Fragment Hotspot Mapping to Drive the Rational Elaboration of Fragment Screening Hits

Mihaela Smilova

Linacre College

University of Oxford

A thesis presented for the degree of

Doctor of Philosophy

Hilary 2022

Acknowledgements

Doing a PhD is not an easy undertaking in the best of times, and the past two years have not been the best of times on a global scale. I have been incredibly fortunate to have the support of a large network of colleagues, friends, and family, without whom this thesis would not have been possible.

I want to thank my supervisors, who kept me and the project on track over the past few years. Brian, for being the best mentor I could have imagined and always making time for me in his busy schedule. Anthony, for showing me how to think on my feet and what working hard looks like. Frank, for always asking the difficult questions. Jason, for introducing me to the subtleties of compchem and for always having excellent suggestions when they were most needed.

I also want to thank Chris and Pete for their amazing support and insightful discussions on fragment hotspot maps and beyond. Charlotte and Garrett, for the warm welcome and the opportunity to learn from their vast knowledge of protein informatics. The SGC code review and the PX lunch gang, for always providing stimulating discussion and rescuing many a bad day.

The computational work presented in this thesis is critically dependent on the quality of the experimental data used for input and validation. This project would not have been possible without the excellent work of Eleanor Williams, James Bennett, Joseph Newman, Lizbé Koekemoer, and Gian Fillippo Ruda.

The past two years have also meant getting to know one's housemates particularly well, and highlighting the importance of the support we give each other during times of uncertainty. Bridge Street crew (and honorary member Zuzana) - thank you for making this time so much more bearable!

Finally, I want to say thank you to my family for giving me everything I needed to set on this path, and to Kalo, for helping me every step of the way.

Pandemic Impact

I have received a funding extension for 1 term from the CSEF fund based on disruptions to the project related to the COVID-19 pandemic.

While the work I have undertaken for this thesis is computational, Chapter 4 involves validating the presented methods and workflows prospectively on live projects at the Centre for Medicines Discovery (CMD) and collaborating departments. The COVID-19 pandemic has meant that the majority of these projects were put on hold, and so I was unable to get the expected amount of data for this chapter during this time. The turnaround times for compound design and testing on these projects are between 6 months to a year under normal circumstances. Before the first lockdown in March 2020, I was involved in designing compounds for two ongoing projects at the CMD. From March to August 2020, all non-coronavirus projects were halted, and from the second half of August 2020, the laboratories have started to operate again, albeit at a reduced capacity. Work on the projects I was involved in resumed in January 2021, and since I have worked on suggesting compounds for these and an additional COVID-related project (NSP13). These projects are highly collaborative and interdisciplinary, with crystallographers, biologists, medicinal chemists and computational scientists working together. Because of the scale and number of people involved, I did not have much influence on the timelines. The compounds shown in Chapter 4 arrived in the beginning of July, but results were only available in the fall/winter of 2021, as the department moved buildings, resulting in a major disruption to experimental work.

Statement on COVID-19 related work

The methods and workflows presented in this thesis are intended for use in the early stages of drug discovery. Specifically, they aim to make the process of developing fragment hits into an optimised lead compound more streamlined, robust, and objective. Medicines discovery is a field that has become even more relevant in the context of the COVID-19 pandemic. While this situation has meant that many existing medicinal chemistry projects were disrupted, work on designing chemical matter against proteins of therapeutic significance to SARS-CoV-2 infection has been undertaken at a fantastic pace.

One of the prospective case studies in Chapter 4 involved designing follow-up compounds based on a crystallographic fragment screen against the NSP13 protein target. This is a viral helicase from the SARS-CoV-2 pathogen, and is essential for the virus' replication. Therefore, molecules that inhibit this protein are of interest both as chemical probes to investigate its function, as well as potential starting points for drug discovery efforts. The work on this target presented in Chapter 4 resulted in compounds with detectable activity (hundreds of micromolar) against NSP13, which was an improvement over the starting fragment hits. Crystallographic data to confirm the compounds' expected binding mode is not yet available, but they could form the basis of further compounds designed against this target.

Abstract

Fragment-based drug discovery (FBDD) has established itself as a powerful tool for developing probe and drug candidates by rationally elaborating small chemical fragment hits into larger, optimised lead compounds. The use of X-ray crystallography as a medium-throughput screening tool for FBDD results in a wealth of structural data on low molecular weight molecules in complex with a protein target. Interpreting this data and distilling it into prioritised suggestions for the elaboration of fragment hits into leads with increased potency and selectivity for the target protein is currently a significant challenge to using the technique. Thus, computational methods and pipelines designed to streamline and automate the process, providing follow-up hypotheses in an objective and high-throughput way, are in high demand.

Fragment hotspot mapping is a computational method that highlights specific interactions within a protein's binding site that drive the binding of small molecule fragments. As crystallographic FBDD campaigns result in an ensemble of structures of the same protein, a method to combine fragment hotspot maps information for these structures into an "ensemble map" for the protein target was developed. A workflow for comparing ensemble maps between a target and a related off-target protein was implemented and extended to allow comparisons across a protein family. This workflow was applied to examples from the well-researched human bromodomain and kinase families, and was able to identify selectivity-determining regions that have been exploited in past drug discovery campaigns.

Dynamic undocking, a steered molecular dynamics method for estimating the structural stability of protein-ligand complexes, was then investigated as a way of characterising specific binding site interactions. To facilitate integration into computational workflows,

an open-source implementation of the method was benchmarked and shown to perform comparably.

A workflow combining the extended fragment hotspot maps method, dynamic undocking, docking and a chemistry recommendation engine was developed and used to suggest follow-up compounds in three ongoing medicinal chemistry projects. The compounds showed detectable binding affinity, a significant improvement from the starting fragment hits, demonstrating the workflow's utility in the initial round of compound elaboration.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

ADMET ..... Absorption, Distribution, Metabolism, Excretion, Toxicity

API ..... Application Programming Interface

CADD ..... Computer-Aided Drug Design

CMD ..... Centre for Medicines Discovery

CSD ..... Cambridge Structural Database

CPU ..... Central Processing Unit

DMSO ..... Dimethyl sulfoxide

DUck ..... Dynamic Undocking

FBDD ..... Fragment-Based Drug Discovery

GPU ..... Graphics Processing Unit

GUI ..... Graphical User Interface

HTS ..... High Throughput Screening

MCSS ..... Multiple Copy Simultaneous Search

MD ..... Molecular Dynamics

MSCS ..... Multiple Solvent Crystal Structures

NMR ..... Nuclear Magnetic Resonance

PDB ..... Protein Data Bank

SAR ..... Structure-Activity Relationship

SBDD ..... Structure-Based Drug Discovery

SGC . . . . . Structural Genomics Consortium

SMARTS . . . . . Smiles Arbitrary Target Specification

SMD . . . . . Steered Molecular Dynamics

SMILES . . . . . Simplified Molecular Line Input Entry System

TR-FRET . . . . . Time Resolved Fluorescence Resonance Energy Transfer

1 | Introduction

The past decade has seen an explosion in the availability of genomic and structural data for a great number of biomolecular disease targets [1, 2]. Rational drug discovery aims to use this knowledge to design chemical and biological agents that modulate the activity of these targets [1]. Despite recent technological advances, the development of these agents is exceedingly expensive, is not routine, and carries the risk of the compound failing after many years of development. This means that there is great need for methods that can streamline and automate the development process, both for drugs, as well as for small molecule probes, the highly selective chemical agents used to investigate the underlying biology of disease [2].

1.1 The discovery process for small molecule drugs

Historically, many small molecule drugs have been discovered through serendipity, with their efficacy and performance in *in vivo* trials being used to guide development. This process is also known as phenotypic screening, or classical or forward pharmacology. As knowledge about the molecular processes underlying disease processes accumulated, the paradigm has shifted towards designing agents that can modulate these mechanisms in rational and deliberate ways. As a result, increasing numbers of drug design projects today start with a hypothesis that influencing the activity of a molecular target or biological pathway will exert a positive effect on a disease state with an unmet therapeutic need [3]. A biomolecular, usually protein, target is chosen as the focus of the drug discovery effort. Validating the importance of this target to the mechanism of disease is crucial, as a great number of drugs fail after many years of development due to a lack of efficacy as

a result of failures in target validation [2]. As many successful drugs have started from phenotypic screens, this method is still widely used. However, even in these cases, target identification and validation is important in establishing the mechanism of action of the hit compound, and any potential metabolic and toxicity risks. Once a suitable target has been identified (for target-driven campaigns), or a hit is found that has an effect on the disease state (phenotypic screening), the campaign progresses to the lead discovery phase (shown in blue in Figure 1.1). Here, an intensive search to find a suitable drug-like candidate is performed. If successful modulation of the disease target is achieved, the candidate then progresses into clinical development, where its safety and efficacy are established.

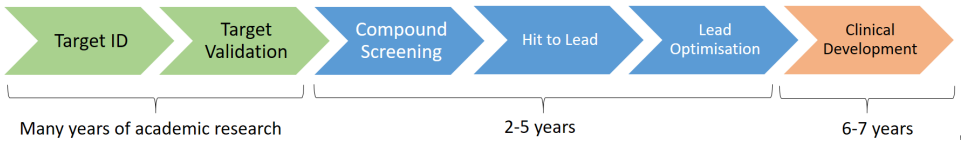


Figure 1.1: Schematic overview of the drug discovery process for small molecule drugs. Basic research on target identification and validation is shown in green. The lead discovery phase, which is focus of this thesis - in blue, and clinical development - in orange.

1.2 Small molecule drugs and chemical probes

Correct target identification is central to the drug discovery process. However, drugs themselves do not require exquisite selectivity for a particular target.

In certain cases, polypharmacology, or the ability to modulate multiple targets, may even be desirable to prevent the development of resistant disease phenotypes [4]. Highly selective tools are needed to probe the underlying biology of disease, however. In recent years, knock-out techniques based on CRISPR-Cas9 gene editing have been used to selectively ablate specific proteins. However, this approach does not work in cases where the gene knockout mutants are not viable. The

effects of ablating genes are also not titrateable: it is difficult to modulate the effect of a gene product to a fraction of its normal activity through these methods. Chemical probes, small molecules with high (less than 100 nM K_i, K_d, or IC_{50}) *in vitro* potency and selectivity for a particular target, are better suited tools for this task [5, 6]. Probes are usually selective for different subsets of proteins within a protein family, while the selectivity of drugs needs to be understood across both the target family and the proteome [7]. Probes also do not need to be safe for use in patients or exert a therapeutic effect, as their purpose is to be used in the pre-clinical discovery phase, and specifically for target validation. Still, the hit to lead development phase for probes and drugs has significant overlap, and chemical probes in the public domain can be used as starting points for drugs [6]. The Target 2035 initiative aims to make such investigational tools available for most human proteins in the next 15 years [8, 6]. This means that tools that speed up and streamline the hit to lead discovery phase for both drugs and chemical probes are greatly needed [8].

1.3 Navigating chemical space

Chemical space refers to the space of all possible chemical compounds [9]. Enumerating chemical space is the process of exhaustively listing the individual molecules that comprise the space. It is also possible to navigate through chemical space, starting at a point (compound) and arriving at a different point (compound) through a series of operations (chemical reactions; it must be noted, however, that it may be the case that a very structurally similar compound cannot be reached with a sequence of achievable chemical reactions). Chemical space can be broken up into subsets: for example, "drug-like" chemical space describes the set of all small molecules with drug-like properties. Drug discovery can then be thought

of as "mining" this space for compounds that are active against the target protein, as well as display favourable ADMET (absorption, distribution, metabolism, excretion, toxicity) properties. Several descriptions of what constitutes "drug-likeness" have been developed over the years, with Lipinski's "Rule of 5" [10] being the most famous. The chemical space corresponding to compounds that fit drug-likeness criteria, however, is prohibitively large: 10^{63} molecules is one of the most often quoted estimates [11]. This was calculated by considering only compounds with up to 30 heavy atoms and made up of C, N, O, and S. However, this estimate may be an overestimation, as the lack of restrictive filters may also include compounds with impossible clashes and strains [12]. Work by Reymond [9] focused on comprehensively enumerating subsets of realistic chemical space consisting of molecules with up to a certain number of heavy atoms. GDB-17 (Generated DataBase) is the largest such set, consisting of 166 billion compounds with up to 17 heavy atoms, including C, N, O, S, and halogens. Analyses of the GDB databases have shown that for each additional heavy atom, the number of potential molecules increases 8-fold [13]. However, the size of these molecules (up to 17 heavy atoms), does not comprise the full set of drug-like molecules. Polishchuk *et al* used the GDB-17 set to extrapolate the number of realistic compounds with up to 36 heavy atoms (roughly corresponding to a molecular weight of 500 Da) that follow Lipinski's rule to be 10^{33} [12]. This is likely a more realistic estimate, as the GDB rules include filters on reactivity, clashes and strains [9]. Even though this is far less than Bohacek's estimate, it is still a prohibitively large number of compounds. The vast size of drug-like chemical space has necessitated the development of methods and technologies for its systematic enumeration and navigation.

1.3.1 High throughput screening and combinatorial chemistry

Combinatorial chemistry (combichem) is a technique that allows the synthesis of large compound libraries, consisting of millions of molecules. In the 1990s, technological advances in robotics meant that these libraries could be experimentally assayed, in a process called High Throughput Screening (HTS). Pharmaceutical companies curated and screened compound libraries consisting of millions of compounds. HTS does not rely on existing structural knowledge of the disease target. Instead, the goal is to find a "hit" within the screening library, which shows activity in the assay and exhibits a drug-like profile. However, HTS hits are rarely fully drug-like, and usually require further optimisation. This can be difficult, as these compounds generally have a molecular weight of 300-400 Da (the binding of larger compounds is more easily detected in biophysical assays), leaving little room for modifications to be made. This can in turn lead to molecules that suffer from "molecular obesity" [14]. This refers to the practice of increasing the potency of a lead by increasing its lipophilic character, which introduces downstream problems with safety and ADMET [14]. In addition, HTS is prone to false positives, such as compounds that interfere with the assay readout, or nonspecific binders that inhibit many targets in diverse protein families [15]. Finally, the size of HTS screening libraries, while vast, is not sufficient to provide an extensive coverage of drug-like space. Analyses of the distribution of chemical descriptors also showed that the compounds output by early combichem approaches were also not representative of this space [16].

1.3.2 Readily synthesisable chemical spaces

A drawback of the compounds generated by computational enumeration methods is that they may not be synthetically accessible. To address this problem, chem-

ical spaces may be enumerated using a procedure known as forward synthetic analysis. This involves generating virtual libraries by enumerating combinations of chemical building blocks, or synthons, with known reactivity. An example of this approach is Enamine's REAL (REadily AccessibLe) database, which currently numbers over a billion enumerated compounds, which can be synthesised quickly and at low cost [17]. The distributions of chemical descriptors in this space were also shown to be lead-like. Such databases are valuable resources for virtual screening campaigns; once a hit has been found, they also allow for the rapid exploration of adjacent chemical space, as closely-related compounds can be supplied quickly and cheaply and used to establish structure activity relationships experimentally. This, in turn, greatly helps accelerate the design-make-test cycles that drive drug discovery efforts.

1.4 Structure-based drug discovery

In addition to having favourable physicochemical properties, a successful lead candidate modulates the function of the disease protein target. A protein's function is determined by its structure, binding partners, and post-translational modifications. Consequently, information about the 3-dimensional structure of a protein can greatly aid the effort to discover agents that modulate its function. In 2021, the Protein Data Bank (PDB), the largest online hub and resource for structural information of biomolecules celebrated its 50^{th} anniversary. Figure 1.2 shows the growth of this resource over time, leading up to over 180000 entries in 2021. They can be grouped into about 100000 clusters with unique sequences, indicating that ensembles of structures have been solved for a large number of targets. If the structure for a disease target is available, it can be used to guide the development of lead candidates in a process referred to as structure-based drug discovery

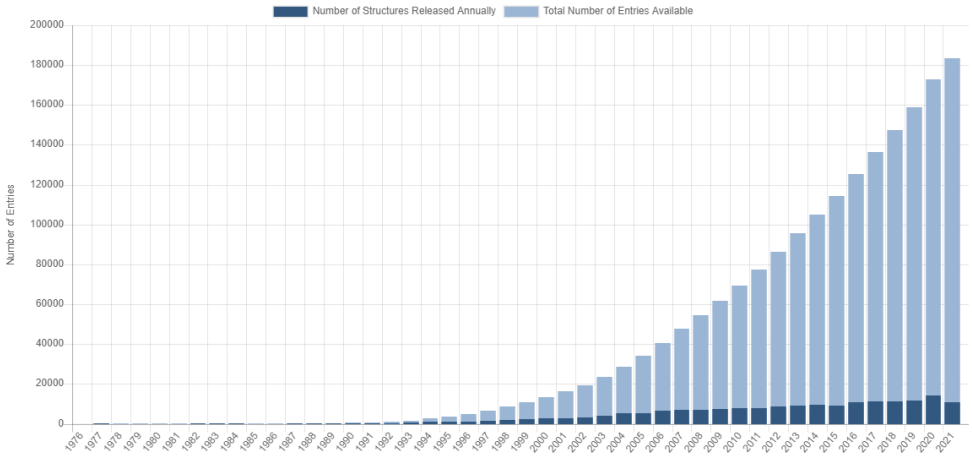


Figure 1.2: Growth of PDB-released structures by year in 2021. Image taken from https://www.rcsb.org/stats/growth/growth-released-structures

.

(SBDD). The starting point for such campaigns can be a apo or natural ligand-bound structure of the receptor, in which case various methods can be used to convert this structural information into suggestions for chemical matter. Even in cases where the starting hit comes from a technique that does not produce 3D structures, such as high-throughput screening, a structure of the protein in complex with the hit may be pursued in order to guide hit to lead optimisation [3, 2].

1.4.1 Fragment-based drug discovery

Fragment based drug discovery (FBDD) has been established as a powerful tool to develop lead compounds, the starting point for probes or drugs [18]. Fragments are small (up to 300 Da), typically weakly binding molecules, with binding affinity in the millimolar to micromolar range [19, 20]. They have been shown to have higher hit rates and sample chemical space more efficiently compared to the larger molecules traditionally used in high throughput screening [21, 13,

22]. By limiting the size of the molecules (and therefore the number of heavy atoms), the number of ways in which these atoms can be arranged into synthesisable molecules also decreases, and consequently the available chemical space is smaller. As an example, the GDB-13 database comprises 100 million compounds with 12 heavy atoms. A library of only 1000 compounds would then cover 0.001 % of the space: a vast improvement on the coverage provided by the millions of compounds in HTS libraries [13]. In a 2014 publication discussing the concept of fragment space [13], Hall *et al* pose the question of navigating fragment chemical space in a more application-focussed way: how big does a library of compounds with a particular size need to be, in order to ensure a sufficient number of hits against any protein target? Featurised models of binding sites and ligands with increasing complexity have been used to explore this idea [21]: the probability of finding a match was found to steadily decrease with increasing complexity. However, low-complexity ligands may form several matches with the binding site. It can be argued that these would be a less useful starting point for drug design, as the binding mode is ambiguous. The likelihood of a "useful" match (one which has a unique binding mode) was found to peak at ligands with complexity 3 (3 potential interactions with the receptor) for binding sites with 12 potential interactions. While the authors note that the model makes significant simplifications and these numbers are not rules or guidelines for library design, they nevertheless illustrate the concept.

However, smaller compounds have weaker binding affinities, often beyond the sensitivity of the biophysical methods used in high throughput screens. The goal of fragment-based drug design (FBDD) is to elaborate these initial, weakly binding hits in a stepwise and rationally-guided manner. This would ultimately lead to a potent and selective lead molecule that contains the most valuable features of the initial hits, retaining the binding mode of the starting fragment. Over the

last decade, fragment-based drug discovery efforts have produced four clinically approved drugs [23, 24, 25, 26], and at least 30 compounds are in various stages of clinical trials [27].

1.4.1.1 Fragment libraries

A key part of the success of these methods lies in the fragment libraries themselves. In 2003, Congreve and colleagues noted that fragment hits seemed to conform to a Rule of Three (molecular weight under $\leq$ 300 Da, number of hydrogen bond donors $\leq$3, number of hydrogen bond acceptors $\leq$3 and cLogP $\leq$3) [19]. But the idea of going even smaller is attractive, as this creates smaller libraries that sample chemical space more efficiently [21]. This has led to the introduction of "very small fragments" (VSFs), with usually less than 11 heavy atoms [28]. Recently, Astex's library of MiniFrags [29] and the FragLites library developed by Wood and colleagues [28] used VSFs to experimentally probe binding sites with crystallography. These fragments showed high hit rates and raised the possibility of being used as a "pre-screen", in order to determine what larger fragments to screen next.

In addition to covering as large a portion of drug-like space as possible with the minimum number of compounds, fragments need to be able to be quickly and easily elaborated into larger and more potent leads [30, 31]. This idea has lead to the design of 'poised' fragment libraries, such as the DSI-Poised library currently used by the XChem facility at Diamond Light Source [31, 32]. These fragments are designed to be easily deconstructed into at least two "synthons", containing at least one functional group that can be used to extend the fragment using a robust and well-characterised reaction [31].

Libraries have also been devised that carry modifications for the specific biophys-

ical technique used in the screen. For example, the FragLites library uses halogenated fragments, which can be easily detected in crystallographic density, allowing X-ray crystallography to be more readily used in the experimental mapping and tractability assessment of protein binding sites [28]. Fluorine and phosphorus-containing fragments have extensively been used in fragment screening by NMR applications [30]. Resnick and colleagues at the Weiszmann Institute and the Structural Genomics Consortium have developed libraries of electrophile fragments for covalent screening [33], which have been used to probe the binding site of the main SARS-CoV-2 protease [34]. The binding of covalent fragments can be detected by mass spectrometry, allowing for fragment screening on a cellular or even proteomic level [35]. The "fully functionalised fragments" described by Parker and co-workers contained a click-chemistry tag, which can be activated by UV light [36], allowing the screening of non-covalent fragments in cells.

1.4.1.2 Experimental hit detection

However, as fragments bind so weakly, binding events are difficult to observe by most biophysical techniques. In addition, biophysical screening assays can be prone to interference by pan-assay interference compounds (PAINS) and colloidal aggregators (molecules that form an aggregate that interacts non-specifically with the target) [15]. Screening by X-ray crystallography has the advantage of being able to detect a large range of ligand binding affinities (sub-nanomolar to millimolar), and is only limited by the availability of crystals for the target and the solubility of the compounds. In addition, this technique also shows the binding mode of the fragments, which is important for their further elaboration [22].

1.4.2 Crystallographic fragment screening

Historically, one of the biggest drawbacks to the method was the complexity, cost, and time-intensiveness of the experiments [22]. In the last few years, a fragment screening pipeline has been implemented at the XChem facility at Diamond Light Source and the Structural Genomics Consortium, which has reduced the time from soaking the crystals to finding hits in the solved structures from three months to around a week [37, 32].

Screening by X-ray crystallography begins by obtaining crystals of the target protein, referred to as the crystal system. This system needs to be robust and reproducible, as well as diffract to a high resolution (the XChem facility currently recommends at least 2.8 Å [32]), in order to allow the screening of the hundreds of compounds in the fragment screening library. There are generally two ways in which a crystal of the protein-ligand complex can be obtained: co-crystallisation, in which the target compound is added to the protein before crystals are grown, and soaking, in which pre-formed apo protein crystals are immersed in a solution containing a high concentration of the screening compound. The solution diffuses through solvent channels in the crystal, allowing ligand molecules to localise in environments that favour their binding. Protein crystallisation can be unpredictable even for well-characterised systems, and introducing different screening compounds to the crystallisation conditions may interfere with the formation of crystals. It is also more difficult to perform at scale. Soaking avoids the co-crystallisation problem by starting with a pre-formed crystal. However, the crystal system must be able to withstand high concentrations of solvent (usually DMSO or ethylene glycol) and screening compounds. To obtain the maximum amount of information from a screen, the highest concentration of screening compound that does not interfere with crystal diffraction is used (500 mM for the XChem

fragment libraries). Pre-screen experiments are used to determine these optimal soaking conditions for each crystal system. The requirements for high resolution and robustness of the crystal system means that small, soluble proteins (many enzymes, for example) that crystallise easily are highly suitable for the platform. Challenging targets, such as membrane proteins, proteins with extensive post-translational modifications, *etc.*, are less well suited to these experiments. In addition, the crystal environment may interfere with the binding of the screen compounds. For example, access to the protein's active site may be blocked by a crystal symmetry mate. Loop rearrangements and other local conformational changes have been observed in soaking experiments. However, in cases where ligand binding introduces significant conformational change, this may be incompatible with the crystal structure. In such cases, further co-crystallisation experiments may be needed. False positive hits may also occur in crystal contacts, however those tend to be easily identified by the crystallographer.

After the crystals have been soaked, they are harvested and diffraction data is collected. Automated pipelines carry out the processing of hundreds of crystallographic datasets. This is followed by the process of detecting hits in the crystal density. However, as the fragment hits tend to have only partial occupancy within the crystal, the signal associated with the binding event tends to be weak, and the electron density - difficult to interpret through standard methods [38]. Dedicated analyses geared towards extracting information on these low-occupancy states, such as the PanDDA [38] method employed at XChem, have been transformative in the field, allowing even very weak binders to be identified.

Ultimately, the crystallographic screen results in an ensemble of crystal structures of the target protein in complex with modelled fragment hits, as shown in Figure 1.3. Currently, the bottleneck comes at the step when the fragment data has been collected and decisions have to be made on which fragments to elaborate further,

and into which areas of the binding site of the target protein. The first step in this process involves summarising and presenting the fragment screening data. The simple superposition of the fragment-bound structures results in a visualisation that is very difficult to interpret even to the trained eye, especially when looking for subtle changes present in a minority of the structures [39]. To deal with this problem, methods such as WONKA, which highlights interesting and unusual features in protein-ligand complexes, and OOMMPPAA, which summarises structural and activity data in the context of the target protein, have been developed [40, 41]. However, these methods do not generate suggestions for future experiments in an automated way. In order to further streamline the process of going from hits to leads, computational chemistry methods that can take this information and automatically generate suggestions for future experiments are needed.

1.4.2.1 The Fragalysis platform

To enable the analysis and progression of fragment screening hits, the XChem facility uses a web-based, open source platform called Fragalysis (`https://fragalysis.diamond.ac.uk/viewer/react/landing`). This tool provides functionality to summarise, present, and share data from fragment screening experiments. In addition, it makes use of an open-source implementation of the Astex fragment network [42] to automatically provide suggestions for purchasable follow-up compounds (the fragment network compound suggestion algorithm is described in more detail in section 4.2.1.5). Fragalysis uses the NGL viewer [43] to present data from the fragment screening experiments. This includes views of the structural ensembles, individual protein-fragment complexes, and non-covalent interactions between the protein and fragment hit. Compound elaboration vectors are mapped on the 3D bound fragment structures. Follow-up selections can be highlighted and downloaded as .csv files. Views and sessions

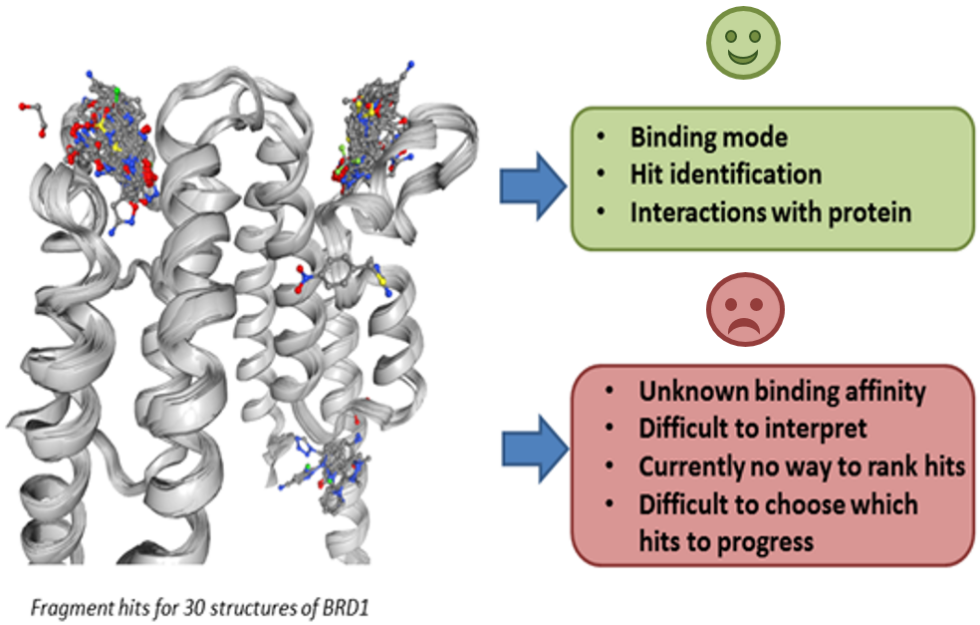


Figure 1.3: The current bottleneck in fragment screening by crystallography. The method has the advantage of providing the binding mode, identity, and interaction profile of the hits, but the results are not easily progressed into suggestions for further experiments.

.

can be saved and shared between users. This functionality made the Fragalysis platform key for the rapid dissemination of structural data and initial chemical matter for a number of protein targets of significance to developing treatments for COVID-19 [34, 44]. The platform is envisioned to be extended to include further algorithms and tools that enable the fast and cheap progression of fragment hits, making them available to users without a background in computational chemistry.

1.5 Fragment hit to lead progression

A typical Design-Make-Test cycle following a successful crystallographic fragment screening campaign is shown in Figure 1.4. After the protein-fragment complexes have been solved, computational follow-up algorithms can be used to distill

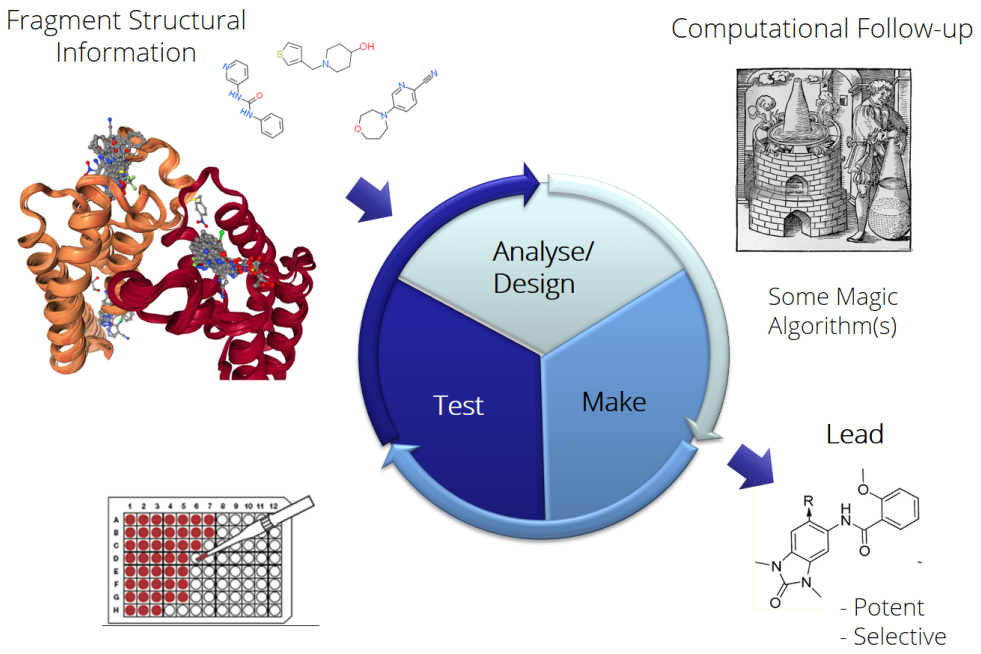


Figure 1.4: A typical Design-Make-Test cycle in a crystallographic fragment screening campaign. Computational methods are used to guide the iterative design and synthesis/purchasing of follow-up compounds, culminating in a lead compound with improved potency and selectivity against the protein target.

this information into testable hypotheses for follow-up design [45]. The size of the search space [12] and the time and resource limits on what is already a very expensive process mean that only a fraction of the possible follow-up compounds can be tested experimentally. *In silico* methods are generally faster and cheaper to employ, and so play a crucial role in prioritising which potential follow-up compounds are then synthesised (or purchased) and tested. Ideally, both structural and binding data for the tested compounds can be obtained and used to drive the next round of suggestions. This process is repeated over multiple iterations, with the goal of obtaining a lead that is both potent and selective for the target protein.

1.5.1 The process of computational follow-up

The process of suggesting follow-up compounds after a crystallographic fragment screen can be split up into three main stages: triaging of the initial hits, enumeration of potential follow-up compounds, and prioritisation of the suggestions.

- **Triaging of the initial fragment hits**

  This step involves choosing which crystallographic fragment screen hits to elaborate and into what regions of the binding site. In the case where experimental binding affinity data is available for the initial fragment hits, a metric such as ligand efficiency [46] can be used to prioritise the initial hits. However, this is often not the case in crystallographic fragment screens, especially when very low-molecular weight probing fragments are used. Even in cases where affinity data is available, complementary metrics such as the robustness of the binding mode to withstand small perturbations [47, 45], or the complementarity (both in shape and electrostatics) between the hit and the protein's binding site [48, 49] may be used to aid the decision making process.

- **Enumeration of follow-up compounds**

  Once a subset of the initial hits have been selected as starting points for optimisation, the next step involves using methods that explore the adjacent chemical space and output potential virtual molecules for further synthesis. Compound suggestions may be generated *de novo*, which has been argued to result in a more complete and unbiased sampling [50], or by searching against databases of purchasable or synthesisable compounds [51, 52, 42]. The latter method may restrict the space of compounds to those that can be readily synthesised by a fairly limited set of commonly-used reactions in

medicinal chemistry [53, 54]. However, the size and variety of commercially available compound libraries is already in the tens of millions and rapidly growing [55, 17]. As of May 2022, the number of synthesisable on-demand compounds in Enamine's REAL Space is 29 billion, OTAVA's CHEMriya numbers around 12 billion, Wuxi LabNetwork's GalaXi - 12 billion, and ChemSpace's Freedom Space - around 150 million [56]. In the case of *de novo* methods, there may be no guarantee that the compounds can actually be synthesised in a suitable time frame [57, 45].

- **Prioritisation of potential follow-up compounds**

The final stage involves selecting which potential follow-up designs to prioritise for synthesis/ purchasing and experimental testing. In crystallographic screens, the structure of the protein and binding mode of the parent fragment are known, so methods that look for structural complementarity between the receptor and the virtual follow-up compound are widely used. The core assumption in FBDD is that the binding mode of the parent fragment remains unchanged upon optimisation. While fragment deconstruction studies have shown that this is not always the case, an analysis of 359 drug-like molecules and their fragment substructures across 51 protein targets found that the fragment binding modes are generally conserved, with the drug-like compounds sharing the polar interactions of the fragment substructures [58]. In addition, making small, conservative chemical changes to the starting fragment (ECFP2 similarity > 0.7) was not found to interfere with binding mode conservation [58]. Protein binding site plasticity was found to be a common feature in cases where the larger compounds differed in binding mode, highlighting the importance of considering this property during fragment optimisation. Another predictor of changes in binding mode was found to be the size of the starting fragment, with frag-

ments smaller than 150 Da being less likely to maintain their position upon elaboration [59, 58].

These points must be taken into consideration when assessing the complementarity between the receptor and the virtual follow-up compound. As the fragment is expected to retain its binding mode, constrained docking approaches are often used to generate 3D poses of potential follow-up compounds [60]. It is then important to be able to measure the overlap between the binding mode of the parent fragment and the docked follow-up ligand. This is most commonly done by calculating the root-mean square deviation (RMSD) between the coordinates of the shared substructure between the initial hit and the docked follow-up molecule, with RMSD values < 2 Å usually considered to have retained the binding mode [61]. However, in cases where a common substructure cannot be easily determined (for example bioisosteres, small changes in the shared chemical scaffold, or pseudosymmetries) shape and colour-based metrics have been shown to perform better than RMSD for measuring binding mode conservation [59, 61]. In 2017, Malhotra and Karanicolas developed the Combined Overlap Score (COS) as a way to measure binding mode similarity between a fragment and a larger structurally similar compound [59]. On open-source (the original COS metric was based on the ROCS [62] algorithm, which is proprietary) implementation of this score, SuCOS [61], was found to perform better than RMSD for evaluating fragment elaboration and docking poses [61]. This metric was used as part of the computational workflows described in Chapter 4 of this thesis. It is described in more detail in Section 4.2.1.7.

In terms of the docking protocols used, rigid receptor docking is common, as it is the least computationally expensive and can be used to assess the

thousands of potential compounds generated in the enumeration phase. To take into account protein plasticity, ensemble docking can be set up, using a selection of conformationally diverse receptors. These can all be derived from the crystallographic fragment screening experiment, or supplemented with structural information from the PDB. Including such external structures can be beneficial (up to a point, in which including large numbers of structures was found to introduce false positives [63, 64]), as information from structures in different crystal groups and conformations not found in a soaking experiment can be included. In such cases, care needs to be taken when compiling the ensemble - this will be further discussed in Chapter 2. Some docking programs also allow side chain flexibility, and limited backbone plasticity to be modelled, at a higher computational cost. The most detailed information on binding site flexibility, especially with regards to concerted pocket changes and the opening of cryptic pockets, is provided by molecular dynamics based methods. These are even more expensive to run (taking days, compared to minutes for most docking programs), but improvements in both the software [65] and hardware available have made such approaches feasible in the context of FBDD pipelines.

1.5.2 Overview of computational fragment-to-lead follow up strategies

A large number of fragment to lead progression campaigns have incorporated the use of *in silico* methods, recently summarised in two excellent reviews by Rachman and colleagues [60] and de Souza Neto *et al.* [66]. Cases in which computational methods have been used to support a particular step of the followup process are far more common than cases in which the methods are used to guide the strategy in an automated or semi-automated way [60, 45]. However, the promise of

such workflows for streamlining and rationalising lead discovery is significant, leading to great interest in their development and several promising developments [60, 45, 66].

1.5.2.1 Computational methods in fragment hit ranking

Choosing which hits to progress is not always a bottleneck in a crystallographic screen: there may be cases when only a single hit is available, as will be discussed in Chapter 4. The medium throughput nature of crystallographic screening can mean that even when multiple hits are present, human judgement can be used to select promising hits for follow-up. This has the advantage of incorporating expert knowledge, but also carries the risk of introducing biases into the decision-making process. For this reason, objective computational methods have been used to complement expert opinion. These can include hotspot and binding site mapping methods (discussed further in the chapter) to choose between sites where fragments are present. For instance, Thomas and colleagues [67] used the fragment hotspot maps method [48] to prioritise fragments in the design of inhibitors against a tRNA-methyltransferase target in *Mycobacterium abcessus*. More recently, the method was used by the same group for structure-guided inhibitor development of the DNA-dependent protein kinase catalytic subunit [68]. Rachman and colleagues used MDMix [69] in selecting target fragment hits making favourable interactions with NUDT21 [45]. Increasingly, molecular dynamics methods have also been used to aid fragment ranking. An example is provided by Bissaro and co-workers, who used their Supervised Molecular Dynamics (SuMD method) to inspect the mechanism of fragment binding hits to the SARS-CoV2 main protease (Mpro) [70]. Steered molecular dynamics methods [47] have also been used to rank fragments by the stability of their binding mode. This set of methods will be discussed in detail in section 1.7.

1.5.2.2 Computational methods for growing, linking, and merging of fragments

Fragment elaboration strategies can be broadly split up into three categories (shown in Figure 1.5): growing, in which a starting fragment is elaborated through the addition of R-groups; linking, in which two fragments located near each other in the binding site are connected via a linker group, and merging, in which two fragments with overlapping moieties are combined into a single molecule. These strategies present distinct computational problems, and so a variety of methods have been developed to address them.

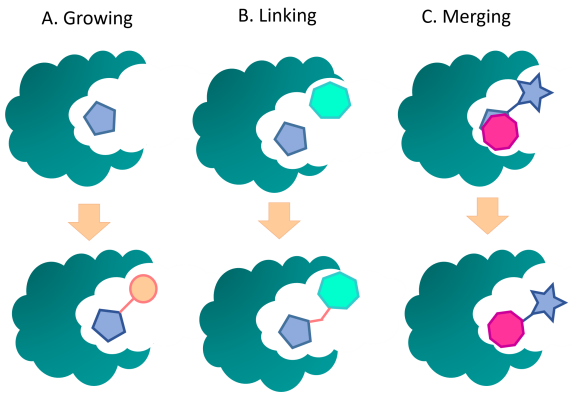


Figure 1.5: Overview of the strategies used in fragment elaboration. A. Growing: the starting fragment (blue pentagon) is grown through the addition of an appropriate R group (orange circle). B. Fragment linking, where two starting fragments (blue pentagon and cyan heptagon) are joined via a linker (orange lines). C. Merging: two overlapping fragments (blue pentagon-star and magenta octagon) are merged into a new compound (octagon-star)

Fragment growing is the most commonly used strategy [71], and is conceptually similar to traditional compound optimisation methods in high throughput screening campaigns [66] and in many cases, the same computational approaches have been used. Potential follow-ups are generated through expert intuition, substructure searches against large commercial databases, or via *de novo* methods. Then,

docking approaches are used to assess the fit to the receptor. These docking-guided optimisations were found to be by far the most common way in which computational methods have been used to drive fragment to lead growing [60]. The programs LUDI [72], SPROUT [73], CONCERTS [74], BREED [75], GANDI [76] are all computational methods that implement fragment growing methodologies, but they do not account for the synthetic accessibility of their predictions, which has limited their application [60]. Recently, deep learning methods that attempt to grow the fragment in 3D space, taking into account the structure of the binding site, have also been developed. Examples are DEVELOP [77], which incorporates pharmacophoric features in its molecule generation procedure, and its successor STRIFE [78], which can also incorporate information from hotspot mapping methods.

Fragment linking has perhaps been the most difficult strategy to implement historically, despite its conceptual simplicity [66]. The introduction of flexible linkers can disrupt the original binding mode of the fragment, or even introduce an entropic penalty causing loss of binding affinity [49, 66]. DeLinker, introduced by Imrie and coworkers, is a deep-learning based method specifically designed for fragment linking, while also taking the 3-dimensional structure of the binding site into account as part of the generative process [79]. In cases where fragments have overlapping moieties, merging can be a simpler strategy than linking, as no flexible linkers need to be added [66].

Scaffold hopping methods have also been used to generate fragment linking and merges [79, 71, 45]. The computational pipeline introduced by Rachman, which also uses a scaffold hopping approach [45], will be discussed in detail in section 1.7.3.3. A similar approach was then used by the same group to elaborate a BRD4 fragment hit via scaffold hopping [80], resulting in chemically diverse, non-obvious suggestions from the neighbouring chemical space, including a com-

pletely novel scaffold [80].

1.6 Binding site hotspots

The structure of a protein is inextricably linked to its function - whether that be catalytic, structural, or any other among the myriad of roles that proteins perform in the cell. All of these different actions are created by the arrangements of a limited amino acid alphabet on several levels of structural complexity (from the primary structure, which corresponds to the sequence of amino acids, to complex quaternary structural arrangements of protein complexes with tens of components). Hence, it is not surprising that the surfaces of proteins are not uniform, and in fact, possess binding site regions that mediate their interactions with substrates, other proteins, inhibitors and effectors [81]. The first step in a rational drug discovery pipeline then involves locating and characterising such sites on the surface of the target protein [81].

1.6.1 Introduction to small molecule binding hotspots

One of the earliest experimental methods for systematic binding site mapping was Multiple Solvent Crystal Structures (MSCS) [82]. It was introduced in 1996 by Allen *et al.,* who solved the structure of porcine pancreatic elastase in diverse organic solvents [82]. Conceptually, this was based on earlier experimental work by Fitzpatrick *et al.,* who used crystal structures of subtilisin Carlsberg in acetonitrile [83], and Yennawar *et al.,* who solved the structure of γ-chymotrypsin in hexane [84], both with the goal of elucidating the structural factors involved in nonaqueous catalysis [82]. In MSCS, the organic solvents were chosen in such a way that the solvent molecules would represent functional groups in a potential larger inhibitor molecule [82]. The authors also envisioned the method as an experimental

counterpart and way of validating the predictions of computational methods such as Multiple Copy Simultaneous Search (MCSS), which places thousands of copies of a small molecule or functional group on the surface of a protein, then minimises them using an empirical potential energy function to generate 'functionality maps' of the protein surface [85, 86]. The structure of porcine elastase in neat acetonitrile was not found to differ significantly from that solved in an aqueous solution, while the crystallographic positions of resolved acetonitrile molecules in the protein's active site were found to coincide with functional groups of similar types in known inhibitor structures [82]. However, the number of sites discovered experimentally via MSCS was far lower than those discovered by the corresponding computational methods [81]. This discrepancy is addressed in the 1995 paper by Dagmar Ringe on what makes a binding site [81], and in a 2001 paper by English and colleagues [87], where they compared MSCS results with computational findings from MCSS and GRID [88] for the protein thermolysin (TLN) from *Bacillus thermoproteolyticus*. In both papers, the differences between the experimental and computational results are mainly attributed to the entropic and solvation effects, which were not explicitly included in the calculations. Not considering the flexibility of the protein was also thought to contribute, albeit to a lesser extent, as the protein targets investigated were known not to show great conformational flexibility [81, 87, 82]. GRID estimates the nonbonded interaction energy between a single atom probe at each *xyz* position and an atom of the protein as the sum of three main components: a Lennard-Jones function, an electrostatic function, and a hydrogen bond function [88]. MCSS uses parameters from the CHARMM 20 forcefield, augmented by information from *ab initio* calculations, to model the protein and the molecular probes used. The interaction energy is then determined using a classical time dependent Hartree (TDH) approximation [86]. Crucially, in both of these methods, only the interactions between the probe groups and the

protein atoms are considered.

1.6.1.1 Characteristics of small molecule binding sites

Proteins function within an aqueous environment, and so in its physiological state, the surface of the protein is covered in shells of water molecules. Any interacting molecules must then be able to disrupt these shells and displace those waters in a way that leads to ligand binding. Ringe and colleagues observed that even when a protein was crystallised in a neat organic solvent, a shell of water molecules remained on the surface of the protein, with certain waters being displaced by solvent molecules that formed highly specific interactions with particular residues [81]. They also noticed that the sites in which organic probes tended to cluster had an unusually hydrophobic character for areas on the protein surface, and that the molecules present in those sites tended to be less well ordered, but still observable crystallographically [81]. They further suggested that displacing such partially ordered waters may assist ligand binding by providing an entropy gain when the partially ordered water molecules are released. Ligand binding regions also tended to be found in depressions on the protein surface [81].

These findings about the nature of binding sites (summarised in Table 1.1) have been experimentally confirmed in the decades since by multiple studies. Notably, in 2005 Hajduk and colleagues used data from heteronuclear NMR-based screening to derive relationships between the 'drugability' of a protein binding site (referring to its ability to bind small molecule ligands) and parameters describing the properties of the binding site [89]. Regardless of binding affinity, nearly 90 % of the NMR screening hits across 23 diverse protein targets were found to bind at known small molecule binding sites, highlighting the general property of ligands to bind in specific small molecule 'hotspots' on the protein surface. In addition, targets with hit rates greater than 0.10 % yielded high affinity (Kd < 300 nM) drug

Table 1.1: Features of small molecule binding sites, as formulated by Ringe [81]
.

Feature	**Description**
Concave	Located in depressions on the protein binding surface
Displaceable waters	Contain partially disordered waters that can be displaced by ligand binding
Hydrophobic	Have more exposed hydrophobic surface area than expected for patches on the protein surface
Specific polar contacts	Contain residues that make specific polar contacts with bound ligands

leads, while no compounds with comparable potency had been reported for the 9 targets without such small molecule binding hotspots [89]. An analysis of the active vs. inactive pockets showed that those with low hit rates had 35 % lower apolar surface area, smaller volumes, and lower geometric complexity [89].

In 2007, Young and coworkers investigated the thermodynamics of two protein active site recognition motifs that were found to lead to particularly high ligand binding affinities [90]. The first of these were strongly hydrophobic cavities that enclose multiple water molecules, while the second motif involved the formation of 1-3 hydrogen bonds between the protein and the ligand within a very hydrophobic local environment. In the latter case, the contribution to binding affinity was greatest when the hydrogen bonds were in close proximity on a ring system of the ligand [90]. Molecular dynamics simulations found that such hydrophobic enclosed environments enhanced the affinity of the protein-ligand complex by perturbing the solvation of the binding cavity. In 2010, Muley and colleagues also investigated cooperativity between hydrophobic interactions and hydrogen bonds

in enclosed hydrophobic environments [91]. These motifs were suggested to be highly relevant targets for drug discovery efforts, as it would be possible to obtain a significant enhancement in potency with a minimal increase in molecular weight of the ligand [90, 91].

1.6.1.2 The thermodynamics of small molecule fragment binding is enthalpy-driven

The thermodynamic measure of ligand binding affinity is the binding free energy, ΔG in equation 1.1, which consists of enthalpy (ΔH) and entropy (ΔS) terms.

$$\Delta G = \Delta H - T\Delta S \tag{1.1}$$

The relative contributions of these components are known to provide information on the specifics of the binding event. The enthalpic term reflects the net change in the number and strength of non-covalent bonds formed upon complex formation [92]. Polar contacts have strong distance and angle dependencies, and interactions between favourably ordered polar groups contribute to favourable binding enthalpy [93, 94]. These favourable changes in enthalpy originating from a good fit between the ligand and protein are considered key to the binding process, as they compensate for the entropic loss of conformational degrees of freedom upon the formation of the protein-ligand complex [93, 92]. In certain cases, introducing a change to a compound meant to increase the enthalpic contribution may be offset by a disadvantageous entropy change arising from the reduced conformational freedom of the system, and so the overall binding affinity of the ligand may remain unchanged. This is known as the "entropy-enthalpy compensation" [93, 95]. On the other hand, gains in entropy can be achieved through a fit of apolar regions in the interacting partners, again leading to complexes with higher affin-

ity [93]. However, this strategy tends to produce large, hydrophobic ligands with low selectivity (and so unwanted off-target effects) and entropy driven binding - this phenomenon has been dubbed "molecular obesity" [14]. A 2008 study using isothermal titration calorimetry (ITC) data across more than 250 distinct protein-ligand interactions found this to be a key difference between naturally occurring biological ligands and the products of medicinal chemistry campaigns: the latter had a proportionally greater entropic contribution, as a reflection of the tendency for these campaigns to increase potency without optimising ΔH [95].

As stated earlier in section 1.4.2, the small size and strong polar character of small molecule fragments make them particularly suitable as starting points for drug discovery campaigns. In 2012, Ferenczy and Keserű investigated the thermodynamics of fragment binding using a data set of high-resolution X-ray data from the PDB [93]. Their results showed that on average, fragments formed two near-optimal geometry hydrogen bonds to their protein targets; these bonds tended to be located in a 5 Å sub-region of the fragment binding site. The total (linear) extent of the fragment binding site was found to be 10 Å [93]. The authors defined fragment binding hotspots as areas that "are able to form a limited number of strong H-bonds in a hydrophobic environment" and noted two key properties of such hotspots. Firstly, the strong hydrogen bonds within these hotspots were conserved in complexes with multiple different fragments and numbered, on average, 2 per hotspot. Secondly, these strong hydrogen bonds tended to be conserved between the fragment and elaborated ligands based on the fragment. An analysis of the binding entropy and enthalpy of the compounds showed that fragment binding was overwhelmingly enthalpy-driven, featuring strong H-bond formation [96]. With increasing ligand size, however, binding was increasingly skewed towards being entropy-driven, relying instead on apolar desolvation. As enthalpy-driven binders present a better starting point for high-potency leads, avoiding ligand obe-

sity, these findings were considered to provide further thermodynamic arguments for the utility of fragment-based drug discovery [96, 93].

1.6.2 Overview of computational hotspot-based methods

Given the importance of fragment hotspots in driving small molecule binding, a great number of computational methods have been developed over the past two decades to detect such sites on the surfaces of proteins. Based on their core premises and hotspot definitions, they can be loosely grouped as follows.

1.6.2.1 Atomic interaction methods

This first group of methods generates interaction maps of single atom probes that sample the protein surface, resulting in what can be considered "atomic" hotspots [48, 97]. GRID [88] and SuperStar [98] are two well-established, grid-based methods used to predict atomic hotspots. Unlike molecular hotspots, earlier defined as regions within the binding pocket that drive fragment binding, atomic hotspots can be thought of as a single favourable interaction, which may not be sufficient to drive the binding of an entire fragment [48, 97, 99]. Molecular hotspots cannot always be identified from the atomic interaction maps, as they are generated by single atom probes and so the context of these atoms within a larger molecule is lost [48]. Consequently, maps output by these methods tend to show favourable density spanning the surface of the whole protein and are not useful for locating putative binding sites (an example is presented by the SuperStar maps in Figure 1.6). However, in cases where the binding site is already known (either from experimental data or other computational methods), they can provide a detailed view of the interactions available for ligand binding.

1.6.2.2 Water-based simulation methods

Following from the observations on the presence of displaceable waters with unfavourable binding energies [81, 90, 100] in fragment hotspot sites, a number of water-centered methods for hotspot detection have been developed [101, 102, 103, 104, 105]. Schrödinger's WaterMap uses inhomogenous solvation theory to calculate the thermodynamics of binding site waters, and so can be used to identify those that are likely to be displaced by fragment binding [90]. WaterDock is an open-source method which predicts the positions and energies of waters in protein binding sites [101] based on AutoDock Vina scoring [106], leading to much shorter calculation times than using molecular dynamics in explicit water. Despite their higher computational costs, such methods provide valuable information on the water molecules present in the binding site, which is difficult to obtain through other experimental and computational methods. While the thermodynamics of water molecules have shown to be key to the binding of small molecules, such methods on their own do not highlight or rank specific interactions that can be exploited by the ligand.

1.6.2.3 Molecular probe binding consensus sites

These methods can be considered as the computational counterparts of experimental consensus mapping methods such as Multiple Solvent Crystal Structures [82], FragLites [28], and MiniFrags [29]. Most of them use a single, rigid receptor and model the solvent as a continuum [107]. The earliest of these was the previously discussed Multiple Copy Simultaneous Search (MCSS) [85]. FTMap is perhaps the most widely used of this group of methods. It ranks hotspots by counting the number of different probe types (from a total of 16 distinct small molecule probes) that bind to a given cluster on the surface of the protein. Probes that have polar atoms can be used to identify specific polar interactions in the

binding site [108, 97]. Astex have developed a protocol for binding site mapping based on their Protein-Ligand Informatics forcefield (PLIff) called PLIMap, which also falls into this category of methods and has been successfully used in drug discovery campaigns [99]. One advantage of such methods is that they can be used to predict small molecule binding hotspots from an apo protein structure. This makes them highly valuable in cases when little experimental information is available for the target protein. A possible limitation of these methods are the underlying models for protein and ligand chemistry. While they have shown to be useful, the approximations used are less sophisticated than the potentials used in MD, for example. In addition, these methods generally do not account for the flexibility of the protein.

1.6.2.4 Mixed cosolvent molecular dynamics methods

Molecular dynamics allows for the consideration of both solvation effects and protein flexibility - two of the factors that early computational hotspot detection methods generally failed to incorporate [81]. In the past decades, a multitude of such methods has emerged, highlighting the utility of such approaches despite their higher computational cost. One of the best known is MDMix [69]. It uses 20 ns MD simulations in the presence of 20 % organic solvent in order to ensure sufficient sampling of the probe molecules, while still retaining the aqueous environment needed to account for the effect of solvation [69]. Simulation coordinates are taken every picosecond and aligned. Then, a 0.5 Å grid is constructed around the protein, and the density of each probe type at each grid point is converted to binding free energy using the inverse Boltzmann relationship [69]. The authors compared their results to GRID [88] and report that MDMix's solvent probes are selective in displacing water molecules that are also displaced by ligands, while the GRID methodology discovered too many potential polar interaction sites [69].

Similar methodologies are MixMD [109], developed by Carlson and colleagues, and SILCS [107], developed by Faller *et al*, have also reported successful applications in ligand design. While recent advances in both hardware and software have lowered the time cost associated with using these methods, they are still not used routinely in high-throughput workflows. Generally, less computationally intensive methods are used in the first instance, with MD-based approaches used to probe specific subsets of input structures, or specific interactions within the binding site.

1.6.3 Fragment Hotspot Mapping

In 2016, Radoux *et al.* [48] introduced a method for fragment hotspot mapping that combined atomic and probe-based methods for hotspot identification and can identify the specific interactions that drive fragment binding. The authors found that not only could the method identify fragment binding sites, but the highest scoring interactions predicted were often those made by the fragment, with other parts of the lead molecule making more moderately scoring interactions. This makes the method a promising avenue for structure-based drug discovery campaigns, as it can highlight both the most important interactions within the binding site , and the moderately scoring areas in which initial fragment hits can be elaborated. As these moderately scoring areas are important for conferring selectivity [89], differences in the positions of "warm spots" in the binding sites of related proteins should indicate interactions that are important for selectivity.

1.6.3.1 Overview

The fragment hotspot mapping method has three main components. Initially, interaction maps for three atomic probes are calculated using the CSD program SuperStar [98]. Then, a protein surface buriedness calculation is performed, as fragments tend to bind in depressions on the protein surface. The atomic and

buriedness maps are then multiplied to give a weighted propensity map. Finally, to take into account the molecular nature of fragments, the weighted maps are sampled with three fragment-like probes. The full workflow is illustrated in Figure 1.6.

1.6.3.2 IsoStar and SuperStar

SuperStar is a method for identifying interaction sites in proteins based on experimental information about non-bonded interactions occurring in small-molecule crystal structures in the Cambridge Structural Database (CSD) and the Protein Data Bank (PDB) [98, 110, 111]. This data is stored in the IsoStar database, in the form of scatterplots which show the distribution of one group (the contact or probe group), around another one (the central group) [112]. The template protein is placed on a 3D grid, then broken up into fragments (IsoStar central groups). The IsoStar scatterplots corresponding to the selected central groups and probe molecule are overlaid over the protein as part of the SuperStar calculation. The scatterplots are converted into density maps in order to normalise them to the same scale. Overlapping maps are then combined and contoured on the final map, which is weighted to give the propensity.

As a second step to calculating hotspot maps, the protonated structures are input into SuperStar, where atomic propensity grids are calculated for three types of probe atoms. The acceptor probe is carbonyl oxygen, the donor probe is an uncharged NH nitrogen, and the apolar probe is an aromatic CH carbon. These are the settings recommended by the creators of the method [48, 97, 113].

1.6.3.3 Cavity Detection

Radoux *et al.* used LIGSITE [114] for cavity detection, but currently a morphology-based method called Ghecom [115] is recommended as a more reliable alternative

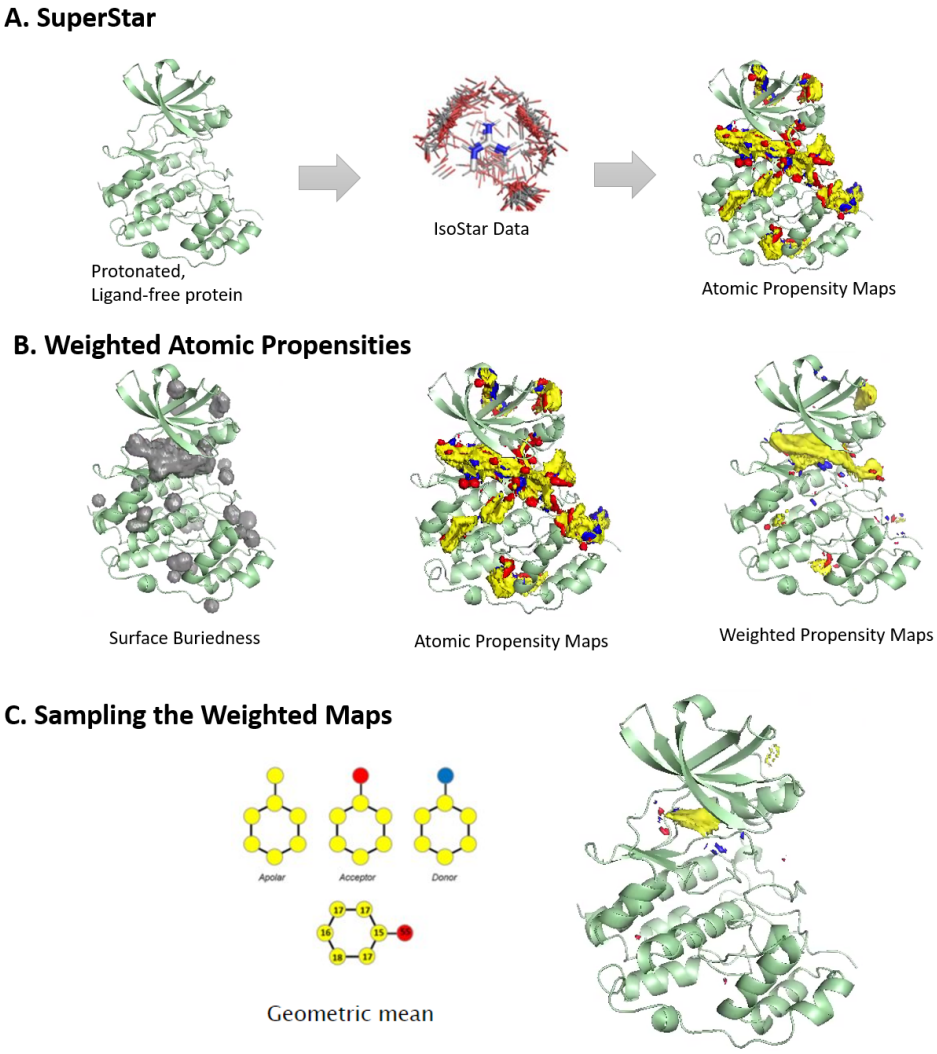


Figure 1.6: **The fragment hotspot maps algorithm.** A. Atomic propensities for the three probe types are calculated in SuperStar. B. The buriedness grid of the protein is generated and multiplied by the atomic propensity maps to give weighted propensity maps. C. The weighted maps are sampled by rotating three fragment-shaped probes on the surface of the protein to give the final hotspot maps. The hotspot scores reflect the geometric mean of the atom scores for the highest scoring pose at each point. Poses that clash with the protein are discarded. Adapted from [97].

[113]. Ghecom is a cavity-finding program developed in 2010 by Takeshi Kawabata. A pocket is defined as a region where a small spherical probe can enter, but a large one cannot. The minimum size of pockets (as defined by the radius of the 'small' probe) is 1.87 Å, or the size of a methyl group. The radius of large probe determines the shallowness of the pocket. The metric that the program outputs is $R_{inaccess}$: the minimal radius of inaccessible spherical probes. The Ghecom settings recommended by the creators of the hotspot mapping method are used. The grid spacing is set to 0.5 Å (to be consistent with the SuperStar grids). The radii of the minimum and maximum large spheres are 2.5 and 9.5 Å, respectively. These values were chosen for the reasons listed below.

1. Different sized ligands prefer different sized cavities, with $R_{inaccess}$ of 3-4 Å being most favourable for small molecule binding [115].

2. To ensure that the output values are similar to LIGSITE results for the purposes of weighing the maps, so that fragment hotspot scores using the two cavity detection methods are comparable and have the same score cutoff values.

1.6.3.4 Sampling

The propensity maps output by SuperStar are aligned to the buriedness grid output by Ghecom and the two are multiplied to give the weighted propensity maps. The highest scoring grid points from the weighted maps are sampled by probes containing either all carbons, or carbons with a single donor or acceptor heteroatom (shown in Figure 1.7). These probes were selected to mimic minimal fragments, which generally have a ring moiety. In addition, it has been observed that the hydrogen bonds contributing to fragment binding are often in the vicinity of a ring atom on the ligand [90]. Coupled with the atomic probes, this gives the advantage

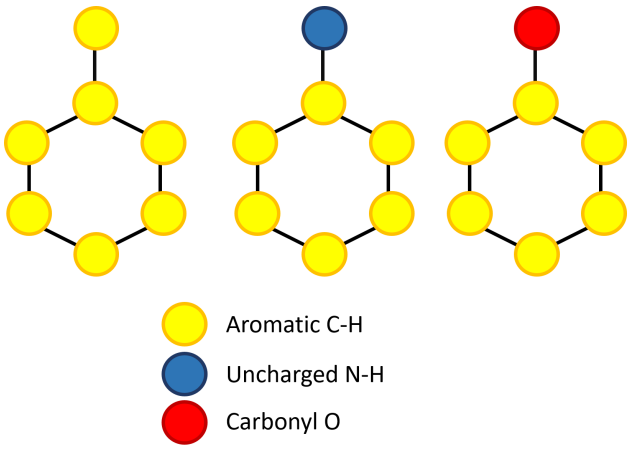


Figure 1.7: **Probes used to sample the weighted atomic propensity maps.** The bond orders of the probes are ignored; however, their geometry is planar. All atom types sample their respective weighted grid (e.g. carbons – apolar, oxygen – acceptor, nitrogen- donor).

of providing a molecular context to the "atomic" hotspot maps output by SuperStar. To sample the maps, the probes undergo a number of rotations, with a default value of 3000 [113], and are then translated so that the probes are centered on the heteroatom (for the polar probes) or the methyl group for the toluene probe. All rotations of a certain probe are placed on points in the weighted grid that are above a user-defined threshold (default value 15). Grid points that contain a probe atom are set to the score of the probe if the type of the probe atom matches the type of the grid (so carbon atoms from polar probes do not contribute to the polar maps). If multiple probes place atoms in the same grid point, the highest score is used. Poses are scored by taking the geometric mean of all atoms, which ensures that no poses that clash with the protein or fall outside the weighted buriedness maps are considered.

1.6.3.5 Validation of the fragment hotspot maps method

A key aspect of the fragment hotpot definition employed by the Radoux method is that fragments that bind within hotspots retain their binding upon being elaborated into larger lead-like compounds [48, 97]. In fact, fragment hotspots were defined as the minimum binding site that has that property [48, 97].The method was validated on 21 pairs of fragments and elaborated leads from a data set assembled by Ichihara and colleagues [100], which contain the fragment as a substructure of the elaborated lead. When scored against the hotspot maps, fragment atoms were found in the highest-scoring areas, with a median score of 97 % [48], and for lead atoms (outside the fragment core): 72 %.

1.6.3.6 Advantages and limitations of the fragment hotspot maps method

Fragment hotspot mapping is a method tailored to the specific challenges associated with FBDD. It is not computationally intensive and can be used not only to locate binding hotspots from a global search of the protein, but also to identify and prioritise the specific interactions that drive fragment binding. In its original implementation, the method could only be run on a single input structure, and so did not incorporate ensemble information or protein plasticity. Chapter 2 of this thesis describes the development of a protocol that extends the fragment hotspot maps method to handle ensemble data. Chapter 2 also describes a procedure that allows the comparison of fragment hotspot maps for closely related binding sites: an application that was not handled by the original method. Fragment hotspot mapping uses the buriedness of a point on the protein's surface as a proxy for modelling the effects of solvation. As buriedness contributes to the final hotspot score, the method may miss interactions in shallow surface pockets, for example those involved in protein-protein interactions. In addition, the toluene-like shape of the probes prevents them from fully sampling narrow, groove-like binding sites.

A set of probes with aliphatic tails is accessible through the Hotspots API [113], and may be used in such cases. However, their use may lead to a higher number of false positive hotspots [97]. Efforts to include a wider range of interactions (for example, positively and negatively charged probes) are also ongoing [113].

1.6.4 Extensions and applications of fragment hotspot mapping

The fragment hotspots method is available as an open source package called the Hotspots API [113], which can be found at `https://github.com/prcurran/hotspots`. In addition to the binding site mapping application described above, the API also includes functionality to extract tractability scores for protein binding sites, extract pharmacophoric features from the hotspot maps, score molecules against the maps, and compile and compare maps for ensembles of structures of the same protein. The ensembles and comparison applications will be discussed in detail in Chapter 2.

The fragment hotspot maps have also been incorporated into a generative model for scaffold elaboration, STRIFE, developed by Hadfield and colleagues in the Deane group [78]. The incorporation of hotspot maps allowed the method to take the protein structure into account when generating suggestions for elaboration. In a large-scale evaluation, STRIFE outperformed existing fragment elaboration workflows, providing highly ligand efficient suggestions for elaboration [78].

1.7 Structural stability and Dynamic Undocking

1.7.1 Looking beyond binding affinity in compound optimisation

Many of the methods introduced earlier in this chapter, and drug discovery efforts in general, focus on optimising binding affinity as a way of achieving potency. Binding affinity is an equilibrium thermodynamic property, and is measured either directly through assays that measure the binding of the compound to its macromolecular target, or indirectly, through the effect of the compound on the bioactivity of the target. However, non-equilibrium properties, such as the drug-target residence time, can also provide valuable information in drug discovery campaigns, and, in certain cases, may even be a better proxy for the *in vivo* efficacy of a compound [116, 117]. Binding affinity depends on the free energy difference between the bound and unbound states, both of which can be readily observed. Kinetic and non-equilibrium effects can be more challenging to study, as the local energy maxima and minima that define them are transient [117]. In addition, the observed macroscopic rate constants may arise from a combination of the contributions of multiple transient states, further complicating the study of these phenomena.

1.7.2 Dynamic Undocking

Computational methods such as molecular dynamics, which explicitly model all atoms and forces in a closed system and on very short timescales, are then particularly suited to studying these short-lived states. Steered molecular dynamics (SMD) methods in particular have been applied to probe the energy landscapes driving receptor-ligand binding and dissociation, as well as to provide a time-

resolved overview of the binding/unbinding process [117].

Steered molecular dynamics is a computational technique inspired by single-molecule pulling experiments [118]. An external, time-dependent force is applied to the system under investigation, pushing it away from equilibrium [119]. This has the effect of accelerating the transitions between minima in the energy landscape, allowing them to be analysed in atomic detail [119, 120]. This makes the method suitable for studying the binding and unbinding mechanisms for protein-small molecule complexes, and a number of studies aiming to exploit SMD as a drug discovery tool have been published over the past decade [121, 94, 122, 123].

In 2016, Ruiz-Carmona *et al.* introduced a method called Dynamic Undocking (DUck), which estimates the structural stability of the protein-ligand complex by measuring the work needed to perturb a ligand from its protein-bound state to a non-equilibrium, "quasi-bound" state in which a key native contact has just been broken [47]. Once a key hydrogen bond contact has been defined, an external force is applied to the participating ligand atoms as part of an SMD simulation. The work exerted by this force in moving the ligand atom a distance of 2.5 Å to 5 Å away from the protein is measured. The quasi-bound state is then defined as the point at which the work done by the pull force reaches a maximum, and is used as a measure of the structural stability of the protein-ligand complex [47].

The structural stability of a complex reflects its "ability to form a robust and precise binding mode" [120, 47]. This property is not a necessary prerequisite for tight binding (cases in which protein-ligand complexes without a defined binding mode show picomolar binding affinity have been reported [124]). However, a structurally stable binding mode can be highly advantageous, and is to some extent necessary, in structure-based drug design [47]. In their 2016 paper, Ruiz-Carmona and colleagues showed that a combination of docking and dynamic undocking is

highly effective in virtual screening applications for fragment binding, resulting in hit rates of up to 40% against the Hsp90 chaperone [47].

1.7.2.1 Structural stability and the quasi-bound state

The origin of structural stability lies in the steepness of the local minimum in the bound state [120], as illustrated in Figure 1.8. The three ligands, 1, 2 and 3 in the figure, will have the same values for the kinetic and thermodynamic constants, as they only depend on the relative energies of the unbound, bound and transition states. However, compound 1 has a much steeper local minimum (around the bound state), which is a profile that was found much more commonly among true binders compared to decoys [47]. The structural stability of a complex also differs from having a slow k_{off}, as the macroscopic off-rate will ultimately depend on a number of intermediate states along the dissociation pathway.

Hydrogen bonds are considered to be major determinants of structural stability, due to their strict distance and angular constraints [47, 125]. As discussed in Section 1.6.1.1, buried hydrogen bonds are a key feature of environments that favour the binding of small molecule fragments [48, 93]. The contribution of a hydrogen bond to the binding free energy of a protein-ligand complex is context-dependent, ranging from 0 to -1.5 kcal/mol [126, 127]. In addition, Schmidtke *et al.* showed that certain hydrogen bonds, originating from "almost buried" atoms on the protein surface, are "highly robust to small structural distortions" [94]. The dissociation of these water-shielded hydrogen bonds was shown to have a high energy transition state, causing them to be exchanged at slower rates [94]. The origin of this effect was postulated to lie in the decoupling of the association and dehydration processes, causing such bonds to act as "kinetic traps" that prevent structural fluctuations [94]. Although the dissociation of the complex may involve multiple steps, the rupture of such water-shielded bonds was shown to be able to

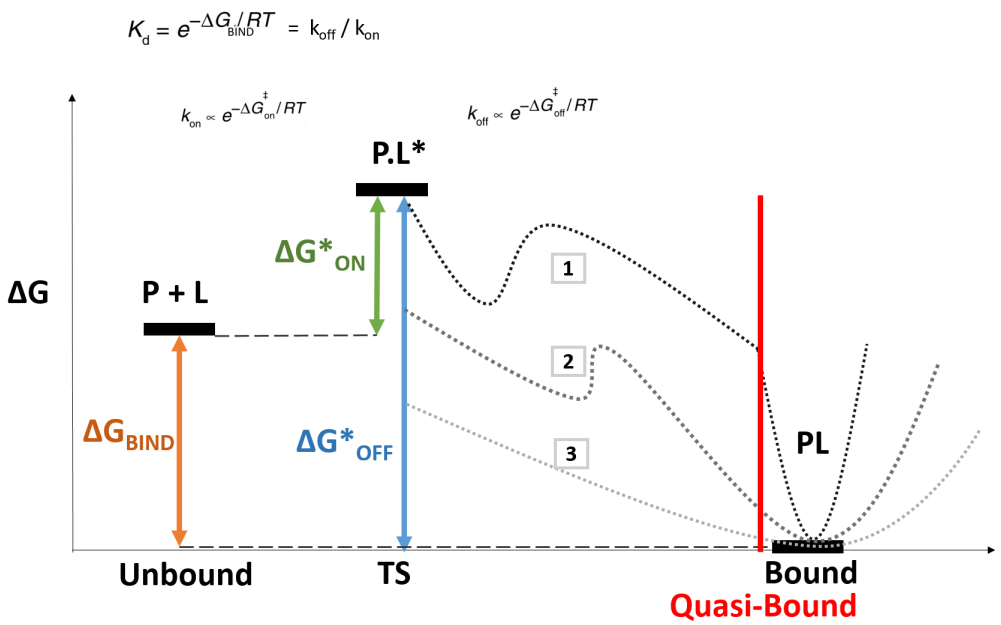


Figure 1.8: The quasi-bound state in the context of ligand association/dissociation. The quasi-bound state is illustrated in the context of a 3-state protein (P) -ligand (L) association model, passing through a high energy transition state (P.L*). The highlighted energy profiles reflect three separate ligands with identical macroscopic binding constants, but different levels of structural stability, as estimated by the energy of the quasi-bound state. In virtual screening applications, true binders have profiles similar to that of Ligand 1. Adapted from [47]

.

affect the rate-limiting step, influencing the dissociation process as a whole. This provided further motivation that the energy of the quasi-bound state could be a useful metric in structure-based drug discovery [47].

1.7.2.2 The DUck calculation

In the steered MD phase of the DUck calculation, an external force is applied on the ligand atom involved in the H-bond interaction under investigation. This force can be thought of as a spring, which pulls the ligand atom to a point located at a set distance (R in Figure 1.9) away from the protein atom [47]. The reaction coordinate is the distance between the pre-defined protein and ligand atoms involved

in the interaction. The pull distance R is varied between 2.5 and 5 Å, which spans the range from a close contact to a broken contact, in which the ligand is displaced by solvent [47]. This distance is divided into equal steps; at each step, the work done by the pull force is calculated (see Figure 1.9, panel A) and the pull distance is updated, effectively 'pulling' the ligand away from the protein. This results in a profile similar to that seen in Figure 1.9; initially, work is applied to bring the ligand to 2.5 Å from the interacting protein atom (labelled as "Start" on the trace in Figure 1.9, panel B). The work exerted by the external force then decreases, reaching a minimum around 3 Å. After this, the work increases again as the ligand atom is being pulled from its equilibrium position. The quasi-bound state is defined as the point at which the work done by the pull force reaches a maximum [47]. When more than one peak is present in the trajectory, the value of the first maximum is taken as the W_{QB} [128] The simulation does not use the full structure of the protein, focusing instead on a "chunk" of binding site residues in a 6-9 Å radius around the interacting protein atom. This considerably reduces the computational time, as well as removes parts of the protein that could impede the ligand's exit trajectory during the steered MD [47, 128].

1.7.2.3 An open-source DUck workflow

The DUck method is currently implemented using two commercially available pieces of software: MOE [129] and AMBER [130]. MOE is used for the chunking and parameterisation of the protein and ligand, while the AMBER molecular dynamics package is used to run the simulation. Both of these require a licence, although AMBER is free for academic use. The authors of the DUck method note that an open-source version of DUck (OpenDUck) is currently being developed. The implementation of the main stages of the DUck workflow using open source Python tools was originally done by Anthony Bradley and Peter Schmidtke, and

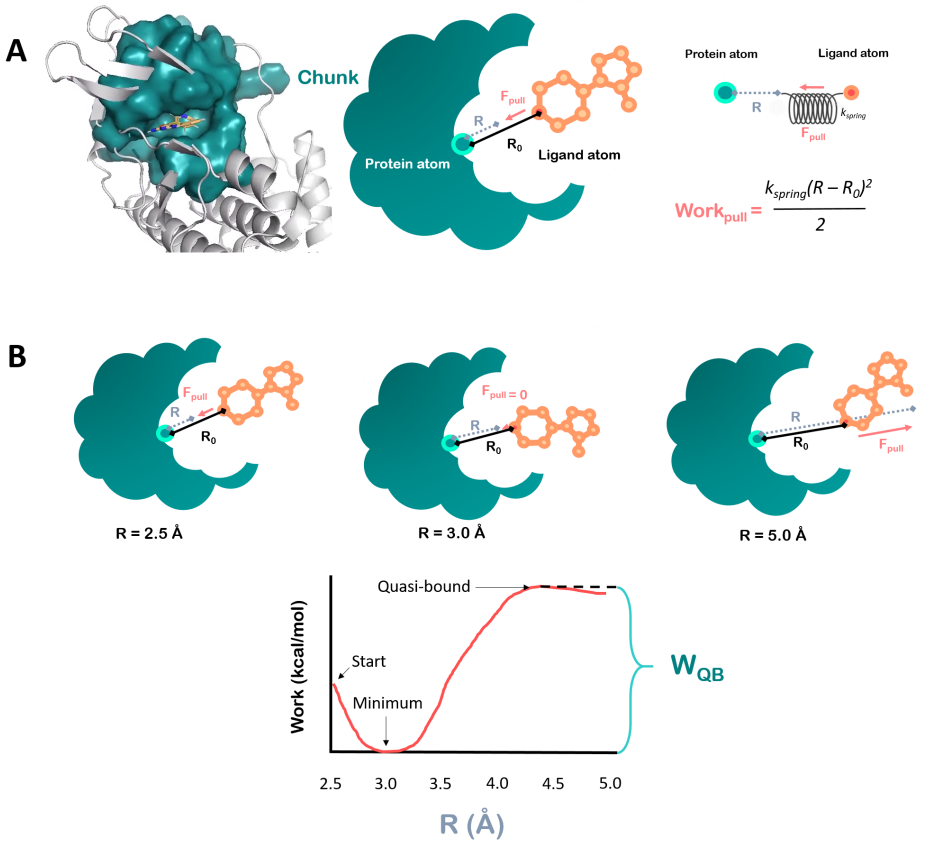


Figure 1.9: The W_{QB} calculation.
A. A "chunk" of residues is selected around the protein atom in the H-bond contact and is used as the receptor model. To measure structural stability, a force is applied to the ligand atom. This force acts as a spring, pulling the ligand atom to a distance R away from the protein. The reaction coordinate is the distance between the protein and the ligand atom (R_0). The work done by the pull force can then be computed according to Hooke's law. B. The pull distance R is varied in small increments between 2.5 and 5 Å, simulating the rupture of the hydrogen bond.At each increment, the work done by the pull force is computed. The quasi-bound state is defined as the point where the work reaches a maximum. W_{QB} is the difference between the minimum and maximum work values observed in the trace.

can be found at https://github.com/xchem/duck. Simon Bray then updated the dependencies and re-packaged the input and output as part of the Galaxy Project [131] https://github.com/galaxycomputationalchemistry/duck. This is the version that will be used, evaluated and extended in this thesis.

OpenDUck has two key advantages over the original implementation. The first is accessibility: all the libraries and packages used are freely available (including the source code) and can be integrated with each other (and other Python-based methods) from within Python. Many of these (for example, the OpenMM [65] and OpenForcefield [132] libraries) are the subject of international academic collaborations and are constantly updated, improved and documented. The availability of the source code and the large user and developer communities in these projects also greatly help with debugging and improving components of the workflow.

OpenDUck's second advantage lies in the MD engine it uses - OpenMM [65]. This is an open-source MD simulation toolkit designed to enable the fast and efficient implementation and extension of MD methods and workflows [65]. It works using a three-layer architecture, of which the top-most (user-facing) layer enables the use of Python scripting for setting up simulation protocols and performing file I/O. This layer also allows for custom forces to be defined, starting from a mathematical expression and using minimal code [65]. This layer communicates with the underlying kernels using a C++ API, in which the forces and integrators are implemented. The underlying code is optimised for running high-performance calculations on both CPUs and GPUs (OpenCL/CUDA). Custom forces are parsed, compiled, and optimised internally by the simulation engine, so minimal development is required to implement new features. Finally, the package supports multiple input pipelines, tools, and formats, allowing it to integrate with existing MD preparation pipelines.

The OpenDUck implementation available at the start of my PhD project already had the input, simulation, and analysis components needed to run a full DUck calculation and arrive at the final W_{QB} value. The core functionality had been implemented by Anthony Bradley and Peter Schmidtke in 2017/2018, then Simon Bray had updated the dependencies to the latest compatible versions in 2020, as well as developed scripts to perform the automated chunking and MD simulations. I built on this work by allowing the chunking script to retain specific waters and residues supplied by the user, as well as developing a script that allows the DUck trajectories to be visualised using VMD [133]. This extended version of OpenDUck, including scripts for the diagnostics and analysis introduced in this thesis (specifically in Chapter 3), can be found at `https://github.com/mihaelasmilova/duck`.

1.7.3 Further applications of Dynamic Undocking

The initial work by Ruiz-Carmona *et al.* showed that the quasi-bound state can be used to usefully separate true binders from decoys in virtual screening applications, and introduced dynamic undocking as an extremely promising tool for structure-based drug discovery. In a subsequent paper from the same group, Majewski and colleagues applied the dynamic undocking method to investigate the structural stability and degree of robustness of protein-ligand complexes across different protein families [134].

1.7.3.1 Using Dynamic Undocking to investigate structural stability across protein families

While the original application only measured the W_{QB} for a single, "key" interaction between the protein and the ligand, this analysis extended to all hydrogen bonds that were detected in 79 complexes (345 detected hydrogen bonds in total)

taken from the Iridium dataset [135] and 27 fragment-protein complexes (with 126 detected hydrogen bonds) from the SERAPhiC dataset [136].

The Iridium dataset [135] was created to provide a highly trustworthy set of structures to benchmark docking programs. It consists of crystallographic structures in complex with drug-sized ligands, extracted from four datasets that had previously been used to validate docking programs [135]. These structures have been re-refined, and show no ambiguities in the binding site region crystal density. This dataset will be discussed further in Section 3.3.1.1.

The SERAPhiC dataset [136] was devised as a benchmarking dataset for *in silico* FBDD methods. It also consists of highly trustworthy X-ray structures of protein-ligand complexes, with ligand mass restricted to be between 78 and 300 Da, excluding common crystallographic additives. The included structures have have a resolution of 2.5 Å or better, and the crystallographic density is freely available. This dataset will be further discussed in Section 3.3.1.2.

Both of these datasets consist of high-quality crystallographic structures, and so the conclusions drawn from this analysis are likely biased towards complexes that crystallise. This makes them highly relevant to structure-based drug design, which greatly relies on crystallographic structures of protein-ligand complexes.

Structurally stable hydrogen bonds are common in protein-ligand complexes

57.4 % of the bonds in the Iridium dataset were classified as 'robust', or able to provide structural stability, while 75 % of the complexes were anchored by at least one robust hydrogen bond. "Robust" hydrogen bonds were defined as having W_{QB} values greater than 6 kcal/mol, while "labile" bonds have W_{QB} values of less than 4 kcal/mol [47, 134]. On average, complexes had 2.5 robust hydrogen bonds, but the distribution varied greatly between

pocket types and protein families. W_{QB} values for the 345 measured bonds varied between 0 and 26 kcal/mol, with the peak of the distribution between 0 and 6 kcal/mol.

In the context of a fragment elaboration workflow, these results show the importance of considering structural stability in, for example, complexes of the receptor with docked follow-up poses. They also indicate the range of values that can be expected, as well as providing W_{QB} thresholds for the detection of stable interactions.

Protein atom buriedness is necessary but not sufficient for structural stability

Previous work from Schmidtke and colleagues had shown that 'water-shielded' hydrogen bonds between a buried protein atom and the ligand can act as 'kinetic traps', providing strong opposition to structural fluctuations. The solvent accessible surface area (SASA) of all protein atoms involved in the 345 interactions of the Iridium set was computed and plotted against the W_{QB} values for those interactions. The result showed that while the interactions with the highest W_{QB} values had low SASA values, not all buried protein atoms were involved in robust interactions. Therefore, protein atom buriedness alone is not sufficient to predict a structurally stable hydrogen bond.

When using DUck as part of a fragment follow-up workflow, calculating the SASA of the protein atoms involved in target interactions can then be used to assess their suitability for the method. This would decrease the overall computational time used by the workflow by indicating interactions where DUck may not be informative.

Structurally robust hydrogen bonds act cooperatively, forming fragment-sized anchors

All hydrogen bonds in a complex were clustered based on their spatial coordinates into fragment-sized atom groups. In 62 % of complexes, the robust hydrogen bonds were located in a single group, with 23 % of ligands forming two structurally distinct anchors (common in carbohydrate binding sites), and three ligands formed three such anchors. Curiously, these three ligands had unrelated functions and structures. Isolated hydrogen bonds generally did not form structurally stable interactions, but when three or more hydrogen bonds clustered together, robust complexes were usually formed. These hydrogen bonds usually had similar W_{QB} values, implying that they function in a concerted and synergistic manner.

This observation implies that ways of combining W_{QB} scores for the fragment-sized anchors tethering a molecule may be a useful metric in a fragment elaboration workflow. For example, they could be used to assess the overall stability of a docked pose, providing more complete information than looking at the stability of a single interaction.

Structurally robust bonds often cluster in binding hotspots

For all kinases and proteases in the Iridium data set, the anchoring cluster of robust hydrogen bonds corresponded to the known binding hotspots for these two protein families. This was also the case for most of the fragments in the SERAPhiC data set. Binding hotspots are areas that tend to have a disproportionate contribution to binding affinity [48], yet there is no correlation between W_{QB} and binding affinity, as previously mentioned [47, 134]. The authors concluded that the origin of the free energy barrier that gives rise to structural stability lies in a transitory dissociation penalty. This dissocia-

tion penalty may arise from a physical decoupling between hydrogen bond rupture and desolvation, as was previously reported by Schmidtke *et al.,* for water-shielded hydrogen bonds. This is supported by the observation that surface buriedness and water-shielding are necessary for a high W_{QB} value, while solvent exposed hydrogen bonds always showed low W_{QB} values.

This observation also presents a strong case for using DUck in tandem with a hotspot mapping method within fragment elaboration pipelines. The hotspot method would be used to identify the key hotspot interactions. The stability of potential follow-up molecules with these key interactions would then be evaluated.

1.7.3.2 Using dynamic undocking to predict ligand binding mode

The work discussed in section 1.7.3.1 showed that while structural robustness is not a necessary condition for binding affinity, cases of labile protein-ligand complexes are rare. From the perspective of structure-based drug design specifically, a stable binding mode that is robust to perturbations is highly desirable. Building on these ideas, in 2020 Majewski and Barril extended the Dynamic Undocking method to consider global structural stability, and applied this approach to predicting the binding mode of protein-ligand complexes with both fragments and larger, drug-like ligands [137]. This involved docking the structures of ligands with known binding modes from the previously discussed Iridium [135] and SERAPhiC [136] data sets, then calculating W_{QB} values for all detected hydrogen bonds for the 5 best scoring unique poses for each ligand (poses were clustered and a pose was defined as unique if its RMSD to the other clusters was greater than a cutoff of 1.5 Å for fragments and 2.0 Å for drug-like ligands).

The above described protocol generated several W_{QB} values per docking pose.

To enable the comparison of docking poses, the individual W_{QB} scores needed to be combined into a single value. As structurally robust bonds tend to cluster in fragment binding hot spots, and fragments generally have fewer rotatable bonds, which would complicate the analysis, the fragments in the SERAPhIC set were initially used to explore this question. Four ways of combining the scores were compared, each testing a different assumption.

Taking the maximum W_{QB} value per pose tests the assumption that the most robust structural bond defines the stability of the complex (DUck_max). Conversely, taking the minimum observed value in a complex tests the assumption that the complex dissociates once the most labile bond is perturbed (DUck_min). The other two assumptions investigated were that the work needed to break the complex corresponds to the average W_{QB} (DUck_ave), or that each hydrogen bond acts independently and their contributions are additive (DUck_sum), favouring poses that form the greatest number of hydrogen bonds. For fragment-sized ligands, DUck_max and DUck_sum were found to outperform docking, while DUck_min showed inferior performance. Consequently, the "weakest link" assumption was not considered in further analyses [137].

Using the observation that for larger ligands, structural stability tends to originate from fragment-sized 'anchors' that form spatially clustered robust interactions (with other parts of the ligand lacking structural stability, or acting as secondary anchors), Majewski and Barril then considered other ways of combining W_{QB} values into a single pose score.

Hydrogen bonds detected for a pose were grouped based on their mutual distances, based on a distance cutoff. In all of the cases, this value was found to be in the 3-4 Å range, reflecting a local (chemical group level) analysis of structural stability, rather than on the atomic or whole molecule level. The best performing way for

combining scores was Gr_ave (taking the average W_{QB} values within a group, then summing the scores of all detected anchoring groups), which had a success rate of 76.6 %, comparable to that of docking [137].

In the context of a fragment elaboration workflow, such combination scores could be used to summarise the structural stability of docked poses. The implementation used for comparing docked poses of the same ligand (as is the case in binding mode prediction) may not be the most suitable for comparing the poses of different follow-up compounds, especially if they have a different number of fragment "anchors". Combining W_{QB} scores as part of a fragment elaboration workflow will be further discussed in Chapter 4.

The work described in this section laid the foundations for defining structural stability and its use in drug discovery campaigns. Building on these ideas, Rachman and colleagues in the Barril group explored ways in which DUck can usefully be integrated into computational drug discovery (CADD) pipelines that combine complementary thermodynamic and structural stability-based computational methods with the goal of progressing fragment hits into lead compounds [45, 80].

1.7.3.3 Integrating Dynamic Undocking into a CADD pipeline

The pipeline described by Rachman performed iterative structure-based scaffold hopping, starting with a similarity search between the starting fragment and the search library. The maximum common substructure between the fragment and the suggested followup was then identified and used as a restraint for tethered docking. Molecular Mechanics using the Generalised Born model and Solvent Accessibility (MMGBSA) was used to assess solvation effects for the docked poses; only the solvation term was considered. DUck along a pre-defined key hydrogen bond was then applied to assess the structural stability of the poses of interest.

In cases where sufficient information for the binding site was not available, the MDMix [69] binding site mapping method was used to identify key interactions to be further investigated with DUck.

This automated platform was applied to NUDT21 in the Nudix protein family. An XChem fragment screen had been performed against this target, yielding over 40 crystallographic hits. However, these had few common interactions in the putative binding site, so MDMix was used to assist the choice of key hydrogen bond contacts to be used in the dynamic undocking step. Three iterations of the automated pipeline yielded a total of 52 compounds that were ordered to be tested by Surface Plasmon Resonance (SPR). Of these, 15 compounds showed dose dependency, with the most potent molecule displaying a Kd of 200 μM [45]. For one of the fragments which had generated no successful followup compounds on the first attempt, a second round of iterations was attempted using more permissive parameters for the maximum common substructure detection step in the scaffold hopping stage. This, combined with increasing the search space from commercially available compounds to synthesisable ('on-demand') molecules, resulted in the most potent hit reported at the time, with a Kd of 10 μM [45].

To test the ability of the pipeline to progress fragment hits for which structural data is not available, four test cases were explored within three target systems: HSP90, BRD4, and DYRK1A. [45]. In the case of BRD4, 11 compounds were ordered and shown to bind by differential scanning fluorimetry at 10 μM. For two of these, crystallographic structures could also be obtained. For one of these compounds, the binding mode had shifted following a methyl to ethyl transition, while the other one had a flipped binding mode compared to the prediction. However, constrained docking using the crystallographically observed pose yielded lower scores compared to the prediction, as did DUck [45]. For one of the Hsp90 test fragments, the pipeline was able to generate compounds with binding observable

by SPR, but none could be crystallised with the target. One of the predicted binding modes for the second Hsp90 fragment generated two ligands with significant improvements in potency and good ligand efficiencies, as well as good overlap with the predicted starting fragment poses. In the case of DYRK1A, two compounds had better affinities than the starting fragments, one of which maintained the same binding mode. From these experiments, Rachman concluded that while the automated platform was able to generate suggestions for active compounds from fragments with an unknown binding mode, it is not suitable for identifying the correct binding mode of the starting fragment [45].

A similar workflow was used by Piticchio *et al.* to perform prospective scaffold hopping on a BRD4 fragment hit, resulting in chemically diverse hits that explored the adjacent chemical space, and were found to be active in biophysical assays [80]. This workflow used a similarity search of compounds in the ZINC15 [55] database that were 2 heavy atoms away from the starting hit. These were then superposed on the starting fragment structure and only those predicted to retain the a conserved interaction to an asparagine residue [138] were retained and docked. DUck was used to prioritise the top 500 highest scoring and most chemically diverse compounds from the docking stage, based on the structural stability of the conserved hydrogen bond. The workflow was used iteratively and succeeded in suggesting compounds that were found to be active in biophysical assays, including one that displayed a completely novel scaffold [80]. The X-ray structure of the hit in complex with BRD4 showed that it assumed the predicted binding mode. The authors note that in addition to being a highly promising method for scaffold hopping, the platform could in future be extended in to drive fragment growing campaigns as well.

Overall, these two workflows showed that pipelining computational methods with increasing levels of rigour could be used successfully for fragment elaboration,

and provided a starting point from which the interplay between dynamic undocking and other computational methods could be leveraged to tackle challenges in fragment hit elaboration.

1.8 Aim and Objectives

Over the past two decades, FBDD has established itself as a powerful tool for developing probe and drug candidates by rationally elaborating small chemical fragment hits into larger, optimised lead compounds. Recent technological advances have enabled the use of X-ray crystallography as a medium-throughput screening tool for FBDD. This results in a wealth of structural data on low molecular weight molecules in complex with a protein target. Interpreting this data and distilling it into prioritised suggestions for the elaboration of fragment hits into leads with increased potency and selectivity for the target protein is currently a significant challenge to using this technique. Consequently, there is a great need for computational methods to address this challenge.

In this thesis, I will investigate the use of existing computational methods in the space of computational drug discovery within the specific context of crystallographic high-throughput fragment screening. Starting from an ensemble of structures of a protein in complex with small molecule ligands, a workflow will be introduced to summarise this multidimensional structural information in a format that is both visually intuitive for human users, as well as amenable to computational analysis and downstream processing. A procedure to quickly compare binding site interaction information for related protein targets will be introduced. The methods presented here aim to streamline and facilitate the process of hypothesis generation after a successful fragment screening campaign, as well as to provide a framework for assessing the poses and interactions made by the suggested virtual

follow-up compounds.

Finally, a workflow that combines these different methods will be presented, along with a guidelines and recommendations for its application to the prioritisation of fragment screening hits, generation of hypotheses for follow-up compounds, and the selection of hypothetical follow-up compounds for experimental assays.

1.9 Thesis Outline

The remainder of this thesis is organised as follows:

Chapter 2 will detail the reasoning behind using the fragment hotspot maps method as a way to capture information about the protein's binding site and prioritise the interactions observed. A workflow for combining information from multiple structures of the same protein will be introduced, as well as a method for comparing closely related binding sites to identify putative selectivity-determining regions. The utility of this workflow in the identification of selectivity-determining regions between closely related proteins will be demonstrated through its application to retrospective examples in two well-understood protein families: bromodomains and kinases.

Chapter 3 will look at a complementary computational method that can be used to prioritise fragment hits and the derived suggestions for followup compounds. The steered molecular dynamics method called Dynamic Undocking will be used to assess the structural stability of the bound fragment hits, as well as to provide insights into the contribution of individual interactions. To facilitate integration into computational workflows, an open-source implementation of the method will be benchmarked, and a simple diagnostic for the output W_{QB} values will be introduced. Ways of using

the method to prioritise closely-related fragment follow-up compounds will also be explored.

Chapter 4 will describe the development and use of an integrated workflow combining the extended fragment hotspot maps methods, dynamic undocking, traditional docking, and a chemistry recommendation engine in prospective medicinal chemistry campaigns starting from crystallographic fragment screens. This work will focus on three target proteins: ACVR1, a kinase target implicated in cancer and connective tissue diseases; NSP13, a viral helicase from the SARS-CoV-2 pathogen, and PARP14, which is a human cancer target. The application of the workflow to suggest follow-up compounds for these projects will be described, and experimental results will be presented for the selected compounds.

2 | Extending the fragment hotspot mapping method

The work presented in this chapter has largely been described in the following paper published in the Journal of Chemical Information and Modeling in 2022.

Smilova, M. D., Curran, P. R., Radoux, C. J., von Delft, F., Cole, J. C., Bradley, A. R., & Marsden, B. D. (2022). Fragment Hotspot Mapping to Identify Selectivity-Determining Regions between Related Proteins. Journal of Chemical Information and Modeling, 62(2), 284–294. https://doi.org/10.1021/acs.jcim.1c00823

2.1 Introduction

A successful crystallographic screening experiment results in an ensemble of structures of the target protein in complex with small molecule fragment hits. While other biophysical methods used for fragment screening output a measure of the hit compounds' binding affinities, which are usually used to prioritise the hits, crystallographic fragment screening output is categorical ("a hit has been observed"). The method's chief advantage lies in determining the binding mode of the hit: its position in the protein's binding site and the protein-ligand interactions made. An analysis of target-fragment complexes can show what interactions of the binding site are exploited, as well as give an idea of what further opportunities might be available. However, not all interactions within the binding site contribute equally to fragment binding. The task of prioritising the fragment hits then equates to selecting those molecules that make the most favourable interactions with the binding site, and provide opportunities for future elaboration. The success of

consensus hotspot methods (MSCS [82], FTMap [108], MDMix [69], *etc.*) shows that prioritising interactions that are shared by multiple fragment hits is a viable strategy. In an experimental setting, it could be the case that not enough hits are observed to make such conclusions. Furthermore, proteins are not static entities, and individual fragment-protein structures may reveal subtle structural rearrangements that lead to dramatic changes in the binding site's interaction profile. Interactions that could confer selectivity over closely related proteins also provide valuable suggestions for early hit optimisation.

To aid with the task of summarising the data from a crystallographic fragment screen, a method was needed that could capture information from multiple structures of the same protein in a way that identifies the most important interactions and structural changes that influence the binding and selectivity of small molecule ligands. The output of this method should be intuitively interpretable to the investigator, as well as amenable to computational analysis and automation.

This chapter will describe the rationale behind choosing the fragment hotspot mapping method as a way to encode binding site information following a crystallographic fragment screening campaign. A workflow for compiling binding site information from an ensemble of structures of the same protein will be presented, as well as a method for comparing the binding sites of two closely related protein ensembles. Case studies between pairs of proteins from two well-researched families, kinases and bromodomains, will be presented, as well as an example of using the method to automate comparisons across a target protein family.

2.1.1 Mapping and comparing binding sites

In Chapter 1, I presented how the use of methods to predict fragment hotspots, regions within the protein's binding site that make a disproportionately large con-

tribution to binding affinity [89, 48], has been reported in the literature both to determine potentially tractable pockets and sub-pockets on the protein surface [108, 109, 69, 85], as well as to guide the rational design of inhibitors [139]. I also introduced the method by Radoux *et al*, which takes the molecular context of fragment binding into account.

This method is a promising avenue for guiding the rational design of inhibitors, as the maps provide an intuitive visual guide to favorable interactions within the binding site and can indicate suboptimal interactions within the original hit. The maps also give an objective numerical understanding of the features important for binding, which allows them to serve as the basis of automated approaches for hit prioritisation and progression. The fast calculation times of this method (around 5-10 minutes per structure on an ordinary laptop) make it applicable to scenarios in which large numbers of protein structures need to be routinely analysed, such as the output of fragment screening campaigns, or for analysing frames from molecular dynamics simulations. The propensity scores used by the fragment hotspot mapping algorithm are also entirely data-driven, based on the high-quality, high-resolution data available in the CSD. This makes it complementary to MD-based methods, which use forcefields (empirical models) to estimate the potential energy between atoms in a system. It should also be free from the underlying biases in the datasets used to train most machine learning-based methods [140, 141].

As the propensity scores are based on interactions present in the CSD, however, the method will be influenced by underlying biases in the types of interaction data present in the database. In 2001, a comparison between using SuperStar with CSD and PDB-based interaction fields revealed that polar interactions are more common in small molecule crystals (CSD data), while hydrophobic contacts are more frequently present in protein-ligand interactions (PDB data) [142]. PDB-derived scatterplots also had no information on the protonation states of glutamate, aspar-

tate, and histidine side chains. In cases where these protonation states are known, the authors concluded that CSD-based scatter plots would provide a more accurate representation of the interactions. However, the CSD-derived scatter plots were also more sensitive to errors and uncertainties in the PDB models for which propensities were calculated [142]. As protein models are sanitised and prepared prior to running the hotspots calculation, and the hydrophobic effect is accounted for through the buriedness calculation, CSD scatter plots are used, and the biases associated with them have been addressed.

The fragment hotspot maps method takes a single protein structure as input. However, the output of a crystallographic fragment screen is an ensemble of structures of the same protein in complex with different ligands. Consequently, the first task I set out to address was determining how the information from multiple hotspot maps calculated for the same protein target can be summarised. A way to look for differences between the ensembles of closely related proteins would then also be needed, in order to address questions of selectivity and polypharmacology within a target family. The desired outcome for this would be a map of the binding site, intuitively showing selective and shared areas, as well as a measure of the confidence in assigning them.

Tools for presenting and analysing ensembles of structures of the same protein have been published previously, both in the context of analysing trajectories from molecular dynamics, or in the context of NMR ensembles [143], as well as specifically for summarising information from crystallographic fragment screens. Of the latter, Polyphony [144] and WONKA [40] are specifically geared towards protein-small molecule complexes, but they do not employ grid-based representations of the binding site. WONKA [40] uses the spatial information for waters, residues, ligand atoms and ligand pharmacophores, which are clustered to give a detailed representation of the persistence of features within the ensemble. A key feature

of the method is its interpretability, as observations can be linked back to the fragment-ligand complexes they originate from. This allows rare, but interesting from a medicinal chemistry perspective, structural features to be readily identified [40]. Polyphony uses Cα-spline curvature and torsion, alongside side-chain conformation, intermolecular interaction fingerprints, and pocket propertied as per-residue descriptors. These are then grouped by position in the sequence alignment [144].

Early work by Österberg *et al.* [145] combined ensembles of Autodock [146] interaction energy grids into a single grid by using Boltzmann-weighted averages for the values at each point in space. These grids showed improved performance in docking relative to those derived by taking the mean or minimum values of the interaction energy. More recently, Volkamer and colleagues [4] used grids generated by the pocket detection and drugability estimation algorithm DogSiteScorer [147], and then compared the frequencies at which points were observed in the target and off-target ensembles. Building on this work, in 2019 Turk *et al.,* used AutoGrid atom-based energy grids, using polar an apolar atom probes, as a key part of a computational pipeline to guide the automated selectivity conversion of an Aurora kinase inhibitor for the TrkA kinase [148]. In terms of extracting information from an ensemble of grids, Schmalhorst and Bergner have developed a method based on SiteMap [149] to identify structures with unique design opportunities within the ensemble [150], providing a further example of the utility and opportunities that can be explored by combining information from grid-based binding site representations.

To address the challenge of usefully combining information for structures of the same protein target, an "ensemble" hotspot map approach was developed. By comparing two ensemble maps, a hotspot selectivity map can be derived. This highlights the structural differences that contribute to the selectivity of a com-

pound for one protein over another. The ensemble and selectivity maps were parameterised using retrospective examples of compounds showing selectivity between proteins in the same family.

The case studies presented in this chapter come from two well-researched human protein families: bromodomains and kinases.

2.1.2 Structural features of bromodomain proteins

Bromodomains are small interaction modules that act as "readers" of lysine acetylation, one of the key histone post-translational modifications that regulate the structure of chromatin [151, 138]. Bromodomains act as acetyl-lysine recognition units in epigenetic regulator complexes that have been linked to a wide range of therapeutic areas, most notably cancer, inflammation, and viral infections [138, 152].

The bromodomain fold consists of a conserved left-handed bundle of four alpha helices, connected by loops of varying length [138], as shown in Figure 2.1, panel B. Acetylated lysines are recognised through an almost universally conserved asparagine residue in a narrow and mostly hydrophobic binding pocket [152]. Additional polar interactions are made between the histone tail and the residues lining the acetyl binding pocket. These surface residues differ between bromodomain proteins and so act as selectivity determinants [138]. Figure 2.1, panel D shows an example of the BRPF1 binding pocket interactions with an acetylated histone peptide.

2.1.3 Structural features of protein kinases

Protein kinases are proteins that transfer the γ-phosphate from ATP onto a the hydroxyl group of a specific residue (tyrosine, serine, or threonine) of a substrate

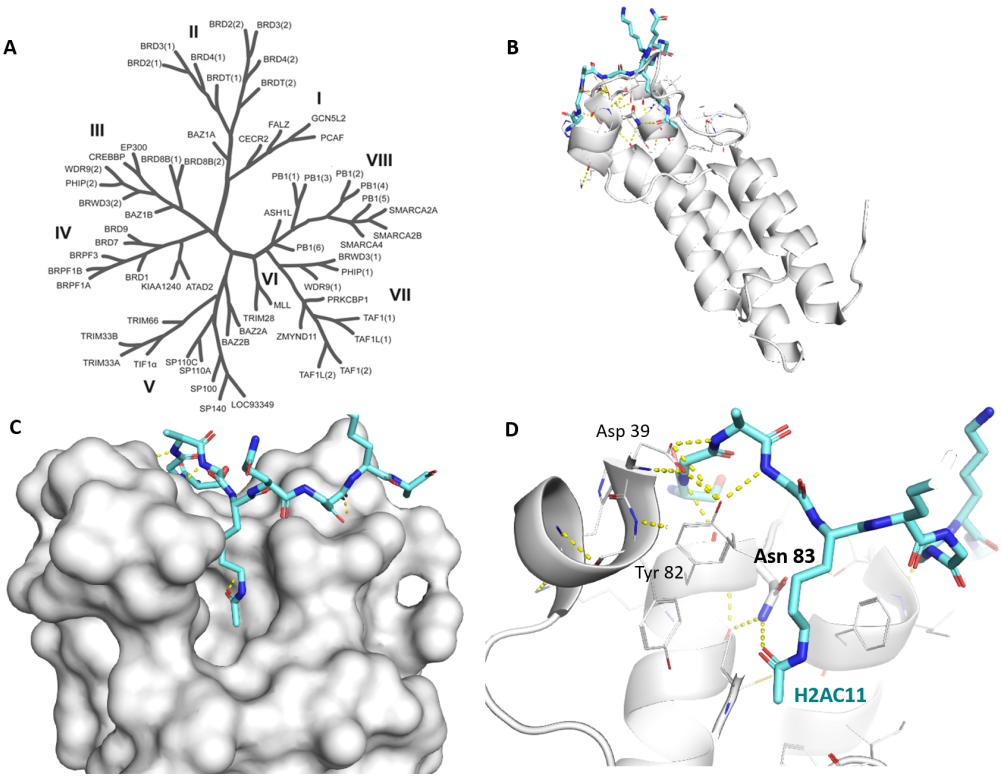


Figure 2.1: Structural features of human bromodomain proteins. A. Bromodomain family tree, adapted from [138]. B. Human BRPF1 in complex with an acetylated histone peptide (PDB ID 4QYL). The protein is shown as a cartoon, and the peptide - as cyan sticks. C. Surface representation of the acetyl-lysine binding pocket of human BRPF1 (PDB ID 4QYL). D. Bromodomain binding site interactions. The same view of 4QYL is used as in C. The key asparagine residue (Asn 83) is shown grey sticks. The interacting peptide is in cyan. Residues along the edge of the binding pocket that make polar contacts with the histone peptide are shown as lines.

protein [153]. They act as regulators in many key cellular signal transduction pathways, which has led to significant pharmaceutical interest in the design of kinase inhibitors [154].

Protein kinases have a characteristic fold (shown in Figure 2.2, Panel A), consisting of a C-terminal domain (made up of α-helices) and an N-terminal domain (consisting mostly of β-sheets), connected by a "hinge", which binds the ATP molecule through a characteristic H-bond pattern [153].

Kinase domains show great flexibility when moving between their active and inactive forms. In the active conformation, a key 3-residue motif (DFG) faces towards the ATP binding pocket (DFG-in conformation, shown in Figure 2.2, panel B), while in the inactive form, it flips to face 'out', opening up a large back portion of the binding site(Figure 2.2, panel C). Inhibitors that target the the active DFG-in conformation are referred to as "Type I" kinase inhibitors, while Type II inhibitors target the inactive DFG-out conformation [154]. Targeting the DFG-out conformation has led to the development of highly specific inhibitors, as the back pocket that becomes accessible in the DFG-out conformation is less conserved between family members.

Another determinant of structural specificity is the so-called "Gatekeeper" residue, located between the hinge and the DFG motif. The size of the Gatekeeper residue controls the accessibility of a hydrophobic back region. As this region is not accessible in all kinases, developing inhibitors that make favourable or unfavourable interactions with the target gatekeeper has been a successful strategy in the design of selective compounds [155].

Other key structural features in the kinase ATP binding site include the glycine-rich loop, a highly flexible region that may undergo ligand-induced conformational changes, and the α-C helix of the N-terminal domain, where a conserved

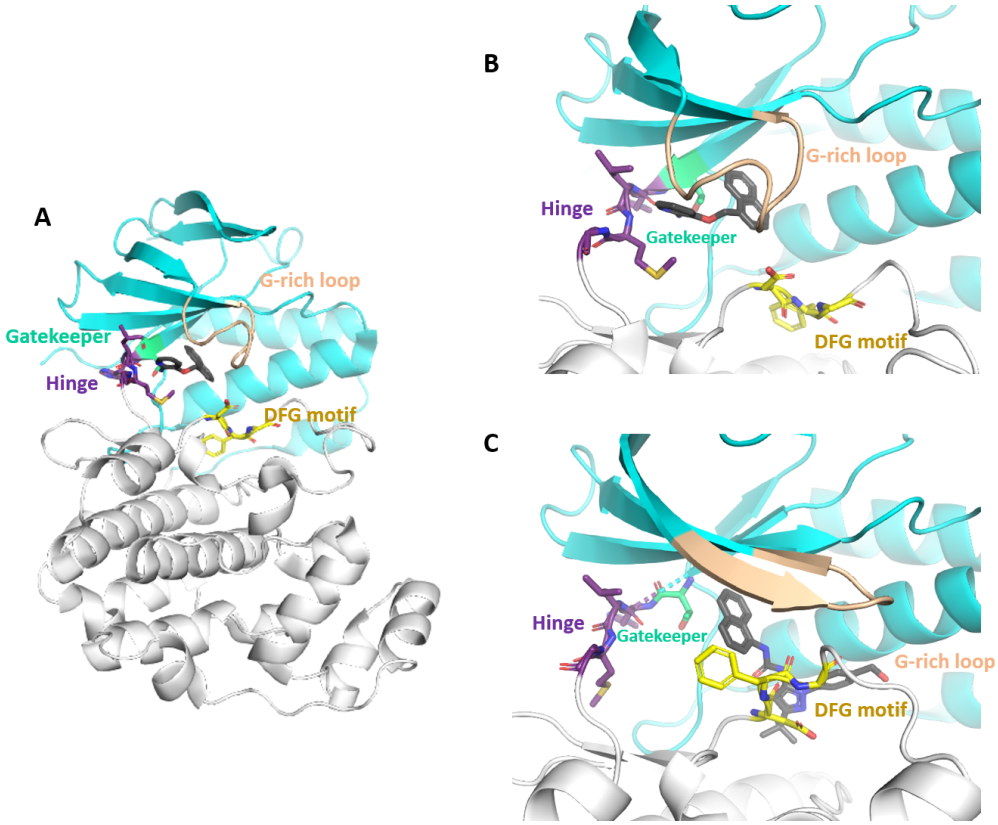


Figure 2.2: Structural features of protein kinases. A. Structure of the p38α kinase in the DFG-IN state (PDB ID 1WBW). The N-terminal lobe is shown in cyan, and the C-terminal - in white. Important structural features discussed in the text are highlighted. B. The DFG-IN conformation. Close up of the binding site of the p38α kinase (PDB ID 1WBW). The ligand is shown as black sticks and overlaps with the adenine binding site. C. The DFG-OUT conformation. The structure of the p38α kinase in complex with a Type II inhibitor (PDB ID 3NNV) shows significant rearrangements in the region of the DFG motif. Conformational flexibility is also displayed by the G-rich loop.

lysine residue forms a salt bridge with the conserved glutamate residue in the DFG motif [153].

2.2 Methods

Figure 2.3 shows the full workflow for generating ensemble and selectivity hotspot maps. After curating the ensemble data, the selected protein structures are aligned in the region of the chosen binding site and prepared for the hotspot maps calculation. Ensemble and selectivity maps are then calculated, as described below.

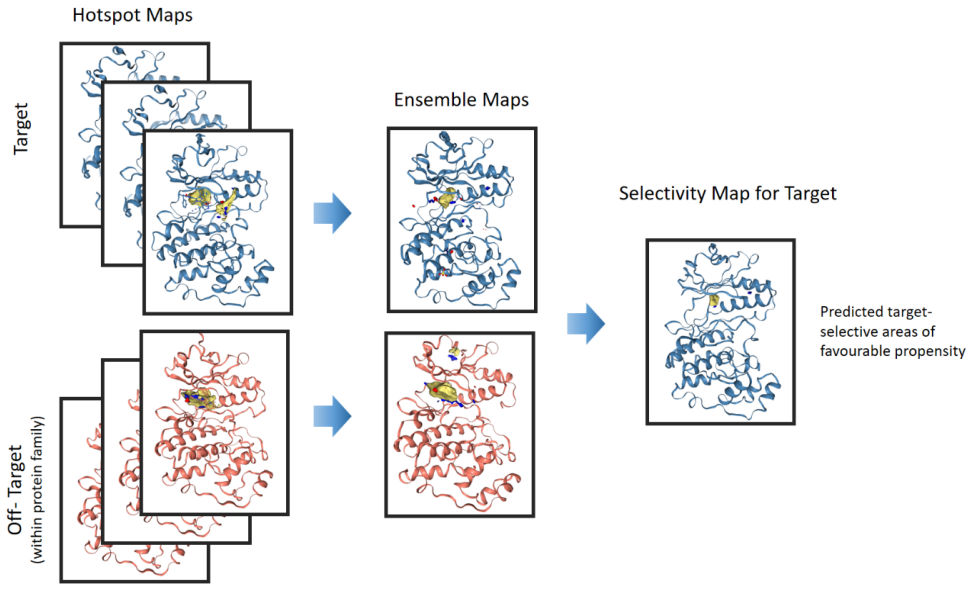


Figure 2.3: Workflow for generating ensemble and selectivity hotspot maps. Protein structures in the ensemble are aligned in the region of the binding site using the CSD Python API, and ligands, metals and waters are removed. Hotspot maps for each structure are then calculated and combined into an ensemble map. The ensemble maps for an on and off-target can then be compared, highlighting areas predicting target-selective interactions. The colour coding is red for the hydrogen bond acceptor maps, blue - for hydrogen bond donors, and yellow for apolar. This colour coding is consistent throughout the thesis.

2.2.1 Data curation and structure preparation

Human bromodomain crystallographic structures were collected using the SIENA tool as made available through the ProteinsPlus webserver [156]. SIENA is an automated workflow for the pre-processing and assembly of ensembles of protein binding site structures from the PDB. In addition to filtering and standardising the input structure annotation (for example, different residue labelling between structures of the same protein), SIENA provides high-quality structural alignments in the region of the binding site [156] through the ASCONA algorithm [157]. It works by fragmenting the query binding site structure into short peptide fragments. Then, a sequence alignment procedure is used to find matching fragments in a set of target binding sites. In the last step, the spatial orientation of all identified fragments is analysed and the target binding sites are reconstructed [157]. The tool also provides automated ways to rationally reduce the size of ensembles and remove redundant structures. This is an important step, as the false positive rate has been shown to increase with ensemble size in docking studies [63]. Methods for clustering ensemble structures based on backbone conformation, as well as by protein-ligand interaction patterns are available as part of the workflow.

The tool was queried through its RESTful API; the full query parameters are presented in table 2.1. For the majority of these parameters, the default values were used as detailed in the original publication [156]. To ensure that only high quality models were used, a resolution cutoff of 2.5 Å was placed, as well as a requirement for all residues in the binding site to be completely modelled. The latter point is crucial, as structures with missing residues can have very different hotspot profiles. To ensure consistency between the ensemble structures, as well as to mimic a crystallographic fragment screening campaign, where all proteins will have same sequence, no mutations were allowed in the binding site (minimalSiteIdentity=1

in Table 2.1). The structures were protonated using the Protoss web server [158].

Table 2.1: SIENA query parameters. All parameters not included were kept at their default values, as detailed in [156]. The fragment_length, flexibility_sensitivity, and fragment_distance parameters control the alignment stage of the SIENA workflow. These values are recommended for use in binding site flexibility analysis by the method creators.

SIENA parameter	**Value**
SiteRadius	6.5
fragment_length	10
flexibility_sensitivity	0.6
fragment_distance	4
minimalSiteIdentity	1
resolution	2.5
complete_residues_only	True

Human kinase structures were downloaded from the KLIFS ("Kinase-Ligand Interaction Fingerprints and Structures") database, accessed in September 2020 [153]. KLIFS contains manually curated, sanitised and aligned kinase-inhibitor structures from the PDB, which are annotated with a large number of kinase-specific descriptors, such as DFG motif orientation, positions and angles of the αC helix and G-rich loops, *etc.* Results can also be filtered by ligand and crystallographic structure properties, as well as interaction fingerprints, and the presence of conserved waters or allosteric ligands. KLIFS also assigns structures a KLIFS quality score, which is a structure quality metric running from 0 (bad) to 10 (flawless) and is based on the number of missing residues and atoms in the binding site, as well as the z-scores of the binding site RMSD [153]. This vast amount of detailed structural information allows for fine-grained control over the selection of ensemble structures. The database can be queried via a RESTful API, allowing for programmatic access [153].

Only structures in the DFG-in conformation with resolution higher than 2.5 Å

and a KLIFS quality score of 7 or above were included in the analysis. No restrictions were placed on the position of the glycine-rich loop or αC helix. All included structures were ligand-bound, with ligands in the main ATP-binding pocket. Structures with allosteric binders were discarded, as were those with mutations in the binding site sequence. In cases with duplicate ligands (based on the ligand canonical SMILES), the highest resolution structure was kept. Structures downloaded from KLIFS are protonated using the Protoss web server and have had alternative conformations removed [153].

Finally, to mimic a prospective crystallographic fragment screening scenario, only structures with ligands (excluding solvents) with molecular weight below 300 Da were included in both the bromodomain and kinase datasets. The full list of structures used in the case studies is provided in the Appendix tables A.1 - A.6. No unliganded structures were included in the analysis. Waters, ions and solvents were also removed prior to calculating the hotspot maps.

2.2.2 Structure alignment

For each structure, the binding site was defined by taking all residues within 5 Å of the binding site ligand. The union of binding site residues from all protein structures within the ensemble then gave the ensemble binding site. The CSD Python API (version 3.0.4) [110] was used to align the ensemble based on the ensemble binding site, using only the Cα atoms.

2.2.3 Fragment hotpot maps

Fragment hotspot maps were calculated using the default parameters as previously described [113, 48]. Release 1.0.5 of the fragment hotspot mapping package (available on PyPI, as well as on the Github repository: `https://github.`

com/prcurran/hotspots) was used. The fragment-like probes sampling was set to 3000 rotations, as recommended in [113]. This value offers sufficient thoroughness of the sampling, while still retaining the fast calculation speed of the hotspot maps. The default 7-atom probes described by Radoux *et al.* were used [48]. Maps were then truncated to the region of the binding site (within 4 Å of any heavy atom in any of the aligned ensemble ligands).

2.2.4 ChEMBL dataset curation

Compound activities were downloaded from the ChEMBL [159] database, release 29 (July 2021), following the protocol described by Bosc *et al.,* [160] with the following modifications. All activities recorded against human proteins BRD1, BRPF1, BRD2 , BRD4, BRD7, and BRD9 (ChEMBL IDs supplied in Table 2.2) were retrieved. Only bioactivities with a standard relation of "=" and standard_flag = True were considered. Mutant sequences, potential duplicates, or data points with data validity comments were dropped. The assay type was restricted to "B" (binding assays), and data sources to src_id = 1 (scientific literature). Only assays with standard units in nM were included, with standard_type = "IC50" or "Kd". Only entries with ChEMBL quality scores of 9 (human targets flagged as "SINGLE PROTEIN") were included. Activities against the second bromodomains of BRD2 and BRD4, as well as against the BRPF1A isoform were removed using a keyword search in the "assay_description" field. When multiple activity values were reported for a compound/target pair, the lowest (most potent) one was taken. Selectivity ratios were calculated by dividing the standard value for the off-target by that for the on-target. Selectivity ratios were calculated only for activities of the same standard type ("IC50" or "Kd"). Finally, only compounds for which activity values for at least two of the targets and a crystal structure in complex with one of the targets were included in the dataset. Crystal structures were retrieved

from the PDB [111], by querying with the compound InChI.

Table 2.2: ChEMBL IDs for the human bromodomains used

Target name	ChEMBL ID
BRD1	CHEMBL2176774
BRPF1	CHEMBL3132741
BRD2	CHEMBL1293289
BRD4	CHEMBL1163125
BRD7	CHEMBL3085622
BRD9	CHEMBL1163125

2.2.5 Detection of hotspot features

The fragment hotspot maps are sparsely-populated grids, with dense clusters of values in areas where hotspot propensity is detected. To computationally detect these features, the density-based clustering algorithm HDBSCAN [161] is used. It uses the distance of a point to the k^{th} nearest neighbour (denoted as the "core distance") as an initial estimate of density (low density points will have high core distances and vice versa). Points with low density are then filtered out using a metric called the mutual reachability distance between two points. For high density points, this will be equivalent to the Euclidean distance between the points. If any of the two points have core distances larger than the distance between them, the metric is set to the largest core distance. This has the effect of "pushing away" points with low density. The matrix of these distances is then used to create a minimum spanning tree, which is converted into a hierarchy of connected components.

Clusters are extracted from this tree by selecting those that pass a minimal cluster size threshold (user defined) and persist through the tree. The minimal cluster size is the only user-provided parameter used by the algorithm, and has a concrete physical meaning. This, along with the fact that the number of clusters is derived

by the algorithm, rather than provided externally, guided the choice this algorithm for hotspot feature detection. In the polar maps, a value of 7 points was chosen as the equivalent of the smallest spherical element with a radius comparable to that of the polar probes. For the apolar maps, a value of 27 points was used to approximate the volume of a methyl group.

2.2.6 Implementing the ensemble and selectivity maps

The code for generating ensemble and selectivity maps is currently integrated into the Hotspots API [113], within the hs_ensembles.py module. The hotspots API is freely available and can be found at the following locations: (https://github.com/prcurran/hotspots,

https://github.com/ccdc-opensource/hotspots/tree/master, but it is based on the CSD Python API [110], which requires a licence.

The main object in the Hotspots API is the Hotspot Result (Figure 2.4), which holds the grid data for the hotspot maps, the input protein as a ccdc.protein.Protein object, grid data corresponding to the buriedness of the the point, and (optionally) the SuperStar maps used to calculate the hotspot maps. This class also acts as input to the rest of the hotspots API functionality, as has been described elsewhere [113]. The ensemble and selectivity maps calculations are handled by the EnsembleResult and SelectivityResult classes, respectively. To integrate into the API, these classes take HotspotResult objects as input(a Python list of HotspotResult objects, in the case of the EnsembleResult), perform calculations internally, and output ensemble and selectivity maps as HotspotResult objects. This means that selectivity maps can be generated from any kind of HotspotResult (both containing ensemble or individual hotspot maps), and that the ensemble and selectivity maps can feed into the rest of the downstream processing workflows available in

the Hotspots API.

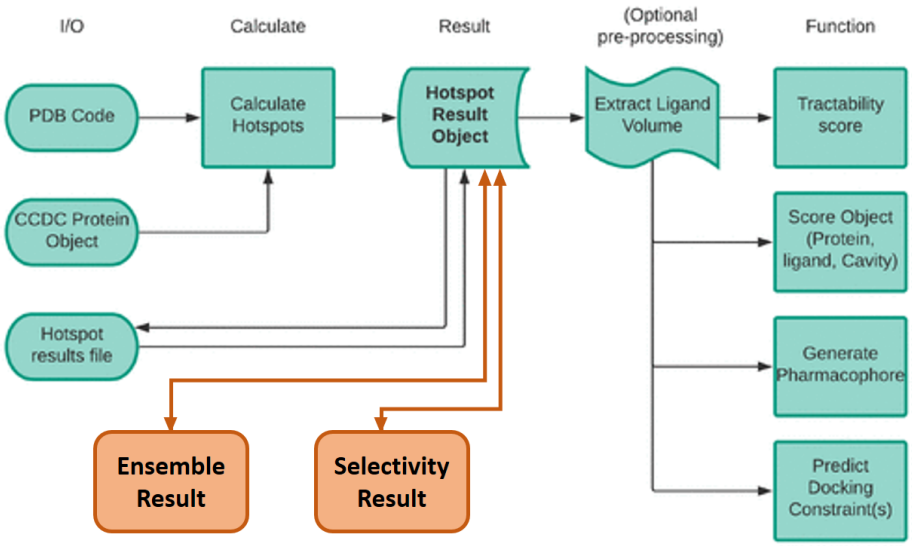


Figure 2.4: Integrating the ensemble and selectivity maps into the Hotspots API. The existing Hotspots API modules and functionality, as presented in Curran *et al.,* 2020 [113], are shown as green shapes. The ensemble and selectivity hotspot maps are calculated by dedicated classes that run calculations internally (shown in orange), and integrate with the other module through the HotspotResult class. This image is modified from Curran *et al.,* 2020 [113].

The ensemble and selectivity map parameters are exposed to the user through nested "Settings" classes. This organisation is consistent with other modules within the Hotspots and CSD Python API, and can be used as shown below.

```
ensemble_settings= EnsembleResult.Settings()

ensemble_settings.combine_mode = 'median'

ensemble_settings.polar_frequency_threshold = 20.0
```

The Settings object is then passed to the constructor of the EnsembleResult object.

```
ensemble = EnsembleResult(hotspot_results_list,
ensemble_id='on_target',
```

```
settings=ensemble_settings)
```

If a Settings object is not explicitly provided, the default values for its parameters are used, as shown in Figure 2.9 (ensemble maps) and Figure 2.12 (selectivity maps) later in this chapter. Ensemble map parameters exposed through the EnsembleResult.Settings class are combine_mode, which controls the way in which ensemble values at each point in space are combined. Currently, this can be set to "max", "median", or "mean". The frequency cutoffs described earlier can be changed through the polar_frequency_threshold and apolar_frequency_threshold attributes of the Settings class.

The EnsembleResult object takes the input Python list of HotspotResult objects and transforms the grid information into Numpy [162] arrays. First, all grids are converted to the same size and coordinates, which are stored by the EnsembleResult object. For each probe type, the information for a single structure is then converted to a 3D Numpy array, and arrays for the same probe type across the ensemble are stacked into a 4D Numpy array, on which further calculations are performed. The Numpy operations are handled by a private class within the hotspots.grid_extension.py module (grid_extension._GridEnsemble).

In the selectivity maps, the SelectivityResult.Settings class exposes parameters that control the minimal median score needed for a cluster to be considered selective (minimal_cluster_score), the minimal distance between the centroids of selective clusters in the target and off-target maps (cluster_distance_cutoff), the minimum points parameter for the HDBSCAN clustering (min_points_cluster_polar, min_points_cluster_apolar), and the percentile of scores at which to threshold the difference maps (apolar_percentile_threshold, with a default value of 95, and polar_percentile_threshold, with a default value of zero, which takes all points in the difference maps).

This interface provides the user with great flexibility and control over the ensemble and selectivity maps calculations, as different target classes and ensemble sizes may require modifications to the default values. Recommendations on selecting these values in a prospective scenario are provided in Section 2.3.5. Scripts showing the application of the ensemble and selectivity maps to the retrospective case studies presented later in this chapter can be found at: https://github.com/CMD-Oxford/hotspotEnsembles.

2.2.7 Interactively visualising hotspots through the NGL viewer

The hotspots API currently includes scripts that generate PyMOL visualisations of the input protein and the three hotspot maps, which have been developed by Peter Curran and Christopher Radoux. To help with integrating hotspot maps data into the Fragalysis platform's visualisations, I worked on developing an interactive web-based visualisation using the NGL viewer [43]. This is a web application for molecular visualisation, written in Javascript. It is very lightweight compared to local molecular viewers and can be used both in a stand-alone way, as well as embedded into other web applications. A very simple visualisation, in which the hotspot maps are loaded as isosurfaces coloured by probe type was introduced (Figure 2.5, panel A). A set of three sliders control the contour level at which the maps are displayed (in hotspot score units). It is also possible to visualise clusters, as shown in Figure 2.5, panel B. Points in each cluster are shown using a "spheres" representation. Darker colours indicate a higher cluster number, which is also displayed when hovering the mouse on top of the spheres. The code for this visualisation is not yet integrated into the Hotspots API, but can be found at https://github.com/xchem/hotspots/tree/master/ngl_viewer.

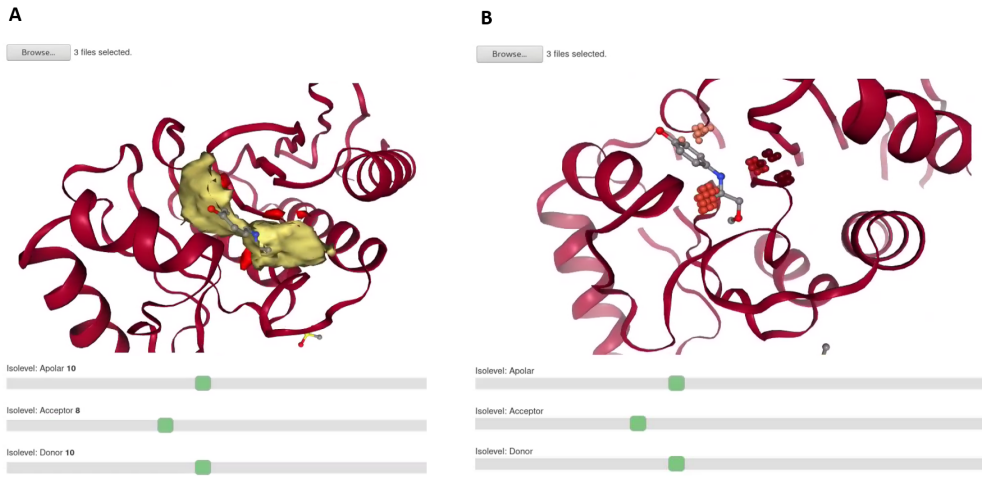


Figure 2.5: Visualising fragment hotspot maps in the NGL viewer. A. Hotspot maps for PARP14 in complex with fragment x0315. The ligand was removed prior to the hotspot maps calculation. The protein is shown using the default cartoon representation, ligands are shown as "ball + stick". The three sliders under the NGL window control the maps' contour level. The colour coding is red for the hydrogen bond acceptor maps, blue - for hydrogen bond donors, and yellow for apolar. B. Acceptor clusters for PARP14. The features are colour-coded according to the number of the cluster detected in the acceptor hotpot maps. Darker reds indicate a higher cluster number (for the donor maps, this is in blue).

2.3 Results and Discussion

2.3.1 Development of an ensemble hotspot map

2.3.1.1 Understanding the ensemble hotspot map data

For a single protein structure, the fragment hotspots algorithm outputs three maps: one for each interaction probe type (donor, acceptor, and apolar). Each of these maps is a 3-dimensional grid with a spacing of 0.5 Å. Consequently, for each point in space, there are three associated hotspot values: one from each probe "channel" (shown in Figure 2.6). When hotspot maps are calculated on an ensemble of pre-aligned structures, each point in space receives hotspot scores from all of the

structures in the ensemble (Figure 2.6, panels B and C).

As an example, the donor map values associated with a single point in an ensemble fragment-bound structures have been plotted as a histogram in Figure 2.6, panel C. It is important to consider the zero values in these histograms - they are derived from structures in which there is no hotspot signal for the particular probe type at that point. Such variations in hotspot score can often be linked back to local movements in the binding site. Recording this negative information allows for a more detailed view into how often a point has an associated nonzero value in the maps, as well as which structures contribute to the hotspot signal at a particular point in space.

Points that consistently score highly in all ensemble structures are naturally of interest in the context of compound design. These indicate areas where an environment favoring a particular fragment probe is available in the majority of the complex structures. The interaction is expected to be readily available for ligand binding, which, combined with a high hotspot score (indicating an interaction expected to contribute disproportionately to fragment binding) makes it an attractive avenue for the design of follow-up compounds predicted to make this interaction. Points that "persist" through structures of complexes with chemically diverse ligands are particularly interesting, as they show that the interaction favours binding to diverse chemotypes, confirming its significance for ligand binding, as well as opens the possibility of changing the scaffold.

Points that show variation in their hotspot scores across the ensemble may be located at the "edge" of a hotspot feature, and so show variation due to the small local motions and imperfections in the alignment. Alternatively, they may be part of a hotspot feature that is present in a subset of conformations. Such features can be interesting from a ligand design perspective, as they may provide ways

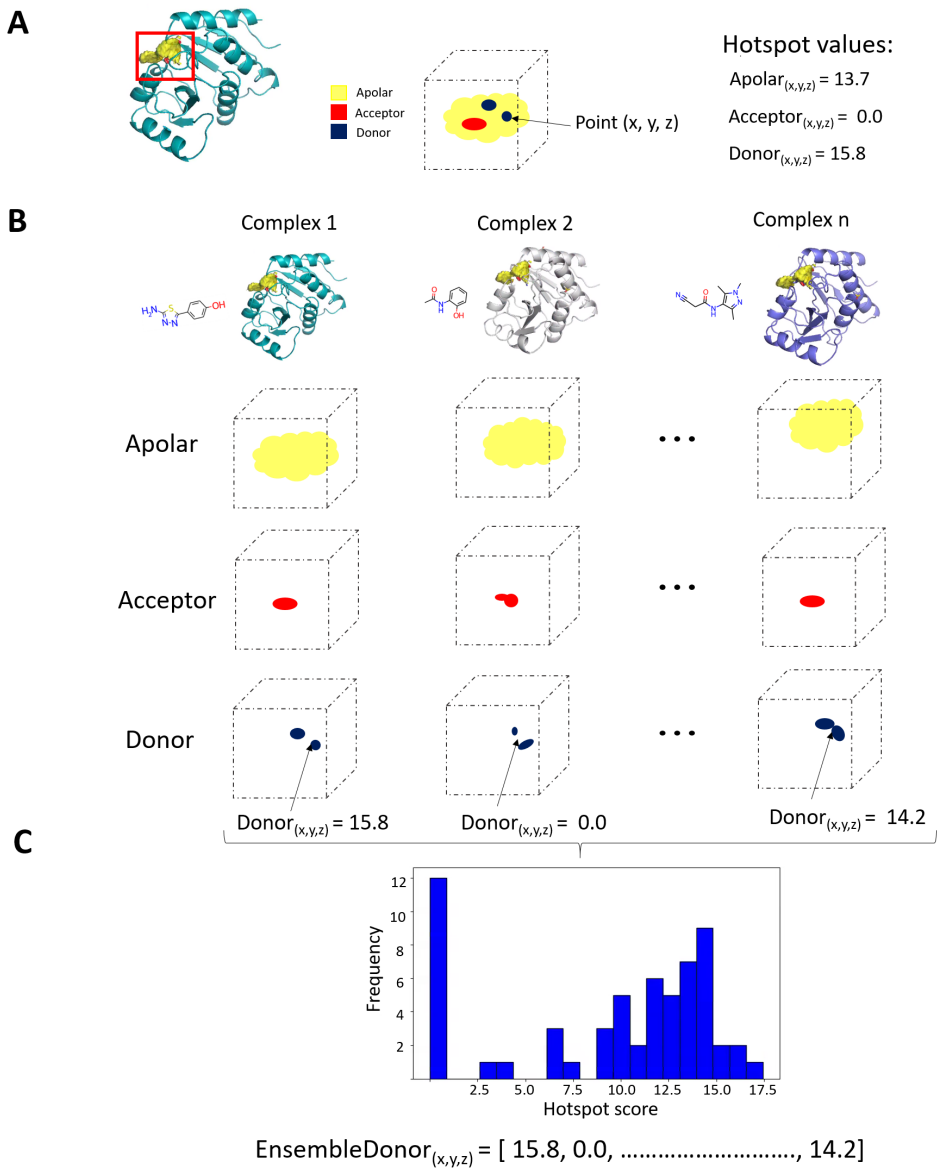


Figure 2.6: Understanding the ensemble maps data. PARP14 is shown as an example target. A. Hotspot maps are calculated for three probe types: donor, acceptor, and apolar. Each point in space has an associated value in each map, which may be zero. A "hotspot" is referred to as the whole area of hotspot density, while "hotspot feature" refers to individual clusters of points within the maps. B. Calculating hotspot maps for a pre-aligned ensemble of structures puts the hotspot maps in the same reference frame to each other. C. The hotspot values from the map of each complex structure (shown here for the donor maps) at a point form the ensemble hotspot values for that point. The histogram shows their distribution for a particular point in one of the donor hotspot features.

of achieving selectivity (for example, if this conformation is inaccessible in a related off-target protein), or novel strategies for inhibition. However, while the hotspot maps indicate that such conformations create an environment that favours fragment binding, they cannot inform on the stability of the conformation, its accessibility to ligand binding, or the stability of the protein-ligand complex once the interaction is formed. In such cases, other computational methods, for example based on molecular dynamics, would be needed to support the hypothesis.

To summarise data in the ensemble that is considered interesting from a ligand design perspective, the ensemble maps then need to be able to both indicate features that are present in the majority of ensemble structures, as well as highlight unusual conformations that lead to transient, but highly scoring hotspot regions. The ensemble maps should also retain the link between a point or a hotspot feature and the structure it originated from.

2.3.1.2 Combining ensemble hotspot information into a single probe map

The generation of hotspot maps with an added layer of information recording hotspot features through the ensemble also presents a challenge for the visualisation and downstream processing of the data. One of the key advantages of the fragment hotspot maps method is the ease with which they can be interpreted by the human eye. Adding an extra layer of information that tracks the 'frequency' with which point appears in the ensemble through, for example, the shade or intensity of the colour, can be distracting. In terms of computational processing, by storing all the values at each point along the ensemble, the maps themselves become 4-dimensional arrays. This causes the computational space used when compiling ensemble maps to increase rapidly ($O(n^3)$). While there are possible workarounds, such as truncating the maps in the region of the binding site, or using sparse matrix formats, the initial focus of this work was to condense the

hotspot information from each probe type in the ensemble to a single value at a point in space.

Initially, two simple ways of combining these values were investigated: taking the median and taking the maximum of the ensemble values at a point. Taking the maximum value is equivalent to a "best case" scenario in the sense that any high scoring interaction observed in the ensemble could be potentially useful for ligand design. This approach presented two major issues to the interpretation and downstream processing of the resulting maps. Firstly, the polar features in the maximum maps cover a much greater volume (hundreds of grid points, whereas the individual polar features are usually in the tens of grid points), as a result of protein plasticity. This can lead to adjacent features merging, creating difficulties in both the visual and computational processing of the maps. Second, the maximum maps lose information about the frequency with which a point is sampled, so information about the persistence of the feature is lost. An important part of the ensemble information is what fraction of the structures are in a particular conformation. A maximum map will give the same weight to a feature that arises from an unusual side chain motion as it would to a persistent hotspot feature that is present in all of the ensemble structures. Both of these kinds of features are interesting from a compound design perspective, but it is important to be able to differentiate between them, as discussed in the previous section 2.3.1.1.

The hotspot scores at each point are not normally distributed, and often feature outliers, as can be seen in the histogram in Figure 2.7. Mean values are more easily skewed by outliers; in cases were zero values are present in the distribution of values at point (x, y, z), the resulting mean score could be artificially lowered, even though the majority of points are more highly scoring. Conversely, a high scoring outlier could result in a false positive value, giving the impression that a cluster is consistently highly scoring. This is why the median, rather than the

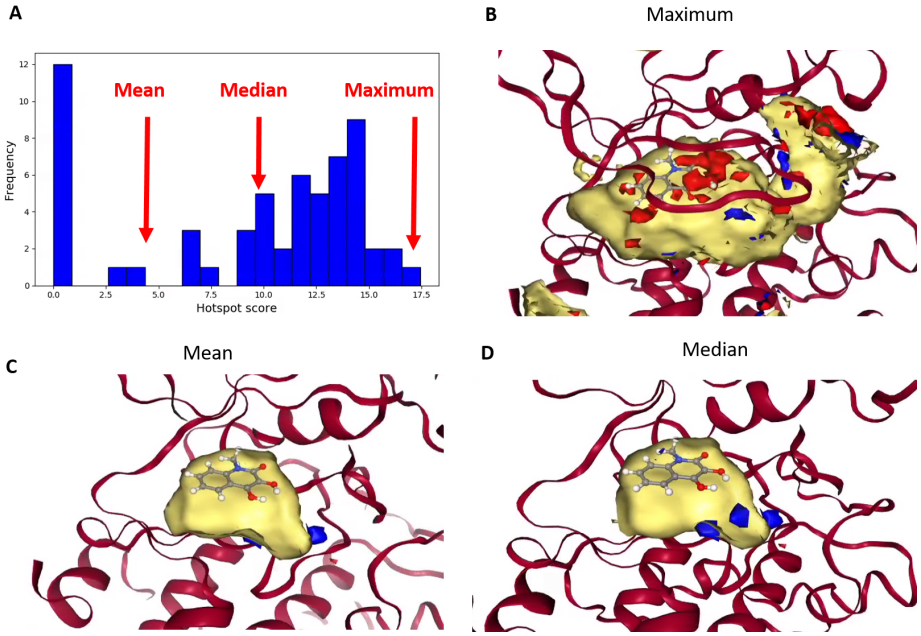


Figure 2.7: Choosing a way of combining the information of multiple hotspot maps. A. The distribution of ensemble hotspot values at a point in the donor maps. The mean, median and maximum values of the distribution are indicated. B. Ensemble maps generated by taking the maximum of the distribution at each point. C. Ensemble maps generated by taking the mean of the distribution at each point. D. Ensemble maps generated by taking the median of the distribution at each point. The protein target is PIM1, PDB ID of the displayed model is 1YXV. The colour coding is red for the hydrogen bond acceptor maps, blue - for hydrogen bond donors, and yellow for apolar.

mean of these values was chosen as a single hotspot value summarising all the point contributions. In cases where a hotspot feature is consistently present in the ensemble, the median is less likely to be influenced by rare conformations. More sophisticated statistical descriptors of the distributions of hotspot values at a point were not considered, as the shape of the distributions is irregular, and the number of data points generally low (tens to hundreds of structures in an ensemble), resulting in low statistical power. The resulting median maps had the opposite problem to the maximum hotspot maps: while the maximum maps were prone to overestimating the importance of individual features, the median scores for points

that have been sampled in less than 50 % will be zero. In ensembles of proteins with greater binding site flexibility, such as the CK2α kinase, this can lead to key hotspot features, such as the hinge acceptor pharmacophore, to be missed (see Figure 2.8).

2.3.1.3 Introducing a point frequency cutoff for the ensemble maps

With the shortcomings of both the maximum and median hotspot maps in mind, it was evident that a better way of combining the ensemble hotspot values was needed. This method would need to be sufficiently sensitive to the information from polar hotspots, while avoiding the introduction of false positives in the downstream analysis.

The first step was to track the 'frequency' with which each point is sampled in the ensemble. This frequency is defined as the fraction of cases (structures in the ensemble) in which that point in space has a score assigned in the maps for a particular probe type (see Figure 2.9. For example, points with over 20 % frequency have at least 20 % of the hotspot values at that position greater than zero. Figure 2.8 shows the effect of thresholding the maps at different frequencies and taking the median of all nonzero points. Three different structural ensembles are presented: one for the bromodomain BRD1, which is large a (n = 23 structures) but structurally rigid ensemble, the kinase p38α, which is smaller (n = 5) and with greater flexibility in the binding site, and one for the kinase target CK2α, which is both a large and flexible ensemble (n = 28). Column A in Figure 2.8 shows the result of calculating the median of all non-zero values at each point in the ensemble. Setting the frequency to this level (at least 1 non-zero value) has similar issues to those associated with the maximum ensemble maps, as all sampled points in the ensemble are represented in the final map. For the bromodomain example, this approach does not introduce additional features compared to the more restrictive

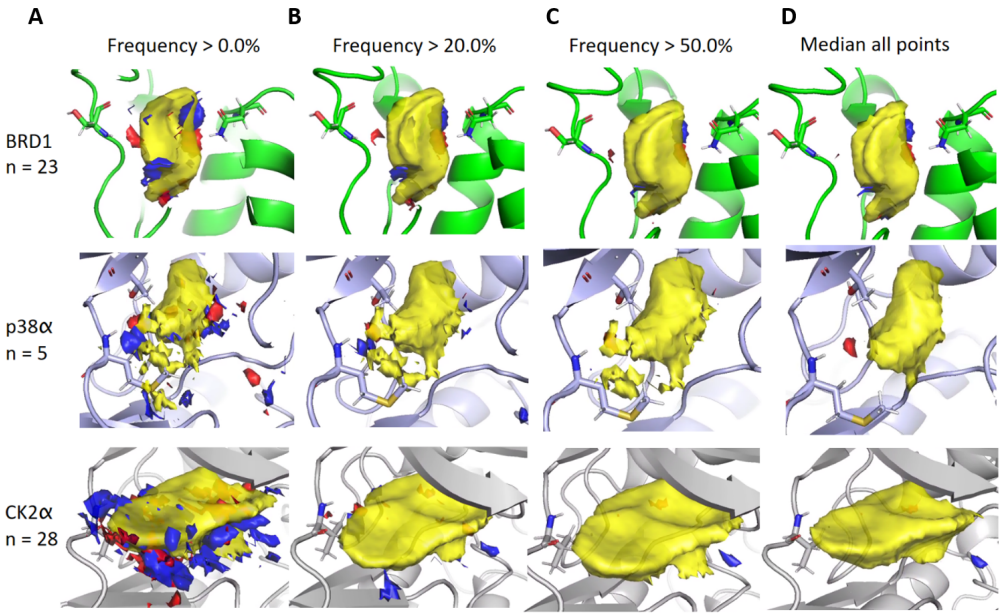


Figure 2.8: Setting the frequency threshold parameter for the ensemble maps. Ensemble maps are compiled by taking the median of values samples at a grid point for all points that have nonzero values in at least the threshold fraction of structures in the ensemble. The first three columns show the effect of setting this parameter to 0% (all points sampled at least once, column A), 20.0 % (column B) and 50.0 % (column C) for the respective protein ensembles. The column D shows the median maps for the ensemble when zero values are also included. The colour coding is red for the hydrogen bond acceptor maps, blue - for hydrogen bond donors, and yellow for apolar.

cutoffs, but the polar features show an increase in volume. In the p38α example, additional polar features are detected compared to the more restrictive cutoffs, due to the plasticity of the binding site. The presence of these features does carry useful information, but as they have only been observed in a single structure, further experiments would be needed to support the design of compounds that target these features. In the case of CK2α, the polar features have merged and there is a dramatic difference between the pattern of polar features at the 0 % (column A, bottom panel) and 20 % (column B, bottom panel) frequency cutoffs.

Conversely, setting the frequency threshold to points that have been sampled in

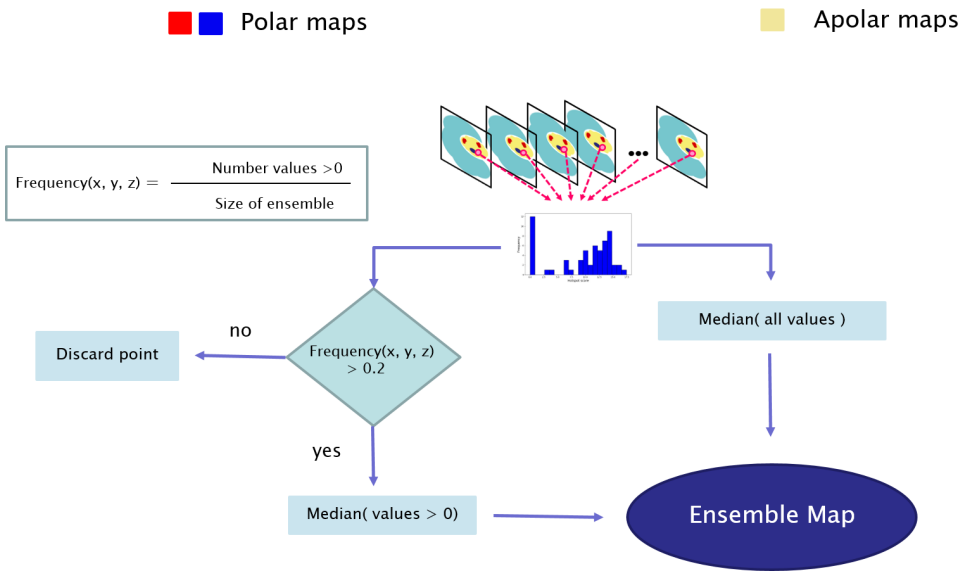


Figure 2.9: Methodology for generating ensemble hotspot maps. The left side of the flowchart shows the procedure for compiling the polar maps (colour coded in red and blue, as throughout the thesis). The right side shows the procedure for compiling information in the apolar maps

at least 50 % of cases or more has the same problems as taking the median of all the values, including zeros (column D in Figure 2.8), in that key pharmacophoric features (such as the kinase hinge acceptor pharmacophore) can be missed. In the case of the apolar maps, apolar hotspots tend to be significantly larger (thousands of grid points in a cluster, compared to fewer than a hundred in polar clusters). This means that points are more likely to have associated values in most of the structures and individual zero values have less of an effect. In Figure 2.8, the apolar maps are visibly less affected by the frequency thresholds, compared to the dramatic differences observed in the polar cases. For this reason, the frequency cutoff values were only used for the polar maps, with a 20 % frequency cutoff as a recommended starting point when calculating maps on a new target ensemble. The full final ensemble maps generation workflow is summarised in Figure 2.9.

2.3.1.4 Automating feature detection in the ensemble maps

The workflow presented in Figure 2.9 creates maps that summarise the key binding site interaction information across an ensemble of protein structures and present it in a visually intuitive way. However, one of the key strengths of the fragment hotspots method is the quantitative nature of the maps, which allows for automated feature processing and prioritisation. The ensemble maps can be used to answer two main questions arising from a successful crystallographic fragment screen: what are the key interactions in the binding site, and which protein-fragment complexes exploit these interactions. The first question can be answered by loading the ensemble maps in an interactive molecule viewer and inspecting the hotspot features. The automated computational equivalent of this involves locating clusters of points in the maps. This also makes answering the second question (which complexes exploit which interactions) easier to answer computationally.

The ensemble maps are sparse grids with occasional areas of high density, which correspond to hotspot features. The simplest way to cluster these features is to use an island finding algorithm, which locates continuous groups of points. Hotspot and ensemble map features, however, are not always continuous islands - as can be seen in panel D in Figure 2.10.

The desired behaviour in such cases is either for the disconnected points to be assigned to the nearest cluster, or if their cluster affiliation cannot be easily determined, to be flagged as noise. This is especially important when detecting features in the ensemble and selectivity maps, as they can be noisier than the hotspot maps for an individual structure. For this reason, a more sophisticated clustering algorithm was needed. HDBSCAN [161] is a recently published density-based clustering algorithm, which has received wide application in similar tasks. It differentiates between points that likely belong to a hotspot feature (areas of high

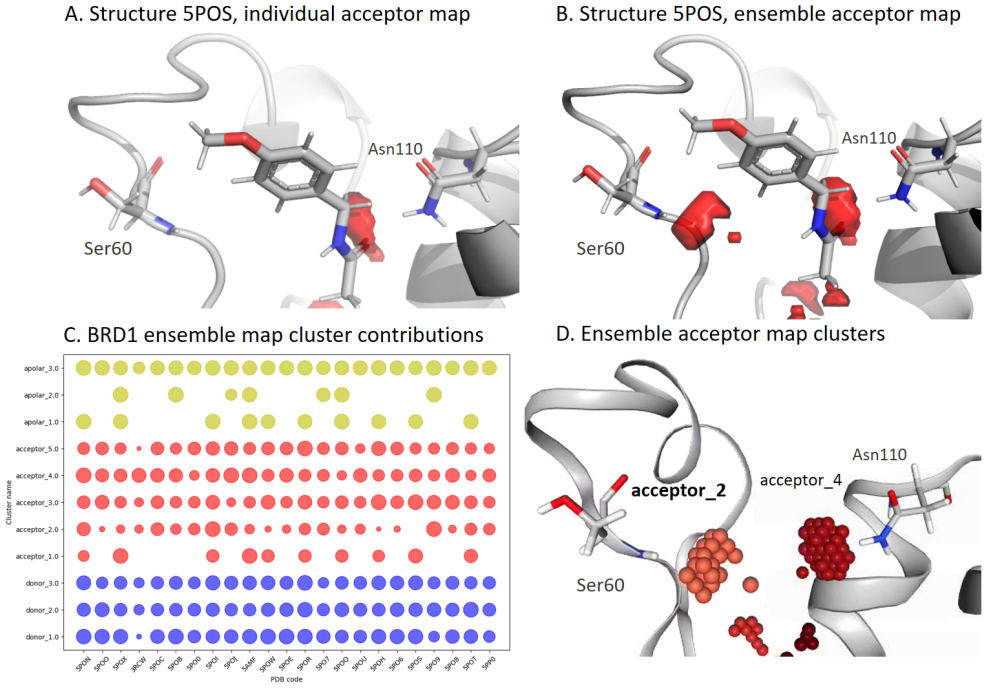


Figure 2.10: Detecting clusters in the ensemble maps. A. Shown as a red surface is the acceptor hotspot map for BRD1 structure 5PO1 at a threshold of 5 hotspot units. This map identifies the acceptor hotspot arising from the asparagine, but not the additional hotspot that corresponds to the serine. B. The ensemble acceptor map (threshold 5) identifies the feature next to Ser 60. C. The ensemble maps were clustered using the HDBSCAN algorithm.The number of points each hotspot map contributes to the ensemble clusters was then plotted. Contributions are represented by circles coloured by probe type. The radius of the circle is proportional to the number of grid points shared between the individual hotspot map and the ensemble map. D. Acceptor_2 corresponds to the Ser 60 hotspot cluster. The plot in section C shows that structure 5POS does not contribute any points toward it.

point density in the maps), and those that are likely part of the noise (areas of low point density in the maps). HDBSCAN does not require any initial estimate on the number of clusters to identify, merely requiring a single clustering parameter: the minimal number of points in a cluster point's vicinity. Here, 7 points for polar maps, and 27 points for apolar maps was used. A value of 7 is equivalent to the smallest spherical element in a voxelised grid with a radius comparable to that of the polar probe atoms (oxygen and nitrogen), and a value of 27 approximates a methyl group as the minimal apolar feature.

Figure 2.10 shows the utility of using a clustering approach to identify the contributions of individual protein-fragment complexes to the ensemble maps. In the case of BRD1, a key backbone acceptor interaction near Ser 60 may not be present in some of the individual ensemble maps, likely due to subtle twisting of the backbone (PDB ID 5POS is one such case, as shown in panel A). The ensemble maps, however, identify this interaction (panel B). A plot such as the one in panel C, which identifies how many points each individual acceptor map places in the cluster, can then be a useful tool to investigate the behaviour of that feature in the ensemble and inform what types of fragments are likely to exploit that interaction.

2.3.2 Developing the hotspot selectivity maps

In addition to obtaining potency, a successful lead compound should exhibit a desired level of selectivity within the target protein family. Comparing the binding sites of protein family members is a way to guide this process. After having developed a way to summarise information across an ensemble of protein-fragment structures, the next step was to extend the method with a workflow for comparing the ensemble maps of closely related proteins.

A trivial way to compare two ensemble maps is to subtract them from each other, producing a difference map. Positive values in the difference map show areas of selectivity for the on-target ensemble, and vice-versa. However, simply subtracting the maps from each other resulted in "selective" features in the region of known conserved pharmacophores, such as the kinase hinge or the conserved bromodomain asparagine (Figure 2.11). Some of these issues were caused by imperfections in the alignment, or areas along the edges of a hotspot feature that is slightly larger in one of the maps (See Figure 2.6, section A). Therefore, further post-processing was needed.

2.3.2.1 Separating signal from noise in the polar selectivity maps

The difference maps are sparse grids with occasional clusters of non-zero values. Unlike the individual and ensemble hotspot maps, the values in difference maps can be either positive or negative, reflecting propensity towards either the target or the off-target protein. In the original fragment hotspot maps publication [48], a hotspot score of 14 or above denotes areas of positive propensity for the polar probes. The difference maps have a very different distribution of values, and so would likely need different score cutoffs.

Figure 2.11 shows that even at the highest thresholds (panel A), slivers of positive density can be seen in non-selective regions, such as around the highlighted asparagine residue that is conserved between the target (BRD1) and off-target (BRPF1) proteins. Extending the selection to all the points with a positive value for the target protein, clusters of points become apparent. These clusters can be detected using the HDBSCAN algorithm and the feature detection procedure described in section 2.3.1.4.

Once the clusters have been identified, they can be scored. The expectation is

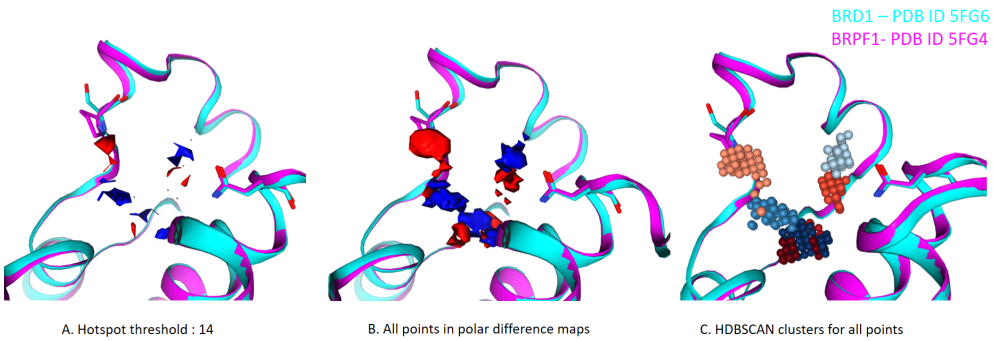


Figure 2.11: Feature detection in the polar selectivity maps. The acceptor (red) and donor (blue) difference maps showing positive propensity for the bromodomain protein BRD1 over the closely related off-target BRPF1. The structures shown are PDB ID 5FG6 (BRD1, in cyan) and 5FG4 (BRPF1, in magenta). The conserved bromodomain asparagine, and a key Ser/Pro substitution in the binding site are shown as sticks. Panel A shows the maps thresholded to the highest hotspot values. Panel B shows all the positive values in the maps, and panel C shows the HDBSCAN-detected clusters based on the points in Panel B.

that clusters with higher scores will be more selective. To score the clusters, the median value of all points belonging to that cluster was chosen. The shapes of the distributions of hotspot values within the clusters do not conform to a particular probability density function, and the number of points (tens to hundreds) is generally too low for more sophisticated statistical methods.

Features that have a great degree of overlap with off-target will have large areas of scores with low values, which drives the cluster median down. Indeed, for the example presented in figure 2.11, the clusters near non-selective areas have median scores between 1 and 6, with the exception of a donor feature close to a conserved bromodomain water, which has an unusually high score of 16. The selective acceptor feature corresponding to the Ser/Pro substitution in the binding site has a median score of 12. In the case of the false positive selective donor feature, an inspection of the reverse selectivity map (areas that confer propensity for the off-target over the on-target) reveals that it is less than 1 Å away from a feature that

is 'selective' for the off-target. Visual inspection revealed that these are likely the same feature, but displaced due to binding site plasticity and imperfect alignment.

A distance cutoff for the selectivity maps was then introduced, as desirable selective clusters need to not only be highly scoring, but also be located away from features of the same probe type that grant propensity for the off-target cluster. In the bromodomain example presented here, this distance cutoff was set to 1.5 Å, as the binding site is narrow and very static, and this value has previously been reported as a distance clustering parameter in BRD1 [40]. Later sections of this chapter will look at the choice of both score and distance cutoffs for targets of different protein classes.

2.3.2.2 Developing selectivity maps for apolar features

In the case of the apolar maps, using HDBSCAN to cluster all positive values in the difference maps failed to provide meaningful clusters, as all areas have similar density. However, as features in the apolar maps tend to be much larger (in the thousands of grid points) and are less dependent on orientation, thresholding the difference maps to values in the top 95^{th} percentile resulted in dense clusters, which could be identified by the feature detection procedure described earlier in the chapter.

Figure 2.12 shows the final selectivity maps generation procedure, which was then validated based on its ability to rationalise observed compound selectivity between members of the same protein family in three example datasets.

2.3.3 Retrospective validation

The ensemble and selectivity maps were envisaged as a tool to guide the elaboration of fragment and lead-like compounds. To validate the ensemble and se-

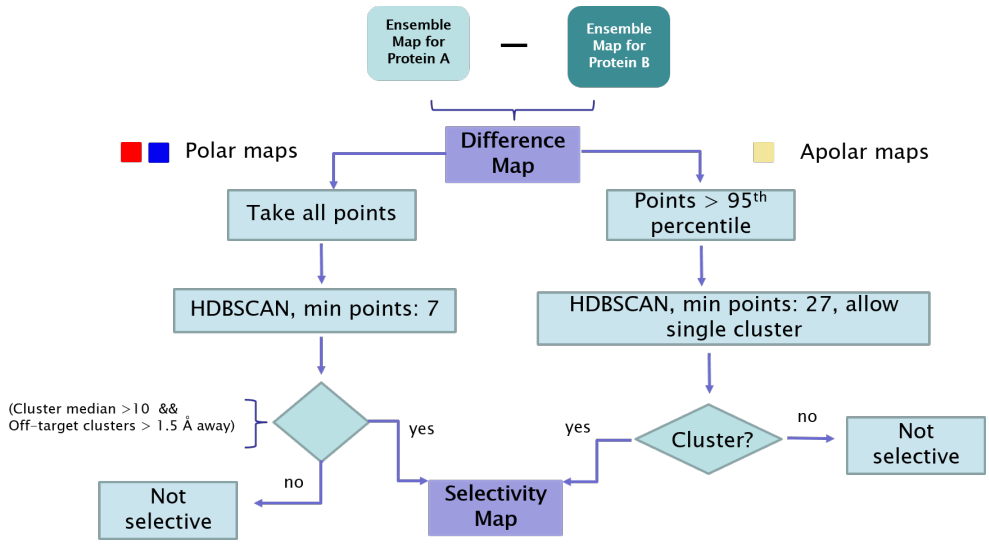


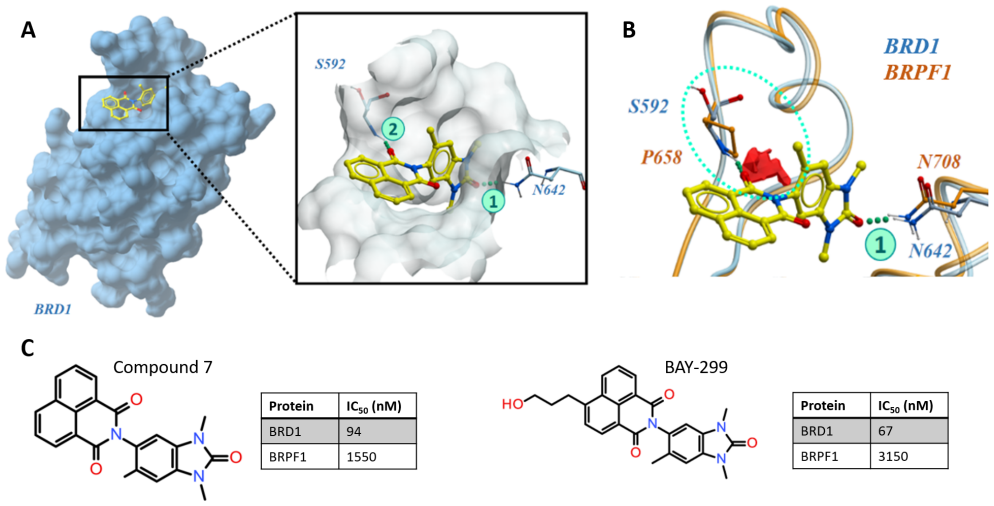
Figure 2.12: Workflow for generating the selectivity maps. The left side of the flowchart shows the procedure for compiling the polar maps (red, blue, corresponding respectively to the donor and acceptor maps), and the right side: for apolar.

lectivity maps generation protocol, it was applied to retrospective examples from well-researched, therapeutically significant human protein families. To mimic the output of a crystallographic fragment screening campaign, only structures with bound ligands smaller than 300 Da were used, as detailed in the Methods section.

2.3.3.1 Selectivity in the human bromodomain BRPF1 subfamily

Bromodomains are a family of proteins involved in epigenetic regulation. They act as readers of histone tail modifications by binding acetylated lysines, usually through a highly conserved asparagine residue in the bromodomain binding site [152]. This protein family has been under active investigation over the past decade as targets for a number of human diseases, most notably certain kinds of cancers [152, 163, 151]. In the human bromodomain BRPF-subfamily, the substitution of a serine in BRD1 to a proline in the closely related BRPF1 (Figure 2.13) is a key binding site difference proximal to the conserved bromodomain

asparagine. The serine backbone can form a hydrogen bond to ligand acceptor atoms within the bromodomain binding pocket, whilst the corresponding proline backbone nitrogen in BRPF1 cannot. Although this substitution can be identified from a sequence alignment and the crystal structures of the two proteins show that the serine backbone is accessible from the binding site, this is not sufficient information to conclude whether this interaction would make a measurable difference on binding affinity and, therefore, whether it can be used as a determinant of selectivity in the design of novel compounds.



Protein	IC_{50} (nM)
BRD1	94
BRPF1	1550

Protein	IC_{50} (nM)
BRD1	67
BRPF1	3150

Figure 2.13: Selectivity in bromodomains: BRD1 over BRPF1. A. Shown is the selective precursor (Compound 7) to the chemical probe BAY-299, which was co-crystallised with BRD1 (PDB ID: 5N49). It makes two hydrogen bonds with the protein in the binding pocket, shown as (1) and (2) in the pop-out. B. In BRPF1 (orange ribbon and sticks) this interaction (2) cannot be made because of the substitution of a proline (P658) at this location. The acceptor selectivity maps for BRD1 over BRPF1 identify this difference, shown by the area of acceptor propensity (red surface). C. 2D structures of Compound 7 and BAY-299. The tables show the compounds' IC_{50} values, measure by TR-FRET assay, for both proteins as reported in Bouché *et al.*, 2017.

To explore whether the ensemble and selectivity hotspot maps can answer this question, ensemble maps for the targets were calculated using 23 BRD1 and 26

BRPF1 fragment-bound structures (the full list of PDB codes can be found in Appendix tables A.1 and A.2). Figure 2.13 shows that the acceptor selectivity maps for BRD1 over BRPF1 identified this difference as a highly scoring cluster in the difference maps.

A review of the literature shows that this feature is, in fact, exploited by the BRD1-selective probe BAY-299, published in 2017 by Bouché *at al.* [164]. The original hit was identified in a high-throughput screen and already exhibited nanomolar affinity for BRD1. No activity was detected against the bromodomains BRPF1, BRPF3, and BRD4 up to 20 μM. The addition of a methyl group at position 6 of the 1,3-dimethylbenzimidazolone core introduced a 4-fold increase in affinity for BRD1, locking the tricyclic group in a bioactive conformation. The structure of this compound in complex with BRD1 (PDB ID: 5N49) showed that one of the carbonyl groups on the naphthalidimide makes the selective hydrogen bond to S592. A TR-FRET assay was used to determine IC_{50} values for the compound against other BRPF1-family bromodomains. This showed that the compound has approximately 15-fold selectivity for BRD1 over BRPF1. As this compound had poor solubility, an alkyl alcohol tail was added at position 4 of the naphthalimide, yielding the final structure of the chemical probe BAY-299. The authors report the hydrogen bond with S592 as a key factor driving the selectivity of these compounds for BRD1 versus BRPF1.

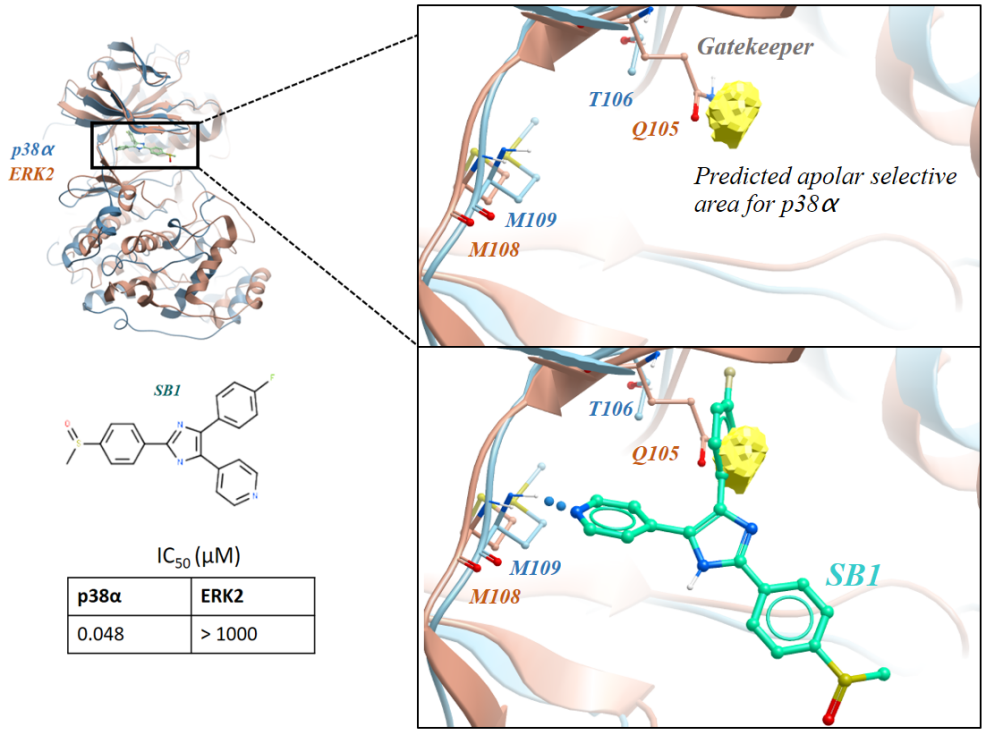
Generating selectivity maps from an ensemble was critical in this case, as the selective feature is not present in all the individual maps from BRD1 structures (an example is structure 5POS, shown in Figure 2.10). This is due to the sensitivity of hydrogen bonds to the orientation of the donor and acceptor groups, where small twisting motions in the backbone can mean that the feature is not detected in a minority of conformations.

2.3.3.2 Designing selectivity between closely-related human kinases: p38α and ERK2

In the family of protein kinases, an ATP-binding pocket residue known as the gatekeeper is an important determinant of selectivity [4]. This is exploited by the p38α-selective inhibitor SB1 (SB203580) shown in Figure 2.14, which exhibits selectivity over related MAPK kinases, notably ERK2 [155, 165]. The apolar selectivity maps calculated from an ensemble of 5 fragment-bound p38α structures versus 17 ERK2 fragment bound structures (full list of PBD codes in Appendix tables A.3 and A.4) can identify and highlight the selective hydrophobic pocket that the inhibitor binds in, as illustrated in Figure 2.14. As with the previous example, all structures in this case study had fragment-sized bound ligands, in order to recreate a prospective fragment-screening scenario.

The fluorophenyl group of SB1 occupies the selective hydrophobic backpocket of p38α, and clashes with the bulkier glutamine gatekeeper in ERK2. The ensemble and selectivity maps were able to identify this feature using only fragment-bound structures as input for the ensemble maps. In fact, this back pocket is often explored by fragment-sized hits in p38α (PDB ID 1W7H is an example). This corroborates the notion that in the case of a fragment screen against p38α, the selectivity maps could be used to suggest hypotheses for which fragments and chemical moieties are selective for p38α over ERK2, and so provide suggestions for achieving selectivity at a very early stage in the compound design process.

Despite using only 5 structures, all the bound ligands in the p38α had unique Murcko scaffolds [166], and they all explored the selective pocket (these are shown in Appendix Figure A.3). This includes the minimal pharmacophore chlorophenol (PDB ID 1WBO), which preferentially binds to the selective area behind the gatekeeper over the canonical kinase hinge hotspot. The ERK2 ensemble contained



IC_{50} (μM)

p38α	ERK2
0.048	> 1000

Figure 2.14: Kinase selectivity: identifying gatekeeper differences. Shown are the overlaid structures of p38α (PDB ID 1A9U, in steel blue) and ERK2 (PDB ID 4QP1, in peach). The close-up shows the binding mode of SB1 (light green). The ligand's fluororophenyl moiety is situated in a hydrophobic back pocket, located behind the gatekeeper residue T106. The apolar selectivity map (yellow surface) for p38α over ERK2 highlights this area as a favourable location to place a selective apolar group, such as the fluorophenyl in SB1. The table shows the compound's IC_{50} values for both proteins, as reported in Wang *et al.*, 1998.

more structures and higher ligand diversity (17 structures, 11 unique scaffolds), and whilst structures such as PDB ID 3ERK (not included in the ensemble, as its bound ligand is above the 300 Da limit) indicate that it is possible to reach that pocket in ERK2, the majority of the ERK2 fragments do not. Overall, despite a relatively low number of structures in the p38α ensemble, and a significant imbalance in the number of structures between the target and off-target ensembles, this approach was still able to identify selective features.

2.3.3.3 CK2α and PIM1: distantly related human kinases that bind the same ligand

The previous two case studies showed that the selectivity maps are able to retrospectively rationalise both polar and apolar selectivity features within protein subfamilies. In this final retrospective case study, an example of achieving selectivity between two kinases from different subfamilies was explored: human CK2α (CK2 sub-family) and human PIM1 (CAMK sub-family).

CX-4945 was originally developed as an ATP-competitive, orally available CK2α inhibitor with nanomolar affinity for its target, but which also displayed off-target activity for the PIM1 kinase [167, 168]. In 2011, Battistutta *et al.* developed a series of CK2α inhibitors, amongst which CX-5279 (Figure 2.15) retained affinity for CK2α whilst achieving selectivity against PIM1 [167].

Ensemble and selectivity maps were calculated for 28 CK2α and 32 PIM1 fragment bound structures, as detailed in the methods section (the full list of PDB IDs can be found in Appendix tables A.5 and A.6). The apolar selectivity map for CK2α over PIM1 shows an area of apolar propensity in CK2α that is inaccessible in PIM1 due to the conformation of residue Phe 49 in the PIM1 structures (Figure 2.15). This difference stems from the dominant conformation of this residue in

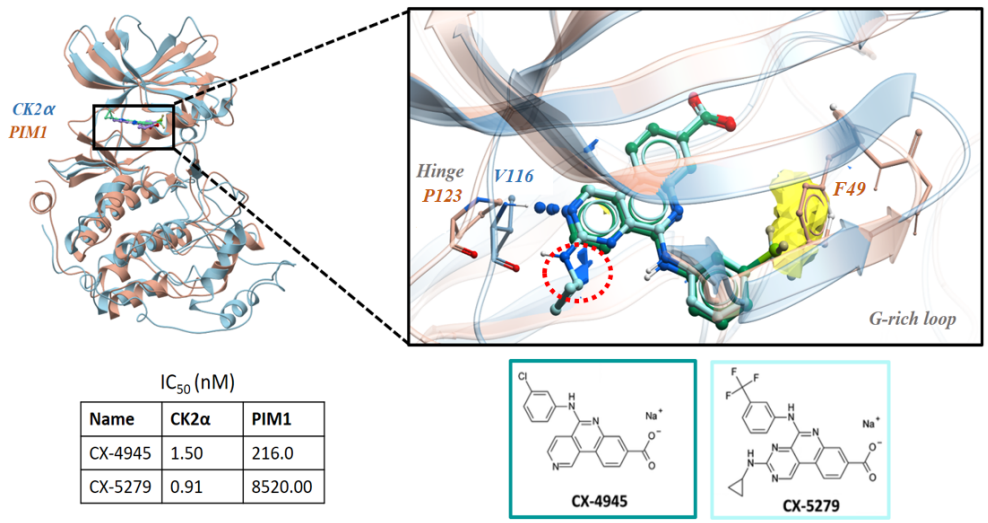


Figure 2.15: Tuning the selectivity of CK2α inhibitors towards PIM1. The binding mode of inhibitors in the ATP-binding pocket of CK2α is shown on the left. The close-up shows the overlay of the CK2α structures of CX-5279 (PDB ID 3ROT, ligand only, shown in aqua) and CX-4945 (PDB ID 3PE1, protein shown in light blue, ligand in dark teal), showing the selective hotspot regions for CK2α. The apolar and donor maps are shown as yellow and blue surfaces, respectively. PIM1 (PDB ID: 2C3I) is shown in peach for comparison. Residues giving rise to the selective features are shown as sticks and labelled. The red circle indicates the donor nitrogen in the cyclopropylamine group of the selective compound and the underlying donor density (blue surface) in the selectivity map. The IC_{50} values shown are as reported in Battistutta *et al.*, 2011.

the majority of ligand-bound PIM1 structures, and so would be nontrivial to predict from the sequence alone and without visual inspection of a large number of off-target structures. The ensemble and selectivity maps automatically highlight this feature without the need for human inspection and subjective judgement. The starting inhibitor, CX-4945, has a chlorophenyl group in this location. The selectivity maps suggested that placing a larger, highly lipophilic substituent on the phenyl ring in that location could improve the selectivity for CK2α over PIM1. In the closely-related inhibitor CX-5279, a trifluoro group has been added at that position on the ring, which resulted in improved selectivity for CK2α (Figure 2.15).

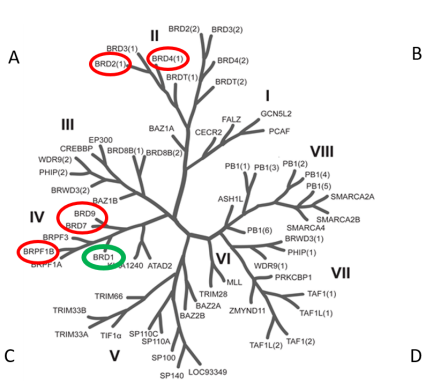
Another difference between CK2α and PIM1 lies in the hinge region and is identifiable by the selectivity maps. PIM1 contains a proline (Pro 123) at the position of the hinge valine (Val 116) in CK2α, making it unable to form one of the hydrogen bonds that CX-4945 makes with CK2α (Figure 2.15). The PIM1 hinge also has an insertion of 2 residues, which, coupled with the valine to proline substitution, prevents the donor nitrogen in the cyclopropylamine group forming a hydrogen bond with the backbone acceptor. The donor selectivity map highlights this feature, as shown in Figure 2.15. In the 2011 paper [167], the SAR for the compound series that includes CX-4945 showed no increase in affinity for CK2α when the cyclopropylamine was added. Furthermore, the influence of this moiety on selectivity against PIM1 is not explicitly shown. However, the resulting ligand is less lipophilic and makes an additional hydrogen bond, which is desirable in the design process. This is a modification that the selectivity maps would have suggested. In the case of the apolar selective feature, a visual inspection of the ensemble structures reveals that Phe 49 consistently adopts the inward-facing conformation that clashes with the selective ligand, increasing confidence in the selectivity of the feature.

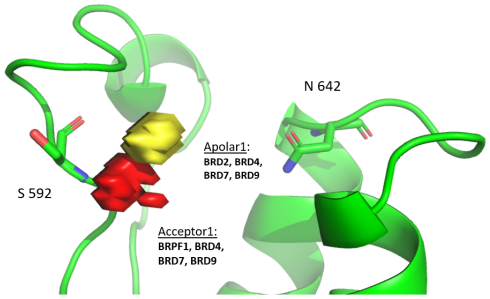
2.3.4 Selectivity maps identify selectivity-determining regions across subsets of targets in the same protein family

After ascertaining that the selectivity maps are able to identify known selectivity features, a procedure was developed that would allow for automated and objective analyses across a target protein family. The focus was again on human bromodomains, due to the wealth of structural and activity data publicly available for validating the method's predictions. In the example presented in Figure 2.16, BRD1 was chosen as the target protein, and was compared to both high sequence identity(BRPF1, BRD7, BRD9) and lower sequence identity (BET bromodomains BRD2(1) and BRD4(1)) off-targets. Ensemble maps were calculated for all the proteins, without applying the 300 Da cutoff on bound ligands, in order to take advantage of all the information available for potential off-targets, as would be the case in a drug discovery project. Selectivity maps were then calculated for BRD1 against each of the off-target ensembles.

The selectivity maps were combined into "summary" selectivity maps (using the same methodology as the ensemble maps, but without applying a frequency cutoff), showing all potential selectivity features observed in the dataset, as well as the off-targets they are predicted to confer selectivity against. As in the previous case studies, a hotspot score cutoff of 10 was used for the selectivity maps. Two prominent features in the summary selectivity maps are shown in Figure 2.16, panel B.

The procedure for querying activities and crystal identified eight compounds for which high-quality data was available for binding to BRD1 and at least one of the off-targets, and which have also been crystallised in complex with one of the proteins of interest. The search was limited to compounds with available crystal structures, in order to avoid introducing noise that would necessarily be created





C

Name	BRPF1	BRD4	BRD9	Acceptor1	Apolar1
5XE	0.2			0	0
9F9	0.12			0	0
8VI	0.01			0	0
LF1	0.0001			0	0
LS8	0.07	80.4	11.3	0	1
8LW	16.5	177.7		1	1
N48	0.03	9.4	0.49	0	1
5XF	0.04	9.12		0	0

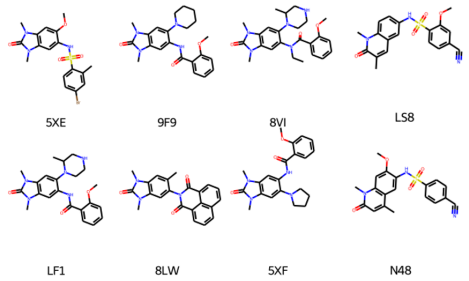


Figure 2.16: Selectivity maps identify selectivity-determining regions across subsets of targets in the same protein family. A. Human bromodomain phylogenetic tree (adapted from https://www.thesgc.org/chemical-probes) showing the target (BRD1) circled in green, and off-targets circled in red. B. Summary maps of the BRD1 selectivity maps against the off-targets. The acceptor feature (Acceptor1) has contributions from the BRPF1, BRD4, BRD7 and BRD9 selectivity maps, and the apolar (Apolar1) – from BRD2, BRD4, BRD7, and BRD9. The apo protein structure shown is of BRD1, PDB ID 3CRW. C. Selectivity ratios for the BRD1 versus the off-targets for the 8 compounds in the dataset. Values above 1 indicate selectivity for BRD1 and are coloured in green. Values below 1 are considered not selective and are coloured in red, and combinations for which information is not available in the dataset are greyed out. The last two columns indicate whether the crystal structure of the compound places a heavy atom in the predicted selective region. D. 2D structures of the compounds in the dataset and their PDB molecule IDs.

with a docking procedure. The structures of the 8 compounds were re-scored against the summary selectivity maps using the Hotspots API (as previously described in [113]), in order to identify substituent groups that interact with the predicted selective areas. A hit was defined if at least one heavy atom in the compound structure scored favourably for a particular feature (last two columns in Figure 2.16, panel C). This analysis shows that compound 8LW, which is selective for BRD1 against both BRPF1 and the more distantly related targets, places appropriate atom types in both clusters. Compounds that hit only the apolar cluster do not exhibit selectivity over BRPF1, as expected. However, compound N48 demonstrates that the selective apolar cluster on its own may not be sufficient to grant selectivity over BRD9. This is not unexpected, as selectivity is a complex phenomenon achieved by the interplay of multiple structural features, hence covering multiple features that are selective against different subsets of the protein family would result in a more selective ligand.

This principle has been used extensively for well-researched protein families with known selectivity determinants, such as kinases [169]. The selectivity maps have shown to be able to detect such features, and as scoring compound poses against the maps is computationally very fast (a few seconds per pose), these types of analyses can be used to objectively score large numbers of docked potential follow-up poses. The previously published Hotspots API [113] also allows for the extraction of pharmacophoric features from the maps, which can then be provided as input to programs such as CrossMiner [170], and used for the growing and merging of compounds.

2.3.5 Considerations for using the ensemble and selectivity maps prospectively

The quality of the ensemble maps is highly dependent on the quality and quantity of the input ensemble data. A key property of the ensemble is the number of structures included. Whilst crystallographic fragment screening experiments can provide up to tens of structures of ligands in complex with the target protein, such data is not always available. The ensemble of p38α structures showed that even a small ensemble with a diverse selection of ligands can be used to identify partially selective features. With smaller ensemble sizes, individual structures contribute more to the ensemble map.

Visual inspection, or plots such as the one in Figure 2.10, can be used to establish which structures contribute to a particular hotspot cluster. In the case of p38α, all the maps contributed to the selective feature discussed above. By reducing the frequency threshold of the ensemble maps, the contribution of individual structures to the maps can be amplified in larger ensembles as well. A complementary computational method can then be used to assess the significance of rare or unique conformations. For instance, molecular dynamics-based methods such as dynamic undocking [47] can be used to assess the stability of the ligand-protein complex and guide the decision whether that would be a desirable structure to include in the ensemble. On the other hand, in an ensemble with tens of structures, even at a 20-30 % frequency threshold, signal arising from rare conformations may be lost in noise from adjacent clusters.

In such cases, the question of what the minimum representative ensemble should be arises. In the examples above, a ligand-based measure for ensemble diversity - the number of unique Murcko scaffolds (the full list of Murcko scaffolds for the ensembles can be found in Appendix tables A.1 - A.6) - was considered to

guide the choice of the minimum number of structures in the ensemble. Protein or protein-and-ligand based methods, such as protein-ligand interaction fingerprints [171, 156], or backbone clustering methods [172] can be used to select a diverse ensemble, or even to generate ensemble maps for a group of structures. For instance, in a lead optimisation campaign, a subset of structures in complex with ligands from the same series might be used to identify unexploited hotspot interactions in this very specific context.

A general recommendation for ensembles with over 10 structures is to build ensemble maps with a frequency threshold of 20 % to identify the most frequent hotspots throughout the ensemble. If a comparison of these maps against an off-target generates no selective clusters, switching to a lower frequency threshold can reveal rarer features. Conversely, if the ensemble maps appear noisy (see Figure 2.8, leftmost column), switching to a higher frequency threshold can reveal the most prominent hotspots.

In the case of selectivity maps, two key parameters to be considered are the minimal distance between the centers of selective clusters in the target and off-target selectivity maps, and the minimum hotspot score a cluster must pass in order to be considered selective. In principle, the ideal selective feature would both score highly and be located a reasonable distance from other features that possess binding propensity for the off-target. The lower limit for the first parameter is twice the step size of the grid: 1 Å between features. The minimum distance is dependent on other factors such as the volume and flexibility of the binding site. For the small and rigid bromodomain binding sites, using a cutoff of 1.5 - 2.0 Å was sufficient to isolate the polar selective features, which is consistent with values published previously [40]. For kinases, a distance threshold of 3 Å was chosen, as their binding sites exhibit greater conformational diversity , so a value of twice the minimal distance was used as a precaution against false positive values. In

cases where little information is available for the target and off-target proteins, prioritising highly scoring clusters at a distance of 3 Å or more from the centre of the closest off-target cluster would be a recommended starting point, lowering this parameter only if no features are detected, with the caveat that features will be identified with lower certainty.

2.3.6 Conclusion

Fragment hotspot mapping has previously been shown to be a promising method for guiding the hit-to-lead phase of drug and probe discovery campaigns. Building upon this approach, ensemble hotspot maps were introduced to summarise important fragment-binding interactions made by an ensemble of structures, extending the original hotspot method to work over multiple conformations of the same protein. Importantly, ensemble hotspot maps can then be used to identify differences between related proteins in the same family. These maps can highlight nuanced differences between protein binding sites. Three case studies from well-understood protein families were shown, in which the selectivity maps were able to prioritise binding site differences used to design selective inhibitors. These examples showed the utility of the ensemble and hotspot maps in comparing the binding sites of a target and an off-target protein in the same family. Provided that suitable ensembles were selected, the target features could be identified even in cases where imbalances in the number of structures between the on and off targets was observed.

In addition to the pairwise comparisons in the case studies, the ensemble and selectivity workflow was extended to allow automated comparisons of multiple binding sites of related proteins. This extended workflow was applied to investigate opportunities for designing compounds selective for BRD1 over five other human bromodomains with various levels of structural similarity and was able to

identify features that are selective against different subsets of the protein family. As selectivity arises from the interplay of multiple structural features, the ability to leverage large amounts of structural data across a target family can be a great asset for modulating a lead's selectivity profile. The extended comparison workflow presents a powerful tool to support this process.

The ensemble and selectivity maps, as well as the extended comparison workflow, will be applied to further protein families to inform the recommended values for the various map parameters. Compiling ensemble maps from snapshots of molecular dynamics trajectories is also being investigated to further understand the behavior of binding site hotspots, especially in cases where few experimental structures are available. Currently, the fragment hotspot mapping method has been validated for apolar, hydrogen bond donor, and hydrogen bond acceptor probes. These interactions represent a large portion of protein-fragment intermolecular contacts, but the method could in future be extended by adding charged or halogen fragment probes, as these are also supported by the underlying SuperStar program. Nevertheless, the ensemble and selectivity maps are a quick and scalable means to summarise and intuitively present structural information from closely related proteins and generate hypotheses on achieving selectivity. Work on automating this workflow and applying it prospectively to on-going projects will be presented in Chapter 4.

3 | Structural stability as a complementary metric for the fragment hotspot maps

3.1 Introduction

This chapter will investigate a complementary computational method that can be used to both prioritise observed fragment hits, as well as score potential follow-up compounds. The Dynamic Undocking [47] steered molecular dynamics method, introduced in Section 1.7 is orthogonal to both fragment hotspot maps and docking in its approach to scoring interactions [47]. Initially, an open-source implementation of the DUck protocol, OpenDUck, will be validated on two large, high-quality benchmarking datasets: Iridium [135] and SeraPHIC[136]. A simple diagnostic developed for detecting outliers in the OpenDUck output will be introduced. OpenDUck will then be used to investigate a retrospective fragment elaboration case study, and its ability to prioritise structurally similar follow-up compounds will be evaluated.

Fragment hotspot maps are a fast, intuitive, and automatable method for identifying the key binding site features that contribute to fragment binding. They are also entirely protein-based: even in cases where the protein-fragment complex is taken as input, the maps are calculated based on the holo protein structure. This means that, apart from ligand-induced conformational changes, the maps are largely "ligand-agnostic". Individual hotspot features can indicate areas where, for example, placing an acceptor group would be favourable, but they cannot differentiate between different acceptor groups.

To address this limitation, the hotspot maps can be used in tandem with docking, as docking scores consider the ligand. However, many docking scoring func-

tions are not parameterised for fragment and follow-up sized ligands [58]. More sophisticated computational methods could add further information towards prioritising both the initial fragments and the suggested followup compounds, particularly with regards to the stability of the binding mode. As discussed in Section 1.7, a combination of docking and dynamic undocking was highly successful in fragment virtual screening [47]. The method's ability to estimate the robustness of particular hydrogen bonds complements the hotspot maps' ability to identify and prioritise key interactions in the binding site, creating the possibility of using the methods as part of a workflow to drive fragment elaboration. To facilitate its integration into computational pipelines, the open-source Python-based implementation of dynamic undocking, OpenDUck, was used. At the start of the PhD project, this version had been implemented, but had not been shown to perform comparably to the original DUck workflow. This comparison was undertaken as a first step towards including OpenDUck in a fragment elaboration workflow.

3.2 DUck and OpenDUck

The main stages of the DUck workflow are outlined in Figure 3.1 and are shared between DUck and OpenDUck (Table 3.1). After a key protein-ligand hydrogen bond has been defined, a "chunk" of residues around the protein atom involved in this interaction is selected. The goal is to retain the "minimal subset of residues that preserve the local environment around the key hydrogen bond" [47, 128, 120]. The purpose of this is twofold: first, removing a large part of the receptor drastically decreases the computational cost of the simulation. Second, it removes residues that may interfere with the ligand as it is steered out of the binding site, which could lead to artificially heightened W_{QB} scores [128]. Conversely, selecting too few residues around the protein atom can artificially increase the solvent

exposure, leading to a much lower W_{QB} estimate [128]. Requirements and guidelines for chunk selection are provided by Majewski *et al.* in a highly informative protocol chapter [128] and will be discussed further in Section 3.2.1.

Table 3.1: Comparison of the DUck and OpenDUck workflows.

Step	**DUck**	**OpenDUck**
Chunking	MOE (interactive)	Python script (no GUI)
Ligand parameters	Prm@frosst	SmirnOFF
Protein parameters	AMBER99sb	AMBER99sb
MD Engine	AMBER	OpenMM
Chunk restraints	Heavy atoms, 1 kcal.mol^{-1}Å^{-2}	Heavy atoms, 1 kcal.mol^{-1}Å^{-2}
Minimisation	1000 cycles	1000 cycles
Heating	Langevin thermostat (4 ps^{-1}, 9 Å)	Langevin thermostat (4 ps^{-1}, 9 Å)
Density equilibration	Berendsen barostat (1 atm, 300 K)	Montecarlo barostat (1 atm, 300 K)
MD	1ns	1 ns
SMD	500 ps, 300/325 K	500 ps, 300/325 K
Spring constant	50 kcal.mol^{-1} Å^{-2}	50 kcal.mol^{-1} Å^{-2}
Analysis	R script	Python script

The next steps involve parameterising the chunk and ligand for the MD runs, the addition of solvent molecules and ions, followed by system equilibration. Short MD runs (1 ns) are performed prior to the SMD runs, in order to create diverse starting points for the SMD. Two SMD runs are started after each bout of MD - one at 300 K, and one at 325 K. The difference in temperature ensures that the trajectories proceed differently [128, 47]. Finally, the W_{QB} values for the SMD runs are calculated.

3.2.1 Creating the chunk

The definition of an appropriate receptor chunk is crucial for the success of the method [120, 128]. As mentioned previously, a chunk that is too large may impede the ligand's exit trajectory, leading to artificially heightened W_{QB} values, whilst one that is too small may expose the interaction to solvent molecules, lowering the W_{QB} estimate. This means that while receptor chunks may be generated

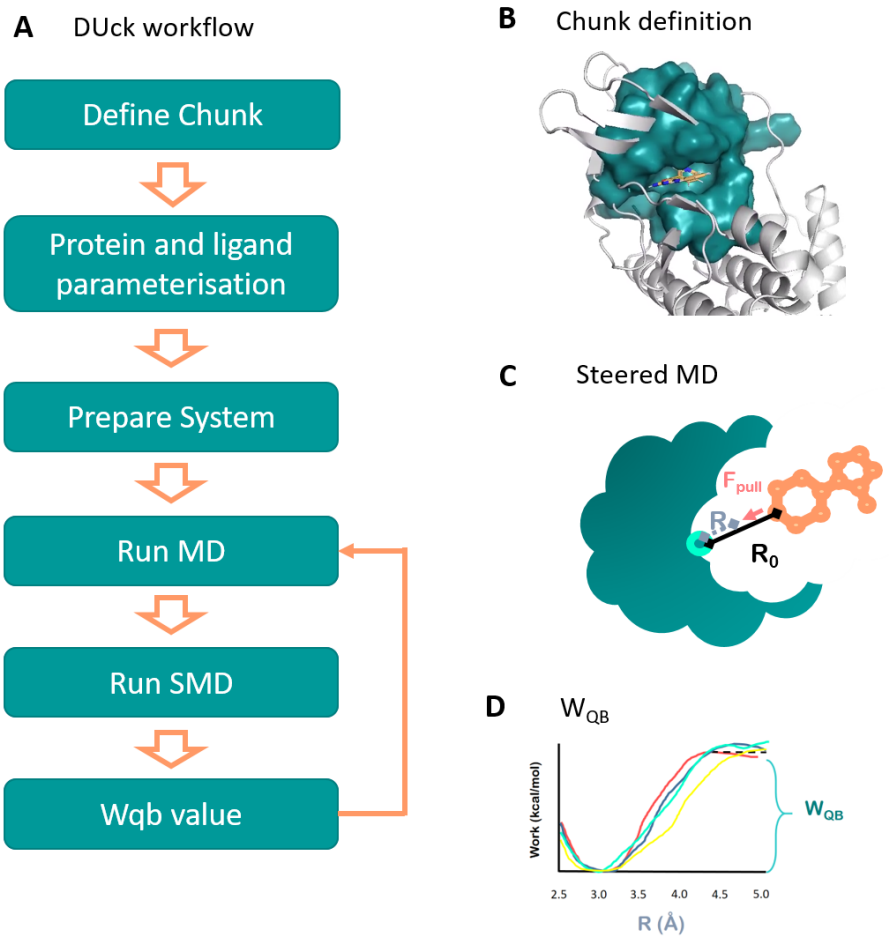


Figure 3.1: Stages of the DUck and OpenDUck workflow. A. Flowchart of the DUck workflow steps once a target interaction has been selected. B. Chunk definition is the process of selecting a subset of protein residues that constitutes the target interaction's local environment. C. The steered MD step that the DUck calculation is based on. D. Calculating the W_{QB} value based on the work needed to reach the quasi-bound state.

automatically, visual inspection of the chunk in a molecular viewer is crucial to the success of the method. The authors of the DUck method make the following recommendations for defining the receptor chunk [128].

1. Correct protonation of the ligand and receptor are key to success. In the case where the starting structure comes from a docking run, it is recommended that the same receptor is used as input for DUck.
2. Structurally important waters should be identified and preserved as part of the chunk.
3. If the chunking process opens channels in the structure through which solvent molecules can reach and disrupt the key interaction, residues that block these channels should be selected as part of the chunk.
4. If the sequence gap between two separate parts of the chunk is less than 3 residues, these connecting residues should be retained.
5. Cropping out the chunk causes the polypeptide chains to split, leaving charged ends which may then interfere with the simulation. These are neutralised by capping with acetyl and N-methyl groups.
6. If an AMBER forcefield is used, the final structure should be saved with residue names adjusted to AMBER templates. This is particularly important for retaining the correct protonation states of histidines and other polarisable residues.

In the original DUck implementation, these steps are executed in MOE through an SVL script. In the OpenDUck version, they are handled in Python through the ParmEd [173] module. The chunk is exported as a .pdb file, which is then visually inspected using a molecular viewer. Updated scripts that can retain or remove specific residues, as well as retain key structural waters, are available at `https:`

//github.com/mihaelasmilova/duck/scripts/chunking.py.

3.2.2 Protein and ligand parameterisation

Dynamic undocking is a method based on molecular dynamics simulations. To model the protein and ligand, force field parameters need to be assigned to both the protein and the ligand. Both DUck and OpenDUck use the AMBER99sb [174] forcefield to model the protein. OpenDUck uses the SmirnOFF99Frosst forcefield (https://github.com/openforcefield/smirnoff99Frosst) for ligand parameterisation, as an open-source alternative to Parm@Frosst, used by DUck. Am1-bcc [175]charges are computed for the ligand prior to parameterisation (OpenDUck uses Antechamber [176] for this step, called through the OpenForceField [132] Python package). The OpenDUck version described in this chapter uses version 0.7.2 of the OpenForceField package and version 7.4.1. of OpenMM.

3.2.3 Preparing the system

In both DUck and OpenDUck, valid ligand and protein atom topologies are generated using AMBER's tLeap module [176]. The protein and ligand are placed in a cubic box, which is filled with TIP3P solvent molecules (pre-defined important structural waters are retained at their positions and parameterised accordingly). OpenDUck uses the PDBFixer [177] Python package for this step, with a padding of at least 20 Å in each direction. Charges are neutralised through the addition of Cl^- and NA^+ ions.

3.2.4 Equilibration

Steered MD methods allow for the investigation of non-equilibrium properties by perturbing the system away from its equilibrium state. To allow this, the equilibrium state must first be reached in the MD simulation. In this state, thermodynamic properties such as volume and temperature fluctuate around stable averages in a way that is resistant to small fluctuations. The process and protocol used to bring the system into this state is referred to as the equilibration step. DUck runs all equilibration, unsteered MD and SMD steps using a Langevin thermostat with collision frequency of 4 ps^{-1} and a non-bonded interaction cutoff of 9 Å [47]. In all the MD steps, harmonic positional restraints with a force constant of 1 $kcal.mol^{-1}Å^{-2}$ are placed on the protein heavy atoms. This reduces conformational flexibility and prevents the chunk polypeptides from dissociating. In addition, a restraint is placed on the interaction between the ligand and protein chunk atom in all non-steered steps, in order to ensure that the interaction is present in all starting points for the steered MD runs. For distances between 3 and 4 Å, this restraint is parabolic with a force constant of 1 $kcal.mol^{-1}$ $Å^{-1}$; for distances beyond 4 Å, it is linear with a force constant of 10 $kcal.mol^{-1}$ $Å^{-1}$.

3.2.4.1 Minimisation

In preparation for the MD calculations, the system undergoes 1000 cycles of minimisation. In OpenDUck, this is done using a Verlet integrator with a time step of 1 fs. These minimised positions are then used as input for the subsequent heating step, using the same restraints.

3.2.4.2 Heating

Random velocities at 100 K are assigned to the atom positions from the end of the minimisation step in both methods. The system is then heated to 300 K in the normal volume and temperature ensemble using a Langevin integrator with a friction coefficient 4 ps^{-1} and a step size of 2 fs. In OpenDUck, this part of the simulation runs for 50000 steps, or 100 ps.

3.2.4.3 Density equilibration

The positions and velocities at the end of the heating step are then used as input for a further density equilibration step of 100 ps, in the normal temperature and pressure ensemble. OpenDUck uses a Monte Carlo barostat with a default pressure of 1 atmosphere and a temperature of 300 K.

3.2.5 MD and SMD runs

After the system has been equilibrated, an unsteered MD run is initiated to generate the starting point for the steered MD simulations. As mentioned previously, a restraint is placed on the protein and ligand atoms involved in the interaction. Two SMD runs are started from each MD checkpoint, one at 300 and one at 325 K. This is done to ensure that the trajectories proceed differently [47]. The length of the unsteered MD runs between SMD runs is set to 1 ns, with each subsequent unsteered run starting from the positions and velocities of the previous. This means that for 40 SMD runs, 20 ns of unsteered MD will be run [47].

During the steered MD, a custom external force is applied to the ligand atom, which acts as a spring, pulling or pushing the ligand atom towards a point located a particular distance away from the protein atom along the vector specified by the ligand and protein atoms at that point in the simulation. In the OpenDUck

implementation, this force is defined through the CustomExternalForce class. The spring constant is a "per particle parameter", which is part of the system definition and cannot be changed during the course of the simulation. The value of the spring constant is set to 50 kcal.mol^{-1} Å^{-2}. In the paper by Ruiz-Carmona *et al.,*, spring constants in the range between 10 and 1000 kcal.mol^{-1} Å^{-1} were tested and were found to not affect the W_{QB} values [47], so the value of 50 is also used by the OpenDUck workflow.

During the course of the steered MD, the distance to which the force pulls the ligand atom is incrementally varied between 2.5 and 5.0 Å away from the protein atom. The size of this increment determines how quickly the ligand is being pulled. A range of velocities was tested in the original DUck implementation, a velocity of 5 Å.ns^{-1} was used, and slower velocities were not found to influence the result [47]. The work exerted by the pull force at each step is calculated according to the equation in Figure 1.9 and recorded for the calculation of W_{QB}.

3.2.6 Calculating W_{QB}

The implementation of OpenDUck used in this chapter uses a Python script originally written by Maciej Majewski to calculate the W_{QB} values from the recorded work. The script finds the first local minimum in the trace by sliding a window of 200 points. The difference between this value and the maximum value observed after the minimum has been reached gives the W_{QB}. When traces are plotted, the minimum work value is subtracted from the measured work.

In the original paper by Ruiz-Carmona *et al.,* the minimum W_{QB} value in any of the SMD runs was taken as the W_{QB} value for the interaction in order to avoid overestimating the structural stability and introducing false positives. This value is referred to as the W_{QB_min} throughout this chapter.While it was shown to be an

effective metric in virtual screening, it is difficult to measure the associated error [134]. In a subsequent paper from the group, Majewski *et al.* ran 40 SMD runs per interaction and randomly split them into four groups of 10 values. The minimum in each group was calculated, and the average of the 4 minima reported as the final W_{QB} estimate for the interactions. In this chapter, this value is referred to as $W_{QB\_mean\_min}$. It has the advantage of providing an estimate of the statistical error in calculating the W_{QB}, however its calculation requires a greater number of SMD runs.

3.2.7 Visualising MD trajectories

While the output of the DUck method is a single numerical value, the ability to visualise the trajectories for the MD and SMD runs is particularly important for assessing the validity of the results, especially when validating the method. As part of the OpenDUck validation work, I have included a VMD [133] script that highlights the ligand and the interacting chunk residue, as well as displays hydrogen bonds and water molecules in the vicinity of the ligand (shown in Figure 3.2). This script is available at `https://github.com/mihaelasmilova/duck/tree/master/scripts/duck_analysis_visualisation.py`.

3.3 OpenDUck Validation

OpenDUck was devised as an open-source version of the DUck protocol, using the same principles and many of the same components. However, at the start of this project, there was no systematic validation showing that DUck and OpenDUck performed comparably on a target dataset. This validation would be a necessary prerequisite for using OpenDUck prospectively. The original implementation of the DUck method had been previously applied to the Iridium [135]

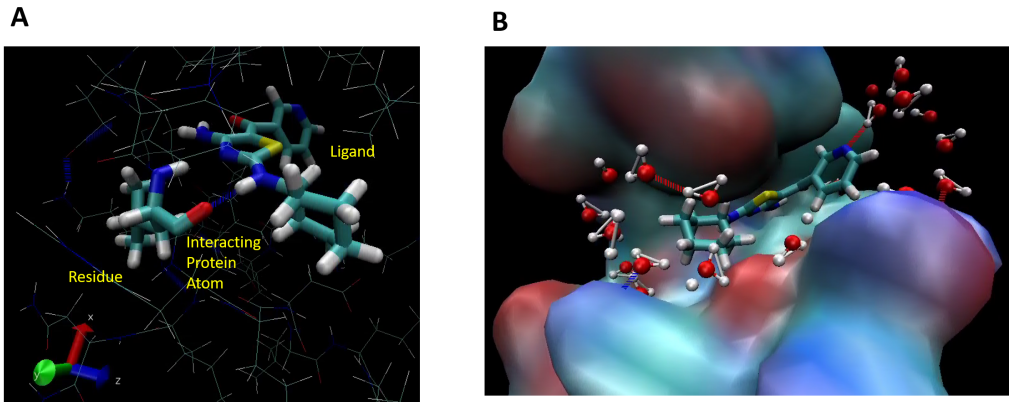


Figure 3.2: Visualising DUck trajectories. A. The default view output by the visualisation script. The interacting protein and residue are shown as licorice, and the hydrogen-bonds between them are displayed. The rest of the chunk is shown as sticks for reference B. A second visualisation, useful for detecting chunking problems. The chunk is shown as a surface (using the QuickSurf representation and a resolution of 0.8), the ligand- as licorice, and water molecules within 5 Å of the ligand are shown as CPK. Hydrogen bonds within 5 Å of the ligand are also displayed.

and SERAPhiC [136] datasets, which consist of diverse and high quality protein-ligand crystallographic structures [134, 137]. The protonated ligand and receptor structures used in the original DUck experiments were also available, making these two datasets an appropriate choice for the retrospective validation of the OpenDUck workflow. A a non-equilibrium property, structural stability is difficult to measure experimentally, and a large-scale data-set of experimental values was not available. Majewski *et al.* showed that even in the absence of experimental data, W_{QB} values can provide useful information in virtual screening [47, 134, 45],target analysis [134], and docking [137]. For these reasons, the W_{QB}values assigned to the Iridium and SERAPhiC datasets by DUck were used to benchmark the performance of OpenDUck. However, it is important to emphasise that both of these computational methods work with models and approximations of the systems being investigated, so neither value is necessarily representative of

the "ground truth".

3.3.1 Datasets

3.3.1.1 Iridium

The Iridium dataset was introduced in 2012 by Warren and co-workers as a highly-trustworthy set of 121 curated structures for benchmarking docking programs [135]. These were compiled from four datasets that had previously been used to validate docking programs, which had a combined total of over 700 unique protein-ligand structures. These were then filtered to only include complexes with available crystallographic density, and those structures were re-refined, removing noise that can occur from different crystallographic refinement and model building pipelines. Clerical errors such as bond orders, ionization states, tautomer states, stereochemistry and atom/group identities were amended. All structures in the dataset have complete crystallographic density in the binding site region and no alternative side chain or ligand conformations.

Not all of the structures in the Iridium set are suitable for DUck experiments, however. In the 2019 paper, Majewski and colleagues used a subset of 78 complexes from Iridium, after removing those that would not be suitable for simulation [134]. Three complexes do not form hydrogen bonds with the ligand, and so were not included in the analysis. A further 28 have coordination complexes with metal ions, and were excluded as the molecular mechanics force field does not faithfully represent them [134]. As this would also be the case for OpenDUck, these complexes were not simulated in the current work, either. Six complexes were deemed not drug-like (large molecular weight, peptide-like), and a further five had additional ligands in the binding site, which would greatly complicate the analysis. The prepared structures of this reduced set, downloaded from the Supplementary Data of

the publication [134], were used in this chapter.

3.3.1.2 SERAPhiC

The SERAPhiC data-set was originally devised as a benchmark dataset specifically for *in silico* fragment-based drug design [136]. It consists of 53 high-quality X-ray models, curated and prepared for docking by the paper authors. Only structures published after the year 2000, with at least one associated PubMed publication, and resolution of 2.5 Å or better were included. Ligand size was limited to between 78 and 300 Da, with common crystallisation additives removed. Only structures for which electron density was available were included, in order to assess the fragment fit.

Majewski *et al.,* purposefully did not include a number of the SERAPhiC structures in their analysis, as they were not suitable for the DUck method. Seven cases made no hydrogen bonds with the ligand, five had additional ligands in the binding site, and 17 had metal ions. This left 26 binding pockets across protein-fragment complexes. Three structures (1e2i, 1ofz, and 2hdq) had two ligands per protein, with 1e2i having 2 enantiomers of the ligand in the same binding pocket. In this case, both isomers were simulated separately [134]. The same subset, downloaded from the Supplementary Data of the publication [134] was used as input for the OpenDUck validation in this chapter.

In both the Iridium and the SERAPhiC sets, individual interactions are named as follows: <protein structure PDB ID >_< name and number of interacting residue >_< interacting protein atom name> (For example, 1k1j_ASP189_OD2 refers to the interaction made by the OD2 oxygen atom of Asp 189 in PDB structure 1k1j). These interactions were originally detected for both datasets by Majewski *et al.* using MOE [129].

3.3.2 Effect of the MD Engine on W_{QB} values

The DUck workflow has many steps and interacting components, so the causes of differences in W_{QB} values between DUck and OpenDUck needed to be carefully investigated and deconvoluted. Maciej Majewski kindly shared the parameterised chunks, ligands, and work values for a subset of the Iridium complexes generated using the original DUck method (referred to as DUck-AMBER throughout this section, as it uses the AMBER MD engine). The .prmtop and .prmcrd files were used as starting points for equilibration, MD and SMD runs using the OpenMM implementation of OpenDUck (referred to as OpenDUck-OpenMM). This allowed for the effect of the MD engine to be isolated as a cause for differences observed between the two methods. As the sets of trajectories will not be identical between the two methods, it is possible that states are sampled in one method but not the other due to differences in sampling, rather than from a systematic difference in the simulation engine. However, such cases should be diagnosable from a visual inspection of the traces and trajectories.

3.3.2.1 Methods

The reference chunks for 30 of the interactions detected in the Iridium dataset were supplied as .prmcrd (containing the coordinates) and .prmtop (containing the parameterisation and topology information) files. In addition, a .pdb file of the chunk and the original protonated files for the full length protein and ligand were also available. A script was generated that uses the ParmEd Python module [173] to parse this information, finds the key interaction between the protein and ligand in chunk, and runs the OpenDUck equilibration, MD and SMD steps. This script can be found at https://github.com/mihaelasmilova/duck/blob/master/scripts/from_amber_input.py. A total of 20 SMD

runs (10 at each temperature) were run for each of the interactions. All other parameters were kept at their default values (MD length 0.5 ns, initial velocity 5 Å.ns^{-1}, pull distance 2.5 Å).

$W_{QB\_mean\_min}$ values for individual SMD runs were calculated as previously described. In order to provide error data, all W_{QB} values were split into 6 randomly selected (without replacement) groups of 4 values. The minima of the groups were averaged, and the standard deviation of the minima is plotted as error bars in Figure 3.3.

3.3.2.2 Results

The comparison between the original DUck method and the OpenDUck MD and SMD started from the same parameterised chunks is presented in Figure 3.3. Overall, the values are highly correlated - this is clear from both the summary plots in Figure 3.3, and from the individual traces shown in Figure 3.4. Of the 27 systems simulated, the 6 shown in Figure 3.3 had a difference between their mean minimal W_{QB} values greater than 1 standard deviation (calculated based on the differences in mean minimal W_{QB} between the pairs of Duck (AMBER) and OpenDUck (OpenMM) runs). Of these, 1k1j_ASP189_OD2 showed the greatest discrepancy in calculated values. The W_{QB} traces for these runs are shown in Figure 3.4.

Even for these outliers, the work profiles have roughly the same shape between the two MD engines, which was encouraging. The sharp peak in the 1f0u_SER190_OG profile is the result of the ligand's amidine moiety flipping during the trajectory to face away from the protein atom. A close inspection of the DUck-AMBER trace shows that this also happens to some (although much lesser) extent in the reference case. The W_{QB} values are still comparable between the cases, in part

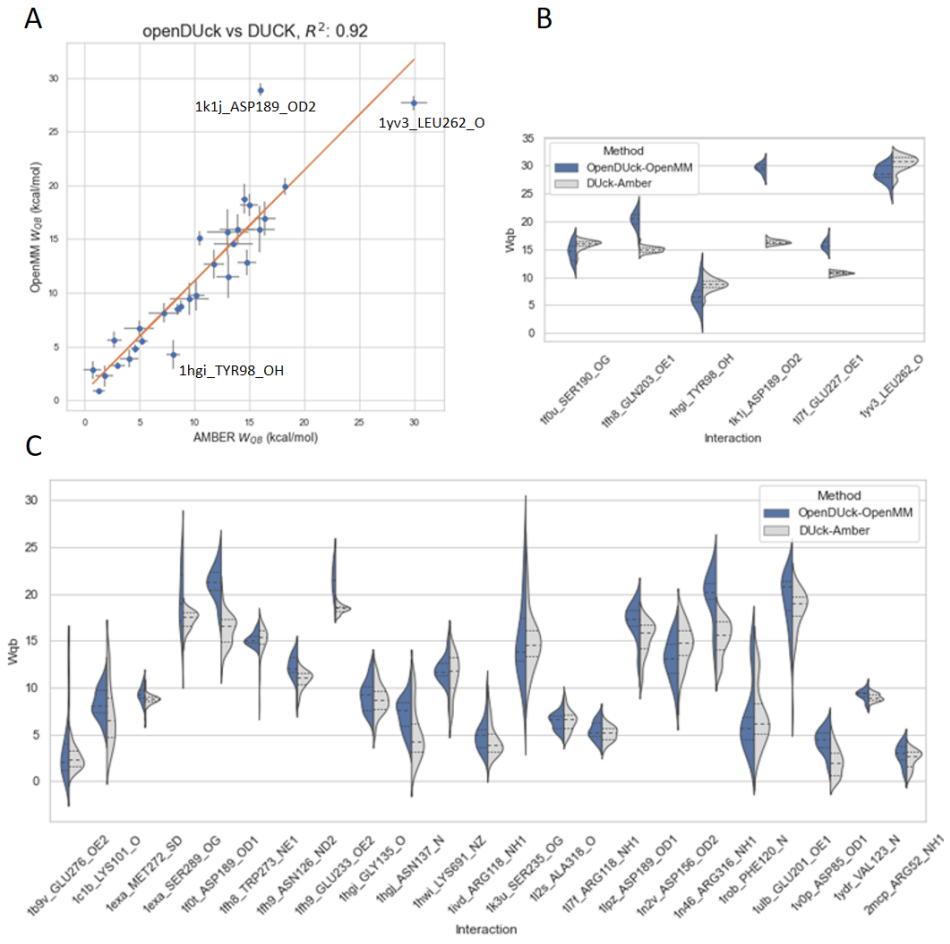


Figure 3.3: Comparing MD engines starting from the same parameterised chunk. A. Correlation between the W_{QB} values calculated by OpenDUck and Amber. B. Systems where the difference in calculated values is greater than one standard deviation (from all the systems calculated). C. Systems where the difference in mean minimal W_{QB} values is less than 1 standard deviation

.

because the work exerted by the pulling force increases somewhat faster in the OpenDUck-OpenMM trajectories.

The case of 1hgi_TYR98_OH in particular required further investigation, as DUck-AMBER classified the interaction as structurally stable according to the $W_{QB\_mean\_min}$ (5.15 if the overall minimum value is taken, which is on the border), while OpenDUck classified it as unstable ($W_{QB\_mean\_min}$ of 3.65 and a (W_{QB_min} of 2.78). Both traces consist of variable trajectories (for example, when compared to 1fh8_GLN203_OE1 or 1lf_GLU227_OE1), and although the OpenDUck traces have a wider distribution overall, the maximum W_{QB} values observed are comparable. Therefore it is possible that in this case, the difference is due to the stochastic nature of the method, and not to a systematic difference in the forces simulated.

Four of the outliers (1f0u_SER190_OG, 1fh8_GLN203_OE1, 1l7f_GLU227_OE1, and 1k1j_ASP189_OD2) feature a charged NH^+ nitrogen as the ligand atom in the key interaction. They all also consistently have higher values for the OpenDUck-OpenMM trajectories than for the corresponding DUck-AMBER ones (the ligand flip in 1f0u_SER190_OG lowers the values, but until this point, the work profile shows a steeper ascent in the OpenMM traces see Figure 3.4). Three of the outlying cases are also paired with a negatively charged atom on the protein (1fh8_GLN203_OE1, 1l7f_GLU227_OE1, and 1k1j_ASP189_OD2). In their 2019 paper [134], Majewski and colleagues reported that charge-reinforced hydrogen bonds generally were not more structurally stable than ones involving neutral atoms. Therefore, it is possible that OpenDUck may slightly overestimate (at least relative to the original DUck implementation) the structural stability in cases where charged nitrogen groups are involved in the interaction. However, in all four cases both methods agree that these interactions are highly stable, and so the difference in values would not impact the outcome of the experiments.

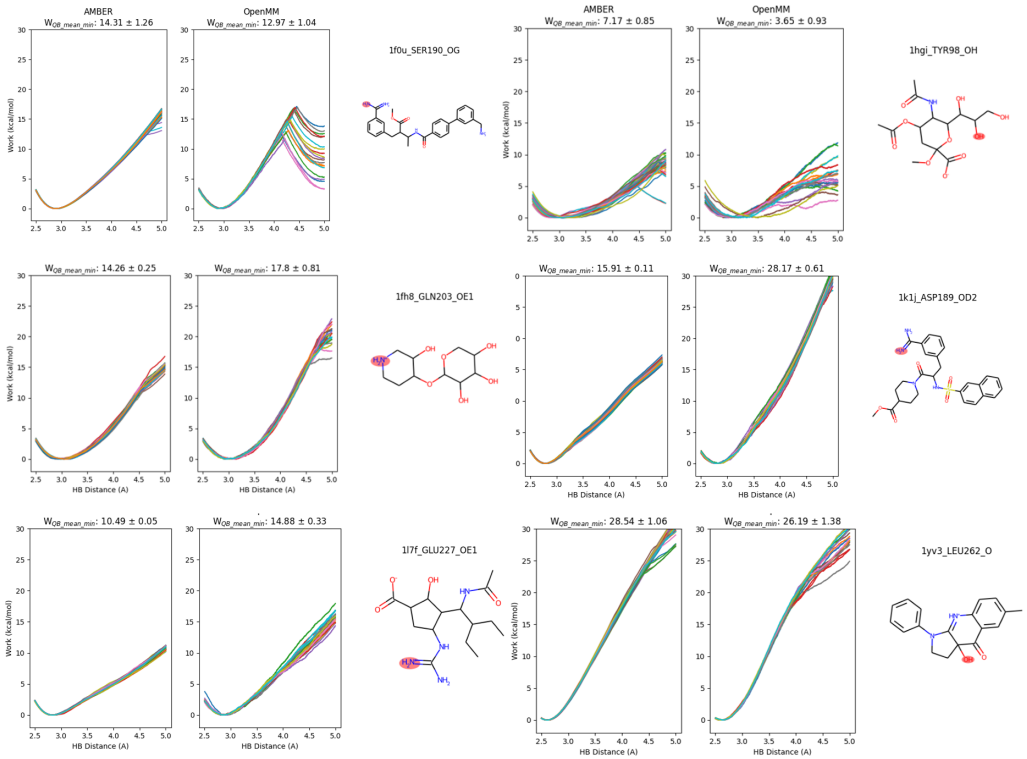


Figure 3.4: W_{QB} traces for the DUck (AMBER) and OpenDUck (OpenMM) outliers. 2D structures of the ligands are shown to the right of the traces, with the ligand atom involved in the key hydrogen bond highlighted in red.

This small overestimation appears to also be the case (albeit to a lesser extent) for the other two interactions involving charged N ligand atoms in the data-set (1lpz_ASP189_OD1, 1f0t_ASP189_OD1, traces shown in Figure 3.5). Panels C and D in Figure 3.5 show other cases involving (uncharged) nitrogen atoms: panel C (1l2s_ALA318_O) shows excellent agreement between the methods in the case of a donor NH on the ligand interacting with a backbone carbonyl, while section D (1ydr_VAL123_N) shows the case of an acceptor N on the ligand interacting with the backbone amide NH, again with very good agreement between the methods.

The final outlier, 1yv3_LEU262_O, again shows a small difference in $W_{QB\_min\_mean}$ values between the two methods, but the traces and general shape of the distributions is comparable. Again, this is ranked as a highly stucturally stable interaction by both of the methods.

Considering that both methods use the same force field, parameters, and structures, it is interesting that a case like 1k1j_ASP189_OD2 (a noticeable and systematic difference between the methods) occurs. However, discrepancies between the energies and forces between different MD engines using the same force fields have been previously reported [173]. In a systematic comparison between commonly used MD programs (AMBER, GROMACS, CHARMM, LAMPS, and DESMOND) on the SAMPL5 blind prediction challenge data-set, Shirts *et al.* found that all of them agreed to better than 0.1 % relative absolute energy for all energy components. However, statistically significant differences between programs were observed in certain cases, with different choices of Coulomb's constant being one of the greatest sources of these discrepancies [173].

A direct comparison between OpenMM and AMBER is presented by Eastman and colleagues, who compared the energies and forces for systems (DHFR and a short hairpin RNA) prepared though the input pipelines of the two programs,

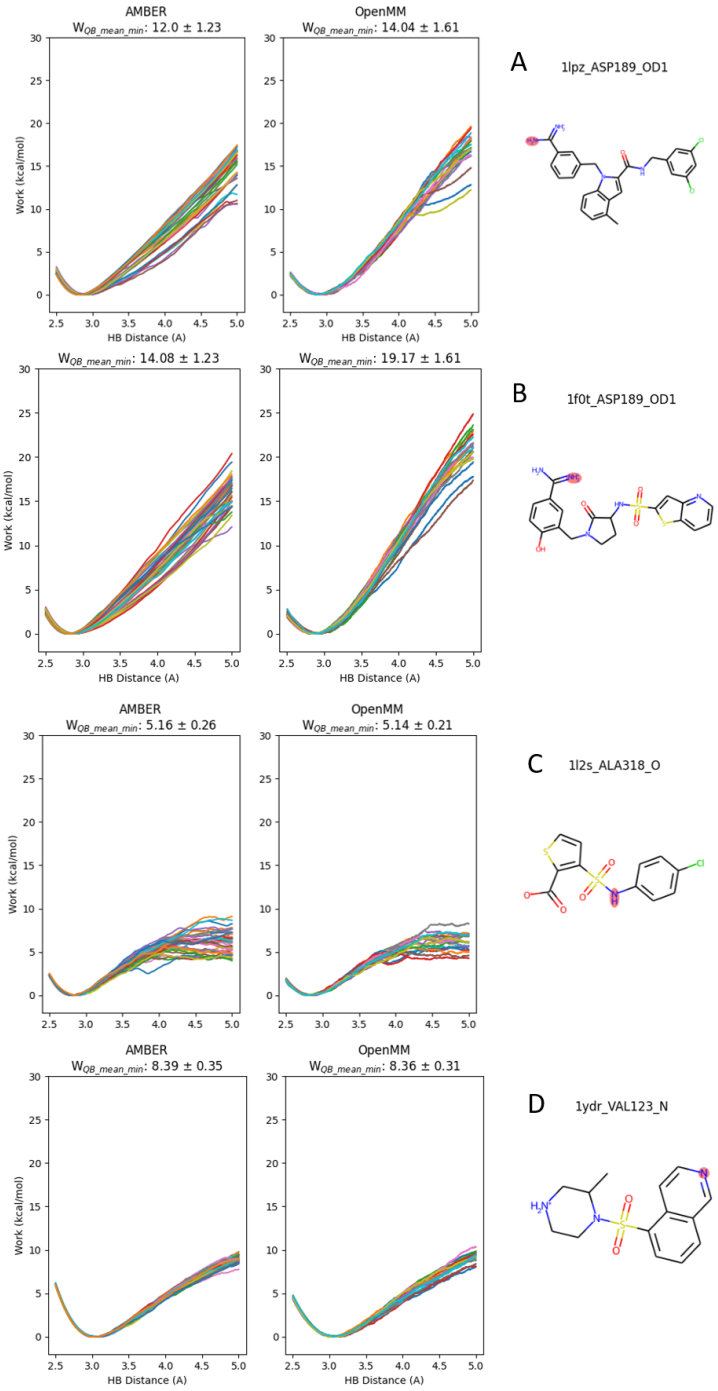


Figure 3.5: W_{QB} traces for the DUck (AMBER) and OpenDUck (OpenMM). A, B. Traces for the remaining charged nitrogen interacting atoms in the data-set. C. The case for a ligand with an uncharged nitrogen donor atom. D. A case with a neutral nitrogen acceptor atom.

using the same force fields and parameterisation [65]. They found that all of the computed energies agreed to 4 significant figures, with the nonbonded interactions showing the largest differences. This was attributed to the larger magnitudes of these energies, as well as to the different ways in which the two programs compute nonbonded interactions (AMBER uses 4^{th} order splines for charge spreading, while OpenMM uses 5^{th} order) [65]. However, this study looked only at systems containing protein and solvent molecules; ligands were not included. In addition, neither of the studies discussed looked at the effects of the different engines on steered molecular dynamics methods.

The excellent correlation between the DUck-AMBER and OpenDUck-OpenMM implementations shown in 3.3 was encouraging enough to proceed with the next phase of the validation - comparing the full workflows, starting from the parameterisation input, with the caveat that charge-reinforced interactions involving NH^+ atoms may be slightly overestimated.

3.3.3 Comparison of the full workflows

Apart from using a different MD engine, the OpenDUck workflow also uses a different force field for parameterising the ligand, and a chunking script that uses different libraries and software (available as Python Conda-distributed packages), despite employing the same underlying logic. To investigate the effects of these changes on the observed W_{QB} values, as well as to test the full OpenDUck workflow from input structure to final value, OpenDUck was run on the interactions measured by Majewski and colleagues in their 2019 publication on structural stability [134].

3.3.3.1 Methods

The full protonated protein (.mol2) and ligand (.sdf) files were downloaded as supplied in the Supplementary Information of Majewski *et. al.,* 2019 [134]. As the OpenDUck chunking and preparation scripts could not easily handle the .mol2 input files for the protein, these were converted to .pdb using PyMOL [178], which also retained the AMBER-style residue naming. The same key interactions were defined as those used by in the 2019 publication. OpenDUck was then run using the same parameters as those used in the original paper, shown in Table 3.2. The value of 20 SMD cycles means that a total of 40 SMD runs were performed per interaction, as two runs (one at 300 and one at 325 K) are initiated from each MD checkpoint. Due to the large number of interactions in the validation data sets, chunks were generated automatically as part of the OpenDUck workflow. However, a visual comparison of the chunks produced by OpenDUck and those generated by DUck's MOE script showed that the OpenDUck chunks tended to be smaller when residues within the same radius (7 Å) were selected as input for the OpenDUck chunking script. For this reason, chunks were generated with a cutoff of 9 Å (shown in Figure 3.6), unless otherwise specified. W_{QB} and $W_{QB\_mean\_min}$ values were calculated as described in Section 3.3.2.1.

Table 3.2: OpenDUck input parameters used in the validation of the full OpenDUck workflow on the Iridium and SERAPhiC data-sets. A move during the SMD stage lasts 200 fs.

Duck parameter	**Value**
equil_len	1 ns
force_constant	50 kcal.mol^{-1}Å^{-2}
cutoff	9.0 Å
distance	2.5 Å
init_velocity	0.00001 Å/move
md_len	0.5
num_smd_cycles	20

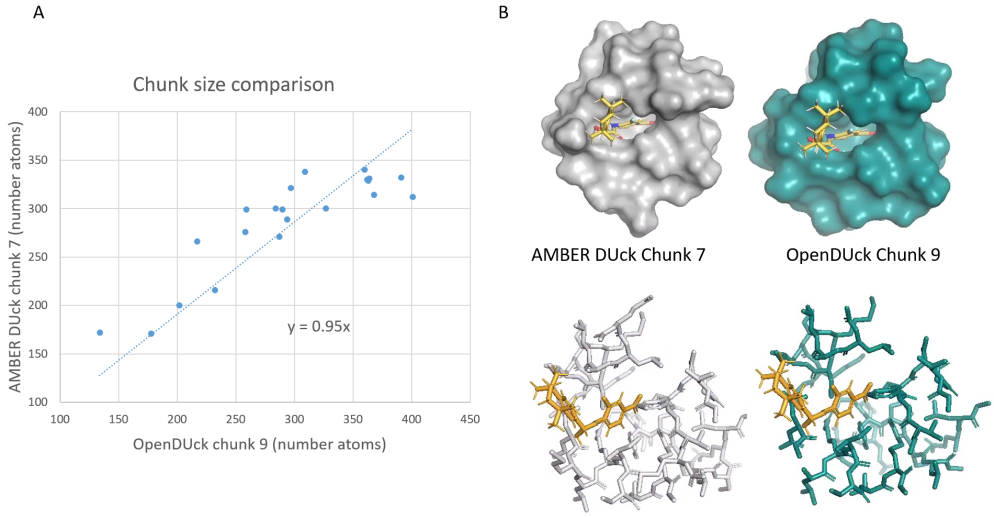


Figure 3.6: Comparing the size of the chunk between AMBER DUck and OpenDUck. A. Scatter plot of the number of heavy atoms in the AMBER chunks (cutoff of 7 Å) and the OpenDUck chunks (cutoff of 9 Å). B. Visual comparison of the chunks generated for 1exa_SER289_OG by the two pipelines.

3.3.3.2 Results

The comparison between DUck and OpenDUck $W_{QB\_mean\_min}$ values for 307 interactions in the Iridium data set is presented in Figure 3.7, and that for 109 interactions in the SERAPhiC set - in Figure 3.8. The values used to generate both of these plots can be found in the Supplementary Data for this thesis (in the files "OpenDUck_validation_iridim.csv" and "OpenDUck_validation_seraphic.csv"). In both data sets, the majority of interactions showed good correlation between the two methods, with the R^2 value greater than 0.75. It is lower than the correlation in the previous section (0.92 in 3.3), as further sources of divergence have been introduced: a different chunking procedure, and different protein and ligand parameterisation. In both data sets, outlying cases in which OpenDUck under- or over-estimated the W_{QB} value were further investigated.

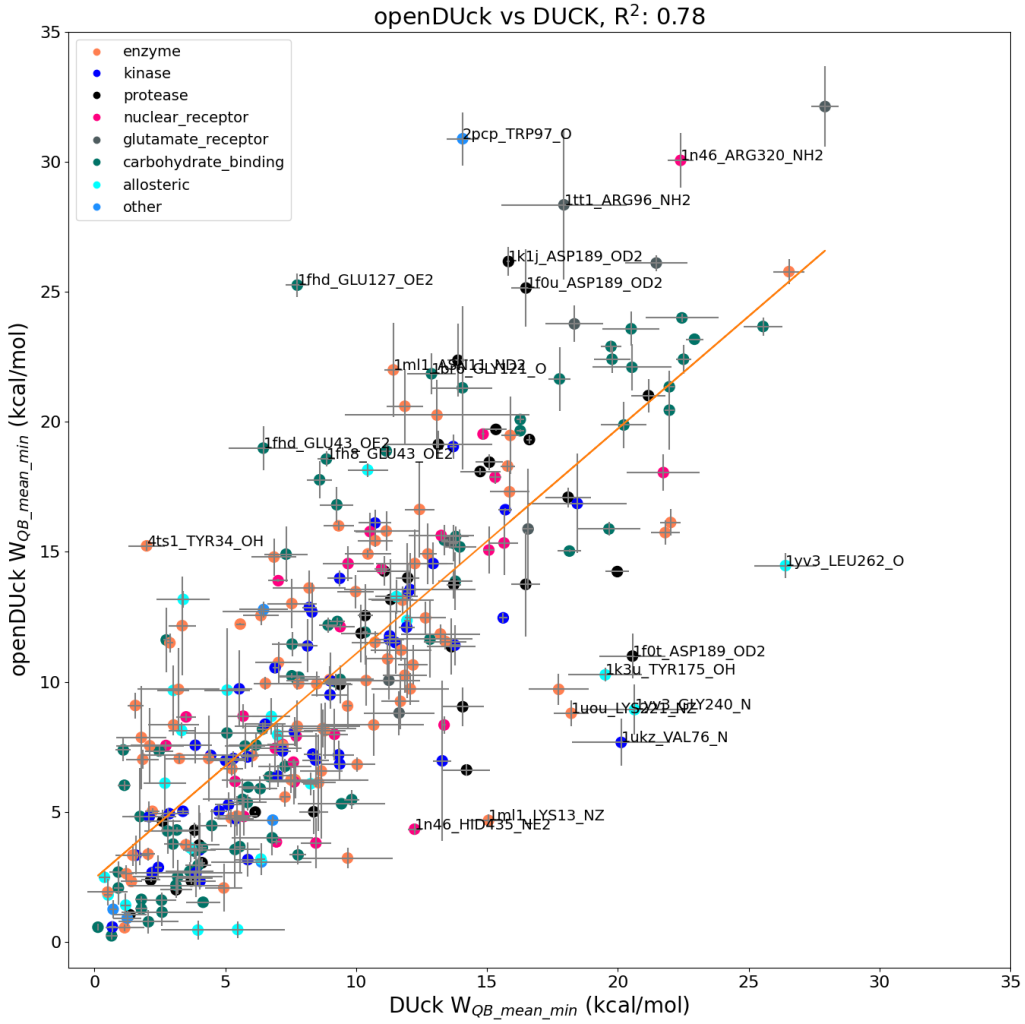


Figure 3.7: OpenDUck validation on the Iridium dataset. Error bars show the standard deviation of the $W_{QB\_mean\_min}$ values, calculated as described in section 3.3.2.1. Outlying interactions are annotated with the interaction names. All interaction points are colour coded based on protein domain type.

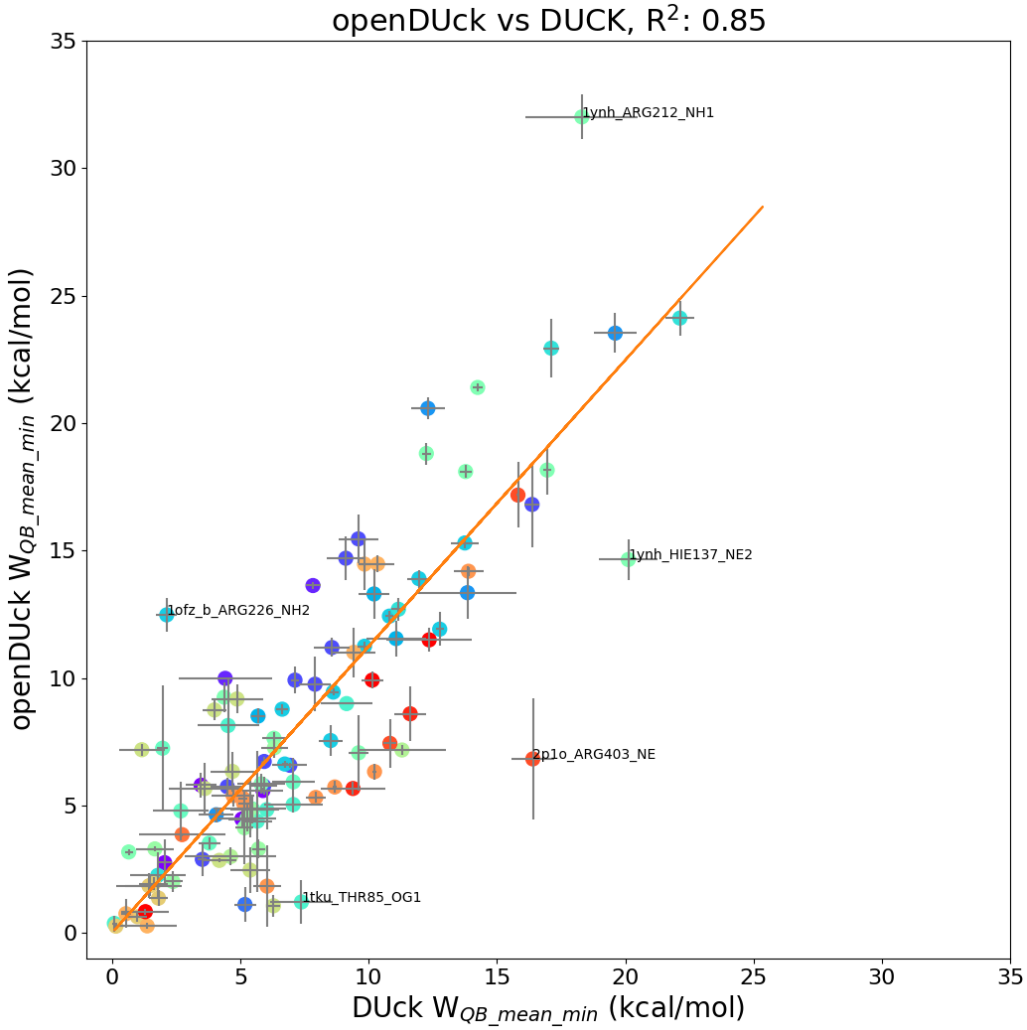


Figure 3.8: OpenDUck validation on the SERAPhiC dataset. Error bars show the standard deviation of the $W_{QB\_mean\_min}$ values, calculated as described in section 3.3.2.1. Outlying interactions are annotated with the interaction names. The interactions are colour coded based on protein target.

3.3.3.3 Cases in which OpenDUck underestimates W_{QB}

Figure 3.7 shows that the outliers in the Iridium set have contributions from various protein target families, with no single protein family being responsible for the observed discrepancies. Of the cases where OpenDUck has underestimated the stability of the interaction, cases like 1n46_HID435_NE2 warrant further investigation, as they show discrepancies in the classification of interactions as structurally stable or unstable between OpenDUck and the DUck benchmark.

A closer inspection of such outliers revealed that the major cause of the observed discrepancies are problems with the automated chunk generation: cases in which the chunking process has changed the local environment of the interaction. In the case of 1n46_HID435_NE2, shown in Figure 3.9, the chunking process opened a solvent channel that allowed water molecules to compete with the interaction, as well as removed the residues that make hydrophobic contacts stabilising the ligand's isopropyl group. Chunk solvent exposure was also a problem with the other outliers in this group, most clearly visible in 1f0t_ASP189_OD2, 1yv3_GLY240_N, 1k1j_GLY219_N, and 1yv3_LEU262_O. The fact that chunks that are too solvent exposed can give rise to artificially lower W_{QB} values has been previously addressed in the publications describing and using the DUck method [47, 128, 134, 137], with the recommendation that all chunks should be manually inspected to prevent such cases.

In the SERAPhiC data set, a similar case was observed among outliers where OpenDUck has underestimated the W_{QB} values. 2p1o_ARG403_NE has a channel open up next to the carboxyl group that participates in the key interaction. One of the other outliers, 1tku_THR85_OG1, presents a case in which contacts that are made as the ligand exits the binding site lower the work needed to 'pull' the ligand, and so contribute to a lower overall W_{QB}. The final outlier, 1ynh_HIE137_NE2

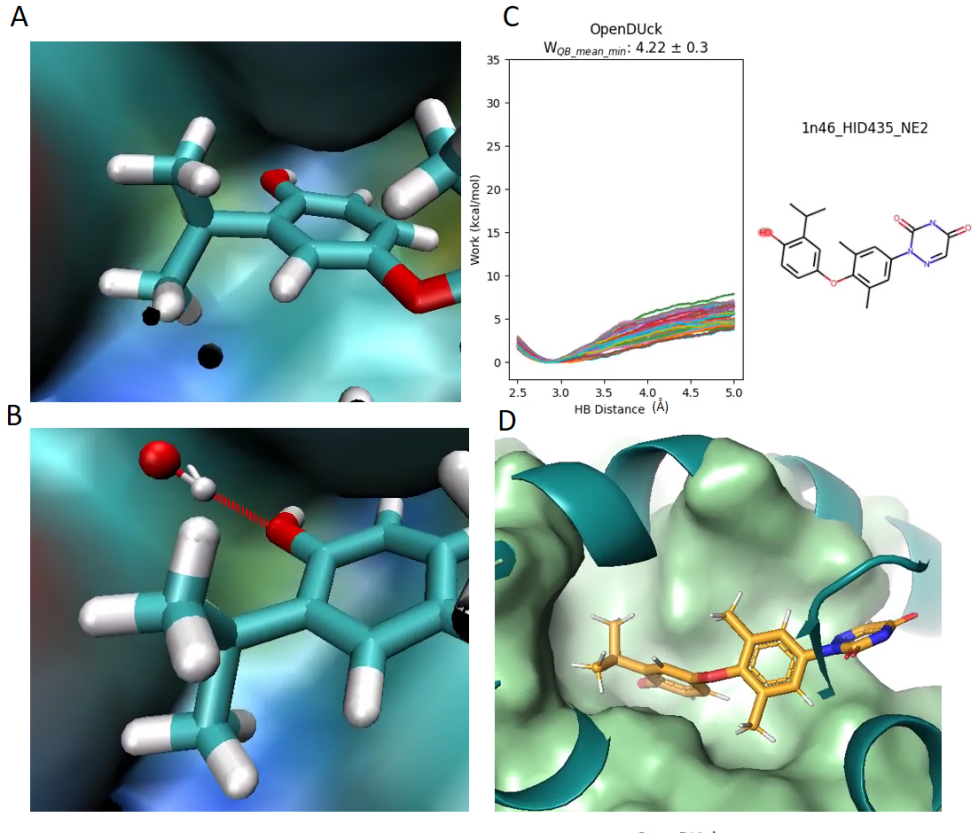


Figure 3.9: Effect of solvent exposure on the chunk in interaction 1n46_HID435_NE2. Panels A and B show a close up surface representation of the protein chunk and a licorice representation of the ligand at the start (around 2.7 Å) and later half (around 4.4 Å) of one of the SMD runs in the interaction. Section B shows a water molecule making a hydrogen bond with the key atom involved in the interaction. The ligand's exit vector is towards the viewer. Panel C shows the work traces and the 2D structure of the ligand, with the interacting atom highlighted in red. Panel D shows the chunk as a light green surface and the full sized protein as a teal cartoon. In the full experimental protein structure, the channel through which the waters enter is blocked, but does not interfere with the ligand's exit vector.

combines the sources of discrepancy, with both increased solvent exposure and stabilisation through interactions made by a different part of the ligand as it exits the site.

3.3.3.4 Cases in which OpenDUck overestimates W_{QB}

Among the outlying cases where OpenDUck provides a higher W_{QB} value, chunking problems were again found to be the main culprit. In the majority of cases, the topology of the chunk includes residues that interfere with the ligand's exit vector. Re-running these interactions with a smaller chunk size (but not small enough to introduce solvent exposure) then led to $W_{QB\_mean\_min}$ estimates that were similar to the reference values in provided by Majewski *et al.* [134]. Figure 3.10 shows the results of these experiments for two of the most extreme outliers: 4ts1_TYR34_OH and 2pcp_TRP97_O.

Table 3.3: Effect of the chunk blocking the ligand exit vector for interactions in 1fhd, 2pcp, and 4ts1. OpenDUck_9 refers to chunks generated with a cutoff 9 Å around the interacting protein atom, and OpenDUck_7 - for 7 Å. Reference DUck values were taken from [134]. All units for the DUck and OpenDUck runs are in kcal/mol

Interaction	**OpenDUck_9**	**OpenDUck_7**	**DUck**
1fhd_GLU43_OE2	18.3 ±1.1	9.0 ±1.4	6.5 ±1.3
1fhd_GLU127_OE2	25.2 ±0.5	8.6 ±0.4	7.8 ±0.5
1fhd_TRP273_NE1	15.6 ±0.2	14.6 ±0.8	13.7 ±0.2
1fhd_HID80_NE2	24.5 ±0.3	21.0 ±1.1	22.4 ±0.3
2pcp_TRP97_O	29.7 ±0.8	15.1 ±1.3	14.1 ±0.6
4ts1_TYR169_OH	19.5 ±1.2	18.0 ±0.7	15.9 ±0.4
4ts1_TYR34_OH	15.3 ±0.1	5.2 ±0.6	2.0 ±0.7

In both cases, there is a difference in the shape of the work profiles in addition to the reduction of the $W_{QB\_mean\_min}$. Table 3.3 shows the comparison with the reference DUck values (as reported in [134]) - in all cases, removing the blocking residues greatly improved the agreement between the methods. Smaller chunks

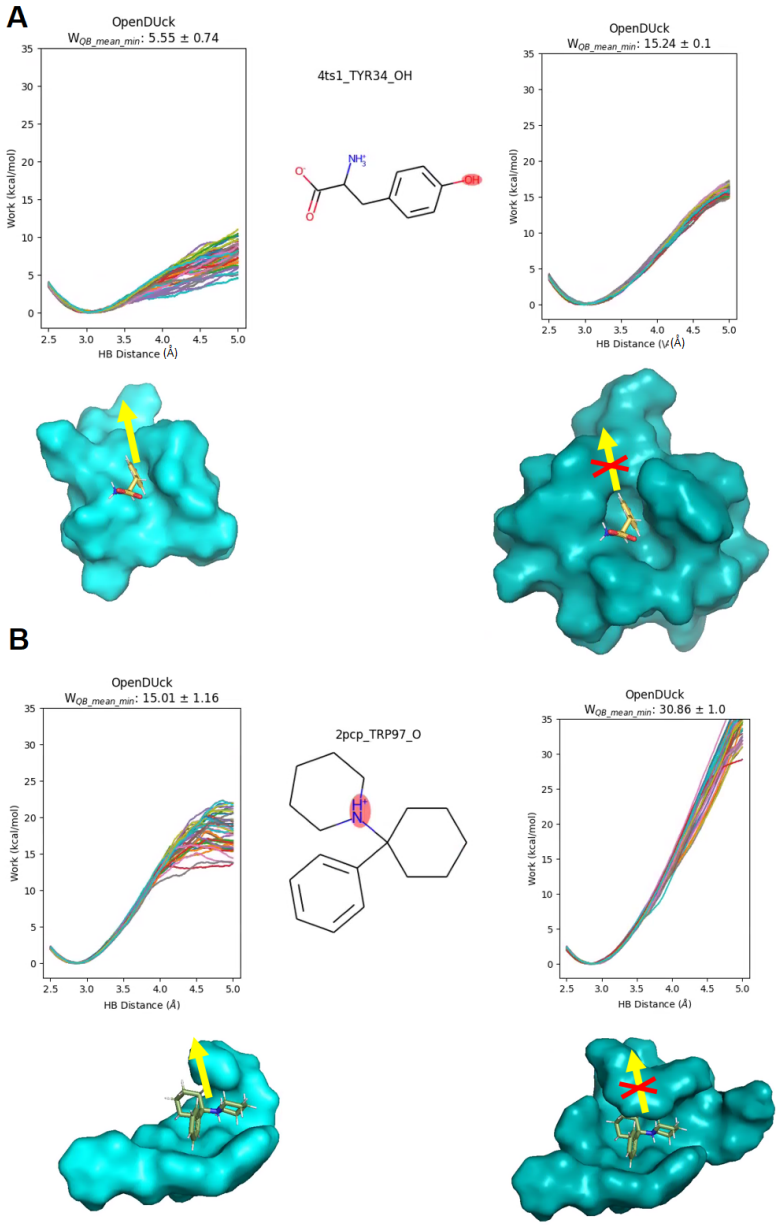


Figure 3.10: Effect of the chunk blocking the ligand exit vector.

were also generated for interactions where the methods showed agreement within these systems (1fhd_TRP273_NE1, 1fhd_HID80_NE2, 4ts1_TYR169_OH) in order to show that this lowering is not merely the effect of increasing the solvent exposure of the chunk.

3.3.3.5 Comparison of the full workflows: a virtual screening perspective

In this chapter, OpenDUck was considered as a complementary computational method to the extended fragment hotspot maps, in the context of fragment elaboration. However, a highly successful application of the original DUck method was found in virtual screening [47]. As an open-source implementation, OpenDUck should perform comparably in the task of identifying structurally stable interactions. A quick analysis of the of the Iridium and SERAPhiC data, presented earlier in this section (3.3.3) is shown in Figure 3.11.

The structural stability threshold of 6 kcal/mol, introduced in [47] was taken as the cutoff for structurally stable interactions. The $W_{QB\_mean\_min}$ values, shown in Figures 3.7 and 3.8 were used in this analysis. The results show that for both datasets, the full workflows are in agreement for over 80 % of the interactions (84.2 % for Iridium and 81.5 % for SERAPhiC). This is an encouraging initial result, especially considering that the chunking discrepancies discussed in Sections 3.3.3.3 and 3.3.3.4 contributed a large part of the false positive and false negative results. In a virtual screening scenario, the same chunk would be used with multiple ligands, so the differences in chunking would not be as problematic. In addition, the error of the $W_{QB\_mean\_min}$ values was not taken into account for this analysis: for example, an interaction with OpenDuck value of 5.9 and an AMBER DUck value of 6.0 would be considered a false negative in this analysis, even if the values are within a standard deviation of each other. This simulates a "worst case" scenario, in which the two workflows were still found to have good agreement.

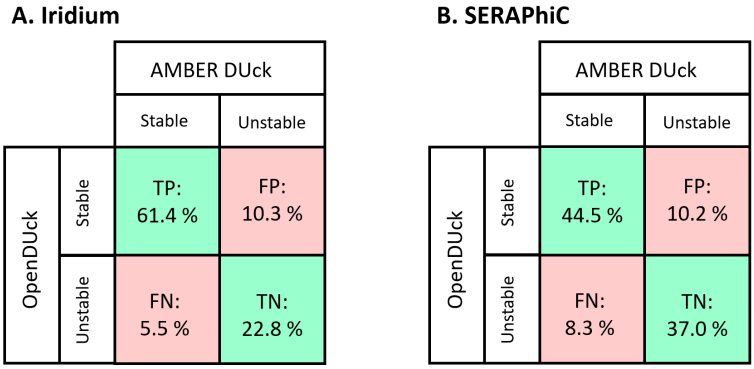


Figure 3.11: Comparison of the full workflows: a virtual screening perspective. Confusion matrices for the agreement between AMBER DUck and OpenDUck for the Iridium (A) and SERAPhiC (B) datasets in classifying interactions as structurally stable. The cutoff value used was 6 kcal/mol. $W_{QB\_mean\_min}$ values for the two methods are the same as shown in Figures 3.7 and 3.8.

As OpenDUck was not envisioned to be used for virtual screening in the workflow presented in Chapter 4, this analysis was not pursued further. However, with a consistent chunking procedure and a way of taking the $W_{QB\_mean\_min}$ standard deviation into account, OpenDUck would likely be a good open-source option for virtual screening with dynamic undocking.

3.3.3.6 Computational performance of OpenDUck

Dynamic undocking is an MD-based method, and so involves a greater computational cost compared to both the hotspot maps and docking. Therefore, it was important to make sure that the open-source implementation has a comparable computational cost to the original. An OpenDUck run with extensive sampling (20 SMDs/ temperature, amounting to 40 SMD runs total and 20 unsteered MD runs in between) requires 2-4 hours (depending on chunk size) on an NVIDIA Tesla p100 GPU. Computational times for the original DUck implementation were reported as 24 minutes/ nanosecond for unsteered MD, and 30 min/ nanosecond (SMD) on a Titan X GPU [47], which is comparable. The validation protocol

used for the OpenDUck saved trajectory snapshots every 2 ps, which is likely to slow down the GPU computation. This could also contribute to the slower times observed when running OpenDUck on an NVIDIA RTX2080 Ti GPU (about 3x slower than the p100), which is a difference compared to that reported in the official OpenMM benchmark (https://openmm.org/benchmarks).

Both AMBER and OpenMM have been benchmarked on the Joint AMBER/CHARMM cellulose dataset using an NVIDIA v100 GPU (OpenMM: https://openmm.org/benchmarks, AMBER: https://hpc.nih.gov/apps/amber/). In the test case, AMBER outperforms OpenMM with 90 ns/day to 65 ns/day. However, we see that AMBER performance drops on two GPUs whereas OpenMM performance grows gradually as usage of the tensorcore and cache memory differs.

Overall, the current performance of the two workflows is comparable. A more extensive comparison, using the same GPU and varying parameters such as the target protein and chunk size, could in future give a more detailed performance benchmark.

3.4 Development of a diagnostic for OpenDUck

Overall, the DUck and OpenDUck methods showed encouragingly good agreement, once issues arising from the chunk topology had been addressed. Creating an appropriate chunk has previously been identified as being key to the success of the method [47, 128], with manual chunk inspection being recommended. However, visually inspecting the individual chunks is not feasible when running dynamic undocking as part of an automated pipeline. In addition, users who are unfamiliar with the technical subtleties of the method but wish to use it as part of a follow-up suggestion and evaluation workflow could benefit from a set of

diagnostics for the output W_{QB} values.

Considering previous work and publications on OpenDUck, as well as the validation studies presented in Section 3.3.3, a useful set of diagnostics would be able to do the following:

1. Identify outlying trajectories when the final W_{QB} or $W_{QB\_mean\_min}$ is calculated.
2. Identify when the chunking process has introduced additional solvent exposure in the vicinity of the key protein-ligand interaction.
3. Automatically detect whether any residues interfere with the ligand's exit vector.

Of these, the first point would be fastest to implement and be of immediate help to new users. Implementing the second and third points will be discussed in Section 3.6.2.

3.4.1 Identifying outlying trajectories

In the original DUck implementation, the minimum W_{QB} value for the set of SMD runs was taken as the stability score for the interaction (W_{QB_min}), in order to avoid overestimating the overall stability due to insufficient sampling [47]. Majewski and colleagues then introduced $W_{QB\_mean\_min}$ as a way to provide further statistical information for the W_{QB_min} value [134]. However, in cases such as those shown in Figure 3.12, a way of automatically identifying systems with unusual work profiles or outlying trajectories would greatly facilitate the analysis, especially when the method is run on a large number of different structures and interactions, as was the case with the OpenDUck validation.

A "good" set of work profiles is shown in Figure 3.12, Panel A. The shape of

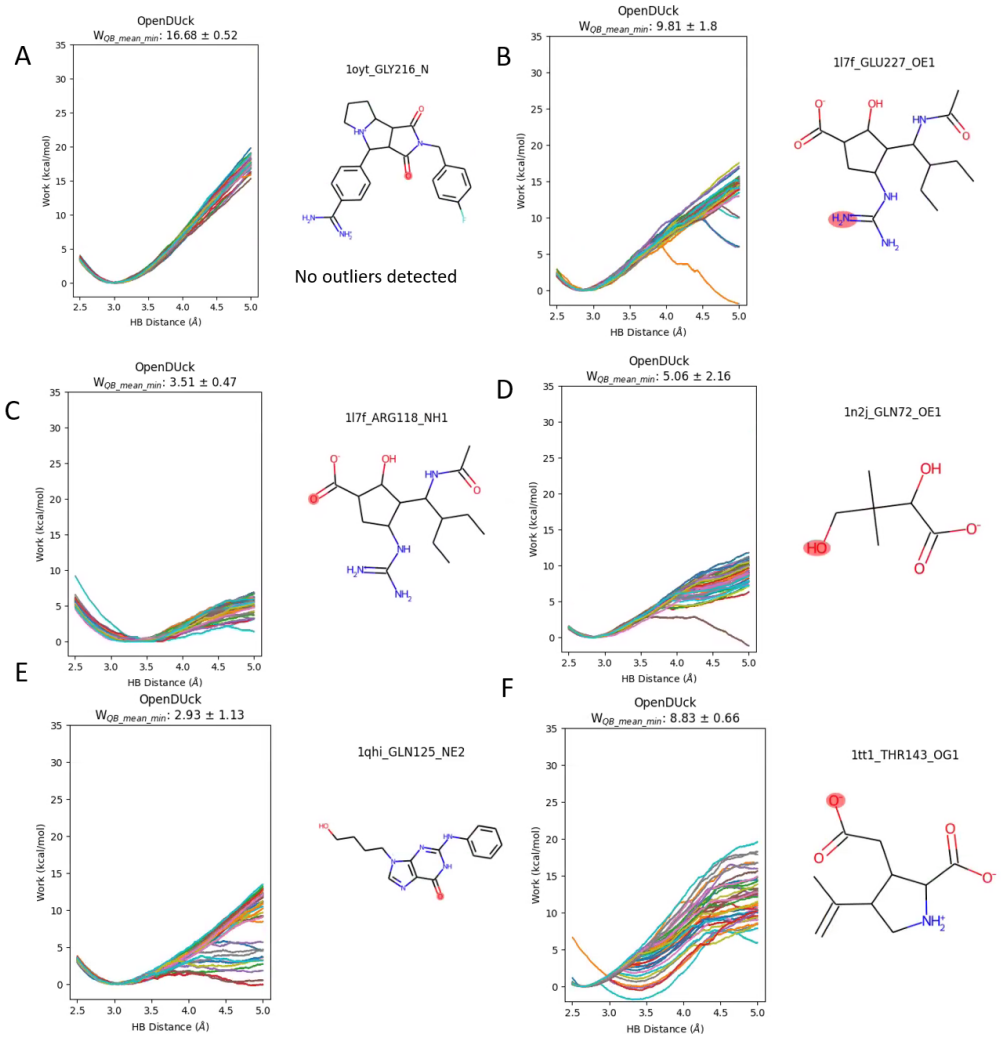


Figure 3.12: Interactions with outlying trajectories in the work profile. Panel A shows an 'ideal' trajectory, for which no outliers should be detected by the diagnostic. The rest show different types of outlying trajectories (in the starting, minimum or final work value) observed in the data set. Ligand atoms involved in the key interaction are highlighted in red

the traces shows a clear minimum around 3Å, and the work steadily increases as the interaction is being pulled, with the values measured at 5Å being in good agreement. Panel B in the same figure shows a subset of trajectories that exhibit a sharp peak, followed by a steady drop in the work value. This is the result of the interacting ligand atom rotating away from the protein atom during the course of the steered MD. The W_{QB} value in such cases is taken as a maximum work (relative to the minimum work value) observed along the trajectory. Panels C and D show cases with single outlying trajectories (panel C in the case of the starting work, and panel D - in the final), which warrant further investigation. Panel E shows trajectories that are split into roughly two groups, in which case automatically identifying trajectories that group together would be useful. Finally, panel F shows a set of profiles with two outlying trajectories in the start and minimum work, and a wide spread of final energies.

To identify outlying cases, three regions of the work trace were monitored for each set of SMD runs (shown in Figure 3.13, panel A): the starting work (W_{start}, defined as the work values measured at the start of the SMD run minus the minimum work observed along each trajectory), the minimum work values observed in the trajectories ($W_{minimum}$), and the final work values (W_{final}). The last value is not necessarily the same as the W_{QB} for a run, as in the example in 3.12, Panel B: for the outlying yellow trace, the W_{QB} will be around 5 kcal/mol (around the location of the peak), while W_{final} is negative.

The values at each of these 3 points were collected for all the SMD runs for a particular interaction. A peak finding algorithm (as implemented in Python's SciPy module [179]) was used to detect peaks in the distribution of values. Peaks were required to be at least 3 kcal/mol apart (governed by the 'distance' parameter in scipy.signal.find_peaks()), If a single peak was detected, outliers were defined as values that are more than 2 standard deviations away from the median value. If

multiple peaks were detected, the largest number of associated values is taken as the reference peak. Values that were more than 2 standard deviations away from the maximum peak value were then classified as outlying. This procedure was implemented as a Python script, which can be accessed at https://github.com/mihaelasmilova/duck/tree/master/scripts/trace_diagnostic.py. Table 3.4 shows the output of the script when applied to the interactions shown in 3.12. As can be seen in the case of 1l7f_ARG118_NH1 smd_9_325.dat, a trajectory may be detected as outlying in more than one part of the trace: this particular trajectory has outlying starting, minimum, and end points.

In the case of 1qhi_GLN125_NE2, a visual inspection of trajectories from the outlying cluster shows that they start from a different position (after the unsteered MD step) compared to the main cluster. The main cluster overall has higher W_{QB} values, as additional hydrogen bonds are present at the start of the SMD run to the oxygen atom of the glutamine residue, as well as to the adjacent ARG 176 (Figure 3.13, panel C). These stabilise the binding mode as the ligand is pulled along the reaction coordinate during the SMD run. In the outlying trajectories, these additional bonds are not present at the start of the SMD run as the ligand has rotated away during the unsteered MD (Figure 3.13, panel D). As fewer contacts are available to stabilise the ligand, the W_{QB} values for these runs are lower. The fact that such outlying trajectories can be identified, grouped, and visualised greatly aided the interpretability of the W_{QB} scores. In a prospective setting, such observations can lead to hypotheses about the contribution of individual interactions towards the stability of the binding mode.

Overall, developing a script for fast outlier detection in the OpenDUck work profiles greatly aided the analysis of trajectories in the validation set, as well as cases in which it was applied to unseen data, as was the case in the following section 3.15.

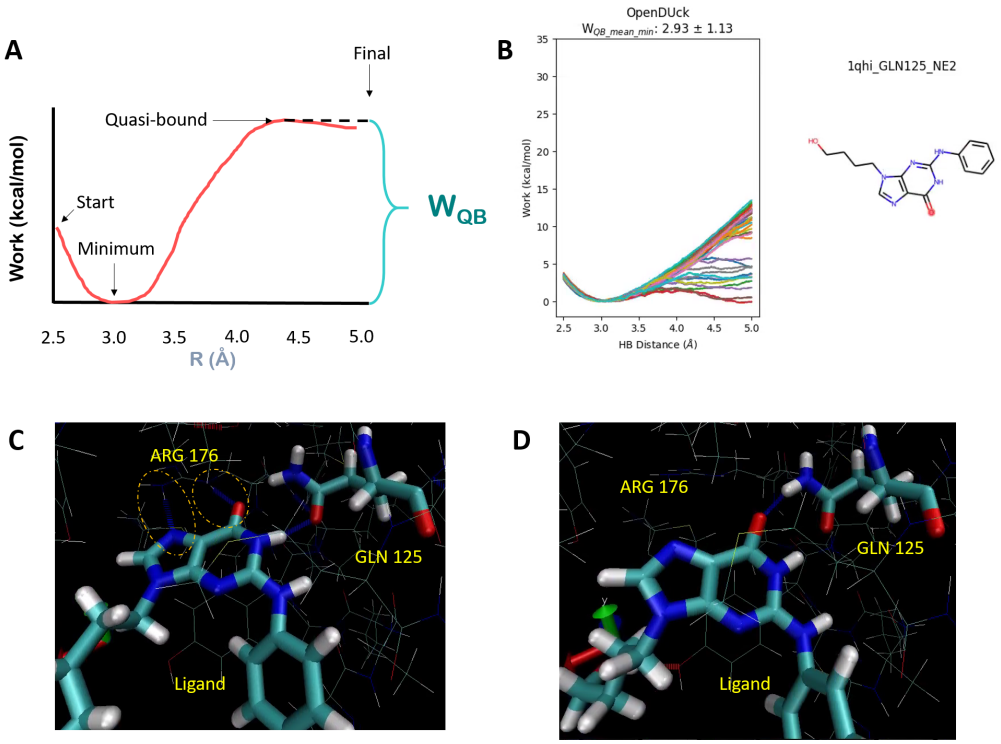


Figure 3.13: A. A typical Duck work profile, with the start, minimum, quasi-bound and final points annotated. B. The W_{QB} traces for 40 SMD runs of 1qhi_GLN125_NE2 form two clusters at the W_{final} point. C. The starting position for one of the SMD runs of the high-scoring cluster for 1qhi_GLN125_NE2. The interacting residue and ligand are shown as licorice, and the rest of the chunk is shown as lines. D. The same view as Panel C, showing the starting position of a trajectory from the low-scoring cluster, in which the ligand makes fewer hydrogen bonds with the chunk.

Table 3.4: Detected outlying trajectories for the interactions presented in 3.12. Example of the output from diagnostic script described in Section 3.4.

Interaction	Trajectory	Outlier Type	Distance to peak/median
1l7f_GLU227_OE1	smd_0_325.dat	W_{final}	-16.16
1l7f_GLU227_OE1	smd_2_300.dat	W_{final}	-8.37
1l7f_GLU227_OE1	smd_5_300.dat	W_{final}	-8.31
1l7f_ARG118_NH1	smd_9_325.dat	W_{start}, $W_{minimum}$, W_{final}	3.53, -3.54, -4.03
1n2j_GLN72_OE1	smd_17_325.dat	W_{final}	-10.18
1qhi_GLN125_NE2	smd_8_300.dat	W_{final}	-8.49
1qhi_GLN125_NE2	smd_15_300.dat	W_{final}	-8.4
1qhi_GLN125_NE2	smd_16_300.dat	W_{final}	-9.38
1qhi_GLN125_NE2	smd_16_325.dat	W_{final}	-12.15
1qhi_GLN125_NE2	smd_17_325.dat	W_{final}	-10.26
1qhi_GLN125_NE2	smd_17_325.dat	W_{final}	-11.55
1qhi_GLN125_NE2	smd_19_300.dat	W_{final}	-8.8
1qhi_GLN125_NE2	smd_19_325.dat	W_{final}	-8.9
1tt1_THR143_OG1	smd_15_325.dat	W_{start}, $W_{minimum}$	6.22, -6.32
1tt1_THR143_OG1	smd_9_325.dat	W_{final}	-6.28
1tt1_THR143_OG1	smd_9_325.dat	W_{final}	7.35

3.5 Applying OpenDUck to fragment follow-up ranking: a retrospective case study

Both the validation and the development of a diagnostic for OpenDUck were undertaken with the end goal of using the method prospectively in the context of a crystallographic fragment screen. There are two main ways in which dynamic undocking could be used in such cases: ranking the initial fragment hits, and ranking prospective follow-up compounds. When paired with fragment hotspot maps, I wanted to determine if dynamic undocking can provide a useful ranking of prospective follow-up compounds that place an appropriate group in a hotspot that is not exploited by the initial fragment hit.

As described earlier in Sections 1.7.3.3 and 1.7.3.2, extensive work has previously been undertaken on the ways in which fragments can be prioritised using DUck [47, 120, 137, 45, 80]. The computational pipelines employed by Rachman and

Piticchio also showed how the method can be used for scaffold hopping, resulting in follow-ups with improved structural stability and a better fit to the binding site. To gain a further understanding of how dynamic undocking can be used in follow-up design, a retrospective case study involving ligands with high structural similarity and large differences in binding affinity was investigated. CDK2 was chosen as the target system, as it has previously been shown to have a suitable binding site that forms clusters of stable interactions, specifically with the kinase hinge residues [134].

3.5.1 Methods

3.5.1.1 Dataset

The set of compounds used in this case study was taken from Schonbrunn *et al.,* 2013 [180]. They form part of a highly potent diaminothiazole series of CDK2 inhibitors developed from a single fragment-like weak binder (IC_{50} = 15 μM) identified in a high throughput screen. The most potent compounds in the set showed 1000-fold improvement in potency; in addition, co-crystal structures were available for the starting compound, as well as for 35 closely related analogues. Of these, only compounds that included additional polar substituents making new hydrogen bonds with the protein were included in the dynamic undocking experiments, along with two "control" compounds that do not (3RKB and 3QTW in Figure 3.14). The final set of compounds used is shown in Figure 3.14. Both the starting fragment and the follow-up compounds form three conserved H-bonds with the hinge region of CDK2 (residues Glu 81 and Leu 83). All of the follow-up compounds have an additional acceptor nitrogen that can hydrogen bond to Lys 33, and three of the compounds make further polar interactions with Lys 83 and Asp 86.

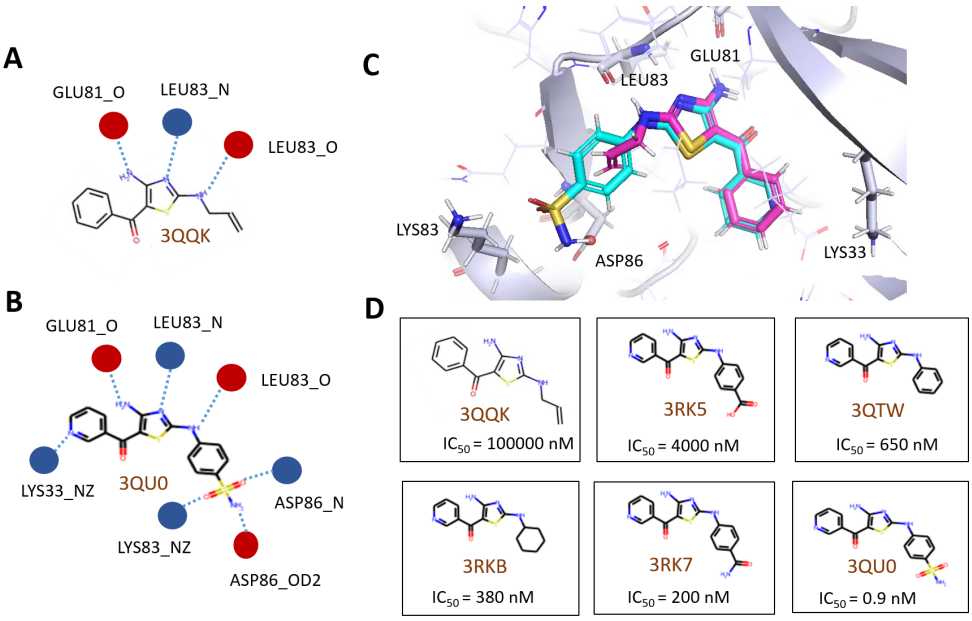


Figure 3.14: Dataset used in the CDK2 retrospective case study. Panel A shows the structure of the initial high throughput screening hit (PDB ID 3QQK for the protein complex) and the key interactions it makes with the kinase hinge residues (Leu 83 and Glu 81). Panel B shows one of the most potent follow-up compounds,in PDB ID 3QU0, which makes additional H-bond interactions with Lys 33, Lys 83 and Asp 86. Panel C shows the 3QQK ligand in magenta and the 3QQK receptor in pale blue. The 3QU0 ligand is shown in cyan. Interacting residues are shown as sticks. Panel D shows the structures, PDB codes and binding affinities (as reported in [180]) of the compounds included in the data set for dynamic undocking with OpenDUck.

3.5.1.2 Structure preparation

Structures were downloaded from the PDB and protonated using Protoss [158]. The CSD Python API [110] was used to separate and save the ligand and receptor files. To mimic a follow-up design scenario, the follow-up compounds were overlaid and complexed with the starting fragment (3QQK) receptor. The resulting complexes were not minimised prior to the OpenDUck runs, but the dynamic undocking workflow contains a minimisation step, as described in section 3.2.4.

3.5.1.3 Dynamic undocking

Interactions were automatically detected using the ODDT Python Toolkit [181]. OpenDUck runs were performed as previously described, using a chunking cutoff of 10 Å, as smaller cutoffs created very open chunks in the region of residues Lys 83 and Asp 86. To recreate the case in which prospective follow-up compounds are being assessed, 6 SMD runs (3 at each temperature) were run for each interaction. The minimum W_{QB} was taken as the stability score of the interaction. With this number of runs, any low-scoring interactions can be disregarded as structurally unstable, allowing higher scoring interactions to be prioritised for further investigation. Interactions that showed W_{QB} values above the structural stability threshold of 6 kcal/mol [47] were further simulated to 20 SMD runs/temperature.

3.5.2 Results

A key feature of structural stability is that it is orthogonal to thermodynamics-based methods and does not correlate with binding affinity [47]. However, successful follow-up compounds are more likely to have both good structural stability and strong binding affinity. The experiment presented in this section had the goal of exploring two questions. First, whether the conserved interactions made by all

the followups vary in structural stability between compounds. Second, whether the W_{QB} values of additional interactions outside the 3 "hotspot" (hinge) hydrogen bonds can inform on the structural stability of the prospective compound in a virtual screen.

3.5.2.1 The conserved hydrogen bonds show little variation between compounds

Figure 3.15 shows the W_{QB} values measured for all the interactions made by the compounds in the set. The interactions with GLU81_O and and LEU83_N vary surprisingly little, even between compounds with great differences in binding affinity. In all cases, these interactions remain well above the structural stability threshold of 6 kcal/mol [47]. The final hinge interaction, LEU83_O, is most structurally stable in all compounds. This was somewhat surprising, as it is located closest to the opening of the binding site pocket and so has the potential to be most solvent exposed. Curiously, this interaction has the greatest variation in W_{QB} values between compounds. This could be due to stabilising (or destabilising) effects from the adjacent substituent.

All of the compounds presented also have the potential to form a hydrogen bond with Lys 33 through the acceptor nitrogen atom on the pyridine ring. However, in all cases this interaction was found to be structurally unstable (see Figure 3.15). This is also supported by the structural data for these compounds with CDK2; Lys 33 is modelled as facing away from the ligand [180].

The interactions made by the compounds in structures 3QU0 and 3RK7 with Lys 83 are also structurally unstable. This residue is very close to the outside of the binding pocket and is also modelled as facing away from the ligand in some of the structures.

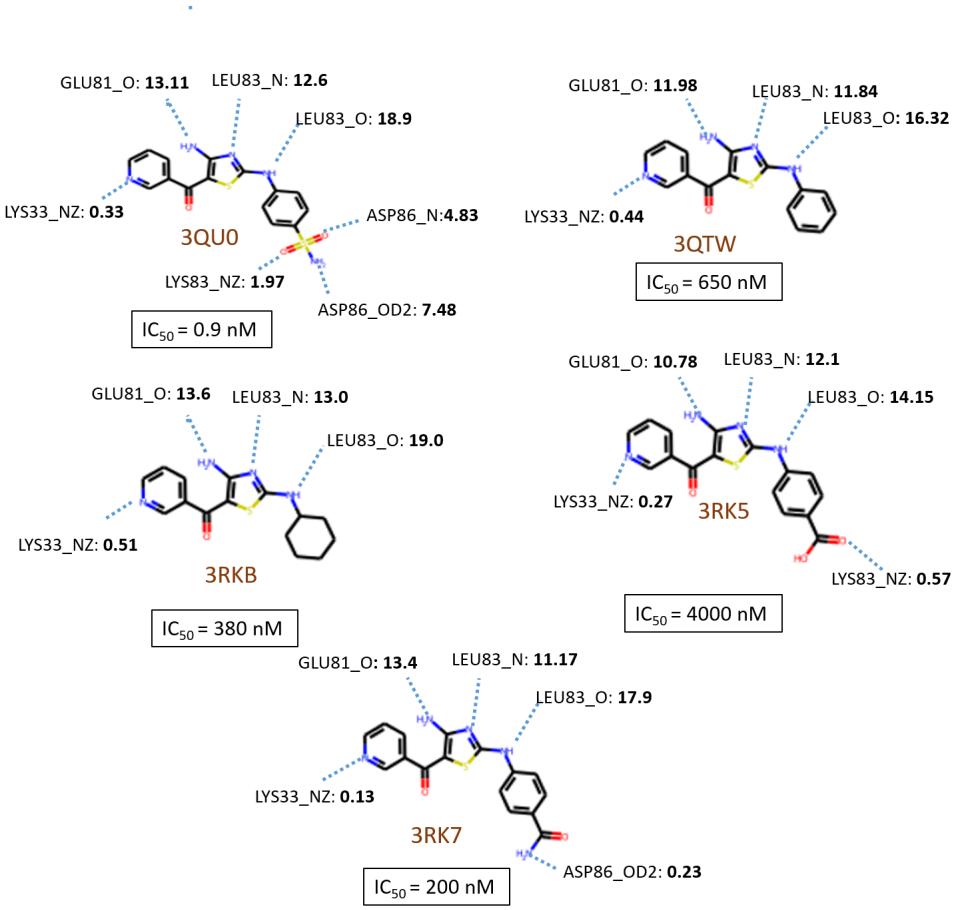


Figure 3.15: Structural stability of interactions in the CDK2 inhibitor set. W_{QB} values are shown as numbers next to the labels of the protein interaction atom. All values are in kcal/mol.

3.5.2.2 Additional hydrogen bonds to Asp 86 show greater variation

The compounds in structures 3QU0 and 3RK7 both form a hydrogen bond with the O2 carboxylate atom of Asp 86. In the case of 3RK7, this interaction is highly structurally unstable (W_{QB} of 0.23 kcal/mol, shown in Figure 3.15). In 3QU0, however, the interaction crosses the structural stability threshold of 6kcal/mol. The sulfonamide group of the compound in 3QU0 also forms more interactions overall with the protein: one of the sulfonamide oxygens interacts with the backbone N of Asp 86 in an interaction that is almost structurally stable. The other makes a contact with Lys 83, which is not, as mentioned in the previous subsection.

3.5.3 Implications for prospective use

Going back to the questions behind this experiment, it appears that for most of the compounds, the key hydrogen bonds do not vary greatly in structural stability, even when the compounds show orders of magnitude difference in binding affinity. Cases in which a key interaction is adjacent to a position where changes are being introduced could show more variation, and could perhaps be used as a way to measure the destabilisation caused by adding large substituents at that location. However, further work would be needed to confirm this, including a greater number of structures across different targets.

Interactions made outside the main hotspot seem to show more variation in this set of highly structurally similar compounds, as shown in the case of Asp86 interaction in 3QU0 and 3RK7. In the 3QU0 case, there may also be a stabilisation effect from the interactions made with the protein. Majewski *et al.* showed that structurally stable interactions tend to form cooperative 'anchor' clusters. In this case, the sulfonamide group in compound 3QU0 forms three interactions with the

protein, two of which are moderately to structurally stable and form a secondary structural anchor, which can be picked up by the method. In a prospective use case, such analysis could be used to prioritise potential follow-ups that make the same interaction with the protein through different chemical groups.

3.6 Discussion

Dynamic undocking was explored as a method that could be used in tandem with the fragment hotspot maps and alongside thermodynamics-based methods. To improve the accessibility of the method and make it more easily extensible, as well as facilitate its integration into workflows for follow-up suggestion, the OpenMM-based implementation of DUck, OpenDUck, was validated.

3.6.1 Validating the OpenDUck workflow

As structural stability is difficult to measure experimentally, the benchmark used for the validation was a set of W_{QB} values that had previously been calculated using the original DUck implementation for the vast number of interactions detected in the Iridium and SERAPhiC datasets.The goal of this validation was to show that OpenDUck performs comparably to the original version of the method. As OpenMM allows the use of input files from various preparation pipelines, the AMBER-parameterised chunks for a subset of structures in the Iridium set could be used as inputs for the OpenMM DUck calculations. In this way, it was possible to isolate the effect of changing the MD engine on the observed W_{QB} values. Encouragingly, the AMBER and OpenMM values showed good agreement overall (R^2 = 0.92). In some interactions involving charged nitrogen atoms, the OpenDUck W_{QB} estimates were consistently higher than the original DUck values. However, these differences did not change whether an interaction was classified

as structurally stable. The fact that for some of the interactions the discrepancy was greater than in other cases (for example, 1k1j_ASP189_OD1 shows a difference of + 15 kcal/mol, while in most of the others the differences are in the order of 3-4 kcal/mol, shown in Figure 3.4) indicates that the effect is to some extent also dependent on the local environment of the interaction.

Once it had been determined that the two MD engines output comparable W_{QB} values, a comparison of the full DUck and OpenDUck workflows was performed using both the Iridium and SERAPhiC interaction sets. Again, the values showed good correlation between the methods (R^2=0.78 for the drug-sized compounds in the Iridium set and 0.85 for the fragment-sized compounds in the SERAPhiC set). The biggest discrepancies came from differences in the automated chunk generation. In the case of AMBER DUck, the chunks are prepared interactively using MOE, and this allows the user to visually inspect the output and correct for all the recommendations outlined in Section 3.2.1. At the start of this work, OpenDUck did not include visualisations as part of the workflow, and chunking was done using an automated Python script. This script retained the set of residues located within the user-defined cutoff distance from the interacting protein atom and performed checks to ensure the connectivity of the resulting chunks. However, it did not allow for the retention of water molecules, or for the addition or removal of specific residues. The addition of this functionality greatly improved the usability of the method, and was used in the ACVR1 case study presented in Section 4.3.

3.6.2 Challenges and potential solutions for automating the chunking step

Duck is an MD-based method, which requires greater computing power than most desktop workstations can provide. This is why DUck calculations are usually

performed on remote GPU clusters, which generally do not have a graphical user interface (GUI) to allow the user to look at the generated chunks. This means that either the chunks have to be generated locally, manually inspected by the user, and then uploaded to the cluster (this can become cumbersome when running large numbers of calculations), or that a set of automated diagnostics be developed that allow the visual inspection stage to be bypassed. These diagnostics would have to be able to detect cases such as those in Figure 3.10, where residues in the chunk block the ligand exit vector, as well as cases in which the chunking process has introduced solvent-accessible channels that lead to the premature formation of water-mediated contacts between the ligand and protein. These are the types of tasks which, while trivial for users (in small quantities), are difficult to automate.

In the case of residues blocking the ligand's exit vector, a check could be implemented that projects the position of the interacting ligand atom 5 Å away from the interacting protein atom along the vector defined by the two atoms' positions in the crystal structure, and looks for clashing residues in the chunk. This would only provide an estimate, however, as the vector along which the force is exerted changes throughout the course of the SMD run, and the final position of the ligand will likely not correspond to the projection. In addition, a "clash" needs to be defined - would only atoms that physically occupy the same space as the ligand count, or would a radius around it need to be defined? In the converse case, where the chunking process opens up solvent-accessible channels, a possibility would be to look at the change in the solvent accessible surface area between the protein and the chunk in the region of the key interaction. Looking at the full change in solvent accessibility of the chunk would not be informative, as in many cases parts of the protein need to be removed to allow the ligand to be steered away during the course of the SMD simulation. So it is possible for a successful chunk to introduce solvent accessibility on one side of the complex, as long as the local environment

around the key interaction is retained. The definition of a "local environment", however, is not clear in this case, and would likely depend on the target. To set these parameters in an objective manner, a set of rules would need to be compiled based on a wide diversity of targets and interactions. One could also consider using machine learning approaches from computer vision, which have recently seen great popularity in structure-based drug design applications [77]. The Iridium and SERAPhiC datasets provide a large number of interactions and diverse target types, however they do not contain enough data (hundreds of data points) to train and validate such methods(which usually require at least tens of thousands of data points). Consequently, visual inspection of the chunks likely remains the most effective solution for the detection of problems in the chunking step. In the case of evaluating followup compounds, chunks only need to be generated for the key interactions in the binding site (for example those highlighted in the hotspot maps), which greatly reduces the amount of time spent in chunk generation.

To streamline the process, an open-source visualisation (for example, using PyMOL or the NGL viewer) could be developed that displays the protein and ligand structures, as well as the hotspot maps. The user would then input a chunking cutoff, and inspect the resulting selection, adding and subtracting residues and water molecules as needed. Once a chunk has been approved, the inputs for the associated DUck run would be automatically generated and saved in a form that could be easily submitted to the GPU cluster (this is similar to the solution already employed in the original DUck workflow through MOE). The next protein would be automatically loaded into the viewer, without breaking the flow. As this implementation is a software engineering challenge, rather than a scientific one, it was outside of the scope of the thesis, but would constitute a natural continuation of the work presented here.

3.6.3 Strategies to automate the assessment of DUck outputs

Once the DUck simulations are complete, outliers in the calculated trajectories can be detected using the diagnostic script presented in Section 3.4. Trajectories flagged as outlying can be informative in diagnosing chunking and setup problems, or in the detection of unusual unbinding events. For example, an if an interaction has a low W_{QB_min}, but that minimum value is the product of an outlying trajectory, it can be visualised using the VMD script shown in Figure 3.2. The user can then compare this to other trajectories in the set and decide whether this is a realistic unbinding event, or is the effect of errors in the setup or chunk generation steps. In both cases, the VMD script allows for a mechanistic rationalisation to be given for the detected outlying trajectories. To integrate this step in a single workflow, the W_{QB} calculations and outlier detection scripts would be run on the remote cluster, where the simulation data is stored. A shortlist of outlying trajectories would then be compiled and downloaded to the local workstation for visual inspection. Currently, the trajectory visualisations are implemented for the VMD molecular graphics program, as this is a widely used solution that many users would be familiar with. However, both PyMOL and the NGL viewer support trajectory visualisations, so a possible next step would be to implement the visualisation in one of these viewers, allowing all the visualisations to be carried out by a single piece of software. This would further facilitate integration into a computational workflow for fragment elaboration.

3.6.4 Potential usage of OpenDUck in guiding fragment growing campaigns

Once it had been confirmed that OpenDUck performed comparably to the original DUck implementation, the method was used to estimate the structural stability of

the interactions made by a set of follow-up compounds for the CDK2 kinase, for which crystallographic structures and binding data were available. The utility of DUck in ranking chemically diverse fragments that make the same interaction had been previously demonstrated [47], as well as its use in scaffold-hopping follow-up suggestion pipelines[45, 80]. The goal of this experiment was to approach the problem from a different angle: what information could be provided by DUck for interactions made by compounds with a conserved core, as would be the case in a fragment growing campaign?

The results showed that even for follow-up compounds with micromolar binding affinities, the interactions made with the kinase hinge region by the conserved core were structurally stable. Additional interactions made by the compounds generally had lower W_{QB_min} values, and showed greater variation between compounds making the same interaction. The most stable of these additional interactions were made by the sulfonamide group of the compound in structure PDB ID 3QU0. In a prospective fragment growing scenario, the stability of conserved "hotspot" interactions made by the fragment core would then be expected to remain structurally stable. If a prospective follow-up resulted in a lowering of the W_{QB} value for such an interaction, then the introduced change would be considered particularly destabilising. On the other hand, measuring the stability of additional polar interactions made by potential follow-ups could help prioritise compounds that make the same interaction through different chemical groups.

3.7 Conclusion

In this chapter, an open source implementation of the dynamic undocking protocol was validated on two data sets in which structural stability had previously been investigated. There were individual cases where the differences in MD en-

gine between AMBER and OpenMM had an influence on the final W_{QB} values, but they generally do not change the estimate of whether an interaction is structurally stable. The largest source of differences between the workflows was the automated chunk generation step. To further aid with the analysis of the output of the OpenDUck workflow, a diagnostic to detect outlying trajectories of interest was developed. Finally, the OpenDUck pipeline was applied to a retrospective case study of closely related CDK2 inhibitors, showing how the method could be used to prioritise potential follow up compounds.

Dynamic undocking had previously shown success in pipelines for scaffold hopping and chemical space exploration [45, 80]. Therefore, the method could be used to prioritise structurally dissimilar compounds making the same interaction with the protein. In the context of a workflow for elaborating crystallographic screening hits, the method could then be used to automatedly prioritise the starting bound fragments, as this presents a similar problem. Fragment growing presents a slightly different case, in which the suggested follow-up compounds have a conserved core making key interactions, and elaborations make additional contacts with other regions of the binding site. In the case of the CDK2 structures, the core hydrogen bonds were found to be structurally stable for compounds with a wide range of activities. Additional interactions with the protein showed greater variation, raising the possibility of using dynamic undocking as a way of prioritising compounds that make such interactions. This would be particularly useful for prioritising ligands that make a contact that was highlighted by the hotspot maps, but is not exploited by the starting fragments. A workflow combining fragment hotspot mapping and OpenDUck could then provide a novel tool for guiding fragment growth in a rational and semi-automated way.

4 | Development and prospective application of a computational workflow to drive fragment elaboration

4.1 Introduction

In the previous chapters, the fragment hotspot maps method was extended to allow for information from multiple experimental structures of a protein target to be combined, and hotspot selectivity maps were developed to allow the comparison of closely related binding sites. A complementary MD-based method, dynamic undocking, was then introduced as a way of measuring the structural stability of the molecule's binding mode, and the open-source implementation of this method, OpenDUck, was validated. Once these components were available in a format that could allow multiple calculations to run in parallel, they were incorporated into a workflow used to suggest follow-up compounds for three on-going projects at the Centre for Medicines Discovery at the University of Oxford. The protein targets of these projects were the human kinase ACVR1, the SARS-CoV-2 helicase NSP13, and the human protein mono-ADP ribosyltransferase PARP14.

4.2 Overview of the computational follow-up pipeline

Figure 4.1 shows the general workflow used to suggest follow-up designs for the projects. Not all of the steps were used in all of the projects, reflecting the different questions posed in each case. Due to the timelines of the projects, OpenDUck had not yet been fully validated at the time when compounds were ordered, and hence those results were not used in the decision making process. It was run

retrospectively on the selected fragments and follow-up compounds to provide further information on individual interactions.

Fragment hotspot mapping, including selectivity maps and ensemble maps, was envisioned as a way to prioritise the starting fragment hits based on their overlap with the detected interactions. Individual interactions could also be assessed using OpenDUck, allowing for structural stability to be used as a further metric to prioritise the starting hits and the interactions highlighted in the hotspot maps. Together, these two methods would form the "triaging" stage of computational follow-up, described in Section 1.5.1. For the subsequent follow-up enumeration stage, the fragment network [42], implemented as part of the Fragalysis platform [32], was used to provide potential elaborations.

While the previous steps process 3-dimensional information from the crystallographic experiment (a crystallographic fragment pose, having both connectivity and spatial information), the enumeration stage produces 1-dimensional SMILES [182] strings of potential follow-ups.

Constrained docking is then used to convert the 1-dimensional SMILES strings into predicted follow-up binding modes, which would then enter the final, "scoring" phase of the pipeline.

Hotspot maps and dynamic undocking would then be used to score the proposed follow-up poses in a manner similar to the initial fragment hits. The combination of these methods would result in a shortlist of follow-up suggestions with improved predicted fit to the hotspot, as well as high predicted structural stability, and which would present high-quality starting points for further optimisation.

The presented workflow uses methods that have been previously developed. However, a number of modifications and additional filters were introduced to several of these in order to allow their application in this context and tailor them to the

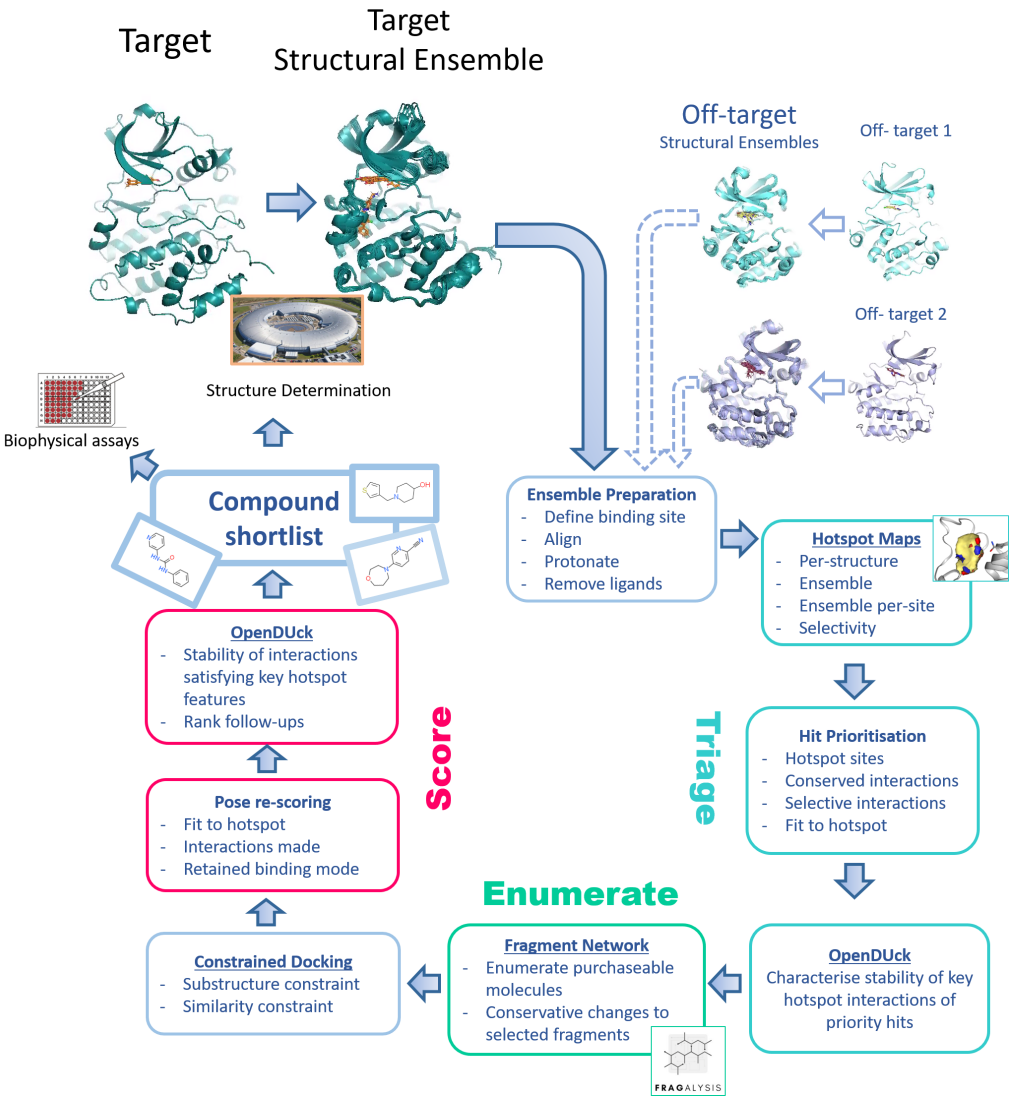


Figure 4.1: Computational pipeline used in the prospective work

specific challenges of crystallographic FBDD.

- To allow a molecule's binding mode to be re-scored using the fragment hotspot maps, a way in which the per-atom scores output by the Hotspots API could be combined into a single, per-molecule score was investigated.
- The elaborations suggested by the graph network should be easily purchased; however, further filters on molecular weight, solubility, drug-likeness, *etc.* were used to triage molecules with unsuitable properties prior to the more computationally expensive follow-up scoring phase.
- A key assumption of FBDD is that the elaborated compound retains the starting fragment's binding mode. While using constrained docking after the enumeration step was performed to ensure that the binding mode of the shared sub-structure was retained, an additional metric for binding mode similarity, SuCOS [61], was used to further triage the output.

The next section describes how these modifications were applied, as well as the default parameters used for all the methods in Figure 4.1. These were the values used for the case studies presented in this chapter, unless otherwise specified in their respective sections.

4.2.1 Methods used in the computational workflow

4.2.1.1 Ensemble Preparation

Structures for the target proteins were downloaded from Fragalysis (`https://fragalysis.diamond.ac.uk/viewer/react/landing/`) and protonated using the Protoss [158] web server. Off-target ensembles were compiled with SIENA [156], using the parameters shown in Section 2.2.1. Ligands, waters, metals and crystallisation additives were removed. Each binding site was defined

by taking the residues within 6 Å of any of the fragments binding in that site. Protein structures were then aligned by the Cα atoms of the binding site residues. Off-target proteins were aligned to the target ensembles using the Cα atoms of their equivalent residues.

4.2.1.2 Fragment Hotspot Maps

- **Individual hotspot maps**

  As a first step in the workflow, fragment hotspot maps were calculated for all targets. Structures were downloaded from Fragalysis (https://fragalysis.diamond.ac.uk/viewer/react/landing/) and prepared as described in Section 4.2.1.1. Fragment hotspot maps were then calculated using the default parameters presented in Section 2.2.3, Ghecom version 20200721, and CSD version 2020.3.

- **Ensemble hotspot maps**

  Individual hotspot maps were combined into ensemble maps, using a polar frequency threshold of 20 % and no threshold for the apolar maps, as recommended in 2.3.5.

  When multiple sites of interest are available (this was the case of NSP13, presented in section 4.4.2.3), separate ensemble maps were generated for each of them. The fragment site mappings were taken from Fragalysis. Each binding site was defined by taking the residues within 6 Å of any of the fragments binding in that site. Proteins were then aligned by the Cα atoms of the binding site residues, and ensemble maps were calculated as described above.

- **Selectivity hotspot maps**

Selectivity hotspot maps for were calculated using a cluster distance cut-off of 3 Å and a minimal cluster score of 10 hotspot units, which are the recommended default settings for the method (see Section 2.3.5).

4.2.1.3 Scoring molecules using the fragment hotspot maps

Fragment hotspot maps carry two main kinds of information about the interactions available in the binding site (as discussed in Section 2.3.1.1). The first type is numerical information, reflecting the propensity of a fragment probe binding at that location. The second type of information is spatial, and relates to the positions of the interactions, particularly in terms of hydrogen bonding. A fragment has good fit with the hotspot if it both scores highly in the maps, and satisfies the highest ranked interactions.

The functionality to assign a hotspot score to each atom of a molecule with 3D coordinates is implemented as part of the Hotspots API [113] and has been previously described in Chapter 2.2.3. In the current workflow, a tolerance of 2 grid points, or 1 Å was used. This means that the highest score from the corresponding probe map within 1 Å in each direction was assigned to the atom. However, these scores on their own are not an intuitive measure of the molecule's fit to the hotspot, or of which interactions are satisfied by the molecule.

1. To enable the ranking of compounds, these per-atom scores needed to be combined into a single per-molecule score.

2. Initial work showed that combining the polar and apolar contributions into a single score could lead to the prioritisation of molecules that make no polar interactions with the protein (this is further elaborated in Section 4.3.2.3 and in Figure 4.3). To prevent this, scores summarising the contributions of only the polar atoms were introduced.

3. The per-atom scores' positional information is lost when they are combined into a single molecule score. Therefore, an additional, "by-feature" scoring scheme was introduced, tracking the specific binding site interactions satisfied by the molecule.

Initially, two simple ways of combining the scores were investigated: taking the sum of the per-atom scores (additive hotspot score), and taking their mean value (mean hotspot score).

- **additive_hotspot_score** The sum of the hotspot scores of all heavy atoms in the molecule.

- **mean_hotspot_score** The mean of the hotspot scores of the heavy atoms in the molecule.

The additive score assumes that each atom contributes independently to binding, and so can give high scores to molecules that place large regions of the molecule outside the hotspot. This may be desirable in certain cases, but does not measure the overall hotspot fit. The mean hotspot score has the disadvantage of deprioritising larger compounds, as they extend towards regions with lower hotspot scores (the initial publication on fragment hotspot maps notes that the starting fragments co-localise with the highest scoring areas of the maps, while elaborations are more moderately scoring [48]). However, between similarly-sized molecules, the mean hotspot score would favour the molecule that better overlaps with the hotspot.

In all cases, the combined hotspot scores do not differentiate between the contributions of polar and apolar atoms. This can lead to high scores being assigned to molecules that make no polar interactions with the protein. This is undesirable, as such highly lipophilic compounds can introduce issues with solubility, specificity and toxicity [14]. To avoid such cases, a separate score was computed to take into

account only the contributions of the polar score:

- **mean_polar_score** The sum of the hotspot scores of donor and acceptor heavy atom scores, divided by the total number of heavy atoms.

The application of these scores is presented and discussed in Section 4.3.2.3, and they are used in all of the case studies presented in this chapter.

The combined hotspot scores also do not contain information on which (and how many) polar interactions within the hotspot are satisfied. To approach this problem computationally, polar features in the ensemble maps first had to be identified, and then a procedure for scoring compounds against individual polar features had to be implemented. To identify clusters in the polar maps, the density-based clustering algorithm HDBSCAN is used, employing the same parameters used to detect features in the selectivity maps (minimal cluster size of 7 points). Once clusters have been identified, the centre of mass of the cluster is calculated, as well as its volume in $Å^3$. The radius of a sphere with this volume is calculated. Compounds that place an appropriate atom within this radius ($\pm$ 1 Å) of the cluster centroid are considered to 'hit' that polar feature. A script implementing this procedure can be found at https://github.com/mihaelasmilova/fragment_workflow/blob/main/by_feature_scoring.py.

4.2.1.4 Dynamic Undocking

Dynamic undocking has previously been used to provide information on the structural stability of individual interactions [47, 134, 45], as well as combined into a single score to give the structural stability of a pose [137].

Ways of combining multiple W_{QB} values into a single, per-molecule score had been previously explored by Majewski et al [137]. For fragment-sized ligands, using the average or sum of the W_{QB} scores of the detected fragments resulted in

improved binding pose prediction relative to the docking program used [137]. For larger ligands, combined scores based on clusters of interactions performed better than using whole-molecule scores and as well as the docking program [137]. The average W_{QB} score for a cluster of interactions was taken, and the sum of all clusters in the molecule was taken as the final score. While these metrics are highly useful in binding mode prediction, where the goal is to select an orientation that best exploits the binding site, they may be less well suited to comparing molecules anchored through different numbers of interaction clusters. In the workflow presented by Rachman [45], a single key interaction is defined and compounds are screened based on the structural stability of their complexes along this bond. This is similar to the way in which DUck was initially applied to virtual screening, presented in [47].

The workflow described in this chapter used W_{QB} values in the following ways:

1. To characterise the stability of the hydrogen-bond interactions made by starting fragments prioritised by the hotspot maps.

2. To compare the stability of hydrogen bonds made by docked follow-up compounds exploiting the same interaction with the protein.

Both of these involve a per-interaction, rather than a per-molecule view of the pose. The overall goal was to obtain more detailed information on the interactions highlighted by the hotspot maps, while taking into account the contribution of the ligand. Ideally, this would allow not only for the selection of compounds and chemotypes that place atoms in the predicted hotspot features, but also to prioritise those that make the most stable interactions with the binding site hotspot.

OpenDUck was used to perform all dynamic undocking steps. The same prepared structures used for calculating the hotspot maps and docking were also used as inputs for the OpenDUck calculations. Hydrogen bonds between the ligand and

the receptor were detected using the ODDT Python module [181]. Appendix table A.7 shows the chunking cutoffs and retained waters for the interactions simulated in the case studies. For the starting fragment interactions, 20 SMD runs/temperature were performed, which is considered to give extensive sampling [47], and the $W_{QB\_mean\_min}$ value was taken.

For the docked follow-ups, 3 SMD/ runs/ temperature were used and the W_{QB_min} was taken as the value for the interaction. In cases where W_{QB_min} was greater than 6 kcal/mol (the structural stability threshold introduced in [47]) after 3 SMD/ runs/ temperature, the interaction was simulated for the full 20 SMD/ runs/ temperature and the $W_{QB\_mean\_min}$ calculated.

4.2.1.5 Compound enumeration using the Fragalysis fragment network

After the starting fragments had been selected, the next step of the workflow involved enumerating compounds adjacent in chemical space to the starting fragment hits. A number of methods have been previously developed to for compound enumeration. These range from substructure searching of databases of commercial compounds, to *de novo* generated suggestions of molecules anywhere within drug-like space [52]. Selecting compounds from commercial databases has the advantage of providing compounds that can be easily synthesised and delivered. The size of these libraries has rapidly expanded over the past few years, increasing the number of compounds purchasable at low cost into the hundreds of millions [17]. Drug-like chemical space is considered to be even larger, with estimates ranging up to 10^{60} possible molecules. Hence, using compounds exclusively from commercial libraries may lead to missed opportunities. However, selecting molecules that are difficult to synthesise and obtain can slow down the timeline of a project, and increase costs. For the case studies presented in this chapter, a medicinal chemistry recommendation engine (the fragment network) was used, which out-

puts compounds with conservative changes from the starting fragment structures, which are also commercially available. This tool is integrated as part of Fragalysis, the computational FBDD platform for the XChem facility, and so was a logical and accessible choice for the enumeration step of the workflow.

The fragment network currently part of the Fragalysis platform is an open-source implementation of a medicinal chemistry recommendation engine first described by Hall *et al.* in 2017 [42]. It consists of a graph database, in which each compound is represented as a set of rings, linkers, and substituents [42]. The nodes and edges of the graph are generated by iteratively removing groups from the parent molecule. Each node is annotated with metadata describing its chemical structure (as a non-isomeric SMILES string and as a simplified graph representation), heavy atom count, and number of ring systems. A SMARTS pattern is used to find all single acyclic bonds to a ring atom. These are broken, creating a set of molecular components, which are then recombined into child nodes. Each of these child nodes excludes exactly one of the parent node's components. Parent and child nodes are linked by graph edges, which contain information about the chemical structures of the excluded parent component and of the rebuilt child molecule, as well as the type of combination that gave rise to the child molecule. When querying the network, relationships between nodes can be automatically detected as deletions, additions, and replacements through a set of rules based on the number of edges between the nodes, as well as the edge annotations. By default, search queries return all available compounds from graph nodes that are 0-2 edges away from the start node. This representation of chemical space allows child compounds to be grouped in a way similar to that in which medicinal chemists consider hit elaboration, as the results are grouped by substitution position [42].

In this chapter, the Fragalysis fragment network was used for hit elaboration.

The Fragalysis RESTful API was used to query the underlying fragment network through a Python script that can be found at https://github.com/mihaelasmilova/fragment_workflow/blob/main/graph_utils.py. This is based on the code supplied in https://github.com/michellab/XChem-examples/tree/master/fragalysis_preproc, written by Rachael Skyner. Follow-up suggestions were generated through either a single "hop" (a single query returning nodes that are 0, 1, or 2 edges away from the starting compound), or a double hop (in which the outputs of the previous query are used as inputs for new queries). This choice depended on the needs of the project, and are further addressed in the case studies presented in this chapter.

Apart from ensuring that the suggested compounds are synthetically accessible and purchasable, the fragment network does not filter the output for favourable physicochemical properties or drug-likeness. In this chapter, the "lead-likeness" guidelines introduced by Teague *et al.* [183] of logP < 3.0 and molecular weight < 350 Da were used. Molecular properties were calculated using the RDKit Python library. Cases where these values were modified according to the specifics of the project, or further filters have been introduced, are further discussed in the case studies presented.

4.2.1.6 Docking

As both the hotspot maps and DUck rely on 3-dimensional information, the 1-dimensional molecular encodings (SMILES strings) generated by Fragalysis would need to be translated into a predicted binding mode. A key assumption of fragment-based drug design is that the elaborated compound retains the starting fragment's binding mode. This is why constrained docking was used for this step, limiting the generated poses to overlap with the starting fragment in the regions of their shared substructures.

Initial 3D structures for the follow-up SMILES strings from the enumeration step were generated using the CSD's Conformer Generator [184] through the CSD Python API [110]. Minimisation was carried out using the CSD API's MoleculeMinimiser class. This uses an algorithm based on the Tripos force field functional form, complemented by information on bond lengths and valence angles taken from the CSD, as described in [184].

GOLD (Genetic Optimisation for Ligand Docking) is an automated ligand docking program [185]. In addition to being one of the most successful and widely used tools in the field, GOLD can also be accessed directly through the CSD Python API's docking module, which facilitated its integration into the fragment elaboration workflow.

GOLD uses a genetic algorithm to drive the exploration of ligand binding modes within the receptor binding pocket. The genetic algorithm works by mimicking the process of evolution among a set of data structures referred to as chromosomes. For the docking case, each chromosome encodes a possible solution to the docking problem and is assigned a "fitness" score, based on a scoring function. For the work presented in this chapter, the ChemPLP scoring function was used. This has been shown to be the best of the scoring functions available in GOLD for virtual screening [186], and is the default option in GOLD [187].

Chromosomes are allowed to reproduce using a set of reproduction operators (crossover, mutation, and migration), each of which is associated with a weight. An initial population of protein and ligand conformations is generated and the fitness calculated in each case. Parent chromosomes are chosen based on a weighted "roulette wheel" selection, using the fitness scores as weights [185]. An operator is also chosen using weighted selection, using the pre-defined operator weights. Child chromosomes are generated by applying the operator to the parent chro-

mosomes, and their fitness scores calculated. Child chromosomes then replace the least fit members of the population (unless the child chromosome is already present in the population). This process continues until a number of operators have been applied [185]. This number is governed by the autoscale parameter, which is user-provided and reflects the search efficiency of the algorithm. An autoscale value of 100 reflects a search efficiency of 100 %, with an associated computational cost [187].

In this workflow, "fast" sampling was used with an autoscale value of 10, as the majority of the atoms will already be constrained to the preferred binding mode. 50 poses were generated per follow-up as a compromise between speed and sampling efficiency. This protocol took about 30 seconds to 1 minute per compound, and can be parallelised across multiple CPU cores.

When docking follow-up compounds generated from a crystallographic fragment hit, the binding mode of the starting fragment is already known. A key assumption of FBDD is that the fragment retains its binding mode upon elaboration. GOLD allows spatial constraints to be placed on the protein and the ligand, which can be used to focus the algorithm's conformational search to solutions where there is good overlap between the docked pose and the starting fragment's binding mode [187]. Two types of constraints were used in the work presented in this chapter. Similarity constraints in GOLD bias the conformations of docked ligands towards a given solution by adding an energy term to the fitness score, which is based on the overlap between the docked ligand and the provided template (modeled as a Gaussian overlap term). The similarity constraint can be applied in three ways, which differ in the way that the overlap between the ligand and the template is calculated. In the docking API, they are controlled through the value of the "overlap" parameter. Setting its value to "donor" uses the overlap between all donor atoms in the template and docked ligand; conversely, "acceptor" uses the

acceptor atoms, and setting it to "all" uses all of the atoms, essentially providing a shape constraint for the ligand.

GOLD has a separate constraint type for scaffold matches ("constraint scaffold"), which places a substructure at an exact position in the binding site, and this geometry is not changed during docking. Unlike the similarity constraint, which adds a term to the scoring function, the scaffold match constraint is applied during the ligand placement step. This constrains scaffold atoms in the matching portion of the ligand to retain the positions of the parent fragment's scaffold location. Constraints in GOLD have a default value of 10. However, previous work by Radoux [97] showed that increasing constraint weights to 100 improves the performance of hotspot-guided docking. Consequently, this value was used in the constrained docking step of the workflow.

4.2.1.7 Measuring fragment binding mode conservation

Using constraints in the docking step will bias solutions towards those in which the constraints are satisfied. However, the top ranked solutions will not necessarily be those with the best overlap to the parent fragment. To get a measure of the similarity in the binding mode between the bound fragment and the elaborated ligand, a dedicated metric was used. An open-source version of the COS (combined overlap score) metric used by Malhotra and Karanicolas in their 2017 publication on detecting changes in fragment binding mode upon elaboration [59] was employed. COS is based on the ROCS algorithm, which uses a summation of Gaussians to represent the shape density function of the molecule, and a "colour" term, which describes the overlap of features of the same type (hydrogen bond donors, acceptors, cations, anions, and aromatic rings). SuCOS, the open source version of COS, uses the RDKit functions *ShapeProtrudeDist* and *ScoreFeats* to compute volume and chemical feature overlap, respectively [61]. It was shown

to outperform RMSD, the traditionally used metric for fragment binding mode conservation, and perform comparably to the original COS metric [61].

4.2.2 Experimental validation

The compounds suggested by the computational workflow were ordered from commercial sources (Enamine and PostEra) and their binding was determined experimentally. Compounds for all of the case studies presented in this chapter were ordered in the spring of 2021, and so experimental results from one case study could not be used to inform decisions on the others.

4.2.2.1 ACVR1 crystal screening

The compounds designed for ACVR1 were purchased, and a crystallographic screen was performed at the XChem facility [32] by Eleanor Williams and Lizbé Koekemoer. ACVR1 was co-crystallised with the inhibitor M4K2117, as the apo form of the protein did not form crystals [188]. The co-crystal form was also used in the original fragment screen [188]. The protein-M4K2117 complex was crystallised in 0.1 M citrate pH 6, 1.4M ammonium sulphate, and 0.2 M sodium/potassium tartrate. Crystals were soaked for 60 seconds at 4 C, with 1 μL of 100 mM compound, 1 μL of mother liquor, 0.5 μL of 20 % ethylene glycol. Crystal data was collected to a resolution of 1.5 Å and processed using PanDDA [38] and XChem Explorer [32].

4.2.2.2 NSP13: ADP-Glo functional assays

Experimental testing of the compounds for activity against NSP13 was performed by Joseph Newman using a Promega ADP-Glo ATPase assay. Inhibitory activity was measured at 500 μM compound concentration in duplicate. The final ATP concentration was 100 μM, and the final DNA concentration - 100 nM. The reac-

tion buffer was made up of 20 mM HEPES pH 7.5, 50 mM NaCl, 5mM MgCl, 1mM DTT, 0.1 mg/ml BSA. The assay controls used were cells with no enzyme, DMSO only, and enzyme but no DNA added.

4.2.2.3 PARP14-MD3: HTRF assays

HTRF (Homogenous Time-Resolved Fluorescence) assays were performed by James Bennett at the Centre for Medicines discovery. Binding of a biotinylated and ADP-ribosylated peptide to PARP14 (ARTK(Bio)QTARK(Aoa-RADP)S) was measured using 250 nM peptide, 50 nM protein, a 1:20000 dilution of 6HisAbEU anti-histag antibody, and 60 nM SA-XL665-labelled streptavidin. The reaction buffer was made up using 25 mM HEPES pH 7.0, 20 mM NaCl, 0.05 % BSA, and 0.05 % TWEEN 20. Compounds were tested at two concentrations: 500 μM and 200 μM. The signal control was ADPR (adenosine diphosphate ribose, the native substrate) at 50 μM, and DMSO only was used as a background control.

4.3 Case Study: ACVR1

The proposed workflow was applied to the human kinase target ACVR1. This protein had been the focus of a Target Enabling Package (TEP) [2] at the Centre for Medicines Discovery, meaning that starting chemical matter in the form of crystallographic fragment screening hits was available.

4.3.1 Target overview

Activin A receptor, Type I (ACVR1) is a target implicated in a rare childhood brain tumour called DIPG (Diffuse Intrinsic Pontine Glioma), as well as in the rare and debilitating disease FOP (Fibrodysplasia ossificans progressiva), in which muscles, tendons and other tissues are replaced by abnormal bone growth [189]. Both

of these diseases are linked to gain-of-function mutations in the *ACVR1* gene, which encode the BMP receptor ALK2 [188, 189]. BMPs (bone morphogenic proteins) and activins are protein growth hormones, whose action is tightly regulated. Mutations in ALK2 lead to the constitutive activation of the receptor, and ultimately the uncontrolled phosphorylation and activation of downstream SMAD1/5/8 transcription factors. These pathogenic mutations have been shown to break interactions that would hold the ALK2 kinase domain in an inactive conformation, inhibiting this downstream activation of transcription factors [189]. Therefore, developing kinase inhibitors that specifically target the ALK2 kinase domain could provide much needed treatments for these two rare and highly debilitating diseases. While a number of drugs have entered clinical trials for FOP, the most advanced of these were stopped due to safety concerns [188]. Crystal structures of ALK2 in its inactive conformation show that it is incompatible with ATP and substrate binding. An XChem crystallographic screen of the inactive crystal form performed by Williams *et al.,* [188] produced 18 fragment hits from the DSPI poised library [31] and a further 50 hits from the MiniFrags [29] library. A single hit from the poised library (fragment x1344_0B) was found to bind in an allosteric site (shown in Figure 4.2) and presented an attractive hit for elaboration.

4.3.2 Results and Discussion

4.3.2.1 Fragment Hotspot Maps

While the ACVR1 fragment screen had resulted in over 60 hits from the DSPI-poised and MiniFrag libraries, the focus of this case study was a single fragment, x1344_0B. This was bound in an allosteric site (designated as Allosteric Site 1 in [188]), located behind the ATP binding site (shown in Figure 4.2, panel A) of the kinase. This starting point was chosen, as allosteric inhibitors are more

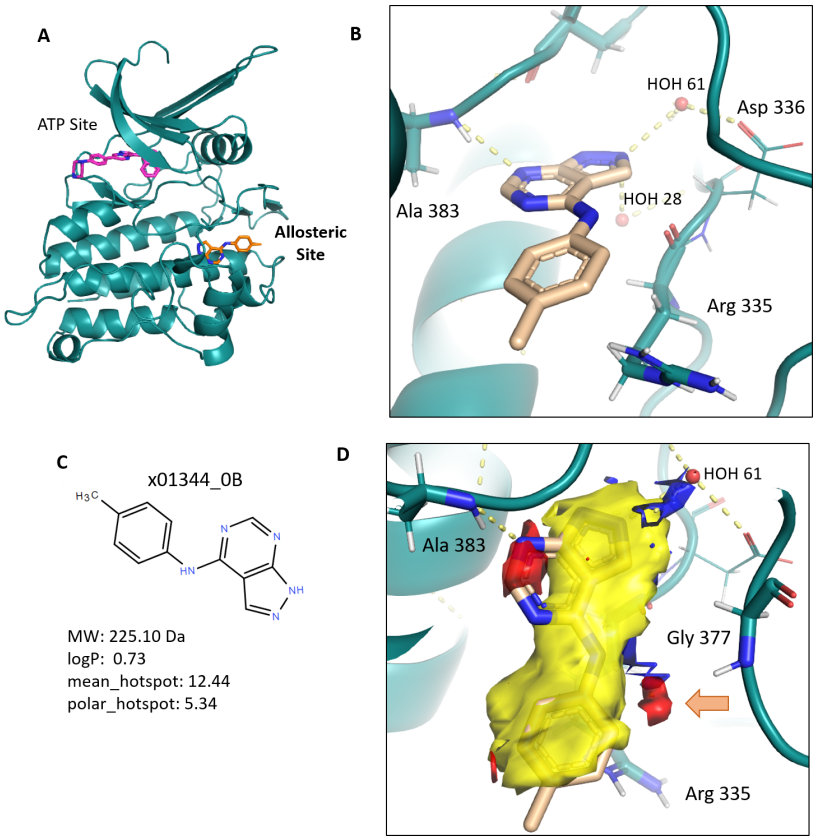


Figure 4.2: Binding mode of fragment x1344_0B. **A** Location of the ATP binding site and Allosteric Site 1 on the surface of ACVR1. **B** Binding mode of fragment x1344_0B within Allosteric Site 1. Interacting residues are shown as sticks. The two structurally important waters are shown as spheres and hydrogen bonds are represented by dotted yellow lines. **C** 2D structure of the starting fragment. **D** Fragment hotspot maps calculated for the holo structure of the complex. The orange arrow shows the unexploited acceptor feature. The maps are colour-coded by probe type: red - acceptor, blue - donor, yellow - apolar.

likely to have good selectivity in a family with a conserved functional site, such as kinases. Because a single structure was available of a fragment in this binding site, ensemble maps were not generated in the first step of the workflow. Figure 4.2, panel B shows the binding mode for x1344_0B. The fragment's purine moiety makes a direct hydrogen bond with the backbone nitrogen of Ala 383 and two water-mediated contacts via HOH 62 and HOH 28. The hotspot maps calculated on the holo complex structure (Figure 4.2, Panel D) show good agreement with the observed binding mode, with a region of acceptor density located across from the Ala 383 backbone N and the positions of the waters coinciding with regions of favourable donor density. Another region of donor density lies across from the Arg 335 backbone, coinciding with the fragment's secondary amine. Interestingly, a strong acceptor region was detected across from the backbone nitrogen of Gly 337 (highlighted with an orange arrow in Figure 4.2, panel D). This is not an interaction exploited by the initial fragment, and so presented an opportunity for the design of follow-up compounds that could make this contact. A further, albeit weaker, acceptor hotspot feature originates from the NH1 nitrogen on the side chain of Arg 335. Both of these unexploited acceptor features were located so that they could be reached by extending from the fragment's phenyl ring.

4.3.2.2 Enumerating potential follow-up compounds using the Fragalysis graph network

The enumeration procedure described in section 4.2.1.5 resulted in 213 unique molecules (from a total of 233 molecules returned). As these would be part of the first round of follow-up compounds, with only conservative changes from the starting fragment, the lead-like cutoffs introduced by Teague *et al.*[183] were applied. As discussed in Section 4.2.1.5, compounds with clogP > 3.0 (as calculated by RDKit's Crippen module [190]), molecular weight > 350 Da, and > 3 rotatable

bonds were discarded (leads tend to have lower complexity and flexibility compared to drug-like compounds [183, 45]). This left 122 unique molecules, which were docked according to the protocol described in section 4.2.1.6. The full list of these compounds can be found in the Supplementary Data for this thesis, in the file "acvr1_all_followups.csv".

4.3.2.3 Docking and re-scoring of the enumerated follow-up compounds

As discussed in section 4.2.1.6, GOLD can perform constrained docking using both substructure and template constraints. In this case, the template similarity to the starting fragment was chosen, in order to retain the fragment's overall good fit with the hotspot, while allowing modifications to the shared substructure. Figure 4.2 shows that fragment x1344_0B has a very good overall fit to the hotspot, placing all of its rings within the apolar hotspot density, and satisfying the acceptor interaction with the backbone of Ala 383. Therefore, similarity with the overall fragment shape is desirable, as restricting only a substructure of the molecule could result in poses that place parts of the molecule away from the targeted hotspot. The GOLD all-atoms similarity constraint is purely shape-based, however. In order to take into account how well the docked poses recapitulate the binding mode of the initial fragment, considering the pattern of donor and acceptor features, SuCOS [61] scores were calculated for all docked poses.

A key assumption of the fragment hotspot maps method, and of fragment-based drug design in general, is that the binding mode of the starting fragment remains unchanged. In fact, using reference fragment structural information to re-score docking poses has been shown to improve performance in the binding mode prediction of fragment-sized ligands [191]. This is why for each follow-up, the pose with the highest SuCOS score was selected.

Docked poses were also re-scored against the fragment hotspot maps, as described in section 4.2.1.3. The mean_hotspot_score is a general measure of how well the fragment overlaps with the hotspot, while the mean_polar_score measures the contribution of fragment atoms that satisfy the polar hotspot features. In both scoring cases, molecules that place atoms outside the hotspot will have poorer scores, even if they fulfill multiple hotspot interactions. In the first round of follow-up enumeration (as was the case here), this ensures that only compounds with a good overall fit to the hotspot volume and interactions are considered. For larger ligands, however, other ways of combining hotspot scores may be more appropriate. These will be further discussed in Section 4.6.

Figure 4.3 shows the distribution of mean and polar hotspot scores for the best SuCOS-ranked follow-up poses. Follow-up 141, shown in panel B in Figure 4.3 (highlighted in purple), is an extreme outlier with a very high mean score, which is dominated by apolar atom contributions. This pose does not make any hydrogen bonds with the receptor, showing the necessity of including information about the overlap with polar hotspot features. The fragment network has also substituted the initial fragment's purine headgroup, resulting in a change of the predicted binding mode. On the other extreme, Follow-up 21 (shown in a magenta box in Fig 4.3) has excellent overlap with the polar hotspot features, but places 2 atoms of the 5-membered ring outside the apolar hotspot, leaving a large amount of unexploited apolar hotspot volume. Panels C and E in Figure 4.3 show borderline cases that were used to define the mean and polar hotspot score thresholds for further selection. Follow-up 15 (Panel C) possessed good overlap with the apolar maps, but at least two of the potential polar interactions (the acceptor hotspot proximal to Gly 377 and the donor hotspot proximal to Arg 335) are not satisfied. Follow-up 36 (panel E) places a methyl group in the acceptor hotspot corresponding to Gly 377, as well as leaving a large amount of apolar hotspot volume unsatisfied

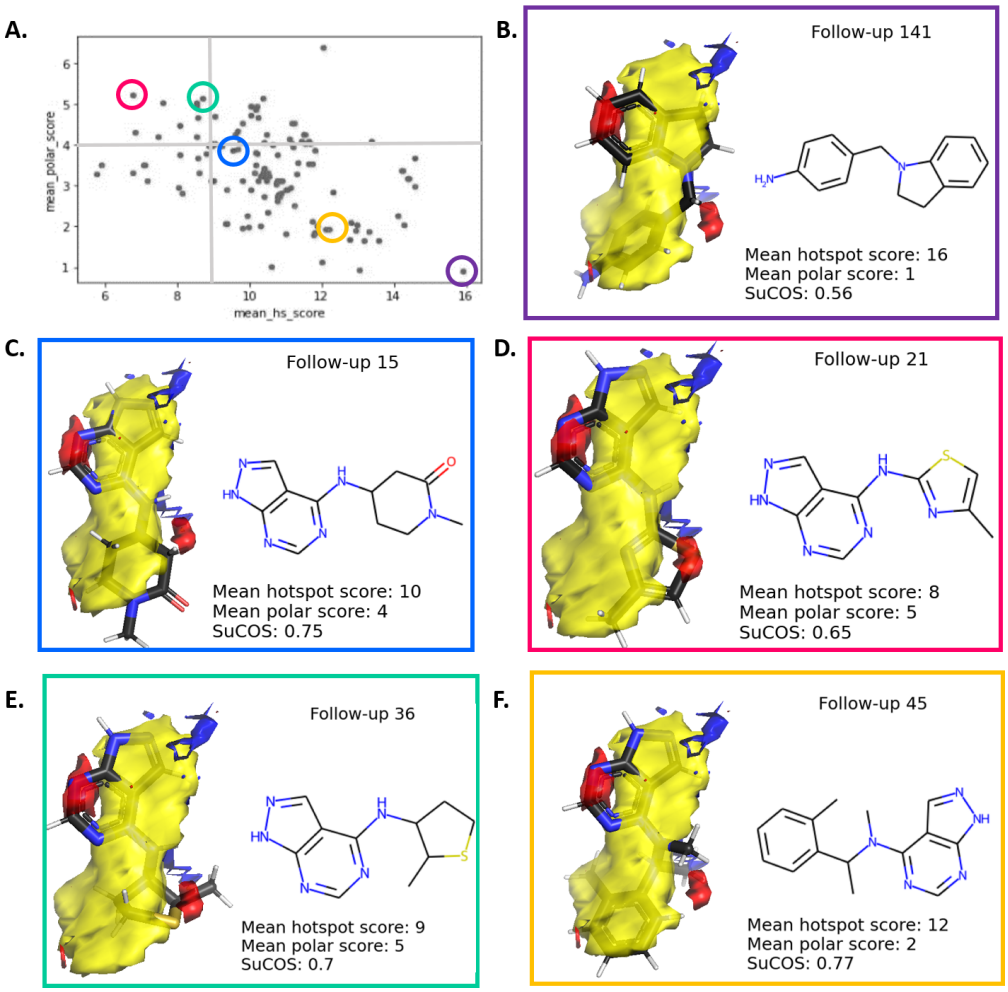


Figure 4.3: Selecting Hotspot scoring thresholds in ACVR1. **A.** Plot of the mean_hotspot_score (x axis) versus the polar_hotspot_score (y axis). The highlighted points represent the structures in panels B-F. The points are colour coded according to their corresponding panels. The 3D compound structures show the docked pose with the highest SuCOS score for the respective followup. Hotspot maps are shown as PyMol isosurfaces. The colour coding is yellow for the apolar maps, red for H-bond acceptor, and blue for the H-bond donor maps.

when compared with the starting fragment. Finally, Follow-up 45 (Panel F) is another example where the favourable mean hotspot score is dominated by the apolar contribution. This ligand also places a tertiary amine in the hydrogen bond donor hotspot feature arising from the Arg 335 backbone oxygen.

Based on these observations of borderline cases, a mean_hotspot_score cutoff of 9.0 and a mean_polar_score cutoff of 4.0 were used (shown as grey lines in Figure 4.3) in order to select only follow-up poses that showed good overlaps with both the polar and apolar hotspot maps. This resulted in a selection of 25 compounds (these are shown in the Appendix, Figure A.7), which were further triaged by selecting those that satisfied the two acceptor interactions (to the Gly 377 backbone nitrogen and to the Arg 335 side chain NH1). This selection was made through the by-feature re-scoring procedure presented in Section 4.2.1.3. Figure 4.4 shows the final set of follow-up compounds that were selected using the script, along with Follow-up 30, which does not make additional polar interactions compared to the starting fragment, but has excellent apolar overlap. Follow-ups 13, 160, and 161 were chosen for placing an acceptor atom (carbonyl oxygen in the case of 13, ether oxygen in 160 and sulfide sulfur in 161) in the acceptor feature associated with Gly 377. Followup-ups 9, 27, and 38 were chosen as they satisfy the hotspot feature arising from the Arg 335 side chain NH1.

4.3.2.4 Dynamic Undocking

Dynamic undocking was set up using the protocol described in 4.2.1.4. Initially, 20 SMD runs/temperature were performed on the interactions made by the parent fragment x01344_0B. The two structurally important waters (HOH28 and HOH61) were retained in the chunk, and the interactions with the backbone NH of Ala 383, the backbone NH1 of Tyr 381, and the backbone O of Arg 335 were investigated. These are shown in Figure 4.5.

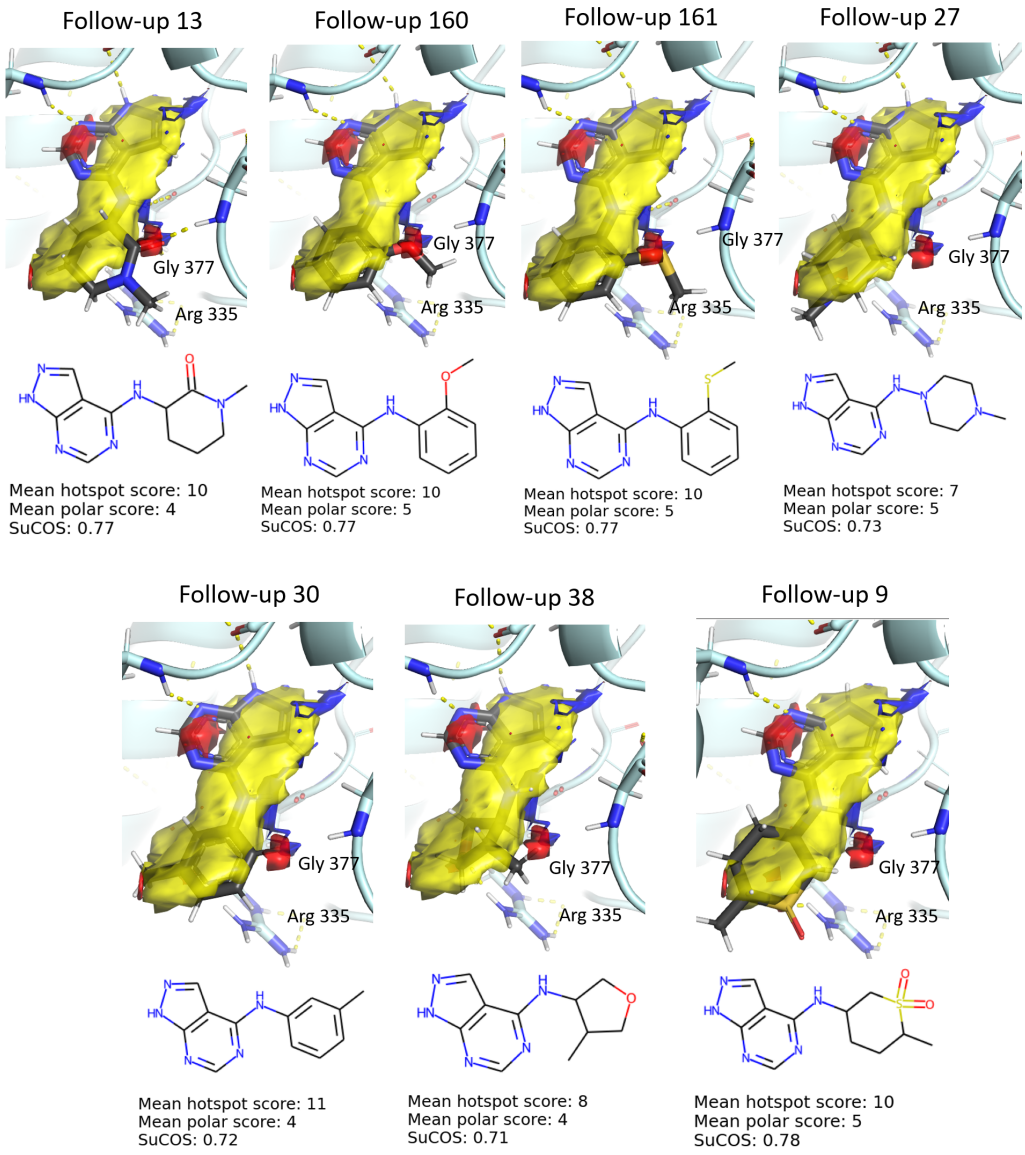


Figure 4.4: Selected compounds for ACVR1. Hotspot maps are shown as PyMol isosurfaces. The colour coding is yellow for the apolar maps, red for H-bond acceptor, and blue for the H-bond donor maps.

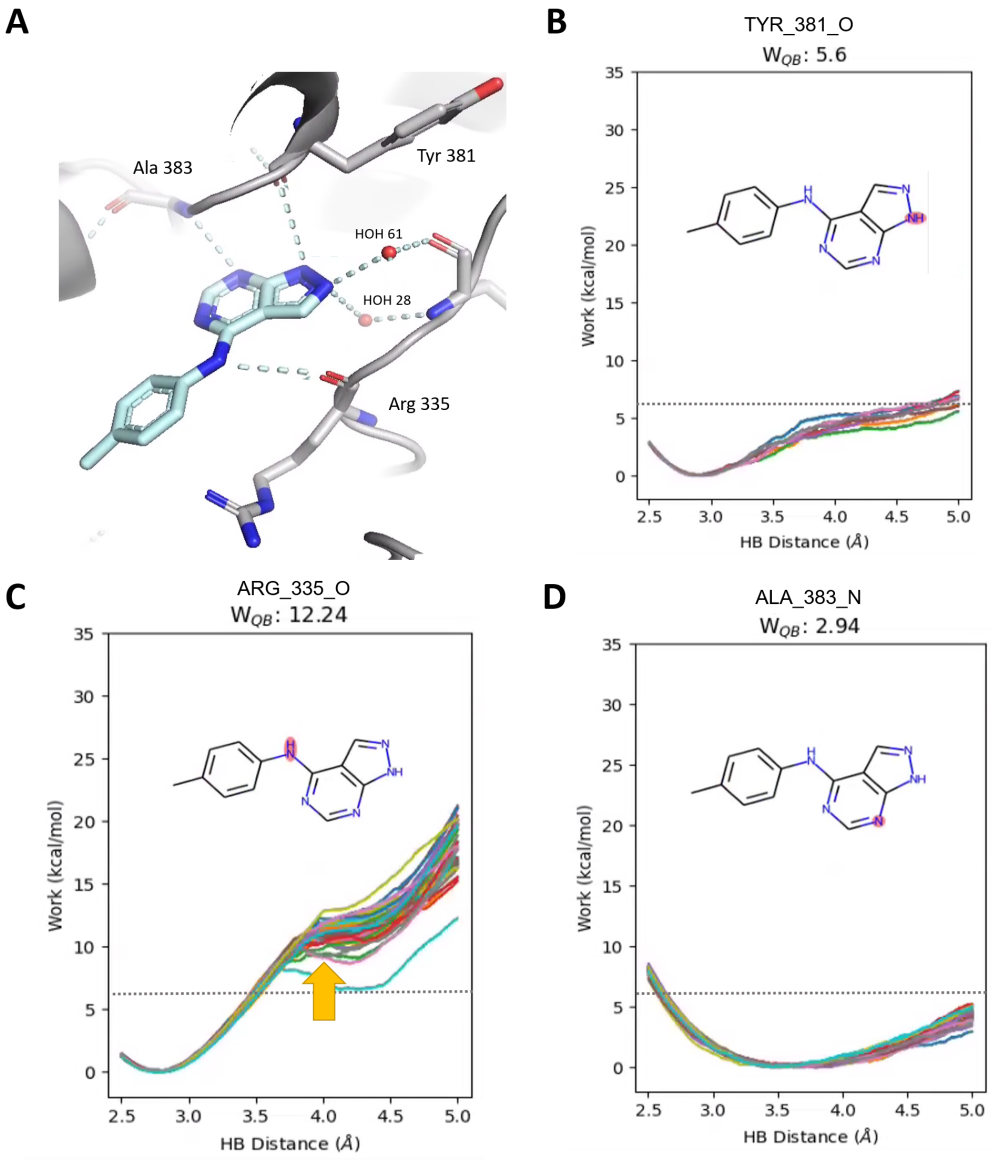


Figure 4.5: DUck profiles of the interactions made by ACVR1-x1344_0B. A. Interactions made by fragment x1344_0B with the binding site residues. B. DUck work traces for 40 SMD runs of the TYR_381_O interaction. C. DUck work traces for 40 SMD runs of the ARG_335_O interaction. The "hump" feature discussed in the text is highlighted with an arrow. D. DUck work traces for 40 SMD runs of the ALA_383_N interaction. The interacting ligand atoms are highlighted by red circles in panels B-D. The grey dotted line corresponds to a work value of 6 kcal/mol, the threshold for structural stability introduced in [47].

Of the three interactions, only that with the backbone of Arg 335 (Figure 4.5, panel C) passed the structural stability threshold of 6 kcal/mol introduced in [47]. The "hump" (highlighted with an arrow in panel C) is the result of the NH group on the ligand rotating during the course of the steered MD run. As hydrogen bonds have strong angular dependencies, this results in the abolition of the hydrogen bond contact with the protein. The W_{QB} value in such cases is taken as the value of the hump.

The interaction with the backbone O of Tyr 381 (Figure 4.5, panel B) is very close to the structural stability threshold, while the interaction with ALA 383 (Figure 4.5, panel D) is structurally labile.

The shortlisted follow-up molecules identified in Section 4.3.2.3 have variable phenyl substituents adjacent to the NH group involved in the interaction with the Arg 335 backbone O atom. OpenDUck was used to give an estimate of whether these changes would disrupt the stability of that interaction. W_{QB} values for the hotspot-suggested hydrogen bonds to Gly 377 and the side chain of Arg 335 were also calculated in order to assess their potential contribution to the stability of the binding mode. Figure 4.6 shows that for all the follow-up compounds, the interaction with the backbone O of Arg 335 remains structurally stable. In the interactions with Gly 377 and the side chain of Arg 335, the follow-ups demonstrated comparable W_{QB} values, none of which were above the structural stability threshold of 6 kcal/mol [47].

This is similar to the retrospective case study presented in Section 3.5.2, where the compounds under investigation had a common core (or 'anchor') making conserved, structurally stable interactions with the kinase hinge hotspot. Individual H-bonds outside such anchoring clusters of stable interactions are less likely to be structurally stable themselves, as shown by Majewski *et al.* [134]. In the

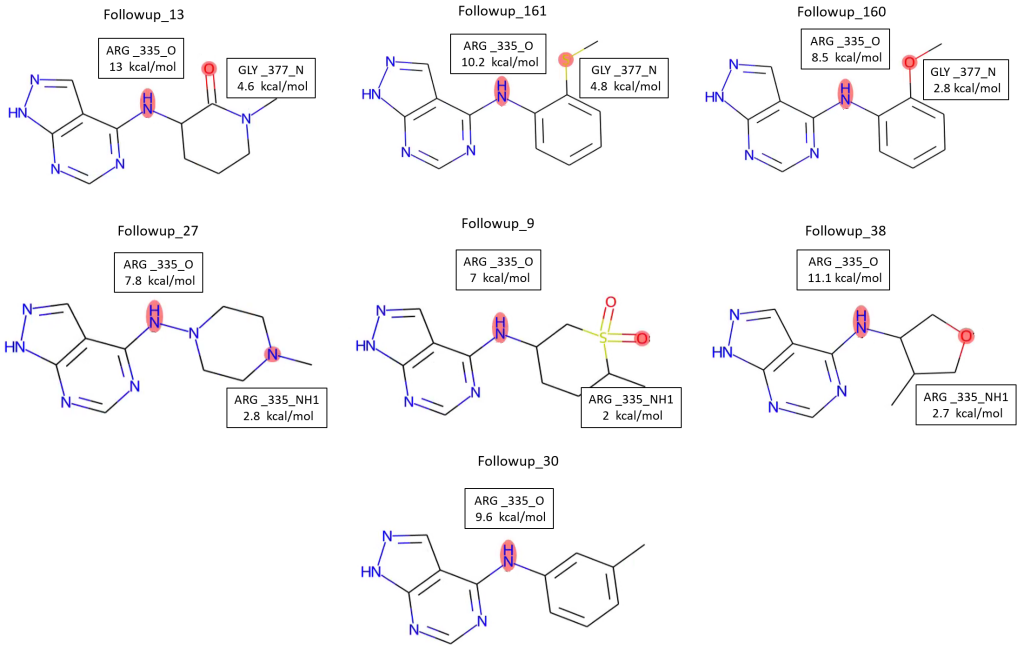


Figure 4.6: W_{QB} values of the selected ACVR1 follow-ups. The minimum W_{QB} observed in 6 runs was reported. The interacting ligand atoms are highlighted by red circles.

CDK2 case, the most potent compound had a sulfonamide group making three contacts with the protein, one of which was structurally stable (Figure 3.15) and which together formed a secondary cluster of stabilising interactions. None of the ACVR1 follow-up molecules formed such secondary clusters. However, the predicted poses all retained the stable interaction with the backbone of Arg 335 and presented an improved fit to the hotspot compared to the initial fragment.

4.3.2.5 Crystal soaking at XChem

The compounds shown in Figure 4.4 were purchased and soaked in ACVR1 crystals at the XChem facility [32] by Eleanor Williams and Lizbé Koekemoer. A total of 22 compounds were screened, including the 7 suggested by the workflow. Crystal data was collected to a resolution of 1.5 Å and processed using the standard XChem pipeline [32] by Lizbé Koekemoer. Further refinement and manual processing was done by Eleanor Williams. PanDDA [38] was able to detect three binding events, which were found to be ambiguous by the crystallographers, indicating that longer soaking times were required. None of the three hits were compounds among those suggested by the workflow, although they shared the adenine moiety with the starting hit and with the follow-ups suggested by the workflow described in this chapter (panels B in Figures 4.7, 4.8, 4.9). Two were in the correct site, but not overlapping the initial fragment hit (follow-up hits x711 and x712, shown in Figure 4.7 and 4.8, respectively). The third was bound in a different site on the protein surface (follow-up x695, shown in Figure 4.9). The crystallographers concluded that the soaking step had not been successful, and that longer soaking times, soaking a higher concentration of the compounds, or co-crystallisation, would be required. Longer soaking would be problematic, however, as even with the current conditions, precipitation and crystal damage was observed during this step. Therefore, co-crystallisation experiments would

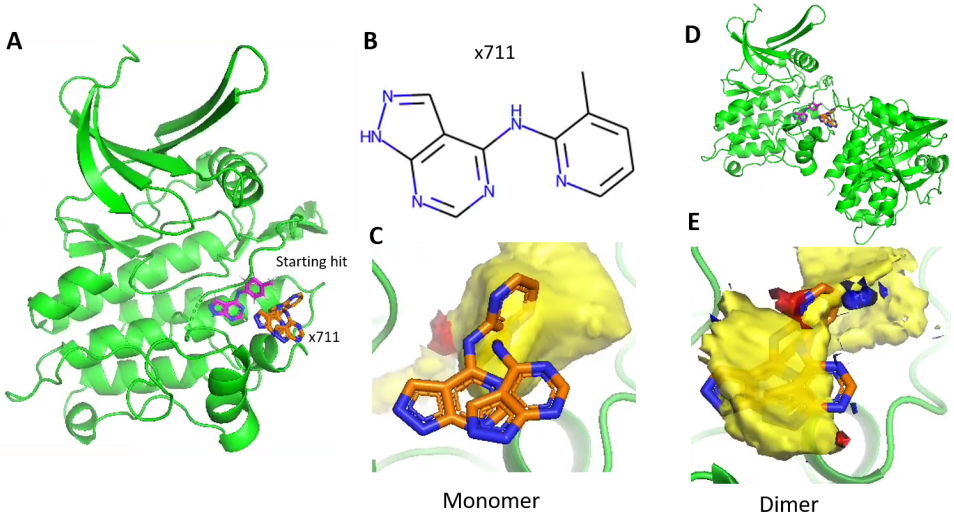


Figure 4.7: ACVR1 follow-up hit x711. A. The binding site of follow-up hit x711 (shown in orange) relative to the starting fragment hit (x1344, shown in magenta). B. 2D structure of x711. C. Hotspot maps for x711 (calculated based on the biological monomer). D. Hotspot maps for x711 (calculated based on the asymmetric unit). E. Binding positions of x711 (orange) in the asymmetric unit and relative to the starting fragment hit(magenta)

be required to test the hypotheses suggested by the workflow.

Hits x711 and x712 were modelled in the vicinity of the starting fragment hit, but did not bind deep enough into the pocket in order to overlap with the starting hit (Panel A in Figures 4.7 and 4.8). Fragment hotspot maps were calculated on the holo protein structures in order to look for interactions and environments that could retain the fragments from fully entering the site. At the time of writing, these structures were not yet processed to a PDB submission standard, but could be used to gain preliminary insights into factors that influence the alternative binding mode. Initially, hotspot maps were calculated based on the ACVR1 monomer, which is the biologically functional unit (panel C in Figures 4.7 and 4.9). These maps showed that the compounds bind mostly outside of the hotspot density. However, the asymmetric unit of ACVR1 is a dimer, shown in panel D

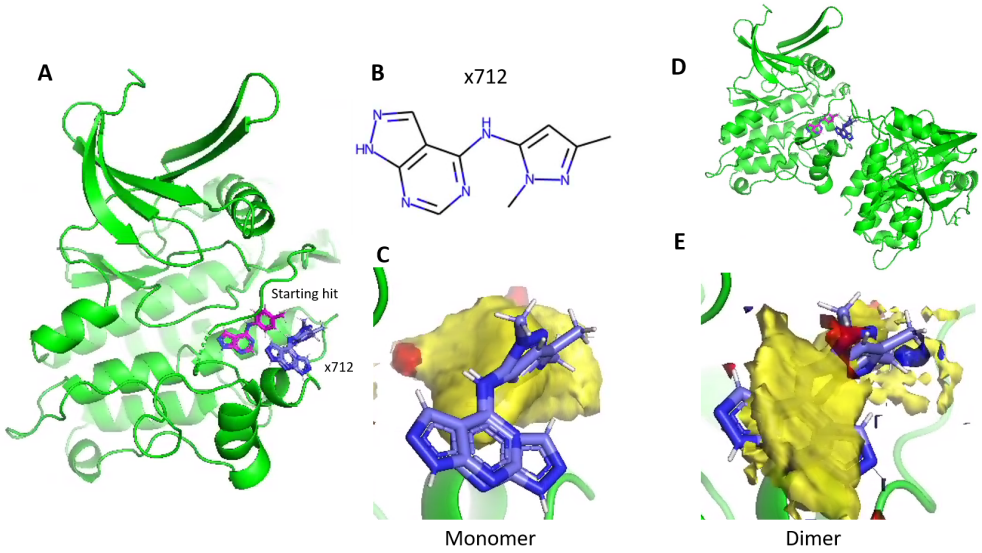


Figure 4.8: ACVR1 follow-up hit x712. A. The binding site of follow-up hit x712 (shown in orange) relative to the starting fragment hit (x1344, shown in magenta). B. 2D structure of x712. C. Hotspot maps for x712 (calculated based on the biological monomer). D. Hotspot maps for x712 (calculated based on the asymmetric unit). E. Binding positions of x712 (orange) in the asymmetric unit and relative to the starting fragment hit(magenta)

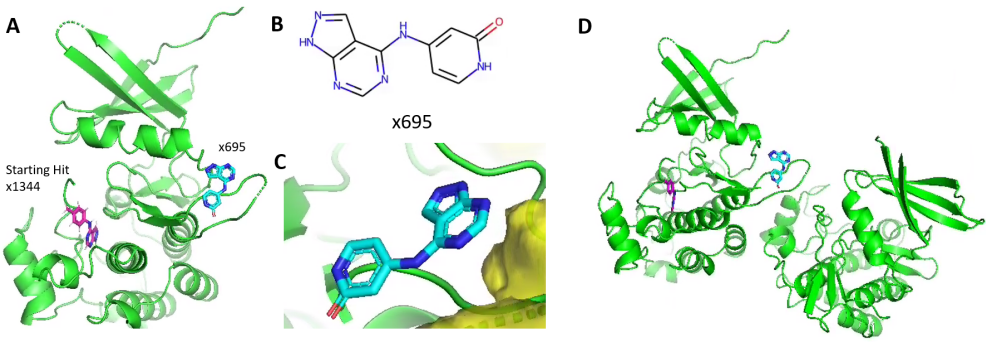


Figure 4.9: ACVR1 follow-up hit x695. A. The binding site of follow-up hit x695 (shown in cyan) relative to the starting fragment hit (x1344, shown in magenta). B. 2D structure of x695. C. Hotspot maps for x695 (calculated based on the asymmetric unit). D. Binding positions of x695 (cyan) in the asymmetric unit and relative to the starting fragment hit(magenta).

of the figures. When hotspot maps were calculated for the full asymmetric unit, mimicking a "crystal" view, rather than a biological one, a strong hotspot appeared that overlaps the modelled follow-ups. In both cases, the adenine moiety falls in a region of strong apolar density, but does not overlap polar hotspot interactions. This likely contributes to the ambiguity in its binding mode in both of the bound follow-up hits.

In the case of follow-up x695, the hotspot maps for both the monomer and the dimer did not show density in the region. Panel D in Figure 4.9 shows that the hit is bound near the surface of the interface between the two copies of the asymmetric unit in a solvent-exposed location. As the fragment hotspot detection is based on the buriedness of the protein surface, this is the most likely reason why no hotspot density was detected. As small molecules tend to bind in pockets, it is also possible that this site is too solvent exposed to drive their binding outside the context of the crystal lattice.

The lack of structural data for the compounds suggested for ACVR1 was disap-

pointing, but shows some of the limitations of using crystallography as a readout. As the crystals were too fragile to allow for sufficiently long soaking, signal from putative binding events was low, even when sophisticated analysis tools such as PanDDA [38] were employed. This does not necessarily mean that the follow-ups suggested by the workflow are non-binders, however. The crystallographers on the project reasoned that a longer contact time, or a higher concentration of compound, could allow cases such as x711 and x712 to overcome the initial interactions made and enter fully into the pocket, as was observed in the original screen. As longer soaking times are likely not possible, higher concentration soaks or co-crystallisation experiments are needed to test the hypotheses generated by the workflow. Orthogonal readouts, such as binding or functional assays, would also be needed to provide information on the binding affinity of the suggestions.

The three hits that were detected by PanDDA were not generated by the workflow, but were suggestions from collaborators. They were structurally similar to the starting fragment and to the follow-ups generated by the workflow, sharing the adenine moiety of the starting fragment hit. For the two hits that bound close to the allosteric site of the initial fragment, hotspot mapping was able to show that the observed binding mode is likely an artefact of the packing of protein chains in the asymmetric unit, and would not be observed in solution. This raises the possibility of using the method as a way to detect such artefacts, which are not uncommon in fragment screening by crystallography. Cases such as fragment x695, where a binding hotspot was not observed in either the biological or the crystallographic assembly, could be automatically annotated for further manual inspection by crystallographers based on such results. Overall, despite the inconclusive results with regards to testing the hypotheses generated by the workflow, the results of the soaking experiments demonstrated a new potential application of the fragment hotspot maps method, which will be explored in the future.

4.4 Case Study: NSP13

Non-structural protein 13 (NSP13) is a viral helicase from the SARS-CoV-2 pathogen. It has been identified as a promising target for antivirals due to its high sequence conservation and essential function in the viral life cycle [44]. Newman *et al.* performed a crystallographic screen against this target at the XChem facility, aiming to identify potential starting points for novel antiviral agents [44]. The screen resulted in 65 fragment hits across 24 sites on the protein's surface.

4.4.1 Target overview

As a superfamily 1B helicase, NSP13 uses energy from the hydrolysis of nucleotide triphosphates (NTPs) to catalyse the unwinding of double stranded nucleic (DNA or RNA) acids in the 5' to 3' direction [44]. The protein has a modular structure with 5 domains, shown in Figure 4.10. The DNA/RNA and nucleotide binding sites are both located in interface regions between adjacent domains (Figure 4.10, panel A). As the protein goes through its catalytic cycle, these domains move relative to each other, enabling the binding and hydrolysis of NTPs, the release of products (NDP and phosphate), unwinding of the helix and translocation along the DNA/RNA molecule. The nucleotide and RNA binding sites are critical to the protein's function and show good conservation among related viral pathogens [44], so designing NTP and RNA-competitive inhibitors presented a viable strategy for the development of novel antivirals.

Figure 4.10, panel B shows that several of the fragment hits bound in interfacial regions between the domains (highlighted in cyan circles), presenting an opportunity for the design of allosteric inhibitors. These could 'lock' the protein in a particular conformation, blocking the inter-domain motions necessary for its

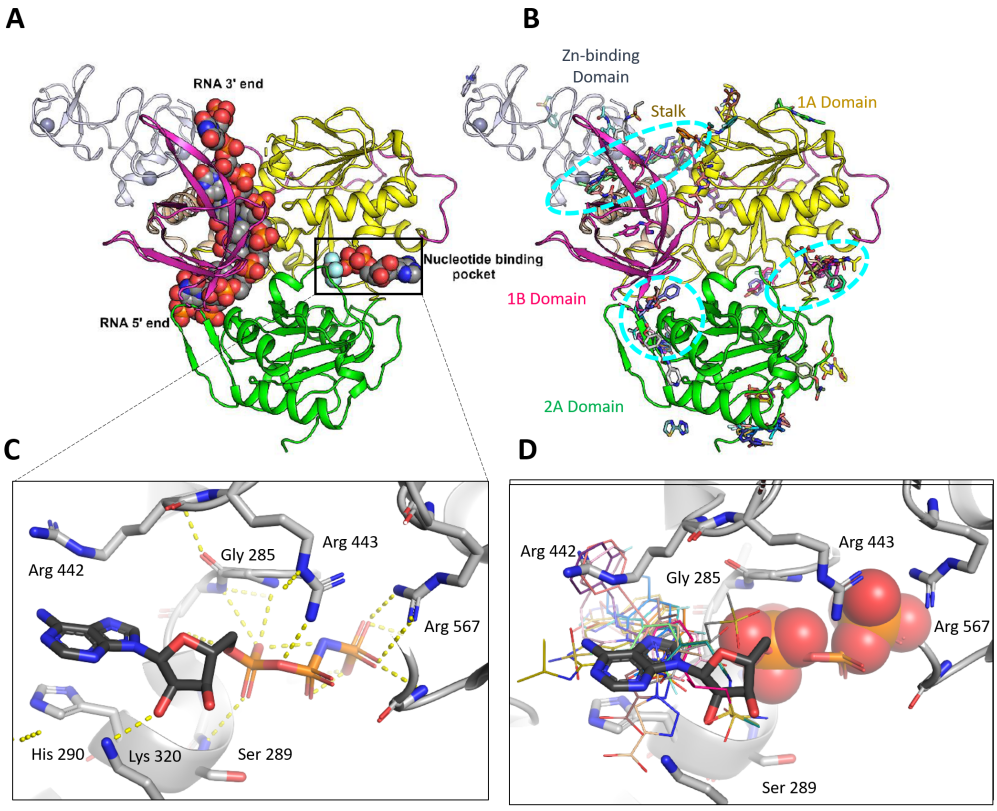


Figure 4.10: NSP13 target overview. A. Model of the NSP13-RNA-NTP complex, based on the structure of the related helicase UPF-1 in complex with RNA (PDB ID 2XZL). Adapted from [192]. B. Crystallographic fragment screening hits in NSP13. The protein is shown as a cartoon and colour coded by structural domain. Bound fragments are shown as sticks. The cyan circles highlight fragments bound in interfacial regions. Adapted from [192]. C. The binding mode of AMP-PNP in the nucleotide binding site (PDB ID 7NN0). The ligand is shown in charcoal, interacting residues are displayed as grey sticks. D. The same view as C, showing the positions of the phosphates and nucleotide site binding fragments relative to the native ligand. Fragments are shown as sticks and phosphates - as spheres.

function in viral replication.

Newman and colleagues solved the structure of the apo form of NSP13 (PDB ID 7NIO), a phosphate-bound form (6ZSL), which was used for the crystallographic screen, and the structure in complex with a non-hydrolysable ATP-analogue (AMP-PNP). Figure 4.10, panel C shows the interactions made by AMP-PNP in the nucleotide binding site. The majority of the hydrogen bond interactions are between the ligand's phosphate groups in arginines 443 and 567. One of the ribose hydroxyls interacts with Lys 320, an interaction that is shared by some of the small molecule fragments that bind in the nucleotide site. The adenine moiety is sandwiched between Arg 442 and His 290. Figure 4.10 panel D shows that most of the bound fragments cluster in the region of the adenine. A subset have sulphonamide moieties, which interact with the proximal phosphate. The two phosphate ions colocalise with the α and γ phosphates on the AMP-PNP ligand. As the release of the distal phosphate ion is part of the helicase's catalytic cycle, inhibitors that prevent its exit from the binding site could potentially be effective antiviral agents.

4.4.2 Results

4.4.2.1 Using the ensemble maps to select which binding site to target

The crystallographic screen against NSP13 had been highly successful, resulting in 65 fragment hits in multiple binding sites on the protein's surface [44]. While the nucleotide binding site was the most populated, and sites binding multiple fragments tend to be more drugable [85, 108, 69, 109], there are other sites in NSP13 that have both biological relevance and are interesting from a pharmacological point of view. However, the crystal structures give no information about the binding affinity of the hits, and the number of hits in a particular site is not, on its own, a sufficient measure of its suitability as an environment for a drug

design campaign (for example, multiple fragments may bind in a crystal contact; particular clusters may also be over or under-populated purely due to a chance combination of factors stemming from the target, fragment library composition, and experimental details). At this stage of the project, no binding affinity data was available for any of the initial fragment hits.

Fragment hotspot maps were used to further help decide which sites and fragments to pursue. Figure 4.11 shows the hotspot maps for the surface of the whole protein (Panel A), as well as for the individual sites with biological relevance (Panels B-F). The RNA 3' site (Figure 4.11 B) showed an interesting fragment merging opportunity. However, the site showed low propensity for the apolar hotspot probes (a strong hydrophobic character is needed for a fragment hotspot, as discussed in Section 1.6.1.1). The site was also provided limited volumes into which the fragments could be grown beyond the fragment merging opportunity.

The RNA 5' site (panel C) showed particularly strong hotspot density in the crevices of the RNA-binding channel. However, the fragments were bound outside the strongest hotspots, and possibly more than one round of elaboration would be needed to reach them. The decision was made to focus on fragment hit for which hotspot-suggested hypotheses could be more rapidly tested, in order to fit within the timelines of the PhD project.

The RNA 5' proximal site showed multiple fragments binding (Figure 4.11, Panel D); however, no hotspot density was observed for the site, and it had a shallow and highly solvent-exposed topology.

The Stalk site is located at the interface between the zinc-binding and stalk domains (Figure 4.10, panel B), close to where the 3' end of the RNA exits the protein. Inhibiting the motion of the Stalk site could potentially block protein motions that are key to its function. The crystal structure shows a long and nar-

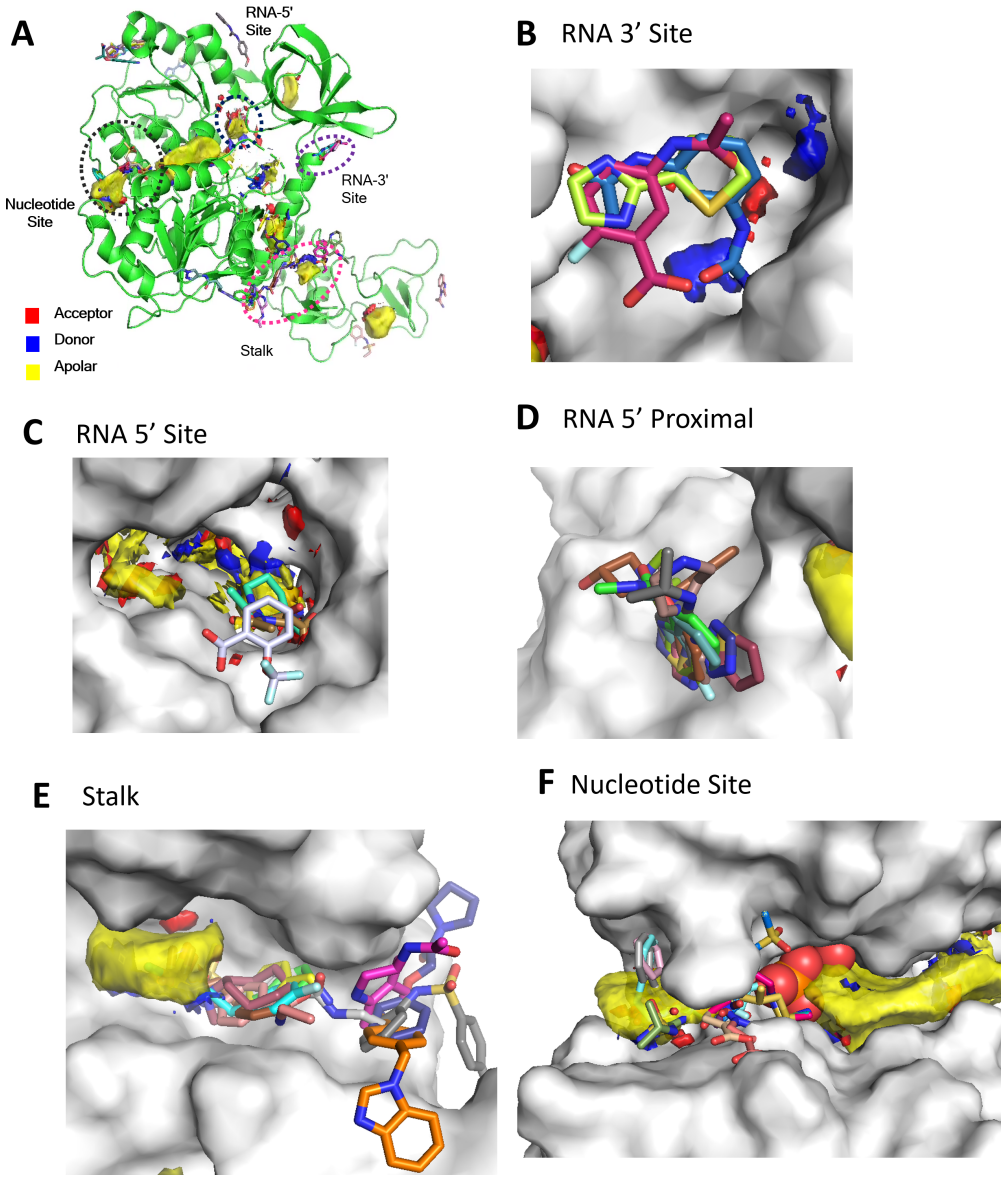


Figure 4.11: Hotspot Maps of the NSP13 fragment sites. A. Ensemble hotspot maps using all structures in the fragment screen. The apolar maps are shown as a yellow surface, hydrogen bond donor: in blue, and acceptor: in red. Maps are shown at a threshold of 14, showing "hot" and "warm" spots[48]. B. Per-site ensemble maps for the RNA 3' site, fragment hits shown as sticks, protein shown as a white surface. C. Per-site ensemble maps for the RNA 5' sites. D. Per-site ensemble maps for the RNA 5' Proximal site. E. Per-site ensemble maps for the Stalk site. F. Per-site ensemble maps for the nucleotide site. The two phosphates are shown as spheres and were not included in the hotspot maps calculation.

row groove (Figure 4.11, panel E) with hotspot density in its innermost portion, as well as a large number of fragment hits. However, only a subset of the hits were located within the hotspot. The fragments bound near the exit of the pocket extended out into solvent and appear to be bound in a crystal contact, rather than being genuine fragment hits.

The nucleotide site (Figure 4.11, panel F) was both most populated (15 fragments, versus 11 in the Stalk site, which is the second most populated), and the majority of these fragments overlap with the hotspot density. There is a second hotspot located after the two phosphate ions in the site, which offered further potential growth vectors. The combination of the site's biological significance, high fragment hit rate, and favourable hotspot profile led to its selection for further hotspot-inspired fragment elaborations.

4.4.2.2 Using the Fragment Hotspot Maps to triage the NSP13 fragment screen hits

Fragment hotspot mapping was used to prioritise the most promising NSP13 fragments for follow-up. Fragments were scored using the Hotspots API against the individual hotspot maps for their parent complex structure. Fragments were sought that possessed good overlap with both the hotspot as a whole and with specific polar interactions, so the mean_hotspot_score and polar_hotspot_score were used to rank the hits, calculated as described in 4.2.1.3. Figure 4.12 shows the relationship between the two scores for all of the fragment hits (across all of the sites).

In the case of NSP13, fragments with a high mean_hotspot_score tended to also have high polar scores. Plots like the one shown in the figure can be output by the workflow to provide an overview of the data and aid in the selection of score

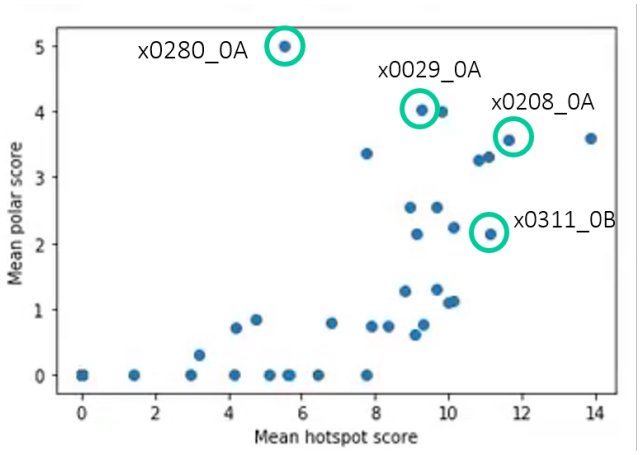


Figure 4.12: Scatter plot of the mean_hotspot_score vs. the mean_polar_score for all of the NSP13 fragment hits. The highest-scoring nucleotide site fragments discussed in the text are highlighted and annotated.

thresholds in the later parts of the workflow. Of the top-scoring fragments by polar score, 4 were located in the Nucleotide Site (highlighted with cyan circles in Figure 4.12). Of these, fragment x0280_0A (shown in Figure 4.13, Panel A) possessed the highest polar score. It makes a hydrogen-bond interaction with Ser 264, satisfying the hydrogen bond acceptor hotspot feature in the hotspot maps at that location. It is also stabilised through a cation-pi interaction with Arg 442, and the thiazole ring is located in a position that would allow elaboration towards the interior of the nucleotide site (and towards the phosphates).

Fragments x0208_0A (Figure 4.13, Panel B) and x0029_0A (Panel C) both make contacts with Ser 264 and Lys 320, satisfying the acceptor hotspots that correspond to these residues. The aromatic ring of x0208_0A is placed centrally within in the apolar hotspot, while in the case of x0029_0A, it is slightly outside the hotspot. It is interesting that while these two fragments have high structural similarity, x0029_0A interacts with Lys 320 through its carbonyl substituent, while in x0208 the carbonyl group interacts with Ser 264, a reversal of the binding mode.

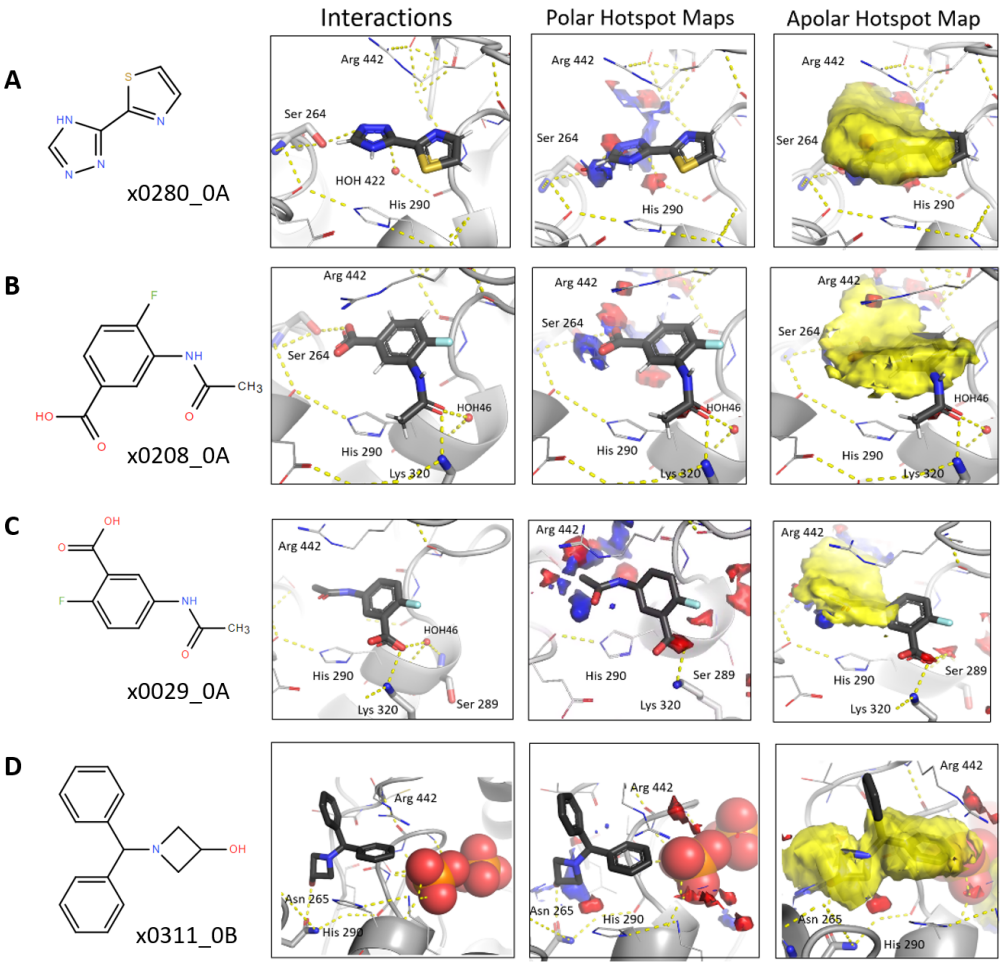


Figure 4.13: Using the hotspot maps to triage NSP13 fragment hits. The four columns show the fragments' 2D structures, binding site interactions, polar hotspot maps, and apolar hotspot maps calculated for the NSP13-fragment holo structures. Rows (panels) A-D correspond to the selected starting fragments discussed in the text.

Fragment x0311_0B induced a displacement of Arg 442, which was also seen in a subset of the other nucleotide site fragment hits. The fragment had good overlap with the apolar hotspot and made a polar interaction with Asn 265, satisfying the donor hotspot at that location.

Together, these four fragments provided a set of starting points with good fit to the nucleotide binding site hotspot features, and provided opportunities for further elaboration towards the phosphate binding region.

4.4.2.3 Using the ensemble NSP13 maps and selectivity maps over UPF-1 to prioritise docked follow-up poses

The selected nucleotide site fragments had good fit to the hotspot maps in the site. However, elaborating towards the phosphate binding sites would mean extending beyond the main nucleotide site hotspot, so taking the mean hotspot scores when re-scoring poses would bias against larger ligands, which may otherwise be extending in the desired direction. To avoid such cases, the by-feature re-scoring procedure presented in Section 4.2.1.3 was used. Clusters were detected within the polar ensemble maps as described in Section 4.2.1.3. The fragment network was queried as described in Section 4.2.1.5; in addition to the cLogP < 3.0 and molecular weight <350 Da criteria for lead-likeness, compounds with molecular weight < 250 Da were discarded. This modification aimed to bias the search towards compounds that were larger than the initial fragments, providing the possibility to grow towards the phosphate binding site. Overall, this resulted in 2964 compounds which proceeded on to the constrained docking step. The full list of compounds is available in the Supplementary Data for this thesis, in the file "nsp13_all_followups.csv".

A scaffold match constraint based on the maximum common substructure between

the starting fragment and the follow-up compound was introduced in the docking step. This would allow additional parts of the ligand to explore areas of the binding site not covered by the starting fragment. The default workflow value of 50 poses/compound was used for the docking step. The pose that scored highest against the Nucleotide Site maps (using the mean_hotspot_score) was selected. While comparing ligands of different sizes by the mean hotspot score would generally de-prioritise the larger ligand, using it to compare different conformations of the same molecule would prioritise those that have better fit overall with with the hotspot, especially with regards to the apolar maps. Poses were then filtered based on the number of polar clusters that they interacted with. The starting fragments interacted with either one (x0029_0A, x0280_0A) or two of the detected polar clusters. To ensure that follow-up compounds had at least equally good fit with the detected interactions, those that interacted with 2 or more clusters were selected. This brought the selection down to 412 compounds.

Another point to consider was the fact that the nucleotide site is conserved between helicases. In the case of antiviral agents, this could enable the design of compounds that are effective against multiple pathogens. However, it also creates the potential for off-target effects in human patients, whose cells also produce and use family 1B helicases. Selectivity maps were used to identify differences in the binding sites of NSP13 and the related [44] human family 1B helicase UPF-1 and inform follow-up fragment designs. The selectivity maps against UPF-1 are shown in panels A and B in Figure 4.14. The selective region is located towards the edge of the nucleotide site, in the vicinity of Ser 264 in NSP13. In UPF-1, this volume is blocked by the position of the α-helix shown in Figure 4.14, panel B. When compared to the ensemble map clusters (panels C and D in the same figure), the selective region co-localised with acceptor clusters 1 and 2 and donor cluster 1 and 2. Acceptor cluster 3 and donor cluster 3 are located in the phosphate binding

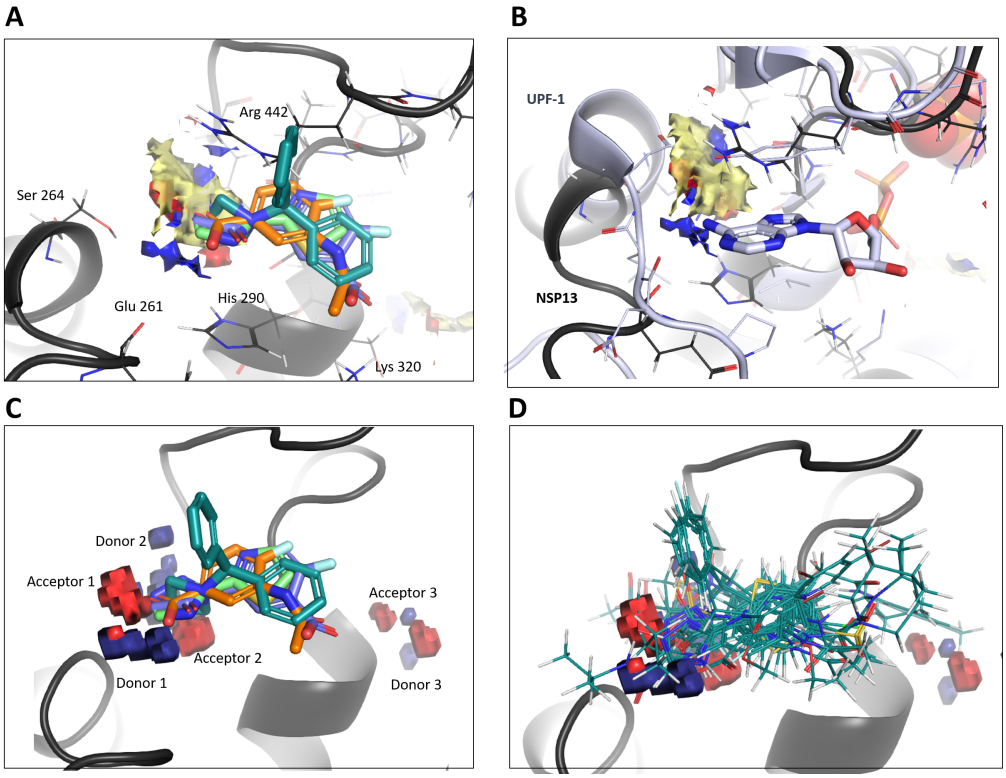


Figure 4.14: Re-scoring poses uses NSP13 ensemble and selectivity maps. A. Selectivity maps for NSP13 over UPF-1. NSP13 is shown as a black cartoon, residues interacting with the bound fragments are shown as sticks. The four starting fragments are shown as coloured sticks B. Selectivity maps for NSP13 over UPF-1. UPF-1 (PDB ID 2GK6) is shown as a light blue cartoon, and the bound AMP molecule - as sticks. Residues interacting with the ligand are shown as lines. C. Polar clusters identified in the ensemble maps for NSP13. Interacting residues are shown as lines, and the four starting fragments - as coloured sticks. D. Selected docked follow-up poses are overlaid and shown as teal sticks. Polar ensemble maps clusters are shown as isosurfaces (same as panel C).

region of the nucleotide site, which is shared with UPF-1 (Figure 4.14, B), and so they are not present in the selectivity maps. An overlay of the bound starting fragments shows that they all interact with at least one of the predicted selective clusters (Panel C). Therefore, it was hypothesised that by maintaining the interactions in this area of the binding site, compounds that bind only to the viral target (due to the unavailability of these interactions in the human target) would be prioritised. Using the selectivity maps to further prioritise the shortlisted compounds resulted in a selection of 29 compounds.

4.4.2.4 Dynamic Undocking on the NSP13 fragments and follow-up compounds

OpenDUck was run on both the interactions made by the selected NSP13 fragments, as well as on the docked poses of the suggested follow-up compounds. Figure 4.15 shows that none of the interactions passed the structural stability threshold of 6 kcal/mol, introduced in [47]. Visual inspection of the nucleotide site shows that it is open to solvent, and so a number of the interactions appear solvent-exposed. Majewski *et al.* observed that protein atoms with a low solvent accessible surface area (SASA) are a necessary but not sufficient condition for the structural stability of an interaction.

This is illustrated in panel A in Figure 4.15, showing the spread of OpenDUck $W_{QM\_mean\_min}$ values for the interactions in the IRIDIUM set. When compared to the profiles of the selected NSP13 fragment hits (panel B) and docked poses of the follow-up compounds (panel C), it can be seen that a number of protein interacting atoms in NSP13 have SASA above 20 Å (LYS_320_NZ, SER_259_OG, SER_259_O, SER_264_OG, ARG_442_NH2), meaning they are unlikely to make structurally stable interactions. In the cases with low SASA atoms, the low $W_{QB\_mean\_min}$ values are not a product of the increased solvent accessibility. It

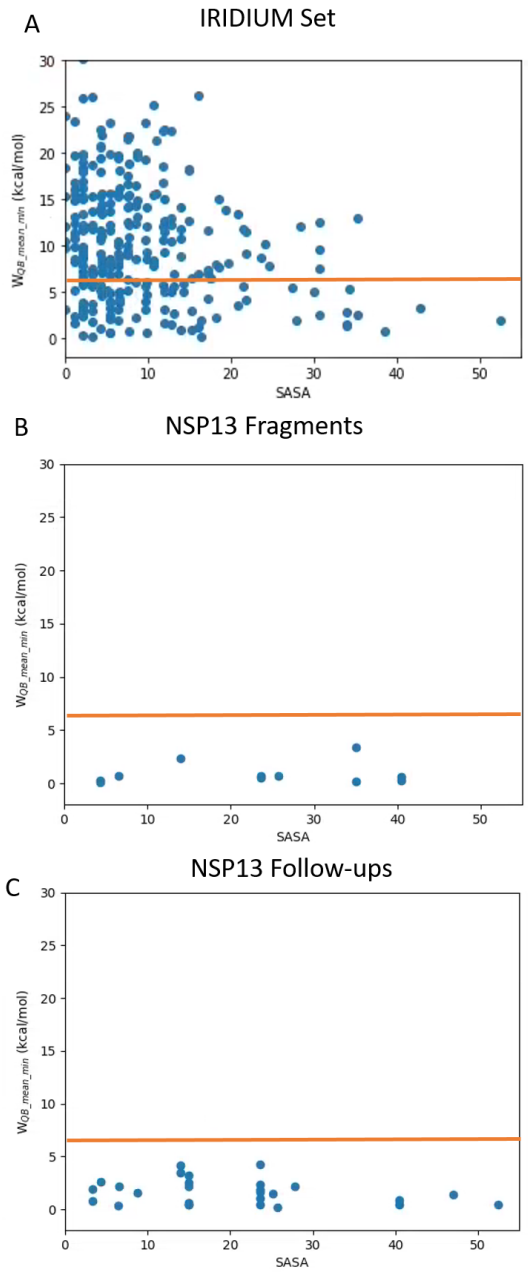


Figure 4.15: Relationship between the $W_{QB\_mean\_min}$ and SASA for interactions in the benchmark IRIDIUM set (A), the selected NSP13 fragments (B) and follow-ups (C). The orange lines indicate the structural stability threshold of 6 kcal/mol [47].

is likely that these interactions do not confer structural stability for other reasons. Unsurprisingly, all but one of the highly solvent exposed protein atoms are in the amino acid side chains, rather than being part of the backbone. Despite the increased solvent accessibility, these interactions do have corresponding hotspot features that are satisfied by the starting fragments.

4.4.2.5 Functional assay results for the ordered follow-up compounds

A total of 29 compounds were selected according to the procedure described in Section 4.2.1.3 and 4.4.2.3 and ordered for purchase. Of these, the 16 shown in Figure 4.16 could be synthesised and were delivered. Activity data was collected by Joseph Newman using an ADP-Glo assay for ATPase activity. Of the 16 compounds tested, only 6 showed detectable inhibition at 500 μM. Four of these followups were elaborated from x0208_0A, one from x0029_0A, and one from x0280_0A.

A dose-response curve was collected for the most potent hit (nsp13-x0208_0A_1, shown in Figure 4.17), showing an IC_{50} of around 250 μM.

The predicted binding mode of this follow-up, as well as that of the other five hits that showed detectable inhibition, are shown in Figure 4.18. Follow-up nsp13-x0208_0A_1 makes an H-bond to Asn 265 through its nitrile group and to Ser 264 via the carbonyl oxygen. It is important to note, however, that the compound's nitrile group is part of an electrophilic pattern that is known to react with cysteine residues and has been used in the development of covalent inhibitors [193]. There are 25 cysteines in NSP13, mostly located in the protein's zinc binding domain. In the absence of a crystal structure, it cannot be ruled out that this follow-up may bind covalently to the protein. This highlights the importance of introducing additional filters for compounds containing such groups, especially when the target

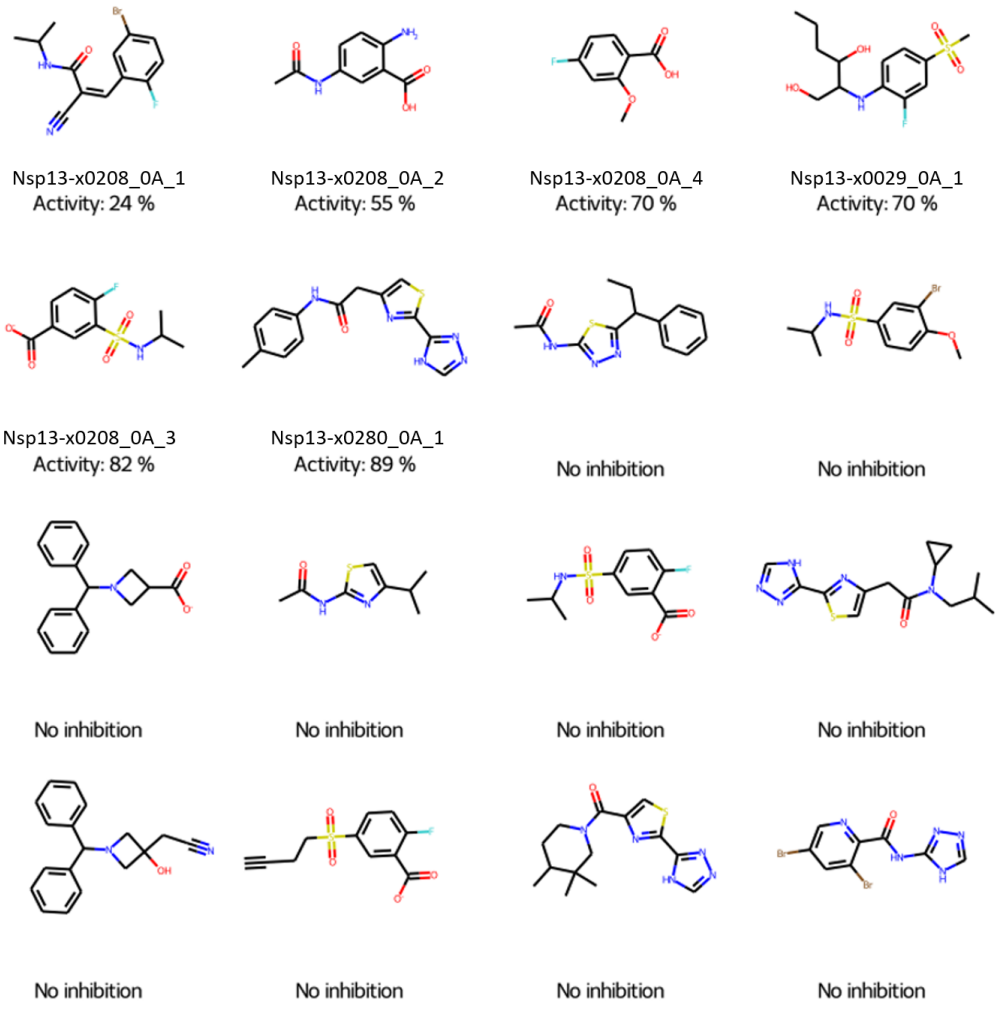


Figure 4.16: ADP-Glo assay results for the ordered NSP13 follow-up compounds.

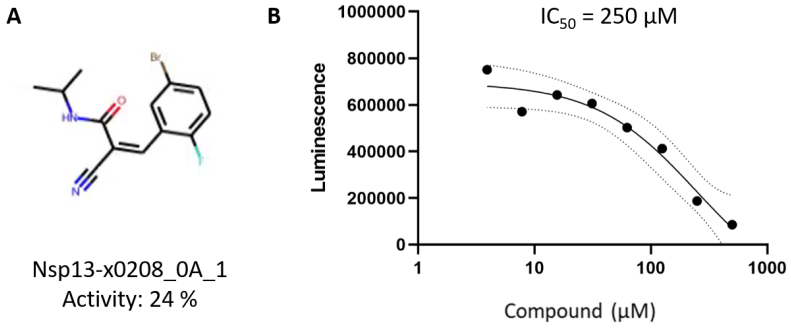


Figure 4.17: Activity and IC_{50} of the top NSP13 followup. Panel A shows the 2D structure of Compound nsp13-x0208_0A_1 and the activity recorded in the ADP-Glo assay. Panel B shows the dose response curve for the same compound

.

contains multiple cysteine residues.

Follow-up nsp13-x0208_0A_2 satisfies the donor hotspot feature through an H-bond with the backbone O of Thr 286 through the ligand's amine group. The carbonyl group makes the same acceptor interaction with Ser 264 as the parent fragment. nsp13-x0208_0A_3 also makes this interaction, with a second hydrogen bond to Lys 320 through the ligand's sulfonamide moiety.

nsp13-x0280_0A_1 also makes an interaction with Ser 264 via the substructure it shares with the parent fragment. This follow-up is also predicted to extend towards the phosphate binding locations within the nucleotide site.

Follow-up Nsp13-x0029_0A_1 is predicted to make interactions with Ser 264 and Asn 265, satisfying the two acceptor clusters in the ensemble maps. Its sulfone moiety extends towards the phosphate binding region, in addition to making a hydrogen bond with Lys 320.

Overall, the follow-up compounds were not strong binders, with binding affinities in the high micromolar range. However, the follow-up compounds are not much larger than the starting fragment hits, and the fact that inhibition can be observed

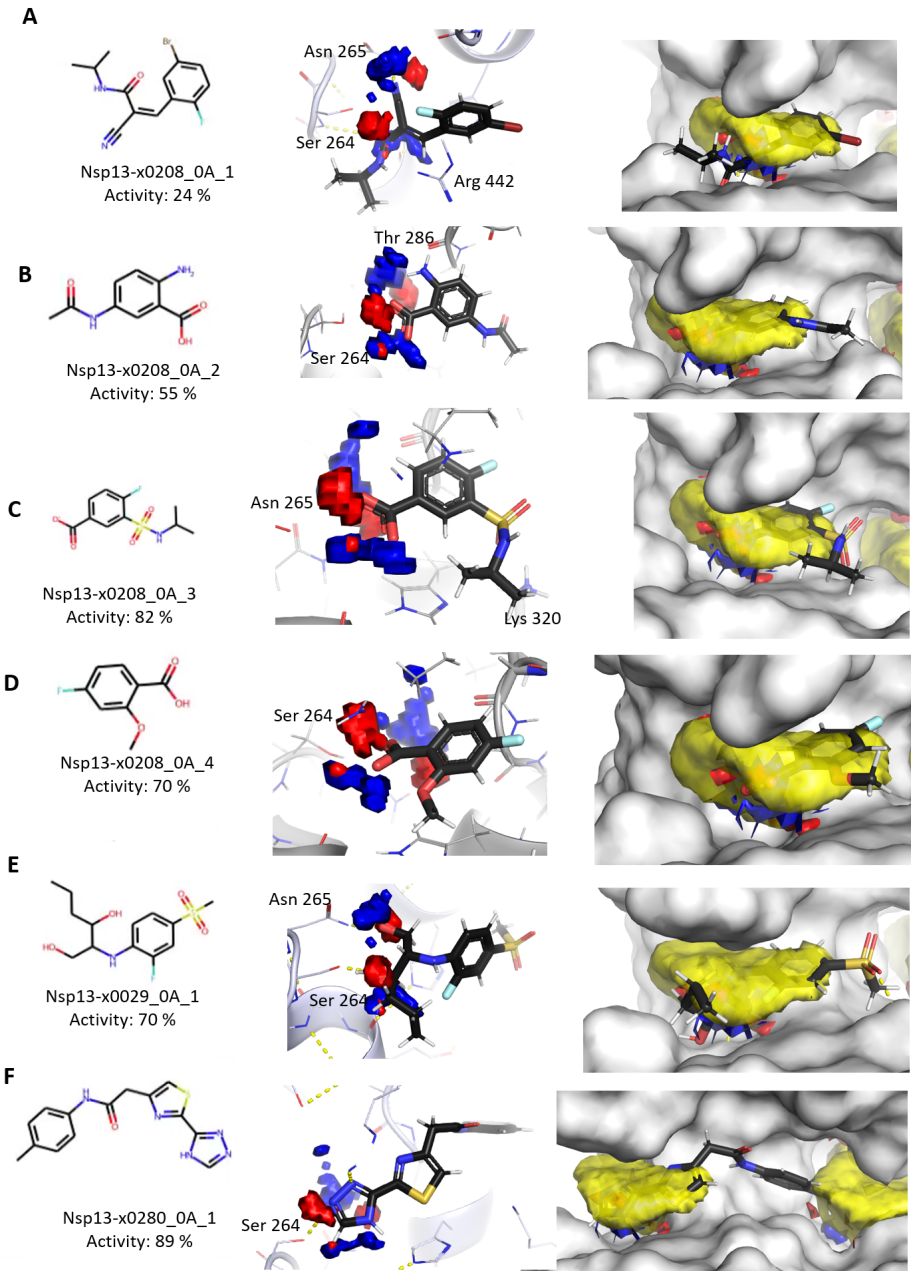


Figure 4.18: Predicted binding modes and hotspot interactions of the NSP13 follow-up hits. The middle column shows the detected polar clusters in the nucleotide ensemble maps

.

at such an early stage was encouraging and presented an improvement over the initial fragment hits. Crystallographic studies for confirming the binding modes of the active hits are planned for the near future.

4.5 Case Study: PARP14

Human Poly(ADP-ribose) Polymerase Family Member 14 (PARP14) has emerged as a promising target for the treatment of tumours and allergic inflammation [194]. A fragment screening campaign against the third macrodomain of PARP14 was carried out at SGC-Oxford in 2017 by Marion Schuller [195].

4.5.1 Target overview

PARP14 is an ADP-ribosyltransferase that mediates mono-ADP ribosylation of glutamate residues on target proteins [194]. ADP-ribosylation is a post-translational modification that acts as signal in multiple stress response pathways, including DNA damage repair and the inflammatory response. PARP14 exerts its action by transferring ADP-ribose (ADPR) using NAD^+ as substrate. Figure 4.19, panel A shows the domain structure of the PARP14 gene. The catalytic PARP domain catalyses the substrate transfer reaction. The protein also has two RNA-recognition motifs(RRMs), a WWE domain, and three macrodomains labelled MD1-3. Macrodomains are readers of mono-ADP ribosylation. Inhibiting their action through an ADPR-competitive inhibitor could modulate PARP14's function in disease-implicated signalling pathways.

The crystallographic fragment screen resulted in 18 hits, located in the 3 sites shown in Figure 4.19. One of the sites was a crystal contact, while the other two co-localised with the adenine and ribose binding sub-pockets within the ADPR binding site. As a general trend, the fragment hits in these sites placed aromatic

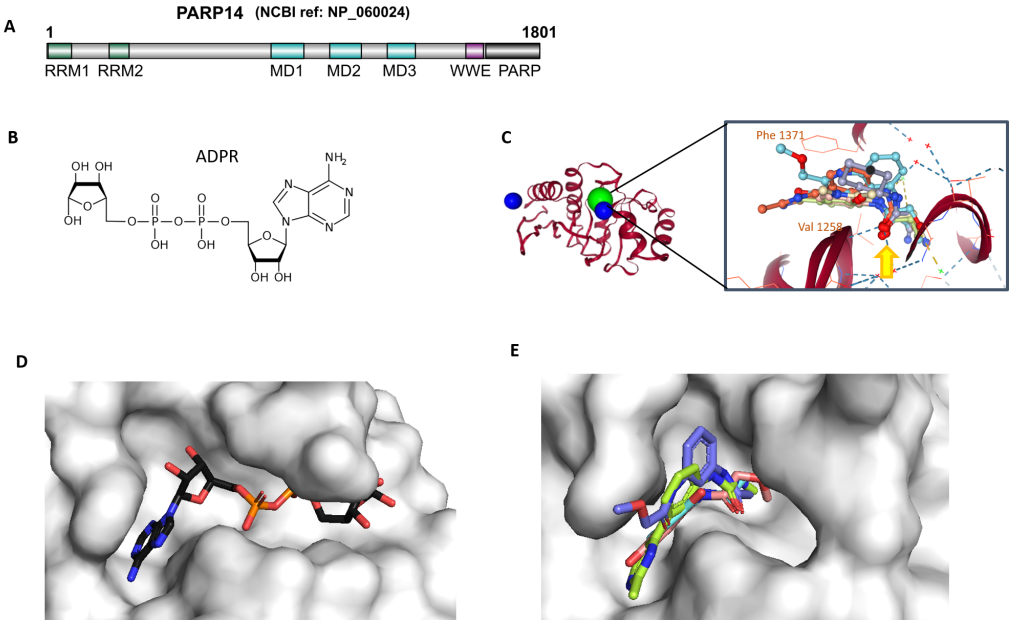


Figure 4.19: Targeting the third macrodomain of PARP14. A. Domain structure of the PARP14 gene. B. 2-D chemical structure of ADPR. C. Fragment binding sites detected in the crystallographic screen. The pop-out shows the binding mode of the fragment hits that made polar interactions with the protein. The yellow arrow highlights the interaction with the backbone of Val 1258, shared by the fragments.D. Crystal structure of PARP14 MD3 in complex with ADPR (PDB ID 4ABK). E. The fragment hits making polar interactions with the protein, using the same view as D. The protein structure shown is the complex with fragment x0161 (PDB ID 5QHU).

moieties stacked between Val 1258 and Phe 1371 (Figure 4.19, panel C). Of the fragments that made polar interactions with the protein, all made a hydrogen bond with the backbone of Val 1258 through a carbonyl oxygen on the ligand (highlighted with an arrow in Figure 4.19, panel C). The phosphate and distal ribose regions of the ADPR binding site showed a level of flexibility, as the fragment-bound structures were more enclosed in that region (Figure 4.19, Panel E) compared to the ADPR complex (Panel D in the same figure). Overall, the results of the fragment screen, in combination with the binding mode of the native ligand, presented a favourable starting point for further elaboration.

Prior to the work presented in this chapter, two attempts had been made to design follow-up compounds based on this fragment screen. In both cases, the resulting compounds were not found to bind in an AlphaScreen assay optimised for the PARP14 macrodomains, which could detect binding up to 200 μM [195]. The first series of compounds were a set of bicyclic amides and ureas designed by Vinnie Fagan in the Brennan group. They were designed to target the adenine and proximal ribose sites shown in Figure 4.19, panel C, while making the fragment-inspired interaction with the backbone NH of Val 1258. The second series was a set of Fragalysis enumerations and catalogue compounds selected by Gian Filippo Ruda (who performed the enumeration and docking) and I (provided the hotspot maps) based on docking scores and the hotspot maps (visual inspection of the fit between the top docking poses and the hotspot features). These compounds were designed to target the adenine site, and included fused rings to maximise overlap with the apolar hotspot. The Fragalysis enumerations were based off fragment x0161, and were generated via one "pass" along the reaction vectors. The lack of binding observed for these two compound series led to the following conclusions.

1. For compounds based on such weakly binding initial hits, crystallographic data could help assess whether the suggested compounds are making the

desired interactions, even if no binding could be detected in the assay. If the follow-up has better overlap with the key hotspot interactions, it would be a better starting point for subsequent rounds of chemistry.

2. In the case of fragment x0161, the modifications suggested by one pass along the fragment network vectors may have been too conservative. To allow for larger, more complex ligands to be suggested, two passes along the vectors were used in subsequent work.

4.5.2 Results

The fragment hits that made polar contacts with the protein all had similar structure and binding modes, making a conserved interaction with Val 1258 (shown in Figure 4.20). The first step in the follow-up process involved prioritising the fragment hits.

4.5.2.1 Using Dynamic Undocking and Fragment Hotspot Maps to prioritise the PARP14 fragment hits.

When previously tested in an AlphaScreen assay optimised for the PARP14 macrodomains, none of the fragment hits had shown detectable activity up to 200 μM compound concentration [195]. Therefore, computational methods were used to prioritise fragments within the ADPR binding pocket. In the first step of the workflow, hotspot maps were calculated for the protein structures and ensemble maps were generated for the binding site. The site was defined by residues placing heavy atoms within 6 Å of any of the bound fragments in the site (x0457, x0712, x0324, x0161, x0505, x0590, x0315, x0334). The ensemble maps (Figure 4.20) showed that the fragments' cyclic moieties overlapped with the highest scoring areas of the apolar maps, while all of the polar interactions satisfied corresponded

to the acceptor feature arising from the backbone NH of Val 1258 (with contributions from the backbone NH on Gly 1334). Since the starting fragments all interacted with the same polar feature along a single interaction (all make the interaction with the backbone NH of Val 1258; x0315 and x0324 make a further interaction with the backbone of Gly1334), I decided to investigate the stability of the complexes along this interaction as a possible way of ranking the fragments.

OpenDUck was used to investigate the structural stability of the interaction with Val 1258 between the fragments of interest. 20 SMD runs/temperature were performed, and the mean minimal $W_{QB\_mean\_min}$ was calculated as described in Section 4.2.1.4. The results in Table 4.1 show that for all of the fragments, the interaction was not predicted to be stable, and the values are mostly within a standard deviation of each other. This is perhaps not surprising, as both the fragment hits and their binding modes are highly similar. The solvent accessible surface area for the Val 1258 backbone NH is not high (under 10 Å in all of the fragment-bound structures), so the low W_{QB} values cannot be attributed to high solvent exposure.

Table 4.1: W_{QB} values for the interaction between the PARP14 fragment hits and Val 1258.

Fragment	**$W_{QB\_mean\_min}$ (kcal/mol)**	**St.Dev.**
PARP14-x0457	2.8	± 0.7
PARP14-x0712	2.2	± 0.9
PARP14-x0161	1.4	± 0.8
PARP14-x0324	1.2	± 1.4
PARP14-x0505	0.9	± 1.0
PARP14-x0590	0.7	± 1.2
PARP14-x0315	0.4	± 1.6
PARP14-x0334	0.3	± 1.3

A combination of fragment hotspot mapping and automated interaction detection was used to further prioritise the fragment hits. Selected fragments made the conserved interaction with Val 1258. In addition, fragments that placed all of

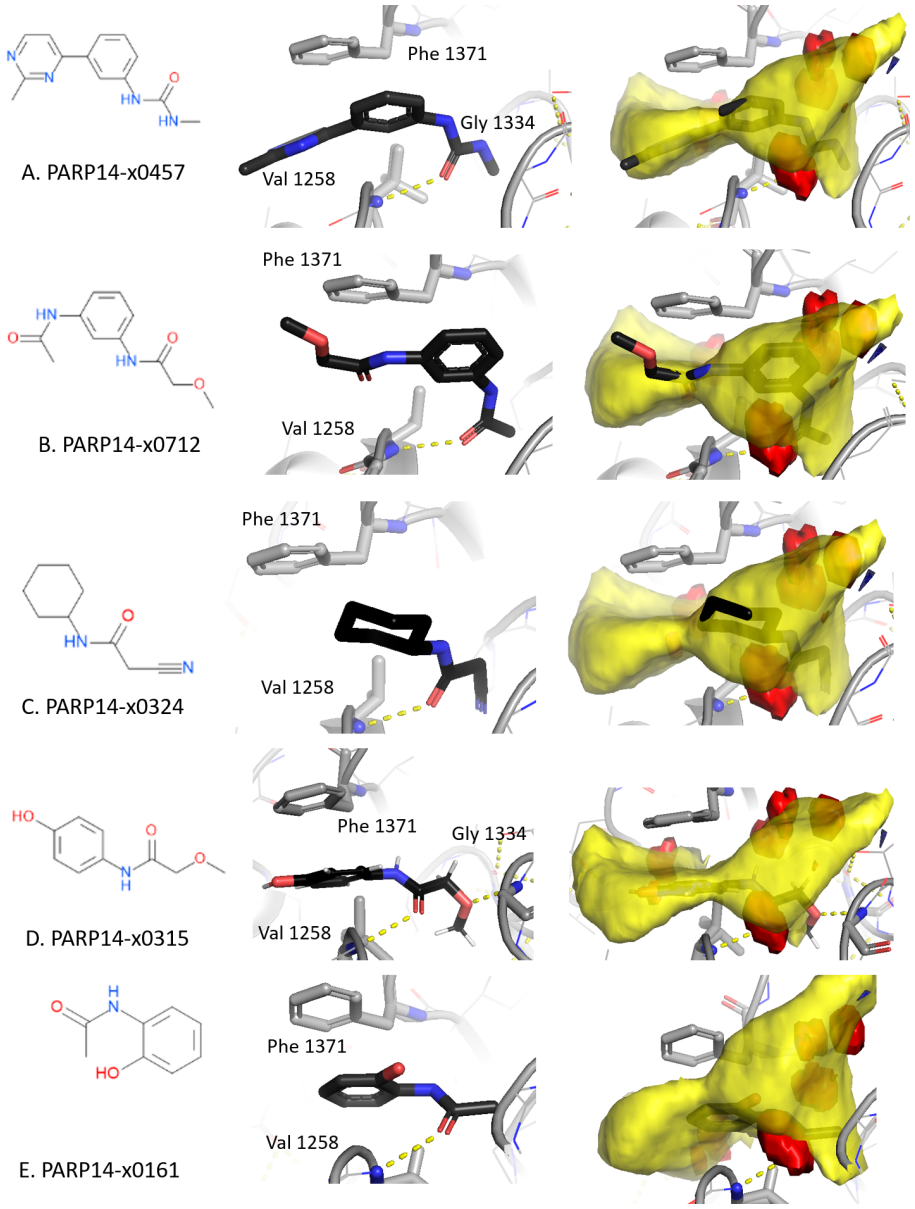


Figure 4.20: Interactions and fragment hotspot map profiles of the selected PARP14 fragments.

their aromatic atoms within the apolar hotspot (x0457, Figure 4.20 , Panel A) or made additional polar interactions detected via the ODDT toolkit [181] (x324 (Panel C), x315 (Panel D) in Figure 4.20) were selected. The binding modes and interactions made by the fragments, as well as their overlap with the ensemble hotspot maps, are shown in Figure 4.20.

All of the fragments occupy the adenine and part of the ribose binding subsites of the ADPR site. The fragments' apolar rings are sandwiched between Phe 1371 and Val 1258. Fragment x0315 also makes a further polar interaction with the backbone of Gly 1334. However, none of the fragments satisfied any of the detected polar clusters, apart from the one corresponding to Val 1258.

4.5.2.2 Using the ensemble map polar features to re-score docked follow-up compounds

Using the procedure described in Section 4.2.1.3, four clusters were located in the ensemble hotspot maps for PARP14. These are shown in Figure 4.21, and co-localise with interactions made by the ADPR native ligand (Figure 4.21, C). Acceptor cluster 1 arises from the backbone of Ile 1236 and constitutes an interaction exploited by the acceptor nitrogen on the ligand's adenosine moiety. Acceptor cluster 2 corresponds to the backbone of Val 1258, the interaction shared by the PARP14 fragment hits. In ADPR, the proximal phosphate makes this interaction. The last two polar clusters, Acceptor 3 and 4, are located in the proximal ribose binding pocket and correspond to interactions with the backbone NH of Phe 1371 (Acceptor 3), and the side chain of Thr 1333 and backbone of Gly 1334 (Acceptor 4).

In the enumeration stage, the fragment network was queried as described in Section 4.2.1.5, using two passes along the fragment vectors. In addition to the cLogP

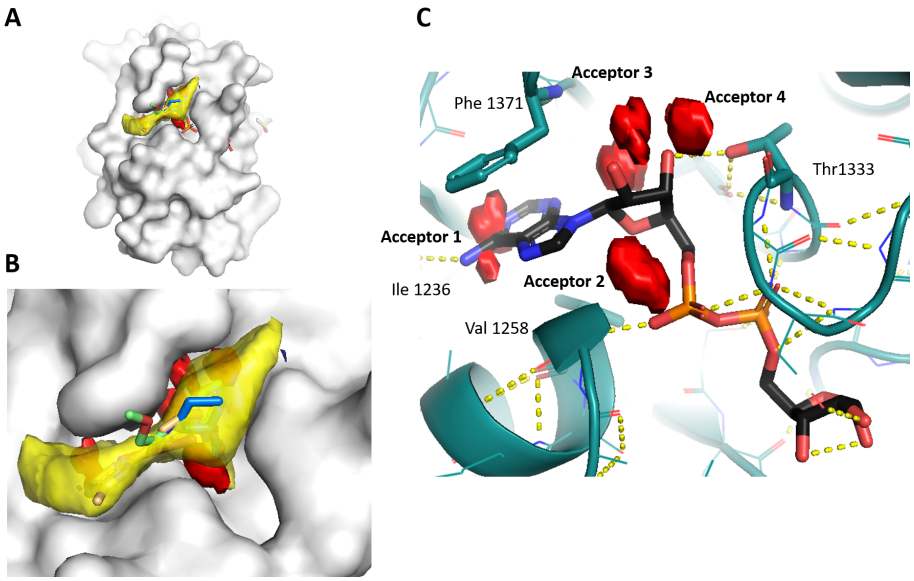


Figure 4.21: PARP14 ensemble maps. A. PARP14 MD3 is shown as a white surface. The ensemble maps highlight the location of the ADPR-binding pocket. B Close-up of the binding pocket with the selected starting fragments overlaid. C. Polar clusters detected in the ensemble maps. The native ADPR ligand is overlaid (PDB ID 4ABK) and shown in black. Key interacting residues are shown as sticks.

< 3.0 and molecular weight <350 Da criteria for lead-likeness, compounds with molecular weight < 250 Da were discarded, as well as any with more than 5 rotatable bonds. These modifications aimed to bias the search towards compounds that were larger than the initial fragments, without introducing excessive flexibility, which would not be desirable in a lead compound [183]. This resulted in a total of 4857 compounds, which progressed to the constrained docking stage. The full list of these compounds is available in the Supplementary Data for this thesis, in the file "PARP14_all_followups.csv".

Docked poses for the PARP14 follow-ups were re-scored using the mean hotspot score, and the highest scoring pose was selected for each follow-up. These were then scored against the polar clusters and ranked by the number of these polar clusters that they interacted with (389 were found to interact with at least 1 cluster). As the starting fragments all interacted with a single cluster (Acceptor 2), only follow-ups that interacted with two or more clusters were selected. This resulted in the set of compounds shown in Figure 4.22, which were purchased and tested in a HTRF assay by James Bennett.

4.5.2.3 Structural stability of the prioritised follow-up compounds

The procedure described in the previous section narrowed the selection down to 22 compounds, which could all be ordered within the allocated budget. The deadline for ordering the compounds was such that OpenDUck could not be run in time to further prioritise the suggested follow-up compounds, as at the time the workflow had not yet been set up to automatically prepare and run follow-up OpenDUck calculations. OpenDUck was run retrospectively on the chosen compounds in order to prioritise those making interactions that satisfied the same hotspot cluster (feature). In the full workflow, the detected interactions would be mapped automatically to their nearest hotspot cluster, with the option for the user to review the

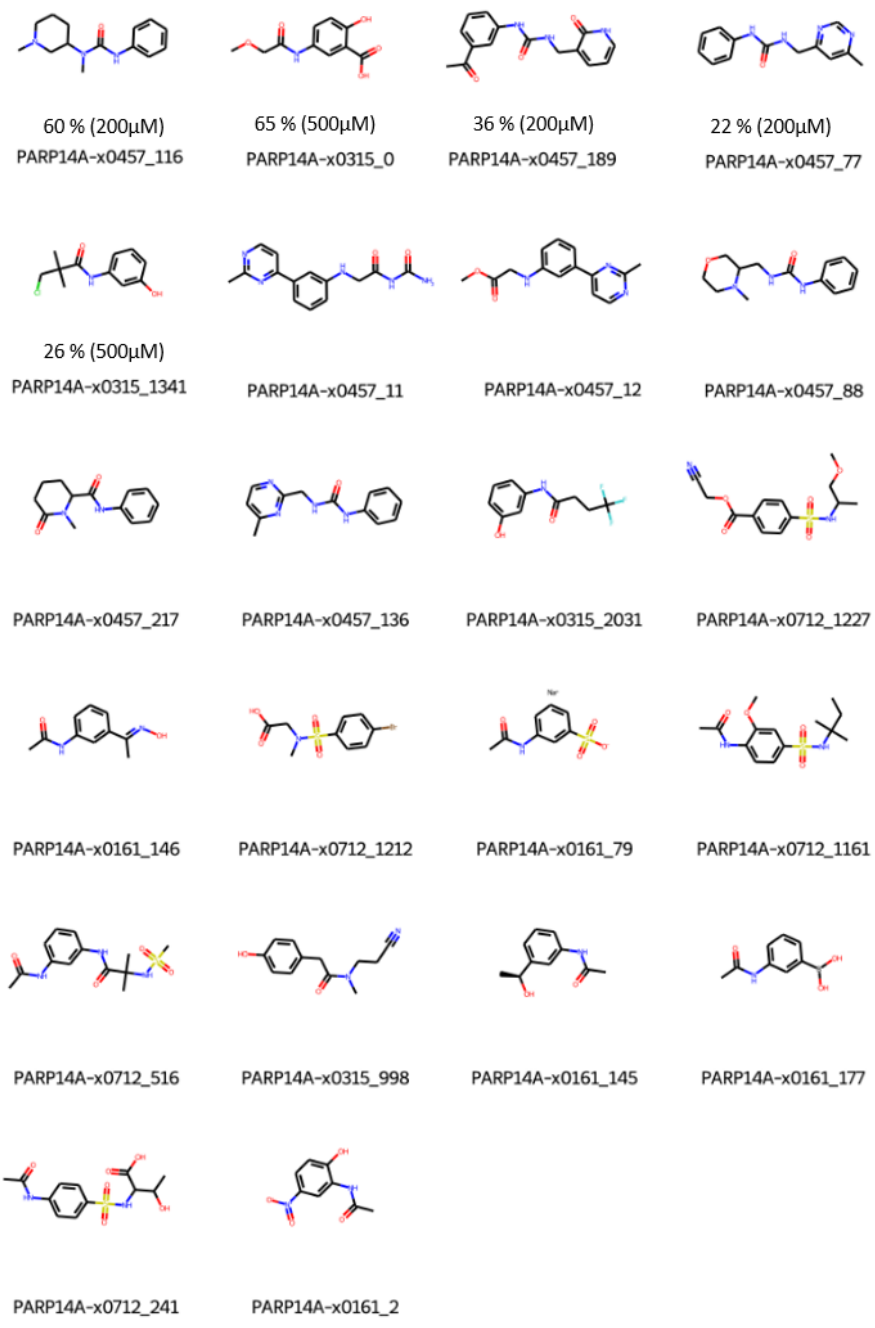


Figure 4.22: Selected follow-up compounds for PARP14.

allocation. OpenDUck would then be used to prioritise compounds and groups that rank highly for a particular interaction.

Despite the low W_{QB} values for the starting fragments, a large number of the follow-up interactions passed the stability threshold. Each of the five compounds with detectable inhibition has at least one high-scoring $W_{QB\_mean\_min}$ interaction in either of the clustering sites (Table 4.2 and Table 4.3). This suggests that in the cases where the pipeline up to this point produces a larger amount of compounds, further shortlisting can be done by taking the highest scoring DUck values for each cluster. Curiously, none of the compounds making interactions in clusters 1 and 3 (shown in Appendix A.8) showed activity in the assays, despite the favourable hotspot scores and predicted structurally stable interactions. Crystallographic data would be needed to show whether the interactions are made; it is possible that as the starting fragments were very weakly bound, the followups may also bind too weakly to be detected, but have improved interaction profiles.

4.5.2.4 Results of the PARP14 binding assays

Of the twenty-two compounds selected and tested against PARP14, five showed detectable inhibition in the binding assay. Figure 4.23 shows their predicted binding modes and hotspot interactions. The compounds follow the same pattern as the interactions made by the initial fragments in the hydrophobic region between Phe 1371 and Val 1258, and all also interact with the main acceptor hotspot (Acceptor cluster 2), corresponding to the Val 1258 backbone.

The elaborated parts of the compounds explored the proximal ribose binding pocket (shown in panels B and D in figure 4.23), or extended towards the phosphate binding region (the compounds shown in panels A, C, E). The top ranked compound is predicted to retain the interaction with the backbone of Gly 1334

Table 4.2: Structural stability ($W_{QB\_mean\_min}$) of interactions for Hotspot cluster 2

Follow-up	**Interaction**	**$W_{QB\_mean\_min}$ (kcal/mol)**	**StDev (kcal/mol)**
PARP14A-x0712_1212	A_VAL_1258_N	5.9	0.7
PARP14A-x0315_1341	A_VAL_1258_N	4.7	1.1
PARP14A-x0315_2031	A_VAL_1258_N	4.6	0.2
PARP14A-x0712_516	A_VAL_1258_N	4.3	2.4
PARP14A-x0712_1257	A_VAL_1258_N	4.2	1.5
PARP14A-x0161_169	A_VAL_1258_N	4.0	0.3
PARP14A-x0457_168	A_VAL_1258_N	3.3	0.8
PARP14A-x0457_189	A_VAL_1258_N	2.9	0.3
PARP14A-x0712_1227	A_VAL_1258_N	2.4	0.6
PARP14A-x0712_534	A_VAL_1258_N	2.3	0.4
PARP14A-x0457_136	A_VAL_1258_N	1.6	0.4
PARP14A-x0457_88	A_VAL_1258_N	1.2	0.3
PARP14A-x0161_146	A_VAL_1258_N	0.9	0.6
PARP14A-x0457_217	A_VAL_1258_N	0.7	0.2
PARP14A-x0457_116	A_VAL_1258_N	0.6	0.2
PARP14A-x0712_1161	A_VAL_1258_N	0.6	0.2
PARP14A-x0315_998	A_VAL_1258_N	0.5	0.3
PARP14A-x0457_113	A_VAL_1258_N	0.4	0.1
PARP14A-x0457_190	A_VAL_1258_N	0.2	0.1
PARP14A-x0457_77	A_VAL_1258_N	0.2	0.1
PARP14A-x0457_168	A_SER_1259_N	0.2	0.0

Table 4.3: Structural stability ($W_{QB\_mean\_min}$) of interactions for Hotspot cluster 4

Follow-up	Interaction	$W_{QB\_mean\_min}$ (kcal/mol)	StDev (kcal/mol)
PARP14A-x0457_189	A_THR_1333_OG1	9.0	2.9
PARP14A-x0712_516	A_THR_1333_OG1	8.1	0.3
PARP14A-x0712_241	A_THR_1333_OG1	7.0	0.7
PARP14A-x0712_1212	A_THR_1333_OG1	7.0	1.1
PARP14A-x0712_534	A_THR_1333_OG1	6.8	1.5
PARP14A-x0712_1257	A_THR_1333_OG1	3.4	0.8
PARP14A-x0457_116	A_THR_1333_OG1	2.2	0.6
PARP14A-x0457_88	A_THR_1333_OG1	2.2	0.0
PARP14A-x0457_11	A_THR_1333_OG1	1.3	1.1
PARP14A-x0315_0	A_THR_1333_OG1	1.2	0.6
PARP14A-x0712_1161	A_THR_1333_OG1	0.8	0.1
PARP14A-x0315_781	A_THR_1333_OG1	0.8	0.3
PARP14A-x0457_4	A_THR_1333_OG1	0.4	0.1
PARP14A-x0457_77	A_THR_1333_OG1	0.2	0.0
PARP14A-x0712_1227	A_THR_1333_OG1	0.2	0.2
PARP14A-x0457_12	A_THR_1333_OG1	0.1	0.0
PARP14A-x0457_136	A_THR_1333_OG1	0.1	0.0
PARP14A-x0457_77	A_GLY_1334_N	12.6	0.7
PARP14A-x0457_136	A_GLY_1334_N	12.5	1.3
PARP14A-x0712_1212	A_GLY_1334_N	9.3	0.0
PARP14A-x0315_781	A_GLY_1334_N	3.3	0.5
PARP14A-x0712_1227	A_GLY_1334_N	3.0	2.0
PARP14A-x0315_0	A_GLY_1334_N	1.6	0.2
PARP14A-x0712_1161	A_GLY_1334_N	0.3	0.1

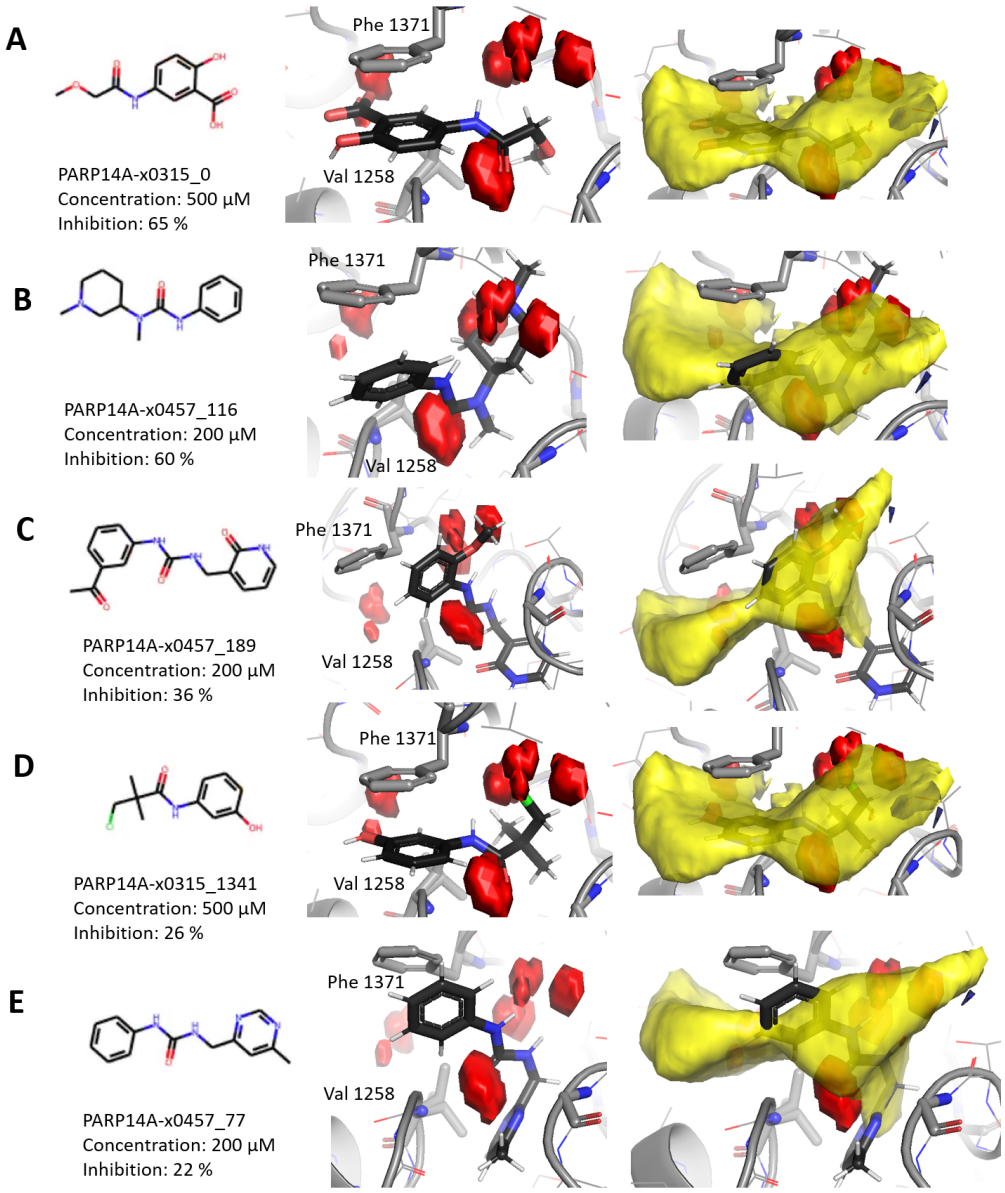


Figure 4.23: Predicted binding modes and hotspot interactions of follow-up compounds with activity against PARP14 MD3

made by the parent fragment x0315. Compound PARP14-x0315_1341 (Figure 4.23) features a chlorine atom that is predicted to bind at the same position in which a chloride ion is observed in some of the fragment crystal structures (towards the top portion of the proximal ribose binding pocket). Compound PARP14-x0457_116 is also predicted to explore this region of the binding pocket, fulfilling the Acceptor 3 interaction with the backbone of Phe 1371.

Overall, the results of the HTRF assays showed that the compounds are weakly binding. However, inhibition was detected at 200 μM compound concentration, which is an improvement on the starting fragments, for which no activity could be detected at that level [195] (the original hits were not tested in the currently used HTRF assay, however). As with the previous two case studies, the follow-up compounds are not much larger than the starting fragments. However, in this case the compounds are predicted to explore more volumes and interactions within the binding site that were not covered by the parent fragments (no interactions satisfying Acceptor cluster 1, or extending towards the ribose and phosphate sites were made by the initial hits) compared to the previous case studies. As in the other cases presented in this chapter, crystallographic experiments would be needed to confirm the binding modes of the predicted follow-up ligands.

4.6 Discussion and Conclusions

For the case studies presented above, the computational workflows using fragment hotspot maps, constrained docking, and dynamic undocking were able to generate suggestions for follow-up compounds with detectable binding affinity. Even though the suggested follow-ups were weakly binding, in this first round of elaboration, the focus was on generating compounds that had overall better fit with the main interactions within the binding site hotspots. For this reason, confirming the

binding mode of the fragments at this stage is perhaps even more critical and a better criterion for success than the measured binding affinity. Unfortunately, due to a combination of the COVID-19 pandemic and the Centre for Medicines Discovery laboratories moving location, crystallographic data could only be collected for the ACVR1 follow-up compounds.

Whilst the case studies presented used similar methodology, modifications to the workflow were made in each of the cases to adapt the pipeline to answer the specific questions relevant to each of the follow-up design campaigns. As compounds for all of these projects were suggested in the spring of 2021, and experimental results became available in the fall/winter of 2021/2022, conclusions from one case study could not be used to inform the procedures used in the others.

In the case of ACVR1, a single fragment hit was available as the starting point, so the question of triaging and prioritising fragment hits was not relevant. The fragment hotspot maps were able to identify features within the main hotspot that introduced additional polar interactions. The mean overall and mean polar hotspot scores were able to prioritise compounds that had improved fit with the polar clusters, while retaining the original fragment's favourable fit to the apolar hotspot maps. Compounds that made interactions with specific features detected in the hotspot maps could also be prioritised using an automated procedure. Dynamic undocking showed that the interaction with Arg 445 was most structurally stable in the initial fragment, and that none of the suggested follow-up poses are expected to induce large changes to the stability of the binding mode along that bond. The additional interactions suggested by the fragment hotspot maps were not found to be structurally stable, but this does not necessarily mean that they will not contribute to the overall stability of the binding mode, as well as to the binding affinity of the follow-up compounds (as structural stability is a property orthogonal to binding affinity [47]). To confirm the latter point, binding affinities for the

ACVR1 follow-up compounds would need to be determined experimentally.

The ACVR1 follow-up compounds suggested by the workflow were soaked into crystals at the XChem facility and diffraction data was collected and processed. Results showed very weak signal from binding events, indicating that the soaking experiment had not been successful, and the results - inconclusive. Repeating the soaking experiment with a longer soak time or higher compound concentration would likely not be feasible due to crystal damage. Hence, to confirm the the binding mode hypotheses output by the workflow, co-crystallisation experiments would be needed. These would ideally be complemented by binding data from other biophysical assays. Despite the inconclusive soaking results, the PanDDA tool was able to detect very weak binding events for three compounds that were structurally similar to the compounds suggested by the workflow. However, none of these retained the binding mode of the initial fragment, and two showed ambiguity in the modelled binding mode. Comparing hotspot maps generated for the biologically relevant monomer structure and for the crystallographic asymmetric units of the hits revealed that these binding modes are likely artifacts of the crystal packing and would not be observed outside a crystal context. Overall, despite the inconclusive soaking results, this data introduced a potential application of the fragment hotspot maps in detecting crystallographic artifacts in fragment screens.

The case of NSP13 presented a very different starting point, in which not only fragments, but also binding sites on the surface of the protein had to be triaged and prioritised prior to the enumeration step, as no binding data was available for any of the starting fragment hits. While estimating the drugability of the sites using fragment hotspot maps was successful at de-prioritising sites with shallow topologies and unfavourable surface properties, the ultimate choice for pursuing the Nucleotide Site was made based on the biological significance of this site. It is not known (as of the moment of writing) whether inhibiting the Stalk site would

have the desired effect on the motions of the protein, even though a large portion of the fragments bound in that site showed good overlap with the hotspot.

The interactions made by the NSP13 fragments and the docked poses of the suggested follow-ups were not found to be structurally stable - they did not pass the stability threshold of 6kcal/mol introduced in the original DUck publication [47]. However, this value was based on the observation that active compounds (binding affinity better than 1 μM) showed enrichment when W_{QB} values greater than 6 kcal/mol were considered. The strongest binder from the NSP13 followups had an IC_{50} around 250 μM, and would not have been considered an active compound by this measure. The large solvent exposure for some of the interacting atoms may also contribute to the lower scores observed.

NSP13 also presented a new challenge in terms of designing for selectivity. In the case of ACVR1, the fragments were bound in an allosteric site, which is generally not conserved between kinases. For NSP13, using hotspot selectivity maps against a human protein from the same helicase family allowed for the identification of hotspot interactions that are unique to the viral target. A workflow that does such comparisons between multiple family members has now also been developed and has been described in [196] and in this thesis.

In the final case study targeting the third macrodomain of PARP14, there was a single binding site of interest, and the fragments within had highly similar structures, binding modes, and interaction profiles. They were also expected to be weakly binding, as no inhibition had been detected using the AlphaScreen assay developed for PARP14 macrodomains. Dynamic undocking showed that the main polar interaction (a hydrogen bond to the backbone N of Val 1258) shared by the fragments was structurally labile, with comparable $W_{QB\_mean\_min}$ values in all of the fragment hits. Interactions with W_{QB} values in this range have been used

previously to rank docked follow-up poses, resulting compounds with improved potency [45], so the interaction was still pursued in subsequent designs. Apart from the interaction with Val 1258 favoured by the initial fragment hits, fragment hotspot mapping identified three main further interactions within the ADPR binding site, and specifically in the vicinity of the bound fragments. These interactions were all exploited by the native ADPR ligand, as shown in Figure 4.21. A workflow that selected poses that interact with at least two of these was able to prioritise compounds that had detectable inhibition of the target at 200 μM concentration, an improvement over the starting fragment hits. When dynamic undocking was performed on the ordered follow-up compounds, the ones that were found to bind also made stable interactions corresponding to at least one of the hotspot clusters. Crystal structures would be needed to confirm the binding mode hypotheses, but this was an encouraging result for using the method for ranking follow-up compounds in fragment growing campaigns.

The mean hotspot scores used to prioritise fragments in this chapter presented a simple way of combining the hotspot information for a compound's crystal structure or docked pose into a single number. However, this approach is limited by the fact that larger compounds will naturally have lower mean hotspot scores, as interactions outside of the main fragment anchors are added to the molecule. When comparing different poses of the same molecule, however, the mean scores can be useful to prioritise conformations that have a better overall fit to the hotspot. The mean score of all atoms can also bias solutions towards highly apolar compounds, as shown in the ACVR1 example, but this can be mitigated by considering poses that score highly in both the mean hotspot and mean polar scores, and by prioritising compounds that make specific polar interactions.

Another limitation of the current workflow lies in the follow-up elaboration stage. While the fragment network proved an excellent source of suggestions for struc-

turally related compounds that could be quickly purchased, it does not easily allow for fragment merges (a modified workflow for this is currently being developed in the von Delft group) or scaffold hopping. For compounds that originate from multiple "hops" along the reaction vectors, the accumulation of multiple rotate-able bonds, or groups that negatively impact the compounds' physico-chemical properties can be problematic. In the work presented here, a set of very simple filters for cLogP, number of rotatable bonds, and molecular weight of the follow-ups were employed. The structures and docked poses of the suggested compounds were visually checked by an experienced medicinal chemist prior to submitting the purchase orders. In future, the workflow would greatly benefit from filters for toxicity, known highly reactive chemical substructures, and, if compound sources other than the fragment network are used, scores to measure synthetic accessibility of the proposed follow-ups.

With regards to automation, the workflow's implementation should allow the user to input a set of aligned fragment-bound structures, producing an output of compound suggestions with 3-dimensional poses based on the most promising starting hits. The case studies presented here, however, show that modifications to the workflow are often needed in order to answer the specific questions relevant to each fragment screen. This is why the goal was not to automate the decision-making process itself, but to make the summary and presentation of information as complete and automated as possible, allowing key decisions to be made by the user in a data-driven way. The implementation of this workflow is currently not complete, as the fragment hotspot maps and dynamic undocking sections have not been integrated within a single environment, and the code has not been containerised. These aspects are currently being worked on, and present challenges from a software development perspective, rather than being purely scientific.

In conclusion, a workflow combining fragment hotspot mapping, including en-

semble and selectivity hotspot maps, constrained docking, dynamic undocking, and the fragment network was used to suggest and score follow-up compounds for three on-going projects at the Centre for Medicines Discovery. This resulted in compounds with measurable binding affinity against two of these targets. Due to external circumstances, crystallographic data could not be obtained for all of these compounds prior to submission of this thesis, and further co-crystallisation trials are needed in the case of ACVR1. Despite these difficulties, the initial experimental results showed the workflow's utility in the first round of compound elaboration.

5 | Conclusions and Outlook

The current pandemic situation has highlighted the importance of speed, efficiency, and reproducibility in drug discovery efforts. After two years, during which safe and effective vaccines were developed and employed around the world, the first small molecules have been approved in the fight against the COVID-19 pandemic. This has prompted efforts to develop starting chemical matter against proteins from other viral families, so that tools, lead compounds and knowledge of the disease targets would already be primed and available in the case of subsequent pandemics. In addition, efforts are already under way to develop chemical matter targeting every human protein as part of the ambitious Target 2035 initiative [8]. This means that more than ever before, there is a pressing need for methods to support these efforts by making the process of arriving at chemical matter cheaper, more robust, and more reproducible.

Fragment-based drug discovery has established itself as a powerful technology for developing small molecule drugs and lead compounds. Crystallographic screening, in particular, can detect even very weakly binding fragments. Performing such experiments at scale was historically unfeasible, but automated platforms and facilities such as XChem [32] have made the process more routine, bringing the time from crystal soaking to finding hits down to a week (down from several months [37]). The utility of this platform received great visibility in early 2020, when a fragment screen against the SARS-CoV-2 main protease was performed and the results made public within about a month of the protein's crystal structure being solved [34]. Currently, the bottleneck in crystallographic fragment screening comes at stage of progressing the hits into leads with improved potency and selectivity for the target protein.

5.1 Summary

The aim of this DPhil project was to develop computational workflows to support the rational elaboration of crystallographic fragment screening hits. Such workflows would be able to summarise vast amounts of structural information for a single target or a family of related protein targets and distill them into testable suggestions for elaborated compounds. By providing automated, objective, and high throughput procedures for data analysis, the time needed to leverage this data in a drug discovery setting would be greatly reduced. As drug discovery is an interdisciplinary effort, these analyses should be intuitively interpretable by users, allowing data-driven hypotheses to be made without an extensive knowledge of structural biology.

To achieve these aims, the interplay between fragment hotspot maps, a binding site mapping procedure developed specifically for fragments, and dynamic undocking, a steered molecular dynamics method developed to assess the stability of protein-ligand complexes, was investigated as the basis of a computational workflow for structure-based fragment elaboration.

Initially, two extensions of the fragment hotspot mapping method were developed to adapt it to the type of data output by crystallographic fragment screens. The ensemble hotspot maps were introduced as a way to summarise information from multiple crystal structures of the same protein. Selectivity hotspot maps were then developed to compare between ensemble maps for related proteins, allowing for insights from structures of related protein targets to contribute hypotheses for the design of selective follow-up compounds. The utility of the ensemble and selectivity maps was first shown through three case studies involving proteins from two well-researched families: bromodomains and protein kinases. The ensemble and

selectivity maps could identify and prioritise structural features that had previously been shown to be important in historical compound design campaigns. The ensemble and selectivity maps' ability to to investigate the selectivity of specific binding site interactions across a subset of the bromodomain protein family was also demonstrated using retrospective examples.

This work showed that the extended hotspot maps could highlight binding site features that are of particular interest to follow-up compound design. However, the fragment hotspot maps are entirely protein-based and do not take the interacting ligand into account. Existing computational methods developed to estimate thermodynamic properties such as binding affinity for a protein-ligand complex could be employed, but are generally not parameterised for fragment-sized ligands. In addition, the conservation of a fragment's binding mode upon elaboration is a key aspect of fragment-based drug design, and few methods have been developed that focus on the selection of compounds with robust binding modes. Dynamic undocking was then chosen to complement the hotspot maps, as it both focuses on the structural stability of the complex, and takes the contribution of the ligand into account.

To facilitate the integration of dynamic undocking into computational pipelines and workflows, an open-source version of the method, OpenDUck, was benchmarked on two datasets of high-quality structures of protein-ligand complexes. The change in MD engine was not found to introduce significant differences in the structural stability estimates, although OpenDUck may assign higher values to interactions featuring charged nitrogen atoms. When the full OpenDUck workflow was compared to the original (AMBER-based) DUck, the majority of the discrepancies in structural stability estimates came from the chunk generation step. A diagnostic was then introduced that allowed for outlying trajectories to be easily identified and visualised. This enabled the generation of mechanistic hypotheses

for the calculated structural stability values, which could be used to inform compound design. Introducing diagnostics for the method was also a key aspect of making the results more interpretable for users without expert knowledge of MD and the tools used to analyse and visualise trajectories in the field.

To evaluate the use of DUck in prioritising structurally similar compounds, a retrospective example taken from a campaign against the human CDK2 kinase was investigated. The results of this showed that even for very weak binders, the interactions made by the common core of the molecules remained structurally stable, while additional interactions made with the protein showed greater variability.

Once the ensemble and selectivity hotspot maps, as well as the OpenDUck pipeline had been validated, these methods were employed as part of a workflow for computational follow-up in three ongoing projects. In the case studies, the computational workflow using fragment hotspot mapping, constrained docking, and dynamic undocking, and the fragment network chemistry recommendation engine, was able to generate suggestions for follow-up compounds with detectable binding affinity. Compounds were suggested simultaneously for all of the projects (due to external constraints) and ordered in the spring of 2021. Due to disruptions from the COVID-19 pandemic and the department moving to a different building, experimental results were not available until the fall/winter of 2021/2022.

In the case of ACVR1, a human kinase target, fragment hotspot mapping was used to identify features in the main hotspot that introduced additional polar interactions. Using the maps to re-score the docking poses for suggested fragment followups prioritised poses that showed improved fit with the polar clusters, while retaining the original fragment's good overlap with the apolar hotspot maps. OpenDUck was able to identify the most stable interaction made by the initial fragment and was used to test whether the suggested followups would be expected to intro-

duce changes to the stability of the binding mode. The compounds suggested by the workflow were soaked into crystals at the XChem facility, but the results of this experiment were inconclusive and co-crystallisation experiments are needed to confirm the predicted binding modes of the suggested follow-ups. Weak binding events could be detected for three structurally related compounds suggested by collaborators, however none of these recapitulated the binding mode of the starting fragment. Comparing hotspot maps for the biologically active monomer with those calculated for the dimer found in the crystallographic asymmetric unit revealed that these binding modes are likely the product of hotspot environments that are present in the crystal, but would not be available outside of a lattice context.

NSP13, a viral helicase from the SARS-CoV-2 pathogen, presented a case where the ensemble hotspot maps were used to find and rank drugable sites on the protein's surface, aiding the choice of which site to pursue. The selectivity hotspot maps for NSP13 versus the human target helicase target UFP-1 were used to identify regions and interactions within the conserved nucleotide binding site that are unique to the viral target.

In the third prospective case study, the human protein target PARP14 had a single binding site of interest, containing fragments with similar 2D structures and binding modes. DUck showed that the main interaction shared by all the fragments was likely structurally labile, while the fragment hotspot maps detected further polar interactions within the binding site. A workflow that prioritised compounds making at least 2 of the interactions highlighted by the ensemble maps were able to suggest compounds that were detected in an HTRF binding assay. The predicted binding modes of these compounds were also found to make structurally stable interactions with the protein.

Overall, a workflow for combining the fragment hotspot maps (including the ensemble and selectivity maps), dynamic undocking, constrained docking, and the fragment network was successfully used to suggest and prioritise potential follow-up compounds for three ongoing projects at the Centre for Medicines Discovery.

5.2 Towards an integrated computational workflow for fragment elaboration using fragment hotspot mapping and DUck

The case studies further showed that the current ways of scoring prospective follow-up poses against the hotspot maps can provide useful suggestions for elaboration. However, more sophisticated re-scoring schemes, for example measuring the overlap with the highest scoring parts of the hotspot in a way similar to SuCOS [61], could be explored. Another way in which the workflow could be improved is the enumeration step. In this work, the fragment network which forms part of XChem's Fragalysis platform was used to enumerate chemical space around the fragment hits. While it showed great utility in generating potential follow-up compounds, it currently does not support merges or ring fusions. The workflow could benefit from steps that run substructure or similarity searches against commercial libraries, allowing for more sophisticated fragment combinations or scaffold-hopping. Machine-learning based methods for follow-up generation, such as DeLinker [79] or STRIFE [78], could also be made available to the user. In cases where specific properties are not filtered by the enumeration methods, a series of filters for toxicity, drug-likeness, and known highly-reactive chemical groups should also be easily accessible.

Although the workflow currently employs a molecular dynamics-based method, it

does not include information on the flexibility of the receptor beyond the structural rearrangements related to fragment binding. In addition to compiling information from crystallographic fragment ensembles, the ensemble maps can also be applied to snapshots from unsteered MD trajectories. Running hotspots on such data could give additional insights not observed in the fragment screening structures, such as the opening of transient subpockets within the binding site. These can provide additional stabilising interactions and regions into which fragments can be elaborated [97]. Methods geared towards detecting transient subpockets, could then also be investigated as ways to further support fragment elaboration campaigns.

A key advantage of an automated workflow is the increase in speed that comes from integrating and parallelising the individual components. A fragment hotspot maps calculation for an individual structure takes 5-20 minutes (depending on the size of the receptor) on a CPU core; an OpenDUck run with extensive sampling (20 SMDs/ temperature, amounting to 40 SMD runs total and 20 unsteered MD runs in between) requires 2-4 hours (depending on chunk size) on an NVIDIA Tesla p100 GPU. Querying the fragment network along all vectors for a fragment takes a few minutes, and running constrained docking on a single follow-up compound takes seconds to a few minutes. Re-scoring the docked poses takes seconds per pose. Consequently, the full workflow could theoretically be run in about a day, assuming all stages can be perfectly parallelised. However, in-between steps such as standardising inputs between methods, and summarising and visualising the results, can add many hours to the overall workflow. As these tasks are then done by the user, the time lost increases rapidly with the number of structures, fragments, and follow-up compounds under investigation. This is why standardising inputs, outputs and visualisations across the methods is crucial for leveraging the vast amount of data output both by the initial crystallographic experiment, and by the subsequent computational methods. Diagnostics such as that intro-

duced in Section 3.4 for identifying outlying OpenDUck trajectories are also key to decreasing the time needed to process and interpret the results of the method. The ACVR1 case study presented in Section 4.3 demonstrated a potential use for the hotspot maps as a diagnostic for binding events that are the product of crystallographic artefacts. This could help mitigate one of the limitations of using crystallography as a readout.

When compounds were suggested for the case studies in Chapter 3, the workflow was still under development, and I was running parts of the input preparation and data analysis manually or through scripts that were specific to the case study at hand. Standardised visualisations and diagnostics for the OpenDUck trajectories were also not implemented at the time. Hence, the time taken to generate the suggestions described in the case studies took a few weeks. The timelines for ordering compounds were externally set and meant that OpenDUck could not be run on all interactions prior to the deadline. This would not have been the case if an integrated workflow had been in place at the start of the project.

In terms of its implementation, the workflow currently comprises of a series of Python scripts, with visualisations of the hotspot maps done through PyMOL and the NGL viewer, and trajectory visualisations through VMD. To improve its usability, the workflow would greatly benefit from a graphical user interface that ties together the stages of the workflow from the user perspective. Due to the large amounts of data that would need to be retrieved (for example, MD trajectories), it may be best suited to a web-based application, rather than a standalone program that is locally installed. Structures for the target and off-target proteins should be made viewable in their aligned, protonated state with the atoms used to define the binding site highlighted. The user should be able to easily add or remove atoms from this selection by clicking on the atoms or residues. To compress the amount of information displayed, protein summary visualisations similar to those used by

Polyphony [144] or WONKA [40] could be employed.

It is crucial that the user can easily switch between protein structures, docked ligand structures, and the various kinds of hotspot maps (per-structure, ensemble, per-site, selectivity maps over various off-targets). Hydrogen bonds between docked compound structures and the displayed receptor should be automatically detected and shown, as well as annotated with the DUck W_{QB} value in cases when it has been calculated. Interactions with outlying DUck trajectories, as detected by the diagnostic described in Section 3.4, can be highlighted as requiring further attention. W_{QB} traces can be displayed when clicking the bond. Ideally, when an outlying trajectory is selected in the trace image, a popup with the corresponding trajectory would be displayed.

Structure preparation and OpenDUck chunking would have dedicated wizards, which show visualisations as the steps are being applied. An exportable summary spreadsheet detailing the methods and parameters used by the workflow, would used for record-keeping and reproducibility. Compound suggestions would be exportable as both 1D (SMILES) and 3D (.sdf) representations. The docked complexes should also be able to be easily exported and downloaded.

5.3 Outlook

In many cases, the challenge of demonstrating the usefulness of automated workflows lies in showing that they outperform human medicinal chemists by designing better, more potent and selective ligands. Rather than taking the decision-making process away from the user however, such workflows can also be used to provide useful, complete, and objective summaries of the available data, allowing the users to make better decisions.

The ensemble and selectivity maps were designed as a way to summarise the available structural data for related protein targets, outputting hypotheses for binding site features or potential interactions that could contribute to the potency and selectivity of follow-up compounds. This automates the process of having to memorise the positions of key features and interactions between structures, allowing for more sophisticated hypotheses to be formed on a protein family level. Dynamic undocking can provide an estimate of the stability of the protein-ligand complex. This is a piece of information that is both difficult to measure experimentally, and is not subject to human intuition. The fragment network can then be used to enumerate the adjacent chemical space, providing readily available suggestions that can be quickly purchased. This has shown to be a great starting point for compound suggestion, but also one of the areas in which the workflow could be improved. For example, allowing particular chemical groups to be interactively added and removed, along with an indication of catalogue availability or synthetic accessibility and fit to the binding site, could further refine the process of suggesting compounds that test specific hypotheses. Integrating machine-learning based methods that implicitly take the shape of the receptor into account, such as STRIFE [78] and deLinker [79], could further aid this application.

Overall, the utility of the methods and workflow developed was shown through their use in on-going projects. The workflow currently focuses on producing compounds with a conserved binding mode and improved fit to the hotspot. As such, it can be further extended by methods that assess binding affinity (for example, free energy perturbation), receptor flexibility, and the physicochemical and ADMET properties of the resulting compounds. In order to make an impact, however, a computational tool such as this needs to be easily accessible to a wide variety of users. In this case, the ability to iteratively and interactively generate compound suggestions will be critical to the usability of the method. This is a software engi-

neering and interface design problem as much as a scientific one. If this problem is addressed, however, I believe that the workflow presented in this thesis would be a valuable addition to platforms for automated and semi-automated fragment to lead optimisation.

Bibliography

[1] Rob.L.M. van Montfort and Paul Workman. "Structure-based drug design: aiming for a perfect fit". In: *Essays In Biochemistry* 61.5 (Nov. 2017), 431 LP –437. DOI: 10.1042/EBC20170052. URL: http://essays.biochemistry.org/content/61/5/431.abstract.

[2] Anthony R. Bradley et al. "The SGC beyond structural genomics: re-defining the role of 3D structures by coupling genomic stratification with fragment-based discovery". In: *Essays In Biochemistry* 61.5 (Nov. 2017), 495 LP –503. DOI: 10.1042/EBC20170051. URL: http://essays.biochemistry.org/content/61/5/495.abstract.

[3] J P Hughes et al. "Principles of early drug discovery". eng. In: *British journal of pharmacology* 162.6 (Mar. 2011), pp. 1239–1249. ISSN: 1476-5381. DOI: 10.1111/j.1476-5381.2010.01127.x. URL: https://pubmed.ncbi.nlm.nih.gov/21091654%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3058157/.

[4] Andrea Volkamer et al. "Identification and Visualization of Kinase-Specific Subpockets". In: *Journal of Chemical Information and Modeling* 56.2 (Feb. 2016), pp. 335–346. ISSN: 1549-9596. DOI: 10.1021/acs.jcim.5b00627. URL: https://doi.org/10.1021/acs.jcim.5b00627.

[5] Peter J Brown and Susanne Müller. "Open access chemical probes for epigenetic targets". eng. In: *Future medicinal chemistry* 7.14 (2015), pp. 1901–1917. ISSN: 1756-8927. DOI: 10.4155/fmc.15.127. URL: https://pubmed.ncbi.nlm.nih.gov/26397018%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4673908/.

[6] Robert B A Quinlan and Paul E Brennan. “Chemogenomics for drug discovery: clinical molecules from open access chemical probes”. In: *RSC Chemical Biology* 2.3 (2021), pp. 759–795. DOI: 10.1039/D1CB00016K. URL: http://dx.doi.org/10.1039/D1CB00016K.

[7] Stephen V Frye. “The art of the chemical probe”. In: *Nature Chemical Biology* 6 (Mar. 2010), p. 159. URL: https://doi.org/10.1038/nchembio.296%20http://10.0.4.14/nchembio.296.

[8] Adrian J Carter et al. “Target 2035: probing the human proteome”. In: *Drug Discovery Today* 24.11 (2019), pp. 2111–2115. ISSN: 1359-6446. DOI: https://doi.org/10.1016/j.drudis.2019.06.020. URL: https://www.sciencedirect.com/science/article/pii/S1359644619301382.

[9] Jean-Louis Reymond. “The chemical space project.” eng. In: *Accounts of chemical research* 48.3 (Mar. 2015), pp. 722–730. ISSN: 1520-4898 (Electronic). DOI: 10.1021/ar500432k.

[10] Christopher A Lipinski et al. “Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings”. In: *Advanced Drug Delivery Reviews* 23.1 (1997), pp. 3–25. ISSN: 0169-409X. DOI: https://doi.org/10.1016/S0169-409X(96)00423-1. URL: https://www.sciencedirect.com/science/article/pii/S0169409X96004231.

[11] Regine S. Bohacek, Colin McMartin, and Wayne C. Guida. “The art and practice of structure-based drug design: A molecular modeling perspective”. In: *Medicinal Research Reviews* 16.1 (1996), pp. 3–50. DOI: https://doi.org/10.1002/(SICI)1098-1128(199601)16:1<3::AID-MED1>3.0.CO;2-6. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/%28SICI%291098-1128%

28199601%2916%3A1%3C3%3A%3AAID-MED1%3E3.0.CO%3B2-6. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/%28SICI%291098-1128%28199601%2916%3A1%3C3%3A%3AAID-MED1%3E3.0.CO%3B2-6.

[12] P G Polishchuk, T I Madzhidov, and A Varnek. "Estimation of the size of drug-like chemical space based on GDB-17 data". In: *Journal of Computer-Aided Molecular Design* 27.8 (2013), pp. 675–679. ISSN: 1573-4951. DOI: 10.1007/s10822-013-9672-4. URL: https://doi.org/10.1007/s10822-013-9672-4.

[13] Richard J Hall, Paul N Mortenson, and Christopher W Murray. "Efficient exploration of chemical space by fragment-based screening". In: *Progress in Biophysics and Molecular Biology* 116.2 (2014), pp. 82–91. ISSN: 0079-6107. DOI: https://doi.org/10.1016/j.pbiomolbio.2014.09.007. URL: https://www.sciencedirect.com/science/article/pii/S0079610714000960.

[14] Michael M. Hann. "Molecular obesity, potency and other addictions in drug discovery". In: *MedChemComm* 2.5 (2011), pp. 349–355. ISSN: 20402503. DOI: 10.1039/c1md00017a.

[15] Courtney Aldrich et al. "The Ecstasy and Agony of Assay Interference Compounds". In: *ACS Central Science* 3.3 (2017), pp. 143–147. ISSN: 1549-9596. DOI: 10.1021/acs.jmedchem.7b00229.

[16] Miklos Feher and Jonathan M Schmidt. "Property Distributions:Differences between Drugs, Natural Products, and Molecules from Combinatorial Chemistry". In: *Journal of Chemical Information and Computer Sciences* 43.1 (Jan. 2003), pp. 218–227. ISSN: 0095-2338. DOI: 10.1021/ci0200467. URL: https://doi.org/10.1021/ci0200467.

[17] Oleksandr O Grygorenko et al. “Generating Multibillion Chemical Space of Readily Accessible Screening Compounds”. In: *iScience* 23.11 (Nov. 2020). ISSN: 2589-0042. DOI: 10.1016/j.isci.2020.101681. URL: https://doi.org/10.1016/j.isci.2020.101681.

[18] Célien Jacquemard and Esther Kellenberger. “A bright future for fragment-based drug discovery: what does it hold?” In: *Expert Opinion on Drug Discovery* 14.5 (May 2019), pp. 413–416. ISSN: 1746-0441. DOI: 10.1080/17460441.2019.1583643. URL: https://doi.org/10.1080/17460441.2019.1583643.

[19] Miles Congreve et al. “A ‘Rule of Three’ for fragment-based lead discovery?” In: *Drug Discovery Today* 8.19 (2003), pp. 876–877. ISSN: 1359-6446. DOI: https://doi.org/10.1016/S1359-6446(03)02831-9. URL: https://www.sciencedirect.com/science/article/pii/S1359644603028319.

[20] Harren Jhoti et al. “The ’rule of three’ for fragment-based drug discovery: where are we now?” In: *Nature Reviews Drug Discovery* 12.8 (2013), p. 644. ISSN: 1474-1784. DOI: 10.1038/nrd3926-c1. URL: https://doi.org/10.1038/nrd3926-c1.

[21] Michael M Hann, Andrew R Leach, and Gavin Harper. “Molecular Complexity and Its Impact on the Probability of Finding Leads for Drug Discovery”. In: *Journal of Chemical Information and Computer Sciences* 41.3 (May 2001), pp. 856–864. ISSN: 0095-2338. DOI: 10.1021/ci000403i. URL: https://doi.org/10.1021/ci000403i.

[22] Disha Patel, Joseph D Bauman, and Eddy Arnold. “Advantages of crystallographic fragment screening: Functional and mechanistic insights from a powerful platform for efficient drug discovery”. In: *Progress in Biophysics and Molecular Biology* 116.2-3 (2014), pp. 92–100. ISSN: 00796107. DOI:

10.1016/j.pbiomolbio.2014.08.004. arXiv: arXiv:1011.1669v3. URL: http://dx.doi.org/10.1016/j.pbiomolbio.2014.08.004.

[23] Gideon Bollag et al. "Vemurafenib: the first drug approved for BRAF-mutant cancer". In: *Nature Reviews Drug Discovery* 11.11 (2012), pp. 873–886. ISSN: 1474-1784. DOI: 10.1038/nrd3847.

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

[25] Shanada Monestime and Dovena Lazaridis. "Pexidartinib (TURALIO™): The First FDA-Indicated Systemic Treatment for Tenosynovial Giant Cell Tumor". eng. In: *Drugs in R&D* 20.3 (Sept. 2020), pp. 189–195. ISSN: 1179-6901. DOI: 10.1007/s40268-020-00314-3. URL: https://pubmed.ncbi.nlm.nih.gov/32617868%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7419392/.

[26] Anthony Markham. "Erdafitinib: First Global Approval". In: *Drugs* 79.9 (2019), pp. 1017–1021. ISSN: 1179-1950. DOI: 10.1007/s40265-019-01142-9. URL: https://doi.org/10.1007/s40265-019-01142-9.

[27] Bradley C Doak, Raymond S Norton, and Martin J Scanlon. "The ways and means of fragment-based drug design". In: *Pharmacology and Therapeutics* 167 (2016), pp. 28–37. ISSN: 1879016X. DOI: 10.1016/j.pharmthera.2016.07.003. URL: http://dx.doi.org/10.1016/j.pharmthera.2016.07.003.

[28] Daniel J Wood et al. "FragLites—Minimal, Halogenated Fragments Displaying Pharmacophore Doublets. An Efficient Approach to Druggability Assessment and Hit Generation". In: *Journal of Medicinal Chemistry* 62.7

(Apr. 2019), pp. 3741–3752. ISSN: 0022-2623. DOI: 10.1021/acs.jmedchem.9b00304. URL: https://doi.org/10.1021/acs.jmedchem.9b00304.

[29] Marc O'Reilly et al. "Crystallographic screening using ultra-low-molecular-weight ligands to guide drug design". In: *Drug Discovery Today* 24.5 (2019), pp. 1081–1086. ISSN: 1359-6446. DOI: https://doi.org/10.1016/j.drudis.2019.03.009. URL: http://www.sciencedirect.com/science/article/pii/S1359644619300017.

[30] Nikolaj S Troelsen and Mads H Clausen. "Library Design Strategies To Accelerate Fragment-Based Drug Discovery". In: *Chemistry – A European Journal* 26.50 (Sept. 2020), pp. 11391–11403. ISSN: 0947-6539. DOI: https://doi.org/10.1002/chem.202000584. URL: https://doi.org/10.1002/chem.202000584.

[31] Oakley B Cox et al. "A poised fragment library enables rapid synthetic expansion yielding the first reported inhibitors of PHIP(2), an atypical bromodomain". In: *Chemical Science* 7.3 (2016), pp. 2322–2330. ISSN: 2041-6520. DOI: 10.1039/C5SC03115J. URL: http://dx.doi.org/10.1039/C5SC03115J.

[32] Alice Douangamath et al. "Achieving Efficient Fragment Screening at XChem Facility at Diamond Light Source". In: *JoVE* 171 (2021), e62414. ISSN: 1940-087X. DOI: doi:10.3791/62414. URL: https://www.jove.com/t/62414.

[33] Efrat Resnick et al. "Rapid Covalent-Probe Discovery by Electrophile-Fragment Screening". In: *Journal of the American Chemical Society* 141.22 (June 2019), pp. 8951–8968. ISSN: 0002-7863. DOI: 10.1021/jacs.9b02822. URL: https://doi.org/10.1021/jacs.9b02822.

[34] Alice Douangamath et al. “Crystallographic and electrophilic fragment screening of the SARS-CoV-2 main protease”. In: *Nature Communications* 11.1 (2020), p. 5047. ISSN: 2041-1723. DOI: 10.1038/s41467-020-18709-w. URL: https://doi.org/10.1038/s41467-020-18709-w.

[35] Keriann M Backus et al. “Proteome-wide covalent ligand discovery in native biological systems”. eng. In: *Nature* 534.7608 (June 2016), pp. 570–574. ISSN: 1476-4687. DOI: 10.1038/nature18002. URL: https://pubmed.ncbi.nlm.nih.gov/27309814%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4919207/.

[36] Christopher G Parker et al. “Ligand and Target Discovery by Fragment-Based Screening in Human Cells”. In: *Cell* 168.3 (Jan. 2017), 527–541.e29. ISSN: 0092-8674. DOI: 10.1016/j.cell.2016.12.029. URL: https://doi.org/10.1016/j.cell.2016.12.029.

[37] John C Spurlino. “Fragment screening purely with protein crystallography”. In: *Methods in Enzymology* 493 (2011), pp. 321–356. ISSN: 00766879. DOI: 10.1016/B978-0-12-381274-2.00013-3. URL: http://dx.doi.org/10.1016/B978-0-12-381274-2.00013-3.

[38] Nicholas M Pearce et al. “A multi-crystal method for extracting obscured crystallographic states from conventionally uninterpretable electron density”. In: *Nature Communications* 8.1 (2017), p. 15123. ISSN: 2041-1723. DOI: 10.1038/ncomms15123. URL: https://doi.org/10.1038/ncomms15123.

[39] “WONKA and OOMMPPAA: Analysis of protein-ligand interaction data to direct structure-based drug design”. In: *Acta Crystallographica Section D: Structural Biology* 73 (2017), pp. 279–285. ISSN: 20597983. DOI: 10.1107/S2059798316009529.

[40] A. R. Bradley et al. “WONKA: objective novel complex analysis for ensembles of protein-ligand structures”. In: *Journal of Computer-Aided Molecular Design* 29.10 (2015), pp. 963–973. ISSN: 15734951. DOI: 10.1007/s10822-015-9866-z.

[41] Anthony R. Bradley et al. “OOMMPPAA: A Tool To Aid Directed Synthesis by the Combined Analysis of Activity and Structural Data”. In: *Journal of Chemical Information and Modeling* 54.10 (2014), pp. 2636–2646. ISSN: 1549-9596. DOI: 10.1021/ci500245d. URL: http://pubs.acs.org/doi/abs/10.1021/ci500245d.

[42] Richard J Hall, Christopher W Murray, and Marcel L Verdonk. “The Fragment Network: A Chemistry Recommendation Engine Built Using a Graph Database”. In: *Journal of Medicinal Chemistry* 60.14 (July 2017), pp. 6440–6450. ISSN: 0022-2623. DOI: 10.1021/acs.jmedchem.7b00809. URL: https://doi.org/10.1021/acs.jmedchem.7b00809.

[43] Alexander S. Rose and Peter W. Hildebrand. “NGL Viewer: A web application for molecular visualization”. In: *Nucleic Acids Research* 43.W1 (2015), W576–W579. ISSN: 13624962. DOI: 10.1093/nar/gkv402.

[44] Joseph A Newman et al. “Structure, mechanism and crystallographic fragment screening of the SARS-CoV-2 NSP13 helicase”. In: *Nature Communications* 12.1 (2021), p. 4848. ISSN: 2041-1723. DOI: 10.1038/s41467-021-25166-6. URL: https://doi.org/10.1038/s41467-021-25166-6.

[45] Moira Rachman. “Fragment-to-Lead Optimisation with Automated and Iterative Virtual Screening”. PhD thesis. Universitat de Barcelona, 2020. DOI: http://hdl.handle.net/10803/671756.

[46] Andrew L Hopkins, Colin R Groom, and Alexander Alex. “Ligand efficiency: a useful metric for lead selection”. In: *Drug Discovery Today* 9.10

(2004), pp. 430–431. ISSN: 1359-6446. DOI: https://doi.org/10.1016/S1359-6446(04)03069-7. URL: https://www.sciencedirect.com/science/article/pii/S1359644604030697.

[47] Sergio Ruiz-Carmona et al. "Dynamic undocking and the quasi-bound state as tools for drug discovery". In: *Nature Chemistry* 9 (Nov. 2016), p. 201.

[48] Chris J. Radoux et al. "Identifying Interactions that Determine Fragment Binding at Protein Hotspots". In: *Journal of Medicinal Chemistry* 59.9 (2016). PMID: 27043011, pp. 4314–4325. DOI: 10.1021/acs.jmedchem.5b01980. eprint: https://doi.org/10.1021/acs.jmedchem.5b01980. URL: https://doi.org/10.1021/acs.jmedchem.5b01980.

[49] Zenon Konteatis. "What makes a good fragment in fragment-based drug discovery?" In: *Expert Opinion on Drug Discovery* 16.7 (July 2021), pp. 723–726. ISSN: 1746-0441. DOI: 10.1080/17460441.2021.1905629. URL: https://doi.org/10.1080/17460441.2021.1905629.

[50] Markus Hartenfeller and Gisbert Schneider. "Enabling future drug discovery by de novo design". In: *WIREs Computational Molecular Science* 1.5 (Sept. 2011), pp. 742–759. ISSN: 1759-0876. DOI: https://doi.org/10.1002/wcms.49. URL: https://doi.org/10.1002/wcms.49.

[51] Markus Boehm et al. "Similarity Searching and Scaffold Hopping in Synthetically Accessible Combinatorial Chemistry Spaces". In: *Journal of Medicinal Chemistry* 51.8 (Apr. 2008), pp. 2468–2480. ISSN: 0022-2623. DOI: 10.1021/jm0707727. URL: https://doi.org/10.1021/jm0707727.

[52] Louis Bellmann, Patrick Penner, and Matthias Rarey. "Topological Similarity Search in Large Combinatorial Fragment Spaces". In: *Journal of Chemical Information and Modeling* 61.1 (Jan. 2021), pp. 238–251. ISSN: 1549-9596. DOI: 10.1021/acs.jcim.0c00850. URL: https://doi.org/10.1021/acs.jcim.0c00850.

[53] Xiwen Jia et al. "Anthropogenic biases in chemical reaction data hinder exploratory inorganic synthesis". In: *Nature* 573.7773 (2019), pp. 251–255. ISSN: 1476-4687. DOI: 10.1038/s41586-019-1540-5. URL: https://doi.org/10.1038/s41586-019-1540-5.

[54] Dávid Péter Kovács, William McCorkindale, and Alpha A Lee. "Quantitative interpretation explains machine learning models for chemical reaction prediction and uncovers bias". In: *Nature Communications* 12.1 (2021), p. 1695. ISSN: 2041-1723. DOI: 10.1038/s41467-021-21895-w. URL: https://doi.org/10.1038/s41467-021-21895-w.

[55] Teague Sterling and John J Irwin. "ZINC 15 – Ligand Discovery for Everyone". In: *Journal of Chemical Information and Modeling* 55.11 (Nov. 2015), pp. 2324–2337. ISSN: 1549-9596. DOI: 10.1021/acs.jcim.5b00559. URL: https://doi.org/10.1021/acs.jcim.5b00559.

[56] *Chemical Spaces*. URL: https://www.biosolveit.de/infiniSee#chemical_spaces (visited on 05/18/2022).

[57] Marwin H S Segler et al. "Generating Focused Molecule Libraries for Drug Discovery with Recurrent Neural Networks". In: *ACS Central Science* 4.1 (Jan. 2018), pp. 120–131. ISSN: 2374-7943. DOI: 10.1021/acscentsci.7b00512. URL: https://doi.org/10.1021/acscentsci.7b00512.

[58] Malgorzata N Drwal et al. “Structural Insights on Fragment Binding Mode Conservation”. In: *Journal of Medicinal Chemistry* 61.14 (July 2018), pp. 5963–5973. ISSN: 0022-2623. DOI: 10.1021/acs.jmedchem.8b00256. URL: https://doi.org/10.1021/acs.jmedchem.8b00256.

[59] Shipra Malhotra and John Karanicolas. “When Does Chemical Elaboration Induce a Ligand To Change Its Binding Mode?” In: *Journal of Medicinal Chemistry* 60.1 (Jan. 2017), pp. 128–145. ISSN: 15204804. DOI: 10.1021/acs.jmedchem.6b00725. URL: https://doi.org/10.1021/acs.jmedchem.6b00725.

[60] Moira Rachman et al. “Fragment-to-lead tailored in silico design”. In: *Drug Discovery Today: Technologies* (2021). ISSN: 1740-6749. DOI: https://doi.org/10.1016/j.ddtec.2021.08.005. URL: https://www.sciencedirect.com/science/article/pii/S1740674921000226.

[61] Susan Leung et al. “SuCOS is Better than RMSD for Evaluating Fragment Elaboration and Docking Poses”. In: *ChemRxiv* (2019). DOI: 10.26434/chemrxiv.8100203.v1.

[62] J. A. GRANT, M. A. GALLARDO, and B. T. PICKUP. “A fast method of molecular shape comparison: A simple application of a Gaussian description of molecular shape”. In: *Journal of Computational Chemistry* 17.14 (1996), pp. 1653–1666. DOI: https://doi.org/10.1002/(SICI)1096-987X(19961115)17:14<1653::AID-JCC7>3.0.CO;2-K. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/%28SICI%291096-987X%2819961115%2917%3A14%3C1653%3A%3AAID-JCC7%3E3.0.CO%3B2-K. URL: https://onlinelibrary.wiley.com/doi/abs/10.

1002/%28SICI%291096-987X%2819961115%2917%3A14%3C1653%3A%3AAID-JCC7%3E3.0.CO%3B2-K.

[63] Xavier Barril and S David Morley. “Unveiling the Full Potential of Flexible Receptor Docking Using Multiple Crystallographic Structures”. In: *Journal of Medicinal Chemistry* 48.13 (June 2005), pp. 4432–4443. ISSN: 0022-2623. DOI: 10.1021/jm048972v. URL: https://doi.org/10.1021/jm048972v.

[64] Oliver Korb et al. “Potential and Limitations of Ensemble Docking”. In: *Journal of Chemical Information and Modeling* 52.5 (May 2012), pp. 1262–1274. ISSN: 1549-9596. DOI: 10.1021/ci2005934.

[65] Peter Eastman et al. “OpenMM 7: Rapid development of high performance algorithms for molecular dynamics”. In: *PLOS Computational Biology* 13.7 (July 2017), e1005659. URL: https://doi.org/10.1371/journal.pcbi.1005659.

[66] Lauro Ribeiro de Souza Neto et al. “In silico Strategies to Support Fragment-to-Lead Optimization in Drug Discovery”. eng. In: *Frontiers in chemistry* 8 (Feb. 2020), p. 93. ISSN: 2296-2646. DOI: 10.3389/fchem.2020.00093. URL: https://pubmed.ncbi.nlm.nih.gov/32133344%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7040036/.

[67] Sherine E Thomas et al. “Fragment-based discovery of a new class of inhibitors targeting mycobacterial tRNA modification”. In: *Nucleic Acids Research* 48.14 (Aug. 2020), pp. 8099–8112. ISSN: 0305-1048. DOI: 10.1093/nar/gkaa539. URL: https://doi.org/10.1093/nar/gkaa539.

[68] Shikang Liang et al. “Structural insights into inhibitor regulation of the DNA repair protein DNA-PKcs”. In: *Nature* 601.7894 (2022), pp. 643–

648. ISSN: 1476-4687. DOI: 10.1038/s41586-021-04274-9. URL: https://doi.org/10.1038/s41586-021-04274-9.

[69] Daniel Alvarez-Garcia and Xavier Barril. "Molecular Simulations with Solvent Competition Quantify Water Displaceability and Provide Accurate Interaction Maps of Protein Binding Sites". In: *Journal of Medicinal Chemistry* 57.20 (2014). PMID: 25275946, pp. 8530–8539. DOI: 10.1021/jm5010418. eprint: https://doi.org/10.1021/jm5010418. URL: https://doi.org/10.1021/jm5010418.

[70] Maicol Bissaro, Mattia Sturlese, and Stefano Moro. "The rise of molecular simulations in fragment-based drug design (FBDD): an overview". In: *Drug Discovery Today* 25.9 (2020), pp. 1693–1701. ISSN: 1359-6446. DOI: https://doi.org/10.1016/j.drudis.2020.06.023. URL: https://www.sciencedirect.com/science/article/pii/S1359644620302518.

[71] Alexandre Bancet et al. "Fragment Linking Strategies for Structure-Based Drug Design". In: *Journal of Medicinal Chemistry* 63.20 (Oct. 2020), pp. 11420–11435. ISSN: 0022-2623. DOI: 10.1021/acs.jmedchem.0c00242. URL: https://doi.org/10.1021/acs.jmedchem.0c00242.

[72] Hans-Joachim Böhm. "LUDI: rule-based automatic design of new substituents for enzyme inhibitor leads". In: *Journal of Computer-Aided Molecular Design* 6.6 (1992), pp. 593–606. ISSN: 1573-4951. DOI: 10.1007/BF00126217. URL: https://doi.org/10.1007/BF00126217.

[73] Valerie J Gillet et al. "SPROUT, HIPPO and CAESA: Tools for de novo structure generation and estimation of synthetic accessibility". In: *Perspectives in Drug Discovery and Design* 3.1 (1995), pp. 34–50. ISSN:

1573-9023. DOI: 10.1007/BF02174466. URL: https://doi.org/10.1007/BF02174466.

[74] David A Pearlman and Mark A Murcko. "CONCERTS:Dynamic Connection of Fragments as an Approach to de Novo Ligand Design". In: *Journal of Medicinal Chemistry* 39.8 (Jan. 1996), pp. 1651–1663. ISSN: 0022-2623. DOI: 10.1021/jm9507921. URL: https://doi.org/10.1021/jm9507921.

[75] Albert C Pierce, Govinda Rao, and Guy W Bemis. "BREED: Generating Novel Inhibitors through Hybridization of Known Ligands. Application to CDK2, P38, and HIV Protease". In: *Journal of Medicinal Chemistry* 47.11 (May 2004), pp. 2768–2775. ISSN: 0022-2623. DOI: 10.1021/jm030543u. URL: https://doi.org/10.1021/jm030543u.

[76] Fabian Dey and Amedeo Caflisch. "Fragment-Based de Novo Ligand Design by Multiobjective Evolutionary Optimization". In: *Journal of Chemical Information and Modeling* 48.3 (Mar. 2008), pp. 679–690. ISSN: 1549-9596. DOI: 10.1021/ci700424b. URL: https://doi.org/10.1021/ci700424b.

[77] Fergus Imrie et al. "Deep Generative Design with 3D Pharmacophoric Constraints". In: *bioRxiv* (Jan. 2021), p. 2021.04.27.441676. DOI: 10.1101/2021.04.27.441676. URL: http://biorxiv.org/content/early/2021/04/28/2021.04.27.441676.abstract.

[78] Thomas E Hadfield et al. "Incorporating Target-Specific Pharmacophoric Information Into Deep Generative Models For Fragment Elaboration". In: *bioRxiv* (Jan. 2021), p. 2021.10.21.465268. DOI: 10.1101/2021.10.21.465268. URL: http://biorxiv.org/content/early/2021/10/22/2021.10.21.465268.abstract.

[79] Fergus Imrie et al. “Deep Generative Models for 3D Linker Design”. In: *Journal of Chemical Information and Modeling* 60.4 (Apr. 2020), pp. 1983–1995. ISSN: 1549-9596. DOI: 10.1021/acs.jcim.9b01120. URL: https://doi.org/10.1021/acs.jcim.9b01120.

[80] Serena G Piticchio et al. “Discovery of Novel BRD4 Ligand Scaffolds by Automated Navigation of the Fragment Chemical Space”. In: *Journal of Medicinal Chemistry* 64.24 (Dec. 2021), pp. 17887–17900. ISSN: 0022-2623. DOI: 10.1021/acs.jmedchem.1c01108. URL: https://doi.org/10.1021/acs.jmedchem.1c01108.

[81] Dagmar Ringe. “What makes a binding site a binding site?” In: *Current Opinion in Structural Biology* 5.6 (1995), pp. 825–829. ISSN: 0959440X. DOI: 10.1016/0959-440X(95)80017-4.

[82] Karen N Allen et al. “An Experimental Approach to Mapping the Binding Surfaces of Crystalline Proteins”. In: *The Journal of Physical Chemistry* 100.7 (Jan. 1996), pp. 2605–2611. ISSN: 0022-3654. DOI: 10.1021/jp952516o. URL: https://doi.org/10.1021/jp952516o.

[83] P A Fitzpatrick, D Ringe, and A M Klibanov. “X-Ray Crystal Structure of Cross-Linked Subtilisin Carlsberg in Water vs Acetonitrile”. In: *Biochemical and Biophysical Research Communications* 198.2 (1994), pp. 675–681. ISSN: 0006-291X. DOI: https://doi.org/10.1006/bbrc.1994.1098. URL: https://www.sciencedirect.com/science/article/pii/S0006291X84710989.

[84] Neela H Yennawar, Hemant P Yennawar, and Gregory K Farber. “X-ray Crystal Structure of -Chymotrypsin in Hexane”. In: *Biochemistry* 33.23 (1994), pp. 7326–7336.

[85] Amedeo Caflisch, Andrew Miranker, and Martin Karplus. “Multiple copy simultaneous search and construction of ligands in binding sites: appli-

cation to inhibitors of HIV-1 aspartic proteinase". In: *Journal of Medicinal Chemistry* 36.15 (1993). PMID: 8340918, pp. 2142–2167. DOI: 10.1021/jm00067a013. eprint: https://doi.org/10.1021/jm00067a013. URL: https://doi.org/10.1021/jm00067a013.

[86] Andrew Miranker and Martin Karplus. "Functionality maps of binding sites: A multiple copy simultaneous search method". In: *Proteins: Structure, Function, and Bioinformatics* 11.1 (Sept. 1991), pp. 29–34. ISSN: 0887-3585. DOI: https://doi.org/10.1002/prot.340110104. URL: https://doi.org/10.1002/prot.340110104.

[87] Andrew C English, Colin R Groom, and Roderick E Hubbard. "Experimental and computational mapping of the binding surface of a crystalline protein". In: *Protein Engineering, Design and Selection* 14.1 (Jan. 2001), pp. 47–59. ISSN: 1741-0126. DOI: 10.1093/protein/14.1.47. URL: https://doi.org/10.1093/protein/14.1.47.

[88] P. J. Goodford. "A Computational Procedure for Determining Energetically Favorable Binding Sites on Biologically Important Macromolecules". In: *Journal of Medicinal Chemistry* 28.7 (1985), pp. 849–857. ISSN: 15204804. DOI: 10.1021/jm00145a002.

[89] Philip J. Hajduk, Jeffrey R. Huth, and Stephen W. Fesik. "Druggability Indices for Protein Targets Derived from NMR-Based Screening Data". In: *Journal of Medicinal Chemistry* 48.7 (2005). PMID: 15801841, pp. 2518–2525. DOI: 10.1021/jm049131r. eprint: https://doi.org/10.1021/jm049131r. URL: https://doi.org/10.1021/jm049131r.

[90] Tom Young et al. "Motifs for molecular recognition exploiting hydrophobic enclosure in protein–ligand binding". In: *Proceedings of the National Academy of Sciences* 104.3 (Jan. 2007), 808 LP –813. DOI: 10.1073/

pnas.0610202104. URL: http://www.pnas.org/content/104/3/808.abstract.

[91] Laveena Muley et al. “Enhancement of Hydrophobic Interactions and Hydrogen Bond Strength by Cooperativity: Synthesis, Modeling, and Molecular Dynamics Simulations of a Congeneric Series of Thrombin Inhibitors”. In: *Journal of Medicinal Chemistry* 53.5 (Mar. 2010), pp. 2126–2135. ISSN: 0022-2623. DOI: 10.1021/jm9016416. URL: https://doi.org/10.1021/jm9016416.

[92] John E Ladbury, Gerhard Klebe, and Ernesto Freire. “Adding calorimetric data to decision making in lead discovery: a hot tip”. In: *Nature Reviews Drug Discovery* 9.1 (2010), pp. 23–27. ISSN: 1474-1784. DOI: 10.1038/nrd3054. URL: https://doi.org/10.1038/nrd3054.

[93] György G Ferenczy and György M Keserű. “Thermodynamics of Fragment Binding”. In: *Journal of Chemical Information and Modeling* 52.4 (Apr. 2012), pp. 1039–1045. ISSN: 1549-9596. DOI: 10.1021/ci200608b. URL: https://doi.org/10.1021/ci200608b.

[94] Peter Schmidtke et al. “Shielded Hydrogen Bonds as Structural Determinants of Binding Kinetics: Application in Drug Design”. In: *Journal of the American Chemical Society* 133.46 (Nov. 2011), pp. 18903–18910. ISSN: 0002-7863. DOI: 10.1021/ja207494u. URL: https://doi.org/10.1021/ja207494u.

[95] Tjelvar S G Olsson et al. “The Thermodynamics of Protein–Ligand Interaction and Solvation: Insights for Ligand Design”. In: *Journal of Molecular Biology* 384.4 (2008), pp. 1002–1017. ISSN: 0022-2836. DOI: https://doi.org/10.1016/j.jmb.2008.09.073. URL: https://www.sciencedirect.com/science/article/pii/S0022283608012473.

[96] György G Ferenczy and György M Keserű. “Enthalpic Efficiency of Ligand Binding”. In: *Journal of Chemical Information and Modeling* 50.9 (Sept. 2010), pp. 1536–1541. ISSN: 1549-9596. DOI: 10.1021/ci100125a. URL: https://doi.org/10.1021/ci100125a.

[97] Chris. J. Radoux. “The Automatic Detection of Small Molecule Binding Hotspots on Proteins: Applying Hotspots to Structure-Based Drug Design”. PhD thesis. University of Cambridge, 2018, pp. 108–128. URL: https://www.repository.cam.ac.uk/handle/1810/275133.

[98] “SuperStar: A knowledge-based approach for identifying interaction sites in proteins”. In: *Journal of Molecular Biology* 289.4 (1999), pp. 1093–1108. ISSN: 00222836. DOI: 10.1006/jmbi.1999.2809.

[99] Prakash Chandra Rathi et al. “Predicting "Hot" and "Warm" Spots for Fragment Binding”. In: *Journal of Medicinal Chemistry* 60.9 (2017), pp. 4036–4046. ISSN: 15204804. DOI: 10.1021/acs.jmedchem.7b00366.

[100] Osamu Ichihara, Yuzo Shimada, and Daisuke Yoshidome. “The importance of hydration thermodynamics in fragment-to-lead optimization.” eng. In: *ChemMedChem* 9.12 (Dec. 2014), pp. 2708–2717. ISSN: 1860-7187 (Electronic). DOI: 10.1002/cmdc.201402207.

[101] Akshay Sridhar, Gregory A Ross, and Philip C Biggin. “Waterdock 2.0: Water placement prediction for Holo-structures with a pymol plugin”. In: *PLOS ONE* 12.2 (Feb. 2017), e0172743. URL: https://doi.org/10.1371/journal.pone.0172743.

[102] Alessio Amadasi et al. “Mapping the Energetics of Water–Protein and Water–Ligand Interactions with the “Natural” HINT Forcefield: Predictive Tools for Characterizing the Roles of Water in Biomolecules”. In: *Journal of Molecular Biology* 358.1 (2006), pp. 289–309. ISSN: 0022-2836. DOI:

https://doi.org/10.1016/j.jmb.2006.01.053. URL: https://www.sciencedirect.com/science/article/pii/S0022283606000830.

[103] Alessio Amadasi et al. "Robust Classification of "Relevant" Water Molecules in Putative Protein Binding Sites". In: *Journal of Medicinal Chemistry* 51.4 (Feb. 2008), pp. 1063–1067. ISSN: 0022-2623. DOI: 10.1021/jm701023h. URL: https://doi.org/10.1021/jm701023h.

[104] Caterina Barillari et al. "Classification of Water Molecules in Protein Binding Sites". In: *Journal of the American Chemical Society* 129.9 (Mar. 2007), pp. 2577–2587. ISSN: 0002-7863. DOI: 10.1021/ja066980q. URL: https://doi.org/10.1021/ja066980q.

[105] Alfonso T García-Sosa, Ricardo L Mancera, and Philip M Dean. "WaterScore: a novel method for distinguishing between bound and displaceable water molecules in the crystal structure of the binding site of protein-ligand complexes". In: *Journal of Molecular Modeling* 9.3 (2003), pp. 172–182. ISSN: 0948-5023. DOI: 10.1007/s00894-003-0129-x. URL: https://doi.org/10.1007/s00894-003-0129-x.

[106] Oleg Trott and Arthur J Olson. "AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading". eng. In: *Journal of computational chemistry* 31.2 (Jan. 2010), pp. 455–461. ISSN: 1096-987X. DOI: 10.1002/jcc.21334. URL: https://pubmed.ncbi.nlm.nih.gov/19499576%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3041641/.

[107] Christina E Faller et al. "Site Identification by Ligand Competitive Saturation (SILCS) simulations for fragment-based drug design". eng. In: *Methods in molecular biology (Clifton, N.J.)* 1289 (2015), pp. 75–87.

ISSN: 1940-6029. DOI: 10.1007/978-1-4939-2486-8_7. URL: https://www.ncbi.nlm.nih.gov/pubmed/25709034%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4685950/.

[108] Ryan Brenke et al. "Fragment-based identification of druggable 'hot spots' of proteins using Fourier domain correlation techniques". In: *Bioinformatics* 25.5 (Jan. 2009), pp. 621–627. DOI: 10.1093/bioinformatics/btp036. eprint: https://academic.oup.com/bioinformatics/article-pdf/25/5/621/16891987/btp036.pdf. URL: https://doi.org/10.1093/bioinformatics/btp036.

[109] Sarah E. Graham, Noah Leja, and Heather A. Carlson. "MixMD Probeview: Robust Binding Site Prediction from Cosolvent Simulations". In: *Journal of Chemical Information and Modeling* 58.7 (2018). PMID: 29905479, pp. 1426–1433. DOI: 10.1021/acs.jcim.8b00265. eprint: https://doi.org/10.1021/acs.jcim.8b00265. URL: https://doi.org/10.1021/acs.jcim.8b00265.

[110] Colin R. Groom et al. "The Cambridge structural database". In: *Acta Crystallographica Section B: Structural Science, Crystal Engineering and Materials* 72.2 (2016), pp. 171–179. ISSN: 20525206. DOI: 10.1107/S2052520616003954.

[111] H M Berman et al. "The protein data bank." In: *Nucleic acids research* 28.1 (2000), pp. 235–242. ISSN: 0305-1048. DOI: 10.1093/nar/28.1.235.

[112] Ian J Bruno et al. "IsoStar: A library of information about nonbonded interactions". In: *Journal of Computer-Aided Molecular Design* 11.6 (1997), pp. 525–537. ISSN: 0920654X. DOI: 10.1023/A:1007934413448.

[113] Peter R Curran et al. “Hotspots API: A Python Package for the Detection of Small Molecule Binding Hotspots and Application to Structure-Based Drug Design”. In: *Journal of Chemical Information and Modeling* (Mar. 2020). ISSN: 1549-9596. DOI: 10.1021/acs.jcim.9b00996. URL: https://doi.org/10.1021/acs.jcim.9b00996.

[114] Manfred Hendlich, Friedrich Rippmann, and Gerhard Barnickel. “LIGSITE: Automatic and efficient detection of potential small molecule-binding sites in proteins”. In: *Journal of Molecular Graphics and Modelling* 15.6 (1997), pp. 359–363. ISSN: 10933263. DOI: 10.1016/S1093-3263(98)00002-3.

[115] Takeshi Kawabata. “Detection of multiscale pockets on protein surfaces using mathematical morphology”. In: *Proteins: Structure, Function and Bioinformatics* 78.5 (2010), pp. 1195–1211. ISSN: 08873585. DOI: 10.1002/prot.22639.

[116] Robert A Copeland, David L Pompliano, and Thomas D Meek. “Drug–target residence time and its implications for lead optimization”. In: *Nature Reviews Drug Discovery* 5.9 (2006), pp. 730–739. ISSN: 1474-1784. DOI: 10.1038/nrd2082. URL: https://doi.org/10.1038/nrd2082.

[117] Albert C Pan et al. “Molecular determinants of drug–receptor binding kinetics”. In: *Drug Discovery Today* 18.13 (2013), pp. 667–673. ISSN: 1359-6446. DOI: https://doi.org/10.1016/j.drudis.2013.02.007. URL: https://www.sciencedirect.com/science/article/pii/S1359644613000627.

[118] Grubmüller Helmut, Heymann Berthold, and Tavan Paul. “Ligand Binding: Molecular Mechanics Calculation of the Streptavidin-Biotin Rupture Force”. In: *Science* 271.5251 (Feb. 1996), pp. 997–999. DOI: 10.1126/

science.271.5251.997. URL: https://doi.org/10.1126/science.271.5251.997.

[119] Jagdish Suresh Patel et al. “Steered Molecular Dynamics Simulations for Studying Protein–Ligand Interaction in Cyclin-Dependent Kinase 5”. In: *Journal of Chemical Information and Modeling* 54.2 (Feb. 2014), pp. 470–480. ISSN: 1549-9596. DOI: 10.1021/ci4003574. URL: https://doi.org/10.1021/ci4003574.

[120] Maciej Majewski. “Implications of Structural Stability for Drug Design Implications of Structural Stability for Drug Design Maciej Majewski”. PhD thesis. Universitat de Barcelona, 2019.

[121] Francesco Colizzi et al. “Single-Molecule Pulling Simulations Can Discern Active from Inactive Enzyme Inhibitors”. In: *Journal of the American Chemical Society* 132.21 (June 2010), pp. 7361–7371. ISSN: 0002-7863. DOI: 10.1021/ja100259r. URL: https://doi.org/10.1021/ja100259r.

[122] Sotomayor Marcos and Schulten Klaus. “Single-Molecule Experiments in Vitro and in Silico”. In: *Science* 316.5828 (May 2007), pp. 1144–1148. DOI: 10.1126/science.1137591. URL: https://doi.org/10.1126/science.1137591.

[123] Barry Isralewitz et al. “Steered molecular dynamics investigations of protein function”. In: *Journal of Molecular Graphics and Modelling* 19.1 (2001), pp. 13–25. ISSN: 1093-3263. DOI: https://doi.org/10.1016/S1093-3263(00)00133-9. URL: https://www.sciencedirect.com/science/article/pii/S1093326300001339.

[124] Alessandro Borgia et al. “Extreme disorder in an ultrahigh-affinity protein complex”. In: *Nature* 555.7694 (2018), pp. 61–66. ISSN: 1476-4687. DOI:

10.1038/nature25762. URL: https://doi.org/10.1038/nature25762.

[125] Caterina Bissantz, Bernd Kuhn, and Martin Stahl. "A Medicinal Chemist's Guide to Molecular Interactions". In: *Journal of Medicinal Chemistry* 53.14 (July 2010), pp. 5061–5084. ISSN: 0022-2623. DOI: 10.1021/jm100112j. URL: https://doi.org/10.1021/jm100112j.

[126] I D Kuntz et al. "The maximal affinity of ligands". eng. In: *Proceedings of the National Academy of Sciences of the United States of America* 96.18 (Aug. 1999), pp. 9997–10002. ISSN: 0027-8424. DOI: 10.1073/pnas.96.18.9997. URL: https://pubmed.ncbi.nlm.nih.gov/10468550%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC17830/.

[127] Alan R Fersht. "The hydrogen bond in molecular recognition". In: *Trends in Biochemical Sciences* 12 (1987), pp. 301–304. ISSN: 0968-0004. DOI: https://doi.org/10.1016/0968-0004(87)90146-0. URL: https://www.sciencedirect.com/science/article/pii/0968000487901460.

[128] Maciej Majewski, Sergio Ruiz-Carmona, and Xavier Barril. "Dynamic Undocking: A Novel Method for Structure-Based Drug Discovery BT - Rational Drug Design: Methods and Protocols". In: ed. by Thomas Mavromoustakos and Tahsin F Kellici. New York, NY: Springer New York, 2018, pp. 195–215. ISBN: 978-1-4939-8630-9. DOI: 10.1007/978-1-4939-8630-9_11. URL: https://doi.org/10.1007/978-1-4939-8630-9%7B%5C_%7D11.

[129] Chemical Computing Group ULC. *Molecular Operating Environment (MOE), 2019.01*. Montreal, 2021.

[130] David A Case et al. “The Amber biomolecular simulation programs”. In: *Journal of Computational Chemistry* 26.16 (Dec. 2005), pp. 1668–1688. ISSN: 0192-8651. DOI: https://doi.org/10.1002/jcc.20290. URL: https://doi.org/10.1002/jcc.20290.

[131] Enis Afgan et al. “The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update”. In: *Nucleic Acids Research* 46.W1 (May 2018), W537–W544. ISSN: 0305-1048. DOI: 10.1093/nar/gky379. eprint: https://academic.oup.com/nar/article-pdf/46/W1/W537/25110642/gky379.pdf. URL: https://doi.org/10.1093/nar/gky379.

[132] Yudong Qiu et al. “Development and Benchmarking of Open Force Field v1.0.0, the Parsley Small Molecule Force Field”. In: *ChemRxiv* (2021). DOI: 10.33774/chemrxiv-2021-l070l-v4.

[133] William Humphrey, Andrew Dalke, and Klaus Schulten. “VMD – Visual Molecular Dynamics”. In: *Journal of Molecular Graphics* 14 (1996), pp. 33–38.

[134] Maciej Majewski, Sergio Ruiz-Carmona, and Xavier Barril. “An investigation of structural stability in protein-ligand complexes reveals the balance between order and disorder”. In: *Communications Chemistry* 2.1 (2019), p. 110. ISSN: 2399-3669. DOI: 10.1038/s42004-019-0205-5. URL: https://doi.org/10.1038/s42004-019-0205-5.

[135] Gregory L Warren et al. “Essential considerations for using protein–ligand structures in drug discovery”. In: *Drug Discovery Today* 17.23 (2012), pp. 1270–1281. ISSN: 1359-6446. DOI: https://doi.org/10.1016/j.drudis.2012.06.011. URL: https://www.sciencedirect.com/science/article/pii/S1359644612002048.

[136] Angelo D Favia et al. “SERAPhiC: A Benchmark for in Silico Fragment-Based Drug Design”. In: *Journal of Chemical Information and Modeling* 51.11 (Nov. 2011), pp. 2882–2896. ISSN: 1549-9596. DOI: 10.1021/ci2003363. URL: https://doi.org/10.1021/ci2003363.

[137] Maciej Majewski and Xavier Barril. “Structural Stability Predicts the Binding Mode of Protein–Ligand Complexes”. In: *Journal of Chemical Information and Modeling* 60.3 (Mar. 2020), pp. 1644–1651. ISSN: 1549-9596. DOI: 10.1021/acs.jcim.9b01062. URL: https://doi.org/10.1021/acs.jcim.9b01062.

[138] Panagis Filippakopoulos and Stefan Knapp. “The bromodomain interaction module”. In: *FEBS Letters* 586.17 (Aug. 2012), pp. 2692–2704. ISSN: 0014-5793. DOI: https://doi.org/10.1016/j.febslet.2012.04.045. URL: https://doi.org/10.1016/j.febslet.2012.04.045.

[139] Valentina Straniero et al. “Benzodioxane-Benzamides as Antibacterial Agents: Computational and SAR Studies to Evaluate the Influence of the 7-Substitution in FtsZ Interaction”. In: *ChemMedChem* 15.2 (2020), pp. 195–209. DOI: https://doi.org/10.1002/cmdc.201900537. eprint: https://chemistry-europe.onlinelibrary.wiley.com/doi/pdf/10.1002/cmdc.201900537. URL: https://chemistry-europe.onlinelibrary.wiley.com/doi/abs/10.1002/cmdc.201900537.

[140] Lieyang Chen et al. “Hidden bias in the DUD-E dataset leads to misleading performance of deep learning in structure-based virtual screening”. In: *PLOS ONE* 14.8 (Aug. 2019), pp. 1–22. DOI: 10.1371/journal.pone.0220113. URL: https://doi.org/10.1371/journal.pone.0220113.

[141] Fatma-Elzahraa Eid et al. “Systematic auditing is essential to debiasing machine learning in biology”. In: *Communications Biology* 4.1 (2021), p. 183. ISSN: 2399-3642. DOI: 10.1038/s42003-021-01674-5. URL: https://doi.org/10.1038/s42003-021-01674-5.

[142] D.Roeland Boer et al. “Superstar: comparison of CSD and PDB-based interaction fields as a basis for the prediction of protein-ligand interactions11Edited by R. Huber”. In: *Journal of Molecular Biology* 312.1 (2001), pp. 275–287. ISSN: 0022-2836. DOI: https://doi.org/10.1006/jmbi.2001.4901. URL: https://www.sciencedirect.com/science/article/pii/S0022283601949010.

[143] Matteo Tiberti et al. “ENCORE: Software for Quantitative Ensemble Comparison”. In: *PLOS Computational Biology* 11.10 (Oct. 2015), pp. 1–16. DOI: 10.1371/journal.pcbi.1004415. URL: https://doi.org/10.1371/journal.pcbi.1004415.

[144] William R Pitt, Rinaldo W Montalvão, and Tom L Blundell. “Polyphony: superposition independent methods for ensemble-based drug discovery”. In: *BMC Bioinformatics* 15.1 (2014), p. 324. ISSN: 1471-2105. DOI: 10.1186/1471-2105-15-324. URL: https://doi.org/10.1186/1471-2105-15-324.

[145] Fredrik Österberg et al. “Automated docking to multiple target structures: Incorporation of protein mobility and structural water heterogeneity in AutoDock”. In: *Proteins: Structure, Function, and Bioinformatics* 46.1 (Jan. 2002), pp. 34–40. ISSN: 0887-3585. DOI: 10.1002/prot.10028.

[146] Garrett M Morris et al. “Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function”. In: *Journal of Computational Chemistry* 19.14 (Nov. 1998), pp. 1639–1662. ISSN: 0192-8651. DOI: https://doi.org/10.1002/(SICI)1096-

987X(19981115)19:14<1639::AID-JCC10>3.0.CO;2-B. URL: https://doi.org/10.1002/(SICI)1096-987X(19981115)19:14%7B%5C%%7D3C1639::AID-JCC10%7B%5C%%7D3E3.0.CO%20http://2-b.

[147] Andrea Volkamer et al. "DoGSiteScorer: a web server for automatic binding site prediction, analysis and druggability assessment". In: *Bioinformatics* 28.15 (May 2012), pp. 2074–2075. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/bts310. URL: https://doi.org/10.1093/bioinformatics/bts310.

[148] Samo Turk et al. "From Cancer to Pain Target by Automated Selectivity Inversion of a Clinical Candidate". In: *Journal of Medicinal Chemistry* 61.11 (June 2018), pp. 4851–4859. ISSN: 0022-2623. DOI: 10.1021/acs.jmedchem.8b00140. URL: https://doi.org/10.1021/acs.jmedchem.8b00140.

[149] Thomas A Halgren. "Identifying and Characterizing Binding Sites and Assessing Druggability". In: *Journal of Chemical Information and Modeling* 49.2 (Feb. 2009), pp. 377–389. ISSN: 1549-9596. DOI: 10.1021/ci800324m. URL: https://doi.org/10.1021/ci800324m.

[150] Philipp S. Schmalhorst and Andreas Bergner. "A Grid Map Based Approach to Identify Nonobvious Ligand Design Opportunities in 3D Protein Structure Ensembles". In: *Journal of chemical information and modeling* 60.4 (2020), pp. 2178–2188. ISSN: 1549960X. DOI: 10.1021/acs.jcim.0c00051.

[151] "Bromodomain inhibitors: What does the future hold?" In: *Clinical Advances in Hematology and Oncology* 16.7 (2018), pp. 504–515. ISSN: 15430790.

[152] Susanne Muller, Panagis Filippakopoulos, and Stefan Knapp. "Bromodomains as therapeutic targets". In: *Expert Reviews in Molecular Medicine* 13.September (2011), e29. ISSN: 1462-3994. DOI: 10.1017/S1462399411001992. URL: http://www.journals.cambridge.org/abstract%7B%5C_%7DS1462399411001992.

[153] Albert J. Kooistra et al. "KLIFS: A structural kinase-ligand interaction database". In: *Nucleic Acids Research* 44.D1 (2016), pp. D365–D371. ISSN: 13624962. DOI: 10.1093/nar/gkv1082.

[154] Fleur M Ferguson and Nathanael S Gray. "Kinase inhibitors: the road ahead". In: *Nature Reviews Drug Discovery* 17 (Mar. 2018), p. 353. URL: https://doi.org/10.1038/nrd.2018.21%20http://10.0.4.14/nrd.2018.21.

[155] "Regulation of stress-induced cytokine production by pyridinylimidazoles inhibition of CSBP kinase". In: *Bioorganic and Medicinal Chemistry* 5.1 (1997), pp. 49–64. ISSN: 09680896. DOI: 10.1016/S0968-0896(96)00212-X.

[156] Stefan Bietz and Matthias Rarey. "SIENA: Efficient Compilation of Selective Protein Binding Site Ensembles". In: *Journal of Chemical Information and Modeling* 56.1 (Jan. 2016), pp. 248–259. ISSN: 1549-9596. DOI: 10.1021/acs.jcim.5b00588. URL: https://doi.org/10.1021/acs.jcim.5b00588.

[157] Stefan Bietz and Matthias Rarey. "ASCONA: Rapid Detection and Alignment of Protein Binding Site Conformations". In: *Journal of Chemical Information and Modeling* 55.8 (Aug. 2015), pp. 1747–1756. ISSN: 1549-9596. DOI: 10.1021/acs.jcim.5b00210. URL: https://doi.org/10.1021/acs.jcim.5b00210.

[158] "Protoss: a holistic approach to predict tautomers and protonation states in protein-ligand complexes". eng. In: *Journal of cheminformatics* 6 (Apr. 2014), p. 12. ISSN: 1758-2946. DOI: 10.1186/1758-2946-6-12. URL: https://www.ncbi.nlm.nih.gov/pubmed/24694216%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4019353/.

[159] Mark Davies et al. "ChEMBL web services: streamlining access to drug discovery data and utilities". In: *Nucleic Acids Research* 43.W1 (July 2015), W612–W620. ISSN: 0305-1048. DOI: 10.1093/nar/gkv352. URL: https://doi.org/10.1093/nar/gkv352.

[160] Nicolas Bosc et al. "Large scale comparison of QSAR and conformal prediction methods and their applications in drug discovery". eng. In: *Journal of cheminformatics* 11.1 (Jan. 2019), p. 4. ISSN: 1758-2946. DOI: 10.1186/s13321-018-0325-4. URL: https://pubmed.ncbi.nlm.nih.gov/30631996%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6690068/.

[161] Leland McInnes, John Healy, and Steve Astels. "hdbscan: Hierarchical density based clustering". In: *The Journal of Open Source Software* 2.11 (2017), p. 205. DOI: 10.21105/joss.00205.

[162] Charles R. Harris et al. "Array programming with NumPy". In: *Nature* 585.7825 (Sept. 2020), pp. 357–362. DOI: 10.1038/s41586-020-2649-2. URL: https://doi.org/10.1038/s41586-020-2649-2.

[163] "Design and Synthesis of a Highly Selective and in Vivo-Capable Inhibitor of the Second Bromodomain of the Bromodomain and Extra Terminal Domain Family of Proteins". In: *Journal of Medicinal Chemistry* 63.17 (2020), pp. 9070–9092. ISSN: 15204804.

[164] "Benzoisoquinolinediones as Potent and Selective Inhibitors of BRPF2 and TAF1/TAF1L Bromodomains". In: *Journal of Medicinal Chemistry* 60.9 (2017), pp. 4002–4022. ISSN: 15204804.

[165] Zhulun Wang et al. "Structural basis of inhibitor selectivity in MAP kinases". In: *Structure* 6.9 (1998), pp. 1117–1128. ISSN: 0969-2126. DOI: https://doi.org/10.1016/S0969-2126(98)00113-0. URL: http://www.sciencedirect.com/science/article/pii/S0969212698001130.

[166] Guy W. Bemis and Mark A. Murcko. "The properties of known drugs. 1. Molecular frameworks". In: *Journal of Medicinal Chemistry* 39.15 (1996), pp. 2887–2893. ISSN: 00222623. DOI: 10.1021/jm9602928.

[167] Roberto Battistutta et al. "Unprecedented Selectivity and Structural Determinants of a New Class of Protein Kinase CK2 Inhibitors in Clinical Trials for the Treatment of Cancer". In: *Biochemistry* 50.39 (Oct. 2011), pp. 8478–8488. ISSN: 0006-2960. DOI: 10.1021/bi2008382. URL: https://doi.org/10.1021/bi2008382.

[168] Fabrice Pierre et al. "Discovery and SAR of 5-(3-Chlorophenylamino)benzo[c][2,6]naphthyridine-8- carboxylic Acid (CX-4945), the first clinical stage inhibitor of protein kinase CK2 for the Treatment of Cancer". In: *Journal of Medicinal Chemistry* 54.2 (2011), pp. 635–654. ISSN: 00222623. DOI: 10.1021/jm101251q.

[169] Vigneshwari Subramanian et al. "Visually Interpretable Models of Kinase Selectivity Related Features Derived from Field-Based Proteochemometrics". In: *Journal of Chemical Information and Modeling* 53.11 (Nov. 2013), pp. 3021–3030. ISSN: 1549-9596. DOI: 10.1021/ci400369z.

[170] Oliver Korb et al. "Interactive and Versatile Navigation of Structural Databases". In: *Journal of Medicinal Chemistry* 59.9 (May 2016), pp. 4257–4266.

ISSN: 0022-2623. DOI: 10.1021/acs.jmedchem.5b01756. URL: https://doi.org/10.1021/acs.jmedchem.5b01756.

[171] Sebastian Salentin et al. "PLIP: Fully automated protein-ligand interaction profiler". In: *Nucleic Acids Research* 43.W1 (2015), W443–W447. ISSN: 13624962. DOI: 10.1093/nar/gkv315.

[172] Antonio Marinho da Silva Neto et al. "A superposition free method for protein conformational ensemble analyses and local clustering based on a differential geometry representation of backbone". In: *Proteins: Structure, Function and Bioinformatics* 87.4 (2019), pp. 302–312. ISSN: 10970134. DOI: 10.1002/prot.25652.

[173] Michael R Shirts et al. "Lessons learned from comparing molecular dynamics engines on the SAMPL5 dataset". In: *Journal of Computer-Aided Molecular Design* 31.1 (2017), pp. 147–161. ISSN: 1573-4951. DOI: 10.1007/s10822-016-9977-1. URL: https://doi.org/10.1007/s10822-016-9977-1.

[174] Viktor Hornak et al. "Comparison of multiple Amber force fields and development of improved protein backbone parameters". eng. In: *Proteins* 65.3 (Nov. 2006), pp. 712–725. ISSN: 1097-0134. DOI: 10.1002/prot.21123. URL: https://pubmed.ncbi.nlm.nih.gov/16981200%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4805110/.

[175] Araz Jakalian, David B Jack, and Christopher I Bayly. "Fast, efficient generation of high-quality atomic charges. AM1-BCC model: II. Parameterization and validation". In: *Journal of Computational Chemistry* 23.16 (Dec. 2002), pp. 1623–1641. ISSN: 0192-8651. DOI: https://doi.org/10.1002/jcc.10128. URL: https://doi.org/10.1002/jcc.10128.

[176] "Automatic atom type and bond type perception in molecular mechanical calculations". In: *Journal of Molecular Graphics and Modelling* 25.2 (2006), pp. 247–260. ISSN: 1093-3263. DOI: https://doi.org/10.1016/j.jmgm.2005.12.005. URL: https://www.sciencedirect.com/science/article/pii/S1093326305001737.

[177] P Eastman et al. "OpenMM 4: A Reusable, Extensible, Hardware Independent Library for High Performance Molecular Simulation". In: *Journal of Chemical Theory and Computation* 9.1 (2013), pp. 461–469. URL: http://pubs.acs.org/doi/abs/10.1021/ct300857j.

[178] Warren L DeLano. "Pymol: An open-source molecular graphics tool". In: *{CCP4} Newsletter On Protein Crystallography* 40 (2002). URL: http://www.ccp4.ac.uk/newsletters/newsletter40/11%7B%5C_%7Dpymol.pdf.

[179] Pauli Virtanen et al. "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python". In: *Nature Methods* 17 (2020), pp. 261–272. DOI: 10.1038/s41592-019-0686-2.

[180] Ernst Schonbrunn et al. "Development of Highly Potent and Selective Diaminothiazole Inhibitors of Cyclin-Dependent Kinases". In: *Journal of Medicinal Chemistry* 56.10 (May 2013), pp. 3768–3782. ISSN: 0022-2623. DOI: 10.1021/jm301234k. URL: https://doi.org/10.1021/jm301234k.

[181] Maciej Wójcikowski, Piotr Zielenkiewicz, and Pawel Siedlecki. "Open Drug Discovery Toolkit (ODDT): a new open-source player in the drug discovery field". In: *Journal of Cheminformatics* 7.1 (2015), p. 26. ISSN: 1758-2946. DOI: 10.1186/s13321-015-0078-2. URL: https://doi.org/10.1186/s13321-015-0078-2.

[182] David Weininger. “SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules”. In: *Journal of Chemical Information and Computer Sciences* 28.1 (Feb. 1988), pp. 31–36. ISSN: 0095-2338. DOI: 10.1021/ci00057a005. URL: https://pubs.acs.org/doi/abs/10.1021/ci00057a005.

[183] Simon J Teague et al. “The Design of Leadlike Combinatorial Libraries”. In: *Angewandte Chemie International Edition* 38.24 (Dec. 1999), pp. 3743–3748. ISSN: 1433-7851. DOI: https://doi.org/10.1002/(SICI)1521-3773(19991216)38:24<3743::AID-ANIE3743>3.0.CO;2-U. URL: https://doi.org/10.1002/(SICI)1521-3773(19991216)38:24%7B%5C%%7D3C3743::AID-ANIE3743%7B%5C%%7D3E3.0.CO%20http://2-u.

[184] Jason C Cole et al. “Knowledge-Based Conformer Generation Using the Cambridge Structural Database”. In: *Journal of Chemical Information and Modeling* 58.3 (Mar. 2018), pp. 615–629. ISSN: 1549-9596. DOI: 10.1021/acs.jcim.7b00697. URL: https://doi.org/10.1021/acs.jcim.7b00697.

[185] Gareth Jones et al. “Development and validation of a genetic algorithm for flexible docking11Edited by F. E. Cohen”. In: *Journal of Molecular Biology* 267.3 (1997), pp. 727–748. ISSN: 0022-2836. DOI: https://doi.org/10.1006/jmbi.1996.0897. URL: https://www.sciencedirect.com/science/article/pii/S0022283696908979.

[186] John W Liebeschuetz, Jason C Cole, and Oliver Korb. “Pose prediction and virtual screening performance of GOLD scoring functions in a standardized test.” eng. In: *Journal of computer-aided molecular design* 26.6 (June 2012), pp. 737–748. ISSN: 1573-4951 (Electronic). DOI: 10.1007/s10822-012-9551-4.

[187] *GOLD User Guide*. Version 1. Oct. 2021. DOI: https://www.ccdc.cam.ac.uk/solutions/csd-discovery/components/gold/. URL: https://www.ccdc.cam.ac.uk/support-and-resources/ccdcresources/GOLD_User_Guide_2020_1.pdf.

[188] Eleanor Williams et al. *Activin A receptor, type I (ACVR1); A Target Enabling Package*. Nov. 2020. DOI: 10.5281/zenodo.4436354.

[189] Apirat Chaikuad et al. "Structure of the bone morphogenetic protein receptor ALK2 and implications for fibrodysplasia ossificans progressiva". eng. In: *The Journal of biological chemistry* 287.44 (Oct. 2012), pp. 36990–36998. ISSN: 1083-351X. DOI: 10.1074/jbc.M112.365932. URL: https://pubmed.ncbi.nlm.nih.gov/22977237%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3481300/.

[190] Scott A Wildman and Gordon M Crippen. "Prediction of Physicochemical Parameters by Atomic Contributions". In: *Journal of Chemical Information and Computer Sciences* 39.5 (Sept. 1999), pp. 868–873. ISSN: 0095-2338. DOI: 10.1021/ci990307l. URL: https://doi.org/10.1021/ci990307l.

[191] Célien Jacquemard et al. "Binding mode information improves fragment docking". In: *Journal of Cheminformatics* 11.1 (2019), p. 24. ISSN: 1758-2946. DOI: 10.1186/s13321-019-0346-7. URL: https://doi.org/10.1186/s13321-019-0346-7.

[192] Joseph Newman et al. *SARS-CoV-2 NSP13; A Target Enabling Package*. Version 1. This document represents version 1 of the TEP datasheet and includes all updates on the project as of November 2020. For more information about TEPs and the TEP Programme, please visit https://www.cmd.ox.ac.uk/TEP.

Nov. 2020. DOI: 10.5281/zenodo.4449959. URL: https://doi.org/10.5281/zenodo.4449959.

[193] Roman Lagoutte, Remi Patouret, and Nicolas Winssinger. "Covalent inhibitors: an opportunity for rational target selectivity". In: *Current Opinion in Chemical Biology* 39 (2017), pp. 54–63. ISSN: 1367-5931. DOI: https://doi.org/10.1016/j.cbpa.2017.05.008. URL: https://www.sciencedirect.com/science/article/pii/S1367593117300601.

[194] Wei Qin et al. "Research Progress on PARP14 as a Drug Target". eng. In: *Frontiers in pharmacology* 10 (Mar. 2019), p. 172. ISSN: 1663-9812. DOI: 10.3389/fphar.2019.00172. URL: https://www.ncbi.nlm.nih.gov/pubmed/30890936%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6411704/.

[195] Marion Schuller. "Investigating strategies to modify PARP14 function through macrodomain inhibition". In: *PhD Thesis, University of Oxford* (2017).

[196] Mihaela D Smilova et al. "Fragment Hotspot Mapping to Identify Selectivity-Determining Regions between Related Proteins". In: *Journal of Chemical Information and Modeling* 62.2 (Jan. 2022), pp. 284–294. ISSN: 1549-9596. DOI: 10.1021/acs.jcim.1c00823. URL: https://doi.org/10.1021/acs.jcim.1c00823.

A | Additional Tables and Figures

Table A.1: List of structures used in the Chapter 2 case study for BRD1

Target	Family	PDB ID	Chain	Resolution (Å)	Ligand MW (Da)
BRD1	Bromodomain	3RCW	C	2.21	99.13
BRD1	Bromodomain	5AMF	A	1.75	194.23
BRD1	Bromodomain	5PO0	A	1.46	164.16
BRD1	Bromodomain	5PO6	A	1.61	238.08
BRD1	Bromodomain	5PO7	A	1.5	236.23
BRD1	Bromodomain	5PO8	A	1.5	176.21
BRD1	Bromodomain	5PO9	A	2.12	241.21
BRD1	Bromodomain	5POB	A	1.78	177.2
BRD1	Bromodomain	5POC	A	1.48	283.08
BRD1	Bromodomain	5POE	A	1.52	174.2
BRD1	Bromodomain	5POH	A	1.61	178.23
BRD1	Bromodomain	5POI	A	2.37	137.14
BRD1	Bromodomain	5POJ	A	1.62	232.32
BRD1	Bromodomain	5PON	A	1.52	205.26
BRD1	Bromodomain	5POO	A	1.5	230.27
BRD1	Bromodomain	5POQ	A	1.97	206.26
BRD1	Bromodomain	5POR	A	1.58	142.16
BRD1	Bromodomain	5POS	A	1.75	179.22
BRD1	Bromodomain	5POT	A	1.63	183.63
BRD1	Bromodomain	5POU	B	1.43	247.34
BRD1	Bromodomain	5POW	A	1.77	150.18
BRD1	Bromodomain	5POX	A	1.75	194.23
BRD1	Bromodomain	5PP0	B	1.61	151.17

Table A.2: List of structures used in the Chapter 2 case study for BRPF1

Target	Family	PDB ID	Chain	Resolution Å	Ligand MW (Da)
BRPF1	Bromodomain	5C85	A	1.7	225.04
BRPF1	Bromodomain	5C87	A	1.55	145.16
BRPF1	Bromodomain	5C89	A	1.65	241.29
BRPF1	Bromodomain	5D7X	A	1.35	253.26
BRPF1	Bromodomain	5DY7	A	1.69	216.16
BRPF1	Bromodomain	5DYA	A	1.65	220.22
BRPF1	Bromodomain	5DYC	A	1.65	225.04
BRPF1	Bromodomain	5E3D	A	1.71	200.24
BRPF1	Bromodomain	5E3G	A	1.65	184.18
BRPF1	Bromodomain	5EM3	A	1.4	145.16
BRPF1	Bromodomain	5EPR	A	1.65	162.19
BRPF1	Bromodomain	5EPS	A	1.47	160.17
BRPF1	Bromodomain	5EQ1	A	1.55	189.24
BRPF1	Bromodomain	5ETB	A	1.33	173.21
BRPF1	Bromodomain	5ETD	A	1.4	159.18
BRPF1	Bromodomain	5EV9	A	1.45	252.27
BRPF1	Bromodomain	5EVA	A	1.45	237.21
BRPF1	Bromodomain	5EWC	A	1.75	234.25
BRPF1	Bromodomain	5EWD	A	1.58	155.2
BRPF1	Bromodomain	5EWH	A	1.63	145.16
BRPF1	Bromodomain	5EWV	A	1.67	197.24
BRPF1	Bromodomain	5EWW	A	1.73	257.33
BRPF1	Bromodomain	5T4U	A	1.5	159.18
BRPF1	Bromodomain	5O4T	A	1.5	190.2
BRPF1	Bromodomain	5OWB	A	1.65	164.16
BRPF1	Bromodomain	5OWE	A	1.7	228.2

Table A.3: List of structures used in the Chapter 2 case study for ERK2

Target	Family	PDB ID	Chain	Resolution Å	Ligand MW (Da)
ERK2	Kinase	4QP1	A	2.7	225.25
ERK2	Kinase	6G91	A	1.8	213.66
ERK2	Kinase	5BUI	A	2.12	294.33
ERK2	Kinase	5BUJ	A	1.85	293.32
ERK2	Kinase	6G92	A	1.99	203.24
ERK2	Kinase	4QP4	B	2.2	217.27
ERK2	Kinase	4QP3	B	2.6	269.3
ERK2	Kinase	4QP8	B	2.45	262.27
ERK2	Kinase	6QAG	A	2.07	138.13
ERK2	Kinase	2OJG	A	2.0	280.32
ERK2	Kinase	6QA3	A	1.57	136.15
ERK2	Kinase	4G6N	A	2.0	232.71
ERK2	Kinase	6Q7T	A	1.6	168.15
ERK2	Kinase	6Q7K	A	1.84	83.09
ERK2	Kinase	6QAW	A	1.84	232.3
ERK2	Kinase	4QP7	B	2.25	185.19
ERK2	Kinase	4QP2	A	2.23	168.58

Table A.4: List of structures used in the Chapter 2 case study for p38α

Target	Family	PDB ID	Chain	Resolution Å	Ligand MW (Da)
p38a	Kinase	1WBO	A	2.16	128.56
p38a	Kinase	1WBW	A	2.41	250.29
p38a	Kinase	1BL6	A	2.5	293.34
p38a	Kinase	1W7H	A	2.21	200.24
p38a	Kinase	1W84	A	2.2	222.28

Table A.5: List of structures used in the Chapter 2 case study for CK2α

Target	Family	PDB ID	Chain	Resolution (Å)	Ligand MW (Da)
CK2a	Kinase	5CT0	B	2.01	276.78
CK2a	Kinase	6Q4Q	B	1.45	121.11
CK2a	Kinase	5MOH	A	1.38	181.17
CK2a	Kinase	6HOP	A	1.55	192.21
CK2a	Kinase	3MB6	A	1.75	294.22
CK2a	Kinase	3WAR	A	1.04	122.1
CK2a	Kinase	3Q9Y	A	1.8	272.21
CK2a	Kinase	6YPJ	A	1.64	295.34
CK2a	Kinase	5MO7	A	2.15	289.78
CK2a	Kinase	3AXW	A	2.5	220.23
CK2a	Kinase	3OWL	A	2.1	282.72
CK2a	Kinase	6HNW	A	2.0	268.22
CK2a	Kinase	3MB7	A	1.65	294.22
CK2a	Kinase	5CSV	A	1.38	136.13
CK2a	Kinase	5CSP	A	1.5	151.14
CK2a	Kinase	6QY7	A	2.1	271.24
CK2a	Kinase	6HOU	A	1.8	152.15
CK2a	Kinase	5MMF	A	1.99	260.78
CK2a	Kinase	5MOT	A	2.09	274.27
CK2a	Kinase	6TEI	B	1.76	296.32
CK2a	Kinase	5MOV	A	2.2	163.15
CK2a	Kinase	5MOW	A	1.86	188.03
CK2a	Kinase	3OWK	A	1.8	282.72
CK2a	Kinase	6HOP	A	1.55	193.18
CK2a	Kinase	6HOT	A	1.5	178.18
CK2a	Kinase	3BQC	A	1.5	270.24
CK2a	Kinase	3WIK	A	2.0	264.26
CK2a	Kinase	3OWJ	A	1.85	278.3

Table A.6: List of structures used in the Chapter 2 case study for PIM1

Target	Family	PDB ID	Chain	Resolution (Å)	Ligand MW (Da)
PIM1	Kinase	3VBT	A	2.23	194.62
PIM1	Kinase	3JPV	A	2.35	234.25
PIM1	Kinase	5KGD	A	1.98	195.22
PIM1	Kinase	5KGE	A	2.23	228.08
PIM1	Kinase	1YXX	A	2.0	238.24
PIM1	Kinase	4MBI	A	2.3	272.33
PIM1	Kinase	5N4R	A	2.13	158.18
PIM1	Kinase	5N4X	A	2.2	299.97
PIM1	Kinase	2XIY	A	2.2	148.16
PIM1	Kinase	5N4N	A	2.09	204.25
PIM1	Kinase	5N5L	A	1.97	239.05
PIM1	Kinase	5N4Z	A	2.26	162.18
PIM1	Kinase	4LL5	A	2.0	297.35
PIM1	Kinase	5KGI	A	2.13	207.08
PIM1	Kinase	2XIZ	A	2.21	148.14
PIM1	Kinase	3C4E	D	1.98	209.25
PIM1	Kinase	3R00	A	2.1	240.03
PIM1	Kinase	3VBX	A	2.03	241.04
PIM1	Kinase	5DHJ	A	2.46	210.23
PIM1	Kinase	5KGG	A	1.95	195.67
PIM1	Kinase	5NDT	A	1.99	159.16
PIM1	Kinase	4MTA	A	2.2	201.2
PIM1	Kinase	5N50	A	1.92	216.69
PIM1	Kinase	5N4V	A	1.85	247.29
PIM1	Kinase	3UIX	A	2.2	295.36
PIM1	Kinase	3VBV	A	2.08	188.18
PIM1	Kinase	1YXV	A	2.0	191.18
PIM1	Kinase	3VC4	A	2.23	273.23
PIM1	Kinase	5N4U	A	2.2	232.26
PIM1	Kinase	2XIX	A	2.4	99.09
PIM1	Kinase	3VBY	A	2.27	211.22
PIM1	Kinase	3JYA	A	2.1	271.12

Table A.7: Chunks used in the prospective case studies

Target	Interaction	Chunking Cutoff (Å)	Waters
ACVR1	B_ARG_335_O	8.0	HOH28, HOH62
ACVR1	B_ARG_335_NH1	8.0	HOH28, HOH62
ACVR1	B_GLY_337_N	8.0	HOH28, HOH62
ACVR1	B_TYR_381_O	8.0	HOH28, HOH62
ACVR1	B_ALA_383_N	9.5	HOH28, HOH62
ACVR1	B_VAL_402_O	8.0	HOH28, HOH62
PARP14	A_ILE_1236_N	9.0	None
PARP14	A_SER_1247_O	9.0	None
PARP14	A_VAL_1258_N	9.0	None
PARP14	A_SER_1259_N	9.0	None
PARP14	A_THR_1333_OG1	9.0	None
PARP14	A_GLY_1334_N	9.0	None
PARP14	A_ASN_1335_N	9.0	None
PARP14	A_PHE_1371_N	9.0	None
NSP13	A_ASN_265_ND2	11.0	HOH46
NSP13	A_HIS_290_ND1	11.0	HOH46
NSP13	A_SER_264_OG	11.0	HOH46
NSP13	A_SER_264_N	11.0	HOH46
NSP13	A_THR_286_O	11.0	HOH46
NSP13	A_LYS_320_NZ	11.0	HOH46
NSP13	B_ASN_260_OD1	11.0	None
NSP13	B_SER_259_O	11.0	None
NSP13	B_SER_259_OG1	11.0	None
NSP13	B_THR_281_O	11.0	None

Table A.8: Structural stability ($W_{QB\_mean\_min}$) of interactions for Hotspot cluster 1

Follow-up	Interaction	$W_{QB\_mean\_min}$ (kcal/mol)	StDev (kcal/mol)
PARP14A-x0161_146	A_ILE_1236_N	7.9	0.1
PARP14A-x0161_169	A_ILE_1236_N	2.6	0.3
PARP14A-x0457_53	A_ILE_1236_N	0.3	0.0
PARP14A-x0161_146	A_ASP_1235_OD1	2.0	0.4
PARP14A-x0161_169	A_ASP_1235_OD1	3.6	0.4

Table A.9: Structural stability ($W_{QB\_mean\_min}$) of interactions for PARP14 ensemble hotspot cluster 3

Follow-up	**Interaction**	**$W_{QB\_mean\_min}$ (kcal/mol)**	**StDev (kcal/mol)**
PARP14A-x0315_1438	A_PHE_1371_N	6.9	0.6
PARP14A-x0457_217	A_PHE_1371_N	6.2	0.3
PARP14A-x0712_516	A_PHE_1371_N	5.9	2.6
PARP14A-x0457_11	A_PHE_1371_N	2.3	1.3
PARP14A-x0712_1212	A_PHE_1371_N	2.1	1.2
PARP14A-x0712_241	A_PHE_1371_N	1.0	0.3
PARP14A-x0712_1227	A_PHE_1371_N	0.9	0.2
PARP14A-x0712_241	A_LEU_1372_N	5.3	0.7
PARP14A-x0457_11	A_LEU_1372_N	5.1	0.7
PARP14A-x0712_516	A_ALA_1330_N	4.4	1.2
PARP14A-x0712_1212	A_ALA_1330_N	1.3	0.1
PARP14A-x0712_241	A_ALA_1330_N	1.1	0.1
PARP14A-x0712_1227	A_ALA_1330_N	0.2	0.0

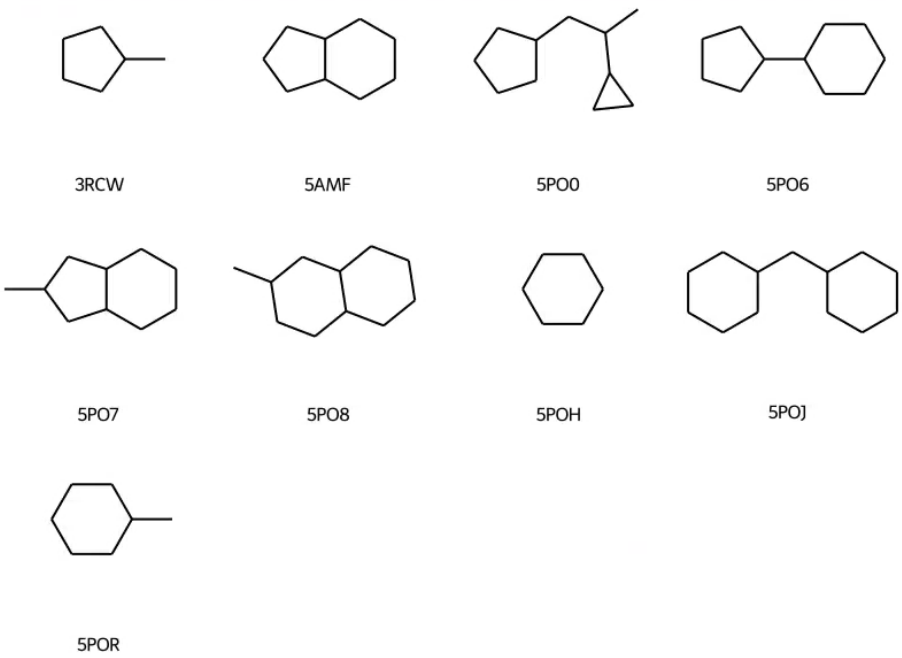


Figure A.1: Unique Murcko scaffolds for the BRD1 ensemble (23 structures) presented in Chapter 2.

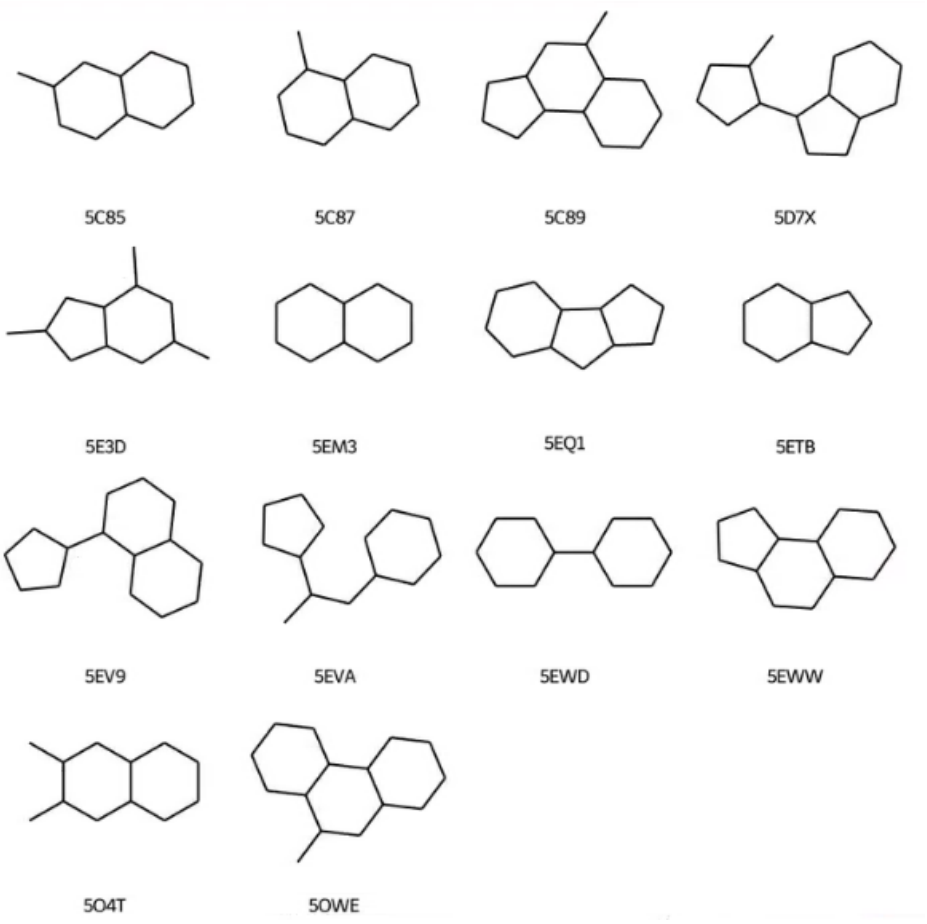


Figure A.2: Unique Murcko scaffolds for the BRPF1 ensemble (26 structures) presented in Chapter 2.

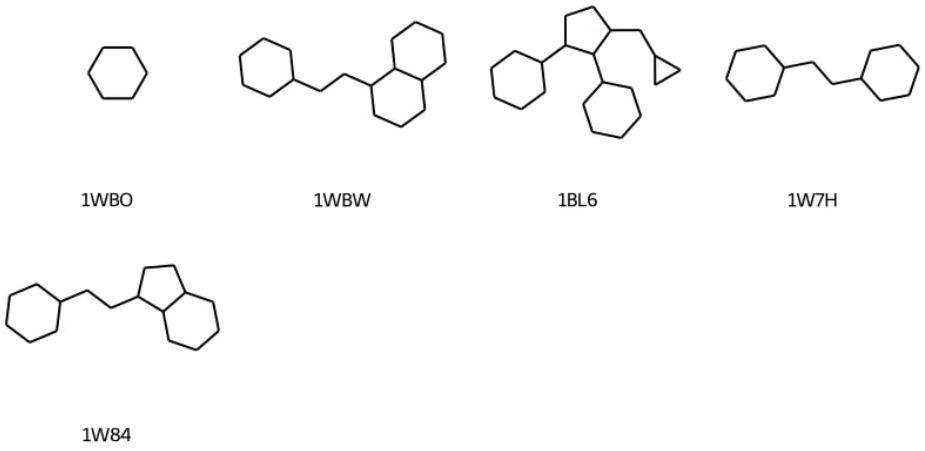


Figure A.3: Unique Murcko scaffolds for the p38α ensemble (5 structures) presented in Chapter 2.

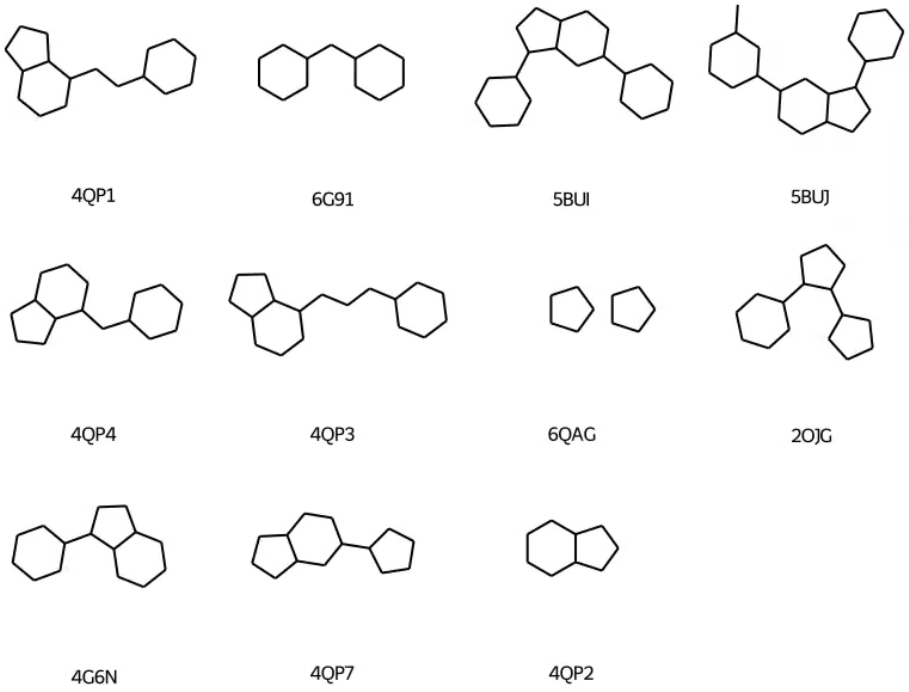


Figure A.4: Unique Murcko scaffolds for the ERK2 ensemble (19 structures) presented in Chapter 2.

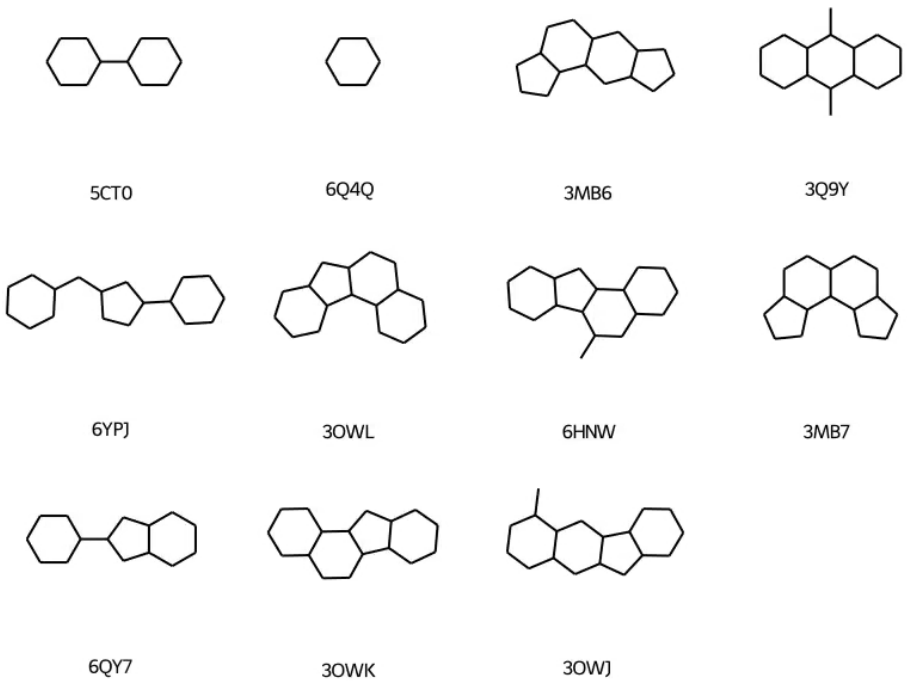


Figure A.5: Unique Murcko scaffolds for the CK2α ensemble (28 structures) presented in Chapter 2.

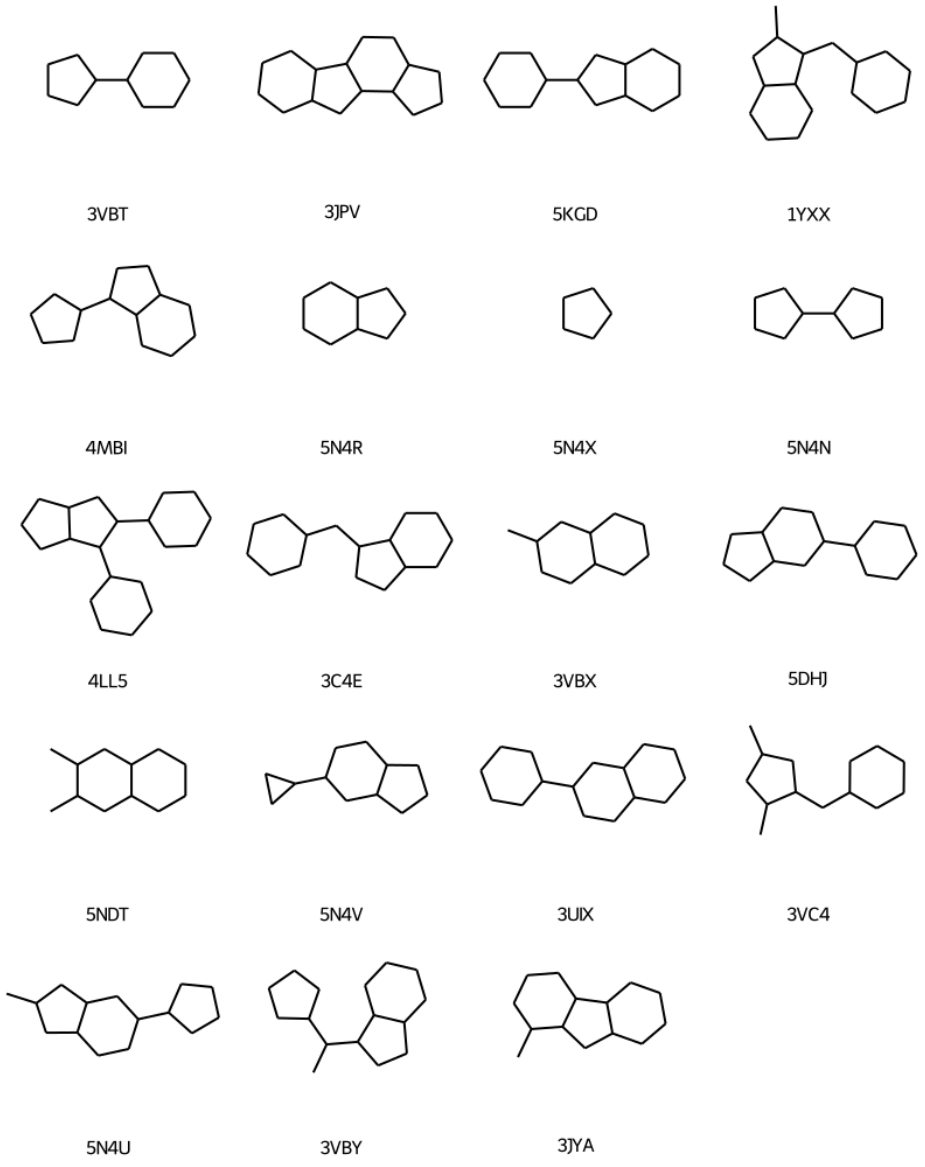


Figure A.6: Unique Murcko scaffolds for the PIM1 ensemble (5 structures) presented in Chapter 2.

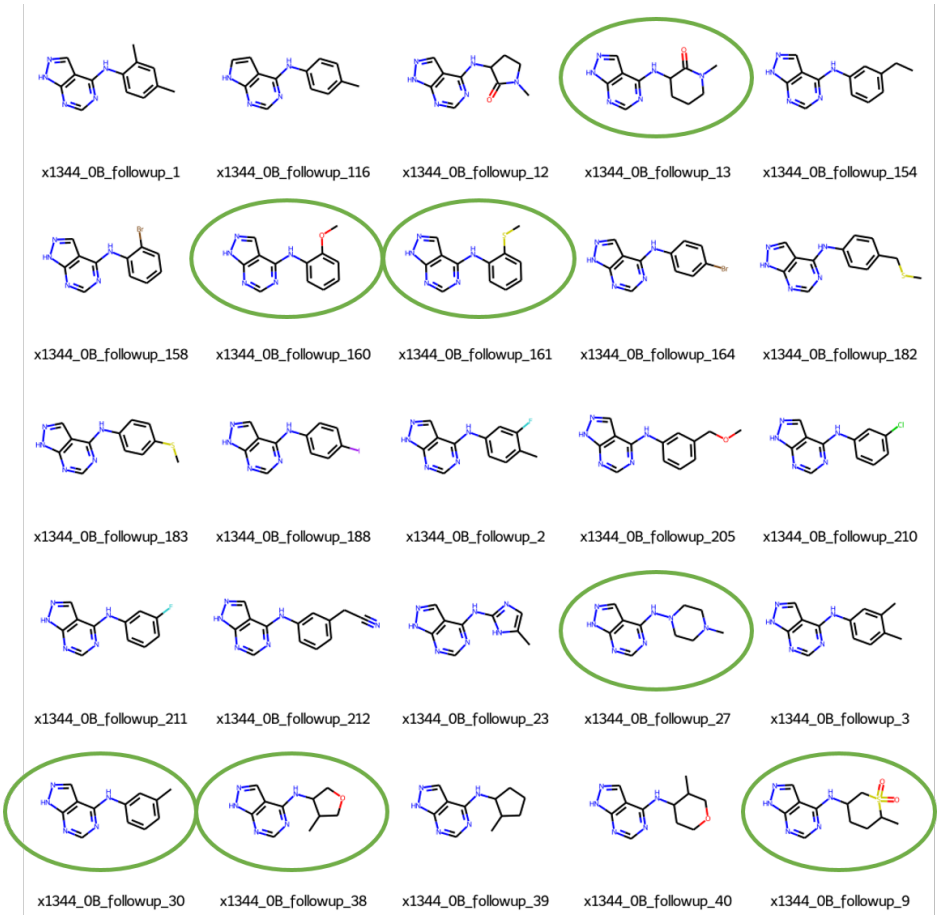


Figure A.7: ACVR1 follow-up compounds shortlisted based on their mean and polar hotspot scores. The compounds ordered after a further selection using by-feature scoring are circled in green.