Machine Learning for Molecular Modelling and Drug Discovery

Lucy Vost

Green Templeton College
University of Oxford

A thesis submitted for the degree of
Doctor of Philosophy

October 2025

Abstract

The process of drug design is notoriously expensive, time-consuming, and inefficient. A central challenge is our limited understanding of the complex function governing the binding between small molecules and proteins, a critical factor for both the design and screening of potential drugs. In recent years, machine learning (ML) has shown promise in extracting complex patterns from data, leading to its application to drug design. This thesis explores ML-driven approaches applied within the early stages of the drug design pipeline.

The first part of this thesis confronts a key challenge in virtual screening: the tendency for models to exploit dataset biases rather than learn the principles of molecular binding. In response, we developed a virtual screening tool designed to be as debiased as possible. Rather than relying on default validation setups prone to leakage, we conducted interpretability experiments, proposing that a desirable model should not only identify high-affinity binders but also pinpoint the functional groups responsible for binding. Our interpretability analysis revealed that the model could indeed identify key groups, a finding corroborated by a physics-based interaction profiler. We further validated this by applying the method to fragment screen data, successfully extracting binding 'hotspots': regions of the pocket which contribute disproportionately to binding. These hotspots were then used to inform fragment elaboration, leading to the generation of promising new binders.

Next, we turn to the more fundamental task of pocket-free, 3D *de novo* molecule generation. While diffusion models have shown promise at this, their performance is often limited to very small molecules rather than larger, drug-like structures. A common technique is property conditioning, where models are trained to generate molecules with specified chemical properties. We experimented with an alternative approach: conditioning on molecule quality. By augmenting the training dataset with synthetic, distorted, low-quality molecules and corresponding labels, we could conditionally sample from the high-quality region of the chemical space. We found this method generated a significantly higher proportion of physically plausible molecules.

We then investigated another method for improving the physical plausibility of generated molecules by modifying the diffusion implementation itself. We replaced the standard atomic diffusion approach with a substructural block-based method. The theoretical benefits are twofold: assembling a molecule from known substructures is a simpler problem than assembling it from individual atoms, and the method could be trained using a catalogue of commercially available substructures to ensure generated molecules are readily synthesisable. However, we encountered implementation challenges, particularly in defining the 3D arrangement of non-uniform blocks in an information-rich and consistent manner. We provide an overview of the strategies we tested and a brief review of subsequent methods that have successfully implemented similar ideas.

Finally, we explore a method for cryo-EM model building that integrates the experimental map within the AlphaFold prediction process. By conditioning the prediction with the map from the outset, we guide the model toward the experimentally observed conformation. Building on recent studies like ROCKET and CryoBoltz, we present an exploration of conditioning strategies using both an AlphaFold 2-based model (OpenFold) and an AlphaFold 3-based model (Boltz-2). We found that interventions late in the inference pipeline (specifically, in the structure module) were insufficient to induce large-scale conformational changes. Inspired by these findings, we introduced an MSA subsampling strategy to inject stochasticity into the Boltz-2 model. This method enabled the model to escape the local minima of its unconstrained prediction and sample alternative conformations that were more consistent with the provided cryo-EM data.

Machine Learning for Molecular Modelling and Drug Discovery



Lucy Vost

Green Templeton College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

October 2025

Dedicated to my grandma, Sheila.
You will never read this—and let's
be honest, you wouldn't have anyway.
I miss you every day.

Acknowledgements

It is with massive relief and gratitude that I acknowledge the people who made this thesis possible. First, my supervisors. To my industry supervisors, Payel and Vijil, thank you for your guidance and kindness. To my main supervisor, Charlotte, thank you for your belief in me, encouragement, and patience; for being as on board for discussing nonsense (eating worms or rubbers, and the scent profiles of various politicians come to mind) as actual science; and giving me definitive proof that you can be extremely successful but still up for a laugh.

Speaking of laughing, I don't think a single day of work in this PhD went by without a giggle. For that, I thank my fantastic, intelligent, and simultaneously completely daft fellow opiglets. To the alumni who showed me the ropes—Tom, Eve, Sarah, Brennan, Tobias, Anna, Alissa, Lewis, Leo, Ruben, and Matteo F—thank you for making stats such a happy place to be. To my greatest of pals in the Cube—Henriette, Isaac, and Gemma—thank you for the support, daily camaraderie, and the distractions that kept me sane. My enormous gratitude also goes to Kate, Yael, Isak, Fabian, Guy, Ollie T, Nik, Nele, Sam, Matteo C, Fergus, Matt, Nick, Henry, King, Sanaz, Eoin and Ben for all being the goodest of eggs. I feel very lucky to know you all!

I am also indebted to the friends who provided much-needed perspective and fun. To Ellie and Ellen, thank you for the crosswords, venting sessions, book recs, and sweet treats. To Stef, Raphael, and Pritha, thank you for letting me tag along with you on the best holidays I've ever been on. To Stephi, thank you for being a constant source of enthusiasm and for referring to me exclusively as 'Dr Vost' long before I had even started. To my assorted pals from undergrad, thank you for the continued friendship, with particular shoutouts to Megan, for introducing me to several musicals that were the soundtrack for the writing of this beast; Noske, my indispensable computery guy; Bee, for being almost as encouraging as your film choices are nightmarish; and Mario, for just being a general dreamboat.

Finally, I have enormous thanks to give to my family. To my grandma, Sheila, an icon and a legend, whose weekly calls helped me preserve something of a northern accent. To my grandad, Trevor, from whom I inherited the sense of humour I needed to get through some of the blunders I've made in the past four years of research. To Alice, thank you for being the first person in history to find said humour funny, and for the choccy squares. To Chris and Emma, thank you for the counsel and pizza, but mostly for raising my two favourite little people, Charlie and Izzy, whose landing on Earth has been easily the highlight of the past four years. To my parents: thank you for the therapy lunches, keeping the voices at bay, and for never once asking when I was going to get a real job. Finally, to my partner, Chris. Thank you for running me baths, getting me outdoors, making me pad thai, and for reminding me at every deadline and crisis that it was all going to be ok. And it is!

Abstract

The process of drug design is notoriously expensive, time-consuming, and inefficient. A central challenge is our limited understanding of the complex function governing the binding between small molecules and proteins, a critical factor for both the design and screening of potential drugs. In recent years, machine learning (ML) has shown promise in extracting complex patterns from data, leading to its application to drug design. This thesis explores ML-driven approaches applied within the early stages of the drug design pipeline.

The first part of this thesis confronts a key challenge in virtual screening: the tendency for models to exploit dataset biases rather than learn the principles of molecular binding. In response, we developed a virtual screening tool designed to be as debiased as possible. Rather than relying on default validation setups prone to leakage, we conducted interpretability experiments, proposing that a desirable model should not only identify high-affinity binders but also pinpoint the functional groups responsible for binding. Our interpretability analysis revealed that the model could indeed identify key groups, a finding corroborated by a physics-based interaction profiler. We further validated this by applying the method to fragment screen data, successfully extracting binding 'hotspots': regions of the pocket which contribute disproportionately to binding. These hotspots were then used to inform fragment elaboration, leading to the generation of promising new binders.

Next, we turn to the more fundamental task of pocket-free, 3D *de novo* molecule generation. While diffusion models have shown promise at this, their performance is often limited to very small molecules rather than larger, drug-like structures. A common technique is property conditioning, where models are trained to generate molecules with specified chemical properties. We experimented with an alternative approach: conditioning on molecule quality. By augmenting the training dataset with synthetic, distorted, low-quality molecules and corresponding labels, we could conditionally sample from the high-quality region of the chemical space. We found this method generated a significantly higher proportion of physically plausible molecules.

We then investigated another method for improving the physical plausibility of generated molecules by modifying the diffusion implementation itself. We replaced the standard atomic diffusion approach with a substructural block-based

method. The theoretical benefits are twofold: assembling a molecule from known substructures is a simpler problem than assembling it from individual atoms, and the method could be trained using a catalogue of commercially available substructures to ensure generated molecules are readily synthesisable. However, we encountered implementation challenges, particularly in defining the 3D arrangement of non-uniform blocks in an information-rich and consistent manner. We provide an overview of the strategies we tested and a brief review of subsequent methods that have successfully implemented similar ideas.

Finally, we explore a method for cryo-EM model building that integrates the experimental map within the AlphaFold prediction process. By conditioning the prediction with the map from the outset, we guide the model toward the experimentally observed conformation. Building on recent studies like ROCKET and CryoBoltz, we present an exploration of conditioning strategies using both an AlphaFold 2-based model (OpenFold) and an AlphaFold 3-based model (Boltz-2). We found that interventions late in the inference pipeline (specifically, in the structure module) were insufficient to induce large-scale conformational changes. Inspired by these findings, we introduced an MSA subsampling strategy to inject stochasticity into the Boltz-2 model. This method enabled the model to escape the local minima of its unconstrained prediction and sample alternative conformations that were more consistent with the provided cryo-EM data.

Contents

List of Figures **xiii**

1 Introduction **1**
1.1 Background and motivation . . . 1
1.2 Drug discovery . . . 6
1.2.1 Small molecule drugs . . . 6
1.2.2 The drug discovery pipeline . . . 7
1.2.2.1 Target identification, validation, and structural characterisation . . . 7
Cryo-EM . . . 8
1.2.2.2 Hit identification . . . 8
1.2.2.3 Hit-to-lead and lead optimisation . . . 10
1.3 Machine learning . . . 11
1.3.1 Supervised learning . . . 12
1.3.2 Neural networks and deep learning . . . 13
1.3.3 Convolutional neural networks . . . 14
1.3.4 Graph neural networks . . . 15
1.3.5 Diffusion models . . . 16
1.4 Machine learning in early-stage drug discovery . . . 17
1.4.1 Target structure resolution . . . 18
1.4.1.1 *In silico* structure prediction . . . 18
AlphaFold2 . . . 19
AlphaFold3 . . . 19
Open-source models and extensions . . . 20
1.4.1.2 Cryo-EM model building . . . 21
Direct map interpretation methods . . . 22
AlphaFold-based methods . . . 23
1.4.2 Virtual screening . . . 24
1.4.3 Generative design . . . 26
1.4.3.1 Unconditional generation . . . 27
1.4.3.2 Conditional *de novo* generation . . . 27
1.4.4 Fragment-based design . . . 29
1.5 Thesis outline . . . 30

2 Fragment elaboration with attribution-informed constraints 33
2.1 Preface . . . 34
2.2 Contributions . . . 35
2.3 Introduction . . . 35
2.4 Methods . . . 37
2.4.1 Model . . . 37
2.4.2 Datasets . . . 38
2.4.2.1 The Redocked set . . . 40
2.4.2.2 PDBBind refined: gninaSet$_{Pose}$ set . . . 40
2.4.2.3 PDBBind General and Core sets . . . 41
2.4.3 Training dataset filtering . . . 41
2.4.4 Attribution . . . 42
2.4.5 Fragment elaboration . . . 45
2.4.5.1 Hotspot Extraction . . . 45
2.4.5.2 Protein targets . . . 45
2.4.5.3 Obtaining fragments . . . 46
2.4.5.4 STRIFE: A generative model for elaboration . . . 46
2.4.5.5 Assessing generated molecules . . . 48
2.5 Results . . . 48
2.5.1 Training and testing PointVS . . . 48
2.5.1.1 Bias in the CASF16 test set . . . 48
2.5.1.2 Performance as a scoring function . . . 49
2.5.2 Attribution: identifying important interactions . . . 52
2.5.2.1 Human Tankyrase-2 inhibitors . . . 53
2.5.2.2 Large-scale attribution tests . . . 56
2.5.3 Hotspot identification using PointVS . . . 57
2.5.3.1 Impact of fragment screen similarity . . . 58
2.5.3.2 Dependency on fragment screen size . . . 60
2.5.4 Fragment elaboration . . . 60
2.6 Conclusions . . . 62

3 Improving Structural Plausibility in 3D Molecule Generation via Property-Conditioned Training with Distorted Molecules 67
3.1 Preface . . . 68
3.2 Introduction . . . 68
3.3 Methods . . . 70
3.3.1 Generation of 3D molecules . . . 70
3.3.2 Conditioning on conformer quality . . . 72
3.3.3 Assessment metrics . . . 73

3.3.4 Datasets . . . 74
3.3.4.1 QM9 . . . 75
3.3.4.2 GEOM . . . 75
3.3.4.3 ZINC . . . 76
3.4 Results and discussion . . . 76
3.4.1 Ablation tests . . . 76
3.4.2 Conditioning on distortion factor . . . 78
3.4.3 Conditioning on internal energy . . . 80
3.4.4 Testing the conditioning method on additional models . . . 82
3.5 Conclusions . . . 83

4 Molecule diffusion with fragment blocks 87
4.1 Preface . . . 87
4.2 Introduction . . . 88
4.3 Methods . . . 90
4.3.1 Model . . . 90
4.3.1.1 Atom-based setup . . . 90
4.3.1.2 Initial model adaptations . . . 91
4.3.2 Data . . . 92
4.3.2.1 Training datasets . . . 92
4.3.2.2 Decomposition . . . 92
Regular BRICS decomposition . . . 92
Decomposition with standardised fragments . . . 93
4.3.2.3 Rotational information . . . 95
4.3.3 Assessment metrics . . . 97
4.4 Results . . . 97
4.4.1 Centroid-only method . . . 97
4.4.2 Plane alignment method . . . 98
4.4.3 Addition of rotational information . . . 99
4.5 Related Work . . . 99
4.6 Conclusions . . . 101

5 Cryo-EM model building with AlphaFold 103
5.1 Preface . . . 103
5.2 Introduction . . . 104
5.3 Methods . . . 106
5.3.1 Calculating and applying gradient-based optimisation . . . 107
5.3.1.1 Alignment . . . 107
5.3.1.2 Simulating a map from the prediction . . . 108
5.3.1.3 Overlap and gradient calculation . . . 108

5.3.1.4 Updating the prediction . . . 108
5.3.1.5 Update scheduling . . . 109
`Block` Scheduling . . . 109
`Global` Scheduling . . . 109
`Block and Global` Scheduling . . . 109
5.3.2 MSA subsampling . . . 110
5.3.3 Datasets and evaluation . . . 110
5.4 Results . . . 111
5.4.1 OpenFold conditioning . . . 111
5.4.1.1 Direct conditioning with a constant scaling factor . 111
5.4.1.2 Introducing gradient scaling schedules . . . 112
5.4.1.3 Implementing iterative updates . . . 115
5.4.2 Boltz-2 conditioning . . . 116
5.4.2.1 Gradient-based optimisation alone . . . 117
5.4.2.2 Adding MSA subsampling . . . 119
Only MSA subsampling . . . 119
Combining with gradient-based optimisation . . . 120
5.5 Conclusions . . . 121

6 Conclusions and outlook 123
6.1 Interpretability in virtual screening . . . 123
6.2 Plausibility in diffusion-based generative design . . . 124
6.3 AlphaFold for cryo-EM model building . . . 125
6.4 Concluding remarks . . . 126

Appendices

A Appendix: chapter 2 129
A.1 Equivariance and Invariance - written by Jack Scantlebury . . . 129
A.2 Hotspots API: calculation and processing of hotspots . . . 130
A.3 Structure IDs . . . 133
A.4 FGFR1 Case Study . . . 133

B Appendix: chapter 3 137
B.1 Ablation test: sampling $D = D_{max}$ . . . 137
B.2 Full PoseBusters outputs . . . 137
B.3 Other metrics . . . 139

References 141

List of Figures

1.1 Examples of drugs of three different types: a small molecule (7MI3, Santarossa et al. 2021), a peptide (6X18, X. Zhang et al. 2020), and an antibody (5DK3, Scapin et al. 2015). The first two are shown bound to their respective targets, shown in grey. 6
1.2 Flowchart showing the different stages of the drug discovery pipeline 7
1.3 A flowchart providing a non-exhaustive overview of the various methods contained within artificial intelligence, with particular focus on the ML branch. 13

2.1 An overall schema of the methods used to debias and test PointVS. We first thoroughly filter the test and train sets (box a), before benchmarking the performance of PointVS on the docking power and scoring power tests (b). We then use attribution to gain insights into important binding regions in the protein pocket (c), which we use for fragment elaboration (d). 39
2.2 Architecture and input format of PointVS. Each of the n atoms in the input to the model (left) is given a one-hot encoded feature vector with a single bit added to indicate whether the atom is from the ligand or the receptor, as well as a position. There are m edges, defined by the edge indices which are the indices of connected atoms, and the corresponding edge attributes. These are one-hot encodings representing ligand-ligand, ligand-protein, and protein-protein edges. There are skip connections between each of the EGNN layers, and the linear and global average final pooling (GAP) layers only act upon the node features f. 40
2.3 Top-N vs N pose ranking performance on the gninaSet$_{Pose}$ set for Autodock Vina, PointVS and gnina. Terms in brackets in the legend refer to the training sets; filtering the training set by protein and ligand similarity results in slightly degraded performance. The Top-1 values for PointVS trained on Redocked\gninaSet$_{Pose80}$ and gnina trained on Redocked are the same (68%), with PointVS trained on Redocked achieving 70%. 49

2.5 Donor and acceptor hotspot maps in the binding pocket of Mpro, coloured purple and orange respectively, and numbered according to their rank. The hotspots on the left (a) were obtained with the Hotspots API, the hotspots in the centre (b) were obtained with PointVS by performing attribution on 152 structures of bound fragments, and the hotspots on the right (c) were obtained by extracting the most common protein atoms identified by PLIP as involved in hydrogen bonds with ligand atoms across the 152 bound structures (in this case identified over five times). 59

2.4 Three Tankyrase-2 inhibitors: 5C5P (a, d, g, j), 4J21 (b, e, h, k), and 4J22 (c, f, i, l). The top row shows the PLIP analysis of each structure, where dark blue lines are hydrogen bonds, dotted grey are hydrophobic interactions, dotted green are π-π interactions, solid green are halogen bonds, and lilac are water bridges. The second row shows the results of performing atom masking with gnina, with green representing a positive attribution score (identified as making a positive contribution to binding) and red a negative score (making a negative contribution). The third row shows the results of bond masking with IGN, with lines between atoms showing the top five highest-scoring edges for each structure, with darker red representing higher scores. The bottom row shows the results of edge-attention attribution performed with PointVS, with the lines also showing the top five highest-scoring edges, and darker pink representing higher scores. 64

2.6 Top 5 PointVS donor and acceptor hotspots (both shown in red) obtained in the pocket of Mpro by performing attribution on 40 randomly generated sets of bound fragments. The number of structures constituting each random set is shown below each image. Darker spheres correspond to hotspots which were included in the top 5 more frequently. 65

3.1 An overall schema of the methods used with 1. the datasets used to train both the unconditional and conditional models, 2. the generation of high energy conformers and their addition to the datasets, 3. the training of the conditional model and 4. the conditional inference 71

3.2 Diagram depicting the process of coordinate distortion for a molecule in three-dimensional space. The process involves the following steps: first, sampling the maximum distortion (D) from a uniform distribution between 0 and D_{max} angstrom. Second, generating random offsets within the range of [-D, D] for each dimension of the original coordinates, C. Third, applying these offsets to the original coordinates of each atom in the molecule, resulting in a distorted conformer, S. 73

4.1 An example of a molecule from the GEOM dataset being decomposed using two different methods; the first, the vanilla BRICS implementation, and the second, a BRICS-based approach that swaps ring systems for canonical rings and breaks linkers down into atoms. . . 94
4.2 All 23 of the block types used in the reduced-library decomposition method 95
4.3 An illustration of the heuristic used to assign positional and rotational descriptors. The fragment's position is its geometric centroid (shown as 'X'). Its orientation is defined by a local coordinate system. The primary axis is the normal vector to a best-fit plane. The secondary, in-plane axis is determined by the vector between the two most distant atoms. As the dotted lines indicate, this choice is frequently not unique for symmetric blocks. To resolve this ambiguity, a single vector was chosen at random. This non-deterministic step assumes that the resulting rotational variance is either negligible due to fragment symmetry or introduces an acceptable level of noise into the model's input. 96
4.4 Examples of generations from sampling the centroid-only model, with oxygen atoms coloured red, and nitrogen atoms coloured dark blue. 98

5.1 8CXO crystal structure (grey) (K. Zhang et al. 2022), and predictions by vanilla OpenFold (green) and conditioned with coordinates (purple). We have highlighted regions where the conditioning method has moved the prediction in the wrong direction, as the closest region of high density is in the opposite direction to the density of the corresponding region. 113
5.2 Heat map showing the results of the most basic implementation of the OpenFold conditioning method, tested on the three representation levels available within the structure module, scaled to avoid structurally disordering the predictions. Relative improvements to the predictions made by the vanilla OpenFold model are coloured in green to purple, with purple representing an improvement, and green a worse prediction post conditioning. 114

5.3 8FEI crystal structure (grey) (Newton et al. 2025), and predictions by vanilla OpenFold (green) and conditioned with coordinates (purple) - using this basic conditioning approach, this is the structure that showed the greatest decrease in RMSD to the crystal structure. We have highlighted some of the regions where you can clearly see the purple conditioned prediction drifting closer to the grey crystal structure relative to the green vanilla prediction, but not reaching it. 114
5.4 Heat map showing the results of the OpenFold conditioning method, tested on the three representation levels available within the structure module, with updates added according to three different schedules. Relative improvements to the predictions made by the vanilla OpenFold model are coloured in green to purple, with purple representing an improvement, and green a worse prediction being made after conditioning. 115
5.5 Heat map showing the results of the OpenFold conditioning method, tested on the three representation levels available within the structure module, with three schedulers tested. Relative improvements to the predictions made by the vanilla OpenFold model are coloured in green to purple, with purple representing an improvement, and green a worse prediction being made after conditioning. 117
5.6 Heatmap showing the results of the most complex implementation of the OpenFold conditioning method, where the gradients were tracked and the predictions were updated until the prediction stopped gaining overlap with the experimental map. Relative improvements to the predictions made by the vanilla OpenFold model are coloured in green to purple, with purple representing an improvement, and green a worse prediction being made after conditioning. 118
5.7 The results of conditioning Boltz-2 on simulated cryo-EM data using the gradient-based optimisation method alone, coloured by relative improvement relative to the vanilla Boltz-2 method. 119
5.8 The results of subsampling MSAs in Boltz-2, coloured by relative improvement relative to the vanilla Boltz-2 method. 120
5.9 The results of subsampling MSAs and conditioning Boltz-2 on simulated cryo-EM data using the gradient-based optimisation method, coloured by relative improvement relative to the vanilla Boltz-2 method. 121

A.1 Left, the fragment used as a starting point in the development of erdafitinib, shown on the right. 134

A.2 Elaboration generated by STRIFE using the PointVS hotspots (yellow), that recovers the same hydrogen bond as erdafitinib, shown in blue (PDB ID: 5EW8 (Patani et al. 2016)). 136

List of Abbreviations

ADME absorption, distribution, metabolism, and excretion.

AI artificial intelligence.

BRICS Breaking of Retrosynthetically Interesting Chemical Substructures.

CASF16 Competitive Assessment of Scoring Functions.

CNN convolutional neural network.

cryo-EM cryogenic electron microscopy.

cryo-ET cryogenic electron tomography.

Da daltons.

DL deep learning.

DMTA design-make-test-analyse.

DUD Database of Useful Decoys.

DUD-E Database of Useful Decoys Enhanced.

EDM E(3) equivariant diffusion model.

EGNN E(n) equivariant graph neural network.

EMDB Electron Microscopy Data Bank.

FBDD fragment-based drug design.

FDA Food and Drug Administration.

GAN generative adversarial network.

GMD generative molecular design.

GNN graph neural network.

H2L hit-to-lead.

HTS high-throughput screening.

LLM large language model.

LO lead optimisation.

LSTM long short-term memory.

Mac1 SARS-CoV-2 non-structural protein 3.

ML machine learning.

MLBSF machine learning-based scoring function.

Mpro SARS-CoV-2 main protease.

MSA multiple sequence alignment.

NMR nuclear magnetic resonance.

NN neural network.

NSP14 SARS-CoV-2 non-structural protein 14.

PCC Pearson correlation coefficient.

PDB Protein Data Bank.

PLIP protein-ligand interaction profiler.

QSAR quantitative structure-activity relationship.

RMSD root-mean-square deviation.

RNN recurrent neural network.

SAR structure-activity relationship.

SBDD structure-based drug design.

SMILES Simplified Molecular Input Line Entry System.

TEM transmission electron microscope.

VAE variational autoencoder.

VS virtual screening.

XTB eXtended Tight Binding.

1 Introduction

Contents

1.1 Background and motivation 1
1.2 Drug discovery 6
1.2.1 Small molecule drugs 6
1.2.2 The drug discovery pipeline 7
1.3 Machine learning 11
1.3.1 Supervised learning 12
1.3.2 Neural networks and deep learning 13
1.3.3 Convolutional neural networks 14
1.3.4 Graph neural networks 15
1.3.5 Diffusion models 16
1.4 Machine learning in early-stage drug discovery 17
1.4.1 Target structure resolution 18
1.4.2 Virtual screening 24
1.4.3 Generative design 26
1.4.4 Fragment-based design 29
1.5 Thesis outline 30

1.1 Background and motivation

A drug is a chemical compound designed to produce a desired effect within a living system, such as preventing disease, restoring health, or relieving symptoms. The field of drug discovery aims to design these molecules to achieve a specific response based on knowledge of a biological target. Most commonly, this target is

a biomolecule like a protein, whose function is modulated by the drug to provide a therapeutic benefit (Stromgaard et al. 2009; Chorghade 2006; Patrick 2013).

For almost all of human history, the discovery of new drugs was guided by a combination of luck and trial-and-error, and relied on the extraction of materials from the local environment: plants, minerals, and animal products [1]. The 19th century marked the advent of modern drug discovery (Patrick 2013). While still mostly relying on the isolation of active ingredients from traditional cures—e.g., morphine (Sertuerner 1817) or quinine (Pelletier 1820)—this era also witnessed the birth of modern chemistry: scientists developed the tools necessary to synthesise molecules in the laboratory, enabling them to modify natural compounds or create entirely new ones from scratch.

This new capability gave rise to the first synthetic drugs, like aspirin (Dreser 1899), and established the modern pharmaceutical industry (Urick et al. 2019). More importantly, though, it marked a fundamental shift in perspective: from merely finding what worked to understanding *why* it worked. A pioneering step in this direction was Paul Ehrlich's early 20th-century concept of a 'magic bullet': the idea that a compound could be designed to kill a pathogen, without harming the host, by selectively targeting a mechanism unique to the microbe. This principle led to the development of arsphenamine (Ehrlich et al. 1910), the first antibiotic, and laid the groundwork for the subsequent antibiotic revolution (Hutchings et al. 2019). This concept of selective targeting was eventually broadened from external pathogens to the body's own molecular machinery, which led to the first drugs designed with specific human biological mechanisms in mind, like beta-blockers and histamine H_2 antagonists (Black, Crowther, et al. 1964; Black, Duncan, et al. 1972).

The development of these drugs constituted the first appearance of rational design: they were purposefully engineered to interact with a known biological target rather than being discovered serendipitously (Stromgaard et al. 2009). Still, a final conceptual leap was required to arrive at the dominant paradigm in modern

[1] This traditional form of medicine is actually what gave us the word drug: it is derived from the Old French 'drogue', which in turn comes from the Middle Dutch 'droge vate', meaning 'dry barrels'; a reference to medicinal plants preserved as dry matter (Tupper 2012).

pharmacology: structure-based drug design (SBDD) aims to design molecules whose shape and chemical properties are complementary to a target's binding site by using its 3D structure as a blueprint (Kuntz 1992). This approach was made possible by advances in biophysical methods, particularly X-ray crystallography, which gave scientists the unprecedented ability to determine the atomic-level structure of proteins. The method was first successfully demonstrated in the late 1980s with the design of dorzolamide (Baldwin et al. 1989; Greer et al. 1994), an inhibitor engineered to treat glaucoma by binding to the enzyme carbonic anhydrase. SBDD's most profound impact, however, came shortly after, with the determination of the HIV-1 protease structure in 1989 (Navia et al. 1989). This sparked a monumental effort that led directly to the development of inhibitors like saquinavir (Roberts et al. 1990) and ritonavir (Kempf et al. 1995), which transformed HIV/AIDS from a fatal disease into a manageable condition.

Despite this enormous progress, developing new drugs remains a challenging process. The expense of bringing a drug from discovery to market is immense, taking on average over ten years, with recent estimates placing the median cost at $1.3 billion (Wouters et al. 2020). This is largely due to the high failure rate of candidate compounds, meaning that each successful drug must offset the financial burden of numerous unsuccessful attempts (Patrick 2013).

The reasons for these failures are multifaceted. A 2019 study (Wong et al. 2019) reported that in Phase II of clinical trials—where a drug's effectiveness is first tested in patients—a lack of efficacy was the primary cause of failure in over 50% of cases. In Phase III, this figure rose to over 60%. Although this statistic may seem to suggest that we are simply bad at designing binders, the reality is slightly more complex; failure can also stem from poor absorption, distribution, metabolism, and excretion (ADME) properties (for instance, the drug may be destroyed by stomach acid or be unable to cross the blood-brain barrier) or the initial target identification may have been flawed, meaning that modulating the chosen biomolecule does not produce the desired therapeutic effect.

The other primary cause of failure is safety. The same 2019 study reported that safety concerns consistently accounted for approximately 20-25% of failures across phases II and III (Wong et al. 2019). These safety issues can be broadly divided into two categories. The first of these is on-target toxicity, where the drug successfully binds to the target, but causes negative side effects. This is an inherent and often predictable consequence of the drug's primary mechanism, and is a common issue that frequently defines the maximum tolerated dose. Off-target toxicity, on the other hand, arises from unintended interactions with other biological molecules, and can produce severe and unexpected adverse reactions that are a frequent cause of a drug's outright termination from development (Cook 2023).

Accounting for every potential point of ADME or safety failure in clinical trials is practically impossible, not only due to the hugely complex nature of biological systems, but also because negative data from clinical trials is rarely disclosed publicly (Dickersin et al. 1993; DeVito et al. 2019; Laviolle et al. 2025). Consequently, there is limited systematic data on the specifics of why and how frequently novel agents fail in late-stage development, making it difficult to learn from failures and reduce them.

A practical strategy to improve the overall success rate is to increase the number of high-quality candidates entering clinical trials. The goal is to start with molecules that are already high-affinity, specific binders to the target of interest, thereby improving the odds of success from the outset. An increasingly popular approach to this challenge, and the one adopted in this thesis, is the use of deep learning (DL) algorithms to screen and design drugs in the early stages of development.

For a deep learning model to be effective at structure-based drug design, it must master two fundamental challenges. First, it must understand the underlying forces of *intramolecular* bonding; the chemistry and physics dictating the possible arrangements atoms can take within a compound. This is necessary to produce novel compounds that are chemically valid and stable, ideally with properties similar to those in databases of existing drugs. Second, it must learn the principles of *intermolecular* binding to predict or design molecules that bind effectively to a biological target.

The chapters of this thesis are organised around these two challenges. Chapters 3 and 4 address the first challenge, focusing on deep learning methods for generating molecules that are chemically valid and replicate the distribution of known compounds. Chapter 2 explores the second challenge of protein-ligand bonding, presenting work on a model for virtual screening and then attempting to extract learned information to inform the generation of new candidates.

A prerequisite for structure-based drug design, however, is the availability of an accurate 3D model of the target protein. While the number of solved protein structures is growing (RCSB-PDB 2025), and prediction methods are improving (Baek et al. 2021; Z. Lin et al. 2023; Jumper et al. 2021; Abramson et al. 2024; Passaro et al. 2025), obtaining high-resolution structures for novel or complex targets remains a bottleneck. This motivates the final results chapter of this thesis: chapter 5, where we describe an adaptation of the inference processes of two structure prediction methods (AlphaFold 2 via OpenFold (Jumper et al. 2021; Ahdritz et al. 2024), and AlphaFold 3 via Boltz-2 (Abramson et al. 2024; Passaro et al. 2025)) to directly incorporate cryogenic electron microscopy (cryo-EM) data, with the goal of providing high-quality predictions that align with experimental data.

First, though, we provide some necessary background and context to frame these research chapters. We begin by summarising the drug discovery process, focusing on the early stages that this thesis falls within. We then provide a brief overview of the machine learning (ML) methods used across the different projects within this thesis, and finally describe how computational, and in particular ML-based, methods have been used to screen and design new compounds up until now. Finally, we outline the structure and key contributions of this thesis.

1.2 Drug discovery

1.2.1 Small molecule drugs

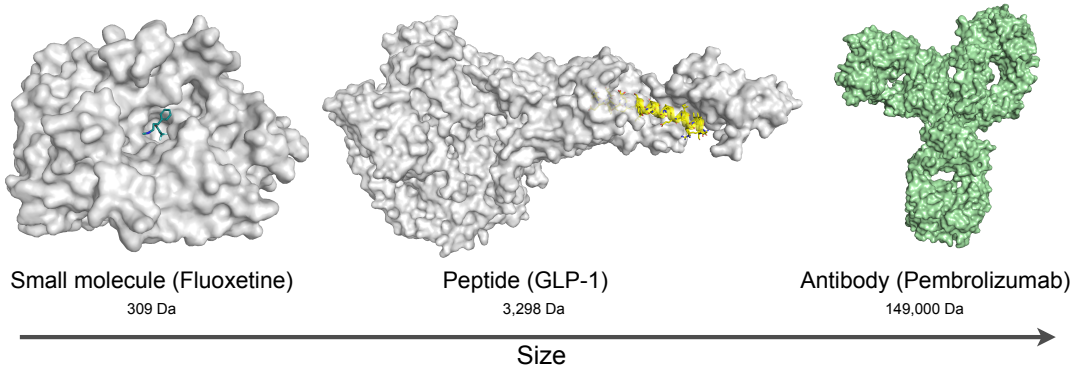


Figure 1.1: Examples of drugs of three different types: a small molecule (7MI3, Santarossa et al. 2021), a peptide (6X18, X. Zhang et al. 2020), and an antibody (5DK3, Scapin et al. 2015). The first two are shown bound to their respective targets, shown in grey.

Small molecule drugs are the most conventional class of pharmaceuticals. They are organic compounds generally defined by a molecular weight of less than 1000 daltons (Da) (Dougherty et al. 2011). Their small size and balanced physicochemical properties allow them to cross cell membranes and reach intracellular targets such as kinases (Chhabra 2021). These features also support oral bioavailability (Patrick 2013). This potential for oral dosing and intracellular target access, combined with their stability, and relatively easy, cheap synthesis, are key reasons for their historical dominance in medicine (Patrick 2013; Makurvet 2021). While the market shares of peptides and biologics are expanding (Healthcare 2022), small molecules still make up a comfortable majority of new drugs; in 2024, for instance, they accounted for 32 of the 50 novel drugs (64%) approved by the Food and Drug Administration (FDA) (Mullard 2025).

Despite being the most simple of the different drug types, each approval marks a successful journey through the lengthy, costly, and high-risk drug discovery and development pipeline. We elaborate on this pipeline below.

1.2.2 The drug discovery pipeline

This section provides a brief overview of the small molecule drug discovery pipeline[2], before delving into the stages most relevant to this thesis. As illustrated in Figure 1.2, the process can be depicted as a linear sequence of stages from target identification through to clinical trials. In practice, this process is highly iterative, and characterised by a high rate of attrition, where the vast majority of projects fail to advance from one stage to the next (Cook 2023).

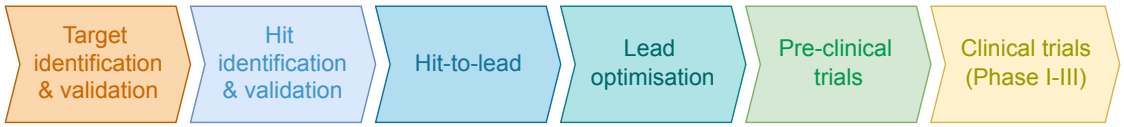


Figure 1.2: Flowchart showing the different stages of the drug discovery pipeline

Briefly, the journey begins with target identification and validation, in which the biological mechanism to be targeted is found and tested. Following this, the hit identification stage aims to find initial molecules, or 'hits', that can modulate the target's activity. These hits then enter the hit-to-lead and lead optimisation stages, where iterative chemical modifications improve key properties. Promising lead compounds then undergo preclinical testing before potentially advancing to human clinical trials.

The work presented in this thesis falls within the target validation, hit identification and hit-to-lead stages of the drug discovery pipeline. We go into more detail on these below.

1.2.2.1 Target identification, validation, and structural characterisation

Target identification generates hypotheses about which molecules have a causal role in a disease, employing methods like genomic studies, proteomics, and transcriptomics (Patrick 2013). This is followed by target validation, a critical step

[2]Whilst some literature distinguishes between 'drug discovery' (as the initial process of identifying lead compounds) and 'drug design' (as the subsequent, more rational process of optimising those leads (Patrick 2013)), the terms are also frequently used interchangeably. In many contexts, 'drug discovery' serves as an umbrella term for the entire preclinical pipeline, from target validation to the selection of a candidate drug (Hughes et al. 2011). This thesis adopts the latter, broader convention; 'drug discovery' and 'drug design' will be used interchangeably to refer to the complete process of identifying and engineering a novel therapeutic agent.

that tests this hypothesis to confirm that modulating the target will indeed lead to a therapeutic effect. Validation also assesses the target's 'druggability': its amenability to chemical modulation (Stromgaard et al. 2009).

The structural characterisation of the target, obtained through methods like X-ray crystallography or cryo-EM, is a pivotal step for enabling structure-based drug design (in contrast to ligand-based approaches, which are used when a structure is unavailable and instead rely on knowledge of known binders). As the work in Chapter 5 centres around cryo-EM model building, we go into more detail on this approach to structural characterisation below.

Cryo-EM Cryogenic Electron Microscopy (cryo-EM) involves flash-freezing a purified protein sample in a thin layer of glass-like ice, which preserves the molecules in their natural shape. A transmission electron microscope (TEM) then takes hundreds of thousands of low-resolution, 2D pictures of these molecules from all possible angles (Gonen et al. 2021).

The first computational step is to find and extract the individual molecule images from the noisy background. These images are then sorted and averaged into clearer 2D class views. From these views, software reconstructs a 3D electron density map; essentially a fuzzy 3D volume of where the atoms in the protein are most likely to be (Kühlbrandt 2014).

Crucially, this density map is still not a 3D structure of the target. The last step is to interpret this raw data by building an atomic model that fits within the 3D map. This data-to-model process is not fully automated and remains a major bottleneck, as it requires complex algorithms and often expert human intervention to produce an accurate final structure (Gonen et al. 2021).

1.2.2.2 Hit identification

This next phase of the drug design pipeline aims to use knowledge of the validated target to identify a viable chemical starting point, or 'hit'. Hit identification strategies can be broadly divided into screening and designing approaches, with fragment-based methods occupying a space between the two.

Screening approaches aim to discover active molecules from large collections. High-throughput screening (HTS) evaluates libraries of hundreds of thousands to millions of compounds against the target in biochemical or cellular assays. While it can reveal unexpected chemotypes, HTS is resource-intensive, generates many false positives, and typically yields very low hit rates (between 0.01% and 0.14% (T. Zhu et al. 2013)). The computational counterpart to HTS, virtual screening (VS), evaluates large *in silico* libraries of molecules to prioritise a smaller, more promising subset for experimental testing (see section 1.4.2 for a more thorough introduction to VS).

In contrast to screening, design approaches instead focus on rationally constructing molecules to complement the target. Medicinal chemists may design new compounds guided by the features of existing binders (ligand-based design) or by directly analysing the target binding pocket (structure-based design). Computational *de novo* design methods parallel these strategies, algorithmically generating novel molecules predicted to have high affinity and selectivity for the binding site (see section 1.4.3).

Fragment-based drug design (FBDD) represents an intermediate strategy that identifies high-quality chemical starting points by screening collections of very small molecules, or fragments, which typically have a molecular weight below 300 Da (Howard et al. 2015). Because these fragments are so small, their binding affinity for a target is usually weak, often in the millimolar (10^{-3}) to micromolar (10^{-6}) range (Zartler et al. 2008). This necessitates the use of highly sensitive biophysical techniques for screening, such as nuclear magnetic resonance (NMR) or X-ray crystallography.

Although they bind weakly, successful fragments are highly efficient binders; they form strong interactions relative to their small size. This high-quality molecular recognition makes them excellent starting points for chemical elaboration. Using the 3D structure of the fragment-protein complex as a guide, chemists evolve these initial hits into more potent lead compounds. This can be achieved by extending the fragment with new functional groups to engage with adjacent regions on the protein surface, a strategy known as fragment growing. Alternatively, two different

fragments which bind in a similar region can be connected through fragment linking, or a single new molecule can be designed that incorporates the key features of multiple overlapping fragments in a process called fragment merging. Since its adoption in the 1990s, FBDD has proven highly successful, leading to multiple approved drugs such as vemurafenib, venetoclax, and erdafitinib (Souers et al. 2013; Tsai et al. 2008; S. Wang, Burgess, et al. 2020).

These hit-identification strategies are not mutually exclusive and are frequently combined: for example, the chemical structures of hits from a high-throughput screen might be used to build a pharmacophoric profile (a 3D map of the essential features a molecule needs to bind to the target) for generative design, or virtual screening used to prioritise designed compounds for synthesis and experimental testing.

1.2.2.3 Hit-to-lead and lead optimisation

Identified, confirmed hits are not yet drugs. They are merely starting points that must be refined through two subsequent phases: hit-to-lead (H2L) and lead optimisation (LO). The H2L phase aims to convert a promising hit into a 'lead' by exploring the surrounding chemical space to establish a robust structure-activity relationship (SAR) and improve initial potency and drug-like properties. The subsequent LO phase is a more intensive, multi-parameter optimisation process that aims to polish the lead compound into a preclinical candidate, balancing high potency and selectivity with a favourable ADME-T profile (Patrick 2013; Stromgaard et al. 2009).

This entire optimisation process is driven by the iterative design-make-test-analyse (DMTA) cycle. In each cycle, new compound ideas are designed, synthesised, and evaluated in a panel of biological and physicochemical assays, and the resulting data is interpreted to inform the next round of design. The design and analysis stages of this cycle rely heavily on the ligand-based and structure-based paradigms discussed previously.

In a ligand-based approach, the DMTA cycle is guided by interpreting the SAR from tested compounds. Quantitative structure-activity relationship (QSAR)

models might be built to predict the potency of new designs, or a pharmacophore model might be refined to better understand the features required for activity: chemists make iterative changes to the molecule and learn from the resulting data.

The cycle becomes more efficient and rational if a high-resolution target structure is available. In this structure-based approach, chemists can directly visualise how molecules interact with the target. For example, if a compound binds weakly, the structure may reveal a nearby hydrophobic pocket that is not being filled. The next design can rationally add a group to occupy that pocket. If a molecule is rapidly metabolised, the structure of a metabolic enzyme might show how to block the site of metabolism without disrupting on-target binding. SBDD thus provides an atomic-level roadmap for solving this multi-parameter optimisation challenge (Patrick 2013; Stromgaard et al. 2009).

The drug discovery stages detailed above present complex challenges. Navigating the vastness of chemical space and understanding intricate structure-activity relationships are major bottlenecks that make this process slow and resource-intensive. These problems are particularly well-suited to ML, which has shown promise at learning from complex data. The following section will introduce the fundamental ML concepts that are applied to these challenges throughout this thesis.

1.3 Machine learning

In this section, we provide a very brief overview of the different machine learning architectures used in this thesis. How they are actually applied to the problems described in section 1.2 will be reviewed in section 1.4.

ML is a subfield of artificial intelligence (AI) focused on the automated detection of patterns in data, which enables computer systems to uncover key insights and predict outcomes. While the field is vast, most approaches share a common strategy: an algorithm learns to identify patterns from input data without being explicitly programmed. The primary objective is for the resulting model to generalise from this experience and accurately interpret new, unseen inputs. The landscape of ML is complex, with methods often stratified by paradigm, architecture, or application.

This can lead to confusion, particularly with terms like 'AI', 'ML', and 'DL' that are often used interchangeably despite their distinct meanings. To help clarify these relationships and provide a conceptual map, we present a flowchart in Figure 1.3 outlining the key methods under the umbrella of AI.

Following this flowchart, we will begin by introducing the foundational supervised learning paradigm. We then move to the fundamental model architecture that forms the basis of modern deep learning: neural networks. From there, we will explore the two architectures most commonly used in SBDD: convolutional neural networks and graph neural networks. Finally, we will shift our focus from predictive models to the generative domain with an introduction to diffusion models. This section is not intended as a comprehensive review of all of these topics; we simply aim to introduce the paradigms, architectures, and applications that are employed in the later chapters.

1.3.1 Supervised learning

Supervised learning is a paradigm where the model learns a mapping function from input variables (X) to an output variable (y) based on a set of labelled training examples, $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ (Goodfellow et al. 2016). The goal is to approximate the true function f such that $y = f(x)$. The model, parametrised by weights θ, learns to make predictions $\hat{y} = f(x; \theta)$ by minimising a loss function $L(\hat{y}, y)$, which quantifies the error between the prediction and the ground truth. This is typically an optimisation problem aimed at finding the optimal parameters θ^* that minimise the expected loss over the entire data distribution.

$$\theta^* = \arg\min_{\theta} \mathbb{E}_{(x,y)\sim\mathcal{D}}[L(f(x;\theta), y)] \tag{1.1}$$

In practice, this expectation is approximated by the empirical mean over the training data, which provides the objective used for gradient-based optimisation. The primary tasks within supervised learning are regression, where y is a continuous value, and classification, where y is a discrete category.

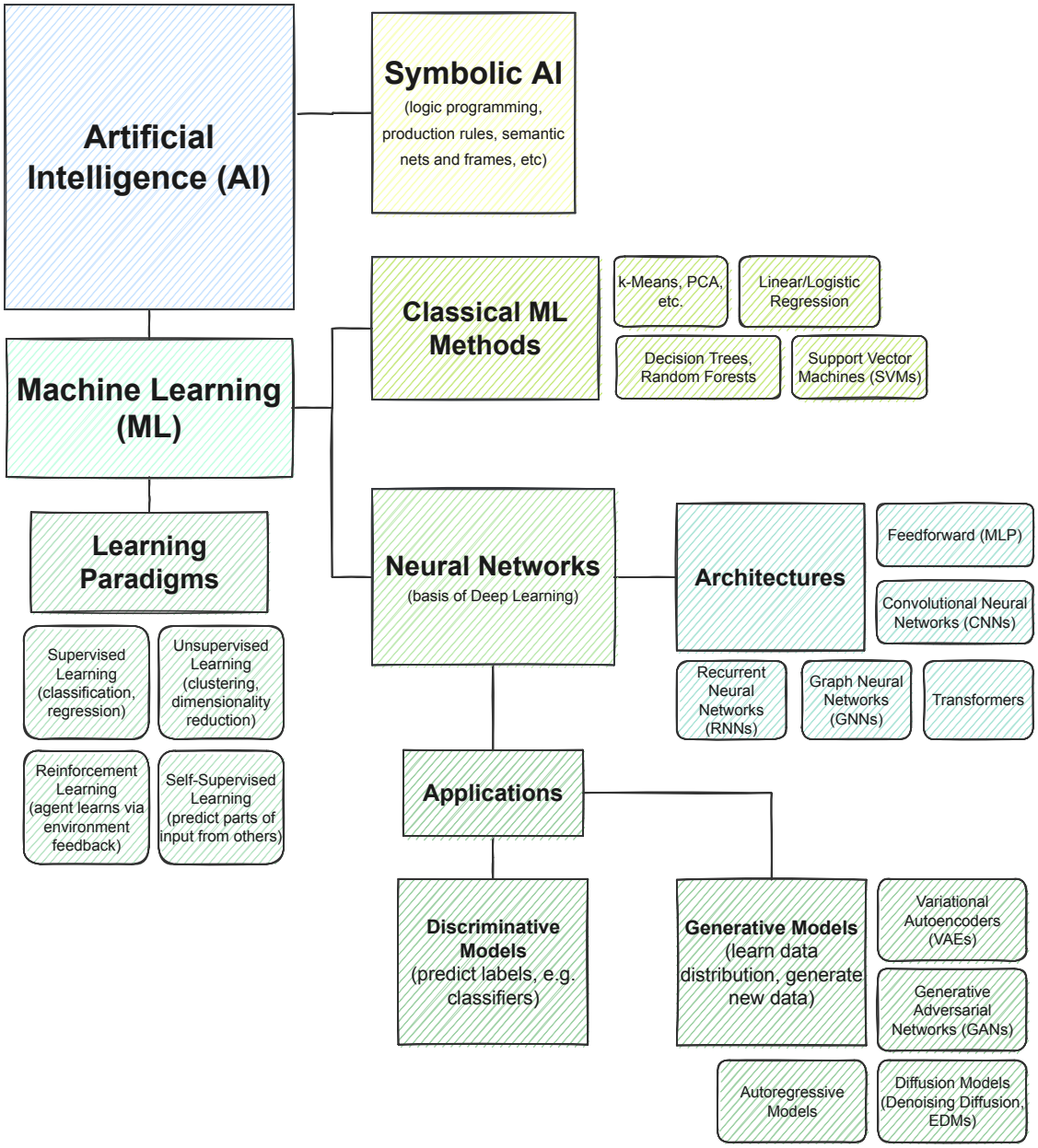


Figure 1.3: A flowchart providing a non-exhaustive overview of the various methods contained within artificial intelligence, with particular focus on the ML branch.

1.3.2 Neural networks and deep learning

Artificial neural networks (NNs) are a class of models inspired by the structure of the biological brain. They are composed of interconnected nodes, or neurons, typically organised in layers. Each neuron computes a weighted sum of its inputs and applies a non-linear activation function, $\sigma(\cdot)$, to produce an output. For a single neuron, this operation is:

$$a = \sigma\left(\sum_{i=1}^{n} w_i x_i + b\right) \tag{1.2}$$

where x_i are the inputs, w_i are the weights, b is a bias term, and a is the neuron's activation output.

DL refers to NNs with many layers (deep architectures), which allows them to learn a hierarchy of features, from simple low-level patterns to complex abstract concepts. These models are trained using optimisation algorithms like stochastic gradient descent, with gradients calculated via the backpropagation algorithm (Goodfellow et al. 2016). Backpropagation efficiently applies the chain rule of calculus to compute the gradient of the loss with respect to every model parameter θ, enabling updates according to the objective defined in eq. 1.1.

1.3.3 Convolutional neural networks

Convolutional neural networks (CNNs) are a specialised type of deep neural network designed for processing grid-like data, such as images (Lecun et al. 1998; Krizhevsky et al. 2012). Their key innovation is the convolutional layer, which leverages the principles of local connectivity and parameter sharing. Instead of each neuron connecting to all neurons in the previous layer, it only connects to a small local region. A filter, or kernel (K), with a set of learnable weights is slid across the input data (I), computing a dot product at each position. This operation, known as a convolution, produces a feature map that highlights specific patterns like edges or textures. The 2D discrete convolution is defined as:

$$(I * K)(i, j) = \sum_{m} \sum_{n} I(i - m, j - n) K(m, n) \tag{1.3}$$

By stacking these layers, CNNs can build a hierarchical representation of visual data.

1.3.4 Graph neural networks

Distinct from CNNs that operate on regular grids, graph neural networks (GNNs) are designed to work directly with data structured as graphs, $G = (V, E)$, where V is a set of vertices (nodes) and E is a set of edges (connections) (Scarselli et al. 2009; J. Zhou et al. 2020). The core principle of GNNs is message passing, where each node aggregates information from its local neighbourhood to update its own state representation. A node's feature vector h_v is updated iteratively across layers. For a node v at layer $l + 1$, a simplified update function is:

$$h_v^{(l+1)} = \sigma\left(W^{(l)} \cdot \mathrm{AGG}\left(\{h_u^{(l)} : u \in \mathcal{N}(v) \cup \{v\}\}\right)\right) \tag{1.4}$$

Here, $\mathcal{N}(v)$ is the set of neighbours of node v, AGG is an aggregation function (e.g., sum or mean), and $W^{(l)}$ is a learnable weight matrix. GNNs are thus able to learn from both node features and the graph's topology.

While standard GNNs operate on abstract graph structures, many scientific applications require reasoning about physical systems in 3D space—like molecules. For these, E(n) E(n) equivariant graph neural networks (EGNNs) are a critical extension. A model is 'equivariant" to a transformation (like rotation or translation) if transforming the input and then passing it through the model yields the same result as passing the original input through the model and then transforming the output (see Appendix A.1 for a more detailed explanation of invariance and equivariance). This property is crucial because the physical properties of a system should not change based on its position or orientation in space. EGNNs achieve this by explicitly incorporating node coordinates $\mathbf{x}_i \in \mathbb{R}^3$ into the message passing mechanism. The updates are designed to be equivariant to isometries of Euclidean space (E(n)), meaning rotations, translations, and reflections. At each layer, an EGNN updates both the node features h_i and the node coordinates $\mathbf{x}_i$:

$$
\begin{aligned}
m_{ij} &= \phi_e(h_i^{(l)}, h_j^{(l)}, ||\mathbf{x}_i^{(l)} - \mathbf{x}_j^{(l)}||^2) \\
h_i^{(l+1)} &= \phi_h\left(h_i^{(l)}, \sum_{j \neq i} m_{ij}\right) \\
\mathbf{x}_i^{(l+1)} &= \mathbf{x}_i^{(l)} + C \sum_{j \neq i} (\mathbf{x}_i^{(l)} - \mathbf{x}_j^{(l)}) \phi_x(m_{ij})
\end{aligned}
\tag{1.5}
$$

where messages m_{ij} are made invariant to rotation by depending only on the squared distance between coordinates. The feature updates ϕ_h and coordinate updates ϕ_x are learned functions, allowing the model to reason about both the graph structure and its physical geometry simultaneously.

1.3.5 Diffusion models

As illustrated in Figure 1.3, diffusion models are not tied to a specific network architecture, but are instead a class of deep generative models. These models aim to approximate the data distribution and generate new samples that are statistically indistinguishable from the training data. Earlier approaches such as adversarial models (where a generator and discriminator compete to produce increasingly realistic samples) dominated generative modelling for several years, but were often limited by unstable training dynamics and mode collapse. In contrast, diffusion models benefit from stable, likelihood-based training and have achieved state-of-the-art performance across a range of domains (Dhariwal et al. 2021; Salimans et al. 2022).

At their core, diffusion models learn to reverse a gradual noising process that transforms data into (approximately) Gaussian noise. The procedure consists of two stages:

1. Forward process (fixed): A Markov chain that progressively corrupts a clean sample x_0 into noise over T timesteps. At each step t, the noised data x_t is sampled from a Gaussian distribution:

$$
q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}\, x_{t-1}, \beta_t \mathbf{I}),
\tag{1.6}
$$

   where β_t is a small, pre-defined variance schedule. After sufficient steps, x_T is close to isotropic Gaussian noise (Sohl-Dickstein et al. 2015; Ho et al. 2020).

2. Reverse process (learned): A neural network (commonly a U-Net, though other architectures such as transformers are also used) is trained to denoise. It predicts the added noise, $\epsilon_\theta(x_t, t)$, and thereby approximates the reverse conditional distribution $p_\theta(x_{t-1}|x_t)$. Training typically minimises a mean-squared error between the true noise ϵ and the prediction ϵ_θ.

Several variants of diffusion models exist, including continuous-time score-based models formulated via stochastic differential equations (Y. Song, Sohl-Dickstein, et al. 2021), deterministic samplers such as denoising diffusion implicit models (J. Song et al. 2022), and alternative parameterisations (e.g. x_0- or v-prediction (Ho et al. 2020; Salimans et al. 2022)). A fuller taxonomy also distinguishes models by their training objectives, sampling strategies, and choice of architecture, but all share the same underlying principle: modelling data by learning to denoise corrupted samples.

1.4 Machine learning in early-stage drug discovery

The previous sections provided a general framework for drug discovery and ML. We now narrow our focus to the central theme of this thesis: applying ML methods in the early stages of the drug design pipeline. Following the chronological progression of the pipeline, we will address three primary areas.

First, we explore target structure resolution, covering both purely *in silico* prediction and the computational refinement of experimental data; both important context for the work presented in Chapter 5. Next, we delve into hit identification and lead optimisation. This begins with a review of virtual screening, in which we consider essential classical computational methods alongside the ML techniques that form the basis of the first half of Chapter 2. Finally, we discuss structure-based molecule generation. This topic is split into two problems: first, the unconditional generation of molecules, a challenge addressed in Chapters 3 and 4, and second, the conditional design of molecules directly within a target's binding site, necessary context for the second half of Chapter 2.

The approach taken in each of these areas will reflect the current state of the respective field. For target structure resolution, where recent progress has been dominated by a few landmark architectures (the AlphaFold series (Jumper et al. 2021; Abramson et al. 2024)), we provide a more in-depth analysis of these specific models. In contrast, for the less monopolised domains of virtual screening, generative design, and fragment-based methods, our review will adopt a broader perspective, giving more of an overview of how ML techniques are applied within them.

1.4.1 Target structure resolution

A protein's biological role is inseparable from its three-dimensional structure. This structure is determined by how its chain of amino acid residues folds, and dictates its ability to interact with specific binding partners such as ions, drugs, or other proteins.

We split the challenge of obtaining protein structures into different approaches: fully *in silico* structure prediction methods, and approaches relying on experimental methods like cryo-EM alongside ML.

1.4.1.1 *In silico* structure prediction

Before turning to structure prediction, it is important to distinguish it from the protein folding problem: the former refers to the computational task of inferring a protein's three-dimensional structure from its amino acid sequence, whereas the latter concerns the physical process by which a polypeptide chain adopts its native conformation. Although the two are often conflated (Lewis 2023), it is structure prediction that has undergone dramatic advances in recent years, largely due to deep learning. These models take as input the amino acid sequence of a protein and produce a predicted structure. The field was transformed by the advent of AlphaFold2, with subsequent progress driven both by its successor AlphaFold3 and by open-source frameworks such as OpenFold and Boltz-2. The core principles behind these approaches are outlined below.

AlphaFold2 The AlphaFold2 (Jumper et al. 2021) process begins with a multiple sequence alignment (MSA), which captures evolutionary relationships between the target protein sequence and similar ones found in nature. This is processed into three key data structures: a single representation encoding features for each individual residue, an MSA representation capturing evolutionary patterns, and a pair representation, a grid where each cell stores information about the spatial relationship between a specific pair of amino acids.

These representations are the inputs for the system's core component, the evoformer. This block iteratively updates both the MSA and pair representations using attention mechanisms. This facilitates a 'cross-talk' where evolutionary information from the MSA informs the geometric predictions in the pair representation, and vice versa, allowing the model to infer which amino acids are likely to be in close contact.

These refined representations then pass to the structure module, which builds the explicit 3D structure. This module operates on a novel frame representation of the protein backbone, where each amino acid residue is treated as a rigid body with its own local coordinate system (a frame, defined by a rotation and translation). This equivariant representation ensures that transformations applied to the input (like rotating the entire molecule) result in an equivalent transformation of the output. The module iteratively refines the positions and orientations of these frames, using the geometric constraints learned by the evoformer to position each residue relative to all others. The final 3D coordinates are then generated from these optimised frames, producing a physically realistic geometry.

AlphaFold3 AlphaFold3 (Abramson et al. 2024) is an architectural evolution of AlphaFold 2, moving from a protein-specific predictor to a general biomolecular structure prediction engine. Its core pipeline consists of an input processing stage, a recycling block built around a new pairformer module, and a generative diffusion module for structure generation. Unlike AlphaFold2, which heavily processed an MSA throughout its main evoformer block, AlphaFold3 distils the evolutionary information into its pair representation early on. This allows the main pairformer

block to focus exclusively on refining the relationship between input components without directly operating on the MSA, simplifying the architecture and centralising the flow of information.

The model's increased generality stems from a fundamental change in its input representation. While AlphaFold2 used tokens for 20 amino acids, AlphaFold3 employs an expanded, unified system that represents proteins, DNA, RNA, and small-molecule ligands. Proteins and nucleic acids are tokenised at the residue/nucleotide level, while ligands and other molecules are represented as a collection of individual heavy-atom tokens. These diverse inputs are processed into a single representation (1D features per token) and a pair representation (2D features for every pair of tokens). This unified framework allows the pairformer to learn interactions between any combination of biomolecules (such as protein-ligand or protein-DNA) by updating the pair representation through triangular attention mechanisms, which enforce geometric consistency.

The most major architectural change in AlphaFold3, however, is the replacement of AlphaFold2's structure module with a diffusion module. Instead of predicting a set of rigid frames for each residue, the diffusion model begins with a random 3D cloud of atoms and iteratively denoises it into a physically plausible structure, conditioned on the refined pair representation from the pairformer. This generative approach directly produces all-atom coordinates without enforcing rotational and translational equivariance, a core principle of AlphaFold2. This allows for more accurate prediction of flexible side chains, chemical modifications, and the precise geometry of non-protein molecules, marking a shift from a template-like construction to a more flexible, generative process for creating the final 3D model.

Open-source models and extensions OpenFold, Boltz-1, and Boltz-2 (Ahdritz et al. 2024; Wohlwend et al. 2024; Passaro et al. 2025) are open-source models created to replicate and expand upon the AlphaFold systems. OpenFold is a reimplementation of AlphaFold2, renowned for releasing a full training pipeline that allowed researchers to retrain and adapt the model.

Following this, Boltz-1 was developed as the first open-source replication of AlphaFold3. Its successor, Boltz-2, introduced the ability to predict binding affinity, meaning it can jointly predict both the structure of a complex and the strength of the interaction between its parts. Boltz-2 also has improved accuracy over Boltz-1, and allows users to specify an experimental type (e.g. X-ray crystallography, NMR, cryo-EM) to bias its sampling towards 3D structures that are more consistent with the types of conformations typically observed by that specific technique.

Despite advances in structure prediction, the direct application of these models in drug discovery is subject to caveats. Even for globally accurate structures, subtle but critical deviations in binding site geometry can be sufficient to mislead virtual screening and generative design algorithms (Scardino et al. 2023; Díaz-Rovira et al. 2023; Holcomb et al. 2023). This problem is particularly acute for novel targets, where over 50% of predictions may deviate more than 2Å root-mean-square deviation (RMSD) from experimental results (Kryshtafovych et al. 2021). A further challenge lies in evaluating the predictive models themselves, as the proprietary nature of their training data prohibits verification against potential train-test data leakage. These issues underscore that while *in silico* predicted structures can serve as helpful starting points, their use necessitates caution.

1.4.1.2 Cryo-EM model building

The advent of highly accurate *in silico* predictors has greatly improved structural biology, providing hypotheses from sequence alone. However, these models complement, rather than replace, experimental methods. Experimental structures remain the gold standard for validation and are essential for capturing the specific, functionally relevant conformations of proteins, especially with large, dynamic complexes, or novel structures outside a model's training scope (Škrinjar et al. 2025).

Given the continued importance of experimental data, a major focus is now on integrating ML directly into experimental workflows to overcome key bottlenecks. A prominent example is in the interpretation of cryo-EM data. While a cryo-EM experiment yields a three-dimensional electron density map, the subsequent

process of building an atomic model into this map is often a challenge; it is time-consuming, requires deep expertise, and can be difficult in areas of the map with low resolution or ambiguity. This crucial step of structure elucidation is where deep learning offers potential.

We group the deep learning tools designed to automate this process into two main categories: those that interpret the map directly and those that use a hybrid approach combining the experimental data with AlphaFold.

Direct map interpretation methods The first category, direct map interpretation, encompasses methods that learn to build an atomic model *de novo* by treating the experimental density map as the primary source of information, often without requiring a predefined sequence. Early deep learning approaches in this area primarily used CNNs to parse the volumetric data. Tools like DeepTracer (Pfab et al. 2021) applied CNNs to detect key features such as Cα atoms and secondary structure elements, which were then assembled into a backbone trace. While these approaches sped up model building, they often relied on separate downstream steps for sequence assignment and struggled in regions with fragmented or ambiguous density (e.g. very noisy or discontinuous map regions). Recent work has suggested that while deep learning methods outperform many of the older physics-based or template-based approaches, especially at higher resolutions, performance drops off sharply at intermediate-to-low resolutions ($\approx$4–8Å) or in local regions of low density (S. Li et al. 2025; C. Zhang et al. 2025).

More recent methods have shifted toward graph-based representations, reflecting the importance of modelling spatial and chemical relationships between atoms. GNN-based frameworks such as ModelAngelo (Jamali et al. 2024) integrate feature detection, residue identification, and chain tracing into a single model. By reasoning over both local density and inter-atomic connectivity, these approaches achieve higher accuracy and completeness, automating much of the previously manual model-building process even at moderate resolutions.

Nevertheless, several challenges remain. ModelAngelo's performance degrades at map resolutions worse than about 3.5Å, particularly for correctly assigning amino acid residues based on side-chain density and for tracing backbone connectivity in regions of weak or ambiguous density. It also struggles with nucleotide chains: distinguishing between similar bases (e.g. the two types of purines or pyrimidines) becomes unreliable unless the map resolution is substantially better than 2.5Å (C. Zhang et al. 2025).

AlphaFold-based methods In this second category, we discuss approaches that aim to condition AlphaFold models during inference. At the time of writing, there are two methods of this type: one built on OpenFold and the other on Boltz-2.

The first method is ROCKET (Fadini et al. 2025). This approach reframes OpenFold as a differentiable prior for experimental structure determination. Rather than refining coordinates directly, ROCKET keeps the pretrained prediction network intact and optimises within its prediction space, adding an experimental likelihood term that quantifies agreement with cryo-EM, cryo-ET, or X-ray data. Because the optimisation is constrained by the structural manifold encoded in the pretrained model, the resulting fits avoid overfitting to noise and recover conformations that remain biophysically plausible, while still maximising consistency with the experimental density.

A more recent approach is CryoBoltz (Raghu et al. 2025), which conditions Boltz-2 by guiding its generative inference process with cryo-EM–derived potentials. Global potentials encourage agreement with the overall fold present in the density, while local potentials bias fine-grained features such as loops and side chains. In doing this, CryoBoltz injects the experimental signal directly into the sampling dynamics of Boltz-2, allowing it to generate ensembles of conformations that reflect both experimental heterogeneity and the priors learned from sequences.

Having considered methods that integrate structural priors with experimental density maps, we now turn to a complementary challenge in drug discovery: identifying small molecules that bind to a given target protein. This problem,

virtual screening, shifts the focus from the target structures to exploring the chemical space of possible binders.

1.4.2 Virtual screening

As introduced in section 1.2.2.2, VS is a computational technique used primarily in the hit identification stage of drug discovery. Much like its experimental counterpart, high-throughput screening, VS is used to search vast libraries of small molecules and prioritise those most likely to bind to a specific protein target. The input for a VS workflow is typically a target protein structure and a collection of candidate molecules. For each molecule, the model generates a classification (binder or non-binder), a binding pose, predicted affinity value, or some combination of these, depending on the specific predictive task it has been trained in. In practice, VS has contributed to a number of successful drug discovery campaigns and is valued for its ability to interrogate chemical libraries on the scale of billions of compounds (Shoichet 2004; Lyu et al. 2019).

Like other hit identification methods, VS approaches can be broadly categorised into ligand-based and structure-based strategies. Ligand-based VS exploits similarities to known active molecules, assuming that compounds sharing physicochemical features with actives are more likely to bind (Riniker et al. 2013). Structure-based VS, by contrast, relies on the three-dimensional structure of the target protein. The most widely used structure-based tools employ docking algorithms, which take as input the coordinates of protein and ligand atoms and generate candidate conformations of the protein–ligand complex (Morris et al. 2009; Friesner et al. 2004). These poses are then evaluated using heuristic scoring functions that approximate binding affinity. While often presented as energy minimisation, these functions are hybrids of empirical, knowledge-based, and physics-inspired terms designed for computational efficiency rather than strict physical accuracy (R. Wang et al. 2003; Trott et al. 2010). Most docking workflows also approximate the receptor as rigid and treat solvation and entropy only crudely, although flexible or ensemble

docking approaches can partially address this at increased computational cost (Carlson 2025; Teodoro et al. 2025).

Before the widespread adoption of ML, docking was the cornerstone of computer-aided drug design, often used alongside methods such as QSAR, pharmacophore modelling, and fragment-based design. These classical approaches laid the foundation for modern virtual screening pipelines but were limited by their reliance on predefined descriptors and simplified physics.

ML has since been applied to VS both as a replacement for and as an augmentation of docking-based scoring functions. Early ML scoring functions were trained on hand-crafted descriptors of protein–ligand complexes using linear models or decision trees (Ballester et al. 2010). With the rise of deep learning, convolutional neural networks operating on voxelised 3D grids became common (Ragoza, Turner, et al. 2017; Jiménez-Rosés et al. 2022), followed by graph neural networks that represent proteins and ligands as graphs and explicitly model their interactions (Lim et al. 2019; Torng et al. 2019). More recent efforts have turned to E(3)-equivariant neural networks that respect the rotational and translational symmetries of 3D space, allowing models to better capture the geometry of binding (Townshend et al. 2022; A. Zhou et al. 2023). Transformer-based architectures have also been explored for integrating protein sequence and structure with ligand representations (S. Luo, T. Chen, et al. 2023). This architectural progression mirrors that in generative molecular design, where the choice of representation (grids, graphs, or full 3D structures) also influences both performance and interpretability.

Despite their promise, ML-based scoring functions are highly sensitive to the data on which they are trained. Chemical datasets reflect a narrow and biased sample of chemical space: although the pharmacologically relevant space is estimated at around 10^{60} molecules (Ruddigkeit et al. 2012), available libraries are shaped by factors such as synthetic accessibility, cost, and historical precedent. As a result, models are often evaluated on benchmarks that do not represent the diversity of possible ligands, making them vulnerable to learning dataset artefacts rather than transferable features of binding (J. Yang et al. 2020; Scantlebury, Vost, et al. 2023).

Biases can also arise during dataset curation. For instance, in the original Database of Useful Decoys (DUD), the distribution of net charges differed substantially between actives and decoys—42% of actives were charged compared to only 15% of decoys—allowing models to inflate performance by simply labelling charged compounds as active (N. Huang et al. 2006). This problem motivated the release of improved datasets (such as Database of Useful Decoys Enhanced (DUD-E) (Mysinger et al. 2012)), although even these remain imperfect (Scantlebury, Vost, et al. 2023; Kretschmer et al. 2025).

Evidence of learned bias is particularly clear in structure-based deep learning models: in some cases, predictive performance remains largely unchanged when the model is provided with only the ligand or only the receptor, suggesting reliance on dataset-specific correlations rather than genuine physical interactions (Wallach et al. 2018; L. Chen et al. 2019; Sieg et al. 2019). While techniques such as ensuring diminished performance in the absence of receptor information can help (Scantlebury, Brown, et al. 2020), ultimately only interpretability methods can come close to establishing whether a model is truly learning the physics of binding.

VS remains a central tool for computational hit identification. Classical docking provides scalability and interpretability, while modern ML architectures offer the potential for richer modelling of molecular interactions. Yet both remain constrained by approximations in scoring functions and biases in datasets. Addressing these limitations is essential if VS is to become a robust, generalisable approach for early-stage drug discovery.

1.4.3 Generative design

An alternative to the VS approach, generative molecular design (GMD) algorithms automate the proposal of molecules rather than screen datasets for them. How different GMD algorithms propose molecules varies from manually specified rule-based methods to complex deep neural network architectures trained on a collection of known molecules (M. Liu et al. 2022; Hoogeboom et al. 2022; H. Zhu et al. 2023).

1.4.3.1 Unconditional generation

Before a model can design molecules for a specific purpose (conditional generation), it must first learn the fundamental rules of what constitutes a valid, stable molecule. This is the domain of unconditional generation, where the objective is to learn the underlying probability distribution of a large corpus of molecules and sample from it to produce novel structures that are chemically plausible. Early approaches operated in 2D, with architectures like recurrent neural networks (RNNs) and long short-term memory (LSTM) networks being used to generate Simplified Molecular Input Line Entry System (SMILES) strings, while GNNs were employed to construct molecular graphs directly by predicting nodes (atoms) and edges (bonds) (Pogány et al. 2019; Y. Yang, Ouyang, et al. 2021).

A more complex and physically relevant task is to generate molecules directly in 3D space. This implicitly involves solving the conformer generation problem: predicting the low-energy 3D arrangement of atoms for a given 2D molecular graph. Unconditional 3D generation models, often built on E(3)-equivariant neural networks or diffusion models (see section 1.3), learn the joint probability distribution over atom types and their 3D coordinates. They learn to place atoms in 3D space, respecting the rotational and translational symmetries of physics. While not directly targeted towards a biological objective, these models are the foundations upon which pocket-based conditional models are built.

1.4.3.2 Conditional *de novo* generation

In conditional generative design, a primary consideration is how to represent the protein target to guide molecule generation effectively. This choice involves a fundamental trade-off. Providing a model with excessive, unfiltered protein information increases the complexity of the learning task, forcing it to learn to filter out irrelevant data. Conversely, providing simplified, pre-processed information risks introducing human bias or inadvertently omitting critical structural features. This challenge has led to several distinct approaches, each defined by its level of structural granularity. The most abstract approaches are shape-based, where

the binding pocket is simplified to its geometric volume, often using a voxel grid or point cloud (Z. Chen et al. 2023). This focuses on generating molecules with good steric fit but lacks chemical specificity. A more chemically-aware strategy is to use pharmacophore-based models (Ziv et al. 2024). Here, the binding site is described by a set of key interaction points (such as hydrogen bond donors, acceptors, and hydrophobic centres) which serve as constraints for the generative model. These pharmacophores can be derived from the protein structure itself or from the poses of known ligands, blending structure-based and ligand-based information. The most detailed and increasingly common approach is to use a full all-atom representation of the binding site, providing the model with the explicit 3D coordinates of all protein atoms (H. Lin et al. 2022). This method theoretically offers the highest potential for discovering novel interactions, relying on the model to learn the underlying physics of binding directly.

De novo design refers to the creation of entirely new molecular structures from scratch. It has evolved to feature various generative models over time, from early architectures like variational autoencoders (VAEs) and generative adversarial networks (GANs) to more recent 3D-aware methods such as CNNs (Ragoza, Masuda, et al. 2022), EGNNs (S. Luo, Guan, et al. 2022), and diffusion models (Schneuing et al. 2024). A foundational concept underpinning many of these approaches is autoregressive generation, where the molecule is constructed sequentially. In 1D, this involves generating a SMILES string character-by-character, with each new character conditioned on the sequence already written. This approach also extends to 3D, where many models build a molecule atom-by-atom in the binding pocket, conditioning the placement of each new atom on the structure built so far. This autoregressive approach has become less common with the proliferation of diffusion models, which generate all atoms simultaneously rather than sequentially.

A major challenge facing the field of generative molecular design is the assessment of binding specificity. Current evaluation usually relies on docking scores, which may inadvertently reward the generation of promiscuous molecules; compounds with generic features that score well against a wide range of targets rather than being truly

selective. As discussed in section 1.2, off-target binding is a major cause for drugs failing clinical trials, so this must be avoided. Finally, the field grapples with the same issues of dataset bias and generalisability that impact virtual screening. While well-reported and publicised in VS, generalisation within generative design models remains largely untested; likely in part because the biased virtual screening tools are often the very methods used for their evaluation. This creates a self-referential loop where models are optimised to satisfy a flawed benchmark, making it nearly impossible to determine if they can generalise to novel, out-of-distribution targets or have simply learned to exploit the biases shared with their evaluation tools.

1.4.4 Fragment-based design

As introduced in section 1.2.2.2, FBDD offers an alternative approach to the 'from-scratch' philosophy of *de novo* generation by building upon experimentally validated starting points. ML has shown some promise in automating FBDD: in this approach, generative models are not tasked with exploring the entirety of chemical space, but are instead used to rationally expand upon small chemical fragments that are already known to bind to the target, thereby constraining the search to more chemically relevant regions. As with virtual screening and *de novo* generative design, early methods used simple 1D and 2D representations like SMILES strings and molecular graphs. Then, the field shifted to using 3D molecular structures with models like CNNs, EGNNs, and diffusion models. Recently, however, the success of LLMs has brought renewed interest back to the simpler 1D text-based approaches (X. Liu et al. 2025).

A primary application of ML in FBDD is fragment elaboration. Here, a generative model is conditioned on the 3D context of a bound fragment and the surrounding protein pocket. New atoms are added sequentially, 'growing' the fragment into unoccupied regions of the binding site to form new, favourable interactions (Hadfield et al. 2022; Z. Gao et al. 2023; Y. Luo et al. 2022).

In scenarios with multiple non-overlapping fragment hits, ML can tackle the challenge of fragment linking. Generative models are tasked with designing a

chemically valid and conformationally stable linker to bridge two distinct fragments. These algorithms must respect the geometric constraints imposed by the fragments' exit vectors and the pocket's shape. Models of this type aim to generate high-quality linkers that actually form interactions with the protein, rather than just acting as simple hydrocarbon spacers (Imrie et al. 2020; Y. Yang, S. Zheng, et al. 2020; Guo et al. 2023; Y. Huang et al. 2022; Igashov et al. 2024).

Finally, fragment merging involves designing a novel scaffold that incorporates the key pharmacophoric features of multiple, overlapping fragment hits. Most ML approaches automate this by first identifying a consensus pharmacophoric profile from the bound fragments. This profile then serves as a set of spatial constraints to guide a generative process, tasking the model with creating a single molecule that connects these key interaction points in an optimised chemical entity (Ghorbani et al. 2023; Fu et al. 2024; Mukaidaisi et al. 2022).

1.5 Thesis outline

This thesis presents a series of studies exploring how ML methods can be developed and adapted to better support the process of drug discovery and structural biology. The overarching theme is the design of computational models that not only make accurate predictions, but also go some way to understanding the underlying forces of intra- and intermolecular bonding.

The first research chapter of the thesis focuses on virtual screening and fragment-based drug design. We develop PointVS, an equivariant graph neural network trained for pose classification and affinity prediction. By introducing stringent dataset filtering and incorporating attribution-based interpretability, we show that PointVS captures binding interactions consistent with physical interaction profiles and can be used to extract pharmacophoric information for fragment elaboration. This enables more effective structure-guided fragment growth compared to traditional hotspot extraction methods.

The second and third research chapters of the thesis address the challenge of generating novel small molecules through diffusion-based *de novo* design. Recognising that these models often struggle to produce physically plausible, druglike structures, we first explore the use of negative training data to help the model delineate valid from invalid regions of chemical space. This approach demonstrates that incorporating both positive and negative examples improves the plausibility of generated molecules. We then test whether simplifying the generative task can further improve results by using chemically valid fragments as building blocks instead of individual atoms. We find that although this method is, in theory, a promising way to reduce the difficulty of *de novo* generation, the block-based approach introduces its own challenges. We provide an overview of work carried out since which has attempted to implement similar, highlighting the key differences, and give an outlook based on this.

The final research chapter turns to structural biology, specifically the problem of fitting protein structures to cryo-EM density maps. While structure prediction tools such as AlphaFold have become invaluable, they often fail on the difficult targets most relevant for cryo-EM. We investigate conditioning strategies for incorporating cryo-EM data into AlphaFold-derived models, first within their structure modules and then through earlier interventions such as MSA subsampling. Our preliminary findings reveal both the limitations of current conditioning strategies and promising directions for future work, particularly for inducing large-scale structural rearrangements.

Taken together, the work in this thesis addresses core challenges in both molecule design and protein structure modelling: ensuring that ML outputs are not only accurate but also physically realistic and practically useful. The conclusions and outlook highlight how these ideas might be developed further; towards ML models that function across data distributions, generate molecules with experimentally sensible properties, and integrate with structural biology workflows.

2

Fragment elaboration with attribution-informed constraints

Contents

2.1 Preface 34
2.2 Contributions 35
2.3 Introduction 35
2.4 Methods 37
2.4.1 Model 37
2.4.2 Datasets 38
2.4.3 Training dataset filtering 41
2.4.4 Attribution 42
2.4.5 Fragment elaboration 45
2.5 Results 48
2.5.1 Training and testing PointVS 48
2.5.2 Attribution: identifying important interactions 52
2.5.3 Hotspot identification using PointVS 57
2.5.4 Fragment elaboration 60
2.6 Conclusions 62

This chapter is based on the work described in the following publication:

Scantlebury J*, **Vost L***, Carbery A, Hadfield TE, Turnbull OM, Brown N, Chenthamarakshan V, Das P, Grosjean H, von Delft F, Deane CM. A Small Step Toward Generalizability: Training a Machine Learning Scoring Function for Structure-Based Virtual Screening. J Chem Inf Model. 2023 May 22.

*Authors contributed equally

This work was carried out in collaboration with Jack Scantlebury, and our individual contributions are covered in section 2.2.

2.1 Preface

In order to meaningfully grow a fragment into a leadlike compound, information about the protein pocket being targeted is important. In the work presented in this chapter we investigated various computational methods for identifying hotspots: areas in protein pockets that contribute disproportionately to binding, and hence can be used to guide fragment elaboration.

This chapter first covers PointVS: an EGNN trained to perform pose classification and binding affinity prediction. During the construction of this model, the most commonly-used benchmark for machine learning-based scoring functions, CASF16, was found to suffer from information leakage concerning almost every test structure. With this in mind, a number of efforts were made to debias PointVS; first, performing a rigorous filtering step on the datasets that it was trained and tested on, and second, implementing an edge-attention attribution step to investigate exactly what PointVS based its predictions on. This last step demonstrated that PointVS identifies binding interactions in at least partial agreement with a distance-based interaction profiler. We subsequently investigated whether the attribution method could be used to identify hotspots in binding pockets when PointVS was supplied with a number of bound structures from fragment screen experiments. We used these hotspots to perform fragment elaboration, and saw improvements in the docking scores of the resulting molecules compared to using structural information from a traditional data-based hotspot extraction method. These results demonstrated that not only is PointVS capable of identifying important binding interactions, but also that it can be used as a tool for extracting said binding pharmacophores from a given protein target.

2.2 Contributions

The work described in this chapter was carried out in contribution with Jack Scantlebury, who designed PointVS, carried out the dataset filtering, and conducted all assessments of PointVS as an MLBSF. I prepared the code for the other attribution methods, carried out the attribution tests, and performed the fragment elaboration tests.

2.3 Introduction

There are growing concerns regarding the generalisability of machine learning models, as recent research into machine learning-based scoring functions (MLBSFs) has highlighted their tendency to learn dataset biases, rather than physical interactions (Boyles et al. 2022; Sieg et al. 2019; L. Chen et al. 2019; Wallach et al. 2018). In these cases, the MLBSFs perform less of an assessment of atomic contributions, and something more analogous to a nearest-neighbours search in ligand-pocket space. This is useful if the application is to proteins or scaffolds with a large amount of previous binding data available, but not for new targets or ligands, as the true functional relationship governing binding is not learned. This problem is often exacerbated by a lack of dataset filtering. One of the most widely used benchmarks for scoring functions, both classical and machine learning-based, is the Competitive Assessment of Scoring Functions (CASF16) (Su et al. 2019). A standard protocol for training an MLBSF is to use the PDBBind (Z. Liu et al. 2017) general set of crystal structures and binding affinities after removing the structures in CASF16. As we found when building our scoring function, this approach suffers from information leakage concerning almost every test structure, resulting in an overestimation of the accuracy of MLBSFs.

There is an active area of research surrounding how machine learning models make predictions (Yosinski et al. 2015; Samek et al. 2015; Poelking et al. 2022; Sharma et al. 2023), a technique known as input attribution. This is a specific application of interpretability which can be applied to MLBSFs, with the idea being

that an ideal MLBSF—one which has learnt to distinguish atomic interactions, rather than just dataset biases—should be able to identify the most important binding interactions taking place in a given bound structure. This should mean that the function predicts well on unseen targets.

In practice, the type of attribution that can be used with an MLBSF depends on its architecture. The two prevalent deep learning architectures used as scoring functions are CNNs (Koes et al. 2013; L. Zheng et al. 2019; Y. Li et al. 2019) and GNNs (Feinberg et al. 2018; Morrone et al. 2020; Jiang et al. 2021). Attribution for CNN-MLBSFs is limited, because spatial relationships are lost as information flows deeper into the network (Hochuli et al. 2018). Masking—a method by which a feature is allocated a score equal to the difference between the model's prediction made when the feature is removed from the input, and the score when it is present—can still be used, but this is computationally expensive.

Another method for performing attribution is attention analysis. Attention mechanisms work by allocating scores to the graph's edges, which can be thought of as indicating how much importance is ascribed to each edge by the model. In contrast to CNNs, GNNs preserve the concept of the atom until the final layers of the network, so edges can be interrogated for information. This is helpful in instances where an attention mechanism has been used to weight the edges, as it can give direct insight into how important the network considers each edge to be. However, none of the current leading models for affinity prediction use attention-weighted edges for intermolecular interactions, so attribution cannot be carried out on them in this way.

If input attribution could be carried out on MLBSFs, it could be used to extract binding insights from a protein, for example, for fragment-based drug discovery (FBDD); a strategy that seeks to identify simple, small molecules (fragments) that interact with protein targets, and grow them into active leads (see chapter 1). In order to meaningfully add atoms to these molecules, structural knowledge about the target needs to be considered. This can be extracted from known binders, but this limits the exploration of chemical space as well as restricting the applicability

of the approach to targets with known ligands. Extracting structural information directly from the protein itself is therefore advantageous.

At the time that this work was carried out, there was only one available generative model for fragment elaboration that relies solely on protein structure (Hadfield et al. 2022). This method uses a data-driven approach (Curran et al. 2020) to identify 'hotspots': regions in the protein pocket that contribute disproportionately to binding. Currently there exists no deep learning-based tool to identify such regions.

In order to have full control over the architecture and training set of the MLBSF we perform attribution on, we built PointVS, an $E(n)$-equivariant graph neural network. It is an MLBSF designed to predict both binding affinity and pose score. We assess its performance on the CASF16 benchmark, and find it achieves comparable results to other leading scoring functions in spite of the additional filtering of the training dataset. We then use attribution to show that the model successfully identifies binding interactions in agreement with a distance-based interaction profiler (Salentin et al. 2015). To further investigate the power of the attribution method, we used it with fragment screen data to obtain information that can be leveraged to perform fragment elaboration. We see that using this information extraction method results in improved docking scores compared to using a data-driven approach. We make our code as well as our unbiased test and train splits available on GitHub at github.com/oxpig/PointVS.

2.4 Methods

2.4.1 Model

PointVS is a lightweight $E(n)$-equivariant graph neural network, consisting of an initial projection to take the number of node features from 12 to 32, then 48 EGNN layers, followed by a global average pooling of the final node embeddings. This is followed by a sigmoid layer which gives a label $y \in [0, 1]$ during the first stage of training on pose prediction. During finetuning on affinity data, this final layer is replaced by a randomly initiated fully connected layer and ReLU activation, which outputs $\mathbf{y} \in (\mathbb{R}^+)^3$ (see Fig 2.2). It also uses a shallow neural network as an

attention mechanism (Vaswani et al. 2017), which learns to score network edges—in this case, representing atomic interactions—by their importance.

Our EGNN takes four inputs: positions, node embeddings, edge indices, and edge embeddings. If there are n atoms in an input structure, the positions tensor is an $n \times 3$ tensor containing the x, y and z coordinates of each atom. The node embeddings are an $n \times 12$ tensor of one-hot encoded OpenBabel atom types, with a bit to distinguish between ligand and receptor atoms.

Edges are generated on-the-fly using a variable distance cutoff. We use a cutoff of 10Å for ligand-protein edges, and 2Å for ligand-ligand and protein-protein edges. In this way, intramolecular edges closely mimic the covalent structure of input molecules, with a much more expansive connectivity to describe intermolecular interactions. The edge tensor is an $3 \times m$ matrix of m edges, with one-hot encoding for ligand-ligand, ligand-receptor, and receptor-receptor interactions.

The binding pocket is defined as the collection of protein atoms within 6Å of any ligand atom; the rest of the protein is ignored.

2.4.2 Datasets

The first of the steps made to debias PointVS was performing a thorough filtering of the datasets it was trained and tested on.

Our training dataset for pose classification was Redocked2020 (Francoeur et al. 2020) (Redocked from here on in), with the test set generated according to McNutt et al. 2021 so that classification performance could be directly compared with their work. This test is similar to the *docking power* test from CASF16 (Su et al. 2019), which we also conducted.

For regression on binding affinity values, we trained a pose classifier on the Redocked set, finetuned on affinity data from the PDBBind General Set, and tested on the PDBBind Core Set. This aimed to replicate the *scoring power* and *ranking power* tests from CASF16. As the Core Set is a subset of the General Set, we removed Core Set structures from the training data.

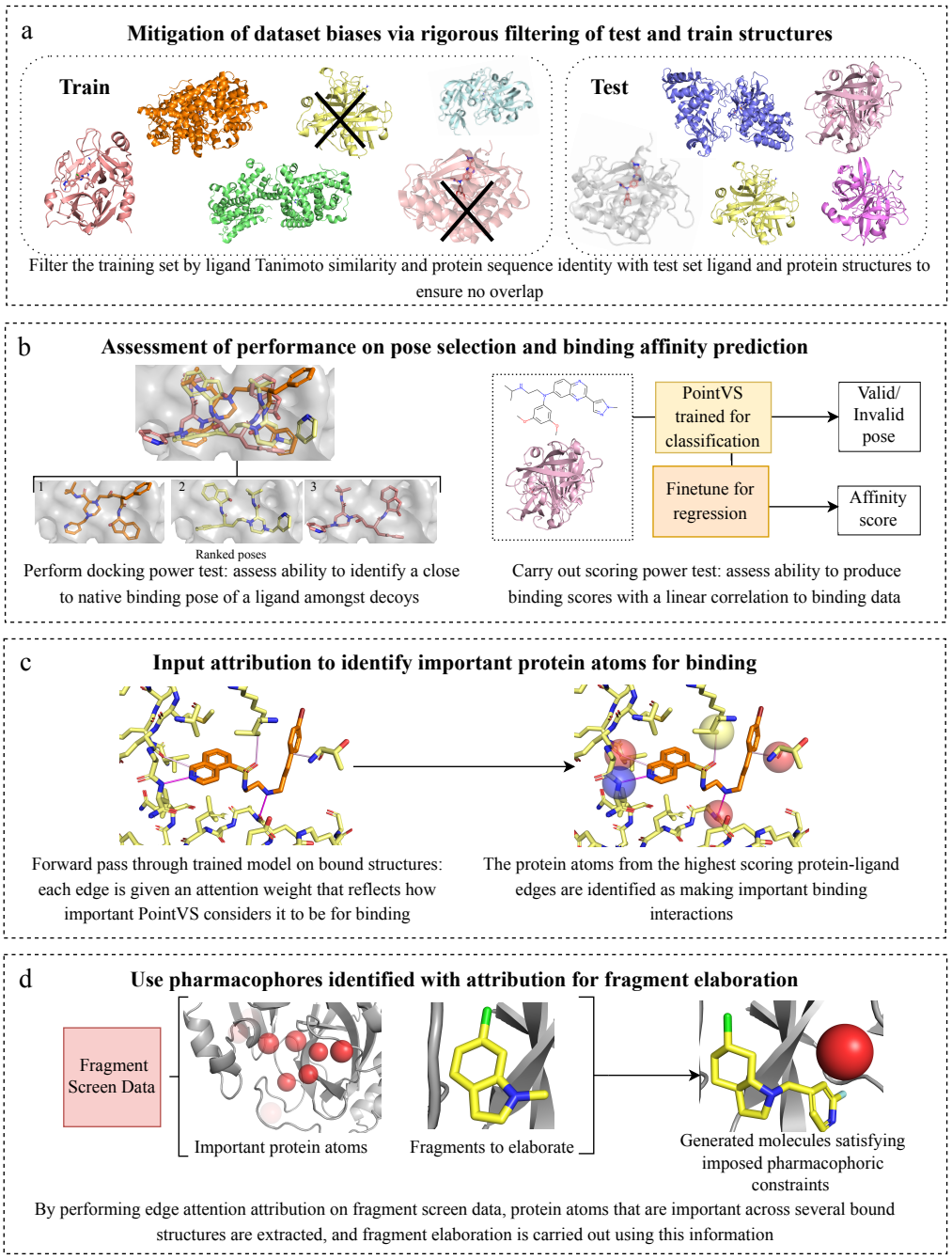


Figure 2.1: An overall schema of the methods used to debias and test PointVS. We first thoroughly filter the test and train sets (box a), before benchmarking the performance of PointVS on the docking power and scoring power tests (b). We then use attribution to gain insights into important binding regions in the protein pocket (c), which we use for fragment elaboration (d).

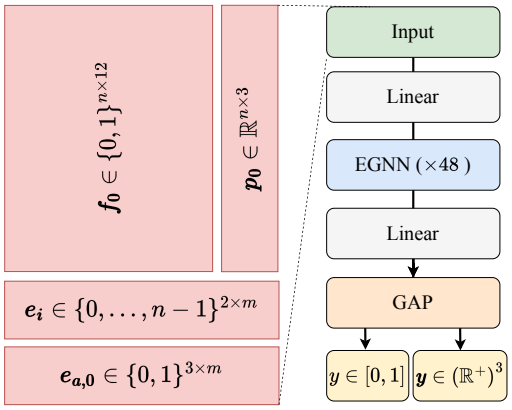


Figure 2.2: Architecture and input format of PointVS. Each of the n atoms in the input to the model (left) is given a one-hot encoded feature vector with a single bit added to indicate whether the atom is from the ligand or the receptor, as well as a position. There are m edges, defined by the edge indices which are the indices of connected atoms, and the corresponding edge attributes. These are one-hot encodings representing ligand-ligand, ligand-protein, and protein-protein edges. There are skip connections between each of the EGNN layers, and the linear and global average final pooling (GAP) layers only act upon the node features f.

2.4.2.1 The Redocked set

Training for pose classification used the same training data as the Redocked model outlined by McNutt et al. 2021. In previous work by the same authors (Ragoza, Hochuli, et al. 2017), ligands from the Pocketome v17.12 (Kufareva et al. 2012) were redocked using AutoDock Vina into their cognate receptors, with an RMSD calculated between the redocked poses and the crystal pose. Sub-2Å poses were labelled as active poses, with the rest labelled as inactive. Further negative examples were generated adversarially: a gnina CNN trained on the original redocked dataset was used to optimise docked poses. Optimised poses which were scored highly (> 0.9) by the trained CNN but were more than 2Å RMSD from the crystal pose were added as inactive poses. In total, there are 52,385 active and 734,417 inactive poses in the version of Redocked2020 used here, which we have shortened to Redocked.

2.4.2.2 PDBBind refined: gninaSet$_{Pose}$ set

The PDBBind refined set v.2019 (Z. Liu et al. 2017) is a carefully selected set of high-quality protein-ligand complexes, complete with binding affinities. In their

gnina 1.0 paper (McNutt et al. 2021), the authors took the 4,852 unique protein-ligand complexes, removing solvents and any other molecules besides the protein and the ligand, as well as removing structures with ligands that do not have a mass between 150 and 1,000 Da. This filtered, preprocessed version of the refined set is hosted by the authors of gnina 1.0 (McNutt et al. 2021).

Of the 4,260 filtered structures, the 441 which are not used in training the 'redocked' gnina model were selected as the test set for consistency with that work, which we call gninaSet$_{Pose}$. Up to 50 poses are reranked by PointVS (some of which may be very similar), after which the standard 1.0Å RMSD filter was applied to get a final pose ranking.

2.4.2.3 PDBBind General and Core sets

The PDBBind v.2020 General set contains protein-ligand complexes and binding affinities curated from the Protein Data Bank (PDB) (Berman et al. 2000). The affinities are in the form of pK_d, pK_i or pIC_{50}, and this was the training set for affinity regression. The PDBBind Core set, also known as the CASF16 test set for affinity regression, comprises 285 high-quality crystal structures, as well as their binding affinities (Su et al. 2019), and is a subset of the General set. This is the test set for regression (*scoring power* and *ranking power*).

2.4.3 Training dataset filtering

Due to the lack of affinity data, we chose to pretrain on a pose classification task in order to learn some of the physics of binding before finetuning on affinity data. As per Fig 2.1, filtered subsets of both the Redocked set and the PDBBind v.2020 General Set were generated for both of these stages of training, as well as random subsets of the original training sets of the same size as the filtered sets. The filtered datasets were constructed by removing proteins and ligands if they met either of the following criteria:

1. Tanimoto similarity of the 2048-bit Morgan fingerprint between the ligand and any of the 285 test set ligands is greater than 0.8

Dataset	Size	Unique PDB IDs	Task/ Split
Redocked	780,572	19,595	C/Train
Redocked\gninaSet$_{Pose80}$	633,047	17,528	C/Train
Redocked\Core$_{80}$	629,963	16,900	C/Train
General	19,157	19,157	R/Train
General\Core$_{80}$	14,555	14,555	R/Train
gninaSet$_{Pose}$	20,411	411	C/Test
Core	285	285	R/Test

Table 2.1: Datasets, their sizes and the number of unique PDB structures in their generation. Redocked includes no structures with the same PDB code as any structures in the gninaSet$_{Pose}$ or Core sets, and the General set shares no PDB codes with the Core set. In the Task/Split column, 'C' refers to pose classification and 'R' refers to affinity regression.

2. Sequence identity between the protein and any of the 285 test set proteins is greater than 0.8

These filters were also applied to the Redocked set with respect to the gninaSet$_{Pose}$ test set. All files which could not be parsed by either RDKit (Landrum et al. 2023) or OpenBabel (O'Boyle et al. 2011) were also discarded.

This process reduced the training set sizes from 780,572 to 633,047 (Redocked\gninaSet$_{Pose80}$) and 629,963 (Redocked\Core$_{80}$) and from 19,157 to 14,555 (General\Core$_{80}$). A full list of datasets is given in Table 2.1. In order to deconvolute the effects of training set size from the effect of the filters, we also constructed training sets of the same size, randomly sampled from the original training sets. These are called Redocked\gninaSet$_{PoseR}$, Redocked \Core$_R$ and General\Core$_R$. Carrying out filtering in this way should ensure the PDBBind Core set performance is a more accurate reflection of accuracy on unseen targets and ligands.

2.4.4 Attribution

A number of previous studies have identified the tendency of MLBSFs to make predictions based on dataset biases rather than an understanding of the physics of binding: dataset clustering and cross-validation have been used to demonstrate that many CNN-based MLBSFs, for instance, are prone to distinguishing binders

from non-binders based not on protein-ligand interactions, but on ligand features alone (J. Yang et al. 2020; Wallach et al. 2018; L. Chen et al. 2019; Sieg et al. 2019). Ensuring that predictive power is diminished in the absence of receptor information in the test set causes at least some receptor information to be used (Scantlebury, Brown, et al. 2020). However, only directly observing which parts of the input space are used in making predictions can give affirmative proof that machine learning models are learning to identify physical interactions that separate active from inactive poses.

Interpretability in machine learning for pose prediction and virtual screening is an ongoing problem; the architecture of most CNNs causes the concept of individual atoms to be lost as information flows deeper into the network, making them fundamentally unsuitable to attribution at the atomic level. By contrast, graph-based architectures such as PointVS maintain distinct nodes (atoms) until the final layers. These, having passed through the entire network, are rich in information and are directly related to the relevant atom in the input. As the node features are directly attributable to their input atom, and the final pose score is a function of the node features, there is a direct relationship between the atoms and the PointVS score, although there is no incentive during training to localise class contributions on particular atoms. The edges between atoms can also be probed for how much importance is ascribed to atomic interactions, as an intuitive way to describe non-covalent bonds. As set out in Fig 2.1, we can also use this knowledge of importance assigned by PointVS to edges to identify important protein atoms for binding.

To check whether PointVS has learnt to identify important binding interactions, and further, whether these can be extracted with attribution, we compare the results of performing attribution on PointVS to those found when using other leading MLBSFs.

We use three approaches for attribution: atom masking, bond masking, and edge attention.

Atom masking

Atom masking is a process by which the importance of different atoms can be ascertained. It is carried out by calculating the difference between the score

given when an atom i is removed from the set of input atoms, X, and the score when it is present. As all architectures of MLBSFs take as input atom information, atom masking can be carried out on any MLBSF.

Bond masking

Similar to atom masking, bond masking is a process by which the importance of different bonds can be ascertained. It is carried out by calculating the difference between the score given when an edge e is removed from the input graph encoding the protein-ligand complex and the score when it is present. Whilst atom masking can be carried out with CNNs and GNNs alike, bond masking can only be carried out with architectures that take as input not only information about the atoms in the complex, but also information about the connections between atoms. Graph representations of molecules, which use nodes and node features to describe atoms, and edges and edge features to describe interactions between them, are well-suited to this type of attribution. CNNs, which take only atom information as input, are not.

Edge attention

GNN-based MLBSFs can be built to include an attention-like mechanism, where the edges are given a score between 0 and 1 by a shallow multilayer perceptron (MLP) which takes as input the edge embeddings themselves. These weights can be thought of as indicating how much each edge is weighted by the network, so attribution can easily be performed by extracting the edge attention weights.

In order to facilitate attribution, PointVS was built to include an attention-like mechanism. After the edges are assigned attention scores, the edge message passed to the i^{th} node at each layer $\mathbf{m}_i$ is the sum of connecting edge embeddings $\mathbf{m}_{ij}$ in the neighbourhood of the node $N(i)$, weighted by this attention score e_{ij}, as in Equation 8 in Satorras *et. al* (Satorras et al. 2022):

$$\mathbf{m}_i = \sum_{j \in N(i)} e_{ij} \mathbf{m}_{ij} \tag{2.1}$$

We carry out attribution with PointVS as well as two other MLBSFs for comparison: a convolution-based method, gnina (McNutt et al. 2021), and another graph-based model, InteractionGraphNet (Jiang et al. 2021).

2.4.5 Fragment elaboration

Having gone some way to demonstrating that PointVS is able to identify important binding interactions—first by showing that it exhibits competitive performance on virtual screening benchmarks even when trained on a rigorously filtered dataset, and second by showing that attribution analysis identifies bonds in some agreement with a distance-based interaction profiler—we investigated whether such analysis could be used to identify hotspots in binding pockets when supplied with a number of bound structures from fragment screen experiments.

2.4.5.1 Hotspot Extraction

We extracted hotspot maps from PointVS by performing edge attention attribution on fragment screen data, and compared the resulting hotspots to those found using a data-driven approach, the Hotspots API (Curran et al. 2020). This implements the algorithm described in Radoux et al. 2016, covered in Appendix A.2—briefly, this involves searching the CCDC library of chemical contacts to deduce areas in the binding pocket that contribute disproportionately to binding. Such areas are known as hotspots. We then process the output of the API as in Hadfield et al. 2022.

2.4.5.2 Protein targets

Our first test used the bound structures available for the SARS-CoV-2 main protease (Mpro). Mpro is involved in processing polyproteins pp1a and pp1b of SARS-CoV-2, which are necessary for replication and transcription (Huff et al. 2022).

The dataset available for this target is far more extensive than most targets will benefit from, as it is comprised of not only fragment-like molecules, but also follow-up compounds (Boby et al. 2023). These follow-up compounds, which have been curated with the specific intention of targetting important protein atoms, consist of closely related compounds with the same basic scaffold. Intuitively, the presence of high-quality compounds should result in PointVS being able to extract higher-quality hotspots, as the bound ligands were designed to interact with protein atoms that have been experimentally validated to be important. We

	Mpro	Mac1	NSP14
Bound structures available	152	58	19
Test fragments generated	109	35	30

Table 2.2: Number of bound structures available and number of test fragments generated for each of the three protein targets used in this work: Mpro, Mac1, and NSP14.

used this dataset as a starting point to verify that, given a selection of bound structures, PointVS can be used to identify these important protein atoms in a way that is useful for fragment elaboration.

To ensure that PointVS hotspots are reliable for proteins other than Mpro, we include two further targets in this analysis: SARS-CoV-2 non-structural protein 3 (Mac1) (Schuller et al. 2021), and SARS-CoV-2 non-structural protein 14 (NSP14) (Imprachim et al. 2022). These both have fragment screen datasets available on fragalysis (*XChem @ Diamond* 2024). In both cases, these are considerably smaller than the Mpro dataset, and do not contain follow-up compounds.

2.4.5.3 Obtaining fragments

We used the fragalysis set of crystal ligand structures to obtain fragments to elaborate on. To do this, we followed the procedure outlined in Hadfield et al. 2022: we enumerated all cuts of acyclic single bonds that are not part of functional groups, and added a 'dummy' atom at the site of the cut. This atom is where atoms will be added to the fragment.

2.4.5.4 STRIFE: A generative model for elaboration

To perform fragment elaboration using the two sets of hotspot maps, we use STRIFE, a graph-based generative model for fragment elaboration. We follow the method outlined for its use in Hadfield et al. 2022. We take as input a fragment, a specified exit atom, and information about one hotspot; namely, its type and coordinates. STRIFE then generates molecules matching the pharmacophoric profile specified by the hotspot.

To achieve this, STRIFE has two main stages. The first of these, exploration, aims to generate a set of elaborations which contain an acceptor or donor in close proximity to a hotspot. These elaborations are then docked using GOLD's constrained docking functionality (Verdonk, Cole, Hartshorn, et al. 2003), and those which successfully place a functional group within a certain distance (2Å for the API hotspots, which are in the protein pocket, and 3Å for PointVS hotspots, which are on protein atoms) of the hotspot being targeted are selected as 'quasi-actives'. In the next stage, refinement, these quasi-actives are used to derive a fine-grained pharmacophoric profile. STRIFE then generates elaborations using those pharmacophoric profiles, and docks them. These docked elaborations constitute its final output.

For a given hotspot, we perform elaboration on a set of fragments. For some fragment-hotspot pairings, elaborations cannot be generated. This can be due to the hotspot in question being too close or too far from the exit point on the fragment, or at an angle that is unfavourable to elaborate along. In these cases, the generation stage will be unable to produce any quasi-actives; in other words, it will be unable to generate any molecules that, when docked, have a hydrogen bond acceptor or donor within a certain distance from the hotspot. An example of this is shown in Table 2.6, which shows the number of fragments successfully elaborated in the pocket of Mpro for every hotspot tested. We see that the hotspot ranked ninth by the Hotspots API, for instance, cannot be reached by elaborating on any of the 109 fragments tested, and hence no molecules are output.

The number of final elaborations generated is dependent on the number of quasi-actives that the first stage, exploration, generates. Although STRIFE is asked to generate 250 elaborations per fragment, there are often fragment-hotspot pairings where the number of quasi-actives identified is insufficient for the refinement phase to produce 250 distinct molecules. Nevertheless, as each hotspot is tested on a number of fragments occupying various positions in the pocket, we still obtain a large number of generated molecules for a given hotspot.

2.4.5.5 Assessing generated molecules

To assess the generated molecules, we docked them using GOLD (Verdonk, Cole, Hartshorn, et al. 2003), the CCDC's protein-ligand docking software. We used the constrained docking functionality—constraining using the fragment, for which we have a crystal structure—from which each molecule receives a score. We then calculated a ligand efficiency by dividing this score by the number of heavy atoms present in the molecule. This was to compensate for docking algorithms tendencies to favour larger molecules. From the ligand efficiency, we derived a standardised ligand efficiency. This was obtained by standardising the ligand efficiencies of the generated molecules and the ground truth molecule for given fragment-hotspot combination to have zero mean and unit variance (here, ground truth molecule refers to the molecule that was segmented to obtain the test fragment). We then sorted the standardised ligand efficiencies of the elaborations from highest to lowest scoring, and took the mean of the top α. In this work, we used α=20: in other words, we only considered the top 20 highest-scoring generated molecules for each fragment-hotspot pairing. Subtracting the ground truth ligand efficiency value from this value then provides us with ΔSLE_{20} for every fragment-hotspot pair, so for a given hotspot, we take the mean over the ΔSLE_{20}s of all the fragments successfully elaborated towards it. We used this metric as it provides an insight into how the elaboration process when using a particular hotspot has impacted the original molecule's docking score: a positive $\Delta\text{SLE}_{\alpha}$ means the top elaborations are improvements on the original molecule, and a negative $\Delta\text{SLE}_{\alpha}$ shows that the elaboration process has decreased the ligand efficiency with respect to the starting molecule.

2.5 Results

2.5.1 Training and testing PointVS

2.5.1.1 Bias in the CASF16 test set

We implemented a rigorous filtering step to ensure that the training set and test set structures used by PointVS have no overlap (see methods). In order to compare

PointVS with other machine learning methods which have not included this filtering step (C. Wang et al. 2017; Lu et al. 2019; McNutt et al. 2021), we also constructed a smaller test set out of the PDBBind Core set by applying the same filters, but excluding test set structures rather than train set structures. However, we found that such a test set would only contain a single structure: 284 of 285 of the Core set proteins have a counterpart with 90% sequence similarity in the General set. Further, 273 of them (95.7%) have a counterpart with an identical sequence, making any fair and unbiased comparison to these methods impossible.

2.5.1.2 Performance as a scoring function

PointVS was compared to a selection of state-of-the-art pose classification and affinity prediction tools—both machine learning and classical—on the CASF16 docking power and scoring power tests.

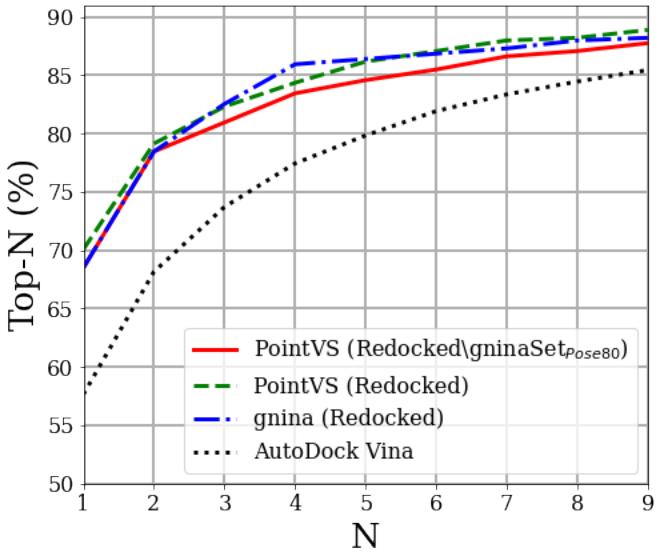


Figure 2.3: Top-N vs N pose ranking performance on the gninaSet$_{Pose}$ set for Autodock Vina, PointVS and gnina. Terms in brackets in the legend refer to the training sets; filtering the training set by protein and ligand similarity results in slightly degraded performance. The Top-1 values for PointVS trained on Redocked\gninaSet$_{Pose80}$ and gnina trained on Redocked are the same (68%), with PointVS trained on Redocked achieving 70%.

Docking power refers to a scoring function's ability to identify a ligand's native binding pose among decoys. It is quantified using Top-N, which is the percentage

of systems with a 'good' pose ranked in the top N. A pose is considered good if the RMSD to the crystal pose is less than 2Å .

Top-N pose ranking performance on the gninaSet$_{Pose}$ set is shown in Fig 2.3. When the training sets are the same, PointVS ranks a good pose at the top in 70% of cases, as opposed to 68% with gnina. When PointVS is trained on the smaller Redocked\gninaSet$_{Pose80}$ set, such that no similar proteins or ligands as the test set are seen during training, this drops to 68%. This suggests that for pose prediction, filtering the dataset to remove overlap between train and test structures only causes a minor decrease in docking power.

The Top-1 pose ranking results for the Core set (the CASF16 docking power set) are shown in Table 2.3, for both the case where the crystal pose is included amongst the poses to be ranked, and the case where it is not.

The difference in performance between all of the methods is minimal, and is not affected significantly by whether or not the training data has been filtered for structural similarity (the without-ligand test set scores are 85.3%, 84.9% and 84.2% for PointVS trained on Redocked, Redocked\Core$_R$ and Redocked\Core$_{80}$, respectively). This lack of dependence on debiasing is in agreement with the docking power results shown in Fig 2.3, although the performance of Autodock Vina is comparable with the machine learning methods in this test.

The **scoring power** is defined by Su *et al.* as "the ability of a scoring function to produce binding scores in a linear correlation with experimental binding data" (Su et al. 2019). This is best measured using the Pearson correlation coefficient (PCC) between the score given by the scoring function and the measured binding affinity data. For training our networks, we once again used both filtered and unfiltered versions of the PDBBind General dataset to test the influence of dataset bias.

We trained PointVS to classify on Redocked\Core$_{80}$, and then switched out the final layer for multi-target regression trained on General\Core$_{80}$. We also used the gnina model trained on the PDBBind General set supplied as part of work by

Model	Top-1	
	Crystal pose included	Crystal pose not included
PointVS (Redocked\Core$_R$)	90.5	84.9
PointVS (Redocked\Core$_{80}$)	91.2	84.2
PointVS (Redocked)	**91.3**	**85.3**
gnina (Redocked)	91.2	83.2
Autodock Vina	90.2	84.6

Table 2.3: Top-1 pose ranking performance for different models and training sets (in brackets) on the PDBBind Core set (the CASF16 docking power test set). The non-machine learning scoring function is shown in italic. The performance is shown for both the case where the crystal structure of the ligand is included in the set of poses being ranked, and the case where it is not included, in which case the native pose is defined as any pose less than 2Å RMSD away from the crystal pose. The highest Top-1 for each case is shown in bold.

	Scoring Function	PCC
Biased	gnina (Redocked)	0.753 ± 0.008
	gnina (General)	**0.816 ± 0.008**
	PointVS (General)	0.805 ± 0.010
	PointVS (General\Core$_R$)	0.803 ± 0.012
	ΔVinaXGB	0.796
	ΔVinaRF	0.732
Debiased	PointVS (General\Core$_{80}$)	**0.754 ± 0.015**
	X-Score	0.631
	Autodock Vina	0.601

Table 2.4: Pearson correlation coefficients between measured and predicted affinity for different scoring functions on the PDBBind Core set. Scoring functions are split into biased and debiased methods according to the overlap between their training sets, which are shown in brackets where applicable, and the Core set. Non-machine learning scoring functions are shown in italic. The highest PCCs obtained with both biased and debiased methods are shown in bold.

McNutt et al. 2021 to find the PCC on the Core set, but again emphasise that the training data for this model includes many proteins and ligands which are similar or identical to those in the Core set. This model achieves a PCC of 0.816. The results from other methods can be seen in Table 2.4. The three best methods without dataset filtering (gnina, PointVS and ΔVinaXGB) achieve very similar performance, and PointVS outperforms the best non-machine-learning methods

from the original CASF16 work (Su et al. 2019), Autodock Vina, significantly (0.805 vs 0.601). However, the performance of PointVS when trained on the filtered dataset General\Core$_{80}$ is a better indicator of its true performance on unseen protein-ligand combinations with no analogues in the training data.

To test whether the drop in the performance of PointVS when similar structures are removed from the training set is related to bias or change in training data size, we trained PointVS on General\Core$_R$. This has the same training set size as General\Core$_{80}$, but can still contain similar structures to those in the test set. PointVS trained on the General\Core$_R$ achieves a PCC of 0.803 on the CASF16 set. This is almost identical to the PCC of PointVS trained on the entire General set (PCC of 0.805), and significantly higher than PointVS trained on the General\Core$_{80}$ set (PCC of 0.754). This once again points to the performance of these methods when trained and tested using biased datasets being boosted. This contrasts with our findings in pose prediction performance, in which we saw very minimal change when filtering was performed.

As PointVS trained and tested in a debiased way still appears to be predictive of binding affinity, we next explored whether it had learnt to identify important interactions in the binding site.

2.5.2 Attribution: identifying important interactions

PointVS appears to offer a substantial improvement in predictive performance of binding affinity over non-machine-learning methods. However, MLBSFs are notoriously difficult to interpret. Good performance when trained on debiased training data is circumstantial evidence that the interatomic interactions involved in binding are being learned, but demonstrating that the attribution scores of PointVS relate to important binding interactions is a more direct measure. We next present several results demonstrating that important interatomic interactions have been learnt by PointVS.

2.5.2.1 Human Tankyrase-2 inhibitors

The protein-ligand interaction profiler (PLIP) (Salentin et al. 2015) uses simple geometric rules to predict interactions at the interface between a protein and a ligand. We used PLIP to identify the 'most important' bonds for three human Tankyrase-2 inhibitors, and examined whether the atoms forming these bonds were highlighted when attribution was carried out on gnina (Koes et al. 2013), IGN (Jiang et al. 2021), and PointVS.

The structures of three tankyrase inhibitors are shown in Fig 2.4. The top row (structures a, b and c) shows the PLIP analysis of each structure. The structures in the other three rows show the results of attribution with gnina (second row, structures d, e and f), IGN (third row, structures g, h and i), and PointVS (bottom row, structures j, k and l). The structures in the second row are coloured by the scores obtained from atom masking with gnina, with atoms contributing negatively to binding being shown in red (negative score), neutral atoms in white, and positively-contributing atoms in green (positive score). Attribution with IGN and PointVS—which was carried out with bond masking and edge attention analysis, respectively—resulted in scores corresponding to interactions between pairs of atoms rather than individual atoms. The five highest-scoring edges for structures using both of these methods are shown in the bottom two rows, with dark red lines representing high scores from IGN in structures g, h and i, and darker pink representing higher scores from PointVS in structures j, k and l.

Considering the first column, the two ligand atoms that form hydrogen bonds with the protein as calculated by PLIP (Fig 2.4a) are scored by gnina as the 6th and 8th most important of 20 (Fig 2.4d). Of the two hydrogen bonding protein atoms, gnina recognises one of them as being extremely important (the highest scoring protein atom), and the other as neutral. The second structure, Fig 2.4e, shows a similar story, with the three important ligand atoms as given by PLIP (Fig 2.4b) scoring 4th, 7th and 8th out of 20, but the protein atoms they respectively bind to scoring negatively, neutrally and very slightly positively. In the third structure, the three bound ligand atoms as identified by PLIP (Fig 2.4c) are scored 11th, 13th and

14th of 30 ligand atoms, with the corresponding protein atoms scoring negatively, weakly positively and moderately positively (Fig 2.4f). Interestingly, the last two structures both contain halogen bonds, and in both cases gnina ranks the ligand atom as important, and the protein atom as a negative contribution. Overall, gnina struggles to identify important individual atoms (both in the ligand and protein) for binding. The atoms being scored individually rather than by pairs as with bond masking or edge attention (see below) makes it more difficult to identify the bonds that are being formed: of the bonds identified by PLIP, only in the first structure was there an example where both the protein and ligand atoms were given a high score by gnina, making it the only case where an important bond was identified.

When we performed atom masking on PointVS, we found that the results were similar to those from gnina, in that the atoms contributing to bonds were scored highly but not necessarily enough to recognise their true importance. However, as described below, we found that edge-based attribution methods provided a clearer insight into important interactions.

The first of the results achieved with edge-based attribution methods is the third row of Fig 2.4, which shows the edges found to be important when carrying out bond masking with IGN. In the first structure, Fig 2.4g, the bond highlighted as most important connects a ligand atom that PLIP highlights as forming a hydrogen bond (Fig 2.4a) with a different protein atom. At first glance, it appears that IGN has incorrectly predicted the formation of a hydrogen bond between these atoms, but it should be noted that the atoms that this edge connects are both in aromatic rings that PLIP identifies as being involved in a π-stacking interaction. IGN also correctly identifies a hydrophobic interaction, which is ranked as the second most important edge. In the second structure (Fig 2.4h), we again see edges connecting rings involved in π-stacking in Fig 2.4b, this time ranked first and fifth most important. However, none of the other interactions shown to be taking place in 2.4b—neither the hydrogen nor the halogen bonds—are recognised. This contrasts the third structure, 2.4i, which shows that IGN has identified the ligand atom that forms a halogen bond as extremely important, with all five of the highest-scoring edges

connecting it to various protein atoms. The edge that connects it to the oxygen atom that PLIP identifies as being the other contributor to this bond (Fig 2.4c) is ranked third most important. In contrast to the first two structures, little importance is ascribed to the edges connecting the aromatic rings involved in π-stacking.

Altogether, the edges identified as important by performing bond masking with IGN show some agreement with PLIP. This is especially true of π-stacking interactions, of which it identifies two of three, and the least true of hydrogen bonds, of which it identifies none. In all three structures, IGN appears to successfully identify one of the interactions taking place (π-stacking in the first two, g and h, and halogen bonding in the third, i) but fails to identify any beyond those.

Considering the first example in the final row of Fig 2.4, we see that the bonds ranked first and second most important by PointVS (Fig 2.4j) correspond to the hydrogen bonds identified by PLIP (Fig 2.4a). Edges between the two aromatic rings involved in π-stacking are also given high importance scores, being ranked fourth and fifth most important. We see a similar pattern for the second structure, where the atoms involved in hydrogen bonds as defined by PLIP (Fig 2.4b) are connected by edges assigned high importance by PointVS (ranked first and second most important, Fig 2.4k). We also again see high scores assigned to the edges connecting the rings involved in π-stacking. PointVS does, however, fail to recognise the edge joining the atoms involved in the halogen bond shown in green in Fig 2.4b. In the third example structure, PointVS also identifies the hydrogen bonds identified by PLIP (Fig 2.4c) as being important, again ranking them as the most and second most important edges (Fig 2.4l). The other edges assigned high importance correspond to hydrophobic interactions.

The results above suggest that of the attribution methods tested, the two edge-based approaches are more effective at identifying important interactions taking place. However, of the two tested, only edge attention analysis with PointVS was able to reliably identify more than one of the bonds that PLIP identified as forming. This highlights that edge attention is a useful method for attribution, but further, that PointVS has gone some way to learning to identify important binding interactions.

2.5.2.2 Large-scale attribution tests

By performing attribution with edge attention scores on a crystal structure of a given protein with a small molecule bound, binding information in the form of an importance score for every protein atom less than 6Å away from any ligand atom is obtained. In cases where a fragment screen has been carried out on a target, this attribution process can then be repeated for several crystal structures. An importance score for each protein atom is then obtained for every bound structure, so we to obtain an average importance score for each of the protein atoms. This list of protein atoms and associated scores can then be used to inform molecule generation.

To verify that performing attribution with PointVS results in important binding interactions being identified, we perform attribution on the PDBBind Core set (the CASF16 test set). We compare the bonds highlighted as important for each structure with those highlighted by PLIP. We then calculated the PCC between the scores assigned to the top five and top ten scoring protein atoms, and the distance between the protein atom and the corresponding ligand atom. This results in a PCC of -0.766 for the top five protein atoms, and -0.844 for the top ten protein atoms. This negative correlation implies that protein-ligand atom pairs that are close together (which in turn PLIP is more likely to consider as forming a bond) are commonly identified by PointVS. We also calculated the correlation coefficient between the ranks of the top five and ten protein atoms and their distance-based rank (this corresponds to the protein atoms ranked by distance to the nearest ligand atom), and obtain 0.719 and 0.788, respectively. This provides evidence that PointVS is identifying bonds that at the very least make sense geometrically, but not relying on interatomic distances entirely.

We carried out this process for gnina: an alternative CNN-based MLBSF (Koes et al. 2013). Due to the architecture of gnina, attribution was carried out with masking in place of edge attention analysis. Masking is a process by which the importance of different atoms can be ascertained. It is carried out by calculating the difference between the score given when an atom i is removed from the set

	PointVS		gnina	
	PCC	ρ	PCC	ρ
Top 5	-0.710	0.640	-0.233	0.234
Top 10	-0.853	0.788	0.094	0.093

Table 2.5: Mean PCCs calculated between the scores of the top five and top ten ranked protein atoms by PointVS, and the distance between them and the nearest ligand atom. Also shown is the mean rank correlation calculated between these values, ρ. The means were calculated over a subset of 20 randomly selected bound structures from the PDBBind Core set (see Appendix A.3 for PDB IDs).

of input atoms, X, and the score when it is present. The importance of atom $i \in X$ is therefore defined as:

$$\text{atom_score}_i = \text{gnina}\,(X) - \text{gnina}\,(X \setminus i)$$

Masking is a significantly more computationally challenging process than edge attention analysis. To account for this, we used only 20 randomly selected structures from the PDBBind Core set. We also performed attribution with PointVS on this subset for comparison. The results of calculating both the PCC and the rank correlation, ρ, are shown in Table 2.5.

On this smaller set we again see a negative correlation between the score assigned by PointVS and the distance between the atoms, and a positive correlation between their rankings by distance and PointVS. When using gnina, by contrast, we see a much weaker correlation for both of these tests. This suggests we might be able to use it to identify important bonds—and therefore atoms—involved in binding.

2.5.3 Hotspot identification using PointVS

Having shown that PointVS is potentially capable of identifying important binding interactions, we next assessed whether it can be used as a method for extracting structural information from a number of bound structures, and hence used to guide molecule design. To do this, we extracted hotspot maps from PointVS by performing edge attention attribution on fragment screen data, and compared the resulting hotspots to those found using a data-driven approach, the Hotspots API (Curran et al. 2020).

Our first test used the non-covalent bound structures available for Mpro. The dataset available for this target is far more extensive than most targets will benefit from, as it is comprised of not only fragment-like molecules, but also follow-up compounds. These follow-up compounds, which have been curated with the specific intention of targetting important protein atoms, consist of closely related compounds with the same basic scaffold. Intuitively, the presence of high-quality compounds should result in PointVS being able to extract higher-quality hotspots, as the bound ligands were designed to interact with protein atoms that have been experimentally validated to be important. We used this dataset as a starting point to verify that, given a selection of bound structures, PointVS can be used to identify these important protein atoms in a way that is useful for fragment elaboration.

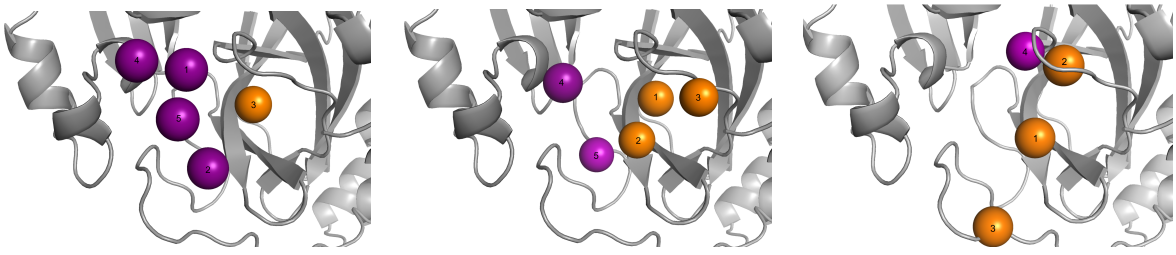
The data was extracted from the fragalysis platform (*XChem @ Diamond* 2024). A full list detailing the structures used is given in Appendix A.3. The hotspots identified by the Hotspots API and attribution with PointVS are shown in Fig 2.5a and b. The hotspots of the two methods will not fall on exactly the same points, as the PointVS hotspots correspond to protein atoms, and the API hotspots to points in the binding pocket. Still, the two methods clearly highlight very separate regions of the binding pocket as important.

2.5.3.1 Impact of fragment screen similarity

As mentioned previously, the Mpro dataset contains a significant number of closely-related follow-up compounds with the same basic scaffold. To ensure that PointVS is providing a novel insight into important binding regions in the pocket of Mpro and not highlighting important protein atoms by simply counting the number of times they are within close proximity of a ligand donor or acceptor atom, we used PLIP to extract the protein atoms that could be forming hydrogen bonds with ligand atoms for every bound structure. We then defined any protein atom that was highlighted more than five times across the 152 bound structures as a PLIP hotspot. This cutoff of five interactions resulted in only four hotspots being extracted. These top

four are shown in Fig 2.5c, and correspond to protein atoms that were highlighted by PLIP 146, 40, 9 and 5 times out of the possible 152.

All but one of the PLIP hotspots are within 2Å of one of the top five PointVS hotspots. However, the top-ranking PLIP hotspot (which corresponds to a protein atom that forms a hydrogen bond in 146 of the 152 bound structures) is ranked second by PointVS, and the second PLIP hotspot (which PLIP identifies as interacting with ligand atoms in 40 structures) is ranked top by PointVS. As the first PLIP hotspot is at most the length of one hydrogen bond away from a ligand acceptor in almost every structure, we would expect a geometry-based method to clearly identify it as the top hotspot, as PLIP does. This is some evidence that PointVS is identifying important protein atoms with a more nuanced approach than simply counting interactions identified with geometric rules. These results do not mean the PointVS hotspot extraction method is immune to being biased when supplied with fragment screens containing clusters of ligand donor and acceptor atoms, but it is a promising sign that it is not entirely biased by such clusters, and that PointVS is still able to look past them and identify other areas in the pocket as important. This is a particularly useful finding, given that recent research into fragment screen libraries has highlighted that even screens designed to be diverse structurally still offer a limited exploration of protein pockets (Carbery et al. 2022).



a b c

Figure 2.5: Donor and acceptor hotspot maps in the binding pocket of Mpro, coloured purple and orange respectively, and numbered according to their rank. The hotspots on the left (a) were obtained with the Hotspots API, the hotspots in the centre (b) were obtained with PointVS by performing attribution on 152 structures of bound fragments, and the hotspots on the right (c) were obtained by extracting the most common protein atoms identified by PLIP as involved in hydrogen bonds with ligand atoms across the 152 bound structures (in this case identified over five times).

2.5.3.2 Dependency on fragment screen size

Given that PointVS identifies sites that are different from a traditional hotspot method and a geometry-based method, we next tested whether the high-scoring sites identified by PointVS are dependent on the size of the input fragment set.

To do this, we performed the same hotspot extraction process as before, but instead of using all of the bound structures of Mpro at once, we randomly sampled smaller sets from the collection. We varied the size of the set sampled between 10 and 80. This random sampling process was carried out 40 times for each dataset size. We then performed the hotspot extraction process as before, and took the top five highest-scoring hotspots generated using PointVS with each randomly sampled set of fragment hits. The results of this are shown in Fig 2.6, where we took samples of size 10, 20, 30, 40, 50, 60, 70 and 80 from the available bound structures. We see that when the hotspot maps are extracted from a smaller number of bound structures, a wider range of protein atoms are identified as being important; PointVS is less capable of consistently picking out the same five protein atoms every time, and increasing the number of bound structures given to PointVS results in more consistent hotspot maps being generated. Nevertheless, we note that even when using the smallest datasets containing only 10 bound structures, the five hotspots that are identified when using the full fragment screen are the most commonly included, hence their darker colour in Fig 2.6. This is a promising sign that attribution with PointVS could still be a useful tool for less well-studied targets.

2.5.4 Fragment elaboration

Having demonstrated that the PointVS hotspots are consistent amongst themselves, and that they highlight different areas in the Mpro active site to the API hotspots, we assessed whether they correctly highlight important areas by performing fragment elaboration. If PointVS is successfully identifying important interactions, molecules generated to satisfy the pharmacophoric constraints imposed by PointVS hotspots should be as good, if not better, than the molecules using the API hotspots.

	PointVS		Hotspots API	
Rank	Fragments successfully elaborated	Elaborations per fragment	Fragments successfully elaborated	Elaborations per fragment
1	57	215	26	189
2	82	182	57	183
3	51	202	59	207
4	14	199	9	158
5	21	223	28	179
6	60	183	23	162
7	46	228	3	220
8	78	216	54	197
9	24	215	0	0
10	47	218	28	179

Table 2.6: The number of fragments that were successfully elaborated on using PointVS and API hotspots for Mpro, and the mean number of elaborations that were generated for each fragment. STRIFE was provided with 109 fragments to elaborate on for each hotspot, and asked to generate 250 elaborations for each one.

	Mpro: 152		Mac1: 58		NSP14: 19	
ΔSLE_{20}	PointVS	API	PointVS	API	PointVS	API
Hotspots 1-5	**0.341**	0.213	**1.749**	1.581	**1.708**	1.240
Hotspots 6-10	0.189	**0.269**	1.727	**1.748**	**1.143**	1.193

Table 2.7: Mean standardised ligand efficiency of the top 20 elaborations (ΔSLE_{20}) made using PointVS hotspots and API hotspots on three different targets: Mpro, Mac1, and NSP14. The numbers after the target names refer to the number of bound fragment structures available for each target. The highest ΔSLE_{20} for both ranges of hotspots is shown in bold.

For every target tested, the highest scoring elaborations by our ΔSLE_{20} metric were generated using the top five PointVS hotspots (Table 2.7). Elaborations made using the hotspots ranked 6-10 by PointVS consistently have worse ΔSLE_{20}s than those ranked 1-5. By contrast, the Hotspots API successfully identified ten high-scoring hotspots (indeed, on average, these top ten hotspots produced elaborations with higher standardised ligand efficiencies than the PointVS hotspots), but was unable to successfully rank them amongst themselves. The accurate ranking of hotspots is important, as in real-world fragment-to-lead campaigns, only a small number of them can easily be targeted.

Within the PointVS hotspot scores, we see that the Mpro hotspots are lower scoring than those for the other two targets. This can be attributed to the 109 starting fragments for Mpro being of higher quality, having been extracted from curated follow-up compounds rather than fragments. The improvement that is made on them by elaborating is thus less significant.

2.6 Conclusions

This chapter covered the development of PointVS, an EGNN-based method for affinity prediction, and the first ML-based method for extracting important binding information from a target for molecule design.

During the development of PointVS, a commonly used benchmarking protocol for MLBSFs, CASF16, was found to overestimate their accuracy. In light of this, PointVS was trained and tested on a filtered dataset, and hence encouraged to learn the rules governing intermolecular binding rather than memorise training data. We showed that when trained on this filtered dataset, PointVS achieves comparable results to other leading scoring functions, providing some evidence that it is successfully identifying important binding interactions. We provide further proof of this by performing attribution, and showing that PointVS is able to identify important interactions in line with those found by PLIP, a distance-based tool for profiling protein-ligand interactions.

Finally, we investigated how knowledge of important binding regions could be leveraged in other stages of the drug development pipeline; namely fragment elaboration. We show that when provided with a set of bound structures, performing attribution on such structures yielded hotspot maps describing important protein atoms. Using these in unison with a fragment elaboration tool, STRIFE, resulted in improved docking scores for the elaborated molecules compared to when hotspots were obtained with a data-based method. This is further evidence that PointVS is not learning to memorise ligand information but to recognise important interactions, and additionally constitutes the first ML-based method of extracting structural information from a protein target in a way that is useful for fragment elaboration.

More broadly, our work demonstrates that attribution techniques, when applied to debiased models, can be useful for extracting structural information in a way that can be useful in molecule generation.

The success of FBDD depends on access to high-quality structural data and suitable fragment libraries. In settings where this data cannot be generated (say, due to cost), or where exploring a broader swathe of chemical space is desirable, researchers use *de novo* design approaches. These methods aim to generate entirely new molecular structures with desirable properties, offering a complementary path to fragment-based strategies (see chapter 1). Among the most promising frameworks for *de novo* design are diffusion models, which have recently emerged as state-of-the-art generative methods in both chemistry and beyond (Hoogeboom et al. 2022; Ramesh et al. 2021). The following chapter explores how diffusion-based methods can be used for molecular design, and in particular, how the plausibility of the molecules they generate can be improved.

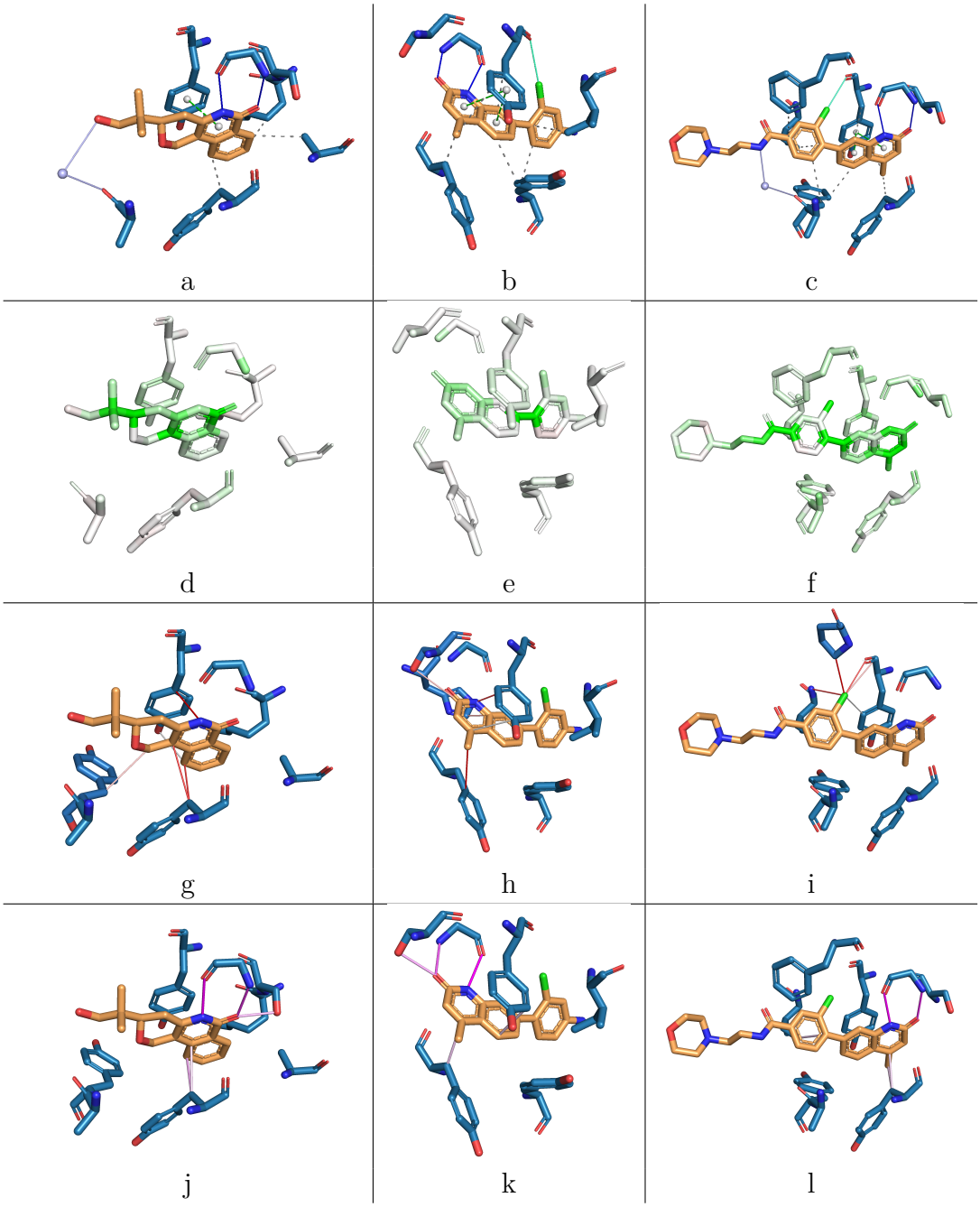


Figure 2.4: Three Tankyrase-2 inhibitors: 5C5P (a, d, g, j), 4J21 (b, e, h, k), and 4J22 (c, f, i, l). The top row shows the PLIP analysis of each structure, where dark blue lines are hydrogen bonds, dotted grey are hydrophobic interactions, dotted green are π-π interactions, solid green are halogen bonds, and lilac are water bridges. The second row shows the results of performing atom masking with gnina, with green representing a positive attribution score (identified as making a positive contribution to binding) and red a negative score (making a negative contribution). The third row shows the results of bond masking with IGN, with lines between atoms showing the top five highest-scoring edges for each structure, with darker red representing higher scores. The bottom row shows the results of edge-attention attribution performed with PointVS, with the lines also showing the top five highest-scoring edges, and darker pink representing higher scores.

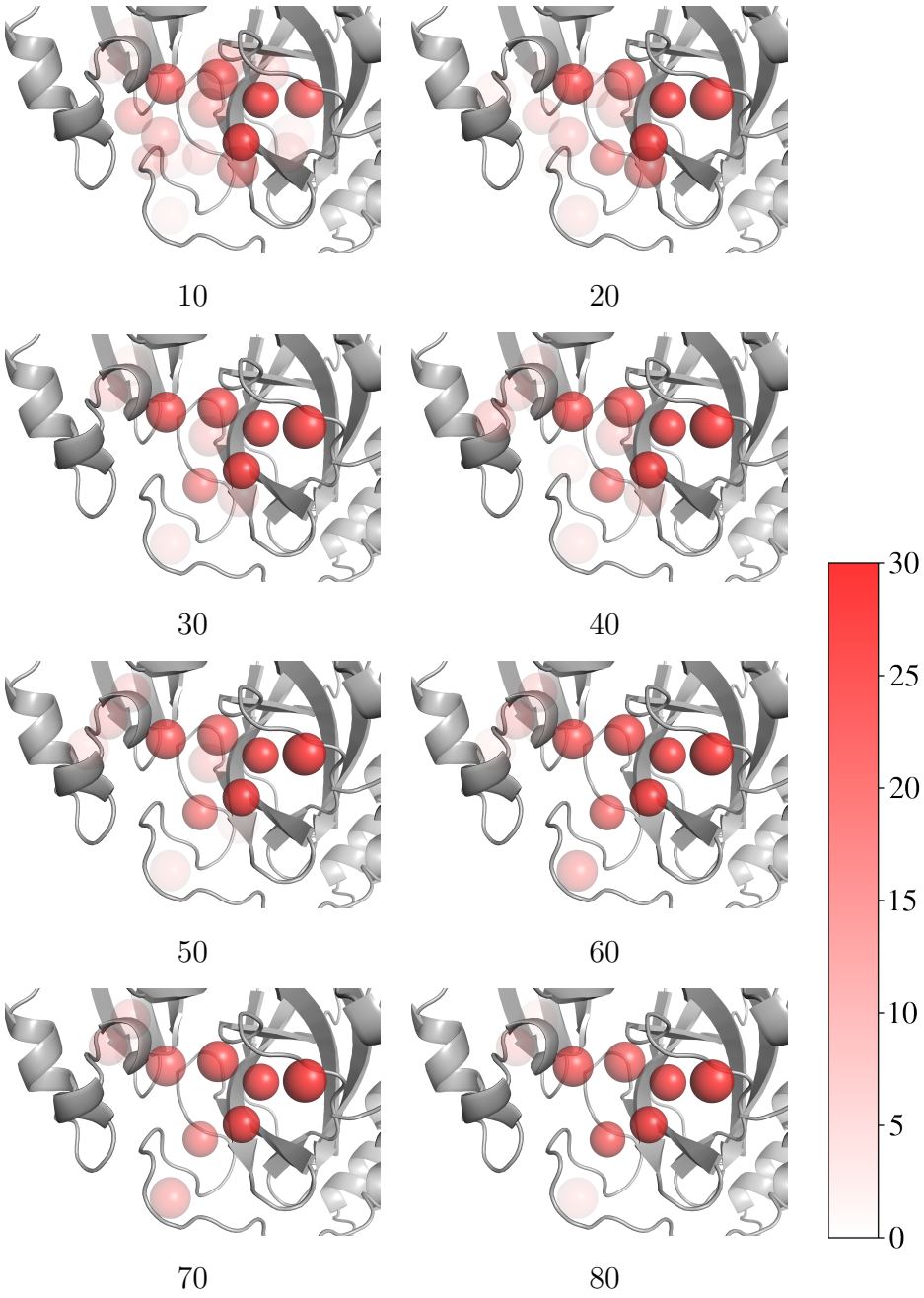


Figure 2.6: Top 5 PointVS donor and acceptor hotspots (both shown in red) obtained in the pocket of Mpro by performing attribution on 40 randomly generated sets of bound fragments. The number of structures constituting each random set is shown below each image. Darker spheres correspond to hotspots which were included in the top 5 more frequently.

3

Improving Structural Plausibility in 3D Molecule Generation via Property-Conditioned Training with Distorted Molecules

Contents

3.1 Preface 68
3.2 Introduction 68
3.3 Methods 70
3.3.1 Generation of 3D molecules 70
3.3.2 Conditioning on conformer quality 72
3.3.3 Assessment metrics 73
3.3.4 Datasets 74
3.4 Results and discussion 76
3.4.1 Ablation tests 76
3.4.2 Conditioning on distortion factor 78
3.4.3 Conditioning on internal energy 80
3.4.4 Testing the conditioning method on additional models . 82
3.5 Conclusions 83

This chapter contains work described in the following publication:

Vost, L.; Chenthamarakshan, V.; Das, P.; Deane, C. M. Improving Structural Plausibility in Diffusion-Based 3D Molecule Generation via Property-Conditioned Training with Distorted Molecules. Digital Discovery 2025, 4 (4), 1092–1099.

3.1 Preface

Unlike fragment elaboration methods like the one covered in the previous chapter, *de novo* drug design involves generating new molecules from scratch. Diffusion models have recently gained popularity as a viable approach to this task thanks to their successful application in fields such as image and video generation (Ramesh et al. 2021; Rombach et al. 2022). However, they have been shown to have issues generating molecules of druglike sizes, struggling in particular with correct bond lengths, angles, and the overall internal energy of the molecules (Harris et al. 2023; Buttenschoen et al. 2024). Addressing the fundamental challenge of generating physically plausible molecules is, therefore, a necessary first step towards effective diffusion-based *de novo* design. In this chapter and the next, we focus on methods to increase the feasibility of druglike molecules generated by diffusion models.

The method we present in this chapter exploits the well-known utility of negative data as a resource in machine learning (Z. Yang et al. 2024; Maloney et al. 2023). By including both positive and negative examples, the dataset teaches the model to recognise the boundaries of the valid chemical space, allowing it to distinguish reasonable structures from physically impossible ones and sample more effectively from the desirable regions.

3.2 Introduction

While many generative molecular design models historically operated in 1D or 2D space (Ross et al. 2024; Chenthamarakshan et al. 2020; Dollar et al. 2021), focus has recently shifted towards developing models capable of directly outputting both atom types and coordinates in 3D. As covered more extensively in Chapter 1, autoregressive models were once prominent in this domain, generating 3D molecules by adding atoms and bonds iteratively (Y. Luo et al. 2022; Spiegel et al. 2020; Peng, S. Luo, et al. 2022). However, such models suffer from an accumulation of errors during the generation process and do not fully capture the complexities of real-world scenarios due to their sequential nature, potentially losing global

context. To address these limitations, recent studies have turned to diffusion models, which iteratively denoise data points sampled from a prior distribution to generate samples. Unlike autoregressive models, diffusion-based methods can simultaneously model local and global interactions between atoms. Nevertheless, diffusion in molecule generation has faced criticism for yielding implausible outputs (Harris et al. 2023; Buttenschoen et al. 2024). There have been ongoing efforts to improve the performance of models trained on small molecules such as those found in the QM9 dataset, and as such the models currently available are capable of reliably generating molecules of this size (Ramakrishnan et al. 2014; Hoogeboom et al. 2022; Morehead et al. 2023; Vignac et al. 2023; L. Huang et al. 2022). However, achieving success in generating larger molecules, as encountered in datasets like GEOM (Axelrod et al. 2022), remains challenging without incorporating additional techniques such as energy minimisation or docking.

In the work presented in this chapter, we focus on enhancing the ability of a diffusion model to generate plausible 3D druglike molecules. To achieve this, we use the property-conditioning method developed by Hoogeboom *et al.* (Hoogeboom et al. 2022). Instead of conditioning a model on pre-existing properties, we condition on conformer quality, training the model to not only generate molecules, but also to distinguish high- and low-quality chemical structures.

To achieve this, we generate distorted versions of each of the three datasets we evaluate the method on: QM9, GEOM, and a subset of ZINC. We sample molecules from each dataset and apply random offsets to their original coordinates, based on a maximum distortion value. Each distorted molecule is assigned a label representing the degree of warping applied and is added back to the dataset. Non-distorted molecules are also labelled, identifying them as high-quality conformers. Using these datasets of molecules with varying levels of quality, we train property-conditioned models, encouraging the model to learn to label molecule validity while simultaneously training it to generate molecules.

First, we evaluate our conditioning method with EDM, the first E(3) equivariant diffusion model for molecule generation (Hoogeboom et al. 2022). We experiment

with two labelling methods: one, a distance-based metric representing the extent of the applied distortion, and two, an internal energy approximation calculated with the eXtended Tight Binding (XTB) program. We then test it on two additional models: a geometry-complete diffusion model (Morehead et al. 2023) and a flow matching method (Y. Song, Gong, et al. 2023), both designed to enhance the structural plausibility of generated molecules. Since existing models already achieve strong performance on QM9, leaving little room or need for improvement, we focus on evaluating our approach on slightly larger, more chemically complex molecules. To this end, we employ two datasets of druglike molecules: the GEOM dataset, and another derived from the ZINC database.

Our findings demonstrate that across the models tested, conditioning a model with low-quality conformers enables it to discern between favourable and unfavourable molecular conformations. This allows us to target the area of the learned space corresponding to high-quality molecules, resulting in an improvement of the validity of generated molecules. More broadly, this demonstrates the potential of supplementing molecule generation methodologies not solely with examples of desired molecules but also with instances exemplifying undesired outcomes.

3.3 Methods

3.3.1 Generation of 3D molecules

Hoogeboom *et al.* (Hoogeboom et al. 2022) introduced the first E(3) equivariant diffusion model (EDM) for generating 3D small molecules. Since then, significant efforts have been made to modify the original EDM, whether to adapt the method for structure-based drug design (Z. Chen et al. 2023; Guan et al. 2023; H. Lin et al. 2022) or to enhance the validity of the generated molecules (Peng, Guan, et al. 2023). Notable examples of the latter include GCDM (Geometry-Complete Diffusion Model) (Morehead et al. 2023) and MolFM (Y. Song, Gong, et al. 2023). GCDM addresses the limitations of diffusion models that rely on molecule-agnostic and non-geometric graph neural networks (GNNs) for 3D graph denoising by introducing a geometry-complete approach. In contrast, MolFM focuses on the

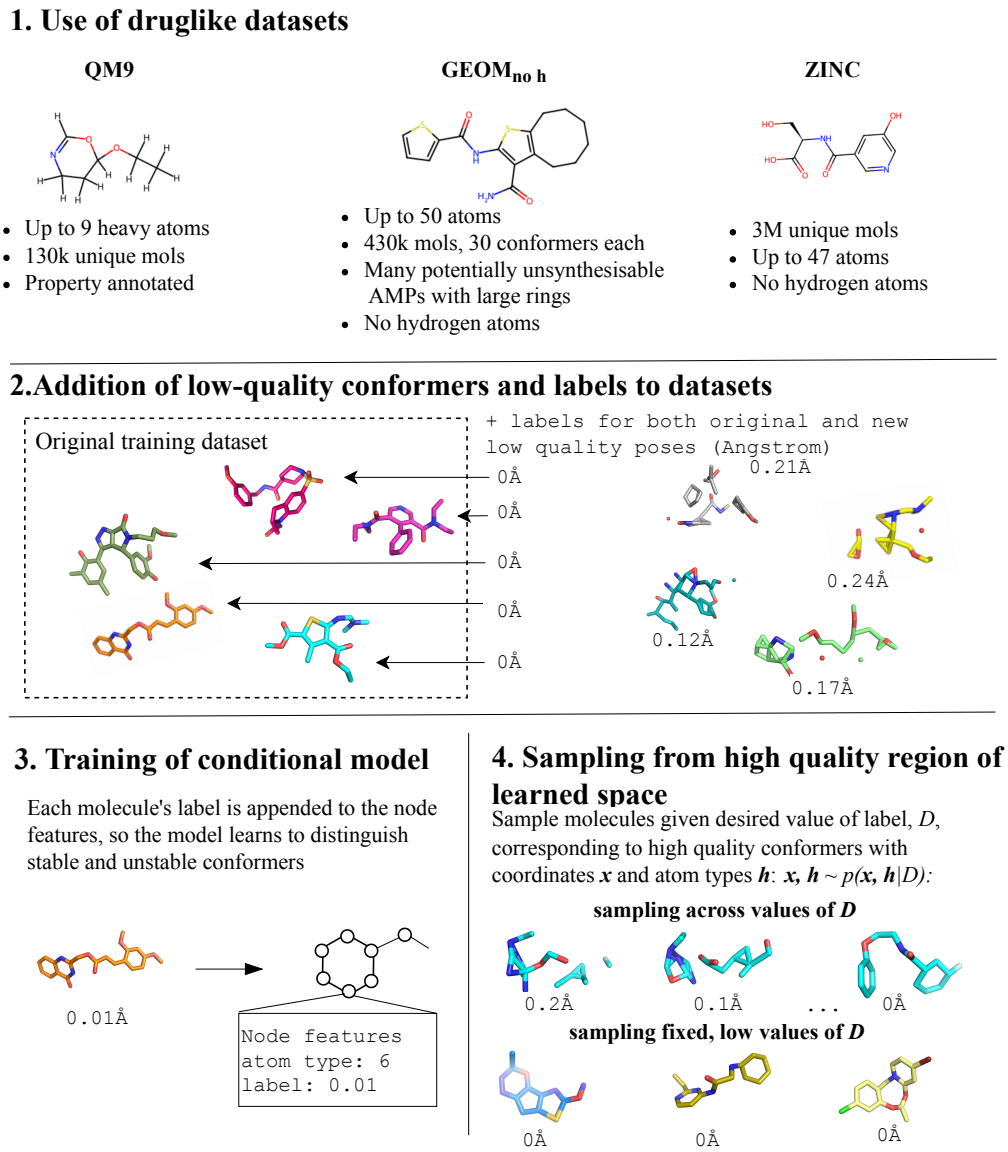


Figure 3.1: An overall schema of the methods used with 1. the datasets used to train both the unconditional and conditional models, 2. the generation of high energy conformers and their addition to the datasets, 3. the training of the conditional model and 4. the conditional inference

issue of unstable probability dynamics in existing diffusion models by incorporating geometric flow matching, merging the advantages of equivariant modelling with stabilised probability dynamics.

3.3.2 Conditioning on conformer quality

The authors of EDM developed an extension to their method to carry out conditional molecule generation. In this instance, property annotations are included alongside each of the molecules in the training dataset, and at inference, molecules can be generated with a desired value of this property. We use this property-conditioning method to train models conditioned on conformer quality. To implement this, we first generated datasets with 3D conformers of molecules of variable quality levels, and corresponding annotations. We generated distorted versions of a subset of molecules from each of the datasets we used. For each molecule, its 3D coordinates, represented as $C = \{(x_i, y_i, z_i)\}$ where i denotes the atom index, were obtained. Subsequently, a random number D within the range of 0 to D_{max} angstroms, labelled as the maximum distortion, was sampled:

$$D \sim U(0, D_{max})\text{Å}$$

This value represents the maximum distance in angstrom that could be added to atoms in that molecule: in other words, the sampled distortion value determines the maximum extent of perturbation to be applied to the molecule's structure. Following this, random offsets were generated within the range of 0 to the sampled distortion, D, for each dimension of every atom's coordinates:

$$\text{offset}_x, \text{offset}_y, \text{offset}_z \sim U(-D, D)$$

These offsets were then applied to the original coordinates:

$$s_{x_i} = x_i + \text{offset}_x;\ s_{y_i} = y_i + \text{offset}_y;\ s_{z_i} = z_i + \text{offset}_z$$

Resulting in a 'distorted' version of the molecule. This distorted molecule, along with its corresponding sampled distortion value D, was subsequently added to the training set. Following the generation of the distorted datasets, we use the property-conditioning training protocol outlined by Hoogeboom *et al.* to train

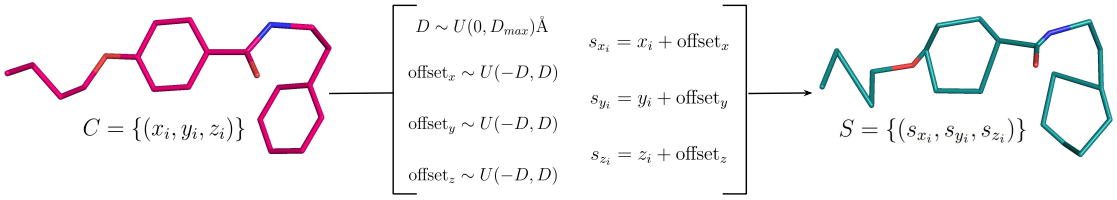


Figure 3.2: Diagram depicting the process of coordinate distortion for a molecule in three-dimensional space. The process involves the following steps: first, sampling the maximum distortion (D) from a uniform distribution between 0 and D_{max} angstrom. Second, generating random offsets within the range of [-D, D] for each dimension of the original coordinates, C. Third, applying these offsets to the original coordinates of each atom in the molecule, resulting in a distorted conformer, S.

on them, using the distortion factor D as the property of interest, and follow the sampling protocol to generate molecules corresponding to $D = 0$Å .

When testing the internal energy conditioning approach, we used the same distorted dataset and employed XTB to calculate approximate internal energy values for all molecules, both distorted and non-distorted. We then conditionally sampled from the model, this time enforcing $D = E_{min}$, where E_{min} is a fixed value for each dataset corresponding to the lowest internal energy annotation of any molecule in it.

3.3.3 Assessment metrics

For each model architecture, we follow their respective guidelines for training and sampling models, with the exception of cases for which pretrained models have been provided (EDM - QM9, GCDM - $\text{GEOM}_{\text{no h}}$, and MolFM - $\text{GEOM}_{\text{no h}}$). We then use each trained model to generate 1000 molecules, with the exception of models in the ablation tests, where we sampled 100 molecules per model to efficiently screen the effects of varying both distortion magnitudes and the ratio of distorted to non-distorted molecules. Since these models output only atom types and coordinates, we adhered to the standard practice (Hoogeboom et al. 2022; Gebauer et al. 2019) of using OpenBabel (O'Boyle et al. 2011) to assign bonds based on interatomic distances. These post-processed molecules were first passed through RDKit's sanitisation checks, giving an RDKit sanitisation pass rate. All molecules were then evaluated using the PoseBusters test suite, which begins with its own sanitisation step - molecules failing this initial check automatically fail all subsequent

tests. PoseBusters then tests the physical validity of the molecule by bond pattern matching, correct tetrahedral chirality specification, and appropriate double bond stereochemistry. The suite also assesses intramolecular validity through multiple geometric checks: bond lengths must fall within 0.75-1.25 times the expected bounds from distance geometry, bond angles must similarly be within 0.75-1.25 times their expected ranges, aromatic rings (5 or 6-membered) must maintain planarity within 0.25 Å , and aliphatic carbon-carbon double bonds must show appropriate planarity with their neighbouring atoms (within 0.25 Å). Additionally, PoseBusters checks for internal steric clashes by ensuring interatomic distances between non-covalently bound atoms exceed 0.7 times their lower geometric bounds. Finally, it evaluates energetic feasibility by comparing the molecule's calculated energy (using RDKit's UFF) against an ensemble of 50 conformations generated via ETKDGV3 and relaxed using force field optimisation. We report both the initial RDKit sanitisation pass rate and the number of molecules that successfully pass all seven non-sanitisation PoseBusters tests. Additionally, we assess structural diversity among the generated molecules using the MOSES framework (Polykovskiy et al. 2020), and check for the presence of undesirable functional groups in each molecule using the REOS (Rapid Elimination of Swill (W. P. Walters, Murcko, et al. 1999; W. P. Walters, Stahl, et al. 1998)) functionality in the `useful_rdkit_utils` toolkit (P. Walters 2025) (see Appendix B.3). For every statistic, we provide 95% confidence intervals computed using SciPy's `stats.bootstrap` with 1,000 samples and the 'percentile' method (Virtanen et al. 2020).

3.3.4 Datasets

To ensure consistency in our comparisons with pretrained baseline models, we use the same dataset splits and versions as those in the EDM paper. For QM9, this corresponds to the train/validation/test split introduced by Anderson et al. (Anderson et al. 2019), based on the 2014 version of QM9 (Ramakrishnan et al. 2014). For GEOM, we follow the split introduced by the EDM authors (Hoogeboom et al. 2022), which uses the 2022 version of the dataset (Axelrod et al. 2022).

To maintain consistency and allow for comparison with the models trained on these datasets, we follow the same splitting regime as proposed by Anderson *et al.* (Anderson et al. 2019) for our ZINC subset.

3.3.4.1 QM9

The QM9 dataset (Ramakrishnan et al. 2014) is a widely used benchmark dataset in quantum chemistry and machine learning research. It consists of quantum-mechanical properties and 3D conformers of 130,000 small organic molecules with an average of 17.5 atoms (8.2 heavy atoms).

The QM9 dataset has been extensively used to develop and validate machine learning models for molecular property prediction. However, it has also recently become the central benchmark for *de novo* molecule generation, particularly in the development of diffusion models (Hoogeboom et al. 2022; Morehead et al. 2023), and as such, new diffusion models are capable of reliably generating molecules similar to those found in QM9.

3.3.4.2 GEOM

While QM9 features only smaller-than-druglike molecules, GEOM (Axelrod et al. 2022) is a larger-scale dataset of molecular conformers. It features 430,000 molecules, of which 317,928 are mid-sized organic molecules from AICures and MoleculeNet (Wu et al. 2018), and 133,258 molecules are from QM9, resulting in an average molecule size of 44.4 atoms (20.1 heavy atoms). For each molecule, a variable number of conformers are given along with their approximate internal energy as calculated with the XTB (Bannwarth et al. 2021) program. From this dataset, Hoogeboom *et al.* (Hoogeboom et al. 2022) retain the 30 lowest energy conformations for each molecule in their work.

Similar to Peng et al. (Peng, Guan, et al. 2023), we use a version of GEOM from which hydrogen have been removed ($\text{GEOM}_{\text{no h}}$), as the positions of hydrogen atoms can often be inferred with a high level of confidence (He et al. 2013). This not only reduces the computational demand of training, but also facilitates more effective learning of heavy atom placements. This leads to the $\text{GEOM}_{\text{no h}}$ dataset

becoming the quickest to train on among the druglike datasets. We therefore use the $\text{GEOM}_{\text{no h}}$ dataset for conducting ablation tests.

3.3.4.3 ZINC

ZINC (Sterling et al. 2015) is a database of commercially-available compounds containing over 230 million purchasable compounds in ready-to-dock, 3D formats.

We generate a training set by selecting a subset of 660,000 molecules from the druglike catalogue of the ZINC database. Unlike GEOM, this subset is curated without repeat conformers. Hydrogen atoms are not included, and the average molecule comprises 26.8 heavy atoms.

3.4 Results and discussion

This section evaluates our proposed quality conditioning method. First, we the present ablation tests carried out on the $\text{GEOM}_{\text{no h}}$ dataset to determine the optimal ratio and distortion level for our negative training examples. Next, using these optimised parameters, we evaluate the performance of a conditional EDM on the QM9, $\text{GEOM}_{\text{no h}}$, and ZINC datasets, comparing it directly to the unconditional model. Finally, we assess the broader applicability of our quality conditioning approach by testing its effectiveness on two different models, GCDM and MolFM.

3.4.1 Ablation tests

To identify the optimal proportion of distorted molecules and the required degree of distortion for effective conditional training, we performed ablation studies using the $\text{GEOM}_{\text{no h}}$ dataset. This dataset was selected due to its inclusion of drug-like molecule sizes, unlike QM9, whilst being a more computationally tractable set to train than the ZINC dataset. For these ablation studies, we sampled 100 molecules per model to efficiently screen the effects of varying both distortion magnitudes and the ratio of distorted to non-distorted molecules.

We introduced varying numbers of distorted molecules at different distortion levels (ranging from 0Å , indicating no distortion, to the maximum distortion,

D_{max}Å) into the original $\text{GEOM}_{\text{no h}}$ dataset. We defined dataset ratios based on the number of distorted and original molecules: for example, a 1:50 ratio indicates one distorted molecule was added for every fifty original molecules. We evaluated each model's performance by training conditioned models and sampling 100 molecules, ensuring that the samples were from the low-distortion-factor region of the learned space (formally, enforcing $D = 0$Å).

The model trained on a dataset with a ratio of 1:50 distorted molecules and a maximum distortion of 0.25Å exhibited the joint highest RDKit parsability rate of 97%, and the highest PoseBusters pass rate at 81%. While several models reached 97% RDKit sanitisation rates (namely 1:20, $D_{max} = 0.5$Å and 1:50, $D_{max} = 0.5$Å), these models exhibited slightly lower PoseBusters pass rates (75% and 78%, respectively). Increasing or decreasing D_{max} further resulted in PoseBusters performance decreasing across all ratios, primarily due to failures in the internal energy test.

This observation suggests that if the training includes molecules that are too distorted, the model does not effectively learn to distinguish between subtly flawed and acceptable molecular structures. Distorted molecules should therefore still bear some resemblance to realistic conformers, albeit with deliberately infeasible bond lengths and angles. On the other hand, insufficient distortion compromises the effectiveness of the conditioning classifier, and the models struggle to distinguish between high-quality and low-quality conformations, leading to poor performance in generating desirable molecules.

These results demonstrate the concept of conditioned training on negative data, and give an idea of the extent of distortion and frequency of distorted molecules to add. We used a ratio of 1:50, and $D_{max} = 0.25$Å for all subsequent tests, but note that any dataset would likely benefit from different exact values of these parameters.

D_{max} (Å)	Ratio of distorted:non-distorted molecules					
	1:20		1:50		1:100	
	RDKit Sanitisation, %	PoseBusters Pass Rate, %	RDKit Sanitisation, %	PoseBusters Pass Rate, %	RDKit Sanitisation, %	PoseBusters Pass Rate, %
0.1	96 (92 –99)	73 (64 –81)	96 (92 –99)	77 (69 –85)	96 (92 –99)	77 (68 –85)
0.25	95 (90 –99)	52 (42 –62)	97 (92 –99)	81 (73 –88)	96 (92 –99)	77 (68 –85)
0.5	97 (93 –100)	75 (66 –83)	97 (93 –100)	78 (70 –86)	95 (90 –99)	68 (59 –77)
1	93 (88 –97)	57 (47 –67)	89 (78–97)	54 (38–70)	62 (52 –71)	8 (3 –14)

Table 3.1: Performance comparison of EDM trained conditionally on $\text{GEOM}_{\text{no h}}$ using a distortion factor, D, and sampled with D=0Å across various ratios of distorted:non-distorted molecules and maximum distortion values in angstrom. 95% confidence intervals are shown in brackets.

We also examined the quality of molecules generated when sampling from the low-quality region of the learned space (formally, $D = D_{max}$Å). The results of this are shown in Appendix B.1. The molecules sampled using D= D_{max}Å are, as expected, worse than both the conditioned models and the baseline model in terms of PoseBusters pass rates, with the highest reaching only 53%. This poor performance is mainly attributed to failures in the internal energy test.

Having established the parameters for conditional training datasets in terms of quantity of distorted molecules and extent of distortion, we move on to testing this approach with the other datasets, QM9 and ZINC.

3.4.2 Conditioning on distortion factor

Following the results of the ablation tests, we generated conditional versions of each dataset using parameters $D_{max} = 0.25$Å and a distorted:non-distorted molecule ratio of 1:50. After training conditional models on these modified datasets, we sampled 1,000 molecules from each and evaluated their quality using RDKit and PoseBusters. The results of this evaluation are presented in Table 3.2.

Our baseline analysis reveals a clear relationship between molecule size and model performance. The QM9 dataset, comprising molecules smaller than 9 heavy atoms had the highest baseline performance with RDKit and PoseBusters pass rates of 92.2% and 81.1%, respectively. The non-conditioned EDM model performed exceptionally well with QM9, surpassing all conditional variants across both evaluation metrics. This superior performance may be attributed to EDM's specific development for QM9, coupled with the dataset's smaller molecular size,

which appears to enable better discrimination between high-quality and low-quality conformers without requiring examples of the latter.

The $\text{GEOM}_{\text{no h}}$ dataset also showed strong baseline performance as assessed with RDKit, with generated molecules achieving a pass rate of 84.7%. However, the more stringent PoseBusters tests presented more of a challenge, with generated molecules presenting a pass rate of 62.2% for PoseBusters. This can particularly be attributed to the tests concerning bond lengths (65.6% pass rate), bond angles (73.8%), connectivity (74.4%) and internal energy (75.2%). Conditional training improved these metrics substantially, reaching pass rates of 96.6% (bond lengths), 93.7% (bond angles), 89.4% (connectivity) and 87.7% (internal energy). Minor improvements were also seen in the steric clash tests (97.8% to 100%).

The baseline model trained on the ZINC subset exhibited markedly lower performance than the other two datasets, with RDKit sanitisation pass rates of 70.6% and PoseBusters pass rates of 40.0%. The most prevalent failures occurred in bond lengths and atom connectivity, with pass rates of only 65.0% and 62.4%, respectively. This performance decline relative to $\text{GEOM}_{\text{no h}}$ may be attributed to two factors. Firstly, the ZINC dataset's increased diversity, featuring unique conformers rather than multiple conformers per molecule (as found in $\text{GEOM}_{\text{no h}}$), may cause the model to prioritise learning atom types over optimising 3D conformer generation. Secondly, and perhaps more significantly, the compositional differences between datasets may play a crucial role. While the ZINC subset exclusively contains medium-sized compounds, $\text{GEOM}_{\text{no h}}$ incorporates the entire QM9 dataset, resulting in a smaller average molecule size.

Conditional training on the ZINC dataset yielded improved RDKit sanitisation rates and PoseBusters scores. The most notable improvement was observed in the ZINC model's atom connectivity pass rate, which increased from 62.4% to 95.7%, but improvements were also seen in the pass rates of the tests assessing bond lengths (65.0% to 94.1%), bond angles (75.3% to 93.7%), internal steric clash (79.5% to 99.0%), and internal energy (78.5% to 89.4%).

These results indicate that the baseline EDM's performance degrades when generating larger molecules, often yielding implausible structures. Our conditioning method at least partially helps mitigate this decline, increasing the output proportion of high-quality molecules. We next investigate the use of a more scientifically grounded labelling of the molecules—specifically, using internal energy approximations—for this conditioning process.

3.4.3 Conditioning on internal energy

The XTB program (Bannwarth et al. 2021) is a computational chemistry software package used for molecular modelling and simulations. It is based on the semi-empirical tight-binding approach, which approximates the electronic structure of molecules using a simplified set of parameters derived from quantum mechanics. XTB extends traditional tight-binding methods by incorporating additional empirical corrections to improve the accuracy of calculated properties. It is well-suited for studying large molecular systems where the computational cost of more accurate methods such as density functional theory (DFT) becomes prohibitive.

We use the default implementation of XTB (GFN2-xTB) to perform singlepoint energy calculations for each of the molecules, both distorted and original, for QM9, $\text{GEOM}_{\text{no h}}$, and our ZINC dataset. For the original GEOM dataset, the original conformers have energy annotations calculated with the same method, so we only carry out this calculation for the distorted molecules. We carry out this process using two different annotation types: a 'distortion factor', a quantity that represents the extent to which the coordinates have been altered, and an internal energy value, obtained by scoring both original and distorted conformers with the extended tight binding program (XTB) (Bannwarth et al. 2021).

We've observed that for medium-sized, drug-like compounds, conditioning on a distance-based distortion factor results in improvements for RDKit sanitisation and PoseBusters tests. To investigate whether using a potentially more meaningful label (specifically, an internal energy value obtained using XTB) improves the conditioned models, we used the datasets generated with the previous distortion

process. The distorted molecules were then passed through an XTB single-point energy calculation. The same calculation was applied to all high-quality, non-distorted molecules in all datasets, excluding the original GEOM dataset, which already has energy annotations calculated with XTB. We then trained each of the models conditionally, and once trained, we sampled 100 molecules from each, this time enforcing $D = E_{min}$, where E_{min} is a fixed value for each dataset corresponding to the lowest internal energy annotation of any molecule in it. Then we once again assessed all 100 molecules using RDKit and PoseBusters (Table 3.3).

As observed when conditioning on distortion factor, the model trained on QM9 generated lower quality molecules when conditioning with internal energy was carried out. Conversely, conditionally training EDM on $\text{GEOM}_{\text{no h}}$ with internal energy resulted in increased performance compared to both the baseline and the conditioned model using the distance distortion factor. The RDKit and PoseBusters pass rates reached 98% and 84%, respectively. The model trained on ZINC, however, exhibited a decrease in both RDKit and PoseBusters pass rate. These pass rates were also lower than those of the molecules generated with the ZINC model trained conditionally on distortion factor.

When using the distance-based distortion factor, we enforced D=0Å when sampling, as this represented molecules that had not undergone any distortion. The distortion factor, ranging from 0Å to D_{max}Å , provides a straightforward measure of how much the structure of a molecule has been altered. In contrast, internal energy values, although physically meaningful, are challenging to compare directly between different molecules. Each molecule can have a low internal energy corresponding to a high-quality conformer and a high one corresponding to a low-quality conformer, but the lowest energy conformers of some molecules may still have higher energy values than the highest energy annotations of others. This inconsistency likely contributed to the observed performance decreases when conditioning on internal energy.

Dataset		RDKit	PoseBusters
QM9	baseline	**91**%	**81**%
	XTB	57%	19%
$GEOM_{no\ h}$	baseline	90%	70%
	XTB	**98**%	**84**%
ZINC	baseline	64%	**40**%
	XTB	**65%**	36%

Table 3.3: Performance comparison of EDM trained on diverse molecular datasets using a baseline (the unconditional EDM), and using a conditioned model, for which the model is trained on an XTB internal energy estimate. The highest performance for each dataset is shown in bold.

3.4.4 Testing the conditioning method on additional models

To evaluate the broader applicability of our distortion-conditioning method, we apply it to two other models: GCDM (Morehead et al. 2023) and MolFM (Y. Song, Gong, et al. 2023). The performance of these models when trained on $GEOM_{no\ h}$ and ZINC is presented in Tables 3.4a and 3.4b, respectively.

		RDKit Sanitisation, %	PoseBusters Pass Rate, %
Baseline	$GEOM_{no\ h}$	100.0 (100.0-100.0)	77.8 (75.2-80.4)
	ZINC	56.3 (53.3–59.3)	40.8 (38–43.6)
Conditional	$GEOM_{no\ h}$	99.9 (99.7-100.0)	79.7 (77.2–82.1)
	ZINC	97.2 (95.8–98.4)	66.5 (62.5–70.4)

(a) GCDM

		RDKit Sanitisation, %	PoseBusters Pass Rate, %
Baseline	$GEOM_{no\ h}$	98.6 (97.5–99.6)	80.8 (77.3–84.1)
	ZINC	72.0 (69.2–74.8)	42.3 (39.2–45.4)
Conditional	$GEOM_{no\ h}$	94.5 (93.0–95.9)	46.8 (43.7–49.8)
	ZINC	93.3 (91.6–94.9)	45.9 (42.6–49.1)

(b) MolFM

Table 3.5: Performance comparison of GCDM and MolFM when trained on $GEOM_{no\ h}$ and our ZINC subset using the default setup (baseline) or conditionally trained on distortion factor using a dataset generated with $D_{max} = 0.25$Å and 1:50 distorted molecules, and sampled with D=0Å . 95% confidence intervals are shown in brackets.

For the GEOM dataset, the GCDM conditional model shows very marginal improvements in PoseBusters performance over the baseline (mostly due to the internal energy test, for which the pass rate increases from 86.2% to 88.6%). However,

since baseline performance is already high, conditioning has limited overall impact. GCDM trained on the ZINC subset, on the other hand, shows a much more substantial improvement with conditioning. The baseline model struggles with the connectivity of molecules, which increases from 63.7% in the baseline model to 94.4% when our conditioning method is applied, resulting in an overall boost in both RDKit sanitisation and PoseBusters pass rate.

The conditional MolFM model showed decreased performance on $\text{GEOM}_{\text{no h}}$, with PoseBusters pass rates declining from 80.8% in the baseline to 46.8% in the conditional model, primarily due to connectivity failures. The reason for this degradation is unclear, though it may relate to fundamental differences between flow matching and diffusion-based architectures in how they learn molecular distributions. Interestingly, MolFM showed modest improvements on ZINC, with pass rates increasing from 42.3% to 45.9%, suggesting the interaction between our conditioning method and flow matching may be dataset-dependent. Understanding why quality conditioning is effective for diffusion models (EDM, GCDM) but inconsistent for flow matching approaches represents an important direction for future work.

3.5 Conclusions

In this work, we have demonstrated the effectiveness of including low-quality conformers in a training set and conditioning a diffusion model on a label representing conformer quality to enhance the generation of high-quality druglike molecules. By leveraging datasets derived from GEOM and ZINC, alongside a conditioning method proposed by Hoogeboom *et al.*, we have successfully improved the validity of generated molecules. Our approach, which focuses on sampling molecules with labels corresponding to low distortion factors, leads to enhancements in RDKit parsability and validity as assessed by PoseBusters for the original EDM, as well as for a subsequent diffusion model, GCDM. We also explored conditioning on internal energy calculated with XTB as an alternative to the distortion factor, but found that the distance-based distortion metric proved more effective, likely due to challenges in comparing absolute energy values across diverse molecular structures.

The method shows strongest improvements for diffusion-based models, particularly those built on the EDM framework, which comprise a significant portion of current 3D molecule generation approaches. Testing on MolFM, a flow-matching architecture, yielded inconsistent results across datasets, suggesting that the conditioning approach may interact differently with non-diffusion-based generative models. The approach is most effective for datasets containing larger, drug-like molecules, as demonstrated by our results with $\text{GEOM}_{\text{no h}}$ and ZINC. This can be explained mechanistically; larger molecules have more complex conformational spaces where explicit examples of invalid states help define the boundary between high-quality and low-quality conformers. Conversely, the method provides limited benefits for datasets like QM9 where molecules are small (fewer than 9 heavy atoms) and have constrained conformational spaces. In these cases, it appears models can effectively learn to distinguish valid from invalid conformations from the training data alone.

Our findings underscore the importance of considering the quality of conformers in molecule generation processes. The results show that by training models to discern between favourable and unfavourable molecular conformations, we can selectively sample from the high-quality region of learned space, resulting in significant improvements in the validity of generated molecules.

Moving forward, further research could explore why quality conditioning is effective for certain architectures but not others, and investigate alternative labelling schemes that might generalise across different generative modelling approaches. Additionally, investigating the applicability of our approach to other areas of molecular design and exploration could yield valuable insights for drug discovery and beyond. Overall, our study provides a promising avenue for generating valid drug-sized molecules efficiently and effectively using diffusion-based architectures.

Having established the effectiveness of this strategy, we next explore a complementary approach that addresses the same challenge from a different angle. The following chapter covers a fragment-based method that aims to enforce validity not by refining the training data, but by directly constructing molecules from pre-validated chemical building blocks during the generation process itself.

Dataset	RDKit Sanitisation Pass Rate, %	Posebusters Pass Rate, %					
		All Atoms Connected	Bond Lengths	Bond Angles	Internal Steric Clash	Internal Energy	All Tests Passed
Baseline							
QM9	92.2 (90.5–93.8)	100.0 (100.0–100.0)	100.0 (100.0–100.0)	99.9 (99.7–100.0)	100.0 (100.0–100.0)	88.1 (85.9–90.1)	81.1 (78.7–83.5)
$GEOM_{no\ h}$	84.7 (82.5–86.9)	74.4 (71.7–77.1)	65.6 (62.5–68.7)	73.8 (70.8–76.7)	97.8 (96.7–98.7)	75.2 (72.3–78.0)	62.2 (58.3–66.1)
ZINC	70.6 (67.8–73.3)	62.4 (59.4–65.3)	65.0 (61.5–68.4)	75.3 (72.2–78.5)	79.5 (76.6–82.4)	78.5 (75.5–81.4)	40.0 (37.0–43.0)
Distortion factor conditioning							
QM9	65.0 (62.0–68.0)	84.0 (81.7–86.2)	96.6 (95.2–98.0)	95.8 (94.3–97.2)	98.9 (98.0–99.7)	91.1 (88.8–93.2)	46.9 (43.9–50.0)
$GEOM_{no\ h}$	92.4 (90.7–94.0)	89.4 (87.5–91.3)	96.6 (95.5–97.8)	93.7 (92.1–95.2)	100.0 (100.0–100.0)	87.7 (85.5–89.7)	68.8 (65.9–71.1)
ZINC	95.3 (93.8–96.6)	95.7 (94.3–96.9)	94.1 (92.5–95.6)	93.7 (92.1–95.2)	99.0 (98.4–99.6)	89.4 (87.4–91.4)	74.5 (71.7–77.3)

Table 3.2: Performance comparison of EDM trained on diverse molecular datasets using the baseline model with no conditioning, and EDM conditionally trained on distortion factor using a dataset generated with $D_{max} = 0.25$Å and 1:50 distorted molecules, and sampled with D=0Å . Pass rates are shown for RDKit sanitisation, five PoseBusters subtasks, and overall PoseBusters pass rates. The two omitted subtasks, which had identical results across all six models, are provided in Appendix B.2. 95% confidence intervals are shown in brackets.

4

Molecule diffusion with fragment blocks

Contents

4.1 Preface 87
4.2 Introduction 88
4.3 Methods 90
4.3.1 Model 90
4.3.2 Data 92
4.3.3 Assessment metrics 97
4.4 Results 97
4.4.1 Centroid-only method 97
4.4.2 Plane alignment method 98
4.4.3 Addition of rotational information 99
4.5 Related Work 99
4.6 Conclusions 101

4.1 Preface

As highlighted in the previous chapter, diffusion models applied to molecule generation tasks frequently produce molecular structures with physically unrealistic bond geometries. Post-processing techniques and filtering have been used as a practical workaround, but do not solve the underlying problem: the inability of diffusion models to learn fundamental chemical principles.

A possible explanation for this is that the generative task itself (arranging

individual atoms in 3D space while respecting physical laws) is too complex for the models to learn. This chapter documents an attempt to test this hypothesis and help resolve the issue by simplifying the problem. The proposed method uses pre-assembled, chemically valid fragments, or 'blocks', as the primary building units instead of individual atoms. The goal was twofold: first, to reduce the geometric complexity the model needed to handle, and second, to potentially improve the synthesisability of the resulting molecules by using known chemical fragments. However, this approach presented its own challenges, particularly in developing an effective spatial representation for these non-symmetrical, inconsistent blocks. This chapter details the different approaches tested, and then provides a brief review of the relevant papers that have been released since this work was carried out.

4.2 Introduction

Generative diffusion models have demonstrated remarkable success in diverse domains, for example, being capable of producing high-fidelity images and text from latent noise (Rombach et al. 2022; Ramesh et al. 2021). However, their application to 3D generative chemistry has been met with challenges. A primary limitation of current atomistic diffusion models is their frequent generation of molecules with unrealistic bond lengths and angles, resulting in chemically and physically implausible structures (Alakhdar et al. 2024). While computational methods such as energy minimisation or redocking can be employed as post-processing steps to correct these geometric inconsistencies, they do not address the fundamental deficiency: that the models have failed to internalise the basic principles of molecular geometry and stereochemistry. This failure raises questions about their utility for more sophisticated tasks, such as SBDD, as a model that cannot generate small molecules obeying the laws of physics is unlikely to effectively extract and utilise the binding information of a protein pocket.

This chapter covers an alternative approach designed to mitigate this core problem by fundamentally simplifying the generative task. We hypothesise that the difficulty of arranging a large number of individual atoms in continuous 3D space is

a primary contributor to the poor quality of generated geometries. Consequently, we propose a shift in the generative unit from atoms to molecular fragments or 'blocks': rigid, common substructures found within druglike small molecules. This fragment-based approach is motivated by two key potential benefits.

The first and most central benefit is the reduction of the problem's dimensionality. By assembling molecules from larger, pre-validated substructures, the model is no longer required to learn the precise bond length and angle for every atomic connection. Instead, it must learn the relative position and orientation of a smaller number of blocks. The ineffectiveness of many atom-based models suggests that the simultaneous optimisation of all atomic coordinates is a task of prohibitive difficulty. Our approach aims to make the problem more tractable by abstracting away the low-level geometric details, allowing the model to focus on the higher-level organisation of the molecule. While this may inherently limit the chemical novelty of the generated outputs, it provides a more constrained and potentially more learnable generative pathway.

The second motivation is grounded in the practicalities of the broader drug discovery pipeline. A persistent criticism of many *in silico* design methods is the poor synthesisability of the molecules they produce, an issue which is in part due to the computational prediction of synthetic accessibility, cost, and complexity remaining a largely unsolved problem (S. Chen et al. 2024; S. Wang, L. Wang, et al. 2023; Bonnet 2012; Abderrahmane et al. 2024; Sanchez-Garcia et al. 2023). A fragment-based method offers a potential avenue to address this; by constructing molecules from a library of known, readily available, and inexpensive building blocks (such as a commercial library like the Enamine REAL database (*Enamine REAL Space - Enamine* 2025)) the resulting compounds could be inherently more amenable to practical synthesis. Although the model could still propose challenging or unfeasible linkages between these blocks, this strategy should, on average, reduce the synthetic cost and complexity of the generated molecular library.

To tackle this task, we adapted the Equivariant Diffusion Model, or EDM (Hoogeboom et al. 2022), the same architecture used in Chapter 3, to operate on

substructural blocks instead of individual atoms. This required several modifications to the modelling pipeline. First, a molecular dataset, GEOM, was processed to deconstruct each entry into a set of constituent fragments. Subsequently, the model's input features were redefined; the categorical variables were expanded to encode the identity of each block type, rather than just atom types. The continuous spatial variables were also modified to represent the position and orientation of these blocks, moving from a simple centroid representation to a more complex descriptor incorporating rotational information. The model was then trained following the same method as the atomic model.

Ultimately, this investigation revealed that replacing the atomistic representation with a substructure-based one is a non-trivial task that presented its own challenges. Despite exploring multiple representational strategies, the highest-performing fragment-based model demonstrated inferior performance compared to the original atomistic baseline. A central difficulty was the geometric complexity of the fragments themselves; they are not simple, symmetric points that can be adequately described by a centroid alone, yet representing their rotation in a consistent and learnable manner across a diverse set of shapes proved difficult. This chapter describes the various approaches tested, a brief overview of the results obtained, and a concluding overview of related work that has since been pursued in this area.

4.3 Methods

4.3.1 Model

4.3.1.1 Atom-based setup

At its core, the EDM model treats a molecule as a collection of atoms, where each atom, i, has two key properties: coordinates, x_i, and categorical information h_i, encompassing discrete features like charge and atom type, represented as one-hot vectors. An entire molecule is therefore a set of these coordinate and feature vectors, and the goal of the model is to learn the joint probability distribution $p(x, h)$.

To learn this distribution, the model uses a diffusion process. It first learns a forward process that systematically corrupts the molecule's two variable types: it

adds Gaussian noise to the continuous coordinates and applies random transitions to the discrete categorical atom types, until only a random cloud remains. The core of the model is an E(3) equivariant neural network trained to reverse this dual noising. Starting from random noise, this network simultaneously denoises the continuous positions and predicts the correct discrete atom types, step-by-step, to generate a new, physically plausible arrangement of atoms.

At inference, the generation of a molecule starts by sampling a cloud of random 3D coordinates from a Gaussian distribution and assigning a random chemical type to each point. This initial noisy state is then iteratively refined by a pre-trained E(3) equivariant network that runs a reverse diffusion process. At each step, this single network takes the molecule's current noisy coordinates and atom types and simultaneously predicts two distinct outputs: the noise vector for the continuous coordinates and the probability distribution of the original, 'clean' discrete atom types. This information is used to update the state to a less noisy version for the next step—calculating the new coordinates and sampling the new atom types from their respective reverse distributions—until the process finishes and yields a complete 3D atomic structure, from which bonds are inferred using Open Babel (O'Boyle et al. 2011).

4.3.1.2 Initial model adaptations

Adapting our diffusion method from an atomistic to a block-based representation involved several key changes.

First, we redefined the coordinates, x_i, to represent block centroids instead of atoms. Our initial model used a simple 3D vector for each block's position, though this was later changed to three 3D vectors to model both position and orientation explicitly.

The feature space was also updated. The one-hot encoding scheme was changed from handling a few atom types (5 different types for QM9 or 16 for GEOM) to at least 50 unique block types (see section 4.3.2.2 for more details), significantly increasing its dimensionality.

The most significant adaptation was to the final reconstruction process. Unlike the atomistic model where bonds are directly inferred from generated atoms using Open Babel (O'Boyle et al. 2011), the block-based model requires an extra step. After sampling, a 3D conformer for each block is generated with RDKit. This conformer is then either translated to the sampled centroid coordinates (centroid-only method) or transformed to align with the plane and in-plane vectors (rotation-aware method) to place and orient it. Finally, after assembling all the positioned blocks, their existing bonds are stripped, and the complete atomistic structure is passed to Open Babel to determine the final molecule.

4.3.2 Data

4.3.2.1 Training datasets

The first decomposition was carried out on the QM9 dataset (Ramakrishnan et al. 2014). However, the small size of QM9's molecules, which average 8.2 heavy atoms (see chapter 3 for an in-depth description of QM9 and GEOM), meant they rarely decomposed into more than two blocks, limiting the model's ability to learn complex arrangements. Consequently, the larger and more structurally diverse GEOM dataset (Axelrod et al. 2022) was adopted. GEOM's inclusion of larger, drug-like compounds provided the necessary molecular complexity to try and train a robust model capable of generating more elaborate multi-block structures.

4.3.2.2 Decomposition

Regular BRICS decomposition The Breaking of Retrosynthetically Interesting Chemical Substructures (BRICS) algorithm (Degen et al. 2008) provides a rule-based method for molecular decomposition by identifying and cleaving 16 types of chemical bonds that correspond to common and synthetically feasible reactions. This procedure breaks down a chemical compound into a set of fragments, which are then classified into 'bricks', significant chemical moieties with at least four heavy atoms, and the 'linkers' that connect them. For this work, the decomposition was implemented using the RDKit library (Landrum et al.

2023). The `BRICS.BreakBRICSBonds` function was first applied to each ligand, followed by `Chem.GetMolFrags` to separate the resulting structure into a list of individual fragment molecules. As per the RDKit implementation, this process leaves isotopically-labelled dummy atoms at the cleavage points. To ensure this additional information wasn't passed into the model, these dummy atoms were removed.

Carrying out BRICS decomposition on the full GEOM dataset results in over 14,000 unique blocks (with just under 30% appearing only once), a prohibitively large number of types to expect the model to learn to place. To try and reduce this, we filtered blocks according to how many times they appeared, ultimately choosing to only keep blocks which arose more than 100 times, resulting in 172 different block types. The molecules in GEOM were then filtered such that any containing blocks outside of these 172 types were removed from the dataset, resulting in a dataset size of 56,281 molecules.

Decomposition with standardised fragments As an initial exploration into a more hierarchical generative strategy, we developed a method to standardise the raw BRICS fragments into a more manageable set of building blocks. The primary motivation for this was to reduce the dimensionality of the categorical feature space for the model's one-hot encoding, avoiding the complexity of treating every unique BRICS fragment as a distinct category. The process began by filtering out molecules containing overly complex fragments, such as those with large ring systems (more than 7 members). All acyclic structures and exocyclic atoms were then decomposed into their constituent individual atoms. Ring systems were preserved as single units but were standardised into a small set of canonical types based on their size and aromaticity. This strategy yielded a very simplified representation where each molecule is described as a collection of these fundamental building blocks (either canonical rings or individual atoms) each retaining its original 3D spatial information. Carrying this process out on the GEOM dataset resulted in a final dataset of 62,746 molecules, described with only 23 block or atom types (shown in Fig 4.2).

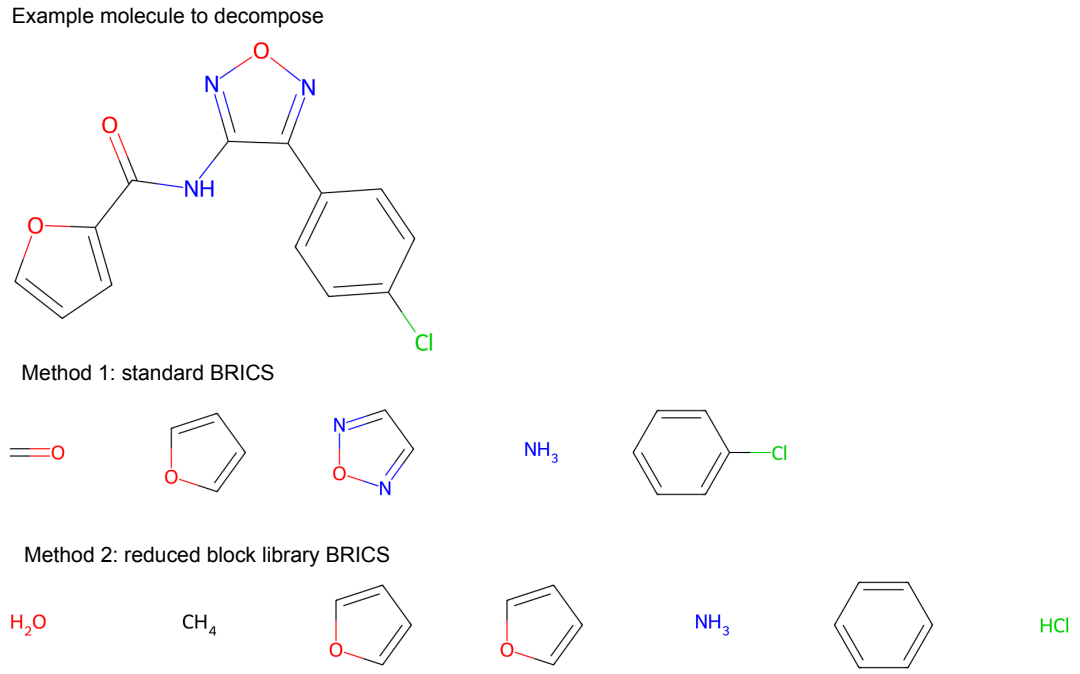


Figure 4.1: An example of a molecule from the GEOM dataset being decomposed using two different methods; the first, the vanilla BRICS implementation, and the second, a BRICS-based approach that swaps ring systems for canonical rings and breaks linkers down into atoms.

This standardisation approach was conceived as the foundation for a potential two-stage generative strategy. In this hierarchical framework, the primary diffusion model would first be responsible for generating a valid molecular scaffold using the simplified vocabulary of canonical rings and individual atoms. A subsequent, specialised process could then be used to refine this basic structure by replacing the generic canonical rings with more specific, chemically diverse analogues. This 'decoration' step could be guided by additional data, such as pharmacophoric hotspot maps from a target protein pocket, to tailor the final molecule to a specific objective. This separation of tasks aims to deconstruct the complex task of molecule generation into the more manageable sub-problems of scaffold assembly and subsequent conditioned refinement.

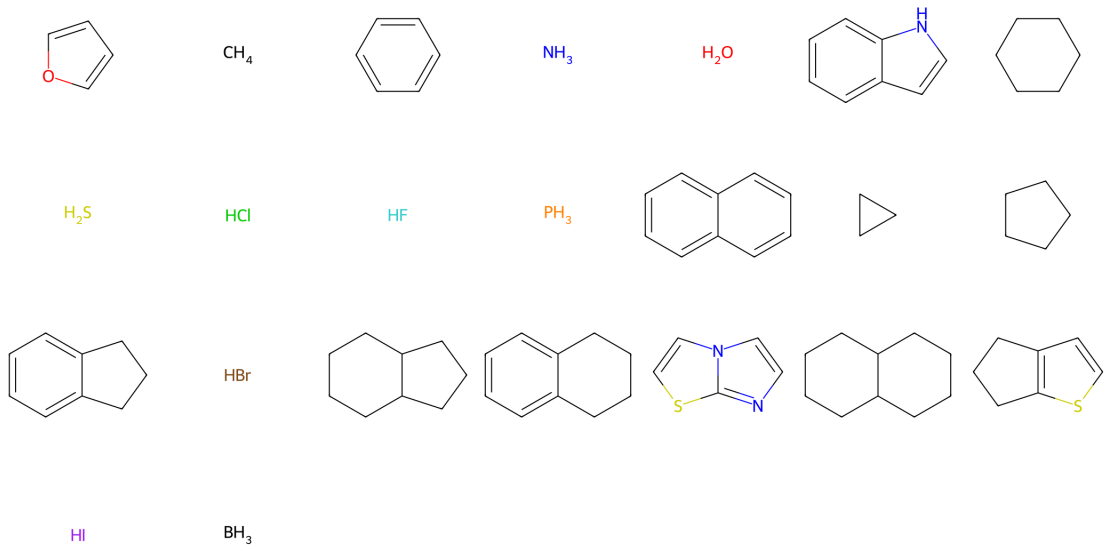


Figure 4.2: All 23 of the block types used in the reduced-library decomposition method

4.3.2.3 Rotational information

To equip the model with a more detailed understanding of each fragment's orientation, we moved beyond a simple centroid representation. Inspired by the local coordinate frames used in models like AlphaFold (Jumper et al. 2021), our goal was to explicitly encode both the position and the 3D orientation of each building block generated by our decomposition strategies.

Each fragment was encoded as three 3D vectors. The first represents the geometric centroid (mean atomic coordinates), while the other two provide an approximate orientation. For fragments with three or more atoms, a best-fit plane was computed from atomic coordinates using the `scikit-spatial` library. The plane normal served as the first orientation vector. To define an in-plane axis, we projected the vector between the two most distant atoms onto the plane. For fused ring systems of different sizes, the direction was defined as pointing from the larger ring to the smaller ring.

This procedure is intentionally simple and comes with limitations. For many fragments there are multiple equally distant atom pairs, so the chosen in-plane axis is not unique. In symmetric cases such as benzene, three equivalent atom pairs

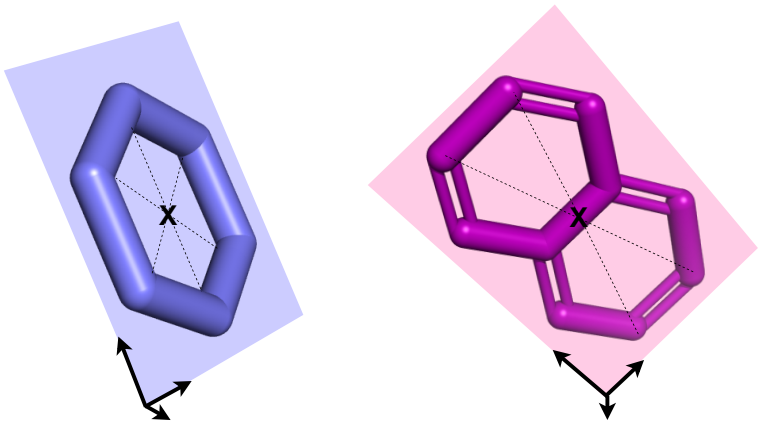


Figure 4.3: An illustration of the heuristic used to assign positional and rotational descriptors. The fragment's position is its geometric centroid (shown as 'X'). Its orientation is defined by a local coordinate system. The primary axis is the normal vector to a best-fit plane. The secondary, in-plane axis is determined by the vector between the two most distant atoms. As the dotted lines indicate, this choice is frequently not unique for symmetric blocks. To resolve this ambiguity, a single vector was chosen at random. This non-deterministic step assumes that the resulting rotational variance is either negligible due to fragment symmetry or introduces an acceptable level of noise into the model's input.

exist, meaning the same fragment can be represented with orientations rotated by 60°from one another. For less symmetric fused systems, two solutions exist, leading to an uncertainty of around 20°. These ambiguities are not unstable but non-deterministic, reflecting the fact that no single canonical orientation can be defined without hard-coded rules for each fragment.

For very small fragments (fewer than three atoms), where no plane can be defined, we assigned placeholder values for the orientation vectors and masked them during training.

Overall, this heuristic is a rough approximation: it does not resolve chirality and introduces rotational ambiguity for many blocks. Nevertheless, it provides a richer description than centroid-only features, enabling the model to access useful information about relative orientation that would otherwise be lost, while remaining tractable for a large and diverse fragment library.

4.3.3 Assessment metrics

To assess the quality of the generated molecules, our initial strategy was to employ the same quantitative metrics used in Chapter 3, namely PoseBusters and RDKit sanitisation checks. However, in practice, the pass rates for both assessments were very low, so most of our analysis shifted to rely predominantly on qualitative visual inspection to diagnose the specific structural and chemical issues responsible for these failures.

4.4 Results

4.4.1 Centroid-only method

Initial experiments were conducted using only the basic BRICS decomposition method, as preliminary results quickly highlighted significant issues arising from the lack of rotational information (see Figure 4.4). To isolate and quantify this problem, we performed a control experiment to establish the theoretical upper bound on reconstruction accuracy for a centroid-only approach. For this, we took a subset of 100 molecules from our pre-processed GEOM dataset and fed their ground-truth BRICS fragments and centroid coordinates directly into our reconstruction pipeline, bypassing the generative model entirely. Even with this idealised input, only 48 of the 100 molecules passed RDKit sanitisation checks.

This 48% success rate reveals a fundamental limitation of the representation itself; even with perfect block placement, the ambiguity in fragment orientation leads to chemically invalid structures more than half the time. This finding provides critical context for the much lower RDKit sanitisation pass rate of approximately 4% observed for molecules generated by the full diffusion model. This lower rate is a product of two compounding factors: the model's own errors in learning the placement task, and the inherent flaws of the centroid-only representation. To address this primary limitation, it became clear that explicitly incorporating rotational information was an essential next step.

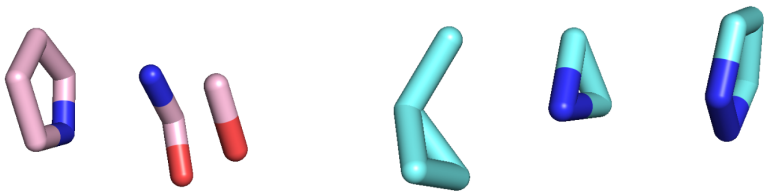


Figure 4.4: Examples of generations from sampling the centroid-only model, with oxygen atoms coloured red, and nitrogen atoms coloured dark blue.

4.4.2 Plane alignment method

As an initial attempt to address the rotational ambiguity observed in the centroid-only approach, we tested a simple geometric post-processing step. Rather than encoding orientation as additional model inputs, this method attempted to standardise fragment orientations during reconstruction.

For each generated fragment, we computed its best-fit plane using the same procedure described above. This plane was then aligned to a global plane—found by fitting a plane to the sampled centroids—using a rigid rotation. The fragment was then translated to place its centroid at the model's predicted coordinates.

This approach aimed to remove rotational degrees of freedom entirely by imposing a standard orientation on all fragments. However, as shown in Table 1, this method achieved only a 10.8% RDKit sanitisation pass rate; an improvement over the 4% of the unaligned centroid-only model, but still substantially worse than the atomic model.

Visual inspection revealed that the primary failure mode was incorrect connectivity between fragments. While individual fragments were now consistently oriented, the fixed canonical orientation was often geometrically incompatible with the correct bonding geometry required to connect adjacent blocks. This result demonstrated that fragment orientation cannot be handled through post-processing alone; the orientation must be learned and predicted by the model as part of the generative process, motivating the development of the full rotation-aware representation tested next.

4.4.3 Addition of rotational information

Following the analysis of the centroid-only models, we incorporated the position and orientation vectors into two existing frameworks: the full BRICS decomposition and the reduced-library BRICS method. To directly measure the impact of this new representation, we first repeated the reconstruction control experiment described previously. Using the ground-truth coordinates and the new orientation vectors for 100 molecules from the full BRICS dataset, our reconstruction pipeline successfully produced 89 molecules that passed RDKit sanitisation checks. This represents a dramatic improvement from the 48% success rate of the centroid-only approach and confirms that our orientation descriptor, while imperfect, provides substantial and valuable information for molecular reconstruction.

Next, we assessed the performance of the full generative models. For each of the two methods, 1000 molecules were sampled and subjected to the same RDKit sanitisation check. The model using the full BRICS library achieved a pass rate of 27.2%. The model trained on the more constrained, reduced-library BRICS dataset performed significantly better, achieving a pass rate of 51.7%.

While the inclusion of rotational information yielded a marked improvement over the centroid-only models, these results still fall well short of the 84.7% pass rate achieved by the fully atomistic model. The large gap suggests that the primary challenge lies not just in the model's ability to place fragments, but in the inherent inconsistency of the fragments themselves. The geometric diversity of the substructures produced by BRICS decomposition appears too great to allow for the definition of a consistent and reliable local coordinate system using simple geometric heuristics.

4.5 Related Work

Since the completion of this work, significant progress has been made in block-based and hierarchical generative models. Several key approaches have emerged that successfully address the challenges identified in our study, which are outlined below.

Representation	Method	RDKit sanitisation Pass Rate (%)
Baseline	Fully atomistic model	84.7
Centroid-only	Reconstruction - control	48.0
	Full diffusion model	4.03
Rotation-aware	Reconstruction - control	89.0
	Plane alignment method	10.8
	Full BRICS library model	27.2
	Reduced-library BRICS model	51.7

Table 4.1: RDKit sanitisation pass rates for the various block-based molecule generation methods tested, including the reconstruction tests.

HierDiff (Qiang et al. 2023) introduced a 'coarse-to-fine' hierarchical strategy that directly addresses the atom-bond inconsistency issues inherent in simpler fragment-based models. Its fragment vocabulary, in contrast to our use of reaction-based BRICS rules, is constructed by decomposing a large chemical dataset at all rotatable, non-ring single bonds. The generative process first uses diffusion to arrange these coarse nodes in 3D space. Subsequently, a decoder module, built with an E(n)-Equivariant Graph Neural Network (EGNN), is tasked with generating the final atomistic structure. This decoder takes the 3D positions of the coarse nodes as input and directly generates the coordinates for all constituent atoms, thereby solving the orientation problem through a learned, end-to-end process rather than a predefined geometric heuristic. This approach proved highly effective, achieving over 94% molecular validity on standard benchmarks when trained on GEOM.

Another successful implementation of a block-based philosophy is found in CMD-GEN (Zou et al. 2025), a structure-based design tool that uses a hybrid, multi-stage pipeline. Similar to our work, it begins by arranging coarse-grained pseudo-atoms in 3D space using a diffusion model; however, rather than representing arbitrary fragments, these represent pharmacophoric points. Following this, the framework switches to an autoregressive model, which takes the fixed pharmacophore points as a blueprint and constructs the 2D molecular graph atom-by-atom. This separation of tasks—using diffusion for the continuous spatial problem and an autoregressive

model for the discrete graph problem—enables CMD-GEN to tackle the complex task of conditioned, in-pocket generation, achieving validity rates of 97.5%.

Finally, the principle of a substructure-aware process was extended beyond *de novo* generation to the task of conformer generation in SubGDiff (J. Zhang et al. 2024). Its method involves adding noise to randomly selected connected subgraphs rather than to individual atoms. By forcing the denoising network to predict the original state from this partial corruption, the model learns the internal geometry of stable chemical sub-units, leading to a more accurate representation of local chemical environments. While this demonstrates the utility of a substructure-aware approach, it is important to note that conformational generation is a distinct task from *de novo* generation, making a direct comparison of success rates with the other methods impossible.

4.6 Conclusions

This work investigated the adaptation of a successful atomistic diffusion model to a block-based framework for molecular generation. Our exploration revealed that the choice of fragment representation is a critical determinant of model performance. Initial models using a simple centroid-only representation were fundamentally capped, unable to exceed a theoretical reconstruction accuracy of 48% due to rotational ambiguity. The introduction of an explicit 9-dimensional orientation vector helped somewhat alleviate this issue, improving reconstruction fidelity and boosting the final model's RDKit sanitisation pass rate to a peak of 51.7% with a reduced BRICS fragment library. Despite these advances, the block-based models did not approach the 85% pass rate of the original atomistic framework. Our analysis indicates that the primary bottleneck is the high geometric variance within the chemically-derived BRICS fragments. This diversity prevents simple geometric heuristics from defining a consistent and reliable local coordinate system, which is essential for accurate placement and reconstruction by the model.

The challenges encountered in this work are reflective of the broader difficulties in fragment-based generation, and the successes of subsequent models offer a clear

path forward. The consistent theme among effective hierarchical models like HierDiff is the use of a dedicated decoder module—typically an equivariant neural network—tasked specifically with translating the coarse-grained representation into a valid atomistic structure. This learned, end-to-end approach to assembly and orientation proves more robust than the predefined geometric heuristics we employed. Therefore, future research in this area should prioritise the development of these hierarchical architectures. Combining a simplified and standardised fragment vocabulary, as explored in our own work, with an equivariant decoder represents a promising direction for realising the full potential of block-based molecular generation.

5

Cryo-EM model building with AlphaFold

Contents

5.1 Preface 103
5.2 Introduction 104
5.3 Methods 106
5.3.1 Calculating and applying gradient-based optimisation . 107
5.3.2 MSA subsampling 110
5.3.3 Datasets and evaluation 110
5.4 Results 111
5.4.1 OpenFold conditioning 111
5.4.2 Boltz-2 conditioning 116
5.5 Conclusions 121

5.1 Preface

The process of fitting an atomic structure to the 3D density map produced by cryo-EM demands high levels of expertise, time and labour. One way to increase speed and improve quality of output can be to use an *in silico* prediction as a starting point, and recent advances in protein structure prediction have made high-quality starting points readily available. However, when structure prediction models perform poorly on targets that are large or outside of their training data distribution, and these types of systems are some of the primary use cases for cryo-EM.

In this chapter, we provide an overview of our preliminary work on conditioning the inference of models based on AlphaFold 2 (OpenFold) and AlphaFold 3 (Boltz-2) with cryo-EM data. For the first of these, we explored interventions within the structure module, targeting various levels of protein representation—from the single representation to the Cartesian coordinates. Ultimately, we found this method to be insufficient in instances where the model's unconstrained prediction was poor, as large-scale rearrangements proved impossible without causing the model to generate disordered or non-physical conformations. Next, we investigated a similar approach using Boltz-2. We again implemented conditioning within the structure module, and found this implementation to be slightly more robust, but still ultimately incapable of making the necessary large-scale rearrangements. We next took inspiration from the work of Fadini et al. 2025 and Raghu et al. 2025, two implementations of similar, cryo-EM conditioning of structure prediction methods, who reported greater improvements when focusing on earlier stages in the pipeline. We therefore introduced an MSA subsampling approach, similar to that of ROCKET (Fadini et al. 2025), which allows randomness to be injected into Boltz-2. While this allows the model to escape local energetic minima and explore conformations more consistent with the experimental data, our method still struggles to induce the large-scale improvements required for certain targets. Our current investigation is also limited by key simplifications, such as the use of simulated maps and pre-alignment with the crystal structure, which must be addressed for a robust assessment. In this chapter, we report our initial findings and highlight the future directions for this project.

5.2 Introduction

As described in Chapter 1, resolving the 3D structure of a protein is important for many applications in biology and medicine, and particularly in the field of small molecule drug design. A high-resolution model of a target protein not only allows researchers to understand the precise geometry of its binding sites, enabling the informed design of small molecule binders, but is also important earlier in the drug design pipeline at the target identification stage: seeing the structure helps

shed light on the protein's mechanism of action, helping researchers to validate its suitability as a therapeutic target.

Cryo-EM has become an increasingly popular method for structure determination, with the number of new structures available in the Electron Microscopy Data Bank (EMDB) growing exponentially (Lawson, Patwardhan, et al. 2016). This experimental technique relies on averaging data from thousands, or even millions, of projection images of identical macromolecular particles frozen in vitreous ice. This averaging process limits the final resolution of the map, making it challenging to resolve the precise location of individual atoms. Huge progress has been made in this resolution challenge (75.7% of the structures reported in 2024 had resolutions better than 4Å (EMDB 2025; Chiu et al. 2021)), but at worse resolutions, placing all atoms correctly is near impossible (Y. Gao et al. 2023). Furthermore, many large and dynamic biological systems may never reach atomic resolution due to inherent flexibility, meaning that methods for model building at resolutions worse than 4Å remain essential.

To move from a low-resolution density map to an atomic structure, experimentalists have traditionally relied on manual intervention. This often involves using a starting model, which is then refined using interactive 3D visualisation software (Emsley et al. 2010; Croll 2018). The appearance of accurate structure predictors, most notably AlphaFold2 (Jumper et al. 2021), has popularised fully automated implementations of this "predict then refine" strategy. In this paradigm, an AlphaFold prediction is used as a starting model, which is then refined to fit into the experimental density (Dai et al. 2023; Giri et al. 2025). However, these methods can face issues like mis-assignment of peptide-bond geometry and local sequence misalignment, particularly if the initial prediction differs substantially from the conformation captured in the map (Lawson, Kryshtafovych, et al. 2021).

An alternative paradigm, which we explore in this work, is to move beyond these serial "predict then refine" methods. Instead of using the experimental map as a final target for refinement, we directly integrate it into the AlphaFold prediction process itself. By conditioning the prediction with the map, we aim to guide the

model towards the experimentally observed conformation from the outset, rather than attempting to correct it post-hoc.

This concept of conditioning the inference stage of AlphaFold with cryo-EM data has been demonstrated by two recent studies. The first of these, ROCKET (Fadini et al. 2025), performs structure optimisation in AlphaFold 2's internal coevolutionary embedding space, rather than Cartesian coordinates, to better sample large-scale conformational changes. The second, Cryo-Boltz (Raghu et al. 2025), guides the sampling trajectory of the diffusion-based structure model using constraints derived from the density map.

We present an early-stage exploration of conditioning strategies using both an AlphaFold 2-based model (OpenFold) and an AlphaFold 3-based model (Boltz-2). We first show that interventions within the structure module, while effective for local refinement, are insufficient to induce the large-scale conformational changes required to match the experimental density, frequently leading to disordered or non-physical structures. Inspired by the recent findings in ROCKET and CryoBoltz that highlight the importance of earlier pipeline stages, we introduce a different approach. By implementing an MSA subsampling strategy similar to that of ROCKET, we inject stochasticity into the Boltz-2 model. We find that this method can help for the model to escape the local minima of its unconstrained prediction, and sample alternative conformations that are consistent with the provided cryo-EM data.

5.3 Methods

We first implemented a gradient-based conditioning step within the structure module of the OpenFold implementation of AlphaFold 2 (Ahdritz et al. 2024; Jumper et al. 2021). We next adapted this strategy to the Boltz-2 implementation of AlphaFold 3 (Passaro et al. 2025; Abramson et al. 2024), and added an MSA subsampling similar to the approach in ROCKET (Fadini et al. 2025). Below, we provide an overview of these two main conditioning methods (structure module gradient-based optimisation and MSA subsampling) and then more in-depth explanations of the implementations of both.

5.3.1 Calculating and applying gradient-based optimisation

The aim of our method was to guide the prediction to better fit a given experimental density map during the model's internal recycling iterations. The first way we attempted to achieve this was through a gradient-based update applied directly within the structure module.

For both the OpenFold and Boltz-2 implementations, upon the initial generation of 3D coordinates in the structure module, we align the prediction to the target cryo-EM map's frame of reference. A simulated density map is then generated from the prediction's 3D coordinates. Using a simple dot-product between this simulated map and the experimental map, we estimate the agreement between the current prediction and the experimental target map. In the OpenFold implementation, the gradient of a loss function, defined by this correlation score, is calculated with respect to one of three key internal representations: the 3D coordinates, the intermediate frame representations, or the single representation. In the Boltz-2 implementation, it is calculated with respect to the coordinates. Finally, this gradient is applied as an update to the chosen representation, nudging the model's internal state towards a conformation with a higher correlation to the experimental data. In the OpenFold method, the update is also passed onto the other non-chosen representations: for example, if the update is calculated and added to the frame representation, single and coordinate representations which correspond to the new frames are also calculated and updated. For the OpenFold implementation, this process takes place at each recycle through the whole structure module, and for the Boltz-2 implementation, at each timestep in the diffusion process. This means that after each push, the updated structure is fed back to the model, which aims to continue to refine it into a plausible protein structure.

Below we cover in more detail the specifics behind this process.

5.3.1.1 Alignment

As this work is in a preliminary stage, the alignment between the prediction and the cryo-EM map is currently performed by using the ground-truth PDB structure

and aligning based on Cα atoms. This approach would not be applicable in a real experimental setting with a novel cryo-EM map, but it is adopted here to keep the procedure simple while the rest of the method is refined.

5.3.1.2 Simulating a map from the prediction

In order to quantify the agreement between the prediction at each stage of the method and the cryo-EM density map, we must first generate a simulated map of each prediction. This process takes the atomic coordinates and types, and creates a 3D density map by treating each atom as a 3D Gaussian distribution in space. For each atom type, the method creates a small 3D kernel that represents the electron density contribution of that atom type. We then place these kernels at each atom's position in a 3D grid, where the occupancy value scales the intensity of each atom's contribution. The coordinates are normalised and denormalised appropriately to ensure correct placement in the grid, and multiple atoms' contributions are accumulated to build up the final density. The result is a simulated electron density map that approximates what we might see in a cryo-EM experiment, where regions of high electron density correspond to the locations of atoms in the protein structure.

5.3.1.3 Overlap and gradient calculation

After generating a simulated density map from the predicted structure, we quantify its overlap with the experimental cryo-EM map by calculating their dot product. Gradient-based optimisation is then performed with respect to the prediction (in either the single representation, frame representation, or coordinate representation) to obtain an update that should be added to the representation of choice to maximise this overlap.

5.3.1.4 Updating the prediction

Once the update is calculated, it must be applied to the chosen representation to adjust the predicted structure so that its simulated density more closely matches the experimental map. However, directly applying the full update to any of the representations led to severe structural distortion. To mitigate this, the update was

scaled by a constant <1 before being added. The maximum tolerable scaling factor was determined empirically by testing different orders of magnitude and selecting the largest value that did not result in structural collapse.

5.3.1.5 Update scheduling

For the OpenFold conditioning method we tested three distinct gradient scaling schedules. These schedules modulated the strength of the updates based on the model's internal processing steps, using the default structure with $n_{\text{recycle}} = 3$ global recycles and $n_{\text{block}} = 8$ blocks per recycle.

`Block` Scheduling The scaling factor, s, decays with the block index i within a single structure module pass. It resets to 1 at the beginning of each new recycle.

$$s = \frac{1}{i}, \quad \text{for } i = 1, \ldots, 8$$

The scaling factor ranges from a maximum of 1 to a minimum of 1/8.

`Global` Scheduling The scaling factor is held constant for all blocks within a recycle pass but decays with the global recycle index r.

$$s = \frac{1}{r}, \quad \text{for } r = 1, 2, 3$$

In this scheme, the scaling is 1 for the first 8 blocks, 1/2 for the next 8, and 1/3 for the final 8, with the value dropping from 1 to 1/3.

`Block and Global` Scheduling The scaling factor decays continuously across the combined block and recycle indices, never resetting.

$$s = \frac{1}{8(r-1) + i}$$

This creates a monotonic decay across all 24 steps, with the scaling factor decreasing from 1 to 1/24.

5.3.2 MSA subsampling

A limitation of the gradient-based optimisation approach is its tendency to become trapped in local minima. To mitigate this, we introduced stochasticity by carrying out inference with five separately sampled starting MSAs in parallel. Each MSA provided a distinct conditioning context, thereby initiating the diffusion process from multiple, diverse starting points. This strategy increased the likelihood of at least some trajectories escaping local minima and progressing toward higher-overlap solutions.

Unlike ROCKET, we randomly sample MSAs at their full size (a maximum of $n_{max\ seqs} = 8192$), half size ($n_{max\ seqs} = 4096$), and quarter size ($n_{max\ seqs} = 2048$), whilst they implement a more varied subsampling strategy designed to systematically weaken the model's evolutionary prior. Their approach involves creating a 'ladder' of MSAs with very shallow depths, often containing as few as 1, 2, 4, or 8 sequences. The goal of this is to make the model's output more dependent on external experimental data, whereas our method intentionally preserves a strong evolutionary signal whilst adding some diversity to each parallel run to guide the search.

As running multiple MSAs in parallel is computationally intensive, we implemented a pruning strategy to focus resources on the most promising candidates. Specifically, after every 20 optimisation steps, the trajectory with the lowest overlap score was discarded, until only one remained. This ensured that computational effort was concentrated on higher-quality candidates while retaining sufficient diversity to avoid premature convergence.

5.3.3 Datasets and evaluation

We selected 10 small monomer targets for which the vanilla OpenFold and Boltz-2 methods produced moderate- to poor-resolution predictions, defined as structures with a Cα Root Mean Square Deviation (Cα-RMSD) between 3–20 Å from the ground truth crystal structure. Our method is designed to improve intermediate-quality predictions. When the initial structure is already accurate (RMSD $<$ 4 Å),

	5YW7	6LLY	6LN8	7N70	7O3D	7RAV	7VLR	7VPL	8CXO	8FEI
OpenFold	5.47	5.63	5.25	6.46	6.76	15.1	8.36	5.76	6.25	3.58
Boltz-2	9.39	4.62	5.48	8.15	4.21	14.7	4.97	7.25	11.7	20.0

Table 5.1: The Cα RMSDs (Å) of the targets chosen when predicted using the pretrained OpenFold and Boltz-2 models without modification

little or no refinement is needed. Conversely, when the prediction is very poor (RMSD > 15Å), our approach alone is unlikely to correct such large-scale errors.

As this research is in its preliminary stages, we are currently testing the method using ideal, simulated density maps generated from the ground-truth structures. This allows us to first establish the method's core viability before addressing the additional challenge of noise and artifacts present in experimental maps. For this initial assessment, performance is being evaluated solely by the Cα-RMSD of the refined model to the native structure, but more comprehensive validation metrics should be incorporated in future work.

5.4 Results

5.4.1 OpenFold conditioning

5.4.1.1 Direct conditioning with a constant scaling factor

We first implemented a direct conditioning approach within the structure module of OpenFold, using the gradient-based optimisation method described in section 5.3.1 with a constant scaling factor. The scaling factors (0.1 for coordinate and frame-level representations, and 0.01 for the single representation) were determined empirically. Specifically, we selected the highest magnitude for each representation that preserved most of the structural elements in the protein's fold, which we assessed by visually inspecting the relaxed outputs for the loss of secondary structure elements.

As shown in Fig. 5.2, this method yielded modest improvements when conditioning was applied at the coordinate and single representation levels for most targets. Conditioning the frame representation produced no discernible improvement. Furthermore, all conditioning methods degraded the predictions for two targets, 7VPL and 8CXO. Visual inspection of 8CXO reveals multiple instances where the

overlap function was trapped in local minima, causing parts of the model to be pulled towards nearby regions of the density map, rather than the density actually corresponding to those parts (Fig. 5.1).

A closer analysis of the resulting structures suggests that most models begin to drift towards the ground truth conformation before stalling. The best example of this is 8FEI, the target with the greatest percentage improvement, where the conditioned prediction progressed only around halfway from the vanilla, non-conditioned prediction to the correct answer (Fig. 5.3).

This is a promising sign that our approach might be able to generate models in agreement with experimental data, if we can apply more of a drifting force: however, increasing the update strength to accelerate this drift, as mentioned above, results in models with poor structural stability. This presents a clear trade-off: the directional guidance from the gradient is beneficial, but its magnitude must be carefully controlled to avoid overpowering the model's learned structural constraints.

5.4.1.2 Introducing gradient scaling schedules

To better balance the external conditioning force with the model's internal potentials, we introduced three scheduling strategies as detailed in section 5.3.1.5: `block`, `global`, and `block and global`. The objective was to apply a strong conditioning force in early recycling iterations and then allow the model's learned geometry to regularise the structure in later cycles.

For these experiments, we observed that the coordinate representation could tolerate an increased scaling constant of 1.0 (up from 0.1) when paired with a scheduler. This increased tolerance is likely attributable to the scheduler, which attenuates the update strength over successive recycling iterations, mitigating the disruptive effect of the large initial scaling factor. Attempts to apply this increased scaling to the frame and single representations resulted in unstable, disordered predictions, so their scaling constants were maintained at 0.1.

The results, summarised in Fig. 5.4, demonstrate that schedulers did not yield a performance breakthrough, with outcomes largely mirroring those of the constant

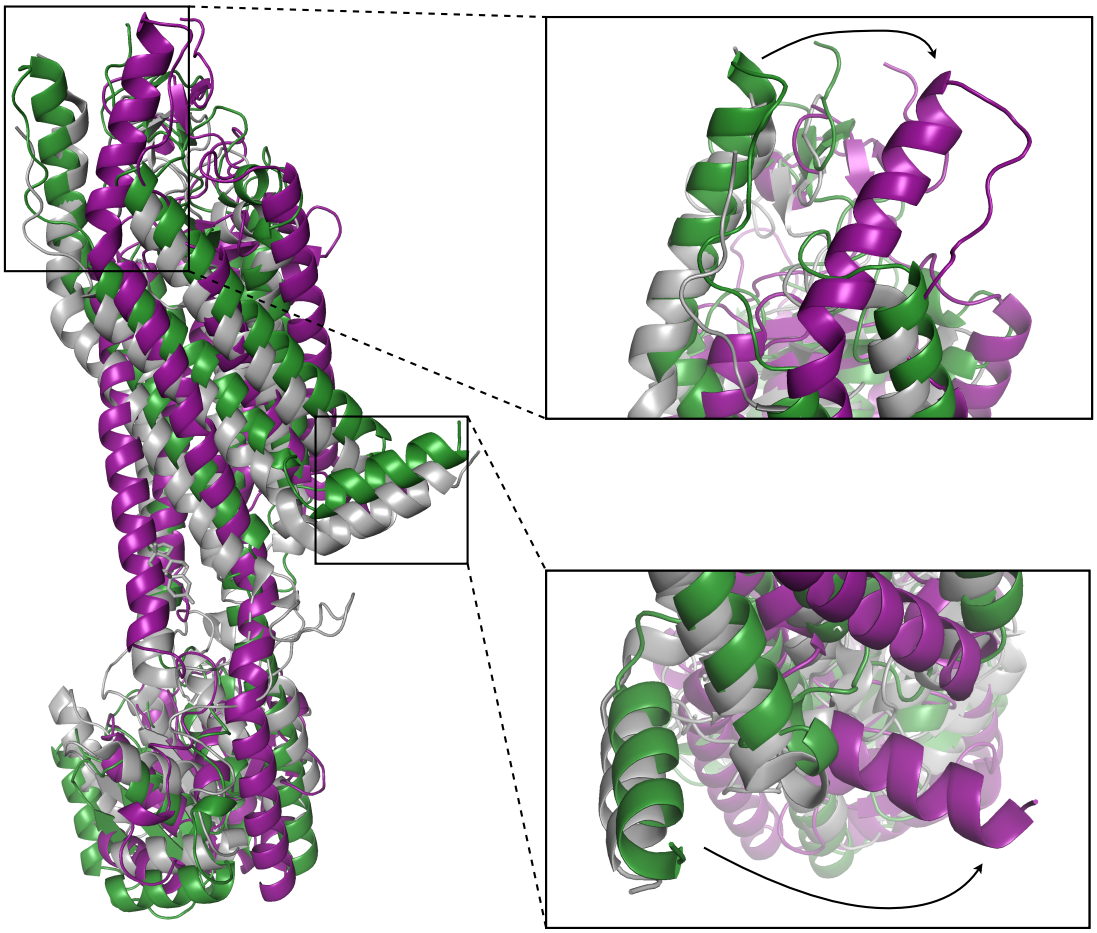


Figure 5.1: 8CXO crystal structure (grey) (K. Zhang et al. 2022), and predictions by vanilla OpenFold (green) and conditioned with coordinates (purple). We have highlighted regions where the conditioning method has moved the prediction in the wrong direction, as the closest region of high density is in the opposite direction to the density of the corresponding region.

scaling approach. The 8FEI target again showed the most consistent improvement, while the method proved detrimental for 8CXO across all configurations. No single combination of representation and scheduler emerged as a superior strategy, though there were some that were better than others: for instance, the `single rep, 0.1, 1, block` method was among the worst performers, only improving predictions for half of the targets, whereas the `coords, 1.0, 1, global` scheme improved results

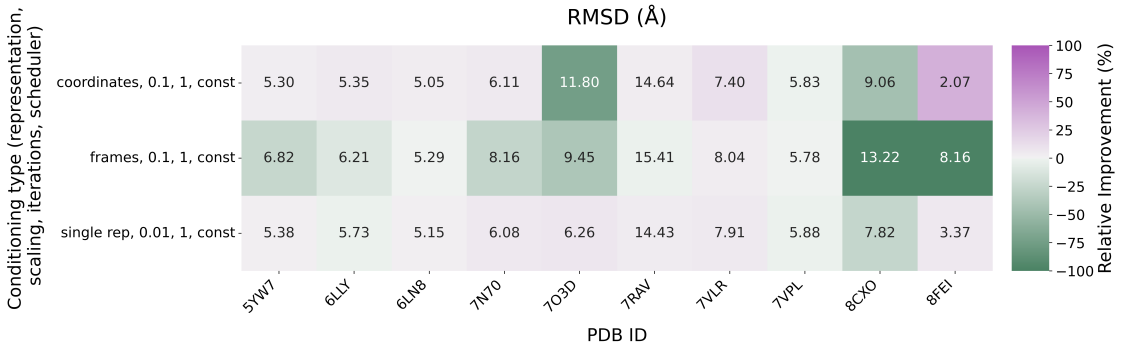


Figure 5.2: Heat map showing the results of the most basic implementation of the OpenFold conditioning method, tested on the three representation levels available within the structure module, scaled to avoid structurally disordering the predictions. Relative improvements to the predictions made by the vanilla OpenFold model are coloured in green to purple, with purple representing an improvement, and green a worse prediction post conditioning.

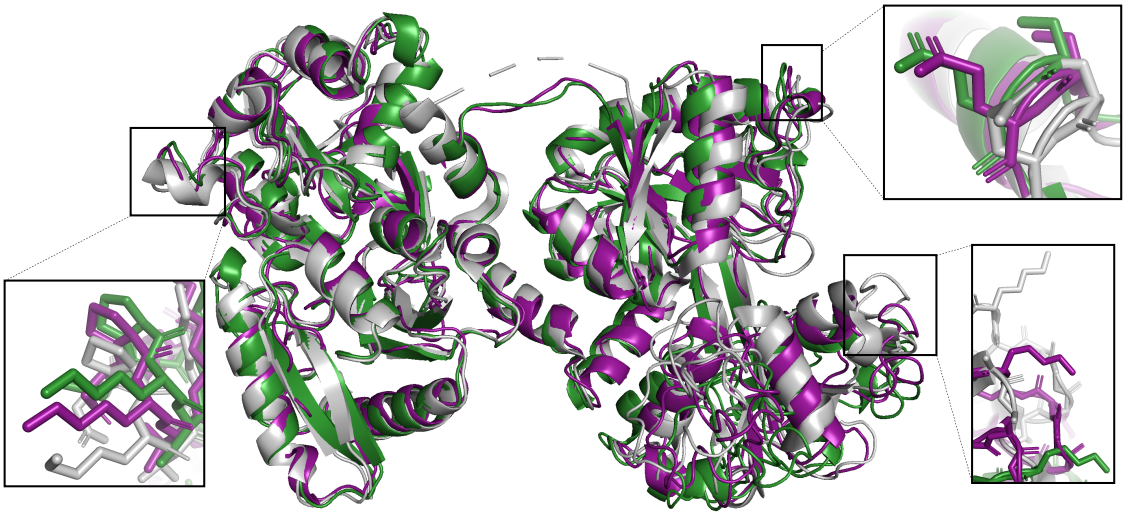


Figure 5.3: 8FEI crystal structure (grey) (Newton et al. 2025), and predictions by vanilla OpenFold (green) and conditioned with coordinates (purple) - using this basic conditioning approach, this is the structure that showed the greatest decrease in RMSD to the crystal structure. We have highlighted some of the regions where you can clearly see the purple conditioned prediction drifting closer to the grey crystal structure relative to the green vanilla prediction, but not reaching it.

most consistently, lowering RMSDs for 8/10 of the targets.

The lack of a clear best strategy and the limited gains across all methods suggests we had not yet found an optimal strategy for balancing our two objectives: on the one hand, the effective incorporation of experimental data, and on the other, the generation of structurally sound models by OpenFold. We therefore

proceeded to test a different approach by applying a series of smaller, more frequent updates within each structure module block.

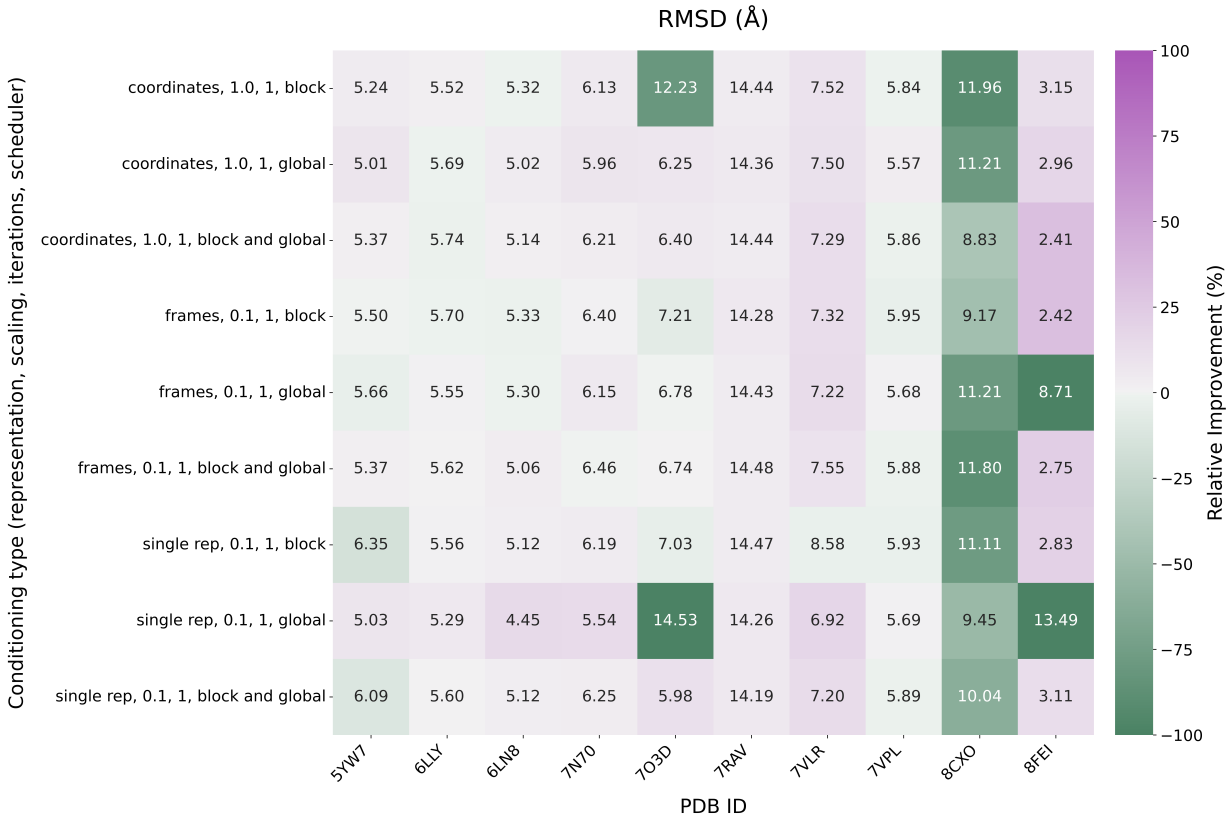


Figure 5.4: Heat map showing the results of the OpenFold conditioning method, tested on the three representation levels available within the structure module, with updates added according to three different schedules. Relative improvements to the predictions made by the vanilla OpenFold model are coloured in green to purple, with purple representing an improvement, and green a worse prediction being made after conditioning.

5.4.1.3 Implementing iterative updates

To test our hypothesis that multiple, smaller updates to predictions might be less disruptive to the OpenFold inference process, we first implemented a fixed-iteration scheme. In this approach, the cycle of map simulation, overlap calculation, gradient computation, and representation update was repeated three times within each structure module block.

This method yielded mixed results (Figure 5.5). While some targets, such as 7VLR, showed consistent minor improvements (a reduction in RMSD of approximately 1 Å on average), the issue of local minima trapping persisted for others

(e.g., 8CXO and 7VPL). This suggests that iterative updates alone cannot overcome poor initialisation without an element of stochasticity.

To address the potential instability of the fixed-iteration approach, we implemented an adaptive strategy. Here, the gradient was applied incrementally, and the process was repeated within a cycle only as long as it improved the map-model overlap score. A maximum of 40 iterations was set as a safeguard, though convergence was typically achieved in 15–25 iterations.

The results, presented in Figure 5.6, demonstrate that this adaptive method is more robust. It produced improvements, reducing the RMSD for 6LN8 from 5.25Å to 4.62Å and for 8FEI from 3.58Å to 2.20Å. However, this adaptive approach also produced worse failures on certain targets, likely because the large number of cumulative updates, even when applied gradually, can still introduce structural distortions.

Overall, no single conditioning strategy emerged as universally superior across all targets. The `frames, 0.1, 3, block and global` scheme was the most consistently positive performer, improving the RMSD for 9 of the 10 targets. Two of the convergence tracking methods (`frames, 0.1, 40, block and global` and `frames, 0.1, 40, block`) and the simpler `coords, 1.0, 1, global` method also performed well, improving 8 out of 10 targets. When considering only the targets that saw improvement, the `frames, 0.1, 1, block` method yielded the highest average improvement of 13.5%. However, all improvements were marginal. This highlights a fundamental challenge: the difficulty of inducing large-scale conformational changes without either getting trapped in local minima or destabilising the model's learned structural priors.

5.4.2 Boltz-2 conditioning

Given the limited success of our OpenFold conditioning experiments we tested an alternative modelling paradigm. The development of Boltz-2 offered an opportunity to address some of the limitations we had encountered: in particular, as a generative diffusion model, Boltz-2's prediction process is inherently more suited to external perturbation. There is strong precedent for guiding, or "drifting", such models

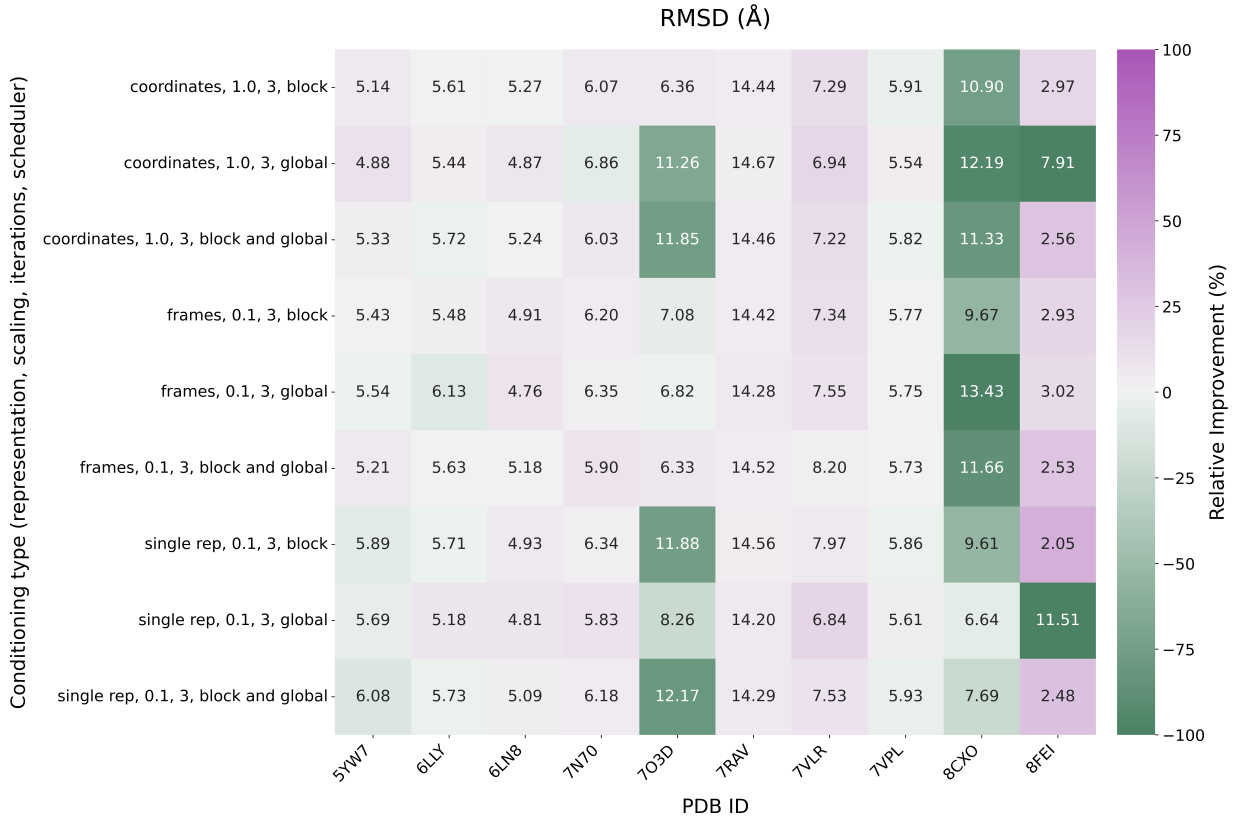


Figure 5.5: Heat map showing the results of the OpenFold conditioning method, tested on the three representation levels available within the structure module, with three schedulers tested. Relative improvements to the predictions made by the vanilla OpenFold model are coloured in green to purple, with purple representing an improvement, and green a worse prediction being made after conditioning.

during inference (Ziv et al. 2024). This suggested that we could apply a more forceful conditioning force to drive conformations towards the experimental data without the structural collapse observed previously. An additional advantage is Boltz-2's ability to model protein-ligand interactions, creating a theoretical pathway to apply our method to cryo-EM maps of bound complexes.

5.4.2.1 Gradient-based optimisation alone

We first applied a direct gradient-based optimisation conditioning scheme, analogous to the one we used in our OpenFold method, within the Boltz-2 structure module. A key advantage of this diffusion-based model is its greater tolerance to external perturbations during the generative process. Unlike OpenFold, the iterative denoising steps could accommodate an external update without structural collapse.

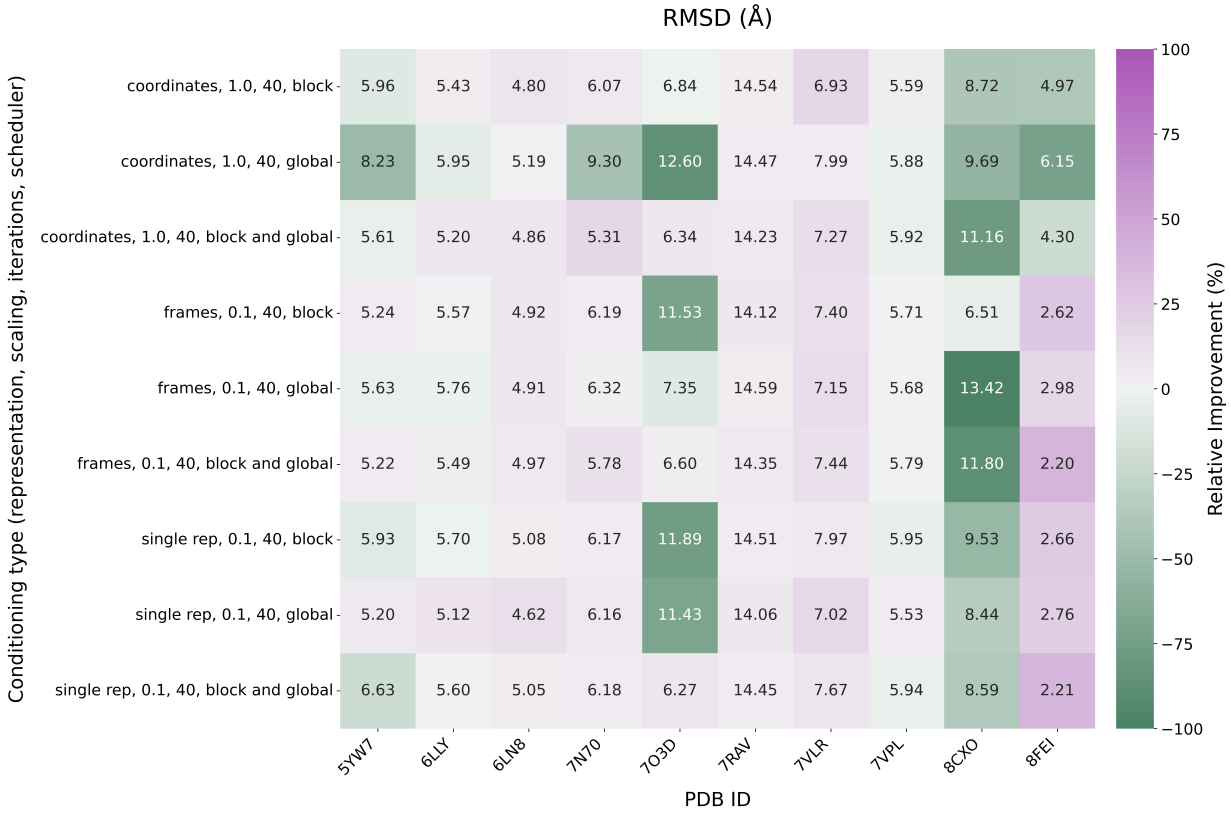


Figure 5.6: Heatmap showing the results of the most complex implementation of the OpenFold conditioning method, where the gradients were tracked and the predictions were updated until the prediction stopped gaining overlap with the experimental map. Relative improvements to the predictions made by the vanilla OpenFold model are coloured in green to purple, with purple representing an improvement, and green a worse prediction being made after conditioning.

The results of this approach, summarised in Fig 5.7, show a mixed pattern of success. While conditioning led to modest improvements over vanilla Boltz-2 for several targets (e.g., 5YW7, 6LN8, and 7N70), it was detrimental to others (e.g., 6LLY and 7VLR). However, this method proved capable of inducing substantial rearrangements that were unachievable with OpenFold. This is best exemplified by 8CXO, where the RMSD was reduced from 11.7Å to 7.7Å, and most strikingly for 8FEI, which improved from a baseline of 20.0Å down to 2.4Å. While these specific successes demonstrate a much higher potential for refinement, the overall inconsistency suggested that this simple implementation was not yet a reliable strategy.

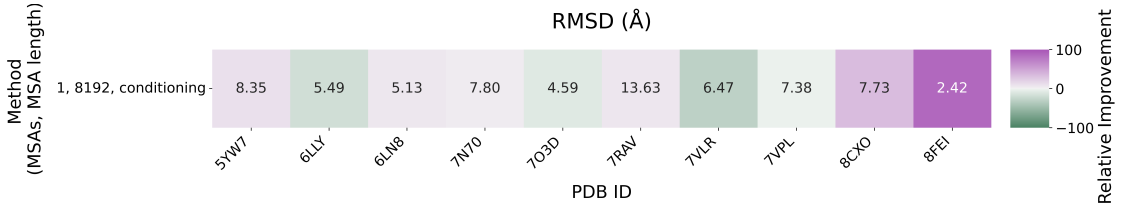


Figure 5.7: The results of conditioning Boltz-2 on simulated cryo-EM data using the gradient-based optimisation method alone, coloured by relative improvement relative to the vanilla Boltz-2 method.

5.4.2.2 Adding MSA subsampling

The inconsistency of the direct gradient-based optimisation approach (both in OpenFold and Boltz-2) suggested that intervening late in the process—during structure refinement—was insufficient for inducing large-scale changes. This led us to explore interventions at an earlier stage of the pipeline. We took inspiration from two recent works: the ROCKET method of Fadini et al. 2025, which demonstrated the potential of MSA subsampling to explore new conformations, and the CryoBoltz method of Raghu et al. 2025. Ablation studies from the latter demonstrated that their global conditioning, applied early in the process, had a more substantial impact than their late-stage structure module conditioning.

Based on these insights, we implemented an MSA subsampling procedure. By providing the model with varied and more sparse subsets of the full MSA, we introduce stochasticity into the prediction process. The goal is to generate a diverse ensemble of initial structures, effectively allowing the model to explore a broader conformational landscape than is possible from a single, deterministic run. We hypothesise that this exploration should increase the likelihood of sampling a conformation closer to the experimental structure, which can then serve as a more effective starting point for refinement.

Only MSA subsampling To first isolate the effect of this stochastic sampling, we generated predictions using subsampled MSAs without any gradient-based conditioning. The results (Fig. 5.8) show that running Boltz-2 with several distinct starting MSAs allows local minima to be escaped more easily. This is shown in the

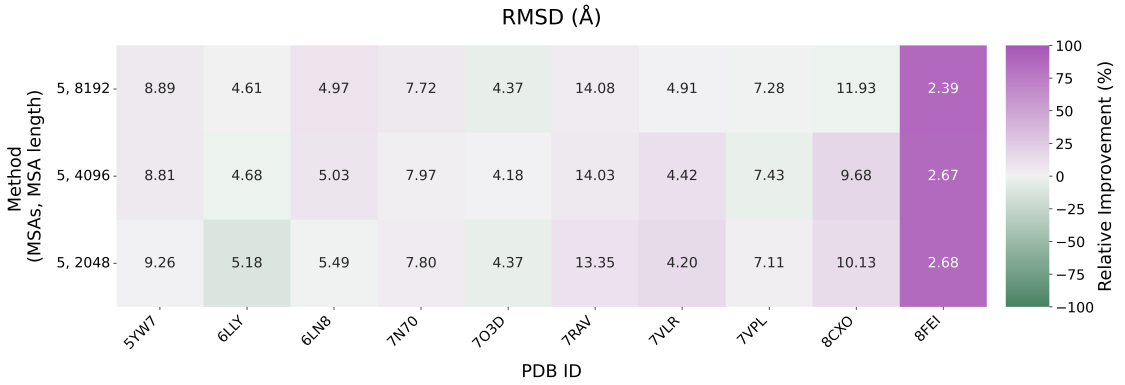


Figure 5.8: The results of subsampling MSAs in Boltz-2, coloured by relative improvement relative to the vanilla Boltz-2 method.

cases of 7VLR and 7VPL, which saw improvements with this method even though the standard gradient-based optimisation approach alone made them worse. However, the method is not consistent, as there are still targets for which subsampling makes the models worse. There is a chance this could be improved by increasing the number of MSAs we start with, but as this is a computationally expensive process, we tested a maximum of five. Further, there is no clear trend regarding which number of sequences yields the best performance: intuitively, including more sequences should make predictions better by providing a more well-informed starting point, but the higher the number of sequences, the less diverse the different MSAs will be. This results in less distinct starting points and therefore fewer opportunities for one to align closely with the map.

Given that this approach seemed promising, we next combined it with the gradient-based optimisation method.

Combining with gradient-based optimisation Finally, we combine the MSA subsampling with the conditioning approach, with the results shown in Fig. 5.9.

As with all other methods tested, this approach had varied success, with the best method for each target improving the RMSDs of 5YW7, 6LN8, 7O3D, 7RAV, 7VLR and 7VPL compared to the gradient optimisation or MSA subsampling methods alone. However, the average RMSD across all targets increased compared to the MSA-subsampling only method, mostly due to a large decrease in quality of

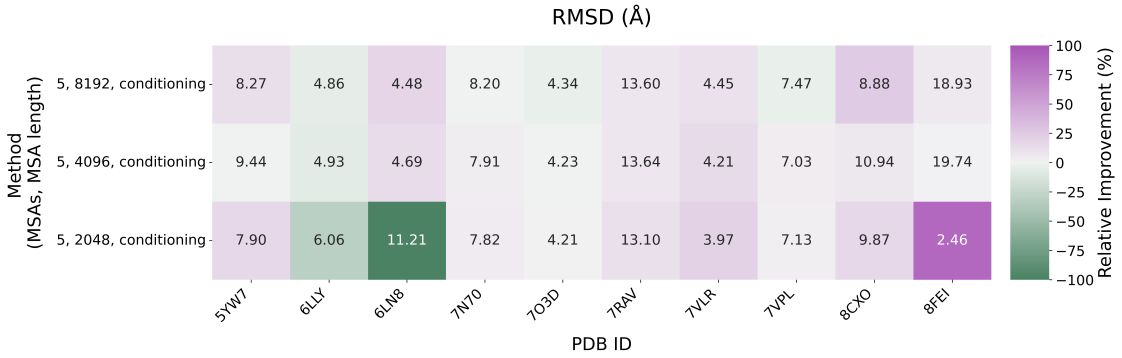


Figure 5.9: The results of subsampling MSAs and conditioning Boltz-2 on simulated cryo-EM data using the gradient-based optimisation method, coloured by relative improvement relative to the vanilla Boltz-2 method.

the predictions for target 8FEI. However, as we showed in Fig 5.3, predictions for this target have a tendency to get trapped in a local minima of nearby incorrect density, explaining why the gradient optimisation approach worsens predictions. It is possible that injecting more stochasticity (say, by subsampling more MSAs) could improve the guidance for this case.

Still, excluding 8FEI, this combination proved to be the most effective approach tested in this chapter. It is still inconsistent, though, and difficult to identify an optimal MSA number, as the performance with different sequence numbers was much better on some targets than others. For example, using 2048 sequences resulted in the most reliable improvements overall, but still only had the lowest RMSD for 6/10 targets. Future directions should accordingly include taking smaller MSA subsamples, as well as carrying out repeat runs to see whether this finding stays consistent, or whether it too is stochastic in nature.

5.5 Conclusions

In this work, we explored several strategies for conditioning *de novo* protein structure prediction models with cryo-EM density maps. Our initial experiments, which applied a gradient-based force within the structure module of OpenFold, highlighted a fundamental limitation: late-stage interventions struggle to induce large-scale

conformational changes without compromising the model's learned structural priors, leading to disordered or non-physical outputs.

Pivoting to Boltz-2, a diffusion-based model, proved to be a more promising direction. While a simple gradient-based optimisation approach still yielded inconsistent results, it demonstrated a far greater capacity for making large structural rearrangements, as evidenced by the dramatic improvements for challenging targets such as 8FEI and 8CXO. Furthermore, our exploration of MSA subsampling confirmed that introducing stochasticity at an early stage is an effective means of broadening the explored conformational landscape and discovering more accurate starting points for refinement.

Future work should investigate the combination of MSA subsampling and gradient-based refinement more systematically, as this approach showed promise but may still be insufficient to escape deep local minima for some targets. Increasing the number of subsamples or incorporating additional stochastic perturbations—such as random noise injection—could further improve sampling diversity and convergence. It will also be essential to evaluate performance across a wider range of targets and to perform multiple runs per target, given the variability observed in current predictions (particularly the large disparity among 8FEI predictions in Figure 5.9).

Finally, the present work still relies on RMSD-based alignment to ground-truth structures and on simulated density maps, which are considerably less noisy than experimental data. Removing these dependencies and extending the approach to real, heterogeneous cryo-EM maps will be crucial steps toward establishing practical, robust conditioning methods. Nevertheless, the results presented here indicate a promising avenue: early stochastic conditioning combined with physically informed gradient guidance can meaningfully steer diffusion-based models toward conformations consistent with experimental data.

6 Conclusions and outlook

Contents

6.1 Interpretability in virtual screening 123
6.2 Plausibility in diffusion-based generative design 124
6.3 AlphaFold for cryo-EM model building 125
6.4 Concluding remarks 126

This thesis has presented contributions across multiple aspects of molecular modelling and drug design, spanning virtual screening, generative design (both fragment-based and *de novo*), and protein structure refinement. While the approaches differ in scope and methodology, they share a central motivation: to improve the reliability of machine learning applications within the early stages of drug discovery. In this chapter, we provide concluding remarks for each application and highlight potential directions for future work.

6.1 Interpretability in virtual screening

In chapter 2, we introduced an EGNN-based scoring function that we designed to be as unreliant as possible on dataset biases. By training and testing on a filtered dataset, we showed that the model was less dependent on memorisation than other leading methods, and instead had started to capture aspects of the real

physics of protein–ligand binding. Through interpretability experiments, we also found that the model based its predictions on realistic protein–ligand interactions, and that these signals could be used to guide fragment elaboration in ways that improved docking outcomes.

From this work, two main points stand out. First, interpretability is not just useful, but should be essential for virtual screening models: without it, there is no clear way to tell if a model is using real binding information or dataset patterns. Although it would be difficult to implement a single standardised interpretability benchmark that works across all models, it would be valuable if new VS methods were expected to demonstrate their performance on a set of such metrics. This could help reduce the flood of publications that report high scores on biased datasets, and help to ensure that reported improvements reflect genuine advances. Second, interpretability methods offer more than just model validation; they can provide structural insights to guide molecule design. A properly debiased model can screen numerous compounds, identify key recurring interactions, and help prioritise new designs that add or preserve these features.

Still, this work also highlights a key issue. With the datasets currently available, it is impossible to train scoring functions that are entirely free from bias. Progress will depend on building large, high-quality public datasets of protein–ligand complexes that are modelled in a consistent way. Such resources would not only support more reliable benchmarking, but also give ML models a better chance to learn binding physics in a way that generalises beyond narrow training sets.

6.2 Plausibility in diffusion-based generative design

Our first piece of work on diffusion models for molecule generation (chapter 3) demonstrated the benefits of incorporating negative data by conditioning on conformer quality. By explicitly including synthetically distorted conformers in the training data and conditioning on so-called distortion labels, we improved the

validity and plausibility of generated molecules, particularly for larger, drug-like compounds with complex conformational landscapes.

We next investigated a substructure-based generative model, adapting atomistic diffusion approaches to a block-based setting (chapter 4). While this direction offers the promise of improved chemical validity through assembly from pre-validated fragments, our results underscore the challenges posed by geometric variance in fragment libraries. The limitations of orientation schemes point to the need for either learned equivariant decoders capable of reconstructing atomistic detail from coarse-grained representations, or hierarchical approaches that get the generation most of the way toward being a molecule with block based approaches, and then refine the atomistic model.

Two primary takeaways can be drawn from this work. First, it serves as a reminder that negative data is a valuable tool for improving the performance of machine learning models, even in domains like unconditional molecule generation where training data may seem abundant. The second, broader takeaway is the importance of mastering foundational principles before applying new, computationally expensive architectures to more complex challenges. Many current applications of diffusion models for small-molecule generation rely heavily on post-hoc filtering to achieve viable results. This raises the question of whether such resource-intensive architectures are justified if they require extensive filtering, or whether efforts should be redirected toward refining these methods or revisiting more established approaches.

6.3 AlphaFold for cryo-EM model building

Finally, we explored conditioning protein structure prediction with cryo-EM density maps. Gradient-based interventions in OpenFold revealed the difficulty of enforcing large-scale conformational changes at late stages without destabilising learned priors. By contrast, Boltz-2 showed greater capacity for structural rearrangements, particularly when combined with stochasticity via MSA subsampling. While the results were mixed, these early experiments point towards promising avenues for

integrating experimental restraints into generative structure prediction. This work has laid important groundwork, perhaps more by clarifying which strategies are ineffective than by arriving at a viable one. Nonetheless, clear next steps have emerged, including sampling more MSAs, experimenting with different conditioning schedulers, and potentially modifying the structure prediction models themselves through fine-tuning to better incorporate experimental data during training. Such advances would bring generative structure prediction closer to practical utility in cryo-EM model building.

6.4 Concluding remarks

Taken together, the work in this thesis illustrates both the promise and the persistent challenges of applying machine learning to drug discovery and structural biology. While none of the methods presented constitute a complete solution, each provides new insights into how data quality, model architecture, and conditioning strategies shape outcomes.

Several key lessons stand out: interpretability is central to ensuring that models learn physics rather than artefacts. Negative data remains a powerful tool for enhancing the robustness of generative models. Block-based generation is an attractive paradigm but requires sophisticated encoding schemes to overcome the geometric inconsistency of substructures. Finally, guiding AlphaFold-like predictions towards experimental constraints is feasible but requires intervention early in the generative process, coupled with sufficient stochasticity to escape local energy minima.

Pursuing these directions will require better datasets, more carefully designed benchmarks, and a stronger collective emphasis on interrogating our models rather than relying solely on benchmark scores. If these challenges can be met, the reward is substantial: machine learning methods that are not only powerful in silico but also robust, interpretable, and truly impactful in real-world biomedical discovery.

Appendices

A

Appendix: chapter 2

Contents

A.1 Equivariance and Invariance - written by Jack Scantlebury 129
A.2 Hotspots API: calculation and processing of hotspots 130
A.3 Structure IDs 133
A.4 FGFR1 Case Study 133

A.1 Equivariance and Invariance - written by Jack Scantlebury

A function $\psi : X \to Y$ is *invariant* with respect to a group of transformations $G : X \to X$ if:

$$\psi(g(x)) = \psi(x) \quad \forall g \in G \quad \forall x \in X \tag{A.1}$$

ψ is *equivariant* with respect to G if there is a transformation $T : Y \to Y$ such that:

$$\psi(g(x)) = T(\psi(x)) \quad \forall g \in G \quad \forall x \in X \tag{A.2}$$

In other words, if ψ is equivariant under G, applying elements of G to x will apply an analogous transformation to the output in Y. The order in which the transformations are applied does not matter to the end result.

While CNNs offer invariance to elements of the group $G = T(n)$, the set of all translations in n-dimensional space, they lack invariance to the group of all roto-translations, $SE(n)$. This means that rotating the protein-ligand input in 3D space may produce different predictions, behaviour which is undesirable. Usually, the way to combat this is to augment training data by assigning random rotations to the inputs, but this does not achieve true rotational invariance. For a more thorough introduction to symmetry groups and equivariance, see (T. Cohen et al. 2016; Finzi et al. 2020).

Group-equivariant neural networks are a new class of deep learning model which offer invariance to various symmetry groups. They do this by using a series of equivariant layers followed by one or more invariant layer. For CNNs, these include Group Equivariant CNNs (T. Cohen et al. 2016) and steerable CNNs (T. S. Cohen et al. 2017). CNNs still have the fundamental problem of a discrete input space and an extremely large input space and parameter count (a $48 \times 48 \times 48$ grid with 22 channels is more than 2.4×10^6 floating point numbers). If we view molecules as graphs, however, we can apply group-equivariant GNNs to the problem of pose predictions.

For these reasons, we use an architecture based on *E(n)*-Equivariant Graph Neural Network (EGNN) layers (Satorras et al. 2022) for pose selection. The $E(n)$ symmetry group contains all elements from $SE(n)$, as well as the group of all reflections; EGNN layers are also permutation-equivariant, meaning that the network is invariant to the order of the input vertices.

A.2 Hotspots API: calculation and processing of hotspots

For a given protein target, the API calculates acceptor, donor, and apolar atomic propensity maps using SuperStar (Verdonk, Cole, and Taylor 1999). This tool defines a grid covering the protein with points spaced 0.5Å apart, and then using data from the Cambridge Structural Database (CSD) (Groom et al. 2016), assigns a propensity for the three atomic probe types—acceptor, donor, and apolar—to each

point. These scores are assigned based on the principle that if a positive interaction between two groups at a certain distance and angle is made, then it will occur more frequently in structures stored in the CSD, and thus be assigned a higher propensity score. These grid point scores are then weighted according to how buried in the protein they are so that more buried sites are favoured.

Further scores are then calculated with small chemical probes consisting of aromatic rings with an acceptor, donor, or apolar atom in the substituent position. The three weighted propensity grids are sampled by their corresponding probe: a given probe is translated to all grid points with scores above a threshold of 15, and randomly rotated 3000 times around the centre of the substituent atom. For each pose, the probe atom scores are read from their corresponding grids, and the probe score is found as the geometric mean of the scores of the probe's atoms. Once complete, the sampled probe scores are assigned to an output grid.

To process the donor and acceptor grids, we follow the protocol outlined in (Hadfield et al. 2022). A greedy clustering algorithm is employed. A cluster is initialised as a random single point, and all points less than 1Å away that are not part of another cluster are added. Then, for each point added, this process is repeated, so all remaining unclustered points within 1Å are added. This process continues until there are no unclustered points within 1Å of the points forming the cluster. Once a cluster is finalised, a new cluster is defined by selecting another single unclustered point. This process is repeated until all points have been assigned to a cluster. The centroids of these clusters are defined as the mean position of the points constituting them. Clusters comprising less than eight points are removed, as well as those with centroids outside of an apolar region (this is assessed using the apolar grid). The remaining clusters are then allocated a score consisting of the number of points comprising them, and output alongside their associated scores as the acceptor and donor hotspot maps for a given protein target.

Mpro			Mac1		NSP14
x10474_0A	x10484_0A	x10494_0A	DLS-EU0569_0A	DLS-X0626_0A	x1203_0D
x10898_0A	P0906_0B	x11543_0A	DLS-X0724_0A	DLS-X0689_0A	x1088_0D
x2964_0A	x11426_0A	x2646_0A	DLS-X0591_0A	DLS-X0711_0A	x1201_0D
x11372_0A	x10604_0A	x10559_0A	DLS-X0681_0A	DLS-EU0034_0A	x1316_0D
x11233_0A	x12682_0A	x11764_0A	DLS-X0895_1A	DLS-EU0087_0A	x1448_0D
x10473_0A	x10889_0A	x2971_0A	DLS-X0910_0A		x1412_0D
x10327_0A	P1470_0A	x11159_0A	DLS-X0587_0A		x1181_0D
x12740_0A	x10723_0A	x2908_0A	DLS-X0421_0A		x1199_0D
x10888_0A	x10248_0A	x11532_0A	DLS-EU0342_0A		x1197_0D
x10371_0A	P0045_0B	x0678_0A	DLS-X0844_0A		x1384_1D
x10606_0A	x10756_0A	x11164_0A	DLS-EU0815_0A		x1486_0D
x11743_0A	x10995_0A	P0045_0A	DLS-EU0291_0A		x1245_0D
x2912_0A	x10555_0A	x3108_0A	DLS-X0299_0A		x1541_0D
x10565_0A	x2572_0A	x11339_0A	DLS-X0471_0A		x1603_0D
x10422_0A	P0906_0A	x11564_0A	DLS-X0895_0A		x1236_2D
x10421_0A	P2007_0B	x10789_0A	DLS-EU0238_0A		x1282_0D
P1073_0A	x3359_0A	P2001_0B	DLS-X0766_0A		x1641_0D
x10488_0A	x10387_0A	x2608_0A	DLS-X0926_0A		x1078_0D
x11427_0A	x0107_0A	x10856_0A	DLS-EU0056_0A		x1212_1D
P0145_0B	x10201_0A	x10314_0A	DLS-X0592_0A		
x10942_0A	x11271_0A	x11513_0A	DLS-X0967_0A		
x10022_0A	x2581_0A	x11540_0A	DLS-EU0144_0A		
x11473_0A	x10377_0A	x2600_0A	DLS-X0722_0A		
x0434_0A	x10395_0A	x10598_0A	DLS-EU0704_0A		
P1470_0B	x10329_0A	x11041_0A	DLS-EU0844_0A		
x12719_0A	x10178_0A	x11417_0A	DLS-EU0481_0A		
x10396_0A	x12300_0A	x10638_0A	DLS-X0969_0A		
z7vh8_0A	x10324_0A	x10334_0A	DLS-EU0241_0A		
x11368_0A	P0884_0A	x10392_0A	DLS-X0962_0B		
P2007_0A	x11346_0A	x11708_0A	DLS-X0742_0A		
x11231_0A	x2649_0A	x11318_0A	DLS-EU0172_0A		
x10906_0A	x11641_0A	x10019_0A	DLS-X0918_0A		
x3080_0A	x10800_0A	x11428_0A	DLS-X1018_0A		
x11560_0A	x10996_0A	x10733_0A	DLS-X0962_0A		
x11013_0A	x10417_0A	x11475_0A	DLS-X0900_0A		
x10566_0A	x10247_0A	x10478_0A	DLS-X0685_0A		
P2001_0A	x3298_0A	x11485_0A	DLS-X0962_1A		
x11809_0A	x3366_0A	x10575_0A	DLS-X0805_0A		
x10679_0A	x2643_0A	P0145_0A	DLS-X0142_0A		
x10801_0A	x11488_0A	x10787_0A	DLS-X0104_0A		
P0925_0A	x11317_0A	x2569_0A	DLS-X0852_0B		
x11225_0A	x10889_1A	x11044_0A	DLS-X0918_1A		
x10338_0A	x10476_0A	x11723_0A	DLS-EU0716_0A		
x10359_0A	x11562_0A	x10322_0A	DLS-EU0123_0A		
x2193_0A	x10900_0A	x11354_0A	DLS-X0905_0A		
x10423_0A	x2562_0A	x12321_0A	DLS-X0910_1A		
x11557_0A	x11801_0A	x11025_0A	DLS-X0933_0A		
x10645_0A	x10237_0A	x10834_0A	DLS-X0837_0A		
x10626_0A	x11011_0A	x11507_0A	DLS-EU0293_0A		
x10236_0A	x10535_0A	x10976_0A	DLS-EU0424_0A		
x11186_0A	x10610_0A		DLS-EU0046_0A		

Table A.1: Fragalysis identification codes of the structures supplied to PointVS to obtain hotspots.

A.3 Structure IDs

The bound fragment structures used for all three targets (Mpro, Mac1 and NSP14) are listed in Table A.1. These datasets are all available to download from the fragalysis platform (*XChem @ Diamond* 2024).

The PDB IDs of the randomly selected subset of 20 structures from the PDBBind core set used in section 3.4.2 are 3URI, 4DE1, 4DLI, 3N76, 4M0Z, 3TWP, 1VSO, 3NQ9, 3UEU, 3AO4, 3JVR, 2BRB, 1H23, 1E66, 2FVD, 2ZB1, 4K18, 3OE5, 3GE7, 3RR4, 3UI7, 4JSZ, 3QQS, 2VKM, 4MGD, 2XII, 2QE4, 1Y6R.

A.4 FGFR1 Case Study

To demonstrate the potential of the PointVS hotspot maps for real-world fragment-to-lead campaigns, we present a case study on a well-studied set of drug targets: the fibroblast growth factor receptors. Aberrant activation of these receptors is associated with a wide variety of cancers, making them a promising druggable target. In 2019, a collaboration between Astex and Newcastle University resulted in erdafitinib (Murray et al. 2019): a drug that has since been FDA approved, making it the third fragment-derived drug to reach this stage.

In the process of designing erdafitinib, the collaborators heavily relied on a fragment screen performed on the target in 2008. A fragment found in this screen, shown in Fig A.1, was found to have high potency, and was used as a starting point in the design process.

In (Patani et al. 2016), two hydrogen bonds are found to form between erdafitinib and the protein. We investigated whether we could identify the protein atoms forming these bonds as important using either PointVS or the API hotspots.

The API placed one donor hotspot 1.5Å away from the location of one of the erdafitinib atoms that forms a hydrogen bond with a protein atom. However, it did not recognise the point the second binding ligand atom is found at as important.

To obtain the PointVS hotspots, we performed attribution on 18 of the bound structures of FGFR1 available in the PDB (Berman et al. 2000). These bound

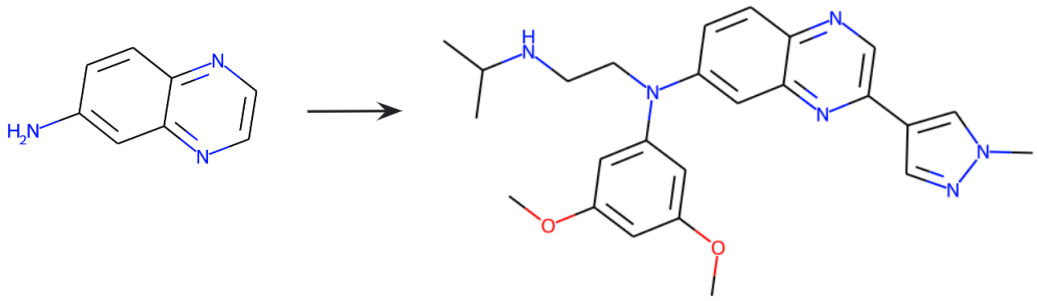


Figure A.1: Left, the fragment used as a starting point in the development of erdafitinib, shown on the right.

structures contain compound-like molecules rather than fragments, so we filtered them to ensure that none had a Tanimoto similarity of greater than 0.7 to erdafitinib. This resulted in us selecting the structures corresponding to the following PDB codes: 4NKS, 4NK9, 1AGW, 3C4F, 4NKA, 4V01, 4V05, 5B7V, 5A46, 4UWB, 3RHX, 1FGI, 5A4C, 4UWC, 5VND, 6P69.

We found that PointVS correctly identified both of the protein atoms that form hydrogen bonds with ligand atoms as important, ranking them as the third and sixth most important atoms for binding.

To provide an example of how knowledge of these two hotspots could be used, we supplied one of them to STRIFE to demonstrate that molecules similar to erdafitinib could be generated. An example of an elaboration that recovered the same binding mode as erdafitinib is shown in Fig A.2.

	PointVS		Hotspots API	
Rank	Fragments successfully elaborated	Elaborations per fragment	Fragments successfully elaborated	Elaborations per fragment
1	21	217	15	189
2	12	187	10	183
3	13	223	11	207
4	11	192	7	158
5	5	207	12	179
6	9	176	15	162
7	6	239	21	220
8	2	222	11	197
9	0	0	0	0
10	1	218	8	210

a

	PointVS		Hotspots API	
Rank	Fragments successfully elaborated	Elaborations per fragment	Fragments successfully elaborated	Elaborations per fragment
1	14	194	2	237
2	3	196	23	218
3	24	215	24	195
4	23	186	14	154
5	14	183	0	0
6	23	186	24	195
7	17	194	0	0
8	5	196	4	116
9	1	89	27	199
10	6	205	13	181

b

Table A.2: The number of fragments that were successfully elaborated on using PointVS and API hotspots for Mac1 (a) and NSP14 (b), and the mean number of elaborations generated per fragment.

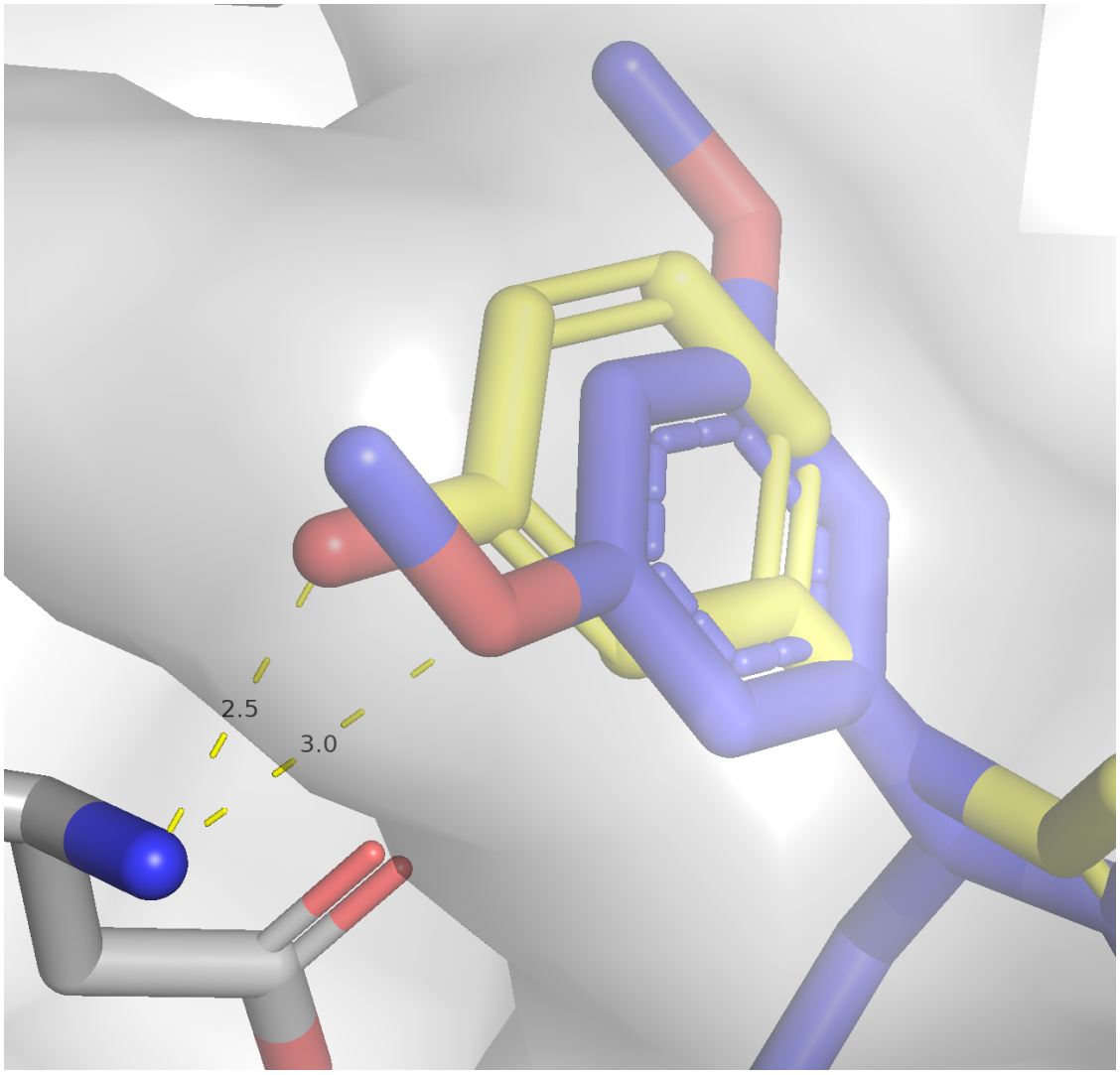


Figure A.2: Elaboration generated by STRIFE using the PointVS hotspots (yellow), that recovers the same hydrogen bond as erdafitinib, shown in blue (PDB ID: 5EW8 (Patani et al. 2016)).

B

Appendix: chapter 3

Contents

B.1 Ablation test: sampling $D = D_{max}$ 137
B.2 Full PoseBusters outputs 137
B.3 Other metrics 139

B.1 Ablation test: sampling $D = D_{max}$

D_{max} (Å)	Ratio of distorted:non-distorted molecules					
	1:20		1:50		1:100	
	RDKit Sanitisation, %	PoseBusters Pass Rate, %	RDKit Sanitisation, %	PoseBusters Pass Rate, %	RDKit Sanitisation, %	PoseBusters Pass Rate, %
0.1	91	46	**96**	**53**	81	26
0.25	81	2	81	2	72	4
0.5	49	0	28	0	38	0
1	41	0	46	0	29	0

Table B.1: Performance comparison 100 molecules sampled with EDM trained conditionally on $\text{GEOM}_{\text{no h}}$ using a distortion factor, D, and sampled with D=D_{max}Å across various ratios of distorted:non-distorted molecules and maximum distortion values in angstrom.

B.2 Full PoseBusters outputs

Below we provide the full outputs for each model's molecules when assessed with PoseBusters.

Dataset	RDKit Sanitisation, %	Posebusters Pass Rate, %							
		All Atoms Connected	Bond Lengths	Bond Angles	Internal Steric Clash	Aromatic Ring Flatness	Double Bond Flatness	Internal Energy	All Tests Passed
No conditioning									
QM9	92.2 (90.5–93.8)	100.0 (100.0–100.0)	100.0 (100.0–100.0)	99.9 (99.7–100.0)	100.0 (100.0–100.0)	100.0 (100.0–100.0)	100.0 (100.0–100.0)	88.1 (85.9–90.1)	81.1 (78.7–83.5)
$GEOM_{no\ h}$	84.7 (82.5–86.9)	74.4 (71.7–77.1)	65.6 (62.5–68.7)	73.8 (70.8–76.7)	97.8 (96.7–98.7)	100.0 (100.0–100.0)	100.0 (100.0–100.0)	75.2 (72.3–78.0)	62.2 (58.3–66.1)
ZINC	70.6 (67.8–73.3)	62.4 (59.4–65.3)	65.0 (61.5–68.4)	75.3 (72.2–78.5)	79.5 (76.6–82.4)	100.0 (100.0–100.0)	100.0 (100.0–100.0)	78.5 (75.5–81.4)	40.0 (37.0–43.0)
Distortion factor conditioning									
QM9	65.0 (62.0–68.0)	84.0 (81.7–86.2)	96.6 (95.2–98.0)	95.8 (94.3–97.2)	98.9 (98.0–99.7)	100.0 (100.0–100.0)	100.0 (100.0–100.0)	91.1 (88.8–93.2)	46.9 (43.9–50.0)
$GEOM_{no\ h}$	92.4 (90.7–94.0)	89.4 (87.5–91.3)	96.6 (95.5–97.8)	93.7 (92.1–95.2)	100.0 (100.0–100.0)	100.0 (100.0–100.0)	100.0 (100.0–100.0)	87.7 (85.5–89.7)	68.8 (65.9–71.1)
ZINC	95.3 (93.8–96.6)	95.7 (94.3–96.9)	94.1 (92.5–95.6)	93.7 (92.1–95.2)	99.0 (98.4–99.6)	100.0 (100.0–100.0)	100.0 (100.0–100.0)	89.4 (87.4–91.4)	74.5 (71.7–77.3)
Ablation tests									
1:20, $D_{max} = 0.1$Å	96 (92 –99)	93 (88 –98)	94 (89 –98)	92 (86 –97)	96 (92 –99)	96 (92 –99)	96 (92 –99)	82 (74 –89)	73 (64 –81)
1:20, $D_{max} = 0.25$Å	95 (90 –99)	73 (64 –82)	88 (81 –94)	90 (84 –95)	95 (90 –99)	95 (90 –99)	95 (90 –99)	82 (74 –89)	52 (42 –62)
1:20, $D_{max} = 0.5$Å	97 (93 –100)	95 (90 –99)	94 (89 –98)	89 (83 –95)	97 (93 –100)	97 (93 –100)	97 (93 –100)	88 (81 –94)	75 (66 –83)
1:20, $D_{max} = 1$Å	93 (88 –97)	88 (81 –94)	85 (78 –92)	85 (78 –92)	92 (86 –97)	93 (88 –97)	93 (88 –97)	76 (67 –84)	57 (47 –67)
1:50, $D_{max} = 0.1$Å	96 (92 –99)	93 (88 –97)	94 (89 –98)	94 (89 –98)	96 (92 –99)	96 (92 –99)	96 (92 –99)	86 (79 –92)	77 (69 –85)
1:50, $D_{max} = 0.25$Å	97 (92 –99)	91 (85 –96)	95 (90 –99)	94 (89 –98)	96 (92 –99)	96 (92 –99)	96 (92 –99)	90 (84 –96)	81 (73 –88)
1:50, $D_{max} = 0.5$Å	97 (93 –100)	90 (84 –95)	94 (89 –98)	96 (92 –99)	97 (93 –100)	97 (93 –100)	97 (93 –100)	91 (85 –96)	78 (70 –86)
1:50, $D_{max} = 1$Å	89.2 (78–97)	81.1 (68–92)	83.8 (70–95)	81.1 (68–92)	89.2 (78–97)	89.2 (78–97)	89.2 (78–97)	73 (57–86)	54.1 (38–70)
1:100, $D_{max} = 0.1$Å	96 (92 –99)	92 (86 –97)	95 (90 –99)	94 (89 –98)	96 (92 –99)	96 (92 –99)	96 (92 –99)	87 (80 –93)	77 (68 –85)
1:100, $D_{max} = 0.25$Å	96 (92 –99)	94 (89 –98)	95 (90 –99)	93 (88 –97)	96 (92 –99)	96 (92 –99)	96 (92 –99)	84 (77 –91)	77 (68 –85)
1:100, $D_{max} = 0.5$Å	95 (90 –99)	86 (79 –92)	94 (89 –98)	94 (89 –98)	95 (90 –99)	95 (90 –99)	95 (90 –99)	81 (73 –88)	68 (59 –77)
1:100, $D_{max} = 1$Å	62 (52 –71)	72 (63 –81)	42 (32 –52)	31 (22 –40)	56 (46 –66)	62 (52 –71)	62 (52 –71)	46 (36 –56)	8 (3 –14)
Energy conditioning									
QM9	57 (47–66)	40 (31–50)	56 (46–65)	57 (47–66)	56 (46–65)	57 (47–66)	57 (47–66)	53 (43–63)	19 (12–27)
$GEOM_{no\ h}$	68 (59–77)	67 (57–76)	68 (59–77)	68 (59–77)	68 (59–77)	68 (59–77)	68 (59–77)	63 (53–72)	39 (30–49)
ZINC	65 (56–74)	21 (13–29)	12 (6–19)	41 (31–51)	53 (43–63)	65 (56–74)	65 (56–74)	39 (30–49)	1 (0–3)
GCDM									
$GEOM_{no\ h}$ baseline	100.0 (100.0–100.0)	93.5 (91.9–95.0)	97.7 (96.7–98.6)	98.8 (98.1–99.4)	98.9 (98.2–99.5)	100.0 (100.0–100.0)	100.0 (100.0–100.0)	86.2 (84.1–88.3)	77.8 (75.2–80.4)
$GEOM_{no\ h}$ conditioned	99.9 (99.7–100.0)	92.9 (91.2–94.4)	97.8 (96.9–98.6)	98.8 (98.1–99.4)	98.4 (97.6–99.1)	100.0 (100.0–100.0)	100.0 (100.0–100.0)	88.6 (86.6–90.5)	79.7 (77.2–82.1)
ZINC baseline	56.3 (53.3–59.3)	63.7 (60.8–66.6)	95.3 (92.8–97.5)	97.5 (95.9–98.9)	91.2 (88.2–93.9)	100 (100–100)	99.7 (99.2–100)	88.2 (84.8–91.2)	40.8 (38–43.6)
ZINC conditioned	97.2 (95.8–98.4)	94.4 (92.5–96.1)	89.0 (86.3–91.5)	95.5 (93.7–97.1)	97.5 (96.0–98.7)	100.0 (100.0–100.0)	100.0 (100.0–100.0)	82.0 (78.7–85.0)	66.5 (62.5–70.4)
MolFM									
$GEOM_{no\ h}$ baseline	100.0 (100.0–100.0)	93.5 (91.9–95.0)	97.7 (96.7–98.6)	98.8 (98.1–99.4)	98.9 (98.2–99.5)	100.0 (100.0–100.0)	100.0 (100.0–100.0)	86.2 (84.1–88.3)	77.8 (75.2–80.4)
$GEOM_{no\ h}$ conditioned	100.0 (100.0–100.0)	92.8 (91.1–94.4)	97.5 (96.5–98.4)	98.7 (98.0–99.4)	98.3 (97.4–99.1)	100.0 (100.0–100.0)	100.0 (100.0–100.0)	87.5 (85.4–89.5)	78.7 (76.2–81.2)
ZINC baseline	72.0 (69.2–74.8)	62.9 (59.9–65.9)	66.9 (63.5–70.4)	76.4 (73.2–79.4)	76.7 (73.5–79.7)	100.0 (100.0–100.0)	100.0 (100.0–100.0)	79.2 (76.1–82.1)	42.3 (39.2–45.4)
ZINC conditioned	93.3 (91.6–94.9)	96.4 (95.1–97.5)	70.4 (67.4–73.5)	86.6 (84.3–88.9)	98.8 (97.9–99.5)	100.0 (100.0–100.0)	100.0 (100.0–100.0)	76.6 (73.8–79.6)	45.9 (42.6–49.1

B.3 Other metrics

In this section we assess structural diversity among the generated molecules using the MOSES framework (Polykovskiy et al. 2020), and check for presence of undesirable functional groups in each molecule using the REOS (Rapid Elimination of Swill (W. P. Walters, Murcko, et al. 1999; W. P. Walters, Stahl, et al. 1998)) functionality in the `useful_rdkit_utils` toolkit (P. Walters 2025). This includes PAINS filters ('PAINS') (Baell et al. 2010), the SureChEMBL Non-MedChem Friendly SMARTS ('SureChEMBL') (*SureChEMBL* 2025), the Bristol-Myers Squibb HTS Deck filters ('BMS') (Pearce et al. 2006), the NIH MLSMR Excluded Functionality filters ('MLSMR') (NIH 2025), the University of Dundee NTD Screening Library Filters ('Dundee') (Brenk et al. 2008), filters of unwanted fragments derived by Inpharmatica Ltd. ('Inpharmatica') (Ehrt et al. 2020), the Pfizer lint filters ('LINT') (Blake 2025), and the Glaxo Wellcome Hard filters ('Glaxo') (Hann et al. 1999).

Dataset	Diversity	REOS Filter Pass Rate, %						
		Glaxo	Dundee	BMS	PAINS	SureChEMBL	MLSMR	Inpharmatica	LINT
No conditioning									
QM9	0.81	70.7 (67.8–73.5)	58.3 (55.3–61.4)	59.5 (56.4–62.6)	95.7 (94.4–96.9)	69.4 (66.5–72.2)	57.7 (54.6–60.7)	65.9 (62.9–68.8)	67.9 (64.9–70.8)
$GEOM_{no\ h}$	0.68	34.4 (31.5–37.3)	10.3 (8.5–12.2)	12.5 (10.5–14.6)	84.6 (82.3–86.8)	28.6 (25.8–31.5)	14.9 (12.7–17.1)	24.1 (21.5–26.7)	20.9 (18.5–23.4)
ZINC	0.63	49.9 (44.8–50.7)	8.7 (6.7–10.0)	2.3 (1.3–3.1)	73.8 (67.8–73.3)	46.0 (41.0–47.0)	19.8 (16.7–21.3)	38.0 (33.5–39.3)	38.8 (34.3–40.0)
Conditioning on distortion factor									
QM9	0.72	44.1 (41.1–47.2)	34.3 (31.4–37.3)	42.6 (39.6–45.8)	86.4 (84.2–88.5)	41.6 (38.6–44.7)	37.7 (34.8–40.8)	41.7 (38.7–44.8)	40.1 (37.2–43.2)
$GEOM_{no\ h}$	0.74	57.7 (54.6–60.8)	23.4 (20.8–26.0)	21.5 (19.0–24.1)	92.4 (90.7–94.0)	49.5 (46.4–52.6)	33.7 (30.8–36.7)	49.3 (46.2–52.4)	50.7 (47.6–53.8)
ZINC	0.77	77.2 (76.9–82.0)	12.9 (11.2–15.4)	4.8 (3.6–6.4)	92.6 (93.8–96.6)	70.6 (69.8–75.4)	23.7 (21.7–27.1)	56.8 (55.2–61.4)	54.4 (52.8–59.1)
GCDM									
$GEOM_{no\ h}$ baseline	0.86	77.5 (74.9–80.0)	29.2 (26.4–32.0)	29.1 (26.3–32.0)	99.6 (99.2–99.9)	65.5 (62.5–68.4)	46.0 (42.9–49.1)	58.5 (55.4–61.6)	60.1 (57.1–63.1)
$GEOM_{no\ h}$ conditioned	0.84	73.5 (66.7–80.2)	37.7 (30.2–45.1)	29.0 (22.2–35.8)	99.4 (98.1–100.0)	68.5 (61.1–75.3)	38.3 (30.9–45.7)	56.2 (48.8–63.6)	59.3 (51.9–66.7)
ZINC baseline	0.61	34.1 (31.2–37.0)	17.6 (15.2–20.0)	26.6 (23.9–29.3)	36.2 (33.2–39.2)	27.3 (24.6–30.1)	9.7 (7.9–11.6)	27.9 (25.2–30.7)	23.5 (20.9–26.1)
ZINC conditioned	0.81	81.1 (77.9–84.2)	18.7 (15.6–21.9)	11.0 (8.6–13.7)	97.2 (95.8–98.4)	71.8 (68.1–75.5)	30.3 (26.6–34.0)	62.9 (58.8–66.7)	58.1 (54.1–62.0)
MolFM									
$GEOM_{no\ h}$ baseline	0.83	66.9 (62.7–71.0)	24.7 (21.0–28.4)	24.3 (20.6–28.0)	98.6 (97.5–99.6)	57.6 (53.3–62.0)	35.9 (31.8–40.0)	55.5 (51.2–59.8)	58.4 (54.1–62.7)
$GEOM_{no\ h}$ conditioned	0.77	58.3 (55.2–61.4)	20.7 (18.2–23.3)	19.6 (17.2–22.1)	94.5 (93.0–95.9)	50.5 (47.4–53.5)	29.2 (26.4–32.0)	50.8 (47.6–53.9)	50.2 (47.1–53.3)
ZINC baseline	0.64	51.3 (48.2–54.4)	7.6 (6.0–9.3)	2.7 (1.8–3.7)	72.0 (69.2–74.8)	46.5 (43.4–49.6)	17.4 (15.1–19.8)	38.4 (35.3–41.4)	40.2 (37.3–43.2)
ZINC conditioned	0.74	72.3 (69.4–75.1)	14.5 (12.3–16.9)	6.6 (5.0–8.3)	93.3 (91.6–94.9)	68.7 (65.7–71.6)	23.6 (20.9–26.4)	57.8 (54.5–61.0)	59.4 (56.2–62.5)

Table B.2: Pass rates for REOS tests assessing baseline and conditional EDM, GCDM and MolFM trained on QM9, $GEOM_{no\ h}$ and ZINC.

References

Abderrahmane, M., Tajmouati, H., Barros Ribeiro Da Silva, V., and Perron, Q. (2024). *Predicting the price of molecules using their predicted synthetic pathways.* doi:10.26434/chemrxiv-2024-8wcfp.

Abramson, J. et al. (2024). "Accurate structure prediction of biomolecular interactions with AlphaFold 3". In: *Nature* 630.8016, pp. 493–500. ISSN: 1476-4687. doi:10.1038/s41586-024-07487-w.

Ahdritz, G. et al. (2024). "OpenFold: retraining AlphaFold2 yields new insights into its learning mechanisms and capacity for generalization". In: *Nature Methods* 21.8, pp. 1514–1524. ISSN: 1548-7105. doi:10.1038/s41592-024-02272-z.

Alakhdar, A., Poczos, B., and Washburn, N. (2024). "Diffusion Models in De Novo Drug Design". In: *Journal of Chemical Information and Modeling* 64.19, pp. 7238–7256. ISSN: 1549-9596. doi:10.1021/acs.jcim.4c01107.

Anderson, B., Hy, T. S., and Kondor, R. (2019). "Cormorant: Covariant Molecular Neural Networks". In: *Advances in Neural Information Processing Systems*. Vol. 32. Curran Associates, Inc.

Axelrod, S. and Gómez-Bombarelli, R. (2022). "GEOM, energy-annotated molecular conformations for property prediction and molecular generation". In: *Scientific Data* 9.1, p. 185. ISSN: 2052-4463. doi:10.1038/s41597-022-01288-4.

Baek, M. et al. (2021). "Accurate prediction of protein structures and interactions using a three-track neural network". In: *Science* 373.6557, pp. 871–876. doi:10.1126/science.abj8754.

Baell, J. B. and Holloway, G. A. (2010). "New Substructure Filters for Removal of Pan Assay Interference Compounds (PAINS) from Screening Libraries and for Their Exclusion in Bioassays". In: *Journal of Medicinal Chemistry* 53.7, pp. 2719–2740. ISSN: 0022-2623. doi:10.1021/jm901137j.

Baldwin, J. J. et al. (1989). "Thienothiopyran-2-sulfonamides: novel topically active carbonic anhydrase inhibitors for the treatment of glaucoma". In: *Journal of Medicinal Chemistry* 32.12, pp. 2510–2513. ISSN: 0022-2623. doi:10.1021/jm00132a003.

Ballester, P. J. and Mitchell, J. B. O. (2010). "A machine learning approach to predicting protein–ligand binding affinity with applications to molecular docking". In: *Bioinformatics* 26.9, pp. 1169–1175. ISSN: 1367-4803. doi:10.1093/bioinformatics/btq112.

Bannwarth, C. et al. (2021). "Extended tight-binding quantum chemistry methods". In: *WIREs Computational Molecular Science* 11.2, e1493. ISSN: 1759-0884. doi:10.1002/wcms.1493.

Berman, H. M. et al. (2000). "The Protein Data Bank". In: *Nucleic Acids Research* 28.1, pp. 235–242. ISSN: 0305-1048. doi:10.1093/nar/28.1.235.

Black, J. W., Crowther, A. F., Shanks, R. G., Smith, L. H., and Dornhorst, A. C. (1964). "A NEW ADRENERGIC: BETA-RECEPTOR ANTAGONIST". In: *The Lancet*

283.7342, pp. 1080–1081. ISSN: 0140-6736, 1474-547X. doi:10.1016/S0140-6736(64)91275-9.

Black, J. W., Duncan, W. a. M., Durant, C. J., Ganellin, C. R., and Parsons, E. M. (1972). "Definition and Antagonism of Histamine H2-receptors". In: *Nature* 236.5347, pp. 385–390. ISSN: 1476-4687. doi:10.1038/236385a0.

Blake, J. F. (2025). "Identification and Evaluation of Molecular Properties Related to Preclinical Optimization and Clinical Fate". In: *http://www.eurekaselect.com* ().

Boby, M. L. et al. (2023). "Open science discovery of potent noncovalent SARS-CoV-2 main protease inhibitors". In: *Science* 382.6671, eabo7201. doi:10.1126/science.abo7201.

Bonnet, P. (2012). "Is chemical synthetic accessibility computationally predictable for drug and lead-like molecules? A comparative assessment between medicinal and computational chemists". In: *European Journal of Medicinal Chemistry* 54, pp. 679–689. ISSN: 0223-5234. doi:10.1016/j.ejmech.2012.06.024.

Boyles, F., Deane, C. M., and Morris, G. M. (2022). "Learning from Docked Ligands: Ligand-Based Features Rescue Structure-Based Scoring Functions When Trained on Docked Poses". In: *Journal of Chemical Information and Modeling* 62.22, pp. 5329–5341. ISSN: 1549-9596. doi:10.1021/acs.jcim.1c00096.

Brenk, R. et al. (2008). "Lessons Learnt from Assembling Screening Libraries for Drug Discovery for Neglected Diseases". In: *ChemMedChem* 3.3, pp. 435–444. ISSN: 1860-7187. doi:10.1002/cmdc.200700139.

Buttenschoen, M., Morris, G. M., and Deane, C. M. (2024). "PoseBusters: AI-based docking methods fail to generate physically valid poses or generalise to novel sequences". In: *Chemical Science* 15.9, pp. 3130–3139. ISSN: 2041-6520, 2041-6539. doi:10.1039/D3SC04185A.

Carbery, A., Skyner, R., Delft, F. von, and Deane, C. M. (2022). "Fragment Libraries Designed to Be Functionally Diverse Recover Protein Binding Information More Efficiently Than Standard Structurally Diverse Libraries". In: *Journal of Medicinal Chemistry* 65.16, pp. 11404–11413. ISSN: 0022-2623. doi:10.1021/acs.jmedchem.2c01004.

Carlson, H. A. (2025). "Protein Flexibility is an Important Component of Structure-Based Drug Discovery". In: *Current Pharmaceutical Design* 8.17 (), pp. 1571–1578. doi:10.2174/1381612023394232.

Chen, L. et al. (2019). "Hidden bias in the DUD-E dataset leads to misleading performance of deep learning in structure-based virtual screening". In: *PLOS ONE* 14.8, e0220113. ISSN: 1932-6203. doi:10.1371/journal.pone.0220113.

Chen, S. and Jung, Y. (2024). "Estimating the synthetic accessibility of molecules with building block and reaction-aware SAScore". In: *Journal of Cheminformatics* 16.1, p. 83. ISSN: 1758-2946. doi:10.1186/s13321-024-00879-0.

Chen, Z., Peng, B., Parthasarathy, S., and Ning, X. (2023). *Shape-conditioned 3D Molecule Generation via Equivariant Diffusion Models*.

Chenthamarakshan, V. et al. (2020). "CogMol: Target-Specific and Selective Drug Design for COVID-19 Using Deep Generative Models". In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin. Vol. 33. Curran Associates, Inc., pp. 4320–4332.

Chhabra, M. (2021). "Chapter 6 - Biological therapeutic modalities". In: *Translational Biotechnology*. Ed. by Y. Hasija. Academic Press, pp. 137–164. ISBN: 978-0-12-821972-0. doi:10.1016/B978-0-12-821972-0.00015-0.

Chiu, W., Schmid, M. F., Pintilie, G. D., and Lawson, C. L. (2021). "Evolution of standardization and dissemination of cryo-EM structures and data jointly by the community, PDB, and EMDB". In: *The Journal of Biological Chemistry* 296, p. 100560. ISSN: 0021-9258. doi:10.1016/j.jbc.2021.100560.

Chorghade, M. S. (2006). *Drug Discovery and Development, Volume 1: Drug Discovery*. John Wiley & Sons. ISBN: 978-0-471-78009-0.

Cohen, T. and Welling, M. (2016). "Group Equivariant Convolutional Networks". In: *Proceedings of The 33rd International Conference on Machine Learning*. PMLR, pp. 2990–2999.

Cohen, T. S. and Welling, M. (2017). "Steerable CNNs". In.

Cook, J. A. (2023). *An Introduction to Clinical Trials*. Oxford University Press. ISBN: 978-0-19-888523-8.

Croll, T. I. (2018). "ISOLDE: a physically realistic environment for model building into low-resolution electron-density maps". In: *Acta Crystallographica Section D: Structural Biology* 74.6, pp. 519–530. ISSN: 2059-7983. doi:10.1107/S2059798318002425.

Curran, P. R. et al. (2020). "Hotspots API: A Python Package for the Detection of Small Molecule Binding Hotspots and Application to Structure-Based Drug Design". In: *Journal of Chemical Information and Modeling* 60.4, pp. 1911–1916. ISSN: 1549-9596. doi:10.1021/acs.jcim.9b00996.

Dai, X., Wu, L., Yoo, S., and Liu, Q. (2023). "Integrating AlphaFold and deep learning for atomistic interpretation of cryo-EM maps". In: *Briefings in Bioinformatics* 24.6, bbad405. ISSN: 1477-4054. doi:10.1093/bib/bbad405.

Degen, J., Wegscheid-Gerlach, C., Zaliani, A., and Rarey, M. (2008). "On the Art of Compiling and Using 'Drug-Like' Chemical Fragment Spaces". In: *ChemMedChem* 3.10, pp. 1503–1507. ISSN: 1860-7187. doi:10.1002/cmdc.200800178.

DeVito, N. J. and Goldacre, B. (2019). "Catalogue of bias: publication bias". In: *BMJ Evidence-Based Medicine* 24.2, pp. 53–54. ISSN: 2515-446X, 2515-4478. doi:10.1136/bmjebm-2018-111107.

Dhariwal, P. and Nichol, A. (2021). *Diffusion Models Beat GANs on Image Synthesis*. doi:10.48550/arXiv.2105.05233.

Díaz-Rovira, A. M., Martín, H., Beuming, T., Díaz, L., Guallar, V., and Ray, S. S. (2023). "Are Deep Learning Structural Models Sufficiently Accurate for Virtual Screening? Application of Docking Algorithms to AlphaFold2 Predicted Structures". In: *Journal of Chemical Information and Modeling* 63.6, pp. 1668–1674. ISSN: 1549-9596, 1549-960X. doi:10.1021/acs.jcim.2c01270.

Dickersin, K. and Min, Y. I. (1993). "NIH clinical trials and publication bias". In: *The Online journal of current clinical trials* Doc No 50, [4967 words, 53 paragraphs]. ISSN: 1059-2725.

Dollar, O., Joshi, N., C. Beck, D. A., and Pfaendtner, J. (2021). "Attention-based generative models for de novo molecular design". In: *Chemical Science* 12.24, pp. 8362–8372. doi:10.1039/D1SC01050F.

Dougherty, T. J. and Pucci, M. J. (2011). *Antibiotic Discovery and Development*. Springer Science & Business Media. ISBN: 978-1-4614-1400-1.

Dreser, H. (1899). "Pharmakologisches über Aspirin (Acetylsalicylsäure)". In: *Archiv für die gesamte Physiologie des Menschen und der Tiere* 76.5, pp. 306–318. ISSN: 1432-2013. doi:10.1007/BF01662127.

Ehrlich, P. and Hata, S. (1910). *Die experimentelle Chemotherapie der Spirillosen*. Berlin, Heidelberg: Springer Berlin Heidelberg. ISBN: 978-3-642-64911-0 978-3-642-64926-4. doi:10.1007/978-3-642-64926-4.

Ehrt, C., Krause, B., Schmidt, R., Ehmki, E. S. R., and Rarey, M. (2020). "SMARTS.plus – A Toolbox for Chemical Pattern Design". In: *Molecular Informatics* 39.12, p. 2000216. doi:10.1002/minf.202000216.

EMDB (2025). *Electron Microscopy Data Bank*.

Emsley, P., Lohkamp, B., Scott, W. G., and Cowtan, K. (2010). "Features and development of Coot". In: *Acta Crystallographica Section D: Biological Crystallography* 66.4, pp. 486–501. ISSN: 0907-4449. doi:10.1107/S0907444910007493.

Enamine REAL Space - Enamine (2025).

Fadini, A. et al. (2025). *AlphaFold as a Prior: Experimental Structure Determination Conditioned on a Pretrained Neural Network*. doi:10.1101/2025.02.18.638828.

Feinberg, E. N. et al. (2018). "PotentialNet for Molecular Property Prediction". In: *ACS Central Science* 4.11, pp. 1520–1530. ISSN: 2374-7943. doi:10.1021/acscentsci.8b00507.

Finzi, M., Stanton, S., Izmailov, P., and Wilson, A. G. (2020). "Generalizing Convolutional Neural Networks for Equivariance to Lie Groups on Arbitrary Continuous Data". In: *Proceedings of the 37th International Conference on Machine Learning*. PMLR, pp. 3165–3176.

Francoeur, P. G. et al. (2020). "Three-Dimensional Convolutional Neural Networks and a Cross-Docked Data Set for Structure-Based Drug Design". In: *Journal of Chemical Information and Modeling* 60.9, pp. 4200–4215. ISSN: 1549-9596. doi:10.1021/acs.jcim.0c00411.

Friesner, R. A. et al. (2004). "Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy". In: *Journal of Medicinal Chemistry* 47.7, pp. 1739–1749. ISSN: 0022-2623. doi:10.1021/jm0306430.

Fu, C., Li, X., Olson, B., Ji, H., and Ji, S. (2024). *Fragment and Geometry Aware Tokenization of Molecules for Structure-Based Drug Design Using Language Models*. doi:10.48550/arXiv.2408.09730.

Gao, Y., Thorn, V., and Thorn, A. (2023). "Errors in structural biology are not the exception". In: *Acta Crystallographica Section D: Structural Biology* 79.3, pp. 206–211. ISSN: 2059-7983. doi:10.1107/S2059798322011901.

Gao, Z., Wang, X., Blumenfeld Gaines, B., Shi, X., Bi, J., and Song, M. (2023). "Fragment-based deep molecular generation using hierarchical chemical graph representation and multi-resolution graph variational autoencoder". In: *Molecular Informatics* 42.5, p. 2200215. ISSN: 1868-1751. doi:10.1002/minf.202200215.

Gebauer, N., Gastegger, M., and Schütt, K. (2019). "Symmetry-adapted generation of 3d point sets for the targeted discovery of molecules". In: *Advances in Neural Information Processing Systems*. Vol. 32. Curran Associates, Inc.

Ghorbani, M., Gendelev, L., Beroza, P., and Keiser, M. J. (2023). *Autoregressive fragment-based diffusion for pocket-aware ligand design*. doi:10.48550/arXiv.2401.05370.

Giri, N. and Cheng, J. (2025). *Atomic Protein Structure Modeling from Cryo-EM Using Multi-Modal Deep Learning and AlphaFold3*. doi:10.1101/2025.03.16.643561.

Gonen, T. and Nannenga, B. L., eds. (2021). *cryoEM: Methods and Protocols*. Vol. 2215. Methods in Molecular Biology. New York, NY: Springer US. ISBN: 978-1-07-160965-1 978-1-07-160966-8. doi:10.1007/978-1-0716-0966-8.

Goodfellow, I., Courville, A., and Bengio, Y. (2016). *Deep learning*. Adaptive computation and machine learning. Cambridge, Massachusetts: The MIT Press. ISBN: 978-0-262-03561-3 978-0-262-33737-3.

Greer, J., Erickson, J. W., Baldwin, J. J., and Varney, M. D. (1994). "Application of the Three-Dimensional Structures of Protein Target Molecules in Structure-Based Drug Design". In: *Journal of Medicinal Chemistry* 37.8, pp. 1035–1054. ISSN: 0022-2623. doi:10.1021/jm00034a001.

Groom, C. R., Bruno, I. J., Lightfoot, M. P., and Ward, S. C. (2016). "The Cambridge Structural Database". In: *Acta Crystallographica Section B: Structural Science, Crystal Engineering and Materials* 72.2, pp. 171–179. ISSN: 2052-5206. doi:10.1107/S2052520616003954.

Guan, J. et al. (2023). "DecompDiff: Diffusion Models with Decomposed Priors for Structure-Based Drug Design". In: *Proceedings of the 40th International Conference on Machine Learning*. PMLR, pp. 11827–11846.

Guo, J. et al. (2023). "Link-INVENT: generative linker design with reinforcement learning". In: *Digital Discovery* 2.2, pp. 392–408. ISSN: 2635-098X. doi:10.1039/D2DD00115B.

Hadfield, T. E., Imrie, F., Merritt, A., Birchall, K., and Deane, C. M. (2022). "Incorporating Target-Specific Pharmacophoric Information into Deep Generative Models for Fragment Elaboration". In: *Journal of Chemical Information and Modeling* 62.10, pp. 2280–2292. ISSN: 1549-9596, 1549-960X. doi:10.1021/acs.jcim.1c01311.

Hann, M., Hudson, B., Lewell, X., Lifely, R., Miller, L., and Ramsden, N. (1999). "Strategic Pooling of Compounds for High-Throughput Screening". In: *Journal of Chemical Information and Computer Sciences* 39.5, pp. 897–902. ISSN: 0095-2338. doi:10.1021/ci990423o.

Harris, C. et al. (2023). *Benchmarking Generated Poses: How Rational is Structure-based Drug Design with Generative Models?* doi:10.48550/arXiv.2308.07413.

He, X., Hatcher, E., Eriksson, L., Widmalm, G., and MacKerell, A. D. (2013). "Bifurcated Hydrogen Bonding and Asymmetric Fluctuations in a Carbohydrate Crystal Studied via X-ray Crystallography and Computational analysis". In: *The journal of physical chemistry. B* 117.25, pp. 7546–7553. ISSN: 1520-6106. doi:10.1021/jp403719g.

Healthcare, G. (2022). *Biologic sales forecast to pass innovative small molecule sales in next five years*.

Ho, J., Jain, A., and Abbeel, P. (2020). "Denoising Diffusion Probabilistic Models". In: *Advances in Neural Information Processing Systems*. Vol. 33. Curran Associates, Inc., pp. 6840–6851.

Hochuli, J., Helbling, A., Skaist, T., Ragoza, M., and Koes, D. R. (2018). "Visualizing convolutional neural network protein-ligand scoring". In: *Journal of Molecular Graphics and Modelling* 84, pp. 96–108. ISSN: 1093-3263. doi:10.1016/j.jmgm.2018.06.005.

Holcomb, M., Chang, Y.-T., Goodsell, D. S., and Forli, S. (2023). "Evaluation of AlphaFold2 structures as docking targets". In: *Protein Science* 32.1, e4530. ISSN: 0961-8368, 1469-896X. doi:10.1002/pro.4530.

Hoogeboom, E., Satorras, V. G., Vignac, C., and Welling, M. (2022). *Equivariant Diffusion for Molecule Generation in 3D*.

Howard, S. and Abell, C. (2015). *Fragment-Based Drug Discovery*. Royal Society of Chemistry. ISBN: 978-1-78262-565-0.

Huang, L., Zhang, H., Xu, T., and Wong, K.-C. (2022). *MDM: Molecular Diffusion Model for 3D Molecule Generation*.

Huang, N., Shoichet, B. K., and Irwin, J. J. (2006). "Benchmarking Sets for Molecular Docking". In: *Journal of Medicinal Chemistry* 49.23, pp. 6789–6801. ISSN: 0022-2623. doi:10.1021/jm0608356.

Huang, Y., Peng, X., Ma, J., and Zhang, M. (2022). "3DLinker: An E(3) Equivariant Variational Autoencoder for Molecular Linker Design". In: *Proceedings of the 39th International Conference on Machine Learning*. PMLR, pp. 9280–9294.

Huff, S. et al. (2022). "Discovery and Mechanism of SARS-CoV-2 Main Protease Inhibitors". In: *Journal of Medicinal Chemistry* 65.4, pp. 2866–2879. ISSN: 0022-2623. doi:10.1021/acs.jmedchem.1c00566.

Hughes, J., Rees, S., Kalindjian, S., and Philpott, K. (2011). "Principles of early drug discovery". In: *British Journal of Pharmacology* 162.6, pp. 1239–1249. ISSN: 0007-1188. doi:10.1111/j.1476-5381.2010.01127.x.

Hutchings, M. I., Truman, A. W., and Wilkinson, B. (2019). "Antibiotics: past, present and future". In: *Current Opinion in Microbiology*. Antimicrobials 51, pp. 72–80. ISSN: 1369-5274. doi:10.1016/j.mib.2019.10.008.

Igashov, I. et al. (2024). "Equivariant 3D-conditional diffusion model for molecular linker design". In: *Nature Machine Intelligence* 6.4, pp. 417–427. ISSN: 2522-5839. doi:10.1038/s42256-024-00815-9.

Imprachim, N., Yosaatmadja, Y., and Newman, J. A. (2022). "Crystal structures and fragment screening of SARS-CoV-2 NSP14 reveal details of exoribonuclease activation and mRNA capping and provide starting points for antiviral drug development". In: *Nucleic Acids Research* 51.1, pp. 475–487. ISSN: 0305-1048. doi:10.1093/nar/gkac1207.

Imrie, F., Bradley, A. R., Schaar, M. van der, and Deane, C. M. (2020). "Deep Generative Models for 3D Linker Design". In: *Journal of Chemical Information and Modeling* 60.4, pp. 1983–1995. ISSN: 1549-9596. doi:10.1021/acs.jcim.9b01120.

Jamali, K., Käll, L., Zhang, R., Brown, A., Kimanius, D., and Scheres, S. H. W. (2024). "Automated model building and protein identification in cryo-EM maps". In: *Nature* 628.8007, pp. 450–457. ISSN: 1476-4687. doi:10.1038/s41586-024-07215-4.

Jiang, D. et al. (2021). "InteractionGraphNet: A Novel and Efficient Deep Graph Representation Learning Framework for Accurate Protein–Ligand Interaction Predictions". In: *Journal of Medicinal Chemistry* 64.24, pp. 18209–18232. ISSN: 0022-2623. doi:10.1021/acs.jmedchem.1c01830.

Jiménez-Rosés, M. et al. (2022). "Combined docking and machine learning identify key molecular determinants of ligand pharmacological activity on β2 adrenoceptor". In: *Pharmacology Research & Perspectives* 10.5, e00994. ISSN: 2052-1707. doi:10.1002/prp2.994.

Jumper, J. et al. (2021). "Highly accurate protein structure prediction with AlphaFold". In: *Nature* 596.7873, pp. 583–589. ISSN: 1476-4687. doi:10.1038/s41586-021-03819-2.

Kempf, D. J. et al. (1995). "ABT-538 is a potent inhibitor of human immunodeficiency virus protease and has high oral bioavailability in humans". In: *Proceedings of the*

National Academy of Sciences of the United States of America 92.7, pp. 2484–2488. ISSN: 0027-8424. doi:10.1073/pnas.92.7.2484.

Koes, D. R., Baumgartner, M. P., and Camacho, C. J. (2013). "Lessons Learned in Empirical Scoring with smina from the CSAR 2011 Benchmarking Exercise". In: *Journal of Chemical Information and Modeling* 53.8, pp. 1893–1904. ISSN: 1549-9596. doi:10.1021/ci300604z.

Kretschmer, F., Seipp, J., Ludwig, M., Klau, G. W., and Böcker, S. (2025). "Coverage bias in small molecule machine learning". In: *Nature Communications* 16.1, p. 554. ISSN: 2041-1723. doi:10.1038/s41467-024-55462-w.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). "ImageNet Classification with Deep Convolutional Neural Networks". In: *Advances in Neural Information Processing Systems*. Vol. 25. Curran Associates, Inc.

Kryshtafovych, A., Schwede, T., Topf, M., Fidelis, K., and Moult, J. (2021). "Critical assessment of methods of protein structure prediction (CASP)—Round XIV". In: *Proteins: Structure, Function, and Bioinformatics* 89.12, pp. 1607–1617. ISSN: 0887-3585, 1097-0134. doi:10.1002/prot.26237.

Kufareva, I., Ilatovskiy, A. V., and Abagyan, R. (2012). "Pocketome: an encyclopedia of small-molecule binding sites in 4D". In: *Nucleic Acids Research* 40.Database issue, pp. D535–D540. ISSN: 0305-1048. doi:10.1093/nar/gkr825.

Kühlbrandt, W. (2014). "The Resolution Revolution". In: *Science* 343.6178, pp. 1443–1444. doi:10.1126/science.1251652.

Kuntz, I. D. (1992). "Structure-Based Strategies for Drug Design and Discovery". In: *Science* 257.5073, pp. 1078–1082. doi:10.1126/science.257.5073.1078.

Landrum, G. et al. (2023). *rdkit/rdkit: 2022_09_4 (Q3 2022) Release*. doi:10.5281/zenodo.7541264.

Laviolle, B. et al. (2025). "Trends of Publication of Negative Trials Over Time". In: *Clinical Pharmacology & Therapeutics* 117.3, pp. 818–825. ISSN: 1532-6535. doi:10.1002/cpt.3535.

Lawson, C. L., Kryshtafovych, A., et al. (2021). "Cryo-EM model validation recommendations based on outcomes of the 2019 EMDataResource challenge". In: *Nature Methods* 18.2, pp. 156–164. ISSN: 1548-7105. doi:10.1038/s41592-020-01051-w.

Lawson, C. L., Patwardhan, A., et al. (2016). "EMDataBank unified data resource for 3DEM". In: *Nucleic Acids Research* 44.D1, pp. D396–D403. ISSN: 0305-1048. doi:10.1093/nar/gkv1126.

Lecun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). "Gradient-based learning applied to document recognition". In: *Proceedings of the IEEE* 86.11, pp. 2278–2324. ISSN: 1558-2256. doi:10.1109/5.726791.

Lewis, T. (2023). *One of the Biggest Problems in Biology Has Finally Been Solved*.

Li, S., Terashi, G., Zhang, Z., and Kihara, D. (2025). "Advancing structure modeling from cryo-EM maps with deep learning". In: *Biochemical Society Transactions* 53.01, pp. 259–265. ISSN: 0300-5127. doi:10.1042/BST20240784.

Li, Y., Rezaei, M. A., Li, C., Li, X., and Wu, D. (2019). *DeepAtom: A Framework for Protein-Ligand Binding Affinity Prediction*.

Lim, J., Ryu, S., Park, K., Choe, Y. J., Ham, J., and Kim, W. Y. (2019). "Predicting Drug–Target Interaction Using a Novel Graph Neural Network with 3D

Structure-Embedded Graph Representation". In: *Journal of Chemical Information and Modeling* 59.9, pp. 3981–3988. ISSN: 1549-9596. doi:10.1021/acs.jcim.9b00387.

Lin, H., Huang, Y., Liu, M., Li, X., Ji, S., and Li, S. Z. (2022). *DiffBP: Generative Diffusion of 3D Molecules for Target Protein Binding.*

Lin, Z. et al. (2023). "Evolutionary-scale prediction of atomic-level protein structure with a language model". In: *Science* 379.6637, pp. 1123–1130. doi:10.1126/science.ade2574.

Liu, M., Luo, Y., Uchino, K., Maruhashi, K., and Ji, S. (2022). "Generating 3D Molecules for Target Protein Binding". In: *Proceedings of the 39th International Conference on Machine Learning*. PMLR, pp. 13912–13924.

Liu, X. et al. (2025). *FragmentGPT: A Unified GPT Model for Fragment Growing, Linking, and Merging in Molecular Design.* doi:10.48550/arXiv.2509.11044.

Liu, Z. et al. (2017). "Forging the Basis for Developing Protein–Ligand Interaction Scoring Functions". In: *Accounts of Chemical Research* 50.2, pp. 302–309. ISSN: 0001-4842. doi:10.1021/acs.accounts.6b00491.

Lu, J., Hou, X., Wang, C., and Zhang, Y. (2019). "Incorporating Explicit Water Molecules and Ligand Conformation Stability in Machine-Learning Scoring Functions". In: *Journal of Chemical Information and Modeling* 59.11, pp. 4540–4549. ISSN: 1549-9596. doi:10.1021/acs.jcim.9b00645.

Luo, S., Chen, T., et al. (2023). *One Transformer Can Understand Both 2D & 3D Molecular Data.* doi:10.48550/arXiv.2210.01765.

Luo, S., Guan, J., Ma, J., and Peng, J. (2022). *A 3D Generative Model for Structure-Based Drug Design.* doi:10.48550/arXiv.2203.10446.

Luo, Y. and Ji, S. (2022). "An Autoregressive Flow Model for 3D Molecular Geometry Generation from Scratch". In.

Lyu, J. et al. (2019). "Ultra-large library docking for discovering new chemotypes". In: *Nature* 566.7743, pp. 224–229. ISSN: 0028-0836, 1476-4687. doi:10.1038/s41586-019-0917-9.

Makurvet, F. D. (2021). "Biologics vs. small molecules: Drug costs and patient access". In: *Medicine in Drug Discovery* 9, p. 100075. ISSN: 2590-0986. doi:10.1016/j.medidd.2020.100075.

Maloney, M. P. et al. (2023). "Negative Data in Data Sets for Machine Learning Training". In: *The Journal of Organic Chemistry* 88.9, pp. 5239–5241. ISSN: 0022-3263. doi:10.1021/acs.joc.3c00844.

McNutt, A. T. et al. (2021). "GNINA 1.0: molecular docking with deep learning". In: *Journal of Cheminformatics* 13.1, p. 43. ISSN: 1758-2946. doi:10.1186/s13321-021-00522-2.

Morehead, A. and Cheng, J. (2023). *Geometry-Complete Diffusion for 3D Molecule Generation and Optimization.*

Morris, G. M. et al. (2009). "AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility". In: *Journal of Computational Chemistry* 30.16, pp. 2785–2791. ISSN: 1096-987X. doi:10.1002/jcc.21256.

Morrone, J. A., Weber, J. K., Huynh, T., Luo, H., and Cornell, W. D. (2020). "Combining Docking Pose Rank and Structure with Deep Learning Improves Protein–Ligand Binding Mode Prediction over a Baseline Docking Approach". In: *Journal of Chemical Information and Modeling* 60.9, pp. 4170–4179. ISSN: 1549-9596. doi:10.1021/acs.jcim.9b00927.

Mukaidaisi, M., Vu, A., Grantham, K., Tchagang, A., and Li, Y. (2022). "Multi-Objective Drug Design Based on Graph-Fragment Molecular Representation and Deep Evolutionary Learning". In: *Frontiers in Pharmacology* 13. ISSN: 1663-9812. doi:10.3389/fphar.2022.920747.

Mullard, A. (2025). "2024 FDA approvals". In: *Nature Reviews Drug Discovery* 24.2, pp. 75–82. ISSN: 1474-1784. doi:10.1038/d41573-025-00001-5.

Murray, C. W., Newell, D. R., and Angibaud, P. (2019). "A successful collaboration between academia, biotech and pharma led to discovery of erdafitinib, a selective FGFR inhibitor recently approved by the FDA". In: *MedChemComm* 10.9, pp. 1509–1511. ISSN: 2040-2511. doi:10.1039/C9MD90044F.

Mysinger, M. M., Carchia, M., Irwin, J. J., and Shoichet, B. K. (2012). "Directory of Useful Decoys, Enhanced (DUD-E): Better Ligands and Decoys for Better Benchmarking". In: *Journal of Medicinal Chemistry* 55.14, pp. 6582–6594. ISSN: 0022-2623. doi:10.1021/jm300687e.

Navia, M. A. et al. (1989). "Three-dimensional structure of aspartyl protease from human immunodeficiency virus HIV-1". In: *Nature* 337.6208, pp. 615–620. ISSN: 1476-4687. doi:10.1038/337615a0.

Newton, T. P., Zimanyi, C., and Kopylov, M. (2025). *RCSB PDB - 8FEI: CryoEM structure of Conalbumin from chicken egg white (sigma-Cas 1391-06-6) - to be published.*

NIH (2025). *MLSMR Excluded Functionality Filters - NIH Molecular Libraries.*

O'Boyle, N. M., Banck, M., James, C. A., Morley, C., Vandermeersch, T., and Hutchison, G. R. (2011). "Open Babel: An open chemical toolbox". In: *Journal of Cheminformatics* 3.1, p. 33. ISSN: 1758-2946. doi:10.1186/1758-2946-3-33.

Passaro, S. et al. (2025). *Boltz-2: Towards Accurate and Efficient Binding Affinity Prediction.* doi:10.1101/2025.06.14.659707.

Patani, H. et al. (2016). "Landscape of activating cancer mutations in FGFR kinases and their differential responses to inhibitors in clinical use". In: *Oncotarget* 7.17, pp. 24252–24268. ISSN: 1949-2553. doi:10.18632/oncotarget.8132.

Patrick, G. L. (2013). *An introduction to medicinal chemistry.* 5th ed. Oxford: Oxford University Press. ISBN: 978-0-19-969739-7.

Pearce, B. C., Sofia, M. J., Good, A. C., Drexler, D. M., and Stock, D. A. (2006). "An Empirical Process for the Design of High-Throughput Screening Deck Filters". In: *Journal of Chemical Information and Modeling* 46.3, pp. 1060–1068. ISSN: 1549-9596. doi:10.1021/ci050504m.

Pelletier, P. J. (1820). *Recherches chimiques sur les quinquinas.* Paris : Crochard.

Peng, X., Guan, J., Liu, Q., and Ma, J. (2023). "MolDiff: Addressing the Atom-Bond Inconsistency Problem in 3D Molecule Diffusion Generation". In: *Proceedings of the 40th International Conference on Machine Learning.* PMLR, pp. 27611–27629.

Peng, X., Luo, S., Guan, J., Xie, Q., Peng, J., and Ma, J. (2022). *Pocket2Mol: Efficient Molecular Sampling Based on 3D Protein Pockets.*

Pfab, J., Phan, N. M., and Si, D. (2021). "DeepTracer for fast de novo cryo-EM protein structure modeling and special studies on CoV-related complexes". In: *Proceedings of the National Academy of Sciences* 118.2, e2017525118. doi:10.1073/pnas.2017525118.

Poelking, C., Chessari, G., Murray, C. W., Hall, R. J., Colwell, L., and Verdonk, M. (2022). *Meaningful machine learning models and machine-learned pharmacophores from fragment screening campaigns.*

Pogány, P., Arad, N., Genway, S., and Pickett, S. D. (2019). "De Novo Molecule Design by Translating from Reduced Graphs to SMILES". In: *Journal of Chemical Information and Modeling* 59.3, pp. 1136–1146. ISSN: 1549-9596. doi:10.1021/acs.jcim.8b00626.

Polykovskiy, D. et al. (2020). "Molecular Sets (MOSES): A Benchmarking Platform for Molecular Generation Models". In: *Frontiers in Pharmacology* 11. ISSN: 1663-9812. doi:10.3389/fphar.2020.565644.

Qiang, B. et al. (2023). *Coarse-to-Fine: a Hierarchical Diffusion Model for Molecule Generation in 3D*. doi:10.48550/arXiv.2305.13266.

Radoux, C. J., Olsson, T. S. G., Pitt, W. R., Groom, C. R., and Blundell, T. L. (2016). "Identifying Interactions that Determine Fragment Binding at Protein Hotspots". In: *Journal of Medicinal Chemistry* 59.9, pp. 4314–4325. ISSN: 0022-2623. doi:10.1021/acs.jmedchem.5b01980.

Raghu, R., Levy, A., Wetzstein, G., and Zhong, E. D. (2025). *Multiscale guidance of AlphaFold3 with heterogeneous cryo-EM data*. doi:10.48550/arXiv.2506.04490.

Ragoza, M., Hochuli, J., Idrobo, E., Sunseri, J., and Koes, D. R. (2017). "Protein–Ligand Scoring with Convolutional Neural Networks". In: *Journal of Chemical Information and Modeling* 57.4, pp. 942–957. ISSN: 1549-9596. doi:10.1021/acs.jcim.6b00740.

Ragoza, M., Masuda, T., and Koes, D. R. (2022). "Generating 3D molecules conditional on receptor binding sites with deep generative models". In: *Chemical Science* 13.9, pp. 2701–2713. ISSN: 2041-6539. doi:10.1039/D1SC05976A.

Ragoza, M., Turner, L., and Koes, D. R. (2017). *Ligand Pose Optimization with Atomic Grid-Based Convolutional Neural Networks*. doi:10.48550/arXiv.1710.07400.

Ramakrishnan, R., Dral, P. O., Rupp, M., and Lilienfeld, O. A. von (2014). "Quantum chemistry structures and properties of 134 kilo molecules". In: *Scientific Data* 1.1, p. 140022. ISSN: 2052-4463. doi:10.1038/sdata.2014.22.

Ramesh, A. et al. (2021). *Zero-Shot Text-to-Image Generation*. doi:10.48550/arXiv.2102.12092.

RCSB-PDB (2025). *PDB Statistics: PDB Data Distribution by Experimental Method and Molecular Type*.

Riniker, S. and Landrum, G. A. (2013). "Open-source platform to benchmark fingerprints for ligand-based virtual screening". In: *Journal of Cheminformatics* 5.1, p. 26. ISSN: 1758-2946. doi:10.1186/1758-2946-5-26.

Roberts, N. A. et al. (1990). "Rational Design of Peptide-Based HIV Proteinase Inhibitors". In: *Science* 248.4953, pp. 358–361. doi:10.1126/science.2183354.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. (2022). *High-Resolution Image Synthesis with Latent Diffusion Models*. doi:10.48550/arXiv.2112.10752.

Ross, J., Belgodere, B., Hoffman, S. C., Chenthamarakshan, V., Mroueh, Y., and Das, P. (2024). *GP-MoLFormer: A Foundation Model For Molecular Generation*.

Ruddigkeit, L., Deursen, R. van, Blum, L. C., and Reymond, J.-L. (2012). "Enumeration of 166 Billion Organic Small Molecules in the Chemical Universe Database GDB-17". In: *Journal of Chemical Information and Modeling* 52.11, pp. 2864–2875. ISSN: 1549-9596. doi:10.1021/ci300415d.

Salentin, S., Schreiber, S., Haupt, V. J., Adasme, M. F., and Schroeder, M. (2015). "PLIP: fully automated protein–ligand interaction profiler". In: *Nucleic Acids Research* 43.Web Server issue, W443–W447. ISSN: 0305-1048. doi:10.1093/nar/gkv315.

Salimans, T. and Ho, J. (2022). *Progressive Distillation for Fast Sampling of Diffusion Models*. doi:10.48550/arXiv.2202.00512.

Samek, W., Binder, A., Montavon, G., Bach, S., and Müller, K.-R. (2015). *Evaluating the visualization of what a Deep Neural Network has learned*. doi:10.48550/arXiv.1509.06321.

Sanchez-Garcia, R. et al. (2023). "CoPriNet: graph neural networks provide accurate and rapid compound price prediction for molecule prioritisation". In: *Digital Discovery* 2.1, pp. 103–111. ISSN: 2635-098X. doi:10.1039/D2DD00071G.

Santarossa, C. C. et al. (2021). "Targeting allostery in the Dynein motor domain with small molecule inhibitors". In: *Cell Chemical Biology* 28.10, 1460–1473.e15. ISSN: 2451-9456, 2451-9448. doi:10.1016/j.chembiol.2021.04.024.

Satorras, V. G., Hoogeboom, E., and Welling, M. (2022). *E(n) Equivariant Graph Neural Networks*. doi:10.48550/arXiv.2102.09844.

Scantlebury, J., Brown, N., Von Delft, F., and Deane, C. M. (2020). "Data Set Augmentation Allows Deep Learning-Based Virtual Screening to Better Generalize to Unseen Target Classes and Highlight Important Binding Interactions". In: *Journal of Chemical Information and Modeling* 60.8, pp. 3722–3730. ISSN: 1549-9596. doi:10.1021/acs.jcim.0c00263.

Scantlebury, J., Vost, L., et al. (2023). "A Small Step Toward Generalizability: Training a Machine Learning Scoring Function for Structure-Based Virtual Screening". In: *Journal of Chemical Information and Modeling* 63.10, pp. 2960–2974. ISSN: 1549-9596. doi:10.1021/acs.jcim.3c00322.

Scapin, G. et al. (2015). "Structure of full-length human anti-PD1 therapeutic IgG4 antibody pembrolizumab". In: *Nature Structural & Molecular Biology* 22.12, pp. 953–958. ISSN: 1545-9985. doi:10.1038/nsmb.3129.

Scardino, V., Di Filippo, J. I., and Cavasotto, C. N. (2023). "How good are AlphaFold models for docking-based virtual screening?" In: *iScience* 26.1, p. 105920. ISSN: 25890042. doi:10.1016/j.isci.2022.105920.

Scarselli, F., Gori, M., Tsoi, A. C., Hagenbuchner, M., and Monfardini, G. (2009). "The Graph Neural Network Model". In: *IEEE Transactions on Neural Networks* 20.1, pp. 61–80. ISSN: 1941-0093. doi:10.1109/TNN.2008.2005605.

Schneuing, A. et al. (2024). "Structure-based drug design with equivariant diffusion models". In: *Nature Computational Science* 4.12, pp. 899–909. ISSN: 2662-8457. doi:10.1038/s43588-024-00737-x.

Schuller, M. et al. (2021). "Fragment binding to the Nsp3 macrodomain of SARS-CoV-2 identified through crystallographic screening and computational docking". In: *Science Advances* 7.16, eabf8711. doi:10.1126/sciadv.abf8711.

Sertuerner (1817). "Ueber das Morphium, eine neue salzfähige Grundlage, und die Mekonsäure, als Hauptbestandtheile des Opiums". In: *Annalen der Physik* 55.1, pp. 56–89. ISSN: 1521-3889. doi:10.1002/andp.18170550104.

Sharma, B. et al. (2023). "Accurate clinical toxicity prediction using multi-task deep neural nets and contrastive molecular explanations". In: *Scientific Reports* 13.1, p. 4908. ISSN: 2045-2322. doi:10.1038/s41598-023-31169-8.

Shoichet, B. K. (2004). "Virtual screening of chemical libraries". In: *Nature* 432.7019, pp. 862–865. ISSN: 1476-4687. doi:10.1038/nature03197.

Sieg, J., Flachsenberg, F., and Rarey, M. (2019). "In Need of Bias Control: Evaluating Chemical Data for Machine Learning in Structure-Based Virtual Screening". In: *Journal of Chemical Information and Modeling* 59.3, pp. 947–961. ISSN: 1549-9596. doi:10.1021/acs.jcim.8b00712.

Škrinjar, P., Eberhardt, J., Durairaj, J., and Schwede, T. (2025). *Have protein-ligand co-folding methods moved beyond memorisation?* doi:10.1101/2025.02.03.636309.

Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., and Ganguli, S. (2015). "Deep Unsupervised Learning using Nonequilibrium Thermodynamics". In: *Proceedings of the 32nd International Conference on Machine Learning*. PMLR, pp. 2256–2265.

Song, J., Meng, C., and Ermon, S. (2022). *Denoising Diffusion Implicit Models*. doi:10.48550/arXiv.2010.02502.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. (2021). *Score-Based Generative Modeling through Stochastic Differential Equations*. doi:10.48550/arXiv.2011.13456.

Song, Y., Gong, J., et al. (2023). *Equivariant Flow Matching with Hybrid Probability Transport*. doi:10.48550/arXiv.2312.07168.

Souers, A. J. et al. (2013). "ABT-199, a potent and selective BCL-2 inhibitor, achieves antitumor activity while sparing platelets". In: *Nature Medicine* 19.2, pp. 202–208. ISSN: 1546-170X. doi:10.1038/nm.3048.

Spiegel, J. O. and Durrant, J. D. (2020). "AutoGrow4: an open-source genetic algorithm for de novo drug design and lead optimization". In: *Journal of Cheminformatics* 12.1, p. 25. ISSN: 1758-2946. doi:10.1186/s13321-020-00429-4.

Sterling, T. and Irwin, J. J. (2015). "ZINC 15 – Ligand Discovery for Everyone". In: *Journal of Chemical Information and Modeling* 55.11, pp. 2324–2337. ISSN: 1549-9596. doi:10.1021/acs.jcim.5b00559.

Stromgaard, K., Krogsgaard-Larsen, P., and Madsen, U., eds. (2009). *Textbook of Drug Design and Discovery*. 0th ed. CRC Press. ISBN: 978-0-429-11124-2. doi:10.1201/9781439882405.

Su, M. et al. (2019). "Comparative Assessment of Scoring Functions: The CASF-2016 Update". In: *Journal of Chemical Information and Modeling* 59.2, pp. 895–913. ISSN: 1549-9596. doi:10.1021/acs.jcim.8b00545.

SureChEMBL (2025).

Teodoro, M. L. and Kavraki, L. E. (2025). "Conformational Flexibility Models for the Receptor in Structure Based Drug Design". In: *Current Pharmaceutical Design* 9.20 (), pp. 1635–1648. doi:10.2174/1381612033454595.

Torng, W. and Altman, R. B. (2019). "Graph Convolutional Neural Networks for Predicting Drug-Target Interactions". In: *Journal of Chemical Information and Modeling* 59.10, pp. 4131–4149. ISSN: 1549-9596. doi:10.1021/acs.jcim.9b00628.

Townshend, R. J. L. et al. (2022). *ATOM3D: Tasks On Molecules in Three Dimensions*. doi:10.48550/arXiv.2012.04035.

Trott, O. and Olson, A. J. (2010). "AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading". In: *Journal of Computational Chemistry* 31.2, pp. 455–461. ISSN: 1096-987X. doi:10.1002/jcc.21334.

Tsai, J. et al. (2008). "Discovery of a selective inhibitor of oncogenic B-Raf kinase with potent antimelanoma activity". In: *Proceedings of the National Academy of Sciences* 105.8, pp. 3041–3046. doi:10.1073/pnas.0711741105.

Tupper, K. W. (2012). "Psychoactive Substances and the English Language: "Drugs," Discourses, and Public Policy". In: *Contemporary Drug Problems* 39.3, pp. 461–492. ISSN: 0091-4509. doi:10.1177/009145091203900306.

Urick, B. Y. and Meggs, E. V. (2019). "Towards a Greater Professional Standing: Evolution of Pharmacy Practice and Education, 1920–2020". In: *Pharmacy* 7.3, p. 98. ISSN: 2226-4787. doi:10.3390/pharmacy7030098.

Vaswani, A. et al. (2017). "Attention is All you Need". In: *Advances in Neural Information Processing Systems*. Vol. 30. Curran Associates, Inc.

Verdonk, M. L., Cole, J. C., Hartshorn, M. J., Murray, C. W., and Taylor, R. D. (2003). "Improved protein–ligand docking using GOLD". In: *Proteins: Structure, Function, and Bioinformatics* 52.4, pp. 609–623. ISSN: 1097-0134. doi:10.1002/prot.10465.

Verdonk, M. L., Cole, J. C., and Taylor, R. (1999). "SuperStar: A Knowledge-based Approach for Identifying Interaction Sites in Proteins". In: *Journal of Molecular Biology* 289.4, pp. 1093–1108. ISSN: 0022-2836. doi:10.1006/jmbi.1999.2809.

Vignac, C., Osman, N., Toni, L., and Frossard, P. (2023). "MiDi: Mixed Graph and 3D Denoising Diffusion for Molecule Generation". In: *Machine Learning and Knowledge Discovery in Databases: Research Track*. Ed. by D. Koutra, C. Plant, M. Gomez Rodriguez, E. Baralis, and F. Bonchi. Cham: Springer Nature Switzerland, pp. 560–576. ISBN: 978-3-031-43415-0. doi:10.1007/978-3-031-43415-0_33.

Virtanen, P. et al. (2020). "SciPy 1.0: fundamental algorithms for scientific computing in Python". In: *Nature Methods* 17.3, pp. 261–272. ISSN: 1548-7105. doi:10.1038/s41592-019-0686-2.

Wallach, I. and Heifets, A. (2018). "Most Ligand-Based Classification Benchmarks Reward Memorization Rather than Generalization". In: *Journal of Chemical Information and Modeling* 58.5, pp. 916–932. ISSN: 1549-9596. doi:10.1021/acs.jcim.7b00403.

Walters, P. (2025). *PatWalters/useful_rdkit_utils*.

Walters, W. P., Murcko, A. A., and Murcko, M. A. (1999). "Recognizing molecules with drug-like properties". In: *Current Opinion in Chemical Biology* 3.4, pp. 384–387. ISSN: 1367-5931. doi:10.1016/S1367-5931(99)80058-1.

Walters, W. P., Stahl, M. T., and Murcko, M. A. (1998). "Virtual screening—an overview". In: *Drug Discovery Today* 3.4, pp. 160–178. ISSN: 1359-6446. doi:10.1016/S1359-6446(97)01163-X.

Wang, C. and Zhang, Y. (2017). "Improving scoring-docking-screening powers of protein–ligand scoring functions using random forest". In: *Journal of Computational Chemistry* 38.3, pp. 169–177. ISSN: 1096-987X. doi:10.1002/jcc.24667.

Wang, R., Lu, Y., and Wang, S. (2003). "Comparative Evaluation of 11 Scoring Functions for Molecular Docking". In: *Journal of Medicinal Chemistry* 46.12, pp. 2287–2303. ISSN: 0022-2623. doi:10.1021/jm0203783.

Wang, S., Wang, L., Li, F., and Bai, F. (2023). "DeepSA: a deep-learning driven predictor of compound synthesis accessibility". In: *Journal of Cheminformatics* 15.1, p. 103. ISSN: 1758-2946. doi:10.1186/s13321-023-00771-3.

Wang, S., Burgess, M., Major, C., English, A., Sweeney, M., and Hartmann, A. (2020). "Identifying fibroblast growth factor receptor genetic alterations using RNA-based assays in patients with metastatic or locally advanced, surgically unresectable, urothelial carcinoma who may benefit from erdafitinib treatment". In: *The Journal of Pathology: Clinical Research* 6.3, pp. 207–214. ISSN: 2056-4538. doi:10.1002/cjp2.163.

Wohlwend, J. et al. (2024). *Boltz-1 Democratizing Biomolecular Interaction Modeling*. doi:10.1101/2024.11.19.624167.

Wong, C. H., Siah, K. W., and Lo, A. W. (2019). "Estimation of clinical trial success rates and related parameters". In: *Biostatistics* 20.2, pp. 273–286. ISSN: 1465-4644. doi:10.1093/biostatistics/kxx069.

Wouters, O. J., McKee, M., and Luyten, J. (2020). "Estimated Research and Development Investment Needed to Bring a New Medicine to Market, 2009-2018". In: *JAMA* 323.9, pp. 844–853. ISSN: 0098-7484. doi:10.1001/jama.2020.1166.

Wu, Z. et al. (2018). "MoleculeNet: a benchmark for molecular machine learning". In: *Chemical Science* 9.2, pp. 513–530. ISSN: 2041-6520, 2041-6539. doi:10.1039/C7SC02664A.

XChem @ Diamond (2024).

Yang, J., Shen, C., and Huang, N. (2020). "Predicting or Pretending: Artificial Intelligence for Protein-Ligand Interactions Lack of Sufficiently Large and Unbiased Datasets". In: *Frontiers in Pharmacology* 11. ISSN: 1663-9812. doi:10.3389/fphar.2020.00069.

Yang, Y., Ouyang, S., Dang, M., Zheng, M., Li, L., and Zhou, H. (2021). "Knowledge Guided Geometric Editing for Unsupervised Drug Design". In.

Yang, Y., Zheng, S., Su, S., Zhao, C., Xu, J., and Chen, H. (2020). "SyntaLinker: automatic fragment linking with deep conditional transformer neural networks". In: *Chemical Science* 11.31, pp. 8312–8322. ISSN: 2041-6539. doi:10.1039/D0SC03126G.

Yang, Z. et al. (2024). *Does Negative Sampling Matter? A Review with Insights into its Theory and Applications*. doi:10.48550/arXiv.2402.17238.

Yosinski, J., Clune, J., Nguyen, A., Fuchs, T., and Lipson, H. (2015). *Understanding Neural Networks Through Deep Visualization*. doi:10.48550/arXiv.1506.06579.

Zartler, E. R. and Shapiro, M. J., eds. (2008). *Fragment-Based Drug Discovery: A Practical Approach*. 1st ed. Wiley. ISBN: 978-0-470-05813-8 978-0-470-72155-1. doi:10.1002/9780470721551.

Zhang, C., Condon, A., and Dao Duc, K. (2025). "A comprehensive survey and benchmark of deep learning-based methods for atomic model building from cryo-electron microscopy density maps". In: *Briefings in Bioinformatics* 26.4, bbaf322. ISSN: 1477-4054. doi:10.1093/bib/bbaf322.

Zhang, J., Liu, Z., Wang, Y., and Li, Y. (2024). *SubGDiff: A Subgraph Diffusion Model to Improve Molecular Representation Learning*. doi:10.48550/arXiv.2405.05665.

Zhang, K., Wu, H., Hoppe, N., Manglik, A., and Cheng, Y. (2022). "Fusion protein strategies for cryo-EM study of G protein-coupled receptors". In: *Nature Communications* 13.1, p. 4366. ISSN: 2041-1723. doi:10.1038/s41467-022-32125-2.

Zhang, X. et al. (2020). "Differential GLP-1R Binding and Activation by Peptide and Non-peptide Agonists". In: *Molecular Cell* 80.3, 485–500.e7. ISSN: 1097-2765. doi:10.1016/j.molcel.2020.09.020.

Zheng, L., Fan, J., and Mu, Y. (2019). "OnionNet: a Multiple-Layer Intermolecular-Contact-Based Convolutional Neural Network for Protein–Ligand Binding Affinity Prediction". In: *ACS Omega* 4.14, pp. 15956–15965. doi:10.1021/acsomega.9b01997.

Zhou, A. et al. (2023). *Permutation Equivariant Neural Functionals*. doi:10.48550/arXiv.2302.14040.

Zhou, J. et al. (2020). "Graph neural networks: A review of methods and applications". In: *AI Open* 1, pp. 57–81. ISSN: 2666-6510. doi:10.1016/j.aiopen.2021.01.001.

Zhu, H., Zhou, R., Cao, D., Tang, J., and Li, M. (2023). "A pharmacophore-guided deep learning approach for bioactive molecular generation". In: *Nature Communications* 14.1, p. 6234. ISSN: 2041-1723. doi:10.1038/s41467-023-41454-9.

Zhu, T. et al. (2013). "Hit Identification and Optimization in Virtual Screening: Practical Recommendations Based on a Critical Literature Analysis". In: *Journal of Medicinal Chemistry* 56.17, pp. 6560–6572. ISSN: 0022-2623. doi:10.1021/jm301916b.

Ziv, Y., Marsden, B., and Deane, C. M. (2024). *MolSnapper: Conditioning Diffusion for Structure Based Drug Design*. doi:10.1101/2024.03.28.586278.

Zou, Y. et al. (2025). "A structure-based framework for selective inhibitor design and optimization". In: *Communications Biology* 8.1, p. 422. ISSN: 2399-3642. doi:10.1038/s42003-025-07840-3.