Designing more efficient compound libraries for crystallographic fragment screening



Anna Carbery
Lady Margaret Hall
University of Oxford

A thesis submitted for the degree of
Doctor of Philosophy

Michaelmas 2023

Acknowledgements

First, I'd like to thank Charlotte, who has now seen me through two degrees and a pandemic. Her consistent guidance, support and patience are hugely appreciated. I'm also immensely grateful to Rachael, who has not only been an invaluable sounding board and a source of a great depth of knowledge, but whose confidence in me has helped me through the highs and lows of this DPhil. Thanks also go to Frank, who ensured my research would always be of use to the wider community.

I would like to thank all OPIGlets past and present for their constant energy and enthusiasm for all things science and otherwise. The long lunches and gossip sessions will be sorely missed, and I hope to make any future workplace as OPIG-like as possible. Special shoutouts go to Lucy, Kate, Gemma, Alissa, Maranga, Olly, Guy and Ísak, whose camaraderie has made the final year of this DPhil the best.

Outside of OPIG, I'm grateful to have close friends who have helped me throughout my eight years in Oxford. I wouldn't have made it to the start of this degree without Jess, Helen, Immie and Jade, and the DPhil years have been made all the better by getting out of Oxford to visit them all. Extra special thanks go to Alice and Agnes for being amazing flatmates and friends who made Oxford really feel like home.

Coaching for LMHBC M1 has been a highlight of my time in Oxford, and I'd like to thank all the generations of crews that I coached for their enthusiastic attempts to put my unintelligible rowing analogies into practice. In particular, thanks go to Harry, Owen, Ben and Will, who made me feel at home from the very beginning.

I am hugely grateful to my family for all their love and support. To John, who was always there to pick up the pieces. To my sister Helen, whose many daily phone calls meant that she was always up to speed. And to my parents, Tony and Maria, who inspired my love of science and empowered me to push myself.

Abstract

Fragment-based drug design aims to explore the binding sites of protein targets and provide information towards bespoke, specific, high-affinity drugs. The small size of fragments facilitates an efficient search of chemical space; only a few efficient interactions need to be made with the target for binding to occur, thus improving the hit rate and reducing the number of fragments that need to be screened in comparison to traditional screening of larger molecules. This approach aims to facilitate a much more efficient structure-based drug design pipeline, decreasing the cost and time taken per drug.

Because of the small size of fragments, in established fragment screening campaigns, even with a good hit rate, the binding site is often not fully characterised. This reduces the number of diverse lead compounds that can be progressed from the initial fragment hits, and makes it more expensive and difficult to develop a viable drug. We hypothesise that this occurs because fragments within screening libraries are traditionally selected for diverse coverage of a broad chemical space, and do not consider functional information such as coverage of potential protein-ligand interactions. This research aims to use retrospective analysis of fragment screening data to develop methods for selecting fragment libraries that will maximise the information obtained from a target's binding site.

We first describe a survey of data availability and quality for XChem fragment screens. The curation procedure for fragment screening data presents many challenges, such as storage of raw crystallographic data and variability within the analysis pipelines. We found that the annotations applied to historic fragment screening data were unreliable due to the complexities of building and validating ligand models. Additionally, locating complete datasets presented major difficulties. To address these issues, we propose the storage of XChem data in a cloud-based object store, coupled with automated backup and curation of crystallographic electron density maps. This facilitates future automated ligand modelling and validation for the generation of large sets of homogenenous fragment screening data.

Next, we use a set of positive (bound structures) and negative (structures for soaked fragments that didn't show clear evidence for binding) data points extracted from the XChem fragment screening data to explore whether the library used is able to make diverse interactions with new protein targets, as intended. Based on these results, we

propose a method for selecting a set of functionally diverse fragments from a larger fragment library. This method calculates the interactions each fragment makes with historic protein targets and ranks the fragment library in order of the number of novel interactions that fragments are able to make. We show that the top-ranked fragments recover information more efficiently from unseen targets compared with traditionally selected fragments.

Finally, we consider what tools are required for the design of target-specific fragment libraries for novel targets. As we are unlikely to have experimental knowledge of a novel target's binding site, we describe a new method for the prediction of ligand binding sites on a protein's surface. This uses a combination of machine learning models and point cloud clustering. For the selection of target-specific fragments, we then explore possible strategies for pocket similarity. We assess the efficacy of a fragment-based binding site similarity tool, and examine the use of learnt representations of proteins for interaction prediction and residue similarity.

In this thesis we examine strategies for the development of more efficient fragment libraries. We use historic fragment screening data to select target-agnostic libraries, and apply recent machine learning methods to explore the potential for target-specific libraries.

Contents

1 Introduction **1**

1.1 Motivations 1

1.2 Fragment based drug design 3

1.2.1 High-throughput Fragment Screening 4

1.2.2 Fragment Screening Libraries 6

1.2.3 Limitations of FBDD 8

1.3 Computer-aided drug design 8

1.3.1 Representations of small molecules 9

1.3.2 Protein-ligand interactions 12

1.3.3 Molecular Dynamics 15

1.3.4 Machine learning 17

1.3.4.1 Supervised learning 17

1.3.4.2 Unsupervised learning 22

1.3.4.3 Neural networks 22

1.3.5 Application of ML to FBDD 24

1.3.5.1 Virtual screening 24

1.3.5.2 Protein structure prediction 25

1.3.5.3 Proteins and transformers 26

1.3.5.4 Data requirements for machine learning on protein structures 26

1.4 Thesis outline 28

2 Analysis of available fragment screening data **31**

2.1 Preface 31

2.2 Introduction 32

2.2.1 Crystallographic fragment screening at XChem 33

2.2.2 Hit identification and ligand modelling in fragment screens . . 35

2.2.3 The Pan-Dataset Density Analysis method 36

2.2.4 Validation of ligand models . . . 37
2.2.5 Availability of fragment screening data . . . 40
2.3 Methods . . . 41
2.3.1 Evaluation of annotations of historic XChem data . . . 41
2.3.2 Data searching and curation . . . 43
2.4 Results . . . 43
2.4.1 Analysis of quality of annotations for XChem fragment screens 43
2.4.2 Structures uploaded to Fragalysis undergo additional validation 46
2.4.3 Curation of electron density maps for XChem crystals . . . 47
2.5 Conclusions . . . 49

3 Fragment libraries designed to be functionally diverse recover information more efficiently from protein targets compared with traditional libraries 52
3.1 Preface . . . 53
3.2 Introduction . . . 53
3.2.1 Design of fragment libraries . . . 54
3.2.2 Definition of desired chemical space . . . 54
3.2.3 Sampling of desired chemical space . . . 55
3.2.4 Influence of HTS library design on fragment library design . . 57
3.2.5 Pharmacophore features . . . 58
3.2.6 Functional activity of fragment libraries . . . 58
3.2.7 Chapter outline . . . 59
3.3 Methods . . . 59
3.3.1 The Diamond-SGC iNext Poised fragment library . . . 59
3.3.2 Comparison of structures in fragment libraries . . . 60
3.3.3 The XChem Dataset . . . 62
3.3.4 Selection of functionally diverse fragments . . . 62
3.3.5 Testing fragment ranking protocol on unseen targets . . . 64
3.4 Results . . . 64
3.4.1 Comparison of fragment libraries . . . 64
3.4.2 Fragment ranking . . . 67
3.4.3 Structurally diverse fragments can make overlapping interactions 68
3.4.4 Ranking of fragments reveals redundancy in interactions . . . 70
3.4.5 Functionally diverse compounds exhibit different chemical properties to non-binding fragments . . . 72

3.4.6 Promiscuous binders are not necessarily the most informative fragments . . . 74
3.4.7 Functionally diverse fragments recover information more efficiently from unseen targets . . . 75
3.4.8 Targets contribute differently to prediction of important fragments for unseen targets . . . 76
3.5 Conclusions . . . 78

4 Learnt representations of proteins can be used for accurate prediction of small molecule binding sites on experimentally determined and predicted protein structures 81
4.1 Preface . . . 82
4.2 Introduction . . . 82
4.2.1 Strategies for protein binding site prediction . . . 83
4.2.2 Machine learning in protein binding site prediction . . . 84
4.2.3 Binding site prediction on predicted structures . . . 86
4.2.4 Binding sites of predicted protein structures . . . 87
4.2.5 Accuracy of predicted protein structures . . . 88
4.2.6 Protein dynamics and binding sites . . . 89
4.2.7 Summary . . . 90
4.3 Methods . . . 91
4.3.1 Datasets . . . 91
4.3.2 Model training . . . 93
4.3.3 Binding site prediction with IF-SitePred . . . 94
4.3.4 Evaluation Criteria . . . 95
4.3.5 Comparison to existing methods . . . 96
4.3.6 Ligand similarity to training sets . . . 97
4.3.7 Molecular dynamics simulations for testing predictions on low-quality structures . . . 97
4.3.7.1 Evaluation . . . 98
4.3.8 Combining predictions by using multiple models on multiple structures . . . 98
4.4 Results . . . 99
4.4.1 IF-SitePred binding site prediction is competitive with state-of-the-art methods . . . 99

4.4.2 Accurate binding site prediction is not dependent on ligand similarity to the training set . . . 103
4.4.3 IF-SitePred outperforms other methods in top-3 binding site prediction on MD structures . . . 104
4.4.4 Combining predictions of multiple models on multiple structures improves predictive power . . . 106
4.5 Conclusions . . . 111

5 Binding site similarity strategies for target-specific fragment library design 113
5.1 Preface . . . 113
5.2 Introduction . . . 114
5.2.1 Overview of binding site similarity approaches . . . 115
5.2.1.1 Alignment-based approaches . . . 115
5.2.1.2 Alignment-free approaches . . . 116
5.2.1.3 Machine learning approaches . . . 117
5.2.2 Identical ligands bound to different proteins . . . 118
5.2.3 Benchmarking of binding site comparison methods . . . 119
5.2.4 Chapter summary . . . 120
5.3 Methods . . . 122
5.3.1 ProCare evaluation . . . 122
5.3.1.1 Dataset . . . 122
5.3.1.2 ProCare Implementation . . . 122
5.3.2 Testing ProCare's alignment and scoring performance . . . 124
5.3.2.1 Pocket similarity of cavities which bind identical ligands 124
5.3.2.2 Alignment of smaller sub-cavities into larger cavities 125
5.3.2.3 ProCare scores of positive and negative sub-cavities aligned into larger cavities . . . 126
5.3.3 ESM embeddings . . . 127
5.3.4 Distance-based pocket similarity comparison . . . 127
5.3.4.1 The Barelier Dataset . . . 127
5.3.4.2 Comparison method . . . 129
5.3.5 Prediction of protein-ligand interactions . . . 129
5.3.5.1 Test and training data . . . 129
5.3.5.2 Protein-ligand interaction calculation . . . 130
5.3.5.3 Model training . . . 130

5.3.5.4 Evaluation . . . 131
5.4 Results . . . 132
5.4.1 ProCare evaluation . . . 132
5.4.1.1 Pocket similarity of cavities which bind identical ligands 132
5.4.1.2 Alignment of smaller sub-cavities into larger cavities 133
5.4.1.3 ProCare scores of positive and negative sub-cavities aligned into larger cavities . . . 136
5.4.2 Use of ESM2 and ESM-IF1 embeddings for pocket similarity . 139
5.4.2.1 Identical ligands do not bind residues with similar embeddings . . . 139
5.4.2.2 ESM embeddings outperform one-hot embeddings in prediction of interaction types . . . 140
5.4.2.3 High-affinity complexes contain slightly stronger signal than low-affinity complexes . . . 142
5.5 Conclusions . . . 143

6 Conclusions and Future Work **147**
6.1 Data availability . . . 147
6.2 Structural diversity and functional diversity . . . 148
6.3 Target-specific library design . . . 149
6.3.1 Binding site prediction on predicted protein structures . . . 149
6.3.2 Pocket similarity for fragment selection . . . 150
6.4 Closing remarks . . . 151

A Fragment libraries designed to be functionally diverse recover information more efficiently from protein targets compared with traditional libraries **152**

B Learnt representations of proteins can be used for accurate prediction of small molecule binding sites on experimentally determined and predicted protein structures **170**

C Binding site similarity strategies for target-specific fragment library design **174**

Bibliography **185**

List of Figures

1.1 The stages in the development of a successful drug to market. 2
1.2 Fragment screening results in a higher hit rate and more specific interactions than screening of traditional drug-sized molecules, but fragment-to-lead development is more complex than hit-to-lead. 4
1.3 Fragments that bind to a protein can be grown, linked or merged to provide follow-up compounds for further screens. 5
1.4 A hypothetical representation of how substructures of a molecule can be encoded to a 10-bit fingerprint. 10
1.5 The substructures of the molecule shown represented in ECFPs of diameters 0, 2 and 4. 11
1.6 The distance and angle requirements of a hydrogen bond. 14
1.7 The training and assembly of a gradient boosting machine. 19

2.1 The pipeline of experimental compound screening using X-ray crystallography. 33
2.2 The hit-finding process used at XChem. 34
2.3 The data analysis pipeline of fragment screens. 37
2.4 RSCC values of the model's expected density and the observed density for PanDDA events for 8,633 crystals do not show a clear correlation between RSCC and annotation as 'bound' or 'not bound'. 45
2.5 An example of mislabelling of modelled structures at the binding sites of two proteins. 45
2.6 The backup and curation of XChem data. 48

3.1 Pharmacophore feature representation in commercially-available fragment libraries. 65
3.2 Representation of families of pharmacophore features in five fragment libraries. 66

3.3 Examples of USRCAT scores obtained from comparing a reference molecule with four other molecules. 67
3.4 The DSiP, 3D diverse and F2X-Universal libraries are slightly more diverse than the High Fidelity and Essential libraries. 68
3.5 Diverse fragments make identical interactions. 69
3.6 Structures of the three fragments shown in Figure 3.5(b) 70
3.7 Ranked fragment libraries recover information at smaller library sizes than other methods. 71
3.8 Fragments that have never bound to a target are more likely to have very small or very large MW and heavy atom counts. 73
3.9 The number of targets bound by fragments in the DSiP library, grouped by their position when ranked. 74
3.10 Ranked fragment libraries show superior information recovery for unseen targets at every library size. 75
3.11 Top-ranked fragments show superior information recovery compared with random fragments and structurally diverse fragments. 76
3.12 The impact of each target on every other target's information recovery at library size of 100 fragments. 77

4.1 ESM-IF1 embeddings were generated for each residue of the protein, and the model labelled each of these as binding or non-binding. . . . 91
4.2 The four pathways implemented to compare the F1 scores achieved when predicting ligand-binding residues on 21 MD structures. 100
4.3 AF2-predicted structures in the HAP set are extremely accurate compared with most free modelling AF2 predictions on novel proteins. . . 105
4.4 Binding site prediction on a target at increasing RMSD. 106
4.5 IF-SitePred outperforms P2Rank and DeepPocket when considering top-3 success on low-accuracy structures. 107
4.6 Using multiple predictive models on multiple protein structures improves IF-SitePred's predictive power. 108
4.7 Predictive power for binding residue annotation improves when predictions for 10 structures were combined for all three methods. 109
4.8 IF-SitePred predicted fewer residues incorrectly as ligand-binding. . . 110

5.1 Pairings of similar and dissimilar pairs of cavities used to develop the ProCare scoring function and those constructed from structures taken from Fragalysis 125

5.2 The Barelier dataset is made up of three classes of pairs of unrelated proteins bound to identical ligands. 128
5.3 Pockets that bind identical ligands do not get higher ProCare scores than those which bind different ligands. 134
5.4 Ligand RMSD values when realigning fragment-based cavities into the predicted cavity of the target are lower when a larger cavity is used. . 135
5.5 The proportion of ligand atoms within 2Å of any predicted cavity point for sub-cavities that were realigned within 2Å of their original position or not within 2Å of their original position slightly reduces realignment success. 136
5.6 ProCare scores for 15 positive, 199 negative, and 297 self pairings of sub-cavities aligned into larger predicted cavities show that positive pairs do not score significantly higher than negative pairs. 137
5.7 As recall increases, precision falls quickly, indicating a poor ability to correctly classify cavity pairings. 138
5.8 Normalised RMSD values show that over 50% of alignments are within 2Å of the optimal alignment for structures. 139
5.9 Residues that bind identical ligands do not exhibit higher cosine similarity or lower euclidean distance of ESM-IF1 embeddings from those which bind different ligands. 140
5.10 Residues that bind identical ligands do not exhibit higher cosine similarity or lower euclidean distance of ESM2 embeddings from those which bind different ligands. 141
5.11 The greatest F1 scores in predicting interaction type can be achieved by using ESM2 and combined embeddings to represent protein residues. 142
5.12 All models trained on ESM embeddings benefit from higher-affinity training data. 144

A.1 Fragments in the DSP library are poised for 11 different categories of reaction. 154
A.2 A comparison of the fragment ranking method with various types of structurally diverse fragment selection methods. 156
A.3 Interactions calculated across all protein-fragment structures used in this analysis. 156
A.4 Interactions calculated for each target show that most targets make significantly more hydrophobic interactions than hydrogen bonds. . . 157

A.5 The numbers of polar interactions made by each fragment in a protein-fragment complex. 157
A.6 The factor of information improvement when using 100 functionally diverse fragments compared with random and structurally diverse libraries. 158

B.1 AlphaFold prediction confidence does not vary between targets with top 1, top 2, top 3 success or failure in binding site prediction by IF-SitePred. 172

List of Tables

1.1 Common types of intermolecular interactions present in protein-ligand complexes. 13

2.1 Five metrics used by PanDDA to assess and validate model quality. . 39
2.2 RSCC-based classification of 8,633 modelled structures compared with their 'bound' or 'not bound' annotations provided by the user shows that many reliable models are missed, and many poor models are labelled as a hit. 44

3.1 The fragment libraries used as comparators. 60
3.2 The pharmacophore feature descriptors used in the comparison of fragment libraries. 61
3.3 Definitions for the various groups of fragments used in this analysis. . 70

4.1 Commonly-used evaluation metrics for protein binding site prediction. 96
4.2 IF-SitePred, P2Rank and DeepPocket are competitive across PDB and AF2 structures, whereas FPocket experiences a significant loss of performance on AF2 structures. 101
4.3 DeepPocket performs worse than average on sites that bind ligands significantly different from the training set. 103

5.1 Interaction types as calculated by ODDT, with their label for training and testing. 130
5.2 Numbers of complexes and residues for each label in high-affinity, low-affinity and full datasets. 131
5.3 The sets of structures used in our analyses of ProCare. 132
5.4 Mean and median values of the best ProCare scores obtained from each set show a small improvement in the scores of positive pairs compared with negative pairs where all cavity sizes are used. 137

A.1 Many of the fragment libraries currently in use. 153

A.2 The eight types of protein-ligand interaction detected by ODDT's InteractionFingerprint module Wójcikowski et al. (2015). 154
A.3 Descriptions of each target used in this study. 155

B.1 Success rate is largely stable to small perturbations in distance thresholds for point cloud generation and clustering. 171
B.2 Error estimations as calculated using bootstrapping of the predictions for PDB structures in the HAP set show that IF-SitePred and P2Rank have more consistent prediction success than FPocket and DeepPocket. The 95% confidence interval of the standard deviation for each success rate was calculated using SciPy's bootstrapping function. In general, error is estimated to be smaller where prediction accuracy is higher. . 173
B.3 Error estimations as calculated using bootstrapping of the predictions for AF2 structures in the HAP set show that IF-SitePred has more consistent prediction success than P2Rank, FPocket and DeepPocket. Confidence intervals were calculated as in Table B.2. 173
B.4 Error estimations as calculated using bootstrapping of the predictions for PDB structures in the HAP-small set show that IF-SitePred and P2rank have more consistent prediction success than FPocket and DeepPocket. Confidence intervals were calculated as in Table B.2. . . 173
B.5 Error estimations as calculated using bootstrapping of the predictions for AF2 structures in the HAP-small set show that IF-SitePred has more consistent prediction success than P2Rank, FPocket and DeepPocket. Confidence intervals were calculated as in Table B.2. 173

List of Abbreviations

AF2 AlphaFold 2

API application programming interface

AUC area under the curve

CADD computer-aided drug design

CASP The Critical Assessment of Structure Prediction

CNN convolutional neural network

Da Daltons

DBSCAN density-based spatial clustering of applications with noise

DCA distance from centre to atom

DCC distance from centre to centre

DLS Diamond Light Source

DSiP Diamond-SGC-iNext Poised

DVO discretised volume overlap

ECFP extended-connectivity fingerprint

ESM evolutionary scale modelling

FBDD fragment based drug design

HTS high-throughput screening

IFP interaction fingerprint

IoU intersection over union

MD molecular dynamics

ML machine learning

NMR nuclear magnetic resonance

ODDT the Open Drug Discovery Toolkit

PanDDA Pan-Dataset Density Analysis

PCC Pearson's correlation coefficient

PDB the Protein Data Bank

PPIs protein-protein interactions

RESOL resolution

RMSD root mean square deviation

RMSZ root-mean square value of the Z-scores

RSCC real-space correlation coefficient

RSZD real-space difference density Z-score

SBDD structure based drug design

SBVS structure based virtual screening

SMILES Simplified Molecular Input Line Entry System

TB terrabytes

USRCAT Ultrafast Shape Recognition with Credo Atom Types

VHELIBS Validation HElper for LIgands and Binding Sites

Chapter 1

Introduction

1.1 Motivations

Historically, drugs were discovered serendipitously, or through identification of the active ingredient within a traditional remedy. Since the 1980s, this paradigm shifted to the targeted and rational design of molecules that can treat specific illnesses (Appelt et al., 1991; Wlodawer and Vondrasek, 1998; Anderson, 2003). This process relies on our understanding of the biomolecules involved in the mechanism of disease, which are usually proteins.

Proteins are large biomolecules consisting of one or more folded chains of amino acid residues. They carry out a wide range of functions within organisms, including the catalysis of biochemical reactions, provision of structure to cells and tissues, replication of DNA, molecular transport and cellular signalling.

The identification of the proteins implicated in a disease of interest is the first stage in the target-based drug design pipeline. This may be a protein exhibiting compromised function in the patient, or a protein involved in the pathogenicity of a virus or bacteria. Once identified, the protein can be isolated and characterised experimentally. The search then begins for a molecule that has an appropriate shape

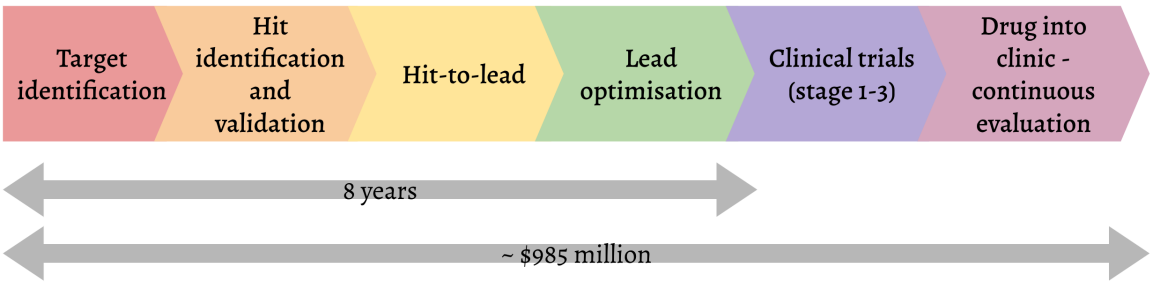


Figure 1.1: The stages in the development of a successful drug to market. This process is slow and expensive, and most potential drugs never make it to market (Wouters et al., 2020). A druggable protein target involved in the disease must first be identified, before molecules which bind to this protein ('hits') can be found and experimentally validated. These candidate molecules are altered to improve binding affinity to the protein, followed by optimisation of the chemical properties of the molecules. At this point, the most promising candidates may be taken forward to clinical trials. The attrition rate is high at every stage of this pipeline, and even drugs that make it to market are continuously evaluated.

and charge profile for binding to the protein, with the intended effect of altering the protein's behaviour such that the disease is treated.

Small molecule drugs are organic compounds that weigh less than 1000 Daltons (Da). Within target-based drug design approaches, they are designed to bind to the protein of interest with high affinity. However, for every additional heavy atom in a molecule, the number of possible compounds increases around ten-fold, and drug-like molecules can contain up to thirty heavy atoms. This results in a drug-like chemical space of up to 10^{60} molecules, so the identification of a molecule which potently binds to the target is expensive and time-consuming (Figure 1.1). Even once a suitable molecule has been identified *in vitro* (in isolated biological material), the compound must be optimised for specificity to the protein of interest, for the required activity *in vivo* (in animals), and for appropriate pharmacological properties (van de Waterbeemd et al., 2001; Leeson et al., 2021). Even drugs selected for clinical trials have a very high attrition rate: only one in 10 candidates are successful at this stage (Sun et al., 2022), and due to the costs of clinical trials, financial implications rocket. Overall, the cost of research and development per product is calculated to

be around $985 million, and on average, 8.3 years of research and development are required before a candidate makes it to clinical trials (Wouters et al., 2020).

Both cost and timescale require improvement: it is important for affordable drugs to be widely available quickly, and this is exacerbated for emergent endemic and pandemic infectious diseases, such as COVID-19. While the clinical stages are the main drivers of the cost of drug discovery (Simoens and Huys, 2021), improving the cost and efficiency of early-stage drug discovery can greatly increase the chances of a successful drug being developed. Fragment based drug design (FBDD) and computational (*in silico*) predictions are two innovations that aim to tackle these issues.

1.2 Fragment based drug design

The design of novel drug compounds currently relies on experimental screening of drug-like molecules. However, the vastness of chemical space makes it nearly impossible to find viable drugs for particular targets without any prior knowledge of what area of chemical space should be sampled.

One solution to sampling more of chemical space without increasing the size of the experiment is to reduce the size of the molecules. Fragments are molecules that are around a third of the size of drug-like molecules (<250 Da), and their use in screening experiments has become a popular starting point in early-stage drug discovery (Kumar et al., 2012). These smaller fragments are each able to form only a small number of interactions with the protein, however this may be all that is required for efficient binding to the target (Scott et al., 2012; Hopkins et al., 2004). In a study examining the role of molecular size in ligand efficiency (the binding free energy for a ligand divided by its molecular size), Reynolds et al. (2007) found that smaller molecules exhibit higher ligand efficiency than larger, more drug-sized molecules, showing that

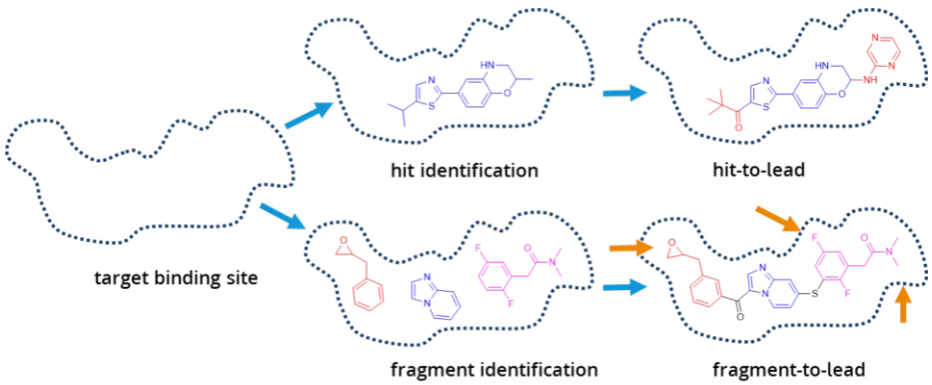


Figure 1.2: Fragment screening (lower) results in a higher hit rate and a more specific interactions (indicated by orange arrows) than screening of traditional drug-sized molecules (upper), but fragment-to-lead development is more complex than hit-to-lead. Adapted from BioSolveIT (2023).

despite the lower binding affinity in protein-fragment complexes, strong interactions are made (Hopkins et al., 2014).

Once the fragments that bind a particular target have been identified, they can be used as building blocks for larger lead molecules which bind more potently and more specifically to the target of interest (van Montfort et al., 2017a). Lead molecules can be developed through linking or merging of multiple fragments that have been found to bind to different parts of the same binding site, or through 'growing' of one or more molecules. These processes are summarised in Figures 1.2 and 1.3. FBDD of this type has become well-established as a powerful approach to early-stage drug discovery, and has a track record of success for difficult targets where other methods have failed (Erlanson et al., 2016; St Denis et al., 2020). At the time of writing, seven approved drugs have originated from FBDD, with the number set to increase further (Rees et al., 2022; Erlanson, 2023).

1.2.1 High-throughput Fragment Screening

Understanding the functional implications of a fragment binding to the protein is crucial in FBDD. This guides the selection of fragments selected for further develop-

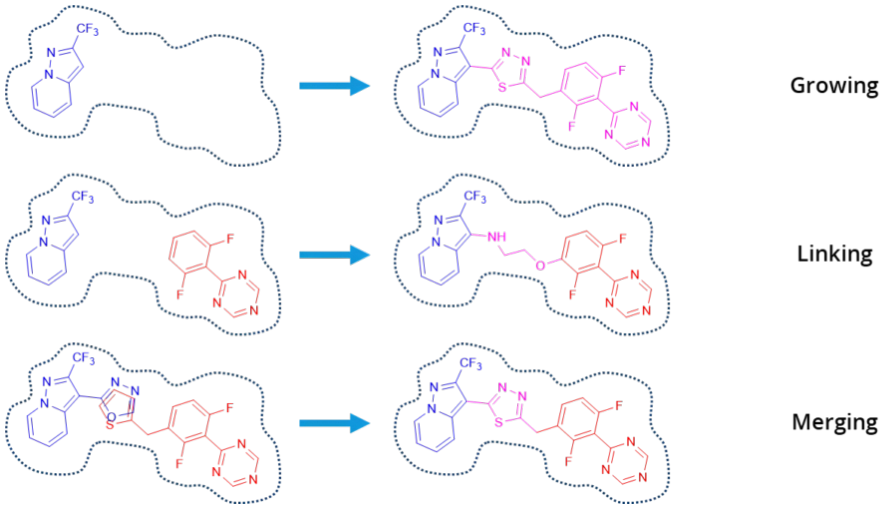


Figure 1.3: Fragments that bind to a protein can be grown, linked or merged to provide follow-up compounds for further screens. This is dependent on the fragments identified as binders in the original screen and synthetic tractability of follow-up compounds. Adapted from BioSolveIT (2023).

ment, as not all binding fragments will be bound to an area of the protein important for function. Furthermore, the specific interactions made between the protein and the fragment are important, as larger lead compounds are designed to retain these interactions, under the assumption of a conserved binding mode (Murray et al., 2012; Kozakova et al., 2015).

X-ray crystallography is the dominant method for determining the structures of protein-ligand complexes (Kirkwood et al., 2015). The first step consists of crystallisation of the protein from a solution into a crystal lattice form which is able to diffract an X-ray beam. This allows the detection of a diffraction pattern which can be processed to determine the density of electrons throughout the crystal. Where small molecules are soaked into the crystal, their binding to the protein can be detected via their contribution to the X-ray diffraction.

Historically, it was difficult to use crystallography for primary screening (Patel et al., 2014; Keserű et al., 2016). Instead, biophysical assays such as surface plasmon resonance (SPR), microscale thermophoresis, and nuclear magnetic resonance (NMR) were used to pre-select fragments that were able to make interactions with a target

protein (Harner et al., 2013; de Souza Neto et al., 2020). Only these fragments were taken forwards to structure determination by X-ray crystallography. However, these methods have lower sensitivity or do not yield important details about the protein-fragment complexes, so many potential interactions may have been missed (Patel et al., 2014; Davies, 2014; van Montfort et al., 2017b).

The efficient screening of fragments has been enabled at XChem at Diamond Light Source (DLS), where the time to screen 2000 compounds (including data collection and processing) has been reduced from over a month to under one week (XChem, 2020b). This makes it possible to screen hundreds of fragments on target proteins using X-ray crystallography with no pre-selection. This both maximises the number of binding events detected and eliminates the need for follow-up structural analyses, as protein-ligand complex structures are already available. This increased number of hits facilitates efficient progression of fragment hits into lead development and optimisation. Many 'users' (academic or industrial groups) that are not crystallographers or experts in structural biology have used the facility at DLS as a starting point for drug design campaigns (e.g. McIntyre et al. (2017); Harding et al. (2017); Foley et al. (2017)). The experimental pipeline used at XChem is detailed by Collins et al. (2018).

1.2.2 Fragment Screening Libraries

To maximise the useful information that can be extracted from each fragment screening experiment, it is important to design the fragment library carefully. Most libraries consist of molecules that adhere to the 'rule of three' (Congreve et al., 2003), which means that the compounds have a molecular weight below 300Da, fewer than three hydrogen-bond donors and acceptors, fewer than three rotatable bonds and a cLogP of three or below. Heavy atom counts tend to be limited at around 20 (Giordanetto et al., 2019). With these rules in mind, libraries are then designed to be as structurally diverse as possible. However, this is not directly predictive of the interactions

a compound is able to make with another molecule (Guha, 2013). As a prediction of functional activity, pharmacophoric features are often used. These are substructures of a compound that are responsible for particular interactions that the compound is able to make, and so give an idea of the ligand's functional activity. Fragments are usually made up of two or three pharmacophoric features (Konteatis, 2021) and in order to maximise a fragment library's predicted functional activity, the pharmacophoric features represented in the library are chosen to match interactions seen with proteins across the Protein Data Bank (PDB) (Berman et al., 2000). For example, hydrophobic interactions and hydrogen-bonding are the two most common types of protein-ligand interactions seen in the PDB (Ferreira de Freitas and Schapira, 2017), and so aromatic and hydrogen-bond donors and acceptors should be more represented in the library than ionic bond-forming pharmacophoric features.

Within these rules, there is variation between fragment libraries to allow some tailoring to types of interaction. For example, the MiniFrag library contains smaller fragments and is designed to identify the most powerful binding interactions (O'Reilly et al., 2019). Other libraries may specialise in covalent linkages, as these have been able to successfully inhibit difficult targets (Resnick et al., 2019). Libraries with increased 'three dimensionality' of fragments increase the potential of the library to explore new areas of chemical space (Morley et al., 2013), but their overall effectiveness is not agreed upon, as initial studies show that the hit rates of 3D fragments are reduced compared to the flatter compounds (Hall et al., 2014). This is likely due to the more complex 3D molecular shapes being suitable for binding only to pockets with specific complementary shapes.

Further tailoring of the library towards specific targets or classes of targets has resulted in some success on difficult targets such as protein-protein interfaces (Valenti et al., 2019) and membrane proteins (Früh et al., 2010), or targets that require a high level of specificity such as kinases (Akritopoulou-Zanze and Hajduk, 2009). Target-

specific design strategies have generally relied on knowledge of previous inhibitors for comparisons of structure or interaction fingerprint (IFP), descriptor-based activity predictions, or activities from biophysical or biochemical pre-screening (Posy et al., 2011). These all require previous characterisation of the target's ligand-binding potential; a technique that is able to exploit the structure of the unbound target is docking, which has been shown to be useful in prioritisation of fragments for experimental screening (Bian and Xie, 2018).

1.2.3 Limitations of FBDD

While FBDD has proved successful on some targets, success is not guaranteed. Where a site is well-explored by diverse fragments that make a range of interactions, leads can be developed using linking and merging of molecular structures that are known to bind and make interactions (de Souza Neto et al., 2020; Wills et al., 2023). However, in many cases the site of interest lacks adequate exploration, meaning that only one fragment is bound, or that multiple fragments make the same interactions with the binding site. This can hinder the development of potent and diverse sets of leads, as only elaboration of a single fragment can be explored, with no further knowledge of molecular structures that bind to the surrounding area. In such cases, it can be more difficult to successfully progress compounds to the lead optimisation stage.

Other challenges and limitations faced in FBDD include difficulties in producing enough protein, low solubility of fragments, lack of protein crystallisation and untargetable protein families (Rees et al., 2022).

1.3 Computer-aided drug design

Computational (*in silico*) analysis of protein and ligand structures provides a cheaper and faster alternative to *in vitro* and *in vivo* experiments. A range of methods

have been developed to analyse protein structures, ligand structures, and how they interact in protein-ligand complexes. These include virtual screening, protein-ligand interaction analysis, molecular dynamics and protein structure prediction (Yu and MacKerell, 2017). In this section we describe key areas of computer-aided drug design (CADD) which are relevant to the work in this thesis.

1.3.1 Representations of small molecules

Small molecules must be presented in a machine-readable format for computational analysis. This facilitates the comparison of the structures and properties of different molecules, and additionally allows the computational prediction of chemical properties.

String formats A common machine-readable representation is the Simplified Molecular Input Line Entry System (SMILES) (Weininger, 1988), which converts the two- or three-dimensional structure of the chemical compound to a string of characters. While this is a simple approach and is a machine-readable format which retains some human readability, SMILES strings can be ambiguous: a single SMILES string could represent several different possible molecules, while a single molecule could be encoded into SMILES strings in several ways. Additionally, atomic coordinates are not contained in this format, which limits our understanding of the three-dimensional conformations of the molecules. International chemical identifier (InChi) keys (Heller et al., 2015) are a more recent format which alleviate most of the ambiguities of SMILES strings but are less human-readable.

Fingerprints An alternative to string-based molecular representations is the 'fingerprint': the molecule is converted to a (usually fixed-length) bit string which describes the two- or three-dimensional structure of the compound (Figure 1.4).

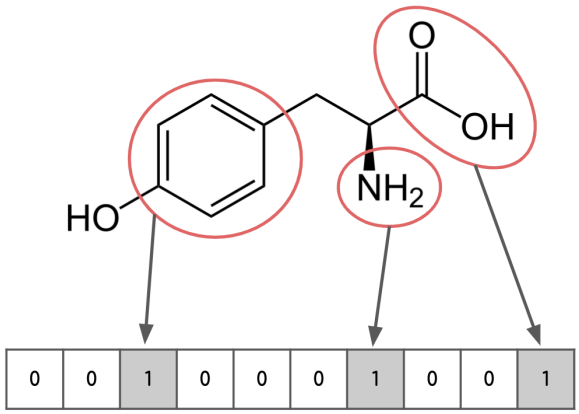


Figure 1.4: A hypothetical representation of how the circled substructures of the molecule shown can be encoded to a 10-bit fingerprint. Adapted from Cereto-Massagué et al. (2015).

Molecular fingerprints generally bypass the limitations faced by SMILES strings as they map the ligand structure deterministically to the binary fingerprint (Tazhigulov et al., 2022), but lack human readability. The primary advantage of fingerprints is that they are easy to compare mathematically, and so the similarities of large numbers of compounds (such as those within a high-throughput screening (HTS) screening library) can be calculated. The most common similarity metric is Tanimoto similarity, which calculates the number of bits which are 'on' in both fingerprints and divides it by the total number of bits which are 'on' across both fingerprints.

However, the similarity between two molecules can vary depending on the type of fingerprint used. Many types of fingerprints have been developed, including path-based, substructure-based and circular fingerprints.

Atom pair fingerprints (Capecchi et al., 2020) are path-based fingerprints, where pairs of atoms along with their shortest path distance are enumerated. Another path-based fingerprint, the topological torsion fingerprint, aims to complement this long-range information by capturing short-range information contained in the torsion angles of a molecule.

Substructure-based fingerprints such as MACCS keys (also known as MDL keys) (Durant et al., 2002) simply count the occurrences of pre-defined substructures present

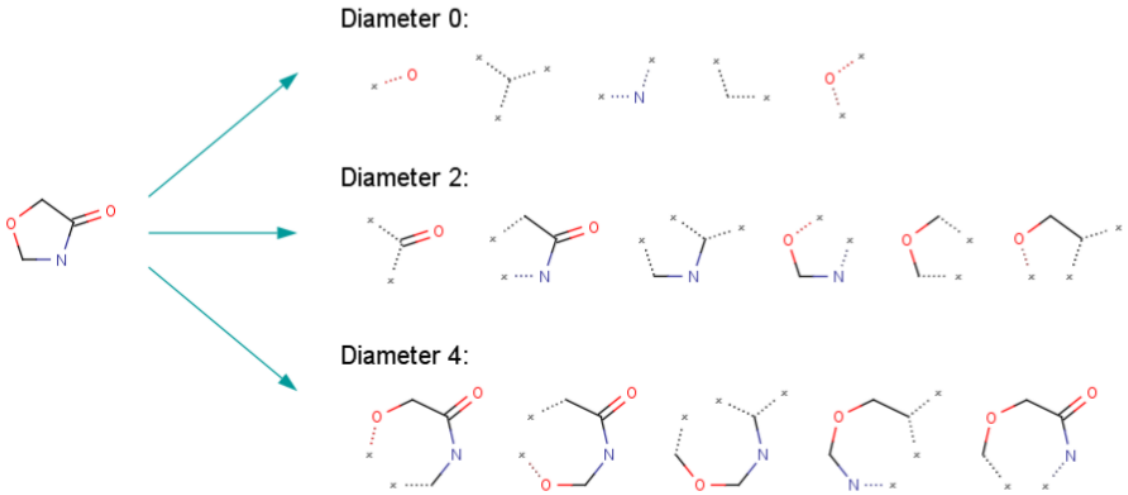


Figure 1.5: The substructures of the molecule shown represented in ECFPs of diameters 0, 2 and 4. Adapted from ChemAxon (2023).

in a molecule. These substructures are 166 pharmacologically relevant functional groups as defined by experts, and while the MACCS fingerprint was developed for substructure searching, it has been reoptimised for use in molecular similarity.

A key example of a circular fingerprint is the extended-connectivity fingerprint (ECFP) (Rogers and Hahn, 2010). This method assigns an integer identifier to each atom, then iteratively updates the identifier to reflect the identifiers of the atom's neighbours. The identifier for each atom at each iteration is added to the fingerprint, and duplicate identifiers are removed. The size of the diameter is predefined: an ECFP2 fingerprint requires 2 iterations and includes the central atom and those that are immediately adjacent, whereas an ECFP6 fingerprint requires 6 iterations (Figure 1.5).

Lists of ECFP identifiers vary in length, but in this format they are still suitable for the measurement of molecular similarity or for the prediction of molecular properties. If the application requires the fingerprints to be fixed-length, a hashing function is applied to 'fold' the identifier lists into a fixed-length bit strings. While this simplifies the comparison of molecules and reduces the required storage space, this has the

potential for clashes, where multiple identifiers map to the same bit, resulting in a loss of information, thus worsening the quality and interpretability of this representation.

Graphs Molecules can also be represented as graphs. Although this is perhaps the most intuitive and human-readable format, graphs present challenges, as they are memory-intensive. However, they are popular representations for use in more complex CADD tasks such as virtual screening and docking (See Section 1.3.4), as they are able to contain most of the information relevant to the chemical structure.

1.3.2 Protein-ligand interactions

Protein-ligand binding is driven by a negative Gibbs' free energy of binding,

$$\Delta G = \Delta H - T\Delta S,$$

where ΔH and ΔS are the changes in enthalpy and entropy respectively, and T is the temperature (Ott and Boerio-Goates, 2000). The entropic contribution is a measure of disorder in the overall system. There is a loss of motion caused by the changes in degrees of freedom of the protein and ligand upon formation of the protein-ligand complex, creating an unfavourable entropic change. Conversely, the binding cavity is partially or completely desolvated upon binding, which leads to a favourable entropic change (Singh and Warshel, 2010). The enthalpic component is dependent on interactions between the protein and the ligand (Bronowska, 2011).

Cofactors can form an additional component in protein-ligand binding: they are non-protein compounds or metal ions which may be required for the protein to have a particular function or to be correctly folded. They may be bound to the protein transiently or permanently. Cofactors can impact on ligand binding in various ways: they may be present within the ligand binding site; they may allosterically cause conformational shifts that make ligand binding more or less favourable; and cofac-

tors with charged components have long-ranging electrostatic effects on surrounding molecules.

Non-covalent protein-ligand interactions, herein referred to as simply protein-ligand interactions, are attractive intermolecular forces between the ligand and protein chemical groups in a complex. As they contribute to the favourable binding of a ligand to a protein, drug design campaigns are guided by previously observed interactions, striving to optimise drug leads by maximising the number of robust interactions.

Class	**Interaction Type**	**Distance requirements**	**Angle requirements**
Hydrophobic	Hydrophobic contact	Carbon to carbon <4Å	-
Aromatic	Pi-stacking face to face	Ring to ring <5Å	parallel
	Pi-stacking edge to face	Ring to ring <5Å	perpendicular
Hydrogen bonding	Hydrogen bond (protein as hydrogen bond donor)	Donor heavy atom <3.9Å from acceptor heavy atom	Shown in Figure 1.6
	Hydrogen bond (protein as hydrogen bond acceptor)	Donor hydrogen <2.5Å from acceptor heavy atom	
Salt Bridges	Salt bridge (protein positively charged)	Acceptor <4Å to hydrogen	Defined by atom's hybridisation
	Salt bridge (protein negatively charged)		
	Salt bridge (ionic bond with metal ion)	Acceptor <4Å to metal	
Halogen	polar interaction of halogen with halogen bond acceptor	<sum of van der Waals of involved atoms	

Table 1.1: Common types of intermolecular interactions present in protein-ligand complexes.

The main types of protein-ligand interactions are summarised in Table 1.1. Many

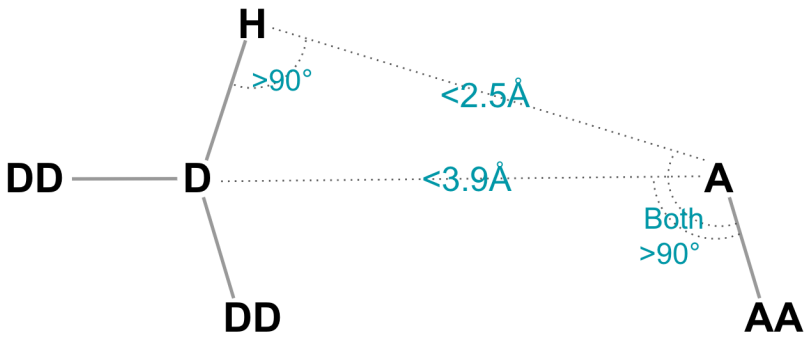


Figure 1.6: The distance and angle requirements of a hydrogen bond as described by McDonald and Thornton (1994). 'A' is the acceptor atom, which has an adjacent atom 'AA'. 'D' is the donor atom, bound to a hydrogen atom and two adjacent atoms 'DD'.

tools have been developed for the calculation and analysis of these intermolecular interactions in protein-ligand complexes (Salentin et al., 2015; Wójcikowski et al., 2015; Bouysset and Fiorucci, 2021). Their rules for calculating interactions are largely as outlined in Table 1.1 and Figure 1.6, however some thresholds and atom typing rules vary. Some tools also include strictness and tolerance options: for example, the Open Drug Discovery Toolkit (ODDT) (Wójcikowski et al., 2015) allows for the calculation of non-strict hydrogen-bonds, where the angle requirements are not applied. Additionally, aromatic and salt bridge interactions have an adjustable angle tolerance, with the default setting being 30 degrees.

A wide range of analyses are facilitated by the calculation of protein-ligand interactions. For example, PDB-wide analyses of interaction types in protein-ligand complexes revealed the overall frequencies of different interactions (Ferreira de Freitas and Schapira, 2017). Comparison of interactions made in different protein-ligand complexes has also been used, via the embedding of the calculated interactions into an interaction fingerprint (IFP), which can then be compared in a similar way to molecular fingerprints (e.g. Weill and Rognan (2010)) (see Section 1.3.1).

FBDD leverages protein-ligand interactions in several different ways. For example, clustering of interactions made in fragment screening experiments highlights the areas

of the binding site that contribute disproportionately to ligand binding - known as hotspots (Hall et al., 2015). By observing interactions made in large databases, hotspots can also be predicted based on target structures by placing virtual chemical probes over the protein's surface (Radoux et al., 2016). Protein-ligand interactions are also used in the scoring of potential binders for proteins of interest, including follow-up compounds subsequent to fragment screens. This is discussed further in Section 1.3.5.

1.3.3 Molecular Dynamics

Proteins are dynamic systems, however only static snapshots of protein structures are experimentally determined. Therefore, traditional fixed-coordinate structures of proteins do not contain all the information relevant to protein function and binding activity. Molecular dynamics (MD) simulations were developed to overcome this limitation and allow the prediction and analysis of physical movements of atoms and molecules. MD simulations of proteins and protein-ligand complexes can greatly expand our knowledge and understanding of these systems (De Vivo et al., 2016; Hollingsworth and Dror, 2018).

Typically, all protein atoms are present in an MD simulation, along with solvent molecules. Depending on the context, ions may be included to balance the charge within the simulation, or biologically-relevant molecules such as ligands or lipids may be present. Interatomic 'bonded' properties (bonds, bond angles and torsional angles) and 'nonbonded' potential energy terms (van der Waals and Coulombic) are summed to give the total potential energy of the system (known as the force field). The derivatives of the potential energy terms provide forces on atoms, which are used to change the atomic coordinates for a given time step (Collier et al., 2020). The resulting series of structures reveal the dynamics of the system, which for a protein can represent anything from small-scale atomic movements to large-scale conformational

changes and folding or unfolding pathways.

These changes can be used to learn more about the system. For example, the protein can be simulated in isolation to observe protein folding (e.g. Karplus and Kuriyan (2005)), or by including the ligands in the system, protein-ligand binding and mechanisms of enzymatic catalysis can also be studied (e.g. Swegat et al. (2003)).

The primary use of MD in FBDD has been in the identification of druggable binding sites as the surface of the protein varies across the dynamic landscape (e.g. Kozlovskii and Popov (2020)). An example of a successful application of MD to binding site prediction is MixMD, which is a mixed-solvent MD simulation that includes organic solvents at high concentrations (Smith and Carlson, 2021; Bissaro et al., 2020). It was shown that these organic molecules clustered at key binding hotspots, and so the MixMD results could be used for binding site identification and druggability assessment.

Despite the wide applicability of MD, a major limitation remains: the simulation of such a system as large as a protein is very slow and so only processes that take place on a short timescale can be properly observed, while also taking up large amounts of computational power (De Vivo et al., 2016). For this reason, it is not practical to run MD simulations on large numbers of systems or for the analysis of slower biological processes.

A way to speed up MD simulations is via the representation of only a subset of the atoms in the protein, known as coarse-graining. Instead of the full number of protein atoms being present, groups of atoms are reduced to a single entity that often represents a functional group. This also allows the simulation of larger systems such as for membrane proteins, where they may require lipid molecules to be present for physiological behaviour to be observed (Loschwitz et al., 2020). However, this approach is not applicable to exploring the binding of small molecules to proteins due to the small scale of protein-ligand interactions and small molecule binding sites.

1.3.4 Machine learning

In drug design, machine learning (ML) is used widely to further our understanding of protein structure and protein-ligand binding amongst many other applications, and has been shown to improve both accuracy and speed in many tasks (Dara et al., 2022; Vemula et al., 2023).

ML is the automated detection of patterns in data, which enables the uncovering of key insights and the prediction of various properties. Whilst there are several approaches to ML, the core strategy remains constant: the computer is presented with a set of inputs, to which an algorithm is applied, allowing the discovery of patterns in the data without explicit pre-specification or instructions. The objective is for the ML model to generalise from this experience, and to accurately interpret new inputs in the context of the previously seen data. The majority of ML strategies fall into one of three broad categories: supervised learning, which aims to map input data to specified labels; unsupervised learning, which is the process of discovering hidden patterns in data without specified outputs; and reinforcement learning, in which the computer interacts with a changing environment to achieve a particular goal and is provided with feedback as it proceeds.

1.3.4.1 Supervised learning

Supervised learning involves the fitting of a function $f : X \rightarrow Y$, where X is the input data and Y is an associated label. Where the labels are integers, they represent classes and so a classifier is trained; where labels are continuous variables, regression is required. Given a set of inputs, the function f attempts to correctly predict their associated labels. This prediction is assessed by a loss function, which quantifies the difference between predicted and actual values. By providing feedback to f, the loss function guides the model's optimisation towards accurate predictions.

Once this training phase has been completed, the model is tested on unseen data,

which is a subset of around 10-20% of the available data. If the model has been successful in generalising from patterns in the original data, prediction accuracy on the test data will be high. However, effective testing of such ML models requires careful selection of training and test data to ensure that these two datasets do not exhibit high levels of similarity.

Common supervised learning algorithms include decision trees, random forests, support vector machines and neural networks. The work in Chapters 4 and 5 uses gradient boosting machines for classification tasks, so we explain here how these algorithms are constructed from decision trees.

Decision trees, random forests and gradient boosting machines Decision trees learn simple decision rules based on the features of the training data. While they are simple, easy to interpret, and relatively efficient, decision tree learners are susceptible to creating over-complex trees that do not generalise well and are sensitive to small changes in the training data (Costa and Pedreira, 2023). Ensembling methods have been used on decision trees to mitigate these weaknesses (Che et al., 2011).

Random forests are ensembles of independent decision trees which are trained on different random samples of the data (Breiman, 2001). The final prediction is determined by a majority vote for classification, or by the mean value predicted by all trees for regression. This prevents overfitting (where the model learns the training data but cannot generalise to new data) and results in more accurate predictions, while still maintaining ease of determination of feature importance (Ali et al., 2012).

Gradient boosting is another extension of the decision tree, however unlike random forests, gradient boosting trees are built sequentially (Freund and Schapire, 1997; Natekin and Knoll, 2013) (Figure 1.7). This additive building is designed such that each tree improves upon the deficiencies of the previous tree, by minimising the loss function between building trees. The final prediction is calculated by aggregating the

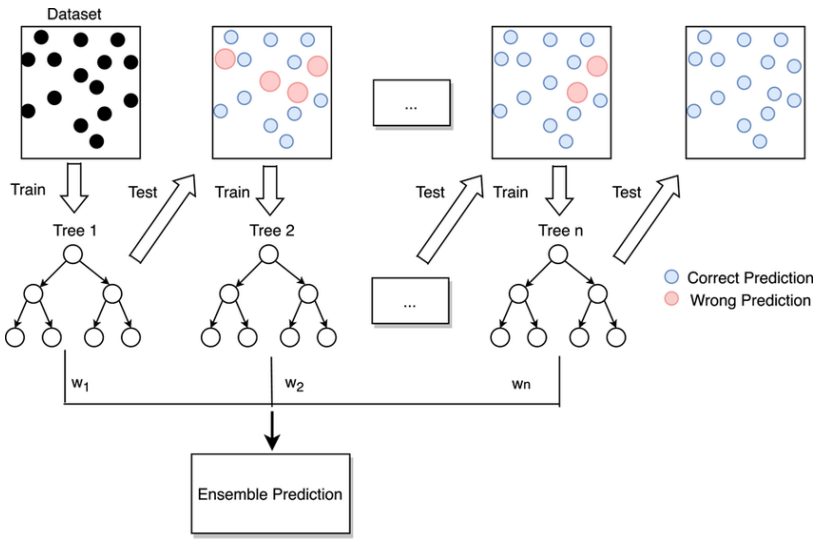


Figure 1.7: The training and assembly of a gradient boosting machine as described in Zhang et al. (2021).

results from each tree along the way. This approach results in models that are more accurate than decision trees or random forests, however gradient boosting machines are still prone to overfitting and so must be evaluated carefully (Mayr et al., 2014; Bentéjac et al., 2021).

Loss functions Loss functions are used to calculate the error of the current predictions of the model during training, and they provide feedback to the model to allow it to improve. The type of loss function used depends on whether the model is performing a classification or regression task, and the type of predictions that incur high loss values and are therefore penalised.

For classification tasks, two of the most commonly used loss functions are binary cross-entropy loss (also referred to as log loss),

$$L = -\frac{1}{m}\sum y_i \cdot \log(\hat{y}_i)$$

and hinge loss,

$$L = \max(0, 1 - y \cdot \hat{y})$$

where y is the actual label, $\hat{y}$ is the predicted label and m is the number of classes. Binary cross-entropy can be used for binary or multi-class classification, and heavily penalises predictions that have a large difference from the actual label (for example, where the true label is 1 and the predicted value is lower than 0.2). Hinge loss is primarily used in support vector machines, can only be used for binary classification tasks, and penalises both incorrect predictions and correct predictions that are not confident (Wang et al., 2022). For example, a (correct) prediction of 0.6 where the true label is 1 will be much more heavily penalised by the hinge loss function, than by the log loss function.

Evaluation For rigorous evaluation of the model, test data must be sufficiently different from the training data, otherwise models that are overfitted will not be detected, and the predictions made on truly novel data may be mistakenly relied upon.

The strategy used for ensuring this is the case depends on the type of data used. For example, where the task is the prediction of chemical properties based on molecular structure, molecular similarity (see Section 1.3.1) may be calculated between training and test molecules. This enables the identification and removal of molecules in the train and test sets that are very similar: for example, in Scantlebury et al. (2023), molecules in the training set with Tanimoto similarity higher than 0.8 (based on Morgan fingerprints with a radius of 3) to those in the test set were removed. For tasks involving protein sequence or structure, protein sequence identity or fold type may be used to split the training and test sets. Proteins with sequence identity higher than 30% usually share some evolutionary features, while those with higher than 40% sequence identity are very likely to share functional similarity (Pearson, 2013). However, some proteins with very low sequence identity may still share some structural features, and so CATH (Sillitoe et al., 2020) or SCOP (Andreeva et al.,

2013, 2019) fold classifications may also be used to remove proteins from the training or test sets.

Various metrics can be calculated to assess the performance of supervised ML models on test data. Accuracy, which measures the proportion of correct predictions, is the most intuitive score, but does not take into account the confidence of predictions, as a single threshold is used. Area under the receiver operating characteristic curve (AUROC or AUC) is able to take this into account by calculating false positive rate and true positive rate at different prediction thresholds. However, both of these metrics can be misleading when the labels in the test set are imbalanced: for example, if 90% of the test data is positive, and the model predicts *everything* as positive, accuracy will be 0.9, and area under the curve (AUC) will also be very high. Balanced accuracy can be used to mitigate this, which is calculated as the average accuracy of each label in the test set (in the example described, the balanced accuracy would be 0.45).

Other metrics that are useful where evaluating on imbalanced test data are precision and recall. Precision is calculated as the proportion of true positives in all true predictions, and recall is the number of true positives divided by the number of all (labelled) positives. These metrics can be combined into single scores in two ways: the area under the precision-recall curve, and the F1 score, which is the harmonic mean of the precision and recall (Hicks et al., 2022). In scenarios in which a small number of the most confident positive predictions are prioritised, precision is often the most useful metric, however if the aim is to rule out the only the most confident negative predictions, a high recall is important. Where the intention is to provide an overall evaluation of a model, F1 score may be the most appropriate.

Once the data that is too similar to the training set has been removed from the test set, the test set is often smaller than sufficient to rigorously assess the model, and so calculating metrics just once do not provide an adequate evaluation. A key

technique used to alleviate this is bootstrapping, in which the data is sampled with replacement several times, allowing a measure of confidence intervals, bias, variance and other metrics that allow us estimate how successful our predictions may be given a larger test set.

1.3.4.2 Unsupervised learning

Unsupervised learning is performed on unlabelled data, and the aim is for the model to discover patterns in the data without guidance. The main unsupervised learning implementations are clustering, association and dimensionality reduction.

Clustering uses the distance calculated between points, and can be performed in various ways. For example, where the number of clusters is known, KMeans clustering can be used (Jin and Han, 2010), or where the distance between 'similar' datapoints is known, density-based spatial clustering of applications with noise (DBSCAN) can be used (Ester et al., 1996).

Association mining is used to learn relationships between points in large datasets, such as correlations or co-occurrences. An example where this can be useful is in diagnostics, where sets of symptoms can be compared to a new patient to determine the most probable diagnosis (Stilou et al., 2001).

Dimensionality reduction is the process of decreasing the number of features in a dataset. It aims to extract important features from the dataset, reducing the number of irrelevant or random features present without compromising the signal contained in the original data (Velliangiri et al., 2019).

1.3.4.3 Neural networks

Artificial neural networks are architectures that can be applied to complex data, such as images (Han et al., 2018). They are formed of layers of input, hidden, and output nodes which are connected by trainable weights. Most architectures are fully con-

nected, and so with more hidden layers, the number of trainable parameters increases. This leads to the ability of so-called deep neural networks to automatically learn features of complex data without explicit featurisation (Shi et al., 2022). However, this also leads to high computational costs and susceptibility to overfitting (Ying, 2019; Thompson et al., 2022).

Where the spatial organisation of the data is important, such as in images, this can be leveraged by using convolution and pooling operations. By incorporating such operations, the convolutional neural network (CNN) is able to capture local patterns within the data while reducing the number of parameters (O'Shea and Nash, 2015).

Transformers Transformers are neural networks which learn distance-independent relationships between features in complex data. To achieve this, the correlations between every feature in the data are calculated, referred to as self-attention. These correlations can be processed in parallel, and can be used as weights to generate a new vector representation for each element. As this learning uses labels built-in to the data, it can be described as unsupervised or self-supervised.

Using the new vector representation as an input, a supervised learning algorithm can then be trained for a specific task. This avoids the requirement for explicit featurisation or definition of rules.

This architecture was motivated by the task of learning language, as words in a sentence have both short- and long-distance relationships. A vector can be produced for each word in a sentence. This approach has been shown to excel in machine translation (Vaswani et al., 2017) and has also been applied to several areas in CADD (see Section 1.3.5.3).

1.3.5 Application of ML to FBDD

ML has been applied to every stage of drug design. It has been used for design purposes, such as the generation of novel small molecules (e.g. Zeng et al. (2022a)). Predictions of ligand properties can made, to understand their physicochemical behaviour, potential toxicity and off-target effects (e.g. Rao et al. (2019); Zeng et al. (2022b); Takai et al. (2023)). Accurate protein structure prediction is also now possible, which facilitates computational analysis of proteins for which an experimental structure has not yet been solved (e.g. Jumper et al. (2021)). As FBDD aims to find diverse hits on novel targets, virtual screening and protein structure prediction are two areas which provide opportunities to improve the efficiency of FBDD campaigns.

1.3.5.1 Virtual screening

Virtual screening is the process of computationally predicting whether a ligand will bind to a protein target. It can be based on the knowledge of other ligands which are known to bind the protein target (ligand-based virtual screening) or alternatively the input data can consist of structural information from the protein (structure-based virtual screening). While ML is not required for virtual screening, many recent methods for virtual screening are ML-based (de Oliveira et al., 2023).

As FBDD is primarily used on targets which do not have many known binders, structure based virtual screening (SBVS) can be used to select fragments that are predicted to bind to the target of interest more efficiently. Usually, this is achieved by first placing the ligand into the pocket and predicting the pose that the ligand takes (docking), followed by the prediction of the binding affinity for this protein-ligand complex. Only compounds that have been predicted to have efficient binding with the protein are then taken forward for experimental screening. If this results in more high-quality hits, the drug design process will be faster and cheaper.

However, recent studies have shown that both the ligand placement stage and the

binding affinity prediction stage are susceptible to error and bias, and are not learning about the biophysics involed in protein-ligand binding (e.g. Scantlebury et al. (2020); Buttenschoen et al. (2023); Durant et al. (2023)). Additionally, these techniques are usually trained on ligands that are much larger than fragments, and so can not be reliably applied to fragments. For this reason, it is important to explore fragment-specific virtual screening and other computer-aided avenues for the efficient design of fragment libraries.

1.3.5.2 Protein structure prediction

In the past three years, protein structure prediction has undergone huge improvements, with the release of several accurate ML-based tools, including AlphaFold (Jumper et al., 2021), RosettaFold (Baek et al., 2021), OmegaFold (Wu et al., 2022) and ESMFold (Lin et al., 2022). These developments provide us with the opportunity to learn from protein structures before the structure has been experimentally solved.

However, there are several important considerations when using predicted protein structures. For novel structures with no known homologues, over 50% of predictions are likely to deviate more than 2Å root mean square deviation (RMSD) from the experimentally-solved structure (see Section 4.2.5) (Kryshtafovych et al., 2021). Even for highly accurate structures, binding site differences are significant enough to make virtual screening significantly more difficult (Díaz-Rovira et al., 2023; Holcomb et al., 2023; Scardino et al., 2023). Additionally, many SBVS tools require the binding site to be specified (Ghersi and Sanchez, 2009; Zhao et al., 2020). A final consideration is the challenge of evaluating the performance of predictive models on protein structure predictions, as much of the viable test data has likely been used in the training of the protein structure prediction tool. We discuss these issues in more detail in Chapter 4.

1.3.5.3 Proteins and transformers

Following the successful implementation of transformers in machine translation tasks (Vaswani et al., 2017) (see Section 1.3.4.3), the transformer architecture has been applied to several areas in CADD. Similar to language, protein sequences are represented as letters. By applying transformer architectures to protein sequences, with each protein residue being treated as a word, the relationships between different residues in a sequence can be extracted.

This can be leveraged in various ways. For example, the contacts made between residues in a protein's three-dimensional structures can be directly inferred from the attention between residues, offering structural insights (Lin et al., 2023). The weights from the network can also be extracted as vector inputs for the training of supervised learning models, such as for the prediction of ligand binding, protein-protein interactions (PPIs) or disease variants (Chandra et al., 2023; Lee and Nam, 2022; Brandes et al., 2022). In Chapter 4 we describe the use of these learnt representations in more depth, and implement them in both Chapters 4 and 5.

1.3.5.4 Data requirements for machine learning on protein structures

Generally, a ML model requires at least ten times as many data points as there are features in the dataset (Chicco, 2017). The large number of parameters in deep neural network models places an additional requirement for large training set size. For context, DeepFace, a deep learning facial recognition system, was trained on four million images (Taigman et al., 2014).

Protein sequence data is abundant: ESM2 (Lin et al., 2023) was trained on around 65 million unique protein sequences (maximum 90% sequence identity). However, the quantities of data available to train ML models for tasks that incorporate protein structure are significantly lower than this. Models for the prediction of protein structures or for virtual screening are usually trained on structures taken from the

PDB (Berman et al., 2000), which as of November 2023 contains around 212,000 structures. These are ligand-free structures or ligand-bound complexes of just 87,000 unique proteins (based on 95% sequence identity). To get around this shortage of data, ESM-IF1 (Hsu et al., 2022), a transformer trained on three-dimensional protein backbone structures, was trained on 12 million predicted protein structures in addition to the ∼16,000 experimental structures used. However, use of predicted structures is not currently possible for learning about protein-ligand complexes, with such tools usually being trained on around 20,000 experimental complexes in total (Corso et al., 2022; Yu et al., 2023b; Scantlebury et al., 2023).

Several studies have explored how the performance of these predictive models improves with more data. Fan and Shi (2022) found that a deep learning-based virtual screening model was able to achieve a Pearson's correlation coefficient (PCC) (Freedman et al., 2007) over 0.8 when trained on 320,000 data points, but this value was reduced to 0.7 when only 100,000 data points were available. Another example of such an effect was shown on a model predicting the effects of mutations on antibody-antigen $\Delta\Delta$G: at least 90,000 mutations were required to achieve a PCC of ∼0.85, while the reduction of training set size to 50,000 mutations reduced the score to ∼0.79 (Hummer et al., 2023). A training set size of 10,000 mutations resulted in a PCC of just ∼0.45.

The training of effective ML models also requires high quality data. Fan and Shi (2022) found that where complexes had multiple associated binding affinity values (over 25% of data), these figures often varied by up to one order of magnitude (pK ±1). However when this level of random error was introduced to the data, the PCC was reduced from 0.80 to 0.74, demonstrating the sensitivity of these ML models to errors in the training data.

Negative data is also an important component of a training dataset. A thorough study of the impact of negative training set size on ML-based virtual screening found

a clear improvement model performance as more negative data was included in the training set (Kurczab et al., 2014). However, the results of screening studies usually include only the positive results, in the form of protein-ligand complexes, and the list of tested molecules which did not bind to the protein of interest is not available. In the absence of negative data, many virtual screening models are trained on positive examples and decoys, which are protein-ligand complexes which have not been observed (Mysinger et al., 2012). However, these have been shown to be commonly biased, allowing methods to exploit these biases instead of learning to perform molecular recognition. For this reason, methods to generate property-matched decoys have been developed (e.g. Imrie et al. (2021)).

Data augmentation, the mining of several databases, synthetic data and the use of pre-trained models are all techniques that can be implemented to attempt to alleviate issues arising from lack of data. However data quantity and quality remain a challenge for the training of accurate ML models for structure based drug design (SBDD).

1.4 Thesis outline

In this chapter, we have outlined the issues faced in the field of SBDD, and introduced two areas of innovation that aim to reduce costs and speed up the design of novel small molecule drugs. We have described the successes and limitations of FBDD and discussed how ML is applied to FBDD.

In Chapter 2, we survey historic fragment screening data generated by by XChem between 2015 and 2021, to evaluate whether there yet exists the quantity and quality of data required for the bulding of effective CADD tools. We demonstrate that there exist significant inaccuracies in model building and annotation of models. Additionally, we discuss how the file storage system resulted in loss of information across many screening campaigns. We describe a new storage and curation system that can be

used for future re-processing and large-scale analysis of the fragment screening data.

In Chapter 3, we compare key structural and pharmacophoric features of commercially available fragment libraries. We then use protein-ligand interactions to show that structurally-diverse fragment libraries exhibit significant structural redundancy. Finally, we propose a strategy for selection of functionally diverse fragments which results in more efficient recovery of interactions from novel targets.

In Chapter 4, we propose a new method for binding site prediction based on the classification of protein residues as represented using an inverse folding model. While comparing our method to existing tools in the prediction of binding sites of predicted structures, we find that the predicted structures in our test set are much more accurate than we expected. To ameliorate this, we reduce the accuracy of these structures using molecular dynamics simulations. On these models, we are able to predict ligand binding sites with higher accuracy than competitors, with this lead opening further where multiple structures are available. This allows us to accurately predict binding sites of novel targets, so that target-specific library design can commence.

In Chapter 5, we explore possible approaches for the design of target-specific fragment libraries via binding site comparison. First, we test whether ProCare (Eguida and Rognan, 2020) is able to identify sites from different proteins that bind identical ligands. We find that while the pocket alignment method is somewhat effective, the scoring strategy is flawed and results in the method being ineffective for target-specific fragment selection. We then study the possibility of using learnt representations of proteins for the same task. While a simple distance-based metric is not a viable strategy for fragment selection, we find that learnt representations of proteins are rich sources of information on protein-ligand interactions, so could be used in future binding site comparison tools.

Finally, in Chapter 6, we summarise the results of this work, discuss the implica-

tions for the field, and suggest future directions.

Chapter 2

Analysis of available fragment screening data

2.1 Preface

Since 2015, XChem has processed over 200,000 crystals at Diamond Light Source, primarily on the I04-1 beamline. Each of these crystals has been soaked with a specific fragment and X-ray diffraction has been used to determine whether the fragment binds to the target of interest. As efficient protein-ligand interactions are required for fragments to bind to the protein target, these structures comprise a rich source of information on the binding of small molecules to proteins (Scott et al., 2012; Hopkins et al., 2004).

To design effective computational tools for fragment library design, large sets of high-quality experimental results are required. Complete results from fragment screens (both the positive binding cases and the negative non-binding cases) are not widely available, so curating historic XChem results would directly facilitate computational analyses and the development of predictive tools. In this chapter we have examined the data processed by XChem to determine the volume of data, what fea-

tures of the data are reliable, and what limitations and biases are present. To explore this, we tracked every crystal processed by XChem (from academic users) through each stage of its journey from diffraction pattern to modelled ligand, with the aim of labelling an outcome. During this process, we gained a deeper understanding of the challenges faced when processing large numbers of complex experiments and how this impacts the quality and availability of data.

We found that for much of the historical data, the annotation of final results was not sufficiently regimented to enable computational tools to draw reliable conclusions from each fragment screen. Additionally, the volume of data generated for each screen resulted in a large proportion of historical data being archived in an inaccessible manner. The more recent data were annotated more thoroughly, but the complex nature of the data makes it difficult to use in a high-throughput learning platform. In this chapter we describe the challenges faced when attempting to assemble an accurate set of positive (bound) and negative (not bound) results from historic XChem experiments.

2.2 Introduction

As discussed in Chapter 1, the training of machine learning models for computer-aided drug design requires large quantities of high-quality data. Compared with traditional high-throughput screening of drug-like molecules, fragment-based drug design results in a much higher number of complexes. Due to the small size of fragments, ligand efficiency is high and so this type of data may contain stronger signals from which machine learning models can learn, compared to structures of a protein bound to a drug-sized ligand. Fragment screens also offer the promise of negative data – lists of fragments which do not bind, something which is very uncommonly collected from drug-sized ligands.

As FBDD is an emerging field, we set out to explore whether there yet exists the required quantity of high-quality data for the development of machine learning and, in particular, deep learning binding prediction methods.

2.2.1 Crystallographic fragment screening at XChem

As described in Chapter 1, macromolecular crystallography is the dominant method of determining protein structures (Shao et al., 2022a). It is more sensitive to weak binding and generates more structural information than assay-based binding detection methods, such as NMR or biochemical assays (Davies, 2014; van Montfort et al., 2017b). This makes X-ray crystallography ideal for the screening of fragments, which bind more weakly than drug-sized molecules.

Traditionally, this was a process that took several months; however, advances within the last ten years have enabled high-throughput fragment screening to take place. XChem have sped up the process of fragment screening reducing the time from over 6 weeks to around a week (Figure 2.1) (XChem, 2020b). Using their improved technology, XChem can screen more than 2000 crystals per week.

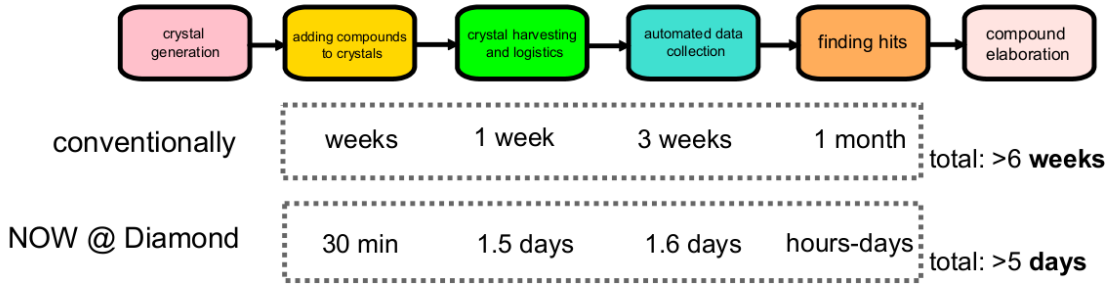


Figure 2.1: The pipeline of experimental compound screening using X-ray crystallography. XChem at Diamond Light Source has been able to significantly accelerate the process of screening and data analysis. The bulk of the experimental screening and data analysis is complete within one week, however model completion remains a difficult task, sometimes taking several months.

While the majority of the experimental screening and the initial data analysis is

complete within one week, the process of building and validating the ligand models can take several months. In this chapter we focus on this 'finding hits' stage (Figure 2.1), which is the process of determining if and where fragments are bound to the target structure (Figure 2.2).

In this hit-finding process, the raw electron densities are first processed by DIMPLE (Wojdyr et al., 2013), which is used to calculate a Sigma-A weighted difference map. Then, Pan-Dataset Density Analysis (PanDDA) (Pearce et al., 2017) is used to calculate event maps which represent areas of outlying electron density that could correspond to a bound ligand. Concurrently, GRADE (Smart et al., 2021) generates ligand restraints so that a SMILES string can be converted into a three-dimensional ligand structure. Finally, this ligand structure is fit to PanDDA event maps by a crystallographer within the XChem Explorer interface (Douangamath et al., 2021).

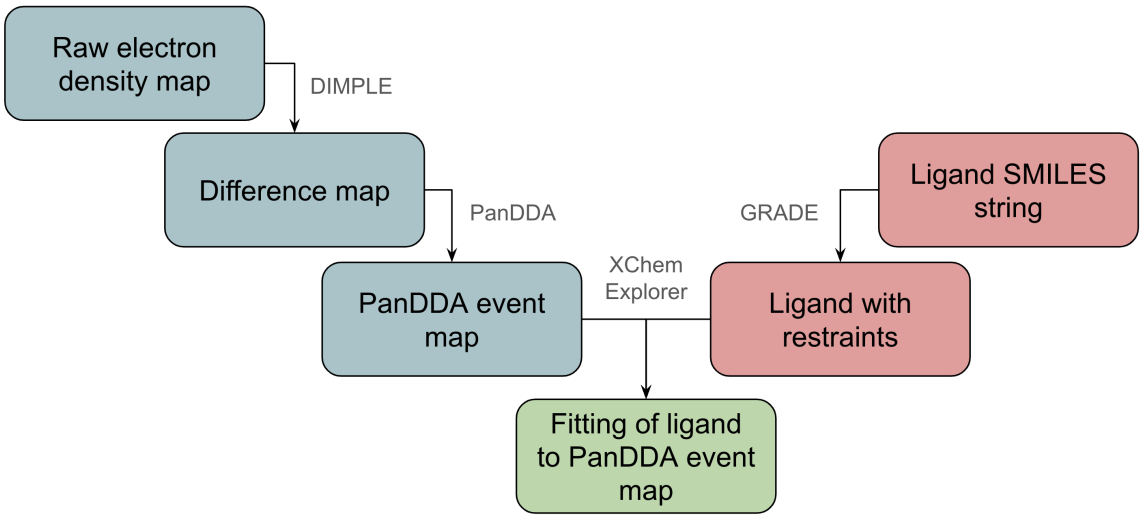


Figure 2.2: The hit-finding process used at XChem. Calculation of difference maps and PanDDA event maps by DIMPLE and PanDDA is automated, however fitting of the ligand structure into the PanDDA event map is carried out by crystallographers.

This method of ligand modelling is highly reliant on the correct interpretation of electron density maps and PanDDA event maps (see Section 2.2.3) and can contain inaccuracies (Pearce et al., 2022). It is therefore important to analyse this step when determining the quality of data produced in a fragment screen, and this is the central

focus of this chapter. We first explain the difficulties faced when identifying fragment hits, then review various strategies for fragment hit identification and ligand model validation.

2.2.2 Hit identification and ligand modelling in fragment screens

Traditionally, hit identification requires the crystallographer to identify areas of electron density that have not been modelled as part of the protein structure. If the density is of a shape that can be explained by the soaked ligand, the ligand is fitted into the density (Davies, 2014). However, because fragments have weaker binding and thus a lower occupancy than drug-sized ligands, the electron density corresponding to fragments is also weaker. The electron density corresponding to a fragment is therefore not always easily identifiable from the original electron density maps, as the signal may be obscured by proteins without a ligand bound and by experimental artefacts (Pearce et al., 2017).

The challenge of identifying fragment-related electron density has been well documented. For example, Schiebel et al. (2016) found that without rigorous refinement of structures, around a quarter of fragment-binding events were missed. An automated refinement workflow was implemented to improve the identification of fragment hits while avoiding the need for a team of crystallographers to manually analyse the quality of the data and reprocess low-quality data. However, manual inspection of the refined maps was still required to identify electron densities which correspond to fragment hits, and to build ligand models into these densities.

The process of building ligand models into electron densities can be further complicated when cocktails of fragments are screened together, rather than a single fragment per crystal. While this approach is used to increase screening throughput, it creates a requirement to first identify which fragment from the cocktail has bound, before modelling the ligand structure (Bauman et al., 2012).

Another important factor in model building is the level to which we can identify accurately the position of ligand atoms, which is generally limited by crystal robustness (Patel et al., 2014). For crystals with high resolution (<2.2Å), we can see many or all of the atoms clearly. However, some crystals may have a lower resolution, making it only possible to understand positions of functional groups or perhaps only a rough position of the whole fragment. Additionally, the fragment can occupy different conformations, which is reflected in the B-factor value assigned (Sun et al., 2019), thus indicating a high level of uncertainty of the atomic positions.

PanDDA (Pearce et al., 2017), used in-house at XChem, is another approach to improve the identification of fragment binding events from difference maps.

2.2.3 The Pan-Dataset Density Analysis method

To extract the signal from electron density maps that corresponds to bound fragments, the Pan-Datset Density Analysis (PanDDA) method was developed (Pearce et al., 2017). It creates an electron density map that corresponds specifically to the 'ground state' (the protein with no ligand bound) by using many crystals of the same protein. This ground state can then be subtracted from any specific observed electron density for a given protein, leaving only outlying densities that correspond to structural change or potential binding events (known as PanDDA events).

These PanDDA event maps can then be inspected by crystallographers or users to determine whether the PanDDA event corresponds to a bound fragment, and if so, model the fragment. Several rounds of refinement are used to ensure the optimal model. Where possible, the final model is then labelled as 'bound', for operational reasons (F. von Delft, private communication). A comparison with traditional ligand modelling is shown in Figure 2.3, which demonstrates how PanDDA has the potential to allow us to identify the positions of specific atoms in the fragments to a significantly higher precision than traditional maps.

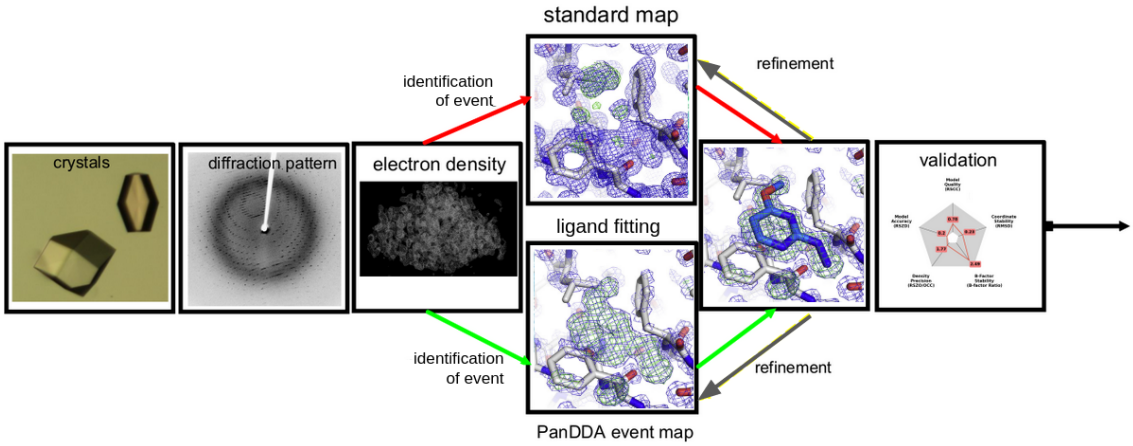


Figure 2.3: The data analysis pipeline of fragment screens. X-ray crystallography produces diffraction patterns, and from this an electron density map is calculated. The protein model is built into the electron density, and areas of density that may correlate with a ligand are located. At these points, a ligand is fitted into the model. A standard map (upper path) and a PanDDA event map (lower path) of the same experiment are shown to compare the clarity of the ligand density. Validation may take place after refinement and before publication of results or elaboration of compounds. Adapted from (Lysozyme.co.uk, 2020; Wikipedia, 2020; Center for Free-Electron Laser Science, 2020; Horowitz, 2020; Wollenhaupt et al., 2020; Pearce, 2020).

2.2.4 Validation of ligand models

It is acknowledged that the building of ligand models in electron density maps is open to bias and error (Koehler, 1993; Deller and Rupp, 2015; Smart et al., 2018; Wlodawer et al., 2018; Pearce et al., 2022). Even where the electron density clearly indicates a true binding event, errors may still arise in the model (Liebeschuetz et al., 2012; Deller and Rupp, 2015). For this reason, several tools have been developed to help experts or non-experts to model ligands into electron density in a reliable way (Pozharski et al., 2013; Smart et al., 2018).

Most tools use a combination of metrics to independently evaluate two factors: the theoretical feasibility of the model and the likelihood that the model is an accurate representation of the electron density map. Scores based on the real-space correlation coefficient (RSCC) (see Table 2.1) are viewed as the best single way to assess a model (Weichenberger et al., 2013), but this score has significant dependence on resolution.

Validation HElper for LIgands and Binding Sites (VHELIBS) (Cereto-Massagué et al., 2013) allows users to validate protein-ligand models based on several metrics derived from the model and the electron density, such as the real-space correlation coefficient (RSCC), real-space difference density Z-score (RSZD) and average occupancy. These are combined to classify the model as 'good', 'dubious' or 'bad'.

The PanDDA package (Pearce et al., 2017) calculates five validation metrics, shown in Table 2.1, but does not produce a single final score that specifies the quality of the model. Twilight (Weichenberger et al., 2013) uses RSCC and resolution (RESOL) to calculate a single 'S' score,

$$S = \frac{2}{(RSCC/0.6) + (RESOL/1.3)},$$

which provides an intuitive ranking for the degree of misfit in the electron density dependent on the resolution of the protein structure. This is used solely to rank models in the user interface of Twilight, and not used in classification of electron density or of models.

Recent research has highlighted the need for validation tools, as it was found that up to 73% of PDB protein-ligands structures may have inaccuracies (Chakraborti et al., 2021). Using VHELIBS, they found that of 280,000 protein-ligand pairings derived from 66,000 PDB entries, just 27% were classed as 'good' while the remainder were 'dubious' or 'bad'. Only 11% of the structures had both binding site residues and ligands modelled in reliable agreement with the experimental electron density. Upon examination of the lowest quality models, it was found that external organisations carried out the building of these models, with little checking or validation by the internal scientists publishing the structures.

To improve data integrity, the PDB has recently implemented quality indicators for ligands modelled in deposited structures, including RSZD, RSCC, RMSZ-bond-length, RMSZ-bond-angle and clash-per-atom (Shao et al., 2022b). They found that

Metric	Acceptable value	Measures
Real-space correlation coefficient (RSCC)	$\geqslant$0.7 (Pearce et al., 2017); $\geqslant$0.8 (Deller and Rupp, 2015)	Similarity between electron-density maps calculated from structural model and experimental data, to indicate model completeness
Real-space difference density Z-score (RSZD)	$\leqslant$3	Difference density peaks in the electron density sphere of an atom, to indicate model correctness
Real-space observed density Z-score (RSZO)	$\geqslant$1	Signal to noise ratio, to indicate model precision
B-factor ratio (ligand:surrounding residues)	not $\gg$2	Local atomic movements in ligand compared to surrounding residues to indicate stability of refinement
Root mean square deviation (RMSD)	$\leqslant$ 1	Difference between atom positions before and after refinement to indicate model coordinate stability

Table 2.1: Five metrics used by PanDDA along with their suggested acceptable values defined in (Pearce et al., 2017) to assess and validate model quality.

2% of ligand structures contain chirality outliers representing major errors that require structure redetermination. Around 51% of ligands in PDB structures were found to exhibit steric clashes. Approximately 6% of PDB structures determined by X-ray crystallography contained ligand structures that were missing non-hydrogen atomic coordinates, mostly due to issues with the signal in the corresponding part of the electron density map. To allow researchers to easily understand the quality and reliability of the ligand modelling, ligands are now given an overall ranking score on experimental data fitting and geometry, based on the five calculated quality metrics. The mean data fitting score of <1Å-resolution structures is 0.75, while structures with resolution between 3.4 and 3.5Å is around 0.20. While the geometry score is similar for different resolutions, it has improved with time: structures modelled in 2000 score on average around 0.35, whereas structures deposited in 2020 have an

average score of 0.60.

As the electron density associated with fragment hits is even less reliable than that of larger ligands, protein-fragment complexes may need to be handled slightly differently. A recent article (Jaskolski et al., 2022) discussed the differences between electron density maps used in the PDB ligand quality assessment and the PanDDA event maps that are used for fragment modelling. It was argued that mass depositions of fragment screening data in the PDB do not conform to the expected quality standards, which could cause issues for biomedical research purposes and for training structure prediction models such as AlphaFold. A response to this commentary (Weiss et al., 2022) argued that by taking into account heterogeneity within crystal systems and low occupancy of fragments, the depositions represent high-quality models, and that barring them from depositing in the PDB would stunt development of the field.

Even where validation tools are used, there remains a need for an expert crystallographer to ensure that electron density is identified correctly and that ligand models are built appropriately. However, user facilities such as XChem are designed to allow non-experts to perform fragment screening on their targets of interest and identify starting points for drug development. The tools available for analysis of electron density and hit validation leave the hit identification and initial model building stages entirely up to the user, so it is important to investigate the integrity of data output from such experiments. The complexity of crystallographic data translates to a high potential for bias; even with use of PanDDA, there is room for hits to be missed or ignored, or poor models to be labelled as a hit.

2.2.5 Availability of fragment screening data

The data made available from fragment screens is often incomplete: fragment screens that lead to promising drug leads are likely to be discussed in publications, but often

only complexes containing the specific fragments involved in the design of the follow-up compounds are made available (Chessari et al., 2021).

A fragment screen frequently contains significantly more information than this: other fragments may have bound to the same or other binding sites; many fragments will not have bound at all; some fragments may have reduced the quality of crystals, leading to failed experiments. While these outcomes may not themselves be clear and and thus make it a challenge to assign unequivocal labels to every protein-fragment combination, all of these pieces are key to understanding FBDD and SBDD. Therefore, making these annotations available to the community would drive improved methodologies.

For example, a validated ligand model can be confidently labelled as 'bound', a failed experiment can be labelled as such, and where crystals contain no density to support a bound ligand, this can also be recorded. Ambiguous density could be labelled as 'unsure', to ensure that these crystals are treated with appropriate caution. While these labels would rely on assumptions that are limited by experimental constraints (such as sensitivity to weakly bound ligands or resolution), they would provide a more complete dataset on which CADD tools could be built.

With the aim of assembling a set of complete results from fragment screens, we first explored the reliability of annotations in the dataset. We then combined the two main database resources at XChem to curate historic data into a format compatible with large-scale analysis of fragment screens.

2.3 Methods

2.3.1 Evaluation of annotations of historic XChem data

In order to use the results from the XChem experiments to understand the binding of fragments to proteins through the machine learning-based analysis, we must first

understand the accuracy and coverage of the experimental outputs. To explore this, we used a set of 8,633 crystals and their modelled structures that had previously been curated by Conor Wild (C. Wild, private communication) to evaluate the reliability of annotations of ligand-bound structures. Structures that had been given a 'bound file' annotation by the user were labelled as bound; structures without this annotation were labelled as 'not bound'.

To evaluate the accuracy of the protein-fragment ('bound') models that were produced, we calculated the RSCC for all 8,633 models. This measures the correlation between the electron density expected from the structure that was modelled with the observed electron density, with a score of 1.0 being complete overlap and a score of 0 showing no correlation. Although RSCC does not measure all aspects of model quality (see Table 2.1), it is an approximation of whether the bound structure modelled corresponds to the density observed or not. To ensure that we did not erroneously penalise models that were almost correct, we lowered the threshold for a positive result from the generally accepted RSCC $\geqslant 0.8$ (Deller and Rupp, 2015) to RSCC $\geqslant 0.7$. Only models with RSCC $\geqslant 0.7$ and labelled by the user as 'bound' or RSCC < 0.7 and labelled as 'not bound' were considered to have been correctly labelled. As RSCC has some resolution dependence, we also collated the resolutions for the 8,633 crystals and compared this with the RSCC values calculated.

Users are able to use the PanDDA 'giant score' functionality to calculate the validation metrics shown in Table 2.1 for each model that they believe represents a ligand bound to the protein. We conducted a preliminary study using 65 protein-ligand complexes which each had these 5 metrics previously calculated, to test which models had acceptable values in all metrics and so would be considered high quality models on all accounts. We set the thresholds for acceptable values at exactly those shown in Table 2.1 (as in Pearce et al. (2017)), except for the B-factor ratio which was not explicitly set: we required the value to be less than 2.5.

2.3.2 Data searching and curation

At DLS, every crystal screened is recorded on the ISpyB database (Delagenière et al., 2011). Raw diffraction data are automatically processed to generate human-interpretable electron density maps for each crystal. Each crystal screened at XChem and its corresponding fragment is also recorded in an SQLite database (SoakDB), which contains all of the information relevant to the campaign from soaking in the lab through structure solution and refinement, and even eventual deposition to the PDB. Relevant file locations and various validation scores that we required for our analysis are also included. We used the ISpyB and SoakDB databases in tandem to determine how many electron density files we were expecting to find on the DLS server for each protein target.

We determined it was necessary to use cloud-based object storage as the server did not have sufficient space (with data storage being a common problem for large datasets, such as those coming from high-throughput crystallography), and we secured access to Echo S3 data storage. In this system we were able to curate electron density maps for crystals screened by XChem.

2.4 Results

2.4.1 Analysis of quality of annotations for XChem fragment screens

We calculated RSCC for 8,633 crystals and their modelled structures. Only models with RSCC $\geqslant 0.7$ were regarded as a true protein-fragment complex. These annotations were compared with the labels of 'bound' or 'not bound' to examine the correlation between a high RSCC value and an annotation of 'bound'. The results are shown in Table 2.2. 5,643 ($\sim$65%) models were correctly labelled, while 2,990 ($\sim$35%)

	RSCC $\geqslant$ 0.7	**RSCC <0.7**
Bound	933	1255
Not Bound	1735	4710

Table 2.2: RSCC-based classification of 8,633 modelled structures compared with their 'bound' or 'not bound' annotations provided by the user shows that many reliable models are missed, and many poor models are labelled as a hit.

were labelled inconsistently with the RSCC value calculated. For simplicity, we refer to these cases as 'incorrectly labelled'. This simple test suggests that the annotations applied by users are not sufficiently accurate for a large-scale analysis such as the training of a machine learning model and that there may be a bias towards labelling low RSCC events as 'bound'.

RSCC is a continuous metric and has significant dependence on resolution. To rule out the possibility that many of the incorrectly labelled models may lie close to the minimum threshold of RSCC = 0.7 we examined RSCC values for each of these models in the context of the resolution of the structures (Figure 2.4).

We observe that while there are many high-resolution high-RSCC structures that are labelled as bound, there are also a significant number of structures which have a high resolution and low RSCC, but are also labelled as 'bound'. Of the 1,255 structures that have been incorrectly labelled as bound, 251 lie within $0.65 \leqslant \text{RSCC} < 0.7$. Of the 1,735 structures that have been incorrectly labelled as not bound, 734 lie within $0.7 \leqslant \text{RSCC} < 0.75$. In addition to this, there are several models which lie far from the threshold and remain incorrectly labelled, such as those shown in Figure 2.5. In Figure 2.5A we show a model labelled as 'not bound' that fits well within the modelled structure, where we can see electron density that clearly correlates to each part of the ligand structure. Conversely, Figure 2.5B shows a ligand labelled as 'bound' by a user in which the ligand model exhibits poor overlap with the observed electron density. However, the shape of the density does correlate with the shape of the fragment. In this case it may be possible to achieve a higher RSCC value by

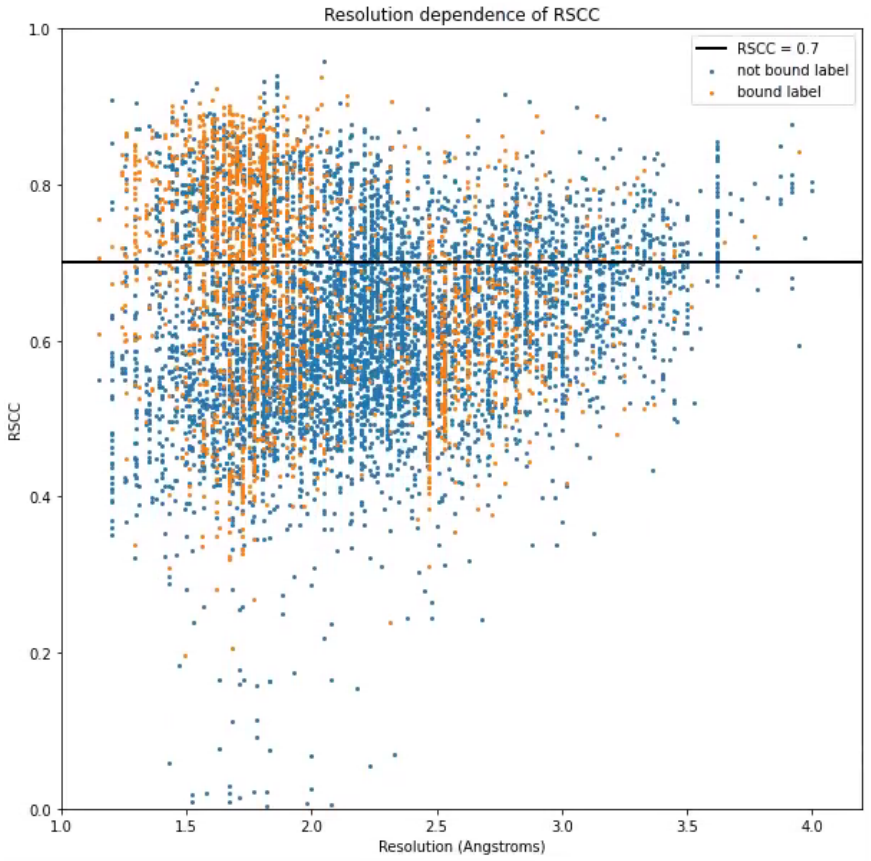


Figure 2.4: RSCC values of the model's expected density and the observed density for PanDDA events for 8,633 crystals do not show a clear correlation between RSCC and annotation as 'bound' or 'not bound'. Structures that have been potentially labelled incorrectly (blue points above RSCC=0.7 (black line) and orange points below RSCC=0.7) are not clustered near to the threshold, so are unlikely to represent ambiguous results.

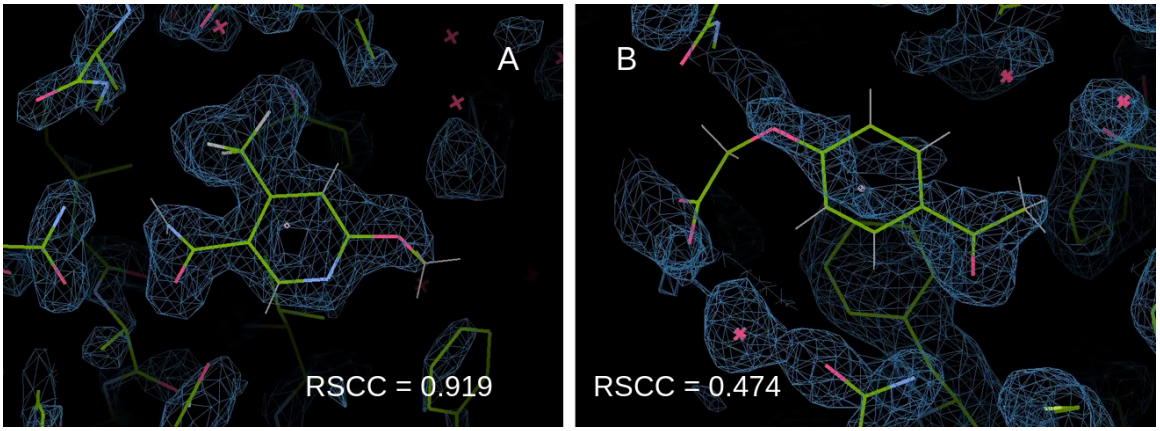


Figure 2.5: An example of mislabelling of modelled structures at the binding sites of two proteins. **A:** a model annotated as 'not bound' by the user; **B:** a model annotated as 'bound' by the user. Blue mesh represents observed electron density in the form of a PanDDA event map (Pearce et al., 2017). Sticks represent modelled structure.

adjusting the position of the fragment.

Of the models for which we calculated RSCC, just 65 (of the total 8,633 models) had been validated using the PanDDA 'giant score' function which calculates all metrics from Table 2.1. Of these models, 63 satisfied the RSCC $\geqslant 0.7$, while the other two had RSCC scores of 0.62 and 0.67. Only 42 of these 63 models produced acceptable values for all five metrics. A further 18 of the 63 were deemed good models by four of the five metrics. If this is representative of the wider picture, applying all five metrics as classifiers could cause over a quarter of models that likely represent a true bound structure but are poorly built to be discarded. However, requiring four of five metrics to be satisfied could constitute a more viable validation regime that takes into account the nuance and complexity of model building.

These findings underline the challenges faced when building and validating ligand models.

2.4.2 Structures uploaded to Fragalysis undergo additional validation

Fragalysis is an online platform on which structures of the protein-ligand complexes from fragment screens at XChem are uploaded. Only protein-ligand structures from academic screening campaigns in which all ligand modelling and refinement was conducted by expert crystallographers are uploaded. Thus, all structures available on Fragalysis are guaranteed to have undergone rigorous validation (F. von Delft, private communication).

However, Fragalysis only contains structures for fragments that bound to the protein target – leaving out the large list of fragments which do not bind to the target of interest. This information has not been explicitly recorded and rather exists implicitly where there is no record of a modelled protein-ligand complex and no record of failure of processing of the crystal or its experimental data. To make full use of

the experimental data generated by XChem for machine learning purposes, the next steps would be to reprocess the dataset to extract as many of the clear positive and negative results as possible.

2.4.3 Curation of electron density maps for XChem crystals

As the annotations contained in the current version of the data were less reliable and complete than expected, we sought to curate a version of the XChem data that would allow for homogenised reprocessing. This would enable the generation of a reliable set of fragment screening data that would be appropriate for large scale data analysis and machine learning.

Upon initial attempts to locate protein-fragment structures for targets screened at XChem, we found that a majority of targets had no associated bound-state structures available on the DLS server. Further investigation showed that entire directories of historic data (everything from initial experiments to final analyses) had been archived, and were stored on tape at DLS. Additionally, some experiments had been analysed locally on the users' computers, so the final results were not available on the DLS server. No records had been kept of which protein-fragment structures had been modelled for each target, or where in the directory the files were kept. Additionally, there was no infrastructure in place to retrieve the full set of files, and at an estimated 30 terrabytes (TB) of data (nearly half of all experiments at XChem from 2015 to 2020), XChem did not have a location in which the data could be stored in an accessible manner.

Staff from DLS developed a protocol for the temporary retrieval of the data in a staggered manner, in which a few TB at a time could be housed for 14 days on the central disk. We moved the data to S3 object storage as it became available on disk. For any new files generated by XChem, we established an automated back-up to the object store to ensure that we would have continuous access to all data. Once the

data had been transferred to the object store, we were able to search through the files.

We prioritised the curation of the electron density maps that are the outputs of DLS automated processing tools. These maps are the input for PanDDA, and by collecting these, it would be possible run the updated versions of PanDDA on older experiments. We searched through the files transferred to the object store to locate an electron density map for each crystal. We found electron density maps for ~81,000 out of the around 220,000 crystals successfully screened at XChem. For each of these crystals, using the ISpyB and SoakDB database files, we made sure that there was an associated reference structure and a recorded valid SMILES string for the fragment soaked, to ensure that additional analysis could be carried out. The directories were restructured so that instead of being organised as 'visits', crystals were grouped by protein target and were stored together with their required reference structures, SMILES strings and other important metadata.

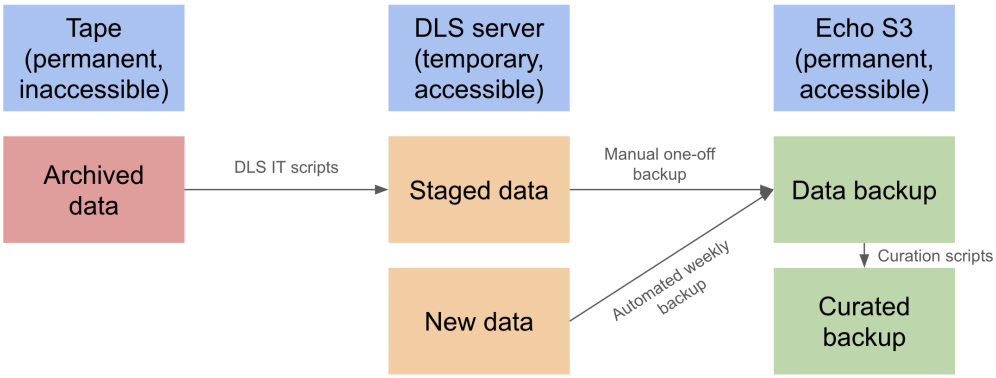


Figure 2.6: The backup and curation of XChem data. Data was originally on tape or on the DLS server, and was backed-up to S3 object storage and curated so that all crystals for a given protein target could be easily located.

The curation of these electron density maps facilitated the development of new methods for processing electron density maps (developed by Conor Wild), which aims to generate more accurate PanDDA event maps, and consequently facilitate better modelling and clearer annotations (C. Wild, private communication). This work also

includes the use of RhoFit (Smart et al., 2014) to build models of fragments into PanDDA event maps in an automated manner and select the ligand pose for each crystal which best explains the electron density. While this has not been tested on all available data, a preliminary study on 10 targets showed that the autobuilding and scoring method generates ligand models within 1Å of the corresponding human-built models (C. Wild, private communication).

With an average fragment hit rate of 5%, proper reprocessing of the data from 81,000 crystals could result in over 4,000 high-quality bound structures. While some experiments may have failed and some results may be ambiguous, the bulk of the remaining results would be classed as negatives. If an appropriate automated pipeline could be developed, a dataset of more than 4,000 bound structures and many thousands of negative results would be an invaluable resource for the training of machine learning models for virtual fragment screening.

2.5 Conclusions

Given the data requirements for the training of machine learning models, our analysis of historic XChem data showed a need for better modelling of fragment structures and more rigorous validation of models. In order to ensure reliability of models, we would suggest that all model building and analysis takes place on the DLS server, and all models undergo validation using PanDDA 'giant score' to facilitate flagging of models with concerning validation scores. However, we also showed that simple classification using 'giant score' metrics would not be sufficient for automated labelling of hits. Other calculated features could help in the validation process, such as protein-ligand interactions for the evaluation of physicochemical feasibility of the modelled complex. By tracking all of these values and files in a robust, backed-up database, XChem could build a huge source of protein-fragment binding information for the whole community

to use.

To facilitate future analyses, we curated electron density maps for 81,000 crystals that had undergone automated processing but had not yet undergone PanDDA analysis or ligand modelling. By rerunning PanDDA on these maps and modelling ligands, we estimate that over 4,000 protein-fragment structures complexes could be found along with tens of thousands of negative results. The only practical way to build such a large number of models would be automated ligand modelling; trials of Conor Wild's automated ligand model building and scoring method have given promising results, with ligand models being within 1Å of high quality human-built models.

The development of an effective method for automated ligand modelling and validation requires a thorough understanding of what annotations are required for the various applications of such data, additional to the basic 'bound' or 'not bound' label. In order to make the data suitable for machine learning analysis, further annotations of the quality of the model would be advantageous: for example, a fragment may have been modelled into poor density but still be regarded as bound, in which case it may not necessarily be treated in the same way as a fragment that has been built into good supporting density. In other cases, only one atom may be in an uncertain position, while the rest of the fragment has strong density showing the atomic positions, and so just the uncertain atom needs to be treated with care. By adding this information, a stronger signal may be extracted from the data when attempting to train machine learning models.

For generalisable ML models to be trained, we would likely require hundreds of thousands of bound structures to be available. To get to this stage, collaboration between the groups that generate this data is crucial, as no one group would be able to generate this quantity of data on a suitable timescale. Additionally, to ensure that all data is produced in a similar enough way that it can be combined and models

trained on it, there must be open discussions between experimental and computational researchers.

As automated ligand modelling and validation is a long-term vision, we decided to only use the high-quality data that had been uploaded on the Fragalysis platform for shorter-term projects. We were able to use this set of protein-fragment structures to develop a new strategy for fragment library design. In the next chapter we describe our exploration of protein-ligand interactions in these structures and how we used them to identify the most functionally informative fragments in the Diamond-SGC-iNext Poised (DSiP) library.

Chapter 3

Fragment libraries designed to be functionally diverse recover information more efficiently from protein targets compared with traditional libraries

This chapter contains work described in the following publication: Carbery, A., Skyner, R., von Delft, F., & Deane, C. M. (2022). Fragment Libraries Designed to Be Functionally Diverse Recover Protein Binding Information More Efficiently Than Standard Structurally Diverse Libraries. *Journal of Medicinal Chemistry*, *65*(16), 11404–11413. The work described in this chapter is my own unless explicitly stated otherwise.

3.1 Preface

Current fragment-based drug design relies on the efficient exploration of chemical space by using structurally diverse libraries of small fragments. However, structurally dissimilar compounds can exploit the same interactions, and thus have similar functions in a protein-ligand complex.

To explore the structural features of traditionally-designed fragment libraries, we compare the composition of pharmacophore features in several commercially available libraries, and find that libraries tend toward similar profiles.

Using 3D structures of many fragments bound to multiple targets, we then examine if there exists a better strategy for selecting fragments for screening libraries. We show that structurally diverse fragments can be described as functionally redundant, often making the same interactions. Ranking fragments by the number of novel interactions they made, we show that functionally diverse selections of fragments substantially increase the amount of information (interactions made) recovered for unseen targets compared to other methods of selection.

Using these results, we design small functionally efficient libraries that are able to give significantly more information about new protein targets than similarly sized structurally diverse libraries. By covering more functional space, more diverse sets of drug leads can be generated.

3.2 Introduction

As described in Chapter 1, FBDD is now well-established as a powerful approach to early-stage drug discovery, and has led to success for targets that proved otherwise intractable (Erlanson et al., 2016). Ideally, a fragment screen obtains information about the molecules or functional groups that bind the protein of interest, and the interactions they make (Giordanetto et al., 2019). Fragments are subsequently elabo-

rated or combined to create larger lead molecules that can be developed into potential drugs. The better the site of interest is explored by fragments, the more insights we have into the key interactions critical for binding. As these key interactions are usually conserved upon generation of larger lead molecules (Chessari et al., 2021), this translates to improved chances of a viable drug candidate being discovered.

3.2.1 Design of fragment libraries

Maximising the useful information that can be extracted from each fragment screening experiment is key, and so design of fragment libraries is a major element of FBDD research (Hann et al., 2001; Congreve et al., 2003; Keserű et al., 2016; Ray et al., 2017; Revillo Imbernon et al., 2022). There are two major aspects to library design: definition of the desired region of chemical space, and the sampling of that region of chemical space. Table A.1 describes the design strategies and size of several major fragment libraries.

3.2.2 Definition of desired chemical space

Fragment libraries tend to be built from molecules that adhere to the 'rule of three' (Congreve et al., 2003): fragments that have a molecular weight below 300Da, fewer than three hydrogen-bond donors and acceptors, fewer than three rotatable bonds and a cLogP of three or below. Heavy atom counts tend to be limited to below 20 (Giordanetto et al., 2019). These rules aim to limit the structural complexity of the fragments so that only one or two efficient interactions with the protein target are required for binding. Compounds containing toxicophores or highly reactive groups are not included (Konteatis, 2021), as these would not be appropriate fragments to use in the development of drug leads. Fragment libraries also tend to prioritise chemical tractability (Cox et al., 2016b) and/or the availability of analogues (Wollenhaupt et al., 2020) to enable fast and easy follow-up experiments. This is different to tra-

ditional HTS libraries as even fragments that bind efficiently to the protein must be combined in order to increase their affinity enough for them to be suitable drugs, along with ensuring that other physicochemical properties are appropriate for use as therapeutics. Such fragments that are ideal for subsequent lead development have recently been referred to as 'social fragments' (St Denis et al., 2020), and are contrasted with 'unsocial fragments' which have limited or no synthetic pathways for elaboration, and no analogues.

Historical experimental results can be incorporated to further guide their definition of desired chemical space. The SpotXplorer library (Bajusz et al., 2021) was designed to contain pharmacophore features (see Section 3.2.5) that have been observed to commonly bind protein hotspots (Ngan et al., 2012), based on a comprehensive analysis of the PDB (Berman et al., 2000). A library designed recently (Bilsland et al., 2021) used past experimental results to develop a machine learning model that generated novel fragments that contain the characteristics of fragments that bind multiple targets, referred to as 'privileged fragments'.

Libraries may also have a particular intent, for example to target certain protein classes or to fulfil specific properties. Examples of such properties include high Fsp3 character (Aimon et al., 2018), 3D shape (Kidd et al., 2020), the ability to form covalent bonds with the target (Resnick et al., 2019), protein-protein interface binding character (Enamine, 2020b) or natural product resemblance (Aimon et al., 2018).

3.2.3 Sampling of desired chemical space

To avoid the synthetic challenges presented by the design of novel compounds, most libraries are made using previously available fragments, whether available commercially or in-house. In these cases, a catalogue of fragments that lie within the desired chemical space is generated, and fragments are selected from this in a way that maximises the structural or shape diversity of the library. A common approach is to

use molecular fingerprints such as ECFP (Rogers and Hahn, 2010), MACCS (Landrum, 2020) or Ultrafast Shape Recognition with Credo Atom Types (USRCAT) (Schreyer and Blundell, 2012) (which compares the shape and pharmacophoric types of the atoms of two molecules). A fingerprint is generated for each fragment, and a maximin-derived algorithm (Johnson et al., 1989) (such as the RDKit MaxMin picker) is used to select the most structure- or shape-diverse fragments. For example, the DSiP library (XChem, 2020a) (the successor to Cox et al. (2016b)) uses USRCAT fingerprints while the F2X libraries (Wollenhaupt et al., 2020) uses MACCS fingerprints to maximise structural and shape diversity.

Another method used to achieve structural diversity is to cluster fragments based on structure or functional groups. Representatives of each cluster can then be selected for the final library. An advantage of this approach is that clusters that cover more attractive chemical space can be sampled more often than those that are less desirable. The 3D shape diverse library (Enamine, 2020a) is an example of a library employing this strategy.

There are a few libraries which consist of novel fragments that were designed and synthesised specifically for the library (Hung et al., 2011; Ray et al., 2017). The most common aim of such libraries is to address the historic uneven coverage of chemical space. Final compounds for the library are generally selected to ensure synthetic feasibility, low similarity to commercially available fragments and to maximise shape diversity of the final library (Downes et al., 2020).

In nearly all cases, chemists are reported to be the final gatekeepers of selection, using visual inspection (Bilsland et al., 2021; Chen and Hubbard, 2009; Lau et al., 2011), which can introduce human biases. This indicates that algorithmic approaches are never trusted to completely select the fragments, and makes it difficult to learn from these final decisions, which are rarely fully documented or quantified.

3.2.4 Influence of HTS library design on fragment library design

To assess the effectiveness of current library design methods, it is useful to first understand how these strategies came about. The emphasis on structural diversity on library design appears to have its origin in the design of HTS libraries (Gordon and Kerwin, 1998), where the large chemical space made computable metrics imperative in the selection of compounds. Experimental approaches like diversity-oriented synthesis (Schreiber, 2009) also built up on this principle. The premises were transferred to fragment library design, but to our knowledge, the underlying assumptions were never rigorously interrogated.

Gordon *et al.* (Gordon and Kerwin, 1998) proposed that HTS screening libraries should be iteratively redesigned based on the results of screening campaigns. This has been applied to several libraries to guide their definition of 'attractive chemical space', however only knowledge of which fragments produced hits in past experiments was utilised, with 3D information from the protein-fragment structures being ignored. For example, an analysis of results from screens using the Astex 2012 fragment library (Hall et al., 2014) showed a need to focus on fragments with 10-14 heavy atoms and ensure that larger fragments are not overly complex. The Vernalis library was analysed using the results of 12 fragment screening campaigns (Chen and Hubbard, 2009) and found that compounds with slightly lower molecular weight had higher hit rates. AstraZeneca observed the high rate of project failure between 2002 and 2008 (Blomberg et al., 2009) and used the results to iteratively improve the library in several ways: fragments that were prone to decomposition, highly reactive or deemed 'unattractive' for follow-up chemistry by medicinal chemists were all removed, while pharmacophoric and structural diversity analyses were employed to 'fill in' gaps in chemical space.

3.2.5 Pharmacophore features

Pharmacophore features are commonly used as proxies for functional behaviour of a compound. Pharmacophores are an abscract description of the geometric arrangement of atoms or functional groups necessary to produce a given biological response (Seidel et al., 2020). An area of a molecule that is predicted to be a pharmacophore is referred to as a pharmacophore feature. The most important pharmacophore features are able to form hydrogen bonds, aromatic interactions, ionic interactions or hydrophobic interactions with the protein residues, and these each have many structural examples.

If a fragment library contains a biased set of pharmacophoric features, then the resulting protein-fragment complexes have the potential to exhibit the same bias. As this could have a negative effect on the lead development for the protein target, it is important to verify that the screening library does not contain obvious biases.

3.2.6 Functional activity of fragment libraries

As using only binary hit or miss results does not tell us whether these frequently hitting fragments are giving us diverse information about a target, this approach may fail in achieving the primary aim of a fragment screen of thoroughly exploring the binding site of a protein of interest. It is known that the molecular structure of a fragment does not accurately predict the interactions it is able to make with a protein (Guha, 2013): similar fragments may have diverse functional activity (e.g. they bind to different protein environments) while structurally diverse fragments may have very similar activity (Martin et al., 2002). Thus, the number of hits cannot be used as a proxy for quantity of information.

Most library designs aim to elucidate as much information as possible about a protein's binding site(s). However, so far it has been difficult to establish which or indeed whether any fragment libraries achieve this. Now that crystallographic

fragment screens are routine, we can use primary data of protein-fragment interactions to assess whether structurally diverse libraries are behaving in a functionally diverse manner.

3.2.7 Chapter outline

We begin by exploring the structural and pharmacophoric diversity of several commercially available libraries to verify that the DSiP library (XChem, 2020b) does not contain significant biases towards or away from particular pharmacophoric features or functional groups.

To establish whether fragment libraries designed to be structurally diverse have functional redundancy, we then examine a set of 10 diverse targets that have all been screened against the majority of fragments in the DSiP library. The same set of fragments have been tested on the same targets, generating full data of what bound and how, as well as which fragments did not bind. Using this data we describe an approach for analysing functional redundancy within a fragment library, and examine the relationship between structurally diverse and functionally diverse fragment libraries. Our findings suggest that structurally diverse fragment libraries do not necessarily exhibit any more functional diversity than randomly selected libraries. On the other hand, by selecting functionally diverse fragment libraries, we show that the information recovered for unseen targets is substantially improved compared with using randomly selected or structurally diverse fragments.

3.3 Methods

3.3.1 The Diamond-SGC iNext Poised fragment library

To ensure that any fragment hit discovered can be rapidly elaborated, a library of fragments poised for reaction was assembled in the form of the Diamond-Structural

Library	No. compounds	MW	cLogP
DSiP (Cox et al., 2016b)(Enamine, 2020a)	768	100-260	-1.5 - 4.5
Essential (Enamine, 2020a)	320	100-250	-1.5 - 2.5
3D Diverse (Enamine, 2020a)	1200	120-300	-1.5 - 3.0
High Fidelity (Enamine, 2020a)	1920	132-198	-1.0 - 2.5
F2X-Universal (Wollenhaupt et al., 2020)	1103	100-350	-3.0 - 6.5

Table 3.1: The fragment libraries used as comparators.

Genomics Consortium (SGC) Poised (DSP) Library (Cox et al., 2016b). The second generation of this library (DSPL2) was composed of 768 fragments which were poised for 11 categories of reaction (Figure A.1), and were determined to be diverse via calculation of USRCAT similarity, which compares the shape and pharmacophoric types of the atoms of two molecules (Schreyer and Blundell, 2012). The DSPL2 was used as a starting point for the construction of the DSiP library which could be supplied by Enamine (Enamine, 2020a) and removed redundant pharmacophore features and any toxicophores. The 768 fragments in the DSiP library have molar weights between 100 and 260, heavy atom counts between 8 and 19, and CLogP values between -1.5 and 4.5.

3.3.2 Comparison of structures in fragment libraries

To explore the structural diversity of the library, we compared structural and pharmacophoric features of the DSiP library with the F2X-Universal fragment library (Wollenhaupt et al., 2020) and several non-targeted fragment libraries that are available from Enamine. Details of all the compounds in the libraries available from Enamine were downloaded from the relevant web pages (Enamine, 2020a), while the F2X-Universal library compound list was available in the supplementary information as SMILES strings (Weininger, 1988). Summaries of all libraries referred to can be found in Table 3.1.

As pharmacophore features are commonly used as a structural proxy that is predictive of functional activity, we compared the representation of different pharma-

Family	**Pharmacophore feature type**	**Descriptors**
Hydrophobic	Aromatics	Arom5, Arom6
	Rings	RH3_3, RH4_4, RH5_5, RH6_6
	End groups	tButyl, iPropyl
	Attaching groups	ChainTwoWayAttach, ThreeWayAttach
Hydrogen-bonding	Electron donors	SingleAtomDonor
	Electron acceptors	SingleAtomAcceptor
Ionisable	Positive ionisable	BasicGroup, Guanidine, Imidazole, PosN
	Negative ionisable	AcidicGroup
Zinc binding	Zinc binding	ZnBinder4, ZnBinder5

Table 3.2: The pharmacophore feature descriptors used in the comparison of fragment libraries. The descriptors are as defined by RDKit (Landrum, 2020).

cophore features in each library in Table 3.1. To generate the pharmacophore feature descriptors for each ligand, we used RDKit (Landrum, 2020). A list of the descriptors used in this analysis and their families is provided in Table 3.2.

To additionally compare the structural diversity of the fragment libraries, we calculated USRCAT (Schreyer and Blundell, 2012) scores for each library. USRCAT is an algorithm which compares shape and pharmacophoric information from the input molecules, and is built-in to RDKit.

For each molecule, a USRCAT moment was calculated. Within each library, a score between 0 and 1 is calculated for every possible pair of compounds. A score of 1 indicates that the molecules compared are identical, while a score of close to 0 indicates that the molecules are structurally very different and it is inferred that they have different capacities for forming interatomic interactions.

3.3.3 The XChem Dataset

For exploration of the functional activity of the DSiP library, structures of 309 protein-fragment complexes were used, from 10 targets and 225 unique fragments, representing results from 4928 individual crystals. Each target had between 13 and 65 fragment-bound structures. The targets were diverse, with a global pairwise sequence identity mean of 12% and maximum of 27%. Sequence identity was calculated using EMBOSS Needle (Needleman and Wunsch, 1970). No two targets shared a CATH class (Sillitoe et al., 2020). Table A.3 contains descriptions of the targets used, and the fragment-bound structures of these are available on the Fragalysis platform (Skyner and von Delft, 2020).

To ensure that our analyses were not biased by fragments only observed a few times, we selected those fragments which had been tested on at least 7 of our 10 targets. This resulted in 520 fragments being included, of which 295 had not bound to any targets. The remaining 225 fragments had bound to between 1 and 4 targets. We selected fragments that had been tested on at least 7 targets as a balance between coverage (7 out of 10) and dataset size. Requiring fragments to have been tested on 8 or more of our targets would have led to the inclusion of only 480 fragments in the dataset, of which only 214 had bound one or more targets. The fragments used in our analysis are included within Appendix A.

3.3.4 Selection of functionally diverse fragments

Definition of functional activity We define the functional activity of a fragment as the interactions it makes with the protein. This definition is used to prioritise our understanding of a fragment's ability to make interactions rather than the structure of the fragment itself. The interactions in each structure within the dataset were calculated using ODDT's InteractionFingerprint module (Wójcikowski et al., 2015), which generates a binary fingerprint for each protein-fragment structure. This

method, referred to in this chapter as 'residue IFP', calculates up to eight types of interaction between the fragment and each residue in the protein (see Table A.2 for details of interaction types). We also adapted the InteractionFingerprint module to output the interactions between the fragment and each atom of the protein, resulting in what we will refer to as the 'atomic IFP' method.

Ranking of fragments based on novelty of functional activity The ranking protocol aims to identify fragments that add the most information about interactions a target can make. We start with a library size of one and append fragments one by one, each time including the fragment that adds the largest amount of novel information to the current library. As several fragments may add identical amounts of information, we repeat this 100 times, randomly shuffling the order of the fragment list before each run. At each library size, we calculate the number of interactions recovered compared to all the interactions from the full screen. The mean fractional recovery rate at each library size is calculated, along with the standard deviation. The code used to rank fragments is available at https://github.com/oxpig/fragment-ranking.

Other methods of fragment ordering for comparison For comparison, we used two other methods of fragment ordering, including only fragments that had bound one or more targets, to match the set of fragments that were possible to rank. The first of these was a random control, where we shuffled the fragments into a random order, repeating this 100 times and calculating the mean and standard deviation of the recovery rate at each library size. We also generated a structurally diverse control, using the technique employed in the selection of the F2X-Entry library (Wollenhaupt et al., 2020): we used RDKit (Landrum, 2020) to calculate the MACCS key for each fragment, and MaxMin picker to select a structurally diverse set of fragments for every library size tested. This was also repeated 100 times with the mean recovery rate and standard deviation taken. This strategy for fragment selection was compared

with other fingerprinting methods, shown in Figure A.2.

3.3.5 Testing fragment ranking protocol on unseen targets

To test the effectiveness of ranking fragments by their functional information, we tested the protocol on each target in the dataset, using a leave one out test.

For each target, we ignored the results of its own screen, ranked the fragments using the other targets, and calculated the recovery of information on this previously ignored target. This was run 100 times, and the mean amount of information recovered was calculated, along with the standard deviation. These values were also calculated for randomly ordered and structurally diverse fragment libraries.

Only fragments that had bound to previously seen targets could be ranked, so once this library size was exceeded, 'dummy fragments' were used, which did not recover any new information from the protein. This resulted in 100 sets of fragment libraries at every size from 1 to 200 for each of the library types. For each method, the mean information recovered was calculated, along with the standard deviation of this information recovery.

3.4 Results

3.4.1 Comparison of fragment libraries

To analyse the representation of important pharmacopohore features within the DSiP library, we compared it with the other libraries summarised in Table 3.1. From the summary information we can see that it is not outlying in terms of the size of the molecules or the cLogP values.

Pharmacophore feature representation in fragment libraries Using RDKit's ligand-based pharmacophore feature generator (Landrum, 2020), we were able to

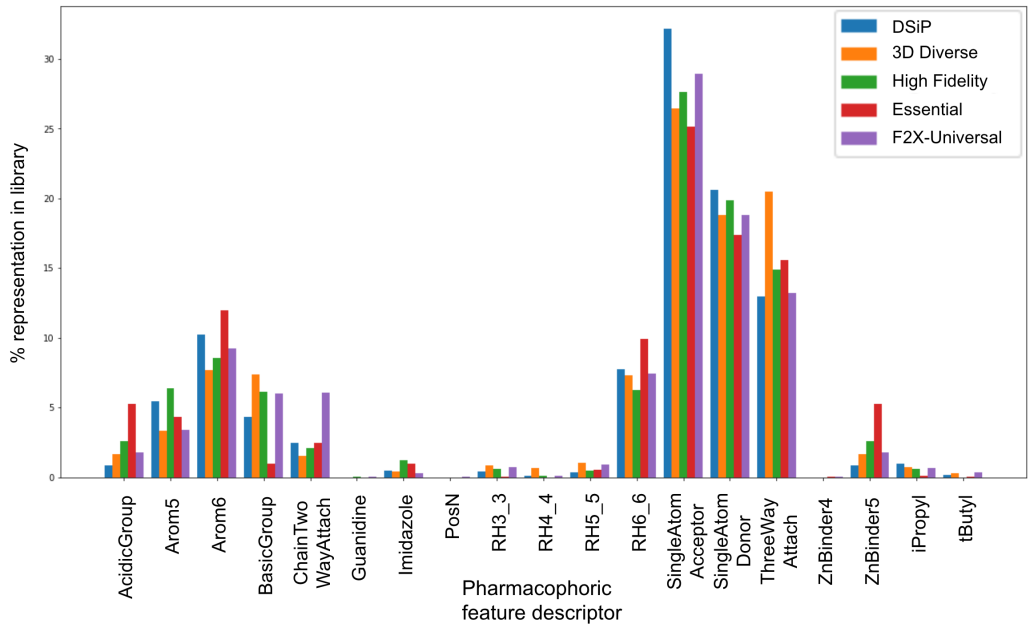


Figure 3.1: Pharmacophore feature representation in commercially-available fragment libraries. The height of the bars is the proportion of total pharmacophore features in the library that are of this type, as a percentage. Pharmacophore features are listed in Supplementary Table 3.2.

examine the representation of different pharmacophore features types in the different libraries. This is shown in Figure 3.1. The overall distribution of different features is very similar between different libraries.

However, there are a few small differences, for example in the over-representation of the 'chain two way attach' feature in the F2X-Universal library and the 'three way attach' feature in the 3D diverse library. These higher levels of the two- and three-way attachments in the F2X-Universal and 3D diverse libraries respectively are expected, as these libraries consist of larger, more complex fragments. The DSiP library has a higher level of SingleAtomAcceptor features compared with other libraries. All libraries have a higher predicted potential to make hydrogen-bonds than hydrophobic interactions (see Table 3.2), of which the DSiP has the greatest disparity (Figure 3.2).

USRCAT analysis of fragment libraries USRCAT moments were calculated for every compound in each library, and within every library, each compound's moment

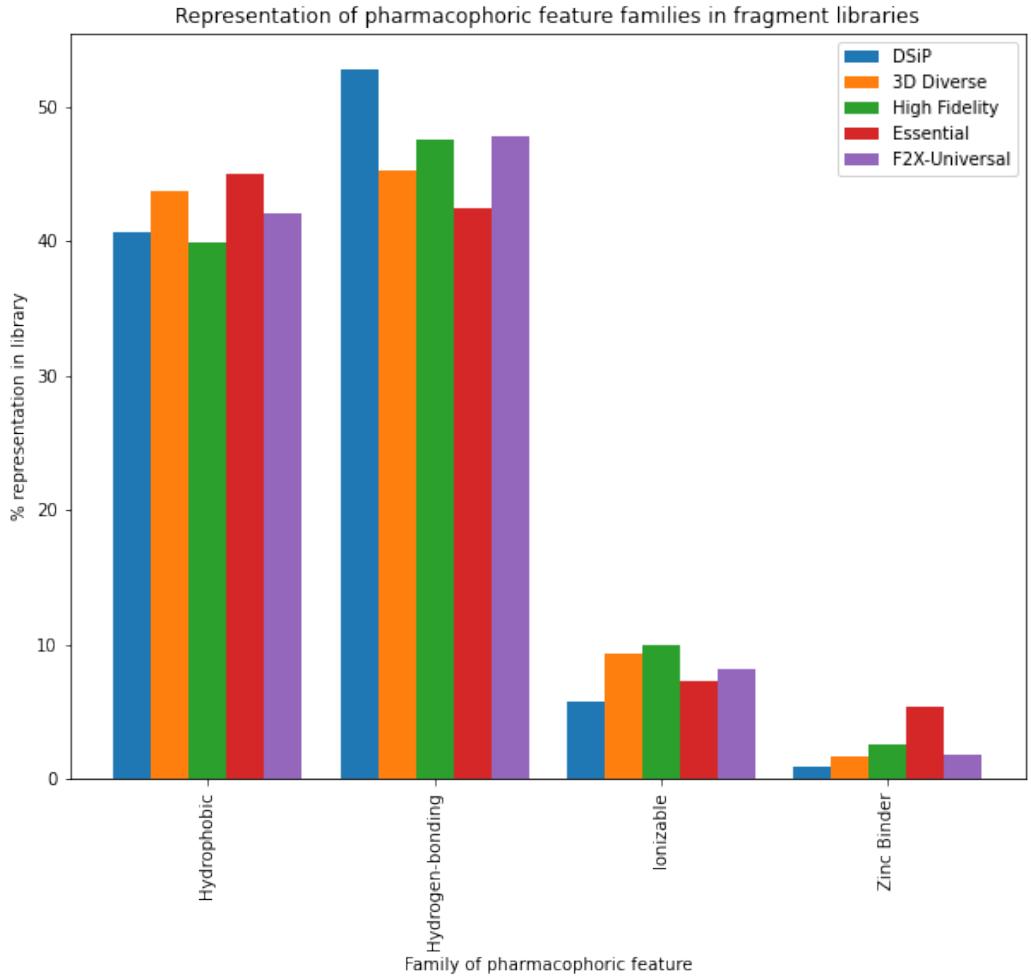


Figure 3.2: Representation of families of pharmacophore features in fragment libraries detailed in Table 3.1. The height of the bars is the proportion of total pharmacophore feature families in the library that are of this type, as a percentage. All libraries have a higher proportion of hydrogen-bond forming pharmacophore features than hydrophobic interaction forming features.

was compared to all others'. The output was a similarity score between 0 and 1, with a score of 0 meaning no shape or pharmacophoric similarity, while a score of 1 indicates that the molecules being compared are identical. Examples of the scores obtained when comparing one molecule with various others are shown in Figure 3.3.

The distribution of USRCAT scores for each library is shown in Figure 3.4, with the total number of scores normalised to that of the largest library. While all libraries are distributed similarly, the DSiP, 3D diverse and F2X-Universal libraries

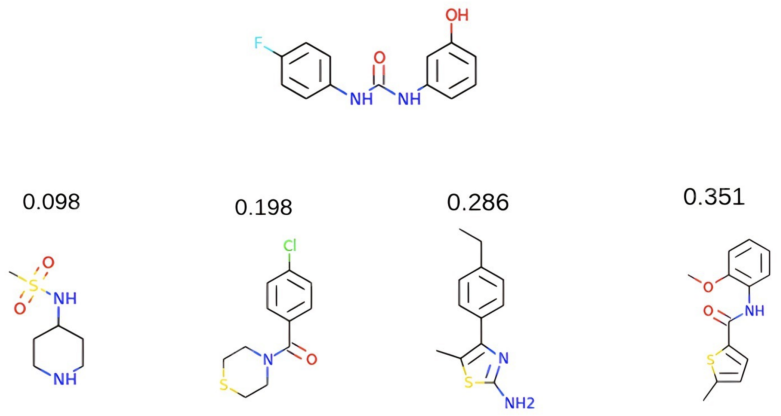


Figure 3.3: Examples of USRCAT scores obtained from comparing the reference molecule (above) with each of the four molecules (below).

have fewer molecules with USRCAT scores between 0.2 and 0.4. This shows that they contain slightly more diverse molecules than the High Fidelity and Essential fragment libraries.

Based on our analysis of the DSiP library, we would not expect the DSiP library to show results biased towards particular functional groups, pharmacophore features or interaction types compared with other available fragment libraries. In the future this could be tested by comparing the proportions of different types of interactions made with protein targets and the proportions of different pharmacophore feature types in a fragment library.

3.4.2 Fragment ranking

To explore the functional activity of the DSiP library, structural data from fragment screens of 10 unrelated protein targets bound to 520 fragments were used in this study. Protein-ligand IFPs were calculated for each structure, between fragment atoms and protein residues (residue IFP) and between fragment atoms and protein atoms (atomic IFP). For both types of IFP the fragments were ranked based on the novel interactions they made with all or a subsection of protein targets (see Section 3.3). These rankings were used to group fragments for analysis, defined in Table 3.3.

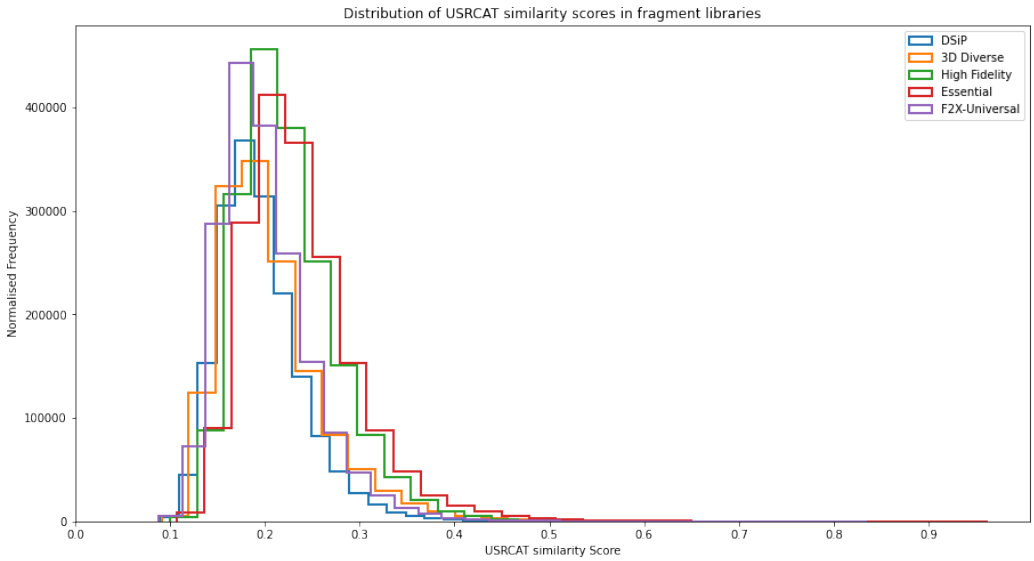


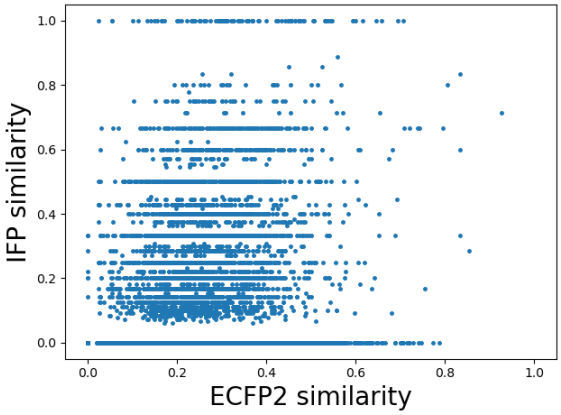
Figure 3.4: The DSiP, 3D diverse and F2X-Universal libraries are slightly more diverse than the High Fidelity and Essential libraries. The total numbers of scores in the distribution were normalised to that of the largest library to ease comparison. Two main clusters can be seen with the DSiP, 3D Diverse and F2X-Universal libraries exhibiting higher levels of diversity compared with the High Fidelity and Essential libraries.

We also examined the types and frequencies of different protein-ligand interactions made, shown in Figures A.3, A.4 and A.5.

3.4.3 Structurally diverse fragments can make overlapping interactions

We compared the molecular similarity (calculated using ECFP2 fingerprints (Rogers and Hahn, 2010)) to the residue IFP (our measure of functional activity) similarity for fragments bound to 10 highly diverse protein targets (Table A.3). Fragments that are structurally very different (ECFP2 similarities as low as 0.02) bound to the same location on a target and made one or more of the same interactions, leading to an IFP similarity above zero (Figure 3.5a). While a single interaction in common may not be indicative of functional similarity in many experimental contexts, for the purposes of this analysis we viewed this as containing some level of redundancy. Across the set we found 44 pairs of structurally distinct fragments that formed identical

(a) Molecular (ECFP2) similarity compared to functional (IFP) similarity



(b) Three distinct fragments making identical interactions

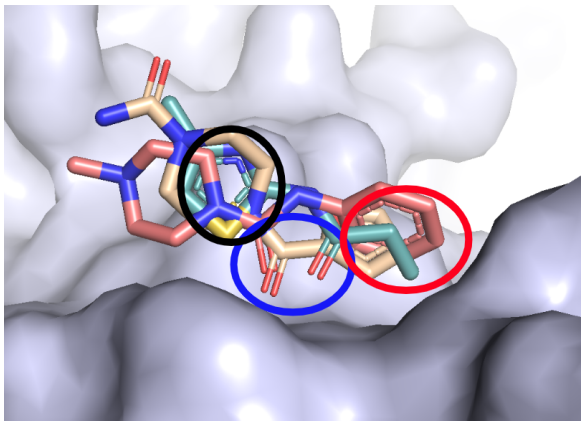


Figure 3.5: Very diverse fragments make identical interactions. **(a):** Molecular (ECFP2) similarity compared to functional (IFP) similarity. Each point represents a pair of fragments that bind to the same target. There is no direct correlation between ECFP2 (molecular) similarity and IFP (functional) similarity, showing that structural diversity of fragments is not predictive of functional diversity. **(b):** An example of three fragments which make the same interactions with target TBXTA (lilac). The atoms circled in bright blue (carbonyl oxygens) are making hydrogen bonds with two residues; atoms circled in black (nitrogens) are making a hydrogen bond with a single residue; atom groups circled in red are making a hydrophobic interaction with a single residue.

interactions. An example of this structural dissimilarity but functional similarity is shown in Figure 3.5b, where three distinct fragments form identical interactions with the protein. The molecular fingerprint (ECFP2) similarity of these three fragments ranges between 0.27 and 0.34, which would be considered appropriately diverse for inclusion in conventional libraries.

While the 44 pairs of fragments that form identical interactions only accounted for 0.82% of all fragment pairs, over a quarter (26.9%) of fragment pairs shared at least one common interaction. To explore the possibility that fewer fragments could make the same interactions with the targets, we assessed which fragments made the most novel interactions, and calculated the minimum number of fragments required to explore all interactions across all targets.

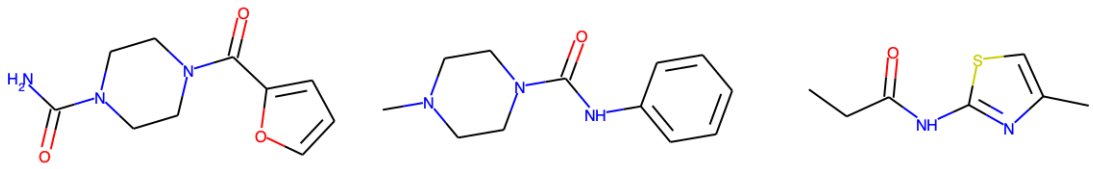


Figure 3.6: Structures of the three fragments shown in Figure 3.5(b).

Group of fragments	Definition
Top 100	The 100 most informative fragments, as determined using the fragment ranking methods. Also can be described as the most functionally diverse selection of fragments
Remaining minimum	The fragments not in the top 100, but that still make novel interactions, and are thus within the minimum number of fragments to make all interactions from the original screen
Remaining bound	The fragments not in the top 100, but that have bound one or more protein targets
Redundant	Fragments that have bound to one or more protein targets yet do not make any novel interactions
Never bound	Fragments that have never been observed to bind a protein target
Structurally diverse	Sets of fragments that have been selected to be functionally diverse using MACCS similarity
Random	Sets of randomly selected fragments

Table 3.3: Definitions for the various groups of fragments used in this analysis.

3.4.4 Ranking of fragments reveals redundancy in interactions

Fragments were ranked by the number of novel interactions (residue-level and atomic-level) that they made with the 10 targets in the dataset (Table A.3). The highest ranked fragment was that which made the most interactions across all targets, and from there fragments were selected by the number of interactions that they made across all targets that had not already been made by higher ranked fragments (termed as novel interactions). Two libraries, one randomly ordered and another structurally diverse, were generated for comparison to the functionally diverse libraries (see Sec-

tion 3.3 for details of how these were prepared). Each of these three methods of ordering fragments was repeated 100 times, and the mean fraction of unique interactions (also referred to as 'information') recovered at each library size across all runs was calculated (Figure 3.7).

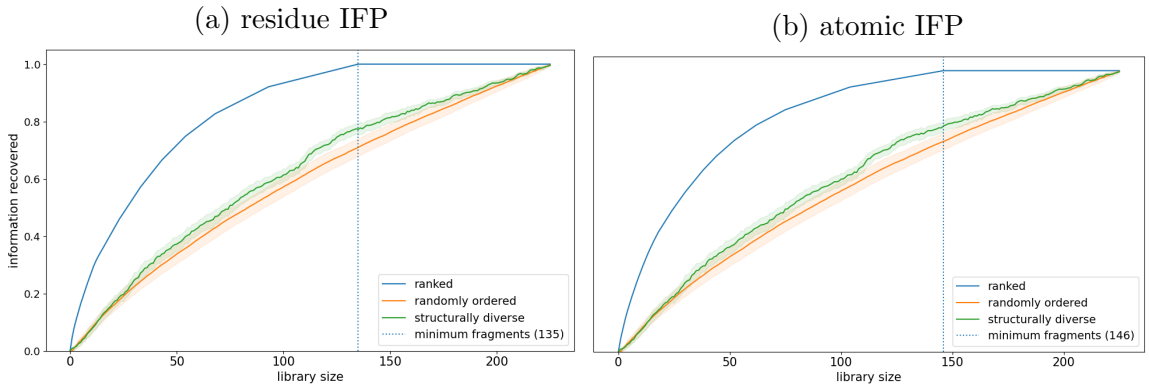


Figure 3.7: Ranked fragment libraries recover information at smaller library sizes than other methods. Blue dotted line indicates minimum number of fragments required to recover all information. **(a):** fragments have been ranked using the residue IFP method. **(b):** fragments have been ranked using the atomic IFP method.

Interactions were recovered at smaller library sizes for the functionally ranked library compared with either the random or the structurally diverse library. On a residue level, all interactions with targets were recovered using only the 135 top-ranked fragments of the 225 that were bound to at least one target (Figure 3.7a); for atomic level of interactions, 146 fragments were required (Figure 3.7b). This result shows that by selecting the fragments in this way, a 135-fragment screen could recover the same information as randomly selected 520-fragment screens on all 10 targets. It was expected that the ranked libraries (using both residue and atomic IFP) would recover information at smaller library sizes due to the method of ranking, however it is notable that structurally diverse libraries do not recover information at library sizes any smaller than the randomly selected libraries. The lack of difference between structurally diverse and random libraries may be because the fragments used are already relatively structurally diverse.

To assess whether different libraries were selecting similar fragments, the mean number of fragments in common between libraries of 100 fragments was calculated. The functionally diverse and the structurally diverse libraries had 49 fragments in common, similar to the 44 fragments in common between the functionally diverse and randomly selected fragments. This indicates that the level of similarity between the functionally diverse and structurally diverse libraries is little more than random. Conversely, the functionally diverse libraries generated using the residue IFP and atomic IFP methods have 83 fragments in common.

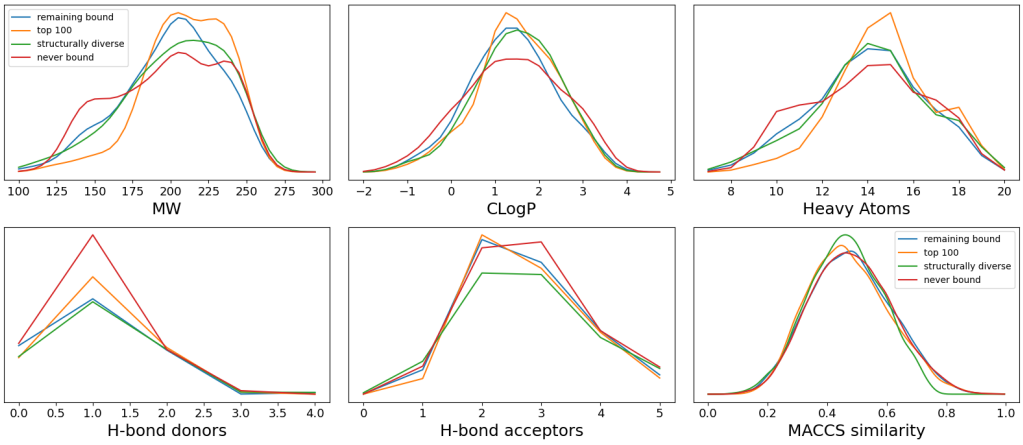
The rankings assigned to each fragment in our analysis are shown in Appendix A.

3.4.5 Functionally diverse compounds exhibit different chemical properties to non-binding fragments

To explore the relationship between particular chemical properties and a fragment's rank, we split the DSiP fragment library into three sets: the top-ranked 100 fragments (those that are most informative), the 125 remaining bound fragments (fragments that had bound to one or more proteins), and the 295 fragments that had never bound to a target (Figure 3.8). This was performed on sets generated by both residue and atomic IFP methods, and as a comparison, chemical properties were also calculated for a structurally diverse set of 100 fragments.

Fragments ranked in the top 100 functionally diverse were more likely to have a molecular weight (MW) around 200 for both residue and atomic IFP methods, while the 'other bound' groups of fragments were slightly more likely to have a MW under 175. For all other properties, there was no substantial difference between groups of fragments that had bound one or more targets. Fragments that had not been observed to bind any target were much more likely to have a low MW (under 175) and a higher heavy atom count (17 or above) than those that had bound one or more targets. The low MW of many 'never bound' fragments suggests that they may be too small for

(a) residue IFP



(b) atomic IFP (below)

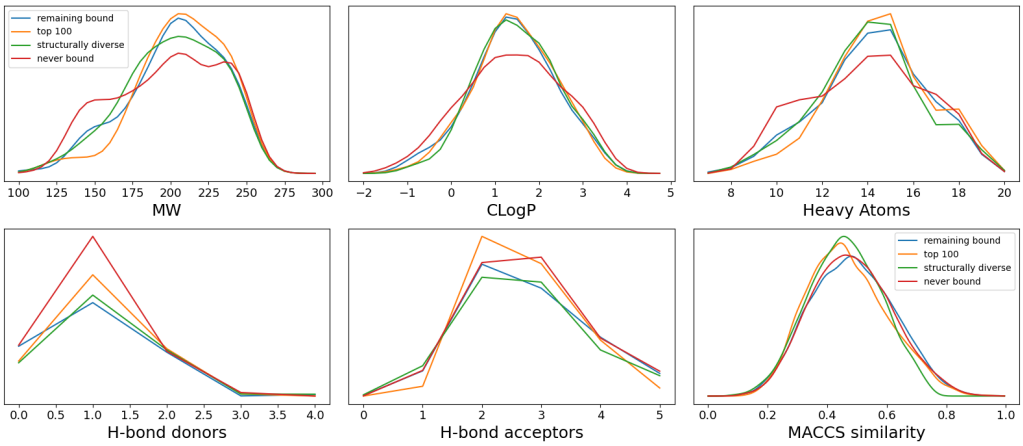


Figure 3.8: Fragments that have never bound to a target (red) are more likely to have very small or very large MW and heavy atom counts. Fragments have been categorised into the 'top 100' and 'remaining bound' groups based on our functional ranking (see Section 3.3). Fragments that bound no targets make up the 'never bound' set, while a set of structurally diverse fragments are shown for comparison. Various properties of these sets of fragments are compared. **(a):** fragments have been ranked using the residue IFP method. **(b):** fragments have been ranked using the atomic IFP method. **(a) and (b):** Medium-sized fragments are most likely to be highly informative.

reliable detection and modelling in many experiments. As expected, the fragments selected as a diverse subset had lower MACCS similarity with each other compared with the overall library and the highly-ranked fragments.

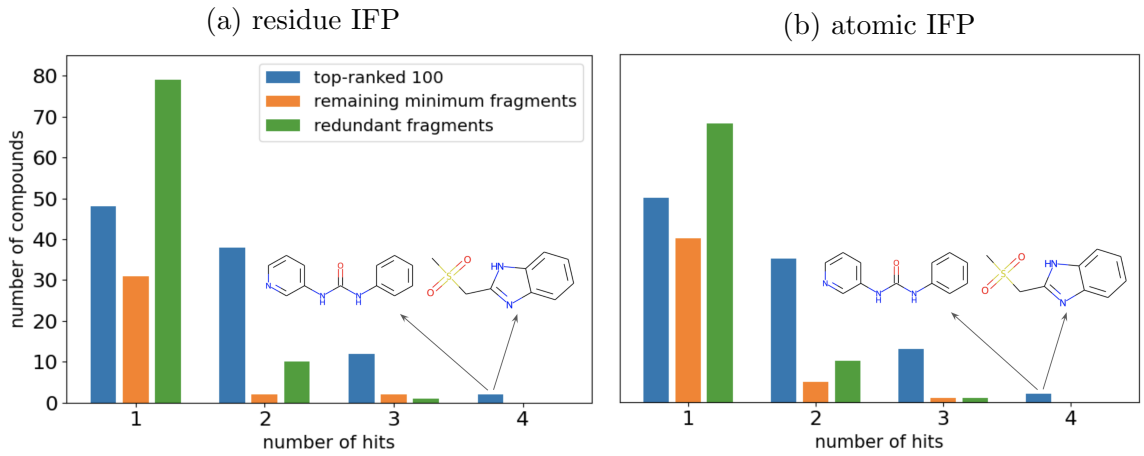


Figure 3.9: The number of targets bound by fragments in the DSiP library, grouped by their position when ranked. The top-ranked 100 fragments are compared to those remaining fragments that give us novel interactions, and to fragments that make only redundant interactions. Promiscuous fragments (e.g. 3 or 4 hits) are not always selected as the most informative, whereas some fragments with only 1 hit are ranked in the top 100. Chemical structures of the two unique fragments that bound 4 targets are pictured within the figures. **(a):** fragments ranked using the 'residue IFP' method. **(b):** fragments ranked using the 'atomic IFP' method.

3.4.6 Promiscuous binders are not necessarily the most informative fragments

To examine whether a fragment that bound multiple targets (a promiscuous fragment) was likely to be selected as a highly informative fragment, we compared the number of hits (targets bound) for three sets of fragments: the 100 top-ranked fragments, the remaining minimum fragments (fragments that would be required to recover all information from the original screen), and the fragments that could be removed without losing any information (redundant fragments). While the top-ranked sets were made up mostly of fragments that had bound multiple targets, some fragments that bound to 3-4 targets were excluded in favour of those that had only bound to a single target (Figure 3.9).

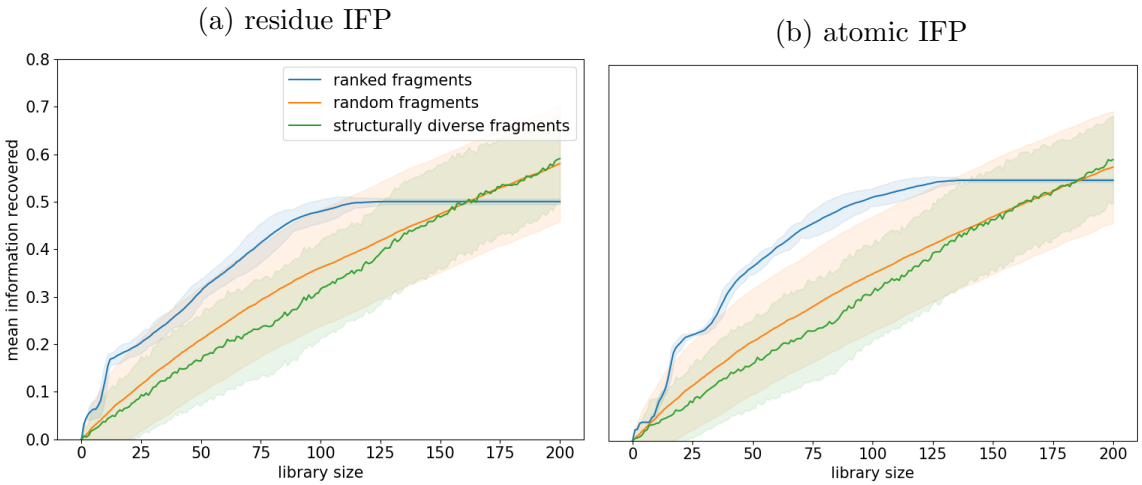


Figure 3.10: Ranked fragment libraries show superior information recovery for unseen targets at every library size. The recovery of information for each target when unseen was calculated 100 times. The mean across each run for each target was calculated, and the mean of these values was taken. This value is shown at each library size, with error clouds showing 1 standard deviation across 100 runs. **(a):** fragments have been ranked using the 'residue IFP' method. **(b):** fragments have been ranked using the 'atomic IFP' method.

3.4.7 Functionally diverse fragments recover information more efficiently from unseen targets

The results above show that a functionally diverse fragment set contains different fragments from a structurally diverse one. We investigated whether using such functionally diverse fragments is an effective strategy to more efficiently get information about unseen targets. To do this, we performed a leave one out test (see Section 3.3 for more details).

To compare the information recovery (in our analysis defined as the fraction of interactions recovered) for each target when using functionally diverse fragments with the other methods of fragment selection, we analysed the information recovered from each target at a library size of 100 fragments (Figure 3.11). We also calculated the fractional improvement when using the functionally diverse fragments compared with random and structurally diverse fragments (Figure A.6). Upon visual inspection, it

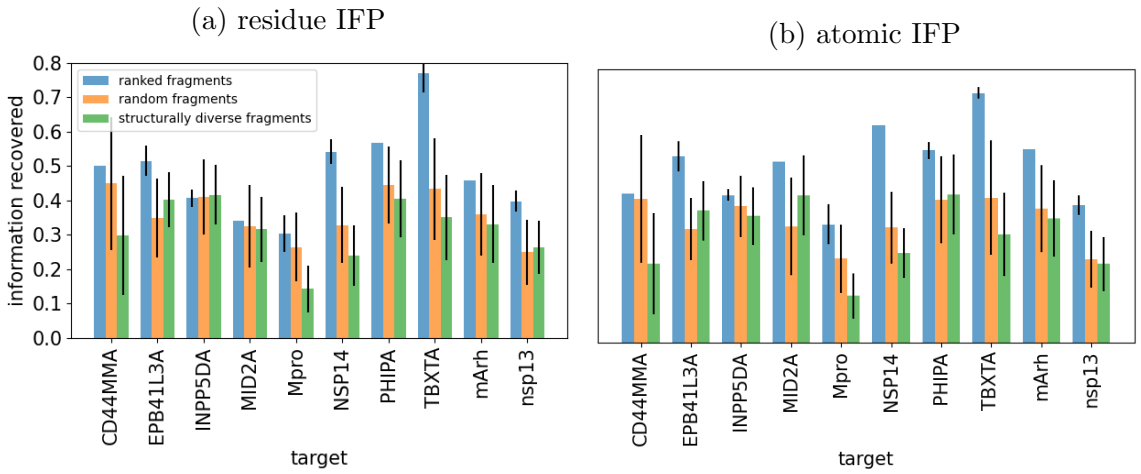


Figure 3.11: Top-ranked fragments show superior information recovery compared with random fragments and structurally diverse fragments. For each target, average information recovery when using the top-ranked 100 fragments over 100 runs is shown. Error bars show 1 standard deviation. Lack of error bars show no variability in the result. **(a):** fragments have been ranked using the 'residue IFP' method. **(b):** fragments have been ranked using the 'atomic IFP' method.

was apparent that the top-ranked fragments created a more uniform coverage over the binding site than other methods of library selection.

We compared the methods of fragment selection across all unseen targets. The mean values of information recovery across all targets are shown in Figure 3.10b. On average, the functional information about the unseen target was recovered more efficiently using the functionally diverse fragments than the random libraries.

3.4.8 Targets contribute differently to prediction of important fragments for unseen targets

The targets in this dataset are very diverse (the maximum pairwise global sequence identity is 27%), but some fragments do make similar interactions with different targets. To assess the impact of each target on the effectiveness of a fragment set for giving information about an unseen target, we removed the previously seen targets one by one, ranked only on the remaining 9 targets, and took the top-ranking 100

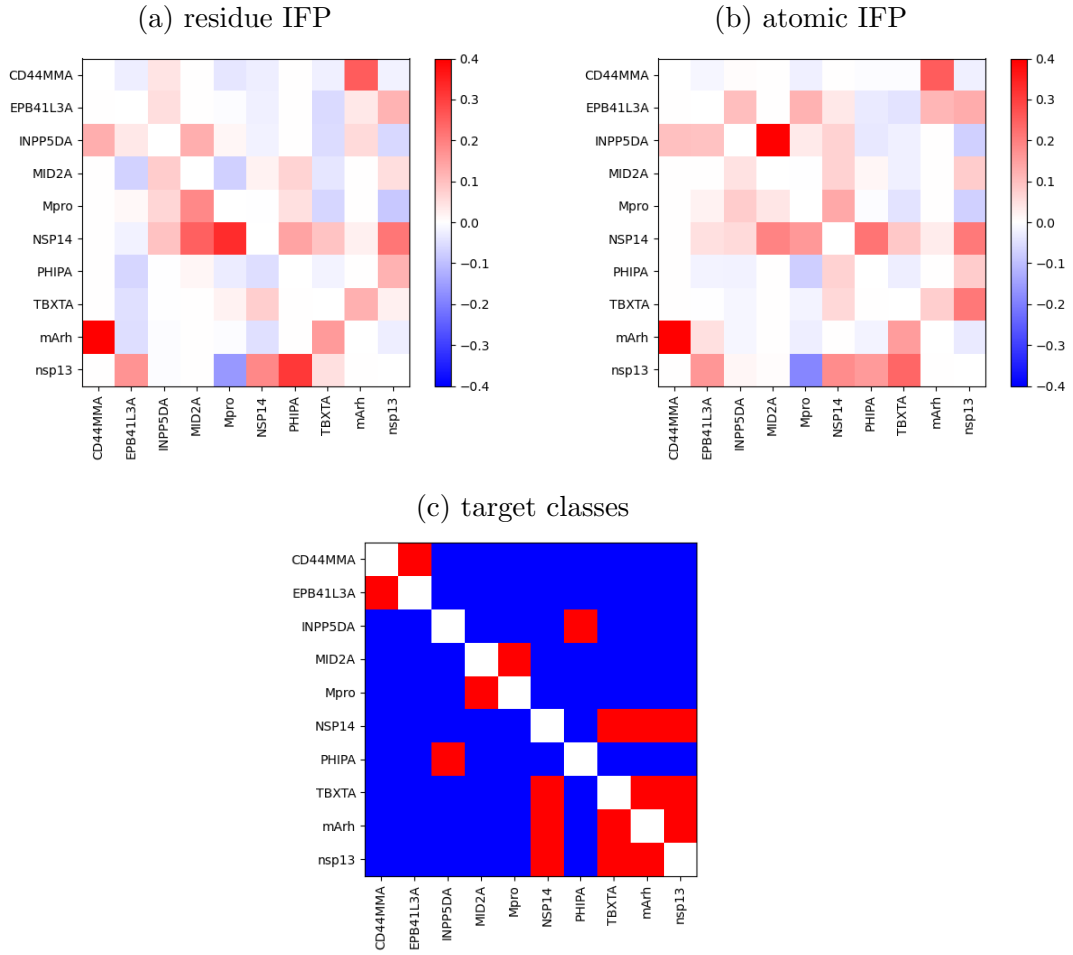


Figure 3.12: The impact of each target on every other target's information recovery at library size of 100 fragments. For example, including mArh improves recovery of information on CD44MMA by 40% compared with only using the remaining 8 targets to rank the fragments; including CD44MMA improves recovery of information on mArh by 27%. **(a):** fragments have been ranked using the 'residue IFP' method. **(b):** fragments have been ranked using the 'atomic IFP' method. **(c):** A heatmap showing whether two targets fall into the same class. Red indicates that two targets are in the same class and blue indicates a different class.

fragments as the functionally diverse library. The recovery of interactions was compared with the original (when results of all 10 targets were used to rank), and the factor that each target impacts the recovery of interactions for every other target was calculated. These impact scores are a measure of the similarity of the most informative fragments between two targets (Figure 3.12a-b). Between the residue IFP method and the atomic IFP method, the scores are consistent in terms of overall effect, but the magnitudes differ. Each target positively impacts some targets while negatively impacting others.

As a high impact score is indicative of similar fragments making interactions with the targets, we assessed which targets were in the same classes (phosphatase/kinase, protease, nucleic or other). This is shown in Figure 3.12c, however it does not appear that there is a strong correlation between these two factors. This confirms that there are no two targets with similar fragment binding activity, and yet by ranking fragments on previously seen targets, we can improve information recovery on unseen unrelated targets.

3.5 Conclusions

We began by comparing the representation of pharmacophore features in the DSiP with other commercially available libraries. Based on our findings, we would not expect the DSiP library to show results biased towards particular functional groups, pharmacophore features or interaction types compared with other available fragment libraries.

In agreement with previous work (Martin et al., 2002), we then found that structurally diverse fragments are able to make similar or even identical interactions. Additionally, we showed that by defining fragments as the interactions they make with all targets, fragment libraries can be selected based on their functional diversity. Such

libraries were able to make more diverse interactions with previously unseen targets, and thus improve the information recovered by an average of 68% and a maximum of 152% across all targets tested, compared with traditionally designed fragment libraries. The atomic IFP method gave slightly more consistent results than the residue IFP method, suggesting that the particular atom within a residue that a fragment interacts with has importance.

We also determined that different targets contribute differently to prediction of important fragments for unseen targets: each target positively impacts some targets while negatively impacting others. Additionally, some targets are mostly negatively impacted by others at a residue level while being unaffected or positively affected at an atomic level. We found that there was no correlation between these effects and protein classes, suggesting that a simple predictor of this behaviour is unlikely to exist.

Currently, the most common strategy for fragment library design is to select the most structurally diverse set of fragments from those that lie in the desired chemical space, without considering structural results from previous fragment screens. In this study we have shown that libraries designed on the basis of functional diversity recover information more efficiently from unseen targets than traditionally designed structurally diverse libraries. Even with a limited dataset, we have proven the potential for functionally diverse fragment selections to substantially improve the information recovered from fragment screens. Every time more experimental data becomes available, it would be possible to quickly re-select functionally diverse fragments that can reliably and significantly outperform traditional library selection methods.

This ability to better explore the interactions in protein binding sites would allow a larger number of diverse lead compounds to be developed, thus improving the chances of a fragment screening campaign producing a viable drug candidate. Additionally, we believe that by collecting comprehensive information on the functional activity of

fragments (by testing them on many targets), we can begin to understand and predict which fragments will be functionally diverse just based on their structure.

Another approach to increasing the information yielded from a fragment screen is to construct the fragment library in a target-specific manner. Historically, this has been achieved using knowledge of ligands that bind to proteins with a similar structure or function. Following the development of AlphaFold and other protein structure prediction tools, we now have the opportunity to extract information from a target's structure prior to experimental structure determination. By understanding the target's binding sites, we can predict which fragments are most likely to bind them. In the next chapter, we develop a binding site prediction method which is able to outperform competitors on protein structures that have been predicted by AlphaFold.

Chapter 4

Learnt representations of proteins can be used for accurate prediction of small molecule binding sites on experimentally determined and predicted protein structures

This chapter contains work described in the following publication: Carbery, A., Buttenschoen, M., Skyner, R., von Delft, F., & Deane, C. M. (2024). Learnt representations of proteins can be used for accurate prediction of small molecule binding sites on experimentally determined and predicted protein structures. *Journal of Cheminformatics*, *16*(1), 32. The work described in this chapter is my own unless explicitly stated otherwise.

4.1 Preface

Protein-ligand binding site prediction is a useful tool for understanding the functional behaviour and potential drug-target interactions of a novel protein of interest (see Chapter 1). However, most binding site prediction methods are tested by providing crystallised ligand-bound (holo) structures as input. This testing regime is insufficient to understand the performance on novel protein targets where experimental structures are not available. An alternative option is to provide computationally predicted protein structures, but this is not commonly tested. However, due to the training data used, computationally-predicted protein structures tend to be extremely accurate, and are often biased toward a holo conformation.

In this study we describe and benchmark IF-SitePred, a protein-ligand binding site prediction method which is based on the labelling of ESM-IF1 protein language model embeddings combined with point cloud annotation and clustering. We show that not only is IF-SitePred competitive with state-of-the-art methods when predicting binding sites on experimental structures, but it performs better on proxies for novel proteins where low accuracy has been simulated by molecular dynamics. Finally, IF-SitePred outperforms other methods if ensembles of predicted protein structures are generated.

4.2 Introduction

A key part of early-stage drug development is building a thorough understanding of the protein target of interest. Identification of potential ligand-binding sites facilitates a host of techniques, such as hit identification, small molecule screening, functional prediction, off-target binding prediction and binding site comparison (Pérot et al., 2010; Özçelik et al., 2022). Many methods have been developed to locate ligand-binding pockets on protein structures. Originally, these methods were designed for use on experimentally determined structures, but the development of AlphaFold (Jumper

et al., 2021) and other accurate protein structure prediction tools (Lin et al., 2022; Baek et al., 2021; Wu et al., 2022) now allows the exploration of the three-dimensional features of the protein, including the prediction or identification of the ligand binding site, from structural models.

4.2.1 Strategies for protein binding site prediction

Various strategies exist for protein binding site prediction. Where the protein of interest has close homologues that have already been crystallised with ligands, the protein binding site can be inferred based on the alignment to these complexes (McGreig et al., 2022; Gao et al., 2016; Lee and Im, 2013). However, if no homologous structural data is available, predictions must be made in other ways. Sequence-based methods use the amino acid sequence of the protein which is sometimes enriched with predicted structural features such as secondary structure, solvent accessibility or hydrophobicity. However, these methods achieve lower success compared with structurally-informed methods, and so they tend to be used for the prediction of specific ligand binding sites, such as carbohydrates (Taherzadeh et al., 2016). Structurally-informed methods include geometry-based (Le Guilloux et al., 2009) and probe-based (Ngan et al., 2012) methods, which use the shape of the protein surface to predict which regions are likely to bind ligands. While many new machine learning-based methods have been developed recently, there are some non-machine learning-based methods that are still commonly used, such as FPocket (Le Guilloux et al., 2009), FTSite (Ngan et al., 2012), and DoGSite3 (Graef et al., 2023). FPocket uses Voronoi tessellations to facilitate alpha-sphere labelling and clustering, followed by partial least squares fitting for ranking of binding sites. FTSite places 16 small molecular probes over the surface of the protein and finds clusters of favourable positions using empirical free energy functions. DoGSite3 uses a difference of Gaussians algorithm for finding binding sites of specific ligands.

4.2.2 Machine learning in protein binding site prediction

The move to machine learning-based methods is driven by the requirement to learn underlying patterns in large sets of data that have proved difficult to learn or represent using physics-based or geometry-based approximations (Kimber et al., 2021). Where a protein of interest has no ligand-bound homologues, binding site prediction requires an understanding of the specific chemical environment needed to bind a ligand. The complexities of protein structure, the large number of data points and the number of features that make up this environment make this an ideal problem to approach via machine learning functions. Additionally, making predictions using machine learning can be significantly faster than other techniques, allowing predictions to be made on large sets of proteins with a lower computational cost.

Proteins can be represented in a variety of ways to facilitate machine learning approaches, which can be split broadly into two groups: featurised representations and learnt representations (Devlin et al., 2018; Rives et al., 2021).

Featurised representations consist of extracted spatiochemical information such as atom type, residue type or euclidean coordinates, usually combined with calculated features such as solvent accessibility, secondary structure, hydrophobicity and charge (Krivák and Hoksza, 2018; Desaphy et al., 2012; Khazanov and Carlson, 2013). These features are then represented by annotation of the atom or residue, usually as part of a point cloud, graph or voxelisation (Zhao et al., 2020). Alternatively, the protein surface itself can be represented using a mesh, Voronoi tessellation or pseudo-atoms annotated with features that describe their chemical environment. The chosen representation is then used to train a machine learning model that predicts on the same type of data. Examples of featurised methods include P2Rank (Jakubec et al., 2022), DeepPocket (Aggarwal et al., 2021), BiteNet (Kozlovskii and Popov, 2020), DeepSurf (Mylonas et al., 2021), NodeCoder (Abdollahi et al., 2023) and PUResNet (Kandel et al., 2021).

Learnt, or non-featurised, representations commonly use a series of vectors to describe a protein. These are often weights from the final layer of a transformer architecture that has been trained to predict masked residues of a protein (Chandra et al., 2023), and are often referred to as embeddings. They represent protein residues as continuous vectors rather than discrete variables, and describe the environment in which a residue exists with respect to neighbouring residues. By using these embeddings along with experimentally-determined labels, machine learning models have been trained to predict features (Chandra et al., 2023) such as ligand binding (Lee and Nam, 2022), PPIs (Si and Yan, 2023), disease variants (Brandes et al., 2022) or structural features (Rives et al., 2021). Unsupervised learning can also be conducted, and has been used for enzyme function prediction (Yu et al., 2023a).

However, these techniques did not incorporate three-dimensional structural information. In 2020, the geometric vector perceptron (GVP) architecture, which does contain this 3D information, was introduced to leverage the proteins' geometric and relational aspects in a sequence recovery task (Jing et al., 2021). This particular architecture was later combined with a generic transformer to create ESM-IF1 (Hsu et al., 2022) which produces embeddings that consist of 512-dimensional vectors for each residue of a protein structure. The ESM-IF1 embeddings have been used for epitope prediction (Høie et al., 2023) and prediction of PPIs (Si and Yan, 2023). Successes such as these suggest that the embeddings contain task-relevant information relating to the protein function and biochemical activity. The embeddings do not take into account side chain positions, but only the protein backbone, which should make the embedding robust to errors in the side chain positioning. This should offer advantages when making predictions on low-accuracy structures, whether experimental or predicted.

4.2.3 Binding site prediction on predicted structures

Most binding site prediction methods are designed to predict binding sites from the ligand-bound (holo) structure of the protein, however this is not necessarily the most useful task. Since mid-2021, highly accurate structure predictions for many proteins have been available, first from AlphaFold (Jumper et al., 2021), and later from others, including RosettaFold (Baek et al., 2021), OmegaFold (Wu et al., 2022) and ESMFold (Lin et al., 2022). This means that for many novel proteins, there is now a structural starting point for protein binding site prediction where previously there was no experimental information available.

To aid this process, AlphaFill (Hekkelman et al., 2023) was developed following the release of AlphaFoldDB (Varadi et al., 2021), a database of predicted protein structures. AlphaFill 'fills in' predicted protein structures with putative ligands by searching for areas of local sequence similarity between predicted proteins and existing complexes in the PDB (Berman et al., 2000). An alignment and 'transplantation' strategy places ligands in potential binding sites of AlphaFold-predicted protein structures (herein referred to as AlphaFold 2 (AF2) structures). For novel proteins that have regions of at least 85 residues with higher than 25% sequence identity to proteins with ligands bound in the PDB, AlphaFill provides a first step to locating ligand-binding sites. Where sequence identity for a minimum of 85 residues is higher than 40%, binding site RMSD is rarely higher than 2Å, suggesting a good match. AlphaFill was able to fill over 59% of proteins in AlphaFoldDB (in February 2022) with ligands. This leaves a large proportion of proteins that do not have close homologues already experimentally solved with ligands bound. These proteins will therefore require template-free binding site prediction.

4.2.4 Binding sites of predicted protein structures

The most commonly used accuracy measurement for protein structure predictions is the global RMSD of their backbone atoms when aligned to the experimentally-solved structure (herein referred to as the PDB structure); a value below 2Å is taken to indicate the model is of high quality. However, this global measure does not provide an assessment of the local quality of the binding site in this predicted structure.

Binding site prediction methods have already been applied to AF2 structures to identify areas of the protein which interact with small molecules such as native binders (e.g. substrates, modulators or inhibitor) or drug-like compounds. FPocket (Le Guilloux et al., 2009) was used to compare volumes of binding pockets in PDB structures and their high-accuracy AF2 counterparts (median all-atom RMSD of 1.54Å), and a 20% reduction in binding pocket volume in AF2 structures was found (Díaz-Rovira et al., 2023). A 2022 study on AF2 structures showed that binding site prediction by AutoSite (Ravindranath and Sanner, 2016) was much less successful where mean residue confidence (pLDDT) for a protein was below 90%, with F-scores reduced by around 80% compared with predictions on holo, apo or high-confidence AF2 structures (Akdel et al., 2022). The same study also found that only 25% of residues are predicted with confidence over 90%, indicating that many AF2 structures may be difficult targets for binding site prediction.

Several studies on docking small molecule ligands into AF2 structures have been published (Díaz-Rovira et al., 2023; Holcomb et al., 2023; Scardino et al., 2023). Despite high accuracy in the test structures (17 of 22 had RMSD lower than 2Å in (Scardino et al., 2023)), docking proved much more difficult for AF2 structures than their PDB counterparts (Scardino et al., 2023). Even when controlling for the accuracy of the predicted structure around the binding site, docking remained a challenge: a recent study found that even proteins with binding site all-atom RMSD as low as 1Å were significantly more difficult to dock into than experimentally-determined struc-

tures (Karelina et al., 2023).

These studies all suggest that even accurate predictions may exhibit significant differences in binding sites. However, in a recent paper (Jakubec et al., 2022), P2Rank was found to have similar levels of accuracy in predicting binding sites on several thousand AF2 structures and PDB structures of the same proteins on two different test sets (HOLO4K (Schmidtke et al., 2010) and COACH420 (Krivák and Hoksza, 2018)).

4.2.5 Accuracy of predicted protein structures

The proteins in commonly-used test sets for binding site prediction (such as HOLO4K and COACH420) are by definition publicly available as protein-ligand complexes. Therefore, it is possible they contain many proteins that are in the training sets for protein structure predictors (Akdel et al., 2022). This would result in predicted structures being much closer to the PDB structure than would be achieved for novel targets without existing close homologues. The consequence of this would be that the datasets used to test tools on 'predicted structures' would not be representative of predicted structures of novel targets, limiting the effectiveness of any evaluation.

The Critical Assessment of Structure Prediction (CASP) (Moult et al., 1995) carries out rigorous blind testing of protein structure prediction methods and evaluation of results by independent assessors (Kryshtafovych et al., 2021). For the CASP iteration in which AlphaFold 2 was first present (2020), the best predicted structure for each target was taken and the fractions of targets predicted at different levels of accuracy were calculated. This provides insight into the accuracy of current protein structure prediction methods based on the availability of structural information of related proteins. Proteins are grouped based on the availability of related proteins with existing structures: of the 'Free Modelling' (FM) group that have no detectable homology to existing protein structures, over 50% of proteins present in the CASP

test set have RMSD values greater than 2Å when aligned to the experimentally-solved structures. These are the proteins that cannot be filled by AlphaFill, so will require template-free binding site prediction.

One of the recent studies looking at docking into AF2 structures specifically selected proteins that were not in the AF2 training set for their test set (Karelina et al., 2023), and found that most of these structures had between 2 and 4Å all-atom RMSD when compared to the experimental structures, further confirming that novel proteins which are targets for template-free binding site prediction are expected to have RMSD values in this range. Therefore, we would expect template-free binding site prediction tools to be evaluated using AF2 structures that have up to 4Å all-atom RMSD.

4.2.6 Protein dynamics and binding sites

Protein structures, whether experimentally-determined or predicted, represent just a single snapshot of dynamic systems. This can limit prediction success, as the protein pocket may not be present in the particular structure that is being used for prediction. An analysis of BiteNet's predictions on a minimisation molecular dynamics (MD) trajectory of an adenosine A2A receptor showed that an allosteric site became detectable by BiteNet at a backbone global RMSD value of just 0.4Å compared to the original structure (Kozlovskii and Popov, 2020). This emphasises how significant structural changes can arise within ligand binding sites at a local scale even where the change is negligible on a global scale. While this has importance in prediction of binding sites of experimental structures, it is even more relevant when using predicted protein structures, as 'highly-accurate' structures (<2Å to the experimental structure) may contain large local differences that make it difficult to identify any binding sites correctly.

Currently, MD simulations remain the only proven way of generating multiple

structures for protein binding site prediction but their usefulness is limited because these simulations are computationally costly. Computationally cheaper generation of multiple protein conformations is an area of much interest (Anand and Achim, 2022; Stein and Mchaourab, 2022; Janson et al., 2023; Jing et al., 2023; Liu et al., 2023; Lane, 2023), however, it is not yet clear if these methods are able to replace effectively information gained from MD simulations.

4.2.7 Summary

Here we describe IF-SitePred, a method for protein binding site prediction using representations obtained from ESM-IF1 (Hsu et al., 2022). We compare the performance of our method to other commonly-used binding site prediction tools on PDB structures taken from the HOLO4K test set (a common set for testing binding site prediction tools) and their equivalent AF2 structures. We assess the accuracy of the predicted structures, and find that the structures are much more accurate than expected, likely due to the presence of these proteins in the training set of AlphaFold. We therefore use molecular dynamics simulations of predicted structures to create lower accuracy protein models (up to 4Å RMSD) and evaluate how binding site prediction success varies with structural accuracy. Finally, we show that by using ensembles of structure predictions, the prediction of binding residues can be greatly improved. We find that IF-SitePred achieves superior binding site prediction compared to commonly-used methods on low-accuracy protein structures, particularly where multiple structures are available.

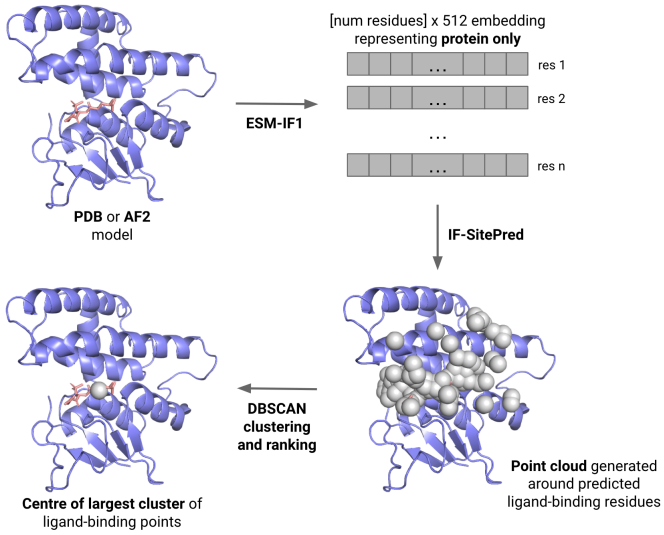


Figure 4.1: ESM-IF1 embeddings were generated for each residue of the protein, and the model labelled each of these as binding or non-binding. A point cloud around predicted binding residues was generated, made up of points that were within 4.5Å of at least three predicted binding residues. These points were clustered using the DBSCAN algorithm. Clusters were ranked in order of number of points, and the mean coordinates of each cluster was taken as the binding site centre.

4.3 Methods

4.3.1 Datasets

We selected HOLO4K (Schmidtke et al., 2010) as our test set to facilitate comparison to other methods, as it is a common set used to test binding site prediction tools. For 4309 proteins in the HOLO4K set, we used the UniProt (Consortium, 2022) ID mapping service to map each PDB code to the AlphaFold Protein Structure Database (Varadi et al., 2021). For the 3914 proteins that appeared to have a corresponding AF2 structure, we verified that the sequence identity between the PDB structure and the prediction was over 90% (100% was not always possible due to the presence of tags or absence of flexible regions in the PDB structure). This removed 1636 protein pairs, leaving 2278 proteins with correctly matched sequences. We then clustered each pair of sequences using MMseqs2 (Steinegger and Söding, 2017) to ensure our

test set did not contain any pairs of proteins more similar than 90%. This resulted in 691 viable pairs of PDB and AF2 structures. To make it possible to evaluate predictions on the AF2 structures, we aligned each prediction to its corresponding PDB structure. Just 14 of the 691 predictions had a backbone RMSD above 2Å and each of these were visually inspected to check whether the binding sites aligned well enough to be included in the analysis. Of these, 11 were retained, including pairs with RMSD values up to 16Å (these contained some large differences in relative domain positions compared with their PDB counterparts, but still had well-aligned binding sites). We named the final set of 688 pairs the HOLO4K-AlphaFold2 Paired (HAP) set. We also extracted a set of 280 pairs (herein referred to as the HAP-small set) which only contained proteins with lower than 25% sequence identity (calculated using Diamond (Buchfink et al., 2021)) to the P2Rank (Krivák and Hoksza, 2018) training set (as we compare to P2Rank). As AlphaFold was most likely trained on structures contained in the HOLO4K set, the quality of structures may be higher than expected. We evaluate this by calculating the RMSD of the predicted structures to the experimentally-determined structures and compare this with the expected level of accuracy as found in CASP (Kryshtafovych et al., 2021).

Our training set consists of structures taken from Binding MOAD (Hu et al., 2005; Ahmed et al., 2014; Smith et al., 2019). The Binding MOAD platform contains 11058 families (clusters) of proteins with each cluster containing a leader (the cluster centre) and members which each have over 90% sequence identity to the leader. We first removed any family for which the leader had greater than 25% sequence identity to any protein in our test set. For each of the remaining 6550 families, we aligned all members to the family leader, and labelled the residues of the leader as follows: any residues within 5Å of the ligand of the leader were labelled as binding residues; any remaining residue within 5Å of any ligand bound to member structures were not used in the dataset, to avoid the potential for false negative annotations; the remainder were

labelled as non-binding. Only residues with relative surface accessibility over 0.02 (calculated using the PyMOL application programming interface (API) (Schrödinger, LLC, 2015)) were included, as we wanted to eliminate only fully-buried residues, and we found that even residues with a relative surface accessibility of 0.025 could have atoms within 4.5Å of a ligand. This resulted in 143,022 binding residues and 1,414,153 non-binding residues for use in training.

4.3.2 Model training

For each protein in the training set an ESM-IF1 embedding (Hsu et al., 2022) was generated. The residue annotations were applied as above, and the residues were treated independently. We generated 40 training sets, made up of a random 80% sample of the binding residues and an equal number of randomly sampled non-binding residues. We initially used AutoML (Microsoft, 2022) to train models on all 40 datasets and found that in the majority of models, the LightGBM model (Ke et al., 2017) had the highest performance on the validation data (randomly taken from the training set). LightGBM is an implementation of the gradient boosting machine (see Chapter 1) which is faster and more efficient than previous implementations as it avoids having to use all the data in the determination of splits within each decision tree.

We trained LightGBM models for all datasets, using a 10% random sample of the input data as a validation set. For all models, the parameters were fixed. A binary objective was used, along with a 'binary_logloss' metric parameter. Based on commonly selected parameters in the AutoML models, gradient-boosted decision trees (GDBT) were used, with no feature pre-filtering and no early stopping round; the number of leaves was set at 200 and 200 iterations were used.

4.3.3 Binding site prediction with IF-SitePred

For a protein in the test set, an ESM-IF1 embedding was generated and each residue was independently predicted by each of the 40 models to be ligand-binding or non-ligand-binding, with a minimum predicted probability of 0.5 for positive labelling. Only if all 40 models predicted a residue to be binding was a positive label applied. Only the relevant chain of the protein was included in the prediction (the chain which had the largest number of residues within 5Å of the ligand). Using the PyMOL API (Schrödinger, LLC, 2015), a point cloud on a 1.5Å grid was generated around the protein, containing only points between 3 and 6Å from any protein atom. For every residue that was labelled as binding, points within 4.5Å of the residue were saved. Points that were saved three or more times (i.e. were within 4.5Å of three or more residues labelled as binding) were clustered using the DBSCAN algorithm from Scikit-learn (Pedregosa et al., 2011), using a 1.7Å cutoff to separate clusters. Clusters were ranked simplistically, using the total number of points in the cluster (including repetitions of the same point), and the centres of the top-ranked clusters were calculated by taking the mean coordinates of all points. This process is summarised in Figure 4.1. All distance thresholds in the point cloud labelling and clustering were selected to maximise success on training set proteins where all residues were labelled correctly (99% success when predicting top 1 pocket). To explore how small changes in these thresholds impact prediction success, we adjusted the thresholds higher and lower within 1Å for all values except for the clustering distance threshold, which was adjusted by 0.1Å (Table B.1).

When predicting on an AF2 structure, the predicted structure was first aligned to its PDB counterpart prior to removal of the entire PDB structure. The point cloud generation and clustering protocol was then applied as above.

For comparison of predictions with a random baseline, the number of residues predicted as ligand-binding was calculated, and the same number of surface residues

were randomly annotated as binding, keeping the same proportions of sub-surface (relative surface area between 0.01 and 0.05) and surface (relative surface area over 0.05) residues. An identical protocol to that described above was used for point cloud generation and clustering.

4.3.4 Evaluation Criteria

Several evaluation strategies have been used for binding site prediction. The traditional metrics of area under the receiver operator characteristic curve (AUROC) or accuracy are not appropriate for imbalanced problems as very high scores can be achieved by predicting all residues as non-binding (Longadge and Dongre, 2013). distance from centre to atom (DCA), distance from centre to centre (DCC), discretised volume overlap (DVO) or atom intersection over union (IoU) (Table 4.1) are commonly used, however DCC, DVO and IoU are based on the assumption that the ligand in the PDB complex is the perfect ligand. While this may be the case, it is not certain, and so we opted to use DCA to evaluate and compare our prediction method, where we measure success by whether the centre of the predicted site is within 4Å of any ligand heavy atom. This avoids reliance upon an over-specific definition of the binding site. Several methods calculate DCC for the top-ranked pocket and top-n ranked pockets (where n is the number of ligands bound to the target protein), however this assumes that the dataset contains the protein bound to all possible correct ligands, and so we opted to use the top-ranked pocket and the top-3 ranked pockets.

As a secondary evaluation method, we used F1 score of residue annotation as ligand-binding or non-ligand-binding. F1 score takes into account precision and recall, thus avoiding the issue created by the imbalanced number of positive and negative labels in the data. A score of 1 indicates perfect prediction, however predicting all residues as non-binding would yield an F1 score of 0.

Metric	Definition
DCA	Distance from site Centre to ligand Atom: the distance in Å from the predicted binding site centre to any ligand heavy atom from the experimentally-determined complex
DCC	Distance from site Centre to ligand Centre: the distance in Å from the predicted binding site centre to the centre of the ligand from the experimentally-determined complex
DVO	Discretized Volume Overlap: intersection over union of the volume of the predicted binding site and the volume of the ligand from the experimentally-determined complex
atom IoU	atom Intersection over Union: for use in methods where protein atoms are individually labelled as ligand-binding or non-ligand-binding, intersection over union of the predicted ligand-binding atoms and the experimentally-observed ligand-binding atoms is calculated

Table 4.1: Commonly-used evaluation metrics for protein binding site prediction.

4.3.5 Comparison to existing methods

We selected three popular methods for comparison, based on the range of techniques used, the availability of command-line software and the compatibility of their training sets with our test set: FPocket (Le Guilloux et al., 2009), P2Rank (Kriválk and Hoksza, 2018) and DeepPocket (Aggarwal et al., 2021). FPocket uses Voronoi tessellations and alpha-sphere clustering for prediction to find clefts on the protein surface that have the correct size and shape for ligand binding. P2Rank annotates points on the solvent-accessible surface area of the protein based on feature-vectors applied to exposed protein atoms, labels each point as ligand-binding or non-ligand-binding using a random forest classifier, and ranks sites according to their cumulative ligand-binding score. DeepPocket is a 3D-CNN based method which re-scores pocket centres identified by FPocket, then elucidates the shapes of the predicted pockets.

For all four methods, we compared binding site prediction success as defined by DCA on the top-ranked pockets and top-3 ranked pockets for PDB and AF2 structures for both the HAP and HAP-small test sets, as defined above.

For FPocket and DeepPocket, default parameters were used for binding site prediction on all structures. For FPocket, the centre of each binding site was calculated by taking the mean coordinates of all points in the pocket. For P2Rank, default

parameters were used for PDB structures, and for predicting on AF2 structures, the AlphaFold-specific configuration was used.

4.3.6 Ligand similarity to training sets

Similar to the dataset splitting strategy for preventing protein sequence bias, we also checked whether the methods displayed a bias towards the ligands present in their training sets as follows. Ligand USRCAT (Schreyer and Blundell, 2012) similarity was used as implemented in RDKit (Landrum, 2020). For each test protein, the USRCAT similarity for each ligand bound was calculated for every ligand bound to a training set protein. The highest value was taken and the top-1 success rate for proteins at varying levels of ligand similarity was compared with the mean top-1 success rate. This analysis was performed for each of IF-SitePred, FPocket, P2Rank and DeepPocket with their respective training sets.

4.3.7 Molecular dynamics simulations for testing predictions on low-quality structures

The AF2 structures in the HAP set are highly accurate, with only 11 of 688 predictions having backbone RMSD values over 2Å to the PDB structure. This represents just 1.6% of the predictions, whereas the analysis of the best prediction for each target in CASP14 (Kryshtafovych et al., 2021) suggested that for free modelling targets (those targets without a known structural homologue), over 50% of predictions are likely to have an RMSD over 2Å. Additionally, a study on G protein-coupled receptors found that most structural predictions that were not in the AF2 training set had RMSD values between 2 and 4Å (Karelina et al., 2023). To explore how binding site prediction success changes as predicted structures become less accurate, we filtered the HAP set for physiologically monomeric proteins with fewer than 230 amino acids,

resulting in a set of 21 proteins for which we conducted MD simulations to generate structures that were representative of low-accuracy protein structure predictions.

We used the AF2 structures of each protein for the MD simulations. For each AF2 structure, pKa values were estimated using h++ (Anandakrishnan et al., 2012; Gordon et al., 2005; Myers et al., 2006) to assign residue protonation states and represent solvent implicitly. Ions were added using tleap (Case et al., 2022) to neutralise the system. We then used the Amber (Case et al., 2022) protein forcefield (fF14sb) within OpenMM (Eastman et al., 2017) to heat the proteins from 298K to 548K, with the system simulated for 20ns for each 10K interval. Using MDAnalysis (Gowers et al., 2016), the trajectory was randomly sampled to extract 10 structures for each 0.25Å interval from RMSD values up to 8Å when aligned to the PDB structure.

4.3.7.1 Evaluation

For IF-SitePred, P2Rank and DeepPocket, we predicted binding sites for structures with RMSD values up to 7Å from the PDB structure. Mean rates of top-1 success and top-3 success for each 0.25Å interval were calculated for each method.

4.3.8 Combining predictions by using multiple models on multiple structures

We tested two ensembling methods. For the first, we trained 40 models on different samples of the training data, and combined the results of these to make our final predictions. For the second, we made predictions on multiple protein structures (these could be generated by different protein structure prediction tools or by molecular dynamics) and combined these.

We tested the improvements made by using multiple models and multiple medium-accuracy protein structure predictions. The protein structures we used to test this were from MD simulations of AF2 structures with RMSD to the PDB structures lower

than 4Å, as this would cover around 82% of predictions on free-modelled proteins in CASP14 (Kryshtafovych et al., 2021).

We implemented four prediction pathways to understand the effects of using single or multiple models on single or multiple structures (Figure 4.2). Residues were annotated as ligand-binding or non-ligand-binding by either 40 models (as previously) or by just one model, and point clouds were generated and clustered as previously. The top 3 binding sites for each structure were used to re-annotate only the residues within 5Å of their points as ligand-binding, as this is a commonly used distance threshold for intermolecular interactions. To view the effect of combining predictions on multiple structures, 9 other MD-generated structures of the same protein were randomly selected, and only residues that were predicted as ligand-binding in at least 3 of the 10 structures were given a final prediction as ligand-binding. This minimum threshold of 3 positive predictions for a residue was arbitrarily selected. The F1 scores for these final predictions were calculated to compare residue annotation for these four prediction pathways.

Additionally, we combined residue labels of 10 structures (as above) 1000 times for 150 randomly-selected frames of each of the 21 proteins for which MD had been performed using IF-SitePred, P2Rank and DeepPocket to compare which of these methods are able to benefit from this ensemble method.

4.4 Results

4.4.1 IF-SitePred binding site prediction is competitive with state-of-the-art methods

The prediction of ligand-binding sites on the surfaces of proteins is a useful step towards understanding the function and druggability of novel targets. It is particularly important in the era of accurate protein structure prediction, where we often

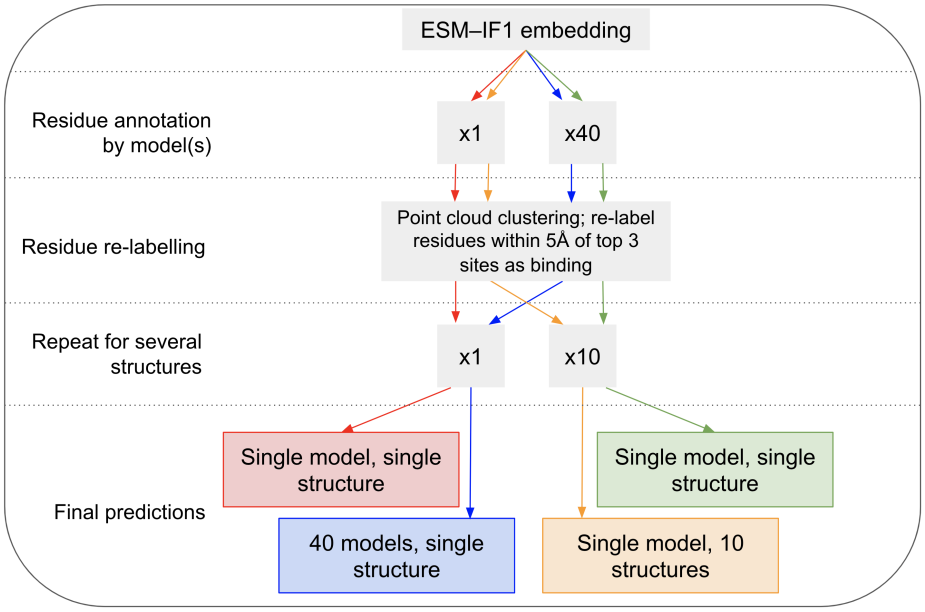


Figure 4.2: The four pathways implemented to compare the F1 scores achieved when predicting ligand-binding residues on 21 MD structures. These allowed us to explore the benefit gained from using multiple predictive models and from making predictions on multiple MD structures.

have a predicted structure before any experimental studies have been carried out. We developed IF-SitePred, a binding site prediction tool which avoids featurisation, and instead uses embeddings generated by ESM-IF1 as the basis for labelling protein residues as ligand-binding or non-ligand-binding. This is followed by point cloud annotation and clustering to determine the three most likely binding sites and their centres. To evaluate our method, we predicted binding sites on hundreds of proteins in the HAP set and compared our rate of success to that of FPocket, P2Rank and DeepPocket. In particular, we compared prediction success on experimental (PDB) structures with those predicted by AlphaFold (AF2). To ensure that we evaluated the tools on a range of proteins that were sufficiently different from the training data from each method, we designed the HAP set to include 688 proteins (both PDB and AF2 structures) which have no more than 90% sequence identity to any other protein in the set and no more than 25% sequence identity to the training sets of IF-SitePred, FPocket and DeepPocket. The HAP-small set is a subset of 280 proteins from the

HAP set, made up of proteins that have no more than 25% sequence identity to the training data used for comparator method P2Rank. Prediction success was measured by top-1 and top-3 DCA, which checks whether the predicted binding site centre is within 4Å of any heavy atom of the experimentally-determined ligand. This measure avoids the assumption that the observed ligand perfectly fills the site.

HAP: 688 proteins

PDB	baseline	IF-SitePred	FPocket	P2Rank	DeepPocket
Top 1	0.12	0.76	*0.75*	**0.81**	0.78
Top 2	0.17	0.89	*0.81*	**0.90**	0.87
Top 3	0.22	**0.93**	*0.83*	**0.93**	0.89

AF2	baseline	IF-SitePred	FPocket	P2Rank	DeepPocket
Top 1	0.09	0.77	*0.50*	**0.81**	0.78
Top 2	0.17	**0.89**	*0.60*	0.88	0.87
Top 3	0.20	**0.94**	*0.67*	0.89	0.90

HAP-small: 280 proteins

PDB	baseline	IF-SitePred	FPocket	P2Rank	DeepPocket
Top 1	0.12	0.75	*0.73*	**0.78**	0.75
Top 2	0.18	**0.89**	*0.80*	0.86	0.85
Top 3	0.24	**0.92**	*0.82*	0.91	0.88

AF2	baseline	IF-SitePred	FPocket	P2Rank	DeepPocket
Top 1	0.10	**0.76**	*0.48*	**0.76**	0.75
Top 2	0.17	**0.88**	*0.58*	0.85	0.86
Top 3	0.20	**0.94**	*0.65*	0.88	0.90

Table 4.2: IF-SitePred, P2Rank and DeepPocket are competive across PDB and AF2 structures, whereas FPocket experiences a significant loss of performance on AF2 structures. Success rates of top-1, top-2 and top-3 binding site prediction as measured using DCA is shown, where success is defined as the centre of the predicted binding site being within 4Å of any ligand heavy atom. We show results for IF-SitePred, FPocket, P2Rank and DeepPocket on two test sets that contain PDB and AF2 structures respectively. For each test set, the highest success rate is shown in bold, and the lowest success rate is shown in italics.

The binding site prediction success rates for IF-SitePred, FPocket, P2Rank, and DeepPocket are shown in Table 4.2. On PDB structures, all methods performed similarly well. On the HAP set, DeepPocket achieved the highest top-1 success rates, but was outperformed by IF-SitePred on top-2 and top-3 success rates. FPocket had the lowest success rates. By using the HAP-small set to compare methods, we observed that P2Rank outperformed all other methods at top-1 success, but shared top performance with IF-SitePred when considering top-3 success. Overall, P2Rank had similar performance on the HAP and HAP-small sets, suggesting the method generalises well. Similar results were observed for IF-SitePred, DeepPocket and P2Rank when predicting binding sites on AF2 structures, with these three methods sharing the highest success rates across the HAP and HAP-small sets. However, FPocket experienced a significant drop-off in performance on AF2 structures. Given that

DeepPocket is a prioritisation method that takes FPocket's predictions as an input, this suggests that the ranking procedure used by FPocket failed on AF2 structures, as opposed to FPocket having difficulties in identifying the ligand-binding sites on the protein's surface. This could be due to the lower pocket volume of AF2 structures (Díaz-Rovira et al., 2023).

In previous studies where different test sets are used, different methods rank differently: in the DeepPocket analysis (Aggarwal et al., 2021), FPocket has a much lower success rate than DeepPocket, and DeepPocket significantly outperforms P2Rank. This indicates that even with large test sets, the composition of the set can make a significant difference to results. To quantify the variability that our results may experience due to the dataset, we used bootstrapping to make an error estimation for each method on each dataset, with the results shown in Tables B.2-B.5 .

For proteins where IF-SitePred failed to make a successful top-1 prediction on the PDB structure, around half were also failures when using the AF2 structure. However, the other half were successfully predicted when using the AF2 structure, suggesting that predicted structures sometimes contain information about ligand binding that is not present in the ligand-bound PDB structure. A similar result was found in the recent P2Rank paper (Jakubec et al., 2022). As it is highly likely that the proteins in this set were present in the AF2 training data, it is interesting that in some cases the predicted structure deviates from the experimentally-predicted structure in a way that may confer more information about the protein's behaviour.

We additionally found that AF2 prediction confidence (pLDDT) has no influence on binding site prediction success by IF-SitePred (Figure B.1).

4.4.2 Accurate binding site prediction is not dependent on ligand similarity to the training set

By removing any protein from the training set with more than 25% sequence identity to any test set protein, we attempted to ensure that our predictions were not based on the model learning the sequences from the training set. However, the model could be learning ligand-based information that was contained in both the training and test sets. To explore this possibility, we investigated whether we were able to predict binding sites more successfully on proteins that have similar ligands to our training set.

For each protein-ligand complex in the HAP-small set, we calculated the maximum ligand similarity to any ligand that bound to proteins in the training set for each of IF-SitePred, FPocket, P2Rank and DeepPocket. Ligand similarity was defined using USRCAT fingerprints, which takes into account the ligand shape and pharmacophoric features. These values were then compared to the mean top-1 success (found in Table 4.1). These results are shown in Table 4.3.

Ligand similarity to training set	IFSitePred	FPocket	P2Rank	DeepPocket
0.2-0.4	*+0.14*	*+0.11*	+0.04	**-0.22**
0.4-0.6	*+0.11*	**-0.36**	-0.06	**-0.40**
0.6-0.8	+0.03	*+0.22*	*+0.14*	+0.05
0.8-1.0	-0.03	-0.01	*+0.14*	+0.06

Table 4.3: DeepPocket performs worse than average on sites that bind ligands significantly different from the training set. We calculated the USRCAT similarity for the most similar ligand in the training set tfrom that binding each protein in the HAP-small set, and calculated the fraction difference between overall top-1 success rate (from Table 4.1) for each 0.2 interval of USRCAT similarity. Success rates differing by over 10% from the mean value are shown in bold (performance loss) or italic (performance gain).

If our predictions were dependent on ligand similarity, we would observe that where ligands bound to the binding site are significantly different from the training set, the binding sites would be predicted with lower success than those where similar ligands were present in the training set. We did not see this trend in the IF-SitePred

results, so we can be confident that our predictions do not rely on ligand similarity to the training set. P2Rank also does not exhibit a significant bias towards proteins that have ligands that bound to training set proteins. However, where a site binds a ligand with only dissimilar ligands in the training set (USRCAT similarity lower than 0.4), DeepPocket and FPocket are significantly less able to correctly predict the binding site at the highest rank.

In our tests IF-SitePred and P2Rank are not affected by similar ligands being available in the training set, whereas DeepPocket and FPocket are, meaning that they may not generalise well to the prediction of binding sites of novel ligands.

4.4.3 IF-SitePred outperforms other methods in top-3 binding site prediction on MD structures

When verifying the binding site alignment of the AF2 counterparts of the test set proteins, we observed that only 11 proteins in the final HAP set had backbone RMSD to the PDB structure above 2Å (Figure 4.3). This represents just 1.6% of the predictions, whereas analysis of CASP14 results (Kryshtafovych et al., 2021) suggested that for free modelling targets, over 50% of predictions had an RMSD over 2Å, over 30 times what we observed in our dataset.

To explore how binding site prediction success changes when the input structures are less accurate, we selected 21 monomeric proteins with fewer than 230 amino acids from the HAP set for which we conducted molecular dynamics simulations, followed by binding site prediction.

We heated each protein from 298K to 548K, simulating for 20ns at each 10K interval, and sampled structures from every tenth frame. In Figure 4.4, we show binding site predictions made by IF-SitePred on an example structure at different stages of the MD trajectory.

Structures were sampled uniformly up to 8Å RMSD (when aligned to the PDB

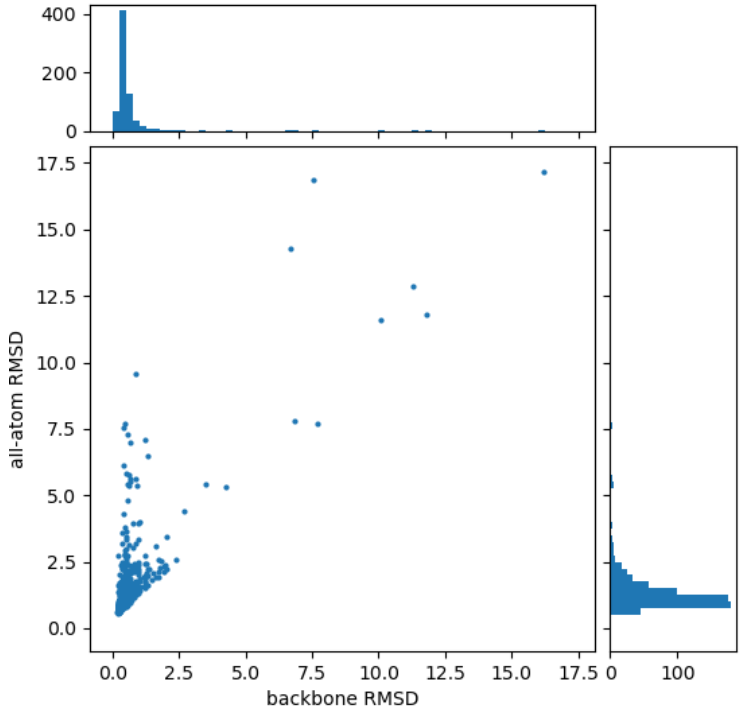


Figure 4.3: AF2-predicted structures in the HAP set are extremely accurate compared with most free modelling AF2 predictions on novel proteins. Backbone RMSD is shown on the x-axis, and all-atom RMSD is shown on the y-axis. Each axis has a corresponding histogram to show the spread of values. Over 98% of AF2 structures in the HAP set have backbone RMSD values below 2Å.

structure) and binding sites were predicted using IF-SitePred, P2Rank and DeepPocket. Using DCA, the probability of success for each 0.25Å RMSD interval was calculated for top-1 and top-3 ranked binding sites (Figure 4.5). The means of these probabilities for each interval were calculated, and a line of best fit was determined. Even at just 1Å RMSD, performance was significantly reduced for all methods, with a reduction in success rates of up to 15% compared to the PDB or original AF2 structures.

When comparing the three methods, we found that P2Rank performed best on structures up to 3Å RMSD, but had a greater loss of performance than IF-SitePred, which attained higher success rates at RMSD values over 4Å. When evaluating top-3

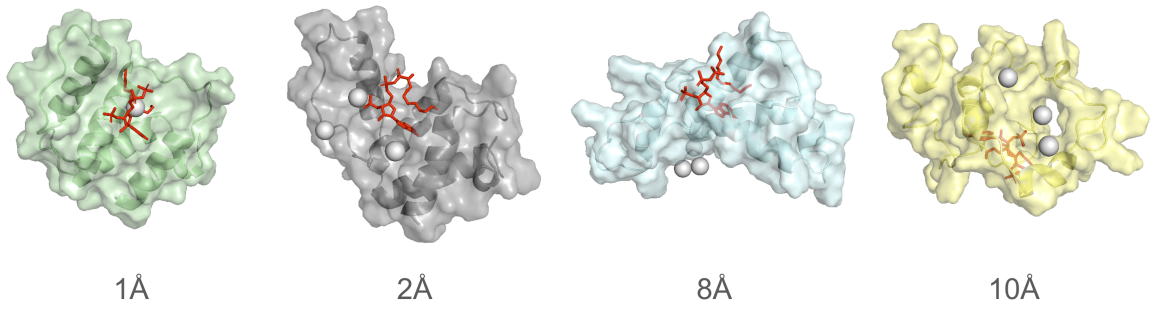


Figure 4.4: An example of binding site prediction on structure 1HBK at increasing RMSD, induced via molecular dynamics simulations. IF-SitePred is able to predict a binding site in all cases, however for the higher RMSD structures it is difficult to discern which would be classed as the 'correct' binding site.

success rates, IF-SitePred achieved higher success rates than P2Rank and DeepPocket across almost all RMSD values. The difference grew at higher RMSD values, with IF-SitePred able to succeed 60% of the time on structures with an RMSD of 5Å, compared to 50% and 47% success for P2Rank and DeepPocket respectively. These results all suggest that IF-SitePred is more robust to errors in the protein structure.

4.4.4 Combining predictions of multiple models on multiple structures improves predictive power

IF-SitePred uses 40 models that were trained independently on 40 different samples of the training data. Due to the imbalanced nature of the number of binding and non-binding residues, this allows the different models to learn the features of different sets of the non-binding residues, thus providing a final prediction informed by more data. Another strategy to maximise use of available data is possible when multiple structures of the same protein are available, whether by generating multiple structure predictions when using a tool like AlphaFold, or by performing molecular dynamics simulations: predictions can be made on multiple structures and combined to give a final prediction.

We implemented four pathways which used a combination of different models and

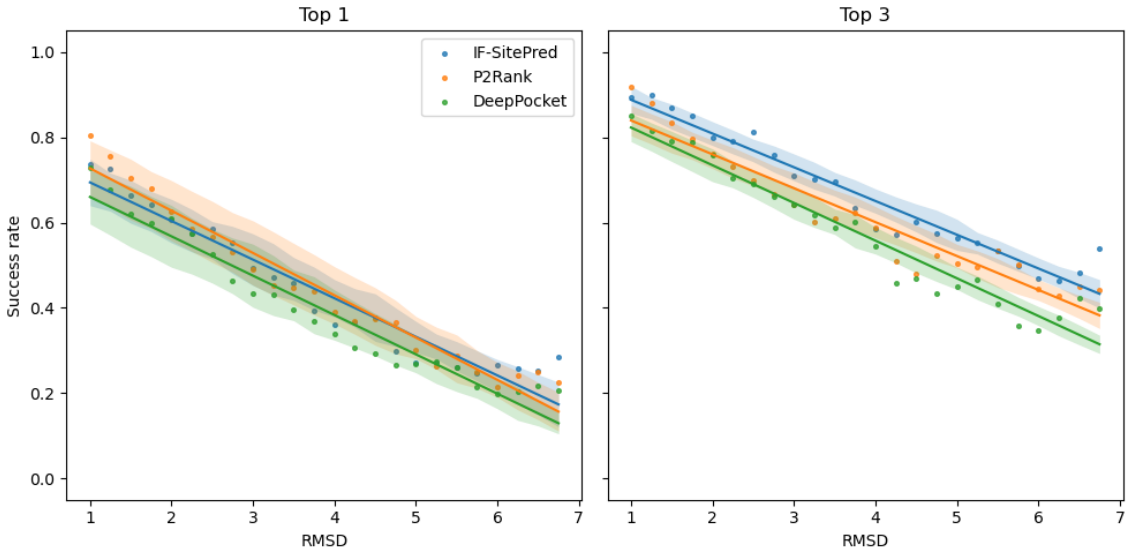


Figure 4.5: IF-SitePred outperforms P2Rank and DeepPocket when considering top-3 success on low-accuracy structures. We predicted binding sites on MD structures at increasing RMSD when aligned to PDB structures. Points represent mean success rate at each RMSD interval (value to value plus 0.25Å) across all 21 targets. Based on these, lines of best fit are calculated and plotted (solid lines of identical colour). Standard error of the mean across all targets is represented by the shaded areas. The three methods performed similarly when considering only the top-ranked binding site (left), with P2Rank performing slightly better at low RMSD values. However, IF-SitePred achieved higher top-3 success than P2Rank and DeepPocket at all RMSD values (right).

structures (see Figure 4.2) to understand the improvements made when combining different predictions. For each prediction on an MD structure with RMSD values lower than 4Å, we calculated the F1 score of our predicted binding residues compared with the binding residues observed in the PDB protein-ligand complex (Figure 4.6).

We first calculated the mean F1 scores of binding residue prediction on PDB structures using a single predictive model (0.62) and using 40 predictive models (0.68). When a single predictive model predicted the binding residues of single low-accuracy protein structures, the mean F1 score was just 0.41. When multiple models were used for the prediction, the mean F1 score rose to 0.43, and a further improvement to 0.59 was seen when multiple protein structures were also included. This represented an improvement of 44% when using two ensemble strategies compared to when only using single models and structures, and showed that by developing methods that take

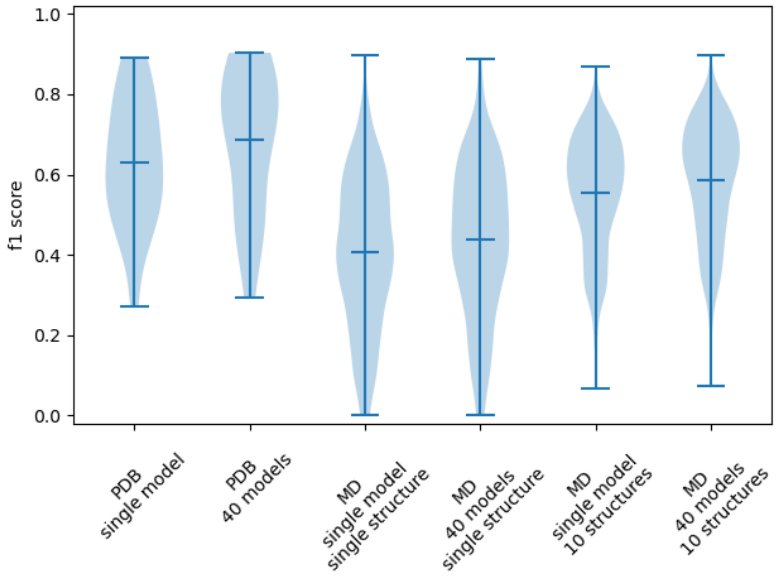


Figure 4.6: Using multiple predictive models on multiple protein structures improves IF-SitePred's predictive power. We compare F1 scores of predictions of binding residues for PDB and MD structures of 21 proteins when using single or multiple predictive models on single or multiple structures (MD only). Combining predictions of multiple models on multiple structures yielded the most accurate prediction of binding residues on MD structures, with F1 scores approaching those of predictions on PDB structures.

into account as much data as possible, we were able to mitigate errors in the data and make good predictions on flawed data that were significantly closer to the accuracy seen when predicting on high-quality data.

We additionally compared the benefits of using many structures to make final predictions for IF-SitePred, P2Rank and DeepPocket. This involved taking predictions from a single structure and comparing the F1 score with the combined prediction of the same structure with nine additional randomly-selected structures. This was repeated 1000 times for each target to ensure that the results were as representative as possible of each method. While the F1 scores do not directly correspond to DCA success rate as the protocols for determination and ranking of protein binding site centres differ, we were able to compare the impact of combining multiple sets of predictions between methods. Additionally, we used the frequency of prediction of each residue as ligand-binding to create a multi-structure prediction probability, and used

this to plot precision-recall curves for each method.

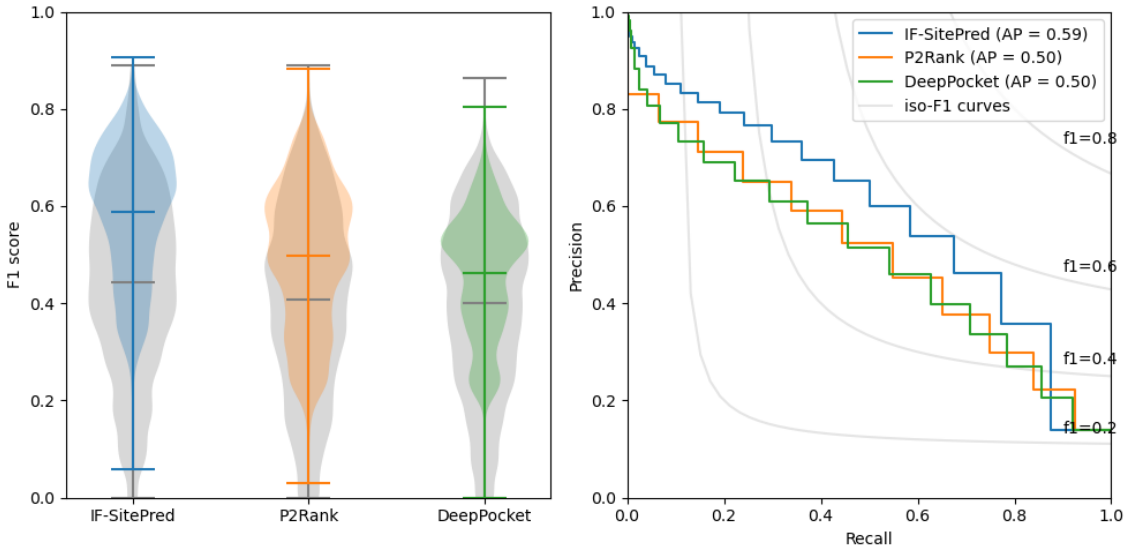


Figure 4.7: **Left:** Predictive power for binding residue annotation improves when predictions for 10 structures were combined for all three methods (blue, orange, green) compared to when predictions for single structures were used (grey). IF-SitePred had a slightly higher original F1 score, and also saw the greatest improvement of the three methods, increasing to 0.59. **Right:** Precision-recall curves for all three methods reveal that IF-SitePred has a higher average precision (AP) (0.59) than P2Rank (0.50) and DeepPocket (0.50). Iso-F1 curves are shown in grey, demonstrating that IF-SitePred achieves higher F1 scores across all probability thresholds.

When just one structure was used to make a prediction, IF-SitePred slightly outperformed P2Rank and DeepPocket, in agreement with DCA-based results (Figure 4.5). When predictions for 10 structures were combined, all three methods had improved predictive power, showing the importance of understanding multiple states of the protein. IF-SitePred was able to benefit the most, with F1 score increasing by around 34%, compared with 22% and 15% improvements from P2Rank and DeepPocket respectively (Figure 4.7 (left)). We expect that this would translate to a higher DCA success rate for IF-SitePred compared to P2Rank and DeepPocket when predicting binding sites on error-prone structures. Additionally, we show that IF-SitePred has a higher average precision (AP) than P2Rank and DeepPocket for multi-structure prediction (Figure 4.7 (right)). While the F1 scores were calculated

based on a binary classification with only residues predicted as ligand-binding in at least 3 of 10 structures regarded as positive, the iso-F1 curves applied to the precision-recall curve show that regardless of threshold used, IF-SitePred will achieve a higher F1 score.

These results suggest that where multiple protein structures are available or can be generated, IF-SitePred is able to take greater advantage of the available data to outperform DeepPocket and P2Rank consistently.

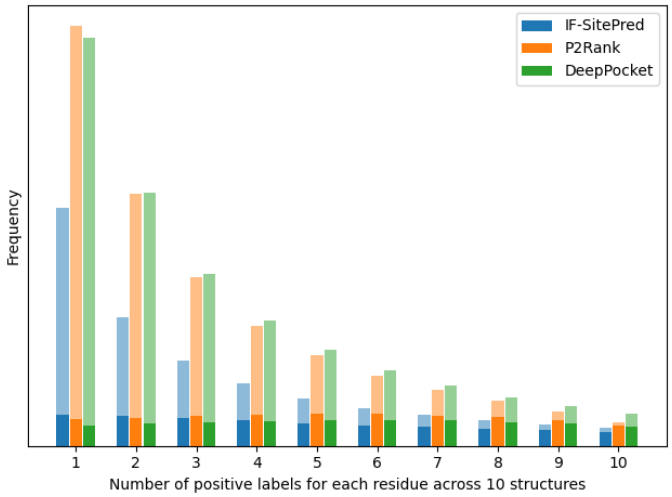


Figure 4.8: IF-SitePred predicted fewer residues incorrectly as ligand-binding. The frequencies of correctly-labelled (solid colour) and incorrectly-labelled (translucent) residues across 10 randomly-selected MD structures of the same protein are shown. Non-ligand-binding residues were most likely to be incorrectly labelled as ligand-binding in 0 or 1 structure, whereas ligand-binding residues had a roughly equal likelihood of being labelled as ligand-binding up to 10 times. IF-SitePred consistently had a lower rate of labelling non-ligand-binding residues as ligand-binding, with only a small reduction in the number of correctly-labelled ligand-binding residues compared to other methods.

To explore why IF-SitePred was able to take advantage of the information contained in multiple structures better than P2Rank and DeepPocket, we examined the rates of true positives (correctly-labelled ligand-binding residues) and false positives (non-ligand-binding residues labelled as ligand-binding) for residues that were predicted as positive in at least one of the 10 randomly-selected MD structures of the same protein (Figure 4.8). We found that the number of false positives was reduced in

residues that were predicted as positive in more structures, while the number of true positives remained similar. However, IF-SitePred had consistently fewer false positives without a significant reduction in rate of true positives. Conversely, DeepPocket had the highest number of false positives of the three methods, with the number of false positives that were predicted in all 10 structures as ligand-binding almost triple those for P2Rank or IF-SitePred.

4.5 Conclusions

As predicted protein structures become the initial input for tools locating ligand binding sites, it is important to evaluate whether these structures are accurate, and develop binding site prediction tools which are robust to errors in structure predictions.

In this study we describe IF-SitePred, a protein binding site prediction tool that is competitive with state-of-the-art tools on high-accuracy AF2 structures. This is despite no explicit featurisation and a simplistic pocket ranking strategy. However, we found that the AF2 models used to evaluate binding site prediction methods had far higher accuracy than would be expected for free modelling targets, which would be the primary targets for template-free binding site prediction. Therefore, we examined how binding site prediction tools perform on structures which are representative of novel protein targets.

To evaluate how the inaccuracies in the protein structure prediction impact binding site prediction, we used MD simulations to generate models of the target proteins with varying accuracy and found that IF-SitePred can consistently outperform competitors when predicting three binding site centres on lower-accuracy structures. By taking the most popular predictions on 10 medium-accuracy structures (1-4Å RMSD), predictions made by IF-SitePred contain a significantly reduced rate of false positives

without serious loss in recall of true positives. This results in a significant improvement in performance of IF-SitePred compared with competitors. By representing the protein as a set of ESM-IF1 embeddings, it is possible to take greater advantage of the diversity in the ensemble of structures compared with the explicit featurisation used for P2Rank and DeepPocket.

For many structures with high-accuracy predictions, P2Rank, DeepPocket and IF-SitePred will yield similarly reliable predictions. However, where the accuracy of a predicted structure for a novel target is thought to be poor, our results suggest that IF-SitePred will provide a considerably more accurate binding site prediction than other tools currently available. This will allow researchers to begin to examine the binding site before any experimental results are available, and although poor structures sometimes are not suitable for some types of analysis (e.g. docking), having some insight into the binding sites greatly facilitates other types of analysis such as binding site comparison and selection of targets for drug design campaigns.

Additionally, where a group of structures can be generated using MD or by making many structural predictions (by using one or multiple structure prediction tools), the accuracy of binding site prediction can be enhanced further. Several tools have been developed specifically to generate conformational ensembles of protein structures, such as idpGAN (Janson et al., 2023), which was trained on MD trajectories.

By accurately locating ligand-binding sites on predicted protein structures, exploration of the ligand-binding behaviour of the protein can commence. In the next chapter, we discuss possible strategies for the comparison of binding sites and how these can be used for the design of target-specific fragment libraries.

Chapter 5

Binding site similarity strategies for target-specific fragment library design

5.1 Preface

Binding site similarity is a major area of interest for structure-based drug design: pinpointing areas of local similarity between proteins allows the prediction of off-target effects and facilitates drug repurposing. In FBDD, where the same fragment library is screened on many targets, the ability to identify binding sites from unrelated proteins which bind the same fragment would allow the creation of smaller, target-specific libraries that thoroughly explore the binding site of the desired target. This strategy could significantly reduce costs, as it demands far fewer experiments to extract the required information from the target protein. However, a comprehensive benchmarking study (Ehrt et al., 2018) showed that binding site identification of unrelated proteins that bind identical ligands remained an unsolved challenge.

In this chapter we first evaluated ProCare (Eguida and Rognan, 2020), a tool for

the scoring of fragment binding site similarity. We found that while it effectively aligns subcavities into larger binding sites with reasonable accuracy, the ProCare scores of cavities binding identical fragments do not exhibit a significant deviation from those binding different fragments. We then explored two distinct approaches for using learnt representations for the identification of binding sites from unrelated proteins which bound identical ligands: distance-based residue similarity and ML-based prediction of protein-ligand interactions. While the distance-based similarity approach fell short in identifying residues from unrelated proteins that bound identical ligands, use of learnt representations for protein-ligand interaction prediction significantly outperformed one-hot encoding.

5.2 Introduction

Identifying similarities and differences in binding sites across different proteins can elucidate the function of uncharacterised proteins and facilitate prediction of off-target effects of drugs, polypharmacology and drug repurposing (Govindaraj and Brylinski, 2018; Ehrt et al., 2018). Binding site similarity can be defined in several ways, including sequence, shape, electrophilicity and ligand-binding behaviour. As binding site similarity tools have been developed for various use cases, it is important to define what type of similarity is required for our problem of interest. For instance, closely-related proteins may exhibit significant differences in their binding sites due to small changes in the amino acid sequence, thereby changing key physicochemical features or ligand-binding behaviour. Conversely, it has been found that unrelated proteins with no obvious similarities in their binding sites can bind identical ligands.

In this chapter, we focus on the latter scenario for use in the design of target-specific fragment libraries. Accurately identifying binding sites that bind identical fragments would enable the leverage of historical fragment screening data to create a

database of ligand-bound pockets. These pockets can then be compared to the protein surface of any novel target of interest. By screening only the ligands bound to pockets that exhibit high similarity to the target binding site, we can simultaneously reduce experimental costs and maximise the information gained from the target.

5.2.1 Overview of binding site similarity approaches

Comparison of binding sites entails the examination of similarities and differences in geometric and physicochemical characteristics. In some cases, global features such as fold, sequence identity, or RMSD are used, but more often than not, scoring functions focus on local features. These include residue composition, shape, buriedness, chemical features, hydrophobicity, local RMSD, van der Waals forces, and electrostatic potentials, among others.

A key consideration when designing a method for binding site similarity is the requirement for alignment of the sites: alignment makes the comparison of binding sites features simpler, however the alignment of binding sites is uncertain and computationally costly. A misalignment of the binding site will heavily reduce the possible similarity score a pair of binding sites can achieve (Weill and Rognan, 2010). Alignment-free tools are more applicable to sites which have very different shapes and chemical compositions or where the ligand has significant conformational differences.

5.2.1.1 Alignment-based approaches

To align binding sites from structurally different proteins, the binding sites must first be defined. The residues involved are usually selected using binding site prediction tools or by including residues within a certain threshold distance of a bound ligand. Often, the binding site is represented using points that characterise the surface of the protein (Shulman-Peleg et al., 2004), important protein functional groups (Konc and Janezic, 2010) or the cavity labelled with pharmacophore types (Chartier and

Najmanovich, 2015; Eguida and Rognan, 2020). These can be left as point clouds or can be used to construct graphs. Alignment algorithms can then be applied, such as geometric hashing (Shulman-Peleg et al., 2004), clique procedures to calculate vertex-to-vertex correspondences (Konc and Janezic, 2010), or random sample consensus (RANSAC) iterations (Fischler and Bolles, 1981; Eguida and Rognan, 2020). The final binding site similarity scores are usually composed of a score for the alignment (e.g. RMSD, alignment fitness, or surface vector angles) and a score for how well the pharmacophoric environments of the binding sites match.

In some cases, the presence of a bound ligand can be leveraged, such as in IsoMIF (Chartier and Najmanovich, 2015) where the ligand features can influence the labelling of the point cloud, or in ProCare (Eguida and Rognan, 2020) where the interaction fingerprints of the protein-ligand complexes can be compared and included in the scoring regime.

5.2.1.2 Alignment-free approaches

To achieve alignment-free binding site similarity scoring, the binding sites must be defined in a format that allows for direct comparison. Such encodings may represent combinations of pharmacophoric features (Weill and Rognan, 2010; Ritschel et al., 2014), distances between critical residue atoms (Yeturu and Chandra, 2008) or combinations of many features including atom type, residue type, hydrophobicity, solvent accessibility and secondary structure (Liu and Altman, 2011). A similarity score can then be quickly calculated between pairs of binding sites, with most methods opting to use a Tanimoto similarity-based scoring metric (Weill and Rognan, 2010; Ritschel et al., 2014; Liu and Altman, 2011; Yeturu and Chandra, 2008).

Some alignment-free methods also utilise information from bound ligands, such as KRIPO (Ritschel et al., 2014) which uses bound ligands to calculate protein-ligand interactions.

5.2.1.3 Machine learning approaches

Machine learning approaches have become common in many areas of drug discovery, including protein-ligand binding affinity prediction (Jones et al., 2021; Scantlebury et al., 2023), protein structure prediction (Jumper et al., 2021), molecular generation (Zeng et al., 2022b), binding site detection (see Chapter 4) and classification of binding sites (Scott et al., 2022; Pu et al., 2019). However, the limited availability of data for identifying similar binding sites renders this task particularly challenging from a machine learning perspective.

Site2Vec (a successor of PocketMatch) (Bhadra and Yeturu, 2020) and DeeplyTough (Simonovsky and Meyers, 2020) are tools that approach this issue with a machine learning strategy. They focus on learning a representation for binding sites that can then be compared using a distance-based metric or additional machine learning models.

Site2Vec initially lists the representation of atoms in a binding site along with pairwise distances of the atoms. A one-dimensional autoencoder deep neural network was trained to convert information extracted from this list into a fixed-length vector. This vector generation step was trained on the entire PLIC dataset (Anand et al., 2014). A fused vector consisting of the vectors of two binding sites is then classified as 'similar' or 'dissimilar' by a binary classifier: random forest, Gradient Boosting and Ada Boost classifiers were trained. For each evaluation task included in the study, these binary classifiers were trained on a random half of the data provided.

DeeplyTough is a convolutional neural network which learns three-dimensional pocket descriptors through the application of a contrastive loss function to pairs of pockets that are labelled as similar or dissimilar. The goal of such an approach is to embed the similar pockets close to one another in euclidean space while keeping the negative descriptors far apart. Descriptor robustness was stimulated by data augmentation via a random rotation of each pocket seen during training. Furthermore,

each descriptor of the same pocket was encouraged to be identical via a stability loss function. The descriptors generated can be compared using a simple calculation of euclidean distance. As a result, the generation of a similarity score is fast and efficient, and may be a more accurate prediction of the similarity between pockets that have not been seen in the training set. DeeplyTough is trained on the TOUGH-M1 (Govindaraj and Brylinski, 2018) dataset, but to prevent any data leakage, proteins with more than 30% sequence identity to the evaluation set are excluded.

5.2.2 Identical ligands bound to different proteins

In fragment screening campaigns, identical fragments often bind to unrelated proteins (see Chapter 3). As this also occurs for larger drug-like molecules, the properties of binding sites from different proteins which bind to the same ligand have been closely examined.

Kahraman et al. (2007) assembled a set of 100 binding sites, each bound to one of nine ligands. It was found that the ligands did not retain the same conformation when bound to different binding sites, and that the binding sites were even more variable in shape than their ligand counterparts. This observation explains why shape-based approaches often struggle to identify binding sites that bind the same ligand.

Kahraman et al. (2010) subjected the same set of 100 binding sites to a more detailed analysis, including the calculation of physicochemical properties and electrostatic potentials. When comparing hydrogen bond donors, hydrogen bond acceptors, hydrophobicity scores and electrostatic potentials, it was found that hydrogen bond patterns varied the most within sets of ligands, wheras van der Waals forces varied the least. When analysing the properties of the ligands' environments, it was noted that complementarity was only observed with some averaged properties within a ligand set. On an individual pair basis, there was no specific complementarity that would enable separation of binding site pairs that bound identical ligands from those which

bound different ligands.

A second dataset was compiled by Barelier et al. (2015), focusing on unrelated proteins bound to identical ligands. A group of 116 complexes was assembled, bound to 59 ligands. This set contains 62 pairs or threes of complexes containing the same ligand, and within each group the proteins had lower than 10% sequence identity, and different PFAM (Mistry et al., 2020), CATH (Sillitoe et al., 2020) and SCOP (Andreeva et al., 2013, 2019) domain descriptions. These 62 groups were further categorised into three classes: in class A complexes, the same ligand functional groups interacted with similar protein environments; in class B complexes, the same ligand functional groups interacted with different protein environments; and in class C complexes, different ligand functional groups interacted with different (or similar) protein environments (Figure 5.2).

Barelier et al. (2015) found that class A complexes shared similar electrostatic energies within groups, and class A and B complexes had relatively similar van der Waals forces within groups. Class C complexes had very different electrostatic energies and van der Vaals forces. Energy minimisation of the protein-ligand complex did not substantially alter any of these patterns. It was concluded that since each ligand functional group was recognised in several unrelated ways, it was unrealistic to expect that evolutionarily unrelated proteins used the same residue patterns to recognise similar or even identical ligands. Therefore, any pattern-matching code for protein-ligand recognition would require a much higher level of complexity than current approaches had adopted.

5.2.3 Benchmarking of binding site comparison methods

A major benchmarking exercise of binding site comparison, known as ProSPECCTs and conducted by Ehrt et al. (2018), evaluated 15 residue-, surface- and interaction-based tools. The datasets used for testing included the Kahraman dataset (Kahraman

et al., 2007) and the Barelier dataset (Barelier et al., 2015). The findings indicated that interaction-based methods outperformed other tools on the Kahraman dataset, but all tools struggled on the Barelier dataset. In fact, only KRIPO (Ritschel et al., 2014), an interaction-based tool that extracts interactions and compared them based on known ligands, was able to achieve a prediction performance significantly better than random on the Barelier dataset. This confirms the importance of protein-ligand interactions in binding site comparison. However this approach is not applicable to novel targets lacking known bound ligands.

Both DeeplyTough and Site2Vec have since tested their method on these datasets. While Site2Vec performed well on the Kahraman dataset, DeeplyTough had much lower success. Both methods encountered significant challenges when applied to the Barelier dataset, and among the 17 methods tested, they yielded some of the least accurate predictions. The results suggest that Site2Vec is able to detect subtle similarities in fold or sequence identity, but fails to detect any similarities for entirely unrelated protein pairs.

As neither of these tools have been trained to specifically detect unrelated proteins which bind identical ligands, it is expected that they may struggle on this task; however, there is not yet enough data to enable robust training of a generalisable machine learning model. Site2Vec adopted a random data splitting strategy when compiling training and test data, so it is possible that data leakage contributed to Site2Vec's impressive performance on the Kahraman dataset.

5.2.4 Chapter summary

Previous studies have shown that neither alignment nor similarity scoring of full binding pockets is likely to produce an effective prediction of which binding sites are able to bind identical ligands, due to the variability in the binding modes observed across such pairs of complexes. Additionally, interaction-based methods appear to make

the most accurate predictions. By combining these observations, we suggest that an interaction- or pharmacophore- based method of comparing sub-pockets could improve our ability to identify binding sites from unrelated proteins which bind identical ligands. If this is the case, this provides an opportunity for the design of target-specific fragment libraries based on historic fragment screening data (Eguida and Rognan, 2022).

As ProCare was developed since the Ehrt et al. (2018) benchmarking exercise and is one of the only methods to be developed to work specifically on fragments, we first explore whether ProCare can discriminate between binding sites from unrelated proteins which bind identical or different fragments from XChem fragment screening data. We found that despite its ability to align sub-cavities into larger cavities with reasonable accuracy, the scoring was not correlated with whether a pair of binding sites bound the same or different ligands.

Having shown that learnt representations contain information on protein-ligand binding (see Chapter 4), we then study the potential for the use of learnt representations of proteins in this area, whether by unsupervised learning or by generating pocket descriptions via supervised learning. The unsupervised learning strategy was not successful, but these learnt representations were successfully used in the prediction of binding site interactions. Our results suggest that with further analysis, descriptors of binding sites could be generated using learnt representations, and could be used for comparison of pockets and prediction of whether pairs of pockets bind identical or different ligands.

5.3 Methods

5.3.1 ProCare evaluation

5.3.1.1 Dataset

A total of 922 ligand-bound structures across 23 different protein targets were downloaded from Fragalysis, using the Fragalysis API (Skyner et al., 2020) (accessed July 2020). The highest level of sequence similarity between two targets was 23%, as calculated using CD-Hit (Fu et al., 2012). The full list of targets and Fragalysis structure IDs is shown in Appendix C.

Water molecules and ions were stripped from each structure. Homopolymeric structures were separated into single chains and aligned. Redundant structures were then removed, leaving 670 unique protein-ligand complexes. Each of these protein-ligand structures were split into protein and ligand files. All protein and ligand file processing was conducted with the PyMOL API (DeLano, 2002).

5.3.1.2 ProCare Implementation

The ProCare code was obtained at `https://github.com/kimeguida/ProCare` (accessed June 2020). IChem (Da Silva et al., 2018) was used to generate the cavities and was downloaded from `http://bioinfo-pharma.u-strasbg.fr/labwebsite/download.html` (accessed June 2020). As the full FragScore metric introduced in the ProCare study uses interaction fingerprint similarity and thus requires knowledge of a bound ligand, is not applicable to novel targets with no known ligands. For this reason, we use the ProCare score throughout, which only scores pocket similarity.

Cavity generation IChem VolSite was used to generate point clouds that described the shape and pharmacophoric features of the protein environment surrounding the ligand. Pockets of various sizes were generated for each ligand using truncation of

points above a certain distance threshold from any ligand heavy atom. The resulting fragment-based cavities described the environments within 4, 6, 8, or 12Å of any ligand atom. The full ligand-bound cavity was also predicted by VolSite (referred to as 'ALL'). These binding sites are subjected to druggability predictions by the VolSite SVM classifier (Desaphy et al., 2012). Only druggable binding sites were converted to point clouds that filled the space in the cavity. This process resulted in 378 unique ligand-based point cloud cavity descriptors, while the remaining 292 structures were discarded due to the ligand being positioned in a 'not druggable' area or outwith a cavity altogether.

We then generated predicted cavities based on only the protein structure. This approach attempts to mimic a possible protocol given a novel protein target with no known ligands. In this context we had multiple structures available for each protein, and so we used all ligand-stripped structures, of which the largest predicted cavity (based on number of points) was taken forward into the cavity alignment methods, to give the best coverage of ligands and thus maximise the number of cavities that could be compared. This was also conducted using IChem VolSite. Of the 378 ligands that were successfully processed into point clouds by VolSite, 336 were within a cavity predicted by the VolSite cavity prediction method.

Within these 336 fragment-based cavities, there were 15 pairs in which the same ligand had bound to two different proteins. These 15 pairs are described as positive results where the same ligand is described by two different cavities. There were 776 negative results (pairs of proteins), in which a ligand bound to one protein, but was experimentally found to not bind the other protein.

Cavity alignment ProCare takes two point clouds as an input, a 'source' cloud and a 'target' cloud. The initial alignment is based on pharmacophoric descriptors, followed by the application of an iterative closest point (ICP) method to refine the

alignment. The quality of the alignment at each stage is estimated by two scores: fitness, which measures the overlap of source, and root-mean square error (RMSE) between corresponding pairs of points in source and target clouds. Once the best alignment has been found, the similarity between the two cavities is scored according to the Tversky similarity metric. This provides the ProCare score used in the evaluation of the method.

5.3.2 Testing ProCare's alignment and scoring performance

We tested ProCare's ability to score pockets that bind identical and different ligands, its ability to correctly place ligands in cavities, and the potential of ProCare to be used as a virtual screening method.

5.3.2.1 Pocket similarity of cavities which bind identical ligands

In order to be a useful tool for compound prioritisation, ProCare must be able to discriminate between cavities to which the same ligand is able to bind and those which bind different ligands. We compared each of the 15 positive pairs twice, using each one as the 'source' cavity once. As the smallest fragment-based cavity (<4Å from any ligand atom) can be regarded as describing only the immediate environment of the ligand, we first obtained scores for only this cavity size. As a comparison, 6 negative fragment-based cavity pairings - which describe binding sites of different ligands - were randomly chosen (from the total set of 776 negative pairs) for each positive result. Any cavity derived from either target in the positive pairing was excluded. Once these filters had been applied, we were able to obtain scores for 11 positive cavity pairings, and a total of 66 negative cavity pairings. We show the protein sequence identity and Tanimoto similarity of ligands in each pairing in Figure 5.1 (fragalysis dataset).

If the cavity describing the ligand's environment is similar in both targets for the

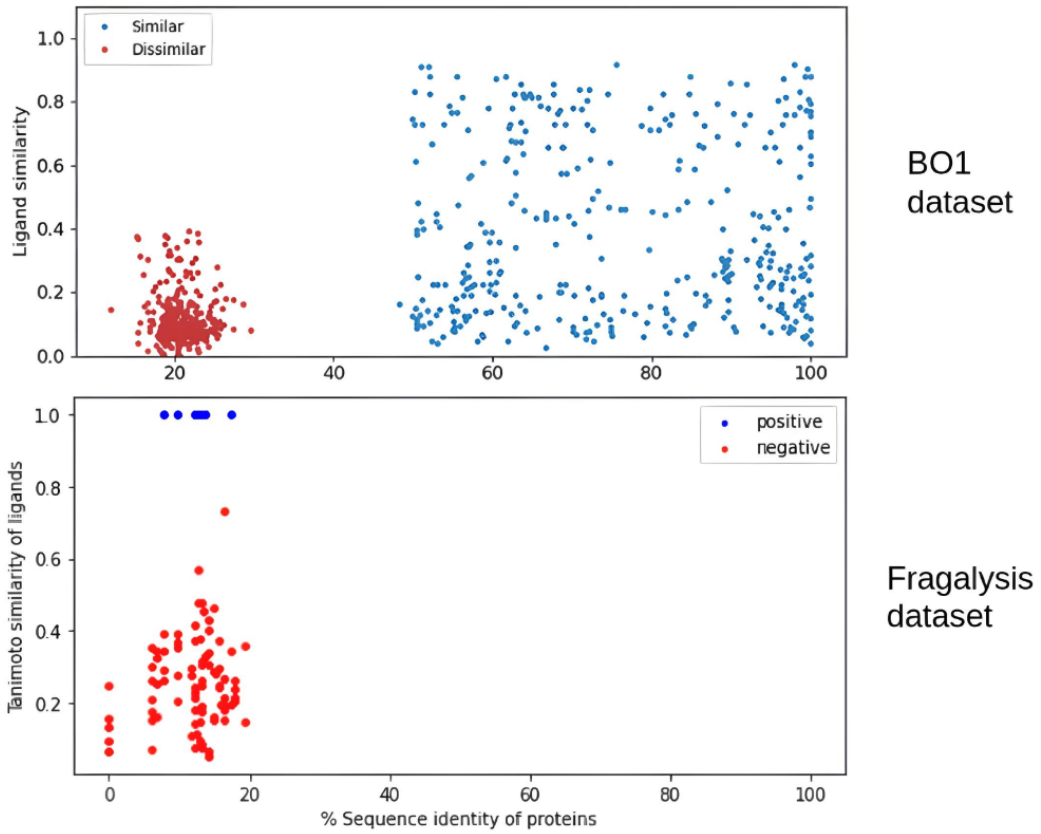


Figure 5.1: **Above:** adapted from (Eguida and Rognan, 2020). Pairings of similar and dissimilar pairs of cavities used to develop the ProCare scoring function, plotted by percentage global sequence identity of target and ligand similarity (Tanimoto coefficient of the Morgan fingerprints of the ligands). **Below:** pairings of cavities which describe the same ligand binding to unrelated targets (positive) and pairings which describe different ligands binding to unrelated proteins (negative), plotted by the same parameters as above.

positive results, then we would expect the ProCare scores for these pairings to be higher than those for the negative results. This experiment was repeated using all cavity sizes, with the best score for each pairing taken as a measure of the similarity of the two fragment-based cavities.

5.3.2.2 Alignment of smaller sub-cavities into larger cavities

To test ProCare's ability to align smaller fragment-based cavities into the larger predicted cavities, we aligned fragment-based cavities with their original target. These are described as self pairings. This allowed the details of the pharmacophoric descrip-

tors of the points and the scoring method to be controlled for, and only the alignment method was tested. First, the smallest fragment-based cavity was aligned into the larger cavity, and the required transformation and translocation was applied to the ligand. The RMSD between the original ligand position and the realigned ligand position was calculated, and cavities were regarded as correctly aligned where ligands were within 2Å of the original position. This was then repeated three times to calculate the variation in RMSD scores when using the same cavity sizes. All other fragment-based cavities (<6Å, <8Å, <12Å and full cavity) were subsequently realigned and the best RMSD was taken to view the improvement upon increased overlap in sizes of cavities, and to test whether there was a larger variation in alignments when changing cavity sizes compared to repetition of the same cavity size alignment.

5.3.2.3 ProCare scores of positive and negative sub-cavities aligned into larger cavities

We obtained ProCare scores for the alignment of fragment-based cavities into the predicted cavities for positive, negative and self pairings. This was first performed using only the smallest fragment-based cavity and then we repeated the process with all cavity sizes, taking only the highest score for each pairing.

If positive pairings obtain higher scores than negative pairings on average, then prioritising fragments with a high ProCare score to the target should increase the hit rate for a screen. This would improve the amount of information that can be obtained about a protein target from a single screen, and facilitate the design of follow-up compounds.

In addition to the ProCare score, we calculated the RMSD between the original position of the ligand in the target and the aligned ligand from the fragment-based cavity of the unrelated protein for positive pairings only.

5.3.3 ESM embeddings

As in Chapter 4, we used ESM-IF1 (Hsu et al., 2022) embeddings for the representation of binding site residues (see Section 4.3.2). We additionally used ESM2 embeddings (Lin et al., 2023) for comparison. These are both learnt vector representations of proteins in a per-residue format (see Section 4.2.2). We selected these embeddings as they contain complementary information regarding the protein sequence and structure: ESM-IF1 does not explicitly contain sequence-related information, and only uses spatial information to describe the environment of each residue; conversely, ESM2 is a sequence-based embedding and so only residue type is explicitly contained.

5.3.4 Distance-based pocket similarity comparison

To evaluate the potential for learnt representations of residues to be used in binding site comparison methods, we calculated the distances between residues that bind identical ligands and compared this to the distances between residues that bind different ligands.

5.3.4.1 The Barelier Dataset

The Barelier dataset is composed of pairs (or threes) of unrelated proteins that bind identical ligands (Barelier et al., 2015). It is formed of three classes which together contain 59 ligands in 116 complexes. In class A pairs, the same ligand functional groups interact with similar protein groups in each site; in class B pairs, the same ligand functional groups interact with different protein groups in each site; and in class C pairs, different ligand functional groups interact with different protein groups in each site (Figure 5.2).

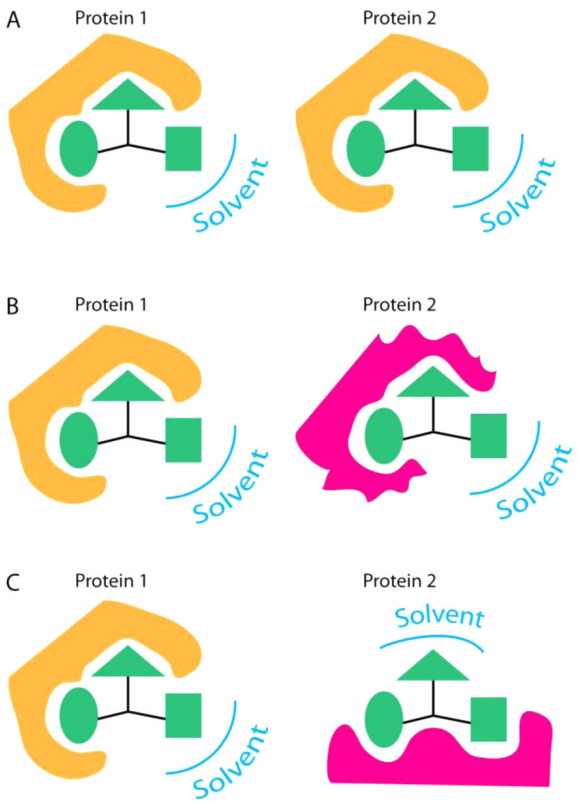


Figure 5.2: The Barelier dataset is made up of three classes of pairs of unrelated proteins bound (yellow, pink) to identical ligands (green). The protein pairs are categorised into different classes based on the similarity of which ligand functional groups are used to bind protein functional groups, and which types of protein functional groups they bind to. Adapted from Barelier et al. (2015).

5.3.4.2 Comparison method

Within each class of protein pairs from the Barelier dataset, we first extracted every binding site residue from every protein, defined as having any heavy atom within 5Å of any ligand heavy atom. We generated the ESM-IF1 and ESM2 embeddings for each of these residues, and calculated the euclidean distance and cosine similarity between every pair of residues that came from different proteins. If the residues came from structures that were in the same pair (i.e. bound to an identical ligand) they were labelled as a 'match'.

5.3.5 Prediction of protein-ligand interactions

As interaction-based methods performed well in the ProSPECCTs analysis (Ehrt et al., 2018), we explored whether ESM-IF1 or ESM2 embeddings can be used for the prediction of which residues make specific interactions with bound ligands. As a baseline comparison we also used one-hot encoding of residue types.

5.3.5.1 Test and training data

As our test set, we used all 279 complexes present in the CASF-2016 core set. We used PDBBind (Liu et al., 2014) as a source of protein-ligand complexes for training. The 2020 version of PDBBind contains 23,496 protein-ligand complexes which have associated binding affinity data, mined from the PDB (Berman et al., 2000). We used only complexes from the dataset used to train PointVS (Scantlebury et al., 2023), as this contains complexes with less than 80% sequence identity and less than 0.8 ligand Tanimoto similarity to the CASF-2016 core set. Our full training set was made up of 14,495 complexes. To explore whether binding affinity of training complexes affects prediction accuracy, we used the median affinity to further split the training set into 7,219 higher-affinity complexes and 7,276 lower-affinity complexes.

Interaction type	Label	Defined by
None	0	within 4Å of ligand heavy atom but not making an interaction
Hydrophobic	1	ODDT interaction
Hydrogen bond donor	2	ODDT interaction
Hydrogen bond acceptor	3	ODDT interaction

Table 5.1: Interaction types as calculated by ODDT, with their label for training and testing. For details of how interaction types are defined see Table 1.1

5.3.5.2 Protein-ligand interaction calculation

As in Chapter 3, the interactions in each structure within the dataset were calculated using ODDT's InteractionFingerprint module (Wójcikowski et al., 2015), which generates a binary fingerprint for each protein-fragment structure. This method calculates up to eight types of interaction between the fragment and each residue in the protein (see Table A.2 for details of interaction types). Based on the calculated interactions, each residue was given a label as shown in Table 5.1. Residues that were within 4Å of any ligand heavy atom that were not making any interactions were given a negative label.

5.3.5.3 Model training

For all training sets and representations we trained multi-label LightGBM classifiers with default parameters. We began by training four models that used the same datasets but with different representations: ESM-IF2, ESM2, combined (ESM-IF1 and ESM2 vectors fused together) and one-hot as a baseline comparison.

To explore the effect of training dataset size on prediction accuracy, we trained models with increasing proportions of the training data. As high-affinity complexes may contain stronger interactions that provide a clearer signal for the models, we additionally trained models on only the higher-affinity half of the complexes and only the lower-affinity half of the complexes (split using the median affinity value in the

	High affinity	Low affinity	Full
Complexes	7219	7276	14495
Binding site residues	112495	93313	205808
Label support (0, 1, 2, 3)	(68999, 29201, 7114, 7181)	(58587, 19830, 7778, 7118)	(127586, 49031, 14892, 14299)

Table 5.2: Numbers of complexes and residues for each label in high-affinity, low-affinity and full datasets.

dataset), again with increasing volumes of training data. We trained 10 models for each training dataset size.

At lower dataset sizes, we sampled equally from each label. Once the training set reached the maximum size for the least represented label (7,114 residues for higher-affinity label 2), we used upsampling (repetition of data points) to retain the balance of different labels in the dataset while enlarging the dataset. This led to a maximum of 14,228 data points being included for each label, where for labels 2 and 3, most data points were repeated. Conversely, for labels 0 and 1, the 14,228 data points were a sample of a much larger dataset. As a result, the largest training sets for higher- and lower-affinity models contained 56,912 residues, and the combined model was trained on a maximum of 113,824 residues. Models were trained using all four representations. For each training dataset size five models were trained using different random samples of the training data and the order of data points was randomised. The mean F1 score was taken for each training dataset size.

5.3.5.4 Evaluation

Our test set consisted of 279 complexes from the CASF-2016 core set. Upon calcuation of interactions using ODDT, the test set contained 1949 residues with label 0, 973 residues with label 1, 355 residues with label 2 and 313 residues with label 3.

We evaluated our models using the balanced F1 score which takes into account the different numbers of data points for different labels in the test set. Our baseline

Name	Structures contained in dataset	Used in
D1	11 positive pairs with ProCare scores for all cavity sizes	Score comparison and ranking for ligand-based cavities
D2	15 positive pairs with ProCare scores for only <4Å cavity sizes	Score comparison for ligand-free cavities; alignment of subcavities into larger cavities
D3	336 proteins with druggable fragment-binding sites	All self and negative pairings for all analyses

Table 5.3: The sets of structures used in our analyses of ProCare.

was set using a random prediction, which had an associated F1 score of 0.22.

5.4 Results

5.4.1 ProCare evaluation

We started with 1,131 protein-ligand complexes, which were reduced to 670 via sequence-based clustering to ensure diversity in our test set. However, the binding site generation procedure (from VolSite) used by ProCare reduced this number to 378, as the ligands in 292 of the complexes were classed as bound to non-druggable binding sites. When generating site descriptors in a ligand-free manner, only 336 of 378 complexes had sites containing the ligands. This resulted in 15 pairs of binding sites that bound to the same ligand (positive pairs), however ProCare was not able to generate scores for four of these pairs when comparing ligand-based cavities, resulting in a final dataset size of just 11 positive pairs for the analysis of ProCare scores of ligand-based cavities. We summarise the different sets of structures used in Table 5.3.

5.4.1.1 Pocket similarity of cavities which bind identical ligands

We obtained ProCare scores from 11 pairs of unique ligand-based cavities, in which the members of each pair were generated from the same ligand binding two unrelated proteins (see Section 5.3 and D1 in Table 5.3).

We would expect the ProCare score for the 11 positive cavities to be higher than for the negative pairs, as ProCare score quantifies pocket similarity, and pockets that bind the same ligands would be expected to be more similar to each other than pockets that bind different ligands. However, the positive pairings did not produce significantly higher scores than negative pairings (Figure 5.3A-B). To determine whether ProCare scores are ligand-dependent and therefore require ligand-specific comparisons, we compared the score for each of the 11 cavities that had positive pairings with 6 negative pairings containing the same cavity. These scores were ranked, and we would expect the positive pairing to get the highest score, and therefore obtain the top rank.

In these comparisons we would expect to gain the best scores when comparing the smallest cavities (<4Å from ligand heavy atoms), as these describe only the micro-environments in which the same ligand has been able to bind to two unrelated proteins. However, positive pairings did not rank any higher than negative pairings (Figure 5.3C). The results obtained when comparing all cavity sizes and taking the best score led to positive pairings ranking higher (Figure 5.3D), which indicates that the model is rewarding areas which we know to bind to different ligands more than areas to which identical ligands have bound.

5.4.1.2 Alignment of smaller sub-cavities into larger cavities

To test ProCare's ability to align sub-cavities into largest pockets, 366 unique fragment-based subcavities were aligned with the larger predicted cavity from the same protein (see Section 5.3 and D3 in Table 5.3). Figure 5.4 shows the RMSD values obtained when realigning the smallest sub-cavity (<4Å from any ligand atom) with the predicted cavity and the best RMSD obtained when realigning all fragment-based subcavities with the predicted cavity.

A single attempt at realigning the smallest cavity was successful for 60.4% of

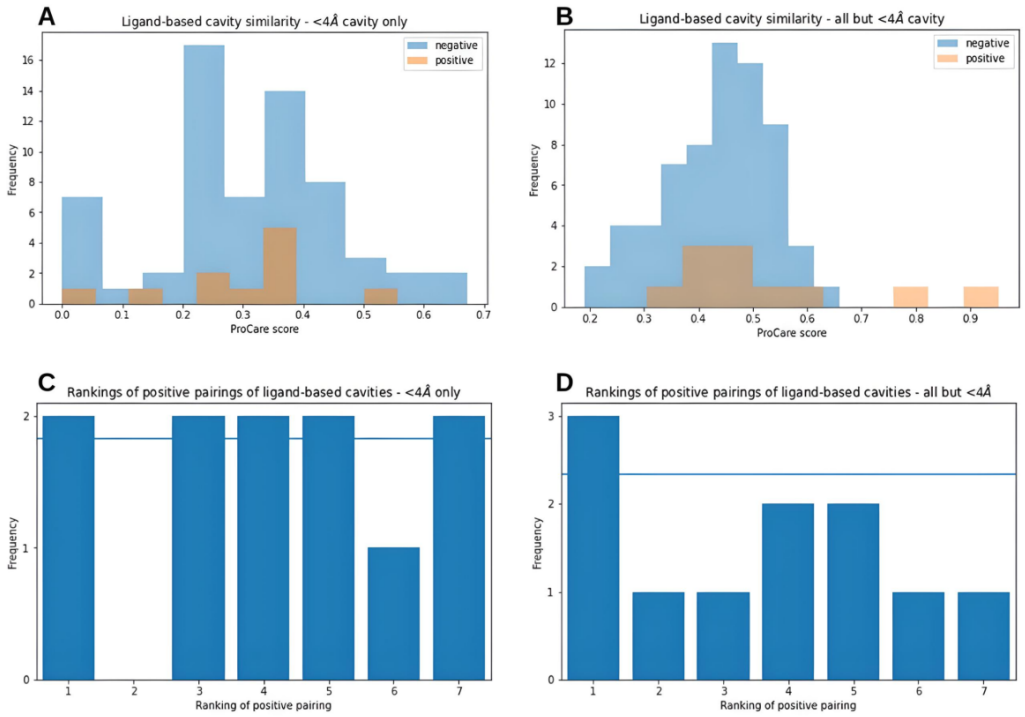


Figure 5.3: Pockets that bind identical ligands do not get higher ProCare scores than those which bind different ligands. ProCare scores obtained by positive (D1) and negative (D1 and D3) pairings when aligning only <4Å cavities against each other **(A)** and when aligning all other cavities against each other **(B)**. The rank obtained by the positive pairing's ProCare score when compared with 6 random negative pairings' ProCare scores, when aligning only <4Å cavities **(C)** and when aligning all other **(D)**. Blue horizontal line represents the mean height expected from a random distribution.

cavities (final ligand position within 2Å of the original position). This rose to 64.4% after four attempts; a small increase which reflects the lack of variability in the alignment method. Single attempts of alignment of all other cavity sizes with the predicted cavity were successful for 75.2% of cavities. As some ligands had as few as half of their atoms within 2Å of the predicted cavity, we checked whether these ligands were more difficult to realign. We compared the proportion of ligand atoms within 2Å of the predicted cavity with the best RMSD value obtained (Figure 5.5). We found that cavities generated from ligands with <65% of their atoms within 2Å of the predicted cavities more difficult to realign, but many ligands that resided fully within the predicted cavity also produced the same issues. A possible explanation for

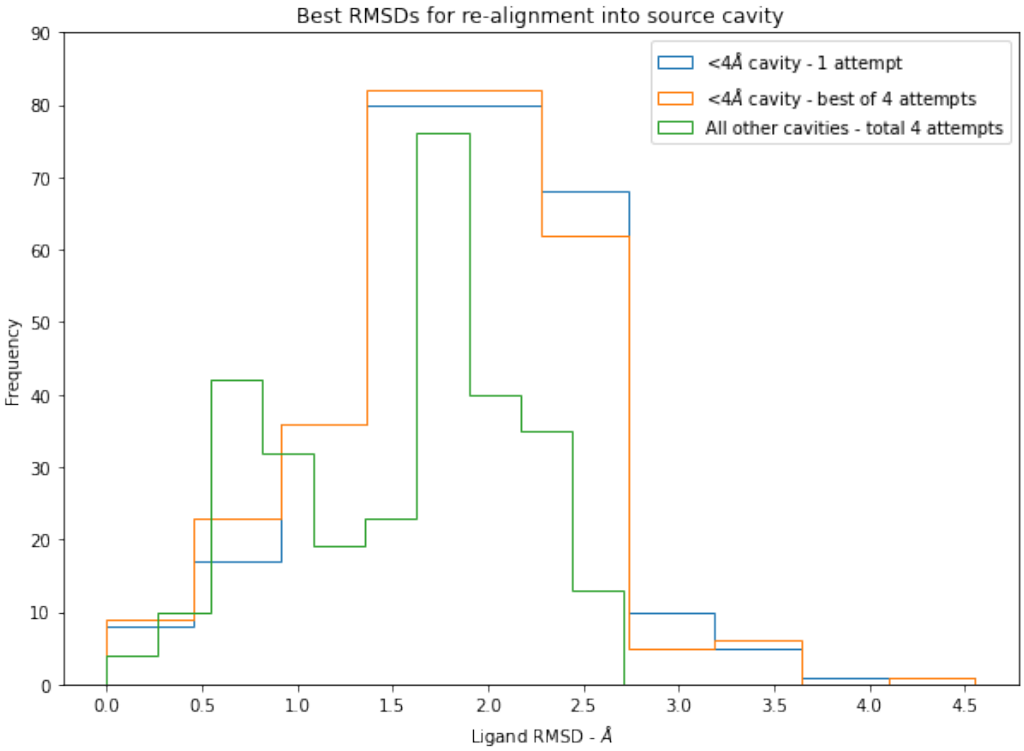


Figure 5.4: Ligand RMSD values when realigning fragment-based cavities into the predicted cavity of the target are lower when a larger cavity is used. Repeated attempts to realign the <4Å cavity into the predicted cavity slightly reduce RMSD values compared with a single attempt. Structures were taken from D3.

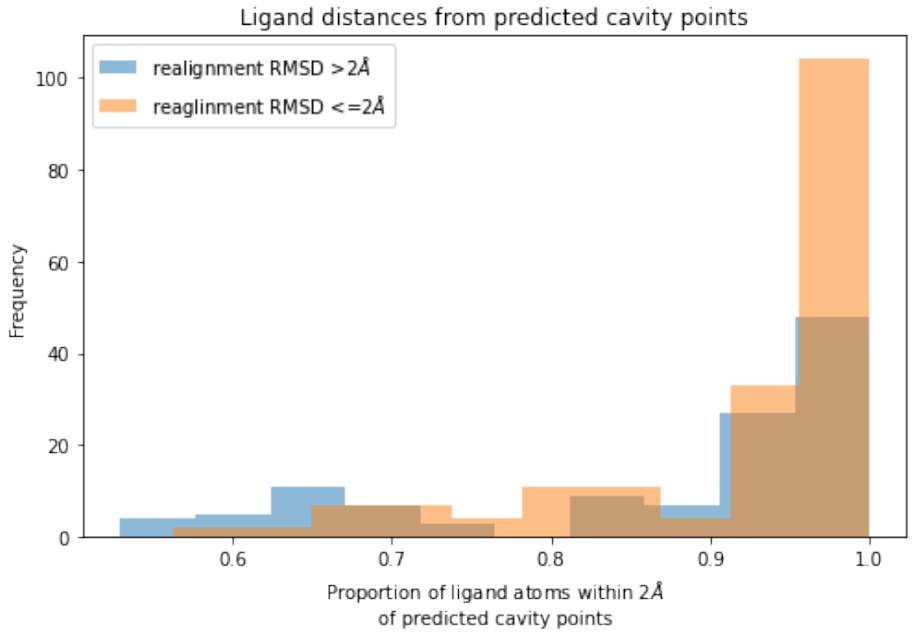


Figure 5.5: The proportion of ligand atoms within 2Å of any predicted cavity point for sub-cavities from D3 that were realigned within 2Å of their original position (orange) or not within 2Å of their original position (blue) slightly reduces realignment success. Many sub-cavities that are completely within the predicted cavity were not aligned within 2Å.

this could be global similarity being rewarded more than small areas of local similarity in the fragment-binding site.

5.4.1.3 ProCare scores of positive and negative sub-cavities aligned into larger cavities

We obtained ProCare scores for the alignment of a fragment-based cavity into a predicted cavity for 15 positive (D2) and 199 negative cavity pairings (taken from D2 and D3) across all 23 targets, and compared them. As a reference, we also obtained the scores from the realignment of the 297 remaining fragment-based cavities from D3 into the predicted version of its source cavity (self pairings, see Table 5.3). This was done both with the smallest fragment-based cavity (<4Å from any ligand heavy atom) and with the best scores from all fragment-based cavity sizes (Figure 5.6). While the self pairings score higher than the positive and negative pairings, the fact

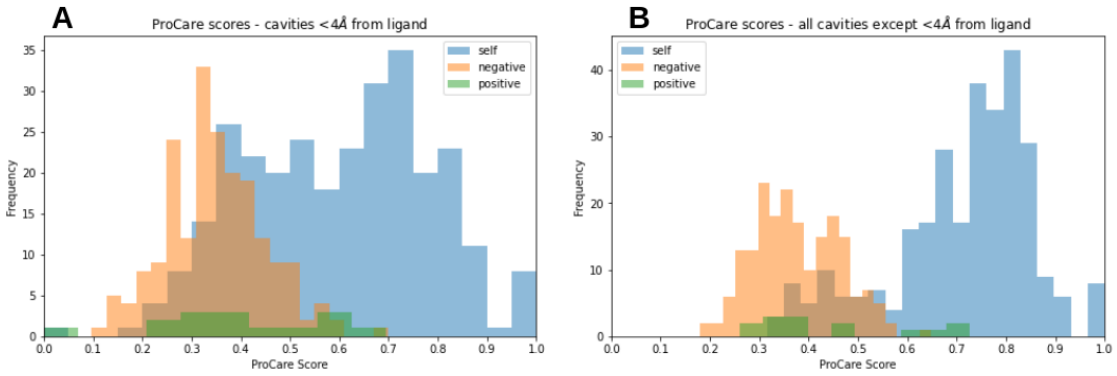


Figure 5.6: ProCare scores for 15 positive (green, D2), 199 negative (orange, D2 and D3), and 297 self (blue, D3) pairings of sub-cavities aligned into larger predicted cavities show that positive pairs do not score significantly higher than negative pairs. This was performed with only the smallest sub-cavity **(A)** and with the best score taken from alignment of all other sub-cavities with the predicted cavity **(B)**.

	ProCare scores			
	Negative median	Positive median	Negative mean	Positive mean
<4Å only	0.34	0.35	0.35	0.40
All but <4Å	0.37	0.37	0.38	0.45
All	0.39	0.43	0.40	0.47

Table 5.4: Mean and median values of the best ProCare scores obtained from each set show a small improvement in the scores of positive pairs compared with negative pairs where all cavity sizes are used.

that they do not consistently score above all positive and negative pairings (e.g. all scores ≥ 0.7) shows that the alignment method is inherently flawed.

In Table 5.4 the mean and median ProCare scores for positive and negative pairings are shown, for <4Å and remaining cavity sizes. The scores for positive pairs are slightly higher than for negative pairs, however the difference is small. Using this full set of data, we generated a precision-recall curve to examine ProCare's ability to reduce the number of possible pairings to screen while retaining the positive pairings (Figure 5.7). This showed that ProCare has a poor ability to correctly classify the cavity pairings in this dataset: this may be due to the small amount of data used, or it could be reflective of ProCare's skill.

We then analysed the alignment of positive pairings of sub-cavities into the larger

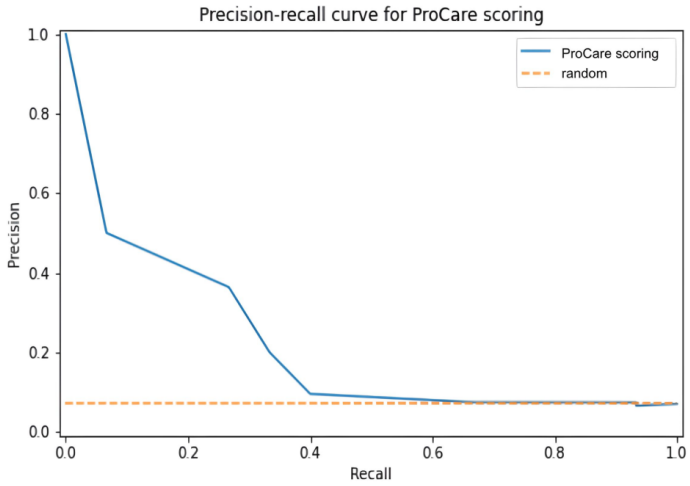


Figure 5.7: As recall increases, precision falls quickly, indicating a poor ability to correctly classify cavity pairings. Precision and recall when using the best ProCare score for each pairing.

predicted cavity for structures from D2 and D3 (see Table 5.3). The full results are shown in Figure 5.8A, in which we compare the RMSD values obtained from positive pairings to those from self pairings. For ease of comparison, we have normalised the frequencies. We observed that 56% of alignments resulted in the ligand being placed within 3Å of the correct position. As there are likely to be some differences in conformation and the ligands are being treated as completely rigid, we took this into account by calculating the RMSD of the best alignment of the two conformations of the ligand, and subtracting this from the RMSD of the ProCare-predicted alignment (Figure 5.8B). This comparison highlights that over 50% of the positive pairings result in alignments of less than 2Å (from the best possible alignment), which is well within the range of values seen for self pairings and therefore can be regarded as a 'correct alignment'. However, RMSD values of over 2.5Å (35% of positive pairings) are likely to indicate significant misalignments.

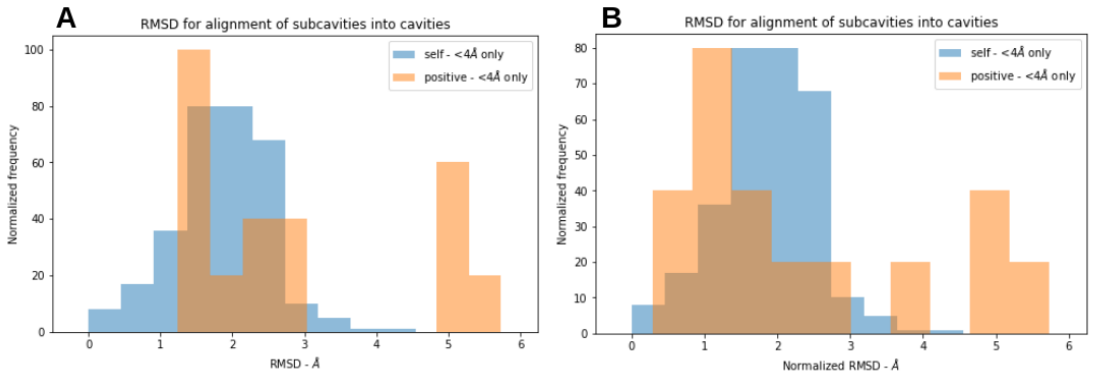


Figure 5.8: Normalised RMSD values show that over 50% of alignments are within 2Å of the optimal alignment for structures from D2 and D3. **A:** The RMSD values of the ligand aligned into the cavity compared with the correct ligand position. **B:** RMSD values for positive pairings have been normalised to take into account the conformational differences between the two structures of the same ligand.

5.4.2 Use of ESM2 and ESM-IF1 embeddings for pocket similarity

5.4.2.1 Identical ligands do not bind residues with similar embeddings

To explore the potential for ESM embeddings to be used for clustering of binding site residues that bind identical ligands, we calculated euclidean distance and cosine similarity between residues from unrelated proteins that bound identical and different ligands. If the embeddings between residues from different proteins which bound identical ligands were closer in space to each other compared with those which bound different ligands, it would enable the prediction of ligand binders for novel targets.

However, we found that no embedding or distance combination that was tested satisfied these criteria (Figures 5.9 and 5.10). For both ESM-IF1 and ESM2 embedding types and for both euclidean distance and cosine similarity, none of the residues that bound identical ligands (orange, labelled as matches) had values outlying from the main distributions of residues that bound different ligands (blue). In particular, we expected that the residues bound to identical substructures of identical ligands would have lower euclidean distances or higher cosine similarities to each other than

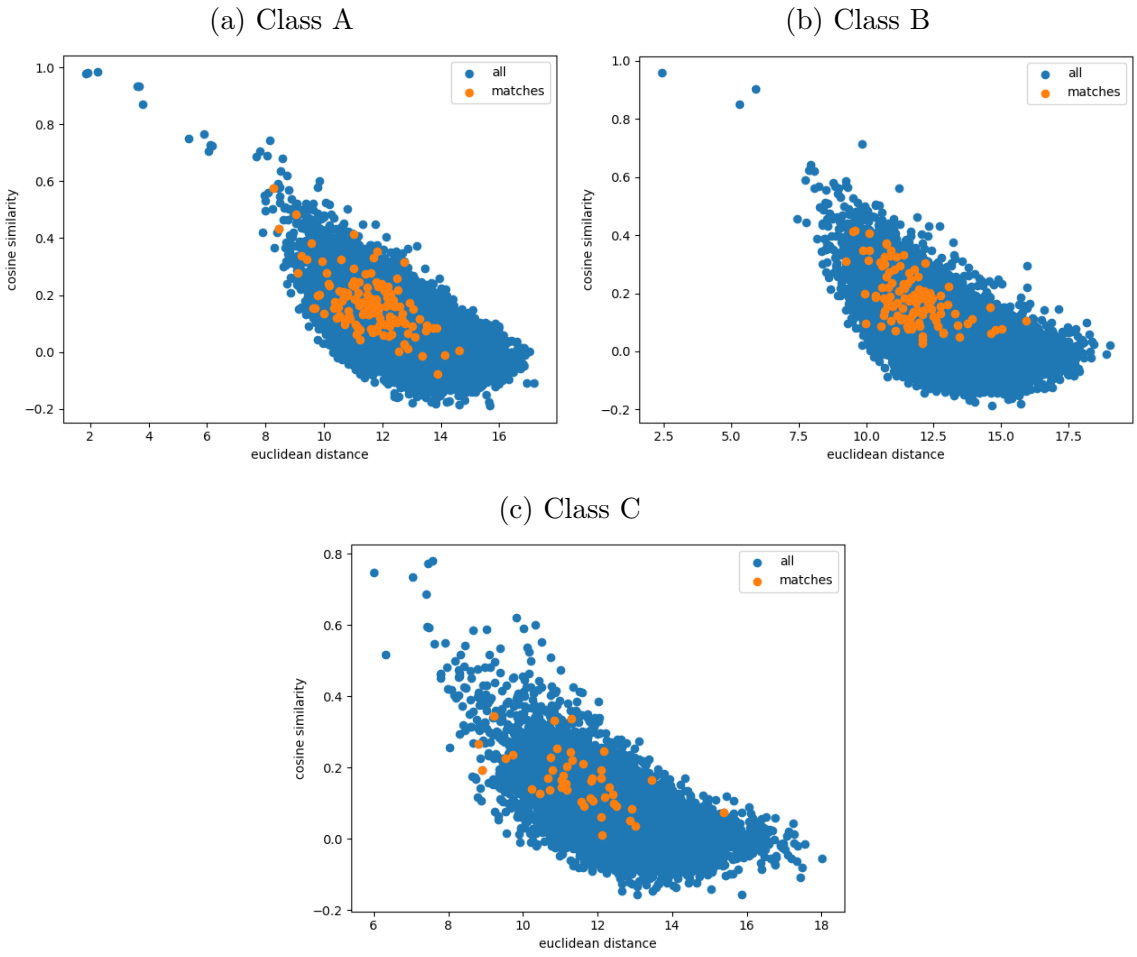


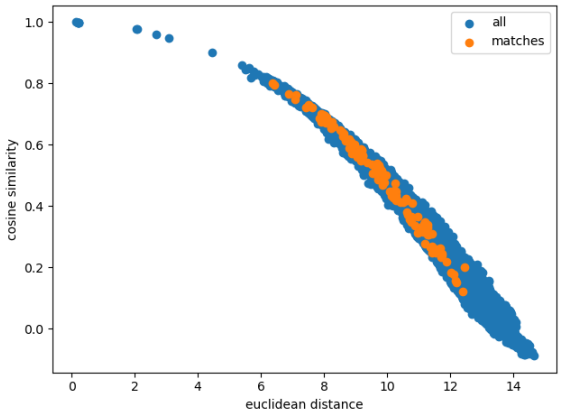
Figure 5.9: Residues that bind identical ligands (orange) do not exhibit higher cosine similarity or lower euclidean distance of ESM-IF1 embeddings than those which bind different ligands (blue). This is the case for all three classes of protein pairs from the Barelier dataset.

residues bound to different ligands, however this was not consistently observed.

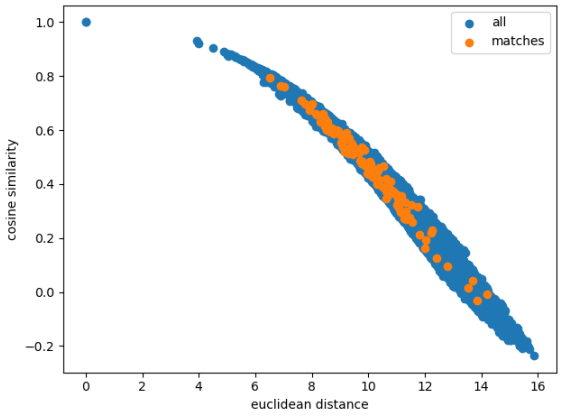
5.4.2.2 ESM embeddings outperform one-hot embeddings in prediction of interaction types

We set out to test whether learnt representations of proteins can be used to predict interaction types in binding pockets as a means of facilitating binding site comparisons. Multi-label classification LightGBM models were trained on one-hot, ESM2, ESM-IF1 and combined (ESM2 and ESM-IF1) embeddings (Figure 5.11). For training set sizes larger than 500 residues, using ESM-based representations led to higher F1

(a) Class A



(b) Class B



(c) Class C

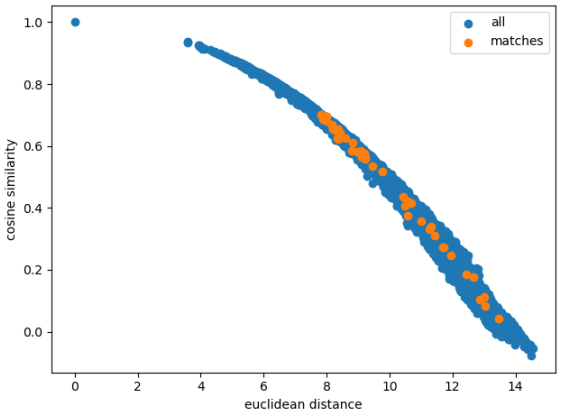


Figure 5.10: Residues that bind identical ligands (orange) do not exhibit higher cosine similarity or lower euclidean distance of ESM2 embeddings than those which bind different ligands (blue). This is the case for all three classes of protein pairs from the Barelier dataset.

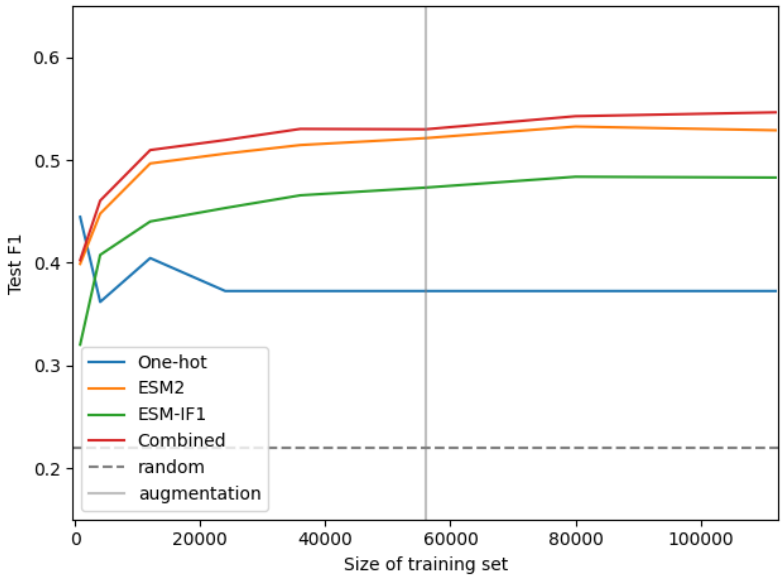


Figure 5.11: The greatest F1 scores in predicting interaction type can be achieved by using ESM2 (orange) and combined embeddings (red) to represent protein residues. The random baseline is shown as a grey dashed line, while the predictions made with one-hot encoding are shown in blue. Training set sizes larger than 56000 residues (right of the grey solid line) contain upsampling of the less supported labels in the dataset.

scores than one-hot representations. ESM2 embeddings outperform ESM-IF1 when used individually, but by combining these representations, we further improved the F1 score. At the largest unaugmented training set size, the combined model achieved an F1 score of 0.53, ESM2 achieved 0.52, ESM-IF1 achieved 0.47 and one-hot achieved 0.37.

All models improved further upon augmentation of the training data (via upsampling of less represented interaction types), with models training on ESM-IF1, ESM2 and combined embeddings achieving F1 scores of 0.49, 0.53 and 0.55 respectively.

5.4.2.3 High-affinity complexes contain slightly stronger signal than low-affinity complexes

By splitting the training data into high-affinity and low-affinity complexes, we explored whether high-affinity complexes contain interactions that translate to a stronger

signal from which ML models can achieve a higher predictive power.

We found that while the model trained on one-hot encodings appeared to extract more signal from lower-affinity complexes, all models trained on ESM representations (ESM-IF1, ESM2 and combined) had a superior prediction performance when using only high affinity complexes. This difference in performance was most pronounced where the training set was larger and where the representation size was largest (ESM2 and combined). When compared with the combined model trained on 28000 datapoints, a model trained on a mixed affinity dataset of the same size had marginally lower performance (F1 scores of 0.53 and 0.52 respectively). The model trained on the largest number of datapoints using the combined representation achieved an F1 score of 0.54, slightly lower than what was achieved using a larger dataset of mixed-affinity complexes (0.55).

While these results suggest that high-affinity complexes can be used to train models that are marginally more accurate, the improvement is not enough to outperform larger mixed-affinity datasets. A recent study of the PDBBind database has shown that the affinity data may be unreliable Li et al. (2023), so models trained on data with more reliable binding affinity values may exhibit different behaviour.

5.5 Conclusions

We began by evaluating ProCare's ability to discriminate between cavities from unrelated proteins that bind identical ligands and those which bind different ligands. The data preparation process significantly reduced the number of usable data points, as the VolSite SVM classifier predicted over 40% of the fragment-bound sites to be undruggable. The druggability prediction uses a SVM classifier on 73 features including cavity size, representation of pharmacophoric features and solvent accessibility, to determine whether the shape and chemistry of the site is suitable for a drug to bind

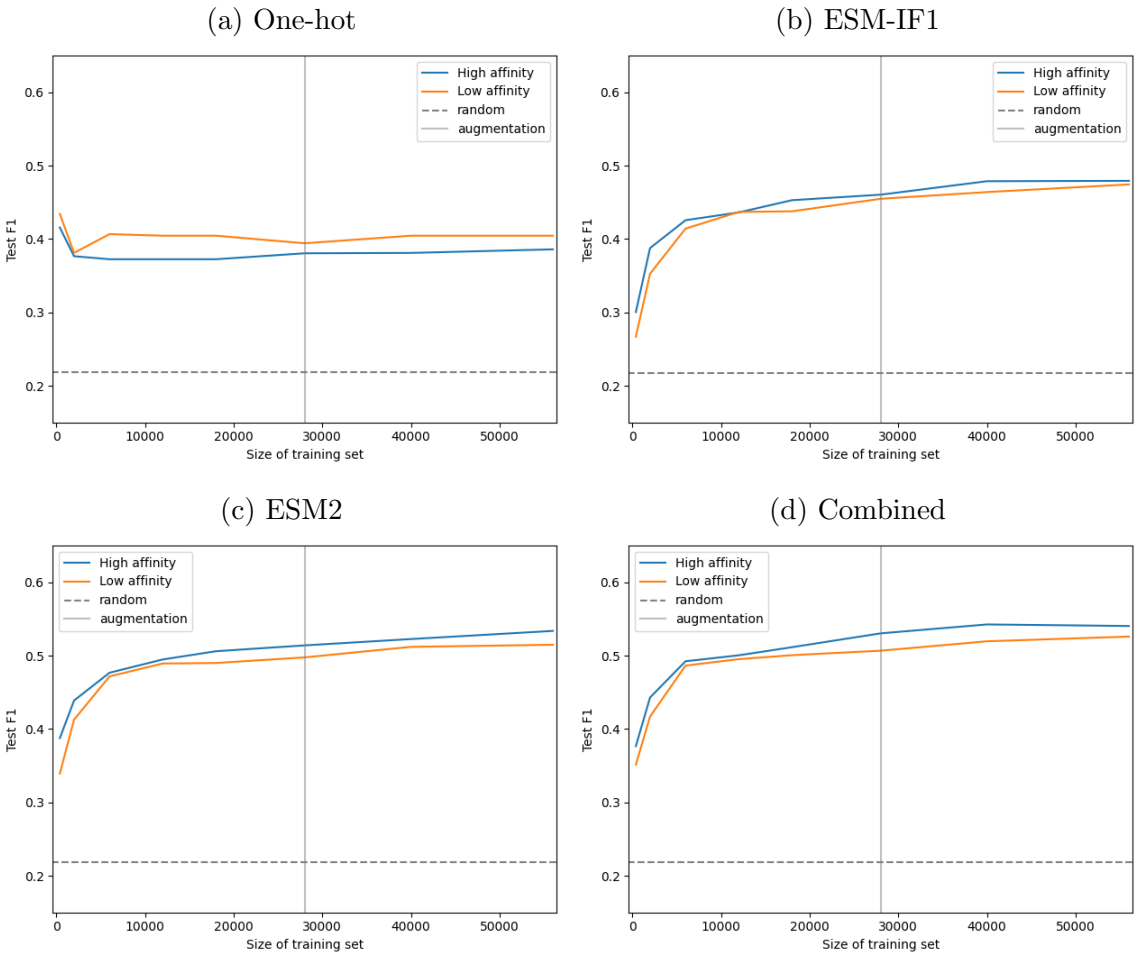


Figure 5.12: Performance comparison of models trained on high-affinity (blue) or low-affinity (orange) complexes shows that all models trained on ESM embeddings benefit from higher-affinity training data. The random baseline is shown as a grey dashed line. Training set sizes larger than 28,000 residues (right of the grey solid line) contain upsampling of the less supported labels in the dataset.

(for example, 'flat' binding sites without adequate volume are not deemed druggable). While ProCare was able to align smaller sub-cavities into larger cavities with reasonable accuracy, the scores assigned to sub-cavities being realigned into the larger cavity from the same site ('self' pairings) were lower than expected, showing some inherent limitations in the alignment method used. When comparing unrelated protein binding sites, we found that ProCare does not assign significantly higher scores to cavities that bind the same ligand than to cavities that bind different ligands. This results in a poor predictive power, meaning that ProCare would not be able to effectively prioritise fragments that are more likely to bind a novel target.

We then calculated the distances between ESM embeddings for residues that bound identical and different ligands to evaluate whether this approach could be used in pocket similarity scoring. However, we found that euclidean distance and cosine similarity were no different for residues bound to identical ligands compared with those that bound to different ligands. This task perhaps requires further representation learning to extract the information relevant to ligand binding, such as that conducted in Gazizov et al. (2023).

Finally, we adopted a similar approach as in Chapter 4 to train a LightGBM model for prediction of binding site features based on ESM embeddings. Here we trained a multi-label classifier to predict whether binding site residues made interactions with bound ligands, and if so, what type of interaction was made. We found that ESM embeddings outperformed one-hot encodings, further reinforcing our previous conclusion that learnt representations of proteins are a rich source of information on ligand-binding activity of residues.

We additionally trained models on only high-affinity or low-affinity complexes from the training set and showed that while training on only high-affinity complexes results in slightly improved performance, a larger dataset (which contains complexes with mixed binding affinities) is superior in this context.

In this chapter we have explored potential approaches for calculating pocket similarity with the aim of selecting fragments which are likely to bind to a novel target, and thus enable target-specific library design. We showed that while learnt representations such as ESM2 and ESM-IF1 are not able to be used in their raw form for identifying pockets from unrelated proteins which bind unrelated ligands, they do contain information on interactions made with ligands, indicating that they are a useful tool in addressing this area of research.

Chapter 6

Conclusions and Future Work

The processs of drug design is slow and expensive, and most potential drugs never make it to market. Fragment-based drug design and machine learning approaches are two promising additions to the field of rational drug design which aim to reduce these issues of time and cost. However, the design of fragment libraries has not been critically analysed in the light of experimental results, and structurally-diverse fragment libraries remain the norm. This thesis has focused on developing strategies for the design of both target-agnostic and target-specific fragment libraries that more thoroughly explore the binding site of the protein of interest.

6.1 Data availability

Due to their small mass, fragments must bind to proteins with high efficiency in order to be detected. This is in contrast to larger compounds, which can be detected by X-ray crystallography even where binding efficiency is low. Therefore, protein-fragment complexes are a much richer source of information about the mechanisms and rules of how small molecules bind to protein binding sites, and so could be used in the training of machine learning-based virtual screening tools.

However, full results from fragment screens are not commonly publicly available.

Even where the protein-fragment complexes arising from a fragment screen are made available, this does not come along with the also important list of fragments which did not bind to the protein. In Chapter 2 we analysed the data preserved from fragment screening projects in the period from 2015 to 2020, with a view to using the resulting data to train machine learning models.

However, we found flaws in both the structures that were accessible internally, and in the data curation and annotation protocol. To address this, we developed a new storage and automated backup procedure for electron density maps, coupled with restructuring of the dataset to facilitate large-scale analyses. This data has already been used in the development of PanDDA2 (a follow-up to Pearce et al. (2017)) (C. Wild private communication), and there are several additional opportunities for further development, such as the automated modelling of ligands into electron density maps and the validation of automatically-built or human-built ligand models.

The generation of a large, high-quality set of protein-fragment structures has potential to facilitate the training of higher-accuracy machine learning models for virtual screening and other CADD applications.

6.2 Structural diversity and functional diversity

The traditional strategy for fragment library design hinges on the philosophy that structurally diverse compounds are the most likely to extract the most diverse binding information from the protein. However, in Chapter 3, we showed that structural diversity of a fragment library does not necessarily confer functional diversity. Moreover, the data-driven selection of functionally diverse fragments results in fragment libraries that recover more diverse information from new, unseen targets.

This work demonstrates the potential for new strategies for the design of efficient target-agnostic fragment libraries. For example, our proposed method could be ap-

plied to larger datasets, which could result in new libraries with even higher efficiency of information recovery. Another avenue for further study in this area is to explore how different definitions of functional diversity impact the effectiveness of the method.

6.3 Target-specific library design

Target-specific fragment library design is a process which uses knowledge of the protein of interest to guide the selection of fragments for screening. This is common practice for proteins that have been previously screened and so ligand-based virtual screening can take place. However, FBDD is a strategy commonly selected for proteins which have been found to be intractable, and so this type of ligand-based information is unlikely to be available. In the absence of this, a crystallographic or predicted protein structure is the main source of information that can be leveraged for the selection of fragments that are likely to bind the target. In the latter chapters of this thesis we focused on potential strategies for this type of target-specific fragment library design.

6.3.1 Binding site prediction on predicted protein structures

Having described a new binding site prediction method that uses learnt representations of protein residues, IF-SitePred, we evaluated it on crystallographic and AlphaFold-predicted structures and compared it to three existing tools. The ML-based tools were all able to maintain accuracy on the AlphaFold structures, but we found that the accuracy of these AlphaFold structures was much higher than expected for novel targets. This can be explained as a form of data leakage: as available test proteins are (by definition) publicly available on the PDB, they (or close homologues) are included in the training data for AlphaFold. This data leakage is an issue likely to be faced by most studies evaluating performance of ML models on AlphaFold or

other computationally-predicted protein structures.

We attempted to address the data leakage by using MD simulations to reduce the accuracy of the predicted structures. Testing of the ML-based binding site prediction tools showed decreases in prediction accuracy consistent with the decrease in structural accuracy. The MD protocol also offered us the opportunity to test the ability of the tools to leverage multiple structures of the same protein: we observed a marked improvement for all tools, but the availability of multiple structures was best exploited by IF-SitePred.

This study revealed that while data leakage is a major issue in the evaluation of tools on predicted protein structures, much of the performance lost due to lower-accuracy structures can be ameliorated with the use of multiple structures. Our results showed that this can be best exploited when using ML-based representations of proteins. A possible extension of this work is the identification of binding hotspots, which could facilitate the targeting of key interactions in initial fragment screens and follow-up experiments.

6.3.2 Pocket similarity for fragment selection

Binding site similarity is broad area of study that is used to understand the conformational shifts, physicochemical properties and evolutionary relationships between proteins in addition to the study of polypharmacology and off-target effects. However, a decisive method for comparing protein binding sites has not yet been established.

Following our success in predicting binding sites based on the transformer embeddings from ESM-IF1, we explored the use of these representations and their sequence-based relatives (ESM2) for the comparison of binding site residues. While unsupervised learning for distance-based pocket similarity was not viable, we found that the embeddings were a rich source of information on protein-ligand interactions, significantly outperforming our baselines and performing even better in combination.

For distance-based binding site similarity to be possible, contrastive learning could be used on the ESM-IF1 or ESM2 embeddings of binding site residues. Given similar and dissimilar protein pairs (those which bind identical and different ligands respectively), contrastive learning algorithms encode the similar pairs nearby in euclidean space while dissimilar pairs are distant. However, the amount of data required for such a task is much greater than what is currently available, whether for fragment-sized or drug-sized ligands.

6.4 Closing remarks

In this thesis we have discussed possible avenues for the design of efficient fragment libraries that will facilitate the development of more diverse lead compounds and thus improve the odds of a viable drug being designed for a disease of interest.

While we have found an effective method for target-agnostic library design and explored potential strategies for target-specific library design, we have found data availability to be a significant limiting factor. Although FBDD produces a much higher quantity of protein-ligand complexes than high-throughput screening of drug-like molecules, there is not yet available a complete set of fragment screening results. We believe that through the collaboration of computational experts with expert crystallographers, the generation of such a dataset would provide a wealth of opportunities for state-of-the-art ML methods to be applied to FBDD and greatly further our ability to design effective fragment libraries.

Appendix A

Fragment libraries designed to be functionally diverse recover information more efficiently from protein targets compared with traditional libraries

Library Name	Chemical space	Sampling strategies	Size	Reference
DSiP	poised	shape diversity (USRCAT)	768	XChem (2020a)
F2X	pharmacophore	ph4 diversity (MACCS)	96	Wollenhaupt et al. (2020)
SpotXplorer	hotsplot exploration	ph4 + interaction clustering	96	Bajusz et al. (2021)
Fraglites	halogenated fragments with paired H-binding motifs		25	Wood et al. (2019)
PepLites	halogenated peptidomimetics		25	Douangamath et al. (2020)
Cambridge 3D	3D and poised	diversity oriented synthesis	137	Kidd et al. (2020)
York 3D	3D, substituted aliphatic heterocycle	shape diversity	106	Downes et al. (2020)
Leeds 3D	3D, natural product-like scaffolds, high sp3			Foley et al. (2017); Aimon et al. (2018)
MiniFrags	Astex's pharmacophores, ultra-low-MW	'chemical diversity'	81	O'Reilly et al. (2019)
Astex	general	chemical and structural diversity	2000	Hall et al. (2014)
CovHetFrags	Small heterocyclic electrophiles	chemical properties	141	Keeley et al. (2020, 2019)
CysElectrophile	Cysteine covalent library		993	Resnick et al. (2019)

Novartis	general	structural diversity, clustering	1408	Kutchukian et al. (2014)
AZ	general xrc	structural diversity (ECFI)	384	Fuller et al. (2016)
Chembridge	large set	subsets: solubility, fluorine, bromine, structurally diverse	15000	Chembridge (2021)
Asinex	'BioDeisgn' - synthetically feasible, NP-like and common NP ph4		20061	Asinex (2021)
Life Chemicals	known bioactives	diversity, Fsp3, target family-specific	8000	Chemicals (2021)
Prestwick	Fragments of approved drugs		1456	Prestwick (2021)
Selcia			1214	Selcia (2021)
Timetec		structural diversity		Timetec (2021)
Zenobia	general inc. PPI	shape diverse	352	Zenobia (2021)
Bilsland	ML-generated privileged fragments	FeCo, novelty, 'examination by medicinal chemists'	741	Bilsland et al. (2021)
Essential	initial screen, frequently reported hits		320	Enamine (2020a)
High fidelity	MedChem tractable	diversity selection using clustering algorithm	1920	Enamine (2020a)
Fluorinated	19F NMR		1000	Enamine (2020a)
Covalent	Diverse covalent		6210	Enamine (2020a)
Fully-functionalized	Photoaffinity labeled fragments		2000	Enamine (2020a)
NP-like	NP-like	scaffold frequency analysis	4160	Enamine (2020a)
3D shape-diverse	3D diverse	K-mean clustering, centroids taken	1200	Enamine (2020a)
PPI	peptidomimetics		3600	Enamine (2020a)
Single ph4	ro3-like	variety of ph4s and scaffolds	1500	Enamine (2020a)
Carboxylic acid	specific protein targets		4000	Enamine (2020a)
Halogen-enriched	halogen bonding motifs	diversity	3000	Enamine (2020a)
Fully-functionalized probe	Efficient exploration of novel protein targets w/ photoaffinity probes		640	Enamine (2020a); Parker et al. (2017)
EU openscreen	availability of parent molecules within larger collection		1056	openscreen (2021)
Maybridge	pharmacophorically-rich	clustering and centriods selected	2500	Maybridge (2021)
Vernalis	general	diversity of ph4 triangles	1063	Chen and Hubbard (2009)
Pfizer	general	diversity and visual inspection	2592	Lau et al. (2011)

Table A.1: Many of the fragment libraries currently in use. Some are used privately, while others are commercially available. Empty cells indicate that the method has not been published.

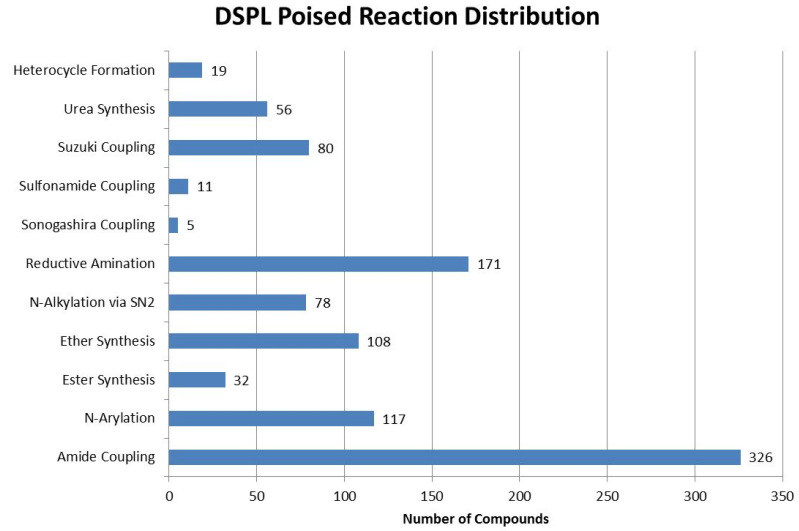


Figure A.1: Fragments in the DSP library are poised for 11 different categories of reaction. From XChem (2020b).

Bit	Interaction type
0	hydrophobic contacts
1	aromatic face to face
2	aromatic edge to face
3	hydrogen bond (protein as hydrogen bond donor)
4	hydrogen bond (protein as hydrogen bond acceptor)
5	salt bridges (protein positively charged)
6	salt bridges (protein negatively charged)
7	salt bridges (ionic bond with metal ion)

Table A.2: The eight types of protein-ligand interaction detected by ODDT's Interaction-Fingerprint module Wójcikowski et al. (2015).

Acronym	Full name	Description of activity	CATH class(es)	Reference
CD44MMA	Hyaluronan binding domain of murine CD44	Binds hyaluronan via a lectin-like fold termed the Link module after appropriate functional activation	3.10.100.10	Banerji et al. (2007)
EPB41L3A	Erythrocyte membrane protein band 4.1-like 3a	Binds CD44 and transmits signals to the cytoskeleton	3.10.20.90, 1.20.80.10, 2.30.29.30	Bradshaw et al. (2020a)

INPP5DA	Src homology 2 domain containing inositol polyphosphate 5-phosphatase 1 (SHIP1)	Phosphatase within the PI3K/AKT/mTOR signallling pathway	3.60.10.10	Bradshaw et al. (2020b)
MID2A	Homo sapiens midline 2	Probable E3 ubiquitin-protein ligase	2.60.120.920	Uniprot (2020); The UniProt Consortium (2020)
Mpro	SARS-CoV-2 Main Protease	Cysteine viral protease that cleaves at 11 sites, essential to viral replication	2.40.10.10, 1.10.1840.10	Douangamath et al. (2020)
NSP14	SARS-CoV-2 nsp14 exoribonuclease	3'-to-5' proofreading exoribonuclease (ExoN) and guanine-N7-methyltransferase	-	Ogando et al. (2020)
PHIPA	Pleckstrin Homology Domain Interacting Protein (second bromodomain)	Mediates transcriptional responses in pancreatic islet cells	1.20.920.10	Cox et al. (2016a)
TBXTA	T-box transcription factor TA	Transcription factor in embryonic tissues	2.60.40.820	Newman et al. (2020a)
mArh	SARS-CoV-2 nonstructural protein 3	counteracts host-mediated antiviral adenosine diphosphate–ribosylation signaling	3.40.220.10	Schuller et al. (2021)
nsp13	SARS-CoV-2 nonstructural protein 13	Superfamily 1 DNA or RNA helicase	3.40.50.300	Newman et al. (2020b)

Table A.3: Descriptions of each target used in this study. CATH classes obtained through sequence-based searching of CATH Sillitoe et al. (2020) at `https://www.cathdb.info/`.

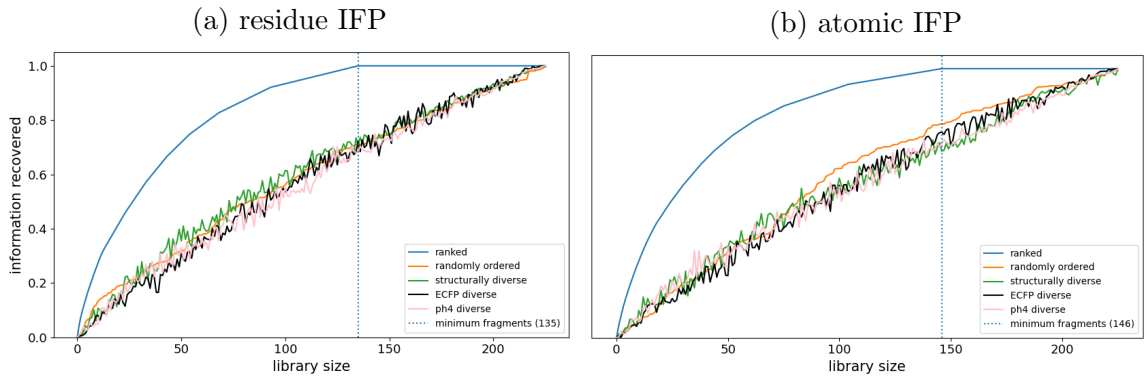


Figure A.2: A comparison of the fragment ranking method (blue) with various types of structurally diverse fragment selection methods. MACCS keys (green), ECFP2 fingerprints (black) and 2D pharmacophoric fingerprints (pink) were generated for each fragment (all using RDKit Landrum (2020)) and diverse selections of fragments based on these fingerprints were selected at each library size 5 times. The average information recovered from all targets at each library size for each type of fingerprint was calculated and this is shown. None of the structurally diverse fragment selection methods significantly outperformed the random control (orange). **(a):** fragments have been ranked using the residue IFP method. **(b):** fragments have been ranked using the atomic IFP method.

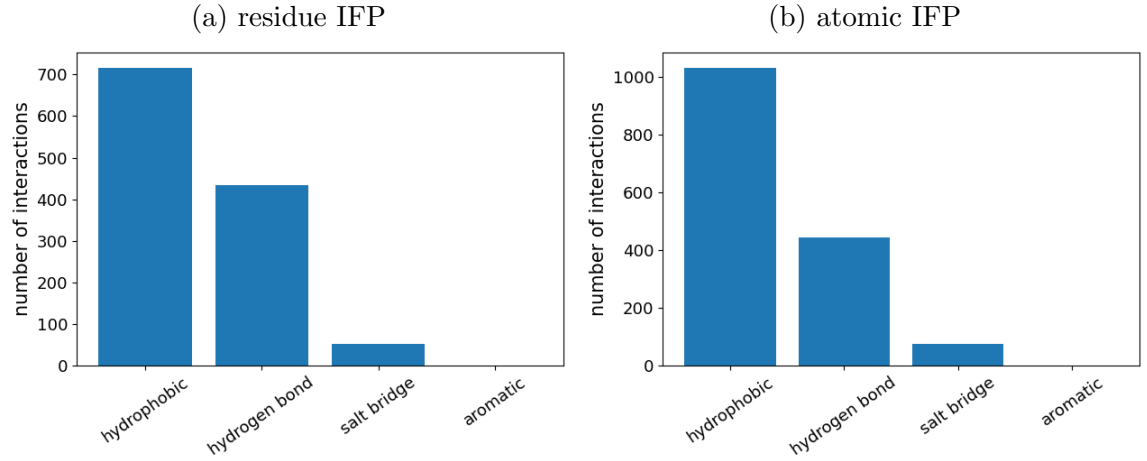


Figure A.3: Interactions calculated across all protein-fragment structures used in this analysis. As in Ferreira de Freitas and Schapira (2017), most interactions are hydrophobic, followed by hydrogen bonds and salt bridges. No aromatic interactions were detected. **(a):** interactions were calculated using the residue IFP method. **(b):** fragments have been calculated using the atomic IFP method.

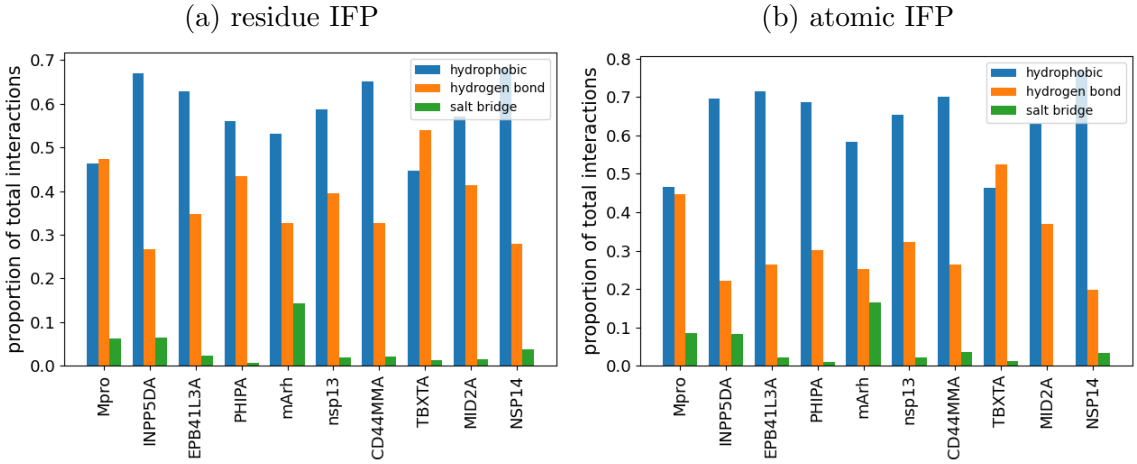


Figure A.4: Interactions calculated for each target. Most targets follow the trend of making significantly more hydrophobic interactions than hydrogen bonds, however Mpro has similar levels of the two interaction types, and TBXTA forms more hydrogen bonds than hydrophobic interactions. **(a):** interactions were calculated using the residue IFP method. **(b):** fragments have been calculated using the atomic IFP method.

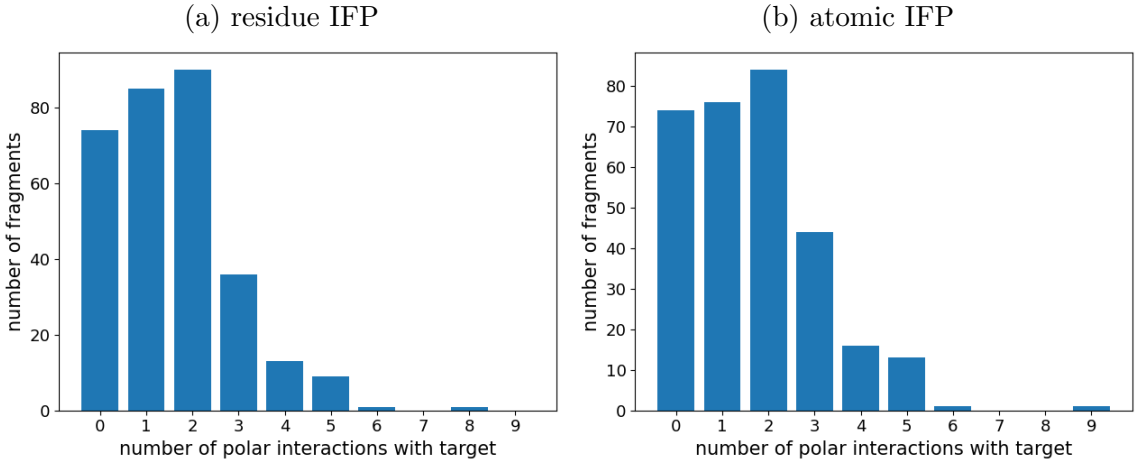


Figure A.5: The numbers of polar interactions made by each fragment in a protein-fragment complex. In comparison with a previous analysis (Giordanetto et al., 2019), this dataset exhibits a larger proportion of complexes in which no polar interactions are made with the protein. This could be due to the fact that water-mediated interactions were excluded from this analysis, or that a larger proportion of fragments in the DSi-poised library are able to make efficient hydrophobic interactions to facilitate binding. **(a):** interactions were calculated using the residue IFP method. **(b):** fragments have been calculated using the atomic IFP method.

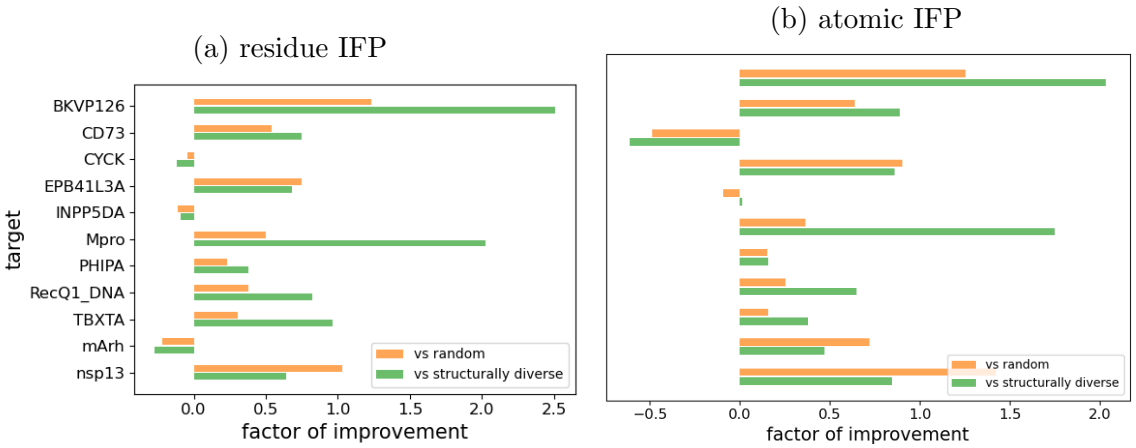


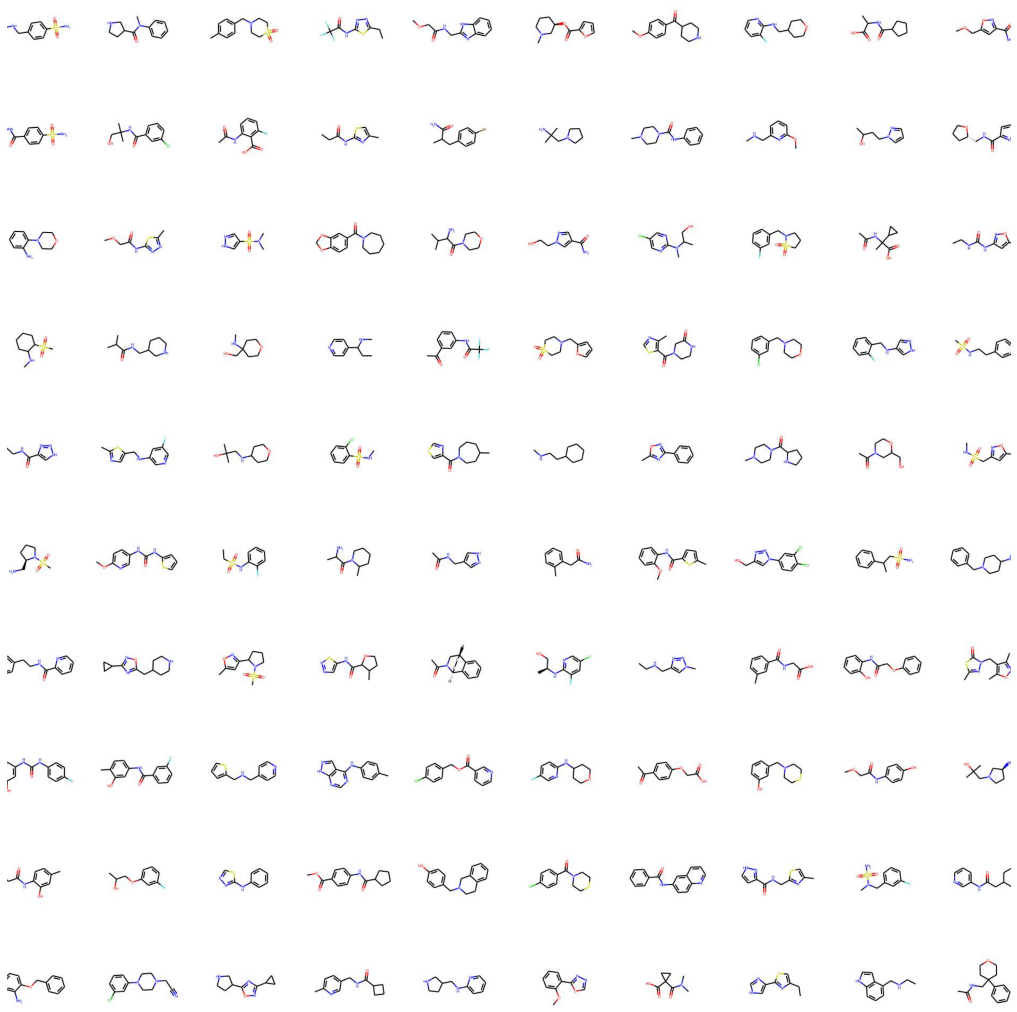
Figure A.6: Factor of information improvement when using 100 functionally diverse fragments compared with random and structurally diverse libraries. **(a):** fragments have been ranked using the residue IFP method. **(b):** fragments have been ranked using the atomic IFP method.

Molecules included in analysis









Ranks for each molecule

SMILES	rank group	hits
c1cc(-c2ccc3c(c2)OCO3)n[nH]1	top 100	1
OC1CN(C(c2ccccc2)c2ccccc2)C1	top 100	2
COC(=O)Nc1ccc(Cc2ccncc2)cc1	top 100	1
CC1C(O)CCCN1Cc1ccccc1	top 100	1
CCNc1ccc(C#N)cn1	top 100	3
COc1ccc2[nH]cc(CN(C)C)c2c1	top 100	3
O=C(NCCc1ccncc1)c1ccccc1F	top 100	3
CS(=O)(=O)Cc1nc2ccccc2[nH]1	top 100	4
Cc1ccc(NC(=O)c2cccc(F)c2)cc1O	top 100	1
CCOC(=O)N1CCOCC1	top 100	1
CCOc1ccc(CC(=O)O)cc1	top 100	1
c1ccc(SCCN2CCOCC2)cc1	top 100	2
NC(=O)c1cnc(C(F)(F)F)nc1	top 100	3
O=C(CN1CCCC1)Nc1ccc(F)cc1	top 100	2
CCN(CC)c1cc(C)nc2ncnn12	top 100	2
CNCc1ccncc1	top 100	1
CS(=O)(=O)NCC1CCNCC1	top 100	1
Fc1cccnc1NCC1CCOCC1	top 100	1
CC(C)(CO)NC(=O)c1cccc(Cl)c1	top 100	2
Cc1nn(C)c(C)c1CC(N)=O	top 100	2
C[C@H](O)CNCc1c(Cl)cccc1Cl	top 100	2
COc1ccc(NC(=O)CN)cc1	top 100	1
O=C(c1ccco1)N1CCN(C(=O)C2CC2)CC1	top 100	2
Cc1cc(CN(C)C(=O)NC2CC2)no1	top 100	2
Cc1nc(C(=O)N2CCCC2)c(C)s1	top 100	1
CC(C)Nc1cccnc1	top 100	1
Cc1occc1C(=O)Nc1ccncc1	top 100	2
CNc1ccccc1S(C)(=O)=O	top 100	2
CNC(=O)c1ccc(S(N)(=O)=O)cc1	top 100	1
CCn1cc(CNC(=O)c2cccnc2)cn1	top 100	1
C1CC(c2nc(C3CC3)no2)CN1	top 100	1
Fc1cccc(CNCc2ccco2)c1	top 100	2
Cc1nn(Cc2c(C)noc2C)c(=O)s1	top 100	2

Ranks for each molecule

COC(=O)c1cccc(NS(C)(=O)=O)c1	top 100	1
CNS(=O)(=O)Cc1cc(C)on1	top 100	1
O=C(c1ccc(Cl)cc1)N1CCSCC1	top 100	1
CCS(=O)(=O)NCc1ccc2c(c1)OCO2	top 100	1
OC1CCN(Cc2ccsc2)CC1	top 100	3
CCN1CCC(Nc2ccccc2F)CC1	top 100	2
NC(=S)Nc1ccccc1OC(F)(F)F	top 100	1
CC1CN(C(=O)Cn2cccn2)CCO1	top 100	1
CC(=O)NCc1ccc(S(N)(=O)=O)cc1	top 100	2
Cc1cc(F)ccc1CS(N)(=O)=O	top 100	2
CC(C)(O)CNC1CCOCC1	top 100	1
CC(C)NC(=O)N(C)C1CCS(=O)(=O)C1	top 100	2
O=C(c1ccc2c(c1)OCO2)N1CCCCCC1	top 100	1
CNS(=O)(=O)c1cccc(C)c1F	top 100	1
Cn1cc(Cl)c(C(=O)NC2CCCC2)n1	top 100	1
COC(=O)c1ccc(S(N)(=O)=O)cc1	top 100	2
Cc1ccc(NC(=O)C2CCCO2)nc1	top 100	1
CC(=O)Nc1ccc(Oc2ncccn2)cc1	top 100	2
CN1CCN(C(=O)Nc2ccccc2)CC1	top 100	3
Fc1ccccc1NCc1cc[nH]n1	top 100	3
CC(=O)Nc1cccc(NC(N)=O)c1	top 100	3
CN(CCc1ccccn1)C(=O)c1ccccc1	top 100	2
CC(C)c1ncc(Cl)c(C(N)=O)n1	top 100	2
CN(CCO)S(=O)(=O)c1ccc(N)cc1	top 100	1
COc1ccc(C)cc1NC(=O)Nn1cnnc1	top 100	2
O=C(CC1CCCCC1)Nn1cnnc1	top 100	1
CC(=O)Nc1ccc(C)c2c1CCCN2	top 100	1
C[C@@H](CO)N(C)c1ncccc1F	top 100	1
CN(Cc1cccc(F)c1)S(N)(=O)=O	top 100	3
CNS(=O)(=O)CC1CCCCC1	top 100	2
CC(=O)N(C)c1ccccc1C(=O)O	top 100	1
O=S1(=O)CCCN1Cc1cccc(F)c1	top 100	1
Cc1cc(N2CCNCC2)nc(C)n1	top 100	2
CC(=O)Nc1cc(C(=O)O)ccc1F	top 100	3

Ranks for each molecule

CNC(=O)c1cccc2cc[nH]c12	top 100	2
CCNCc1cn(C)nn1	top 100	1
COCC(=O)Nc1ccccc1C(N)=O	top 100	1
N#Cc1ccc(Cn2cnc3ccccc32)cc1	top 100	1
N#Cc1ccc(CNCc2ccc(F)cc2)cc1	top 100	1
Cc1noc(C)c1C(=O)NCCC1=CCCCC1	top 100	2
CCn1cc(NC(=O)C2CCC2)cn1	top 100	3
CC(NC(=O)C1CCCC1)C(=O)O	top 100	1
Oc1ccc(CN2CCc3ccccc3C2)cc1	top 100	1
CCN(C)C(=O)c1cccc(N)c1	top 100	1
O=C(COc1ccccc1)N1CCCC1	top 100	1
CC(N)C(=O)N1CCCCC1C	top 100	1
CC(=O)c1ccc(OCC(=O)O)cc1	top 100	1
COc1ccc(C(=O)C2CCNCC2)cc1	top 100	1
CS(=O)(=O)NCCc1ccc(F)cc1	top 100	3
CN1CCC(Oc2cccc(F)c2)C1=O	top 100	1
CNC(C)c1sc(C)nc1C	top 100	2
Nc1ccc(S(=O)(=O)Nc2ccccn2)cc1	top 100	2
O=C(CC1CCCCC1)Nc1cccnc1	top 100	2
O=C(c1c(F)cccc1F)N1CCCC1	top 100	2
COCC(=O)Nc1ccc(O)cc1	top 100	2
Cc1cccc(C(=O)N2CCC(C(N)=O)CC2)c1	top 100	3
O=C(Nc1ccccc1)Nc1cccnc1	top 100	4
CC(C)Nc1nc2ccccc2[nH]1	top 100	1
CC(=O)N1CCN(CCCc2ccccc2)CC1	top 100	1
Cc1ccc(CS(N)(=O)=O)cc1F	top 100	2
CCS(=O)(=O)Nc1ccccc1F	top 100	2
CC(Oc1ccc(C#N)cc1)C(N)=O	top 100	2
CC(=O)Nc1ccc(Nc2ccccc2)cc1	top 100	1
COc1ccc2sc(N)nc2c1	top 100	2
CC(C)N(C)S(=O)(=O)N1CCNC(=O)C1	top 100	1
Cc1csc(CNC(=O)c2cc[nH]n2)n1	top 100	1
CN(C)c1ccncc1	top 100	1
CN1CCC[C@@H](OC(=O)c2ccco2)C1	remaining mi	1

3

Ranks for each molecule

Cc1ccc(OCC(=O)N2CCN(C)CC2)cc1	remaining mi	1
Cc1nnc(CN2CCC=C(F)C2)s1	remaining mi	1
c1csc(-c2nnc[nH]2)n1	remaining mi	1
CC(C)c1ccc(CN2CCC(O)CC2)cc1	remaining mi	1
CC1CCCN(C(=O)c2cscn2)CC1	remaining mi	1
COC(=O)[C@@H]1C[C@@H](O)CN1C(=O)(	remaining mi	1
COc1cccc(C(=O)Nc2ccccc2F)c1	remaining mi	1
CCC(=O)Nc1ccc(S(N)(=O)=O)cc1	remaining mi	1
O=C(COc1ccccc1)Nc1cccc(F)c1	remaining mi	2
Cc1cc(F)cc(S(N)(=O)=O)c1	remaining mi	1
CC(N)c1ccc(NC(=O)C2CC2)cc1	remaining mi	1
CN1CCN(C(=O)c2ccc(F)c(F)c2)CC1	remaining mi	3
CC(=O)N1CCCC(O)(C(=O)O)C1	remaining mi	1
CS(=O)(=O)CC(O)c1ccccc1	remaining mi	1
c1cnc(N2CCC3(CCOC3)C2)cn1	remaining mi	1
CC(C)NCc1ccc(N(C)C)cc1	remaining mi	1
Cc1cccc(NC(=O)[C@@H]2CCCN2)n1	remaining mi	1
CS(=O)(=O)NCCc1ccccc1	remaining mi	1
CNS(=O)(=O)c1ccccc1Cl	remaining mi	1
CC(C)N(C)c1ncnc2c1cnn2C	remaining mi	1
CCN(CC)c1cccc(O)c1	remaining mi	1
CC(=O)N1CCCC2(CCCCCC2)C1	remaining mi	1
Clc1snnc1CN1CCOCC1	remaining mi	1
O=C(C1CC1)N1CCN(c2ccc(F)cc2)CC1	remaining mi	2
CC(C)(C)CC(=O)Nc1ccc(F)cc1F	remaining mi	1
CC(=O)Nc1ccncc1C	remaining mi	1
C[C@@H](CO)Nc1ncc(Cl)cc1F	remaining mi	2
CCNCc1cccc2[nH]ccc12	remaining mi	1
Cc1ccc(OCC(=O)N2CCOCC2)c(C)c1	remaining mi	1
Cc1cccc(C(=O)N(C)C)c1C	remaining mi	2
CC(NS(C)(=O)=O)c1ccccc1Cl	remaining mi	1
Cc1cc(C)nc(Nc2ccccc2)n1	remaining mi	1
CC1(C)CN(c2ncccc2F)CC1O	remaining mi	1
CN(C1CCCCCC1)S(C)(=O)=O	remaining mi	1

Ranks for each molecule

Cc1nc(CC(=O)N2CCOCC2)cs1	remaining mi	1
COCc1cc(C(N)=O)no1	remaining mi	1
Cc1cc(C(=O)NCC2CCCO2)ns1	remaining mi	1
O=C(CCCc1ccccc1)N1CCOCC1	remaining mi	1
Cc1ncc(CNc2cncc(F)c2)s1	remaining mi	1
O=C(CCC1CCCCC1)N1CCOCC1	remaining mi	1
CC(=O)NC1CNc2ccccc2C1	remaining mi	2
OCC1CN(Cc2ccccc2)CCO1	remaining mi	1
c1ccc(CNC2CCNC2)cc1	remaining mi	1
CC(CO)(CO)NC(=O)Nc1ccccc1	remaining mi	1
CCOC(=O)CN1CCS(=O)(=O)CC1	remaining mi	1
Cc1noc(C)c1CC(C)C(=O)N(C)C	redundant	1
CC(=O)NCc1cn[nH]c1	redundant	1
NC(=O)c1cnn(CCO)c1	redundant	1
CCCC(=O)Nc1cccnc1Cl	redundant	1
Cc1cccc(C(=O)NCC(=O)O)c1	redundant	2
CC(=O)Nc1cnccc1C	redundant	1
CNCc1cnn(-c2ccccc2)c1	redundant	1
Fc1ccc(CNCc2ccccc2)cc1	redundant	1
CN(CC1CCOC1)c1ncncc1Cl	redundant	1
N#Cc1ccc(CNC(=O)N2CCOCC2)cc1	redundant	1
CC1CN(C(=O)c2cnsn2)C(C)CO1	redundant	1
COCc1nnc(N)s1	redundant	1
COc1ccc(NC(=O)Nc2cccs2)cn1	redundant	2
CN1CCN(C(=O)Nc2ccc(F)cc2)CC1	redundant	2
Cn1cc(CNC(=O)c2ccco2)cn1	redundant	1
Cn1nccc1C(=O)NCc1cccs1	redundant	1
Cc1ccc(CC(=O)NC[C@@H]2CCCO2)cc1	redundant	1
CC(=O)Nc1cccc(F)c1C(=O)O	redundant	1
CN1CCCc2ccc(S(N)(=O)=O)cc21	redundant	1
CNC(=O)c1cccc(C(=O)O)n1	redundant	1
CN(C)C(=O)c1ccc(F)cc1Br	redundant	1
CC(=O)N1CCCNCC1	redundant	1
C1CC(Cc2nc(C3CC3)no2)CCN1	redundant	1

Ranks for each molecule

CC1CCN(C(=O)c2cc3ccccc3o2)CC1	redundant	1
CN(C(=O)C1CCNC1)c1ccccc1	redundant	1
CN(C)c1ccc(CNn2cnnc2)cc1	redundant	2
CCNC(=O)c1c[nH]nn1	redundant	1
CC(Cc1ccc(Br)cc1)C(N)=O	redundant	1
CCOC(=O)C1CCN(c2cnccn2)CC1	redundant	1
NC(=O)c1c[nH]c(C2CC2)n1	redundant	1
CN(C)C(=O)N1CCOCC12CCOC2	redundant	1
c1csc(CNCc2ccncc2)c1	redundant	1
CN(C(=O)[C@H]1CC1(C)C)c1cccnc1	redundant	1
Cc1cncc(NCc2ccncc2)c1	redundant	1
O=C(CCc1ccccc1)Nc1ccc(O)cc1	redundant	1
Clc1cccc(CN2CCOCC2)c1	redundant	2
CC(N)c1ccc(N2CCOCC2)cc1	redundant	1
CC(=O)Nc1cccc2nonc12	redundant	1
CC(Oc1ccccc1F)C(=O)O	redundant	1
CCC(=O)Nc1nc(C)cs1	redundant	1
NC(=O)N1CCN(C(=O)c2ccco2)CC1	redundant	2
CN(C)C(=O)OCCn1ccnc1	redundant	1
COc1ccc2nc(NC(=O)C3CC3)sc2c1	redundant	1
CCOC(=O)c1cn[nH]c1	redundant	1
CC(=O)Nc1ccc(F)c(C(=O)O)c1	redundant	1
c1ccc(CN2CCC(n3cccc3)CC2)cc1	redundant	1
Cc1cc(N(C)C2CC(O)C2)nc(C)n1	redundant	1
O=C(Cc1ccc(Cl)cc1)Nc1ncccn1	redundant	1
Fc1ccccc1CNc1cn[nH]c1	redundant	1
CC(=O)Nc1ccccc1O	redundant	2
Cc1ccsc1CNC(C)(C)CO	redundant	2
O=C(Nc1cccc(Cl)c1)C(F)(F)F	redundant	1
CN(C)Cc1ccc(C#N)cc1	redundant	1
c1ccc(CCNc2nc3ccccc3[nH]2)cc1	redundant	1
CC(C)N1CCN(C(=O)c2cccs2)CC1	redundant	1
Cc1n[nH]c(C)c1-c1ccccc1N	redundant	1
NC(=O)N1CCN(c2ccc(F)cc2)CC1	redundant	3

Ranks for each molecule

CCNC(=O)Nc1cc(C)on1	redundant	1
O=C(NCc1nc2ccccc2[nH]1)c1ccco1	redundant	1
CC(C)C(=O)Nc1cccc(C#N)c1	redundant	2
CNCC(=O)Nc1cc(C)ccn1	redundant	1
Cc1ccc(CNC(=O)C2CCC2)cn1	redundant	1
CNC1CCCN(c2ccccc2)C1=O	redundant	1
C1CN(CC2CCOCC2)CCN1	redundant	1
CCCC(=O)NCc1nc2ccccc2[nH]1	redundant	1
c1ccc(NCC2CCNC2)nc1	redundant	1
CS(=O)(=O)NC1CCNCC1	redundant	1
CS(=O)(=O)N1CCC(O)CC1	redundant	1
OCc1cn(-c2ccc(Cl)cc2)nn1	redundant	1
COc1ccccc1N1CCN(CC#N)CC1	redundant	1
CN(C)C(=O)CNc1ccccc1	redundant	1
Cc1nccn1Cc1cccc(C#N)c1	redundant	1
CCn1cc(CCO)cn1	redundant	1
CNc1ncccn1	redundant	1
OCC1CCN(c2ncccn2)CC1	redundant	2
CC(CS(N)(=O)=O)c1ccccc1	redundant	1
CS(=O)(=O)NC1CCCC1	redundant	1
CN1CCN(c2ncccc2F)CC1=O	redundant	1
CC(Oc1ccc(Cl)cc1)C(N)=O	redundant	1

Appendix B

Learnt representations of proteins can be used for accurate prediction of small molecule binding sites on experimentally determined and predicted protein structures

min distance to protein	max distance to protein	distance for relabelling	clustering distance	success rate
2.5	5.5	4	1.6	0.99
2.5	5.5	4	1.7	0.99
2.5	5.5	4	1.8	0.99
2.5	5.5	4.5	1.6	0.99
2.5	5.5	4.5	1.7	0.99
2.5	5.5	4.5	1.8	0.99
2.5	5.5	5	1.6	0.99
2.5	5.5	5	1.7	0.99
2.5	5.5	5	1.8	0.99
2.5	6	4	1.6	0.99
2.5	6	4	1.7	0.99
2.5	6	4	1.8	0.99
2.5	6	4.5	1.6	0.99
2.5	6	4.5	1.7	0.99
2.5	6	4.5	1.8	0.99
2.5	6	5	1.6	0.99
2.5	6	5	1.7	0.99
2.5	6	5	1.8	0.99
2.5	6.5	4	1.6	0.99
2.5	6.5	4	1.7	0.99
2.5	6.5	4	1.8	0.99
2.5	6.5	4.5	1.6	0.99
2.5	6.5	4.5	1.7	0.99
2.5	6.5	4.5	1.8	0.99
2.5	6.5	5	1.6	0.99
2.5	6.5	5	1.7	0.99
2.5	6.5	5	1.8	0.99
3	5.5	4	1.6	0.97
3	5.5	4	1.7	0.97
3	5.5	4	1.8	0.97
3	5.5	4.5	1.6	0.99
3	5.5	4.5	1.7	0.99
3	5.5	4.5	1.8	0.99
3	5.5	5	1.6	1.0
3	5.5	5	1.7	1.0
3	5.5	5	1.8	1.0
3	6	4	1.6	0.97
3	6	4	1.7	0.97
3	6	4	1.8	0.97
3	6	4.5	1.6	0.99
3	6	4.5	1.7	0.99
3	6	4.5	1.8	0.99
3	6	5	1.6	1.0
3	6	5	1.7	1.0
3	6	5	1.8	1.0
3	6.5	4	1.6	0.97
3	6.5	4	1.7	0.97
3	6.5	4	1.8	0.97
3	6.5	4.5	1.6	0.99
3	6.5	4.5	1.7	0.99
3	6.5	4.5	1.8	0.99
3	6.5	5	1.6	1.0
3	6.5	5	1.7	1.0
3	6.5	5	1.8	1.0
3.5	5.5	4	1.6	0.93
3.5	5.5	4	1.7	0.93
3.5	5.5	4	1.8	0.93
3.5	5.5	4.5	1.6	0.99
3.5	5.5	4.5	1.7	0.99
3.5	5.5	4.5	1.8	0.99
3.5	5.5	5	1.6	0.99
3.5	5.5	5	1.7	0.99
3.5	5.5	5	1.8	0.99
3.5	6	4	1.6	0.93
3.5	6	4	1.7	0.93
3.5	6	4	1.8	0.93
3.5	6	4.5	1.6	0.99
3.5	6	4.5	1.7	0.99
3.5	6	4.5	1.8	0.99
3.5	6	5	1.6	0.99
3.5	6	5	1.7	0.99
3.5	6	5	1.8	0.99
3.5	6.5	4	1.6	0.93
3.5	6.5	4	1.7	0.93
3.5	6.5	4	1.8	0.93
3.5	6.5	4.5	1.6	0.99
3.5	6.5	4.5	1.7	0.99
3.5	6.5	4.5	1.8	0.99
3.5	6.5	5	1.6	0.99
3.5	6.5	5	1.7	0.99
3.5	6.5	5	1.8	0.99

Table B.1: Continued on following page.

Table B.1: Success rate is largely stable under small perturbations in distance thresholds for point cloud generation and clustering. Where the distance for relabelling residues as ligand-binding was lowered (4Å), success rate was reduced as low as 0.93, however all other combinations of thresholds yielded success rates of 0.99-1.00.

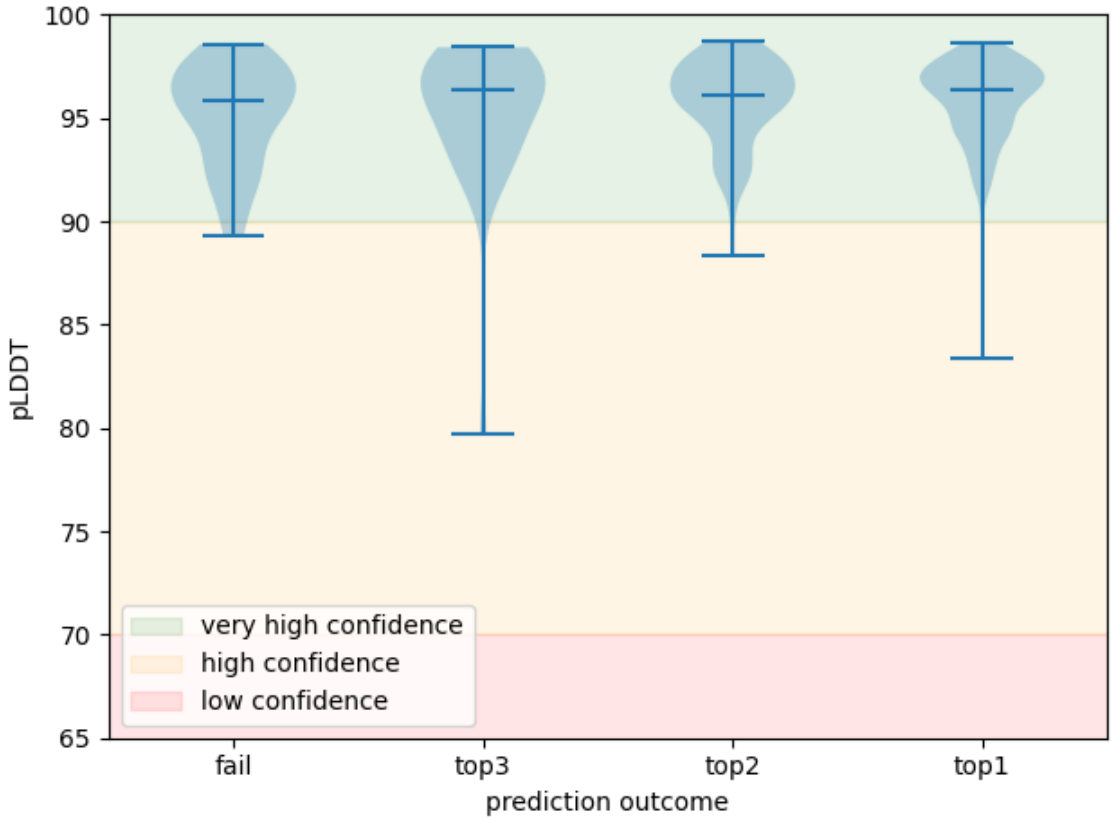


Figure B.1: AlphaFold prediction confidence (pLDDT) does not vary between targets with top 1, top 2, top 3 success or failure in binding site prediction by IF-SitePred. While there are some outlying values, mean pLDDT is similar between levels of prediction success. The prediction confidence is very high compared with what would be expected for novel protein predictions.

	Baseline	IF-SitePred		FPocket		P2Rank		DeepPocket	
	Success rate	Success rate	STD (95% CI)	Success rate	STD (95% CI)	Success rate	STD (95% CI)	Success rate	STD (95% CI)
Top 1	0.12	0.76	0.39-0.44	0.75	0.41-0.45	**0.81**	0.37-0.41	0.78	0.40-0.43
Top 2	0.17	0.89	0.25-0.32	0.81	0.37-0.42	**0.90**	0.28-0.34	0.87	0.31-0.36
Top 3	0.22	**0.93**	0.22-0.28	0.83	0.35-0.40	**0.93**	0.22-0.28	0.89	0.28-0.33

Table B.2: Error estimations as calculated using bootstrapping of the predictions for PDB structures in the HAP set show that IF-SitePred and P2Rank have more consistent prediction success than FPocket and DeepPocket. The 95% confidence interval of the standard deviation for each success rate was calculated using SciPy's bootstrapping function. In general, error is estimated to be smaller where prediction accuracy is higher.

	Baseline	IF-SitePred		FPocket		P2Rank		DeepPocket	
	Success rate	Success rate	STD (95% CI)	Success rate	STD (95% CI)	Success rate	STD (95% CI)	Success rate	STD (95% CI)
Top 1	0.09	0.77	0.40-0.44	0.50	0.50-0.50	**0.81**	0.36-0.41	0.78	0.40-0.44
Top 2	0.17	**0.89**	0.28-0.34	0.60	0.48-0.50	0.88	0.29-0.36	0.87	0.31-0.37
Top 3	0.20	**0.94**	0.19-0.26	0.67	0.46-0.48	0.89	0.28-0.34	0.90	0.27-0.33

Table B.3: Error estimations as calculated using bootstrapping of the predictions for AF2 structures in the HAP set show that IF-SitePred has more consistent prediction success than P2Rank, FPocket and DeepPocket. Confidence intervals were calculated as in Table B.2.

	Baseline	IF-SitePred		FPocket		P2Rank		DeepPocket	
	Success rate	Success rate	STD (95% CI)	Success rate	STD (95% CI)	Success rate	STD (95% CI)	Success rate	STD (95% CI)
Top 1	0.12	0.76	0.39-0.45	0.73	0.41-0.47	**0.78**	0.38-0.45	0.75	0.40-0.46
Top 2	0.18	**0.90**	0.26-0.35	0.80	0.36-0.43	0.86	0.30-0.39	0.85	0.32-0.40
Top 3	0.24	**0.91**	0.23-0.33	0.82	0.34-0.42	**0.91**	0.24-0.33	0.88	0.26-0.35

Table B.4: Error estimations as calculated using bootstrapping of the predictions for PDB structures in the HAP-small set show that IF-SitePred and P2rank have more consistent prediction success than FPocket and DeepPocket. Confidence intervals were calculated as in Table B.2.

	Baseline	IF-SitePred		FPocket		P2Rank		DeepPocket	
	Success rate	Success rate	STD (95% CI)	Success rate	STD (95% CI)	Success rate	STD (95% CI)	Success rate	STD (95% CI)
Top 1	0.10	0.74	0.41-0.46	0.48	0.50-0.50	**0.76**	0.39-0.46	0.75	0.40-0.46
Top 2	0.17	**0.87**	0.30-0.38	0.58	0.48-0.50	0.85	0.31-0.39	0.86	0.30-0.39
Top 3	0.20	**0.94**	0.19-0.30	0.65	0.46-0.49	0.88	0.28-0.37	0.90	0.26-0.35

Table B.5: Error estimations as calculated using bootstrapping of the predictions for AF2 structures in the HAP-small set show that IF-SitePred has more consistent prediction success than P2Rank, FPocket and DeepPocket. Confidence intervals were calculated as in Table B.2.

Appendix C

Binding site similarity strategies for target-specific fragment library design

Structures included in analysis

ATAD2A	NUDT7A	TBXTA	DCP2B		FALZA	OXA10OTA	NUDT5A	PHIPA
ATAD2A-x1803	NUDT7A-x0140	TBXTA-x0368	DCP2B-x1475	DCP2B-x0033	FALZA-x0177	OXA10OTA-x0059	NUDT5A-x1291	PHIPA-x9178
ATAD2A-x1835	NUDT7A-x0159	TBXTA-x0992	DCP2B-x1042	DCP2B-x1374	FALZA-x0402	OXA10OTA-x0074	NUDT5A-x1214	
ATAD2A-x1716	NUDT7A-x0254	TBXTA-x0835	DCP2B-x1424	DCP2B-x1214	FALZA-x0172	OXA10OTA-x0106	NUDT5A-x1024	
ATAD2A-x1740	NUDT7A-x0286	TBXTA-x0409	DCP2B-x1217	DCP2B-x1502	FALZA-x0085	OXA10OTA-x0049	NUDT5A-x0685	
ATAD2A-x1754	NUDT7A-x0151	TBXTA-x0786	DCP2B-x1246	DCP2B-x1386	FALZA-x0079	OXA10OTA-x0051	NUDT5A-x0171	
ATAD2A-x1712	NUDT7A-x0149	TBXTA-x0530	DCP2B-x1362	DCP2B-x1235	FALZA-x0309	OXA10OTA-x0038	NUDT5A-x0463	
ATAD2A-x1719	NUDT7A-x0299	TBXTA-x0167	DCP2B-x1254	DCP2B-x0042	FALZA-x0271	OXA10OTA-x0108	NUDT5A-x0333	
	NUDT7A-x0304	TBXTA-x0114	DCP2B-x1179	DCP2B-x1361	FALZA-x0438	OXA10OTA-x0079	NUDT5A-x0600	
	NUDT7A-x0303	TBXTA-x0753	DCP2B-x1035	DCP2B-x1524		OXA10OTA-x0073	NUDT5A-x0299	
	NUDT7A-x0374	TBXTA-x0338	DCP2B-x1367	DCP2B-x1247		OXA10OTA-x0053	NUDT5A-x0262	
	NUDT7A-x0220	TBXTA-x0258	DCP2B-x1257	DCP2B-x1527			NUDT5A-x0673	
		TBXTA-x0567	DCP2B-x1044	DCP2B-x1036			NUDT5A-x0114	
		TBXTA-x0096	DCP2B-x1033	DCP2B-x1532			NUDT5A-x1211	
		TBXTA-x0773	DCP2B-x1382	DCP2B-x0340			NUDT5A-x0339	
		TBXTA-x0465	DCP2B-x0067	DCP2B-x1025			NUDT5A-x0692	
		TBXTA-x0820	DCP2B-x1021	DCP2B-x1372			NUDT5A-x0605	
		TBXTA-x0093	DCP2B-x1047	DCP2B-x1479			NUDT5A-x0552	
		TBXTA-x0128	DCP2B-x1519	DCP2B-x1523			NUDT5A-x1004	
		TBXTA-x0244	DCP2B-x1518				NUDT5A-x0627	
		TBXTA-x0491	DCP2B-x1396				NUDT5A-x0158	
		TBXTA-x0229	DCP2B-x0996				NUDT5A-x0392	
		TBXTA-x0545	DCP2B-x1039				NUDT5A-x0412	
		TBXTA-x0938	DCP2B-x1022				NUDT5A-x1066	
		TBXTA-x1079	DCP2B-x1259				NUDT5A-x0244	
		TBXTA-x0830	DCP2B-x1526				NUDT5A-x1083	
		TBXTA-x0390	DCP2B-x1038				NUDT5A-x1213	
		TBXTA-x0986	DCP2B-x1008				NUDT5A-x0574	
		TBXTA-x0204	DCP2B-x1535				NUDT5A-x0526	
		TBXTA-x0113	DCP2B-x1561				NUDT5A-x0256	
		TBXTA-x1161	DCP2B-x1529				NUDT5A-x1298	
		TBXTA-x0164	DCP2B-x1010				NUDT5A-x0320	
		TBXTA-x0076	DCP2B-x0533				NUDT5A-x0403	
		TBXTA-x0249	DCP2B-x1393				NUDT5A-x1217	
		TBXTA-x0464	DCP2B-x1225				NUDT5A-x0319	
		TBXTA-x0552	DCP2B-x1521				NUDT5A-x0176	
		TBXTA-x0144	DCP2B-x1040				NUDT5A-x0681	
		TBXTA-x0948	DCP2B-x1019				NUDT5A-x0554	
		TBXTA-x0124	DCP2B-x1364				NUDT5A-x0469	
		TBXTA-x0776	DCP2B-x1264				NUDT5A-x0637	
		TBXTA-x1164	DCP2B-x1256				NUDT5A-x1093	
		TBXTA-x0353	DCP2B-x1508				NUDT5A-x1223	
		TBXTA-x0744	DCP2B-x1012				NUDT5A-x1219	
		TBXTA-x0316	DCP2B-x1046				NUDT5A-x1039	
		TBXTA-x1174	DCP2B-x0628				NUDT5A-x1235	
		TBXTA-x0179	DCP2B-x0020				NUDT5A-x1078	
		TBXTA-x0563	DCP2B-x1344				NUDT5A-x0125	
		TBXTA-x0506	DCP2B-x1048				NUDT5A-x0651	
		TBXTA-x0500	DCP2B-x1032				NUDT5A-x0286	
		TBXTA-x0024	DCP2B-x1276				NUDT5A-x1242	
			DCP2B-x1037				NUDT5A-x1028	
			DCP2B-x1023				NUDT5A-x1212	

Structures included in analysis

			DCP2B-x0466				NUDT5A-x0169	
			DCP2B-x1020				NUDT5A-x0177	
			DCP2B-x0290				NUDT5A-x0525	
			DCP2B-x1538				NUDT5A-x1072	

Structures included in analysis

STAG1A	DCLRE1AA	CAMK1DA	BRD1A	MUREECA	PARP14A
STAG1A-x0996	DCLRE1AA-x0139	CAMK1DA-x0168	BRD1A-x0904	MUREECA-x0198	PARP14A-x0590
STAG1A-x0237	DCLRE1AA-x0233	CAMK1DA-x0146	BRD1A-x0905	MUREECA-x0164	PARP14A-x0202
STAG1A-x0488	DCLRE1AA-x0172	CAMK1DA-x0208	BRD1A-x0900	MUREECA-x0114	PARP14A-x0473
STAG1A-x0583	DCLRE1AA-x0928	CAMK1DA-x0177	BRD1A-x0902	MUREECA-x0206	PARP14A-x0505
STAG1A-x0245	DCLRE1AA-x0150	CAMK1DA-x0067	BRD1A-x0933	MUREECA-x0215	PARP14A-x0334
STAG1A-x0515	DCLRE1AA-x0195	CAMK1DA-x0290	BRD1A-x0924	MUREECA-x0234	PARP14A-x0137
STAG1A-x0894	DCLRE1AA-x0135	CAMK1DA-x0253	BRD1A-x0907	MUREECA-x0244	PARP14A-x0637
STAG1A-x0531	DCLRE1AA-x0209	CAMK1DA-x0145	BRD1A-x0922	MUREECA-x0172	PARP14A-x0324
STAG1A-x0521	DCLRE1AA-x0968	CAMK1DA-x0218	BRD1A-x0927	MUREECA-x0169	PARP14A-x0238
STAG1A-x0720	DCLRE1AA-x0183	CAMK1DA-x0107			PARP14A-x0161
STAG1A-x0921	DCLRE1AA-x0965	CAMK1DA-x0049			PARP14A-x0690
STAG1A-x0457	DCLRE1AA-x0164	CAMK1DA-x0083			PARP14A-x0266
STAG1A-x1055	DCLRE1AA-x0234	CAMK1DA-x0297			PARP14A-x0412
STAG1A-x0627	DCLRE1AA-x0193	CAMK1DA-x0321			PARP14A-x0315
	DCLRE1AA-x1013	CAMK1DA-x0091			PARP14A-x0457
	DCLRE1AA-x0169	CAMK1DA-x0204			PARP14A-x0484
	DCLRE1AA-x1015	CAMK1DA-x0226			PARP14A-x0567
	DCLRE1AA-x1085	CAMK1DA-x0089			PARP14A-x0712
	DCLRE1AA-x1087	CAMK1DA-x0299			
	DCLRE1AA-x1010				
	DCLRE1AA-x1074				
	DCLRE1AA-x0157				
	DCLRE1AA-x1012				
	DCLRE1AA-x0925				
	DCLRE1AA-x0128				
	DCLRE1AA-x0194				
	DCLRE1AA-x0972				

Structures included in analysis

Mpro				
Mpro-5c5n-2020-04-Akaji	Mpro-x2581.pdb	Mpro-4mds-2020-04-Mesecar	Mpro-x10396:JANGHE-83b26c96-23	
Mpro-x0305.pdb	Mpro-x2754:LONWEI-120e5cf5-5	Mpro-x1478.pdb	Mpro-x0398:AARPOS-0daf6b7e-28	
Mpro-x2912.pdb	Mpro-x0540:AARPOS-d2a4d1df-12	Mpro-x1402:AARPOS-0daf6b7e-15	Mpro-x1308.pdb	
Mpro-1q2w-2020-04-Bonanno	Mpro-6m2n-2020-04-Su	Mpro-x3303.pdb	Mpro-x1311:AARPOS-0daf6b7e-1	
Mpro-x0967.pdb	Mpro-x0336:AARPOS-0daf6b7e-45	Mpro-x10377.pdb	Mpro-5c5o-2020-04-Akaji	
Mpro-x0390.pdb	Mpro-x3305.pdb	Mpro-x3324:DAVCRI-3edb475e-4	Mpro-4wy3-2020-04-Akaji	
Mpro-x3110:LONWEI-adc59df6-2	Mpro-3tit-2020-04-Zhu	Mpro-x3298.pdb	Mpro-x1392:AARPOS-0daf6b7e-14	
Mpro-x0336.pdb	Mpro-x10082.pdb	Mpro-x10403.pdb	Mpro-x10201:EDGMED-0da5ad92-2	
Mpro-x0887:AARPOS-0daf6b7e-22	Mpro-x1086.pdb	Mpro-x0499:AARPOS-0daf6b7e-41	Mpro-x1308:AARPOS-d2a4d1df-44	
Mpro-x2572:TRYUNI-714a760b-20	Mpro-x3117.pdb	Mpro-x1336.pdb	Mpro-x2649.pdb	
Mpro-x0786:AARPOS-d2a4d1df-38	Mpro-x1077.pdb	Mpro-6y2f-2020-04-Zhang	Mpro-x1237.pdb	
Mpro-x3351.pdb	Mpro-x0434.pdb	Mpro-x0692.pdb	Mpro-x10049:DUNNEW-f8ce3686-14	
Mpro-x10334:EDJMED-49816e9b-2	Mpro-3sne-2020-04-Zhu	Mpro-x1386:AARPOS-0daf6b7e-13	Mpro-x1336:AARPOS-0daf6b7e-3	
Mpro-x0350:AARPOS-0daf6b7e-36	Mpro-x3077:DANLON-a5fc619e-8	Mpro-x2964.pdb	Mpro-x2764.pdb	
Mpro-x0072:AARPOS-d2a4d1df-1	Mpro-x0770.pdb	Mpro-5n5o-2020-04-Zhang	Mpro-3v3m-2020-04-Jacobs	
Mpro-x1249.pdb	Mpro-2amd-2020-04-Yang	Mpro-x1385:AARPOS-0daf6b7e-12	Mpro-x1493:AARPOS-0daf6b7e-21	
Mpro-x1163:AARPOS-0daf6b7e-30	Mpro-x3124:LONWEI-adc59df6-3	Mpro-x1374:AARPOS-0daf6b7e-7	Mpro-x0305:AARPOS-d2a4d1df-5	
Mpro-x1382:AARPOS-0daf6b7e-10	Mpro-x0397:AARPOS-d2a4d1df-9	Mpro-x1418:AARPOS-0daf6b7e-17	Mpro-x2964:DARDIA-23aa0b97-13	
Mpro-x0705:AARPOS-d2a4d1df-25	Mpro-x2754.pdb	Mpro-x10314.pdb	Mpro-x1384.pdb	
Mpro-x1002.pdb	Mpro-x2646.pdb	Mpro-x2563:DARDIA-23aa0b97-20	Mpro-x0398.pdb	
Mpro-x10371:RALMED-2de63afb-1	Mpro-x3303:MATPOS-916a2c5a-4	Mpro-x2600.pdb	Mpro-x0759.pdb	
Mpro-x3324.pdb	Mpro-x2608:DARDIA-842b4336-3	Mpro-x2764:BARCOM-4e090d3a-49	Mpro-x10280.pdb	
Mpro-x0770:AARPOS-d2a4d1df-35	Mpro-2z9j-2020-04-Lee	Mpro-x10392.pdb	Mpro-x2912:TRYUNI-714a760b-22	

Structures included in analysis

Mpro				
Mpro-x2971.pdb	Mpro-x3113.pdb	Mpro-x3080:GAB-REV-70cc3ca5-8	Mpro-x0478.pdb	
Mpro-x0771.pdb	Mpro-x0731:AAR-POS-d2a4d1df-27	Mpro-x3080	Mpro-7buy-2020-04-Zhao	
Mpro-x0387:AAR-POS-d2a4d1df-7	Mpro-x10255.pdb	Mpro-x0194:AAR-POS-0daf6b7e-33	Mpro-x0995.pdb	
Mpro-x10324:JAG-UCB-a3ef7265-20	Mpro-x10178.pdb	Mpro-x3366.pdb	Mpro-x2776:LON-WEI-b8d98729-8	
Mpro-x0177:AAR-POS-0daf6b7e-32	Mpro-x1119:AAR-POS-0daf6b7e-42	Mpro-2gz8-2020-04-Lu	Mpro-x1402.pdb	
Mpro-x0731.pdb	Mpro-x0820.pdb	Mpro-x3348:SIM-SYN-f15aaa3a-1	Mpro-x1226:AAR-POS-0daf6b7e-34	
Mpro-x0946:AAR-POS-d2a4d1df-15	Mpro-x1012:AAR-POS-0daf6b7e-24	Mpro-x0708.pdb	Mpro-x1311.pdb	
Mpro-x10392:JAN-GHE-83b26c96-20	Mpro-5n19-2020-04-Zhang	Mpro-x10306.pdb	Mpro-2gz7-2020-04-Lu	
Mpro-x10049.pdb	Mpro-x1187:AAR-POS-0daf6b7e-23	Mpro-x0691.pdb	Mpro-3tnt-2020-04-Zhu	
Mpro-x0830.pdb	Mpro-x10395.pdb	Mpro-x10338.pdb	Mpro-x1351:AAR-POS-0daf6b7e-5	
Mpro-x10334.pdb	Mpro-x10237.pdb	Mpro-x2052:AAR-POS-5507155c-2	Mpro-x10417.pdb	
Mpro-x0107.pdb	Mpro-x10022:AAR-POS-0daf6b7e-37	Mpro-2a5k-2020-04-Lee	Mpro-x0434:AAR-POS-d2a4d1df-11	
Mpro-x1093:AAR-POS-d2a4d1df-20	Mpro-x3325.pdb	Mpro-x0691:AAR-POS-d2a4d1df-23	Mpro-x10419.pdb	
Mpro-x0354:AAR-POS-d2a4d1df-6	Mpro-x2562:BAR-COM-4e090d3a-39	Mpro-x3115:LON-WEI-adc59df6-26	Mpro-x1380:AAR-POS-0daf6b7e-9	
Mpro-3d62-2020-04-Bacha	Mpro-x10172:DAV-CRI-3edb475e-6	Mpro-x2097.pdb	Mpro-x10242.pdb	
Mpro-x0397.pdb	Mpro-x2910:MAT-POS-916a2c5a-2	Mpro-6y2g-2020-04-Zhang	Mpro-x1458:AAR-POS-0daf6b7e-19	
Mpro-x2540.pdb	Mpro-x3298:BAR-COM-4e090d3a-57	Mpro-x10306:DAN-LON-a5fc619e-3	Mpro-x0426:AAR-POS-d2a4d1df-10	
Mpro-x1386.pdb	Mpro-x0967:AAR-POS-d2a4d1df-16	Mpro-x2705:LON-WEI-b8d98729-20	Mpro-2d2d-2020-04-Yang	
Mpro-x10150.pdb	Mpro-x0769:AAR-POS-d2a4d1df-34	Mpro-x0669.pdb	Mpro-2gtb-2020-04-Lee	
Mpro-x2649:TRY-UNI-714a760b-18	Mpro-x3115.pdb	Mpro-2z9l-2020-04-Lee	Mpro-x0165.pdb	
Mpro-x2193.pdb	Mpro-x2540:LON-WEI-b8d98729-23	Mpro-x0425.pdb	Mpro-x1358.pdb	
Mpro-6w63-2020-04-Mesecar	Mpro-x0395.pdb	Mpro-x3124.pdb	Mpro-x10423:JAN-GHE-83b26c96-8	
Mpro-x3325:TOB-UNK-c2aba166-1	Mpro-x2694:LON-WEI-b8d98729-43	Mpro-x0978:AAR-POS-d2a4d1df-42	Mpro-x0755.pdb	

Structures included in analysis

Mpro				
Mpro-x1382.pdb	Mpro-3tiu-2020-04-Zhu	Mpro-x0734:AAR-POS-d2a4d1df-28	Mpro-x2694.pdb	
Mpro-x10422.pdb	Mpro-x1384:AAR-POS-0daf6b7e-11	Mpro-3szn-2020-04-Zhu	Mpro-x10314:TRY-UNI-2eddb1ff-1	
Mpro-x1226.pdb	Mpro-x2119.pdb	Mpro-x10359.pdb	Mpro-x3333.pdb	
Mpro-x0387.pdb	Mpro-x1077:AAR-POS-d2a4d1df-19	Mpro-x1358:AAR-POS-0daf6b7e-6	Mpro-x2910.pdb	
Mpro-x1187.pdb	Mpro-x0177.pdb	Mpro-x10019:GAB-REV-70cc3ca5-4	Mpro-x3305:BAR-COM-4e090d3a-47	
Mpro-x1392.pdb	Mpro-x1412.pdb	Mpro-x0390:AAR-POS-0daf6b7e-25	Mpro-x3359.pdb	
Mpro-2z9k-2020-04-Lee	Mpro-x2659.pdb	Mpro-x0376:AAR-POS-0daf6b7e-43	Mpro-x0705.pdb	
Mpro-x1093.pdb	Mpro-x0786.pdb	Mpro-x2703:LON-WEI-b8d98729-18	Mpro-x0759:AAR-POS-d2a4d1df-33	
Mpro-x0426.pdb	Mpro-x0874:AAR-POS-d2a4d1df-14	Mpro-x0981.pdb	Mpro-3sn8-2020-04-Zhu	
Mpro-x0887.pdb	Mpro-x0769.pdb	Mpro-x0464.pdb	Mpro-x1493.pdb	
Mpro-x0749.pdb	Mpro-x0195.pdb	Mpro-x1348:AAR-POS-0daf6b7e-4	Mpro-x2659:AAR-POS-fca48359-19	
Mpro-x3077.pdb	Mpro-x0499.pdb	Mpro-x0774.pdb	Mpro-x0708:AAR-POS-d2a4d1df-26	
Mpro-x0161.pdb	Mpro-2qiq-2020-04-Grum-Tokars	Mpro-x0991:AAR-POS-d2a4d1df-17	Mpro-x1012.pdb	
Mpro-2alv-2020-04-Ghosh	Mpro-x1334:AAR-POS-0daf6b7e-2	Mpro-4twy-2020-04-Akaji	Mpro-x0736.pdb	
Mpro-x0161:MAT-POS-7dfc56d9-1	Mpro-x0464:AAR-POS-0daf6b7e-26	Mpro-x3333:DAN-PUR-6788a628-2	Mpro-x10155.pdb	
Mpro-x0752.pdb	Mpro-x10248.pdb	Mpro-x2646:TRY-UNI-714a760b-6	Mpro-x1334.pdb	
Mpro-x10396.pdb	Mpro-2v6n-2020-04-Verschueren	Mpro-x3108.pdb	Mpro-x1478:AAR-POS-0daf6b7e-20	
Mpro-x10296.pdb	Mpro-2zu5-2020-04-Hsu	Mpro-x10247.pdb	Mpro-x0107:AAR-POS-d2a4d1df-3	
Mpro-x2572.pdb	Mpro-x3110.pdb	Mpro-x2097:AAR-POS-f650c5f2-3	Mpro-x10327:JAN-GHE-83b26c96-9	
Mpro-x2600:ANN-UNI-26382800-5	Mpro-x10327.pdb	Mpro-x0072.pdb	Mpro-x0195:AAR-POS-d2a4d1df-4	
Mpro-x10387.pdb	Mpro-x0749:AAR-POS-d2a4d1df-30	Mpro-x0831:AAR-POS-d2a4d1df-41	Mpro-x1418.pdb	
Mpro-x1351.pdb	Mpro-x2929:DUN-NEW-d8381160-1	Mpro-x2608.pdb	Mpro-x0669:AAR-POS-0daf6b7e-29	
Mpro-x1237:AAR-POS-0daf6b7e-39	Mpro-x3108:GAB-REV-70cc3ca5-13	Mpro-2vj1-2020-04-Verschueren	Mpro-x1348.pdb	
Mpro-x10424.pdb	Mpro-x1101:AAR-POS-0daf6b7e-40	Mpro-x10421.pdb	Mpro-x2569.pdb	

Structures included in analysis

Mpro				
Mpro-x3366:GAB-REV-70cc3ca5-18	Mpro-x0425:AAR-POS-0daf6b7e-27	Mpro-x1002:AAR-POS-0daf6b7e-37	Mpro-x2193:AAR-POS-5507155c-1	
Mpro-2gx4-2020-04-Hsu	Mpro-x0874.pdb	Mpro-x10355:MAT-POS-590ac91e-32	Mpro-4tww-2020-04-Akaji	
Mpro-x1375:AAR-POS-0daf6b7e-8	Mpro-x3359:LON-WEI-adc59df6-30	Mpro-3tns-2020-04-Zhu	Mpro-x3113:LON-WEI-adc59df6-27	
Mpro-x2643.pdb	Mpro-x10019.pdb	Mpro-x2562.pdb	Mpro-x3080.pdb	
Mpro-x10423.pdb	Mpro-x2705.pdb	Mpro-x0946.pdb	Mpro-2zu4-2020-04-Hsu	
Mpro-x1458.pdb	Mpro-x2052.pdb	Mpro-x0978.pdb	Mpro-x1412:AAR-POS-0daf6b7e-16	
Mpro-x0734.pdb	Mpro-x0820:AAR-POS-d2a4d1df-39	Mpro-x10359:ALP-POS-95b75b4d-2	Mpro-7bqy-2020-04-Liu	
Mpro-x0692:AAR-POS-d2a4d1df-24	Mpro-x1101.pdb	Mpro-x10172.pdb	Mpro-x0991.pdb	
Mpro-x2119:AAR-POS-f650c5f2-2	Mpro-1wof-2020-04-Yang	Mpro-x10236:JOR-UNI-2fc98d0b-12	Mpro-x3305	
Mpro-x3117:LON-WEI-adc59df6-39	Mpro-x1086:AAR-POS-0daf6b7e-46	Mpro-2a5i-2020-04-Lee	Mpro-x1235:AAR-POS-0daf6b7e-35	
Mpro-x0689:AAR-POS-d2a4d1df-22	Mpro-x0194.pdb	Mpro-x2929.pdb	Mpro-x10322:RAL-MED-2de63afb-2	
Mpro-x0981:AAR-POS-d2a4d1df-43	Mpro-x10236.pdb	Mpro-x0104:AAR-POS-d2a4d1df-2	Mpro-x0395:AAR-POS-d2a4d1df-8	
Mpro-x0752:AAR-POS-d2a4d1df-31	Mpro-x1374.pdb	Mpro-x0831.pdb	Mpro-x0678:AAR-POS-d2a4d1df-13	
Mpro-x1375.pdb	Mpro-x1163.pdb	Mpro-x0755:AAR-POS-d2a4d1df-32	Mpro-2z94-2020-04-Lee	
Mpro-x2569:DAR-DIA-23aa0b97-17	Mpro-x10022.pdb	Mpro-x0165:AAR-POS-0daf6b7e-44	Mpro-x3351:DAR-DIA-eace69ff-36	
Mpro-x10237:JOR-UNI-2fc98d0b-6	Mpro-x10338:JAG-UCB-a3ef7265-3	Mpro-x2779.pdb	Mpro-x0350.pdb	
Mpro-x1101	Mpro-x1249:AAR-POS-d2a4d1df-21	Mpro-x2581:ALV-UNI-7ff1a6f9-47	Mpro-x0540.pdb	
Mpro-x10355.pdb	Mpro-x0678.pdb	Mpro-x1425:AAR-POS-0daf6b7e-18	Mpro-x0376.pdb	
Mpro-x1235.pdb	Mpro-x10395:JAN-GHE-83b26c96-10	Mpro-x0830:AAR-POS-d2a4d1df-40	Mpro-x10371.pdb	
Mpro-x2971:BEN-DND-362d364a-10	Mpro-x0689.pdb	Mpro-x0736:AAR-POS-d2a4d1df-29	Mpro-x0354.pdb	
Mpro-x2563.pdb	Mpro-x1380.pdb	Mpro-2op9-2020-04-Craik	Mpro-x10377:JAN-GHE-83b26c96-3	
Mpro-x10422:JAN-GHE-83b26c96-22	Mpro-x10247:JAN-GHE-83b26c96-14	Mpro-x1132.pdb	Mpro-x1425.pdb	
Mpro-x0104.pdb	Mpro-x2908:TRY-UNI-714a760b-12	Mpro-x1119.pdb	Mpro-x0771:AAR-POS-d2a4d1df-36	

Structures included in analysis

Mpro				
Mpro-x10322.pdb	Mpro-3snd-2020-04-Zhu	Mpro-x3351	Mpro-x2908.pdb	
Mpro-x2643:DARDIA-842b4336-13	Mpro-x2779:LONWEI-8f408cad-5	Mpro-x3348.pdb	Mpro-x2776.pdb	
Mpro-x10324.pdb	Mpro-x2703.pdb	Mpro-x0478:AARPOS-0daf6b7e-31		
Mpro-x1132:AARPOS-0daf6b7e-38	Mpro-x10201.pdb	Mpro-x0774:AARPOS-d2a4d1df-37		
Mpro-x10387:RALMED-2de63afb-14	Mpro-x1385.pdb	Mpro-x0995:AARPOS-d2a4d1df-18		

Structures included in analysis

NUDT4A	FAM83BA	XX02KALRNA	TNCA	EPB41L3A	mArh	NUDT21A	VIM2
NUDT4A-x0159	FAM83BA-x0131	XX02KALRNA-x1453	TNCA-x0100	EPB41L3A-x0306	mArh-x0969.pdb	NUDT21A-x0798	VIM2-MB-105
NUDT4A-x0688	FAM83BA-x0285	XX02KALRNA-x1505	TNCA-x0334	EPB41L3A-x0267	mArh-x0173.pdb	NUDT21A-x0584	
NUDT4A-x0634	FAM83BA-x0786	XX02KALRNA-x1480	TNCA-x0179	EPB41L3A-x0429	mArh-x0591.pdb	NUDT21A-x0703	
NUDT4A-x0400	FAM83BA-x0813	XX02KALRNA-x1492	TNCA-x0192	EPB41L3A-x0305	mArh-x0628.pdb	NUDT21A-x0976	
NUDT4A-x0556	FAM83BA-x0837	XX02KALRNA-x1483	TNCA-x0291	EPB41L3A-x0558	mArh-x0301.pdb	NUDT21A-x1047	
NUDT4A-x0690	FAM83BA-x0958	XX02KALRNA-x1490	TNCA-x0323	EPB41L3A-x0249	mArh-x0950.pdb	NUDT21A-x0503	
	FAM83BA-x0168	XX02KALRNA-x1389	TNCA-x0084	EPB41L3A-x0475	mArh-x0159.pdb	NUDT21A-x0974	
	FAM83BA-x0183	XX02KALRNA-x1377	TNCA-x0107	EPB41L3A-x0330	mArh-x0598.pdb	NUDT21A-x0498	
	FAM83BA-x0963	XX02KALRNA-x1388	TNCA-x0141	EPB41L3A-x0455	mArh-x0253.pdb	NUDT21A-x0942	
		XX02KALRNA-x1376	TNCA-x0245	EPB41L3A-x0141	mArh-x0849.pdb	NUDT21A-x0969	
		XX02KALRNA-x1449	TNCA-x0146	EPB41L3A-x0161	mArh-x0852.pdb	NUDT21A-x1046	
		XX02KALRNA-x1463	TNCA-x0228	EPB41L3A-x0274	mArh-x0128.pdb	NUDT21A-x0203	
			TNCA-x0296	EPB41L3A-x0229	mArh-x0685.pdb	NUDT21A-x0404	
			TNCA-x0162	EPB41L3A-x0499	mArh-x0837.pdb	NUDT21A-x0502	
			TNCA-x0062	EPB41L3A-x0580	mArh-x0855.pdb	NUDT21A-x0967	
			TNCA-x0290	EPB41L3A-x0494	mArh-x0655.pdb	NUDT21A-x0159	
			TNCA-x0063	EPB41L3A-x0630	mArh-x0967.pdb	NUDT21A-x0896	
				EPB41L3A-x0235	mArh-x0895.pdb	NUDT21A-x0368	
				EPB41L3A-x0364	mArh-x0900.pdb	NUDT21A-x0964	
				EPB41L3A-x0076	mArh-x0689.pdb	NUDT21A-x0659	
				EPB41L3A-x0296	mArh-x0766.pdb	NUDT21A-x0935	
				EPB41L3A-x0150	mArh-x0469.pdb	NUDT21A-x0750	
				EPB41L3A-x0401	mArh-x0844.pdb	NUDT21A-x0936	
				EPB41L3A-x0104	mArh-x0926.pdb	NUDT21A-x0930	
				EPB41L3A-x0179	mArh-x0574.pdb	NUDT21A-x1068	
				EPB41L3A-x0154	mArh-x0592.pdb	NUDT21A-x0266	
				EPB41L3A-x0608	mArh-x0962.pdb	NUDT21A-x1065	
				EPB41L3A-x0168	mArh-x0465.pdb	NUDT21A-x0948	
				EPB41L3A-x0601	mArh-x0471.pdb	NUDT21A-x0743	
				EPB41L3A-x0316	mArh-x0104.pdb	NUDT21A-x0581	
				EPB41L3A-x0544	mArh-x0421.pdb	NUDT21A-x0489	
				EPB41L3A-x0617	mArh-x0681.pdb		
				EPB41L3A-x0182	mArh-x0722.pdb		
				EPB41L3A-x0551	mArh-x1018.pdb		
				EPB41L3A-x0520	mArh-x1015.pdb		
				EPB41L3A-x0331	mArh-x0822.pdb		
				EPB41L3A-x0575	mArh-x0792.pdb		
				EPB41L3A-x0457	mArh-x0496.pdb		
				EPB41L3A-x0325	mArh-x0805.pdb		
				EPB41L3A-x0538	mArh-x0724.pdb		
				EPB41L3A-x0403	mArh-x0737.pdb		
				EPB41L3A-x0162	mArh-x0216.pdb		
				EPB41L3A-x0176	mArh-x0436.pdb		
				EPB41L3A-x0463	mArh-x0299.pdb		
				EPB41L3A-x0224	mArh-x0228.pdb		
				EPB41L3A-x0105	mArh-x0587.pdb		
				EPB41L3A-x0556	mArh-x0142.pdb		
				EPB41L3A-x0485	mArh-x0814.pdb		
				EPB41L3A-x0419	mArh-x0516.pdb		
				EPB41L3A-x0212	mArh-x0423.pdb		
				EPB41L3A-x0303	mArh-x0259.pdb		
				EPB41L3A-x0527	mArh-x0158.pdb		

Structures included in analysis

NUDT4A	FAM83BA	XX02KALRNA	TNCA	EPB41L3A	mArh	NUDT21A	VIM2
				EPB41L3A-x0453	mArh-x0729.pdb		
				EPB41L3A-x0337	mArh-x0905.pdb		
				EPB41L3A-x0547	mArh-x0548.pdb		
				EPB41L3A-x0202	mArh-x0631.pdb		
				EPB41L3A-x0209	mArh-x1092.pdb		
				EPB41L3A-x0550	mArh-x0371.pdb		
				EPB41L3A-x0598	mArh-x0334.pdb		
				EPB41L3A-x0244	mArh-x0796.pdb		
				EPB41L3A-x0171	mArh-x1012.pdb		
				EPB41L3A-x0509	mArh-x0864.pdb		
				EPB41L3A-x0283	mArh-x0910.pdb		
				EPB41L3A-x0175	mArh-x0438.pdb		
				EPB41L3A-x0413	mArh-x0290.pdb		
				EPB41L3A-x0473	mArh-x0711.pdb		
				EPB41L3A-x0366	mArh-x0524.pdb		
					mArh-x0417.pdb		
					mArh-x0918.pdb		
					mArh-x0676.pdb		
					mArh-x0727.pdb		
					mArh-x0742.pdb		
					mArh-x0091.pdb		
					mArh-x0600.pdb		
					mArh-x0380.pdb		
					mArh-x0282.pdb		
					mArh-x0662.pdb		
					mArh-x0933.pdb		
					mArh-x0797.pdb		
					mArh-x0626.pdb		

Bibliography

N. Abdollahi, S. A. M. Tonekaboni, J. Huang, B. Wang, and S. MacKinnon. NodeCoder: a graph-based machine learning platform to predict active sites of modeled protein structures. *arXiv*, Feb. 2023. doi: 10.48550/arxiv.2302.03590. URL http://arxiv.org/abs/2302.03590.

R. Aggarwal, A. Gupta, V. Chelur, C. V. Jawahar, and U. Deva Priyakumar. DeepPocket: Ligand Binding Site Detection and Segmentation using 3D Convolutional Neural Networks. *Journal of Chemical Information and Modeling*, 62(21):5069–5079, Aug. 2021. doi: 10.1021/acs.jcim.1c00799.

A. Ahmed, R. D. Smith, J. J. Clark, J. Dunbar James B., and H. A. Carlson. Recent improvements to Binding MOAD: a resource for protein–ligand binding affinities and structures. *Nucleic Acids Research*, 43(D1):D465–D469, 2014. ISSN 0305-1048. doi: 10.1093/nar/gku1088. URL https://doi.org/10.1093/nar/gku1088.

A. Aimon, G. Karageorgis, J. Masters, M. Dow, P. G. E. Craven, M. Ohsten, A. Willaume, R. Morgentin, N. Ruiz-Llamas, H. Lemoine, T. Kalliokoski, A. J. Eatherton, D. J. Foley, S. P. Marsden, and A. Nelson. Realisation of small molecule libraries based on frameworks distantly related to natural products. *Org. Biomol. Chem.*, 16(17):3160–3167, 2018. doi: 10.1039/C8OB00688A. URL http://dx.doi.org/10.1039/C8OB00688A.

M. Akdel, D. E. V. Pires, E. P. Pardo, J. Jänes, A. O. Zalevsky, B. Mészáros, P. Bryant, L. L. Good, R. A. Laskowski, G. Pozzati, A. Shenoy, W. Zhu, P. Kundrotas, V. R. Serra, C. H. M. Rodrigues, A. S. Dunham, D. Burke, N. Borkakoti, S. Velankar, A. Frost, J. Basquin, K. Lindorff-Larsen, A. Bateman, A. V. Kajava, A. Valencia, S. Ovchinnikov, J. Durairaj, D. B. Ascher, J. M. Thornton, N. E. Davey, A. Stein, A. Elofsson, T. I. Croll, and P. Beltrao. A structural biology community assessment of AlphaFold2 applications. *Nature Structural & Molecular Biology*, 29(11):1056–1067, 2022. ISSN 1545-9985. doi: 10.1038/s41594-022-00849-w. URL https://doi.org/10.1038/s41594-022-00849-w.

I. Akritopoulou-Zanze and P. J. Hajduk. Kinase-targeted libraries: the design and synthesis of novel, potent, and selective kinase inhibitors. *Drug discovery today*, 14(5-6):291–297, Mar. 2009. ISSN 1878-5832 (Electronic). doi: 10.1016/j.drudis.2008.12.002.

J. Ali, R. Khan, N. Ahmad, and I. Maqsood. Random forests and decision trees. *International Journal of Computer Science Issues (IJCSI)*, 9(5):272, 2012.

N. Anand and T. Achim. Protein Structure and Sequence Generation with Equivariant Denoising Diffusion Probabilistic Models. *arXiv*, May 2022. URL https://arxiv.org/abs/2205.15019.

P. Anand, D. Nagarajan, S. Mukherjee, and N. Chandra. PLIC: protein-ligand interaction clusters. *Database : the journal of biological databases and curation*, 2014 (0):bau029, 2014. ISSN 1758-0463 (Electronic). doi: 10.1093/database/bau029.

R. Anandakrishnan, B. Aguilar, and A. V. Onufriev. H++ 3.0: automating pK prediction and the preparation of biomolecular structures for atomistic molecular modeling and simulations. *Nucleic acids research*, 40(Web Server issue):W537–41, July 2012. ISSN 1362-4962 (Electronic). doi: 10.1093/nar/gks375.

A. C. Anderson. The Process of Structure-Based Drug Design. *Chemistry & Biology*, 10(9):787–797, 2003. ISSN 1074-5521. doi: https://doi.org/10.1016/j.chembiol.2003.09.002. URL https://www.sciencedirect.com/science/article/pii/S1074552103001947.

A. Andreeva, D. Howorth, C. Chothia, E. Kulesha, and A. G. Murzin. SCOP2 prototype: a new approach to protein structure mining. *Nucleic Acids Research*, 42 (D1):D310–D314, 2013. ISSN 0305-1048. doi: 10.1093/nar/gkt1242. URL https://doi.org/10.1093/nar/gkt1242.

A. Andreeva, E. Kulesha, J. Gough, and A. G. Murzin. The SCOP database in 2020: expanded classification of representative family and superfamily domains of known protein structures. *Nucleic Acids Research*, 48(D1):D376–D382, 2019. ISSN 0305-1048. doi: 10.1093/nar/gkz1064. URL https://doi.org/10.1093/nar/gkz1064.

K. Appelt, R. J. Bacquet, C. A. Bartlett, C. L. J. Booth, S. T. Freer, M. A. M. Fuhry, M. R. Gehring, S. M. Herrmann, and E. F. Howland. Design of enzyme inhibitors using iterative protein crystallographic analysis. *Journal of Medicinal Chemistry*, 34(7):1925–1934, jul 1991. ISSN 0022-2623. doi: 10.1021/jm00111a001. URL https://doi.org/10.1021/jm00111a001.

Asinex. Asinex Fragment Library, 2021. URL https://www.asinex.com/?page_id=97. Accessed: June 2021.

M. Baek, F. DiMaio, I. Anishchenko, J. Dauparas, S. Ovchinnikov, G. R. Lee, J. Wang, Q. Cong, L. N. Kinch, R. D. Schaeffer, C. Millán, H. Park, C. Adams, C. R. Glassman, A. DeGiovanni, J. H. Pereira, A. V. Rodrigues, A. A. van Dijk, A. C. Ebrecht, D. J. Opperman, T. Sagmeister, C. Buhlheller, T. Pavkov-Keller, M. K. Rathinaswamy, U. Dalwadi, C. K. Yip, J. E. Burke, K. C. Garcia, N. V. Grishin, P. D. Adams, R. J. Read, and D. Baker. Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, 373(6557):871–876, 2021. doi: 10.1126/science.abj8754. URL https://www.science.org/doi/abs/10.1126/science.abj8754.

D. Bajusz, W. S. Wade, G. Satała, A. J. Bojarski, J. Ilaš, J. Ebner, F. Grebien, H. Papp, F. Jakab, A. Douangamath, D. Fearon, F. von Delft, M. Schuller, I. Ahel, A. Wakefield, S. Vajda, J. Gerencsér, P. Pallai, and G. M. Keserű. Exploring protein hotspots by optimized fragment pharmacophores. *Nature Communications*, 12(1): 3201, 2021. ISSN 2041-1723. doi: 10.1038/s41467-021-23443-y. URL `https://doi.org/10.1038/s41467-021-23443-y`.

S. Banerji, A. J. Wright, M. Noble, D. J. Mahoney, I. D. Campbell, A. J. Day, and D. G. Jackson. Structures of the Cd44–hyaluronan complex provide insight into a fundamental carbohydrate-protein interaction. *Nature Structural & Molecular Biology*, 14(3):234–239, 2007. ISSN 1545-9985. doi: 10.1038/nsmb1201. URL `https://doi.org/10.1038/nsmb1201`.

S. Barelier, T. Sterling, M. J. O'Meara, and B. K. Shoichet. The Recognition of Identical Ligands by Unrelated Proteins. *ACS chemical biology*, 10(12):2772–2784, Dec. 2015. ISSN 1554-8937. doi: 10.1021/acschembio.5b00683. URL `https://pubmed.ncbi.nlm.nih.gov/26421501https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4684825/`.

J. D. Bauman, D. Patel, and E. Arnold. Fragment screening and HIV therapeutics. *Topics in current chemistry*, 317:181–200, 2012. ISSN 0340-1022 (Print). doi: 10.1007/128_2011_232.

C. Bentéjac, A. Csörgő, and G. Martínez-Muñoz. A comparative analysis of gradient boosting algorithms. *Artificial Intelligence Review*, 54(3):1937–1967, 2021. ISSN 1573-7462. doi: 10.1007/s10462-020-09896-5. URL `https://doi.org/10.1007/s10462-020-09896-5`.

H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. The protein data bank. *Nucleic Acids Research*, 28(1):235–242, 2000. ISSN 0305-1048. doi: 10.1093/nar/28.1.235. URL `https://doi.org/10.1093/nar/28.1.235`.

A. Bhadra and K. Yeturu. Site2Vec: a reference frame invariant algorithm for vector embedding of protein–ligand binding sites. *Machine Learning: Science and Technology*, 2(1):015005, Dec. 2020. ISSN 2632-2153. doi: 10.1088/2632-2153/abad88. URL `https://iopscience.iop.org/article/10.1088/2632-2153/abad88`.

Y. Bian and X.-Q. S. Xie. Computational Fragment-Based Drug Design: Current Trends, Strategies, and Applications. *The AAPS journal*, 20(3): 59, Apr. 2018. ISSN 1550-7416. doi: 10.1208/s12248-018-0216-7. URL `https://pubmed.ncbi.nlm.nih.gov/29633051https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6618289/`.

A. E. Bilsland, K. McAulay, R. West, A. Pugliese, and J. Bower. Automated generation of novel fragments using screening data, a dual SMILES autoencoder, transfer learning and syntax correction. *Journal of Chemical Information and Modeling*, 61

(6):2547–2559, June 2021. ISSN 1549-9596. doi: 10.1021/acs.jcim.0c01226. URL https://doi.org/10.1021/acs.jcim.0c01226.

BioSolveIT. BioSolveIT: Fragment-based drug discovery, 2023. URL https://www.biosolveit.de/application-academy/fbdd-fragment-based-drug-discovery/. Accessed: November 2023.

M. Bissaro, M. Sturlese, and S. Moro. The rise of molecular simulations in fragment-based drug design (FBDD): an overview. *Drug discovery today*, 25(9):1693–1701, sep 2020. ISSN 1878-5832 (Electronic). doi: 10.1016/j.drudis.2020.06.023.

N. Blomberg, D. A. Cosgrove, P. W. Kenny, and K. Kolmodin. Design of compound libraries for fragment screening. *Journal of Computer-Aided Molecular Design*, 23 (8):513–525, 2009. ISSN 1573-4951. doi: 10.1007/s10822-009-9264-5. URL https://doi.org/10.1007/s10822-009-9264-5.

C. Bouysset and S. Fiorucci. ProLIF: a library to encode molecular interactions as fingerprints. *Journal of Cheminformatics*, 13(1):72, 2021. ISSN 1758-2946. doi: 10.1186/s13321-021-00548-6. URL https://doi.org/10.1186/s13321-021-00548-6.

W. Bradshaw, V. Katis, and O. Gileadi. EPB41L3; A Target Enabling Package, Nov. 2020a. URL https://zenodo.org/record/4429638.

W. Bradshaw, R. Priestley, J. Obst, H. Hall-Roberts, A. Cederbalk, P. B. , E. Mead, E. D. Daniel, J. Davis, and O. Gileadi. SH2-containing-inositol-5-phosphatases (INPP5D); A Target Enabling Package, Nov. 2020b. URL https://zenodo.org/record/4429262.

N. Brandes, G. Goldman, C. H. Wang, C. J. Ye, and V. Ntranos. Genome-wide prediction of disease variants with a deep protein language model. *bioRxiv*, 2022. doi: 10.1101/2022.08.25.505311. URL https://www.biorxiv.org/content/early/2022/08/26/2022.08.25.505311.

L. Breiman. Random Forests. *Machine Learning*, 45(1):5–32, 2001. ISSN 1573-0565. doi: 10.1023/A:1010933404324. URL https://doi.org/10.1023/A:1010933404324.

A. K. Bronowska. Thermodynamics of Ligand-Protein Interactions: Implications for Molecular Design. In J. C. Moreno-Pirajan, editor, *Thermodynamics*, chapter 1. IntechOpen, Rijeka, 2011. doi: 10.5772/19447. URL https://doi.org/10.5772/19447.

B. Buchfink, K. Reuter, and H.-G. Drost. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nature Methods*, 18(4):366–368, 2021. ISSN 1548-7105. doi: 10.1038/s41592-021-01101-x. URL https://doi.org/10.1038/s41592-021-01101-x.

M. Buttenschoen, G. M. Morris, and C. M. Deane. PoseBusters: AI-based docking methods fail to generate physically valid poses or generalise to novel sequences. *arXiv*, 2023.

A. Capecchi, D. Probst, and J.-L. Reymond. One molecular fingerprint to rule them all: drugs, biomolecules, and the metabolome. *Journal of Cheminformatics*, 12 (1):43, 2020. ISSN 1758-2946. doi: 10.1186/s13321-020-00445-4. URL https://doi.org/10.1186/s13321-020-00445-4.

D. Case, K. Aktulga, H.M Belfon, I. Ben-Shalom, J. Berryman, S. Brozell, D. Cerutti, T. Cheatham III, G. Cisneros, V. Cruzeiro, T. Darden, R. Duke, G. Giambasu, M. Gilson, H. Gohlke, A. Goetz, R. Harris, S. Izadi, S. Ismailov, and P. Kollman. Amber 22, 2022.

Center for Free-Electron Laser Science. Femtosecond crystallography, 2020. URL cid.cfel.de/research/femtosecond_crystallography. Accessed: June 2020.

A. Cereto-Massagué, M. J. Ojeda, R. P. Joosten, C. Valls, M. Mulero, M. J. Salvado, A. Arola-Arnal, L. Arola, S. Garcia-Vallvé, and G. Pujadas. The good, the bad and the dubious: VHELIBS, a validation helper for ligands and binding sites. *Journal of cheminformatics*, 5(1):36, July 2013. ISSN 1758-2946. doi: 10.1186/1758-2946-5-36. URL https://pubmed.ncbi.nlm.nih.gov/23895374https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3733808/.

A. Cereto-Massagué, M. J. Ojeda, C. Valls, M. Mulero, S. Garcia-Vallvé, and G. Pujadas. Molecular fingerprint similarity search in virtual screening. *Methods*, 71:58–63, 2015. ISSN 1046-2023. doi: https://doi.org/10.1016/j.ymeth.2014.08.005. URL https://www.sciencedirect.com/science/article/pii/S1046202314002631.

S. Chakraborti, K. Hatti, and N. Srinivasan. 'All That Glitters Is Not Gold': High-Resolution Crystal Structures of Ligand-Protein Complexes Need Not Always Represent Confident Binding Poses. *International Journal of Molecular Sciences*, 22(13), 2021. ISSN 1422-0067. doi: 10.3390/ijms22136830. URL https://www.mdpi.com/1422-0067/22/13/6830.

A. Chandra, L. Tünnermann, T. Löfstedt, and R. Gratz. Transformer-based deep learning for predicting protein properties in the life sciences. *eLife*, 12:e82819, Jan. 2023. ISSN 2050-084X. doi: 10.7554/eLife.82819. URL https://doi.org/10.7554/eLife.82819.

M. Chartier and R. Najmanovich. Detection of Binding Site Molecular Interaction Field Similarities. *Journal of Chemical Information and Modeling*, 55(8):1600–1615, aug 2015. ISSN 1549-9596. doi: 10.1021/acs.jcim.5b00333. URL https://doi.org/10.1021/acs.jcim.5b00333.

D. Che, Q. Liu, K. Rasheed, and X. Tao. *Decision Tree and Ensemble Learning Algorithms with Their Applications in Bioinformatics*, pages 191–199. Springer New York, New York, NY, 2011. ISBN 978-1-4419-7046-6. doi: 10.1007/

978-1-4419-7046-6_19. URL https://doi.org/10.1007/978-1-4419-7046-6_19.

ChemAxon. Extended Connectivity Fingerprint ECFP, 2023. URL https://docs.chemaxon.com/display/docs/extended-connectivity-fingerprint-ecfp.md#src-1806333-extendedconnectivityfingerprintecfp-fig-2. Accessed: December 2023.

Chembridge. Chembridge Fragment Library, 2021. URL https://www.chembridge.com/screening_libraries/fragment_library/. Accessed June 2021.

L. Chemicals. Life Chemicals Fragment Libraries, 2021. URL https://lifechemicals.com/screening-libraries/fragment-libraries. Accessed: June 2021.

I.-J. Chen and R. E. Hubbard. Lessons for fragment library design: analysis of output from multiple screening campaigns. *Journal of Computer-Aided Molecular Design*, 23(8):603–620, 2009. ISSN 1573-4951. doi: 10.1007/s10822-009-9280-5. URL https://doi.org/10.1007/s10822-009-9280-5.

G. Chessari, R. Grainger, R. S. Holvey, R. F. Ludlow, P. N. Mortenson, and D. C. Rees. C–H functionalisation tolerant to polar groups could transform fragment-based drug discovery (FBDD). *Chem. Sci.*, 12(36):11976–11985, 2021. doi: 10.1039/D1SC03563K. URL http://dx.doi.org/10.1039/D1SC03563K.

D. Chicco. Ten quick tips for machine learning in computational biology. *BioData Mining*, 10(1):35, 2017. ISSN 1756-0381. doi: 10.1186/s13040-017-0155-3. URL https://doi.org/10.1186/s13040-017-0155-3.

T. A. Collier, T. J. Piggot, and J. R. Allison. *Molecular Dynamics Simulation of Proteins*, pages 311–327. Springer US, New York, NY, 2020. ISBN 978-1-4939-9869-2. doi: 10.1007/978-1-4939-9869-2_17. URL https://doi.org/10.1007/978-1-4939-9869-2_17.

P. M. Collins, A. Douangamath, R. Talon, A. Dias, J. Brandao-Neto, T. Krojer, and F. von Delft. Chapter Eleven - Achieving a Good Crystal System for Crystallographic X-Ray Fragment Screening. In C. A. B. T. M. i. E. Lesburg, editor, *Modern Approaches in Drug Discovery*, volume 610, pages 251–264. Academic Press, 2018. ISBN 0076-6879. doi: https://doi.org/10.1016/bs.mie.2018.09.027. URL http://www.sciencedirect.com/science/article/pii/S0076687918303902.

M. Congreve, R. Carr, C. Murray, and H. Jhoti. A 'rule of three' for fragment-based lead discovery? *Drug discovery today*, 8(19):876–877, Oct. 2003. ISSN 1359-6446 (Print). doi: 10.1016/s1359-6446(03)02831-9.

T. U. Consortium. UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Research*, 51(D1):D523–D531, 2022. ISSN 0305-1048. doi: 10.1093/nar/gkac1052. URL https://doi.org/10.1093/nar/gkac1052.

G. Corso, H. Stärk, B. Jing, R. Barzilay, and T. Jaakkola. DiffDock: Diffusion Steps, Twists, and Turns for Molecular Docking. *arXiv*, oct 2022. URL https://arxiv.org/abs/2210.01776.

V. G. Costa and C. E. Pedreira. Recent advances in decision trees: an updated survey. *Artificial Intelligence Review*, 56(5):4765–4800, 2023. ISSN 1573-7462. doi: 10.1007/s10462-022-10275-5. URL https://doi.org/10.1007/s10462-022-10275-5.

O. B. Cox, T. Krojer, P. Collins, O. Monteiro, R. Talon, A. Bradley, O. Fedorov, J. Amin, B. D. Marsden, J. Spencer, and et al. Human Pleckstrin Homology domain Interacting Protein (PHIP); A Target Enabling Package, June 2016a. URL https://zenodo.org/record/4134700.

O. B. Cox, T. Krojer, P. Collins, O. Monteiro, R. Talon, A. Bradley, O. Fedorov, J. Amin, B. D. Marsden, J. Spencer, F. von Delft, and P. E. Brennan. A poised fragment library enables rapid synthetic expansion yielding the first reported inhibitors of PHIP(2), an atypical bromodomain. *Chemical science*, 7(3):2322–2330, Mar. 2016b. ISSN 2041-6520 (Print). doi: 10.1039/c5sc03115j.

F. Da Silva, J. Desaphy, and D. Rognan. IChem: A Versatile Toolkit for Detecting, Comparing, and Predicting Protein–Ligand Interactions. *ChemMedChem*, 13(6):507–510, mar 2018. ISSN 1860-7179. doi: 10.1002/cmdc.201700505. URL https://doi.org/10.1002/cmdc.201700505.

S. Dara, S. Dhamercherla, S. S. Jadav, C. M. Babu, and M. J. Ahsan. Machine Learning in Drug Discovery: A Review. *Artificial intelligence review*, 55(3):1947–1999, 2022. ISSN 0269-2821 (Print). doi: 10.1007/s10462-021-10058-4.

D. R. Davies. *Screening Ligands by X-ray Crystallography*, pages 315–323. Springer New York, New York, NY, 2014. ISBN 978-1-4939-0354-2. doi: 10.1007/978-1-4939-0354-2_23. URL https://doi.org/10.1007/978-1-4939-0354-2_23.

T. A. de Oliveira, M. P. da Silva, E. H. B. Maia, A. M. da Silva, and A. G. Taranto. Virtual Screening Algorithms in Drug Discovery: A Review Focused on Machine and Deep Learning Methods. *Drugs and Drug Candidates*, 2(2):311–334, 2023. ISSN 2813-2998. doi: 10.3390/ddc2020017. URL https://www.mdpi.com/2813-2998/2/2/17.

L. R. de Souza Neto, J. T. Moreira-Filho, B. J. Neves, R. L. B. R. Maidana, A. C. R. Guimarães, N. Furnham, C. H. Andrade, and F. P. Silva. In silico Strategies to Support Fragment-to-Lead Optimization in Drug Discovery. *Frontiers in Chemistry*, 8, 2020. ISSN 2296-2646. doi: 10.3389/fchem.2020.00093. URL https://www.frontiersin.org/articles/10.3389/fchem.2020.00093.

M. De Vivo, M. Masetti, G. Bottegoni, and A. Cavalli. Role of Molecular Dynamics and Related Methods in Drug Discovery. *Journal of Medicinal Chemistry*, 59(9):

4035–4061, may 2016. ISSN 0022-2623. doi: 10.1021/acs.jmedchem.5b01684. URL https://doi.org/10.1021/acs.jmedchem.5b01684.

S. Delagenière, P. Brenchereau, L. Launer, A. W. Ashton, R. Leal, S. Veyrier, J. Gabadinho, E. J. Gordon, S. D. Jones, K. E. Levik, S. M. McSweeney, S. Monaco, M. Nanao, D. Spruce, O. Svensson, M. A. Walsh, and G. A. Leonard. ISPyB: an information management system for synchrotron macromolecular crystallography. *Bioinformatics (Oxford, England)*, 27(22):3186–3192, nov 2011. ISSN 1367-4811 (Electronic). doi: 10.1093/bioinformatics/btr535.

W. L. DeLano. Pymol: An open-source molecular graphics tool. *CCP4 Newsletter On Protein Crystallography*, 40:82–92, 2002.

M. C. Deller and B. Rupp. Models of protein-ligand crystal structures: trust, but verify. *Journal of computer-aided molecular design*, 29(9):817–836, Sept. 2015. ISSN 1573-4951 (Electronic). doi: 10.1007/s10822-015-9833-8.

J. Desaphy, K. Azdimousa, E. Kellenberger, and D. Rognan. Comparison and Druggability Prediction of Protein–Ligand Binding Sites from Pharmacophore-Annotated Cavity Shapes. *Journal of Chemical Information and Modeling*, 52 (8):2287–2299, Aug. 2012. ISSN 1549-9596. doi: 10.1021/ci300184x. URL https://doi.org/10.1021/ci300184x.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv*, oct 2018. URL http://arxiv.org/abs/1810.04805.

A. M. Díaz-Rovira, H. Martín, T. Beuming, L. Díaz, V. Guallar, and S. S. Ray. Are Deep Learning Structural Models Sufficiently Accurate for Virtual Screening? Application of Docking Algorithms to AlphaFold2 Predicted Structures. *Journal of Chemical Information and Modeling*, 63(6):1668–1674, Mar. 2023. ISSN 1549-9596. doi: 10.1021/acs.jcim.2c01270. URL https://doi.org/10.1021/acs.jcim.2c01270.

A. Douangamath, D. Fearon, P. Gehrtz, T. Krojer, P. Lukacik, C. D. Owen, E. Resnick, C. Strain-Damerell, A. Aimon, P. Ábrányi-Balogh, J. Brandaõ-Neto, A. Carbery, G. Davison, A. Dias, T. D. Downes, L. Dunnett, M. Fairhead, J. D. Firth, S. P. Jones, A. Keely, G. M. Keserü, H. F. Klein, M. P. Martin, M. E. M. Noble, P. O'Brien, A. Powell, R. Reddi, R. Skyner, M. Snee, M. J. Waring, C. Wild, N. London, F. von Delft, and M. A. Walsh. Crystallographic and electrophilic fragment screening of the SARS-CoV-2 main protease. *Nature Communications*, 11(5047), 2020. doi: https://doi.org/10.1038/s41467-020-18709-w. URL https://www.nature.com/articles/s41467-020-18709-w.

A. Douangamath, A. Powell, D. Fearon, P. M. Collins, R. Talon, T. Krojer, R. Skyner, J. Brandao-Neto, L. Dunnett, A. Dias, A. Aimon, N. M. Pearce, C. Wild, T. Gorrie-Stone, and F. von Delft. Achieving Efficient Fragment Screening at XChem Facility

at Diamond Light Source. *JoVE*, page e62414, 2021. ISSN 1940-087X. doi: doi: 10.3791/62414. URL https://www.jove.com/t/62414.

T. D. Downes, S. P. Jones, H. F. Klein, M. C. Wheldon, M. Atobe, P. S. Bond, J. D. Firth, N. S. Chan, L. Waddelove, R. E. Hubbard, D. C. Blakemore, C. De Fusco, S. D. Roughley, L. R. Vidler, M. A. Whatton, A. J.-A. Woolford, G. L. Wrigley, and P. O'Brien. Design and synthesis of 56 shape-diverse 3D fragments. *Chemistry – A European Journal*, 26(41):8969–8975, 2020. doi: https://doi.org/10.1002/chem.202001123. URL https://chemistry-europe.onlinelibrary.wiley.com/doi/abs/10.1002/chem.202001123.

G. Durant, F. Boyles, K. Birchall, B. Marsden, and C. M. Deane. Robustly interrogating machine learning-based scoring functions: what are they learning? *bioRxiv*, 2023. doi: 10.1101/2023.10.30.564251. URL https://www.biorxiv.org/content/early/2023/11/02/2023.10.30.564251.

J. L. Durant, B. A. Leland, D. R. Henry, and J. G. Nourse. Reoptimization of MDL Keys for Use in Drug Discovery. *Journal of Chemical Information and Computer Sciences*, 42(6):1273–1280, nov 2002. ISSN 0095-2338. doi: 10.1021/ci010132r. URL https://doi.org/10.1021/ci010132r.

P. Eastman, J. Swails, J. D. Chodera, R. T. McGibbon, Y. Zhao, K. A. Beauchamp, L.-P. Wang, A. C. Simmonett, M. P. Harrigan, C. D. Stern, R. P. Wiewiora, B. R. Brooks, and V. S. Pande. OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. *PLoS computational biology*, 13(7):e1005659, July 2017. ISSN 1553-7358 (Electronic). doi: 10.1371/journal.pcbi.1005659.

M. Eguida and D. Rognan. A Computer Vision Approach to Align and Compare Protein Cavities: Application to Fragment-Based Drug Design. *Journal of Medicinal Chemistry*, 63(13):7127–7142, July 2020. ISSN 0022-2623. doi: 10.1021/acs.jmedchem.0c00422. URL https://doi.org/10.1021/acs.jmedchem.0c00422.

M. Eguida and D. Rognan. Estimating the Similarity between Protein Pockets. *International Journal of Molecular Sciences*, 23(20), 2022. ISSN 1422-0067. doi: 10.3390/ijms232012462. URL https://www.mdpi.com/1422-0067/23/20/12462.

C. Ehrt, T. Brinkjost, and O. Koch. A benchmark driven guide to binding site comparison: An exhaustive evaluation using tailor-made data sets (ProSPECCTs). *PLOS Computational Biology*, 14(11):e1006483, Nov. 2018. URL https://doi.org/10.1371/journal.pcbi.1006483.

Enamine. Enamine Fragment Collection, 2020a. URL https://enamine.net/compound-libraries/fragment-libraries/. Accessed: June 2020.

Enamine. Enamine PPI fragment library, 2020b. URL https://enamine.net/compound-libraries/fragment-libraries/ppi-fragments. Accessed: June 2020.

D. Erlanson. Capivasertib: the seventh approved fragment-derived drug, 2023. URL https://practicalfragments.blogspot.com/2023/11/capivasertib-seventh-approved-fragment.html. Accessed: December 2023.

D. A. Erlanson, S. W. Fesik, R. E. Hubbard, W. Jahnke, and H. Jhoti. Twenty years on: the impact of fragments on drug discovery. *Nature reviews. Drug discovery*, 15 (9):605–619, Sept. 2016. ISSN 1474-1784 (Electronic). doi: 10.1038/nrd.2016.109.

M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD'96, pages 226–231. AAAI Press, 1996.

F. J. Fan and Y. Shi. Effects of data quality and quantity on deep learning for protein-ligand binding affinity prediction. *Bioorganic & Medicinal Chemistry*, 72:117003, 2022. ISSN 0968-0896. doi: https://doi.org/10.1016/j.bmc.2022.117003. URL https://www.sciencedirect.com/science/article/pii/S0968089622003960.

R. Ferreira de Freitas and M. Schapira. A systematic analysis of atomic protein-ligand interactions in the PDB. *MedChemComm*, 8(10):1970–1981, Oct. 2017. ISSN 2040-2503 (Print). doi: 10.1039/c7md00381a.

M. A. Fischler and R. C. Bolles. Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography. *Commun. ACM*, 24(6):381–395, jun 1981. ISSN 0001-0782. doi: 10.1145/358669.358692. URL https://doi.org/10.1145/358669.358692.

D. J. Foley, P. G. E. Craven, P. M. Collins, R. G. Doveston, A. Aimon, R. Talon, I. Churcher, F. von Delft, S. P. Marsden, and A. Nelson. Synthesis and demonstration of the biological relevance of sp3-rich scaffolds distantly related to natural product frameworks. *Chemistry – A European Journal*, 23(60):15227–15232, Oct. 2017. ISSN 0947-6539. doi: 10.1002/chem.201704169. URL https://doi.org/10.1002/chem.201704169.

D. Freedman, R. Pisani, and R. Purves. Statistics (international student edition). *Pisani, R. Purves, 4th edn. WW Norton & Company, New York*, 2007.

Y. Freund and R. E. Schapire. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997. ISSN 0022-0000. doi: https://doi.org/10.1006/jcss.1997.1504. URL https://www.sciencedirect.com/science/article/pii/S002200009791504X.

V. Früh, Y. Zhou, D. Chen, C. Loch, E. Ab, Y. N. Grinkova, H. Verheij, S. G. Sligar, J. H. Bushweller, and G. Siegal. Application of fragment-based drug discovery to membrane proteins: identification of ligands of the integral membrane enzyme DsbB. *Chemistry & biology*, 17(8):881–891, Aug. 2010. ISSN 1879-1301.

doi: 10.1016/j.chembiol.2010.06.011. URL https://pubmed.ncbi.nlm.nih.gov/20797617https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4104428/.

L. Fu, B. Niu, Z. Zhu, S. Wu, and W. Li. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics (Oxford, England)*, 28(23):3150–3152, Dec. 2012. ISSN 1367-4811. doi: 10.1093/bioinformatics/bts565. URL https://pubmed.ncbi.nlm.nih.gov/23060610https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3516142/.

N. Fuller, L. Spadola, S. Cowen, J. Patel, H. Schönherr, Q. Cao, A. McKenzie, F. Edfeldt, A. Rabow, and R. Goodnow. An improved model for fragment-based lead generation at AstraZeneca. *Drug Discovery Today*, 21(8):1272–1283, 2016. ISSN 1359-6446. doi: https://doi.org/10.1016/j.drudis.2016.04.023. URL https://www.sciencedirect.com/science/article/pii/S1359644616301490.

J. Gao, Q. Zhang, M. Liu, L. Zhu, D. Wu, Z. Cao, and R. Zhu. bSiteFinder, an improved protein-binding sites prediction server based on structural alignment: more accurate and less time-consuming. *Journal of Cheminformatics*, 8(1):38, 2016. ISSN 1758-2946. doi: 10.1186/s13321-016-0149-z. URL https://doi.org/10.1186/s13321-016-0149-z.

A. Gazizov, A. Lian, C. Goverde, S. Ovchinnikov, and N. F. Polizzi. AF2BIND: Predicting ligand-binding sites using the pair representation of AlphaFold2. *bioRxiv*, 2023. doi: 10.1101/2023.10.15.562410. URL https://www.biorxiv.org/content/early/2023/10/18/2023.10.15.562410.

D. Ghersi and R. Sanchez. Improving accuracy and efficiency of blind protein-ligand docking by focusing on predicted binding sites. *Proteins*, 74(2):417–424, feb 2009. ISSN 1097-0134 (Electronic). doi: 10.1002/prot.22154.

F. Giordanetto, C. Jin, L. Willmore, M. Feher, and D. E. Shaw. Fragment hits: what do they look like and low do they bind? *Journal of Medicinal Chemistry*, 62(7):3381–3394, Apr. 2019. ISSN 0022-2623. doi: 10.1021/acs.jmedchem.8b01855. URL https://pubmed.ncbi.nlm.nih.gov/30875465https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6466478/https://pubs.acs.org/doi/10.1021/acs.jmedchem.8b01855.

E. M. Gordon and J. F. Kerwin. *Combinatorial chemistry and molecular diversity in drug discovery*. Wiley-Liss, New York ; Chichester, 1998. ISBN 9780471155188.

J. C. Gordon, J. B. Myers, T. Folta, V. Shoja, L. S. Heath, and A. Onufriev. H++: a server for estimating pKas and adding missing hydrogens to macromolecules. *Nucleic acids research*, 33(Web Server issue):W368–71, July 2005. ISSN 1362-4962 (Electronic). doi: 10.1093/nar/gki464.

R. G. Govindaraj and M. Brylinski. Comparative assessment of strategies to identify similar ligand-binding pockets in proteins. *BMC bioinformatics*, 19(1):91, mar 2018. ISSN 1471-2105. doi: 10.1186/

s12859-018-2109-2. URL https://pubmed.ncbi.nlm.nih.gov/29523085https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5845264/.

R. Gowers, M. Linke, J. Barnoud, T. Reddy, M. Melo, S. L. Seyler, J. Domański, D. Dotson, S. Buchoux, I. Kenney, and O. Beckstein. MDAnalysis: A Python Package for the Rapid Analysis of Molecular Dynamics Simulations. In *Proceedings of the 15th Python in Science Conference*, pages 98–105, 01 2016. doi: 10.25080/Majora-629e541a-00e.

J. Graef, C. Ehrt, and M. Rarey. Binding Site Detection Remastered: Enabling Fast, Robust, and Reliable Binding Site Detection and Descriptor Calculation with DoGSite3. *Journal of Chemical Information and Modeling*, 63(10):3128–3137, may 2023. ISSN 1549-9596. doi: 10.1021/acs.jcim.3c00336. URL https://doi.org/10.1021/acs.jcim.3c00336.

R. Guha. On exploring structure-activity relationships. *Methods in molecular biology (Clifton, N.J.)*, 993:81–94, 2013. ISSN 1940-6029 (Electronic). doi: 10.1007/978-1-62703-342-8_6.

D. R. Hall, D. Kozakov, A. Whitty, and S. Vajda. Lessons from Hot Spot Analysis for Fragment-Based Drug Discovery. *Trends in pharmacological sciences*, 36(11):724–736, nov 2015. ISSN 1873-3735 (Electronic). doi: 10.1016/j.tips.2015.08.003.

R. J. Hall, P. N. Mortenson, and C. W. Murray. Efficient exploration of chemical space by fragment-based screening. *Progress in Biophysics and Molecular Biology*, 116(2-3):82–91, Nov. 2014. ISSN 00796107. doi: 10.1016/j.pbiomolbio.2014.09.007. URL https://linkinghub.elsevier.com/retrieve/pii/S0079610714000960.

S.-H. Han, K. W. Kim, S. Kim, and Y. C. Youn. Artificial Neural Network: Understanding the Basic Concepts without Mathematics. *Dementia and neurocognitive disorders*, 17(3):83–89, sep 2018. ISSN 2384-0757 (Electronic). doi: 10.12779/dnd.2018.17.3.83.

M. M. Hann, A. R. Leach, and G. Harper. Molecular complexity and its impact on the probability of finding leads for drug discovery. *Journal of chemical information and computer sciences*, 41(3):856–864, 2001. ISSN 0095-2338 (Print). doi: 10.1021/ci000403i.

R. J. Harding, R. Ferreira de Freitas, P. Collins, I. Franzoni, M. Ravichandran, H. Ouyang, K. A. Juarez-Ornelas, M. Lautens, M. Schapira, F. von Delft, V. Santhakumar, and C. H. Arrowsmith. Small Molecule Antagonists of the Interaction between the Histone Deacetylase 6 Zinc-Finger Domain and Ubiquitin. *Journal of Medicinal Chemistry*, 60(21):9090–9096, Nov. 2017. ISSN 0022-2623. doi: 10.1021/acs.jmedchem.7b00933. URL https://doi.org/10.1021/acs.jmedchem.7b00933.

M. J. Harner, A. O. Frank, and S. W. Fesik. Fragment-based drug discovery using NMR spectroscopy. *Journal of biomolecular NMR*, 56(2):65–75, jun 2013. ISSN 1573-5001 (Electronic). doi: 10.1007/s10858-013-9740-z.

M. L. Hekkelman, I. de Vries, R. P. Joosten, and A. Perrakis. AlphaFill: enriching AlphaFold models with ligands and cofactors. *Nature Methods*, 20(2):205–213, 2023. ISSN 1548-7105. doi: 10.1038/s41592-022-01685-y. URL https://doi.org/10.1038/s41592-022-01685-y.

S. R. Heller, A. McNaught, I. Pletnev, S. Stein, and D. Tchekhovskoi. InChI, the IUPAC International Chemical Identifier. *Journal of Cheminformatics*, 7(1):23, 2015. ISSN 1758-2946. doi: 10.1186/s13321-015-0068-4. URL https://doi.org/10.1186/s13321-015-0068-4.

S. A. Hicks, I. Strümke, V. Thambawita, M. Hammou, M. A. Riegler, P. Halvorsen, and S. Parasa. On evaluation metrics for medical applications of artificial intelligence. *Scientific reports*, 12(1):5979, apr 2022. ISSN 2045-2322 (Electronic). doi: 10.1038/s41598-022-09954-8.

M. H. Høie, F. S. Gade, J. M. Johansen, C. Würtzen, O. Winther, M. Nielsen, and P. Marcatili. DiscoTope-3.0 - Improved B-cell epitope prediction using AlphaFold2 modeling and inverse folding latent representations. *bioRxiv*, page 2023.02.05.527174, Jan. 2023. doi: 10.1101/2023.02.05.527174. URL http://biorxiv.org/content/early/2023/02/05/2023.02.05.527174.abstract.

M. Holcomb, Y.-T. Chang, D. S. Goodsell, and S. Forli. Evaluation of AlphaFold2 structures as docking targets. *Protein Science*, 32(1):e4530, 2023. doi: https://doi.org/10.1002/pro.4530. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/pro.4530.

S. A. Hollingsworth and R. O. Dror. Molecular Dynamics Simulation for All. *Neuron*, 99(6):1129–1143, 2018. ISSN 0896-6273. doi: https://doi.org/10.1016/j.neuron.2018.08.011. URL https://www.sciencedirect.com/science/article/pii/S0896627318306846.

A. L. Hopkins, C. R. Groom, and A. Alex. Ligand efficiency: a useful metric for lead selection. *Drug Discovery Today*, 9(10):430–431, 2004. ISSN 1359-6446. doi: https://doi.org/10.1016/S1359-6446(04)03069-7. URL https://www.sciencedirect.com/science/article/pii/S1359644604030697.

A. L. Hopkins, G. M. Keserü, P. D. Leeson, D. C. Rees, and C. H. Reynolds. The role of ligand efficiency metrics in drug discovery. *Nature Reviews Drug Discovery*, 13(2):105–121, 2014. ISSN 1474-1784. doi: 10.1038/nrd4163. URL https://doi.org/10.1038/nrd4163.

S. Horowitz. Electron density maps, 2020. URL www.nanowerk.com/nanotechnology-news/newsid=44730.php. Accessed: June 2020.

C. Hsu, R. Verkuil, J. Liu, Z. Lin, B. Hie, T. Sercu, A. Lerer, and A. Rives. Learning inverse folding from millions of predicted structures. In K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 8946–8970. PMLR, 2022. URL https://proceedings.mlr.press/v162/hsu22a.html.

L. Hu, M. L. Benson, R. D. Smith, M. G. Lerner, and H. A. Carlson. Binding MOAD (Mother Of All Databases). *Proteins: Structure, Function, and Bioinformatics*, 60(3):333–340, 2005. doi: https://doi.org/10.1002/prot.20512. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/prot.20512.

A. M. Hummer, C. Schneider, L. Chinery, and C. M. Deane. Investigating the Volume and Diversity of Data Needed for Generalizable Antibody-Antigen $\delta\delta$G Prediction. *bioRxiv*, 2023. doi: 10.1101/2023.05.17.541222. URL https://www.biorxiv.org/content/early/2023/05/19/2023.05.17.541222.

A. W. Hung, A. Ramek, Y. Wang, T. Kaya, J. A. Wilson, P. A. Clemons, and D. W. Young. Route to three-dimensional fragments using diversity-oriented synthesis. *Proceedings of the National Academy of Sciences*, 108(17):6799–6804, 2011. ISSN 0027-8424. doi: 10.1073/pnas.1015271108. URL https://www.pnas.org/content/108/17/6799.

F. Imrie, A. R. Bradley, and C. M. Deane. Generating property-matched decoy molecules using deep learning. *Bioinformatics*, 37(15):2134–2141, 2021. ISSN 1367-4803. doi: 10.1093/bioinformatics/btab080. URL https://doi.org/10.1093/bioinformatics/btab080.

D. Jakubec, P. Skoda, R. Krivak, M. Novotny, and D. Hoksza. PrankWeb 3: accelerated ligand-binding site predictions for experimental and modelled protein structures. *Nucleic Acids Research*, 50(W1):W593–W597, 2022. ISSN 0305-1048. doi: 10.1093/nar/gkac389. URL https://doi.org/10.1093/nar/gkac389.

G. Janson, G. Valdes-Garcia, L. Heo, and M. Feig. Direct generation of protein conformational ensembles via machine learning. *Nature Communications*, 14(1): 774, 2023. ISSN 2041-1723. doi: 10.1038/s41467-023-36443-x. URL https://doi.org/10.1038/s41467-023-36443-x.

M. Jaskolski, A. Wlodawer, Z. Dauter, W. Minor, and B. Rupp. Group depositions to the Protein Data Bank need adequate presentation and different archiving protocol. *Protein Science*, 31(4):784–786, 2022. doi: https://doi.org/10.1002/pro.4271. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/pro.4271.

X. Jin and J. Han. *K-Means Clustering*, pages 563–564. Springer US, Boston, MA, 2010. ISBN 978-0-387-30164-8. doi: 10.1007/978-0-387-30164-8_425. URL https://doi.org/10.1007/978-0-387-30164-8_425.

B. Jing, S. Eismann, P. Suriana, R. J. L. Townshend, and R. Dror. Learning from Protein Structure with Geometric Vector Perceptrons, 2021.

B. Jing, E. Erives, P. Pao-Huang, G. Corso, B. Berger, and T. Jaakkola. EigenFold: Generative Protein Structure Prediction with Diffusion Models. *arXiv*, Apr. 2023. URL http://arxiv.org/abs/2304.02198.

M. Johnson, M. Lajiness, and G. Maggiora. Molecular similarity: a basis for designing drug screening programs. *Progress in clinical and biological research*, 291:167–171, 1989. ISSN 0361-7742 (Print).

D. Jones, H. Kim, X. Zhang, A. Zemla, G. Stevenson, W. F. D. Bennett, D. Kirshner, S. E. Wong, F. C. Lightstone, and J. E. Allen. Improved Protein–Ligand Binding Affinity Prediction with Structure-Based Deep Fusion Inference. *Journal of Chemical Information and Modeling*, 61(4):1583–1592, apr 2021. ISSN 1549-9596. doi: 10.1021/acs.jcim.0c01306. URL https://doi.org/10.1021/acs.jcim.0c01306.

J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli, and D. Hassabis. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, 2021. ISSN 1476-4687. doi: 10.1038/s41586-021-03819-2. URL https://doi.org/10.1038/s41586-021-03819-2.

A. Kahraman, R. J. Morris, R. A. Laskowski, and J. M. Thornton. Shape Variation in Protein Binding Pockets and their Ligands. *Journal of Molecular Biology*, 368(1):283–301, 2007. ISSN 0022-2836. doi: https://doi.org/10.1016/j.jmb.2007.01.086. URL http://www.sciencedirect.com/science/article/pii/S0022283607001647.

A. Kahraman, R. J. Morris, R. A. Laskowski, A. D. Favia, and J. M. Thornton. On the diversity of physicochemical environments experienced by identical ligands in binding pockets of unrelated proteins. *Proteins*, 78(5):1120–1136, apr 2010. ISSN 1097-0134 (Electronic). doi: 10.1002/prot.22633.

J. Kandel, H. Tayara, and K. T. Chong. PUResNet: prediction of protein-ligand binding sites using deep residual neural network. *Journal of Cheminformatics*, 13 (1):65, 2021. ISSN 1758-2946. doi: 10.1186/s13321-021-00547-7. URL https://doi.org/10.1186/s13321-021-00547-7.

M. Karelina, J. J. Noh, and R. O. Dror. How accurately can one predict drug binding modes using AlphaFold models? *bioRxiv*, 2023. doi: 10.1101/2023.05.18.541346. URL https://www.biorxiv.org/content/early/2023/05/23/2023.05.18.541346.

M. Karplus and J. Kuriyan. Molecular dynamics and protein function. *Proceedings of the National Academy of Sciences*, 102(19):6679–6685, 2005. doi: 10.1073/pnas.0408930102. URL https://www.pnas.org/doi/abs/10.1073/pnas.0408930102.

G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 3149–3157, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

A. Keeley, P. Ábrányi-Balogh, and G. M. Keserű. Design and characterization of a heterocyclic electrophilic fragment library for the discovery of cysteine-targeted covalent inhibitors. *Med. Chem. Commun.*, 10(2):263–267, 2019. doi: 10.1039/C8MD00327K. URL http://dx.doi.org/10.1039/C8MD00327K.

A. Keeley, L. Petri, P. Ábrányi-Balogh, and G. M. Keserű. Covalent fragment libraries in drug discovery. *Drug Discovery Today*, 25(6):983–996, 2020. ISSN 1359-6446. doi: https://doi.org/10.1016/j.drudis.2020.03.016. URL https://www.sciencedirect.com/science/article/pii/S1359644620301239.

G. M. Keserű, D. A. Erlanson, G. G. Ferenczy, M. M. Hann, C. W. Murray, and S. D. Pickett. Design principles for fragment libraries: maximizing the value of learnings from pharma fragment-based drug discovery (FBDD) programs for use in academia. *Journal of Medicinal Chemistry*, 59(18):8189–8206, Sept. 2016. ISSN 0022-2623. doi: 10.1021/acs.jmedchem.6b00197. URL https://doi.org/10.1021/acs.jmedchem.6b00197.

N. A. Khazanov and H. A. Carlson. Exploring the Composition of Protein-Ligand Binding Sites on a Large Scale. *PLOS Computational Biology*, 9(11):1–14, 2013. doi: 10.1371/journal.pcbi.1003321. URL https://doi.org/10.1371/journal.pcbi.1003321.

S. L. Kidd, E. Fowler, T. Reinhardt, T. Compton, N. Mateu, H. Newman, D. Bellini, R. Talon, J. McLoughlin, T. Krojer, A. Aimon, A. Bradley, M. Fairhead, P. Brear, L. Díaz-Sáez, K. McAuley, H. F. Sore, A. Madin, D. H. O'Donovan, K. V. M. Huber, M. Hyvönen, F. von Delft, C. G. Dowson, and D. R. Spring. Demonstration of the utility of DOS-derived fragment libraries for rapid hit derivatisation in a multidirectional fashion. *Chem. Sci.*, 11(39):10792–10801, 2020. doi: 10.1039/D0SC01232G. URL http://dx.doi.org/10.1039/D0SC01232G.

T. B. Kimber, Y. Chen, and A. Volkamer. Deep Learning in Virtual Screening: Recent Applications and Developments. *International journal of molecular sciences*, 22(9), apr 2021. ISSN 1422-0067 (Electronic). doi: 10.3390/ijms22094435.

J. Kirkwood, D. Hargreaves, S. O'Keefe, and J. Wilson. Analysis of crystallization data in the Protein Data Bank. *Acta crystallographica. Section F, Structural biology communications*, 71(Pt 10):1228–1234, oct 2015. ISSN 2053-230X (Electronic). doi: 10.1107/S2053230X15014892.

J. J. Koehler. The Influence of Prior Beliefs on Scientific Judgments of Evidence Quality. *Organizational Behavior and Human Decision Processes*, 56(1):28–55, 1993. ISSN 0749-5978. doi: https://doi.org/10.1006/obhd.1993.1044. URL http://www.sciencedirect.com/science/article/pii/S0749597883710447.

J. Konc and D. Janezic. ProBiS algorithm for detection of structurally similar protein binding sites by local structural alignment. *Bioinformatics (Oxford, England)*, 26(9):1160–1168, may 2010. ISSN 1367-4811 (Electronic). doi: 10.1093/bioinformatics/btq100.

Z. Konteatis. What makes a good fragment in fragment-based drug discovery? *Expert Opinion on Drug Discovery*, 16(7):723–726, 2021. doi: 10.1080/17460441.2021.1905629. URL https://doi.org/10.1080/17460441.2021.1905629.

D. Kozakova, D. R. Hall, S. Jehle, L. Luo, S. O. Ochiana, E. V. Jones, M. Pollastri, K. N. Allen, A. Whitty, and S. Vajda. Ligand deconstruction: Why some fragment binding positions are conserved and others are not. *Proceedings of the National Academy of Sciences of the United States of America*, 112(20):E2585–E2594, may 2015. ISSN 10916490. doi: 10.1073/pnas.1501567112.

I. Kozlovskii and P. Popov. Spatiotemporal identification of druggable binding sites using deep learning. *Communications Biology*, 3(1):618, 2020. ISSN 2399-3642. doi: 10.1038/s42003-020-01350-0. URL https://doi.org/10.1038/s42003-020-01350-0.

R. Krivák and D. Hoksza. P2Rank: machine learning based tool for rapid and accurate prediction of ligand binding sites from protein structure. *Journal of Cheminformatics*, 10(1):39, 2018. ISSN 1758-2946. doi: 10.1186/s13321-018-0285-8. URL https://doi.org/10.1186/s13321-018-0285-8.

A. Kryshtafovych, T. Schwede, M. Topf, K. Fidelis, and J. Moult. Critical assessment of methods of protein structure prediction (CASP)–Round XIV. *Proteins: Structure, Function, and Bioinformatics*, 89(12):1607–1617, 2021. doi: https://doi.org/10.1002/prot.26237. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/prot.26237.

A. Kumar, A. Voet, and K. Y. J. Zhang. Fragment based drug design: from experimental to computational approaches. *Current medicinal chemistry*, 19(30):5128–5147, 2012. ISSN 1875-533X (Electronic). doi: 10.2174/092986712803530467.

R. Kurczab, S. Smusz, and A. J. Bojarski. The influence of negative training set size on machine learning-based virtual screening. *Journal of Cheminformatics*, 6(1):32, 2014. ISSN 1758-2946. doi: 10.1186/1758-2946-6-32. URL https://doi.org/10.1186/1758-2946-6-32.

P. S. Kutchukian, A. M. Wassermann, M. K. Lindvall, S. K. Wright, J. Ottl, J. Jacob, C. Scheufler, A. Marzinzik, N. Brooijmans, and M. Glick. Large scale meta-analysis of fragment-based screening campaigns: privileged fragments and complementary

technologies. *Journal of Biomolecular Screening*, 20(5):588–596, Dec. 2014. ISSN 1087-0571. doi: 10.1177/1087057114565080.

G. Landrum. RDKit: Open-source cheminformatics, 2020. Accessed: April 2020.

T. J. Lane. Protein structure prediction has reached the single-structure frontier. *Nature Methods*, 20(2):170–173, 2023. ISSN 1548-7105. doi: 10.1038/s41592-022-01760-4. URL https://doi.org/10.1038/s41592-022-01760-4.

W. F. Lau, J. M. Withka, D. Hepworth, T. V. Magee, Y. J. Du, G. A. Bakken, M. D. Miller, Z. S. Hendsch, V. Thanabal, S. A. Kolodziej, L. Xing, Q. Hu, L. S. Narasimhan, R. Love, M. E. Charlton, S. Hughes, W. P. van Hoorn, and J. E. Mills. Design of a multi-purpose fragment screening library using molecular complexity and orthogonal diversity metrics. *Journal of Computer-Aided Molecular Design*, 25(7):621, 2011. ISSN 1573-4951. doi: 10.1007/s10822-011-9434-0. URL https://doi.org/10.1007/s10822-011-9434-0.

V. Le Guilloux, P. Schmidtke, and P. Tuffery. Fpocket: an open source platform for ligand pocket detection. *BMC bioinformatics*, 10:168, June 2009. ISSN 1471-2105 (Electronic). doi: 10.1186/1471-2105-10-168.

H. S. Lee and W. Im. Ligand binding site detection by local structure alignment and its performance complementarity. *Journal of chemical information and modeling*, 53(9):2462–2470, sep 2013. ISSN 1549-960X (Electronic). doi: 10.1021/ci4003602.

I. Lee and H. Nam. Sequence-based prediction of protein binding regions and drug–target interactions. *Journal of Cheminformatics*, 14(1):5, 2022. ISSN 1758-2946. doi: 10.1186/s13321-022-00584-w. URL https://doi.org/10.1186/s13321-022-00584-w.

P. D. Leeson, A. P. Bento, A. Gaulton, A. Hersey, E. J. Manners, C. J. Radoux, and A. R. Leach. Target-Based Evaluation of "Drug-Like" Properties and Ligand Efficiencies. *Journal of Medicinal Chemistry*, 64(11):7210–7230, jun 2021. ISSN 0022-2623. doi: 10.1021/acs.jmedchem.1c00416. URL https://doi.org/10.1021/acs.jmedchem.1c00416.

J. Li, X. Guan, O. Zhang, K. Sun, Y. Wang, D. Bagni, and T. Head-Gordon. Leak Proof PDBBind: A Reorganized Dataset of Protein-Ligand Complexes for More Generalizable Binding Affinity Prediction., aug 2023. ISSN 2331-8422 (Electronic).

J. Liebeschuetz, J. Hennemann, T. Olsson, and C. R. Groom. The good, the bad and the twisted: a survey of ligand geometry in protein crystal structures. *Journal of computer-aided molecular design*, 26(2):169–183, Feb. 2012. ISSN 1573-4951. doi: 10.1007/s10822-011-9538-6. URL https://pubmed.ncbi.nlm.nih.gov/22246295https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3292722/.

Z. Lin, H. Akin, R. Rao, B. Hie, Z. Zhu, W. Lu, N. Smetanin, R. Verkuil, O. Kabeli, Y. Shmueli, A. dos Santos Costa, M. Fazel-Zarandi, T. Sercu, S. Candido, and

A. Rives. Evolutionary-scale prediction of atomic level protein structure with a language model. *bioRxiv*, 2022. doi: 10.1101/2022.07.20.500902. URL https://www.biorxiv.org/content/early/2022/12/21/2022.07.20.500902.

Z. Lin, H. Akin, R. Rao, B. Hie, Z. Zhu, W. Lu, N. Smetanin, R. Verkuil, O. Kabeli, Y. Shmueli, A. dos Santos Costa, M. Fazel-Zarandi, T. Sercu, S. Candido, and A. Rives. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023. doi: 10.1126/science.ade2574. URL https://www.science.org/doi/abs/10.1126/science.ade2574.

J. Liu, Z. Guo, T. Wu, R. S. Roy, C. Chen, and J. Cheng. Improving AlphaFold2-based Protein Tertiary Structure Prediction with MULTICOM in CASP15. *bioRxiv*, 2023. doi: 10.1101/2023.05.01.538929. URL https://www.biorxiv.org/content/early/2023/05/01/2023.05.01.538929.

T. Liu and R. B. Altman. Using multiple microenvironments to find similar ligand-binding sites: application to kinase inhibitor binding. *PLoS computational biology*, 7(12):e1002326, dec 2011. ISSN 1553-7358 (Electronic). doi: 10.1371/journal.pcbi.1002326.

Z. Liu, Y. Li, L. Han, J. Li, J. Liu, Z. Zhao, W. Nie, Y. Liu, and R. Wang. PDB-wide collection of binding data: current status of the PDBbind database. *Bioinformatics*, 31(3):405–412, 2014. ISSN 1367-4803. doi: 10.1093/bioinformatics/btu626. URL https://doi.org/10.1093/bioinformatics/btu626.

R. Longadge and S. Dongre. Class Imbalance Problem in Data Mining Review. *arXiv*, may 2013. URL http://arxiv.org/abs/1305.1707.

J. Loschwitz, O. O. Olubiyi, J. S. Hub, B. Strodel, and C. S. Poojari. Computer simulations of protein-membrane systems. *Progress in molecular biology and translational science*, 170:273–403, 2020. ISSN 1878-0814 (Electronic). doi: 10.1016/bs.pmbts.2020.01.001.

Lysozyme.co.uk. Lysozyme, 2020. URL lysozyme.co.uk/crystallization.php. Accessed: June 2020.

Y. C. Martin, J. L. Kofron, and L. M. Traphagen. Do structurally similar molecules have similar biological activity? *Journal of Medicinal Chemistry*, 45(19):4350–4358, Sept. 2002. ISSN 00222623. doi: 10.1021/jm020155c. URL https://doi.org/10.1021/jm020155c.

Maybridge. Maybridge Fragment Library, 2021. URL https://www.thermofisher.com/uk/en/home/industrial/pharma-biopharma/drug-discovery-development/screening-compounds-libraries-hit-identification/maybridge-fragment-libraries.html. Accessed: June 2021.

A. Mayr, H. Binder, O. Gefeller, and M. Schmid. The Evolution of Boosting Algorithms. *From Machine Learning to Statistical Modelling*, 53(06):419–427, 2014. ISSN 0026-1270. doi: 10.3414/ME13-01-0122. URL http://www.thieme-connect.de/products/ejournals/abstract/10.3414/ME13-01-0122.

I. K. McDonald and J. M. Thornton. Satisfying Hydrogen Bonding Potential in Proteins. *Journal of Molecular Biology*, 238(5):777–793, 1994. ISSN 0022-2836. doi: https://doi.org/10.1006/jmbi.1994.1334. URL https://www.sciencedirect.com/science/article/pii/S0022283684713349.

J. E. McGreig, H. Uri, M. Antczak, M. J. E. Sternberg, M. Michaelis, and M. N. Wass. 3DLigandSite: structure-based prediction of protein–ligand binding sites. *Nucleic Acids Research*, 50(W1):W13–W20, 2022. ISSN 0305-1048. doi: 10.1093/nar/gkac250. URL https://doi.org/10.1093/nar/gkac250.

P. J. McIntyre, P. M. Collins, L. Vrzal, K. Birchall, L. H. Arnold, C. Mpamhanga, P. J. Coombs, S. G. Burgess, M. W. Richards, A. Winter, V. Veverka, F. von Delft, A. Merritt, and R. Bayliss. Characterization of Three Druggable Hot-Spots in the Aurora-A/TPX2 Interaction Using Biochemical, Biophysical, and Fragment-Based Approaches. *ACS Chemical Biology*, 12(11):2906–2914, Nov. 2017. ISSN 1554-8929. doi: 10.1021/acschembio.7b00537. URL https://doi.org/10.1021/acschembio.7b00537.

Microsoft. AutoML, 2022. URL https://github.com/microsoft/FLAML. Accessed: October 2022.

J. Mistry, S. Chuguransky, L. Williams, M. Qureshi, G. A. Salazar, E. L. L. Sonnhammer, S. C. E. Tosatto, L. Paladin, S. Raj, L. J. Richardson, R. D. Finn, and A. Bateman. Pfam: The protein families database in 2021. *Nucleic Acids Research*, 49(D1):D412–D419, 2020. ISSN 0305-1048. doi: 10.1093/nar/gkaa913. URL https://doi.org/10.1093/nar/gkaa913.

A. D. Morley, A. Pugliese, K. Birchall, J. Bower, P. Brennan, N. Brown, T. Chapman, M. Drysdale, I. H. Gilbert, S. Hoelder, A. Jordan, S. V. Ley, A. Merritt, D. Miller, M. E. Swarbrick, and P. G. Wyatt. Fragment-based hit identification: thinking in 3D. *Drug Discovery Today*, 18(23):1221–1227, 2013. ISSN 1359-6446. doi: https://doi.org/10.1016/j.drudis.2013.07.011. URL http://www.sciencedirect.com/science/article/pii/S1359644613002456.

J. Moult, J. T. Pedersen, R. Judson, and K. Fidelis. A large-scale experiment to assess protein structure prediction methods. *Proteins: Structure, Function, and Bioinformatics*, 23(3):ii–iv, 1995. doi: https://doi.org/10.1002/prot.340230303. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/prot.340230303.

C. W. Murray, M. L. Verdonk, and D. C. Rees. Experiences in fragment-based drug discovery. *Trends in Pharmacological Sciences*, 33(5):224–232, 2012. ISSN 01656147. doi: 10.1016/j.tips.2012.02.006. URL https://www.sciencedirect.com/science/article/pii/S0165614712000326.

J. Myers, G. Grothaus, S. Narayanan, and A. Onufriev. A simple clustering algorithm can be accurate enough for use in calculations of pKs in macromolecules. *Proteins*, 63(4):928–938, June 2006. ISSN 1097-0134 (Electronic). doi: 10.1002/prot.20922.

S. K. Mylonas, A. Axenopoulos, and P. Daras. DeepSurf: a surface-based deep learning approach for the prediction of ligand binding sites on proteins. *Bioinformatics*, 37(12):1681–1690, 2021. ISSN 1367-4803. doi: 10.1093/bioinformatics/btab009. URL https://doi.org/10.1093/bioinformatics/btab009.

M. M. Mysinger, M. Carchia, J. J. Irwin, and B. K. Shoichet. Directory of Useful Decoys, Enhanced (DUD-E): Better Ligands and Decoys for Better Benchmarking. *Journal of Medicinal Chemistry*, 55(14):6582–6594, jul 2012. ISSN 0022-2623. doi: 10.1021/jm300687e. URL https://doi.org/10.1021/jm300687e.

A. Natekin and A. Knoll. Gradient boosting machines, a tutorial. *Frontiers in neurorobotics*, 7:21, 2013. ISSN 1662-5218 (Print). doi: 10.3389/fnbot.2013.00021.

S. B. Needleman and C. D. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48(3):443–453, 1970. ISSN 0022-2836. doi: https://doi.org/10.1016/0022-2836(70)90057-4. URL https://www.sciencedirect.com/science/article/pii/0022283670900574.

J. Newman, A. Gavard, H. Aitkenhead, D. Drewry, C. Lin, and O. Gileadi. Human T-box transcription factor T (Brachyury); A Target Enabling Package, Nov. 2020a. URL https://zenodo.org/record/4265383.

J. Newman, Y. Yosaatmadja, A. Douangamath, F. V. Delft, and O. Gileadi. SARS-CoV-2 NSP13; A Target Enabling Package, Nov. 2020b. URL https://zenodo.org/record/4449959.

C. H. Ngan, T. Bohnuud, S. E. Mottarella, D. Beglov, E. A. Villar, D. R. Hall, D. Kozakov, and S. Vajda. FTMAP: extended protein mapping with user-selected probe molecules. *Nucleic acids research*, 40(Web Server issue):W271–W275, July 2012. ISSN 1362-4962. doi: 10.1093/nar/gks441. URL https://pubmed.ncbi.nlm.nih.gov/22589414https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3394268/.

N. S. Ogando, J. C. Zevenhoven-Dobbe, Y. van der Meer, P. J. Bredenbeek, C. C. Posthuma, E. J. Snijder, and T. Gallagher. The Enzymatic Activity of the nsp14 Exoribonuclease Is Critical for Replication of MERS-CoV and SARS-CoV-2. *Journal of Virology*, 94(23):e01246–20, 2020. doi: 10.1128/JVI.01246-20.

E. openscreen. EU openscreen Fragment Library, 2021. URL https://www.eu-openscreen.eu/. Accessed: June 2021.

M. O'Reilly, A. Cleasby, T. G. Davies, R. J. Hall, R. F. Ludlow, C. W. Murray, D. Tisi, and H. Jhoti. Crystallographic screening using ultra-low-molecular-weight

ligands to guide drug design. *Drug discovery today*, 24(5):1081–1086, May 2019. ISSN 1878-5832 (Electronic). doi: 10.1016/j.drudis.2019.03.009.

K. O'Shea and R. Nash. An Introduction to Convolutional Neural Networks, 2015.

J. B. Ott and J. Boerio-Goates. Chapter 1 - Introduction. In J. B. Ott and J. Boerio-Goates, editors, *Chemical Thermodynamics: Principles and Applications*, pages 1–36. Academic Press, London, 2000. ISBN 978-0-12-530990-5. doi: https://doi.org/10.1016/B978-012530990-5/50002-X. URL https://www.sciencedirect.com/science/article/pii/B978012530990550002X.

R. Özçelik, D. van Tilborg, J. Jiménez-Luna, and F. Grisoni. Structure-based drug discovery with deep learning, Dec. 2022. URL https://arxiv.org/abs/2212.13295.

C. G. Parker, A. Galmozzi, Y. Wang, B. E. Correia, K. Sasaki, C. M. Joslyn, A. S. Kim, C. L. Cavallaro, R. M. Lawrence, S. R. Johnson, I. Narvaiza, E. Saez, and B. F. Cravatt. Ligand and target discovery by fragment-based screening in human cells. *Cell*, 168(3):527–541.e29, Jan. 2017. ISSN 0092-8674. doi: 10.1016/j.cell.2016.12.029. URL https://doi.org/10.1016/j.cell.2016.12.029.

D. Patel, J. D. Bauman, and E. Arnold. Advantages of crystallographic fragment screening: Functional and mechanistic insights from a powerful platform for efficient drug discovery. *Progress in Biophysics and Molecular Biology*, 116(2):92–100, 2014. ISSN 0079-6107. doi: https://doi.org/10.1016/j.pbiomolbio.2014.08.004. URL http://www.sciencedirect.com/science/article/pii/S0079610714000790.

N. Pearce. PanDDA Tutorials, 2020. URL https://pandda.bitbucket.io/index.html. Accessed: June 2020.

N. M. Pearce, T. Krojer, A. R. Bradley, P. Collins, R. P. Nowak, R. Talon, B. D. Marsden, S. Kelm, J. Shi, C. M. Deane, and F. von Delft. A multi-crystal method for extracting obscured crystallographic states from conventionally uninterpretable electron density. *Nature communications*, 8:15123, Apr. 2017. ISSN 2041-1723 (Electronic). doi: 10.1038/ncomms15123.

N. M. Pearce, R. Skyner, and T. Krojer. Experiences From Developing Software for Large X-Ray Crystallography-Driven Protein-Ligand Studies. *Frontiers in Molecular Biosciences*, 9, 2022. ISSN 2296-889X. doi: 10.3389/fmolb.2022.861491. URL https://www.frontiersin.org/articles/10.3389/fmolb.2022.861491.

W. R. Pearson. An introduction to sequence similarity ("homology") searching. *Current protocols in bioinformatics*, Chapter 3:3.1.1–3.1.8, jun 2013. ISSN 1934-340X (Electronic). doi: 10.1002/0471250953.bi0301s42.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.*, 12(null):2825–2830, Nov. 2011. ISSN 1532-4435.

S. Pérot, O. Sperandio, M. A. Miteva, A.-C. Camproux, and B. O. Villoutreix. Druggable pockets and binding site centric chemical space: a paradigm shift in drug discovery. *Drug Discovery Today*, 15(15):656–667, 2010. ISSN 1359-6446. doi: https://doi.org/10.1016/j.drudis.2010.05.015. URL https://www.sciencedirect.com/science/article/pii/S1359644610001911.

S. L. Posy, M. A. Hermsmeier, W. Vaccaro, K.-H. Ott, G. Todderud, J. S. Lippy, G. L. Trainor, D. A. Loughney, and S. R. Johnson. Trends in Kinase Selectivity: Insights for Target Class-Focused Library Screening. *Journal of Medicinal Chemistry*, 54 (1):54–66, Jan. 2011. ISSN 0022-2623. doi: 10.1021/jm101195a. URL https://doi.org/10.1021/jm101195a.

E. Pozharski, C. X. Weichenberger, and B. Rupp. Techniques, tools and best practices for ligand electron-density analysis and results from their application to deposited crystal structures. *Acta Crystallographica Section D*, 69(2):150–167, Feb. 2013. doi: 10.1107/S0907444912044423. URL https://doi.org/10.1107/S0907444912044423.

Prestwick. Preswick Drug-Fragment Library, 2021. URL https://www.prestwickchemical.com/screening-libraries/prestwick-drug-fragment-library/. Accessed: June 2021.

L. Pu, R. G. Govindaraj, J. M. Lemoine, H.-C. Wu, and M. Brylinski. DeepDrug3D: Classification of ligand-binding pockets in proteins with a convolutional neural network. *PLOS Computational Biology*, 15(2):1–23, 2019. doi: 10.1371/journal.pcbi.1006718. URL https://doi.org/10.1371/journal.pcbi.1006718.

C. J. Radoux, T. S. G. Olsson, W. R. Pitt, C. R. Groom, and T. L. Blundell. Identifying Interactions that Determine Fragment Binding at Protein Hotspots. *Journal of Medicinal Chemistry*, 59(9):4314–4325, 2016. doi: 10.1021/acs.jmedchem.5b01980. URL https://doi.org/10.1021/acs.jmedchem.5b01980.

M. S. Rao, R. Gupta, M. J. Liguori, M. Hu, X. Huang, S. R. Mantena, S. W. Mittelstadt, E. A. G. Blomme, and T. R. Van Vleet. Novel Computational Approach to Predict Off-Target Interactions for Small Molecules. *Frontiers in big data*, 2:25, 2019. ISSN 2624-909X (Electronic). doi: 10.3389/fdata.2019.00025.

P. A. Ravindranath and M. F. Sanner. AutoSite: an automated approach for pseudo-ligands prediction-from ligand-binding sites identification to predicting key ligand atoms. *Bioinformatics (Oxford, England)*, 32(20):3142–3149, Oct. 2016. ISSN 1367-4811 (Electronic). doi: 10.1093/bioinformatics/btw367.

P. C. Ray, M. Kiczun, M. Huggett, A. Lim, F. Prati, I. H. Gilbert, and P. G. Wyatt. Fragment library design, synthesis and expansion: nurturing a synthesis and training platform. *Drug Discovery Today*, 22(1):43–56, 2017. ISSN 1359-6446. doi: https://doi.org/10.1016/j.drudis.2016.10.005. URL https://www.sciencedirect.com/science/article/pii/S1359644616303816.

D. C. Rees, A. K. H. Hirsch, and D. A. Erlanson. Introduction to the themed collection on fragment-based drug discovery. *RSC Med. Chem.*, 13(12):1439, 2022. doi: 10.1039/D2MD90037H. URL http://dx.doi.org/10.1039/D2MD90037H.

E. Resnick, A. Bradley, J. Gan, A. Douangamath, T. Krojer, R. Sethi, P. P. Geurink, A. Aimon, G. Amitai, D. Bellini, J. Bennett, M. Fairhead, O. Fedorov, R. Gabizon, J. Gan, J. Guo, A. Plotnikov, N. Reznik, G. F. Ruda, L. Díaz-Sáez, V. M. Straub, T. Szommer, S. Velupillai, D. Zaidman, Y. Zhang, A. R. Coker, C. G. Dowson, H. M. Barr, C. Wang, K. V. M. Huber, P. E. Brennan, H. Ovaa, F. von Delft, and N. London. Rapid covalent-probe discovery by electrophile-fragment screening. *Journal of the American Chemical Society*, 141(22):8951–8968, June 2019. ISSN 0002-7863. doi: 10.1021/jacs.9b02822. URL https://doi.org/10.1021/jacs.9b02822.

J. Revillo Imbernon, C. Jacquemard, G. Bret, G. Marcou, and E. Kellenberger. Comprehensive analysis of commercial fragment libraries. *RSC Med. Chem.*, pages –, 2022. doi: 10.1039/D1MD00363A. URL http://dx.doi.org/10.1039/D1MD00363A.

C. H. Reynolds, S. D. Bembenek, and B. A. Tounge. The role of molecular size in ligand efficiency. *Bioorganic & Medicinal Chemistry Letters*, 17(15):4258–4261, 2007. ISSN 0960-894X. doi: https://doi.org/10.1016/j.bmcl.2007.05.038. URL https://www.sciencedirect.com/science/article/pii/S0960894X07005914.

T. Ritschel, T. J. Schirris, and F. G. Russel. KRIPO - a structure-based pharmacophores approach explains polypharmacological effects. *Journal of cheminformatics*, 6(Suppl 1):O26–O26, mar 2014. ISSN 1758-2946. doi: 10.1186/1758-2946-6-S1-O26. URL https://pubmed.ncbi.nlm.nih.gov/24765125https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3980168/.

A. Rives, J. Meier, T. Sercu, S. Goyal, Z. Lin, J. Liu, D. Guo, M. Ott, C. L. Zitnick, J. Ma, and R. Fergus. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences of the United States of America*, 118(15):e2016239118, Apr. 2021. ISSN 1091-6490 (Electronic). doi: 10.1073/pnas.2016239118. URL https://www.pnas.org/doi/abs/10.1073/pnas.2016239118.

D. Rogers and M. Hahn. Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling*, 50(5):742–754, May 2010. ISSN 1549-9596. doi: 10.1021/ci100050t. URL https://doi.org/10.1021/ci100050t.

S. Salentin, S. Schreiber, V. J. Haupt, M. F. Adasme, and M. Schroeder. PLIP: fully automated protein-ligand interaction profiler. *Nucleic acids research*, 43(W1):W443–7, jul 2015. ISSN 1362-4962 (Electronic). doi: 10.1093/nar/gkv315.

J. Scantlebury, N. Brown, F. Von Delft, and C. M. Deane. Data Set Augmentation Allows Deep Learning-Based Virtual Screening to Better Generalize to Unseen

Target Classes and Highlight Important Binding Interactions. *Journal of Chemical Information and Modeling*, 60(8):3722–3730, aug 2020. ISSN 1549-9596. doi: 10.1021/acs.jcim.0c00263. URL https://doi.org/10.1021/acs.jcim.0c00263.

J. Scantlebury, L. Vost, A. Carbery, T. E. Hadfield, O. M. Turnbull, N. Brown, V. Chenthamarakshan, P. Das, H. Grosjean, F. Von Delft, and C. M. Deane. A Small Step Toward Generalizability: Training a Machine Learning Scoring Function for Structure-Based Virtual Screening. *Journal of Chemical Information and Modeling*, 63(10):2960–2974, 2023. ISSN 1549960X. doi: 10.1021/acs.jcim.3c00322.

V. Scardino, J. I. Di Filippo, and C. N. Cavasotto. How good are AlphaFold models for docking-based virtual screening? *iScience*, 26(1):105920, Jan. 2023. ISSN 2589-0042 (Electronic). doi: 10.1016/j.isci.2022.105920.

J. Schiebel, S. G. Krimmer, K. Röwer, A. Knörlein, X. Wang, A. Y. Park, M. Stieler, F. R. Ehrmann, K. Fu, N. Radeva, M. Krug, F. U. Huschmann, S. Glöckner, M. S. Weiss, U. Mueller, G. Klebe, and A. Heine. High-Throughput Crystallography: Reliable and Efficient Identification of Fragment Hits. *Structure*, 24(8):1398–1409, 2016. ISSN 0969-2126. doi: https://doi.org/10.1016/j.str.2016.06.010. URL http://www.sciencedirect.com/science/article/pii/S0969212616301356.

P. Schmidtke, C. Souaille, F. Estienne, N. Baurin, and R. T. Kroemer. Large-scale comparison of four binding site detection algorithms. *Journal of chemical information and modeling*, 50(12):2191–2200, dec 2010. ISSN 1549-960X (Electronic). doi: 10.1021/ci1000289.

S. L. Schreiber. Molecular diversity by design. *Nature*, 457(7226):153–154, 2009. ISSN 1476-4687. doi: 10.1038/457153a. URL https://doi.org/10.1038/457153a.

A. M. Schreyer and T. Blundell. USRCAT: real-time ultrafast shape recognition with pharmacophoric constraints. *Journal of Cheminformatics*, 4(1):27, 2012. ISSN 1758-2946. doi: 10.1186/1758-2946-4-27. URL https://doi.org/10.1186/1758-2946-4-27.

Schrödinger, LLC. The PyMOL Molecular Graphics System, Version.8. Nov. 2015.

M. Schuller, G. J. Correy, S. Gahbauer, D. Fearon, T. Wu, R. E. Díaz, I. D. Young, L. C. Martins, D. H. Smith, U. Schulze-Gahmen, T. W. Owens, I. Deshpande, G. E. Merz, A. C. Thwin, J. T. Biel, J. K. Peters, M. Moritz, N. Herrera, H. T. Kratochvil, null null, A. Aimon, J. M. Bennett, J. B. Neto, A. E. Cohen, A. Dias, A. Douangamath, L. Dunnett, O. Fedorov, M. P. Ferla, M. R. Fuchs, T. J. Gorrie-Stone, J. M. Holton, M. G. Johnson, T. Krojer, G. Meigs, A. J. Powell, J. G. M. Rack, V. L. Rangel, S. Russi, R. E. Skyner, C. A. Smith, A. S. Soares, J. L. Wierman, K. Zhu, P. O'Brien, N. Jura, A. Ashworth, J. J. Irwin, M. C. Thompson, J. E. Gestwicki, F. von Delft, B. K. Shoichet, J. S. Fraser, and I. Ahel. Fragment binding to the Nsp3 macrodomain of SARS-CoV-2 identified through crystallographic screening and computational docking. *Science Advances*, 7(16):eabf8711, 2021. doi: 10.1126/sciadv.abf8711.

D. E. Scott, A. G. Coyne, S. A. Hudson, and C. Abell. Fragment-Based Approaches in Drug Discovery and Chemical Biology. *Biochemistry*, 51(25):4990–5003, June 2012. ISSN 0006-2960. doi: 10.1021/bi3005126. URL https://doi.org/10.1021/bi3005126.

O. B. Scott, J. Gu, and A. W. E. Chan. Classification of Protein-Binding Sites Using a Spherical Convolutional Neural Network. *Journal of Chemical Information and Modeling*, 62(22):5383–5396, nov 2022. ISSN 1549-9596. doi: 10.1021/acs.jcim.2c00832. URL https://doi.org/10.1021/acs.jcim.2c00832.

T. Seidel, O. Wieder, A. Garon, and T. Langer. Applications of the Pharmacophore Concept in Natural Product inspired Drug Design. *Molecular informatics*, 39(11): e2000059, nov 2020. ISSN 1868-1751 (Electronic). doi: 10.1002/minf.202000059.

Selcia. Selcia Fragment Library, 2021. URL https://www.cambridgemedchemconsulting.com/DDResources/Hit_iden/frag_coll_profiles_files/SelciaFragmentLibraryPoster_FBLD2010_low%20res.pdf. Accessed: June 2021.

C. Shao, S. Bittrich, S. Wang, and S. K. Burley. Assessing PDB macromolecular crystal structure confidence at the individual amino acid residue level. *Structure*, 30(10):1385–1394.e3, 2022a. ISSN 0969-2126. doi: https://doi.org/10.1016/j.str.2022.08.004. URL https://www.sciencedirect.com/science/article/pii/S0969212622003197.

C. Shao, J. D. Westbrook, C. Lu, C. Bhikadiya, E. Peisach, J. Y. Young, J. M. Duarte, R. Lowe, S. Wang, Y. Rose, Z. Feng, and S. K. Burley. Simplified quality assessment for small-molecule ligands in the Protein Data Bank. *Structure (London, England : 1993)*, 30(2):252–262.e4, feb 2022b. ISSN 1878-4186 (Electronic). doi: 10.1016/j.str.2021.10.003.

Z. Shi, J. Wei, and Y. Liang. A Theoretical Analysis on Feature Learning in Neural Networks: Emergence from Inputs and Advantage over Fixed Features, 2022.

A. Shulman-Peleg, R. Nussinov, and H. J. Wolfson. Recognition of Functional Sites in Protein Structures. *Journal of Molecular Biology*, 339(3):607–633, 2004. ISSN 0022-2836. doi: https://doi.org/10.1016/j.jmb.2004.04.012. URL https://www.sciencedirect.com/science/article/pii/S0022283604004139.

Y. Si and C. Yan. Protein language model embedded geometric graphs power inter-protein contact prediction. *bioRxiv*, 2023. doi: 10.1101/2023.01.07.523121. URL https://www.biorxiv.org/content/early/2023/01/08/2023.01.07.523121.

I. Sillitoe, N. Bordin, N. Dawson, V. P. Waman, P. Ashford, H. M. Scholes, C. S. M. Pang, L. Woodridge, C. Rauer, N. Sen, M. Abbasian, S. Le Cornu, S. D. Lam, K. Berka, I. H. Varekova, R. Svobodova, J. Lees, and C. A. Orengo. CATH: increased structural coverage of functional space. *Nucleic Acids Research*, 49(D1):

D266–D273, Nov. 2020. ISSN 0305-1048. doi: 10.1093/nar/gkaa1079. URL https://doi.org/10.1093/nar/gkaa1079.

S. Simoens and I. Huys. R&D Costs of New Medicines: A Landscape Analysis. *Frontiers in Medicine*, 8, 2021. ISSN 2296-858X. doi: 10.3389/fmed.2021.760762. URL https://www.frontiersin.org/articles/10.3389/fmed.2021.760762.

M. Simonovsky and J. Meyers. DeeplyTough: Learning Structural Comparison of Protein Binding Sites. *Journal of Chemical Information and Modeling*, 60(4):2356–2366, Apr. 2020. ISSN 1549-9596. doi: 10.1021/acs.jcim.9b00554. URL https://doi.org/10.1021/acs.jcim.9b00554.

N. Singh and A. Warshel. A comprehensive examination of the contributions to the binding entropy of protein-ligand complexes. *Proteins*, 78(7):1724–1735, may 2010. ISSN 1097-0134 (Electronic). doi: 10.1002/prot.22689.

R. Skyner and F. von Delft. Fragalysis, 2020. URL https://fragalysis.diamond.ac.uk/. Accessed: August 2020.

R. Skyner, A. Carbery, M. Mokaya, G. Booth, T. H. Olsen, A. Dale-Evans, W. Thomson, T. Gorrie-Stone, and F. von Delft. The Fragalysis API, 2020. URL https://github.com/xchem/fragalysis-api. Accessed: August 2020.

O. Smart, T. Womack, A. Sharff, C. Flensburg, P. Keller, W. Paciorek, C. Vonrhein, and G. Bricogne. RHOFIT, version 1.2.4, 2014.

O. S. Smart, V. Horsky, S. Gore, R. Svobodova Vavrekova, V. Bendova, G. J. Kleywegt, and S. Velankar. Validation of ligands in macromolecular structures determined by X-ray crystallography. *Acta Crystallographica Section D*, 74(3):228–236, mar 2018. doi: 10.1107/S2059798318002541. URL https://doi.org/10.1107/S2059798318002541.

O. S. Smart, T. O. Womack, A. Sharff, C. Flensburg, P. Keller, W. Paciorek, C. Vonrhein, and G. Bricogne. grade, version 1.2.20. https://www.globalphasing.com/, 2021.

R. D. Smith and H. A. Carlson. Identification of Cryptic Binding Sites Using MixMD with Standard and Accelerated Molecular Dynamics. *Journal of Chemical Information and Modeling*, 61(3):1287–1299, Mar. 2021. ISSN 1549-9596. doi: 10.1021/acs.jcim.0c01002. URL https://doi.org/10.1021/acs.jcim.0c01002.

R. D. Smith, J. J. Clark, A. Ahmed, Z. J. Orban, J. B. Dunbar, and H. A. Carlson. Updates to Binding MOAD (Mother of All Databases): Polypharmacology Tools and Their Utility in Drug Repurposing. *Journal of Molecular Biology*, 431(13):2423–2433, 2019. ISSN 0022-2836. doi: https://doi.org/10.1016/j.jmb.2019.05.024. URL https://www.sciencedirect.com/science/article/pii/S0022283619302967.

J. D. St Denis, R. J. Hall, C. W. Murray, T. D. Heightman, and D. C. Rees. Fragment-based drug discovery: opportunities for organic synthesis. *RSC medicinal chemistry*, 12(3):321–329, Dec. 2020. ISSN 2632-8682 (Electronic). doi: 10.1039/d0md00375a.

R. A. Stein and H. S. Mchaourab. SPEACH_AF: Sampling protein ensembles and conformational heterogeneity with Alphafold2. *PLOS Computational Biology*, 18 (8):1–16, 2022. doi: 10.1371/journal.pcbi.1010483. URL https://doi.org/10.1371/journal.pcbi.1010483.

M. Steinegger and J. Söding. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature Biotechnology*, 35(11):1026–1028, 2017. ISSN 1546-1696. doi: 10.1038/nbt.3988. URL https://doi.org/10.1038/nbt.3988.

S. Stilou, P. D. Bamidis, N. Maglaveras, and C. Pappas. Mining association rules from clinical databases: an intelligent diagnostic process in healthcare. *Studies in health technology and informatics*, 84(Pt 2):1399–1403, 2001. ISSN 0926-9630 (Print).

D. Sun, W. Gao, H. Hu, and S. Zhou. Why 90% of clinical drug development fails and how to improve it? *Acta pharmaceutica Sinica. B*, 12(7):3049–3062, jul 2022. ISSN 2211-3835 (Print). doi: 10.1016/j.apsb.2022.02.002.

Z. Sun, Q. Liu, G. Qu, Y. Feng, and M. T. Reetz. Utility of B-Factors in Protein Science: Interpreting Rigidity, Flexibility, and Internal Motion and Engineering Thermostability. *Chemical Reviews*, 119(3):1626–1665, feb 2019. ISSN 0009-2665. doi: 10.1021/acs.chemrev.8b00290. URL https://doi.org/10.1021/acs.chemrev.8b00290.

W. Swegat, J. Schlitter, P. Krüger, and A. Wollmer. MD simulation of protein-ligand interaction: formation and dissociation of an insulin-phenol complex. *Biophysical journal*, 84(3):1493–1506, mar 2003. ISSN 0006-3495 (Print). doi: 10.1016/S0006-3495(03)74962-5.

G. Taherzadeh, Y. Zhou, A. W.-C. Liew, and Y. Yang. Sequence-Based Prediction of Protein–Carbohydrate Binding Sites Using Support Vector Machines. *Journal of Chemical Information and Modeling*, 56(10):2115–2122, Oct. 2016. ISSN 1549-9596. doi: 10.1021/acs.jcim.6b00320. URL https://doi.org/10.1021/acs.jcim.6b00320.

Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. DeepFace: Closing the Gap to Human-Level Performance in Face Verification. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1701–1708, 2014. doi: 10.1109/CVPR.2014.220.

T. Takai, B. D. Jeffy, S. Prabhu, and J. D. Cohen. Retrospective Analysis of Chemical Structure-Based in silico Prediction of Primary Drug Target and Off-Targets.

Computational Toxicology, 26:100273, 2023. ISSN 2468-1113. doi: https://doi.org/10.1016/j.comtox.2023.100273. URL https://www.sciencedirect.com/science/article/pii/S2468111323000142.

R. N. Tazhigulov, J. Schiller, J. Oppenheim, and M. Winston. Molecular Fingerprints for Robust and Efficient ML-Driven Molecular Generation, 2022.

The UniProt Consortium. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Research*, 49(D1):D480–D489, 2020. ISSN 0305-1048. doi: 10.1093/nar/gkaa1100. URL https://doi.org/10.1093/nar/gkaa1100.

N. C. Thompson, K. Greenewald, K. Lee, and G. F. Manso. The Computational Limits of Deep Learning, 2022.

Timetec. Timetec Fragment Library, 2021. URL https://www.timtec.net/fbl-fragment-based-library.html. Accessed: June 2021.

Uniprot. MID2 - Uniprot, 2020. URL https://www.uniprot.org/uniprot/Q9UJV3. Accessed: February 2022.

D. Valenti, J. F. Neves, F.-X. Cantrelle, S. Hristeva, D. Lentini Santo, T. Obšil, X. Hanoulle, L. M. Levy, D. Tzalis, I. Landrieu, and C. Ottmann. Set-up and screening of a fragment library targeting the 14-3-3 protein interface. *MedChemComm*, 10(10):1796–1802, 2019. ISSN 2040-2503. doi: 10.1039/C9MD00215D. URL http://dx.doi.org/10.1039/C9MD00215D.

H. van de Waterbeemd, D. A. Smith, K. Beaumont, and D. K. Walker. Property-Based Design: Optimization of Drug Absorption and Pharmacokinetics. *Journal of Medicinal Chemistry*, 44(9):1313–1333, apr 2001. ISSN 0022-2623. doi: 10.1021/jm000407e. URL https://doi.org/10.1021/jm000407e.

R. L. M. van Montfort, P. Workman, B. Lamoree, and R. E. Hubbard. Current perspectives in fragment-based lead discovery (FBLD). *Essays in Biochemistry*, 61 (5):453–464, 2017a. ISSN 0071-1365. doi: 10.1042/EBC20170028. URL https://doi.org/10.1042/EBC20170028.

R. L. M. van Montfort, P. Workman, A. J. Price, S. Howard, and B. D. Cons. Fragment-based drug discovery and its application to challenging drug targets. *Essays in Biochemistry*, 61(5):475–484, 2017b. ISSN 0071-1365. doi: 10.1042/EBC20170029. URL https://doi.org/10.1042/EBC20170029.

M. Varadi, S. Anyango, M. Deshpande, S. Nair, C. Natassia, G. Yordanova, D. Yuan, O. Stroe, G. Wood, A. Laydon, A. Žídek, T. Green, K. Tunyasuvunakool, S. Petersen, J. Jumper, E. Clancy, R. Green, A. Vora, M. Lutfi, M. Figurnov, A. Cowie, N. Hobbs, P. Kohli, G. Kleywegt, E. Birney, D. Hassabis, and S. Velankar. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Research*, 50(D1):D439–D444, 2021. ISSN 0305-1048. doi: 10.1093/nar/gkab1061. URL https://doi.org/10.1093/nar/gkab1061.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

S. Velliangiri, S. Alagumuthukrishnan, and S. I. Thankumar joseph. A Review of Dimensionality Reduction Techniques for Efficient Computation. *Procedia Computer Science*, 165:104–111, 2019. ISSN 1877-0509. doi: https://doi.org/10.1016/j.procs.2020.01.079. URL https://www.sciencedirect.com/science/article/pii/S1877050920300879.

D. Vemula, P. Jayasurya, V. Sushmitha, Y. N. Kumar, and V. Bhandari. CADD, AI and ML in drug discovery: A comprehensive review. *European Journal of Pharmaceutical Sciences*, 181:106324, 2023. ISSN 0928-0987. doi: https://doi.org/10.1016/j.ejps.2022.106324. URL https://www.sciencedirect.com/science/article/pii/S0928098722002093.

Q. Wang, Y. Ma, K. Zhao, and Y. Tian. A Comprehensive Survey of Loss Functions in Machine Learning. *Annals of Data Science*, 9(2):187–212, 2022. ISSN 2198-5812. doi: 10.1007/s40745-020-00253-5. URL https://doi.org/10.1007/s40745-020-00253-5.

C. X. Weichenberger, E. Pozharski, and B. Rupp. Visualizing ligand molecules in Twilight electron density. *Acta crystallographica. Section F, Structural biology and crystallization communications*, 69(Pt 2):195–200, Feb. 2013. ISSN 1744-3091. doi: 10.1107/S1744309112044387. URL https://pubmed.ncbi.nlm.nih.gov/23385767https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3564628/.

N. Weill and D. Rognan. Alignment-Free Ultra-High-Throughput Comparison of Druggable Protein-Ligand Binding Sites. *Journal of Chemical Information and Modeling*, 50(1):123–135, jan 2010. ISSN 1549-9596. doi: 10.1021/ci900349y. URL https://doi.org/10.1021/ci900349y.

D. Weininger. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences*, 28(1):31–36, Feb. 1988. ISSN 0095-2338. doi: 10.1021/ci00057a005. URL https://pubs.acs.org/doi/abs/10.1021/ci00057a005.

M. S. Weiss, J. Wollenhaupt, G. J. Correy, J. S. Fraser, A. Heine, G. Klebe, T. Krojer, M. Thunissen, and N. M. Pearce. Of problems and opportunities–How to treat and how to not treat crystallographic fragment screening data. *Protein Science*, 31(9):e4391, 2022. doi: https://doi.org/10.1002/pro.4391. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/pro.4391.

Wikipedia. X-ray crystallography, 2020. URL en.wikipedia.org/wiki/X-ray_crystallography. Accessed: June 2020.

S. Wills, R. Sanchez-Garcia, T. Dudgeon, S. D. Roughley, A. Merritt, R. E. Hubbard, J. Davidson, F. von Delft, and C. M. Deane. Fragment Merging Using a

Graph Database Samples Different Catalogue Space than Similarity Search. *Journal of Chemical Information and Modeling*, 63(11):3423–3437, jun 2023. ISSN 1549-9596. doi: 10.1021/acs.jcim.3c00276. URL https://doi.org/10.1021/acs.jcim.3c00276.

A. Wlodawer and J. Vondrasek. INHIBITORS OF HIV-1 PROTEASE: A Major Success of Structure-Assisted Drug Design. *Annual Review of Biophysics and Biomolecular Structure*, 27(1):249–284, 1998. doi: 10.1146/annurev.biophys.27.1.249. URL https://doi.org/10.1146/annurev.biophys.27.1.249.

A. Wlodawer, Z. Dauter, P. J. Porebski, W. Minor, R. Stanfield, M. Jaskolski, E. Pozharski, C. X. Weichenberger, and B. Rupp. Detect, correct, retract: How to manage incorrect structural models. *The FEBS journal*, 285(3):444–466, Feb. 2018. ISSN 1742-4658 (Electronic). doi: 10.1111/febs.14320.

M. Wójcikowski, P. Zielenkiewicz, and P. Siedlecki. Open drug discovery toolkit (ODDT): a new open-source player in the drug discovery field. *Journal of Cheminformatics*, 7(1):26, 2015. ISSN 1758-2946. doi: 10.1186/s13321-015-0078-2. URL https://doi.org/10.1186/s13321-015-0078-2.

M. Wojdyr, R. Keegan, G. Winter, and A. Ashton. *DIMPLE - a pipeline for the rapid generation of difference maps from protein crystals with putatively bound ligands. Acta Crystallographica Section A*, 69(a1):s299, aug 2013. doi: 10.1107/S0108767313097419. URL https://doi.org/10.1107/S0108767313097419.

J. Wollenhaupt, A. Metz, T. Barthel, G. M. A. Lima, A. Heine, U. Mueller, G. Klebe, and M. S. Weiss. F2X-Universal and F2X-Entry: structurally diverse compound libraries for crystallographic fragment screening. *Structure*, 2020. ISSN 0969-2126. doi: https://doi.org/10.1016/j.str.2020.04.019. URL http://www.sciencedirect.com/science/article/pii/S0969212620301404.

D. J. Wood, J. D. Lopez-Fernandez, L. E. Knight, I. Al-Khawaldeh, C. Gai, S. Lin, M. P. Martin, D. C. Miller, C. Cano, J. A. Endicott, I. R. Hardcastle, M. E. M. Noble, and M. J. Waring. FragLites–minimal, halogenated fragments displaying pharmacophore doublets. An efficient approach to druggability assessment and hit generation. *Journal of Medicinal Chemistry*, 62(7):3741–3752, Apr. 2019. ISSN 0022-2623. doi: 10.1021/acs.jmedchem.9b00304. URL https://doi.org/10.1021/acs.jmedchem.9b00304.

O. J. Wouters, M. McKee, and J. Luyten. Estimated Research and Development Investment Needed to Bring a New Medicine to Market, 2009-2018. *JAMA*, 323(9):844–853, Mar. 2020. ISSN 1538-3598 (Electronic). doi: 10.1001/jama.2020.1166.

R. Wu, F. Ding, R. Wang, R. Shen, X. Zhang, S. Luo, C. Su, Z. Wu, Q. Xie, B. Berger, J. Ma, and J. Peng. High-resolution de novo structure prediction from primary sequence. *bioRxiv*, 2022. doi: 10.1101/2022.07.21.500999. URL https://www.biorxiv.org/content/early/2022/07/22/2022.07.21.500999.

D. L. S. XChem. The Diamond-SGC iNext Poised library, 2020a. URL https://www.diamond.ac.uk/Instruments/Mx/Fragment-Screening/Fragment-Libraries0/DSi-Poised-Library.html. Accessed: June 2020.

D. L. S. XChem. XChem Fragment Screening, 2020b. URL https://www.diamond.ac.uk/Instruments/Mx/Fragment-Screening.html. Accessed: June 2020.

K. Yeturu and N. Chandra. PocketMatch: A new algorithm to compare binding sites in protein structures. *BMC Bioinformatics*, 9(1):543, 2008. ISSN 1471-2105. doi: 10.1186/1471-2105-9-543. URL https://doi.org/10.1186/1471-2105-9-543.

X. Ying. An Overview of Overfitting and its Solutions. *Journal of Physics: Conference Series*, 1168(2):22022, feb 2019. doi: 10.1088/1742-6596/1168/2/022022. URL https://dx.doi.org/10.1088/1742-6596/1168/2/022022.

T. Yu, H. Cui, J. C. Li, Y. Luo, G. Jiang, and H. Zhao. Enzyme function prediction using contrastive learning. *Science*, 379(6639):1358–1363, 2023a. doi: 10.1126/science.adf2465. URL https://www.science.org/doi/abs/10.1126/science.adf2465.

W. Yu and A. D. J. MacKerell. Computer-Aided Drug Design Methods. *Methods in molecular biology (Clifton, N.J.)*, 1520:85–106, 2017. ISSN 1940-6029 (Electronic). doi: 10.1007/978-1-4939-6634-9_5.

Y. Yu, S. Lu, Z. Gao, H. Zheng, and G. Ke. Do Deep Learning Models Really Outperform Traditional Approaches in Molecular Docking? *arXiv*, feb 2023b. URL http://arxiv.org/abs/2302.07134.

X. Zeng, F. Wang, Y. Luo, S.-G. Kang, J. Tang, F. C. Lightstone, E. F. Fang, W. Cornell, R. Nussinov, and F. Cheng. Deep generative molecular design reshapes drug discovery. *Cell reports. Medicine*, 3(12):100794, dec 2022a. ISSN 2666-3791 (Electronic). doi: 10.1016/j.xcrm.2022.100794.

X. Zeng, H. Xiang, L. Yu, J. Wang, K. Li, R. Nussinov, and F. Cheng. Accurate prediction of molecular properties and drug targets using a self-supervised image representation learning framework. *Nature Machine Intelligence*, 4(11):1004–1016, 2022b. ISSN 2522-5839. doi: 10.1038/s42256-022-00557-6. URL https://doi.org/10.1038/s42256-022-00557-6.

Zenobia. Zenobia Discovery-ZEN Fragment Library, 2021. URL https://www.zenobiafragments.com/product-page/discovery-zen-library-1. Accessed: June 2021.

T. Zhang, W. Lin, A. Vogelmann, M. Zhang, S. Xie, Y. Qin, and J. Golaz. Improving convection trigger functions in deep convective parameterization schemes using machine learning. *Journal of Advances in Modeling Earth Systems*, 13, 05 2021. doi: 10.1029/2020MS002365.

J. Zhao, Y. Cao, and L. Zhang. Exploring the computational methods for protein-ligand binding site prediction. *Computational and structural biotechnology journal*, 18:417–426, 2020. ISSN 2001-0370 (Print). doi: 10.1016/j.csbj.2020.02.008.