thank you for your friendship and patience throughout my prolonged stay in academia. Having now matched my years at university with those spent at school, I trust that I have finally acquired sufficient qualifications. Most of all, I thank Emelia, who read this thesis at least three times, and whose love, patience, encouragement, and unwavering support have meant more to me than I can adequately express.

I thank my brothers Andreas and Jonas for setting a daunting precedent by collecting doctoral degrees before me. I could probably have done without one myself, but I was reluctant to remain the only member of the family without the requisite academic accolades.

Vor allem danke ich meinen Eltern. Sie haben in mir und meinen Brüdern die Neugier gefördert und uns damit die Welt eröffnet. Von ihnen habe ich gelernt, dass Lernen zu den größten Freuden des Lebens gehören kann. Von meinem Vater habe ich die Überzeugung geerbt, dass sich mit Geduld und Beharrlichkeit nahezu alles verstehen, bauen, reparieren oder lösen lässt. Ob in der Fürsorge für seine Patienten, beim Anbau von Tomaten, bei der Reparatur des Autos oder als er eigenhändig einen Teil des Daches öffnete, um größere Dachfenster einzubauen; jeder Herausforderung begegnete er mit Entschlossenheit und der Bereitschaft dazuzulernen. Von meiner Mutter habe ich die Neugier auf die Welt und die Liebe zu den Menschen in ihr. Ob sie Patienten unterschiedlichster kultureller Herkunft mit Würde und Respekt begegnete, eine neue Sprache lernte, um mit Verwandten oder den Mitarbeitenden eines spanischen Baumarkts sprechen zu können, oder sich mit unermüdlicher Fürsorge um unsere Familie kümmerte – stets zeigte sie Offenheit für Neues und Mitgefühl für andere. Gemeinsam haben meine Eltern uns vorgelebt, dass das Leben ein Abenteuer ist. Sie füllten unser Zuhause mit Büchern, Ideen, Ermutigung und unzähligen Möglichkeiten, die Welt zu entdecken – in meinem Falle auch auf eigene Faust in Quebec, BC, Saskatchewan, Spanien, Brasilien, und der Schweiz. Diese Dissertation ist daher nicht zuletzt Ausdruck des Fundaments, das sie gelegt haben, und der Werte, die sie uns mit auf den Weg gegeben haben. Von ganzem Herzen danke ich ihnen.

Contents

1	Introduction	1
1.1	Structure-based drug discovery	1
1.2	Docking	3
1.3	Deep learning	7
1.4	Motivation	9
1.5	Structure of the thesis	11
2	Benchmarking the models	13
2.1	Introduction	14
2.2	Methods	16
2.2.1	Docking methods	16
2.2.2	The PoseBusters test suite	19
2.2.3	Quality of fit	22
2.2.4	Sequence identity	23
2.2.5	Molecular mechanics energy minimisation	23
2.2.6	Data	23
2.3	Results	24
2.3.1	Results on the Astex Diverse set	24
2.3.2	Results on the PoseBusters Benchmark set	28
2.3.3	Results with post-docking energy minimisation	30
2.4	Discussion	30
2.5	Conclusion	33
3	Evaluating the training data	35
3.1	Introduction	35
3.2	Methods	38
3.2.1	Data sets	38
3.2.2	Assessment	39
3.3	Results	40
3.4	Discussion	43
3.5	Conclusion	46
4	Mining the PDB for guidance	47
4.1	Introduction	47
4.2	Methods	49
4.2.1	Selection and preparation of protein-ligand complexes	49
4.2.2	Pairing and alignment	50
4.2.3	Clustering by sequence identity and ligand similarity	50
4.2.4	Measurements and annotations	51
4.3	Results	54
4.3.1	Correlation between 2D and 3D similarity	54
4.3.2	Error analysis on identical pairs	54
4.3.3	Clustering by sequence identity and ligand similarity	56
4.3.4	Fragment binding mode conservation	56

4.3.5	Binding mode conservation upon elaboration	57
4.3.6	Beyond elaborations using maximum common substructures	59
4.3.7	Interactions replace colour and shape score	61
4.4	Discussion	61
4.5	Conclusion	66
5	Guiding diffusion for docking	67
5.1	Introduction	67
5.2	Methods	69
5.2.1	Base model	69
5.2.2	Guidance	70
5.2.3	Loss functions	72
5.2.4	Experiments	73
5.2.5	Data	74
5.2.6	Metrics	75
5.2.7	Baselines	75
5.2.8	Tuning	76
5.3	Results	77
5.3.1	Physical plausibility in re-docking	77
5.3.2	Binding site-specific re-docking	77
5.3.3	Cross-docking into holo structures	77
5.3.4	Substructure constrained docking	81
5.4	Discussion	81
5.5	Conclusion	83
6	Conclusion	85
6.1	Key findings and contributions	85
6.2	Limitations and recommendations for future research	87
6.3	Closure	88
	References	91
A	Technical background	107
A.1	Cheminformatics	107
A.1.1	String representations of molecules	107
A.1.2	Molecular fingerprints	107
A.1.3	Root mean square deviation	107
A.1.4	Resolution of molecular structures	108
A.2	Deep learning	108
A.2.1	Supervised machine learning	108
A.2.2	Neural networks	108
A.2.3	Convolutional networks	109
A.2.4	Graph neural networks	109
A.2.5	Geometric graph neural networks	109
A.2.6	Transformers	110
A.2.7	Diffusion models	110
B	Detailed methods and further results for Chapter 2	111
B.1	Software and data availability	111
B.2	Docking protocols	111
B.3	PoseBusters Benchmark set procurement	113
B.4	Comparison and description of the data sets	115
B.5	Cofactor prevalence	116
B.6	PDB and CCD identifiers of data sets	116
B.7	Overall performance of each method	118
B.8	Analysis of the severity of test failures	120
B.9	Alternative binding site definitions for Uni-Mol	133

C	Further results for Chapter 3	135
C.1	Failure modes of the PoseBusters test set	135
C.2	Failure modes of the Astex Diverse set	136
C.3	Failure modes of the DockGen test set	137
C.4	Failure modes of the CASF-2016 test set	138
C.5	Failure modes of the PDBbind General training set	142
D	Extended methods and further results for Chapter 4	151
D.1	CCD Block list	151
D.2	Interaction comparison and interaction statistics	153
D.3	MCS comparison with chemical features	155
D.4	Illustrations of conserved and non-conserved pairs of poses	156
E	Technical preliminaries and further results for Chapter 5	159
E.1	Score-based diffusion models	159
E.2	DiffDock-L benchmark	160
E.3	Further experimental results	160
E.4	Further details on data sets	163
E.4.1	Selection of pairs of PDB structures for the GuideDock data set	163
E.4.2	Malhotra data set updates	163
E.4.3	DockGen data set updates	164
E.4.4	Malhotra data set identifiers	164
E.4.5	GuideDock data set identifiers	164
E.4.6	Ligand block list	164

Chapter 1

Introduction

I present a study of deep learning (DL) approaches for docking small molecules into protein receptors. This is a mainstay computational problem in structure-based drug discovery, but surprisingly, recent advances in DL techniques have not yet achieved their projected potential for docking. This chapter introduces structure-based drug discovery (Section 1.1), docking (Section 1.2), and DL (Section 1.3), as well as the motivation for this work (Section 1.4) and the structure of the thesis (Section 1.5).

1.1 Structure-based drug discovery

Structure-based drug discovery is a paradigm for developing molecules, which when small are often called “ligands”, that bind to other molecules, which are called “targets” or “receptors”, using knowledge of the three-dimensional structures of both molecules. The central idea is that binding requires the ligand’s shape and chemical features to be complementary to the receptor. Whether the resulting binding event produces the intended therapeutic effect that makes this compound a drug candidate is determined by the biological target involved, the binding site of the ligand, and the strength of their interaction, in addition to a range of pharmacological constraints (Patrick, 2023, ch. 1). The ligand’s target specificity, in turn, influences its potential to cause side effects. Understanding the structural characteristics of target molecules and how to design ligands with complementary features unique to a selected target is a cornerstone of rational drug discovery.

Structure-based drug discovery is only possible because we know the structures of many biomolecules. Thanks to the continuous improvement of the experimental techniques X-ray crystallography, nuclear magnetic resonance spectroscopy, and cryogenic electron microscopy over the past 80 years, we have the structures of increasingly large and biologically important molecules—from small compounds such as penicillin (Crowfoot et al., 1949, 17 atoms) to peptides such as insulin (M. J. Adams et al., 1969, 788 atoms) and even large proteins such as the 2019-nCoV S protein (Wrapp

et al., 2020, 4634 atoms). Access to these structures has led to breakthroughs in drug discovery and production, elucidation of the mechanisms of action, and the very field of rational drug design (Hägg & Crowfoot Hodgkin, 1999).

Small molecules, namely organic molecules with up to 100 atoms, which can be administered orally, are a very important and the most used drug modality, and proteins are the most common targets (Ritter, 2024). Newer drug modalities such as macrocycles, peptides, or antibodies (Valeur et al., 2017) and other targets such as RNA and DNA provide other avenues for drug discovery. For my thesis I focus on small molecule drugs that target proteins.

Currently, the two most important repositories of experimentally determined structures for proteins and small molecules are the Protein Data Bank (PDB) (Berman et al., 2000) and the Cambridge Structural Database (CSD) (F. H. Allen et al., 1991) respectively. The PDB serves as a *de facto* legal deposit library for protein structures, since most journals require authors to submit structural data upon publication, making it the largest source of protein and protein-complex structures. The CSD similarly aggregates crystal structures from researchers worldwide, with a strong emphasis on crystallographic data, and contains nearly all known drugs as well as a wide range of drug-like molecules. In December 2025, the PDB and the CSD contained 246 715 and 1 376 589 unique, experimentally resolved structures, respectively.¹

A drug must be produced, stored, and administered. It must reach its target in a patient and provide a therapeutic benefit with minimal side-effects (Young, 2009, ch. 2). Because other important properties of drugs have been directly linked to the structures of small molecules, structure-based drug discovery has become a multi-objective optimisation in which favourable interactions are only one criterion. Examples of computational proxies that link structure to pharmacological properties are estimates of physicochemical parameters such as prediction of $\log P$ as an estimate of lipophilicity (Wildman & Crippen, 1999), Lipinski’s ‘rule of 5’ for oral administration effectiveness (Lipinski et al., 1997), and drug-likeness scores (Bickerton et al., 2012). As we shall see in the next section, docking methods are designed to model molecular interactions and do not help optimise pharmacological properties beyond binding (Young, 2009, ch. 12).

Beyond biological properties, there are other chemical and legal limitations that affect drug design. Our current chemical knowledge limits which molecules can be synthesised and tested, often delineated by synthetic accessibility estimators (Ertl & Schuffenhauer, 2009), supplier catalogues, synthesis planning tools (Segler et al., 2018), and expert chemists. Patents further limit the design space of small molecules, as one would sensibly exclude all structures patented by others. Lastly, economic, social, and practical constraints, while also important in drug development, are not directly linked to

¹Statistics retrieved on 17 December 2025 from <https://www.rcsb.org/stats> for the PDB and <https://www.ccdc.cam.ac.uk/solutions/about-the-csd/csd-statistics/> for the CSD.

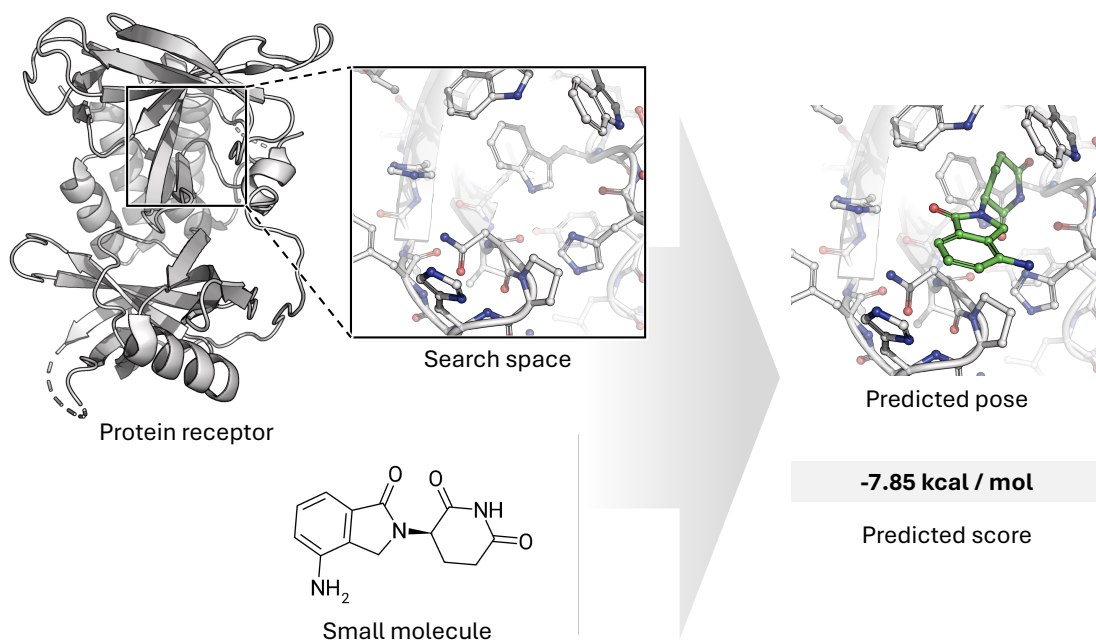


Figure 1.1: Illustration of the inputs and outputs of docking. Given the 3D structure of a protein receptor, a description or starting conformation of the ligand, and a search space, docking methods predict the pose of the protein-ligand complex and a score. Some methods' scores approximate the binding affinity. The structure shown here is a crystal structure of the small molecule drug lenalidomide bound to the protein cereblon-midi (PDB 8RQA) (Kroupova et al., 2024). Docking methods ideally predict the same structure as observed in experimental methods.

structure-based drug discovery, although the favourable mode of administration, *i.e.* orally administered, the relatively low price of treatment, and long shelf life are certainly important aspects of small molecule drugs.

1.2 Docking

Docking is a computational approach to predict the binding mode of a ligand when bound to a receptor. Docking algorithms require as input a 3D structure of the receptor, a description of the ligand, and a search space. Given these inputs, docking algorithms generate as output the 3D structure of the receptor-ligand binding complex (Figure 1.1).

The first input, the 3D receptor structure, should be as close as possible to the *holo* (ligand-bound) form of the protein. Using the *apo* (unbound) form may be appropriate when the structural differences between the *apo* and *holo* states are small. For highly flexible proteins in which the side chains or the backbone undergo conformational changes upon ligand binding, the *holo* form of the receptor bound to a related ligand may be a more suitable starting point. In practice, PDB structures are often used, but care must be taken to prepare them, for example by selecting plausible protonation states of the side

chains in the binding site (Bender et al., 2021). In certain cases, the use of multiple receptor structures is advantageous, and the use of snapshots generated from molecular dynamics simulations has been demonstrated (Dunn et al., 2024).

The second input, the ligand molecule, is typically provided with a 3D starting conformation. Ideally, an unbiased conformation of a ligand is used, one that has not undergone, for example, the minimisation procedure used to generate experimental protein-ligand structures, which would affect the ligand's bond lengths, angles, and rings. The PDB provides idealised versions of ligand structures as part of the Chemical Component Dictionary (Westbrook et al., 2015). Some docking software allow as input 1D string representations (see Appendix A, Section A.1.1) because the software includes a conformation generation module that can generate or regenerate 3D coordinates. Most docking software requires the user to carefully prepare the ligand by defining the stereochemistry, charges, protonation, and tautomeric state; these are all properties that are typically modelled as fixed during docking (Torres et al., 2019).

The third input, the search space restricts the set of solutions that a docking algorithm can output. Typically, ligand placement is restricted to a defined region of the receptor that encompasses a binding site. In some cases, the search space is expanded to include the entire protein to perform blind docking. The search space may be reduced by positional constraints on specific ligand substructures and enlarged by allowing some movement of protein side chains. Most targets do not undergo large conformational changes and modelling only side-chain flexibility is sufficient (J. J. Clark et al., 2019).

The two primary outputs of docking methods are the predicted ligand coordinates and an associated docking score. In principle, a single predicted pose may correspond to any of the many microstates between which a ligand binding mode fluctuates; in practice, however, docking results are typically evaluated by comparison with experimentally determined structures (Paggi et al., 2024). Most docking tools, therefore, report a set of ligand poses paired with corresponding scores. These scores are used by docking programmes to rank poses of the same ligand and externally by users to prioritise a variety of molecular candidates. Some methods report scores in the units of binding affinity (kcal/mol); for example, AutoDock Vina (Trott & Olson, 2010). However, docking-derived affinity estimates are generally considered too inaccurate to reliably predict true binding affinities (Pantsar & Poso, 2018), although improving their accuracy remains an active area of research (G. Durant et al., 2025). In this thesis, I focus on pose generation.

The main applications of docking are the discovery of potentially new ligands, the optimisation of lead compounds, and the investigation of the mode of action (Paggi et al., 2024). Lead discovery by virtual screening with docking proceeds by predicting the binding poses of a (large) set of molecules from a library, measuring the best fit of each molecule, and then ranking the candidates and selecting the top ones as leads. In ligand optimisation, medicinal chemists inspect lead compounds and their

fit with the receptor in order to propose modifications to the molecular structure with the goal of optimising the binding strength or other molecular properties, often iteratively. For target function analysis, ligands known to elicit biological responses are docked and the differences in the target structure are analysed.

One specialised application of docking explored later in this thesis is the use of docking with constraints in fragment-based drug discovery. Fragments, ligands with a molecular mass of less than 300 Da, can be used to explore the binding site of a protein effectively, and can be combined into regular sized ligands by growing, linking, and merging (AlKharboush et al., 2025). Docking with substructure constraints derived from fragments can assist in designing ligand elaborations by helping to determine the fit of the added substructures.

Other more specialised applications of docking related to structure-based drug discovery include fitting ligands to electron densities (Roseman, 2000), uncovering reaction mechanisms of enzymes (Das et al., 2019), and studying host-guest systems (Casbarra & Procacci, 2021). Dedicated software packages, for example PHENIX (P. D. Adams et al., 2010), which provide pipelines for experimental structure solution, include docking routines to place substructures of the model onto electron maps (Liebschner et al., 2019). Improvements in docking algorithms therefore impact many aspects of structural biology, and not just drug discovery directly.

The three main requirements for docking algorithms are speed, accuracy, and generalisability. Speed is needed because virtual screening libraries can be very large, with up to a billion molecules for ‘ultra-large screens’ (Gorgulla et al., 2020); the requirement for speed is further exacerbated by the need to consider multiple structural snapshots of the target when docking into a rigid receptor. Accuracy in terms of recovering the true binding mode is important for decision making in drug discovery (Warren et al., 2006) and to avoid taking “duds” further down the drug discovery pipeline. Generalisation to new protein targets and new ligands is a must because docking for drug discovery, by definition, must predict unknown protein-ligand interactions.

Classical docking algorithms typically break the problem of predicting binding modes into two subroutines: sampling and scoring (Young, 2009, ch. 12). The sampling subroutine generates coordinates for the ligand atoms, and the scoring subroutine grades these coordinates. Often there is a feedback loop from the scoring to the sampling routine.

In the last 43 years, more than 100 docking methods have been developed (Pagadala et al., 2017), and some examples are listed in Table 1.1. The large number of docking methods reflect the diversity of computational approaches that have been applied to the docking problem: Monte Carlo simulated annealing is a stochastic method from physics that is used for docking in the AutoDock series (Goodsell & Olson, 1990; Morris et al., 1998, 2009; Trott & Olson, 2010). Genetic algorithms are stochastic search algorithms that are used in GOLD (Jones et al., 1997) and AutoDock 2 (Morris et al., 1998). Fragment-

Table 1.1: Some classical docking methods. Since the publication of the eponymous DOCK in 1982 more than 100 docking methods have been published. With 31 methods this list is incomplete and only provides a short overview. The list illustrates the diversity in methods and the importance of the docking problem.

Docking program	Creators	Year
DOCK	Kuntz et al.	1982
AutoDock	Goodsell and Olson	1990
HOOK	Eisen et al.	1993
DIVALI	K. P. Clark and Ajay	1995
Genetic algorithm-based docking	Jones et al.	1995
AutoDock 2.4	Morris et al.	1996
FlexX	Rarey et al.	1996
Hammerhead	Welch et al.	1996
AutoDock 3	Morris et al.	1998
Flexible docking with Tabu-search	Baxter et al.	1998
MCDOCK	M. Liu and Wang	1999
SEED	Majeux et al.	1999
EUDOC	Pang et al.	2001
Gaussian docking function	M. R. McGann et al.	2003
rDock	L. Li et al.	2003
GOLD	Verdonk et al.	2003
Surflex-Dock	Jain	2003
ZDOCK	R. Chen et al.	2003
Glide	Friesner et al.	2004
MolDock	Thomsen and Christensen	2006
RosettaLigand	Meiler and Baker	2006
AutoDock 4	Morris et al.	2009
AutoDock Vina	Trott and Olson	2009
FRED	M. McGann	2011
SwissDock (EADock)	Grosdidier et al.	2011
MOE-Dock	Corbeil et al.	2012
DOCK 3.7	Coleman et al.	2013
LeDock	Zhao and Caflisch	2013
Smina	Koes et al.	2013
DOCK 6	W. J. Allen et al.	2015
POSIT	Kelley et al.	2015

based incremental construction builds solutions step-wise starting from small fragments and is used in HOOK (Eisen et al., 1994), FlexX (Rarey et al., 1996), and SEED (Majeux et al., 1999). Systematic search approaches exhaustively enumerate poses by discretising the search space and are used by Hammerhead (Welch et al., 1996), EUDOC (Pang et al., 2001), and in combination with pruning also by Glide (Friesner et al., 2004). The diversity of published approaches highlights the amount of algorithmic experimentation done and the maturity of classical docking methods.

Despite the large number of docking approaches, docking as a paradigm remains limited in some important ways. Being structure-based, the methods require a receptor structure as input. Rigid docking, that is assuming a rigid receptor structure, is the most common and most reliable approach. Flexible docking, allowing some movement of the receptor, is slow and often does not perform better than rigid docking (Paggi et al., 2024). Further limitations are that the protonation and charge patterns of the ligand and receptor are fixed during the docking (Petukh et al., 2013) and water molecules are only selectively modelled (Forli & Olson, 2012). Despite these limitations, docking has been and remains an important tool in structure-based drug design.

Possible improvements of docking methods are faster predictions, better control, and richer models. The first, faster predictions, is for docking ligands into flexible receptors which currently often requires some use of molecular dynamics or docking to a variety of crystal structures of the same target. The second, better control, is the methods' flexibility for users to test hypotheses, around, for example, the binding site, substructure positions, and desired interactions (*e.g.* H-bond, covalent). The last category, richer models, includes improved modelling of waters, receptor flexibility, charges, protonation, and tautomeric states. These factors are known to be important in protein-ligand binding and docking programs should more accurately model them.

1.3 Deep learning

Deep learning (DL) is a subfield of machine learning. Machine learning is an approach to build functions, called models, that do not require a programmer to implement all rules by hand. Rather, the programmer defines parameters that are left to be inferred from data in a process called "training". Once the parameters have been determined by training, the model can be used like a regular function.

Supervised learning is one approach to train a machine learning model. Given pairs of input and desired output, the model is trained to reproduce these data (see Appendix A, Section A.2.1). The goal is to find parameters that both reproduce the training data and lead to sensible outputs for new, previously unseen inputs that were not shown during training, so that the model can be used to make predictions. If a model can make satisfactory predictions on new data points, it generalises; a model that does not generalise is said to be overfitting to its training data (Goodfellow et al., 2016, ch. 5).

To assess generalisation when building a model, machine learning engineers split the example data into a training set and a test set. The training set is used to find the parameters of the model and the test set is used to assess the generalisability of the model. If there are modelling decisions that cannot be made by the programmer or the model has parameters that cannot be learnt, so-called “hyper-parameters”, then a third split, the validation set, is used to select these

To quantify the agreement between a predicted output and ground truth data, a loss function must be specified for supervised training. Given a data point and the model’s prediction, it returns a positive real number, the loss. A loss of 0 means that the two data points (actual and predicted) are identical. The definition of a loss function is an important consideration in supervised machine learning; an underspecified loss function can lead to surprising failure modes of the model. Often, loss functions combine multiple objectives; for example, by adding different loss components linearly to obtain one main loss. The loss function is essential because it defines what constitutes good performance on a task.

DL is the sub-set of machine learning that uses artificial neural networks with a large number of hidden layers to learn patterns directly from data. These models typically have a large number of parameters (see Appendix A, Section A.2.2). DL typically does not rely on human-chosen data representations but uses raw representations of the input to directly learn patterns from the data, including intermediate representations at every layer (Goodfellow et al., 2016, ch. 1).

Neural networks are trained by making small adjustments to the parameters iteratively. A typical training algorithm involves setting the initial parameter values according to a random initialisation scheme, for example, Xavier initialisation (Glorot & Bengio, 2010). Then batches of the data (inputs and outputs) are drawn from the training set. The model makes predictions on the inputs, and these predictions are compared to the ground truth outputs using the loss function. The back-propagation algorithm (Rojas, 1996), using the chain rule of calculus, is then used to calculate how much each parameter affects the loss function. These derivatives are then used to update the parameters, using, for example, gradient descent (Bishop & Bishop, 2024, ch. 7) or more sophisticated stochastic optimisation methods such as Adam (Kingma & Ba, 2014). This procedure is repeated by passing over all the training data multiple times; one pass is called an “epoch”.

Neural networks are highly flexible and, in principle, can approximate any function with enough neurones and hidden layers (Funahashi, 1989; Hornik et al., 1989). This is particularly useful for non-tabular data for which few other universal modelling approaches exist. However, this universality does not mean that DL models can be used to solve any problem. Indeed, no learning algorithm is, in general, the best, when considering all possible problems (Wolpert, 1996); what makes neural networks so useful is that they contain important inductive biases, for example smoothness, that are useful for a large set of relevant problems (Bishop & Bishop, 2024, ch. 9). In practice, it is possible to include other problem-specific inductive biases such as translation invariance or permutation equivariance,

and that makes DL so useful. However, the main downside of DL models is that they require large amounts of training data to learn from.

In many domains, including structural biology, the input data have an intrinsic structure. For example, proteins consist of a small set of identical amino acids connected in sequence, forming chains that mostly fold up in repeating patterns. Understanding the internal structure of the data, the necessary degrees of freedom of a task, and task-specific mathematics have led to new inductive biases, *e.g.* triangle inequality inspired triangular updates (Jumper et al., 2021), which have led to breakthroughs. Choosing a DL framework that follows the structure of the data generating process constitutes an inductive bias that helps a model learn a task more efficiently and using fewer data. Other examples of common network architectures with strong inductive biases that we will encounter later include convolutional networks, graph neural networks, and diffusion models. Section A.2 provides technical background on these and other DL-based models that have been applied to docking.

A particularly successful DL model architecture is the transformer (Bahdanau et al., 2015; Vaswani et al., 2017). It is suitable for all types of data, for example sequential data in conjunction with positional encodings (see Appendix A, Section A.2.6). The transformer has become a universal building block, and many tasks that were solved with more dedicated architectures have now been solved with transformers. The two main advantages of transformers are that they are well suited for parallel processing and that they are effective at transfer learning. This has led some researchers to state that, like search, DL represents a paradigm shift and that training (a transformer) with sufficient data is sufficient to build a supervised model for any task, and even reach artificial general intelligence (Bishop & Bishop, 2024, ch. 12). Whether this is true or not is not yet known. I can only note that DL-based modelling is still relatively young in comparison to purely mathematical or physics-based modelling.

1.4 Motivation

Deep learning (DL) techniques have led to major breakthroughs in protein structure prediction, recognised by the award of the 2024 Nobel Prize in Chemistry (Nobel Foundation, 2024). John Jumper and Demis Hassabis were honoured for developing AlphaFold 2 (Jumper et al., 2021), a DL-based method capable of generating highly accurate models of protein structures. AlphaFold 2 won the fourteenth biennial Critical Assessment of Structure Prediction competition (CASP14) in 2020 (Jumper et al., 2021), and two years later, DL-based methods inspired by AlphaFold 2 dominated CASP15, establishing DL as the dominant method in protein structure prediction (Kryshtafovych et al., 2023).

These advances in computational protein structure prediction are significant for structure-based drug discovery because they can provide structures of more potential drug targets. While the experimentally determined protein structures in the PDB cover about 35 % of the genes of the human genome

Table 1.2: List of early deep learning-based docking methods. GNINA was the first method to use a DL model, a convolutional neural network, in its scoring function. The first to use DL for pose generation was DeepDock. With 20 methods this list is incomplete and only provides a short overview. The list illustrates a renewed interest in docking and the promise that DL holds for docking.

Docking program	Creators	Year
NNScore	Durrant and McCammon	2010
NNScore 2.0	Durrant and McCammon	2011
GNINA	McNutt et al.	2021
DeepDock	Méndez-Lucio et al.	2021
TankBind	Lu et al.	2022
EquiBind	Stärk et al.	2022
DiffDock	Corso et al.	2023
DiffDock-Site	Guo et al.	2023
DiffDock-Pocket	Plainer et al.	2023
FusionDock	Masters et al.	2023
KarmaDock	X. Zhang et al.	2023
PLANTAIN	Brocidiacono et al.	2023
Uni-Mol	Zhou et al.	2023
DeltaDock	Yan et al.	2024
DiffDock-L	Corso et al.	2024
DynamicBind	Lu et al.	2024
QuickBind	Treyde et al.	2024
DockFormer	Shor and Schneidman-Duhovny	2025
FlowDock	Morehead and Cheng	2025
SurfDock	Cao et al.	2025

(Tunyasuvunakool et al., 2021), the computationally generated structures in the AlphaFold Data Base (Varadi et al., 2024) cover more than 98 % of human proteins. This increase unlocks many new targets, for example the computational structure of glucose-6-phosphatase protein is amenable to structure-based drug discovery and Tunyasuvunakool et al. (2021) speculate that a ligand that binds to a pocket near the protein’s binding face may help regulate blood sugar levels.

The aforementioned success of DL in protein structure prediction and the implications for structure-based drug discovery motivated me to investigate DL applied in docking. My guiding hypothesis is that DL-based docking methods are still early on the innovation S-curve (Asthana, 1995), and are less mature than the incumbent methods in functionality and performance, but also that the technology has the potential to replace the incumbent methods. The first successful application of DL techniques applied to docking that I am aware of is the DL-based scoring function in NNScore (Durrant & McCammon, 2010) that uses a convolutional neural network and takes as input a voxel grid. The first approaches that used DL also for pose generation include DeepDock (Méndez-Lucio et al., 2021) that learnt a statistical potential. Table 1.2 lists twenty DL-based methods, four of which had been published when I started this work in 2022. The large number of publications of DL-based docking approaches over the following years demonstrates not only the timeliness of my work but also, in general, the field’s belief that DL could have a large impact on docking. However, none of the approaches published by 2022 had the level of impact on docking that AlphaFold 2 had on protein structure prediction—and this marks the starting point of my thesis.

1.5 Structure of the thesis

In Chapter 2, I describe a benchmark of classical and DL-based docking methods to investigate their failure modes and make suggestions for their improvement. In Chapter 3, I assess the quality of the training and test data sets on which the DL-based docking methods have been trained in an attempt to understand the failure modes of the models relative to the data. In Chapter 4, I use experimental data to measure to what extent similar ligands adopt similar poses upon binding. In Chapter 5, I explore diffusion guidance using molecular mechanics force field-inspired terms to fix the observed failure modes of a leading DL-based docking method, and I add further terms to constrain substructures and interactions using the insights I gathered in Chapter 4. Lastly, Chapter 6 concludes my thesis and introduces the new paradigm of co-folding methods and provides an outlook on DL-based techniques for small molecule docking.

Chapter 2

Benchmarking the models

As I discussed in Chapter 1, the last few years have seen the development of numerous deep learning (DL)-based protein-ligand docking methods. If AlphaFold 2 is any guide, they offer huge promise in terms of speed and accuracy. However, despite claims of state-of-the-art performance in terms of root-mean-square deviation (RMSD) to experimentally determined poses, upon closer inspection, I found that they often produce physically implausible molecular structures. It is therefore not sufficient to evaluate these methods solely by RMSD to a crystallographic native binding mode. It is vital, particularly for DL-based methods, that they are also evaluated in terms of steric and energetic criteria. I present PoseBusters (Buttenschoen et al., 2024), a Python package that performs a series of standard quality checks, built using the well-established cheminformatics toolkit RDKit (Landrum et al., 2025).

The PoseBusters test suite validates chemical and geometric consistency of a ligand including its stereochemistry, and the physical plausibility of intra- and intermolecular measurements such as the planarity of aromatic rings, standard bond lengths, and protein-ligand clashes. Only methods that both pass these stereochemical and energetic checks *and* predict native-like binding modes should be classed as having “state-of-the-art” performance.

I used PoseBusters to compare five DL-based docking methods: (i) DeepDock (Méndez-Lucio et al., 2021); (ii) DiffDock (Corso et al., 2023); (iii) EquiBind (Stärk et al., 2022); (iv) TankBind (Lu et al., 2022); and (v) Uni-Mol (Zhou et al., 2023); and two well-established standard docking methods: (i) AutoDock Vina (Trott & Olson, 2010), and (ii) CCDC GOLD (Verdonk et al., 2003). I also consider their performance with and without an additional post-prediction energy minimisation step using a

The chapter is based on my paper: Buttenschoen, M., Morris, G. M., & Deane, C. M. (2024). PoseBusters: AI-based docking methods fail to generate physically valid poses or generalise to novel sequences. *Chemical Science*, 15(9), 3130–3139. <https://doi.org/10.1039/D3SC04185A> I performed all the computational experiments, created the PoseBusters software, and wrote the manuscript.

molecular mechanics force field. I will show that in terms of both physical plausibility and the ability to generalise to examples that are distinct from the training data, none of the DL-based methods I tested outperformed classical docking tools. In addition, I will show that molecular mechanics force fields contain docking-relevant physics that appears to be missing from DL methods. PoseBusters allows practitioners to assess docking and molecular generation methods and will help drive innovation in new inductive biases still required to improve DL-based methods and ultimately the development methods that generate more accurate and more realistic predictions.

2.1 Introduction

As described in Chapter 1, molecular docking, an essential step in structure-based drug discovery (De Ruyck et al., 2016), is the task of predicting the predominant binding mode(s) of a protein-ligand complex given an experimentally solved or computationally modelled protein structure and a ligand structure (Morris & Lim-Wilby, 2008). The predicted complexes are often used in a virtual screening workflow to help select small molecules from a large library of possible candidates (Lionta et al., 2014); or directly by medicinal chemists to understand the binding mode and to decide whether a small molecule is a suitable drug candidate (Patrick, 2023, ch. 13).

Docking methods are designed with the understanding that binding is enabled by interactions between protein and ligand structures. But due to the complexity of this, methods tend to strike a balance between fast calculation and accuracy (Patrick, 2023, ch. 17). DL, as described in Chapter 1, promises to disrupt the dominant design principle of classical docking software, and DL-based docking methods promise to unlock fast and accurate virtual screening for drug discovery. To this end, several different DL-based docking methods have already been proposed (Corso et al., 2023; Lu et al., 2022; Méndez-Lucio et al., 2021; Stärk et al., 2022; Zhou et al., 2023); see also Table 1.2 for a more complete list.

Non-DL-based “classical” docking methods include within their search and scoring functions terms that ensure chemical consistency and physical plausibility; for example, limiting the degrees of movement in the ligand to only the rotatable bonds in the ligand and including penalties if the protein and ligand clash (Goodsell & Olson, 1990; Jones et al., 1995; Trott & Olson, 2010). In this chapter, I show that some current DL-based docking methods still lack such key “inductive biases”, resulting in the creation of unrealistic chemical structures and binding modes despite obtaining root-mean-squared deviation (RMSD) values (see Appendix A, Section A.1.3) from the experimental binding mode that are less than the widely-used 2 Å threshold (Cole et al., 2005). To assess the output of docking methods, I created an independent test suite called PoseBusters (Buttenschoen et al., 2024) that checks the chemical consistency and physical plausibility alongside established metrics, such as the binding mode RMSD.

The problem of assessing the physical plausibility of docking predictions is akin to the structure

validation of ligand data in the Protein Data Bank (PDB) (Berman et al., 2000; Shao et al., 2022). Structure validation assesses the agreement of each of the ligands' bond lengths and angles with those observed in related chemical structures and the presence of steric clashes both within the ligand and between it and its surroundings (Shao et al., 2022). While these tests were designed for users to select those ligand crystal structures which are likely to be correct (Shao et al., 2022), docking methods are evaluated on their ability to recover crystal structures, so their output should pass the same physical plausibility tests.

Physical plausibility checks are also part of some workflows for conformation generation (Friedrich, de Bruyn Kops et al., 2017; Hawkins et al., 2010). Friedrich, de Bruyn Kops et al. (2017) use geometry checks performed by NAOMI (Urbaczek et al., 2011) which measures—like the PDB tests just mentioned—the deviation from known optimal values for bond lengths and bond angles, and also tests for divergence from planarity for aromatic rings.

In addition to physical checks, chemical checks are also needed (Warren et al., 2012). Chemical checks proposed for checking PDB structures include the identification of mislabelled stereo assignment, inconsistent bonding patterns, missing functional groups, and unlikely ionisation states (Warren et al., 2012). The problem of checking chemical plausibility has also come up in *de novo* molecule generation, where Brown et al. (2019) proposed a test suite including checks for the chemical validity of any proposed molecule. For docking, the focus is less on stability and synthetic accessibility of a generated molecular structure as it is hoped that these have been tested prior to attempting docking, but more on chemical consistency and physical realism of the predicted bound conformation.

Some comparisons of docking methods have included additional metrics based on volume overlap (Warren et al., 2006) or protein-ligand interactions (Ciancetta et al., 2014) to supplement pose accuracy-based metrics such as RMSD of atomic positions and run time measurements, but the majority of comparisons of docking methods are predominantly based on binding mode RMSD (Bolcato et al., 2019; Cole et al., 2005; Onodera et al., 2007; Plewczynski et al., 2011).

The current standard practice of comparing the output conformations of docking methods using on RMSD-based metrics alone also extends to recent new methods. The five DL-based docking methods that I test in this chapter (Corso et al., 2023; Lu et al., 2022; Méndez-Lucio et al., 2021; Stärk et al., 2022; Zhou et al., 2023) all claimed better performance than standard docking methods but these claims rested entirely on the sole criterion on RMSD. None of these methods tested their outputs for physical plausibility.

In this chapter, I present PoseBusters, a test suite that is designed to identify implausible chemical structures, conformations and ligand poses. I used PoseBusters to evaluate the predicted binding modes generated by the five DL-based docking methods described earlier and listed in Table 2.1 and two standard non-DL-based “classical” docking methods AutoDock Vina (Trott & Olson, 2010) and GOLD

Table 2.1: Selected DL-based docking methods. The selection includes five methodologically different DL-based docking methods published over the last two years.

Method	Authors	Date	Search space
DeepDock	Méndez-Lucio et al.	Dec 2021	pocket
DiffDock	Corso et al.	Feb 2023	blind
EquiBind	Stärk et al.	Feb 2022	blind
TANKBind	Lu et al.	Oct 2022	blind
Uni-Mol	Zhou et al.	Feb 2023	pocket

(Verdonk et al., 2003). These poses were generated by re-docking the cognate ligands of the 81 protein-ligand crystal complexes in the Astex Diverse set (Hartshorn et al., 2007) and 308 ligands of the protein-ligand crystal complexes in the PoseBusters Benchmark set, a new set of complexes released from 2021 onwards, into their cognate receptor crystal structures. On the commonly-used Astex Diverse set, the DL-based docking method DiffDock appears to perform best in terms of RMSD alone but when taking physical plausibility into account, AutoDock Vina and GOLD perform best. On the PoseBusters Benchmark set, a test set that is potentially harder for methods that might have overfit to their training data because it contains only complexes that the DL methods have not been trained on, GOLD and AutoDock Vina are the best methods in terms of RMSD alone *and*, when taking physical plausibility into account, or when proteins with novel sequences are considered. The DL-based methods make few valid predictions for the “unseen” complexes and their performance degrades the more dissimilar the task is from their training data. Overall, I show that none of the DL-based methods I considered yet outperforms standard docking methods when physical plausibility is taken into account.

2.2 Methods

Five DL-based and two classical docking methods were used to re-dock known ligands into their respective proteins, and the predicted ligand poses were evaluated with the PoseBusters test suite. The following section describes the docking methods, the data sets, and the PoseBusters test suite for checking physicochemical consistency and structural plausibility of the generated poses.

2.2.1 Docking methods

The selected five DL-based docking methods (Corso et al., 2023; Lu et al., 2022; Méndez-Lucio et al., 2021; Stärk et al., 2022; Zhou et al., 2023) cover a wide range of DL-based approaches for pose prediction. Table 2.1 lists the methods and their publications. In order to examine the ability of standard non-DL-based methods to predict accurate chemically and physically valid poses, I also included the well-established docking methods AutoDock Vina (Forli et al., 2016) and GOLD (Verdonk et al., 2003).

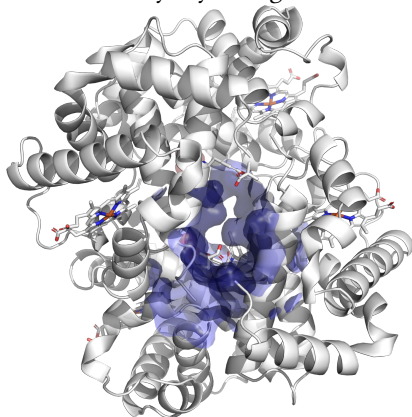
Table 2.2: Data sets used to train the selected five machine learning-based docking methods. All five DL-based methods were trained on subsets of the PDBBind 2020 General Set.

Method	Training and validation set
DeepDock	PDBBind 2019 General Set without complexes included in CASF-2016 or those that fail pre-processing—16 367 complexes
DiffDock, EquiBind	PDBbind 2020 General Set keeping complexes published before 2019 and without those with ligands found in test set—17 347 complexes
TankBind	PDBbind 2020 General Set keeping complexes published before 2019 and without those failing pre-processing—18 755 complexes
Uni-Mol	PDBBind 2020 General Set without complexes where protein sequence identity (MMSeq2) with CASF-2016 is above 40% and ligand fingerprint similarity is above 80%—18 404 complexes

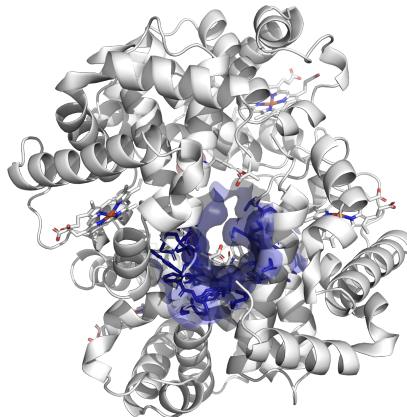
The five DL-based docking methods can be summarised as follows. Full details of each can be found in their respective references. DeepDock (Méndez-Lucio et al., 2021) learns a statistical potential based on the distance likelihood between ligand heavy atoms and points of the mesh of the surface of the binding pocket. DiffDock (Corso et al., 2023) uses equivariant graph neural networks in a diffusion process for blind docking. EquiBind (Stärk et al., 2022) applies equivariant graph neural networks for blind docking. TankBind (Lu et al., 2022) is a blind docking method that uses a trigonometry-aware neural network for docking in each pocket predicted by a binding pocket prediction method. Uni-Mol (Zhou et al., 2023) carries out docking with SE(3)-equivariant transformers. All five DL-based docking methods are trained on subsets of the PDBbind 2020 General Set (Z. Liu et al., 2017) as detailed in Table 2.2. DeepDock is trained on v2019 and the other four are trained on v2020. It should be noted that I used the DL models as trained by the respective authors without further hyper-parameter tuning.

The docking protocols that were used to generate predictions with each method and the software versions used are given in Appendix B, Section B.2. Table 2.3 lists the search space definitions that I used for each method. DeepDock and Uni-Mol require the definition of a binding site while DiffDock, EquiBind, and TankBind are ‘blind’ docking methods that search the entire protein. I used the default search spaces for the DL-based methods but larger than default search spaces for AutoDock Vina and GOLD, and Uni-Mol such to make them more comparable with the blind docking methods. Figure 2.1 shows the search spaces for one example protein-ligand complex. I show results for Uni-Mol across a range of binding site definitions starting from their preferred definition of all residues with an atom within 6 Å of a heavy atom of the crystal ligand. Under this tight pocket definition Uni-Mol performs better than any of the blind docking methods (Appendix B, Figure B.18).

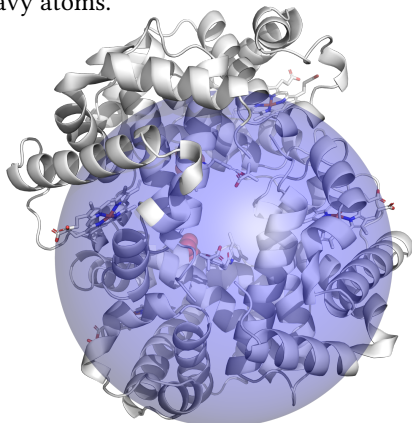
(a) DeepDock: Protein surface mesh nodes within 10 Å of any crystal ligand atom.



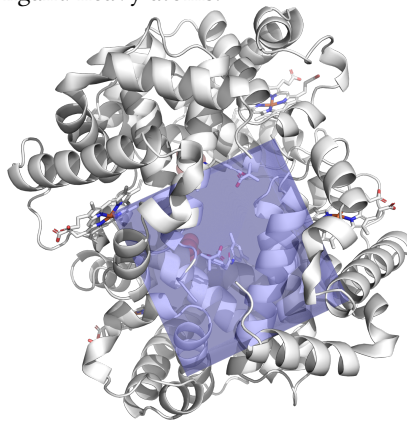
(b) Uni-Mol: Protein residues within 8 Å of any crystal ligand heavy atom.



(c) GOLD: Sphere of radius 25 Å centred on the geometric centre of the crystal ligand heavy atoms.



(d) AutoDock Vina: Cube with side length 25 Å centered on the geometric centre of crystal ligand heavy atoms.



(e) DiffDock, Equibind, TankBind: The entire protein crystal structure.

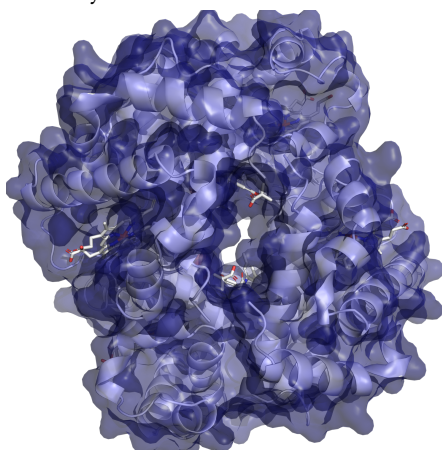


Figure 2.1: Search spaces of the docking methods illustrated on PDB entry 1G9V (Safo et al., 2001) for ligand RQ3. These are the default search spaces of DeepDock, DiffDock, EquiBind, and TankBind but larger than the usual search spaces for AutoDock Vina, GOLD, and Uni-Mol. Larger search spaces were selected for those methods to make them more comparable with the blind docking methods.

Table 2.3: Search spaces of the docking methods used. These are the default search spaces of DeepDock, DiffDock, EquiBind, and TankBind but larger than the usual search spaces for AutoDock Vina, GOLD, and Uni-Mol. Larger search spaces were selected for those methods to make them more comparable with the blind docking methods.

Method	Search space
<i>Classical docking methods</i>	
GOLD	Sphere of radius 25 Å centered on the geometric centre of the crystal ligand heavy atoms
Vina	Cube with side length 25 Å centered on the geometric centre of crystal ligand heavy atoms
<i>DL-based docking methods</i>	
DeepDock	Protein surface mesh nodes within 10 Å of any crystal ligand atom
Uni-Mol	Protein residues within 8 Å of any crystal ligand heavy atom
<i>DL-based blind docking methods</i>	
DiffDock	Entire crystal protein
EquiBind	Chains of crystal protein which are within 10 Å of any crystal ligand heavy atom
TankBind	Pockets identified by P2Rank (Krivák & Hoksza, 2018)

2.2.2 The PoseBusters test suite

The PoseBusters test suite is organised into three groups of tests. The first checks chemical validity and contains tests for the chemical validity and consistency relative to the input. The second group checks intramolecular properties and tests for the ligand geometry and the ligand conformation’s energy computed using the Universal Force Field (UFF) (Rappe et al., 1992). The third group considers intermolecular interactions and checks for protein-ligand and ligand-cofactor clashes. Descriptions of all the tests PoseBusters performs in the three sections are listed in Table 2.4. Molecule poses which pass all tests in PoseBusters are ‘PB-valid’.

For evaluating docking predictions, PoseBusters requires three input files: an SDF file containing the re-docked ligands, an SDF file containing the true “reference” ligand(s), and a PDB file containing the protein with any cofactors. The three files are loaded into RDKit molecule objects with the sanitisation option turned off.

Chemical validity and consistency

The first test in PoseBusters checks whether the ligand passes the RDKit’s sanitisation. The RDKit’s sanitisation processes information on the valency, aromaticity, radicals, conjugation, hybridization, chirality tags, and protonation to check whether a molecule can be represented as an octet-complete

Table 2.4: Description of the checks used in the PoseBusters test suite.

Test name	Description
<i>Chemical validity and consistency</i>	
File loads	The input molecule can be loaded into a molecule object by RDKit.
Sanitisation	The input molecule passes RDKit's chemical sanitisation checks.
Molecular formula	The molecular formula of the input molecule is the same as that of the true molecule.
Bonds	The bonds in the input molecule are the same as in the true molecule.
Tetrahedral chirality	The specified tetrahedral chirality in the input molecule is the same as in the true molecule.
Double bond stereochemistry	The specified double bond stereochemistry in the input molecule is the same as in the true molecule.
<i>Intramolecular validity</i>	
Bond lengths	The bond lengths in the input molecule are within 0.75 of the lower and 1.25 of the upper bounds determined by distance geometry.
Bond angles	The angles in the input molecule are within 0.75 of the lower and 1.25 of the upper bounds determined by distance geometry.
Planar aromatic rings	All atoms in aromatic rings with 5 or 6 members are within 0.25 Å of the closest shared plane.
Planar double bonds	The two carbons of aliphatic carbon-carbon double bonds and their four neighbours are within 0.25 Å of the closest shared plane.
Internal steric clash	The interatomic distance between pairs of non-covalently bound atoms is above 0.8 of the lower bound determined by distance geometry.
Energy ratio	The calculated energy of the input molecule is no more than 100 times the average energy of an ensemble of 50 conformations generated for the molecule. The energy is calculated using the UFF (Rappe et al., 1992) and the conformations are generated with ETKDGv3 followed by force field relaxation using the UFF with up to 200 iterations.
<i>Intermolecular validity</i>	
Minimum protein-ligand distance	The distance between protein-ligand atom pairs is larger than 0.75 times the sum of their van der Waals radii.
Minimum distance to organic cofactors	The distance between ligand and organic cofactor atoms is larger than 0.75 times the sum of their van der Waals radii.
Minimum distance to inorganic cofactors	The distance between ligand and inorganic cofactor atoms is larger than 0.75 times the sum of the pairs covalent radii.
Volume overlap with protein	The share of ligand volume that intersects with the protein is less than 7.5 %. The volumes are defined by the van der Waals radii around the heavy atoms scaled by 0.8.
Volume overlap with organic cofactors	The share of ligand volume that intersects with organic cofactors is less than 7.5 %. The volumes are defined by the van der Waals radii around the heavy atoms scaled by 0.8.
Volume overlap with inorganic cofactors	The share of ligand volume that intersects with inorganic cofactors is less than 7.5 %. The volumes are defined by the van der Waals radii around the heavy atoms scaled by 0.5.

Lewis dot structure (Landrum et al., 2025). Passing the RDKit’s sanitisation is a commonly-used test for chemical validity in cheminformatics, for example in *de novo* molecular generation (Brown et al., 2019).

The next test in PoseBusters checks for docking-relevant chemical consistency between the predicted and the true ligands by generating ‘standard InChI’ strings (Heller et al., 2015) for the input and output ligands after removing isotopic information and neutralising charges by adding or removing hydrogens where possible. InChI is the *de facto* standard for molecular comparison (Goodman et al., 2021), and the ‘standard InChI’ strings generated include the layers for the molecular formula (/), molecular bonds (/c), hydrogens (/h), net charge (/q), protons (/p), tetrahedral chirality (/t), and double bond stereochemistry (/b). Standardisation of the ligand’s protonation and charge state is needed because the stereochemistry layer is dependent on the hydrogen (/h), net charge (/q) and proton (/p) layers. These can unexpectedly change during docking even though most docking software considers the charge distribution and protonation state of a ligand as fixed (Torres et al., 2019). The normalisation protocol also removes the stereochemistry information of double bonds in primary ketimines which only depends on the hydrogen atom’s ambiguous location.

Intramolecular validity

The first set of physical plausibility tests in the PoseBusters test suite validates bond lengths, bond angles, and internal distances between non-covalently bound pairs of atoms in the docked ligand against the corresponding limits in the distance bounds matrix obtained from the RDKit’s Distance Geometry module. To pass the tests, all molecular measurements must lie within the user-specified tolerances. The tolerance used throughout this manuscript is 25 % for bond lengths and bond angles and 30 % for non-covalently bound pairs of atoms *e.g.*: if a bond is less than 75 % of the Distance Geometry bond length lower bound, it is treated as anomalous. This was selected as all but one of the crystal ligands in the Astex Diverse set and all of those in the PoseBusters Benchmark set pass at this threshold.

The PoseBusters test for flatness checks that groups of atoms lie in a plane by calculating the closest plane to the atoms and checking that all atoms are within a user-defined distance from this plane. This test is performed for 5- and 6-membered aromatic rings and non-ring non-aromatic carbon-carbon double bonds. The chosen threshold of 0.25 Å admits all Astex Diverse and PoseBusters Benchmark set crystal structures by a wide margin, and as with all other thresholds, can be adjusted by the user.

The final test for intramolecular physicochemical plausibility carried out by PoseBusters is an energy calculation to detect highly strained or unlikely conformations. Our metric for this is the ratio of the energy of the docked ligand conformation to the mean of the energies of a set of 50 generated unconstrained conformations as in Wills et al. (2023). The conformations are generated

using the RDKit’s ETKDGv3 conformation generator (Riniker & Landrum, 2015) followed by a force field relaxation using the UFF (Rappe et al., 1992) and up to 200 iterations. The test suite rejects conformations for which this ratio is larger than a user-specified threshold. Wills et al. (2023) set a ratio of seven based on the value where 95 % of the crystal ligands in the PDBbind 2020 General set are considered plausible (Wills et al., 2023). I selected a less strict ratio of 100 where only one structure each from the Astex Diverse and PoseBusters Benchmark set is rejected.

Intermolecular validity

Intermolecular interactions are evaluated by two sets of tests in the PoseBusters test suite. The first set checks the minimum distance between molecules and the second checks the share of overlapping volume. Both sets of tests report on intermolecular interactions of the ligand with four types of molecules: the protein, organic cofactors, and inorganic cofactors.

For the distance-based intermolecular tests PoseBusters calculates the ratio of the pairwise distance between pairs of heavy atoms of two molecules and the sum of the two atoms’ van der Waals radii. If this ratio is smaller than a user-defined threshold then the test fails. The default threshold is 0.75 for all pairings. For inorganic cofactor-ligand pairings the covalent radii are used. All crystal structures in the Astex Diverse set and all but one in the PoseBusters Benchmark set pass at this threshold. The van der Waals and covalent radii used are from the Blue Obelisk open data project (O’Boyle, Guha et al., 2011) that the RDKit uses.

For the second set of intermolecular checks, PoseBusters calculates the share of the van der Waals volume of the heavy atoms of the ligand that overlaps with the van der Waals volume of the heavy atoms of the protein using the RDKit’s ShapeTverskyIndex function. The tests have a configurable scaling factor for the volume-defining van der Waals radii and a threshold that defines how much overlap constitutes a clash. A threshold is necessary because many crystal structures already contain clashes. For example, Verdonk et al. (2003) found that 81 out of 305 selected high-quality protein-ligand complexes from the PDB contained steric clashes (Verdonk et al., 2003). The overlap threshold is 7.5 % for all molecule pairings and the scaling factor is 0.8 for protein-ligand and organic cofactor-ligand pairings and 0.5 for inorganic cofactor-ligand pairings.

2.2.3 Quality of fit

PoseBusters calculates the minimum heavy-atom symmetry-aware root-mean-square deviation (RMSD) between the predicted ligand binding mode and the closest crystallographic ligand using the RDKit’s GetBestRMS function. Coverage, a metric often used for testing docking methods, is the share of predictions that are within a user adjustable threshold which by default is 2 Å RMSD. This value is

arbitrary but commonly-used and recommended for regular-size ligands (Cole et al., 2005).

2.2.4 Sequence identity

In this chapter, sequence identity between two amino acid chains is the number of exact residue matches after sequence alignment divided by the number of residues of the query sequence. The sequence alignment used is the Smith–Waterman algorithm (T. Smith & Waterman, 1981) implemented in Biopython (Cock et al., 2009) using an open gap score of -11 and an extension gap score of -1 and the BLOSUM62 substitution matrix. Unknown amino acid residues are counted as mismatches.

2.2.5 Molecular mechanics energy minimisation

Post-docking energy minimisation of the ligand structure in the binding pocket was performed using the AMBER ff14sb force field (Maier et al., 2015) and the Sage small molecule force field (Boothroyd et al., 2023) in OpenMM (Eastman et al., 2017). The protein files were prepared using PDBfixer (Eastman et al., 2017) and all protein atom positions were fixed in space only allowing updates to the ligand atoms' positions. Minimisation was performed until energy convergence within 0.01 kJ mol^{-1} per iteration.

2.2.6 Data

Astex Diverse set

The Astex Diverse set (Hartshorn et al., 2007) published in 2007 is a set of hand-picked, relevant, diverse, and high-quality protein-ligand complexes from the PDB (Berman et al., 2000). The complexes were downloaded from the PDB as MMTF files (Bradley et al., 2017) and PyMOL (DeLano, 2002) was used to remove solvents and all occurrences of the ligand of interest from the complexes before saving the proteins with the cofactors in PDB files and the ligands in SDF files.

PoseBusters Benchmark set

I created the PoseBusters Benchmark set, a new set of carefully-selected publicly-available crystal complexes from the PDB. It is a diverse set of recent high-quality protein-ligand complexes which contain drug-like molecules. It only contains complexes released since 2021 and therefore does not contain any complexes present in the PDBbind 2020 General set used to train many of the methods. Table B.1 lists the steps used to select the 308 unique proteins and 308 unique ligands in the PoseBusters Benchmark set. The complexes were downloaded from the PDB as MMTF files and PyMOL was used

to remove solvents and all occurrences of the ligand of interest before saving the proteins with the cofactors in PDB files and the ligands in SDF files.

2.3 Results

The following section presents the analysis of the PoseBusters test suite on the re-docked ligands of five DL-based docking methods and two standard non-DL-based docking methods on the 85 ligands of the Astex Diverse set and the 308 ligands of the PoseBusters Benchmark set into the receptors crystal structures.

2.3.1 Results on the Astex Diverse set

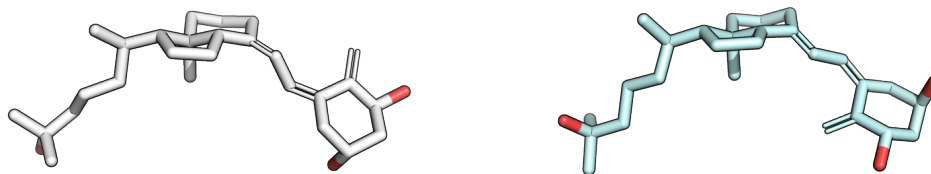
Figure 2.3 shows the overall results on the Astex Diverse set in ocean green for each of the seven methods: AutoDock Vina (Trott & Olson, 2010), GOLD (Verdonk et al., 2003), DeepDock (Méndez-Lucio et al., 2021), DiffDock (Corso et al., 2023), EquiBind (Stärk et al., 2022), TankBind (Lu et al., 2022), and Uni-Mol (Zhou et al., 2023). The striped bars show the performance only in terms of RMSD coverage ($\text{RMSD} \leq 2 \text{ \AA}$) and the solid bars show the performance after also considering physical plausibility, *i.e.*, only predictions which in addition pass all tests in PoseBusters and are therefore PB-valid.

The Astex Diverse set is a well-established and commonly-used benchmark for evaluating docking methods. Good performance on this set is expected because the five DL-based methods evaluated here have been trained on most of these complexes. 47 of the 81 complexes in the Astex Diverse set are in the PDBbind 2020 General Set and 67 out of the 81 of the Astex Diverse set proteins have more than 95 % sequence identity with proteins found in PDBbind 2020 General set. AutoDock Vina may also perform well on this data set because the linear regression model behind the scoring function was trained on an earlier version of PDBbind (Trott & Olson, 2010) which already included most of the Astex Diverse set.

The RMSD criterion alone (striped green bars in Figure 2.3) gives the impression that DiffDock (72 %) performs better than TankBind (59 %), GOLD (67 %), AutoDock Vina (58 %) and Uni-Mol (45 %). However, accepting only ligand binding modes that are physically sensible, *i.e.*, those predictions that pass all PoseBusters tests and are therefore PB-valid (solid green bars in Figure 2.3), many of the apparently impressive DL predictions are removed. The best three methods when considering RMSD and physical plausibility are GOLD (64 %), AutoDock Vina (56 %), and DiffDock (47 %) followed by Uni-Mol (12 %), DeepDock (11 %) and TankBind (5.9 %). DiffDock is therefore the only DL-based method that has comparable performance to the standard methods on the Astex Diverse set when considering physical plausibility of the predicted poses.

All five DL-based docking methods struggle with physical plausibility, but even the poses produced by the classical methods GOLD and AutoDock Vina do not always pass all of the PoseBusters checks.

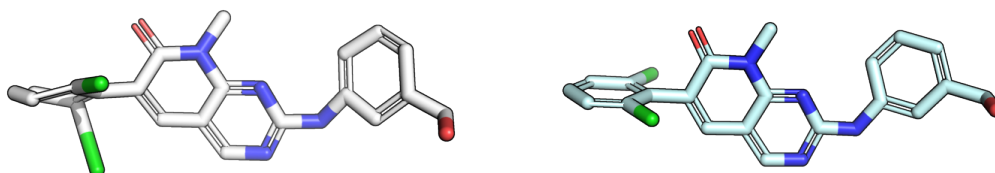
(a) Double bond stereochemistry not preserved. DiffDock prediction (left) for ligand VDX of protein-ligand complex 7QPP. RMSD 1.9 Å:



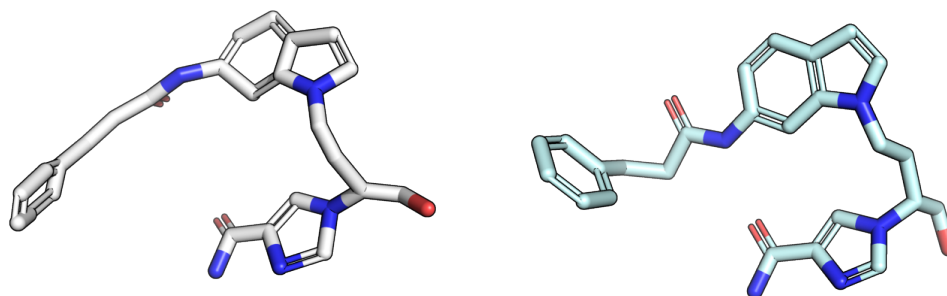
(b) Tetrahedral chiral centre not preserved. DeepDock prediction (left) for ligand GLF of protein-ligand complex 7DKT. RMSD 3.6 Å:



(c) Bond lengths too long. Uni-Mol prediction (left) for ligand P16 of protein-ligand complex 1OPK. RMSD 1.5 Å:



(d) Bond angles too extreme. Uni-Mol prediction (left) for ligand FR4 of protein-ligand complex 1UML. RMSD 1.4 Å:



(e) Internal clash. DeepDock prediction (left) for ligand BDI of protein-ligand complex 1N2V. RMSD 1.6 Å:



Continued on next page

Continued from previous page

(f) Aromatic rings not flat. TankBind prediction (left) for ligand CRZ of protein-ligand complex 1TOW. RMSD 2.2 Å:



(g) Double bond not flat. TankBind prediction (left) for ligand DBQ of protein-ligand complex 1U4D. RMSD 1.7 Å:



(h) Energy ratio too high. AutoDock Vina prediction (left) for ligand IFM of protein-ligand complex 7LOU. RMSD 1.9 Å:



(i) Clash with protein. DiffDock prediction (left) for ligand XQ1 of protein-ligand complex 7L7C. RMSD 1.6 Å:

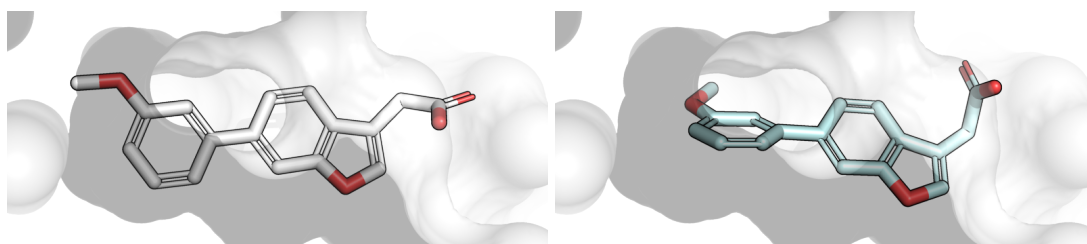


Figure 2.2: Examples of failure modes that PoseBusters is able to detect. Predictions are shown on the left with white carbons and the crystal structures on the right have cyan carbons. Oxygen atoms are red, nitrogen atoms are dark blue, chlorine atoms are green. Most of the predictions shown have a RMSD within 2 Å but all are physically invalid.

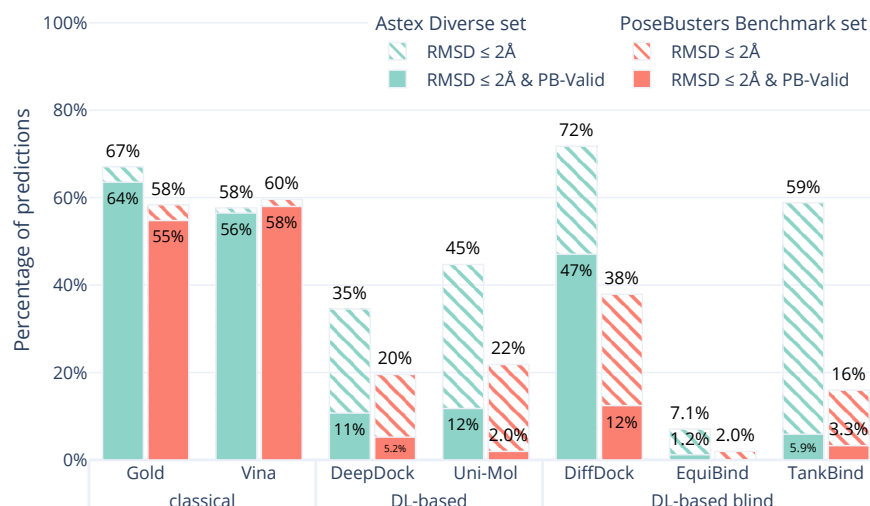


Figure 2.3: Comparative performance of the docking methods. The Astex Diverse set (85 cases) was chosen as an easy test set containing many complexes the five DL-based methods were trained on while the PoseBusters Benchmark set (308 cases) was chosen to be a difficult test set containing complexes none of the methods was trained on. The striped bars show the share of predictions of each method that have an RMSD within 2 Å and the solid bars show the subset that in addition have valid geometries and energies, *i.e.*, pass all PoseBusters tests and are therefore ‘PB-Valid’. DiffDock appears to outperform the classical methods on the Astex Diverse set when only binding mode RMSD is considered (striped teal bars). However, when physical plausibility is also considered (solid teal bars) or when presented with the PoseBusters Benchmark set (coral pink bars), AutoDock Vina and GOLD outperform all DL-based methods.

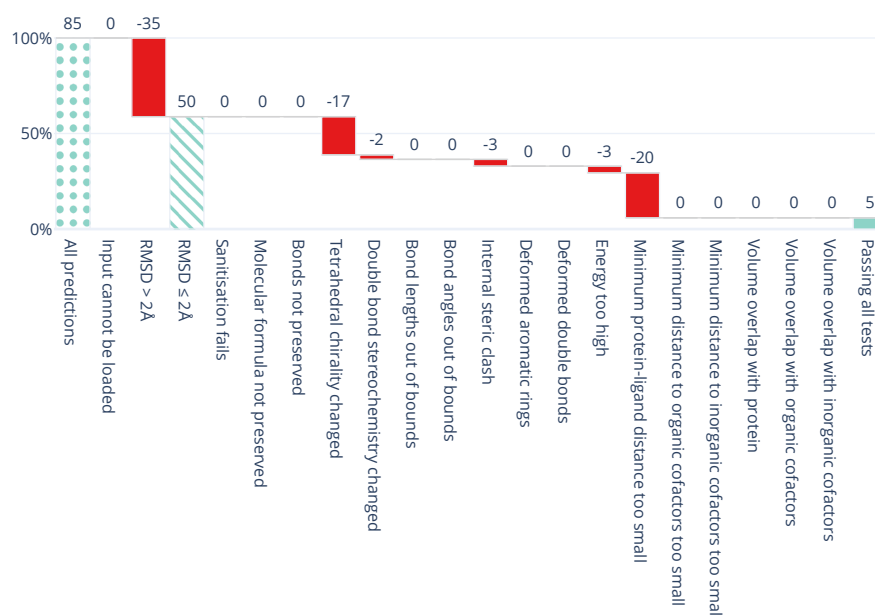


Figure 2.4: Waterfall plot showing the PoseBusters tests as filters for the TankBind predictions on the Astex Diverse data set. The tests in the PoseBusters test suits are described in Table 2.4. The leftmost (dotted) bar shows the number of complexes in the test set. The red bars show the number of predictions that fail with each additional test going from left to right. The rightmost (solid) bar indicates the number of predictions that pass all tests, *i.e.* those that are ‘PB-Valid’. For the 85 test cases in the Astex Diverse set 50 (59 %) predictions have RMSD within 2 Å RMSD and 5 (5.9 %) pass all tests. Figures B.3 and B.4 show waterfall plots for all methods and both data sets.

Figure 2.4 shows a waterfall plot that indicates how many predicted binding modes fail each test. The waterfall plots for the remaining methods are shown in Figure B.3 (Appendix B, Section B.7). The DL-based methods fail in different ways. TankBind habitually overlooks stereochemistry, Uni-Mol very often fails to predict valid bond lengths, and EquiBind tends to produce protein-ligand clashes. The classical methods GOLD and AutoDock Vina pass most tests but also generated a few protein-ligand clashes. Figure 2.2 shows examples of poses generated by the methods illustrating various failure modes.

The results on the Astex Diverse set suggest that despite what the $\text{RMSD} \leq 2 \text{ \AA}$ criterion would imply, none of the DL-based methods analysed here outperform classical docking methods when the physical plausibility of the ligand binding mode is taken into account. However, DiffDock in particular appears capable of making a large number of useful predictions.

2.3.2 Results on the PoseBusters Benchmark set

The results of the seven (AutoDock Vina, GOLD, DeepDock, DiffDock, EquiBind, TankBind, Uni-Mol) docking methods on the PoseBusters Benchmark set are shown in coral pink in Figure 2.3. The striped bars show the performance only in terms of RMSD within $2 \text{ \AA}$, while the solid bars show the performance in terms of PB-validity (passing all PoseBusters tests).

The PoseBusters Benchmark set was designed to contain no complexes that are in the training data for any of the methods. Performing well on this data set requires a method to be able to generalise well.

All methods perform worse on the PoseBusters Benchmark set than the Astex Diverse set. GOLD (48 %) and AutoDock Vina (51 %) perform the best out of the seven methods with (solid coral pink bars) and without (striped coral pink bars) considering PB-validity. On the PoseBusters Benchmark set, the best performing DL method, DiffDock (14 %), does not compete with the two standard docking methods. GOLD and AutoDock Vina again pass the most tests but for a few protein-ligand clashes.

The waterfall plots in Figure B.4 (Appendix B) show which tests fail for each method on the PoseBusters Benchmark set. Again, the methods have different merits and shortcomings. Out of the five DL-based methods, DiffDock still produces the most physically valid poses but few predictions lie within the $2 \text{ \AA}$ RMSD threshold. EquiBind, Uni-Mol and TankBind generate almost no physically valid poses that pass all tests. Uni-Mol has a relatively good RMSD score (23 %) but struggles to predict planar aromatic rings and correct bond lengths.

Figure 2.5 shows the results of the docking methods on the PoseBusters Benchmark set but stratified by the target protein receptor’s maximum sequence identity with the proteins in the PDBbind 2020 General set (Z. Liu et al., 2017). As the DL-based methods were all trained on subsets of the PDBbind 2020 General set, this roughly quantifies how different the test set protein targets are from those that

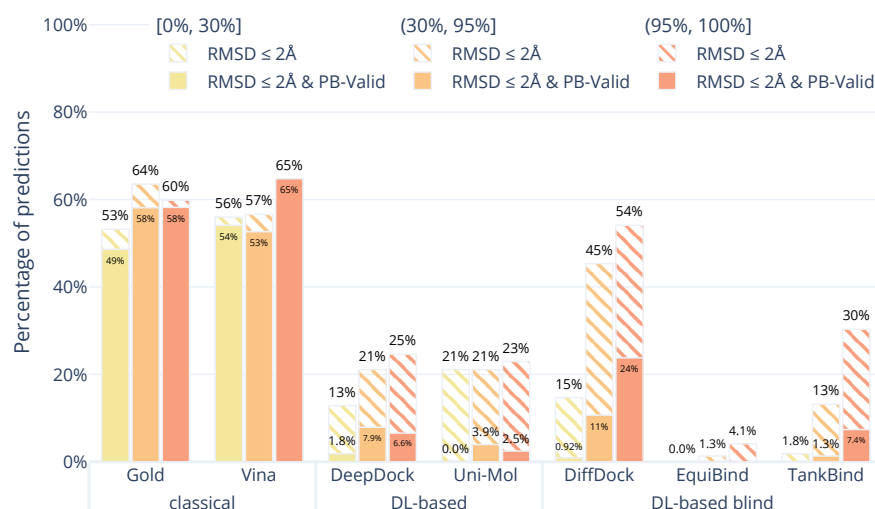


Figure 2.5: Comparative performance of docking methods on the PoseBusters Benchmark set stratified by sequence identity relative to the PDBBind General Set v2020. The sequence identity is the maximum sequence identity between all chains in the PoseBusters test protein and all chains in the PDBBind General Set v2020. The striped bars show the share of predictions of each method that have an RMSD within 2 Å and the solid bars show those predictions which in addition pass all PoseBusters tests and are therefore PB-valid. The DL-based methods perform far better on proteins that are similar to those they were trained on.

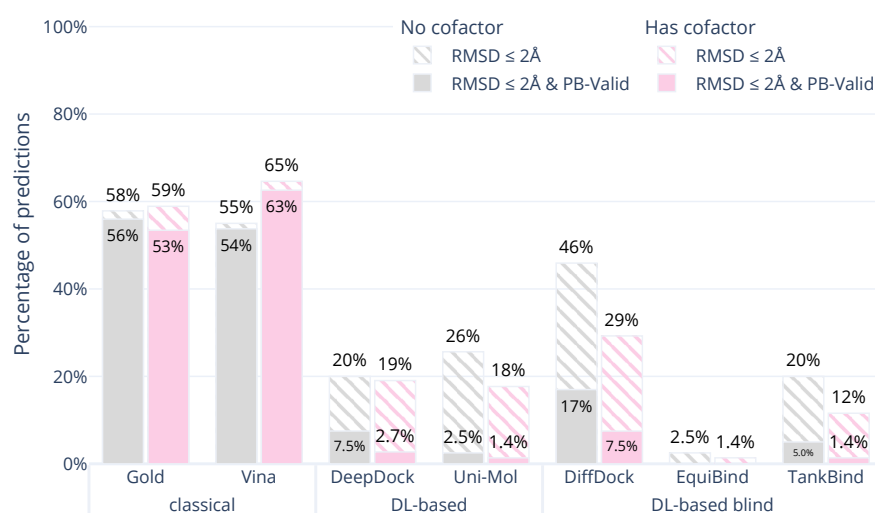


Figure 2.6: Comparative performance of docking methods on the PoseBusters Benchmark set stratified by the presence of cofactors. Cofactors are loosely defined as non-protein non-ligand compounds such as metal ions, iron-sulfur clusters, and organic small molecules that are within 4.0 Å of any ligand heavy atom. The striped bars show the share of predictions of each method that have an RMSD within 2 Å and the solid bars show those predictions which in addition pass all PoseBusters tests and are therefore PB-valid. The classical docking methods perform better on those systems with cofactors present while the DL-based methods perform worse on those systems.

the methods were trained on. I bin the test cases into three categories low [0, 30 %], medium (30 %, 90 %], and high (90 %, 100 %] maximum percentage sequence identity. Without considering physical plausibility (striped bars), the classical methods appear to perform as well on the three protein similarity bins while the DL-based methods perform worse on the proteins with lower sequence identity. This suggests that the DL-based methods are overfitting to the protein targets in their training sets.

I also compared the performance of the docking methods on the PoseBusters Benchmark set stratified by whether protein-ligand complexes contain cofactors (Figure 2.6). Here, I loosely define cofactors as non-protein non-ligand compounds such as metal ions, iron-sulfur clusters, and organic small molecules in the crystal complex within 4.0 Å of any ligand heavy atom. About 45% of protein-ligand complexes in the PoseBusters Benchmark set have a cofactor (Appendix B, Figure B.2). The classical methods perform slightly better when a cofactor is present while the DL-based docking methods perform worse on those systems.

2.3.3 Results with post-docking energy minimisation

In order to examine whether the outputs of the DL-based methods can be made physically plausible, I performed an additional post-docking energy minimisation of the ligand structures in the binding pocket for the PoseBusters Benchmark set (Figure 2.7). Again, striped bars indicate predictions with $\text{RMSD} \leq 2 \text{ \AA}$ while solid bars indicate which of those are also PB-valid *i.e.*, pass all PoseBusters tests. The figure shows that post-docking energy minimisation significantly increases the number of physically plausible structures of the DL-based methods DiffDock, DeepDock, TankBind, and Uni-Mol but does not improve the poses predicted by AutoDock Vina and GOLD. I also performed energy minimisation on the Uni-Mol results for the minimal 6 Å pocket (Appendix B, Figure B.19). The number of poses that pass the tests increases to about the same level as DiffDock. The fact that energy minimisation is able to repair many of the DL methods predicted poses and increase $\text{RMSD} \leq 2 \text{ \AA}$ performance shows that at least some force field-defined physics is missing from DL-based docking methods. An example of a predicted pose that was fixed is shown in Figure 2.8. However, even with the energy minimisation step, the best DL-based docking method DiffDock still performs worse than the classical methods GOLD and AutoDock Vina.

2.4 Discussion

In this chapter, I discuss PoseBusters, a test suite designed and built to identify chemically inconsistent and physically implausible ligand poses predicted by protein-ligand docking and molecular generation methods. I show the results of applying the PoseBusters test suite to the output of seven different

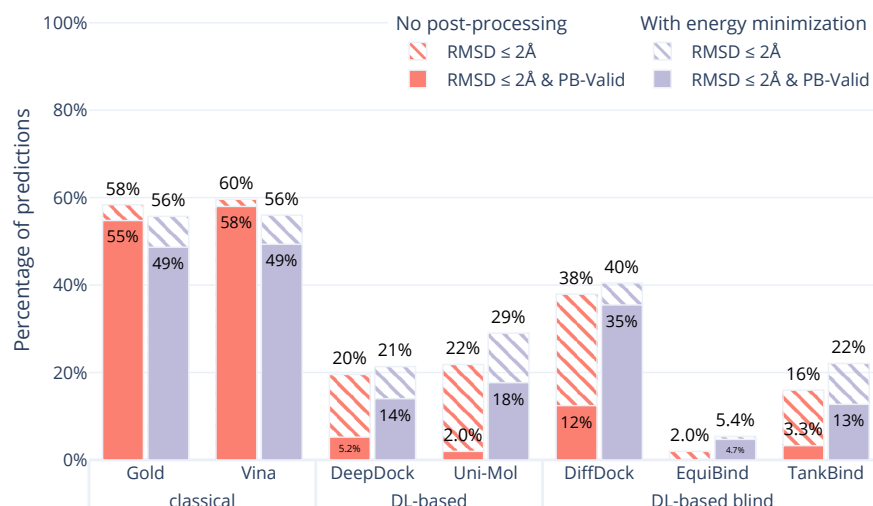
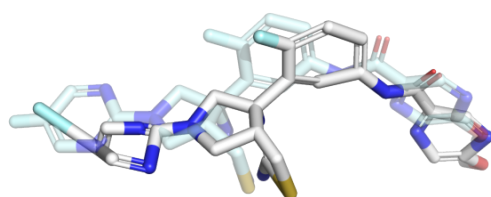


Figure 2.7: Comparative performance of docking methods with post-docking energy minimisation of the ligand (while keeping the protein fixed) on the PoseBusters Benchmark set. The striped bars show the share of predictions of each method that have an RMSD within 2 Å of the crystal pose and the solid bars show those predictions which in addition pass all PoseBusters tests and are therefore PB-valid. Post-docking energy minimisation significantly improves the relative physical plausibility of the DL-based methods' predictions. This indicates that force fields contain docking-relevant physics which is missing from DL-based methods.

(a) Predicted pose



(b) Optimised pose

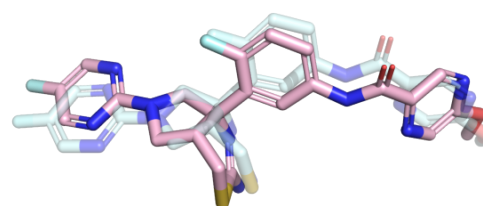
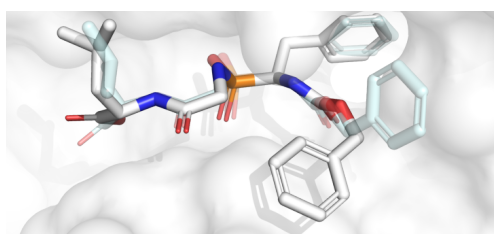


Figure 2.8: Example of a prediction that was fixed by the post-docking energy minimisation. (a) The Uni-Mol prediction (RMSD 2.0 Å) is shown in white. (b) The optimised prediction (RMSD 1.1 Å) is shown in pink. The crystal ligand is shown as a reference in light blue in both panels. Note how the aromatic rings are flattened and the leftmost bond is shortened by the optimisation, making the prediction pass all PoseBusters checks.

(a) Predicted pose



(b) Optimised pose

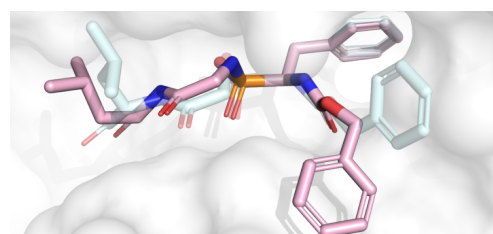


Figure 2.9: Example of a prediction that was 'destroyed' by the energy minimisation. (a) The prediction by Autodock Vina passes all PoseBusters checks and has a RMSD of 1.9 Å and is shown in white. (b) The optimised predicted ligand has a RMSD of 2.2 Å and is shown in pink. The crystal ligand is shown in light blue as a reference in both panels.

docking methods, five current DL-based docking methods (DeepDock, DiffDock, EquiBind, TankBind, and Uni-Mol) and two classical docking methods (AutoDock Vina and GOLD).

As mentioned earlier, none of the DL-based docking methods assessed here yet outperforms classical docking methods when *both* physical plausibility and binding mode RMSD is taken into account. Our work demonstrates the need for physical plausibility to be taken into account when assessing docking tools because it is possible to perform well on an RMSD-based metric while predicting physically implausible ligand poses (Figure 2.2). Using the tests in the PoseBusters test suite as an additional criterion when developing DL-based docking methods will help improve methods and the development of more accurate and realistic predictions.

In addition, the individual tests in the PoseBusters test suite highlight docking-relevant failure modes. The results show that Uni-Mol, for example, predicts non-standard bond lengths and TankBind creates internal ligand clashes. As mentioned earlier, the ability to identify such failure modes in predicted ligand poses makes PoseBusters a helpful tool for developers to identify inductive biases that could improve their binding mode prediction methods.

Our results also show that, unlike classical docking methods, DL-based docking methods do not generalise well to novel data. The performance of the DL-based methods on the PoseBusters Benchmark set overall was poor and the subset of the PoseBusters Benchmark set with low sequence identity to PDBbind 2020 General set revealed that DL-based methods are prone to overfitting to the proteins they were trained on. Our analysis of the targets with sequence identity lower than 30 % to any member of PDBbind 2020 General set revealed that across all of the DL-based docking methods almost no physically valid poses were generated within the 2 Å threshold.

The most commonly-used train-test approach for building DL-based docking models is time-based, *e.g.*, complexes released before a certain date are used for training and complexes released later for testing. Based on our results, I argue that this is insufficient for testing generalisation to novel targets and the sequence identity between the proteins in the training and test must be reported on.

Post-docking energy minimisation of the ligand using force fields can considerably improve the docking poses generated by DL-based methods. However, even with an energy minimisation step, the best DL-based method, DiffDock, does not outperform classical docking methods like GOLD and AutoDock Vina. This shows that at least some key aspects of chemistry and physics encoded in classical docking methods and force fields are missing from DL models.

The PoseBusters test suite provides a new criterion, PB-validity, beyond the traditional “RMSD $\leq$ 2 Å” criterion to evaluate the predictions of new DL-based methods, and hopefully will help to identify inductive biases needed for the field to improve docking and molecular generation methods, ultimately resulting in more accurate and realistic predictions. The next generation of DL-based docking methods should aim to outperform standard docking tools on both RMSD criteria and in terms of chemical

consistency, physical plausibility, and generalisability.

2.5 Conclusion

This chapter introduced the PoseBusters test suite which can check the generated output of docking methods for chemical consistency and physical plausibility. The main results from our analysis of five DL-based docking methods are that no DL-based method yet outperforms standard docking methods when considerations of physical plausibility are taken into account, that the DL-based docking methods tested do not generalize to novel data, and that force fields contain docking specific information that is lacking from these DL-based docking methods.

The PoseBusters test suite is a new standard to test for physical plausibility alongside native binding mode accuracy. Since the publication of the results in this chapter (Buttenschoen et al., 2024), many new DL based docking models, but more broadly also other molecular modelling software, like co-folding methods (Abramson et al., 2024; Boitreaud et al., 2024; Krishna et al., 2024; Wohlwend et al., 2024) and *de novo* molecular generation methods (Buttenschoen et al., 2025; M. Li et al., 2025) have been assessed using the PoseBusters test suite. The tests have become a tool for method developers to check that their methods do not egregiously violate stereochemistry and physics, and have helped identify the inductive biases still required to improve DL-based docking methods, as shown by recent improvements (Corso et al., 2024).

Since the publication of the first version of the PoseBusters software, additional tests have been added and bugs have been fixed based on suggestions and feedback from various users. Maria Kadukova contributed settings for a non-flatness test for aliphatic rings. Guy Durant identified a bug that prevented passing molecule objects into the main function. In addition, checks for the connectedness of molecules, the absence of radicals, and clashes with water molecules have been added. Various stability improvements were made for users running the PoseBusters checks on extremely wonky molecules produced by *de novo* molecular generators.

Other publications have also advanced assessments of protein-ligand models beyond the metrics mentioned here. For example, Errington et al. (2025) have expanded the set of metrics by monitoring the conservation of intermolecular interactions, showing another way to possibly improve DL-based protein-ligand models, namely by training models directly to learn intermolecular interactions.

This chapter assessed DL-based methods for docking. The next chapter takes a look at the data behind those models, its quality, the patterns within it, and what that might actually tell us about how good DL models can be.

Chapter 3

Evaluating the training data

In the previous chapter, I benchmarked deep learning (DL)-based docking models. In this chapter, I describe how I used the PoseBusters tests developed in the previous chapter to also assess the models' training and test data. The models' failure modes are present in the commonly used training data set PDBbind 2020 General set, but not as strongly pronounced. The high quality of the established test sets, the Astex Diverse set, CASF-2016, and PoseBusters Benchmark set, is reconfirmed. This demonstrates that the models might have learnt these failure modes as biases from the data, but the data do not fully explain them.

3.1 Introduction

A well-known adage in computer science is “garbage in, garbage out.” In the context of machine learning, this principle implies that low-quality training data will tend to produce low-quality models. This concern is even more pronounced for deep learning, which seeks to learn meaningful latent representations directly from data. Consequently, a comprehensive evaluation of deep learning-based docking models must also consider the quality of the experimental structures used in training.

Experimentally-solved structures of protein-ligand complexes can vary significantly in quality, and this includes the resolution of the experimental structures (see Appendix A, Section A.1.4), the quality of model fit, and the completeness of the ligand models. The Protein Data Bank (PDB) (Berman et al., 2000), introduced in Chapter 1 as the most important source of experimentally-derived protein-ligand structures, has been continuously growing and now, at the end of 2025, it contains approximately 170 000 protein-ligand structures. However, not all of these structures are suitable for method development, which is why subsets of relevant, representative, and high quality structures from the PDB are typically selected. Some data sets that have been used for the training and evaluation of docking methods are listed in Section 3.1. These apply different selection criteria for different

Table 3.1: Some data sets that have been or are still used to train or evaluate docking methods. The size is the number of unique protein-ligand complexes in the data set. Some complexes contain multiple ligand poses which are sometimes counted instead. Most of these data sets are subsets of the PDB either because they have been derived from the PDB or the data has since been added to the PDB.

Name	Creators	Year	Size
AutoDock 3	Morris et al.	1998	30
Ligand-Protein DataBase	Roche et al.	2001	195
CCDC/Astex test set	Nissink et al.	2002	305
PDBbind	R. Wang et al.	2004	5671
PDBbind 2003	R. Wang et al.	2004	5897
BindingMOAD	Hu et al.	2005	5331
Astex Diverse Set	Hartshorn et al.	2007	85
BindingMOAD 2007	Benson et al.	2007	9836
AutoDock 4	Morris et al.	2009	188
CASF-2007	Cheng et al.	2009	195
sc-PDB	Meslamani et al.	2011	8166
DUD-E	Mysinger et al.	2012	22 886
CASF-2013	Y. Li et al.	2014	195
Binding MOAD 2014	Ahmed et al.	2014	23 269
PDBbind 2014	Z. Liu et al.	2015	8302
CASF-2016	Su et al.	2016	285
PDBbind 2016	Z. Liu et al.	2016	11 839
Platinum Data Set	Friedrich, Meyder et al.	2017	4626
BindingMOAD 2019	R. D. Smith et al.	2019	32 747
CrossDocked	Francoeur et al.	2020	18 450
PDBbind 2020	†	2021	19 443
Roche PDE10A set	Tosstorff et al.	2022	77
BindingMOAD 2023	Wagle et al.	2023	41 409
PoseBusters	Buttenschoen et al.	2024	308
DockGen	Corso et al.	2025	16 881
PDBbind 2020 R1	†	2025	19 443
PDBbind 2024	†	2025	27 385

† No publication specific to this version yet.

purposes. For example, Binding MOAD (Wagle et al., 2023), applies a resolution threshold of 2.5 Å to collect as many structures as possible.

Not only the quality, but also the volume of the training data is important. The first method to use linear regression to predict binding affinity in docking was AutoDock 3, which used 30 protein-ligand complexes (Morris et al., 1998), for a model with 5 parameters. DL-based methods, with billions of parameters to be fitted, require far more training data, but it is not yet clear how much data is needed. The largest data set I have seen used to train DL-base docking models in Section 3.1 is BindingMOAD, which contains 41 409 protein-ligand complexes.

Not only data selection, but also data preparation is important for many docking workflows. Ligand preparation includes creating 3D coordinates, adding hydrogen atoms, setting charges, assigning proper bond orders, and generating accessible tautomers, while receptors are typically prepared by removing atom clashes, filling in missing atoms and residues, placing waters, and optimising hydrogen bonds (Sastry et al., 2013). For classical docking, data preparation has been shown to critically influence docking success rates, with small preparation errors causing fluctuations of more than 20 % in docking success rates (Corbeil et al., 2012).

The representation and parsing of molecular data is not a trivial problem, in part because several cheminformatics file formats in common use are not open standards. For example, the widely used MOL2 file format is defined by a closed-source implementation in proprietary software, and no complete public specification exists. As a result, developers of independent software have had to infer or reconstruct parts of the format, which has led to incompatibilities. For this reason, the most commonly used open-source cheminformatics library, the RDKit, does not support saving molecules in the MOL2 format. Another example is the PDB file format, which allows molecules to be stored without explicit bond order information or atomic charges. This missing information can cause difficulties when parsing molecular structures upon loading. When DL-based models are trained on such poorly specified data, it is therefore prudent to verify the molecular structures that are actually loaded from the data set. If problems are identified, the choice of training data and the associated quality criteria should be re-examined.

Benchmarking the RDKit is important because it is the *de facto* standard cheminformatics library. It is used by all five DL-based docking methods evaluated in the previous chapter. The RDKit's strengths include its robustness, active maintenance, rich feature set, high-performance C++ back-end, user-friendly Python API, and extensive documentation. However, users must exercise caution. The RDKit performs strict parsing of molecular files and does not attempt to infer missing information, such as atomic charges, bond orders, or valence states, that are not explicitly specified in the input. While this behaviour increases reliability and transparency, it also means that data sets that cannot be loaded correctly by the RDKit warrant careful scrutiny, particularly given the widespread adoption

of the library.

3.2 Methods

In this chapter, I take a look at the quality of five data sets that have been used to train DL-based docking models. The data sets were downloaded from the original source, loaded with the open source cheminformatics package RDKit, and tested with the PoseBusters test suite.

All PDB files were loaded with the `MolFromPDBFile` function, SDF and MOL files with the `MolFromMolFile` function, and MOL2 files with the `MolFromMol2File` function.

3.2.1 Data sets

Astex Diverse The Astex Diverse data set contains 85 protein-ligand complexes and was created to assess classical docking methods (Hartshorn et al., 2007). It was obtained from the CCDC website¹. The MOL files of the ligands are appended the extra line “M END”² and the PDB files of the receptors are loaded without further processing.

Astex Diverse PB The Astex Diverse PB data set was created, as described in Chapter 2, using the PDB identifiers of the Astex Diverse set, but using CIF structure files obtained from the PDB. The SDF files of the ligands and the PDB files of the receptors were loaded without further processing.

CASF-2016 The CASF-2016 data set was created to compare the scoring functions of docking methods (Su et al., 2019). It is also a subset of the PDBbind 2020 General data set described below. It was downloaded from the PDBbind website³. The ligands are provided as SDF and MOL2 files, and the receptors are provided in PDB and MOL2 files. All files were loaded without further processing.

DockGen The DockGen data set was introduced as a training, validation, and test set for the DL-based docking method DiffDock-L (Corso et al., 2024). It is a processed version of Binding MOAD 2020 with data splits based on the CATH protein structures classification (Sillitoe et al., 2021). The DockGen data set was downloaded from Zenodo⁴. The PDB files of the ligands were loaded with PyMOL and saved as SDF files and the receptor PDB files were loaded without further processing.

¹Retrieved on 21 September 2022 from <https://ccdc.cam.ac.uk/support-and-resources/downloads>

²While not strictly needed this stops the RDKit parser from complaining about an incomplete file.

³Retrieved on 14 March 2022 from <http://www.pdbbind-cn.org>.

⁴Retrieved on 22 March 2025 from <https://zenodo.org/records/10656052>

PDBbind 2020 General The PDBbind 2020 General set (Z. Liu et al., 2015) contains 19 443 protein-ligand complexes from the PDB and binding affinity data. There are two subsets: the Refine set ($n = 5316$) that applies extra quality filters, and the Core ($n = 285$) set that is used in the CASF-2016 assessment and was just described above. The PDBbind 2020 General set was downloaded from the old PDBbind website⁵. The SDF of the ligands and the PDB files of the receptors were loaded without further processing.

PDBbind 2020 General R1 The 2024 release of the PDBbind data set includes updates to the data processing pipeline. According to the new PDBbind+ website⁶, these changes improve the preparation of the protein-ligand complexes, especially in terms of the ligands' protonation and tautomeric states. The PDBbind 2020 General R1 release contains the same protein-ligand complexes as the PDBbind 2020 General release but they have been processed with the new pipeline. The PDBbind 2020 General R1 set was downloaded from the PDBbind+ website⁷. The SDF of the ligands and the PDB files of the receptors were loaded without further processing.

PoseBusters The PoseBusters Benchmark set was introduced to compare classical and DL-based docking methods in Chapter 2. It contains 308 protein-ligand complexes from the PDB and is available online⁸. The SDF of the ligands and the PDB files of the receptors were loaded without further processing.

3.2.2 Assessment

The PoseBusters tests, introduced in the previous chapter (Chapter 2), are used to test the data. Here, the testing scenario is slightly different. Rather than testing re-docked ligands, as in the previous chapter, the ligand fit in the crystal structures is assessed. Two changes are made to the test suite to facilitate this scenario: First, the RMSD-based metric is omitted. Second, rather than checking the identity between the crystal and predicted ligands, the identity of the crystal ligand is checked against its entry in the PDB's Chemical Component Dictionary (CCD) (Feng et al., 2004). For ligands which do not have entries in the CCD, the chemical similarity tests are skipped. For both Astex Diverse sets, the CASF-2016 set, and the PoseBusters Benchmark set, all ligands are in the CCD; however, the PDBbind 2020 General set has 13.4 % (2611), the PDBbind 2020 General R1 has 13.7 % (2611), and the DockGen set has 7.0 % (23) ligand entries that are not in the CCD.

⁵Retrieved on 22 September 2022 from <http://www.pdbbind-cn.org>.

⁶Retrieved on 30 November 2025 from <https://www.pdbbind-plus.org.cn/>.

⁷Retrieved on 30 November 2025 from <https://www.pdbbind-plus.org.cn/download>.

⁸Available from Zenodo <https://zenodo.org/records/8278563>.

3.3 Results

All data sets had some failure cases. The test set with the highest percentage of passing structures was PoseBusters (99.4 %), followed by Astex Diverse PB (91.7 %), and PDBbind 2020 General R1 (78.3 %). The training set PDBbind 2020 General R1 had 63.0 % passing structures showing an improvement over PDBbind 2020 General (54.5 %). Table 3.2 summarises the results grouping the intermolecular, intramolecular, and identity checks together; the individual test passing rates are provided in Appendix C. The largest category of errors were intermolecular (*e.g.* protein-ligand clash) followed by intramolecular (*e.g.* sanitisation of the ligand).

The commonly used training data set PDBbind 2020 General contains 19 443 protein-ligand complexes (unique PDB identifiers) and 86.6 % (16 832) passed the intermolecular checks, 69.9 % (13 600) passed the intramolecular checks, and 54.5 % (10 590) structure passed all test (Table 3.2). The greatest failure modes, as shown in Figure 3.1, were sanitization (2825, 14.5 %), protein-ligand clash (1860, 9.6 %), mismatching molecular formulas (1099, 5.6 %), and clashes with water molecules (732, 3.8 %). Not all ligands in this data set had entries in the CCD because 2611 ligands are represented as multimers of CCD entries, for example sugars and small peptides. The PDBbind 2020 General R1 set had similar failure modes (Figure C.5b), but more structures passed overall (67 %, 12 804) than for PDBbind 2020 General (54.5 %, 10 590); this improvement was mainly due to lower failure rates in the RDKit sanitization test, while the number of protein-ligand clashes remained the same.

The CASF-2016 set, a subset of the PDBbind 2020 General set, contained 285 protein-ligand complexes and 263 (91.2 %) passed all tests (Figure C.4). The failure modes of the MOL2 ligand and PDB receptor files were mismatches in molecular formula (10, 3.5 %), ligand internal strain (5, 1.8 %), and stereochemistry mismatches (5, 1.8 %). Two (0.7 %) entries failed due to protein-ligand clashes (1O3F, 4TMN). The results for the other file types are summarised in Table C.1; none of the MOL2 receptor files could be loaded correctly and the SDF and PDB files yielded 4.2 % passing structures.

The Astex Diverse set contains 85 complexes, and 51 (60 %) passed all checks (Figure C.2a). 29 (34 %) complexes did not sanitise, two (2.4 %) had internal clash (1JD0, 1T9B), one (2.2 %) had a mismatching molecular formula (1GM8), and two (2.4 %) had double bond stereochemistry mismatches with the CCD (1S19, 1YVF). For the recreated Astex Diverse PB set 79 (93 %) complexes passed all checks (Figure C.2b), one (2.2 %) failed due to internal clash (1JD0), one (2.2 %) failed due to internal strain (1YGC), two (2.4 %) had double bond stereochemistry mismatches with the CCD (1S19, 1YVF), and three (3.6 %) had tetrahedral chirality mismatches with the CCD (1GPK, 1S3V, 1OF6). Further inspection of the files revealed that the MOL files were missing charges which the RDKit does not fill in, and as a result the sanitisation fails, for example for molecules with four-valent nitrogens. The structures of the other data set that was used to assess docking models in the previous chapter, PoseBusters, pass most tests.

Table 3.2: Data set pass rates on subsets and all of the PoseBusters tests. The table shows the size of each data set and the percentage of structures that passed all intermolecular tests, all intramolecular tests, all identity tests, and all tests overall. The intermolecular tests contain all tests relative to molecular structures in the receptor; the intramolecular tests contain the tests that only require the crystal ligand file; the identity tests contain the checks against the CCD entry; and all tests is all tests. The CASF-2016 results are for the MOL2 ligand and PDB receptors. The DockGen (bonds) is the DockGen data set but with bonds filled in by PyMOL. All data sets have some problems, but the ‘high quality’ test sets have less failure modes.

Data set	Size	% intermolecular	% intramolecular	% identity	% all
Astex Diverse	85	100.0	63.5	62.4	60.0
Astex Diverse PB	85	100.0	97.6	94.1	91.8
CASF-2016	285	99.3	96.8	93.7	91.2
DockGen	330	91.8	0.0	7.3	0.0
DockGen (bonds)	330	91.8	63.9	31.8	28.5
PDBbind 2020	19 443	86.6	81.1	73.0	63.0
PDBbind 2020 R1	19 037	87.1	95.6	89.9	78.2
PoseBusters	308	99.7	99.7	100.0	99.4

Table 3.3: The methods’ PoseBusters tests pass rates for re-docking the ligands of the Astex Diverse set. The table recapitulates the results from Chapter 2; shown here are the percentage of structures that passed all intermolecular tests, all intramolecular tests, all identity tests, and all tests overall. The intermolecular tests contain all tests relative to molecular structures in the receptor; the intramolecular tests contain the tests that only require the crystal ligand file; the identity tests contain the checks against the input molecule identity; and all tests is all tests. As discussed in Chapter 2, the five DL-based docking methods DeepDock, DiffDock, EquiBind, TankBind, and UniMol used the PDBbind General set for training (Table 2.2), and the Astex Diverse set is mostly contained in the PDBbind General 2020 set. The failure rates of these five methods are much higher than those observed in the training data set. For example, the pass rate of the intermolecular poses is no more than 60 % for those five methods, but the training data has 87 % intermolecular validity.

Method	% intermolecular	% intramolecular	% identity	% all
AutoDock Vina	100.0	98.8	100.0	98.8
DeepDock	17.6	54.1	71.8	10.6
DiffDock	56.5	100.0	100.0	48.2
EquiBind	4.7	81.2	100.0	1.2
GOLD	96.5	100.0	100.0	96.5
TankBind	22.4	65.9	58.8	7.1
UniMol	24.7	30.6	70.6	11.8

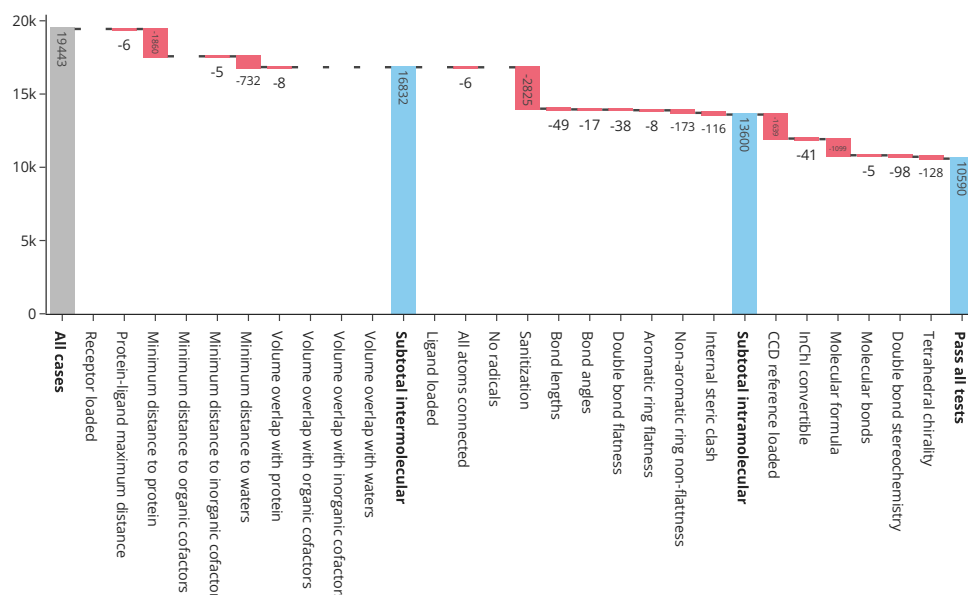


Figure 3.1: PDBbind General 2020 PoseBusters failure mode waterfall plots. The plot shows the number of protein-ligand complexes that fail at each test; the remaining ones pass all tests. The PDBbind data set was combined structure and binding affinity data and it has been used to train many DL-based docking methods, including all 5 that were benchmarked in Chapter 2. Overall, 54 % (10590) of the 19443 protein-ligand complexes pass all tests. The largest failure mode related to the receptor structures is the protein-ligand clash (9.5 %, 1860). The largest failure mode related to the ligand structures is the RDKit sanitization of the ligand (14.5 %, 2825). Note that 2736 of the ligands in this data set are multimers and not single entries in the CCD.

306 (99 %) of the 308 ligand poses pass all the tests (Figure C.1). There are two failure cases, one protein-ligand clash (7T0D) and one aromatic ring that is not flat (7V3S).

Lastly, the DockGen data set contains 330 protein-ligand complexes. Inspection of the identifiers reveals that there are two protein-ligand complexes (6XB3, 6O70) that are both in the validation set and in the test set. None of the complexes pass the tests because there is no bond information in the PDB files (Figure C.3a) and the RDKit does not automatically retrieve this information from the CCD. Using PyMOL to insert the bonds using the CCD and saving the ligands as SDF files fixes this problem (Figure C.3b) but this approach does not fully create chemically correct RDKit molecule objects: 102 (31 %) ligands fail the RDKit sanitization check and 73 (22 %) fail the InChI inter-convertibility check. The intermolecular failure modes are 25 (7.2 %) protein-ligand clashes and two occasions (4ZO7, 6MT9) where the ligand is more than 5 Å away from the protein.

The severity of the protein-ligand clashes and the intramolecular ligand clashes in the data sets is shown as box plots in Figure 3.2. The majority of complexes in all data sets pass both tests; however, the training data set PDBbind 2020 General and the testing data set DockGen set contain outliers with severe clashes of both kinds.

Next, the training data are compared against the predictions made by the models trained on it. Recall that the five DL-based methods were all trained on most of the PDBbind 2020 General set

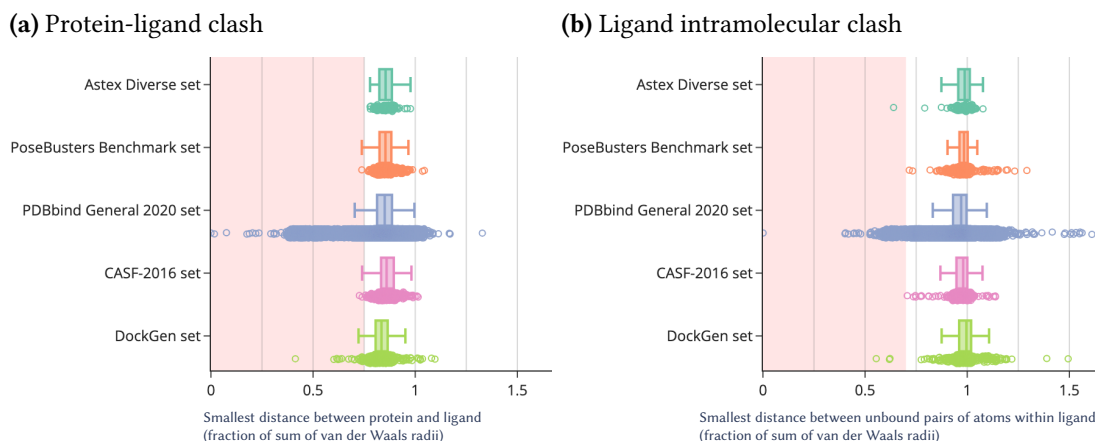


Figure 3.2: Comparison of the severity of protein-ligand and intramolecular ligand clashes in the crystal structures of the data sets. The ends of the boxes indicate the first and third quartile, the line within marks the median, and the fences indicate the last data point within 1.5 times the interquartile range measured from the ends of the boxes. The red areas mark the zone where each test fails. (a) Protein-ligand clash is measured as the smallest relative distance between all pairs of protein-ligand atoms as a fraction of the sum of the van der Waals radii and distances above 0.75 pass. (b) Intramolecular ligand clash is measured as the smallest relative distance between all pairs of non-covalently interacting ligand atoms as a fraction of the van der Waals radii and distances above 0.70 pass. Most of the data passes both tests; the training data PDBbind General 2020 and the DockGen set contain outliers with severe clashes of both kinds.

(Chapter 2, Table 2.2). The PoseBusters test results of the poses generated by the DL-based docking methods from Chapter 2 are recapitulated in Table 3.3. Although most of the complexes in the PDBbind 2020 General training data pass the tests, most of the predicted poses of the DL-based models fail the intermolecular tests (Table 3.3), in particular because they contain protein-ligand clashes and intramolecular ligand clashes (Figure C.7). Figure 3.3 further shows that on the more difficult PoseBusters Benchmark set, more than three quarters of the predicted poses have protein-ligand clashes; this is a much higher rate than observed in the training data.

3.4 Discussion

First, I consider what the results show in regards to the ‘garbage in, garbage out’ hypothesis. As the results show, the PDBbind 2020 General training data set does have a large number of structures that do not pass the tests. The most severe structural failure mode is protein-ligand clash (Appendix C, Figure C.6a). This is shared with the five previously benchmarked DL-based models that were trained on subsets of the PDBbind 2020 General General set (Chapter 2, Table 2.2). However, when comparing failure rates of the methods (Figures C.7 and C.8) to those in the data sets, I observe higher failure rates (Appendix C, Tables 3.2 and 3.3). In particular, the methods generate poses with much higher protein-ligand clash rates from the DL-based methods compared to the training data (Figure 3.3) and while the

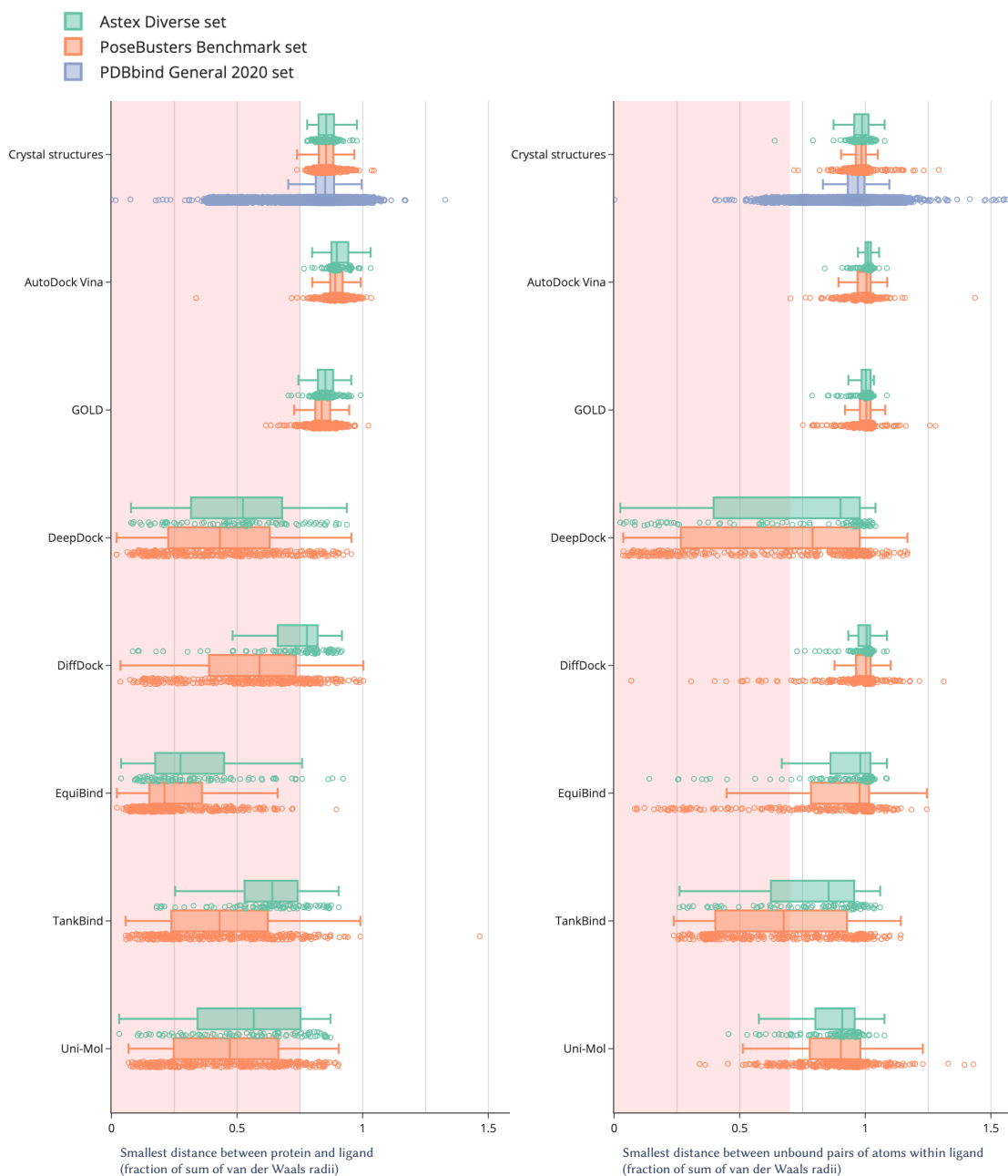
(a) Protein-ligand clash**(b) Ligand intramolecular clash**

Figure 3.3: Comparison of the severity of protein-ligand and intramolecular ligand clashes in the training and test data and the predicted poses of the docking models. The ends of the boxes indicate the first and third quartile, the line within marks the median, and the fences indicate the last data point within 1.5 times the inter-quartile range measured from the ends of the boxes. The red areas mark the zone where each test fails. (a) Protein-ligand clash is measured as the smallest relative distance between all pairs of protein-ligand atoms as a fraction of the sum of the van der Waals radii and distances above 0.75 pass. (b) Intramolecular ligand clash is measured as the smallest relative distance between all pairs of non-covalently interacting ligand atoms as a fraction of the van der Waals radii and distances above 0.70 pass. The training data PDBbind General 2020 contains complexes with severe clashes of both kinds; however, these are outliers. Most of the data passes the thresholds since the whiskers of the blue box do not touch the red zone. Almost three quarters of all the predicted poses of the DL-based models contain protein-ligand clashes and many contain ligand intramolecular clashes. The training data contains systematic structural artefacts but these data issues alone cannot explain the observed error rates in the predicted poses.

training data does contain systematic structural artefacts, these data issues alone cannot explain the observed error rates in the predicted poses. In conclusion, there is some evidence that these DL-based models learn protein-ligand clashes from the data, but the quality of the training data does not fully explain the much higher failure rates observed in the models. To discern the influence of problematic data on model performance, it is not sufficient to examine the training data alone. One could retrain the models on filtered clean versions of the training data set to see what the effect of the faulty data really is.

However, it is likely that incorrect handling and loading of the data might have caused some of the models to learn from faulty molecular representations. Indeed, the results show that loading molecular data sets with the RDKit is not trivial. Depending on which files are chosen and on what options are used, one might get 91 % passing protein-ligand complexes or none when loading the molecules of the CASF-2016 data set. In general, there is no molecular file type from which the RDKit is guaranteed to load valid molecules. All file formats admit underspecified or wrong molecular data, and the RDKit parser, as noted, although known for particularly strict parsing, cannot follow the unwritten conventions of other software when dealing with omitted information.

When the RDKit cannot sanitise a molecule, an error is raised. However, when no error is raised, a molecule can still be wrong. This is because sanitization does not guarantee that the molecule was correctly loaded. These silent bugs are logic errors, and they cause molecules to be parametrised wrongly. This happens when loading the CASF-2016 files. When loading the SDF files of the ligands instead of the MOL2 files, there are some sanitisation errors, but other ligands load without error but have radicals, non-flat double bonds, and flat non-aromatic rings and their internal energies are much too large (Appendix C, Figure C.4). This happens despite the two molecular file formats containing identical coordinates and atom types. The reason is that two different molecules end up being loaded by the RDKit: one is correct, and the other is incorrect. The wrong chemical representation causes a faulty parametrisation of the UFF in the energy ratio test, leading to a downstream error caused by the file being misinterpreted by the RDKit. Clearly, the chemical representation of the molecules matters a lot, and much care must be taken to load molecules correctly with the RDKit, because learning from chemically faulty data possibly limits the methods' performance. In summary, special care must be taken when loading molecular files with the RDKit - some substructures, such as four valent nitrogens, are never correctly loaded and could lead to systematic blind spots.

Obtaining the latest data from the PDB in mmCIF format and regenerating the data set, as was done for the Astex Diverse, was successful. The molecular structures loaded from the original files with RDKit were only 60 % PB-valid, but the recreated Astex Diverse set was 92 % PB-valid. It is apparent that high quality filters (*e.g.* maximum resolution of 2.5 Å), for example those used to select structures for the Astex Diverse set, the CASF-2016 set, and the PoseBusters Benchmark set, did indeed yield higher quality data selections, even though the PoseBusters tests were not used directly as criteria.

As the PDB continually expands and the RDKit is continually improved, regenerating the data sets periodically will be important.

For future work, it would be interesting to study the diversity and uniqueness of the data. The PDBbind 2020 General contains 19 443 protein-ligand complexes from the PDB, of which 155 are redundant leaving 19 288 complexes. For training generative models, it is important not only to have diverse data that do not repeat, but also to have data that cover as much of the input space as possible. Describing the distribution of the data in terms of diversity measured by protein sequence identity or ligand similarity will be important.

Further work for understanding the trade-off between using all data and selecting only high quality data would be a useful step. This is because showing that the training data have the same failure modes as the models is only a necessary but not sufficient condition to show that the models have learnt the problems from the data. Retraining the models with cleaner data would allow this to be measured. It would be interesting to see what effect retraining the DL-based docking methods on cleaner data sets would have. This would give a signal of whether efforts to improve the structure models of the PDB, via refitting the models on the electron densities, could be worthwhile for model developers.

3.5 Conclusion

In conclusion, the results showed that data sets of molecular structures obtained from the PDB should be carefully inspected, filtered, and errors corrected. Although the failure modes of the DL-based docking methods resembled those of the training data, I find that the training data were of better quality than the model predictions would make us believe. To fully discern the influence of problematic data on model performance, the models could be retrained on filtered versions of the data set and their results be compared. Further, I recommend that all practitioners benchmark and carefully select the training data they choose to build DL-based docking models, maybe less is more in some cases.

Chapter 4

Mining the PDB for guidance

In this chapter, I explore common patterns in structure-based drug design using a large collection of protein-ligand complex pairs from the PDB. The chapter describes the process of selecting high-quality complexes from the PDB, aligning the pairs of PDB structures around ligands bound in the same binding site of the same protein receptors, the measurements calculated on the structures, and the patterns observed in the data. The chapter provides extensive experimental evidence for an important concept in medicinal chemistry, which is that molecules with similar structures have similar binding modes. This provides a starting point for the guidance of docking models introduced in the following chapter.

4.1 Introduction

Structure-based drug discovery often relies on the tacit knowledge acquired by practitioners over multiple projects with specific protein-ligand systems. This knowledge includes the binding modes of existing lead series, the flexibility of the protein and its binding pocket, and the effect of structural changes in the ligand on binding mode and affinity. Central to this is the dogma that similar molecules with similar structures have similar properties (Deane, 1990). Molecular modellers and medicinal chemists with relevant experience can recognise familiar patterns in interactions and substructures, which they use to generate new ideas to improve the binding and properties of lead compounds (Bissantz et al., 2010).

The medicinal chemistry literature makes some of this knowledge explicit. Individual case studies of success stories in structure-based drug discovery include those of approved drugs such as vemurafenib (AlKharboush et al., 2025) and current candidates in clinical studies such as AMG 510 (Lanman et al., 2020) and are important reading for practitioners working on the same targets.

Data science approaches have also been applied to make medicinal chemistry knowledge explicit in the context of structure-based drug discovery. For example, Ferreira de Freitas and Schapira (2017)

studied the frequency distribution of non-covalent interactions in a set of 11 016 PDB structures and highlighted how the addition of one interaction may cause structurally similar ligands with conserved binding poses to differ significantly in binding affinity. Malhotra and Karanicolas (2017) found that only in 14 % of 297 pairs of assembled smaller and larger ligands the binding mode definitely changes upon chemical elaboration, and in many of these the change is due to an expansion of the smaller molecule of the pair into the space occupied by the receptor. Giangreco et al. (2013) collected 121 molecular overlays from the PDB that support the pharmacophoric hypothesis, namely that structurally diverse molecules with chemical features arranged similarly in space can bind to the same target. In the context of fragment-based drug design, Drwal et al. (2018) found that in a set of 3672 structures containing fragments and their drug-like superstructures, smaller chemical changes lead to retention of the binding mode more frequently than larger ones.

Although such insights are valuable and can guide practitioners, this is familiar knowledge to expert medicinal chemists. What such experts need are the right tools to put their structural hypotheses to the test when optimising compounds. To this end, classical pose prediction tools allow users to bias the predictions (Paggi et al., 2024). For example, the Python API of the docking program GOLD (Verdonk et al., 2003) allows the specification of distance constraints for individual atoms. However, many recently developed DL-based docking tools still lack such a capability.

In order to address this limitation, and in preparation for training and testing DL-based methods on the structure-based hypotheses that medicinal chemists employ, I set out to collect a large number of pairs of ligands bound to the same binding site from the PDB. Although I could start with a data set curated for studying protein-ligand interaction patterns and binding modes, such as the sc-PDB (Kellenberger et al., 2006) which contains 16 034 carefully protonated and chemically-complete complexes in its latest iteration (Desaphy et al., 2015), I start with the PDB as it is the largest and fastest growing data set of experimentally resolved protein-ligand structures. By the end of 2017, which is the year of the latest release of the sc-PDB, the PDB contained 111 874 structures obtained by X-ray crystallography and this number has nearly doubled and currently contains 246 715 structures, as noted in Chapter 1, including more than 450 000 protein-ligand interactions (Durairaj et al., 2024).

In this chapter, I explore the data set with a hypothesis-driven data science approach to test whether the data make sense and to quantify the degree to which molecules with similar structures have similar binding modes. In the next chapter, I will test and develop machine learning-based drug development tools using these data.

4.2 Methods

This chapter’s methods are organised into four parts: the selection of protein-ligand complexes from the PDB (Section 4.2.1), the pairing and aligning of these structures (Section 4.2.2), clustering (Section 4.2.3), and the metrics used to describe the final data set (Section 4.2.4). The Python library RDKit (Landrum et al., 2025) was used for all cheminformatics tasks unless otherwise specified.

4.2.1 Selection and preparation of protein-ligand complexes

First, I selected all high-quality structures of protein-ligand complexes from the protein data bank (PDB) (Berman et al., 2000) via the PDB search API, selecting structures that were solved using X-ray diffraction crystallography, contained at least one non-polymer entity, had at least one protein chain of length 30 amino acids, were deposited between 1 Jan 2000 and 30 June 2025, and had a resolution of 2.5 Å or better, a cut-off used in previous works (Ferreira de Freitas & Schapira, 2017; Giangreco et al., 2013; Kellenberger et al., 2006).

The PDB’s Chemical Component Dictionary (CCD) (Westbrook et al., 2015) was used to filter the structures further. Selected ligands were those with 3 to 55 non-hydrogen atoms, at least one carbon atom, and containing only the elements H, B, C, N, O, F, P, S, Cl, Br, and I. Ions, simple sugars, amino acid residues, nucleotide residues, buffers, precipitants, reducing agents, solvents, unidentified ligands, and common cofactors were excluded using a combination of block lists (Jubb et al., 2017; Radoux et al., 2016; Shao et al., 2022; Verdonk et al., 2008; J. Yang et al., 2012) (Section D.1 lists the combined block list CCD identifiers). Lastly, ligands for which neither the SMILES or InChI record in the CCD yielded a valid RDKit molecule object were removed.

Next, I filtered the structures using the PDB’s ligand quality reports and selected only ligands with a real space *R*-factor of at most 0.4 to avoid ‘poor fit or low data resolution’ (Smart et al., 2018), a real space correlation coefficient of at least 0.95 to select structures which ‘normally indicates a very good fit’ (Smart et al., 2018), a 100 % complete ligand model, and no stereochemical errors. Ligands without available quality reports were also removed.

Next, I downloaded the structure files (mmCIF format) from the PDB and prepared the structures in PyMOL (DeLano, 2002). For each structure, the biologically relevant assembly was selected and all possible alternative structures were enumerated. One structure per PDB and CCD identifier combination was selected by choosing the first, if available, structure for which any heavy ligand atom is at most within 0.2 Å of any other atom; not within 5 Å of any unknown marked atoms, maximally within 5 Å of any protein atom, and not within 20 Å of any atom belonging to a nucleic acid polymer. At this stage, covalent ligands, *i.e.*, those with a bond between the ligand and any polymer and crystal

contacting ligands were removed. Crystal contacting ligands were identified by generating all of a structure's symmetry mates with any atoms up to within 10 Å of the symmetry unit, and the rule that all of the heavy ligand atoms must be at least 5 Å away from any of the protein's crystal symmetry mates. Lastly, only complexes whose ligands passed all the sanitization checks of the RDKit were kept. All processed structures were saved as mmCIF files removing alternate location markers.

Next, I performed post-processing of all structures. Missing atoms and protonated residues were added using PDBFixer (Eastman et al., 2017) and ligands using OpenBabel (O'Boyle, Banck et al., 2011) at the physiological pH of 7.4. The binding site, that is the protein residues and metal ions within 8.0 Å of the ligand, was saved to PDB files¹, capping amino acids with the uncharged COOH and NH₂ groups at all places where peptide bonds were broken. The ligands were stored separately in SDF files. The hydrogen coordinates were optimised using an energy minimisation with the MMFF94s (Halgren, 1996) force field, designed for recreating time-averaged experimentally determined structures, and available in the RDKit (Tosco et al., 2014). Complexes for which the ligand and surrounding binding pocket could not be processed and extracted were dismissed.

4.2.2 Pairing and alignment

The structures were paired by protein chains that share a PDB 70 % cluster identifier², a UniRef 90 % cluster identifier (Suzek et al., 2015), or UniProt accession number (Bateman et al., 2024) that had small molecule ligands bound to the same binding site. Ligands within 5 Å after the following alignment are considered to bind within the same binding site. I aligned the protein chains using PyMOL's align algorithm which performs a sequence alignment followed by a structural superposition. The structural superposition was performed on the α carbons of the paired chains. When multiple chain mappings or alternate structures were available for a pair of PDB complexes, only the alignment resulting in the highest overlap of the two ligands, measured with the RDKit shape and colour score SC_{RDKit} , was chosen. The alignment matrices were stored for further analysis. Only one pair of structures per PDB CCD—PDB CCD identifier combination was chosen, namely the pair with the highest SC_{RDKit} score.

4.2.3 Clustering by sequence identity and ligand similarity

I created ligand and receptor clusters to test whether there are groups of receptors and ligands for which larger numbers of pairs were created than for others and how such clusters would affect our results. For receptors, the PDB 90 % sequence identity clusters of the chain closest to the ligand were used. For the ligands, Butina clusters (Butina, 1999) were calculated using ECFP4 2048-bit fingerprints.

¹PDB files can store the required residue information and the RDKit can write and read them.

²The PDB provides protein chain clusters computed by the MMSeqs2 algorithm (Steinegger & Söding, 2017).

4.2.4 Measurements and annotations

For all pairs of ligands bound to the same protein binding pocket obtained, various metrics were calculated that describe the ligands, receptors, and their interactions. The metrics based on individual complexes are listed in Table 4.1. The ligand metrics included their molecular weight, number of heavy atoms, number of hydrogen bond acceptors and donors, number of rotatable bonds, and the quantitative estimate of drug-likeness (QED) score (Bickerton et al., 2012). Receptor metrics included the number of residues and the number of chains in the binding pocket. The interaction metrics included the number of interactions of each type obtained by the ProLIF interaction profiler (Bouysset & Fiorucci, 2021), including hydrophobic, hydrogen bond, ionic, π -ionic, π - π stacking and van der Waals interactions. I used ProLIF (Bouysset & Fiorucci, 2021) with more forgiving distance thresholds for hydrogen bonds (3.7 Å instead of 3.5 Å), cation- π interactions (5.5 Å instead of 4.5 Å), and ionic interactions (5 Å instead of 4.5 Å) following Errington et al. (2025) and also switched to using the default van der Waals radii defined in the RDKit. In addition, ligands were labelled according to the following definitions: A ligand is a ‘fragment’ if its molecular weight is at most 300 Da. A ligand is ‘drug-like’ if its QED score is at least 0.7.

The metrics based on pairs of complexes are listed in Table 4.2 and include similarity metrics on the structure (2D) and pose (3D) of the ligands, metrics on changes of the binding pocket, and interaction similarity metrics. The ligand similarity metrics included the difference in molecular weight, the difference in the number of heavy atoms, the Tanimoto (Jaccard) similarity on Morgan fingerprints with radius 2 (ECFP4 2048-bit vectors), and the number of atoms in the maximum common substructure (MCS). The MCS of the ligand pairs were identified using the RDKit’s flexible MCS (FMCS) algorithm with the default settings. The ligand 3D similarity metrics included the RDKit shape and colour score, SC_{RDKit} , identical to SuCOS (Leung et al., 2019), which is similar to the pose similarity metric COS used by Malhotra and Karanicolas (2017). The similarity metrics on binding pockets included the number of identical residues between the binding pockets, the RMSD of the C_α atoms of the conserved residues, and the RMSD of the side chains of the conserved residues. The interaction similarity metrics included the number of conserved ProLIF interactions and the Tanimoto (Jaccard) similarity of the ProLIF interactions. Because protein chains in PDB structures do not use a standardised residue numbering scheme, residues were mapped between structures using the sequence alignments generated previously for the structural alignments. The pairs of ligands were categorised according to the following definitions: A pair of ligands is *identical* if the ligand has the same CCD identifier in the PDB file or if the standard InChI strings (Heller et al., 2015) generated for the molecules are identical. A pair of ligands is an *elaboration* if the MCS is one of the ligands. Following Malhotra and Karanicolas (2017), a pair of ligand poses is *conserved* when SC_{RDKit} is at least 80 %.

Table 4.1: Metrics used to describe protein-ligand complexes.

Metric	Description
<i>Ligand description</i>	
MolWeight	Molecular weight in Da.
NumHeavy	Number of all heavy (non-hydrogen) atoms.
NumHAcc	Number of hydrogen bond acceptors.
NumHDon	Number of hydrogen bond donors.
NumRot	Number of rotatable bonds.
QED	Quantitative estimate of druglikeness (QED) score.
IsFragment	Molecular weight is at most 300 Da.
IsDrugLike	QED is at least 0.7.
<i>Receptor description</i>	
NumRes	Number of protein residues within 6.0 Å of the ligand.
NumChain	Number of protein chains within 6.0 Å of the ligand.
<i>Interaction description</i>	
NumHydrophobic	Number of hydrophobic interactions.
NumHBondAcc	Number of hydrogen bonds where ligand is acceptor.
NumHBondDon	Number of hydrogen bonds where ligand is donor.
NumHBond	Number of hydrogen bonds. Sum of NumHBondAcc and NumHBondDon.
NumCationic	Number of ionic interactions with a cation.
NumAnionic	Number of anionic interactions with an anion.
NumIonic	Number of ionic bonds. Sum of NumCationic and NumAnionic.
NumCationPi	Number of cation- π interactions.
NumPiCation	Number of π -cation interactions.
NumPiStacking	Number of π - π stacking interactions.
NumVDWContacts	Number of van der Waal contacts.

Table 4.2: Metrics used to describe pairs of protein-ligand complexes.

Metric	Description
<i>Ligand chemical similarity</i>	
NumHeavyDiff	Absolute difference in the number of non-hydrogen atoms.
NumMCS	Number of atoms in the maximum common substructure.
FingerprintSim	Tanimoto similarity of ECFPs (2048 bits, radius 2).
IsIdentical	The ligands either have the same CCD identifier or the same standard InChI string.
IsElaboration	One ligand is the substructure of the other.
<i>Ligand pose similarity</i>	
SC _{RDKit}	Pharmacophoric feature and shape overlap score of the ligand poses.
MCSRMSD	RMSD of maximum common substructure.
PoseConserved	SC _{RDKit} is at least 0.8.
<i>Pocket pose similarity</i>	
NumRes	Number of identical binding site residues.
PocketScRMSD	Mean RMSD between the identical pocket side-chains.
PocketCaRMSD	Mean RMSD between the alpha carbons of the identical pocket residues.
<i>Interaction similarity</i>	
SimHydrophic	Tanimoto similarity of hydrophobic interactions per residue.
SimHBondAcc	Tanimoto similarity of hydrogen bonds where ligand is acceptor per residue.
SimHBondDon	Tanimoto similarity of hydrogen bonds where ligand is donor per residue.
SimHBond	Tanimoto similarity of hydrogen bonds per residue.
SimCationic	Tanimoto similarity of ionic interactions with a cations per residue.
SimAnionic	Tanimoto similarity of anionic interactions with anions per residue.
SimIonic	Tanimoto similarity of ionic bonds. Sum of NumCationic and NumAnionic per residue.
SimCationPi	Tanimoto similarity of π -cation interactions per residue.
SimPiCation	Tanimoto similarity of cation- π interactions per residue.
SimPiStacking	Tanimoto similarity of π -stacking interactions per residue.
SimVDWContacts	Tanimoto similarity of van der Waal contacts per residue.

4.3 Results

Overall, 232 665 unique pairs of PDB structures with ligands bound to the same protein and the same binding pocket were collected, and 53 048 (23 %) of these are pairs of identical ligands bound to the same receptor (but still in different PDB structures). Of the 179 617 pairs containing non-identical ligands, 99 319 (55 %) pairs contain fragments, 3698 (2.1 %) pairs are elaborations, and 2587 (1.4 %) contain fragment elaborations. Further, 92 255 (51 %) of the non-identical pairs contain drug-like molecules and 46 159 (26 %) pairs contain a pair of a fragment and a drug-like ligand.

4.3.1 Correlation between 2D and 3D similarity

Binding poses are more commonly conserved in pairs of chemically similar ligands. Figure 4.1 shows the shape and pharmacophoric similarity, measured by SC_{RDKit} , of all pairs of non-identical pairs of ligands (179 617) versus chemical similarity, measured via chemical fingerprints. The plot shows that, for more chemically similar pairs of ligands, the binding pose is conserved more commonly. It also shows a high similarity of H-bond interactions (yellow) for chemically dissimilar ligands. Figure 4.2 shows the same set of non-identical pairs (179 617) binned by chemical similarity and in terms of the percentage of poses conserved ($SC_{RDKit} \geq 0.8$). This plot shows the trend more clearly; clearly, there is a steady increase in the percentage of poses conserved as a function of chemical similarity. For ligand pairs with chemical similarity larger than 80 %, around 90 % of the observed pairs in the data set have conserved binding modes.

4.3.2 Error analysis on identical pairs

Of the 53 048 pairs of identical ligands, 99.0 % are pairs that conserve the binding mode ($SC_{RDKit} \geq 0.8$) and 1.0 % (527) do not conserve binding mode ($SC_{RDKit} < 0.8$). Six of those pairs of structures containing identical ligands (identical CCD identifiers) with non-conserved poses are shown in Figures D.3 and D.4. Three of these poses are distinctly different (Figures D.3a and D.4c and Section D.4). The publications of the PDB structures reveal that ligand DGT in structures 4TNX and 4TO6 is an allosteric binder, ligand DIF binds is shown in different species (6HN0 is ovine and 6HN1 is caprine serum albumin), and ligand OCF receptors 8AYJ and 8ONN differ by a point mutation. This might explain some pairs of poses scoring lower than expected (Figure D.3b). Similarly, three of the pairs of ligands with identical CCD identifiers have non-identical Fingerprints (2048-bit ECFP4 Tanimoto similarity less than 1, for example Figure D.4c). The ECFPs (2048-bit, radius 2) used here also distinguish ligands with different protonation patterns, resulting in some ligand pairs with identical CCD identifiers scoring less than one in terms of fingerprint similarity.

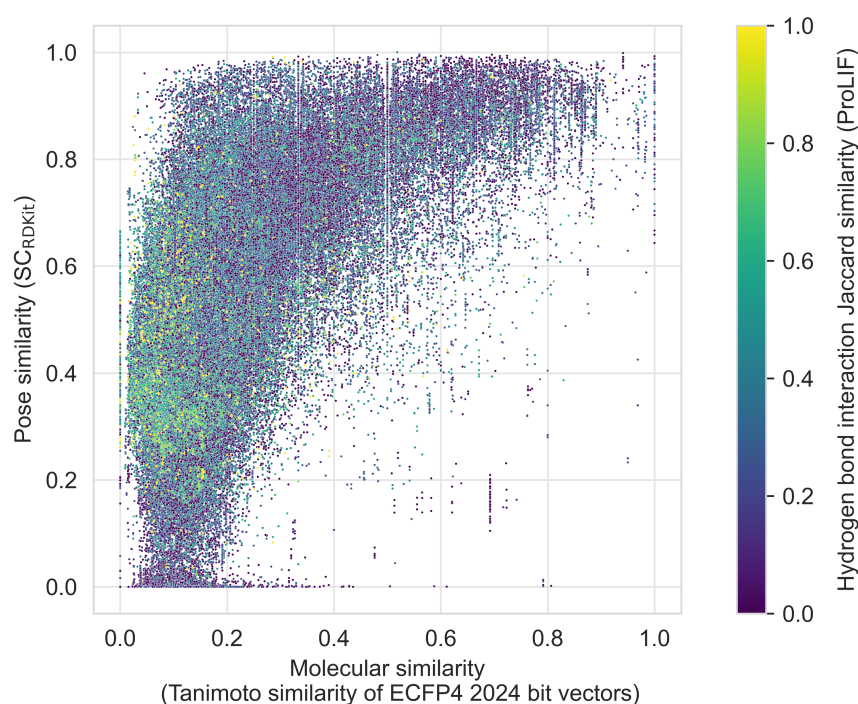


Figure 4.1: Scatter plot of all pairs showing pose similarity versus molecular similarity. The pose similarity is measured by SC_{RDKit}, the molecular similarity by ECFP (2048-bit radius 2) Tanimoto similarity, and the colours show the Tanimoto similarity of Hydrogen bonds per residue. Molecules with similar molecular structures tend to have similar binding modes.

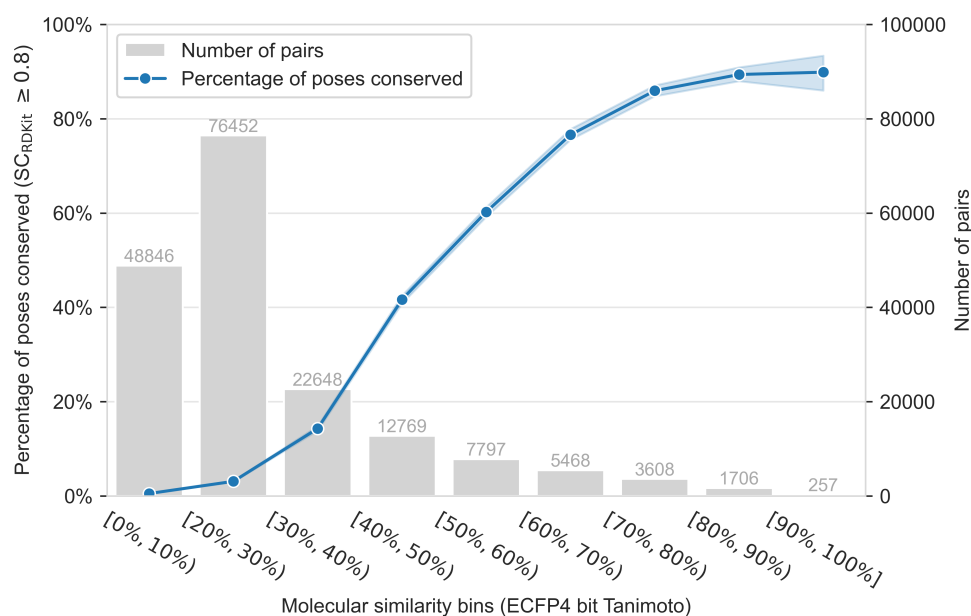


Figure 4.2: Percentage of conserved poses amongst all pairs relative to the molecular similarity of the ligands. This plot aggregates the data in the previous plot after binning. Conserved poses are those with SC_{RDKit} ≥ 0.8 and chemical similarity bins are by ECFP (2048-bit radius 2) Tanimoto similarity. The plot shows that the binding poses are conserved in about 90 % of the 257 pairs of protein-ligand structures where ligands have a similarity of at least 90 %. Molecules with similar molecular structures tend to have similar binding modes.

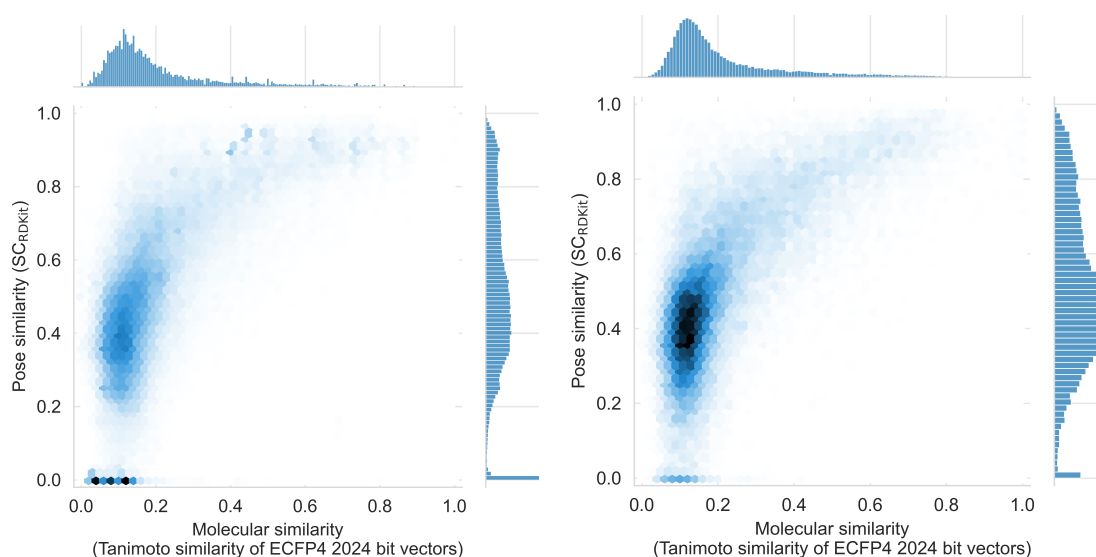


Figure 4.3: Pose similarity versus chemical similarity for all pairs (left) and after clustering (right). Pairs of protein-ligand complexes were clustered by receptor sequence similarity (PDB sequence clusters) and ligand chemical similarity (Butina clusters). For each cluster of ligands, the mean value of the pose similarity and fingerprint similarity was taken before binning. The clustering smoothens the overall distribution but the observed trend is the same after clustering.

4.3.3 Clustering by sequence identity and ligand similarity

Clustering pairs of protein-ligand complexes by chemical similarity and protein identity conserves the correlation between chemical similarity and binding mode conservation. Figure 4.3 shows pose similarity versus molecular similarity before and after clustering, where the metrics were averaged over protein and ligand clusters. The histograms on the margins and the hexbin plot appear smoother after clustering, which indicates that there are a few large clusters. The trends in the data distributions appear to remain the same. Performing the same analysis for chemical versus pose similarity as for all pairs results in Figure 4.4 which shows the percentage of poses conserved after clustering. The line for the non-clustered data that is shown in Figure 4.2 is shown as reference. Clearly, the correlation between chemical similarity and binding mode conservation is the same.

4.3.4 Fragment binding mode conservation

Fragments, typically small organic molecules with molecular weights less than 300 Da, are sometimes used as starting points for drug discovery. It has been reported that fragments can adopt multiple poses in the same binding site. Plots showing the distributions of pose similarity for identical ligands (53 048) split into fragments and non-fragments are shown in Figure 4.5. They show no difference between fragments and non-fragments in this subset, 99.7 % of fragments, and 99.4 % of non-fragments conserve their pose. For the non-identical pairs, the percentage of conserved poses versus chemical

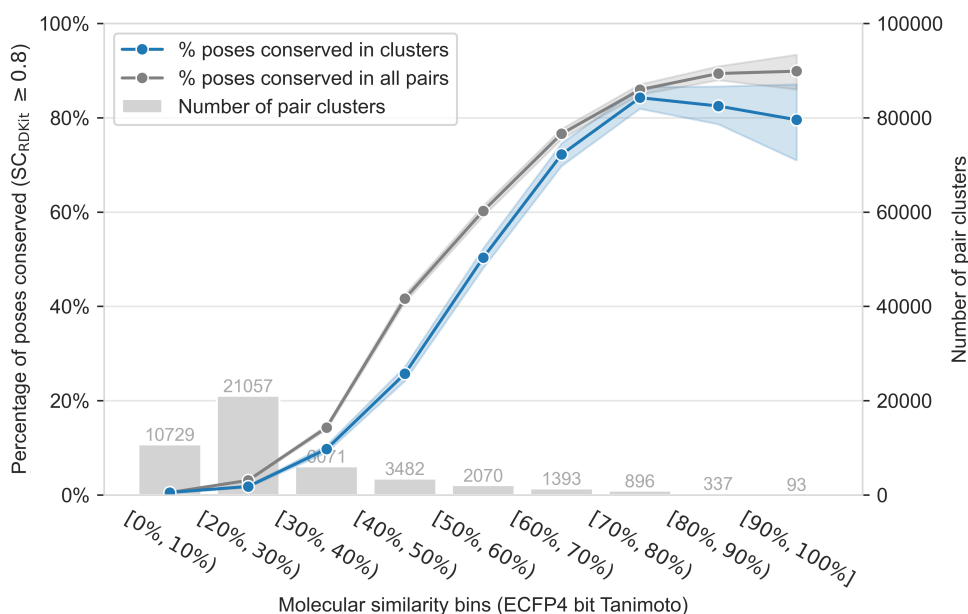


Figure 4.4: Pose conservation as a function of chemical similarity calculated on the clustered pairs. Pairs of protein-ligand complexes were clustered by receptor sequence similarity (PDB sequence clusters) and ligand chemical similarity (Butina clusters). For each cluster of ligands, the mean value of the pose similarity and fingerprint similarity was taken before binning. The overall trend is the same for the clusters than the trend for all pairs.

similarity is plotted in Figure 4.6; the line for all pairs shown in Figure 4.2 is provided for reference in grey. For chemical similarity between 20 % and 70 %, there is slightly higher pose conservation than in all pairs, whereas in between 70 % and 100 % there is lower pose conservation. Overall, the results do not show significant differences in fragments versus all pairs in the collected data.

4.3.5 Binding mode conservation upon elaboration

Malhotra and Karanicolas (2017) found that in 86 % of 297 pairs of assembled smaller and larger ligands the binding mode is conserved upon chemical elaboration. Figure 4.8 shows pose conservation as a function of chemical similarity for ligand pairs where one ligand is the substructure of the other. Across all bins of chemical similarity, pose conservation is higher than for all pairs, as shown by the reference line for all pairs. For elaborations with chemical similarity of at least 50 %, more than 80 % of poses are conserved. Next, I tested whether elaborations starting with fragments behave differently. As shown in Figure 4.7 this is not the case; elaboration and fragment elaborations preserve poses at similar rates. For both of these analyses, the number of pairs for the bins of high chemical similarity is relatively low, with less than 30 examples of fragment elaborations of chemical similarity of at least 80 %. One common elaboration strategy is to grow molecules a few atoms at a time. Figures 4.9 and 4.10 shows the proportion of conserved poses as a function of the difference in the number of

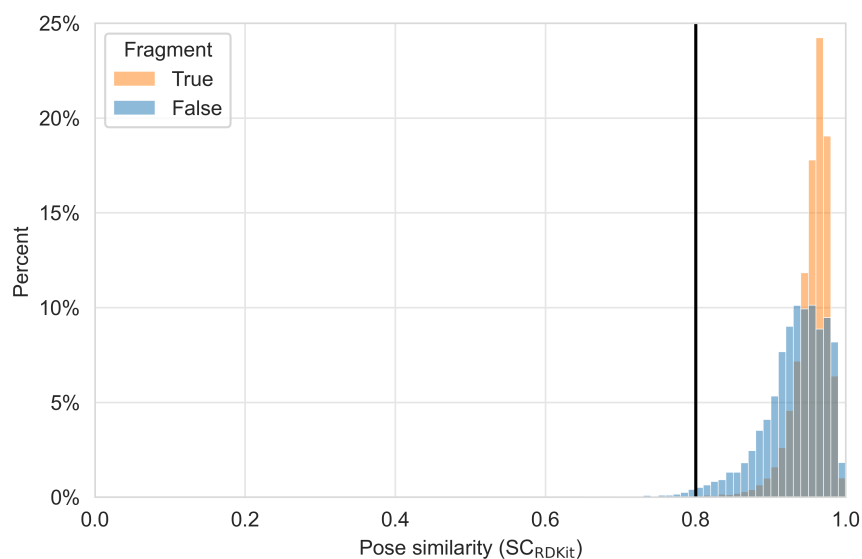


Figure 4.5: Distributions of pose similarity of identical ligands split into fragments and non-fragments. Pose similarity is measured in SC_{RDKit} for pairs of identical ligands (53 048) out of which 70.9 % are fragments and 29.1 % are non-fragments. Each distribution is normalized such that the length of the bars adds up to 1. The results illustrate that 99.7 % of fragments and 99.4 % of non-fragments conserve their pose ($SC_{RDKit} \geq 0.8$, as shown by the vertical black line). This confirms the consistency of binding events observed in crystal structures.

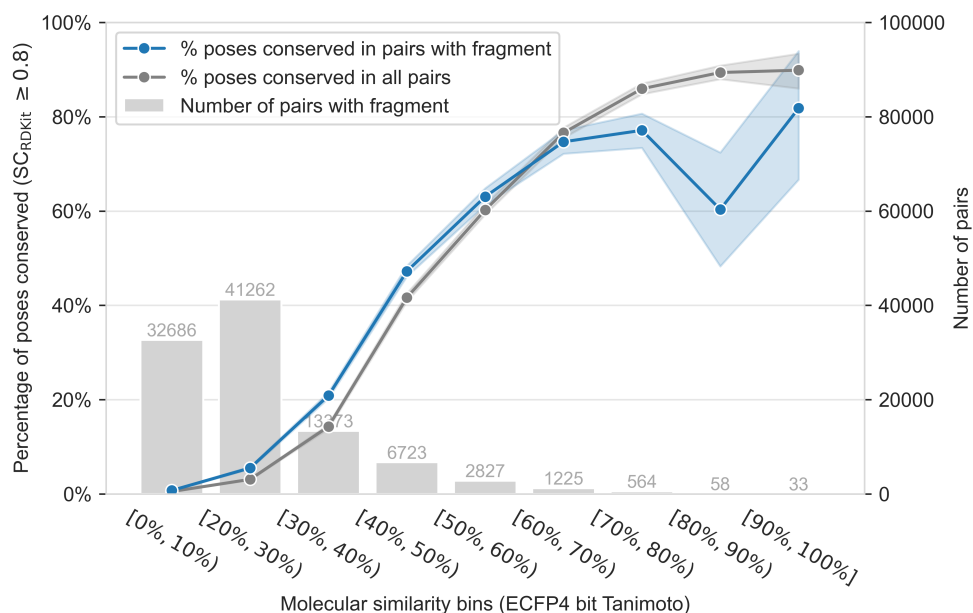


Figure 4.6: Pose conservation of protein-ligand structure pairs with at least one fragment as a function of chemical similarity. This blue line shows the subset of non-identical pairs of ligands where one is a fragment (molecular weight of at most 300 Da), and the gray line shows the line of all pairs that was shown in Figure 4.2. Approximately, pairs with fragments show similar binding modes at the same rates as those without fragments.

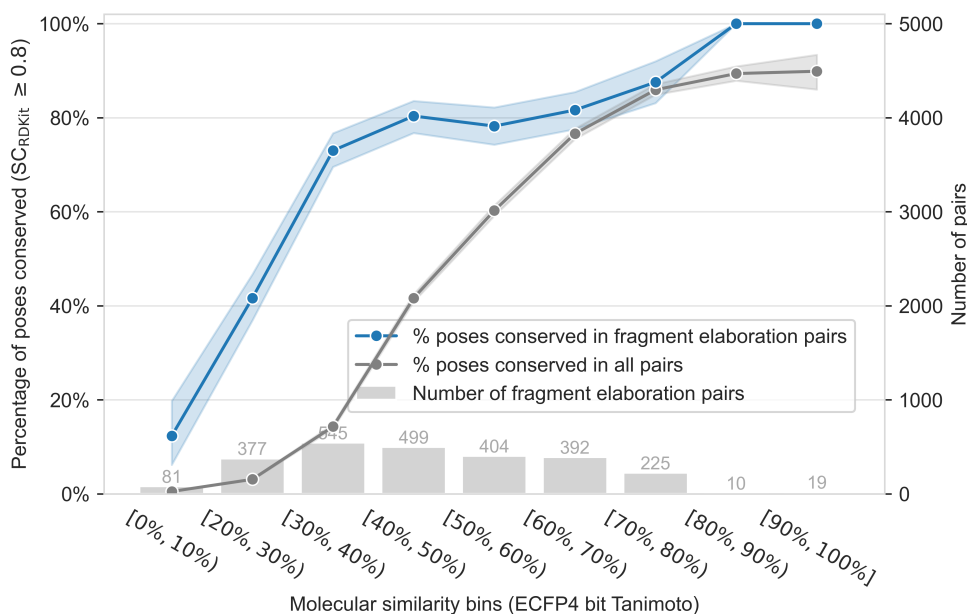


Figure 4.7: Pose conservation as a function of molecular similarity for the subset of pairs of fragment elaborations. Fragment pairs are those with ligand of molecular weight of at most 300 Da. Fragment elaborations where some chemical similarity (at least 40 % ECFP4 2048-bit Tanimoto similarity) is maintain conserve their binding mode in more than 80 % of all observed cases.

atoms for all pairs in Figure 4.9 and for fragment elaborations Figure 4.10, for pairs with up to 20 atom-difference. For pairs with up to 13 atoms difference, the percentage of poses preserved is more than 80 %. It appears that the MCS between the smaller and larger elaborated ligand in our data set maintains its pose very frequently.

4.3.6 Beyond elaborations using maximum common substructures

There are many pairs in which atoms would be removed before others are added, and these are not captured in the elaborations. These pairs were analysed using their MCS. Figure D.2 shows the distributions and pairwise distribution of the absolute size of the MCS, the size of the MCS relative to each ligand, the ligand sizes and the RMSD of the MCS for all 168 630 pairs of non-identical, non-elaboration ligands with an MCS of at least 3 atoms. The pairwise plots show that larger MCS sizes correlate with higher pose similarity, but also that pairs with larger substructures have higher chemical similarity (Tanimoto similarity on ECFP4 2048-bit fingerprints). In addition, the atomic positions in MCS of larger size appear to be more conserved (the RMSD of the atoms in MCS is lower).

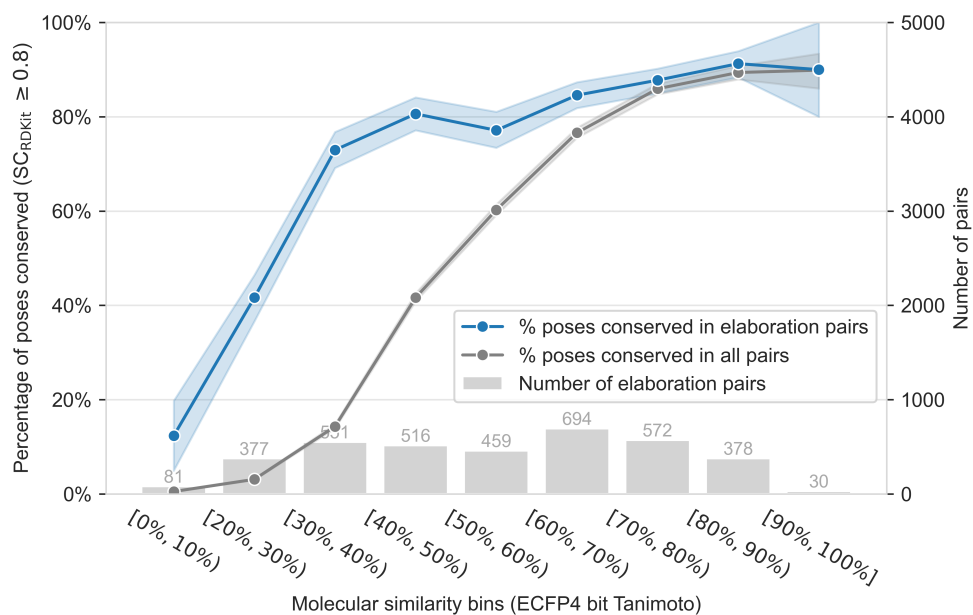


Figure 4.8: Subset of pairs of all elaborations. Elaborations are statistically highly conserved. For most cases where some chemical similarity is maintained (ECFP4 2048-bit Tanimoto similarity of at least 40 %), more than 80 % of observed elaborations maintain the binding mode.

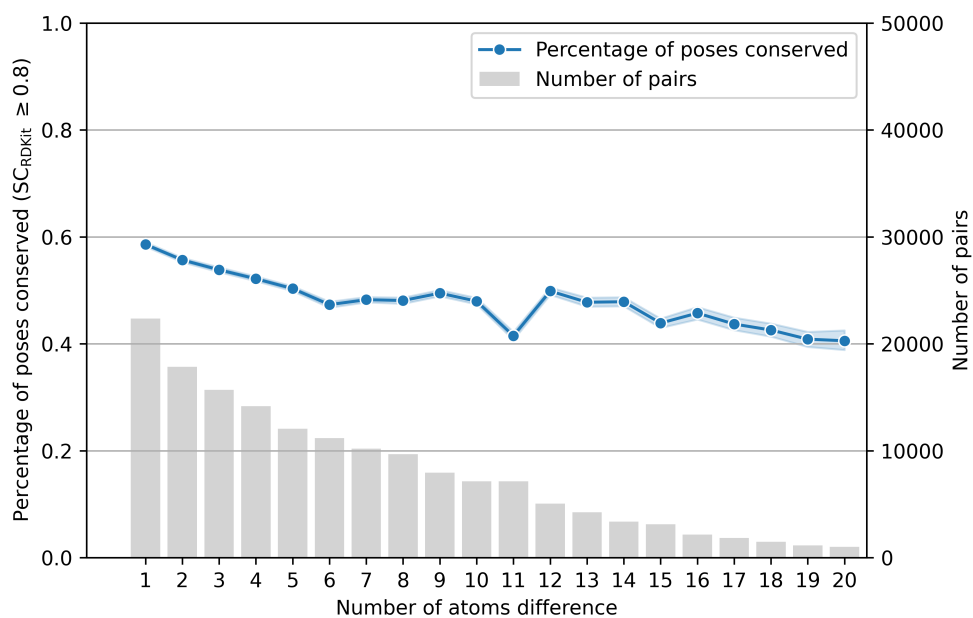


Figure 4.9: All pairs of non-identical ligands in the data set from 1 to 20 atoms difference. Overall, less than 60 % of pairs maintain their binding mode. Approximately for every atom added, the frequency of persevered binding modes drops by 1 %.

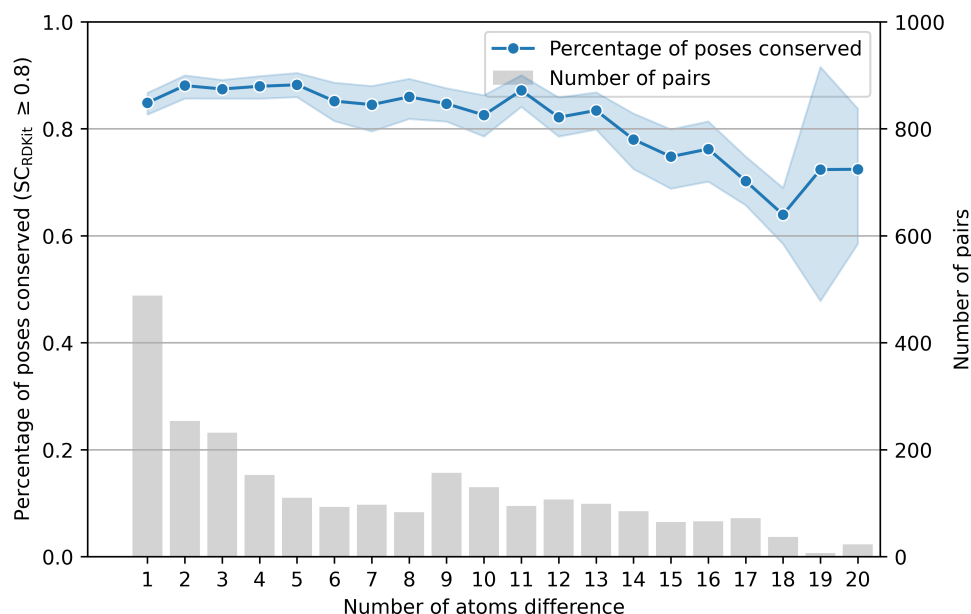


Figure 4.10: All elaboration pairs in the data set up to a 20 atom difference. More than 80 % of pairs with up to thirteen heavy atoms added maintain the binding mode. This is much higher than the rate for non-elaboration pairs Figure 4.9. Elaborations statistically maintain their binding modes.

4.3.7 Interactions replace colour and shape score

Interactions are another major consideration, in addition to pharmacophores, in structural medicinal chemistry. As a first check, I counted the number of interactions and the number of structures with interactions present for each interaction type (Table D.2 and Figure D.1). These values agree with the literature (Ferreira de Freitas & Schapira, 2017), validating the approach of adding hydrogens and optimising their positions. Next, the interaction similarity measured with the Jaccard similarity versus the pose similarity is shown as a set of bivariate histograms for each interaction type in Figure 4.13. The Spearman rank correlations are indicated in each plot. There are almost no positive correlations. Figure 4.14 shows the conservation of the hydrogen bond versus the chemical similarity, similar to before for SD_{RDKit} . There is a small correlation between the chemical similarity of ligands and the rate of conserved hydrogen bonds.

4.4 Discussion

My analysis shows that chemical similarity correlates strongly with pose conservation, as shown in Figures 4.1 and 4.2. This agrees with the approach often used in structure-based drug design. Small changes in molecular structures often maintain the binding modes of the original structures more often than larger ones. In particular, elaborations of molecules maintain binding modes very often; more than 80 % of all observed elaborations maintain the binding pose. This agrees with the main principle of

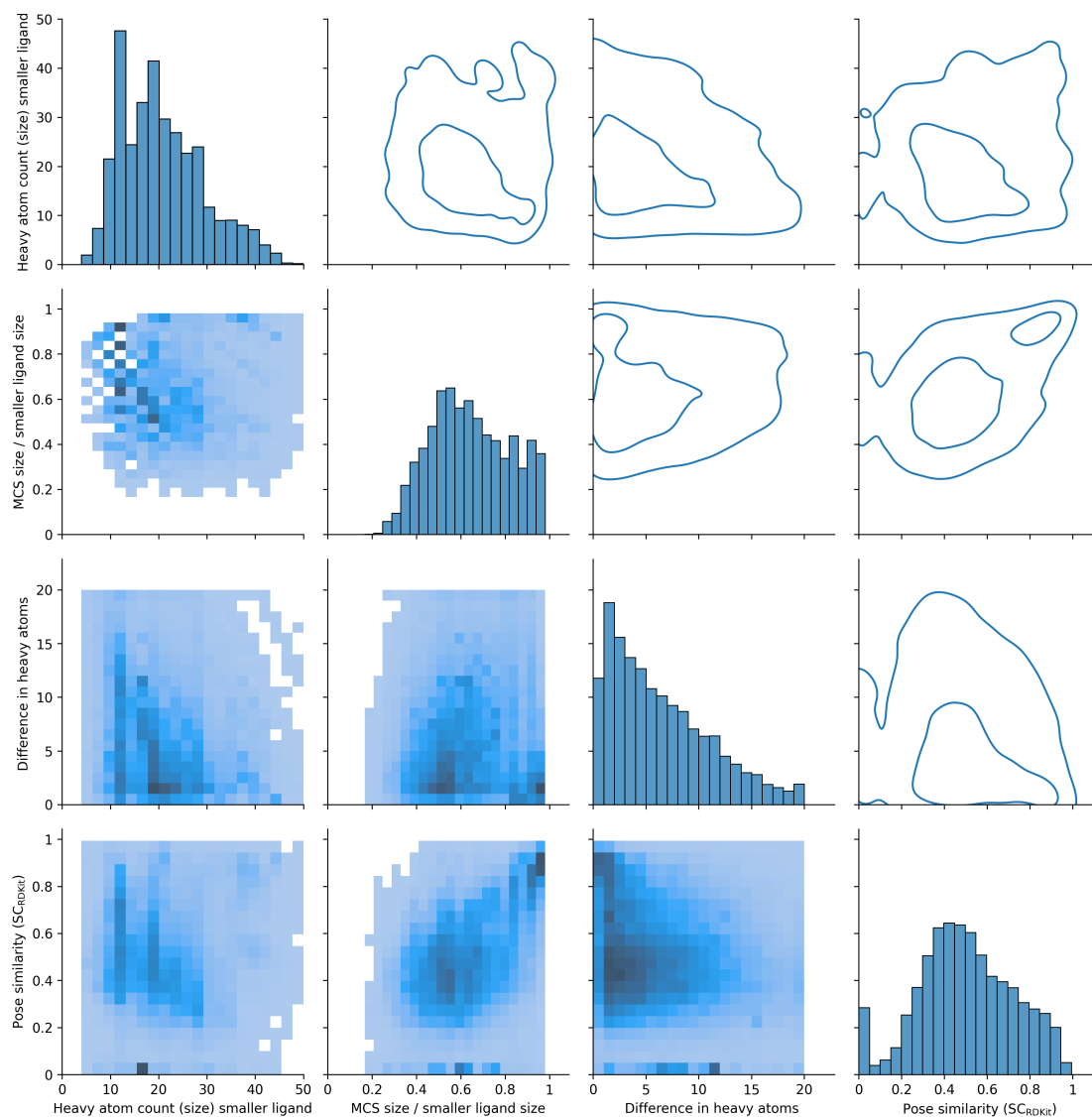


Figure 4.11: Joint plot for the dependent and independent variables used in the generalized additive model describing the influence of the MCS size on pose conservation. Histograms for each variable are shown on the diagonal. The bottom left (top right) off-diagonal panels illustrate the bivariate relationships between variables using histogram plots (contours of kernel density estimates).

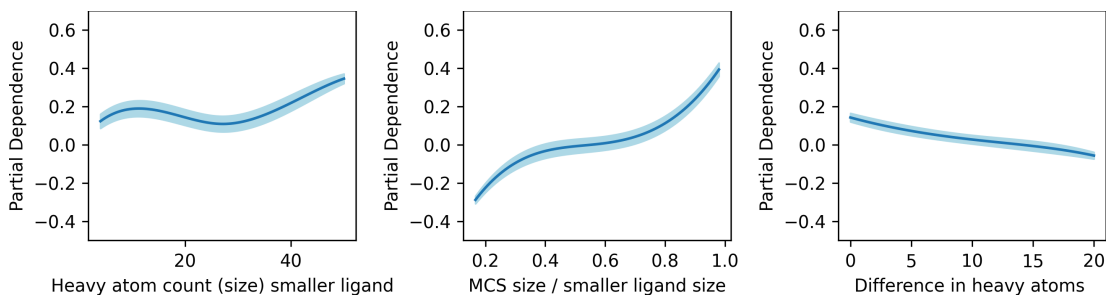


Figure 4.12: Partial dependence plots of the independent variables on the dependent variable of the generalised additive model that describes the relative influence of the MCS on pose conservation (measured by SC_{RDKit}). The shaded band shows the region between the 1st and 99th percentile. Larger common substructures and smaller differences in atom count are statistically associated with a higher conservation rate of binding modes.

structural medicinal chemistry, that is making small changes to lead compounds and existing scaffolds.

The statistic that 99.0 % identical ligands conserve their binding pose in the same pocket of different PDB structures confirms how consistent the binding events observed in the crystal structures are. However, the remaining pairs (1.0 %, 527) include pairs with significant changes in the binding mode when bound to the same receptor (Section D.4). Slightly different protonation patterns affected the ECFP4 fingerprints, but spot checks revealed that point mutations, difference in pH of the crystal, receptors from different species, and allosteric binding might cause these differences. The 90 % sequence identity threshold for pairing protein receptors does not filter pose-relevant point mutations.

Identical fragments have been observed elsewhere to adopt multiple binding poses (Malhotra & Karanickolas, 2017). Initially, perhaps surprisingly, in our data, pairs of identical fragments have a high rate (99.7 %) of pose conservation. However, this does not mean that fragments do not adopt multiple poses, because in our pair selection process, I only choose the combination of ligand poses that leads to the highest SC_{RDKit} . When there were multiple fragments in a binding site, I chose the pair with the pose that was the most similar. This explains the high rate of pose conservation in identical fragments. Fragment elaboration pairs showed a similar high rate of pose conservation as all elaboration pairs.

The trends observed for the MCS also agree with intuition: If more of a ligand is conserved, and if fewer atoms are added, then the more similar the smaller and larger ligand poses are. However, the fitted model on smaller ligand size, relative MCS size, and number of atoms added only explained about 33 % of the variance observed in pose similarity (Pseudo $R^2 = 0.33$). The missing variation might be explained by the amount of chemical complementarity removed and added, which was not represented in these models.

Intermolecular interactions are directly related to binding affinity. For this reason, medicinal chemists count interactions and add up their contributions when designing structures, as can be gleaned from average interaction strength tables (Burnell et al., 2016, Table 3.1, p. 32). The correlations

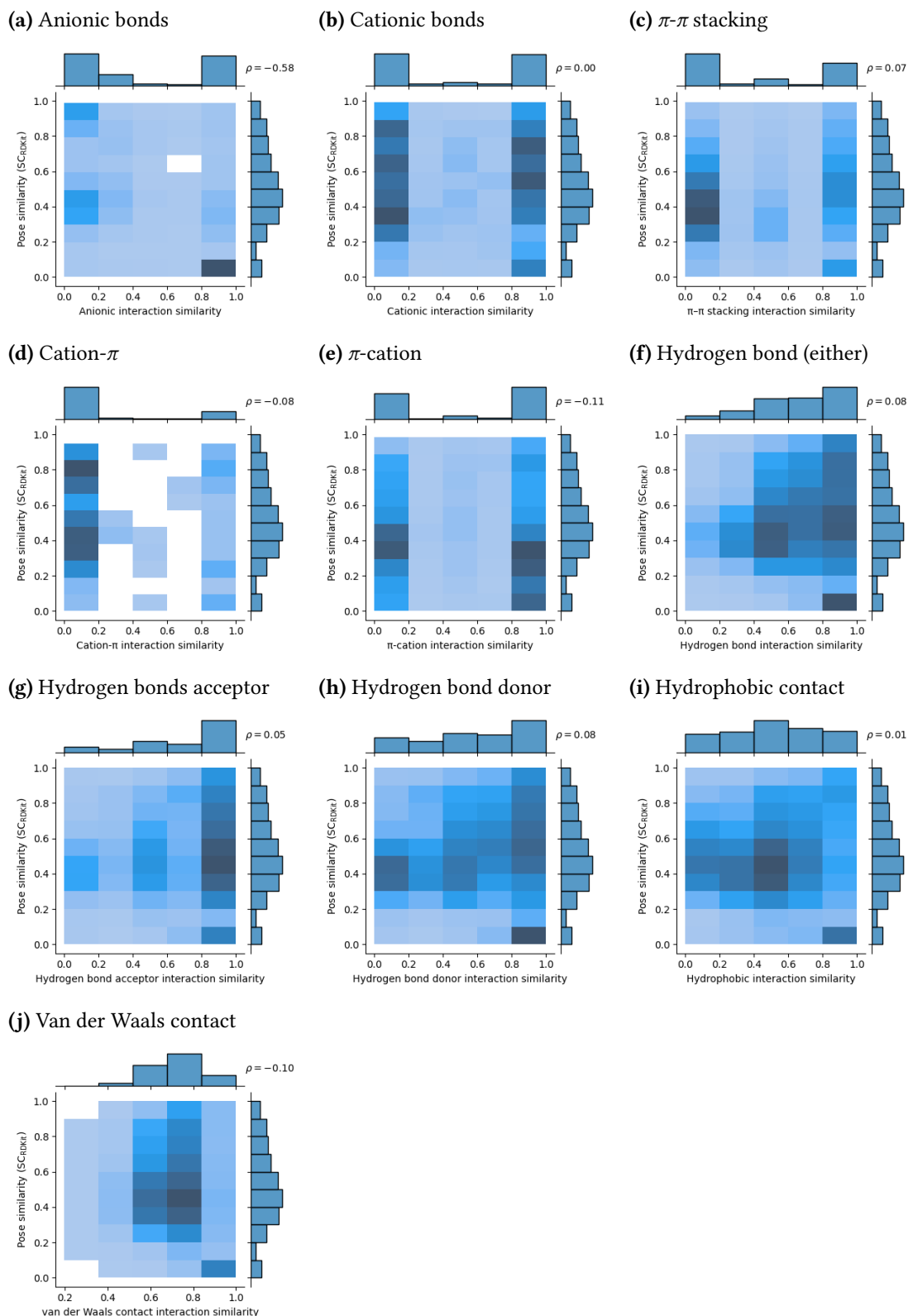


Figure 4.13: Interaction similarity (Tanimoto similarity of ProLIF interactions per residue) versus pose similarity (SC_{RDKit}) per interaction type. The interaction names are relative to the ligand, *i.e.* anionic means the ligand has the negatively charged side, cation- π means the ligand has the cation, H-bond acceptor means the ligand has the electronegative atom with the lone electron pair. Out of all interaction types, hydrogen bonds and van der Waals contacts correlate with pose similarity the most, but the correlation is small with a Spearman rank correlation coefficient of $\rho = 0.08$ for Hydrogen bonds.

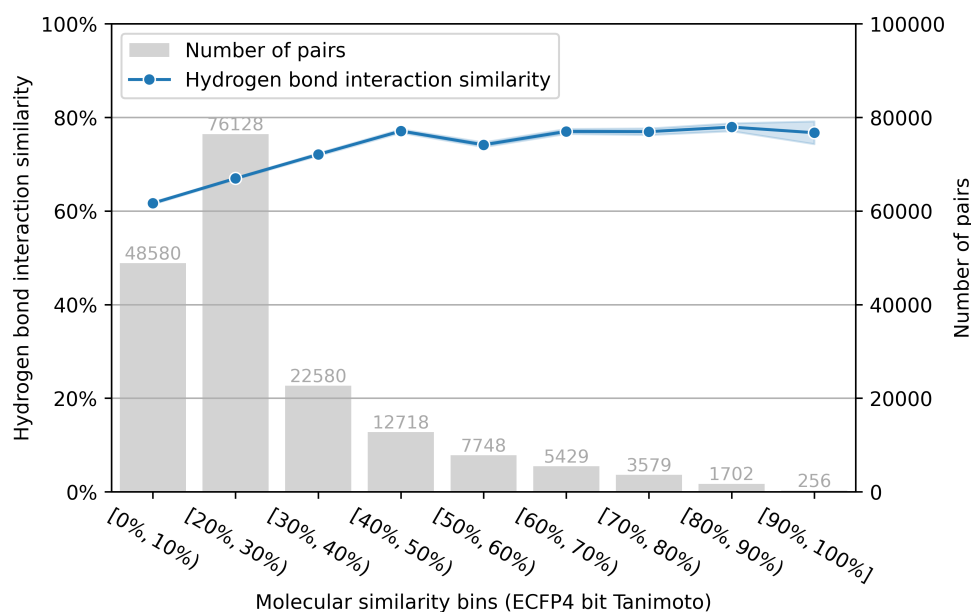


Figure 4.14: Conservation of hydrogen bonds (Tanimoto similarity of ProLIF interactions per residue) versus pose conservation (SC_{RDKit}). Overall, more than 60 % of hydrogen bonds are identical across all pairs of ligands. For chemically more similar ligands, this rises to about 80 %. Overall, hydrogen bonds are highly conserved, independent of ligand similarity.

between interactions and pose conservation were lower than expected. Almost no positive correlation is observed (Figure 4.13); the hydrogen bond interaction type has the highest Spearman correlation, which is almost negligible at $\rho = 0.08$. This suggests that pharmacophores and interactions contain orthogonal information and should both be considered when designing.

The data set used in this study has a *positive selection bias*, as I have selected pairs of protein-ligand complexes for which experimentally resolved structures of both ligands bound within the same pocket of the same protein exist. A modification of a molecule that failed to lead to a binding event in the same binding site (or did not bind at all) would not be represented in our data.

Another bias in our study is that the chemical space represented here is the one explored by structural experiments that led to the submission of structures to the PDB. These experiments do not explore the proteomic and chemical space uniformly, and only a very small subspace of combinations of ligands and proteins can ever be explored.

A further limitation of this analysis is that most PDB structures do not contain biologically accurate protonation patterns of receptors and ligands. Similarly, the hydrogen bonds in water networks of binding sites are mostly unresolved in experimental structures. Although I know that the solvation and protonation patterns strongly contribute to protein-ligand binding, I did not have the data to analyse these here.

4.5 Conclusion

In this chapter, I conducted a study of the protein-ligand binding patterns of pairs of structures of similar ligands bound to the same pocket in the same protein from the PDB. My analysis shows that ligands with high chemical similarity conserve binding modes more frequently and that elaborations, on average, conserve poses more frequently. Interaction metrics are found to correlate with chemical similarity, but not with pharmacophoric measurements. This chapter provided some insight into the types of biases that medicinal chemists use when performing structure-based drug design. It also provides a large data set for the study of protein-ligand patterns that can be further explored. In the next chapter, I use a diverse subset of these paired data to steer a DL-based docking model to conserve user-chosen interactions and substructures.

Chapter 5

Guiding diffusion for docking

Until now, I have explored protein-ligand interactions through docking models (Chapter 2) and paired crystallographic data (Chapter 4). Now we turn to applying our insights and improving deep learning (DL)-based docking. In this chapter, I start with the DL-based model that performed best in the benchmark in Chapter 2, namely DiffDock (Corso et al., 2023), and apply diffusion guidance to inject user constraints and physical information into the diffusion process. I first demonstrate on a series of test setups that this approach only produces predictions in the binding site when given the binding pocket location (a user constraint), that the force field information (physical information) improves physical plausibility, and that substructure and interaction constraints (user constraints) can be met with high precision. In our experiments, I also find that gradient-based guidance performs slightly better than gradient-free guidance. Next, a realistic docking benchmark is set up in which ligands cross-dock into cognate receptor structures of chemically similar ligands, as is custom in lead optimisation. I demonstrate the performance of our approach on this task using a new data set of public PDB structures and on a set of unpublished cereblon binding molecular glues with unpublished interactions and an unpublished target. The resulting method is named GuideDock.

5.1 Introduction

When medicinal chemists use docking for virtual screening, they formulate design hypotheses and select ligands accordingly; for example, they might select ligands that share a certain substructure with a lead molecule for which an experimental protein-ligand complex is available. For some hypotheses, the medicinal chemists might be able to define constraints; in the previous example, this would be to constrain the atoms of the shared substructures in the same positions during cross-docking. Such constraints allow a user to test their design hypotheses, but they might also lead to better predicted poses. Although classical docking methods such as AutoDock 4 (Morris et al., 2009) and GOLD (Ver-

donk et al., 2003) support atom-wise position constraints suitable for this example, dedicated methods such as EnzyDock (Das et al., 2019), EDock (W. Zhang et al., 2020), and FitDock (X. Yang et al., 2022) can extract these constraints from template ligands. DL-based docking models do not yet support any constraints, despite their potential practicality, and even though well-informed constraints can improve docking results.

To test how well DL-based methods might perform with constraints, I integrate constraints into the best DL-based model according to our comparison in Chapter 2. The chosen base model is DiffDock, an SE(3) Riemannian diffusion model. For diffusion models, in general, there are multiple approaches to add constraints. Due to computational constraints, I restricted myself to training-free approaches, which only modify the inference procedure and avoid the computationally costly training of models by re-using model weights fitted by others. Of course, this excludes approaches that modify the functional form of the model (Didi et al., 2023) or propose alternative loss functions (Didi et al., 2023). At the time of starting this work, training-free approaches had not yet been implemented for biological structural models, but I became aware of them in the domain of image generation (Bansal et al., 2023; Lugmayr et al., 2022; Ye et al., 2024) and I sought to adopt that approach for DiffDock. However, in the meantime, approaches similar to ours have been implemented in co-folding methods (Boitreau et al., 2024; Wohlwend et al., 2024).

Defining constraints when modelling biological structures is a well known procedure in the adjacent field of molecular dynamics. Here, distance constraints are often used to prevent or enforce interactions, while position constraints can be used to anchor a structure in place to prevent it from drifting out of the frame (Eastman et al., 2024). In molecular dynamics simulations, soft constraints are typically implemented as additional terms in the force field, for example, as an additional harmonic term with its minimum at a desired atom location. Force fields are differentiable scalar functions of the positions of the atoms, and they are parametrised using the molecular structures in the system of interest. The energy functions in classical force fields are linear functions that sum up terms of various energetic contributions such as bond stretching and steric clashes. Rather than just adopting the constraint terms, I build a system that can flexibly include any differentiable scalar function on atom positions, which includes all force fields by definition. Our motivation is two-fold. First, I hope that guidance with the non-constraint terms will help reduce the number of physically implausible predictions, a problem I identified in Chapter 2, and second, I hope that this setup will make our model easier to use, as users will be familiar with defining constraints on force fields.

To summarise, I tackle three use cases with inference time guidance, illustrated in Figure 5.1: The first is to enable binding site-specific docking with the blind docking model, the second is to allow users to define position and distance constraints, and the third is to improve the physical validity of the predicted poses.

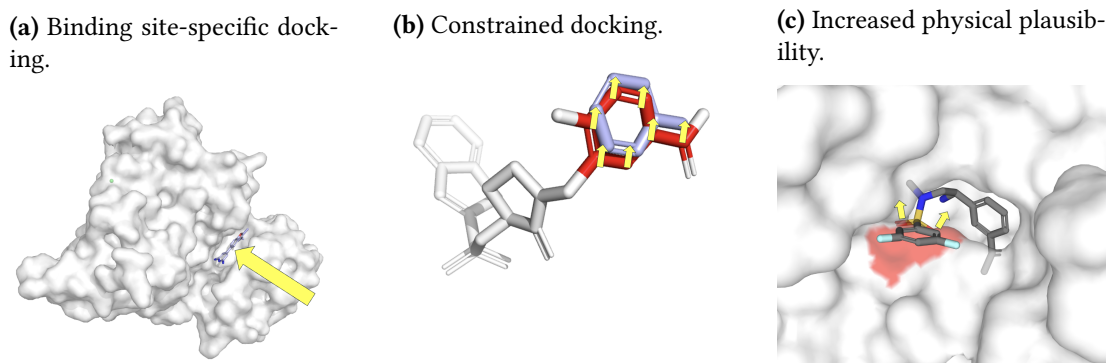


Figure 5.1: Illustration of the three main design goals of GuideDock. Diffusion guidance is used to enable binding site-specific docking, constrained docking, and improving physical plausibility using the Diffdock deep generative model for blind docking.

5.2 Methods

5.2.1 Base model

I choose DiffDock-L (Corso et al., 2024), the follow-up method to DiffDock that I benchmarked, as the base model. DiffDock-L is a blind DL-based docking method that decomposes the docking task like most classical docking methods: into a pose generation model for sampling poses, and a scoring model to rank the generated poses. During inference, the pose generation model is applied 20 times to sample and transform random noise into a ligand binding pose. This is done in batch to generate 20 ligand poses. The scoring model is then applied to rank these predictions.

The pose generation model is an $SE(3)$ -equivariant Riemannian diffusion model that defines a diffusion process over the roto-translations and torsions of the typical docking state-space, and it is trained by score-matching with an exploding variance schedule (Y. Song et al., 2021). The score model has a very similar architecture, but is trained as a classifier that predicts whether a generated pose is within 2 Å RMSD of the crystal pose. Both models use graph neural networks with equivariant tensor products (Geiger & Smidt, 2022). As input, the pose generation model only takes the alpha carbon locations of the receptor proteins into account, ignoring side chains and cofactors, while the scoring model evaluates the poses using all heavy atoms of the protein, also ignoring any cofactors.

Another important limitation, in addition to ignoring anything but protein chains, is the hard-coded limit of 1022 amino acids for the length of the receptor protein chains due to the period of rotary positional embeddings being 1024 for the ESM2 model (Lin et al., 2023), which is used to generate embeddings for residues.

In our benchmark of docking methods (Chapter 2), DiffDock was the best DL-based model. It showed high physical plausibility in terms of bond lengths, bond angles, and stereochemistry, but

protein-ligand clash rates were higher than those of classical docking methods and the main physical failure mode. DiffDock-L represents an improvement and has 27 % of poses within 2 Å RMSD and PB-valid, relative to 12 % of the poses of DiffDock. Figure E.1 shows a waterfall plot of the failure modes of DiffDock-L with the most prominent being protein-ligand clash, as it was for DiffDock (Figure B.4e).

In what follows, I apply guidance to the pose generation model of DiffDock-L to control the sampling and improve the quality of the samples generated. I make no alterations to the pose scoring model. For both models, I maintain the default inference settings and use the publicly available weights provided by the authors for DiffDock v1.1.3 on the project’s GitHub page. These weights were generated by Corso et al. (2024) training on BindingMOAD2020 (Hu et al., 2005) which includes most protein-ligand complexes released to the Protein Data Bank (Berman et al., 2000) before 2020.

5.2.2 Guidance

Guidance is a framework for conditioning generative diffusion models on additional information at inference time. Various authors have already described guidance for different types of diffusion models for various domains, such as images (Bansal et al., 2023; Lugmayr et al., 2022), audio processing (Ye et al., 2024), and molecular systems (Didi et al., 2023). For example, in image generation, guidance can be used to condition on a text prompt (Epstein et al., 2023).

Various guidance frameworks for score-based diffusion models have been proposed recently. Recall that score-based diffusion models (Y. Song et al., 2021) are trained to approximate the score $\nabla_x \log p_t(x)$ of a noisy version of the data. One robust way to create a conditional generative model is to separately train a classifier model on paired data of the form (x, y) and derive the score $\nabla_x \log p_t(y | x)$ from it; the model’s task effectively being to predict the desired property y from noisy versions of the data (Dhariwal & Nichol, 2021). By Bayes’ theorem, the sum of the two scores $\nabla_x \log p_t(x)$ and $\nabla_x \log p_t(y | x)$ is the conditional score for $\nabla_x \log p_t(x | y)$. Another approach, known as classifier-free guidance, trains on the unconditional and conditional task simultaneously via masking (Ho & Salimans, 2021); this effectively performs score-matching for $\nabla_x \log p_t(x | y)$ and $\nabla_x \log p_t(x)$. A disadvantage of these approaches is that they require training and that all particular classes of labels that the model should work with must be specified from the beginning. Training-free guidance, yet another type of guidance, does not change the model but modifies the iteratively generated intermediate predictions during inference. Essentially, this form of guidance adds an optimisation task at each intermediate (Bansal et al., 2023; Ye et al., 2024). When the conditional score $\nabla_x \log p_t(y | x)$ is tractable or good proxies exist, conditioning can be performed using only the unconditional model (Bansal et al., 2023; Chung et al., 2023; Ye et al., 2024). However, while I adopt this approach for being training-free, it has been mentioned that the training-based forms of guidance lead to better results.

The training-free guidance framework used here is akin to Diffusion Posterior Sampling (Chung et al., 2023), Universal Guidance (Bansal et al., 2023), and Training-Free Guidance (Ye et al., 2024) but adapted to DiffDock’s latent space and noise schedule. The three cited frameworks have been developed for variance preserving score-models (J. Song et al., 2021) in Euclidean spaces, but here I apply the same concept to the sampling model of DiffDock-L, a score model with a variance exploding schedule (Ho et al., 2020) operating on a non-Euclidean state space. Our goal is not to mathematically derive the general validity of the resulting approach, but to practically show that this approach is possible and to evaluate the merits of guidance for DL-based docking.

A sketch of our approach, using the generic notation of score-based models (Y. Song et al., 2021), follows. The main idea is that I perform an optimisation of the posterior mean obtained from the latent variables at each inference step. To clarify, the posterior mean is a point estimate in the target space $\hat{x}_0$ obtained from the latent x_t at each inference step using the formula $E[x_0 | x_t] = x_t + \sigma_t^2 \nabla_{x_t} \log p(x_0 | x_t)$, also known as Tweedie’s formula (Efron, 2011). I obtain the estimator by plugging-in the score model that was trained to approximate the score, which in mathematical notation is $S(x_t, t) \approx \nabla_{x_t} \log p(x_0 | x_t)$. This way I obtain the formula for our estimator $\hat{x}_0$ which is

$$\hat{x}_0 = x_t + \sigma_t^2 S(x_t, t). \quad (5.1)$$

This estimator is at the centre of the optimisations in the training-free guidance frameworks cited above.

The next step to setup the optimisations is to define a loss function ℓ which can be any differentiable scalar function on the target space of the model. All force fields for protein-ligand dynamics including constraints fit this description. The force fields used for our experiments are described in Section 5.2.3.

I investigate two optimisation schemes. The first is ‘backward’ guidance, in which $\hat{x}_0$ is optimised without differentiating through the model. In mathematical notation, the backward guidance optimisation task is

$$x_0^* = \hat{x}_0 + \phi^* \quad \text{where} \quad \phi^* = \arg \min_{\phi} \ell(\hat{x}_0 + \phi). \quad (5.2)$$

The second optimisation scheme is ‘forward’ guidance in which differentiates through the score model. This requires evaluating the score model and its gradient and is therefore more costly computationally. However, I investigate this approach because it might lead to better results. In mathematical notation, the ‘forward’ guidance optimisation task is

$$x_0^* = x_t + \sigma_t^2 S(x_t + \psi^*, t) \quad \text{where} \quad \psi^* = \arg \min_{\psi} \ell(x_t + \sigma_t^2 S(x_t + \psi, t)). \quad (5.3)$$

In practice, optimisation might not be performed at every inference step due to efficiency considerations and numerical stability. To illustrate the latter, during early time steps, the mean conditional expectation often does not resemble a valid ligand pose, and the ligand might clash with the receptor, leading to very large values of an energy function due to terms that divide by the distance between atoms, *e.g.* clash terms based on the Lennard-Jones potential.

An important technical detail for implementing this form of guidance for DiffDock-L is that I require an ‘adapter’ function that converts between updates to the internal coordinates and the external coordinates. This is because when I optimise $\hat{x}_0$, I want to optimise using the internal coordinates (to remain on the manifold), but the force field is written in terms of the external coordinates. DiffDock-L already includes a function $\mathcal{A}_c$ that maps elements of the tangent space of the internal coordinates and a reference ligand conformation to the external coordinates. Therefore, I only need to implement its inverse, a new function $\mathcal{P}_c$, which is a function that maps external coordinates and a reference ligand pose to changes in the internal coordinates. This function may use the same reference conformation c that DiffDock-L uses to define its state space, that is, the space spanned by the roto-translations and torsions applied to the reference pose c . Clearly, this function must be a projection onto the tangent space of the internal coordinates (the tangent space of the manifold) because the external coordinates are of higher dimensions than the internal ones. Based on the definition of the translation, rotation, and torsions, we can define $\mathcal{P}_c$ as the inverse of $\mathcal{A}_c$ such that $\mathcal{P}_c \cdot \mathcal{A}_c = \text{I}$. Let $\mathcal{P}_c$ be defined as applying the same operations of $\mathcal{A}_c$ but in reverse order: first undo the torsions¹ in exactly the reverse order, then use the Kabsch algorithm (Kabsch, 1976) to remove the translation and rotation components induced by the torsions, and then recover the rotation and translation with another application of the Kabsch algorithm.

For all optimisations, the LBFGS optimiser (Byrd et al., 1995) of PyTorch (Ansel et al., 2024) was used with a set maximum of 10 iterations and a tolerance of 10^{-5} as an early stopping criterion.

5.2.3 Loss functions

For the guidance loss function ℓ , I test the terms listed in Table 5.1 including a simple distance geometry force field to address ligand internal clash, a protein-ligand van der Waals term to address intermolecular clash, a centroid bias term for binding site steering, and an atom-wise constraint term for binding pose steering.

Here I note that the parametrisation with the Open Force Field 2.0.0 (Boothroyd et al., 2023) requires considerably more rigorous structure pre-processing. While the more robust DiffDock-L method tolerates almost all models of the PDB without pre-processing, including those with missing

¹The torsions must be applied in exactly the reverse order. For DiffDock-L this means rotating the side of the molecule with the fewer atoms.

Table 5.1: Guidance loss function terms

Name	Description
Ligand distance geometry	Squared violation of the distance geometry lower and upper bounds obtained from the RDKit (Landrum et al., 2025).
Protein-ligand clash	Mean squared overlap of the scaled van der Waals radii of protein atoms with ligand atoms.
Simple physics	Sum of the ligand distance geometry and protein-ligand clash terms.
Centroid constraints	Squared distance of the ligand centroid from an input location beyond a chosen threshold.
Atom-wise constraints	Mean squared distance (beyond a chosen threshold) over the provided atom locations.
Molecular dynamics force field	Open Force Field 2.0.0 (Boothroyd et al., 2023) which was developed for molecular dynamics involving small molecules.

residues, missing atoms, clashes, and even stereochemical misspecifications, the molecular dynamics force field requires a chemically complete description of the protein-ligand system including hydrogens and charges. This requirement is often not met by PDB crystal structures, so I prepare them using OpenMM PDBFixer (Eastman et al., 2024).

A technical point, and a bit of a warning to the potential user, is that I compute the molecular dynamics force field using the Python API of OpenMM while DiffDock-L is implemented in PyTorch. While both OpenMM and PyTorch can compute on GPUs, I do not know of any method to share the coordinates between the two frameworks directly on the GPU and I resort to copying the atom coordinates, energies, and forces between the two via CPU NumPy arrays. Moving these data between the CPU and GPU repeatedly for every optimisation potentially creates a computational bottleneck.

Further technical warnings are that the force field parametrisation relies on the AmberTools (Case et al., 2023) method `am1bcc` (Jakalian et al., 2002) to calculate partial charges which are conformation dependent, and that the parametrisation of the force field takes up to an hour for large protein-ligand complexes. I measure the time of the inference process without the pre-processing.

5.2.4 Experiments

I investigated the application of training-free guidance to DiffDock for three different goals: constrained docking, binding-site specific docking, and improving the physical validity of the base model. For constrained and pocket-specific docking, I dock a novel ligand into a related ligand’s cognate receptor, which is similar to how docking is used in lead optimisation for drug discovery. For constrained docking, I provide a set of atom coordinates obtained from the related ligand in its cognate receptor. These are within 2 Å of the pose of the aligned new ligands and should provide useful constraints. For pocket-specific docking, I provide the centroid of the aligned novel ligand as a constraint. I run

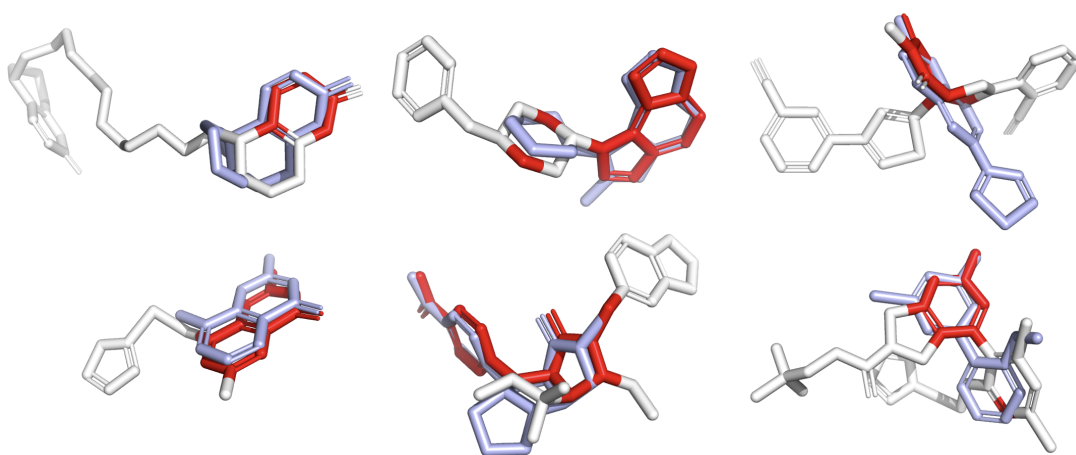


Figure 5.2: Illustration of the poses and constraints generated for six ligand pairs in the Malhotra data set. The ligand poses are shown without receptors after aligning via the receptors. The novel ligand (white) is to be docked using the information in the template ligand (blue) and its receptor.

the constrained and pocket specific docking on an updated Malhotra data set and the new GuideDock data set described in the next section. For the physical plausibility experiments, I perform re-docking on the PoseBusters data set (Buttenschoen et al., 2024). All experiments were performed on a single NVidia A100 GPU with 80GB of memory.

5.2.5 Data

To test the GuideDock, six data sets were used. The ligands of the Astex Diverse set, the PoseBusters Benchmark set, and the DockGen set were used for re-docking. The Malhotra set and the GuideDock data set were used for cross-docking. Further, the molecular glues in the CRBN data set are used for constrained docking.

Malhotra data set The Malhotra data set (Malhotra & Karanickolas, 2017) contains those 256 (86%) pairs of protein-ligand complexes, out of the authors’ original 297 pairs, which are binding mode preserving according to the authors’ manual annotation. Note that in 136 (53%) of the 256 pairs, the ligand pair is an elaboration, that is, the smaller ligand is an exact substructure of the larger. The list of PDB identifiers was obtained from the supplement of the paper, two obsolete PDB identifiers were updated, and typos in the ligand CCD identifiers were corrected. All changes made are listed in Table E.4 (Appendix E). Note that the training data of DiffDock, BindingMoad2020, includes most of the complexes in the Malhotra data set.

GuideDock data set The GuideDock data set contains 148 pairs of protein-ligand complexes from the PDB for which one protein-ligand complex serves as the binding pose template and the other

complex a novel ligand bound to the same receptor. The data set is a selection of the pairs that I processed in the previous chapter. Details for the selection of the GuideDock data set can be found in Section E.4.1 (Appendix E).

PoseBusters data set The PoseBusters data set was introduced in Chapter 2 and contains 308 protein-ligand complexes published between 2021 and 2024.

DockGen data set The DockGen data set (Corso et al., 2024) consists of protein-ligand complexes from the PDB. A subset of the complexes was selected, as described in Section E.4.3 (Appendix E), removing data points not suitable for docking with classical docking methods.

CRBN data set The CRBN data set consists of 27 protein-complexes from the PDB that contain chemical series of Cereblon E3 ligase modulators, also known as immunomodulatory imide drugs. The binding pose of the anchor as shown in Figure 5.7b.

5.2.6 Metrics

RMSD within 2 Å The root mean square deviation (RMSD) of the predicted pose relative to the reference pose was calculated using the symmetry-aware RMSD calculation in the RDKit (Landrum et al., 2025) that also considers symmetries in conjugated terminal functional groups. The threshold of 2 Å is arbitrary, but is commonly used for regular sized ligands (Cole et al., 2005).

Physical plausibility The percentage of poses that are physically valid is determined by running the PoseBusters tests (PB-validity) introduced in Chapter 2 that check chemical consistency and physical plausibility; this includes checking bond lengths, bond angles, and internal ligand distances against threshold values, consistency of stereochemistry of double bonds and tetrahedral centres, and protein-ligand distance thresholds.

The mean of each metric for the top-1 predicted poses over each data set is reported. Bias-corrected and accelerated bootstrap confidence intervals are calculated using 10 000 resamples at a 95 % confidence level with the `scipy.stats.bootstrap` function .

5.2.7 Baselines

Alignment using receptor protein chains The first baseline is an alignment of the ligands using the protein chains of the receptor. The template complex was aligned to the novel complex using a protein sequence alignment followed by a structural superposition in PyMOL (DeLano, 2002) using the `align` command using protein chains within 6 Å of each ligand. When multiple chain combinations

were available, the combination leading to the alignment with the largest colour and shape overlap (Leung et al., 2019; Malhotra & Karanicolas, 2017) of the ligands was selected.

DiffDock I used DiffDock v1.1.3 and the method’s default settings for generating and ranking 10 ligand poses for each complex. The top ranked pose was picked for evaluation.

DiffDock with energy minimisation The top ranked prediction of DiffDock is minimized in a rigid receptor with OpenMM (Eastman et al., 2024) using the Open Force Field 2.0.0 (Boothroyd et al., 2023) until convergence within 0.01 kJ mol⁻¹. The receptors are prepared with PDBFixer (Eastman et al., 2024), finding and replacing missing or non-standard residues and missing atoms.

GOLD without constraints The CSD Python API version 3.3.1 and public documentation was used to dock with GOLD (Verdonk et al., 2003). The ligands were prepared using the provided LigandPreparation script and default settings; this includes adding missing hydrogens, removing unknown atoms, and a rule-based protonation of the ligand. The receptor was prepared by loading the protein and co-factors separately from the PDB files and adding all hydrogens. A settings file was created for each complex using the Docker class default settings with rescore function plp, autoscale 100%, and early termination off. The binding site was defined using the crystal ligand centroid with the BindingSiteFromPoint function and a radius of 25 Å. 40 ligand poses were generated and the top-ranked pose was selected for evaluation.

GOLD with fragment region constraints Constrained docking with GOLD was performed as described in the previous paragraph but constraints for the substructures were added atom-wise using the RegionConstraint function that adds a spherical constraint around each template atom location with a user-defined radius of 2 Å. Other radii (0.5 Å, 1 Å and 3 Å) were also tested but did not yield better results (Table E.1, Appendix E).

5.2.8 Tuning

I determined the guidance parameters including the guidance schedule and loss term weights for re-docking using the DockGen validation set and for constrained docking using the Malhotra set, keeping the PoseBusters set and GuideDock sets as test sets. I tested only few different parameters and did not perform a systematic grid-search for the best parameters due to computational constraints. The default settings of DiffDock were not changed.

5.3 Results

5.3.1 Physical plausibility in re-docking

The re-docking experiments show that physical plausibility can be improved to the same extent by a post-prediction energy minimisation, by post-prediction with the simple force field, guidance with the simple force fields, and guidance with molecular dynamics. Figure 5.3 shows a comparison of the failure modes of DiffDock-L, DiffDock-L with post-prediction energy minimisation, and guidance with the simple clash terms. The figure shows that energy minimisation and physical guidance have virtually the same effect: The physical validity is mostly improved due to reducing protein-ligand clashes. A full list of our experiments re-docking the cognate ligand into the rigid receptor in the PoseBusters data set is provided in Table 5.2. Guidance with OpenFF 2.0.0 and the simpler clash terms yield the same improvement in accuracy and physical validity ($45.8 \pm 2.9\%$ RMSD within 2 Å and PB-valid) as the post-prediction minimisation ($45.8 \pm 2.8\%$) applied to DiffDock-L ($26.6 \pm 2.5\%$ RMSD within 2 Å and PB-valid). Performing the post-processing minimisation before or after the scoring makes no difference. The forward guidance ($45.8 \pm 2.9\%$ RMSD within 2 Å and PB-valid) and reverse guidance ($45.1 \pm 2.8\%$ RMSD within 2 Å and PB-valid) schemes perform equally well. However, forward guidance, which differentiates through the model, takes more time (2 min per protein-ligand complex) than the reverse guidance scheme (1.3 min), both significantly longer than DiffDock-L (12 s).

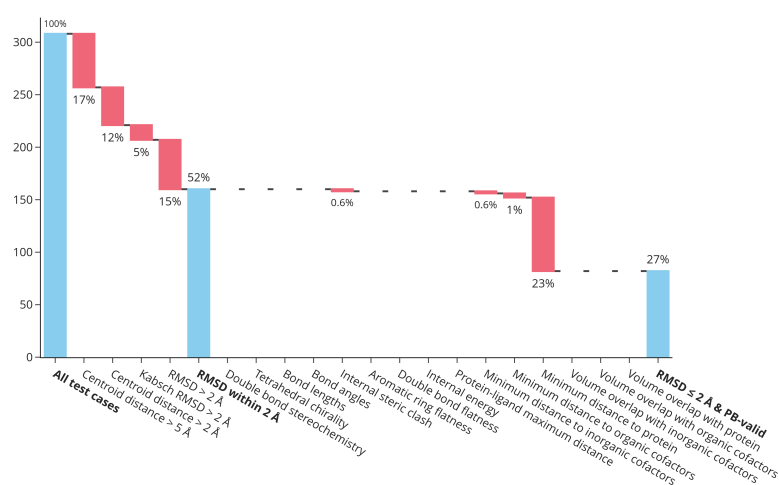
5.3.2 Binding site-specific re-docking

Table 5.2 also shows that, as expected, providing the centroid location of the docked ligand through guidance improves the accuracy ($63.6 \pm 2.8\%$ RMSD within 2 Å) over DiffDock-L ($51.9 \pm 2.8\%$ RMSD within 2 Å). Providing the centroid and running the clash terms results in the overall highest accuracy and physical validity ($59.7 \pm 2.8\%$ RMSD within 2 Å and PB-valid). However, even though the poses are now all within 5 Å RMSD, the 2 Å RMSD threshold is satisfied by only 63.6 % of the generated poses. Those poses failing this threshold have the centroid in the right location, but the binding pose is still not correct.

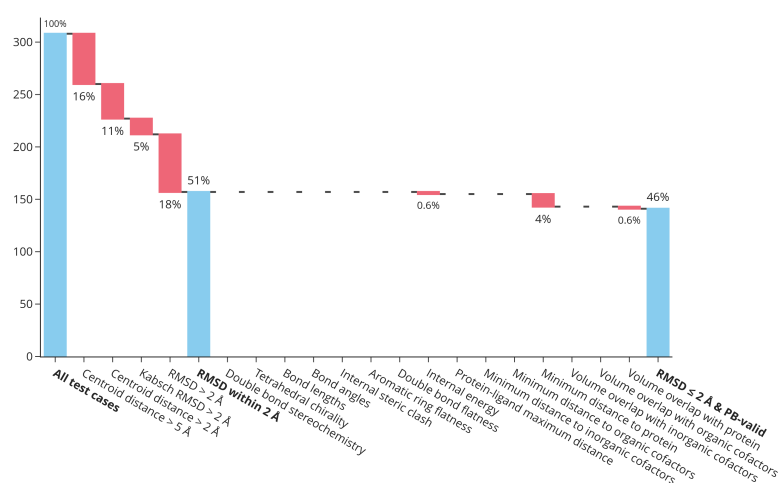
5.3.3 Cross-docking into holo structures

In our cross-docking experiments, ligands are docked into the crystal structure of the paired ligand, and not the cognate receptor structure. The results for cross-docking on the Malhotra set are visualized in Figure 5.5 and summarized in Table E.3. On the Malhotra data set (n=297), the raw performance of DiffDock-L is 27.7 % (RMSD within 2 Å and PB-valid poses). Using only clash terms ('simple physics'),

(a) DiffDock-L.



(b) DiffDock-L with post-prediction energy minimization.



(c) GuideDock with physical guidance.

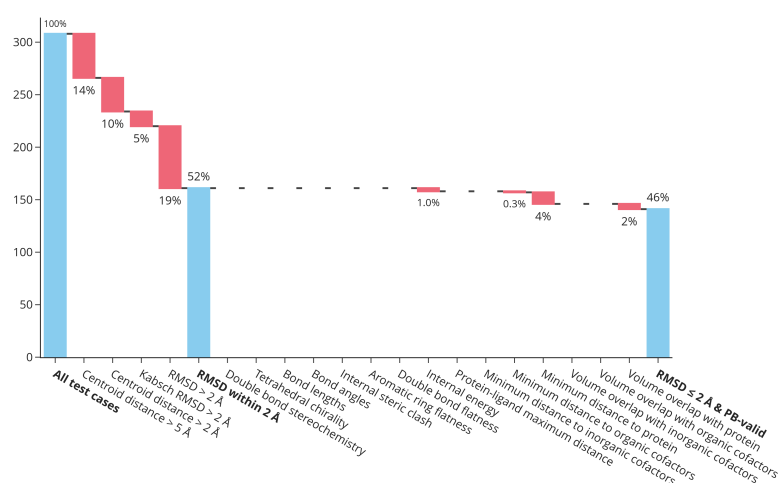


Figure 5.3: Failure modes of DiffDock-L and GuideDock on re-docking the PoseBusters data set. (a) For DiffDock-L, protein-ligand clash is the largest failure mode. (b) A post-prediction energy minimisation fixes all clashes. (c) GuideDock with physical guidance terms also fixes all clashes. Both the energy minimisation and the physical guidance improve the physical plausibility of the methods. The resulting performance is identical.

Table 5.2: Re-docking the ligand into the cognate rigid receptor for the PoseBusters data set (n=308). The metrics are the percentage of top ranked poses (out of 308) that are within 2 Å RMSD of the reference pose, that are PB-valid (pass all PoseBusters tests), and that are simultaneously within 2 Å RMSD and PB-valid. The numbers are percentages and the confidence intervals are 95 % bootstrap estimates. DiffDock post-processing energy minimization generated 45.8% of PB-valid poses with $\text{RMSD} \leq 2 \text{ \AA}$ which is about the same as for the simple physics guidance.

	$\text{RMSD} \leq 2 \text{ \AA}$	PB-valid	$\text{RMSD} \leq 2 \text{ \AA}$ and PB-valid
DiffDock (no guidance)	51.9 ± 2.8	32.8 ± 2.7	26.6 ± 2.5
DiffDock followed by energy minimisation	51.0 ± 2.8	76.0 ± 2.4	45.8 ± 2.8
GuideDock: simple physics	52.3 ± 2.9	78.9 ± 2.3	45.8 ± 2.9
GuideDock: centroid and simple physics	63.6 ± 2.8	91.9 ± 1.6	59.7 ± 2.8
GOLD	60.7 ± 2.8	91.1 ± 1.6	56.7 ± 2.8
Crystal structure	100.0 ± 0.0	99.7 ± 0.3	99.7 ± 0.3

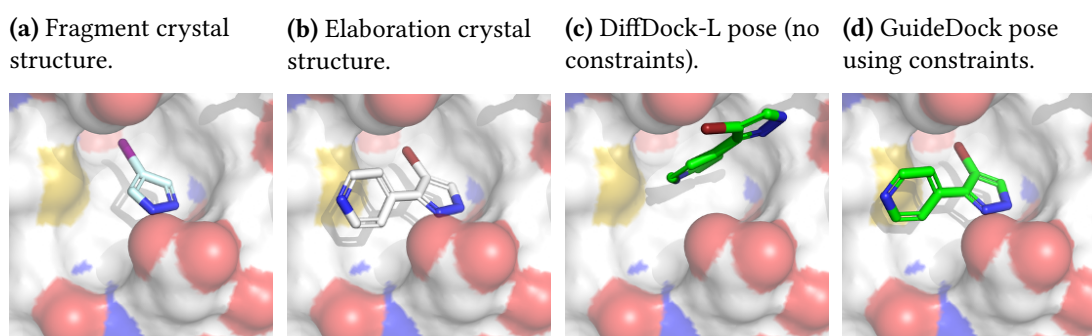


Figure 5.4: Positive example for the use of constraints. For this pair of ligands in the Malhotra data set, GuideDock predicts the correct binding mode due to the additional constraints while DiffDock-L predicts a different binding mode.

the performance of GuideDock is 58.2 % ($\text{RMSD} \leq 2 \text{ \AA}$ and PB-valid poses). This increase is due to the physical validity metric increasing from 32.8 % to 81.2 % despite accuracy dropping from 79.7 % to 66.8 %. For comparison, GOLD without constraints generates 16.0 % PB-valid poses within 2 Å RMSD on this data set. The main failure modes that remain are RMSD being too large (79.7 %) followed by too small protein-ligand distances (3.9 %). With region constraints using the oracle fragment information, the performance rises to 33.6 % PB-valid poses within 2 Å RMSD. The main failure modes remain RMSD being too large (58.3 %) followed by too small protein-ligand distances (7.8 %). Finally, the alignment via protein chains does lead to clashes with the receptor and only 41.0 % of generated protein-ligand poses are PB-valid. This is mostly due to protein-ligand clashes 51.9 % but also 5.5 % of the large ligand crystal conformations have too much steric strain and do not pass the internal energy check.

The results for cross-docking on the GuideDock (n=148) are shown in Figure 5.6 and summarized in Table 5.3. Docking the modified ligands into the rigid receptors of the related ligands DiffDock yielded

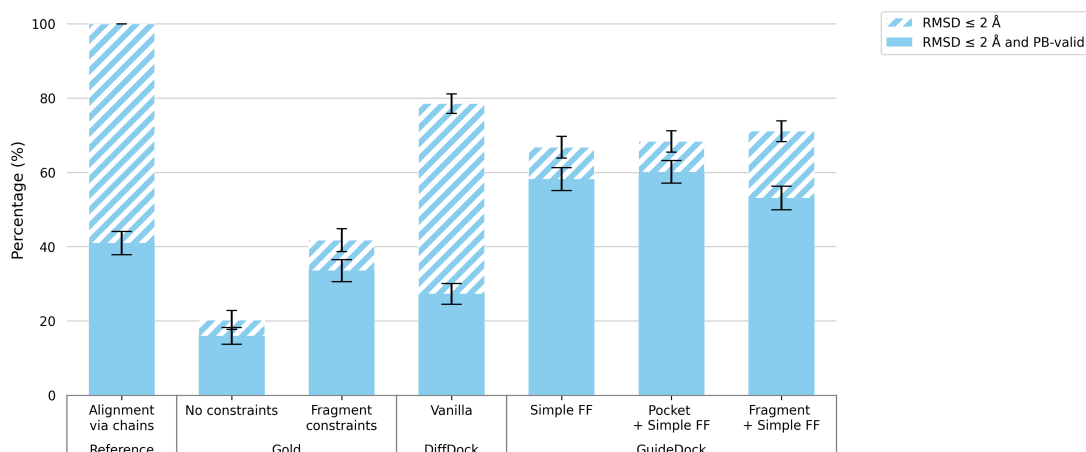


Figure 5.5: Cross-docking ligands into the paired holo structures of the Malhotra data set. DiffDock-L recovers almost 80 % of the aligned poses but fewer than 30 % are PB-valid. GuideDock with simple physics-based terms yields almost 60 % PB-valid poses within 2 Å of the aligned pose. On this data set, DiffDock-L and GuideDock outperform the classical docking method GOLD; however, most of these complexes were in the method’s training set.

22.3 ± 3.4 % PB-valid poses within 2 Å RMSD. Adding simple physics guidance increases this number to 33.8 ± 3.9 %. Adding binding site information in terms of a centroid in the guidance further increases the performance to 51.0 ± 4.1 %. In comparison, GOLD without constraints and a spherical binding site with a radius of 25 Å yields 55.6 ± 4.1 % PB-valid poses within 2 Å RMSD. Adding atom-wise region constraints for the substructure information increases the performance of GOLD to 71.4 ± 3.7 %. The baseline alignment via the protein chains yields less PB-valid poses (66.9 ± 3.9 %) than GuideDock (71.6 ± 3.7 %) or GOLD 92.4 ± 2.2 %, due to the generation of more clashes with the receptor.

Table 5.3: Docking ligands of the GuideDock data set (n=148) into the receptors of the paired ligands. The metrics are the percentage of top ranked poses (out of 148) that are within 2 Å RMSD of the reference pose, which are PB-valid (pass all PoseBusters tests) and that are simultaneously within 2 Å RMSD and PB-valid. The RMSD is calculated relative to the position of the ligand aligned via the protein chains. The numbers are percentages and the confidence intervals are 95 % bootstrap estimates.

	RMSD ≤ 2 Å	PB-valid	RMSD ≤ 2 Å and PB-valid
DiffDock (no guidance)	41.2 ± 4.1	37.8 ± 4.0	22.3 ± 3.4
GuideDock: simple physics	37.2 ± 3.9	71.6 ± 3.7	33.8 ± 3.9
GuideDock: pocket & simple physics	53.1 ± 4.1	92.5 ± 2.2	51.0 ± 4.1
GuideDock: fragment & simple physics	74.3 ± 3.7	72.3 ± 3.7	61.5 ± 4.0
GOLD: no constraints	58.3 ± 4.1	92.4 ± 2.2	55.6 ± 4.1
GOLD: fragment region constraints	76.9 ± 3.5	89.1 ± 2.6	71.4 ± 3.7
Alignment via protein chains	100.0 ± 0.0	66.9 ± 3.9	66.9 ± 3.8

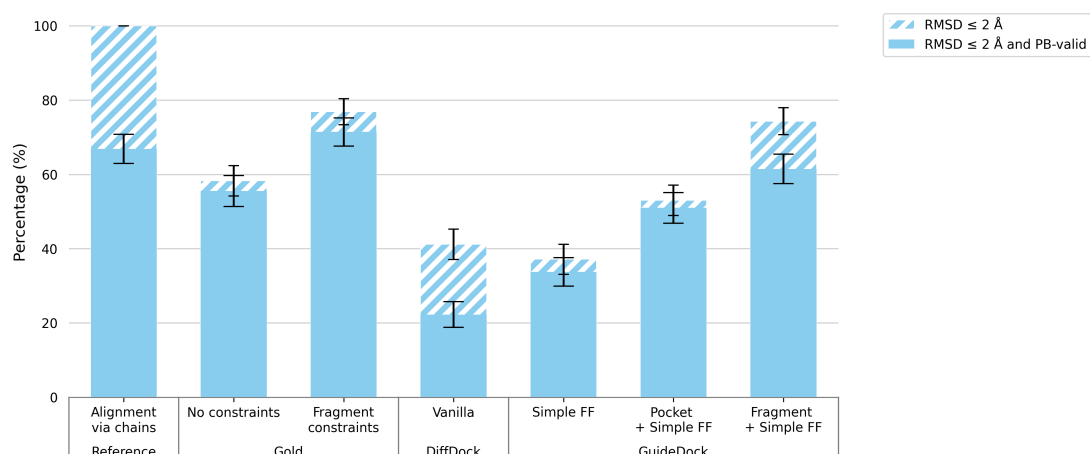


Figure 5.6: Cross-docking ligands into the paired receptor structures of the GuideDock data set. In comparison to the performance on the Malhotra data set Figure 5.5, DiffDock-L recovers only half, that is 40 % of the aligned poses and only around 20 % are PB-valid. GuideDock with simple physics-based terms yields almost 40 % PB-valid poses within 2 Å of the aligned pose, with pocket guidance, this rises to around 50 %, and with fragment guidance to about 60 % PB-valid and RMSD within 2 Å. GOLD performs relatively better on this data set, and obtains almost 60 % of valid poses without constraints. Overall, DiffDock-L and GuideDock do not outperform GOLD with and without constraints.

5.3.4 Substructure constrained docking

For a chemical series of molecular glues (Cereblon E3 ligase modulators, also known as immunomodulatory imide drugs), the substructure of the ‘anchor’ is believed to be highly conserved in place. Adding position constraints for this highly conserved motif successfully guides and recapitulates the binding mode, as shown in Figure 5.7.

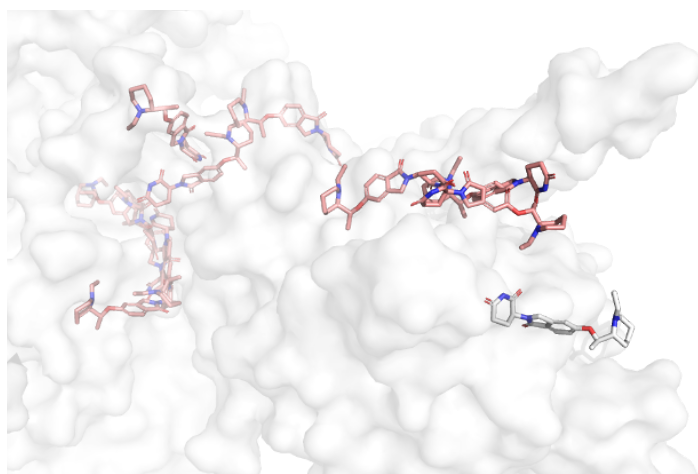
GuideDock was tested by my industrial collaborator, Giorgio Tamo, on a set of proprietary structures of tertiary protein-ligand complexes. The resultant performance is shown in Figure E.2 (Appendix E). Once the target and binding pose are deposited in the PDB, the protein-ligand image will be revealed. Overall, the guidance improved the prediction accuracy by 10 % from around 49 % (DiffDock-L RMSD within 2 Å) to 59 % (GuideDock RMSD within 2 Å).

5.4 Discussion

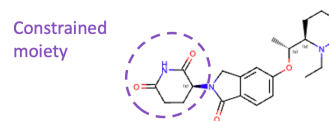
By implementing simple guidance in DiffDock-L, I demonstrated that the predictions of diffusion-based molecular structure models can be steered in a controlled manner. Using simple loss functions for guidance was sufficient to increase the physical plausibility of generated poses, as evaluated by PoseBusters, while avoiding the expensive parametrisation required for energy minimisation with a full molecular dynamics force field.

I further showed that guidance improves the performance of DiffDock-L not only in terms of phys-

(a) Top 10 poses predicted by DiffDock-L.



(b) Molecular structure with constrained moiety highlighted.



(c) GuideDock top 1 prediction.

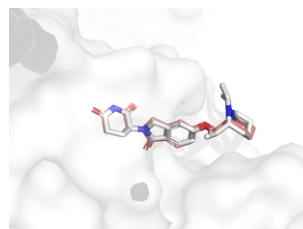


Figure 5.7: Positive example of substructure constraints. Predicted poses are shown in red, and crystal structures are shown in white. (a) None of the top 10 predictions of the unguided model is in the right binding site. (b) Constraints can be defined for the atoms of the ‘anchor’ of the molecular glue because their locations are highly conserved relative to the receptor. (c) A prediction incorporating the constraint is close to the crystal structure showing that the guided model predicted the remaining degrees of freedom correctly. This example illustrates how constraints can be useful for docking.

ical plausibility but also in accuracy, as measured by RMSD within 2 Å. Incorporating prior information from related ligands, sometimes referred to as templates, unsurprisingly improved performance for cross-docking, but only when the underlying assumptions of this approach were satisfied. Crucially, this requires that the imposed constraints do not interfere with docking and that the rigid receptor does not sterically block the correct ligand pose.

Using guidance to enable pocket-specific and constrained docking was shown to work well in practice. Revealing additional information about the ligand and its binding pose to the model during inference improved performance relative to the base model. I showed that the same base model can be applied across these different tasks. However, when compared to the classical docking method GOLD, our models still performed worse in terms of physical plausibility, though not in terms of RMSD accuracy. Training the base model to generate only PoseBusters-valid poses may therefore be a promising direction.

Overall, the performance of the resulting method, GuideDock, is at best comparable to that of the classical docking method GOLD, both with and without additional constraints. A more fundamental paradigm shift, such as moving from docking to co-folding, may be required to surpass traditional physical docking methods in a meaningful way.

Our approach uniquely demonstrates how physical terms derived from force fields and user-defined biases can be integrated with state-of-the-art DL methods for rigid small-molecule docking. The terms used here improve results by incorporating information that the model either did not

learn to exploit (physics) or did not have access to (user biases). Introducing user biases also enables practitioners to directly test their hypotheses within the docking process.

5.5 Conclusion

In this chapter I presented GuideDock, a DL-based docking method that innovates by adding diffusion guidance to the DiffDock-L blind docking method. This allows reusing the pre-trained model to perform constrained docking and pocket-specific docking, while simultaneously also improving the physical plausibility of the generated poses in comparison to the unguided DiffDock. Although GuideDock does not surpass classical docking methods, in particular the classical docking method GOLD, the experiments demonstrate that repurposing a single blind docking model for multiple different tasks is possible.

Since I started working on GuideDock, other researchers have implemented guidance terms in biological structure prediction models. The ‘steering’ approach that was added to the second version of the preprint of Boltz (Wohlwend et al., 2024) this summer is very similar to our method. I am excited that I was on the right track with this work, but I suspect that the machine learning community now views co-folding as having superseded docking.

Chapter 6

Conclusion

The aim of this thesis was to investigate deep learning (DL)-based protein–ligand docking. Motivated by the landmark success of AlphaFold2 in protein structure prediction and by the observation that DL-based docking has not yet achieved a comparable impact, I sought to contribute to the advancement of this field. In particular, this work aimed to assess the current capabilities of existing methods, identify opportunities for improvement, and support the maturation of DL-based docking as a practical technology.

6.1 Key findings and contributions

My research improves our understanding of DL-based protein-ligand docking by investigating the shortcomings of common evaluation metrics, training data, and testing setups. In addition, my research explored the interplay of physics-based and DL-based approaches and showed that diffusion guidance can inject physical biases and user knowledge.

My first key finding is that the established RMSD metric of being within 2 Å of the crystallographic binding mode is insufficient to assess DL-based protein-ligand docking. Using examples, I showed that the 2 Å RMSD threshold admits physically impossible pretzel-shaped conformations and stark protein-ligand clashes, among others. As a remedy, I implemented a new metric, PB-validity, which consists of a series of physical and chemical plausibility tests. This new metric has been adopted by the research community to benchmark docking, *de novo* molecular generation, and co-folding models.

My most significant contribution is that I have systematically shown that DL-based docking methods have not yet outperformed classical docking methods. I did this in two independent ways: first, by creating a new, more difficult test set, the PoseBusters test set; and second, by introducing the new metric: PB-validity. Both provide unique insights into the performance of the models: The PoseBusters tests enumerate chemical and physical failure modes, telling the user what exactly has

gone wrong and how badly. Independently, using a novel out-of-distribution training data test set and splitting it by sequence identity revealed that DL-based methods generalise less well than classical docking methods. The impact of this contribution has seen a renewed effort of model developers to improve the physical plausibility of docking as well as co-folding models.

Also important is my finding that the very training data, the crystal structures in the PDBbind 2020 General set (and its successor PDBbind 2020 General R1), are polluted with the same failure modes with which DL-based methods struggle. Filtering, pre-processing, or remediating the training data is rarely done but should be; this includes double-checking that the molecular representations loaded from the data sets are correct. A simple way to do this is to use the PoseBusters tests also to inspect the training data, but that alone does not guarantee problem-free structures.

Regarding time-based training-test data splits, I found that time-splits are insufficient to demonstrate generalisation because they omit use case-dependent considerations of target and pocket similarity, ligand similarity and known binding modes of congeneric series, and interaction conservation. Splitting my test set by maximum sequence identity to the training data, I showed that DL-based models perform much better close to the training data and worse on more dissimilar proteins—using the entire PoseBusters tests set without subdivision captures only average performance, which is the weakness of time-based splits. The lesson for practitioners is to consider receptor sequence identity, interaction similarity, ligand chemical similarity, and pocket similarity of protein-complexes of interest and report such metrics on data splits when benchmarking protein-ligand structure prediction methods.

My thesis also makes a methodological contribution. The usefulness of guidance for adding physical and user biases to diffusion-based structure models is shown. The GuideDock approach provides a convenient interface for influencing model predictions via force fields, already familiar to practitioners. I show that any force field term can be used including soft constraints and clash terms. GuideDock demonstrates that guidance can improve the physical validity of the predictions and allow the user to define a binding site and substructure constraints. The same principles are applicable to co-folding methods as well.

I have also shown that the DL-based models analysed here have not yet learnt (neither explicitly or implicitly) some of the rules encoded in the force fields. This suggests that physics and DL-based methods might still be combined successfully. I demonstrated two ways of doing this: (i) adding an energy minimisation after pose prediction; and (ii) using diffusion guidance with physics-inspired terms. Both improve the physical plausibility of the generated poses of the methods. Further ways of combining force fields and diffusion models are easy to think of and should be explored.

Another practical contribution to the field are two new docking test sets. The first is the PoseBusters test set that includes data both similar to and different from the training data (PDBbind

2020 General), to highlight the current shortcomings of DL-based models and the way they are benchmarked. The second is the GuideDock test set, which contains pairs of protein-ligand complexes of small molecules that adhere to the same binding pockets on the same protein chain. The PoseBusters Benchmark set has already been adopted by the scientific community as a test set; the GuideDock set will no doubt be useful for testing template-based and constrained protein-ligand prediction methods.

Finally, I quantified the relationship between the chemical similarity of related ligands and the similarity of their binding poses; this relationship was found to be even stronger for chemical elaborations of smaller ligands. This finding is of particular relevance to medicinal chemists, as it provides an empirical basis for the substructure constraints used in constrained docking and, more broadly, in fragment-based drug design.

6.2 Limitations and recommendations for future research

Although my thesis focuses on DL-based protein-ligand docking, it opens opportunities for others to study the newly emerging co-folding methods in a similar way. In 2024, AlphaFold 3 (Abramson et al., 2024) showed, and at first without publishing source code, that DL enables not only protein structure prediction, but also protein-ligand structure prediction. To my delight, the creators of AlphaFold 3 used PoseBusters to assess and develop their method. However, while this thesis does not cover co-folding methods, other academic researchers have demonstrated that co-folding methods can be developed and researched in an academic setting: Wohlwend et al. (2024) released an AlphaFold 3 clone, called Boltz-1, almost within the same year. This is faster than the time it took OpenFold (Ahdriz et al., 2024) to make a methodology inspired by AlphaFold 2 (Jumper et al., 2021) publicly available. And it did not stop there. In the last two years, at least eight co-folding methods have been described in publications (Table 6.1). Early benchmarks of co-folding models show promising results (Nittinger et al., 2025; Škrinjar et al., 2025), and it seems that co-folding methods are well positioned to take over some docking tasks. However, currently, we do not know to what extent their conceptual advantage of being able to model induced fit will translate into virtual screening success in practice.

A limitation of my work is also its theoretical depth. This was mitigated by extensive computational studies. However, developing a theoretical basis for (i) search versus generative modelling, (ii) different guidance schemes, and (iii) a diffusion perspective on physical force fields remains an interesting problem in the docking and co-folding context. Is generative modelling fundamentally better suited than search for docking because it scales better to the billions of possible molecules that could be suitable drug candidates? Is differentiating through the model necessary when taking more steps in the reverse diffusion process during inference? How can force fields provide a training signal and replace some data during training? In my opinion, answering these questions can lead to

Table 6.1: List of early co-folding methods. AlphaFold 3 was the first deep learning-based model to model complete protein-ligand complexes. The co-folding paradigm might replace docking for some uses cases.

Co-folding program	Creators	Year
AlphaFold 3	Abramson et al.	2024
Boltz-1	Wohllwend et al.	2024
Chai-1	Boitreaud et al.	2024
HelixFold3	L. Liu et al.	2024
NeuralPlexer	Qiao et al.	2024
RoseTTAFold All-Atom	Krishna et al.	2024
Boltz-2	Passaro et al.	2025
OpenFold 3	The OpenFold3 Team	2025
Protenix	X. Chen et al.	2025

improvements in protein-ligand structure prediction.

Once DL-based docking methods are fast enough, it will be important to benchmark them on virtual screening, another very important application of docking, which is assessed in terms of the enrichment factor that methods provide given large data sets of potential candidates. Although speed is currently the bottleneck for DL-based methods to be applied to large virtual screening studies, this will be overcome, and then their performance must be established. Ultimately, this will also include prospective experimental validation. Although my research has been limited to computation studies, there is much to learn from *in silico* drug discovery methods *in vitro* and prospectively. Prospective experimental validation is ultimately needed as evidence of the performance of computational models in drug discovery. This thesis began by noting GNINA, one of the first DL-based docking models, and on this note, it seems fitting to highlight that GNINA (McNutt et al., 2021) has recently become the top ranked method in CACHE (Dunn et al., 2024): an independent prospective study that included experimental validation. Other workflows using DL-based docking will follow soon.

6.3 Closure

My research has explored and helped improve both DL-based docking and co-folding methods, especially in terms of chemical and physical plausibility. It has also shown that DL-based protein-ligand docking has not yet reached its potential; however, personally, given the current rate of progress, I believe that DL-based approaches are very likely to become competitive within three years. I am delighted to have contributed directly to the field by adding new metrics, tools, and benchmark test data sets, providing insights into training data, model performance, different ways to guide diffusion-

based models, and ultimately a new guidance model.

References

- Abramson, J., Adler, J., Dunger, J., Evans, R., Green, T., Pritzel, A., Ronneberger, O., Willmore, L., Ballard, A. J., Bambrick, J., Bodenstein, S. W., Evans, D. A., Hung, C.-C., O'Neill, M., Reiman, D., Tunyasuvunakool, K., Wu, Z., Žemgulytė, A., Arvaniti, E., ... Jumper, J. M. (2024). Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature*, 630(8016), 493–500. <https://doi.org/10.1038/s41586-024-07487-w>
- Adams, M. J., Blundell, T. L., Dodson, E. J., Dodson, G. G., Vijayan, M., Baker, E. N., Harding, M. M., Hodgkin, D. C., Rimmer, B., & Sheat, S. (1969). Structure of rhombohedral 2 zinc insulin crystals. *Nature*, 224(5218), 491–495. <https://doi.org/10.1038/224491a0>
- Adams, P. D., Afonine, P. V., Bunkóczi, G., Chen, V. B., Davis, I. W., Echols, N., Headd, J. J., Hung, L.-W., Kapral, G. J., Grosse-Kunstleve, R. W., McCoy, A. J., Moriarty, N. W., Oeffner, R., Read, R. J., Richardson, D. C., Richardson, J. S., Terwilliger, T. C., & Zwart, P. H. (2010). PHENIX: A comprehensive Python-based system for macromolecular structure solution. *Acta Crystallographica Section D Biological Crystallography*, 66(2), 213–221. <https://doi.org/10.1107/S0907444909052925>
- Ahdritz, G., Bouatta, N., Floristean, C., Kadyan, S., Xia, Q., Gerecke, W., O'Donnell, T. J., Berenberg, D., Fisk, I., Zanichelli, N., Zhang, B., Nowaczynski, A., Wang, B., Stepniewska-Dziubinska, M. M., Zhang, S., Ojewole, A., Guney, M. E., Biderman, S., Watkins, A. M., ... AlQuraishi, M. (2024). OpenFold: Retraining AlphaFold2 yields new insights into its learning mechanisms and capacity for generalization. *Nature Methods*, 21, 1514–1524. <https://doi.org/10.1038/s41592-024-02272-z>
- Ahmed, A., Smith, R. D., Clark, J. J., Dunbar, James B., Jr & Carlson, H. A. (2014). Recent improvements to binding MOAD: A resource for protein–ligand binding affinities and structures. *Nucleic Acids Research*, 43(D1), D465–D469. <https://doi.org/10.1093/nar/gku1088>
- AlKharboush, D. F., Kozielski, F., Wells, G., & Porta, E. O. (2025). Fragment-based drug discovery: A graphical review. *Current Research in Pharmacology and Drug Discovery*, 9, 100233. <https://doi.org/10.1016/j.crphar.2025.100233>
- Allen, F. H., Davies, J. E., Galloy, J. J., Johnson, O., Kennard, O., Macrae, C. F., Mitchell, E. M., Mitchell, G. F., Smith, J. M., & Watson, D. G. (1991). The development of versions 3 and 4 of the Cambridge Structural Database System. *Journal of Chemical Information and Computer Sciences*, 31(2), 187–204. <https://doi.org/10.1021/ci00002a004>
- Allen, W. J., Balias, T. E., Mukherjee, S., Brozell, S. R., Moustakas, D. T., Lang, P. T., Case, D. A., Kuntz, I. D., & Rizzo, R. C. (2015). DOCK 6: Impact of new features and current docking performance. *Journal of Computational Chemistry*, 36(15), 1132–1156. <https://doi.org/10.1002/jcc.23905>
- Ansel, J., Yang, E., He, H., Gimelshein, N., Jain, A., Voznesensky, M., Bao, B., Bell, P., Berard, D., Burovski, E., Chauhan, G., Chourdia, A., Constable, W., Desmaison, A., DeVito, Z., Ellison, E., Feng, W., Gong, J., Gschwind, M., ... Chintala, S. (2024). PyTorch 2: Faster machine learning through dynamic Python bytecode transformation and graph compilation. *29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2 (ASPLOS '24)*. <https://doi.org/10.1145/3620665.3640366>
- Asthana, P. (1995). Jumping the technology S-curve. *IEEE Spectrum*, 32(6), 49–54. <https://doi.org/10.1109/6.387142>
- Bahdanau, D., Cho, K., & Bengio, Y. (2015). Neural Machine Translation by Jointly Learning to Align and Translate. *3rd International Conference on Learning Representations*. <https://doi.org/10.48550/arXiv.1409.0473>

- Bansal, A., Chu, H.-M., Schwarzschild, A., Sengupta, S., Goldblum, M., Geiping, J., & Goldstein, T. (2023). Universal guidance for diffusion models. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 843–852. <https://doi.org/10.1109/CVPRW59228.2023.00091>
- Bateman, A., Martin, M.-J., Orchard, S., Magrane, M., Adesina, A., Ahmad, S., Bowler-Barnett, E. H., Bye-A-Jee, H., Carpentier, D., Denny, P., Fan, J., Garmiri, P., Gonzales, L. J. D. C., Hussein, A., Ignatchenko, A., Insana, G., Ishtiaq, R., Joshi, V., Jyothi, D., ... Zhang, J. (2024). UniProt: The universal protein knowledgebase in 2025. *Nucleic Acids Research*, 53(D1), D609–D617. <https://doi.org/10.1093/nar/gkae1010>
- Baxter, C. A., Murray, C. W., Clark, D. E., Westhead, D. R., & Eldridge, M. D. (1998). Flexible docking using tabu search and an empirical estimate of binding affinity. *Proteins: Structure, Function, and Genetics*, 33(3), 367–382. [https://doi.org/10.1002/\(SICI\)1097-0134\(19981115\)33:3<367::AID-PROT6>3.0.CO;2-W](https://doi.org/10.1002/(SICI)1097-0134(19981115)33:3<367::AID-PROT6>3.0.CO;2-W)
- Bender, B. J., Gahbauer, S., Lutten, A., Lyu, J., Webb, C. M., Stein, R. M., Fink, E. A., Balias, T. E., Carlsson, J., Irwin, J. J., & Shoichet, B. K. (2021). A practical guide to large-scale docking. *Nature Protocols*, 16(10), 4799–4832. <https://doi.org/10.1038/s41596-021-00597-z>
- Benson, M. L., Smith, R. D., Khazanov, N. A., Dimcheff, B., Beaver, J., Dresslar, P., Nerothin, J., & Carlson, H. A. (2007). Binding MOAD, a high-quality protein–ligand database. *Nucleic Acids Research*, 36, D674–D678. <https://doi.org/10.1093/nar/gkm911>
- Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., & Bourne, P. E. (2000). The protein data bank. *Nucleic Acids Research*, 28(1), 235–242. <https://doi.org/10.1093/nar/28.1.235>
- Bickerton, G. R., Paolini, G. V., Besnard, J., Muresan, S., & Hopkins, A. L. (2012). Quantifying the chemical beauty of drugs. *Nature Chemistry*, 4(2), 90–98. <https://doi.org/10.1038/nchem.1243>
- Bishop, C. M., & Bishop, H. (2024). *Deep learning: Foundations and concepts*. Springer International Publishing. <https://doi.org/10.1007/978-3-031-45468-4>
- Bissantz, C., Kuhn, B., & Stahl, M. (2010). A medicinal chemist’s guide to molecular interactions. *Journal of Medicinal Chemistry*, 53(14), 5061–5084. <https://doi.org/10.1021/jm100112j>
- Boitreaud, J., Dent, J., McPartlon, M., Meier, J., Reis, V., Rogozhnikov, A., & Wu, K. (2024). *Chai-1: Decoding the molecular interactions of life* (preprint). <https://doi.org/10.1101/2024.10.10.615955>
- Bolcato, G., Cuzzolin, A., Bissaro, M., Moro, S., & Sturlese, M. (2019). Can we still trust docking results? An extension of the applicability of DockBench on PDBBind database. *International Journal of Molecular Sciences*, 20(14), 3558. <https://doi.org/10.3390/ijms20143558>
- Boothroyd, S., Behara, P. K., Madin, O. C., Hahn, D. F., Jang, H., Gapsys, V., Wagner, J. R., Horton, J. T., Dotson, D. L., Thompson, M. W., Maat, J., Gokey, T., Wang, L.-P., Cole, D. J., Gilson, M. K., Chodera, J. D., Bayly, C. I., Shirts, M. R., & Mobley, D. L. (2023). Development and benchmarking of open force field 2.0.0: The sage small molecule force field. *Journal of Chemical Theory and Computation*, 19(11), 3251–3275. <https://doi.org/10.1021/acs.jctc.3c00039>
- Bouysset, C., & Fiorucci, S. (2021). ProLIF: A library to encode molecular interactions as fingerprints. *Journal of Cheminformatics*, 13(1), 72. <https://doi.org/10.1186/s13321-021-00548-6>
- Bradley, A. R., Rose, A. S., Pavelka, A., Valasatava, Y., Duarte, J. M., Prlić, A., & Rose, P. W. (2017). MMTF—an efficient file format for the transmission, visualization, and analysis of macromolecular structures. *PLOS Computational Biology*, 13(6), e1005575. <https://doi.org/10.1371/journal.pcbi.1005575>
- Brocchiacono, M., Popov, K. I., Koes, D. R., & Tropsha, A. (2023). *PLANTAIN: Diffusion-inspired pose score minimization for fast and accurate molecular docking* (preprint). <https://doi.org/10.48550/arXiv.2307.12090>
- Brown, N., Fiscato, M., Segler, M. H., & Vaucher, A. C. (2019). GuacaMol: Benchmarking models for de novo molecular design. *Journal of Chemical Information and Modeling*, 59(3), 1096–1108. <https://doi.org/10.1021/acs.jcim.8b00839>
- Buchfink, B., Reuter, K., & Drost, H.-G. (2021). Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nature Methods*, 18(4), 366–368. <https://doi.org/10.1038/s41592-021-01101-x>
- Burnell, E. S., Jane, D. E., & Thatcher, R. J. (2016). Ligand-based drug design. In *Textbook of drug design and discovery* (5th ed., pp. 29–44). Chapman and Hall/CRC.

- Butina, D. (1999). Unsupervised data base clustering based on daylight's fingerprint and tanimoto similarity: A fast and automated way to cluster small and large data sets. *Journal of Chemical Information and Computer Sciences*, 39(4), 747–750. <https://doi.org/10.1021/ci9803381>
- Buttenschoen, M., Morris, G. M., & Deane, C. M. (2024). PoseBusters: AI-based docking methods fail to generate physically valid poses or generalise to novel sequences. *Chemical Science*, 15(9), 3130–3139. <https://doi.org/10.1039/D3SC04185A>
- Buttenschoen, M., Ziv, Y., Morris, G. M., & Deane, C. (2025). An evaluation of unconditional 3D molecular generation methods. *ICLR 2025 Workshop on Generative and Experimental Perspectives for Biomolecular Design*. <https://openreview.net/forum?id=ZU1m9Y4tYy>
- Byrd, R. H., Lu, P., Nocedal, J., & Zhu, C. (1995). A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, 16(5), 1190–1208. <https://doi.org/10.1137/0916069>
- Cao, D., Chen, M., Zhang, R., Wang, Z., Huang, M., Yu, J., Jiang, X., Fan, Z., Zhang, W., Zhou, H., Li, X., Fu, Z., Zhang, S., & Zheng, M. (2025). SurfDock is a surface-informed diffusion generative model for reliable and accurate protein–ligand complex prediction. *Nature Methods*, 22(2), 310–322. <https://doi.org/10.1038/s41592-024-02516-y>
- Casbarra, L., & Procacci, P. (2021). Binding free energy predictions in host-guest systems using Autodock4. A retrospective analysis on SAMPL6, SAMPL7 and SAMPL8 challenges. *Journal of Computer-Aided Molecular Design*, 35(6), 721–729. <https://doi.org/10.1007/s10822-021-00388-4>
- Case, D. A., Aktulga, H. M., Belfon, K., Cerutti, D. S., Cisneros, G. A., Cruzeiro, V. W. D., Forouzes, N., Giese, T. J., Götz, A. W., Gohlke, H., Izadi, S., Kasavajhala, K., Kaymak, M. C., King, E., Kurtzman, T., Lee, T.-S., Li, P., Liu, J., Luchko, T., ... Merz, K. M. (2023). AmberTools. *Journal of Chemical Information and Modeling*, 63(20), 6183–6191. <https://doi.org/10.1021/acs.jcim.3c01153>
- Chen, R., Li, L., & Weng, Z. (2003). ZDOCK: An initial-stage protein-docking algorithm. *Proteins: Structure, Function, and Bioinformatics*, 52(1), 80–87. <https://doi.org/10.1002/prot.10389>
- Chen, X., Zhang, Y., Lu, C., Ma, W., Guan, J., Gong, C., Yang, J., Zhang, H., Zhang, K., Wu, S., Zhou, K., Yang, Y., Liu, Z., Wang, L., Shi, B., Shi, S., & Xiao, W. (2025). *Protenix — advancing structure prediction through a comprehensive AlphaFold3 reproduction* (preprint). <https://doi.org/10.1101/2025.01.08.631967>
- Cheng, T., Li, X., Li, Y., Liu, Z., & Wang, R. (2009). Comparative assessment of scoring functions on a diverse test set. *Journal of Chemical Information and Modeling*, 49(4), 1079–1093. <https://doi.org/10.1021/ci9000053>
- Chung, H., Kim, J., Mccann, M. T., Klasky, M. L., & Ye, J. C. (2023). Diffusion posterior sampling for general noisy inverse problems. *The Eleventh International Conference on Learning Representations*. <https://openreview.net/forum?id=OnD9zGAGT0k>
- Chupakhin, V., Marcou, G., Gaspar, H., & Varnek, A. (2014). Simple ligand–receptor interaction descriptor (SILIRID) for alignment-free binding site comparison. *Computational and Structural Biotechnology Journal*, 10(16), 33–37. <https://doi.org/10.1016/j.csbj.2014.05.004>
- Ciancetta, A., Cuzzolin, A., & Moro, S. (2014). Alternative quality assessment strategy to compare performances of GPCR-ligand docking protocols: The human adenosine A_{2A} receptor as a case study. *Journal of Chemical Information and Modeling*, 54(8), 2243–2254. <https://doi.org/10.1021/ci5002857>
- Clark, J. J., Benson, M. L., Smith, R. D., & Carlson, H. A. (2019). Inherent versus induced protein flexibility: Comparisons within and between apo and holo structures. *PLOS Computational Biology*, 15(1), e1006705. <https://doi.org/10.1371/journal.pcbi.1006705>
- Clark, K. P., & Ajay. (1995). Flexible ligand docking without parameter adjustment across four ligand–receptor complexes. *Journal of Computational Chemistry*, 16(10), 1210–1226. <https://doi.org/10.1002/jcc.540161004>
- Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., & De Hoon, M. J. L. (2009). Biopython: Freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25(11), 1422–1423. <https://doi.org/10.1093/bioinformatics/btp163>

- Cole, J. C., Murray, C. W., Nissink, J. W. M., Taylor, R. D., & Taylor, R. (2005). Comparing protein-ligand docking programs is difficult. *Proteins: Structure, Function, and Bioinformatics*, 60(3), 325–332. <https://doi.org/10.1002/prot.20497>
- Coleman, R. G., Carchia, M., Sterling, T., Irwin, J. J., & Shoichet, B. K. (2013). Ligand pose and orientational sampling in molecular docking. *PLOS One*, 8(10), e75992. <https://doi.org/10.1371/journal.pone.0075992>
- Corbeil, C. R., Williams, C. I., & Labute, P. (2012). Variability in docking success rates due to dataset preparation. *Journal of Computer-Aided Molecular Design*, 26(6), 775–786. <https://doi.org/10.1007/s10822-012-9570-1>
- Corso, G., Deng, A., Fry, B., Polizzi, N., Barzilay, R., & Jaakkola, T. (2024). Deep confident steps to new pockets: Strategies for docking generalization. *The Twelfth International Conference on Learning Representations*. <https://openreview.net/forum?id=UfBlxpTK10>
- Corso, G., Stärk, H., Jing, B., Barzilay, R., & Jaakkola, T. (2023). DiffDock: Diffusion steps, twists, and turns for molecular docking. *The Eleventh International Conference on Learning Representations*. https://openreview.net/forum?id=kKF8_K-mBbS
- Crowfoot, D., Bunn, C. W., Rogers-Low, B. W., & Turner-Jones, A. (1949). The X-ray crystallographic investigation of the structure of penicillin. In H. T. Clarke (Ed.), *Chemistry of Penicillin* (pp. 310–366). Princeton University Press. <https://doi.org/10.1515/9781400874910-012>
- Da, C., & Kireev, D. (2014). Structural protein–ligand interaction fingerprints (SPLIF) for structure-based virtual screening: Method and benchmark study. *Journal of Chemical Information and Modeling*, 54(9), 2555–2561. <https://doi.org/10.1021/ci500319f>
- Daigavane, A., Ravindran, B., & Aggarwal, G. (2021). Understanding convolutions on graphs. *Distill*, 6(8), 10.23915/distill.00032. <https://doi.org/10.23915/distill.00032>
- Das, S., Shimshi, M., Raz, K., Nitoker Eliaz, N., Mhashal, A. R., Ansbacher, T., & Major, D. T. (2019). EnzyDock: Protein–ligand docking of multiple reactive states along a reaction coordinate in enzymes. *Journal of Chemical Theory and Computation*, 15(9), 5116–5134. <https://doi.org/10.1021/acs.jctc.9b00366>
- De Ruyck, J., Brysbaert, G., Blossey, R., & Lensink, M. (2016). Molecular docking as a popular tool in drug design, an in silico travel. *Advances and Applications in Bioinformatics and Chemistry*, 9, 1–11. <https://doi.org/10.2147/AABC.S105289>
- Deane, P. M. (1990). Molecular recognition: The measurement and search for molecular similarity in ligand–receptor interaction. In M. A. Johnson & G. M. Maggiora (Eds.), *Concepts and applications of molecular similarity* (pp. 211–238). Wiley.
- DeLano, W. L. (2002). PyMOL: An open-source molecular graphics tool. *CCP4 Newsletter on Protein Crystallography*, 40(1), 82–92.
- Desaphy, J., Bret, G., Rognan, D., & Kellenberger, E. (2015). {Sc-PDB}: A 3D-database of ligandable binding sites—10 years on. *Nucleic Acids Research*, 43(D1), D399–D404. <https://doi.org/10.1093/nar/gku928>
- Dhariwal, P., & Nichol, A. (2021). Diffusion Models Beat GANs on Image Synthesis. *Advances in Neural Information Processing Systems*, 34, 8780–8794. <https://openreview.net/forum?id=AAWuCvzaVt>
- Didi, K., Vargas, F., Mathis, S., Dutordoir, V., Mathieu, E., Komorowska, U. J., & Lio, P. (2023). A framework for conditional diffusion modelling with applications in motif scaffolding for protein design. *NeurIPS 2023 Workshop on New Frontiers of AI for Drug Discovery and Development*. <https://openreview.net/forum?id=k2PA7CUUJH>
- Drenth, J. (2007). Crystallizing a Protein. In J. Drenth (Ed.), *Principles of Protein X-Ray Crystallography* (pp. 1–20). Springer. https://doi.org/10.1007/0-387-33746-6_1
- Drwal, M. N., Bret, G., Perez, C., Jacquemard, C., Desaphy, J., & Kellenberger, E. (2018). Structural insights on fragment binding mode conservation. *Journal of Medicinal Chemistry*, 61(14), 5963–5973. <https://doi.org/10.1021/acs.jmedchem.8b00256>
- Dunn, I., Pirhadi, S., Wang, Y., Ravindran, S., Concepcion, C., & Koes, D. R. (2024). CACHE challenge #1: Docking with GNINA is all you need. *Journal of Chemical Information and Modeling*, 64(24), 9388–9396. <https://doi.org/10.1021/acs.jcim.4c01429>
- Durairaj, J., Adeshina, Y., Cao, Z., Zhang, X., Oleinikovas, V., Duignan, T., McClure, Z., Robin, X., Studer, G., Kovtun, D., Rossi, E., Zhou, G., Veccham, S., Isert, C., Peng, Y., Sundareson, P.,

- Akdel, M., Corso, G., Stärk, H., ... Naef, L. (2024). *PLINDER: The protein-ligand interactions dataset and evaluation resource* (preprint). <https://doi.org/10.1101/2024.07.17.603955>
- Durant, G., Boyles, F., Birchall, K., Marsden, B., & Deane, C. M. (2025). Robustly interrogating machine learning-based scoring functions: What are they learning? *Bioinformatics*, 41(2), btaf040. <https://doi.org/10.1093/bioinformatics/btaf040>
- Durant, J. L., Leland, B. A., Henry, D. R., & Nourse, J. G. (2002). Reoptimization of MDL keys for use in drug discovery. *Journal of Chemical Information and Computer Sciences*, 42(6), 1273–1280. <https://doi.org/10.1021/ci010132r>
- Durrant, J. D., & McCammon, J. A. (2010). NNScore: A Neural-Network-Based Scoring Function for the Characterization of Protein-Ligand Complexes. *Journal of Chemical Information and Modeling*, 50(10), 1865–1871. <https://doi.org/10.1021/ci100244v>
- Durrant, J. D., & McCammon, J. A. (2011). NNScore 2.0: A Neural-Network Receptor-Ligand Scoring Function. *Journal of Chemical Information and Modeling*, 51(11), 2897–2903. <https://doi.org/10.1021/ci2003889>
- Eastman, P., Galvelis, R., Peláez, R. P., Abreu, C. R. A., Farr, S. E., Gallicchio, E., Gorenko, A., Henry, M. M., Hu, F., Huang, J., Krämer, A., Michel, J., Mitchell, J. A., Pande, V. S., Rodrigues, J. P., Rodriguez-Guerra, J., Simmonett, A. C., Singh, S., Swails, J., ... Markland, T. E. (2024). OpenMM 8: Molecular dynamics simulation with machine learning potentials. *The Journal of Physical Chemistry B*, 128(1), 109–116. <https://doi.org/10.1021/acs.jpcc.3c06662>
- Eastman, P., Swails, J., Chodera, J. D., McGibbon, R. T., Zhao, Y., Beauchamp, K. A., Wang, L.-P., Simmonett, A. C., Harrigan, M. P., Stern, C. D., Wiewiora, R. P., Brooks, B. R., & Pande, V. S. (2017). OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. *PLOS Computational Biology*, 13(7), e1005659. <https://doi.org/10.1371/journal.pcbi.1005659>
- Efron, B. (2011). Tweedie's formula and selection bias. *Journal of the American Statistical Association*, 106(496), 1602–1614. <https://doi.org/10.1198/jasa.2011.tm11181>
- Eisen, M. B., Wiley, D. C., Karplus, M., & Hubbard, R. E. (1994). Hook: A program for finding novel molecular architectures that satisfy the chemical and steric requirements of a macromolecule binding site. *Proteins: Structure, Function, and Bioinformatics*, 19(3), 199–221. <https://doi.org/10.1002/prot.340190305>
- Epstein, D., Jabri, A., Poole, B., Efros, A., & Holynski, A. (2023). Diffusion self-guidance for controllable image generation. *Advances in Neural Information Processing Systems*, 36, 16222–16239. <https://openreview.net/forum?id=qgv56R2YJ7>
- Errington, D., Schneider, C., Bouysset, C., & Dreyer, F. A. (2025). Assessing interaction recovery of predicted protein-ligand poses. *Journal of Cheminformatics*, 17(1), 76. <https://doi.org/10.1186/s13321-025-01011-6>
- Ertl, P., & Schuffenhauer, A. (2009). Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *Journal of Cheminformatics*, 1(1), 8. <https://doi.org/10.1186/1758-2946-1-8>
- Feng, Z., Chen, L., Maddula, H., Akcan, O., Oughtred, R., Berman, H. M., & Westbrook, J. (2004). Ligand Depot: A data warehouse for ligands bound to macromolecules. *Bioinformatics*, 20(13), 2153–2155. <https://doi.org/10.1093/bioinformatics/bth214>
- Ferreira de Freitas, R., & Schapira, M. (2017). A systematic analysis of atomic protein–ligand interactions in the PDB. *MedChemComm*, 8(10), 1970–1981. <https://doi.org/10.1039/C7MD00381A>
- Forli, S., Huey, R., Pique, M. E., Sanner, M. F., Goodsell, D. S., & Olson, A. J. (2016). Computational protein–ligand docking and virtual drug screening with the AutoDock suite. *Nature Protocols*, 11(5), 905–919. <https://doi.org/10.1038/nprot.2016.051>
- Forli, S., & Olson, A. J. (2012). A force field with discrete displaceable waters and desolvation entropy for hydrated ligand docking. *Journal of Medicinal Chemistry*, 55(2), 623–638. <https://doi.org/10.1021/jm2005145>
- Francoeur, P. G., Masuda, T., Sunseri, J., Jia, A., Iovanisci, R. B., Snyder, I., & Koes, D. R. (2020). Three-dimensional convolutional neural networks and a cross-docked data set for structure-based drug design. *Journal of Chemical Information and Modeling*, 60(9), 4200–4215. <https://doi.org/10.1021/acs.jcim.0c00411>

- Friedrich, N.-O., de Bruyn Kops, C., Flachsenberg, F., Sommer, K., Rarey, M., & Kirchmair, J. (2017). Benchmarking commercial conformer ensemble generators. *Journal of Chemical Information and Modeling*, 57(11), 2719–2728. <https://doi.org/10.1021/acs.jcim.7b00505>
- Friedrich, N.-O., Meyder, A., de Bruyn Kops, C., Sommer, K., Flachsenberg, F., Rarey, M., & Kirchmair, J. (2017). High-quality dataset of protein-bound ligand conformations and its application to benchmarking conformer ensemble generators. *Journal of Chemical Information and Modeling*, 57(3), 529–539. <https://doi.org/10.1021/acs.jcim.6b00613>
- Friesner, R. A., Banks, J. L., Murphy, R. B., Halgren, T. A., Klicic, J. J., Mainz, D. T., Repasky, M. P., Knoll, E. H., Shelley, M., Perry, J. K., Shaw, D. E., Francis, P., & Shenkin, P. S. (2004). Glide: A new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *Journal of Medicinal Chemistry*, 47(7), 1739–1749. <https://doi.org/10.1021/jm0306430>
- Fuchs, F. B., Worrall, D. E., Fischer, V., & Welling, M. (2020). SE(3)-Transformers: 3D Roto-Translation Equivariant Attention Networks. *Proceedings of the 34th Conference on Neural Information Processing Systems*. <https://doi.org/10.48550/arXiv.2006.10503>
- Funahashi, K.-I. (1989). On the approximate realization of continuous mappings by neural networks. *Neural Networks*, 2(3), 183–192. [https://doi.org/10.1016/0893-6080\(89\)90003-8](https://doi.org/10.1016/0893-6080(89)90003-8)
- Geiger, M., & Smidt, T. (2022). E3NN: Euclidean neural networks (preprint). <https://doi.org/10.48550/arXiv.2207.09453>
- Giangreco, I., Cosgrove, D. A., & Packer, M. J. (2013). An extensive and diverse set of molecular overlays for the validation of pharmacophore programs. *Journal of Chemical Information and Modeling*, 53(4), 852–866. <https://doi.org/10.1021/ci400020a>
- Glorot, X., & Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, 249–256. <https://proceedings.mlr.press/v9/glorot10a.html>
- Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.
- Goodman, J. M., Pletnev, I., Thiessen, P., Bolton, E., & Heller, S. R. (2021). InChI version 1.06: Now more than 99.99% reliable. *Journal of Cheminformatics*, 13(1), 40. <https://doi.org/10.1186/s13321-021-00517-z>
- Goodsell, D. S., & Olson, A. J. (1990). Automated docking of substrates to proteins by simulated annealing. *Proteins: Structure, Function, and Bioinformatics*, 8(3), 195–202. <https://doi.org/10.1002/prot.340080302>
- Gorgulla, C., Boeszoermenyi, A., Wang, Z.-F., Fischer, P. D., Coote, P. W., Padmanabha Das, K. M., Malets, Y. S., Radchenko, D. S., Moroz, Y. S., Scott, D. A., Fackeldey, K., Hoffmann, M., Iavniuk, I., Wagner, G., & Arthanari, H. (2020). An open-source drug discovery platform enables ultra-large virtual screens. *Nature*, 580(7805), 663–668. <https://doi.org/10.1038/s41586-020-2117-z>
- Grosdidier, A., Zoete, V., & Michielin, O. (2011). SwissDock, a protein-small molecule docking web service based on EADock DSS. *Nucleic Acids Research*, 39, W270–W277. <https://doi.org/10.1093/nar/gkr366>
- Guo, H., Liu, S., Hu, M., Lou, Y., & Bingyi, J. (2023). DiffDock-site: A novel paradigm for enhanced protein-ligand predictions through binding site identification. *Generative AI and Biology (GenBio) Workshop*.
- Hägg, G., & Crowfoot Hodgkin, D. (with Nobel Foundation). (1999). Chemistry 1964. In *Nobel lectures including presentation speeches and laureates' biographies: Chemistry, 1963-1970* (pp. 65–94). World Scientific.
- Halgren, T. A. (1996). Merck molecular force field. I. basis, form, scope, parameterization, and performance of MMFF94. *Journal of Computational Chemistry*, 17(5–6), 490–519. [https://doi.org/10.1002/\(SICI\)1096-987X\(199604\)17:5/6<490::AID-JCC1>3.0.CO;2-P](https://doi.org/10.1002/(SICI)1096-987X(199604)17:5/6<490::AID-JCC1>3.0.CO;2-P)
- Hartshorn, M. J., Verdonk, M. L., Chessari, G., Brewerton, S. C., Mooij, W. T. M., Mortenson, P. N., & Murray, C. W. (2007). Diverse, high-quality test set for the validation of protein-ligand docking performance. *Journal of Medicinal Chemistry*, 50(4), 726–741. <https://doi.org/10.1021/jm061277y>
- Hawkins, P. C. D., Skillman, A. G., Warren, G. L., Ellingson, B. A., & Stahl, M. T. (2010). Conformer generation with OMEGA: Algorithm and validation using high quality structures from the

- protein databank and cambridge structural database. *Journal of Chemical Information and Modeling*, 50(4), 572–584. <https://doi.org/10.1021/ci100031x>
- Heller, S. R., McNaught, A., Pletnev, I., Stein, S., & Tchekhovskoi, D. (2015). InChI, the IUPAC international chemical identifier. *Journal of Cheminformatics*, 7(1), 1–34. <https://doi.org/10.1186/s13321-015-0068-4>
- Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 34. <https://doi.org/10.48550/arXiv.2006.11239>
- Ho, J., & Salimans, T. (2021). Classifier-free diffusion guidance. *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*. <https://doi.org/10.48550/arXiv.2207.12598>
- Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5), 359–366. [https://doi.org/10.1016/0893-6080\(89\)90020-8](https://doi.org/10.1016/0893-6080(89)90020-8)
- Hu, L., Benson, M. L., Smith, R. D., Lerner, M. G., & Carlson, H. A. (2005). Binding MOAD (Mother Of All Databases). *Proteins: Structure, Function, and Bioinformatics*, 60(3), 333–340. <https://doi.org/10.1002/prot.20512>
- Jain, A. N. (2003). Surflex: Fully automatic flexible molecular docking using a molecular similarity-based search engine. *Journal of Medicinal Chemistry*, 46(4), 499–511. <https://doi.org/10.1021/jm020406h>
- Jakalian, A., Jack, D. B., & Bayly, C. I. (2002). Fast, efficient generation of high-quality atomic charges. am1-bcc model: Ii. parameterization and validation. *Journal of Computational Chemistry*, 23(16), 1623–1641. <https://doi.org/10.1002/jcc.10128>
- Jones, G., Willett, P., & Glen, R. C. (1995). Molecular recognition of receptor sites using a genetic algorithm with a description of desolvation. *Journal of Molecular Biology*, 245(1), 43–53. [https://doi.org/10.1016/S0022-2836\(95\)80037-9](https://doi.org/10.1016/S0022-2836(95)80037-9)
- Jones, G., Willett, P., Glen, R. C., Leach, A. R., & Taylor, R. (1997). Development and validation of a genetic algorithm for flexible docking. *Journal of Molecular Biology*, 267(3), 727–748. <https://doi.org/10.1006/jmbi.1996.0897>
- Jubb, H. C., Higuero, A. P., Ochoa-Montano, B., Pitt, W. R., Ascher, D. B., & Blundell, T. L. (2017). Arpeggio: A web server for calculating and visualising interatomic interactions in protein structures. *Journal of Molecular Biology*, 429(3), 365–371. <https://doi.org/10.1016/j.jmb.2016.12.004>
- Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Židek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., ... Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873), 583–589. <https://doi.org/10.1038/s41586-021-03819-2>
- Kabsch, W. (1976). A solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A*, 32(5), 922–923. <https://doi.org/10.1107/S0567739476001873>
- Kellenberger, E., Muller, P., Schalon, C., Bret, G., Foata, N., & Rognan, D. (2006). Sc-PDB: An annotated database of druggable binding sites from the protein data bank. *Journal of Chemical Information and Modeling*, 46(2), 717–727. <https://doi.org/10.1021/ci050372x>
- Kelley, B. P., Brown, S. P., Warren, G. L., & Muchmore, S. W. (2015). POSIT: Flexible Shape-Guided Docking For Pose Prediction. *Journal of Chemical Information and Modeling*, 55(8), 1771–1780. <https://doi.org/10.1021/acs.jcim.5b00142>
- Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *3rd International Conference on Learning Representations*. <https://doi.org/10.48550/arXiv.1412.6980>
- Kleywegt, G. J., & Jones, T. A. (1997). Model building and refinement practice. In *Methods in Enzymology* (pp. 208–230, Vol. 277). Academic Press. [https://doi.org/10.1016/S0076-6879\(97\)77013-7](https://doi.org/10.1016/S0076-6879(97)77013-7)
- Koes, D. R., Baumgartner, M. P., & Camacho, C. J. (2013). Lessons learned in empirical scoring with smina from the CSAR 2011 benchmarking exercise. *Journal of Chemical Information and Modeling*, 53(8), 1893–1904. <https://doi.org/10.1021/ci300604z>
- Krishna, R., Wang, J., Ahern, W., Sturm, P., Venkatesh, P., Kalvet, I., Lee, G. R., Morey-Burrows, F. S., Anishchenko, I., Humphreys, I. R., McHugh, R., Vafeados, D., Li, X., Sutherland, G. A., Hitchcock, A., Hunter, C. N., Kang, A., Brackenbrough, E., Bera, A. K., ... Baker, D. (2024). Generalized biomolecular modeling and design with RoseTTAFold all-atom. *Science*, 384(6693), ead12528. <https://doi.org/10.1126/science.ad12528>

- Krivák, R., & Hoksza, D. (2018). P2Rank: Machine learning based tool for rapid and accurate prediction of ligand binding sites from protein structure. *Journal of Cheminformatics*, 10(1), 39. <https://doi.org/10.1186/s13321-018-0285-8>
- Kroupova, A., Spiteri, V. A., Rutter, Z. J., Furihata, H., Darren, D., Ramachandran, S., Chakraborti, S., Haubrich, K., Pethe, J., Gonzales, D., Wijaya, A. J., Rodriguez-Rios, M., Sturbaut, M., Lynch, D. M., Farnaby, W., Nakasone, M. A., Zollman, D., & Ciulli, A. (2024). Design of a Cereblon construct for crystallographic and biophysical studies of protein degraders. *Nature Communications*, 15(1), 8885. <https://doi.org/10.1038/s41467-024-52871-9>
- Kryshtafovych, A., Schwede, T., Topf, M., Fidelis, K., & Moult, J. (2023). Critical assessment of methods of protein structure prediction (CASP)—round XV. *Proteins: Structure, Function, and Bioinformatics*, 91(12), 1539–1549. <https://doi.org/10.1002/prot.26617>
- Kuntz, I. D., Blaney, J. M., Oatley, S. J., Langridge, R., & Ferrin, T. E. (1982). A geometric approach to macromolecule-ligand interactions. *Journal of Molecular Biology*, 161(2), 269–288. [https://doi.org/10.1016/0022-2836\(82\)90153-X](https://doi.org/10.1016/0022-2836(82)90153-X)
- Landrum, G., Paolo, T., Kelley, B., Rodriguez, R., Cosgrove, D., Vianello, R., Riniker, S., Gedeck, P., Jones, G., Kawashima, E., Schneider, N., Nealschneider, D., Dalke, A., Swain, M., Cole, B., Turk, S., Savelev, A., Hurst, T., Vaucher, A., ... Maeder, N. (2025). *RDKit: Open-source cheminformatics*. <https://doi.org/10.5281/zenodo.591637>
- Lanman, B. A., Allen, J. R., Allen, J. G., Amegadzie, A. K., Ashton, K. S., Booker, S. K., Chen, J. J., Chen, N., Frohn, M. J., Goodman, G., Kopecky, D. J., Liu, L., Lopez, P., Low, J. D., Ma, V., Minatti, A. E., Nguyen, T. T., Nishimura, N., Pickrell, A. J., ... Cee, V. J. (2020). Discovery of a Covalent Inhibitor of KRAS^{G12C} (AMG 510) for the Treatment of Solid Tumors. *Journal of Medicinal Chemistry*, 63(1), 52–65. <https://doi.org/10.1021/acs.jmedchem.9b01180>
- Leung, S. H., Bodkin, M. J., & Morris, G. M. (2019). *SuCOS is better than RMSD for evaluating fragment elaboration and docking poses* (preprint). <https://doi.org/10.26434/chemrxiv.8100203.v1>
- Li, L., Chen, R., & Weng, Z. (2003). RDOCK: Refinement of rigid-body protein docking predictions. *Proteins: Structure, Function, and Bioinformatics*, 53(3), 693–707. <https://doi.org/10.1002/prot.10460>
- Li, M., Song, K., He, J., Zhao, M., You, G., Zhong, J., Zhao, M., Li, A., Chen, Y., Li, G., Kong, Y., Wei, J., Wang, Z., Zhou, J., Yang, H., Ma, S., Zhang, H., Méliá, I. L., Lin, W., ... Zhang, J. (2025). Electron-density-informed effective and reliable de novo molecular design and optimization with ED2Mol. *Nature Machine Intelligence*, 7(8), 1355–1368. <https://doi.org/10.1038/s42256-025-01095-7>
- Li, Y., Liu, Z., Li, J., Han, L., Liu, J., Zhao, Z., & Wang, R. (2014). Comparative assessment of scoring functions on an updated benchmark: 1. Compilation of the test set. *Journal of Chemical Information and Modeling*, 54(6), 1700–1716. <https://doi.org/10.1021/ci500080q>
- Liebschner, D., Afonine, P. V., Baker, M. L., Bunkóczi, G., Chen, V. B., Croll, T. I., Hintze, B., Hung, L.-W., Jain, S., McCoy, A. J., Moriarty, N. W., Oeffner, R. D., Poon, B. K., Prisant, M. G., Read, R. J., Richardson, J. S., Richardson, D. C., Sammito, M. D., Sobolev, O. V., ... Adams, P. D. (2019). Macromolecular structure determination using X-rays, neutrons and electrons: Recent developments in Phenix. *Acta Crystallographica Section D: Structural Biology*, 75(10), 861–877. <https://doi.org/10.1107/S2059798319011471>
- Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., dos Santos Costa, A., Fazel-Zarandi, M., Sercu, T., Candido, S., & Rives, A. (2023). Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637), 1123–1130. <https://doi.org/10.1126/science.ade2574>
- Lionta, E., Spyrou, G., Vassilatis, D., & Cournia, Z. (2014). Structure-based virtual screening for drug discovery: Principles, applications and recent advances. *Current Topics in Medicinal Chemistry*, 14(16), 1923–1938. <https://doi.org/10.2174/1568026614666140929124445>
- Lipinski, C. A., Lombardo, F., Dominy, B. W., & Feeney, P. J. (1997). Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced Drug Delivery Reviews*, 23, 3–25. [https://doi.org/10.1016/s0169-409x\(00\)00129-0](https://doi.org/10.1016/s0169-409x(00)00129-0)
- Liu, L., Zhang, S., Xue, Y., Ye, X., Zhu, K., Li, Y., Liu, Y., Zhang, X., & Fang, X. (2024). *Technical Report of HelixFold3 for Biomolecular Structure Prediction* (preprint). <https://doi.org/10.48550/arXiv.2408.16975>

- Liu, M., & Wang, S. (1999). MCDock: A Monte Carlo simulation approach to the molecular docking problem. *Journal of Computer-Aided Molecular Design*, 13(5), 435–451. <https://doi.org/10.1023/A:1008005918983>
- Liu, Z., Li, Y., Han, L., Li, J., Liu, J., Zhao, Z., Nie, W., Liu, Y., & Wang, R. (2015). PDB-wide collection of binding data: Current status of the pdbind database. *Bioinformatics*, 31(3), 405–412. <https://doi.org/10.1093/bioinformatics/btu626>
- Liu, Z., Su, M., Han, L., Liu, J., Yang, Q., Li, Y., & Wang, R. (2017). Forging the basis for developing protein–ligand interaction scoring functions. *Accounts of Chemical Research*, 50(2), 302–309. <https://doi.org/10.1021/acs.accounts.6b00491>
- Lu, W., Wu, Q., Zhang, J., Rao, J., Li, C., & Zheng, S. (2022). TANKBind: Trigonometry-aware neural networks for drug-protein binding structure prediction. *Advances in Neural Information Processing Systems*, 35, 7236–7249. <https://openreview.net/forum?id=MSBDFwGYwwt>
- Lu, W., Zhang, J., Huang, W., Zhang, Z., Jia, X., Wang, Z., Shi, L., Li, C., Wolynes, P. G., & Zheng, S. (2024). DynamicBind: Predicting ligand-specific protein-ligand complex structure with a deep equivariant generative model. *Nature Communications*, 15(1), 1071. <https://doi.org/10.1038/s41467-024-45461-2>
- Lugmayr, A., Danelljan, M., Romero, A., Yu, F., Timofte, R., & Van Gool, L. (2022). RePaint: Inpainting using denoising diffusion probabilistic models. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 11461–11471.
- Maier, J. A., Martinez, C., Kasavajhala, K., Wickstrom, L., Hauser, K. E., & Simmerling, C. (2015). ff14SB: Improving the accuracy of protein side chain and backbone parameters from ff99SB. *Journal of Chemical Theory and Computation*, 11(8), 3696–3713. <https://doi.org/10.1021/acs.jctc.5b00255>
- Majeux, N., Scarsi, M., Apostolakis, J., Ehrhardt, C., & Caflisch, A. (1999). Exhaustive docking of molecular fragments with electrostatic solvation. *Proteins: Structure, Function, and Bioinformatics*, 37(1), 88–105. [https://doi.org/10.1002/\(SICI\)1097-0134\(19991001\)37:1<88::AID-PROT9>3.0.CO;2-O](https://doi.org/10.1002/(SICI)1097-0134(19991001)37:1<88::AID-PROT9>3.0.CO;2-O)
- Malhotra, S., & Karanickolas, J. (2017). When does chemical elaboration induce a ligand to change its binding mode? *Journal of Medicinal Chemistry*, 60(1), 128–145. <https://doi.org/10.1021/acs.jmedchem.6b00725>
- Masters, M. R., Mahmoud, A. H., & Lill, M. A. (2023). FusionDock: Physics-informed diffusion model for molecular docking. *ICML 2023 Workshop on Computational Biology*.
- McGann, M. (2011). FRED Pose Prediction and Virtual Screening Accuracy. *Journal of Chemical Information and Modeling*, 51(3), 578–596. <https://doi.org/10.1021/ci100436p>
- McGann, M. R., Almond, H. R., Nicholls, A., Grant, J. A., & Brown, F. K. (2003). Gaussian docking functions. *Biopolymers*, 68(1), 76–90. <https://doi.org/10.1002/bip.10207>
- McNutt, A. T., Francoeur, P., Aggarwal, R., Masuda, T., Meli, R., Ragoza, M., Sunseri, J., & Koes, D. R. (2021). GNINA 1.0: Molecular docking with deep learning. *Journal of Cheminformatics*, 13(1), 13–43. <https://doi.org/10.1186/s13321-021-00522-2>
- Meiler, J., & Baker, D. (2006). RosettaLigand: Protein–small molecule docking with full side-chain flexibility. *Proteins: Structure, Function, and Bioinformatics*, 65(3), 538–548. <https://doi.org/10.1002/prot.21086>
- Meli, R., & Biggin, P. C. (2020). Spyrmsd: Symmetry-corrected RMSD calculations in Python. *Journal of Cheminformatics*, 12(1). <https://doi.org/10.1186/s13321-020-00455-2>
- Méndez-Lucio, O., Ahmad, M., del Rio-Chanona, E. A., & Wegner, J. K. (2021). A geometric deep learning approach to predict binding conformations of bioactive molecules. *Nature Machine Intelligence*, 3(12), 1033–1039. <https://doi.org/10.1038/s42256-021-00409-9>
- Meslamani, J., Rognan, D., & Kellenberger, E. (2011). Sc-PDB: A database for identifying variations and multiplicity of ‘druggable’ binding sites in proteins. *Bioinformatics*, 27(9), 1324–1326. <https://doi.org/10.1093/bioinformatics/btr120>
- Morehead, A., & Cheng, J. (2025). FlowDock: Geometric flow matching for generative protein–ligand docking and affinity prediction. *Bioinformatics*, 41, i198–i206. <https://doi.org/10.1093/bioinformatics/btaf187>
- Morris, G. M., Goodsell, D. S., Halliday, R. S., Huey, R., Hart, W. E., Belew, R. K., & Olson, A. J. (1998). Automated docking using a Lamarckian genetic algorithm and an empirical binding free

- energy function. *Journal of Computational Chemistry*, 19(14), 1639–1662.
[https://doi.org/10.1002/\(SICI\)1096-987X\(19981115\)19:14<1639::AID-JCC10>3.0.CO;2-B](https://doi.org/10.1002/(SICI)1096-987X(19981115)19:14<1639::AID-JCC10>3.0.CO;2-B)
- Morris, G. M., Goodsell, D. S., Huey, R., & Olson, A. J. (1996). Distributed automated docking of flexible ligands to proteins: Parallel applications of AutoDock 2.4. *Journal of Computer-Aided Molecular Design*, 10(4), 293–304. <https://doi.org/10.1007/BF00124499>
- Morris, G. M., Huey, R., Lindstrom, W., Sanner, M. F., Belew, R. K., Goodsell, D. S., & Olson, A. J. (2009). AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *Journal of Computational Chemistry*, 30(16), 2785–2791.
<https://doi.org/10.1002/jcc.21256>
- Morris, G. M., & Lim-Wilby, M. (2008). Molecular Docking. In A. Kukol (Ed.), *Molecular Modeling of Proteins* (pp. 365–382). Humana Press. https://doi.org/10.1007/978-1-59745-177-2_19
- Mysinger, M. M., Carchia, M., Irwin, J. J., & Shoichet, B. K. (2012). Directory of useful decoys, enhanced (DUD-E): Better ligands and decoys for better benchmarking. *Journal of Medicinal Chemistry*, 55(14), 6582–6594. <https://doi.org/10.1021/jm300687e>
- Nissink, J. W. M., Murray, C., Hartshorn, M., Verdonk, M. L., Cole, J. C., & Taylor, R. (2002). A new test set for validating predictions of protein-ligand interaction. *Proteins: Structure, Function, and Bioinformatics*, 49(4), 457–471. <https://doi.org/10.1002/prot.10232>
- Nittinger, E., Yoluk, Ö., Tibo, A., Olanders, G., & Tyrchan, C. (2025). Co-folding, the future of docking – prediction of allosteric and orthosteric ligands. *Artificial Intelligence in the Life Sciences*, 8, 100136. <https://doi.org/10.1016/j.ailesci.2025.100136>
- Nobel Foundation. (2024). *The Nobel Prize in Chemistry 2024*. (Press release). Retrieved November 26, 2025, from <https://www.nobelprize.org/uploads/2024/10/press-chemistryprize2024-3.pdf>
- O’Boyle, N. M., Banck, M., James, C. A., Morley, C., Vandermeersch, T., & Hutchison, G. R. (2011). Open Babel: An open chemical toolbox. *Journal of Cheminformatics*, 3(1), 33.
<https://doi.org/10.1186/1758-2946-3-33>
- O’Boyle, N. M., Guha, R., Willighagen, E. L., Adams, S. E., Alvarsson, J., Bradley, J.-C., Filippov, I. V., Hanson, R. M., Hanwell, M. D., Hutchison, G. R., James, C. A., Jeliazkova, N., Lang, A. S. I. D., Langner, K. M., Lonie, D. C., Lowe, D. M., Pansanel, J., Pavlov, D., Spjuth, O., ... Murray-Rust, P. (2011). Open data, open source and open standards in chemistry: The blue obelisk five years on. *Journal of Cheminformatics*, 3(1).
<https://doi.org/10.1186/1758-2946-3-37>
- Onodera, K., Satou, K., & Hirota, H. (2007). Evaluations of molecular docking programs for virtual screening. *Journal of Chemical Information and Modeling*, 47(4), 1609–1618.
<https://doi.org/10.1021/ci7000378>
- Pagadala, N. S., Syed, K., & Tuszynski, J. (2017). Software for molecular docking: A review. *Biophysical Reviews*, 9(2), 91–102. <https://doi.org/10.1007/s12551-016-0247-1>
- Paggi, J. M., Pandit, A., & Dror, R. O. (2024). The art and science of molecular docking. *Annual Review of Biochemistry*, 93, 389–410. <https://doi.org/10.1146/annurev-biochem-030222-120000>
- Pang, Y.-P., Perola, E., Xu, K., & Prendergast, F. G. (2001). EUDOC: A computer program for identification of drug interaction sites in macromolecules and drug leads from chemical databases. *Journal of Computational Chemistry*, 22(15), 1750–1771.
<https://doi.org/10.1002/jcc.1129>
- Pantsar, T., & Poso, A. (2018). Binding affinity via docking: Fact and fiction. *Molecules*, 23(8), 1899.
<https://doi.org/10.3390/molecules23081899>
- Passaro, S., Corso, G., Wohlwend, J., Reveiz, M., Thaler, S., Somnath, V. R., Getz, N., Portnoi, T., Roy, J., Stark, H., Kwabi-Addo, D., Beaini, D., Jaakkola, T., & Barzilay, R. (2025). *Boltz-2: Towards accurate and efficient binding affinity prediction* (preprint).
<https://doi.org/10.1101/2025.06.14.659707>
- Patrick, G. L. (2023). *An introduction to medicinal chemistry* (7th ed.). Oxford University Press.
- Petukh, M., Stefl, S., & Alexov, E. (2013). The role of protonation states in ligand-receptor recognition and binding. *Current Pharmaceutical Design*, 19(23), 4182–4190.
<https://doi.org/10.2174/1381612811319230004>
- Plainer, M., Toth, M., Dobers, S., Stark, H., Corso, G., Marquet, C., & Barzilay, R. (2023). DiffDock-pocket: Diffusion for pocket-level docking with sidechain flexibility. *NeurIPS 2023 Workshop on New Frontiers of AI for Drug Discovery and Development*.
<https://openreview.net/forum?id=5Wt7qNRc5u>

- Plewczynski, D., Łaźniewski, M., Augustyniak, R., & Ginalski, K. (2011). Can we trust docking results? Evaluation of seven commonly used programs on PDBbind database. *Journal of Computational Chemistry*, 32(4), 742–755. <https://doi.org/10.1002/jcc.21643>
- Qiao, Z., Nie, W., Vahdat, A., Miller, T. F., & Anandkumar, A. (2024). State-specific protein–ligand complex structure prediction with a multiscale deep generative model. *Nature Machine Intelligence*, 6(2), 195–208. <https://doi.org/10.1038/s42256-024-00792-z>
- Radoux, C. J., Olsson, T. S. G., Pitt, W. R., Groom, C. R., & Blundell, T. L. (2016). Identifying interactions that determine fragment binding at protein hotspots. *Journal of Medicinal Chemistry*, 59(9), 4314–4325. <https://doi.org/10.1021/acs.jmedchem.5b01980>
- Rappe, A. K., Casewit, C. J., Colwell, K. S., Goddard, W. A., & Skiff, W. M. (1992). UFF, a full periodic table force field for molecular mechanics and molecular dynamics simulations. *Journal of the American Chemical Society*, 114(25), 10024–10035. <https://doi.org/10.1021/ja00051a040>
- Rarey, M., Kramer, B., Lengauer, T., & Klebe, G. (1996). A fast flexible docking method using an incremental construction algorithm. *Journal of Molecular Biology*, 261(3), 470–489. <https://doi.org/10.1006/jmbi.1996.0477>
- Riniker, S., & Landrum, G. A. (2013). Open-source platform to benchmark fingerprints for ligand-based virtual screening. *Journal of Cheminformatics*, 5(1), 26. <https://doi.org/10.1186/1758-2946-5-26>
- Riniker, S., & Landrum, G. A. (2015). Better informed distance geometry: Using what we know to improve conformation generation. *Journal of Chemical Information and Modeling*, 55(12), 2562–2574. <https://doi.org/10.1021/acs.jcim.5b00654>
- Ritter, J. M. (2024). How drugs act. In H. P. Rang & M. M. Dale (Eds.), *Rang & Dale's pharmacology* (10th ed., pp. 18–94). Elsevier.
- Roche, O., Kiyama, R., & Brooks, C. L. (2001). Ligand-Protein DataBase: Linking Protein-Ligand Complex Structures to Binding Data. *Journal of Medicinal Chemistry*, 44(22), 3592–3598. <https://doi.org/10.1021/jm000467k>
- Rogers, D., & Hahn, M. (2010). Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling*, 50(5), 742–754. <https://doi.org/10.1021/ci100050t>
- Rojas, R. (1996). The backpropagation algorithm. In *Neural Networks* (pp. 149–182). Springer. https://doi.org/10.1007/978-3-642-61068-4_7
- Roseman, A. M. (2000). Docking structures of domains into maps from cryo-electron microscopy using local correlation. *Acta Crystallographica Section D Biological Crystallography*, 56(10), 1332–1340. <https://doi.org/10.1107/S0907444900010908>
- Safo, M. K., Moure, C. M., Burnett, J. C., Joshi, G. S., & Abraham, D. J. (2001). High-resolution crystal structure of deoxy hemoglobin complexed with a potent allosteric effector. *Protein Science*, 10(5), 951–957. <https://doi.org/10.1110/ps.50601>
- Salentin, S., Schreiber, S., Haupt, V. J., Adasme, M. F., & Schroeder, M. (2015). PLIP: Fully automated protein–ligand interaction profiler. *Nucleic Acids Research*, 43(W1), W443–W447. <https://doi.org/10.1093/nar/gkv315>
- Sanchez-Lengeling, B., Reif, E., Pearce, A., & Wiltchko, A. (2021). A gentle introduction to graph neural networks. *Distill*, 6(8), 10.23915/distill.00033. <https://doi.org/10.23915/distill.00033>
- Sastry, G. M., Adzhigirey, M., Day, T., Annabhimoju, R., & Sherman, W. (2013). Protein and ligand preparation: Parameters, protocols, and influence on virtual screening enrichments. *Journal of Computer-Aided Molecular Design*, 27(3), 221–234. <https://doi.org/10.1007/s10822-013-9644-8>
- Satorras, V. G., Hoogeboom, E., & Welling, M. (2021). *E(n) equivariant graph neural networks* (preprint). <http://arxiv.org/abs/2102.09844>
- Schake, P., Bolz, S. N., Linnemann, K., & Schroeder, M. (2025). PLIP 2025: Introducing protein–protein interactions to the protein–ligand interaction profiler. *Nucleic Acids Research*, 53(W1), W463–W465. <https://doi.org/10.1093/nar/gkaf361>
- Segler, M. H. S., Preuss, M., & Waller, M. P. (2018). Planning chemical syntheses with deep neural networks and symbolic AI. *Nature*, 555(7698), 604–610. <https://doi.org/10.1038/nature25978>
- Shao, C., Westbrook, J. D., Lu, C., Bhikadiya, C., Peisach, E., Young, J. Y., Duarte, J. M., Lowe, R., Wang, S., Rose, Y., Feng, Z., & Burley, S. K. (2022). Simplified quality assessment for small-molecule ligands in the protein data bank. *Structure*, 30(2), 252–262.e4. <https://doi.org/10.1016/j.str.2021.10.003>

- Shor, B., & Schneidman-Duhovny, D. (2024). DockFormer: Affinity prediction and flexible docking with pair transformer. *PRX Life*, 3(3), 033024. <https://doi.org/10.1103/1862-rs2k>
- Sillitoe, I., Bordin, N., Dawson, N., Waman, V. P., Ashford, P., Scholes, H. M., Pang, C. S. M., Woodridge, L., Rauer, C., Sen, N., Abbasian, M., Le Cornu, S., Lam, S. D., Berka, K., Varekova, I. H., Svobodova, R., Lees, J., & Orengo, C. A. (2021). CATH: Increased structural coverage of functional space. *Nucleic Acids Research*, 49(D1), D266–D273. <https://doi.org/10.1093/nar/gkaa1079>
- Škrinjar, P., Eberhardt, J., Durairaj, J., & Schwede, T. (2025). *Have protein-ligand co-folding methods moved beyond memorisation?* (preprint). <https://doi.org/10.1101/2025.02.03.636309>
- Smart, O. S., Horský, V., Gore, S., Svobodová Vařeková, R., Bendová, V., Kleywegt, G. J., & Velankar, S. (2018). Validation of ligands in macromolecular structures determined by X-ray crystallography. *Acta Crystallographica Section D Structural Biology*, 74(3), 228–236. <https://doi.org/10.1107/S2059798318002541>
- Smith, R. D., Clark, J. J., Ahmed, A., Orban, Z. J., Dunbar, J. B., & Carlson, H. A. (2019). Updates to binding MOAD (mother of all databases): Polypharmacology tools and their utility in drug repurposing. *Journal of Molecular Biology*, 431(13), 2423–2433. <https://doi.org/10.1016/j.jmb.2019.05.024>
- Smith, T., & Waterman, M. (1981). Identification of common molecular subsequences. *Journal of Molecular Biology*, 147(1), 195–197. [https://doi.org/10.1016/0022-2836\(81\)90087-5](https://doi.org/10.1016/0022-2836(81)90087-5)
- Song, J., Meng, C., & Ermon, S. (2021). Denoising diffusion implicit models. *The Ninth International Conference on Learning Representations*. <http://arxiv.org/abs/2010.02502>
- Song, Y., & Ermon, S. (2019). Generative modeling by estimating gradients of the data distribution. *Advances in Neural Information Processing Systems*, 32. <https://openreview.net/forum?id=B1lcYrBgLH>
- Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., & Poole, B. (2021). Score-based generative modeling through stochastic differential equations. *The Ninth International Conference on Learning Representations*. <https://doi.org/10.48550/arXiv.2011.13456>
- Stärk, H., Ganea, O., Pattanaik, L., Barzilay, R., & Jaakkola, T. (2022). EquiBind: Geometric deep learning for drug binding structure prediction. *International Conference on Machine Learning*, 162, 20503–20521. <https://proceedings.mlr.press/v162/stark22b.html>
- Steinegger, M., & Söding, J. (2017). MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature Biotechnology*, 35(11), 1026–1028. <https://doi.org/10.1038/nbt.3988>
- Su, M., Yang, Q., Du, Y., Feng, G., Liu, Z., Li, Y., & Wang, R. (2019). Comparative assessment of scoring functions: The CASF-2016 update. *Journal of Chemical Information and Modeling*, 59(2), 895–913. <https://doi.org/10.1021/acs.jcim.8b00545>
- Suzek, B. E., Wang, Y., Huang, H., McGarvey, P. B., Wu, C. H., & the UniProt Consortium. (2015). UniRef clusters: A comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics*, 31(6), 926–932. <https://doi.org/10.1093/bioinformatics/btu739>
- The OpenFold3 Team. (2025). *OpenFold3-preview* (Version 0.1.0). <https://doi.org/10.5281/zenodo.17485509>
- Thomsen, R., & Christensen, M. H. (2006). MolDock: A New Technique for High-Accuracy Molecular Docking. *Journal of Medicinal Chemistry*, 49(11), 3315–3321. <https://doi.org/10.1021/jm051197e>
- Torres, P. H. M., Sodero, A. C. R., Jofily, P., & Silva-Jr, F. P. (2019). Key topics in molecular docking for drug design. *International Journal of Molecular Sciences*, 20(18), 4574. <https://doi.org/10.3390/ijms20184574>
- Tosco, P., Stiefl, N., & Landrum, G. (2014). Bringing the MMFF force field to the RDKit: Implementation and validation. *Journal of Cheminformatics*, 6(1), 37. <https://doi.org/10.1186/s13321-014-0037-3>
- Tosstorff, A., Rudolph, M. G., Cole, J. C., Reutlinger, M., Kramer, C., Schaffhauser, H., Nilly, A., Flohr, A., & Kuhn, B. (2022). A high quality, industrial data set for binding affinity prediction: Performance comparison in different early drug discovery scenarios. *Journal of Computer-Aided Molecular Design*, 36, 753–765. <https://doi.org/10.1007/s10822-022-00478-x>
- Treyde, W., Kim, S. C., Bouatta, N., & AlQuraishi, M. (2024). QuickBind: A light-weight and interpretable molecular docking model. *Proceedings of the 19th Machine Learning in*

- Computational Biology Meeting*, 261, 129–152.
<https://proceedings.mlr.press/v261/treyde24a.html>
- Trott, O., & Olson, A. J. (2010). AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of Computational Chemistry*, 31, 455–461. <https://doi.org/10.1002/jcc.21334>
- Tunyasuvunakool, K., Adler, J., Wu, Z., Green, T., Zielinski, M., Židek, A., Bridgland, A., Cowie, A., Meyer, C., Laydon, A., Velankar, S., Kleywegt, G. J., Bateman, A., Evans, R., Pritzel, A., Figurnov, M., Ronneberger, O., Bates, R., Kohl, S. A. A., ... Hassabis, D. (2021). Highly accurate protein structure prediction for the human proteome. *Nature*, 596(7873), 590–596. <https://doi.org/10.1038/s41586-021-03828-1>
- Urbaczek, S., Kolodzik, A., Fischer, J. R., Lippert, T., Heuser, S., Groth, I., Schulz-Gasch, T., & Rarey, M. (2011). NAOMI: On the almost trivial task of reading molecules from different file formats. *Journal of Chemical Information and Modeling*, 51(12), 3199–3207. <https://doi.org/10.1021/ci200324e>
- Valeur, E., Guéret, S. M., Adihou, H., Gopalakrishnan, R., Lemurell, M., Waldmann, H., Grossmann, T. N., & Plowright, A. T. (2017). New modalities for challenging targets in drug discovery. *Angewandte Chemie International Edition*, 56(35), 10294–10323. <https://doi.org/10.1002/anie.201611914>
- Varadi, M., Bertoni, D., Magana, P., Paramval, U., Pidruchna, I., Radhakrishnan, M., Tsenkov, M., Nair, S., Mirdita, M., Yeo, J., Kovalevskiy, O., Tunyasuvunakool, K., Laydon, A., Židek, A., Tomlinson, H., Hariharan, D., Abrahamson, J., Green, T., Jumper, J., ... Velankar, S. (2024). AlphaFold protein structure database in 2024: Providing structure coverage for over 214 million protein sequences. *Nucleic Acids Research*, 52(D1), D368–D375. <https://doi.org/10.1093/nar/gkad1011>
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 31. <http://arxiv.org/abs/1706.03762>
- Verdonk, M. L., Cole, J. C., Hartshorn, M. J., Murray, C. W., & Taylor, R. D. (2003). Improved protein–ligand docking using GOLD. *Proteins: Structure, Function, and Bioinformatics*, 52(4), 609–623. <https://doi.org/10.1002/prot.10465>
- Verdonk, M. L., Mortenson, P. N., Hall, R. J., Hartshorn, M. J., & Murray, C. W. (2008). Protein-ligand docking against non-native protein conformers. *Journal of Chemical Information and Modeling*, 48(11), 2214–2225. <https://doi.org/10.1021/ci8002254>
- Wagle, S., Smith, R. D., Dominic, A. J., DasGupta, D., Tripathi, S. K., & Carlson, H. A. (2023). Sunsetting Binding MOAD with its last data update and the addition of 3D-ligand polypharmacology tools. *Scientific Reports*, 13(1), 3008. <https://doi.org/10.1038/s41598-023-29996-w>
- Wang, R., Fang, X., Lu, Y., & Wang, S. (2004). The PDBbind database: Collection of binding affinities for protein-ligand complexes with known three-dimensional structures. *Journal of Medicinal Chemistry*, 47(12), 2977–2980. <https://doi.org/10.1021/jm030580l>
- Wang, R., Fang, X., Lu, Y., Yang, C.-Y., & Wang, S. (2005). The PDBbind database: Methodologies and updates. *Journal of Medicinal Chemistry*, 48(12), 4111–4119. <https://doi.org/10.1021/jm048957q>
- Wang, S., Witek, J., Landrum, G. A., & Riniker, S. (2020). Improving conformer generation for small rings and macrocycles based on distance geometry and experimental torsional-angle preferences. *Journal of Chemical Information and Modeling*, 60(4), 2044–2058. <https://doi.org/10.1021/acs.jcim.0c00025>
- Warren, G. L., Andrews, C. W., Capelli, A.-M., Clarke, B., LaLonde, J., Lambert, M. H., Lindvall, M., Nevins, N., Semus, S. F., Senger, S., Tedesco, G., Wall, I. D., Woolven, J. M., Peishoff, C. E., & Head, M. S. (2006). A critical assessment of docking programs and scoring functions. *Journal of Medicinal Chemistry*, 49(20), 5912–5931. <https://doi.org/10.1021/jm050362n>
- Warren, G. L., Do, T. D., Kelley, B. P., Nicholls, A., & Warren, S. D. (2012). Essential considerations for using protein–ligand structures in drug discovery. *Drug Discovery Today*, 17(23–24), 1270–1281. <https://doi.org/10.1016/j.drudis.2012.06.011>

- Weininger, D. (1988). SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences*, 28(1), 31–36. <https://doi.org/10.1021/ci00057a005>
- Welch, W., Ruppert, J., & Jain, A. N. (1996). Hammerhead: Fast, fully automated docking of flexible ligands to protein binding sites. *Chemistry & Biology*, 3(6), 449–462. [https://doi.org/10.1016/S1074-5521\(96\)90093-9](https://doi.org/10.1016/S1074-5521(96)90093-9)
- Westbrook, J. D., Shao, C., Feng, Z., Zhuravleva, M., Velankar, S., & Young, J. (2015). The chemical component dictionary: Complete descriptions of constituent molecules in experimentally determined 3D macromolecules in the Protein Data Bank. *Bioinformatics*, 31(8), 1274–1278. <https://doi.org/10.1093/bioinformatics/btu789>
- Wildman, S. A., & Crippen, G. M. (1999). Prediction of physicochemical parameters by atomic contributions. *Journal of Chemical Information and Computer Sciences*, 39(5), 868–873. <https://doi.org/10.1021/ci9903071>
- Wills, S., Sanchez-Garcia, R., Dudgeon, T., Roughley, S. D., Merritt, A., Hubbard, R. E., Davidson, J., Von Delft, F., & Deane, C. M. (2023). Fragment merging using a graph database samples different catalogue space than similarity search. *Journal of Chemical Information and Modeling*, 63(11), 3423–3437. <https://doi.org/10.1021/acs.jcim.3c00276>
- Wohlwend, J., Corso, G., Passaro, S., Getz, N., Reveiz, M., Leidal, K., Swiderski, W., Atkinson, L., Portnoi, T., Chinn, I., Silterra, J., Jaakkola, T., & Barzilay, R. (2024). *Boltz-1 democratizing biomolecular interaction modeling* (preprint). <https://doi.org/10.1101/2024.11.19.624167>
- Wójcikowski, M., Kukiela, M., Stepniewska-Dziubinska, M. M., & Siedlecki, P. (2019). Development of a protein–ligand extended connectivity (PLEC) fingerprint and its application for binding affinity predictions. *Bioinformatics*, 35(8), 1334–1341. <https://doi.org/10.1093/bioinformatics/bty757>
- Wójcikowski, M., Zielenkiewicz, P., & Siedlecki, P. (2015). Open drug discovery toolkit (ODDT): A new open-source player in the drug discovery field. *Journal of Cheminformatics*, 7(1), 26. <https://doi.org/10.1186/s13321-015-0078-2>
- Wolpert, D. H. (1996). The lack of a priori distinctions between learning algorithms. *Neural Computation*, 8(7), 1341–1390. <https://doi.org/10.1162/neco.1996.8.7.1341>
- Wrapp, D., Wang, N., Corbett, K. S., Goldsmith, J. A., Hsieh, C.-L., Abiona, O., Graham, B. S., & McLellan, J. S. (2020). Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. *Science*, 367(6483), 1260–1263. <https://doi.org/10.1126/science.abb2507>
- Yan, J., Zhang, Z., Zhu, J., Zhang, K., Pei, J., & Liu, Q. (2024). DeltaDock: A unified framework for accurate, efficient, and physically reliable molecular docking. *Advances in Neural Information Processing Systems*, 38, 113282–113310. <https://openreview.net/forum?id=dao67XTSPd>
- Yang, J., Roy, A., & Zhang, Y. (2012). BioLiP: A semi-manually curated database for biologically relevant ligand–protein interactions. *Nucleic Acids Research*, 41(D1), D1096–D1103. <https://doi.org/10.1093/nar/gks966>
- Yang, X., Liu, Y., Gan, J., Xiao, Z.-X., & Cao, Y. (2022). FitDock: Protein–ligand docking by template fitting. *Briefings in Bioinformatics*, 23(3), bbac087. <https://doi.org/10.1093/bib/bbac087>
- Ye, H., Lin, H., Han, J., Xu, M., Liu, S., Liang, Y., Ma, J., Zou, J., & Ermon, S. (2024). TFG: Unified training-free guidance for diffusion models. *Advances in Neural Information Processing Systems*, 38. <http://arxiv.org/abs/2409.15761>
- Young, D. C. (2009). *Computational drug design: A guide for computational and medicinal chemists*. Wiley.
- Zhang, W., Bell, E. W., Yin, M., & Zhang, Y. (2020). EDock: Blind protein–ligand docking by replica-exchange monte carlo simulation. *Journal of Cheminformatics*, 12(1), 37. <https://doi.org/10.1186/s13321-020-00440-9>
- Zhang, X., Zhang, O., Shen, C., Qu, W., Chen, S., Cao, H., Kang, Y., Wang, Z., Wang, E., Zhang, J., Deng, Y., Liu, F., Wang, T., Du, H., Wang, L., Pan, P., Chen, G., Hsieh, C.-Y., & Hou, T. (2023). Efficient and accurate large library ligand docking with KarmaDock. *Nature Computational Science*, 3(9), 789–804. <https://doi.org/10.1038/s43588-023-00511-5>
- Zhao, H., & Caflisch, A. (2013). Discovery of ZAP70 inhibitors by high-throughput docking into a conformation of its kinase domain generated by molecular dynamics. *Bioorganic & Medicinal Chemistry Letters*, 23(20), 5721–5726. <https://doi.org/10.1016/j.bmcl.2013.08.009>

Zhou, G., Gao, Z., Ding, Q., Zheng, H., Xu, H., Wei, Z., Zhang, L., & Ke, G. (2023). Uni-Mol: A universal 3D molecular representation learning framework. *The Eleventh International Conference on Learning Representations*. <https://openreview.net/forum?id=6K2RM6wVqKu>

Appendix A

Technical background

A.1 Cheminformatics

A.1.1 String representations of molecules

Molecules are often represented as graphs, but can be encoded as linear strings. The Simplified Molecular Input Line Entry System (SMILES) is a widely adopted standard that represents atoms sequentially, using parentheses to denote branching and numeric labels to specify ring closures (Weininger, 1988).

A more expressive line notation is the International Chemical Identifier (InChI), an international standard that encodes molecular structures in a layered, hierarchical manner. The representation begins with the molecular formula and progressively incorporates connectivity, hydrogen atom placement, formal charges, and protonation states. The InChI string is a human-readable identifier, while the InChIKey is a fixed-length hashed derivative designed for efficient indexing and searching in databases (Heller et al., 2015).

A.1.2 Molecular fingerprints

Extended Connectivity Fingerprints (ECFP) represent molecules as a fixed-length binary vector that captures the local atomic environments. They are constructed starting from each atom in a molecular graph, iteratively updating the representation of each atom using the next set of neighbours and a hash function (Rogers & Hahn, 2010). Typically, two iterations are performed, so that every atom contains information of all atoms up to two bonds away, representing an atomic environments of diameter 4; the resultant fingerprint is called ECFP4.

Other molecular fingerprints such as MACCS (J. L. Durant et al., 2002) exist, but these were highly correlated with ECFP4 (Riniker & Landrum, 2013) and so I will only use ECFP4.

Fingerprints are usually compared using Tanimoto (Jaccard) similarity. Let two molecules be represented by binary fingerprint vectors $\mathbf{a}, \mathbf{b} \in \{0, 1\}^n$ where 1 and 0 indicate the presence and absence of a characteristic, respectively, and n is the length of the fingerprint. The Tanimoto similarity is

$$\text{Tanimoto}(\mathbf{a}, \mathbf{b}) = \frac{|\mathbf{a} \cap \mathbf{b}|}{|\mathbf{a}| + |\mathbf{b}| - |\mathbf{a} \cap \mathbf{b}|}$$

where $|\mathbf{a}| = \sum_{i=1}^n a_i$ is the number of features (bit count), $|\mathbf{a} \cap \mathbf{b}| = \sum_{i=1}^n (a_i \wedge b_i)$ is the number of shared features (intersection), and $|\mathbf{a} \cup \mathbf{b}| = \sum_{i=1}^n (a_i \vee b_i)$ is the number of total features (union).

A.1.3 Root mean square deviation

The root mean square deviation (RMSD) is a commonly used distance metric for molecular poses, especially in docking. Let $\mathbf{A}$ and $\mathbf{B}$ be the $n \times 3$ matrices of the atomic coordinates of two conformers

of the same molecule with n atoms. The RMSD between the two conformers is

$$\text{RMSD}(\mathbf{A}, \mathbf{B}) = \sqrt{\frac{1}{n} \sum_{i=1}^n \sum_{j=1}^3 \|\mathbf{A}_{i,j} - \mathbf{B}_{i,j}\|^2}$$

RMSD values are expressed in length units; in structural biology, the unit Angstrom ($\text{\AA}$) is used and $1 \text{ \AA}$ equals 10^{-10} m . For ligands with symmetries, the RMSD is reported using the injective map between equivalent atoms that produces the lowest value (Meli & Biggin, 2020). For a given atom mapping, the Kabsch algorithm (Kabsch, 1976) can be used to find the rigid transformation that minimises the RMSD between two sets of points.

A.1.4 Resolution of molecular structures

The quality of the molecular models in the PDB and CSD varies. This is because molecular structures are modelled on experimental data; in X-ray crystallography, for example, atomic models are fitted to electron-density maps derived from diffraction patterns produced by a crystallised molecule (Drenth, 2007).

The resolution describes the smallest experimentally distinguishable distance and reflects the precision with which atomic positions are defined. High-quality crystals, composed of uniformly aligned molecules, produce detailed diffraction patterns and high resolution, whereas crystals exhibiting mosaicity, multiple orientations, or anisotropic regions yield poorer diffraction data and lower resolution, complicating model building and refinement. High-resolution structures (up to around $1.5 \text{ \AA}$) resolve individual atoms clearly, whereas lower-resolution structures (from around $3 \text{ \AA}$) reveal only basic contours, requiring atomic positions to be inferred (Kleywegt & Jones, 1997).

A.2 Deep learning

This section provides a brief overview of the models referenced in this thesis. The textbooks of Goodfellow et al. (2016) and Bishop and Bishop (2024) serve as primary references for the material presented. The notation adopted is that of the latter.

A.2.1 Supervised machine learning

In supervised machine learning, one has a finite set of n pairs of data $\mathbb{D} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^n$ where $\mathbf{x}_i \in \mathcal{X}$ are elements of some input space and $\mathbf{y}_i \in \mathcal{Y}$ are elements of some output space. The goal is to learn a function f that maps from the input to the output space:

$$f_{\theta}: \mathcal{X} \rightarrow \mathcal{Y}$$

We write f_{θ} with θ to remind us that f is a function with parameters θ that must be learnt from the data $\mathbb{D}$.

Ideally, the expected value of a task-specific loss function L would be minimised over the true probability distribution of the data. However, typically only samples $\mathbb{D}$ from the data distribution are available, and the empirical risk is minimised instead, giving equal weight to all data points:

$$\mathbb{E}_{\mathbf{x}, \mathbf{y} \sim \mathbb{D}} L(f_{\theta}(\mathbf{x}), \mathbf{y}) = \frac{1}{n} \sum_{i=1}^n L(f_{\theta}(\mathbf{x}_i), \mathbf{y}_i)$$

where $\mathbb{D} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^n$ is the training data set, f_{θ} is the model parametrised by θ , and L is the loss function that quantifies the prediction error (Goodfellow et al., 2016, cf. Equation 8.3).

A.2.2 Neural networks

Neural networks are a general class of models composed of layers of interconnected units, called neurones, that apply learnt affine transformations followed by non-linear activation functions. Stacking

many such layers, neural networks can represent complex non-linear mappings. A standard feed-forward network with K layers computes for each $k = 1, \dots, K$ layer the representation $\mathbf{z}^{(k)}$ following

$$\mathbf{z}^{(k)} = h^{(k)}(\mathbf{W}^{(k)}\mathbf{z}^{(k-1)})$$

where $\mathbf{W}^{(k)}$ are the learnt weights and biases (parameters), and $h^{(k)}$ are the non-linear activation functions of layer k (Bishop & Bishop, 2024, cf. Equation 6.19). Another name for a feed-forward neural network is multilayer perceptron (MLP).

A.2.3 Convolutional networks

Convolutional neural networks process grid-like data using convolutional layers; these are sparsely connected and share weights. Each neurone is connected to a local set of neighbouring neurones in the previous layers, and all neurons within a layer share the same input weights. For a two-dimensional input, a single convolutional layer computes the activation values by

$$Z_{i,j}^{(k)} = h\left(\sum_m \sum_n Z_{m,n}^{(k-1)} \cdot \mathbf{W}_{i-m,j-n}^{(k)}\right)$$

where $Z^{(k-1)}$ denotes the two-dimensional input matrix, $\mathbf{W}$ is a $n \times m$ two-dimensional learnt convolutional kernel (parameters), h is the activation function, and $Z^{(k)}$ is the resulting output representation (Goodfellow et al., 2016, cf. Equation 9.4).

A.2.4 Graph neural networks

Graph neural networks operate on graph data where relationships between entities are encoded as edges E connecting nodes V . By iteratively aggregating and transforming information from the local neighbourhood of a node, graph neural networks learn representations that respect the underlying graph topology and are permutation invariant (Daigavane et al., 2021; Sanchez-Lengeling et al., 2021).

Consider a graph $G = (V, E)$ with node features $\mathbf{h}_i \in \mathbb{R}^d$ for each node $i \in V$. Graph neural networks can be viewed as aggregating messages from surrounding nodes

$$\mathbf{m}_i^{(k)} = \text{Aggregate}\left(\{\mathbf{h}_j^{(k)} : j \in \mathcal{N}_i\}\right)$$

and then updating the local node embeddings

$$\mathbf{h}_i^{(k+1)} = \text{Update}\left(\mathbf{h}_i^{(k)}, \mathbf{m}_i^{(k)}\right)$$

where $\mathcal{N}_i$ denotes the neighbourhood of node i , Aggregate is a permutation-invariant aggregation function like sum or mean, and Update is a learnt weighted linear combination with a non-linear activation (Bishop & Bishop, 2024, cf. Equations 13.10 and 13.11). Extensions that include edge and global features enable further applications.

A.2.5 Geometric graph neural networks

Geometric graph neural networks can incorporate inductive biases for symmetries such as rotation, translation, and reflection by choosing update and aggregation functions that are invariant and equivariant in these actions (Satorras et al., 2021).

Consider a graph $G = (V, E)$ with node features $\mathbf{h}_i \in \mathbb{R}^d$ and coordinates $\mathbf{x}_i \in \mathbb{R}^n$ for each node $i \in V$ and edge features $\mathbf{e}_{ij}$ for each edge $ij \in E$. The following aggregation and update functions proposed by

Satorras et al. (2021) maintain the aforementioned symmetries in addition to permutation invariance:

$$\begin{aligned} \mathbf{e}_{ij}^{(k+1)} &= \text{Update}_{\text{edge}} \left(\mathbf{e}_{ij}^{(k)}, \mathbf{h}_i^{(k)}, \mathbf{h}_j^{(k)}, \|\mathbf{x}_i^{(k)} - \mathbf{x}_j^{(k)}\|^2 \right) \\ \mathbf{x}_i^{(k+1)} &= \mathbf{x}_i^{(k)} + \sum_{j \in \mathcal{N}_i} (\mathbf{x}_i - \mathbf{x}_j) \phi \left(\mathbf{e}_{ij}^{(k+1)} \right) \\ \mathbf{m}_i^{(k+1)} &= \text{Aggregate}_{\text{node}} \left(\{ \mathbf{e}_{ij}^{(k+1)} : j \in \mathcal{N}_i \} \right) \\ \mathbf{h}_i^{(k+1)} &= \text{Update}_{\text{node}} \left(\mathbf{h}_i^{(k)}, \mathbf{m}_i^{(k+1)} \right) \end{aligned}$$

where $\mathcal{N}_i$ denotes the neighbourhood of node i and ϕ is a learnt scalar function (Bishop & Bishop, 2024, cf. Equations 13.18 through 13.41). Note that the coordinate updates only combine learnt weighted sums of the neighbouring coordinators ensuring their equivariance to translations and orthogonal transformations.

A.2.6 Transformers

Transformers are neural network architectures built around attention mechanisms that learn to assign contextual weights to input representations. Transformers can model long-range dependencies and complex interactions within sequences. The core computational component is scaled dot-product self-attention:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax} \left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}} \right) \mathbf{V},$$

where queries $\mathbf{Q}$, keys $\mathbf{K}$, and values $\mathbf{V}$ are linear projections learnt from input representations, and d denotes the key dimensionality (Vaswani et al., 2017, cf. Equation 1). In multi-head attention, multiple sets of $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ are learnt in parallel, enabling the model to capture different types of relationships.

In combination with positional encodings, the attention mechanism allows each element of a sequence to selectively focus on other elements when forming its representation, enabling context-dependent processing of information. Transformers scale efficiently with parallel computation, can model complex interactions, and have achieved strong performance across language, vision, and multimodal tasks (Bishop & Bishop, 2024, ch. 12).

Geometric transformers extend the transformer architecture to data embedded in geometric spaces, such as Euclidean space, by incorporating inductive biases for rotational, translational, and reflective symmetries directly into the attention mechanism. By designing attention operations that are invariant or equivariant to these transformations, geometric transformers enable the principled modelling of structured data arising in molecular systems, robotics, and 3D perception (Fuchs et al., 2020).

A.2.7 Diffusion models

Diffusion models are generative models that learn to sample from data distributions by learning to reverse a noising process. The model trains to undo a forward diffusion process that incrementally corrupts data with noise and the model learns the reverse process that de-noises the data step by step (Bishop & Bishop, 2024, ch. 20).

The forward process typically uses Gaussian noise:

$$q(\mathbf{z}_t | \mathbf{z}_{t-1}) = \mathcal{N}(\mathbf{z}_t; \sqrt{1 - \beta_t} \mathbf{z}_{t-1}, \beta_t \mathbf{I}),$$

and a neural network μ_θ is trained to reverse the process:

$$p(\mathbf{z}_{t-1} | \mathbf{z}_t) = \mathcal{N}(\mathbf{z}_{t-1}; \mu_\theta(\mathbf{z}_t, t), \beta_t \mathbf{I}),$$

where β_t is a user defined noise schedule and μ_θ is a neural network parametrised by θ (Bishop & Bishop, 2024, cf. Equations 20.4 and 20.18). Inference is done by starting from Gaussian noise and sampling sequentially from the conditional distributions $p(\mathbf{z}_{t-1} | \mathbf{z}_t)$.

Appendix B

Detailed methods and further results for Chapter 2

B.1 Software and data availability

PoseBusters is made available as a Python package on the Python Package Index (PyPI) (<https://pypi.org/project/posebusters>) and as open source code under the MIT license on Github (github.com/maabuu/posebusters). Data for this paper, including the Astex Diverse set and PoseBusters Benchmark set, as well as the individual tabulated test results for each docking are available at Zenodo at <https://zenodo.org/rteccords/8278563>.

B.2 Docking protocols

The following protocols detail how the seven docking methods were used to re-dock the ligands into the crystal structures of the Astex Diverse set and the PoseBusters Benchmark set. Methods that require an initial ligand conformation were given identical starting conformations generated with the RDKit's ETKDGv3 conformer generator (S. Wang et al., 2020) followed by an energy minimisation using the universal force field (UFF) (Rappe et al., 1992; Tosco et al., 2014). All docking protocols were given receptors prepared without waters because none of the DL-based methods supports docking with waters.

AutoDock Vina

Software version Vina 1.2.3, Meeko 0.4.0, Reduce 4.9.210817, ADFRsuite 1.0, RDKit 2022.09.1

Ligand preparation The initial ligand conformations described above were prepared with Meeko using standard settings.

Protein preparation Hydrogen atoms were added with reduce and then the PDBQT files were generated with the ADFR prepare_receptor script.

Parameters A bounding box with side-length 25 Å was created around the centroid of the crystal ligand. Vina was used to create 40 poses with an exhaustiveness setting of 32 and the top-ranked pose was selected.

CCDC GOLD

Software version CCDC Python API version 3.0.14

Ligand preparation The initial ligand conformations described above were prepared with LigandPreparation using the default settings which include adding missing hydrogens, removing unknown atoms, and rule-based protonation of the ligand.

Protein preparation The protein and co-factors were loaded from separate files and all hydrogens were added.

Parameters A settings file was created for each complex using the Docker class default settings. The binding site was defined around the crystal ligand centroid using `BindingSiteFromPoint` with radius 25 Å. The settings used are rescore function `plp`, autoscale 100 %, and early termination off. After generating 40 poses only the top-ranked pose was saved.

DeepDock

Software version DeepDock commit hash 54a2a64 from authors' public code repository <https://github.com/OptiMaL-PSE-Lab/DeepDock>, MSMS 2.6.1, PDB2PQR 2.1.1, APBS 3.4.1

Ligand preparation The generated starting ligand conformations were used without further processing.

Protein preparation The steps in example notebook `Docking_example.ipynb` were used to generate protein surface meshes. The function `compute_inp_surface` generated binding site surfaces using the crystal ligands and the crystal protein structures with a distance threshold of 10 Å.

Parameters The protocol and settings in notebook `Docking_example.ipynb` in the DeepDock repository were used for docking.

DiffDock

Software version DiffDock commit hash fff8f0b from authors' public code repository <https://github.com/gcorso/DiffDock>

Ligand preparation The generated starting ligand conformations were used without further processing.

Protein preparation ESM was used to generate FASTA files.

Parameters The protocol in `README.md` was used to generate ESM embeddings and then to do inference. 40 poses were sampled using 20 inference steps with no noise on the final step. The top-ranked pose was selected.

EquiBind

Software version EquiBind commit hash 41bd00f from authors' public code repository <https://github.com/HannesStark/EquiBind>, Reduce 3.3.160602, Open Babel 3.1.0, RDKit 2022.09.1

Ligand preparation The generated starting ligand conformations were processed with Open Babel and then with the RDKit to add missing hydrogens.

Protein preparation The receptors were processed with Open Babel. Then reduce was used to correct receptor residues and to add hydrogens. Then the protein chains which have at least one residue within 10 Å of the crystal ligand were selected.

Parameters The configuration file `configs_clean/inference.yml` in the repository was used.

TankBind

Software version TANKBind commit hash 804e9fc from authors' public code repository <https://github.com/luwei0917/TankBind>, p2rank 2.3

Ligand preparation The notebook `prediction_example_using_PDB_6hd6.ipynb` was used to renumber the ligand atoms and generate features from the ligands.

Protein preparation The notebook `prediction_example_using_PDB_6hd6.ipynb` was used to generate features from the crystal protein structures.

Parameters The steps in the notebook `prediction_example_using_PDB_6hd6.ipynb` were used for inference. The steps were running p2rank to generate a list of binding pockets and then docking using the TankBind model.

Uni-Mol

Software version Uni-Mol commit hash b962451 from authors’ public code repository <https://github.com/dptech-corp/Uni-Mol>

Ligand preparation The ligands were generated according to the protocol described in the README.md file in the top folder of the Uni-Mol repository.

Protein preparation The binding pockets residues were those within 8 Å of any crystal ligand heavy atom.

Parameters The default arguments (recycling=3, batch_size=8, dist_threshold=8.0) were used.

B.3 PoseBusters Benchmark set procurement

Table B.1: Selection process of the PDB entries and ligands for the PoseBusters Benchmark set. The filters are based on the PDB meta data, the PDB quality reports, and the PDB structure data. The final PoseBusters Benchmark set consists of 308 unique PDB entries containing 308 unique ligands.

Selection step	Unique PDB IDs	Unique CCD IDs
PDB entries with a protein and ‘ligand of interest’ released from 1 January 2021 to 30 May 2023	10 537	6635
Ligands weighing from 100 Da to 900 Da	10 537	6424
Ligands with at least 3 heavy atoms	10 537	6374
Ligands containing only H, C, O, N, P, S, F, Cl atoms	10 537	6271
Ligands that are not covalently bound to protein	7247	4891
Structures with no unknown atoms (<i>e.g.</i> element X)	7218	4881
X-ray structure high resolution limit at most 2 Å	4686	3314
Ligand real space R-factor is at most 0.2	3800	2572
Ligand real space correlation coefficient is at least 0.95	1849	1054
Ligand model completeness is 100%	1820	1039
Ligand starting conformation could be generated with ETK-DGv3 (Riniker & Landrum, 2015)	1733	1019
All ligand SDF files can be loaded with RDKit (Landrum et al., 2025) and pass its sanitization	1706	994
PDB ligand report reports no stereochemical errors	1706	994
PDB ligand report reports no atomic clashes	1256	844
Select single protein-ligand conformation ¹	1256	844
Intermolecular distance between the ligand(s) of interest and the protein is at least 0.2 Å	1237	834
Intermolecular distance between ligand(s) of interest and other small organic molecules is at least 0.2 Å	1237	834
Intermolecular distance between the ligand(s) of interest and ion metals in complex is at least 0.2 Å	1232	832
Blocklist for PDB entries ²	1227	827
Blocklist for CCD entries ³	1223	823

Continued on next page

Table B.1: Selection process of the PDB entries and ligands for the PoseBusters Benchmark set. The filters are based on the PDB meta data, the PDB quality reports, and the PDB structure data. The final PoseBusters Benchmark set consists of 308 unique PDB entries containing 308 unique ligands. (Continued)

Randomly select PDB entries to get unique ligands	809	823
Randomly select ligands to get unique PDB entries	809	809
Select representative PDB entries of clustered protein sequences ⁴	428	428
Remove ligands which are within 5.0 Å of any protein symmetry mate	308	308

¹ The first conformation containing the ligand of interest was chosen when multiple conformations containing the ligand were available in the PDB entry.

² The blacklist for the PDB entries (by PDB identifier) contains entries removed due to bad ligand conformations (7X48, 7UYC), ligands forming polymers (7WJD, 7DB4), racemic mixtures of ligands where the stereoisomer has a different CCD identifier (6ZYU, 7W2W), and structures containing elements Te and Yb which AutoDock Vina does not support by default (7ZSQ, 8AVA).

³ The blacklist for the ligands (by CCD identifier) contains the four entries I8P, 5A3, U71, and UEV. These four are omitted because they are highly symmetric and the substructure search yields many possible atom-atom mappings between conformations negatively affecting the RMSD calculation time.

⁴ Clustering with Diamond (Buchfink et al., 2021) is done with an identity cutoff for the clustering of 0% and a minimum coverage of the cluster member sequences by the representative sequences of 100% and otherwise default values which includes the BLOSUM62 substitution matrix.

B.4 Comparison and description of the data sets

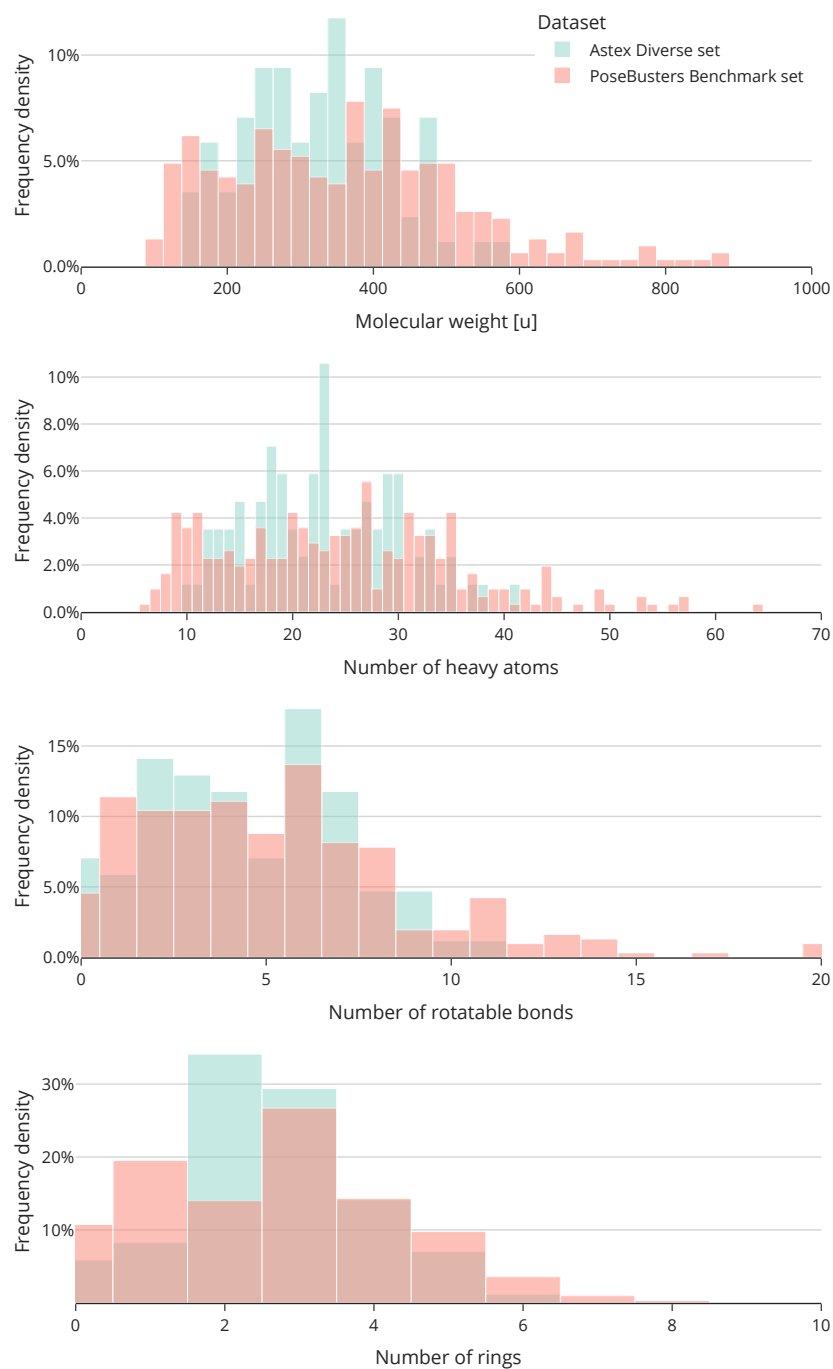


Figure B.1: Comparison of the 85 ligands in the Astex Diverse set and the 308 ligands in the PoseBusters Benchmark set in terms of molecular weight, number of heavy atoms, number of rotatable bonds, and number of rings.

B.5 Cofactor prevalence

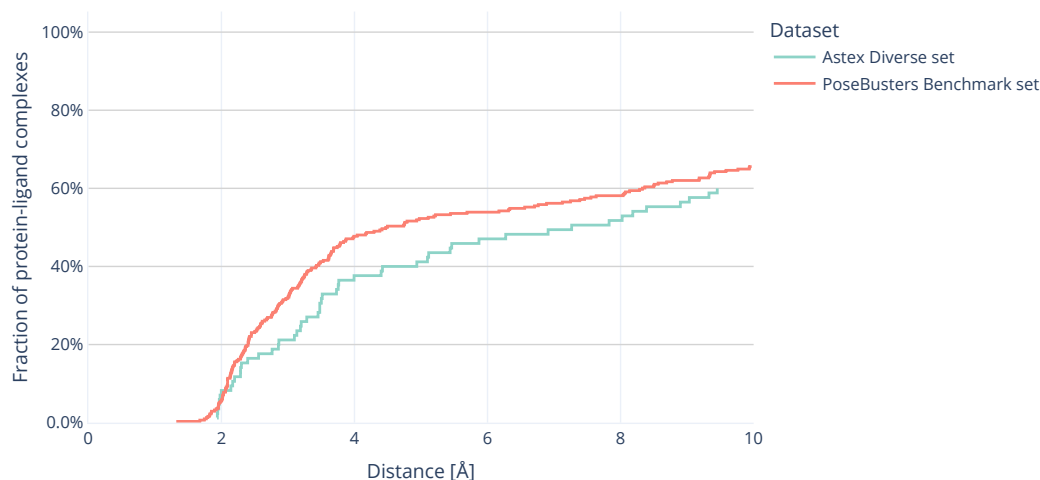


Figure B.2: Fraction of protein-ligand complexes with a cofactor close to the ligand as a function of the distance threshold used. Here a cofactor is any other compound present in the crystal structure besides the ligand of interest, the protein and solvent. This includes metal ions, iron-sulfur clusters, and small organic molecules. 46% of the protein-ligand complexes in the PoseBusters Benchmark set have a cofactor within 4 Å of the ligand.

B.6 PDB and CCD identifiers of data sets

The following sections list the protein database (Berman et al., 2000) (PDB) codes and chemical component dictionary (Westbrook et al., 2015) (CCD) codes for the protein-ligand complexes and the corresponding ligands of interest for the two data sets used.

Astex Diverse set

1G9V RQ3, 1GKC NFH, 1GM8 SOX, 1GPK HUP, 1HNN SKF, 1HP0 AD3, 1HQ2 PH2, 1HVY D16, 1HWI 115, 1HWW SWA, 1IA1 TQ3, 1IG3 VIB, 1J3J CP6, 1JD0 AZM, 1JJE BYS, 1JLA TNK, 1K3U IAD, 1KE5 LS1, 1KZK JE2, 1L2S STC, 1L7F BCZ, 1LPZ CMB, 1LRH NLA, 1M2Z DEX, 1MEH MOA, 1MMV 3AR, 1MZC BNE, 1N1M A3M, 1N2J PAF, 1N2V BDI, 1N46 PFA, 1NAV IH5, 1OF1 SCT, 1OF6 DTY, 1OPK P16, 1OQ5 CEL, 1OWE 675, 1OYT FSN, 1P2Y NCT, 1P62 GEO, 1PMN 984, 1Q1G MTI, 1Q41 IXM, 1Q4G BFL, 1R1H BIR, 1R55 097, 1R58 A05, 1R90 FLP, 1S19 MC9, 1S3V TQD, 1SG0 STL, 1SJ0 E4D, 1SQ5 PAU, 1SQN NDR, 1T40 ID5, 1T46 STI, 1T9B 1CS, 1TOW CRZ, 1TT1 KAI, 1TZ8 DES, 1U1C BAU, 1U4D DBQ, 1UML FR4, 1UNL RRC, 1UOU CMU, 1V0P PVB, 1V48 HA1, 1V4S MRK, 1VCJ IBA, 1W1P GIO, 1W2G THM, 1X8X TYR, 1XM6 5RM, 1XOQ ROF, 1XOZ CIA, 1Y6B AAX, 1YGC 905, 1YQY 915, 1YV3 BIT, 1YVF PH7, 1YWR LI9, 1Z95 198, 2BM2 PM2, 2BR1 PFP, 2BSM BSM

PoseBusters Benchmark set

5SAK ZRY, 5SB2 1K2, 5SD5 HWI, 5SIS JSM, 6M2B EZO, 6M73 FNR, 6T88 MWQ, 6TW5 9M2, 6TW7 NZB, 6VTA AKN, 6WTN RXT, 6XBO 5MC, 6XCT 478, 6XG5 TOP, 6XHT V2V, 6XM9 V55, 6YJA 2BA, 6YMS OZH, 6YQV 8K2, 6YQW 82I, 6YR2 T1C, 6YRV PJ8, 6YSP PAL, 6YT6 PKE, 6YYO Q1K, 6Z0R Q4H, 6Z14 Q4Z, 6Z1C 7EY, 6Z2C Q5E, 6Z4N Q7B, 6ZAE ACV, 6ZC3 JOR, 6ZCY QF8, 6ZK5 IMH, 6ZPB 3D1, 7A1P QW2, 7A9E R4W, 7A9H TPP, 7AFX R9K, 7AKL RK5, 7AN5 RDH, 7B2C TP7, 7B94 ANP, 7BCP GC0, 7BJJ TVW, 7BKA 4JC, 7BMI U4B, 7BNH BEZ, 7BTT F8R, 7C0U FGO, 7C3U AZG, 7C8Q DSG, 7CD9 FVR, 7CIJ G0C, 7CL8 TES, 7CNQ G8X, 7CNS PMV, 7CTM BDP, 7CU0 PHB, 7D5C GV6, 7D60 MTE, 7DKT GLF, 7DQL 4CL, 7DUA HJ0, 7E4L MDN, 7EBG J0L, 7ECR SIN, 7ED2 A3P, 7ELT TYM, 7EPV FDA, 7ES1 UDP, 7F51 BA7, 7F5D EU0, 7F8T FAD, 7FB7 8NF, 7FHA ADX, 7FRX O88, 7FT9 4MB, 7JG0 GAR, 7JHQ VAJ, 7JMV 4NC,

7JXX VP7, 7JY3 VUD, 7K0V VQP, 7KB1 WBJ, 7KC5 BJZ, 7KM8 WPD, 7KQU YOF, 7KRU ATP, 7KZ9 XN7,
 7L00 XCJ, 7L03 F9F, 7L5F XNG, 7L7C XQ1, 7LCU XTA, 7LEV 0JO, 7LJN GTP, 7LMO NYO, 7LOE Y84,
 7LOU IFM, 7LT0 ONJ, 7LZD YHY, 7M31 TDR, 7M3H YPV, 7M6K YRJ, 7MFP Z7P, 7MGT ZD4, 7MGY ZD1,
 7MMH ZJY, 7MOI HPS, 7MSR DCA, 7MWN WI5, 7MWU ZPM, 7MY1 IPE, 7MYU ZR7, 7N03 ZRP, 7N4N 0BK,
 7N4W P4V, 7N6F 0I1, 7N7B T3F, 7N7H CTP, 7NF0 BYN, 7NF3 4LU, 7NFB GEN, 7NGW UAW, 7NLV UJE,
 7NP6 UK8, 7NPL UKZ, 7NR8 UOE, 7NSW HC4, 7NU0 DCL, 7NUT GLP, 7NX0 UU8, 7O0N CDP, 7O1T 5X8,
 7ODY DGI, 7OEO V9Z, 7OFF VCB, 7OFK VCH, 7OLI 8HG, 7OMX CNA, 7OP9 06K, 7OPG 06N, 7OSO 0V1,
 7OZ9 NGK, 7OZC G6S, 7P1F KFN, 7P1M 4IU, 7P2I MFU, 7P4C 5OV, 7P5T 5YG, 7PGX FMN, 7PIH 7QW,
 7PJQ OWH, 7PK0 BYC, 7PL1 SFG, 7POM 7VZ, 7PRI 7TI, 7PRM 81I, 7PT3 3KK, 7PUV 84Z, 7Q25 8J9,
 7Q27 8KC, 7Q2B M6H, 7Q5I I0F, 7QE4 NGA, 7QF4 RBF, 7QFM AY3, 7QGP DJ8, 7QHG T3B, 7QHL D5P,
 7QPP VDX, 7QTA URI, 7R3D APR, 7R59 I5F, 7R6J 2I7, 7R7R AWJ, 7R9N F97, 7RC3 SAH, 7RH3 590,
 7RKW 5TV, 7RNI 60I, 7ROR 69X, 7ROU 66I, 7RSV 7IQ, 7RWS 4UR, 7RZL NPO, 7SCW GSP, 7SDD 4IP,
 7SFO 98L, 7SIU 9ID, 7SUC COM, 7SZA DUI, 7T0D FPP, 7T1D E7K, 7T3E SLB, 7TB0 UD1, 7TBU S3P,
 7TE8 P0T, 7TH4 FF0, 7THI PGA, 7TM6 GPJ, 7TOM 5AD, 7TS6 KMI, 7TSF H4B, 7TU0 KL9, 7TXK LW8,
 7TYP KUR, 7U0U FK5, 7U3J L6U, 7UAS MBU, 7UAW MF6, 7UJ4 OQ4, 7UJ5 DGL, 7UJF R3V, 7ULC 56B,
 7UMW NAD, 7UQ3 02U, 7USH 82V, 7UTW NAI, 7UXS OJC, 7UY4 SMI, 7UYB OK0, 7V14 ORU, 7V3N AKG,
 7V3S 5I9, 7V43 C40, 7VB8 STL, 7VBU 6I4, 7VC5 9SF, 7VKZ NOJ, 7VQ9 ISY, 7VWF K55, 7VYJ CA0,
 7W05 GMP, 7W06 ITN, 7WCF ACP, 7WDT NGS, 7WJB BGC, 7WKL CAQ, 7WL4 JFU, 7WPW F15, 7WQQ 5Z6,
 7WUX 60I, 7WUY 76N, 7WY1 D0L, 7X5N 5M5, 7X9K 8OG, 7XBV APC, 7XFA D9J, 7XG5 PLP, 7XI7 4RI,
 7XJN NSD, 7XPO UPG, 7XQZ FPF, 7XRL FWK, 7YZU D07, 7Z1Q NIO, 7Z20 IAJ, 7Z7F IF3, 7ZCC OGA,
 7ZDY 6MJ, 7ZF0 DHR, 7ZHP IQY, 7ZL5 IWE, 7ZOC T8E, 7ZTL BCN, 7ZU2 DHT, 7ZXV 45D, 7ZZW KKW,
 8A1H DLZ, 8A2D KXY, 8AAU LH0, 8AEM LVF, 8AIE M7L, 8AP0 PRP, 8AQL PLG, 8AUH L9I, 8AY3 OE3,
 8B8H OJQ, 8BOM QU6, 8BTI RFO, 8C3N ADP, 8C5M MTA, 8CNH V6U, 8CSD C5P, 8D19 GSH, 8D39 QDB,
 8D5D 5DK, 8DHG T78, 8DKO TFB, 8DP2 UMA, 8DSC NCA, 8EAB VN2, 8EX2 Q2Q, 8EXL 799, 8EYE X4I,
 8F4J PHO, 8F8E XJI, 8FAV 4Y5, 8FLV ZB9, 8F05 Y4U, 8G0V YHT, 8G6P API, 8GFD ZHR, 8HFN XGC,
 8H00 3ZI, 8SLG G5A

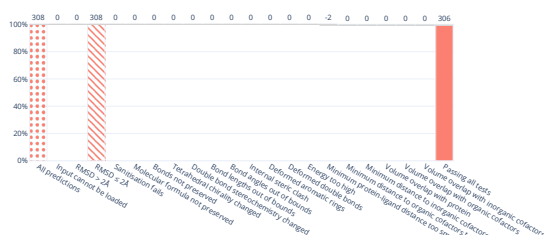
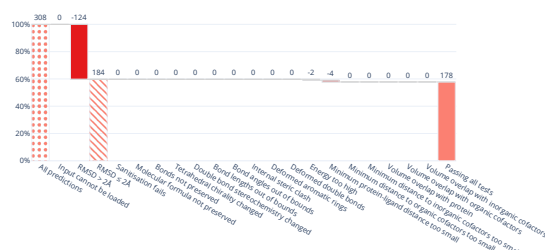
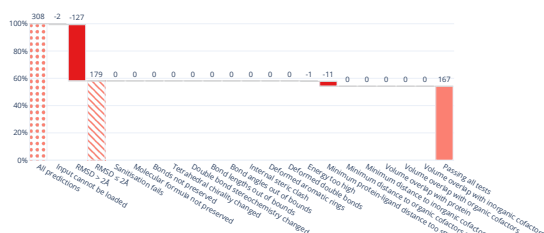
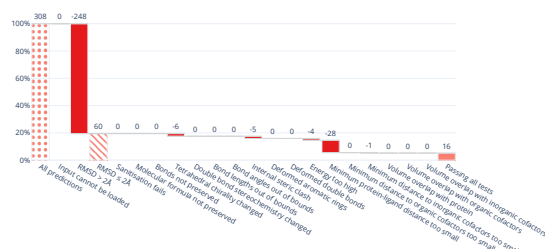
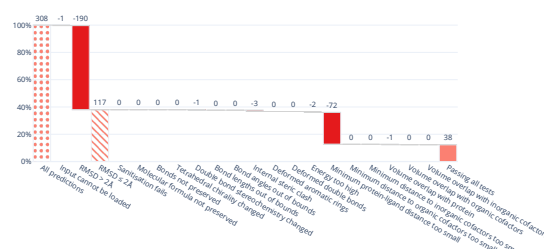
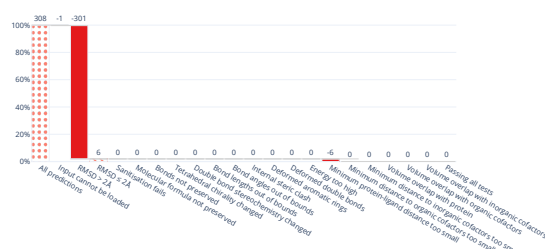
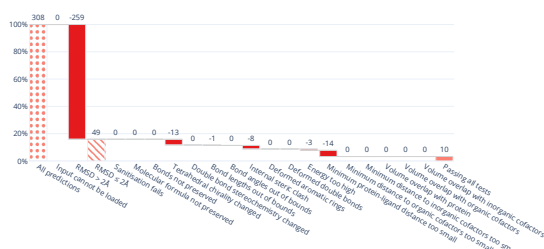
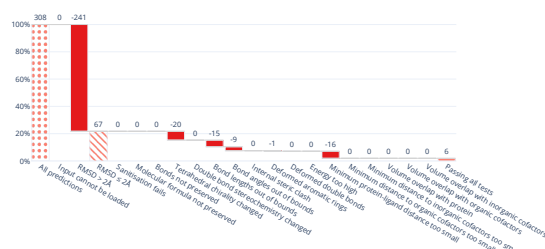
(a) Crystal Structures**(b) AutoDock Vina****(c) CCDC GOLD****(d) DeepDock****(e) DiffDock****(f) EquiBind****(g) TankBind****(h) Uni-Mol**

Figure B.4: Waterfall plots showing test results for the PoseBusters Benchmark set. The leftmost (dotted) bar shows the number of complexes in the test set. The red bars show the number of predictions that fail with each additional test going from left to right. The right most (solid) bar indicates the number of predictions that pass all tests. Refer to the main article for a description of each test. As a reading example, panel (a) shows that out of AutoDock Vina's 308 predictions, 200 are not within 2 Å RMSD, three clash with the protein and 1 clashes with an organic cofactor leaving 224 prediction with a low RMSD passing all tests. AutoDock Vina and CCDC GOLD pass the most tests.

B.8 Analysis of the severity of test failures

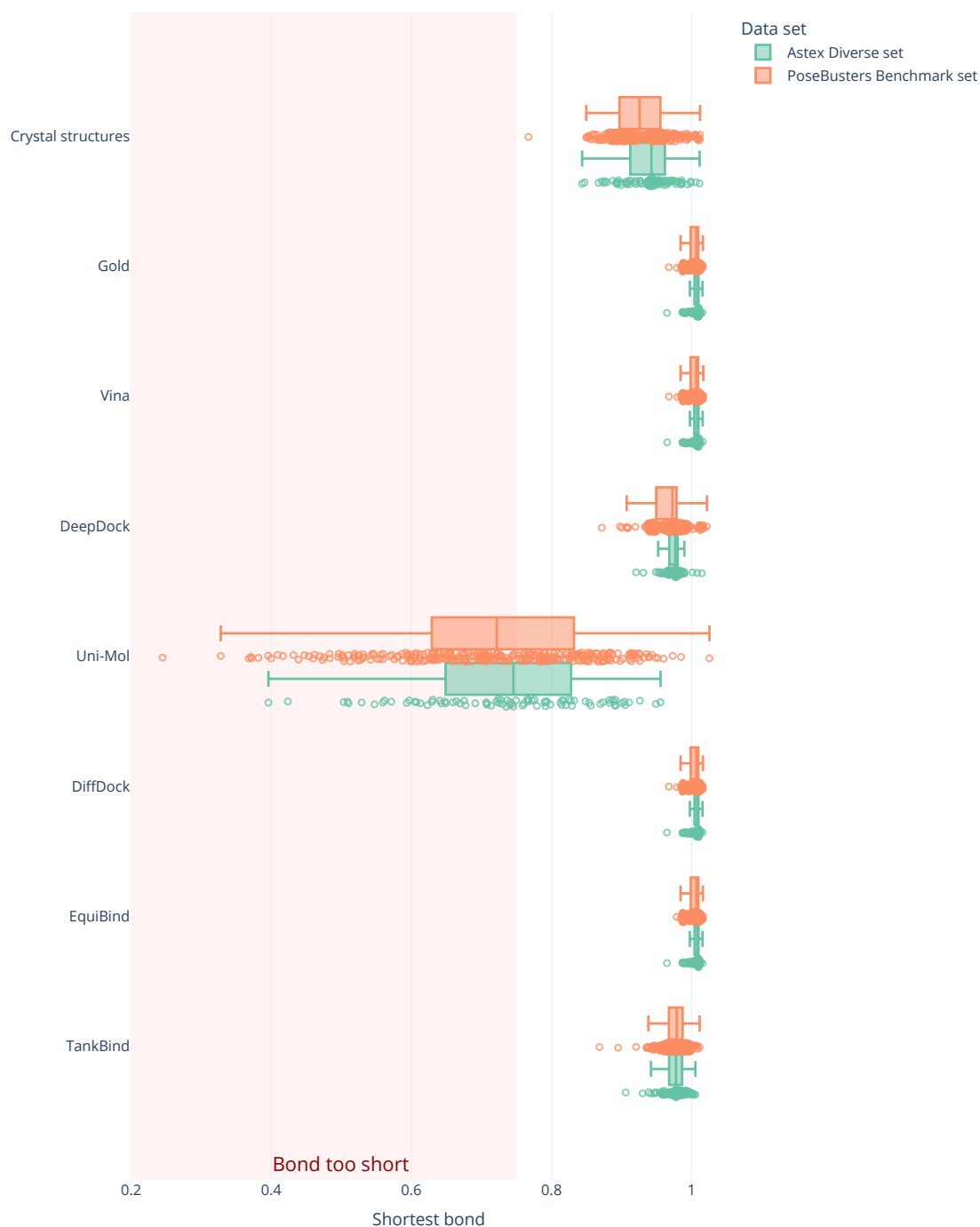


Figure B.5: Distribution of shortest bond lengths. Shown are the shortest bonds of each predicted ligand for each method and data set. The bond length is normalized by the lower bound for bond length obtained from Distance Geometry (DG). The lower bound correspond to one. A data point to the left of 0.75 indicates that the relatively shortest bond was more than 25% shorter than the DG lower bound. All methods except TankBind and Uni-Mol take the bond lengths from the provided ligand starting conformation. Uni-Mol and TankBind generate the bond lengths.



Figure B.6: Distribution of longest bond lengths. Shown are the longest bonds of each predicted ligand for each method and data set. The bond length is normalized by the upper bound for bond length obtained from Distance Geometry (DG). The upper bound correspond to one. A data point to the right of 1.25 indicates that the relatively shortest bond was more than 25 % longer than the DG upper bound. All methods except TankBind and Uni-Mol take the bond lengths from the provided ligand starting conformation. Uni-Mol and TankBind generate the bond lengths.



Figure B.7: Distribution of relative bond angles. Shown are the most extreme angles of each predicted ligand for each method and data set. Each bond angle is normalized by the corresponding bond length bounds obtained from Distance Geometry (DG). The upper bound corresponds to one. A data point to the right of 1.25 indicates that an angle is more than 25 % larger or shorter than the DG bounds permit. All methods except TankBind and Uni-Mol take the bond angles from the provided ligand starting conformation. Uni-Mol and TankBind generate the angles.

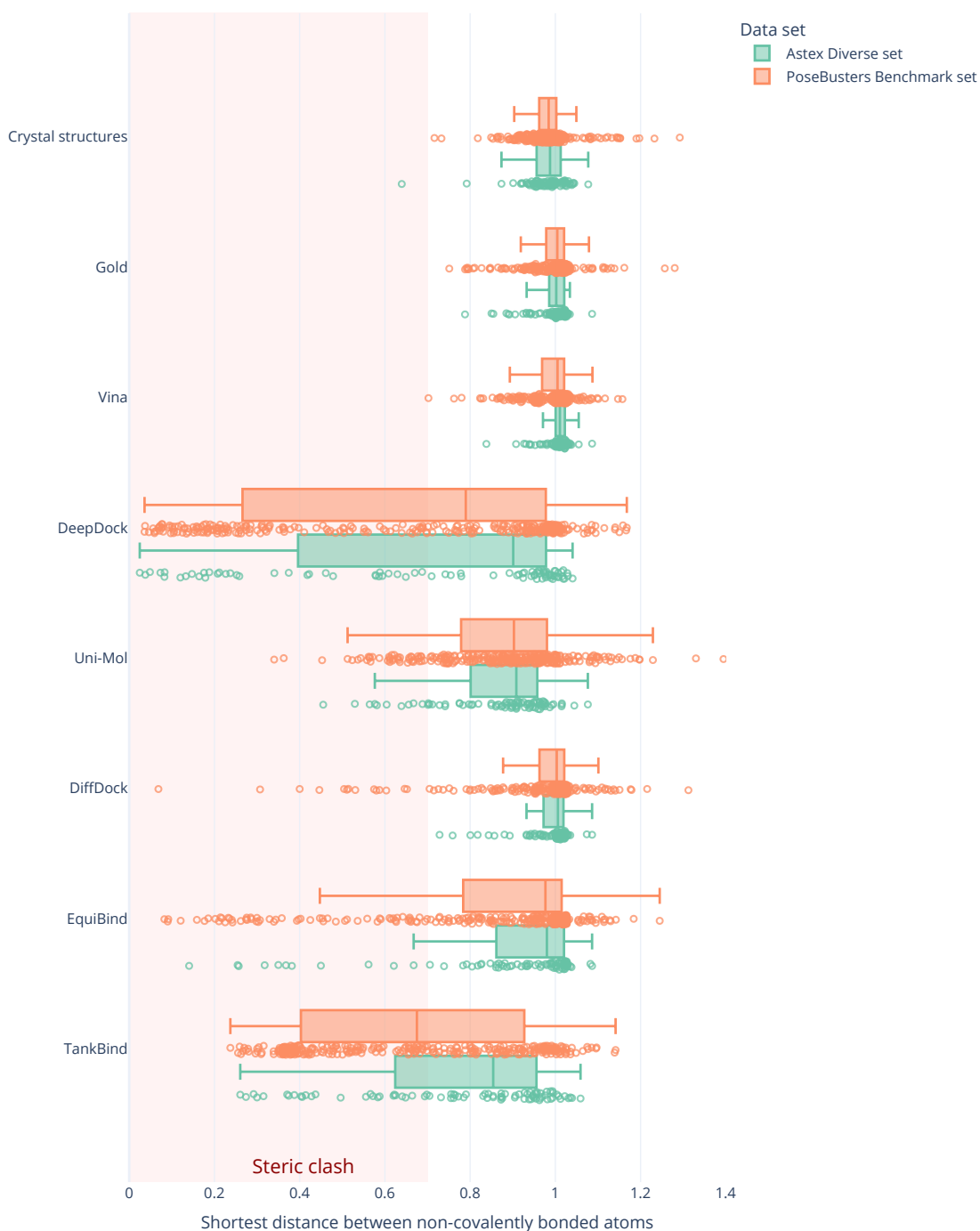


Figure B.8: Distribution of distances between unbounded atoms. The distribution shows the closest pair of unbounded atoms in each predicted ligand for each method and data set. Each distance is normalized to the Distance Geometry bounds. The lower bound corresponds to one. A data point to the left of 0.7 indicates that a distance was more than 30% shorter than the lower bound and was counted as a clash.

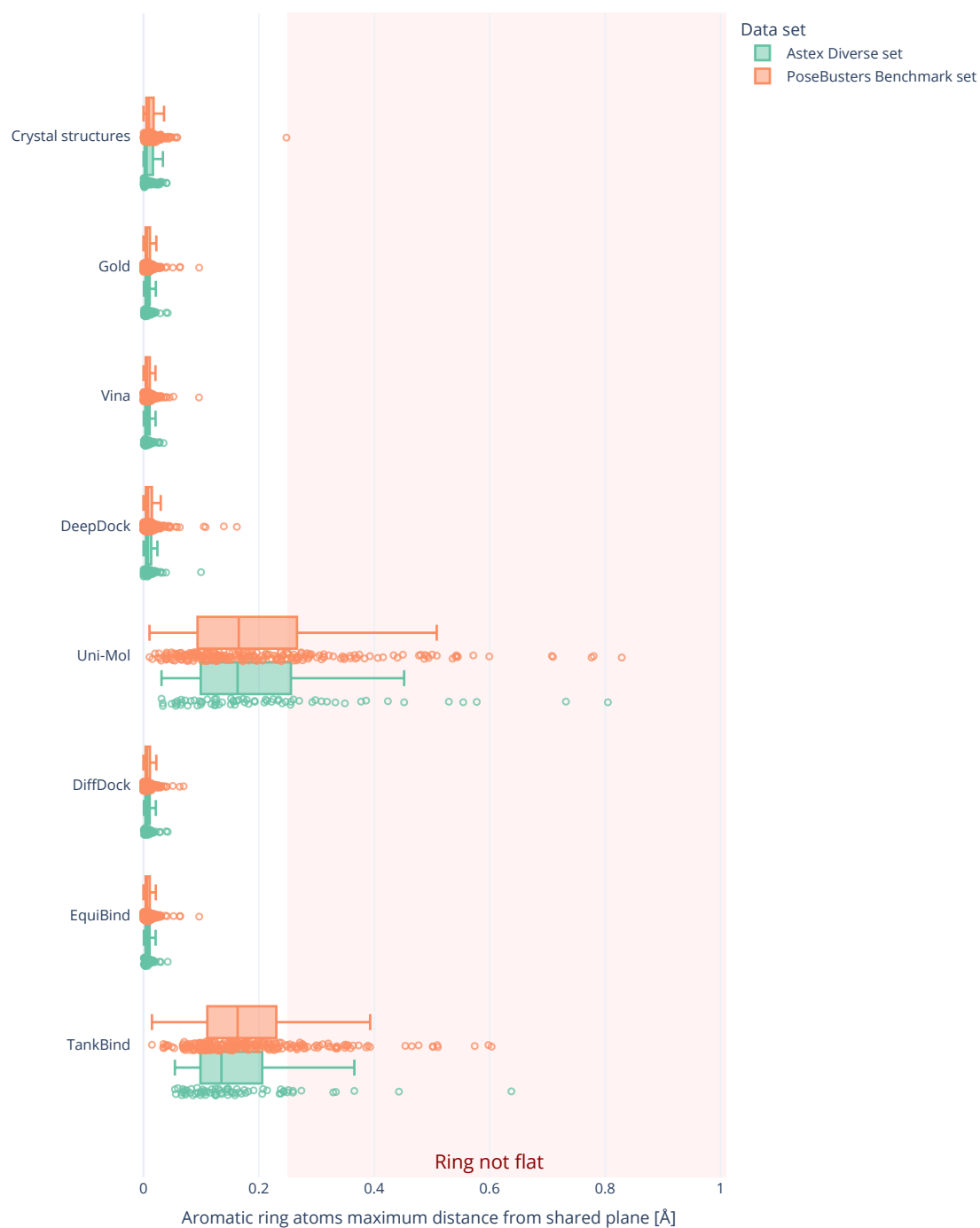


Figure B.9: Distance from shared plane of atoms in 5- or 6-membered aromatic rings. The largest distance in Angstrom from the shared plane is shown for each protein ligand complex. If a ligand has no rings it is not shown. TankBind and Uni-Mol generate non-flat rings.

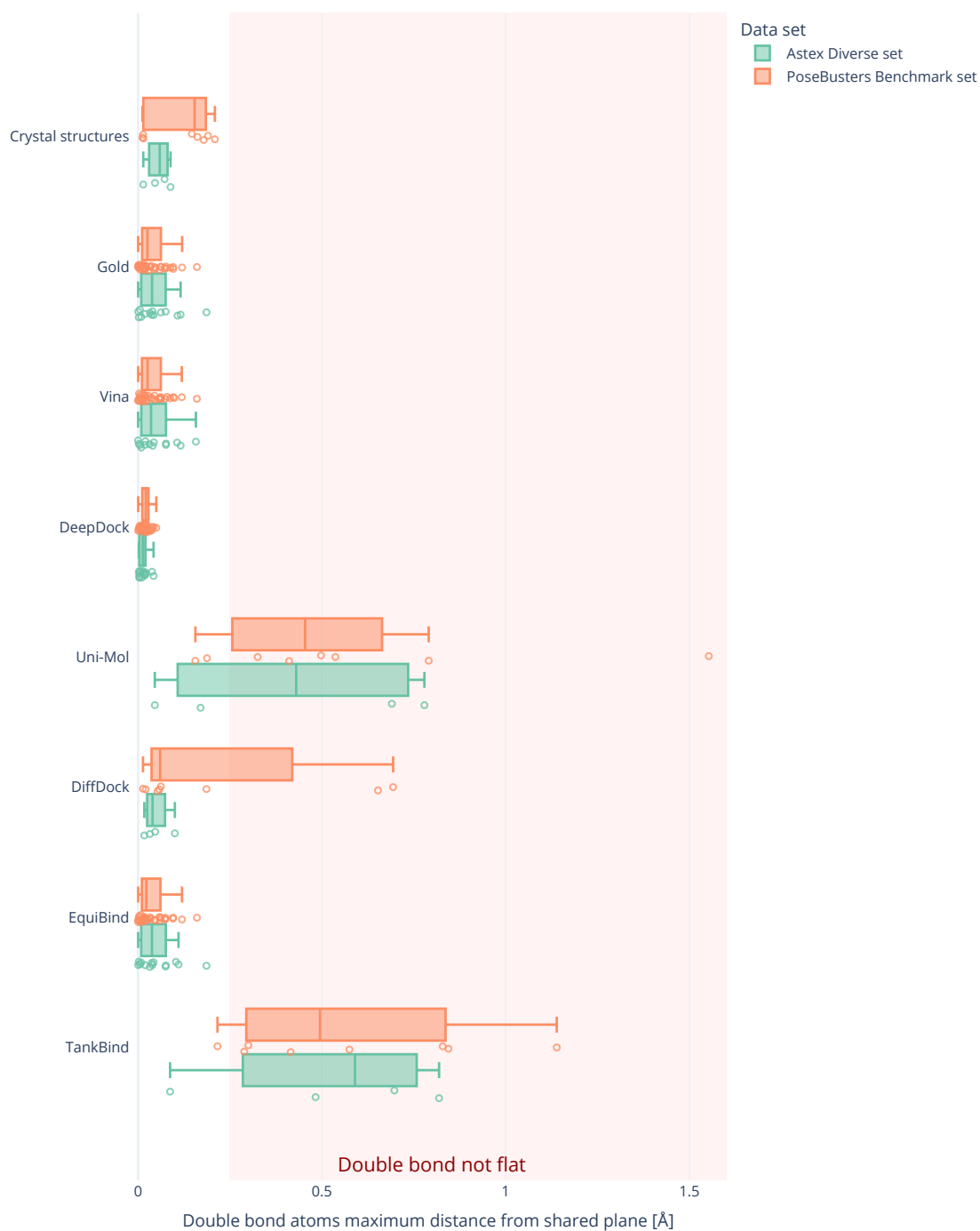


Figure B.10: Distance from shared plane of atoms around aliphatic carbon-carbon double bonds. The largest distance in Angstrom from the shared plane of the two carbons and their four neighbours is shown for each protein ligand complex. If a ligand has no aliphatic carbon-carbon double bonds it is not shown. TankBind and Uni-Mol generate non-flat double bonds.

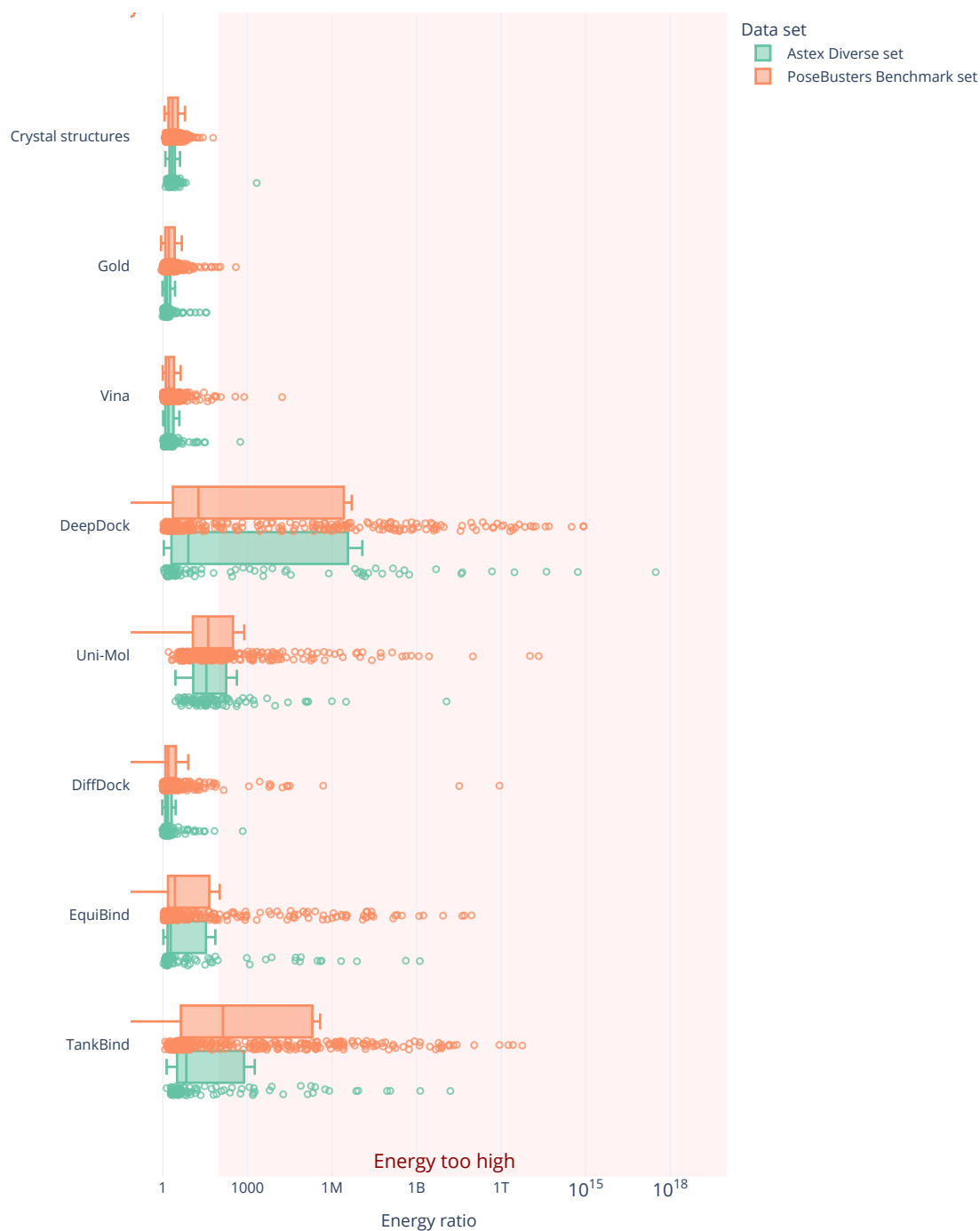


Figure B.11: Energy ratio distributions. The ratio is the energy of the predicted ligand conformation over the average energy of an ensemble of 50 conformations generated with ETKDGv3. The UFF implemented in RDKit was used. The dashed red line shows the cutoff value of 100. There is only one crystal ligand in each data set with a higher energy ratio than the cutoff but all docking methods generate multiple high energy conformations above the cutoff.



Figure B.12: Minimum distances between protein and ligand. Distance is the smallest pairwise distance of heavy atoms of the ligand and protein normalized by their sum of van der Waals radii. The red area highlights the rejection zone below the cutoff of 0.75

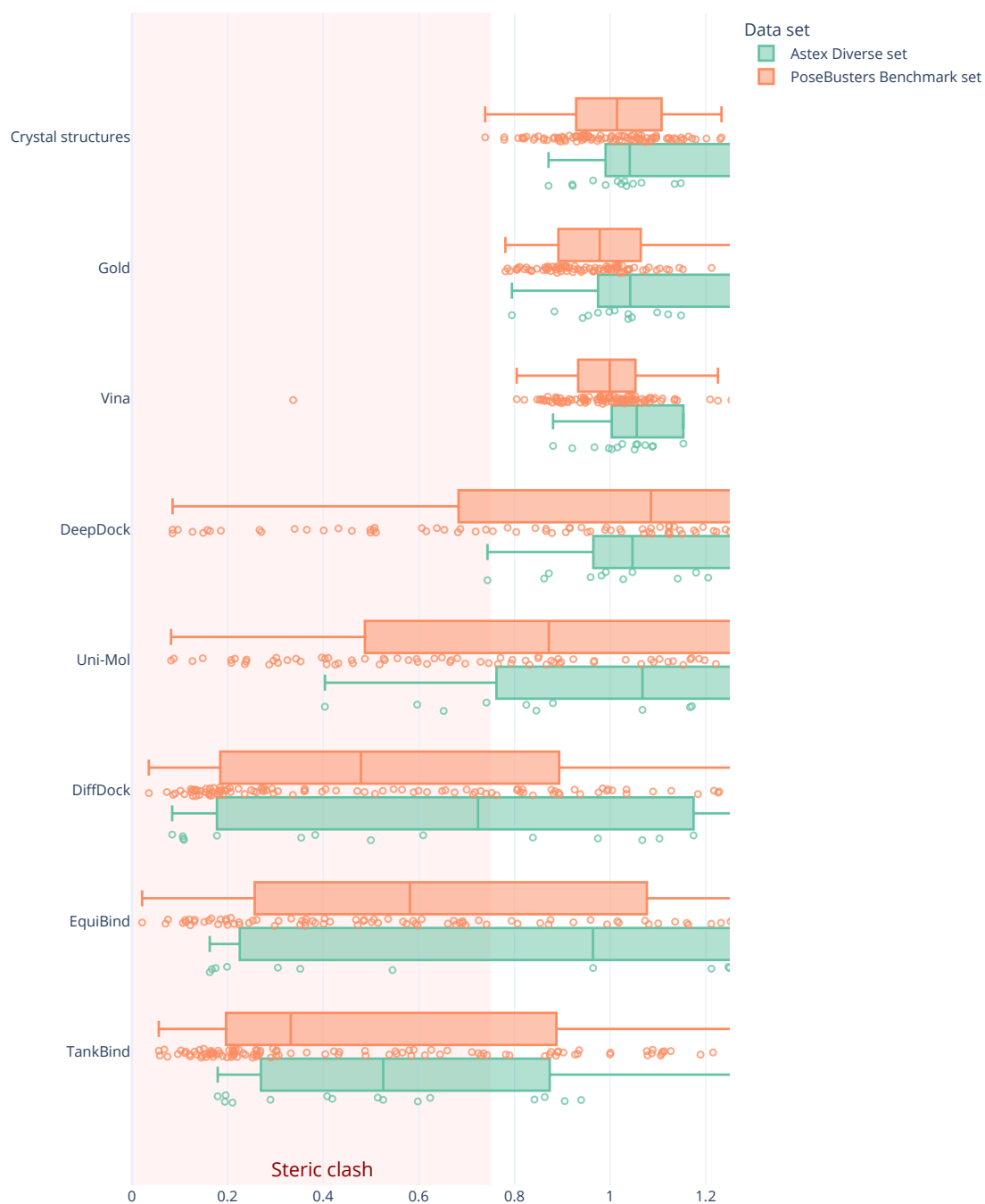


Figure B.13: Minimum distances between ligand and organic molecules. Distance is the smallest pairwise distance of heavy atoms of the ligand and organic molecules normalized by their sum of van der Waals radii. The red area highlights the rejection zone below the cutoff of 0.75

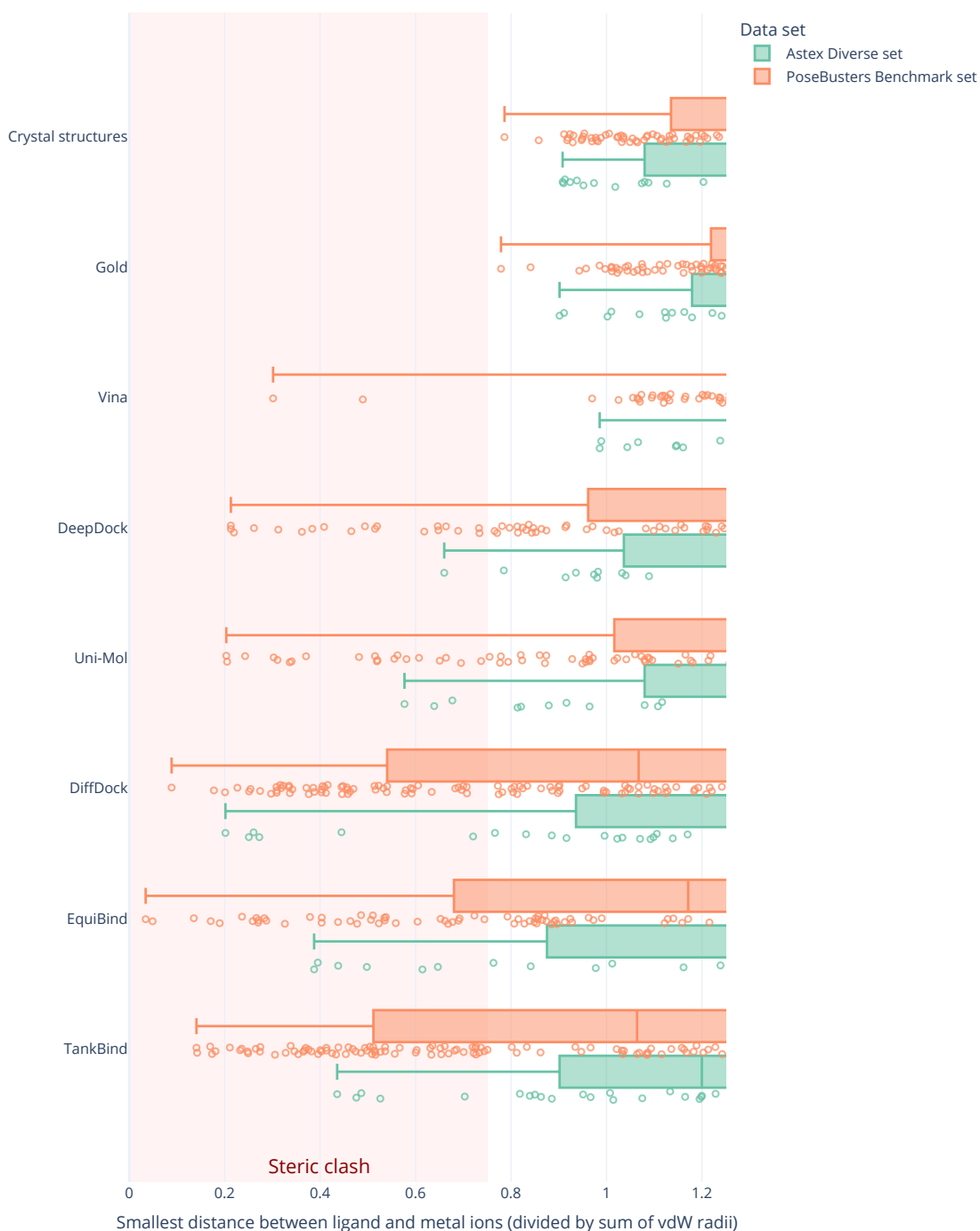


Figure B.14: Minimum distances between ligand and inorganic cofactors. Distance is the smallest pairwise distance of heavy atoms of the ligand and inorganic cofactors normalized by their sum of covalent radii. The red area highlights the rejection zone below the cutoff of 0.75

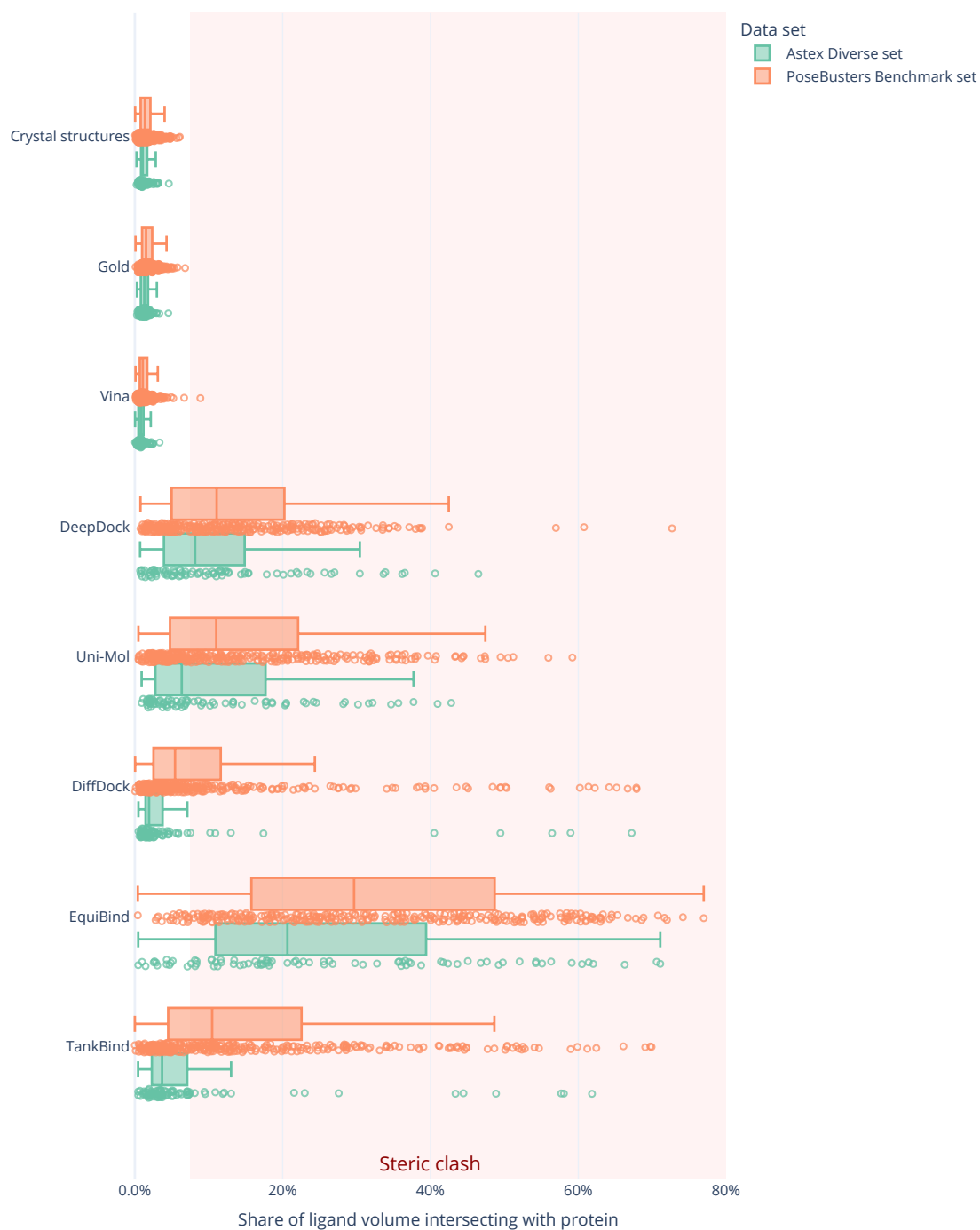


Figure B.15: Volume overlap of protein and ligand. Volume overlap is the percentage of the ligand that overlaps with the protein. The volumes are the van der Waals volumes of the heavy atoms of the ligand and protein. The van der Waals radii are scaled by 0.8. The red dashed line shows the used cutoff of 5%.

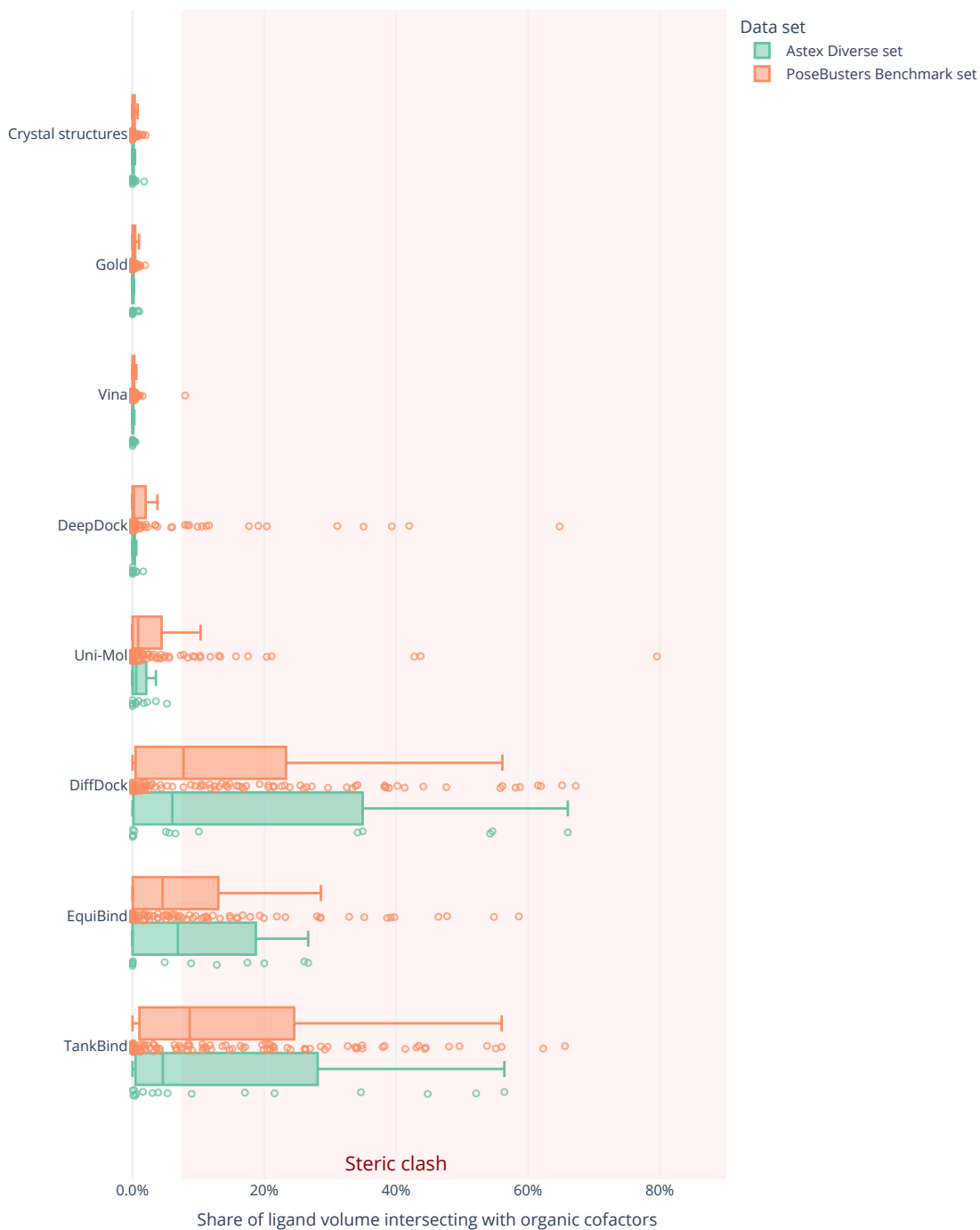


Figure B.16: Volume overlap of ligand and organic molecules. Volume overlap is the percentage of the ligand that overlaps with the organic molecules. The volumes are the van der Waals volumes of the heavy atoms of the ligand and organic molecules. The van der Waals radii are scaled by 0.8. The red dashed line shows the used cutoff of 5%.

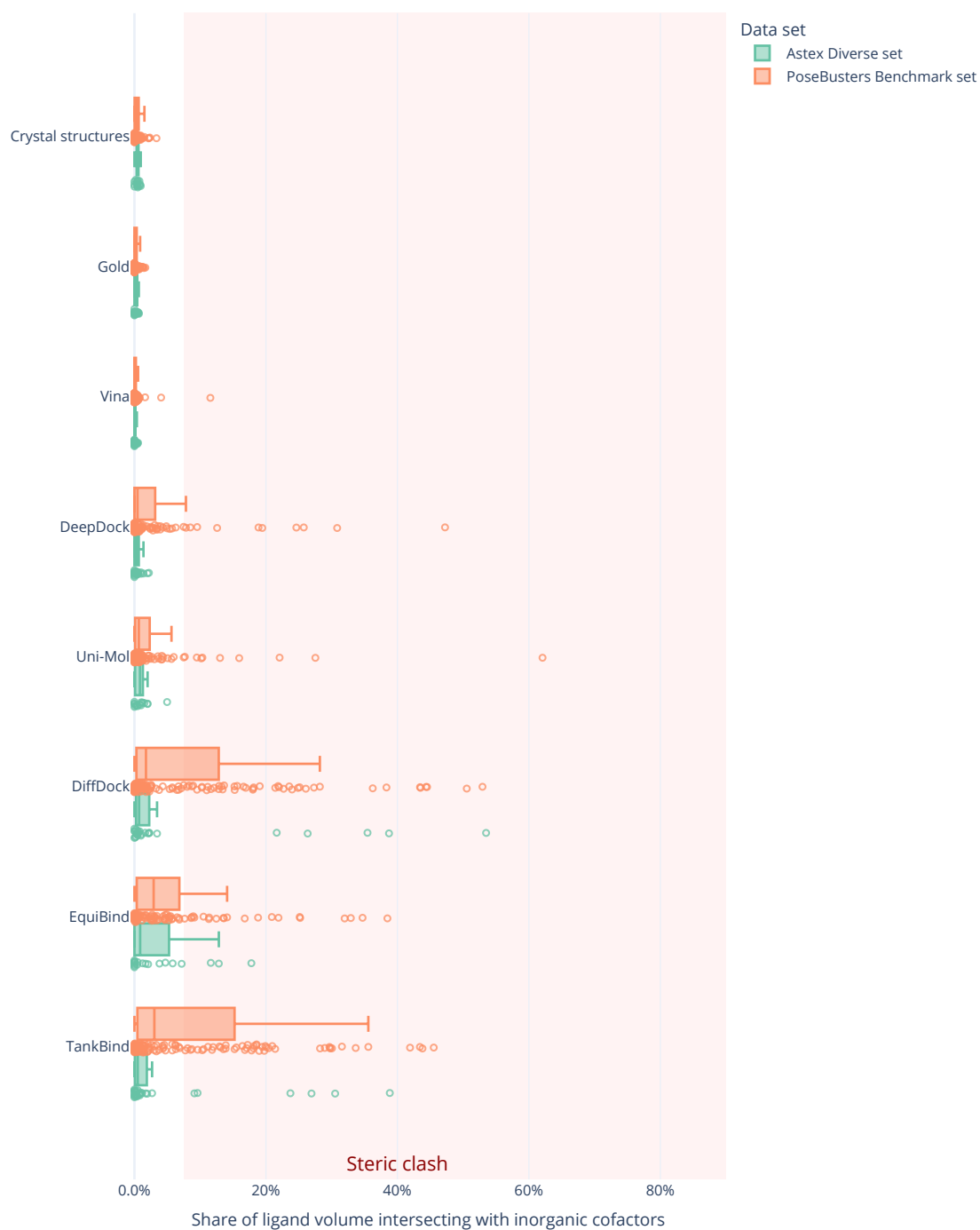


Figure B.17: Volume overlap of ligand and metal ions. Volume overlap is the percentage of the ligand that overlaps with the metal ions. The volumes are the van der Waals volumes of the heavy atoms of the ligand and metal ions. The van der Waals radii are scaled by 0.8. The red zone shows the used cutoff of 20%.

B.9 Alternative binding site definitions for Uni-Mol

This section explores a range of binding site definitions starting from Uni-Mol’s preferred definition of all residues with an atom within 6 Å of a heavy atom of the crystal ligand. The Uni-Mol results shown here were generated using the docking procedure described in Section B.2 but with different thresholds for selecting residues around the crystal ligand. The additional CCDC GOLD results were generated using the procedure described in Section B.2 but with a binding site selection using `BindingSiteFromLigand` with different thresholds and always selecting the entire residues.

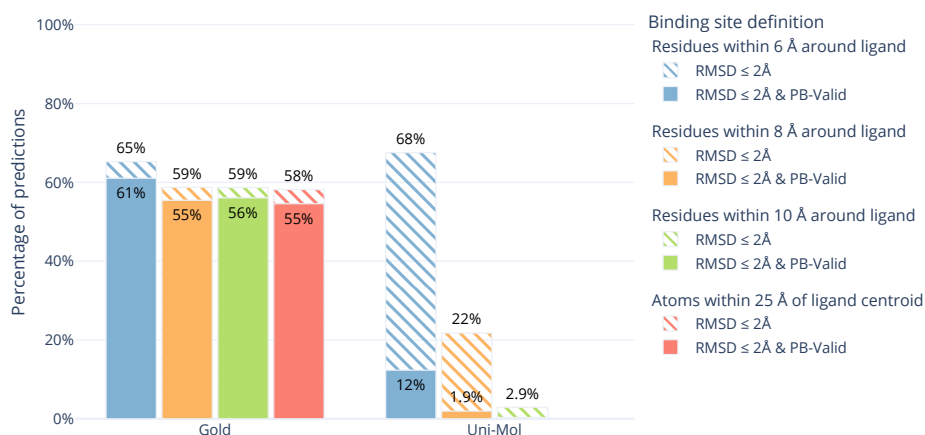


Figure B.18: Performance of CCDC GOLD and Uni-Mol on the PoseBusters Benchmark set on a range of binding site definitions starting from Uni-Mol’s preferred definition of all residues with an atom within 6 Å of a heavy atom of the crystal ligand. The striped bars show the share of predictions of each method that have an RMSD within 2 Å and the solid bars show the subset that in addition have valid geometries and energies, *i.e.*, pass all PoseBusters tests and are therefore ‘PB-Valid’. Under the tight pocket definition of 6 Å Uni-Mol performs better than any of the blind docking methods. The main text (Figure 2.3) shows the results for GOLD shown here in red defining the search space as a sphere with a radius of 25 Å centred on the crystal ligand’s centroid and the results for Uni-Mol are the results shown here in orange defining the binding site as all residues with an atom within 8 Å of a crystal ligand heavy atom. These search spaces were selected to make both methods more comparable to the blind docking methods.

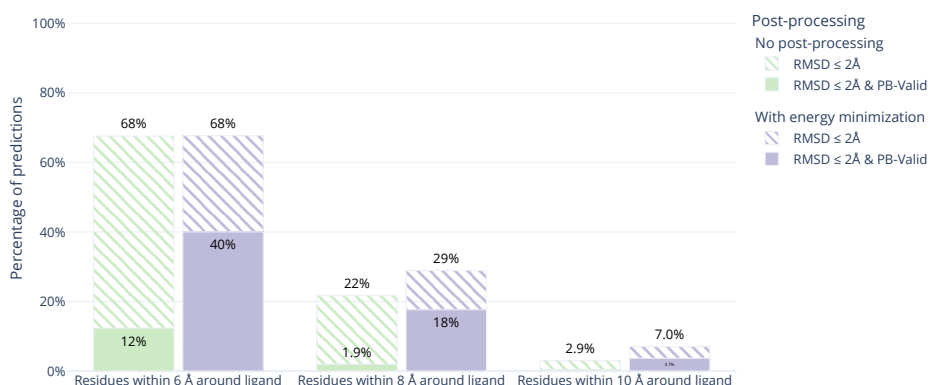


Figure B.19: Performance of Uni-Mol on the PoseBusters Benchmark set for a range of binding site definitions with and without energy minimization as an additional post-processing step. The striped bars show the share of predictions of each method that have an RMSD within 2 Å of the crystal pose and the solid bars show those predictions which in addition pass all PoseBusters tests and are therefore PB-valid. Under the tight binding site definition the number of poses after energy minimisation that pass the tests increases to about the same level as DiffDock (35 %, Figure 2.3).

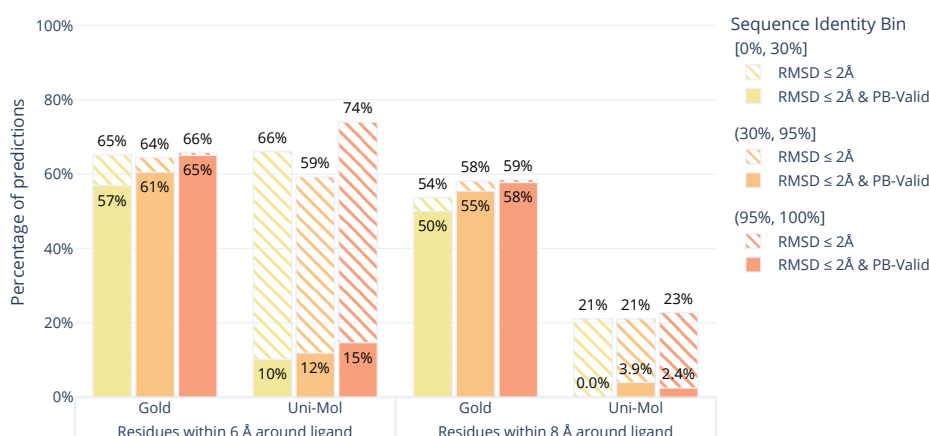


Figure B.20: Comparative performance of CCDC GOLD and Uni-Mol on the PoseBusters Benchmark set stratified by sequence identity relative to the PDDBind General Set v2020. The sequence identity is the maximum sequence identity between all chains in the PoseBusters test protein and all chains in the PDDBind General Set v2020. The striped bars show the share of predictions of each method that have an RMSD within 2 Å and the solid bars show those predictions which in addition pass all PoseBusters tests and are therefore PB-valid. The DL-based methods perform far better on proteins that are similar to those they were trained on.

Appendix C

Further results for Chapter 3

C.1 Failure modes of the PoseBusters test set

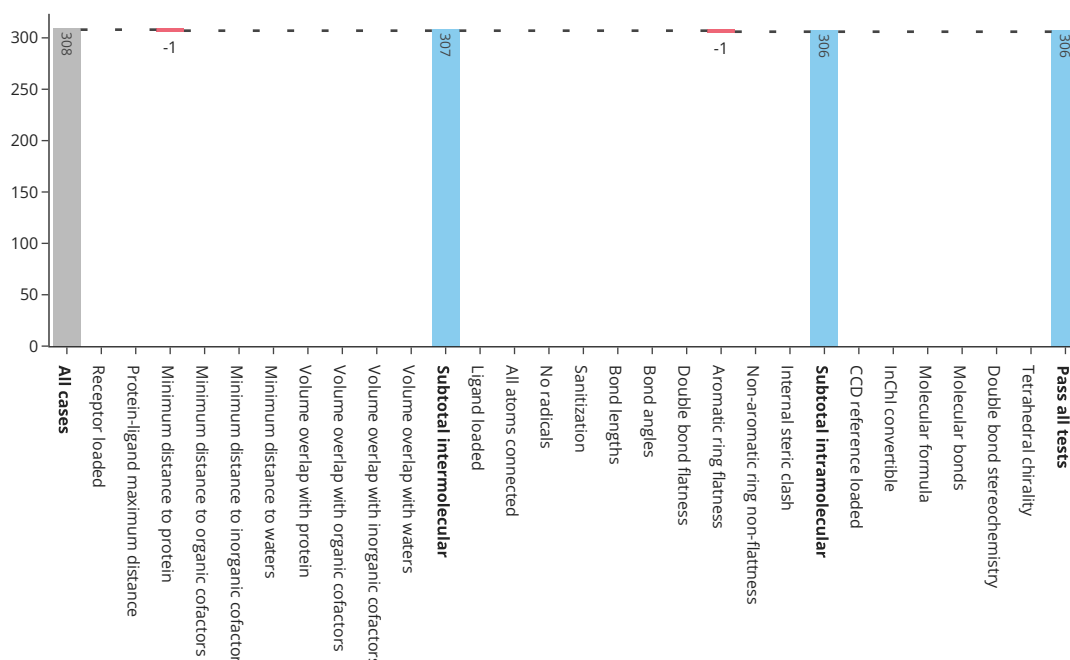
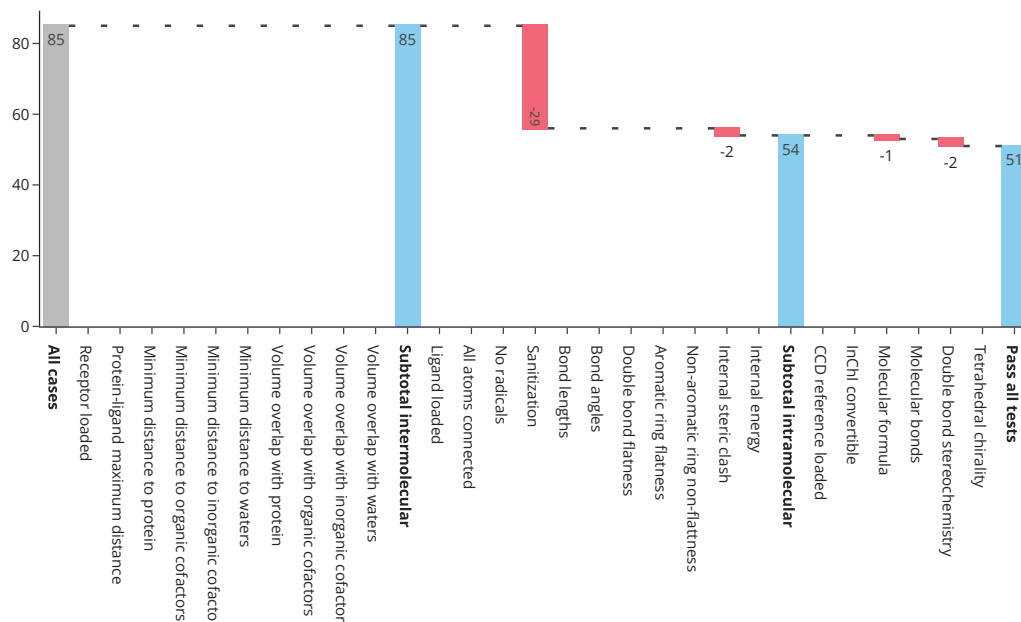


Figure C.1: Failure-mode waterfall for PoseBusters set depicting the cumulative elimination of protein–ligand complexes across successive tests. The gray bar shows the total number of complexes tested, the red bars show the successive failures, and the blue bars show the remaining complexes passing all tests to the left. The PoseBusters set was introduced in Chapter 2 to compare protein–ligand docking methods. Overall, 306 (99 %) of the 308 protein–ligand complexes pass all tests. There are only two failures: one protein–ligand clash and one bent aromatic ring.

C.2 Failure modes of the Astex Diverse set

(a) Original files (Astex Diverse)



(b) Regenerated files from the PDB (Astex Diverse PB)

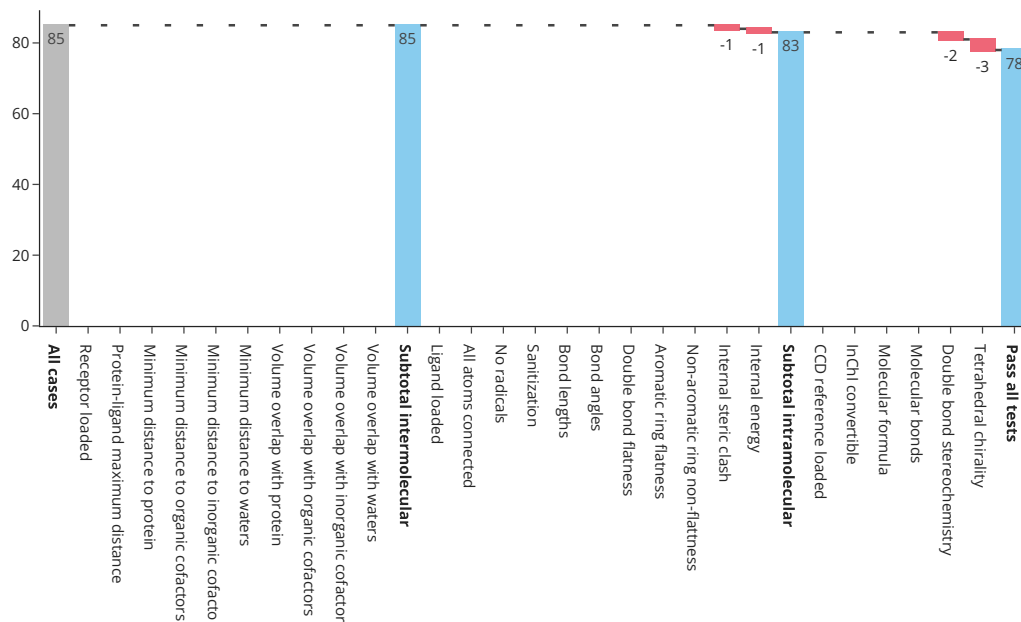
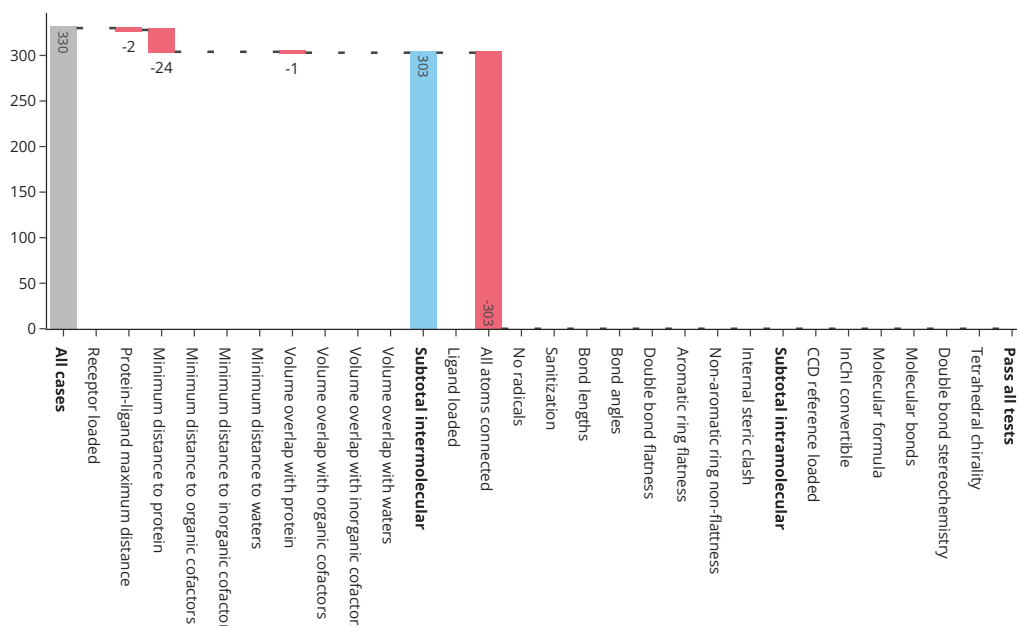


Figure C.2: Astex Diverse set failure mode waterfall plots. The plot shows the number of protein-ligand complexes that fail in each test; the remaining ones pass all tests. The Astex Diverse set was curated by Hartshorn et al. to compare docking methods. Of the 85 protein-ligand complexes, 51 (63 %) pass all tests; the main failure mode is the RDKit ligand sanitization cheque. The Astex Diverse PB set was re-created using structure files from the PDB. Of the 85 protein-ligand complexes, 79 (93 %) pass all tests. The ligand files generated from the PDB structures all pass the sanitisation checks. However, the stereochemistry of 5 of the ligands does not agree with the CCD entries.

C.3 Failure modes of the DockGen test set

(a) Original files



(b) Ligand PDB files with filled-in bond information

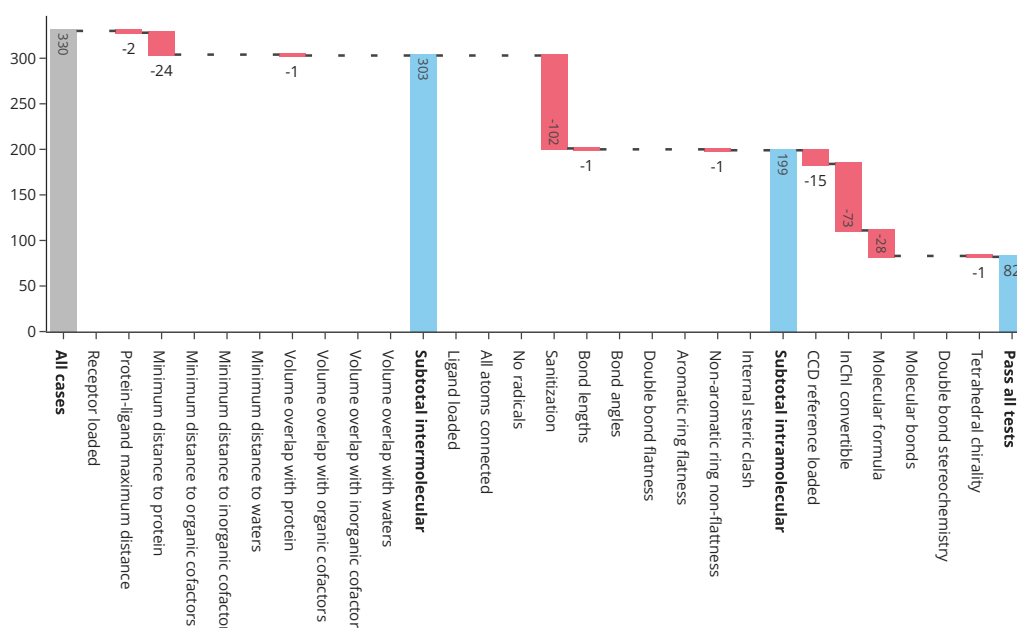


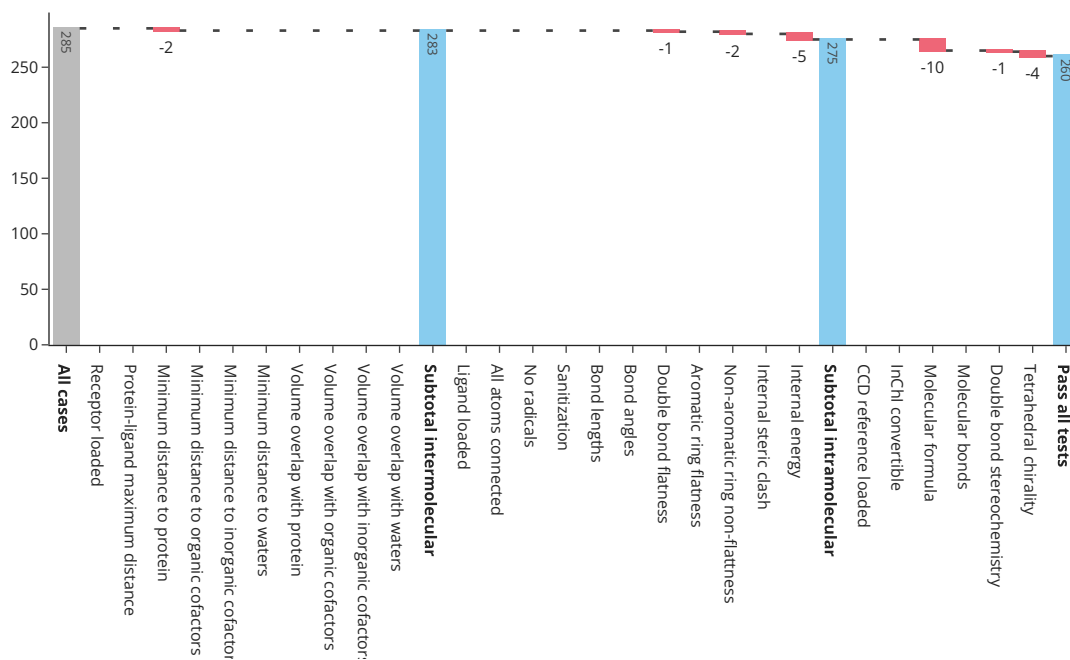
Figure C.3: DockGen failure mode waterfall plots for the original PDB files and the PDB with the filled-in bonds. The plot shows the number of protein-ligand complexes that fail at each test; the remaining ones pass all tests. The DockGen data set was introduced as the validation and testing set in DiffDock-L (Corso et al., 2024). Overall, 0 (0 %) out of the 330 protein-ligand complexes pass all tests. This is because the PDB files do not contain bond information and the RDKit does not fill these in automatically using the residue names that match molecular substructures in the CCD. In the chapter, another waterfall plot (Figure C.3b) shows the results after filling in the bond information using the CCD entries with PyMOL.

C.4 Failure modes of the CASF-2016 test set

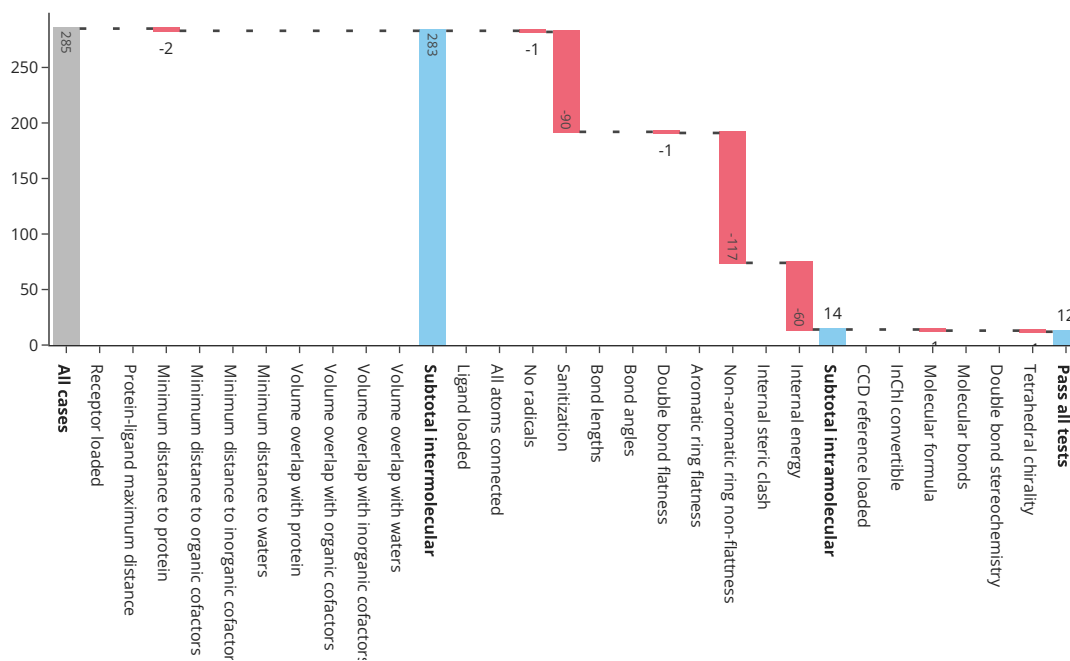
Table C.1: CASF-2016 results of the PoseBusters checks on the different file types. The data sets' molecule files were loaded with the RDKit and tested with the PoseBusters tests. The table shows the percentage of the 285 protein-ligand complexes that passed all intermolecular tests, all intramolecular tests, all identity tests, and all tests overall. The intermolecular tests contain all tests relative to molecular structures in the receptor; the intramolecular tests contain the tests that only require the crystal ligand file; the identity tests contain the checks against the CCD entry; and all tests is all tests. All data sets have some problems, but the smaller and 'high quality' data sets have less failure modes.

Ligand	Receptor	% intermolecular	% intramolecular	% identity	% all tests
MOL2	MOL2	0.0	96.8	93.7	0.0
MOL2	PDB	99.3	96.8	93.7	91.2
SDF	MOL2	0.0	25.3	4.9	0.0
SDF	PDB	99.3	25.3	4.9	4.2

(a) MOL2 ligand and PDB receptor



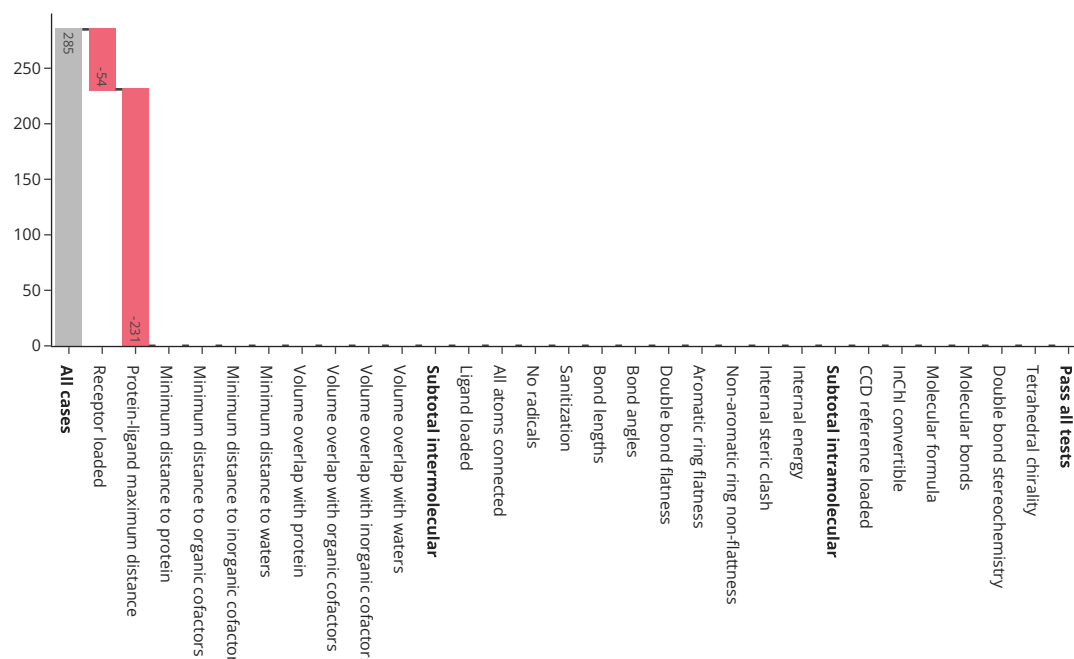
(b) SDF ligand and PDB receptor



Continued on next page

Continued from previous page

(c) SDF ligand and MOL2 receptor



(d) MOL2 ligand and MOL2 receptor

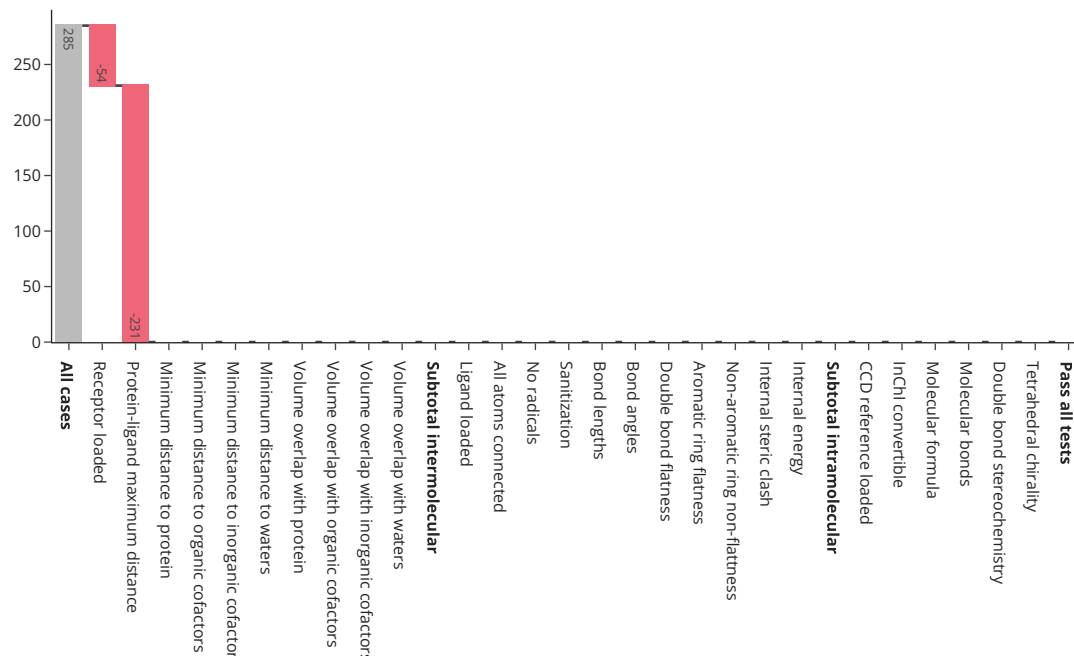
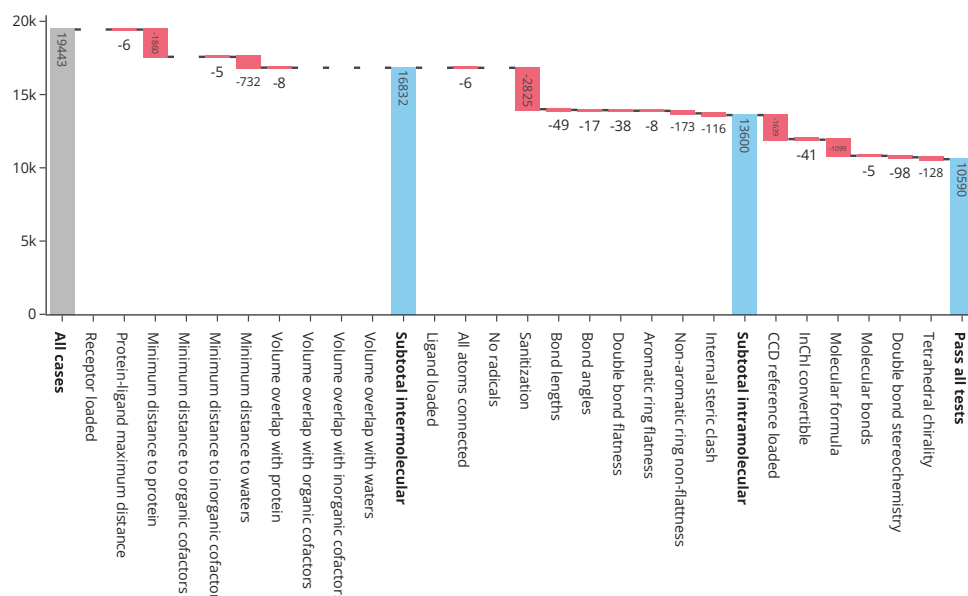


Figure C.4: CASF-2016 failure mode waterfall plots for loading the structures from the different file formats. The plot shows the number of protein-ligand complexes that fail in each test; the remaining ones pass all tests. In the main text, the waterfall for the MOL2 ligand and PDB receptor files is shown (Figure C.4a). The RDKit correctly processes the MOL2 ligand files and the PDB receptor files, but not the MOL2 receptor files or the SDF ligand files.

C.5 Failure modes of the PDBbind General training set

(a) PDBbind General 2020



(b) PDBbind General 2020 R1.

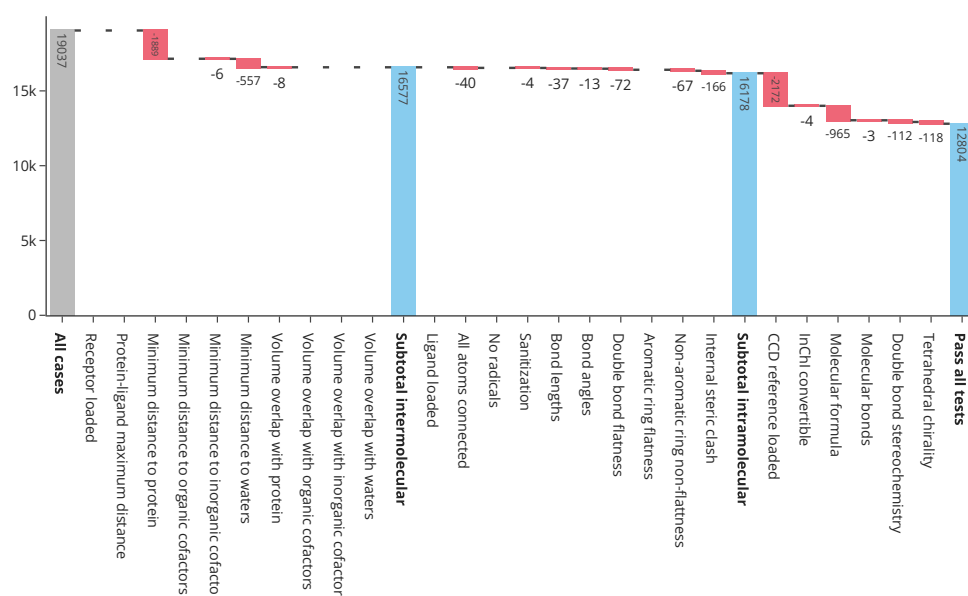
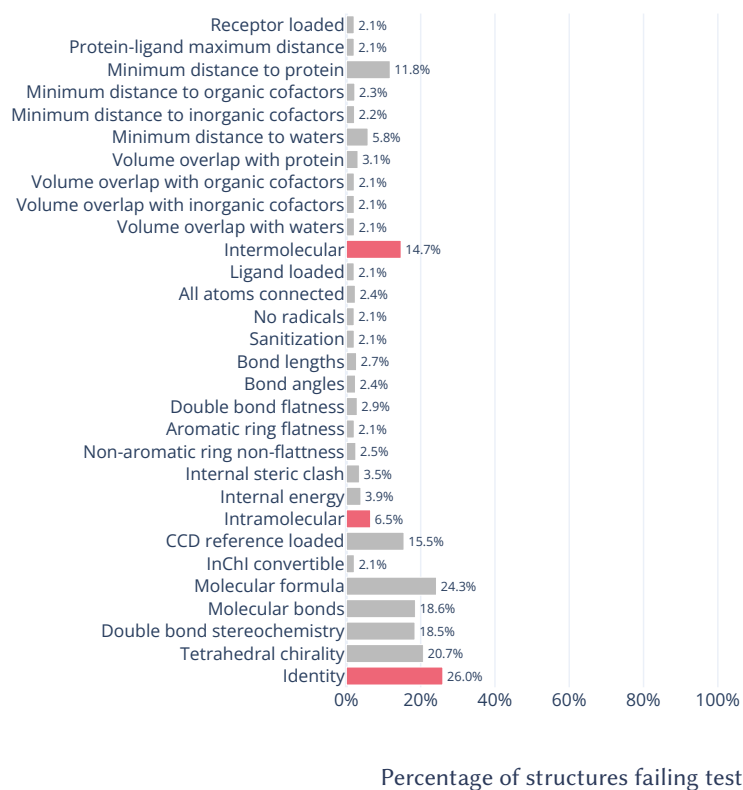


Figure C.5: PDBbind General 2020 PoseBusters failure mode waterfall plots. The plot shows the number of protein-ligand complexes that fail at each test; the remaining ones pass all tests. The PDBbind data set was combined structure and binding affinity data and it has been used to train many DL-based docking methods, including all 5 that were benchmarked in Chapter 2. Overall, 54 % (10590) of the 19443 protein-ligand complexes pass all tests. The largest failure mode related to the receptor structures is the protein-ligand clash (9.5 %, 1860). The largest failure mode related to the ligand structures is the RDKit sanitization of the ligand (14.5 %, 2825). These failure modes. Note that 2736 of the ligands in this data set are multimers and not single entries in the CCD.

(a) PDBbind General 2020



Continued on next page

Continued from previous page

(b) PDBbind General 2020 R1

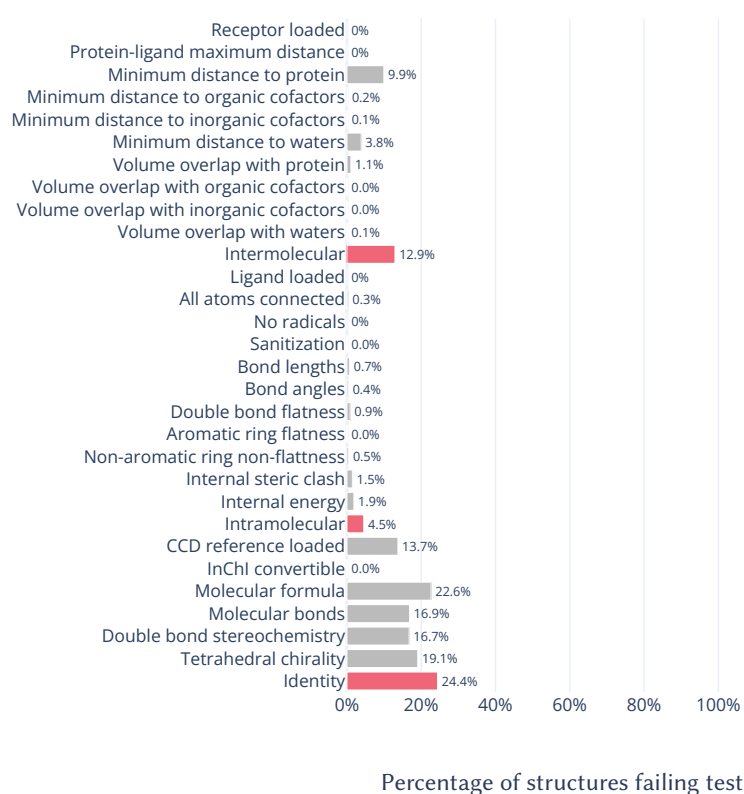
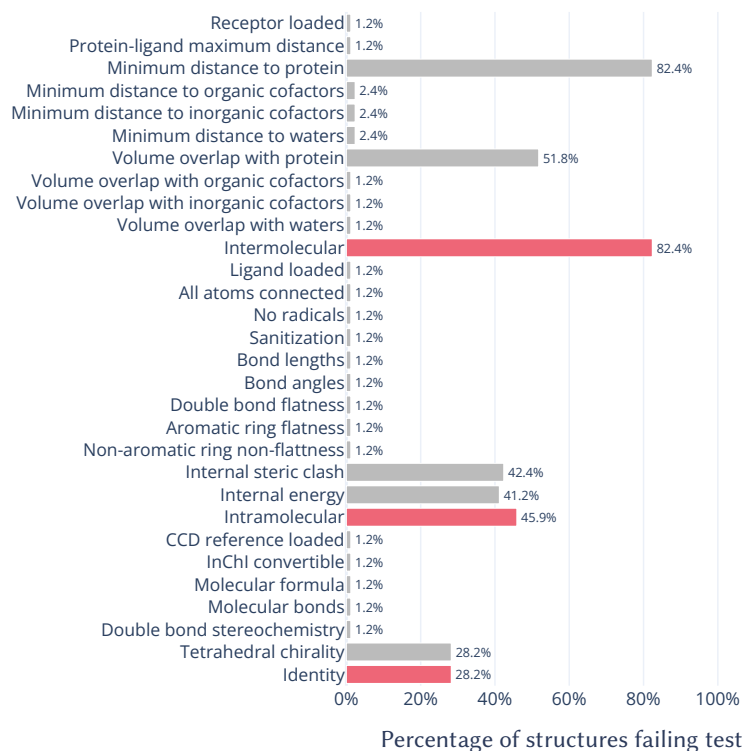
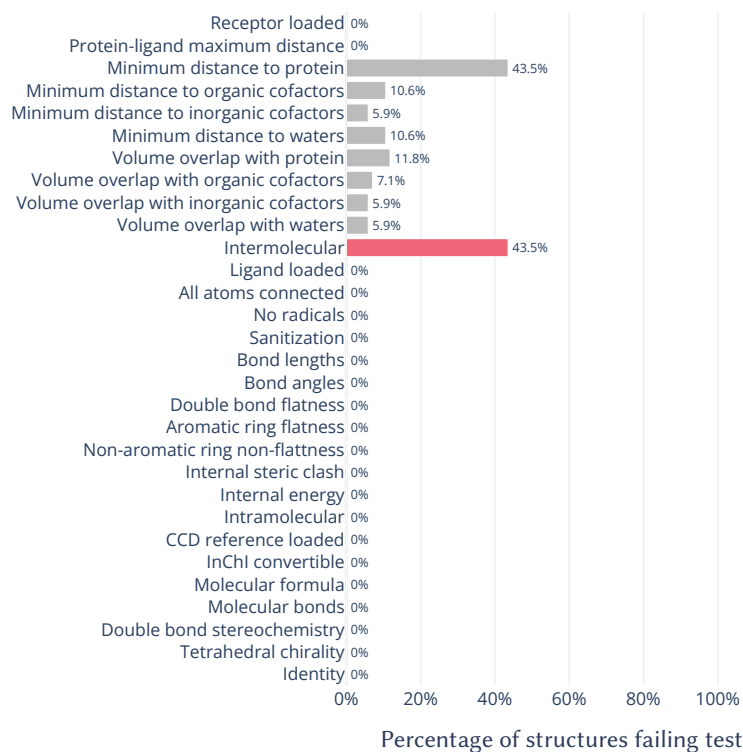


Figure C.6: Failure modes of the PDBbind General 2020 and PDBbind General 2020 R1 training data sets. The PDBbind General set was used for the training of all the DL-based methods tested in Chapter 2. The failure modes of the methods are shown in the same format for re-docking the ligands of the Astex Diverse set in Figure C.7 and the PoseBusters set in Figure C.8. The training data contain systematic structural artefacts and share some failure modes with the models. However, the failure rates of the methods are much higher than those observed in the training data set. For example, all methods generated more than 40 % poses with protein-ligand clashes, but the training data has less than 12 %.

(a) DeepDock



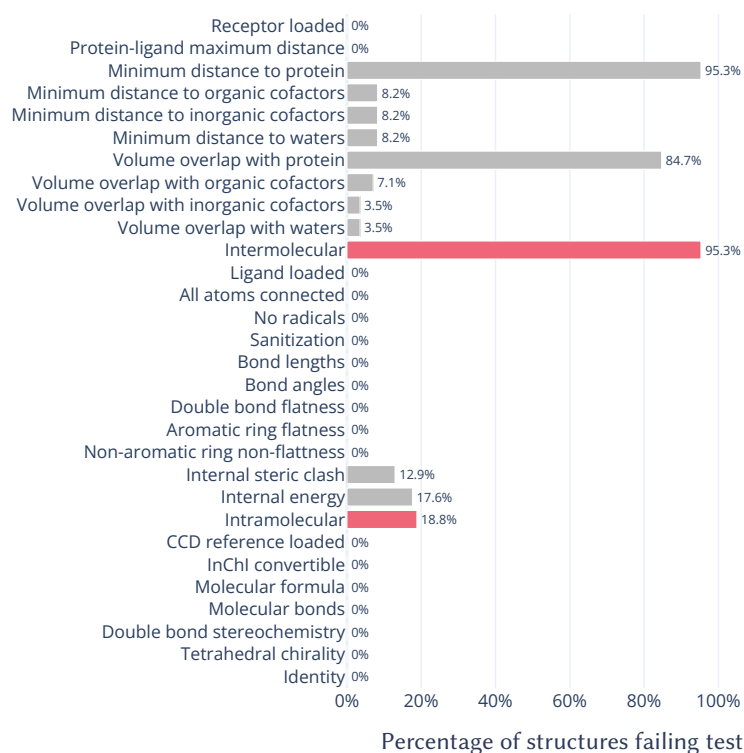
(b) DiffDock



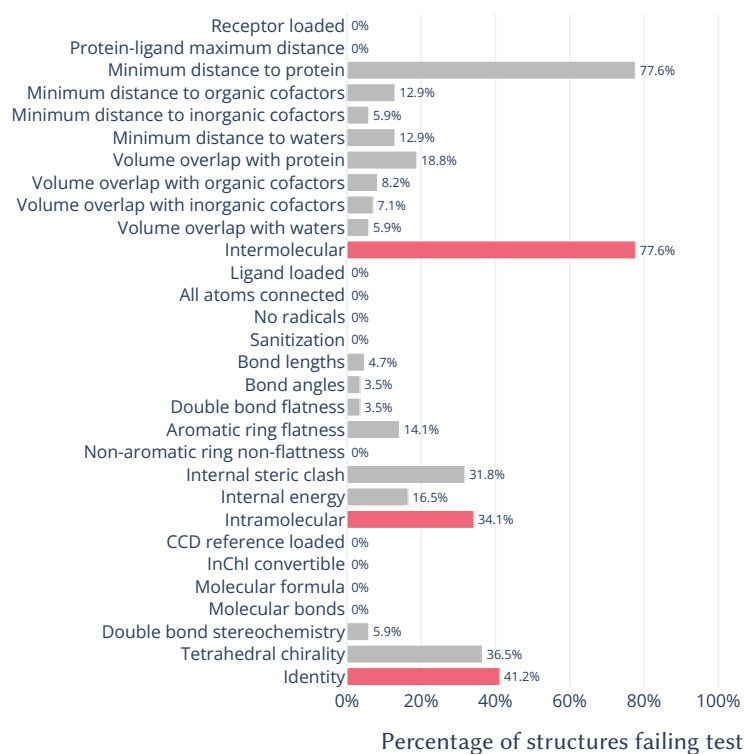
Continued on next page

Continued from previous page

(c) EquiBind



(d) TankBind



Continued on next page

Continued from previous page

(e) UniMol

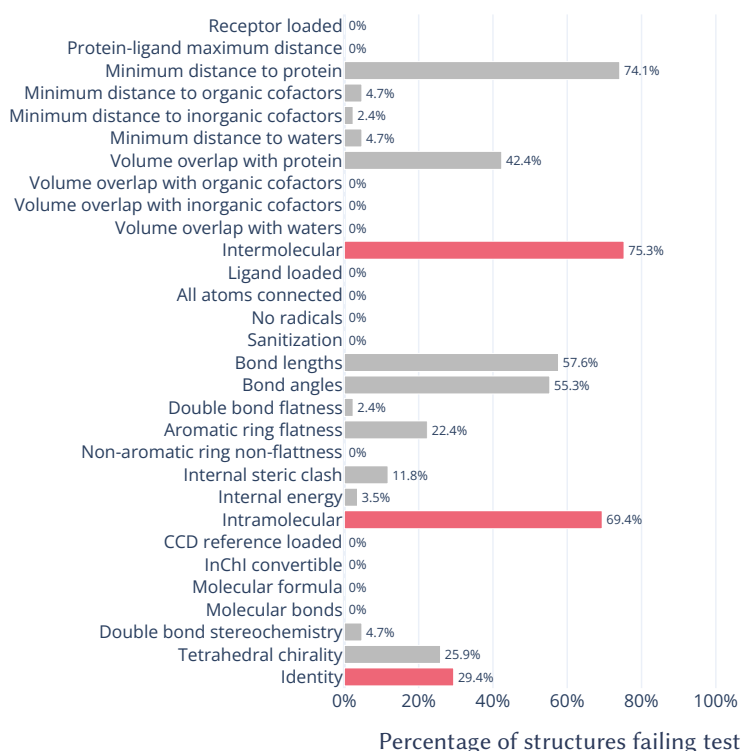
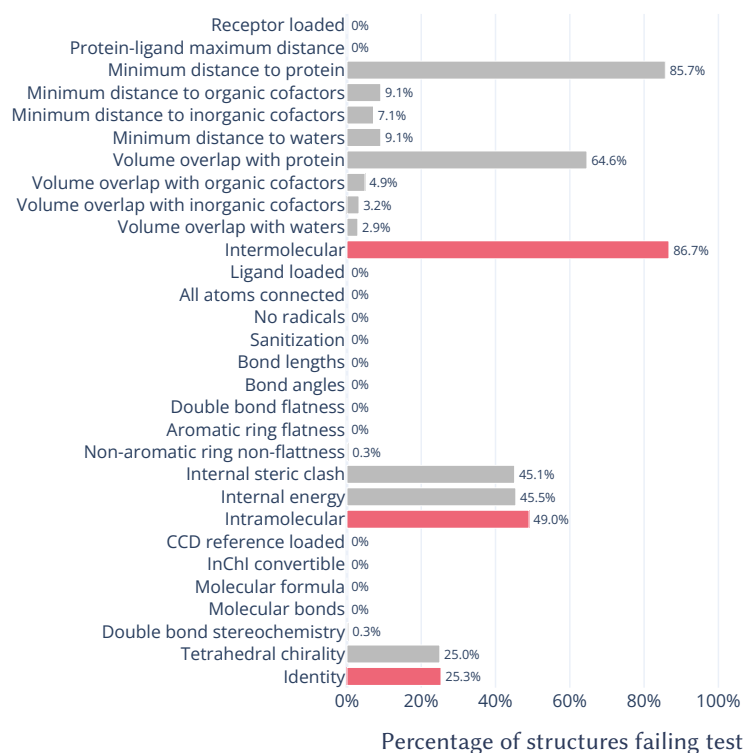
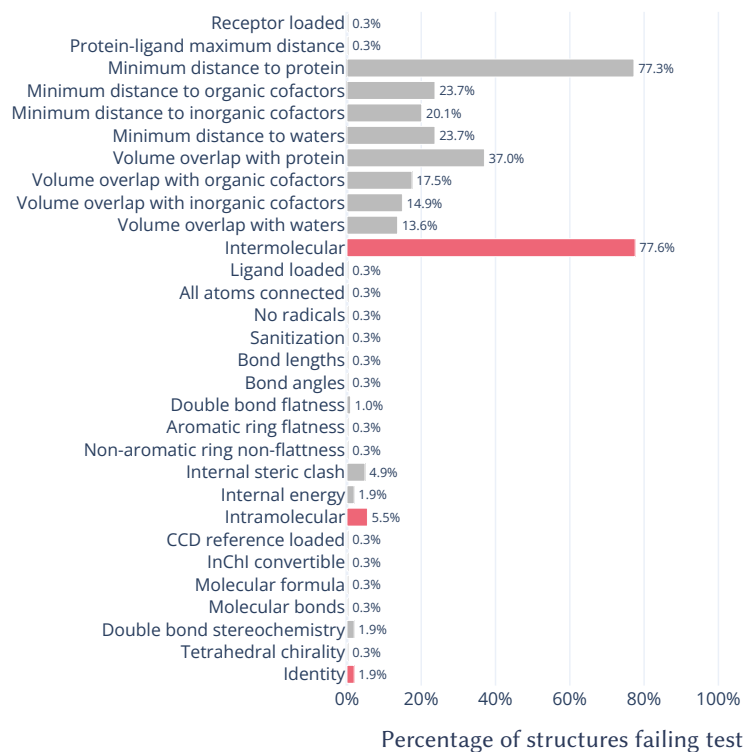


Figure C.7: PoseBusters test suite failure rates for re-docking the Astex Diverse set. The bars indicate the failure rates independently to enable comparison to the training data set failure rates shown in Figure C.6; the results are the same as in Figure B.3 and are described in Chapter 2. The failure rates of the methods are much higher than those observed in the training data set. For example, all methods generated more than 40 % poses with protein-ligand clashes, but the training data has less than 12 % (Figure C.6).

(a) DeepDock



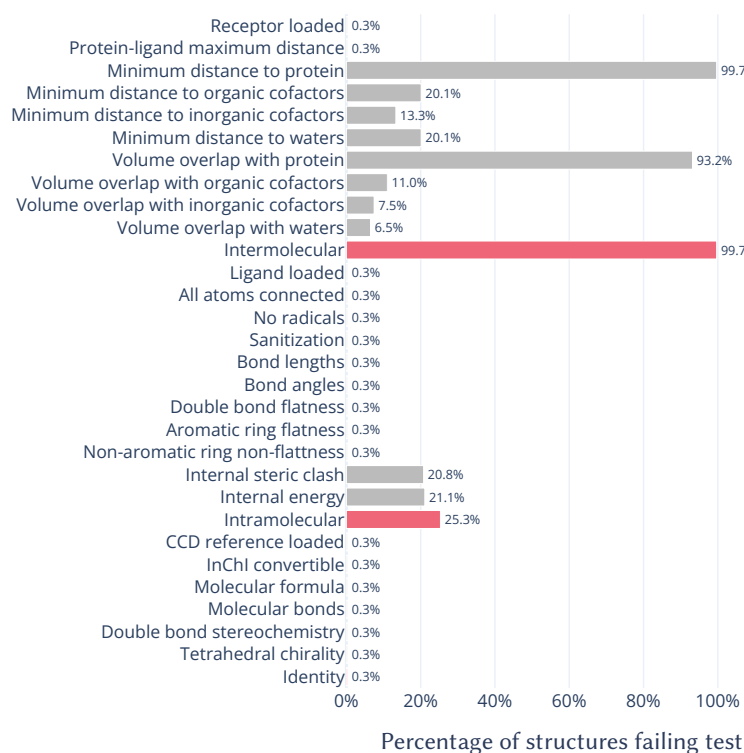
(b) DiffDock



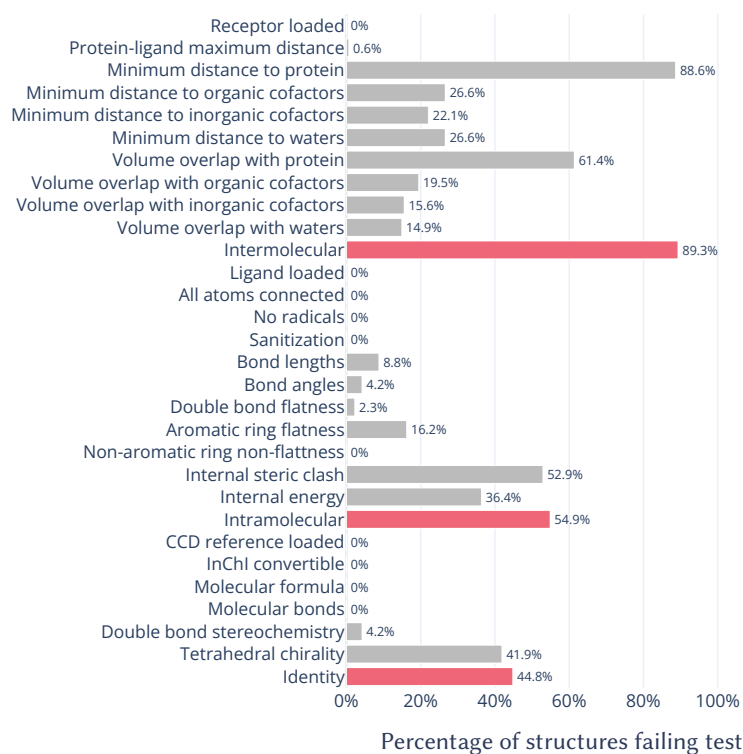
Continued on next page

Continued from previous page

(c) EquiBind



(d) TankBind



Continued on next page

Continued from previous page

(e) UniMol

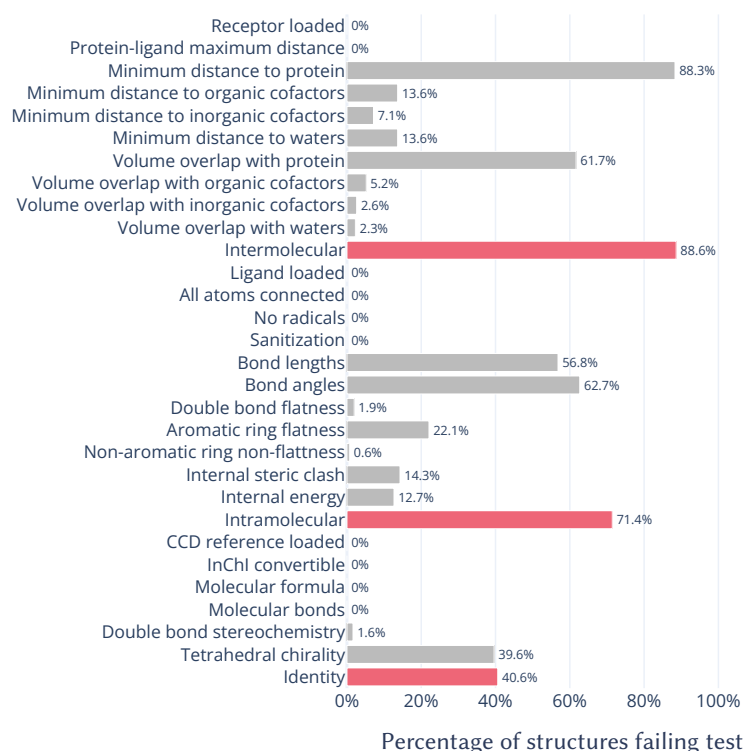


Figure C.8: PoseBusters test suite failure rates for re-docking the PoseBusters set. The bars indicate the failure rates independently to enable comparison to the training data set failure rates shown in Figure C.6; the results are the same as in Figure B.4 and are described in Chapter 2. The failure rates of the methods are much higher than those observed in the training data set. For example, all methods generated more than 40 % poses with protein-ligand clashes, but the training data has less than 12 % (Figure C.6).

Appendix D

Extended methods and further results for Chapter 4

D.1 CCD Block list

The combined block list contains the following 1989 CCD identifiers: 02U, 045, 05L, 07E, 07Y, 08U, 09X, 0BD, 0BE, 0H0, 0HX, 0KA, 0LP, 0MK, 0NZ, 0QE, 0UB, 0V4, 0WK, 0XY, 0YT, 10M, 118, 119, 12E, 12P, 13P, 144, 145, 147, 149, 14T, 15L, 15P, 16A, 16D, 16F, 16G, 16O, 16P, 17T, 18D, 18O, 1AG, 1AL, 1B0, 1CF, 1CP, 1CU, 1EM, 1FH, 1FT, 1GL, 1GN, 1LL, 1PE, 1PG, 1PS, 1S3, 1S4, 1SD, 1X4, 202, 20S, 20X, 220, 22S, 23V, 24S, 25E, 260, 27C, 289, 291, 293, 2BM, 2DG, 2DP, 2DR, 2F2, 2F8, 2FG, 2FH, 2FK, 2FL, 2FU, 2GL, 2GS, 2H5, 2HA, 2HP, 2JC, 2M4, 2M5, 2M8, 2ME, 2MO, 2NO, 2OF, 2OP, 2OS, 2PA, 2PE, 2PN, 2PO, 2T8, 2WP, 2WS, 320, 330, 34V, 38J, 3BR, 3BU, 3CN, 3CO, 3DO, 3DY, 3FM, 3GR, 3HD, 3HR, 3J3, 3J4, 3LJ, 3LR, 3MG, 3MK, 3MT, 3NI, 3OF, 3PE, 3PG, 3PH, 3PO, 3R3, 3S6, 3SA, 3SY, 3YW, 40J, 42D, 445, 44S, 46D, 46Z, 475, 48Z, 491, 49A, 49S, 49T, 49V, 4AG, 4AM, 4CQ, 4GC, 4GL, 4GP, 4IP, 4JA, 4MO, 4N2, 4NN, 4OA, 4PU, 4QY, 4R1, 4RS, 4SG, 4TI, 4UZ, 4V5, 50A, 51N, 543, 56N, 57S, 5A3, 5AD, 5GF, 5GO, 5II, 5KQ, 5KS, 5KT, 5KV, 5L3, 5LS, 5LT, 5MM, 5N6, 5QP, 5SP, 5TH, 5TJ, 5TK, 5TM, 61J, 62I, 64K, 660, 6BG, 6BP, 6C2, 6DM, 6GB, 6GP, 6GR, 6JZ, 6K3, 6KH, 6KL, 6KS, 6KU, 6KW, 6LA, 6LS, 6LW, 6MJ, 6MN, 6MO, 6PE, 6PH, 6PL, 6PZ, 6S2, 6UD, 6UL, 6WO, 6YR, 6ZC, 73E, 79J, 7CV, 7D1, 7E8, 7E9, 7GP, 7JZ, 7K2, 7K3, 7N5, 7NU, 7PE, 7PG, 7PH, 82N, 83Y, 89Y, 8AR, 8B7, 8B9, 8CL, 8EX, 8GA, 8GG, 8GP, 8I4, 8LR, 8OQ, 8P8, 8PE, 8PK, 8S0, 8YV, 95Z, 960, 98U, 9AM, 9C1, 9CD, 9FO, 9GP, 9JE, 9KJ, 9MR, 9OK, 9PE, 9PG, 9QG, 9S7, 9SG, 9SJ, 9SM, 9SP, 9T1, 9T7, 9VP, 9WJ, 9WN, 9WZ, 9YW, A, A0K, A1Q, A2G, A4P, A5C, A6P, AAE, AAL, ABA, ABD, ABE, ABF, ABL, AC1, ACA, ACE, ACM, ACN, ACR, ACT, ACU, ACX, ACY, ADA, ADP, AE3, AE4, AEM, AF1, AF3, AFD, AFO, AFP, AG, AGA, AGC, AGL, AH2, AH8, AHG, AHM, AHR, AIG, AKR, AL, ALA, ALF, ALL, ALX, ALY, AM, AMG, AMN, AMP, AMU, AMV, ANA, ANL, ANP, AOG, APC, AQA, AR, ARA, ARB, ARF, ARG, ARI, ARS, ART, ARW, ASC, ASG, ASN, ASO, ASP, AST, ATH, ATO, ATP, AU, AU3, AUC, AXP, AXR, AY9, AZC, AZI, B0D, B12, B16, B1H, B1N, B2G, B3H, B3P, B4G, B6D, B7G, B8D, B9D, BA, BAM, BBK, BBU, BBV, BBX, BCD, BCN, BCR, BCT, BDF, BDG, BDN, BDP, BDR, BE, BE7, BEF, BEM, BEN, BEQ, BET, BEZ, BF2, BF4, BFN, BG6, BG8, BGC, BGL, BGN, BGP, BGS, BHG, BM3, BM7, BMA, BME, BMX, BND, BNG, BNX, BNZ, BO1, BO3, BO4, BOG, BQY, BR, BRJ, BRM, BRO, BRP, BS3, BS7, BSY, BTB, BTC, BTG, BTN, BTU, BU1, BU2, BU3, BUQ, BW3, BWG, BXC, BXC, BXP, BXX, BXY, BZD, C, C10, C14, C15, C10, C2C, C20, C3B, C3G, C3X, C4B, C4W, C5P, C5X, C8E, CA, CAC, CAD, CAQ, CAT, CB3, CB5, CBF, CBG, CBI, CBK, CBM, CBX, CCH, CCN, CD, CD1, CD3, CD4, CD5, CDL, CDP, CDR, CE, CE1, CE5, CE6, CE8, CE9, CEG, CEQ, CEZ, CF, CF0, CFM, CFN, CF0, CFT, CGF, CHD, CHM, CHT, CHX, CIT, CJB, CKB, CKP, CL, CLF, CLN, CL0, CLP, CLR, CLX, CM, CM5, CME, CMJ, CM0, CMP, CN, CN1, CN3, CN6, CNB, CNF, CNM, CNN, CNP, CNS, CNV, CO, CO2, CO3, CO5, COA, COB, COH, COM, CON, CP2, CP3, CPS, CPT, CR, CR1, CR6, CRA, CRY, CS, CSB, CSD, CT3, CTO, CTP, CTR, CTT, CU, CU1, CU2, CU3, CUA, CUL, CUM, CUN, CU0, CUZ, CXE, CXS, CYC, CYN, CYS, D01, D10, D12, D1D, D1M, D5E, D6G, D6N, D8U, DA, DAF, DAG, DAN, DAO, DC, DCE, DD9, DDA, DDH, DDL, DDQ, DDR, DEG, DEL, DEP, DET, DFR, DFX, DG, DG0, DGA, DGD, DGG, DGO, DGS, DGU, DHB, DHD, DHE, DHJ, DIA, DIO, DIS, DJB, DJE, DK4, DKA, DKX, DKZ, DL6, DLD, DLF, DLG, DME, DMF, DMI, DMN, DMR, DMS, DMU, DMX, DN, DNO, D08, DOD, DOM, DOX, DPC, DPG, DPO, DPR, DQR, DR2, DR3, DR5, DR6, DRI, DSC, DSR, DT, DT6, DTD, DTI, DTN, DTT, DTU, DTV, DUD, DUM, DVC, DVT, DXE, DXG, DY, DYM, DZZ, E1H, E3M, E4N, E5G, EAG, EAP, EBG, EBQ, EDN, ED0, EDR, EEE, EEN, EEQ, EGA, EGL, EHN, EMC, EMP, EMZ, ENC, EOH, EOM, EOX, EPE, EPG, EQP, EQV, ER3, ERE, ERI, ETA, ETE, ETF, ETI, ETN, ETT, ETX, ETZ, EU, EU3, EUS, F, F09, F1P, F1X, F20, F3S, F42, F50, F55, F58, F6P, F8X, FAD, FAR, FBP, FCA, FCB, FCL, FCO, FCT, FCY, FDD, FDE, FDP, FDQ, FE, FE2, FEA, FEC, FEL, FE0, FES, FFC, FFX, FIF, FII, FJ0, FK9, FKD, FLC, FLH, FLO, FMF, FMI, FMN, FMO, FMS, FMT,

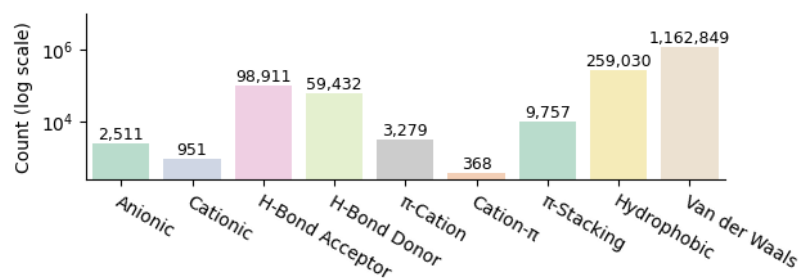
FNE, FNG, FNY, FOL, FOR, FP0, FPP, FRU, FS1, FS2, FS3, FS4, FSA, FSI, FSM, FS0, FSW, FSX, FTT, FU1, FUB, FUC, FUD, FUF, FUL, FUY, FVQ, FW5, FX1, FYJ, G, G0S, G16, G1P, G20, G28, G2F, G3F, G3I, G4D, G4S, G6D, G6P, G6S, G7P, G8Z, GA, GAA, GAC, GAD, GAF, GAI, GAL, GAT, GBH, GBL, GC1, GC4, GC9, GCB, GCD, GCN, GC0, GCP, GCS, GCT, GCU, GCV, GCW, GD, GD3, GDA, GDL, GDP, GE1, GE3, GEP, GFP, GGD, GIV, GL0, GL1, GL2, GL4, GL5, GL6, GL7, GL9, GLA, GLC, GLD, GLF, GLG, GLN, GLO, GLP, GLS, GLT, GLU, GLV, GLY, GM0, GMB, GMH, GMP, GMT, GMZ, GN1, GN4, GNS, GNX, GOL, GP0, GP1, GP4, GPH, GPK, GPM, GPO, GPQ, GPU, GPV, GPW, GPX, GQ1, GRF, GRX, GS1, GS9, GSH, GTK, GTM, GTP, GTR, GTT, GU0, GU1, GU2, GU3, GU4, GU5, GU6, GU7, GU8, GU9, GUF, GUL, GUN, GUP, GUZ, GVT, GXL, GXV, GYE, GYF, GYG, GYP, GYU, GYV, GZL, H1M, H1S, H2P, H2S, H3S, H53, H6Q, H6Z, HAE, HAI, HBZ, HC4, HCA, HCS, HD4, HDD, HDN, HDZ, HE5, HE6, HEA, HEB, HEC, HED, HEG, HEM, HES, HEV, HEX, HEZ, HFM, HFP, HG, HG2, HGC, HGI, HGP, HGX, HIS, HLT, HNI, HNV, HNW, HO, HO3, HOH, HP6, HSG, HSH, HSJ, HSQ, HSW, HSX, HSY, HT3, HTG, HTH, HTM, HTO, HVC, I, I42, I7P, I8P, IAB, IBM, ICE, ICG, ICH, ICI, ICS, ICT, ICZ, IDC, IDF, IDG, IDO, IDR, IDS, IDT, IDU, IDX, IDY, IEM, IHP, IHS, ILE, IMD, IMP, IN, IN1, IOD, IOH, IPA, IPH, IPL, IPT, IR, IR3, IRI, ISD, ISL, ISX, IUM, IXD, J5B, JEF, JFZ, JHM, JLT, JRV, JSV, JV4, JVA, JVS, JZR, K, K5B, K99, KAI, KBA, KBG, KCX, KD5, KDA, KDB, KDD, KDE, KDF, KDJ, KDM, KDN, KDO, KDR, KEN, KFN, KG1, KGM, KHP, KME, KO1, KO2, KO4, KOT, KR, KTU, L0W, L1L, L1P, L2C, L2P, L3P, L44, L4P, L6S, L6T, LA, LAC, LAG, LAH, LAI, LAK, LAO, LAT, LB2, LBS, LBT, LCN, LCO, LCP, LDA, LDY, LEC, LER, LEU, LFA, LFC, LFR, LGC, LGU, LHG, LI, LI1, LIO, LKA, LKS, LM2, LMG, LMO, LMR, LMT, LMU, LNK, LNV, LOG, LOX, LPP, LRH, LTG, LU, LUT, LVO, LVZ, LXB, LXC, LXZ, LYS, LZ0, M1F, M1P, M2F, M2M, M3M, M3N, M55, M6D, M6P, M7B, M7P, M8C, MA1, MA2, MA3, MA4, MA8, MAB, MAC, MAE, MAF, MAG, MAL, MAN, MAT, MAV, MAW, MB3, MBE, MBF, MBG, MBN, MBO, MBR, MC3, MCH, MCR, MCU, MDA, MDD, MDP, ME2, MEE, MEG, MES, MET, MFB, MFU, MG, MG5, MG8, MGC, MGE, MGF, MGL, MGS, MH2, MH3, MHA, MHM, MJJ, MLA, MLB, MLI, MLP, MLR, MLT, MM4, MMA, MMC, MN, MN0, MN3, MN5, MN6, MNA, MNC, MNH, MNR, MO, MO1, MO2, MO3, MO4, MO5, MO6, MO7, MOH, MOO, MOS, MOW, MP1, MPD, MPO, MPR, MPS, MQD, MQG, MQT, MRD, MRH, MRP, MRY, MSE, MSF, MSM, MSX, MTD, MTE, MTF, MTG, MTL, MTO, MTT, MUB, MUR, MVC, MVP, MW1, MW2, MW3, MXE, MXY, MXZ, MYG, MYQ, MYR, MYS, MYY, N, N1L, N2O, N3U, N4M, N8E, N9S, NA, NA1, NA2, NA5, NA6, NAA, NAD, NAG, NAO, NAP, NAW, NBG, NBN, NBX, NBY, NBZ, NC, NCO, NCP, ND4, NDG, NDP, NEH, NER, NET, NEX, NFB, NFC, NFE, NFG, NFO, NFR, NFS, NG1, NG6, NGA, NGC, NGD, NGE, NGK, NGN, NGR, NGS, NGY, NGZ, NH2, NH3, NH4, NHE, NHF, NHY, NI, NI1, NI2, NI3, NIK, NLC, NM6, NM9, NME, NML, NMO, NMU, NNG, NO, NO2, NO3, NOB, NOE, NONE, NPF, NRU, NSP, NSQ, NT1, NTF, NTO, NTP, NVP, NXD, NYT, O, O2, O4B, O4M, O81, OAA, OAK, OC1, OC2, OC3, OC4, OC5, OC6, OC7, OC8, OCL, OCM, OCN, OCO, OCT, OEC, OES, OF0, OF1, OF2, OF3, OGA, OH, OHE, OI7, OLA, OLB, OLC, OMB, OME, OMO, OPC, OPM, OPY, ORO, OS, OS4, OSM, OSU, OTE, OTG, OTN, OTU, OUT, OX, OX2, OXA, OXE, OXL, OXM, OXY, P15, P10, P22, P25, P2K, P33, P3G, P4C, P4G, P4K, P53, P6G, P6P, P8E, PA1, PA8, PAM, PAV, PB, PBM, PC1, PC2, PC3, PC4, PC6, PC7, PC8, PCF, PCL, PCW, PD, PD7, PDO, PDT, PDV, PDX, PE3, PE4, PE5, PE7, PE8, PEE, PEF, PEG, PEH, PEJ, PEK, PEO, PEP, PER, PEU, PEV, PEX, PG0, PG4, PG5, PG6, PG8, PGE, PGF, PGL, PG0, PGQ, PGR, PGT, PGV, PH5, PHB, PHE, PHQ, PI, PIG, PIN, PIS, PKM, PL9, PLC, PLD, PLM, PLP, PMS, PNA, PNG, PNJ, PNL, PNW, PO2, PO3, PO4, POL, PON, POP, POR, POV, PP9, PPC, PPF, PPI, PPK, PPM, PPS, PPV, PQ9, PQN, PQQ, PR, PRO, PRP, PS5, PSC, PSG, PSL, PSV, PT, PT4, PTL, PTN, PTQ, PTY, PUF, PUT, PVO, PX2, PX4, PX6, PX8, PYJ, PYR, PZU, QDK, QIF, QKH, QPS, QV4, R16, R1P, R1X, R2B, R2G, RAE, RAF, RAM, RAO, RB, RB5, RBL, RCD, RE, REO, RER, RF5, RG1, RGG, RGI, RH, RH3, RHA, RHC, RHD, RI2, RIB, RIP, RM4, RMO, ROP, RP3, RP5, RP6, RR7, RRJ, RRY, RST, RTC, RTG, RTV, RU, RU7, RUG, RUU, RV7, RVG, RVM, RVP, RWI, RY7, RZM, S, S0H, S7P, S81, SA0, SAH, SAL, SAM, SAR, SB, SBE, SBO, SBT, SBY, SCG, SCN, SCR, SDS, SDY, SE, SE4, SEA, SEJ, SEK, SEO, SEP, SER, SF3, SF4, SF6, SF9, SFL, SFN, SFO, SFU, SG4, SG5, SG6, SG7, SGA, SGC, SGD, SGM, SGN, SHB, SHD, SHG, SIA, SID, SIN, SIO, SIZ, SLB, SLM, SLT, SM, SMD, SMO, SN5, SNG, SO2, SO3, SO4, SOE, SOG, SOH, SOL, SOM, SOR, SPD, SPH, SPJ, SPK, SPM, SPW, SPZ, SQD, SQU, SR, SR1, SRM, SRT, SSG, SSH, STE, STO, STU, STW, STZ, SUC, SUL, SUP, SUS, SWE, SX, SYL, SZZ, T1A, T3A, T68, T6D, T6P, T6T, TA0, TA6, TAG, TAM, TAR, TAS, TAU, TB, TBA, TBR, TBU, TCA, TCB, TCH, TCN, TDG, TDP, TE, TEA, TEE, TEP, TEU, TF0, TF4, TFA, TFH, TFP, TFU, TG1, TGA, TGK, TGL, TGR, TGY, TH, TH1, THB, THE, THJ, THM, THR, TL, TLA, TM5, TM6, TMA, TME, TMP, TMR, TMX, TNX, TOA, TOC, TOE, TOU, TP7, TPO, TPP, TQY, TRA, TRD, TRE, TRP, TRS, TRT, TRV, TS8, TSD, TSE, TSM, TT7, TTP, TTV, TU4, TUG, TUJ, TUP, TUR, TVD, TVG, TVM, TVS, TVV, TVY, TW7, TWA, TWD, TWG, TWJ, TWT, TWY, TXB, TYR, TYS, TYV, TZZ, U, U1, U10, U1Y, U2A, U2D, U63, U6J, U71, U8V, U97, U9A, U9D, U9G, U9J, U9M, UAP, UBH, UBO, UDC, UDP, UEA, UEV, ULI, UMP, UMQ, UND, UNK, UNL, UNX, UPL, URE, UTP, UVW, V, V3M, V3P, V40, V70, V71, VA3, VAL, VER, VG1, VJ1, VJ4, VKN, VN3, VO3, VO4, VSO, VTB, VX, VXA, W, W9T, WAT, WCC, WIA, WO2, WO4, WO5, WOO, WUN, WZ1, WZ2, X0X, X1P, X1X, X2F, X2Y, X34, X6X, X6Y, XAT, XCC, XDX, XE, XGP, XIL, XKJ, XL1, XLF, XLS, XMM, XMP, XP4, XPA, XPE, XS2, XXM, XXR, XXX, XYF, XYL, XYP, XYS, XYT, XYZ, Y1, YAN, YB, YB2, YBT, YDR, YF3, YH, YIO, YJM, YKR, Y05, YT3, YUT, YX0, YX1, YYB, YYH, YYJ, YYK, YYM, YYQ, YZ0, Z0F, Z15, Z16, Z2D, Z2T, Z3K, Z3L, Z3Q, Z3U, Z4K, Z4R, Z4S, Z4U, Z4V, Z4W, Z4Y, Z57, Z5J, Z5L, Z61, Z6H, Z6J, Z6W, Z8H, Z8T, Z9D, Z9E, Z9H, Z9K, Z9L, Z9M, Z9N, Z9W, ZB0, ZB1, ZB2, ZB3, ZCD, ZCM, ZCZ, ZD0, ZDC, ZDO, ZEE, ZEL, ZEM, ZGE, ZH3, ZM2, ZMR, ZN, ZN1, ZN2, ZN3, ZNH, ZNO, ZO3, ZOW, ZR, ZRC

D.2 Interaction comparison and interaction statistics

Table D.2: Interaction summary statistics of each interaction type for the 25 924 ligand poses in all of the 13 657 PDB structures from which the pairs were assembled. Interactions were measured as described in the methods using ProLIF (Bouysset & Fiorucci, 2021). For each interaction, descriptive statistics are reported including the mean frequency, standard deviation, and quartiles (minimum, first quartile, median, third quartile, and maximum). For example, there are $2.29 + 3.82 = 6.11$ hydrogen bonds per protein-ligand interaction on average in the data set, with zero being the lowest and $24 + 15 = 39$ being the highest number of H-bonds observed in a single structure.

Interaction	Mean	Std.Dev.	Min	Q1	Q2	Q3	Max
Anionic	0.10	0.93	0	0	0	0	31
Cationic	0.04	0.37	0	0	0	0	14
Cation- π	0.01	0.15	0	0	0	0	4
H-Bond Acceptor	3.82	3.42	0	1	3	6	24
H-Bond Donor	2.29	2.02	0	1	2	4	15
Hydrophobic	9.99	13.00	0	0	5	15	114
π -Cation	0.13	0.52	0	0	0	0	8
π -Stacking	0.38	0.68	0	0	0	1	6
Van der Waals	44.86	23.80	0	27	39	57	182

(a) Number of interactions over all structures.



(b) Number of structures where each interaction is present.

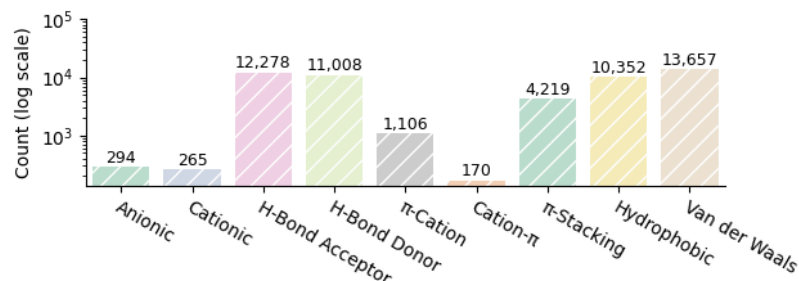
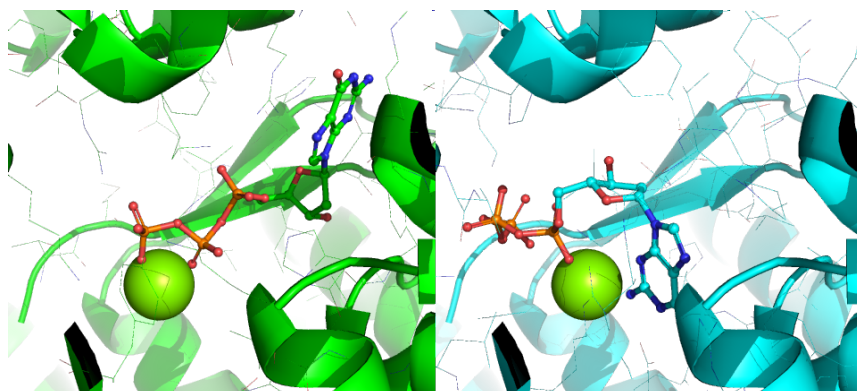


Figure D.1: Interaction counts in the 25 924 ligand poses in the 13 657 PDB structures from which the pairs were assembled in terms of (a) the total number of interactions over all structures, and (b) the number of structures with at least one of each type of interactions. Overall, van der Waals and hydrophobic interactions are the most prevalent, followed by hydrogen bonds and π -cation stacking.

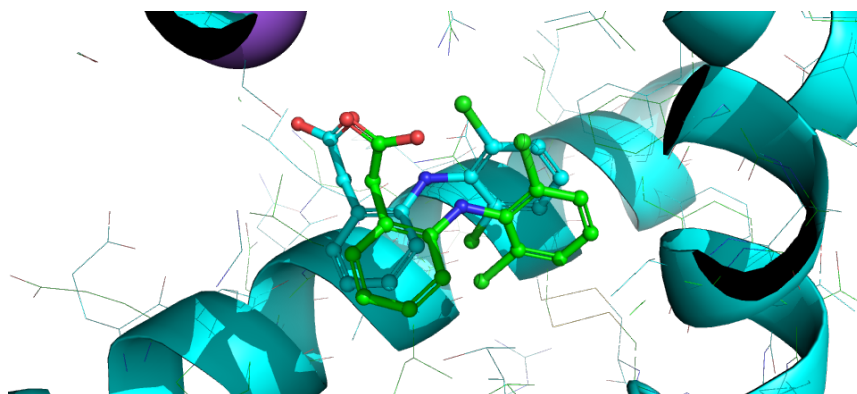
Table D.1: Comparison of explicit interaction types detected by different interaction profilers. Excluded from the list are covalent interactions, water bridges with degree greater than one, halogen-water interactions, weak hydrogen bonds, and other interactions involving metals. The ticks reflect ProLIF (Bouysset & Fiorucci, 2021) v2.0.3, PLIP (Salentin et al., 2015; Schake et al., 2025) v3.0.0, Arpeggio (Jubb et al., 2017) v2015, and SILIRID (Chupakhin et al., 2014) from the open drug development toolkit v0.5 (Wójcikowski et al., 2015). SPLIF (Da & Kireev, 2014) and PLEC (Wójcikowski et al., 2019) do not explicitly encode interaction types; they use a hashing procedure similar based on extended connectivity fingerprints (Rogers & Hahn, 2010) and are not shown in the table.

Interaction	ProLIF	PLIP	Arpeggio	SILIRID
hydrogen bond	✓	✓	✓	✓
salt bridge	✓	✓	✓	✓
π -stacking	✓	✓	✗	✓
π -cation interaction	✓	✓	✓	✓
halogen bond	✓	✓	✓	✓
metal coordination	✓	✓	✓	✓
hydrophobic	✓	✓	✓	✓
van der Waals contact	✓	✗	✓	✗
water bridge (degree 1)	✓	✓	✓	✗

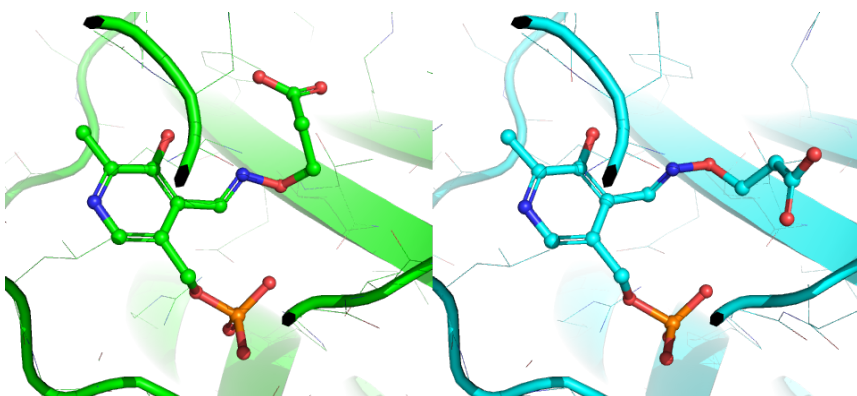
D.4 Illustrations of conserved and non-conserved pairs of poses



(a) Ligand DGT in PDB structures 4TNX (left, green) and 4TO6 (right, blue). CS_{RDKit} is 0.15 and ECFP4 2048-bit Tanimoto is 1.0.

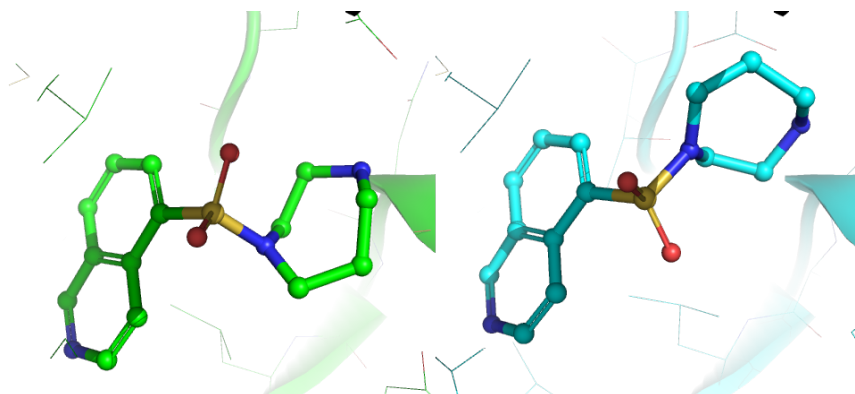


(b) Ligand DIF in PDB structures 6HNO (green) and 6HN1 (blue). CS_{RDKit} is 0.34 and ECFP4 2048-bit Tanimoto is 1.0.

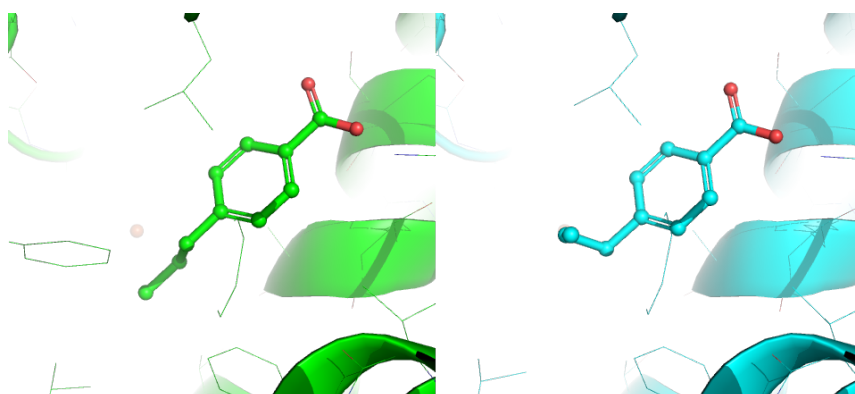


(c) Ligand OCF in PDB structures 8AYJ (left, green) and 8ONN (right, blue). CS_{RDKit} is 0.68 and ECFP4 2048-bit Tanimoto is 1.0.

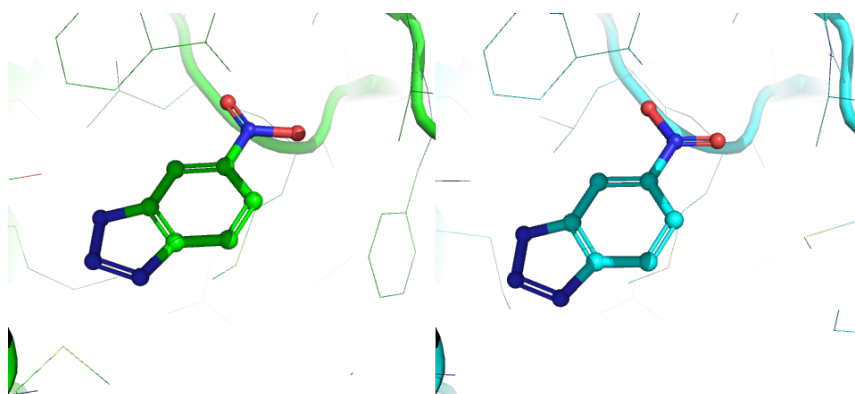
Figure D.3: Pairs of identical (same CCD code) ligands bound to the same binding site in different PDB structures with non-conserved binding poses ($CS_{\text{RDKit}} < 0.8$) and ECFP4 2048-bit Tanimoto is 1.0.



(a) Ligand M77 in PDB structures 2GNI (left, green) and 5LCP (right, blue). CS_{RDKit} is 0.45 and ECFP4 2048-bit Tanimoto is 0.67.



(b) Ligand 8ZU in PDB structures 7R8S (left, green) and 7UDF (right, blue). CS_{RDKit} is 0.78 and ECFP4 2048-bit Tanimoto is 0.67.



(c) Ligand 6NT in PDB structures 5RGA (left, green) and 8USH (right, blue). CS_{RDKit} is 0.70 and ECFP4 2048-bit Tanimoto is 0.70.

Figure D.4: Pairs of identical (same CCD code) ligands bound to the same binding site in different PDB structures with non-conserved binding poses ($SC_{\text{RDKit}} < 0.8$) and ECFP4 2048-bit Tanimoto less than 1.0.

Appendix E

Technical preliminaries and further results for Chapter 5

E.1 Score-based diffusion models

A score-based diffusion model S trained by score-matching approximates the score of a noisy data distribution at different noise levels. Written mathematically, that is $S(x_t, t) \approx \nabla_{x_t} \log p(x_0 | x_t)$. The noisy data is typically generated by adding isotropic Gaussian noise according to a noise schedule $\sigma^2(t)$ and the score-matching objective used in training is

$$\mathbb{E}_{\text{data}} \mathbb{E}_{t \sim \mathcal{U}(0,1)} \arg \min_{\theta} \mathcal{L}_{\theta} = \arg \min_{\theta} \mathbb{E} \left\| \nabla_{x_t} \log p(x_0 | x_t) - S_{\theta}(x_t, t) \right\|^2 \quad (\text{E.1})$$

where $x_t = x_0 + \sigma(t)\varepsilon$ with $\varepsilon \sim \mathcal{N}(0, 1)$ (Ho et al., 2020; Y. Song & Ermon, 2019; Y. Song et al., 2021). Inference is performed by applying the score model to a Gaussian reference distribution repeatedly and re-adding less and less Gaussian noise. This is a discretization of the reverse diffusion process, with the score model in place of the score:

$$x_{n-1} = x_n + \Delta t \cdot g(n \cdot \Delta t)^2 \cdot S_{\theta}(x_n, t) + g(n \cdot \Delta t) \cdot \sqrt{\Delta t} \cdot z_n \quad (\text{E.2})$$

where $z_n \sim \mathcal{N}(0, 1)$.

E.2 DiffDock-L benchmark

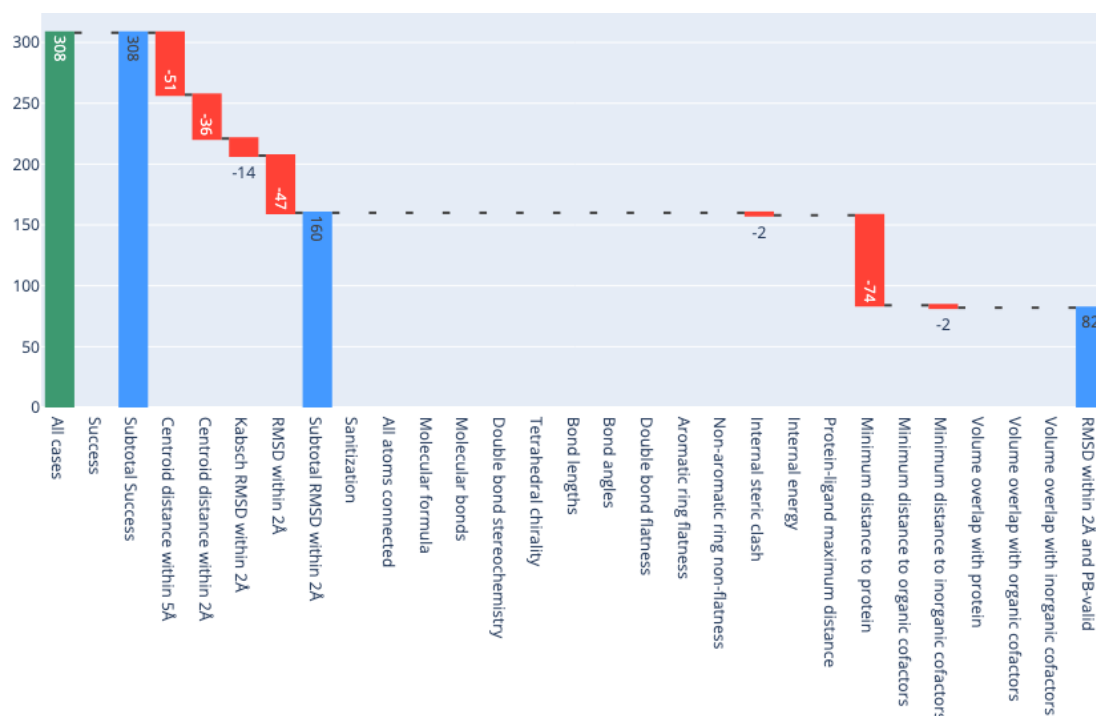


Figure E.1: Waterfall plot showing the PoseBusters tests as filters for DiffDock-L (v1.1.3) predictions on the PoseBusters data set. The performance of DiffDock-L (27 % of poses within 2 Å RMSD and PB-valid) is more than double that of DiffDock (12 % of poses within 2 Å RMSD and PB-valid) which I benchmarked in Chapter 2.

E.3 Further experimental results

Table E.2: Re-docking the ligand into the cognate rigid receptor for the PoseBusters data set (n=308) comparing the performance of forward and backward guidance. The numbers are percentages and the confidence intervals are 95 % bootstrap estimates. The two guidance formulations perform similarly well.

	RMSD $\leq$ 2 Å	PB-valid	RMSD $\leq$ 2 Å and PB-valid
GuideDock: simple physics (reverse guidance)	50.6 $\pm$ 2.9	78.2 $\pm$ 2.4	45.1 $\pm$ 2.8
GuideDock: simple physics (forward guidance)	52.3 $\pm$ 2.9	78.9 $\pm$ 2.3	45.8 $\pm$ 2.9
GuideDock: centroid and simple physics (reverse guidance)	62.3 $\pm$ 2.8	91.6 $\pm$ 1.6	59.7 $\pm$ 2.8
GuideDock: centroid and simple physics (forward guidance)	63.6 $\pm$ 2.8	91.9 $\pm$ 1.6	59.7 $\pm$ 2.8

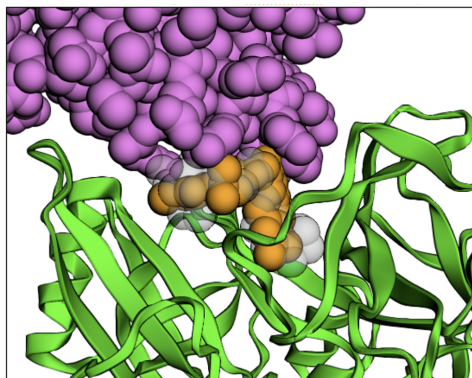
Table E.1: Re-docking the ligand into the cognate rigid receptor for the PoseBusters data set (n=308) testing different constraints for GOLD. The numbers are percentages and the confidence intervals are 95 % bootstrap estimates. The best constraints use a 2.0 Å radius around each constraint atom.

	RMSD ≤ 2 Å	PB-valid	RMSD ≤ 2 Å and PB-valid
GOLD (unconstrained)	58.3 $\pm$ 4.1	92.4 $\pm$ 2.2	55.6 $\pm$ 4.2
GOLD: fragment region constraints 0.5 Å radius	70.7 $\pm$ 3.8	88.4 $\pm$ 2.7	65.3 $\pm$ 3.9
GOLD: fragment region constraints 1.0 Å radius	70.7 $\pm$ 3.8	88.4 $\pm$ 2.6	65.3 $\pm$ 3.9
GOLD: fragment region constraints 2.0 Å radius	77.6 $\pm$ 3.5	89.8 $\pm$ 2.5	72.8 $\pm$ 3.6
GOLD: fragment region constraints 3.0 Å radius	76.9 $\pm$ 3.4	89.1 $\pm$ 2.5	71.4 $\pm$ 3.8
Crystal structure	100.0 $\pm$ 0.0	66.9 $\pm$ 3.8	66.9 $\pm$ 3.8

Table E.3: Docking the larger ligand into the smaller ligand’s receptor for the binding mode preserving pairs of the binding pose preserving pairs of the Malhotra data set (n=256). The RMSD is calculated relative to the large ligand aligned via the protein chains. The numbers are percentages and the confidence intervals are 95 % bootstrap estimates.

	RMSD ≤ 2 Å	PB-valid	RMSD ≤ 2 Å and PB-valid
DiffDock (no guidance)	79.7 $\pm$ 2.5	32.8 $\pm$ 2.9	27.7 $\pm$ 2.8
GuideDock: simple physics	66.8 $\pm$ 3.0	81.2 $\pm$ 2.5	58.2 $\pm$ 3.1
GuideDock: pocket & simple physics	68.4 $\pm$ 2.9	80.9 $\pm$ 2.5	60.2 $\pm$ 3.1
GuideDock: fragment & simple physics	71.1 $\pm$ 2.9	67.6 $\pm$ 2.9	53.1 $\pm$ 3.1
GOLD: no constrains	20.3 $\pm$ 2.5	86.7 $\pm$ 2.1	16.0 $\pm$ 2.3
GOLD: fragment region constrains	41.8 $\pm$ 3.1	79.7 $\pm$ 2.5	33.6 $\pm$ 3.0
Alignment via protein chains	100.0 $\pm$ 0.0	41.0 $\pm$ 3.1	41.0 $\pm$ 3.1

(a) Compound docking within 2 Å of Protein X-CRBN complex



(b) Docking performance on BMS CRBN ternary structures

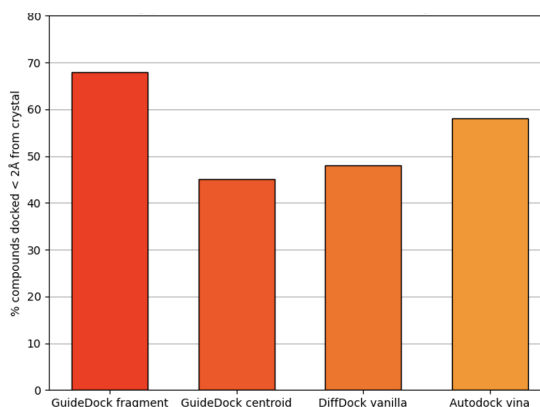


Figure E.2: Figure created by my industrial partner Giorgio Tamo. (a) Placeholder showing the GuideDock pose (RMSD within 2 Å) of a molecular glue bound at the interface of protein X and CRBN. (b) Overall docking success rate of using GuideDock to constrain the warhead substructure, GuideDock with a pocket constraint, DiffDock, and AutoDock Vina. The name of the ligand and identity of protein X will be revealed once the manuscript of our industrial partner describing this particular complex has been accepted or the structure is published in the PDB. GuideDock with fragment constraints performed best overall.

E.4 Further details on data sets

E.4.1 Selection of pairs of PDB structures for the GuideDock data set

The criteria for the high-quality protein-ligand complexes were:

- Non-obsolete PDB entries containing a protein and ‘ligand of interest’ released from 1 January 2000 to 31 March 2025.
- The structure is X-ray crystallography determined with a higher resolution upper limit of at most 2 Å, a ligand real space R-factor of at most 0.2, a real space correlation coefficient of at least 0.95, a model completeness of 100%, and no reported ligand stereochemical errors.
- The protein has an amino acid chain of at least 30 residues.
- The ligand has between 3 and 55 heavy atoms and must have at least one carbon atom, may only contain the elements H, C, O, N, F, S, Cl, P, Br, I, and B, and is not covalently bonded to any other entity.

The criteria for the pairs of protein-ligand complexes were as follows:

- Both novel and template ligands are bound to protein chains that are in the same PDB protein chain cluster (identity threshold 90%), that is, their sequence identity is at least 90%.
- The novel ligand was first added to the PDB in 2022 or later and has a quantitative estimate of drug-likeness score (Bickerton et al., 2012) of at least 0.7.
- The ligands share a common substructure with at least 3 paired atoms within 2 Å of each other.

We further filter the list of pairs by requiring that each PDB protein chain sequence cluster (sequence identity threshold 70%) may occur at most 3 times in the data set. The pairs were aligned using the same procedure as for the Malhotra data set, and when multiple biological assemblies or chain alignment combinations were available, the pair with the highest colour and shape overlap was selected. A list of the 148 resultant paired PDB and CCD identifiers is provided in Section E.4.5.

E.4.2 Malhotra data set updates

Table E.4: Updates to the PDB and CCD identifiers of the protein-ligand complexes reported in the supplementary information provided by Malhotra and Karanicolas (2017).

Previous		Updated	
2WIR	WIR	2WIR	ZZ3
3ORN	3OR	3ORN	3OR
1YHS	STO	1YHS	STU
3NNV	MOL	3NNV	437
4FK6	4AL	4FK6	OUJ
2XPB	4GF	2XPB	4GE
3BG8	INH	3BG8	GOR
4F08	JAK	4F08	1RS
1HXK	DDM	1HXK	DMJ
1C5P	BAM	1C5P	BEN
1C5Z	BAM	1C5Z	BEN
3V0Y	STL	4QER	STL
3GEV	SAQ	4Q4Q	SAQ

E.4.3 DockGen data set updates

The following updates to the DockGen data set (Corso et al., 2024) were made to ensure ligands suitable for docking with classical docking methods:

- Remove two entries from the test set that are included in both the test and validation set: 6070_ACK, 6XB3_9BG.
- Rename 3S5X_BMA-MAN-MAN-MAN-MAN to 3S5X_BMA-MAN-MAN-MAN to match the ligand in the complex.
- Remove two complexes containing covalent ligands: 2BPB_MSS, 2WDY_COA.
- Remove complexes with ligands containing rare elements: 1SIJ_PCD, 2O5M_MNR, 5U82_ZN0, 6A71_9UX, 6A72_9UX.
- Remove ligands with less than 5 heavy atoms.
- Remove ligands with a weight of less than 200 Daltons.
- Remove ligands with more than 20 rotatable bonds.
- Remove CCD IDs blocklist for ligands (fatty acids): 2ZE9_PD7, 4G0S_MYR, 4X9X_OLA, 5EZU_MYR, 5F2T_PLM, 5W00_MYR.
- Get complexes from PDB. Choose the same chains as in the DockGen receptor files (pick new chains for 4Z07_GLC-BGC). Choose chains near ligand in DockGen ligand files (add symmetry chains present in Dockgen receptor files if near ligand). Create new ligand and receptor files with cofactors and missing chains.
- Remove crystal contact ligands (SILIRID interactions (Chupakhin et al., 2014) with protein crystal chains not present in the receptor).
- The number of interactions (SILIRID) of any kind between the ligand and the receptor is at least 3.
- Drop complexes with ligands with stereochemical errors (as reported by the PDB): 1HW8_114, 1UF8_ING, 3NFD_COA, 3O02_JN3, 5FIU_Y3J, 6EGU_43Y.
- Drop complexes with incomplete ligand models (as reported by the PDB): 2XTA_ACO, 3IQF_E4M, 3QE2_NAP, 3S6A_ANP, 4Q05_NAG, 4Y7C_NAP, 4Y9R_NAP, 5FA6_NAP, 6N19_K8V, 6RMS_AMP.

E.4.4 Malhotra data set identifiers

The updated Malhotra data set contains the following 256 binding-pose-preserving pairs of protein-ligand complexes from the PDB. The keys are constructed such that from left to right, the first PDB identifier (even number of characters) is the complex containing the ligand that is to be docked and which is indicated by the first CCD identifier (odd number of characters), and the second PDB identifier contains the receptor that is to be docked into, and also constrains the template binding pose in the ligand that has the second CCD identifier.

E.4.5 GuideDock data set identifiers

The GuideDock data set contains 148 pairs of protein-ligand complexes from the PDB. The identifiers are constructed as for the Malhotra data set. The keys are constructed as for the updated Malhotra data set.

E.4.6 Ligand block list

The block list comprises the same 1,989 CCD identifiers listed in Section D.1.



Figure E.3: Molecular graphs of all the ligands to be docked in the GuideDock data set. These were chosen to have $\text{QED} \geq 0.7$ and a first deposition date in a PDB structure since 2021.