Integrative Approaches in Fragment-Based Drug Discovery

Submitted by

Harold Grosjean

A thesis submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy in Structural Biology.
Trinity 2023

University of Oxford, Department of Biochemistry

Diamond Light Source

Linacre College

The work presented in this thesis was carried out between October 2018 and April 2023 under the shared supervision of Professor Philip C. Biggin and Professor Frank von Delft. All the work is my own unless otherwise stated and has not been submitted previously for any other degree at this, or any other, university.

Harold Grosjean

April 2023

I dedicate this thesis to my grandfather, Dr Philippe Speeckaert (†),

godson, Melchior Bozada

and goddaughter, Alice Grosjean.

Sincere appreciation extends to my supervisors, Prof Philip Biggin, for his exemplary guidance, care, patience, and strategic vision, and Prof Frank von Delft, for his boundless and sometimes overwhelming enthusiasm for scientific research. Profoundly grateful, I acknowledge their complementary approaches providing a balanced research experience.

Recognising the invaluable contributions of the Structural Biochemistry and Structural Biology Research Unit members is essential, particularly Dr Irfan Alibay, Dr William (Zhiyi) Wu, Dr Rocco Meli, Charlie Cook, and Cameron Anderson. Indispensable has been their expertise and camaraderie. Notably, XChem group members deserve appreciation, including Dr Tobias Krojer, Dr Lizbé Koekemoer, Dr Anthony Aimon, Dr Warren Thomson, Dr Ruben Sanchez-Garcia, Dr William Bradshaw, Dr Conor Wild, Dr Mihaela Smilova, Dr Charlie Tomlinson, Dr Romain Talon, Kate Fieseler, and the I04-1 beamline scientists for their steadfast support and guidance. Gratitude goes to Dr James Race for reviewing parts of the manuscript.

Beyond the lab, numerous collaborators and mentors have enriched my scientific education, including Prof John Spencer, Prof David Mobley, Dr Mehtap Işık, Thomas Grimes, Dr Jack Scantlebury, Prof John Chodera, Dr Oleg Fedorov, Binisha Karki, Erica Goldberger, and Dr Edward Fitzgerald. Respectfully acknowledging my transfer and confirmation examiners, Prof Paul Brennan and Dr Brian Marsden, for their invaluable feedback. Extending my thanks to Prof Matthew Higgins and Louise Samson for welcoming and guiding me through the DPhil program in Structural Biology. Appreciation is due for the funding provided by the Diamond Light Source and the Wellcome Trust, as well as the world-class infrastructure offered by Linacre College, the University of Oxford, the Department of Biochemistry, the Diamond Light Source, the Structural Genomic Consortium, the Science and Technology Facilities Council, and the Centre for Medicine Discovery, which made my DPhil experience truly enriching. Kindly, I acknowledge my examiners Prof Sir Tom Blundell and Prof Charlotte Dean for assessing my thesis.

Benefiting from the inspiration and guidance of Dr Emanuelle Thiran, Dr Alejandro Herrero Molina, Dr Timothy Simpson, Dr Boyan Bonev, Dr Ernesto Cota, Prof Michael Sternberg, and Dr Nessim Kichik, my scientific journey has been shaped. Relentless support from friends and family throughout my DPhil journey earns my deepest gratitude. Valuing everything, it has been a truly remarkable journey.

Harold Grosjean, April 2023

Abstract

This thesis combines experimental and computational methods to investigate aspects of fragment identification and elaboration in fragment-based ligand design, a promising approach for identifying small molecule drugs, to target the pharmacologically relevant bromodomain PHIP(2). The research covers various aspects of the process, from initial crystallographic fragment screening to validation of follow-up compounds.

Chapters 1 and 2 provide an overview of relevant perspectives and methodologies in fragment-based drug discovery. Chapter 3 reports a crystallographic fragment screening against PHIP(2), resolving 47 fragments at the acetylated-lysine binding site, and evaluates the abilities of crowdsourced computational methods to replicate fragment binding and crystallographic poses. This chapter highlights the challenges associated with using computational methods for reproducing crystallographic fragment screening results with submissions performing relatively weakly. Chapter 4 demonstrates the advantages of X-ray crystallographic screening of crude reaction mixtures generated robotically, showcasing reduced time, solvent, and hardware requirements. Soaking crude reaction mixtures maintains crystal integrity which led to the identification of 22 binders, 3 with an alternate pose caused by a single methyl addition to the core fragment and 1 hit in assays. It demonstrates how affordable methods can generate large amounts of crystallographic data of fragment elaborations. Chapter 5 develops an algorithmic approach to extract features associated with crystallographic binding, deriving simple binding scores using data from Chapter 4. The method identifies 26 false negatives with binding scores enriching binders over non-binders. Employing these scores prospectively in a virtual screening demonstrated how binding features can be exploited to select further follow-up compounds leading to low micromolar potencies. Chapter 6 attempts to integrate more computationally intensive methods to identify fragment follow-up compounds with increased potency through virtual screening enhanced with free energy calculations. Only two out of six synthesised follow-up compounds showed weak binding in assays, and none were resolved in crystal structures.

This thesis tackles critical challenges in follow-up design, synthesis, and dataset analysis, underlining the limitations of existing methods in advancing fragment-based drug discovery. It emphasises the necessity of integrative approaches for an optimised “design, make, test” cycle in fragment-based drug discovery.

Table of Contents

Abbreviations

AKT: Protein kinase B

AUC: Area Under the Curve

BACE1: Beta-Secretase 1

BCS: Bit Conservation Score

BCL2: B-Cell Lymphoma 2

BRAF: B-Raf Proto-Oncogene

CBB: Conserved Binding-Bits

ClogP: Calculated Octanol-Water Partition Coefficient

CNB: Conserved Non-binding Bits

CPU: Central Processing Unit

CRM: Crude Reaction Mixture

ΔG: Gibbs Free Energy of Binding

ΔH: Enthalpy Change

ΔS: Entropy Change

DMA: Dimethylacetamide

DMSO: Dimethyl Sulfoxide

ECFP: Extended-Connectivity Fingerprint

EG: Ethylene Glycol

EM: electron microscopy

FDA: Food and Drug Administration

FBDD: Fragment-Based Drug Discovery

FN: Fragment Network

GAFF: Generalised Amber Force Field

GCI: Grating-Coupled Interferometry

H4K91ac: Acetylated-Lysine 91 on Histone 4

HIF1α: Hypoxia-inducible factor-1 alpha

HTS: High-Throughput Screening

IC_{50}: Half Maximal Inhibitory Concentration

IE: Interaction Entropy

IRS: Insulin Receptor Substrate

IRS-1: Insulin Receptor Substrate-1

ITC: Isothermal Titration Calorimetry

Iters: Iterations

K_a: Association Constant

k_a: Association Rate

K_d: Association Constant

k_d: Dissociation Rate

LCMS: Liquid Chromatography-Mass Spectrometry

LDH5: Lactate-dehydrogenase 5

MD: Molecular Dynamics

MMGBSA: Molecular Mechanics Generalised Born Surface Area

MCS: Maximum Common Substructure

MSCheck: Semi-Automated LCMS Analyser Tool

M-Si: Manifold Similarity

M-Su: Manifold Substruture

NBS: Negative Binding Scores

NMR: Nuclear Magnetic Resonance

NPT: Constant Number of particles (N), Pressure (P), and Temperature (T)

NVT: Constant Number particles (N), Volume (V), and Temperature (T)

PanDDA: Pan-Dataset Density Analysis

PBS: Positive Binding Scores

PDB: Protein Data Bank

PDK1: Pyruvate Dehydrogenase Kinase 1

PHIP: Pleckstrin-homology domain interacting protein

PHIP(1): Pleckstrin-homology domain interacting protein’s first bromodomain

PHIP(2): Pleckstrin-homology domain interacting protein’s second bromodomain

PR: Precision-Recall

PTPN1: Protein Tyrosine Phosphatase Non-receptor Type 1

QC: Quality Control

QSAR: Quantitative Structure-Activity Relationship

RMSD: Root-Mean-Square Deviation

SAMPL: Statistical Assessment of the Modelling of Proteins and Ligands

SAR: Structure-Activity Relationships

SASA: Solvent-Accessible Surface Area

SID: Submission Identification Number

SMILES: Simplified Molecular Input Line Entry System

SPR: Surface Plasmon Resonance

SuMD: Supervised Molecular Dynamic Simulations

TC: Tanimoto Coefficient

TLN1: Talin-1

t-SNE: t-Distributed Stochastic Neighbour Embedding

VEGF: Vascular Endothelial Growth Factor

WD40: Beta-Transducin Repeat

XRC: X-Ray Crystallography

Chapter 1. Introduction

1.1. A concise journey through small molecule drug discovery

Drug discovery is the process of identifying, designing, and optimising therapeutic agents that aim to correct or suppress a pathological phenotype, ultimately contributing to the safe management of medical conditions. Prior to therapeutic intervention, identifying an appropriate target is crucial, as it underpins the selection of pathways involved in disease progression, enabling specific therapeutic interventions.

Various classes of therapeutic drugs exist, each with distinct properties and targets. Broadly speaking, they are two classes of drug agents: biologics are larger molecules with molecular weights typically greater than 1 KDa predominantly derived from living organisms, encompassing a range of agents such as cells, viruses, proteins, and oligonucleotides.[1] In contrast, small molecules are smaller organic compounds, with molecular weights typically lower than 1 KDa, often synthesised through chemical processes or derived from natural sources.[1]

Although both classes have specific advantages and drawbacks, small molecules tend to exhibit higher permeability and lower immunogenicity due to their smaller size, which can also enable oral bioavailability. Additionally, small molecules are typically less expensive to produce and have a more straightforward manufacturing process than biologics, whilst demonstrating greater stability and longer shelf life. Biologics, however, tend to be more specific than small molecules. For a small molecule drug to be both effective and safe, it needs to reach and bind to its target in a specific and potent manner.[2] This necessitates that the intermolecular interactions between the drug and the target exhibit complementary profiles, while remaining distinct enough to prevent the drug from binding to unintended sites, which could potentially result in adverse side effects.[3]

The identification of materials with therapeutic virtues for alleviating pain or combating diseases has been an integral aspect of human history (**Fig 1.1**). Notable examples include the use of willow plant extracts containing salicylic acid, a precursor of aspirin, for its analgesic and antipyretic properties around 4000 BCE (**Fig 1.1**).[4] Fast forward, in the late 19th and early 20th centuries, significant advances in chemical synthesis and analytical techniques, coupled with an expanding repertoire of therapeutic substances defined medicinal chemistry as a distinct discipline (**Fig 1.1**).[5] The conceptualisation of the chemoreceptor principle gave rise to pharmacology and early screening methodologies, with notable discoveries including arsphenamine. The mid-20th century saw the rise of *in vivo* testing and phenotypic evaluation, which, however, operated as a slow black box process (**Fig 1.1**).[5]

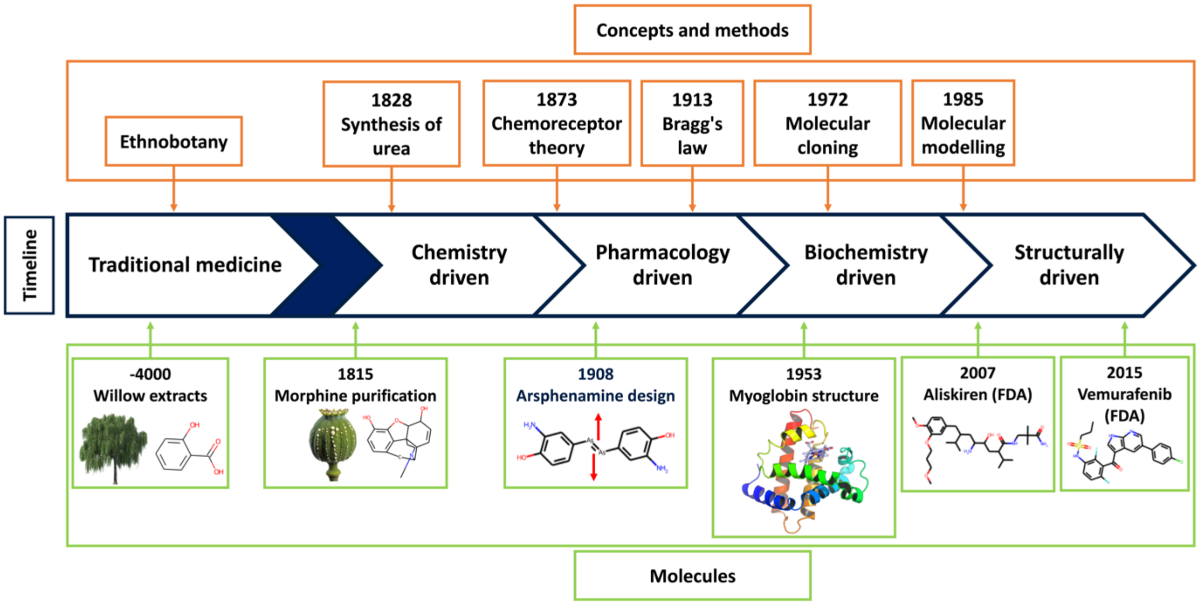


Figure 1.1: A concise timeline of pivotal discoveries and advancements in drug design. The central dark blue timeline features labels denoting major areas of influence. Key methodological scientific breakthroughs are displayed in the upper row with orange highlights, while applied discoveries, illustrated in green and with molecular representations, are depicted in the lower row. Years of Food and Drug Administration (FDA) approval are also indicated. This figure was inspired from Pina et al., (2009). The timeline above is unscaled.

The second half of the 20th century saw remarkable advancements in technological and biochemical techniques, leading to for the development of high-throughput screening methods (**Fig 1.1**).[6] These approaches involve testing large libraries, comprising thousands to millions of samples, for biological activity which can be done across various levels, from organisms to proteins.[5] A major limitation of traditional high-throughput screenings (HTS) lies in the size of the compounds composing the libraries (**Fig 1.2**). Typically, these compounds are drug-sized, with heavy atom counts ranging between 20 and 100 atoms. As the atom count increases, there is a combinatorial explosion of possible molecular combinations, reaching levels that are unattainable by experiments (**Fig 1.2**).[7] Consequently, screening large libraries are necessary, but these still cover relatively poorly the space of molecules with pharmacologic potential.[6] The pharmacological potential of small molecules is typically defined by Lipinski's Rule of 5, which stipulates that there should be no more than 5 hydrogen bond donors, no more than 10 hydrogen bond acceptors, a molecular mass less than 500 Daltons, and a calculated octanol-water partition coefficient (Clog P) that does not exceed 5.[8] However, screening of such large libraries yields a low hit-rate.[6] Smaller libraries tailored against specific systems are achievable, but they necessitate prior knowledge of the target, which is often not available in the context of drug discovery.

1.2. A fragment-based approach alternative to high-throughput screenings

The advent of high-throughput approaches underscored the significance of an intriguing drug discovery paradigm, in which the target, a biological entity, serves as a filter for the chemical space. This space can be defined as the collection of all potential pharmacologically active molecules. Given the vastness of the chemical space paired with target selectivity, strategies are needed to maximise the recovery of the active chemical space while enabling downstream experimental and synthetic efforts (**Fig 1.2**).[9]

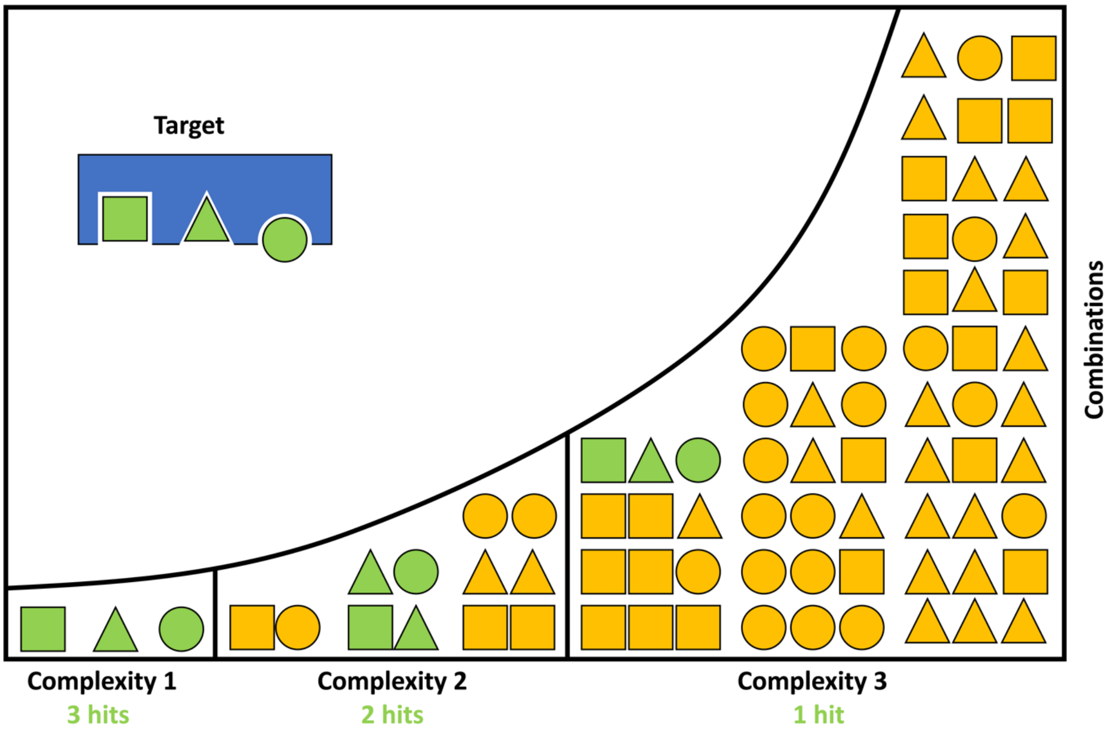


Figure 1.2: The impact of increasing molecular complexity on library combinatorial explosion and decreasing hit rate. The target, depicted in blue at the top left corner, features a binding site that accommodates three distinct binding modes. At a theoretical molecular complexity of 1, all building blocks can interact with the target. However, as complexity increases, fewer compounds can bind due to less frequent alignment with the binding site modes. Binding and non-binding building blocks/combinations are represented in green and yellow, respectively. This figure was adapted from Dr Yuliya Dubianok's doctoral thesis.

Fragment-based approaches present a conceptually elegant solution to this problem. Fragments are small molecules adhering to the Rule of 3, which states that their molecular weight should be less than 300, ClogP should be less than or equal to 3, and the number of rotatable bonds, hydrogen bond donors, and acceptors should each be less than equal to 3. As a result, fragments are smaller than drug-like molecules used in traditional high-throughput screenings. This implies a relatively smaller, and therefore more tractable chemical space to be sampled (**Fig 1.2**). These small-sized compounds tend to bind to the target at relatively low potencies, ranging from millimolar to micromolar concentrations.[9] However, fragments also exhibit high ligand efficiencies, which are determined by the strength of the interaction with the target divided by the number of heavy atoms composing the small molecule. Fragments tend to bind by forming a few localised, higher-quality and enthalpically driven interactions, such as hydrogen bonds, halogen bonds or π-π stackings.[10] Additionally, fragments display increased promiscuity compared to larger molecules, resulting in higher hit rates in identification experiments compared to high-throughput screening approaches (**Fig 1.2**).

In most instances, the targets are proteins identified as key modulators of the phenotype to be corrected (**Fig 1.8**). Fragments are typically screened against a pure sample of the target obtained through a combination of recombinant expression and purification steps. Several techniques have been developed to categorise protein-ligand interactions and binding, which can be applied to the identification of fragments against a particular protein target. However, due to the intrinsically weak potency of fragments, these detection tools must be highly sensitive to the binding signal and exhibit a high signal-to-noise ratio (**Fig 1.7**).[9]

The selection of an appropriate set of fragments for screening, termed a fragment library, is a critical phase in the process. This selection ultimately delineates the span of the chemical space covered and identifies potential hits. Therefore, fragment libraries should be chemically and functionally in order to be capable of sampling a broad spectrum of interaction patterns available within binding sites.[11] In addition, other considerations may come into play, such as the inclusion of halogen atoms to facilitate density fitting,[12] or the presence of chemical groups that can be modified by common reactions.[13]

Once identified, fragments must be expanded into more potent and selective compounds that achieve the desired phenotypic modulation. This is normally accomplished by adding chemical groups to the initial hits to enhance interactions with the target binding site (**Fig 1.5**). Consequently, designing fragment libraries that facilitate subsequent synthetic steps is crucial. Once synthesised, fragment elaboration must be validated for improved potency. This combination of elaboration, synthesis, and validation seldom yields small molecules potent enough and often necessitates further iterations (**Fig 1.3**).[9]

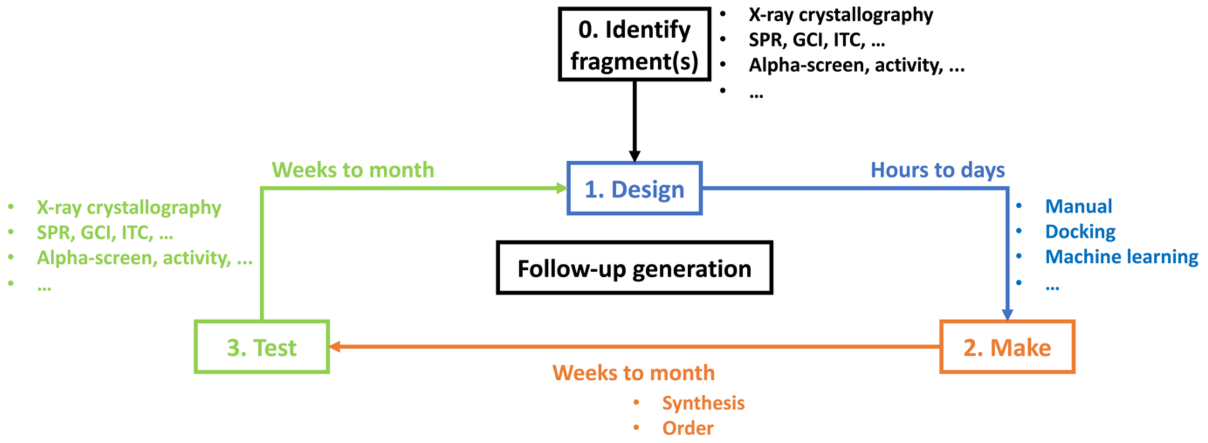


Figure 1.3: Fragment identification facilitates iterative follow-up design. Initial fragments are identified, serving as a basis for conceptual design through manual or computational methods. The chosen designs must then be acquired via synthesis or ordering, allowing validation through biophysical, structural, and biochemical approaches. Data from experimental validation informs subsequent design stages. The typical time range, along with exemplary experimental methods associated with each step, are highlighted with colours corresponding to cycle step.

Thus, fragment-based drug discovery is an iterative process (**Fig 1.3**) in which fragments serve as high-quality, low-potency seeds derived from a more constrained chemical ensemble (**Fig 1.2**). Fragment identification against a target enables local chemical space sampling of potentially active and chemically accessible follow-up molecules, with elaboration primarily aimed at increasing the compound's potency (**Fig 1.5**).

1.2.1. Fragment identification and synergies with structure-based approaches

Fragment-based drug discovery projects begin with the identification of fragment binders against the target through sensitive techniques. Various methods can be employed for this purpose, with most providing information about the interaction strength between the target and the compound.[9]

Although such information is vital, protein-ligand binding is ultimately governed by the structural arrangement of the protein target (**Fig 1.9**), which dictates if and how compounds interact with the binding site. Consequently, understanding protein and binding site structures is essential for making informed decisions about fragment elaboration (**Fig 1.9**).

X-ray crystallography has emerged as one of the leading techniques in this context, as it provides atomic-scale structural information that can support drug discovery by rationalising ligand binding and support optimisation decisions.[14] The first protein structure was resolved in 1958 with the discovery of the myoglobin structure (**Fig 1.1**).[15] Initially, protein crystallography was an extremely laborious process, unable to keep pace with drug discovery efforts. However, methodological and technological improvements led to the first structure-based drug discovery campaigns from the late 1970s (**Fig 1.1**). Notable examples include the rational design of hemoglobin ligands,[16] which sparked significant interest from drug discovery companies. Nowadays, crystallography and structure-based drug discovery are integral parts of the pharmaceutical landscape.[17]

The primary advantage of X-ray crystallography over alternative identification methods is its ability to provide high-resolution structures that precisely resolve side-chain orientations, water molecule positions, and overall protein conformation. Additionally, it directly informs about the position and conformation of the compound, ensuring that it binds to the site of interest, thus providing a three-dimensional template for compound optimisation (**Fig 1.9**).[14]

Recent advances in automation and synchrotron radiation have enabled large-scale screening, with hundreds of potential ligands being screened within hours, yielding bound structures at resolutions suitable for drug discovery.[18,19] The XChem is a leading platform in this regards as it provides integrated infrastructures, for crystal making, targeting,[20] soaking,[21] harvesting,[22] diffraction and structural model building,[23,24] enabling researchers to screen hundreds of small molecules within days (**Fig 1.4**).[25] X-ray crystallography is also label-free, preventing artefacts associated with immobilisation or tagging required by most other methods.

X-ray crystallography also has limitations inherent to the technique. Acquiring reproducible, high-quality, and resilient crystals can be a laborious and challenging task, primarily dependent on trial and error.[26] High quality structure production also require access to specialised beamlines at synchrotron which can be limited depending on the location of the investigating laboratory. Additionally, lattice effects might confine protein conformations to unphysiological states, which may confound subsequent structure-based drug efforts.[27] The resulting crystal structures do not directly convey binding affinity, necessitating the use of complementary approaches to fully rationalise binding.

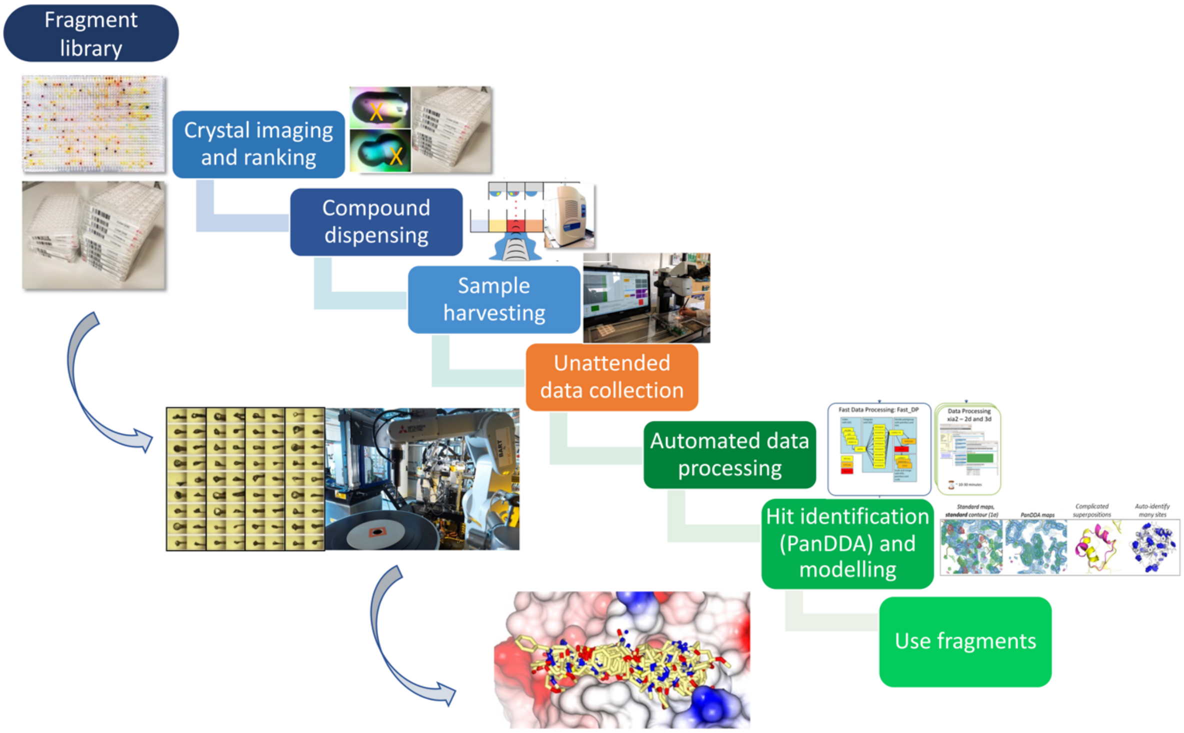

Fragment library
Crystal imaging and ranking
Compound dispensing
Sample harvesting
Unattended data collection
Automated data processing
Fast Data Processing: Fast_DP
Data Processing xia2 – 2d and 3d
~ 10-30 minutes
Hit identification (PanDDA) and modelling
Standard maps, standard contour (1σ)
PanDDA maps
Complicated superpositions
Auto-identify many sites
Use fragments

Figure 1.4: Fragment identification through X-ray crystallographic screening at XChem. The various steps involved in determining fragment-bound protein structures using high-throughput crystallography are outlined in the accompanying box. Initially, appropriate crystals are identified and subsequently soaked with fragment stocks using acoustic dispensing. The soaked crystals are harvested and exposed to specialised beamlines for data collection. Finally, the diffraction data is automatically processed, followed by model building and refinement of the co-crystal structure. This figure was adapted from Douangamath et al., (2021).

A further benefit of using crystals is their enhanced resistance to solvent exposure relative to proteins in solution, which enables compound evaluation at relatively more elevated concentrations directly from solvent stocks. These higher concentrations reveal weaker ligands more efficiently by saturating the binding signal, which is particularly important in fragment-based drug discovery due to the low potencies of fragments.[28]

However, when weak fragments are resolved crystallographically due to unphysiologically high soaking concentrations, the resulting hit, pose or interactions may be artificially present in the crystal, yet fail to bind in solution, thus yielding no signal in assays. Consequently, orthogonal strategies are required to validate these fragment hits. While these often involve solution assays, they come with the drawback of necessitating further experiments. Hotspot mapping provides a convenient alternative, offering increased confidence in the resolved fragments.[29] It use 3D structures to identify pharmacophoric regions, and an overlap between hotspots and fragment pose suggests that the crystallographic interactions are not coincidental, making them more likely to result in binding during assays.

Furthermore, the high-throughput capabilities of X-ray crystallography facilitates the analysis of vast amounts of data, which can collectively be used to unmask even weaker binding events (**Fig 1.4**).[24] Thus, fragment-based approaches pair well with crystallography, due to its scalability and the ability to expose target protein crystals against high fragment concentrations.

Obtaining and optimising protein crystals for drug discovery purposes can be challenging. They must be ligand tolerant, sufficiently reproducible, and capable of diffracting at a high resolution. Cryo-electron microscopy (EM) is becoming an increasingly prevalent alternative to X-ray crystallography for structural elucidation of protein-fragment complexes. Cryo-EM involves blotting a sample onto a grid, cryo-cooling it, and using electrons for scattering, from which the electron density is inferred for structural model building. Consequently, cryo-EM does not necessitate crystal formation, potentially offering an alternative to X-ray crystallography in terms of sample preparation, potentially enhancing throughput, and reducing associated labour.[30] While some fragment studies have been conducted using cryo-EM, these often suffer from comparatively lower resolution than methods relying on crystallography.[31]

1.2.2. Strategies in fragment follow-up design

Fragment libraries are screened against protein crystals, yielding fragment-bound structures that serve as starting points for synthetic elaboration and chemical space sampling (**Fig 1.4**). High-resolution structures provide templates for rational compound elaboration strategies aiming at increasing potency by targeting additional interactions with the binding site. Designing and selecting fragment follow-up compounds is a crucial step in developing more potent and specific small molecules against a target.[28]

Three types of fragment elaboration exist: growing, linking, and merging (**Fig 1.5**). Growing involves expanding a single fragment with chemical groups at appropriate elaboration vectors to establish new interactions with the target binding site (**Fig 1.5**).[28]Growing is a more tractable approach since using a single fragment scaffold places fewer restrictions on the chemistry. However, it has the disadvantage of focusing on a single hit, thereby neglecting the remaining experimental information generated from the initial screening experiment. Linking and merging, on the other hand, involve combining two or more fragments into a single molecule (**Fig 1.5**). In linking, fragments are connected through a linker, whereas in merging, fragments are fused together. These approaches are more synthetically challenging. A merged scaffold may be non-obvious, and engineering a linker group could distort the energy landscape associated with the binding poses.[28] Nevertheless, linking and merging make better use of the available experimental information, potentially leading to more effective small molecules against the target protein.

There has been significant success associated with fragment-based drug discovery with half a dozen drugs approved by the FDA and about 40 other candidates undergoing trials.[32] Significant examples include Venetoclax, a selective BCL2 inhibitors used against leukaemia was discovered through fragment linking. Vemurafenib was the first FDA approved drug (**Fig. 1.1**) discovered from fragment-based methods, initially optimised identified through a growing strategy, and targets mutant BRAF kinase involved in melanoma. No fragment mergers have, to the best of the available knowledge been approved by the FDA but many success stories include the inhibition of targets such as PTPN1 or PDK1.[33]

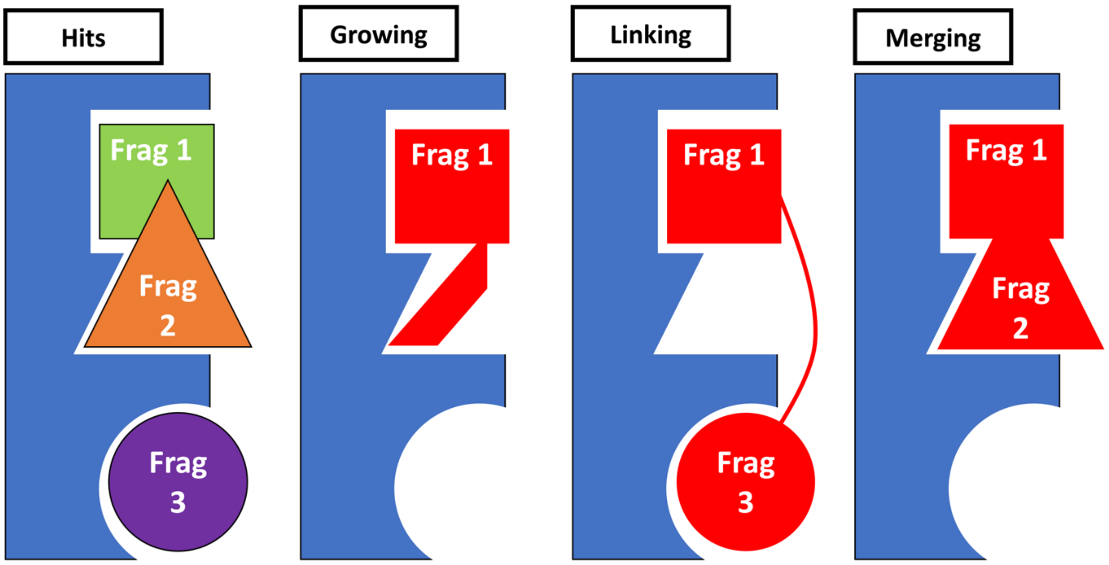


Figure 1.5: Fragment elaboration strategies in follow-up design. Upon identifying fragment hits, they can be developed through growing, linking, or merging. Growing involves adding chemical groups to engage new binding site interactions, linking entails designing chemical linkers to connect two or more fragment hits, and merging involves fusing overlapping fragment hit scaffolds.

The use of *in silico* methods is now a ubiquitous part of drug discovery efforts as they offer rapid and objective ways of generating and selecting compounds. *In silico* methods also pair well with fragment-based approach where both chemical and structural information can be exploited in follow-up compound design.[34]

A hierarchical approach utilises computationally inexpensive methods for processing a high volume of compounds, gradually shifting to more resource-intensive techniques for smaller subsets, ultimately aiming to enhance predictions quality for compounds selection (**Fig 1.6**).[35] Ligand-based approaches are on the cheaper end and can serve to build a virtual library of compounds. Single or multiple fragments can be employed to filter libraries of chemically accessible compounds, with the goal of identifying follow-up molecules that retain features observed experimentally. This can be achieved through straightforward chemical similarity and substructure analyses or more intricate multi-fragment pharmacophoric searches (**Fig 1.6**).[36]

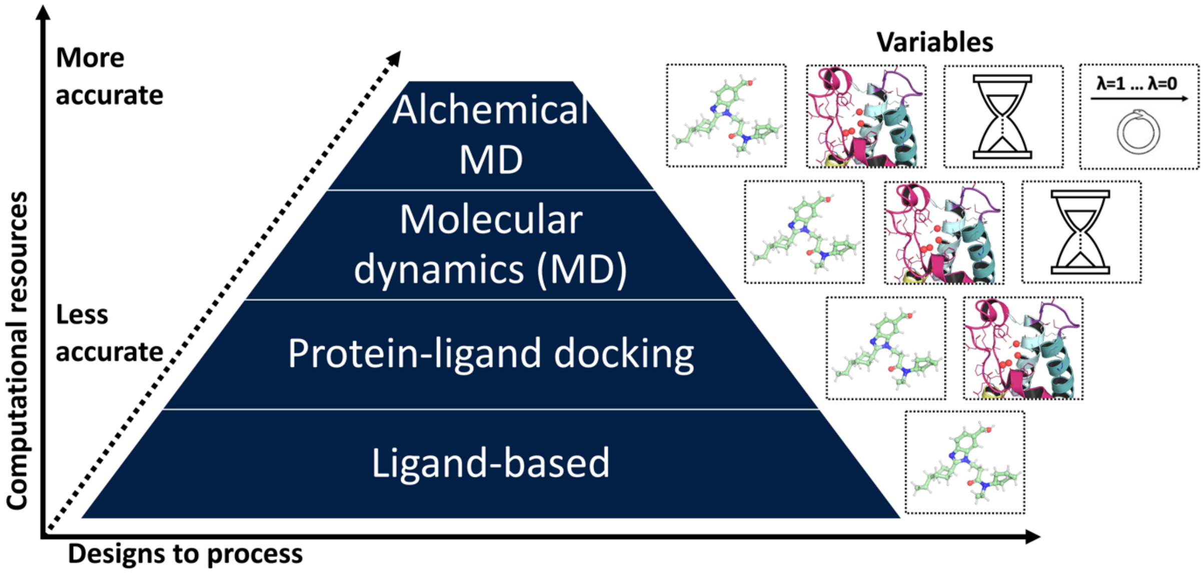


Figure 1.6: Hierarchical application of computational methods in follow-up design. Less demanding methods, which generally provide less accuracy due to an incomplete system evaluation, are employed to process a larger number of compounds. More demanding methods handle fewer compounds but offer greater accuracy because of a more comprehensive system description. Symbols on the right signify the description levels provided by computational methods, with the hourglass indicating ligand and receptor simulations over time, and the Ouroboros symbol representing multiple alchemical windows.

Structural information can then be utilised to score a library by predicting ligand binding poses against the binding site, from which binding affinity can be estimated (**Fig 1.6**). This process, known as docking, can be carried out using various approaches that are generally driven by physics-based modelling or machine learning.[35] Rescoring tools can also be employed to select docking poses that overlap with experimentally resolved fragments or recapitulate a specific interaction motif with the binding site.[37] Docking methods offer the advantage of being relatively inexpensive, but their ability to account for conformational motions and water dynamics is limited (**Fig 1.6**). This can lead to inaccurate predictions when these motions impact binding, such as when the low-energy protein solution state deviates from the crystal structure.[38]

Molecular dynamic simulations can account for such motions by estimating the temporal behaviour of protein-ligand complexes in solution. These simulations can be initiated using bound conformations generated from docking, thus providing more detailed information about the quality of the interaction between the target binding site and candidate follow-up (**Fig 1.6**).[39] However, molecular dynamic simulations are significantly more computationally expensive than docking predictions and, as a result, require more specialised computer infrastructure.[35] These simulations can also be used to estimate binding free energies for protein-ligand complexes. Less rigorous but more affordable approaches, such as Molecular Mechanics Generalised Born Surface Area (MMGBSA) and Interaction Entropy (IE), can employ single trajectories.[40] In contrast, alchemical binding free energies are more expensive, as they necessitate simulating non-physical intermediates across multiple trajectories, but generally have better predictive power (**Fig 1.6**).[39]

Thus, fragment follow-up enumerations and selection can be achieved through various means, ranging from human decision-making to more computationally expensive physics-based modelling. Ultimately, the factors governing fragment elaboration include the initial data, access to computational resources, and expertise.

1.2.3. Large scale chemical synthesis as a bottleneck

Follow-up design can be a relatively swift process, provided that all necessary infrastructure and software are already in place (**Fig 1.3**). Once suitable follow-up candidates are identified, they must be synthesised for experimental validation. This step can be the rate-limiting factor of the cycle, as synthesis demands considerable labour and resource investment for specific molecules (**Fig 1.3**). Consequently, it is crucial for virtual screening libraries to consist of easily obtainable molecules.

Following synthesis, the follow-up products must be purified from the reaction mixtures, often via chromatographic methods, which is another laborious and wasteful process. While follow-ups can sometimes be ordered directly from vendors, reducing labour, they often incur higher costs and lengthier shipping times, further delaying experimental validation.[41] Thus, the ability to perform rapid and cost-effective chemistry to generate fragment follow-ups in the vicinity of experimental validation infrastructure is a significant consideration. However, such infrastructures often necessitate advanced robotic systems for synthesis, purification, and quality control, which can be prohibitively expensive for academic laboratories.[42] Nonetheless, there has considerable methodological improvements in more affordable high-throughput synthesis methods, with the range of achievable reactions also expanding. These progresses enables better chemical space sampling around fragments.[42]

As the number of synthesised compounds increases, so too does the number of required purifications, as assays are performed using pure products, thus further slowing down the follow-up validation process. Recently, it was demonstrated that X-ray crystallography could be conducted by soaking crude reaction mixtures directly onto protein crystals, thereby bypassing purification.[43] However, this method was applied to fewer than 100 ligands. Similarly, dissociation rate (k_d) measurements can be screened from crude reaction mixtures using surface plasmon resonance (SPR), which are concentration-independent thereby informing the binding strength between the ligand and the protein.[44] However, this approach has not yet been expanded to high-throughput levels. Consequently, current validation approaches struggle to keep pace with large follow-up libraries generated via automated and purification-free synthetic approaches.

1.2.4. Experimental validation of fragment follow-up compounds

Synthesising follow-up designing enables downstream validation using experimental techniques. In most instances, validation necessitates testing pure follow-up stocks against the target, which is crucial for assessing if and how designs achieve inhibition of the target, therefore moving towards phenotypic modulation (**Fig 1.7**). A comprehensive understanding of the follow-up's interaction with the target demands validation from three perspectives: structural, biophysical, and biochemical (**Fig 1.7**).[45]

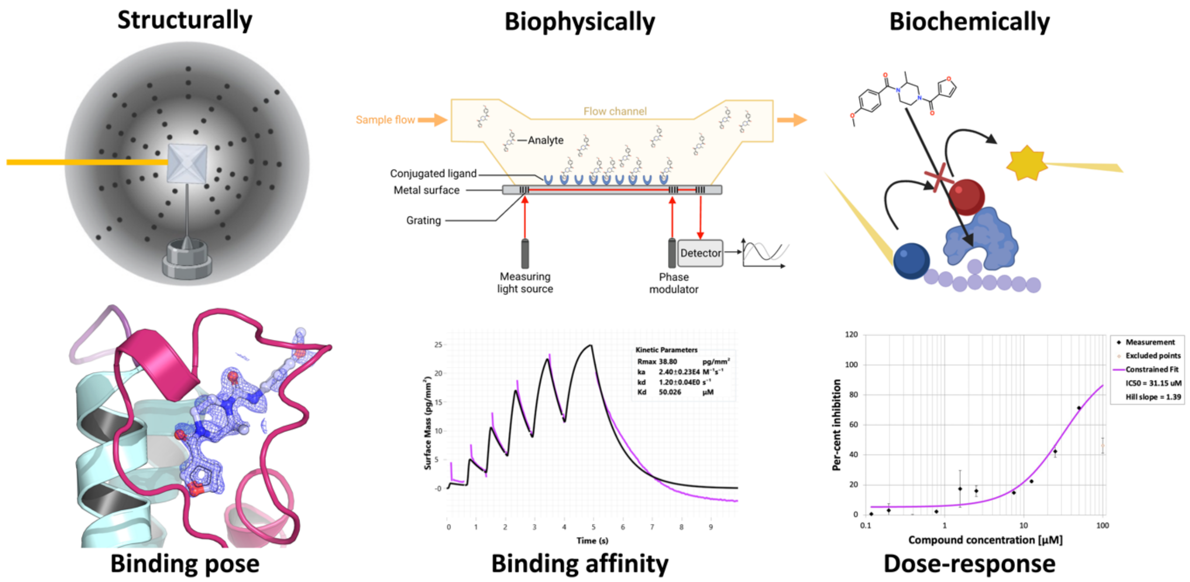


Figure 1.7: Follow-up validation requires a 3-fold approach. Follow-up designs need structural validation, typically using X-ray crystallography, to determine the protein-ligand binding pose. Biophysical validation estimates the interaction strength between the protein and ligand through affinity measurements. Biochemical validation assesses the compound's capacity to modulate the target's biochemical function. Method illustrations were made using available BioRender templates.

Structural validation assists in determining if the fragment elaboration preserved the initial binding pose and provides information about the interactions made, protein conformational motions, and water displacement, which contribute to the resulting binding affinity through enthalpic, entropic, or kinetic changes (**Fig 1.7**). Alternative methods to X-ray crystallography, such as Nuclear Magnetic Resonance Spectroscopy (NMR) spectroscopy or cryo-Electron Microscopy, can also be used for this purpose, although they may be less scalable or have lower resolution.[45]

The second aspect of the validation process is probing the strength of the interaction between the follow-up and the target. This is typically measured with biophysical assays in a solution that mimics pseudo-physiological salt and pH conditions. Numerous tools have been developed for this purpose, generally falling into two categories. Kinetic methods measure the association and dissociation rates between the ligand and target, allowing for the inference of equilibrium constants and informing binding affinity (**Fig 1.7**). A variety of tools for kinetic measurements exist, with the choice of validation assay often dependent on expertise and device availability.[45] Nevertheless, robotic methods have been developed for Surface Plasmon Resonance (SPR), Nuclear Magnetic Resonance (NMR) spectroscopy, and Grating-Coupling Interferometry (GCI), facilitating larger-scale measurements. Thermodynamic methods, such as Isothermal Titration Calorimetry (ITC), enable the deduction of entropic and enthalpic contributions to binding but tend to require large amounts of protein for testing.[46]

Ultimately, follow-ups must be biochemically validated to ensure they achieve the desired effect on the target's function. For example, when addressing a protein-protein interaction facilitator, it is essential to confirm that the follow-up compound effectively disrupts the interaction responsible for the phenotypic effect (**Fig 1.7**).[47] When working with pure protein samples, this is frequently accomplished via optical assays. These assays detect and quantify light in the form of absorbance, fluorescence, or luminescence. The resulting signal is generally converted to half-maximal inhibitory concentration (IC_{50}) values, which represent the concentration of the compound required for 50% inhibition of the targeted biochemical function (**Fig 1.7**).[48]

In drug discovery, validating compounds while considering relevant pharmacodynamic properties that yield a good drug is essential. For instance, designing compounds with fast association constants (K_a) and slow dissociation constants (K_d) enables rapid absorption and extended release, providing a swift and lasting therapeutic effect.[49] Examining and validating multiple compounds is vital for understanding structure-activity relationships (SAR) in drug discovery. SAR analysis offers insights into how specific chemical modifications affect the potency, selectivity, and pharmacodynamic properties of compounds. This valuable information can be applied in future iterations of compound design, guiding the optimisation process to produce more effective therapeutic candidates. However, the analysis and application of SAR data is non-trivial, necessitating the use of cheminformatics and/ or machine learning methods. This is especially true in modern drug discovery projects, where large numbers of compounds are processed, yielding equally extensive datasets that are unanalysable manually. As methods generate increasingly complex datasets, it is crucial to continue the development of such approaches to support drug discovery efforts.[50]

1.3. PHIP(2): a relevant target for fragment-based drug discovery

In this work, the Pleckstrin-homology domain interacting protein (PHIP), specifically its second bromodomain (PHIP(2)), will be investigated. Initially, PHIP was identified as interacting with the Pleckstrin homology (PH) domain of insulin receptor substrate-1 (IRS-1), a tyrosine kinase involved in insulin-mediated processes such glucose transport (**Fig 1.8**).[51] Later, a larger PHIP isoform was discovered in the nuclei of pancreatic beta-cells, where it positively regulates cell growth and survival.[52] PHIP deficiency in young mice results in delayed body growth and anaemia, emphasising its important physiological function.[53] These findings link PHIP to the insulin signalling pathway and suggest its involvement in tumorigenesis (**Fig 1.8**). Studies confirmed that increased PHIP copy number positively regulates metastasis in melanoma tumours lacking mutations in the three most common oncogenic genes.[54]

This revelation opened new avenues for therapies targeting BRAF-negative melanomas, which constitute a significant proportion of all human melanomas and lack effective treatments. Further support for this strategy emerged when research showed that suppressing PHIP inhibits the proliferation and invasion of "driver-negative" melanoma, breast, and lung tumours.[55] PHIP also enhances tumour cell mobility in glioblastoma cancer cells by acting on the focal adhesion complex, a key regulator of actin cytoskeleton organisation and dynamics (**Fig 1.8**).[56] However, the precise molecular and structural mechanisms by which this multi-domain protein operates remain unclear, despite growing evidence of its significant role in lethal cancers (**Fig 1.9**).

PHIP is a multifunctional protein with multiple subcellular locations, comprising eight WD40 repeats and two bromodomains, PHIP(1) and PHIP(2) (**Fig 1.9**). WD40 repeat domains participate in various cellular processes through molecular recognition such as protein-protein or protein-DNA interactions.[57]

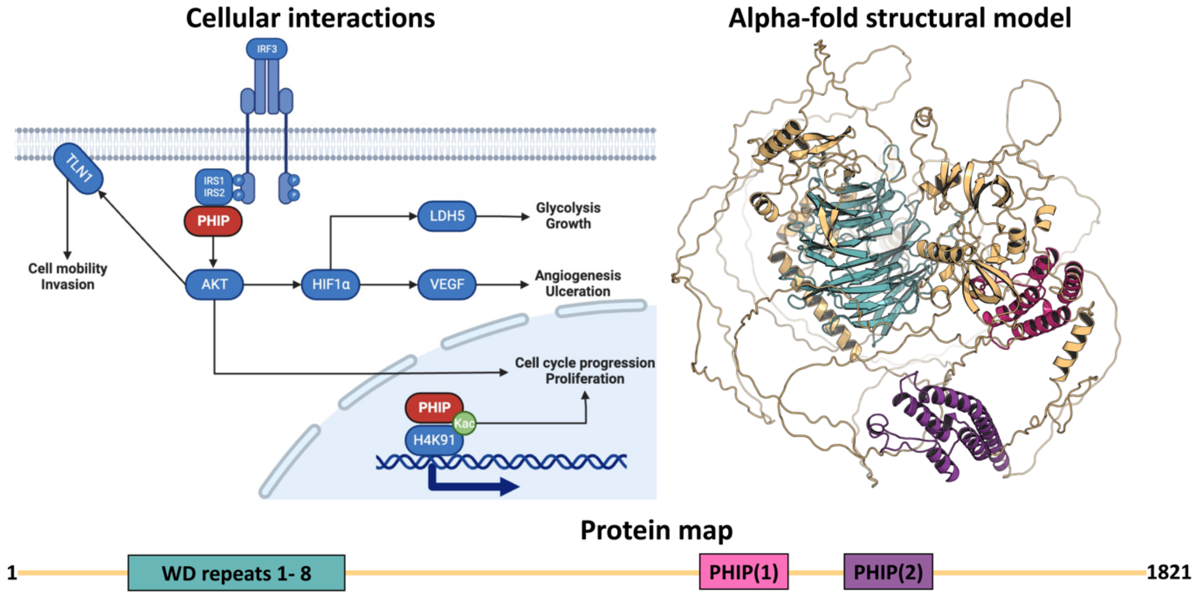


Figure 1.8: PHIP is a multidomain protein that regulates various oncologic cellular pathways. The left panel illustrates PHIP's involvement in cellular pathways and their subcellular locations. In the cytoplasm, PHIP interacts with the insulin signalling pathway, potentially contributing to cancer invasion, growth, and ulceration. In the nucleus, PHIP binds to histones, which may result in cancer cell proliferation. Structurally, PHIP comprises eight consecutive WD repeats, two bromodomains, and substantial unstructured regions. Cellular pathways were illustrated using available BioRender templates. The cellular pathways were adapted from Kashani-Sabet et al., (2018).

Bromodomains are typically involved in transcriptional regulation or chromatin remodelling and feature a conserved α-helical bundle fold, including four α-helices (αZ, αA, αB, and αC) connected by flexible loops defined as the ZA, AB, and BC loops (**Fig 1.9**).[58] Bromodomains possess a conserved network of four water molecules at the core of the fold, facilitating the binding of an acetylated-lysine (**Fig 1.9**). These post-translationally modified amino acids are commonly found on histone tails, where they serve as epigenetic markers (**Fig 1.8**). Bromodomains act as specific "readers," with acetylated-lysines often being central to more complex peptide interactions, where neighbouring residues can also bear other post-translational modifications like methylation or phosphorylation.[59] Through this interaction, bromodomains recruit other factors necessary for cellular function. Pathologically, bromodomain-containing proteins are implicated in numerous cancer types, but their well-defined acetylated-lysine binding site's role as a key mediator of protein-protein interactions renders them attractive drug targets. Several bromodomain inhibitors have been reported, with ongoing clinical trials.[58] Moreover, chemical probe development offers an alternative to extensive laboratory work, such as engineered animal models, required to understand bromodomain-containing proteins' cellular functions.

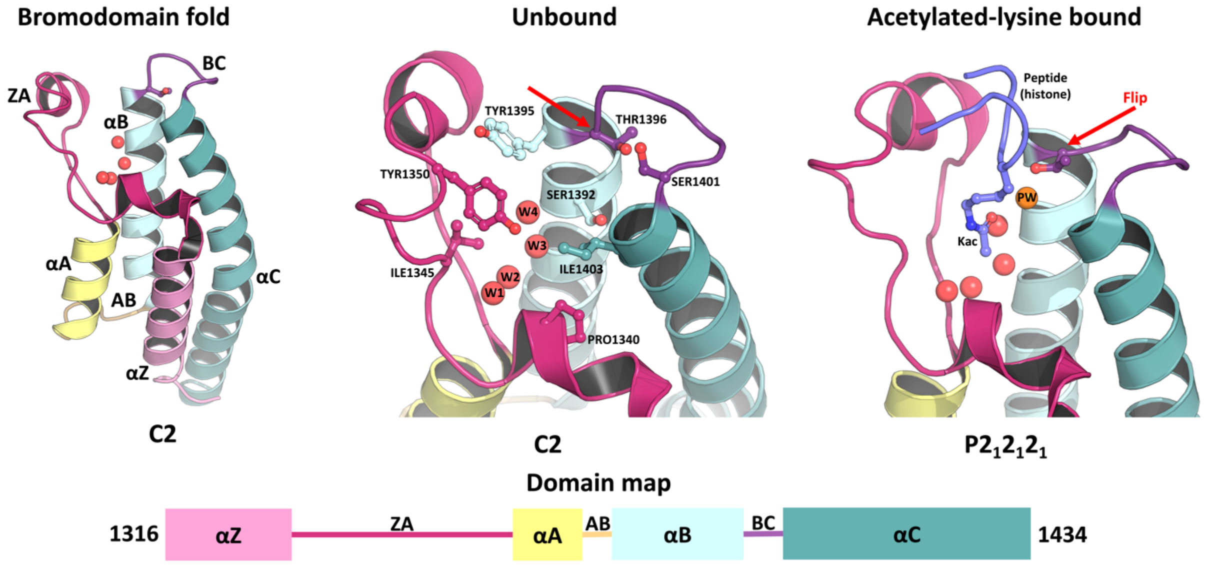


Figure 1.9: PHIP(2) exhibits a typical bromodomain fold with an atypical acetylated-lysine binding site. The left panel displays the characteristic bromodomain fold, with helical and loop regions coloured according to the domain map shown in the bottom panel. The central panel illustrates the pharmacologically relevant acetylated lysine binding site, featuring the conserved bromodomain 4-water molecule network (depicted as red spheres) and side chains involved in ligand binding (displayed as sticks). The right panel presents the structure of PHIP(2) bound to an acetylated-lysine containing peptide, mimicking histone binding. Peptide binding is associated with the introduction of an additional water molecule (shown in orange, labelled PW) and conformational changes of the non-conserved threonine that is typically an asparagine in other bromodomains.

A small molecule probe specific against PHIP could open opportunities for novel and broad-based chemotherapy against non-targetable tumours and/or facilitate understanding the biology underlying these cancers. Computational, biochemical, and proteomic data indicate that PHIP2 binds to acetylated lysine 91 on histone 4 (H4K91ac), supporting the notion that PHIP(2) functions as a histone reader in the context of epigenetics (**Fig 1.8**).[55] PHIP(2) belongs to the third bromodomain family and features an atypical acetylated-lysine binding site due to the highly conserved asparagine (**Fig 1.9**), between the αB helix and the BC-loop, being replaced by a threonine. This substitution is observed in only 21% of known human bromodomains.[13] This special feature of PHIP(2) contributes to its distinct binding properties and may offer opportunities for the development of selective inhibitors, potentially aiding in the advancement of novel cancer therapies.

1.4. Advancing fragment-based methods with PHIP(2)

Previous *in-house* studies have employed PHIP(2) as a model in the development of fragment-based approaches. For instance, the DSI-poised library is a collection of fragments enabling synthetic elaborations by displaying chemical groups targetable via general synthetic reactions, enabling rapid chemistry.[13] The years of research preceding this work laid solid foundations for further methodological and applied research. An effective PHIP(2) expression system, purification method, and multiple crystal forms obtained at the Structural Genomic Consortium provides excellent material for subsequent crystallographic screening and assay studies.[26]

Consequently, this work focuses on high-throughput crystallographic and computational studies of fragment identification and follow-up compounds design and evaluation. Fragment-based methods hold considerable potential for the discovery of new drugs and chemical probes. However, fragment-based ligand discovery is a rapidly expanding field with new challenges remaining to be addressed.

Accordingly, this thesis will investigate fragment-based methods using PHIP(2) as a model system. To facilitate the understanding of subsequent results, methodological concepts relevant to this work will first be introduced in Chapter 2. Crystallographic screening remains a laborious and expensive process, while computational methods are becoming increasingly precise. Therefore, Chapter 3 will assess the ability of computational methods to replicate crystallographic fragment screening data through a community exercise. Follow-up purification represents a bottleneck in fragment-based drug discovery, while protein crystals can tolerate soaking of crude reaction mixtures. As such, Chapter 4 will examine the chemical and structural data resulting from the crystallographic evaluation of crude reaction mixtures generated from robot synthesis to evaluate the method's viability. High-throughput methodologies produce large and complex datasets that can be challenging to understand manually. In Chapter 5, a computational method will be developed to rationalise binding events from the extensive crystallographic evaluation of crude reaction mixtures performed in the previous chapter. Fragments can be used to generate libraries of follow-up compounds, but these must be effectively prioritised for experimental validation. Chapter 6 demonstrates how the PHIP(2) fragments, resolved in Chapter 3, can be used to generate a library of fragment elaborations, which was subsequently filtered with a range of computational methods in an effort to identify more potent binders. Finally, the main lessons and conclusions drawn from this work will be summarised.

Chapter 2. Concepts and experiments

2.1. Experimental concepts

2.1.1. Basic principles of protein-ligand binding

Protein-ligand binding plays a pivotal role in biochemical processes and has emerged as a central focus in drug discovery. This binding process is typically driven by non-covalent interactions, which can be classified into polar and non-polar types. Polar interactions, such as hydrogen bonding and other electrostatic interactions, are localised and specific, taking place between molecules with partially charged regions in proximity (**Fig 2.1**). These interactions facilitate the precise positioning and orientation of the ligand within the protein's binding site. In contrast, non-polar interactions, including van der Waals and hydrophobic interactions, are less specific and directional than polar interactions (**Fig 2.1**). These interactions take place between non-polar regions of the protein and ligand, which prefer to associate with each other to evade water molecules in the surrounding aqueous environment (**Fig 2.1**).[60]

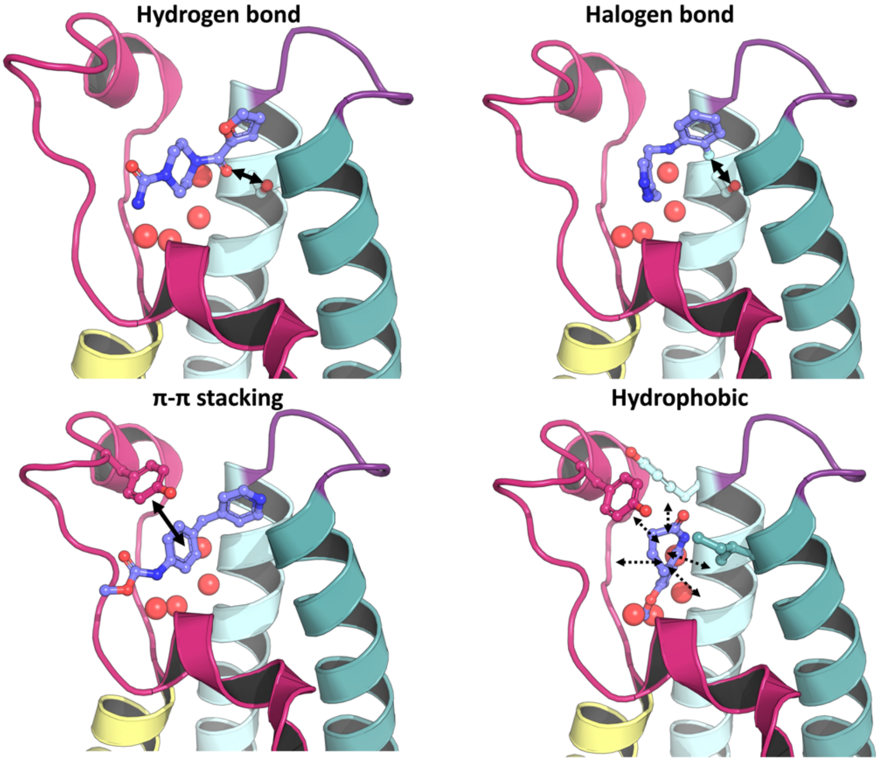


Figure 2.1: Typical molecular interactions observed in protein-ligand binding. Examples from *in-house* data illustrates hydrogen bonding, halogen bonding, π-π stacking, and less specific hydrophobic interactions. The presence of water molecules in hydrophobic interactions highlights that hydrophobic forces arise from the non-specific aggregation of molecules repelled by water's polar forces.

The dynamic nature of protein-ligand binding is dictated by the various conformations that both protein and ligand adopt across the binding event. A binding event is steered by a lower energy bound conformation being favoured over unbound states of the protein and ligand (**Fig 2.2**). Upon ligand binding to a target receptor, both may undergo conformational changes to form a stable complex (**Fig 2.2**). This can involve induced fit, where the protein alter its conformation to accommodate the ligand, or conformational selection, where the protein and ligand sample different conformations in the unbound states, and the binding event takes place when the most suitable conformations are present.[61] Eventually, an equilibrium is established between the bound and unbound states. Favourable interactions and energetics in the protein-ligand complex shift the equilibrium towards the bound state, facilitating complex formation, whereas unfavourable interactions shift the equilibrium towards the unbound states (**Fig 2.2**).[61]

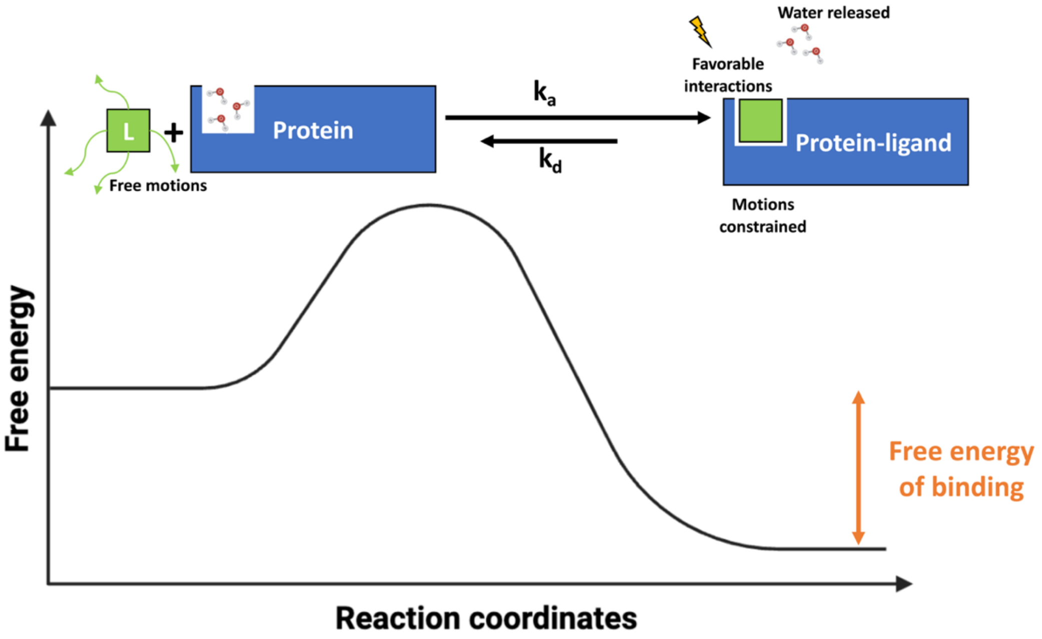


Figure 2.2: Favourable protein-ligand interactions promote complex formation over free ligand and protein, resulting in energy release. The top panel illustrates how ligand binding can restrict molecular motions and release water molecules from the binding site to the solvent. This favourable process is associated with a negative change in the binding free energy of the system.

Overall, the non-covalent and reversible association between a protein and a ligand can be described as followed:

Protein-ligand binding $[P] + [L] \Leftrightarrow [PL]$ **Equation 2.1**

Where [P], [L] and [PL] indicate the concentration for the free protein, free ligand and protein-ligand complex, respectively. Protein-ligand binding kinetics are described by the rate constants for association (k_a) and dissociation (k_d) which are indicative of how quickly the ligand bind and unbinds, respectively.

Kinetic association rates $$k_a[P][L] = k_d[PL]$$ **Equation 2.2**

At equilibrium the ratio of concentrations for the bound complex against the unbound protein and ligand against the bound complex defines the association constant (K_a) which is indicative of the interaction's strength.

Association constant $$K_a = \frac{[PL]}{[P][L]} = \frac{1}{K_d} = \frac{k_a}{k_d}$$ **Equation 2.3**

Conversely, the dissociation constant (K_d) is the ratio of concentrations for the unbound protein and ligand against the bound complex and represents as the ligand concentration required to occupy half of the available protein binding sites.

Dissociation constant $$K_d = \frac{[P][L]}{[PL]} = \frac{1}{K_a} = \frac{k_d}{k_a}$$ **Equation 2.4**

The concepts discussed above relate to the kinetics of the protein-ligand binding process. However, when characterising protein-ligand binding, it is also important to consider the energetics of the system. In particular, the thermodynamics of ligand binding can be described by changes in enthalpy (ΔH) and entropy (ΔS), which together define the Gibbs free energy (ΔG) of binding.

Gibbs free energy $$\Delta G = \Delta H - T\Delta S$$ **Equation 2.5**

Where T represents temperature. A negative ΔG suggests a favourable protein-ligand association process (**Fig 2.2**), while a positive ΔG implies an unfavourable binding process. The enthalpy change ΔH represents the energy released or absorbed during binding, while the entropy change ΔS signifies the alteration in molecular freedom upon binding (**Fig 2.2**). For instance, when a hydrogen bond forms between the protein and ligand, the enthalpy change ΔH will be negative, signifying that energy is released during bond formation. Simultaneously, the entropy change ΔS will likely be negative, indicating that the degree of freedom of the molecules decreases upon binding (**Fig 2.2**). The Gibbs free energy (ΔG) can also be related to the kinetic equilibrium constants, K_d or K_a.[61]

Gibbs free energy $\Delta G = -RT\ln(K_a) = RT\ln(K_d)$ **Equation 2.6**

Comprehending the kinetic and thermodynamic aspects of protein-ligand binding (**Fig 2.2**), along with the underlying interactions and structural information (**Fig 2.1**), is crucial for a thorough understanding of molecular recognition and advancing structure-based drug discovery strategies.

2.1.2. Protein-ligand crystallography

Protein crystallography is a technique used to determine the three-dimensional structure of proteins at atomic resolution, providing valuable insights into protein function and ligand binding. The process begins with obtaining a pure and stable protein sample through recombinant expression, where a gene encoding the protein of interest is introduced into a suitable expression system, such as bacteria (**Fig 2.2**). This approach facilitates the obtainment of large protein quantities, enabling downstream purification steps to isolate the target from cellular contaminants.[62]

The next step is to find suitable conditions for protein crystal formation, a critical step as the quality of the resulting crystals directly impacts diffraction data quality and the subsequent structural model (**Fig 2.2**). To achieve this, a wide range of conditions are often screened, varying parameters such as pH, temperature, and precipitant concentrations, identifying those that promote well-ordered protein crystal growth (**Fig 2.2**). These crystals should be large enough, have minimal defects, and possess high internal order to ensure high-quality X-ray diffraction data (**Fig 2.2**).[26] Ligand incorporation in the workflow helps understand whether and how it bind to the protein target. This can be achieved through co-crystallisation or soaking, where crystals are either formed in presence of the ligand in the crystallisation buffer or directly exposed from ligand stocks to the preformed crystal, respectively. Soaking is often more convenient and accessible by allowing rapid screening of many ligands onto preformed crystals.[26]

Once crystals are obtained, they undergo X-ray diffraction experiments at beamlines, using high-intensity X-ray sources to probe the crystal's internal structure and generate diffraction patterns (**Fig 2.2**). The collected diffraction data contain information on the reflection and intensities of scattered X-rays, used to deduce the protein's 3D structure (**Fig 2.2**).[26]

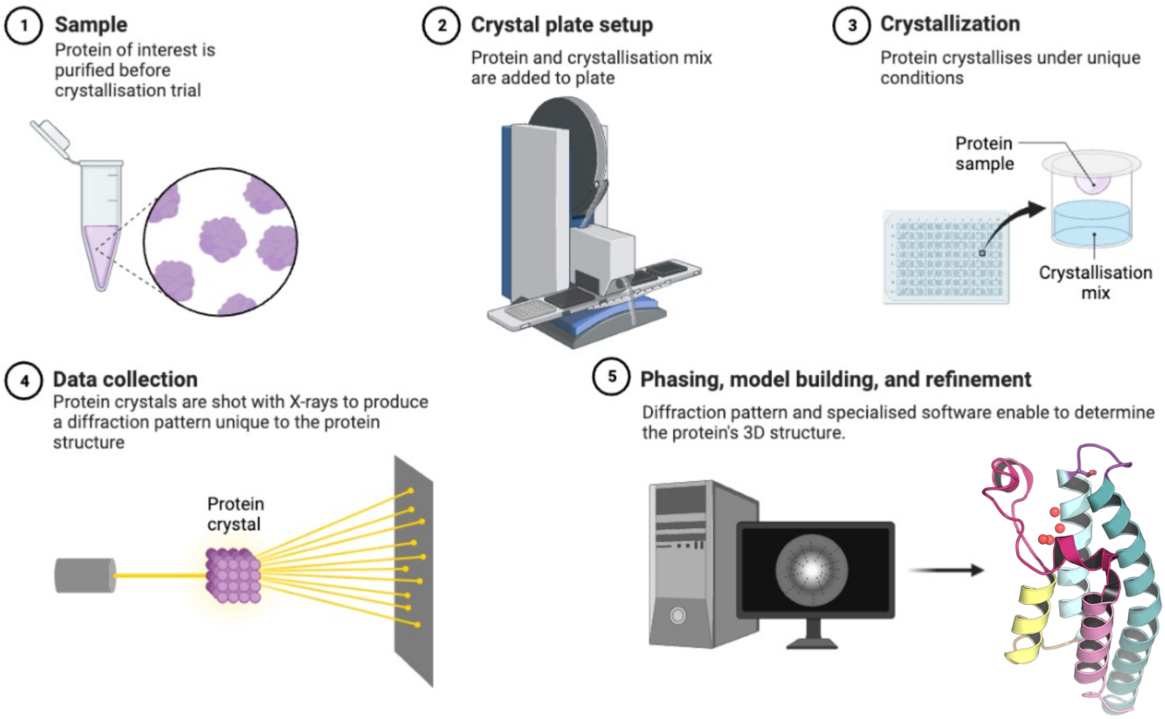

1 Sample
Protein of interest is purified before crystallisation trial
2 Crystal plate setup
Protein and crystallisation mix are added to plate
3 Crystallization
Protein crystallises under unique conditions
Protein sample
Crystallisation mix
4 Data collection
Protein crystals are shot with X-rays to produce a diffraction pattern unique to the protein structure
Protein crystal
5 Phasing, model building, and refinement
Diffraction pattern and specialised software enable to determine the protein's 3D structure.

Figure 2.2: Protein crystallography enables 3D structure determination. The crystallographic process comprises multiple steps across various disciplines. Firstly, pure protein samples are obtained and plated against a crystallisation mix, allowing crystal formation. The crystals are then exposed to high-intensity X-rays, producing diffraction patterns. To generate an electron density map, diffraction data must be processed to recover the missing phase information, which in turn provides material for model building and refinement. The figure was adapted from and available BioRender template.

Solving protein structures is not straightforward due to missing phase information in diffraction data, essential for determining electron density maps. One common approach to overcome this phase problem is molecular replacement, using a known homologous protein structure as a starting model to estimate phase information. This method involves a computational search for the best-fitting orientation and position of the starting model within the target protein crystal's unit cell.[63] Once phase information is obtained, the electron density map can be generated, and the protein structure, including the ligand, can be modelled. Fitting the protein's amino acid sequence into the electron density map and refining atomic positions to minimise discrepancies between observed and calculated diffraction data result in the final protein-ligand complex structure (**Fig 2.2**). This reveals detailed information about the protein conformation and ligand interactions.

Weak ligands, such as fragments, can lead to incomplete electron density as only a small fraction of proteins composing the lattice are in the bound state, making the interpretation of the electron density difficult or if not impossible (**Fig 2.3**). Pan-Dataset Density Analysis (PanDDA) addresses this issue by subtracting the ground state density, representing the unbound state, from individual crystal density maps, generating difference maps that can reveal weak ligand-binding regions (**Fig 2.3**).[24]

Although PanDDa is considered as useful tool in deconvoluting weak binding events, such fragments may also be the result of artefactual binding caused by high stock concentration or unstable conformation and may therefore not bind in solution assays. The correct treatment of structural heterogeneity and water molecules can also be challenging when relying on panda for model building. Indeed, PanDDa assumes structural homogeneity within the crystal and uses the ground state reference model for background correction.[64] However, protein crystals often present some form of structural heterogeneity, and such effect is even more prevalent when ligands are present due to conformational variation between the apo and bound states. Thus, the event map resulting from the subtraction of the ground state to a conformationally different state may generate ambiguous densities or water molecules at locations conflicting with protein or ligand conformations.[65]

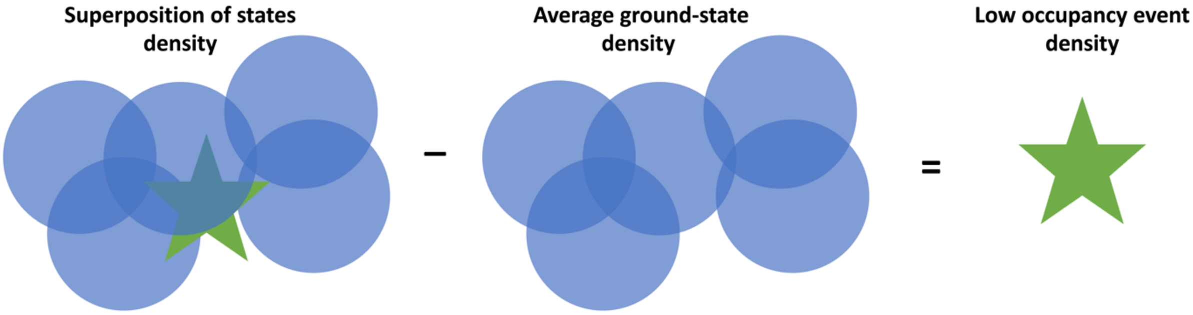

Superposition of states density
–
Average ground-state density
=
Low occupancy event density

Figure 2.3: Background correction uncovers low-occupancy crystallographic states from electron density maps. A conventional electron density map may conceal a low-occupancy state, represented by the green star. The background, equivalent to the average density across multiple unbound datasets, represented with blue circles, is subtracted from the target dataset, thereby revealing low-occupancy states for further processing and analysis. The figure was inspired from Pearse et al., (2017).

2.1.3. Grating-coupled interferometry measurement of binding kinetics

Grating-coupled interferometry (GCI) is an optical biosensing technique that enables the study of biomolecular interactions, such as protein-ligand binding, in real-time. GCI shares similarities with Surface Plasmon Resonance (SPR) as they both measure the accumulation of ligand on a sensor chip as a result of biomolecular binding. GCI differ from SPR in its underlying detection mechanism, as it relies on changes in the effective refractive index upon molecular binding rather than measuring alterations in the plasmon resonance angle that occur due to these interactions.

The fundamental principle of GCI is based on the detection of changes in the refractive index near the sensor surface, which occur as a result of molecular binding events (**Fig 2.4**). The GCI platform employs a nanostructured grating on the sensor chip to couple the incident light into surface-bound electromagnetic waves (**Fig 2.4**). As molecules bind to the sensor surface, the localised electromagnetic wave's phase and intensity are altered, generating an interferometric signal that can be quantitatively analysed to determine binding kinetics and affinity (**Fig 2.4**).[66]

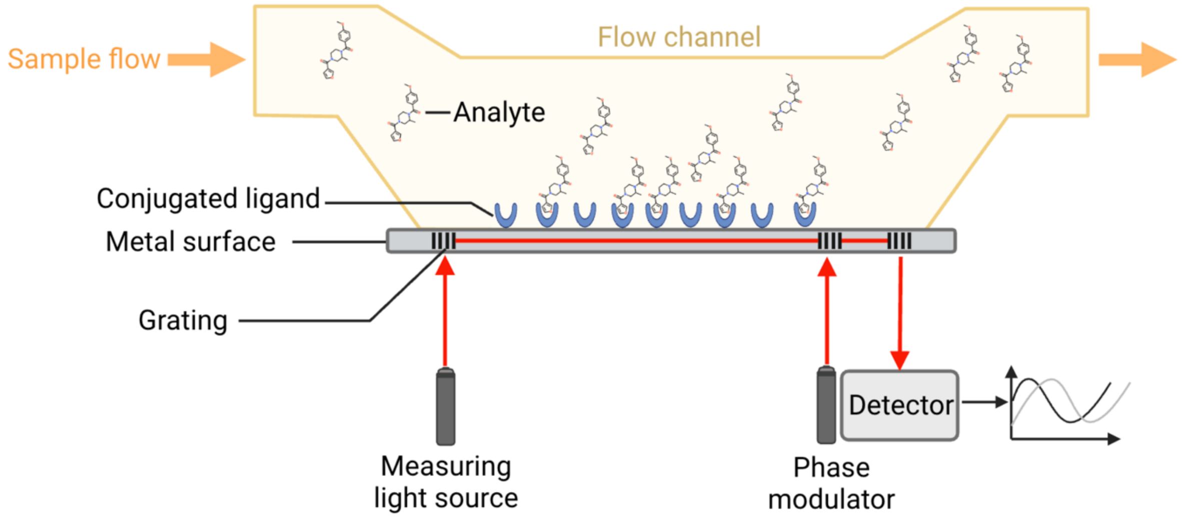


Figure 2.4: Grating-coupled interferometry setup for measuring protein-ligand binding kinetic in real time. The analyte (here a small molecule) is injected through the flow channel and associating with the immobilised protein receptor (here denotated as ligand). The association changes the light signal coupled with metal surface through grating technology. Another reference light source is coupled via grating after the initial light samples association thus modulating the phase from which light phase changes can be calculated allowing binding kinetic determination. The figure was adapted from an available BioRender template.

The first step in GCI experiments is immobilising the protein of interest onto the sensor chip. This can be achieved through amide coupling, which involves the formation of a covalent bond between the carboxyl groups on the sensor surface and the amine groups on the protein. The surface is typically activated with a mixture of N-hydroxysuccinimide and 1-ethyl-3-(3-dimethylaminopropyl) carbodiimide, which facilitates covalent coupling between the receptor and the sensor chip. The amount of protein immobilised on the sensor surface defines the maximum theoretical response, as it determines the maximum number of binding sites available for interaction with the analyte.[67]

Following protein immobilisation, the analyte is introduced to the sensor surface by controlled injection (**Fig 2.4**). The binding of the analyte to the immobilised protein causes a change in the local refractive index, which is detected by the GCI system (**Fig 2.5**). The association phase begins when the analyte is allowed to flow over the sensor surface, leading to interactions between the analyte and the immobilised protein (**Fig 2.5**). The GCI system continuously monitors changes in the refractive index, which are proportional to the amount of analyte bound to the protein (**Fig 2.4**).

After a predetermined period, the flow of analyte is stopped, and a buffer solution is introduced to wash away any unbound analyte (**Fig 2.5**). During the dissociation phase, the remaining bound analyte gradually dissociates from the immobilised protein, and the GCI system measures the changes in the refractive index (**Fig 2.4**).[68]

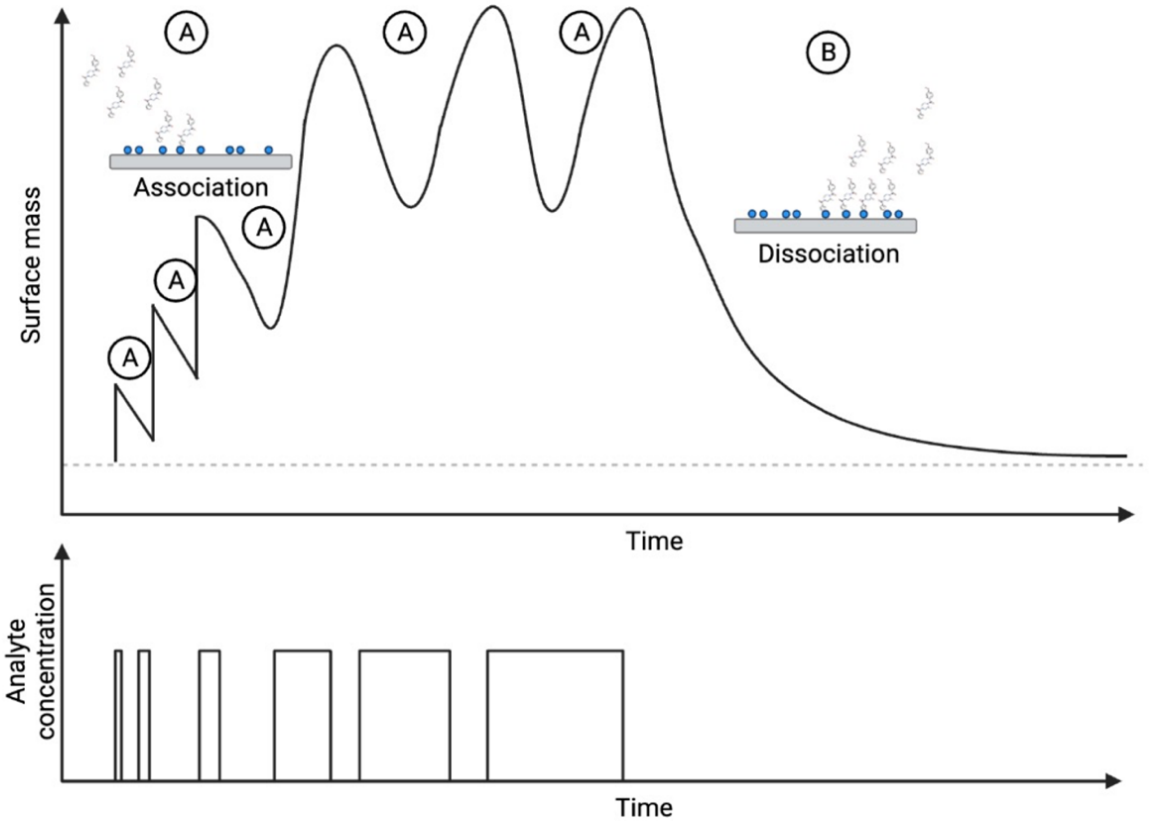


Figure 2.5: Grating-coupled interferometry assay allows binding kinetic measurement by varying single-concentration pulses. The association constant is determined from multiple analyte pulses of increasing duration (A). The dissociation constant is measured during a single phase, where the analyte departs from the target, resulting in a correlated decrease in surface mass (B). The figure was adapted from available BioRender templates.

The data analysis step involves fitting the association and dissociation curves obtained from the GCI measurements to a suitable mathematical model to extract kinetic parameters, such as the association rate constant (k_a), dissociation rate constant (k_d), and equilibrium dissociation constant (K_d) (**Fig 2.1**). The analysis uses specialised software that allows for the comparison of the obtained data with the blank measurements, providing valuable insights into the binding interaction and its characteristics.[68] The association rate constant (k_a) is concentration dependent and is inferred from the accumulation of mass onto the sensor chip upon injection of the ligand-containing solution. Reversibly, the dissociation rate constant (k_d) is concentration independent and is inferred from the depletion of mass from the sensor chip being caused by ligand unbinding (**Fig 2.5**).

2.2. General experimental methods

2.2.1. Protein expression, purification and crystallisation

BL21 cells containing a pNIC28-Bsa4 vector coding for PHIP2 were taken from a glycerol stock (kindly provided by Dr Tobias Krojer). 2 mL of Luria Broth pre-culture with 50 µM kanamycin were inoculated into 1 L Terrific Broth media with 2% glycerol (v/v), 0.01% (v/v) of 10% (v/v) sigma Antifoam 204 in ethanol, 50 µM $FeCl_3$, 20µM $CaCl_2$, 10 µM $MnCl_2$, 10 µM $ZnSO_4$ and 2 µM of $CoCl_2$, $CuCl_2$, $NiCl_2$, Na_2MoO_4, Na_2SeO_3 and H_3BO_3, 2 mM $CaCl_2$, 25 mM$(NH_4)_2SO_4$, 2.77mM glucose and 50 µM kanamycin. The cultures were grown for 6 h at 37 °C at 250 rpm. PHIP(2) expression was induced overnight at 18 °C with 0.1 mM IPTG. Cultures were centrifuged at 4000 g for 30 minutes at 4°C.

Pellets were resuspended in lysis buffer (10mM HEPES, 500mM NaCl, 5% glycerol, 0.5mM TCEP, 0.5 mg/mL Lysozyme, 1 µg/mL Benzonase, pH 7.5). The solution was vortexed and left at room temperature for 30 min before. 2% (v/v) triton-X- and 20-mM imidazole finale concentrations were added to the mixture before being centrifuged at 4000 g for 30 mins at 4°C. The supernatant was applied onto a 1 mL His GraviTrap columns (GE healthcare) fitted with a LabMate extender. The columns were washed twice with wash buffer (10 mM HEPES, 500 mM NaCl, 5% Glycerol, 0.5 mM TCEP, 20 mM imidazole, pH 7.5). The columns were slotted PD10 columns fitted with LabMate extenders. The proteins were eluted by applying 2.5 mL of elution buffer (10 mM HEPES, 500 mM NaCl, 5% glycerol (v/v), 0.5 mM TCEP, 500 mM Imidazole, pH 7.5) onto each GraviTrap column. 3.5 mL of wash buffer was applied onto each PD10 column and elutions were collected. 1 OD_{280} unit of TEV protease per PHIP(2) 10 OD_{280} units was added to the elutions and incubated at 4°C. The solutions were run back over His GraviTrap columns as mentioned above. The fractions were concentrated by 20-fold and applied onto a Yarra SEC 2000 pre-equilibrated with wash buffer. The fractions containing the protein were collected using either a biorad C-9 or a Cytiva ALIAS. The fractions were concentrated to about 15 mg/mL of protein and flash-frozen in liquid nitrogen.

PHIP(2) was crystallised in space group C2 at 4°C by vapour diffusion in 230 nL sitting drops, by mixing 100 nL protein in wash buffer with 100 nL reservoir buffer (20% PEG8000 and 40 mM potassium phosphate) and 30 nL seeds of the same composition than reservoir with final pH measured to be about 5.6.

2.2.2. Crystallographic screening of small molecules and model building

Small molecule crystallographic screening was performed at the XChem (Harwell, UK).[18] Crystals suitable for soaking were located in the plates with TexRank.[20] Suitable crystals were soaked with a 20% (v/v) ethylene glycol final concentration in the drops. Compound stock dispensing was performed with an ECHO acoustic liquid handler dispenser.[21] The crystals were incubated at 5 °C and harvested with a SHIFTER[22] before being plunged into liquid nitrogen and shot at the i04-1 or i03 beamlines located at the Diamond Light Source (Harwell, UK).

The XChemXplorer[23] was used for crystallographic workflow management and paralleling. Molecular replacements and initial refinements were performed with DIMPLE.[69] Pandda[24] was used to identify low occupancy binding events. Ligands were fitted in Coot[70] and the structures refined with Buster[71] and/or Refmac[72] before deposition on the protein data bank (PDB).

2.2.3. Grating-coupled interferometry assay

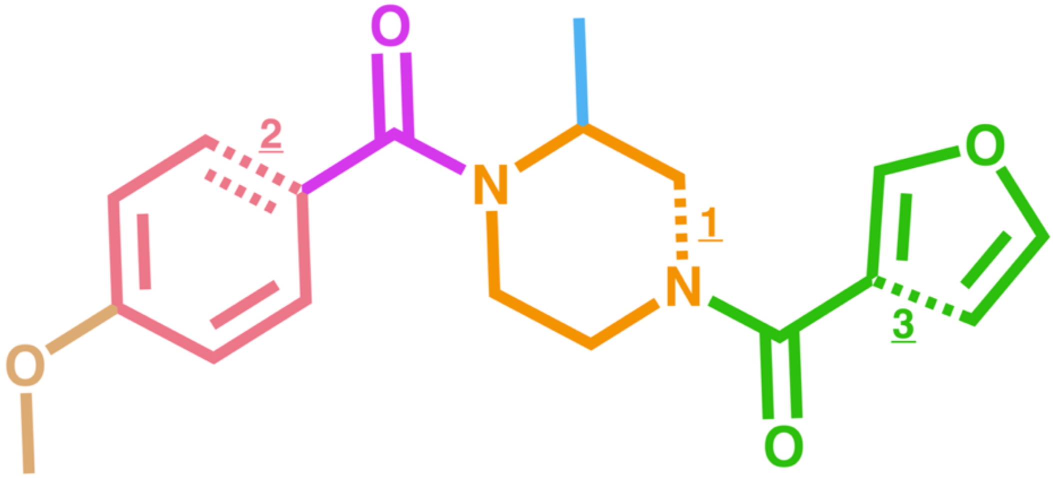
Pulsed single-concentration surface-based biophysical measurements of binding kinetics were performed using a Creoptix®WAVE system (Creoptix, AG).[68] PHIP(2) was immobilised on a 4PCH biosensor surface using amine coupling. Briefly, a sensor chip was conditioned using injections of borate buffer (10 mM sodium tetraborate pH 9, 1 M NaCl). The sensor chip was activated using 1:1 mixture of 400 mM EDC/100 mM NHS for 420 s at 10 µL/min. PHIP(2) was diluted in sodium acetate buffer (10 mM, pH 5.0) and injected over the active surface at a flow rate of 10 µL/min for immobilisation. The surface was then deactivated with ethanolamine-HCl (1.0 M pH 8.5) for 420 s.

Kinetic analysis for PHIP(2) and small molecules was performed using a pulsed injection scheme (waveRAPID)[68] at 25°C with a 5s association and 20s dissociation at a top concentration of 200 µM for all compounds. Blank samples of the running buffer were injected during the measurements every fifth cycle. Compounds were applied to the immobilised surface and a reference channel. Data analysis and visualisation were performed using the WAVEcontrol software 4.5.13 (correction applied: X and Y offset; DMSO calibration; and double referencing). Kinetic parameters were calculated using the Direct Kinetics fitting engine with 1:1 kinetic binding model.

2.3. Introduction to computational concepts

2.3.1. Molecular representation, fingerprinting, and similarity calculations

The ability to represent molecules as machine-readable data structures are essential for computational processing of chemical entities. One such representation is the Simplified Molecular Input Line Entry System (SMILES), a versatile notation for encoding the molecular structure as a single character string (**Fig 2.6**). SMILES strings are generated from generating a series of path within the molecule which correspond to specific substrings which collectively define the molecule. Their generation is governed by a set of rules that dictate the traversal of a molecule's atoms and bonds in a canonical order. These rules cover various aspects of molecular structure. Firstly, they include atomic symbols, which represent elements. Secondly, they define bond types, such as single, double, or triple bonds. Thirdly, they describe ring closures, which use numerical identifiers to connect atoms in a ring via a bond. Lastly, they address branching by using parentheses to denote separate paths within the structure (**Fig 2.6**). The orderly application of these rules aim for a SMILES representation to be unique for a given molecular structure, facilitating effective comparison and manipulation of chemical entities (**Fig 2.7**).[73]



O=C(N1CC(C)N(C(C2=CC=C(OC)C=C2)=O)CC1)C3=COC=C3

Figure 2.6: SMILES string representation encodes molecular atom content, connectivity, and bonding. The image demonstrates how a molecule can be converted into a SMILES string from various substructures. The substructures composing the example molecule are coloured, with the dotted lines corresponding to the number of connecting atoms in the string.

Despite the utility of SMILES in representing molecular structures, there are limitations to consider. Small differences in the molecular structure can significantly alter the resulting SMILES, making molecules with minor differences appear quite dissimilar at a SMILES level. This is particularly problematic as molecules that have similar properties can look very different based solely on their SMILES representations. Furthermore, the generation rules can differ between software implementations, leading to inconsistencies. This variation could potentially confound further processing and analysis.[74,75] Given that the same molecule can yield different SMILES strings depending on the generation procedure used, this lack of standardisation may pose challenges for comparing and integrating cheminformatic data from diverse sources.

SMILES strings can sometimes require conversion to representation suitable for numerical processing. The Morgan fingerprint, also known as the circular fingerprint, is particularly recognised for its ability to capture localised chemical environments around each atom in the molecule. Morgan fingerprints are essentially a set of bit vectors, with each bit corresponding to a chemical group that can be either activated or deactivated for a given molecule. The generation of a Morgan fingerprint is an iterative process, in which atom environments are hashed in concentric circles, with each iteration corresponding to a wider radius of atom neighbours (**Fig 2.7**).[76]

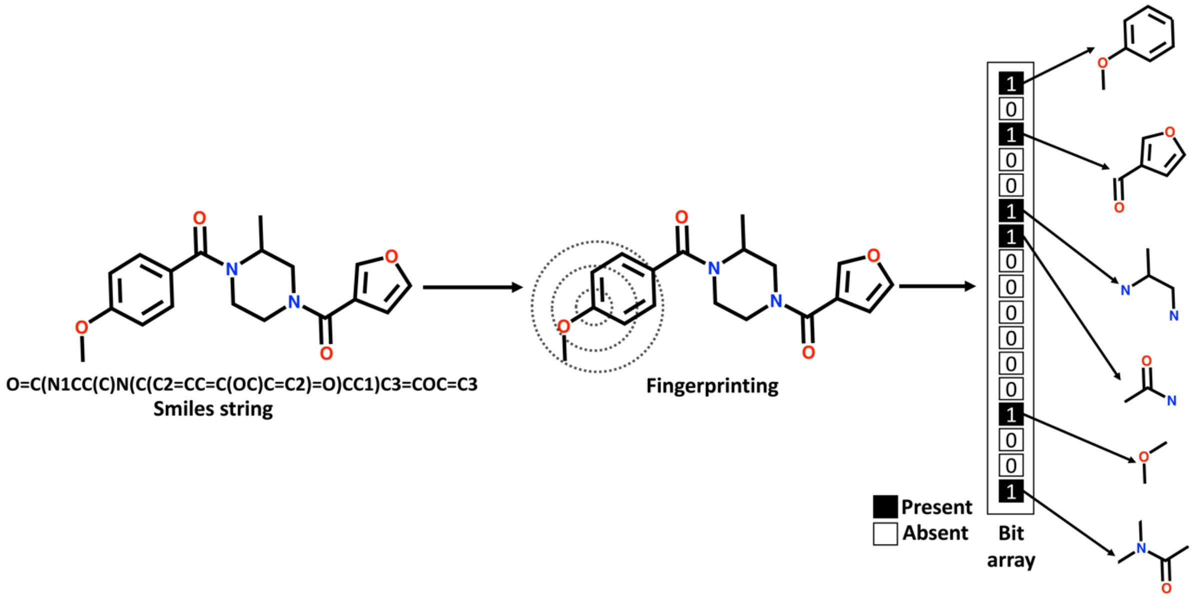


Figure 2.7: Fingerprinting generates machine-readable bit array vectors from SMILES strings for further analysis. The SMILES string serves as input for the fingerprinting process, which decomposes the molecules into substructures while capturing molecular atom content, connectivity, and bonding. The resulting substructures are then folded into a fixed-length fingerprint, with 1 and 0 indicating the presence or absence of a specific substructure within the input enabling comparison with other molecules.

Extended-Connectivity Fingerprints (ECFPs) are a variant of the Morgan fingerprint, sharing the same underlying concept of capturing localised chemical environments around each atom in a molecule. ECFPs differ from the Morgan fingerprint in their hashing method and additional inclusion of topological distance information. This difference results in ECFPs being more sensitive to the structural and topological aspects of a molecule, providing a more refined fingerprint that can better differentiate between similar yet distinct chemical structures. The depth of the fingerprint, often referred to as the radius, controls the extent of the local environment considered.[77] At each radius, the algorithm collects information about the atom types, bond types, and other molecular features, generating a unique identifier based on these properties. This process culminates in the assignment of a unique binary code to each atom, which is subsequently combined into a fixed-length binary vector representing the entire molecule (**Fig 2.7**).

However, ECFPs also have limitations. A key drawback is the potential for different molecules to have identical fingerprints, also known as hash collisions. This can occur when distinct structural features generate the same hash value in the fixed-length binary vector. Furthermore, ECFP does not include 3D spatial information, leading to a loss of chirality, which can be crucial for understanding molecular interactions and function. Consequently, while ECFPs are a valuable tool for molecular representation, these limitations could confound data interpretation.[78]

The utilisation of molecular fingerprints enables the calculation of molecular similarity and diversity through the employment of distance metrics. A prominent metric in this context is the Tanimoto coefficient, also known as the Jaccard index, which measures the degree of overlap between two ensembles. The Tanimoto coefficient ranges from 0 to 1, with values closer to 1 indicating higher similarity between the molecules being compared. Mathematically, the Tanimoto coefficient (TC) is calculated as the ratio of the number of shared features between 2 molecules to the total number of features.[79]

$$\textbf{Tanimoto coefficient} \qquad \mathrm{TC} = \frac{|\mathrm{F(A)} \cap \mathrm{F(B)}|}{|\mathrm{F(A)} \cup \mathrm{F(B)}|} \qquad \textbf{Equation 2.7}$$

Where |F(A)| and |F(B)| are the number of elements in the fingerprints F(A) and F(B), respectively, and |F(A) ∩ F(B)| is the number of elements that are present in both F(A) and F(B). The Tanimoto coefficient can then be translated into the Tanimoto distance.

$$\textbf{Tanimoto distance} \qquad \mathrm{TD} = 1 - \mathrm{TC} \qquad \textbf{Equation 2.8}$$

Molecular fingerprints are also of high dimensionality which renders visualisation and clustering tasks challenging. Dimensionality reduction techniques are instrumental in addressing this challenge, enabling the projection of high-dimensional fingerprint data into lower-dimensional spaces while preserving the essential relationships between data points. One such technique is the t-Distributed Stochastic Neighbour Embedding (t-SNE), a nonlinear dimensionality reduction algorithm that effectively captures both local and global structure in the data.[80] Molecular fingerprints, can also be used as input for, machine learning applications, such as Random Forest classifiers to predict specific properties in molecules, like crystallographic binding. By constructing multiple decision trees and aggregating their results, the classifier distinguishes between positive and negative outcomes, refining its predictions based on features and outcomes.[81]

2.3.2. Binding pose predictions, scoring and filtering

Molecular docking techniques aim to predict the orientation of a ligand within a receptor's binding site (**Fig 2.8**), offering valuable insights for structure-based drug design. Initially, various ligand conformations are generated to explore the conformational space (**Fig 2.8**).

Genetic algorithms, a type of stochastic optimisation technique inspired by natural selection, can then be utilised to identify favourable binding poses. These algorithms enable the optimisation of ligand conformations and orientations within the receptor binding site by simulating evolutionary processes. Genetic algorithms involve the iterative generation of an initial population of candidate solutions (binding poses), selecting the fittest candidates based on their scores (defined by the fitness function), executing crossover (recombination) to produce offspring, and mutating offspring to introduce diversity (**Fig 2.8**). In the context of binding pose prediction, diversity is often defined as the ligand’s (and sometimes protein side chains) rotational and translation degrees of freedom in cartesian coordinates thus defining the orientation and position of the pose.

The fitness or scoring function typically considers factors such as van der Waals interactions, hydrogen bonding, electrostatic interactions, and sometimes desolvation effects (**Fig 2.8**).[82] The selection process, driven by conformational sampling and guided by the genetic algorithm's fitness criteria, ensures that the best candidates are chosen for crossover and mutation in subsequent iterations (**Fig 2.8**). When a predetermined number of docked conformations is achieved and converged by the algorithm, the conformer rank is also determined by the scoring function.[83]

However, these poses can be rescored with other tools, such as machine learning scoring functions, developed through training on databases of protein-ligand complexes with known affinities.[84] Shape and colour overlay rescoring of docked poses serve as complementary methods to traditional docking approaches, filtering "unbiased" docking poses based on experimental information. In this process, a ligand, such as an experimentally resolved fragment, serves as a template pose for scoring. The docked pose is then compared to the template, with only overlapping regions considered. This approach consists of two main components: volume and feature overlap scores. The volume score evaluates the volumetric overlap between the docked and template pose, while the feature overlap score assesses the overlap of chemical features, considering crucial elements such as hydrogen bond donors and acceptor regions.[37]

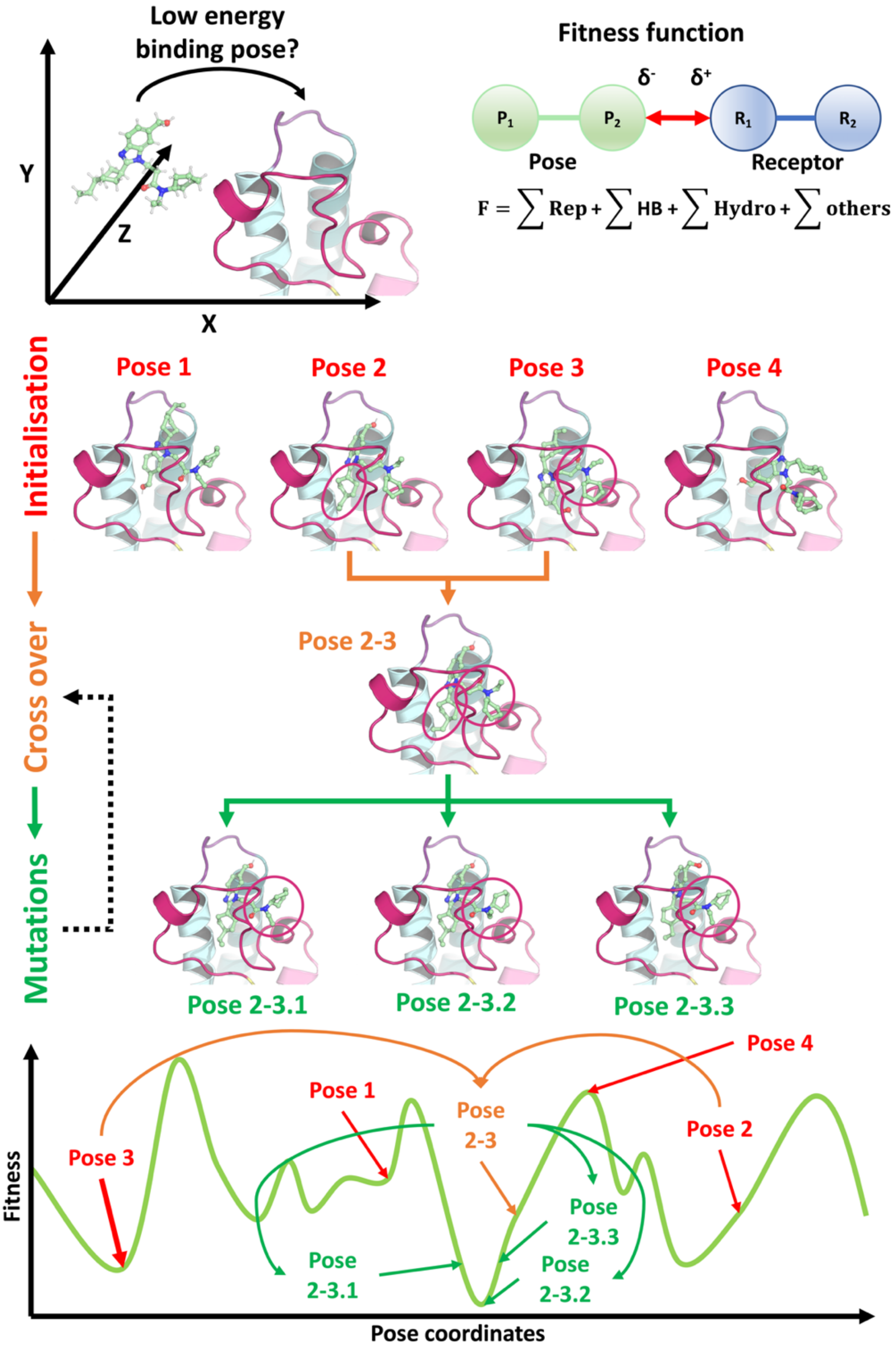


Figure 2.8: Genetic algorithm guides docking prediction to identify low-energy ligand binding poses against a receptor structure. Starting conformations are scored with the fitness function, providing a preliminary sampling of the energy landscape. Poses are then combined, generating offspring, that possess conformational features from both parents. Small mutations, corresponding to variations in rotational and translational degrees of freedom, are introduced to sample local regions of the landscape. The process of crossover and mutations can be iteratively repeated to identify energy minima defined by the fitness function. Fitness is typically defined by a scoring function that assesses the non-bonded potential between the pose and the receptor, sometimes including solvation and intramolecular torsional energies.

Constrained docking, another approach to binding pose prediction, incorporates prior knowledge of the ligand binding pose into the docking process.[85] Fragmenstein,[86] a newly developed software, exemplifies a tool that applies constrained docking. This technique predicts a ligand pose for a receptor by using experimentally resolved information directly, such as the positions of one or more ligands in complex with the target receptor. The maximum common substructure (MCS) of the ligand maps to the positions of the pre-existing ligand poses. A conformer is positioned based on this mapping, which often leads to an unphysical pose due to conflicting atom positions. To resolve this, the system (protein and unphysical ligand) undergoes minimisation to restore normal molecular geometries. This process adjusts the ligand pose to reduce steric clashes and optimise protein-ligand interactions, resulting in a more accurate binding pose.

2.3.3. Molecular dynamic simulations

Molecular dynamics (MD) simulations represent a physics-based methodology for probing the conformational space and temporal behaviour of proteins, ligands, and complexes. This approach enables researchers to investigate the behaviour of biomolecules in contexts approximating physiological settings, yielding valuable insights into their dynamics. The procedure of performing MD simulations encompasses a sequence of predetermined stages, including model construction, parameterisation, minimisation, unbiased simulation, and trajectory analysis. Crystal structures are often used to seed molecular dynamic simulations with the ligand either experimentally resolved or modelled via docking.

Model building is a pivotal phase in MD simulations, establishing the groundwork for subsequent steps. This stage necessitates the preparation of both the protein and the ligand. Accurate protonation states must be assigned to each, frequently via computational pKa predictions, to accurately represent the physiological context. The protein model often necessitates the addition of missing loops and atoms that are not present in experimentally resolved structures. This is carried out employing techniques such as homology modelling to predict protein structures based on sequence resemblance to known structures. Furthermore, it is vital to solvate the protein with water and ion molecules to replicate physiological conditions.[87]

Force fields play an integral role in MD simulations, delineating the mathematical functions and parameters that characterise a system's potential energy, encompassing atomic interactions such as bond stretching, angle bending, torsional interactions, and non-bonded interactions **(Fig 2.9)**. Numerous force fields have been specifically designed for proteins. One such is the Amber ff99SB,[88] which brought novel protein backbone dihedral angle parameters for better alignment with high-resolution protein crystal structures. Amber ff99SB-ILDN,[89] the subsequent update, further refined these parameters and accurately accounted for side-chain torsion potentials of isoleucine, leucine, aspartic acid, and asparagine residues. Later, Amber ff14SB made further enhancements to both backbone and side-chain torsional parameters based on experimental data and high-level quantum mechanical calculations.[90] These force fields contain predefined parameters suitable for classical biological and solvent molecules, thus highlighting their integral role in the accuracy and predictability of MD simulations

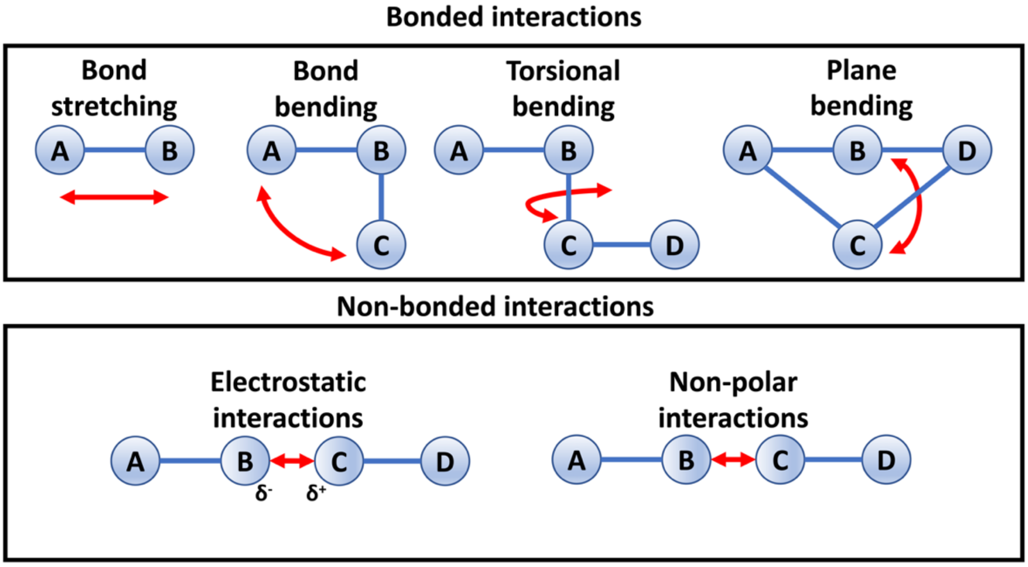


Figure 2.9: Typical force field terms used in molecular dynamics simulations. There are two broad classes of terms related to bonded and non-bonded molecular interactions. The bonded terms describe the potential forces associated with bond stretching, angle bending, dihedral rotation, and plane bending. The non-bonded terms define the attractive and repulsive forces generated from Coulombic and van der Waals interactions.

Conversely, ligands, often small molecules with diverse chemical structures, are not covered by standard protein force fields. Therefore, assigning force field parameters to ligands typically necessitates additional procedures, including parameterisation or derivation of specific parameters for the unique functional groups present in the ligand. This process may entail quantum mechanical calculations or the employment of specialised force fields for small molecules, such as the Generalised Amber Force Field (GAFF).[87]

Before initiating the actual simulation, the system must undergo a series of minimisation steps to eliminate steric clashes, mitigate unfavourable conformations, and equilibrate the solvent molecules within the simulation box and around the protein. Additionally, the system's temperature and pressure are stabilised through dedicated equilibration steps conducted in NVT (constant number of particles, volume, and temperature) and NPT (constant number of particles, pressure, and temperature) ensembles, respectively.[87]

Upon suitable equilibration of the system, unbiased MD simulations can be executed. This process involves integrating the equations of motion over a predefined time interval, derived from Newton's second law of motion. Numerical integration algorithms, such as the Verlet or leapfrog method, are utilised to iteratively compute the positions and velocities of atoms over discrete time increments. During these calculations, hydrogen to heavy atom bonds are frequently constrained using algorithms like LINCS, facilitating longer time steps in the femtosecond range while preserving structural integrity. The MD simulation trajectory is generated by updating the positions and velocities of atoms in the system at each time increment.[87]

After the simulation, the resulting trajectories must be processed and analysed to extract significant information regarding the system's dynamics and energetics. A prevalent technique for analysis is the root-mean-square deviation (RMSD) calculation, which quantifies the average distance between atoms in a reference structure and the atoms in the simulated structures. By monitoring RMSD values over time, it is possible to evaluate structural stabilities within the system, thereby offering atomistic and dynamic information relevant to aspects such as protein-ligand binding.

MD simulations, when complemented with suitable analysis tools, can serve as effective instruments in unveiling conformational states that may remain obscured within experimental structures. An illustrative example is a study that employed extensive MD simulations, which led to the consistent revelation of an allosteric site across bromodomains.[91] Interestingly, this allosteric site is absent in a vast majority of crystal structures, which may provide an explanation for its previous unnoticed existence until persistently delineated by MD simulations. Nevertheless, it is essential to exercise caution as it can be challenging, if not entirely impossible, for classical MD simulations to accurately sample conformational states in their correct proportions. This is primarily because force fields are typically fitted against only a very limited portion of possible conformational ensembles.[88] Moreover, there are additional factors that classical force fields might not account for, such as polarisation effects,[92] potentially leading to imprecise results. Therefore, it is advisable to use MD simulations judiciously, in combination with experimental data, thereby facilitating a comprehensive understanding of molecular behaviour and interactions.

2.3.4. MMGBSA and Interaction Entropy

The Molecular Mechanics Generalised Born Surface Area (MMGBSA) method and Interaction Entropy (IE) method are two widely employed computational approaches to investigate protein-ligand binding free energies, which provide insights into the underlying energetics driving the formation of protein-ligand complexes (**Fig 2.10**).

Both methods rely on pre-processing of molecular dynamics (MD) trajectories, which involves solvent removal, and frame selection, to yield a series of representative snapshots that describe the structural ensemble of the complex under investigation (**Fig 2.10**). Overall, the enthalpy (ΔH) and entropy (-TΔS) of binding are estimated with MMGBSA[40] and Interaction Entropy,[93] respectively, and combined to estimate the Gibbs free energy (ΔG) (**Fig 2.10**).

Gibbs free energy estimation

$$\Delta \mathrm{G}_{\mathrm{est}} = \Delta \mathrm{H}_{\mathrm{mmgbsa}} - \mathrm{T}\Delta \mathrm{S}_{\mathrm{IE}}$$

Equation 2.9

The MMGBSA methodology can involve separating the complex, protein, and ligand from an individual MD trajectory, and calculating the binding enthalpy (ΔH) as the difference between the energies of the protein-ligand complex (H_{PL}) and those of the isolated protein (H_P) and ligand (H_L).

Binding enthalpy estimation

$$\Delta \mathrm{H}_{\mathrm{calc}} = \langle \mathrm{H}^{\mathrm{PL}} - \mathrm{H}^{\mathrm{P}} - \mathrm{H}^{\mathrm{L}} \rangle_{\mathrm{PL}}$$

Equation 2.10

The energies are averaged over the number of frames with the protein and ligand trajectories extracted from protein-ligand complex's (PL) trajectory . The energies for all 3 components (H^x) are computed by summing the molecular mechanical energy (ΔE_{MM}) within individual components, and estimated solvation free energy of transferring the component from the gas to solvent (ΔG_{sol}) (**Fig 2.10**).

Enthalpy energy $$\mathrm{H}^{x} = \Delta \mathrm{E}_{\mathrm{MM}}^{x} + \Delta \mathrm{G}_{\mathrm{sol}}^{\mathrm{x}}$$ **Equation 2.11**

The molecular mechanical energy is the sum of the bonded and non-bonder interaction energies in gas phase. The bonded energies (ΔE_B) are the intramolecular interactions which are composed of bond (ΔE_b), angular (ΔE_{ag}) and dihedral (ΔE_{dhl}) energies (**Fig 2.9**). The non-bonded energies (ΔE_{NB}) are the intermolecular electrostatic (ΔE_{ele}) and van der Waals (ΔE_{vdW}) interactions between the protein and the ligand and these are equal to zero for the ligand and protein are considered alone.

Molecular mechanical energy

$$\begin{aligned}\Delta \mathrm{E}_{\mathrm{MM}} &= \Delta \mathrm{E}_{\mathrm{B}} + \Delta \mathrm{E}_{\mathrm{NB}} \\ &= \left(\Delta \mathrm{E}_{\mathrm{bd}} + \Delta \mathrm{E}_{\mathrm{ag}} + \Delta \mathrm{E}_{\mathrm{dhl}}\right) \\ &\quad + (\Delta \mathrm{E}_{\mathrm{ele}} + \Delta \mathrm{E}_{\mathrm{vdW}})\end{aligned}$$

Equation 2.12

Consequently, most of the energetic contribution for each component to the final binding energies is derived from the free energy of solvation calculations. These calculations attempt to account for solvation effects, such as structural deformation or cavitation. The free energy of solvation is estimated from two terms: the polar (ΔG_p) and non-polar contributions (ΔG_{NP}).

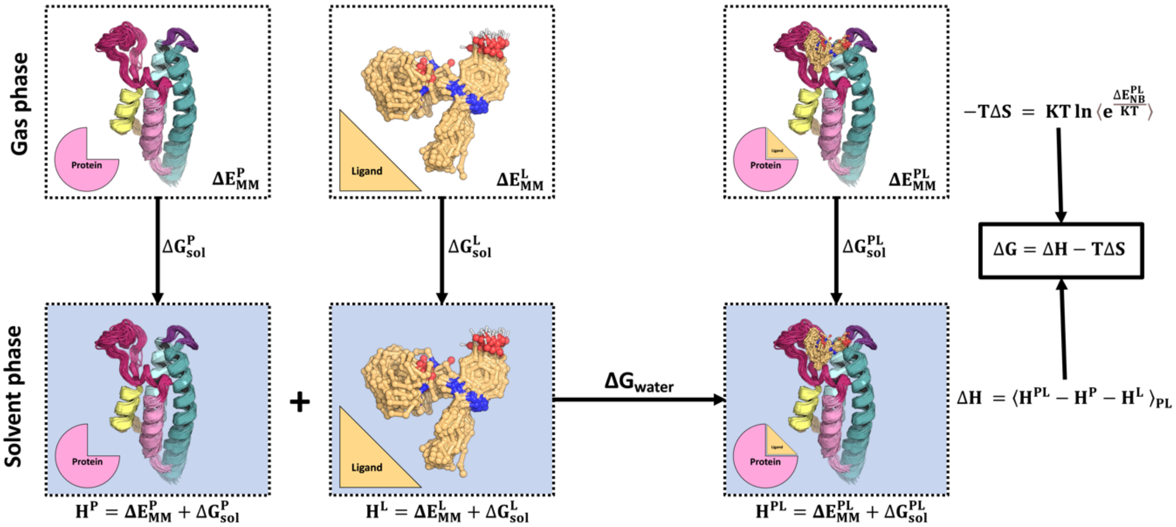
Free energy of solvation

$$\begin{aligned}\Delta \mathrm{G}_{\mathrm{sol}} &= \Delta \mathrm{G}_{P} + \Delta \mathrm{G}_{NP} \\ &= \Delta \mathrm{G}_{\mathrm{GB}} + \Delta \mathrm{G}_{SASA}\end{aligned}$$

Equation 2.13

The Generalised Born (GB) approximation (ΔG_{GB}) estimates the polar solvation energy by representing the solvent as a dielectric continuum. It calculates the electrostatic interactions between the solute and solvent using the Born equation, which relates the solvation energy to the charge distribution of the solute and the dielectric properties of the solvent. The solvent-accessible surface area (SASA) method (ΔG_{SASA}) estimates the non-polar solvation energy, which is related to the hydrophobic effect. It calculates the surface area of a molecule that is accessible to solvent molecules, and the non-polar solvation energy is then derived from this surface area using a linear relationship with a proportionality constant, representing the solute-solvent interaction energy per unit area.[40]

The precision of MMGBSA calculations varies upon a variety of factors. These include the type and charge of protein and ligand under investigation, the choice of force field, the protocol followed, and the corrections implemented.[94] As such, it is recommended to carry out comprehensive benchmarking prior to the execution of MMGBSA calculations when experimental data is available. Expertly applied, MMGBSA can augment the effectiveness of virtual screening processes.[95] This is further evidenced by MMGBSA computations reaching a coefficient of determination value of 0.63 when juxtaposed against a curated version of the PDBbind database.[96] This indicates that whilst MMGBSA can effectively predict broader trends, it may not always provide an exact quantification of binding affinities. Importantly, MMGBSA often demonstrates a higher proficiency in ranking compounds based on their binding affinities as opposed to delivering precise numerical predictions for these affinities.



Gas phase
Solvent phase
Protein
Ligand
ΔE_{MM}^{P}
ΔE_{MM}^{L}
ΔE_{MM}^{PL}
ΔG_{sol}^{P}
ΔG_{sol}^{L}
ΔG_{sol}^{PL}
+
ΔG_{water}
$-T\Delta S = KT\ln\langle e^{\frac{\Delta E_{NB}^{PL}}{KT}}\rangle$
$\Delta G = \Delta H - T\Delta S$
$\Delta H = \langle H^{PL} - H^{P} - H^{L}\rangle_{PL}$
$H^{P} = \Delta E_{MM}^{P} + \Delta G_{sol}^{P}$
$H^{L} = \Delta E_{MM}^{L} + \Delta G_{sol}^{L}$
$H^{PL} = \Delta E_{MM}^{PL} + \Delta G_{sol}^{PL}$

Figure 2.10: Molecular dynamics simulations provide material required for free energy calculations using Molecular Mechanics with Generalised Born and Surface Area Solvation and Interaction Entropy calculations. The trajectories stripped of water are used to calculate the energies for the protein, ligand, and complex. The solvation free energy is then estimated using generalised Born and surface area solvation approximations assuming implicit solvent for polar and non-polar terms, respectively. The protein and ligand enthalpies are subtracted from the protein complex enthalpy, thus estimating the binding enthalpy. The entropy is estimated from the protein-ligand trajectory and combined with the enthalpy to estimate the binding free energy.

The Interaction Entropy (IE) method presents an interesting paradigm for estimating the entropic cost of binding (**Fig 2.10**). It estimates such quantity with the exponential averaging of the protein-ligand intermolecular non-bonded forces across simulation frames.

$$\textbf{Interaction Entropy} \qquad -T\Delta S = KT\ln\langle e^{\frac{\Delta E_{NB}^{PL}}{KT}}\rangle \qquad \textbf{Equation 2.14}$$

Where ΔE_{NB}^{PL} are the non-bonded electrostatic and van der Waals interactions between the protein and the ligand. T and K are the temperature and gas constant, respectively. Together with the MMGBSA estimation of binding enthalpy, the Gibbs free energy of binding can be estimated from MD trajectories (**Fig 2.10**).[93]

Chapter 3. A community evaluation of computational methods for fragment screening

3.1. Credits

The research and manuscript presented in this chapter were conducted and authored by Harold Grosjean. Harold Grosjean was responsible for protein expression, purification, and crystallisation, as well as crystallographic fragment screening, structure refinement, challenge organisation, submission evaluation, and cheminformatic analysis. Additionally, Harold Grosjean prepared the manuscript and created the figures. The PHIP(2) expression system and crystal form were provided by Dr Tobias Krojer. Dr Anthony Aimon, Dr Romain Talon and Dr Alice Douangamath assisted with the crystallographic screening and analysis under the supervision of Prof Frank von Delft. The initial coding and statistical analysis infrastructure were established by Dr Mehtap Işık. Dr Tim Dudgeon generated follow-up database subsets with the fragment network. Prof John Chodera offered guidance in conceptualising the project, while Philip C. Biggin and Prof David Mobley aided in refining and reviewing the manuscript.

3.2. Publication

The research conducted in this chapter led to a publication, which can be accessed using the following reference:

Grosjean, H. *et al.* (2022) "SAMPL7 protein-ligand challenge: A community-wide evaluation of computational methods against fragment screening and pose-prediction," *Journal of Computer-Aided Molecular Design*, 36(4), pp. 291–311.

Available at: https://doi.org/10.1007/s10822-022-00452-7.

3.3. Introduction

Chapter 1 and 2 outlined the basic concepts and challenges associated with fragment identification and elaboration. Computer-assisted drug discovery aims to computerise experimental procedures to reduce labour, time and costs, therefore increasing turnover. Fragment identification is not exempt from this paradigm. However, there is little in the way of systematic validation specifically for fragment-based approaches. To address this issue the SAMPL challenges prospectively assess the predictive power of methods involved in computer-aided drug design.

In this chapter a large crystallographic fragment screen was performed against the therapeutically relevant second bromodomain of the Pleckstrin-homology domain interacting protein that revealed 52 different fragments bound across 4 distinct sites, 47 of which were bound to the pharmacologically relevant acetylated-lysine binding site. These data were kept undisclosed and used to assess computational screening, binding pose prediction and follow-up enumeration in a blinded fashion. All submissions performed randomly for screening. Pose prediction success rates (defined as less than 2 Å root mean squared deviation against heavy atom crystal positions) ranged between 0 and 25% and only a very few follow-up compounds were deemed viable candidates from the manual analysis of common molecular descriptors.

The tighter deadlines imposed here, compared to similar challenges, led to a small number of submissions suggesting that the accuracy of rapidly responsive workflows remains limited. In addition, the application of these methods to reproduce crystallographic fragment results still appears to be very challenging. This work shows that there is room for improvement in the development of computational tools, particularly when applied to fragment-based drug design. In addition, the results suggest that experimental approach are still required for identifying fragments against a target.

3.4. Results

3.4.1. Crystallographic fragment screening against PHIP(2) reveals novel binders

The second bromodomain of the Pleckstrin-homology domain interacting protein (PHIP(2)) was initially crystallised in a C2 space group that diffracts to a resolution of approximately 1.2 Å by Dr Tobias Krojer. This crystal form is easily reproducible making it ideal for an XChem screen.[97] The DSI-poised[98] and FragLites[12] libraries were screened against the C2 crystals. The former library is composed of 768 fragments and was designed to ease follow-up chemistry. FragLites is composed of 31 halogenated fragments, all of which have a paired H-bond donor/acceptor motif to probe interaction doublet in the binding site while the halogen atoms assist electron density fitting.

Out of these 799 fragments, no binding was observed for 707 of them despite the acquisition of adequate diffraction data sets. To minimise the number of missed fragments, all FragLites were screened in duplicate. The remaining fragments were re-screened if the diffraction dataset did not display the expected C2 space group or had a resolution lower or equal to 2 Å or if the R_{cryst} and R_{free} values were lower than 0.23 and 0.25, respectively. This led to the re-soaking of 202 fragments that resulted in the identification of 10 additional hits including 8 acetylated-lysine binding site hits (**Fig. 3.1**).

In total, 52 hits were identified across 4 sites (**Fig. 3.1**), achieving a global hit rate of 6.51%. 47 fragment hits are bound to the pharmacologically relevant acetylated-lysine binding site (**Fig. 3.1**) and are summarised in **Table 3.1**. 7 hits were resolved at a small, solvent-exposed cavity located between α-helices C and Z and 4 out of these also bind to the main binding site, implying dual binding. 2 additional hits were found, 1 behind the BC-loop and 1 in between the α-helices A and B (**Fig. 3.1**). The fragments binding away from the main acetylated-lysine binding site are largely solvent exposed and contact other proteins in the lattice and hence may be artefacts of crystal contacts. Thus, only the acetylated-lysine (Kac) binding site hits were considered going forward.

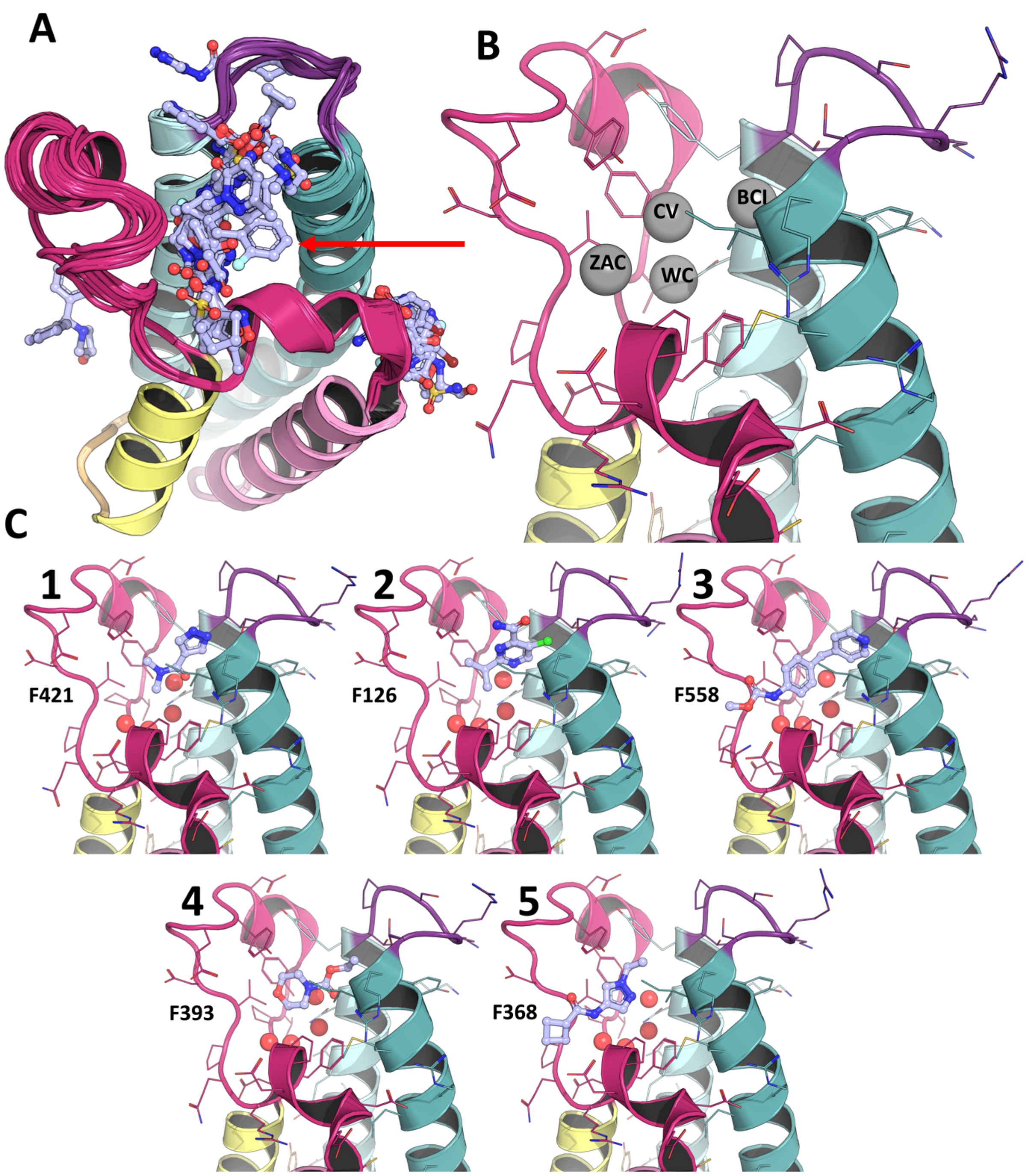
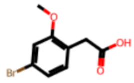
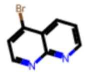
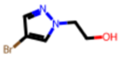
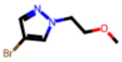
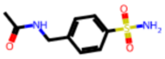
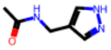
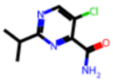
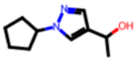
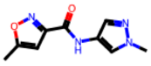
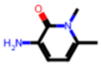
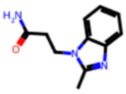
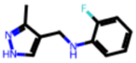
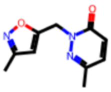
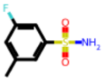
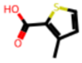
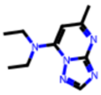
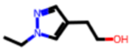
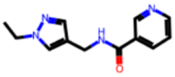
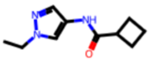
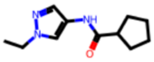
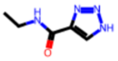
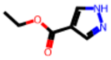
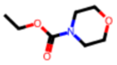
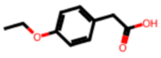
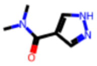
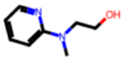
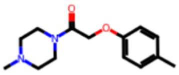
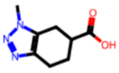
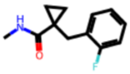
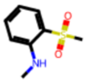
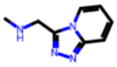
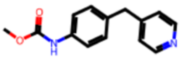
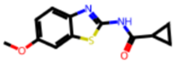
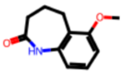
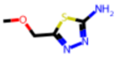
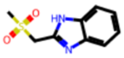
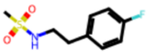
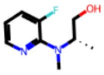
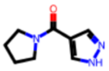
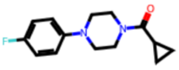
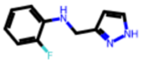
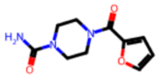
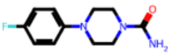
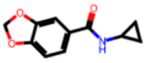
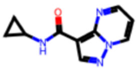
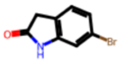
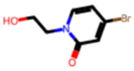


Figure 3.1: Overview of the fragment hits against the C2 crystal form. (A) Overlay of the 52 structures resulting from the XChem screen. The Kac binding is the most populated site. The other sites can be seen on the right and left of the Kac binding site. An additional fragment hit is located behind the purple BC-loop. B illustrates the 4 fragment binding subsites with BCI, CV, ZAC and WC denoting the BC-interface, central void, ZA-channel, and the water cavity, respectively, as grey spheres. C shows selected fragment binding poses to illustrate the interactions and binding regions identified. C1 shows F421 that forms an H-bond with serine 1392 at the BC-interface. C2 shows F126 that forms a halogen bond with threonine 1396 and serine 1401 at the BC-interface. C3 shows F558 with 2 aromatic 6-membered rings that form perpendicular π-π stacking with tyrosine 1350 and tyrosine 1395. C4 shows F393 which has a morpholine moiety that fills the central void. Finally, C5 shows F368 which forms H-bonds on the ZA-channel with backbone nitrogen and oxygen of Aspartic acid 1346 and proline 1350, respectively. The red arrow in panel A indicates the point of view used in panels B and C.

Table 3.1: The 47 Fragments identified in the acetylated-lysine binding pocket of PHIP(2). The fragment IDs and corresponding PDB accession numbers are showed, below each 2D molecular representation on the left and right, respectively.

 F5: 5RJQ	 F11: 5RJJ	 F13: 5RJM	 F14: 5RJN	 F95: 5RK0	 F96: 5RJY
 F126: 5RKE	 F179: 5RKB	 F199: 5RJT	 F205: 5RK3	 F217: 5RK8	 F229: 5RKQ
 F274: 5RKP	 F275: 5RKX	 F309: 5RJW	 F362: 5RK4	 F366: 5RKF	 F367: 5RKK
 F368: 5RKN	 F369: 5RKJ	 F374: 5RJZ	 F389: 5RKG	 F393: 5RK7	 F400: 5RKV
 F421: 5RK9	 F467: 5RK2	 F488: 5RKH	 F501: 5RKD	 F503: 5RKY	 F529: 5RJX
 F535: 5RKC	 F558: 5RJV	 F579: 5RKW	 F584: 5RK1	 F600: 5RK5	 F603: 5RKA
 F616: 5RJR	 F618: 5RKM	 F650: 5RKT	 F666: 5RKO	 F687: 5RKL	 F709: 5RKI
 F710: 5RK6	 F740: 5RKU	 F741: 5RJS	 F760: 5RJO	 F763: 5RJP	

Inspection of fragments binding to the acetylated-lysine binding site suggests this pocket can be divided into 4 sub-sites (**Fig. 3.1**). These were named: i) The BC interface (BCI), which includes interactions with α-helices B and C and, the BC-loop (BCL) ii) the water cavity (WC), which is defined by the location of the 4 water-molecule network iii) the ZA channel (ZAC), which is the part of the ZA loop that forms a semi-circle iv) the central void which lies at the centre of the 3 other sub-cavities. The subsites and exemplar fragments are displayed in **Figure 3.1**, respectively.

The screen probed the binding site well with fragments contacting most all side chains composing its surface (**Fig. 3.1**). The most frequently occurring H-bond forming amino acids were side chains of serine 1392, threonine 1396 and serine 1401, all of which are located at the BC-interface as well as the backbone nitrogen of aspartic acid 1346 and the backbone oxygen of proline 1340, which are located on the ZA channel. Halogen bonds show the same interaction pattern around the BC-loop but penetrate deeper between with α-helices B and C to interact with multiple side chains simultaneously. Tyr1350 seals the top of the central void, whilst Tyr1395 is located on the BC-interface. These side chains are positioned in such a way that fragments form perpendicular π-π stackings. Other hydrophobic residues that frequently contact the fragments are valine 1345, isoleucine 1403, which are in the central void. The PHIP(2) water appears to be easily displaceable as all fragments that interact at the BC interface displace it (**Fig. 3.1**).

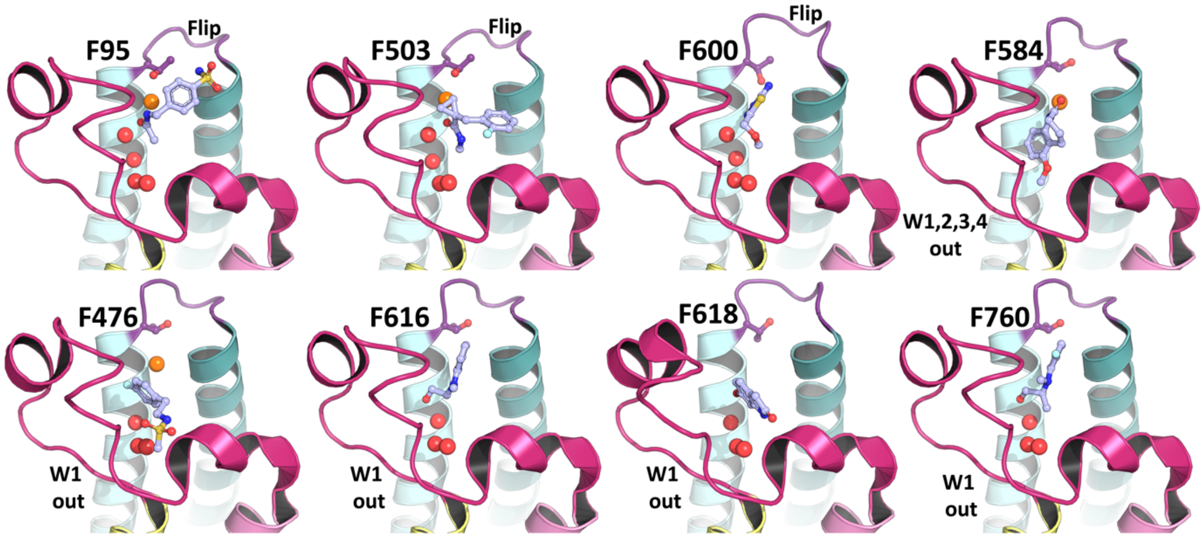
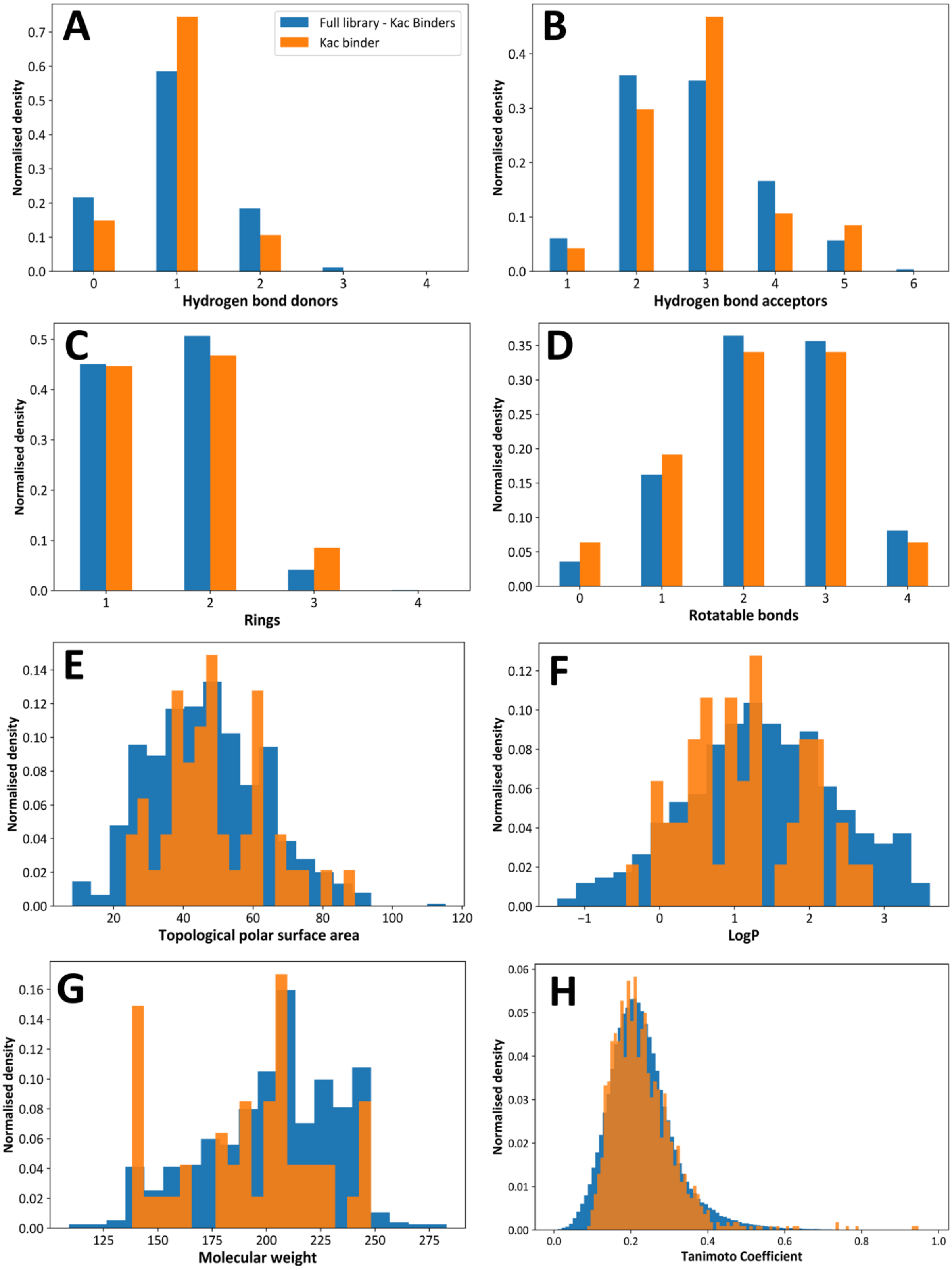


Figure 3.1: Special acetylated-lysine site binder cases. The annotation “flip” indicates that the fragment induces a re-arrangement of threonine 1396 into a peptide-bound like conformation, which is paired with a change of the BC-loop conformation. The annotation “Wn out” indicates that the fragment displaces the Nth bromodomain water.

Overall, the C2 crystal form appears to be rigid with few fragments inducing protein motions. F760 induces the largest protein motions by relaxing the ZA- and BC-loops and bringing tyrosine 1395 closer to the binding site core. Fragments F95, F503 and F600 cause the re-arrangements of threonine 1396 into a peptide-bound conformation (**Fig. 3.1**). In F95 and F503, this is paired with the formation of a water bridge between the fragment and threonine 1396 whilst F600 contacts this side chain directly (**Fig. 3.2**). In addition, the 2 former fragments (F95 and F503) rotate isoleucine 1403 away from the binding site in a parallel orientation with α-helix C. 5 fragments disrupt the 4-water network (**Fig. 3.2**). F584 is the most remarkable hit by displacing the 4 water molecules network to locate itself deep in the binding cavity where its benzene moiety interacts with tyrosine 1353, and its 7-membered ring fills the hydrophobic space of the central void. F467, F616, F618 and F760 all displace the fourth water of the network.

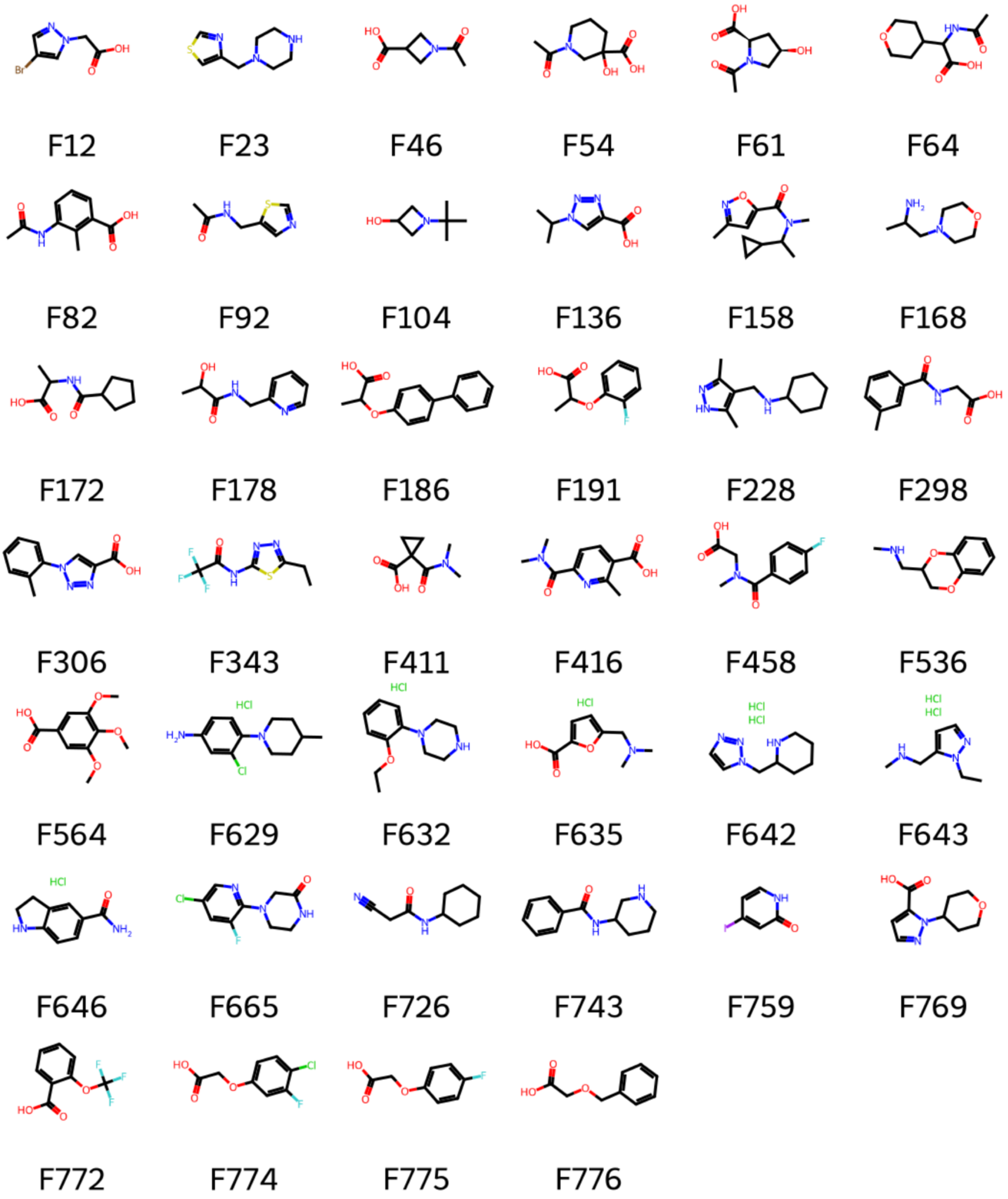
Cheminformatic analysis of the fragments bound to the acetylated-lysine binding site suggests they tend to be smaller in molecular weight and more hydrophobic than the library average and most of them have 1 and 3 H-bond donors and acceptors, respectively (**Table 3.2**).

Table 3.2: Comparison of common molecular descriptor between the screened fragments (excluding the acetylated-lysine site binders) and acetylated-lysine site fragment binders. The plots show the density of molecules at a given descriptor value normalised by the number of molecules composing the set. The full library excluding Kac site binders and Kac binders are displayed in blue and orange, respectively. A, B, C, D, E F, G show the distributions of H-bond donors, H-bond acceptors, rings, and rotatable bonds counts, respectively. Panel H shows the distribution of Tanimoto coefficients for all against all molecules within each set.



No usable data could be collected for 1 FragLites (F12) and 39 other DSI-poised fragments led to consistently damaged crystals (despite repeated soaking) resulting in the absence of diffraction data for these molecules (**Table 3.3**). The exact cause(s) of such degradation remain, unclear and could be diverse but may be due to relative crystal tolerance to individual ligands at the soaking concentration or defective ligand stock solutions.

Table 3.3: Molecular structure of screened fragments for which no usable diffraction data could be obtained.



3.4.2. Stage 1: Discrimination of binders from non-binders

The aim of the first stage was to discriminate binders from non-binders identified at the acetylated-lysine binding site from screening presented above. The participants were provided with the SMILES strings of the 799 screened fragments. In addition, they were also provided with a PHIP(2) apo-structure in the C2 space group. The location of the acetylated-lysine binding site was also indicated by a dummy noble gas atom positioned in the PDB file. The experimental protocol was also supplied to reproduce the crystallographic results computationally. Experimental information provided included: crystallisation conditions, pH, and the fragment soaking concentration.

The participants were asked to categorise each fragment listed with Boolean values: True and False for binders and non-binder, respectively. Ranking, scoring, or confidence metric were not asked to be reported. Participants were given 1 month to submit entries. Predictions were requested for all four binding sites seen in the screen. However, most participants only submitted solutions for the main acetylated-lysine binding site; thus, these peripheral sites further were not investigated further. **Table 3.4** provides a summary of submissions, which were expanded in detail via the individual submission identification numbers (SIDs).

Table 3.4: Overview of Stage 1 virtual screening submissions. 9 sets of predictions were submitted for Stage 1. The first, second, third and fourth columns correspond to the submission identification number (SID), the affiliation of the participants, the method's name and the list of software used. The last column shows the ratio of predicted non-binders over predicted binders.

SID	Participant Affiliation	Method Name	Method category	Software used	Non-Binders to binders ratio
38	Akiyama Lab, Department of Computer Science, Tokyo Institute of Technology	AutodockVina-VirtualScreening	Docking	Autodock Vina 1.1.2, MOE 2019.0102, Openbabel 3.0.0, AutodockTools 1.5.6	0.001
55	Not provided	dock_score	Docking, (MD)	OpenEye docking toolkit v1.1	7.528
56	The University of Tokyo, Japan	Template docking and similarity search	Docking, Ligand-based	Molegro Virtual Docker (7.0.0), Schrödinger suite (2019-3), PubChem (current)	10.859
52	Institut de Chimie des Substances Naturelles, CNRS, Gif-sur-Yvette, France	docking-S1-G2-chemscore-top5perc	Docking	Schrodinger LigPrep v48012, CACTVS Chemoinformatics Toolkit V3.4.6.26, CORINA v4.2.0, CCDC GOLD v5.7.1 (CSDS-2019-1)	17.975
35	University of Oxford, UK	Molecular Docking with MM-GBSA Scoring	Docking, MD	OpenBabel v3.0, UCSF Chimera v1.12, PDB2PQR v2.1, SPORES v1.3, PLANTS v1.2, DOCK v6.8, AmberTools v2019,	20.686
37	Universitat Pompeu Fabra, Spain	ECFP4rdkit-RF	Ligand-based, ML	MOE 2016.08, python 3.6.9, rdkit 2019.09.1.0, scikit-learn 0.21.3,	68.000
43	Institute of Systems Biomedicine, Peking University	Deep Learning; k-deep; docking	Docking, ML	Python, RDKIT,	68.000
44	University of Barcelona, Spain	DUck-aided Virtual Screening (DaViS)	Docking, MD	rDock 2013.1, Amber16, AmberTools16, MDmix 0.2.0, Gromacs version 2018.1, LigPrep version 46013, MOE 2019.0102, Prime version 5.6 (r012)	25.172

SID 38 used a typical workflow, whereby ligand conformations and protein protonation states were generated by MOE, which was followed by docking with AutoDock Vina.[82] A threshold of ≤ 4.0 kcal/mol for the best scoring pose was used to categorise a fragment as a binder.

SID 55 used OpenEye Omega Toolkit[99] for conformer generation and docking, respectively with docks being scored with the Chemgauss4 scoring function. The participants behind SID 55 did not provide the threshold employed to discriminate binders from non-binder.

SID 33 corresponds to a negative control random submission where compounds were selected randomly of compounds were selected as binders using an in-house python script generated by Prof David Mobley.

SID 56 employed two separated protocols. The fragment-hits from Cox *et al.*[98] were considered as binders and used in a PubChem and 3D fingerprints similarity searches. Then, the remaining fragments were protonated at crystallographic pH then template docking was employed against 2 fragment-bound structures in a different ($P2_12_12$) symmetry group. The participants mentioned that they considered docking scores, similarity scores and ligand efficiency when defining binders and non-binders without further explanations.

SID 52 first built a validation set of 1499 compounds from assay and crystallographic data which includes 45 fragments belonging to the provided library. Those 45 compounds were classified as active. 3D conformers for each validation active were generated with Corina[100] and protonated at physiological. Gold was used with 4 scoring function (GoldScore, ChemScore, ASP and ChemPLP) to dock the 1499 validation actives onto 3 distinct conformers identified from 8 fragment-bound structures.[101] They then looked at the ranking of the 45 active fragments as well as RMSDs to all 8 fragment-bound PDBs. ChemScore with the second conformer (out of the 3) yielded the best scores. They used those parameters to subsequently dock the 799-fragment library and defined the top 5% as binders.

SID 35 kept 5 water molecules during the preparations step with DockPrep tool.[102] Protein and ligand were protonated with PDB2PQR and OpenBabel[103], respectively before performing docking with PLANTS[104] using the ChemPLP as scoring function. The poses were further scored using DOCK6[102], which performs minimisation, Molecular Dynamic simulation and generalised Born/Surface Area (GB/SA) continuum model scoring in Amber.

SID 37 was the only submission that did not employ docking at any stage of the predictions. Instead, they built a random forest classifier based on data from publications, ChEMBL and the PDB. For the training set, they identified a total of 10 and 448 binders and non-binders, respectively. An additional 1214 structures were obtained from the PDB by searching for other bromodomain family member and the associated ligands were classified as binders. 70,000 additional molecules were retrieved from ChEMBL and classified as binders based on activity measurements. Model training was performed against chemical descriptors and Morgan fingerprints. Finally, they classified the provided SMILES and they were predicted to be binders if their probability score was higher than 0.5.

SID 43 did not provide their protocol describing the method employed for predictions despite multiple requests. A combination of machine learning and docking was seemingly employed.

SID 44 employed different MD-based methods and docking in their workflow. The stability of the water network was, first, assessed with MDmix[105] which concluded only water-4 to be easily displaceable. Classical MD simulations were then employed against the provided Apo state and ligand bound structures identified from Cox *et al.*[98] to evaluate protein motions. They observed that the ZA-loop is the most flexible region by visiting opened, semi-opened and closed states and that the presence of ligands stabilises the protein in an opened state. Thus, another structure that displays a more open state of the ZA-loop was selected for further processing instead of the provided Apo state. They then used rDock[106] with pharmacophore restraints to generate poses. 3 different protocols, which used different sets of restraints and water molecules, were applied onto the provided smiles after tautomer and protomer enumeration at pH 5.6 +/- 1.0 For each fragment, the resulting poses were first clustered and then finally selected based on the protein-ligand interaction score. The docking poses resulting from each 3 protocols were evaluated with a Dynamic Undocking-like[107] procedure which extracts the ligand and nearby binding-site environment, performs steered-MD and measures work to assess interaction strength. A threshold was used to categorise binders from non-binders with binders scoring lower than -2 kcal/mole. Further refinement was applied to the predicted binders list which excluded the docking poses that were unstable in Duck or MD windows.

To assess the performance of each submission the sensitivity and specificity were calculated, which are the True Positive and True Negative rates, respectively. Since the dataset is largely populated of non-binders, balanced accuracy was to evaluate the results. The balanced accuracy is the simple arithmetic average of True Positive and True Negative rates and a value of 0.5 indicates a random prediction. The predictions were compared to ground truth values as defined by the outcome of the crystallographic fragment screening experiment.

Here an experimental positive was defined when an interpretable electron density, in which a fragment can be fitted, and a negative where such signal is absent. It is difficult if not impossible to discriminate non-binding events that are caused by structural (i.e binding allowed by receptor's conformation(s) and/ or energetics) or chemical factors (such as crystal or ligand stock conditions). Negatives therefore correspond to experimental negatives, irrespectively of their underlying nature. No usable diffraction data could be obtained for the fragments listed in **Table 3.3**; therefore, these were removed from the subsequent analysis.

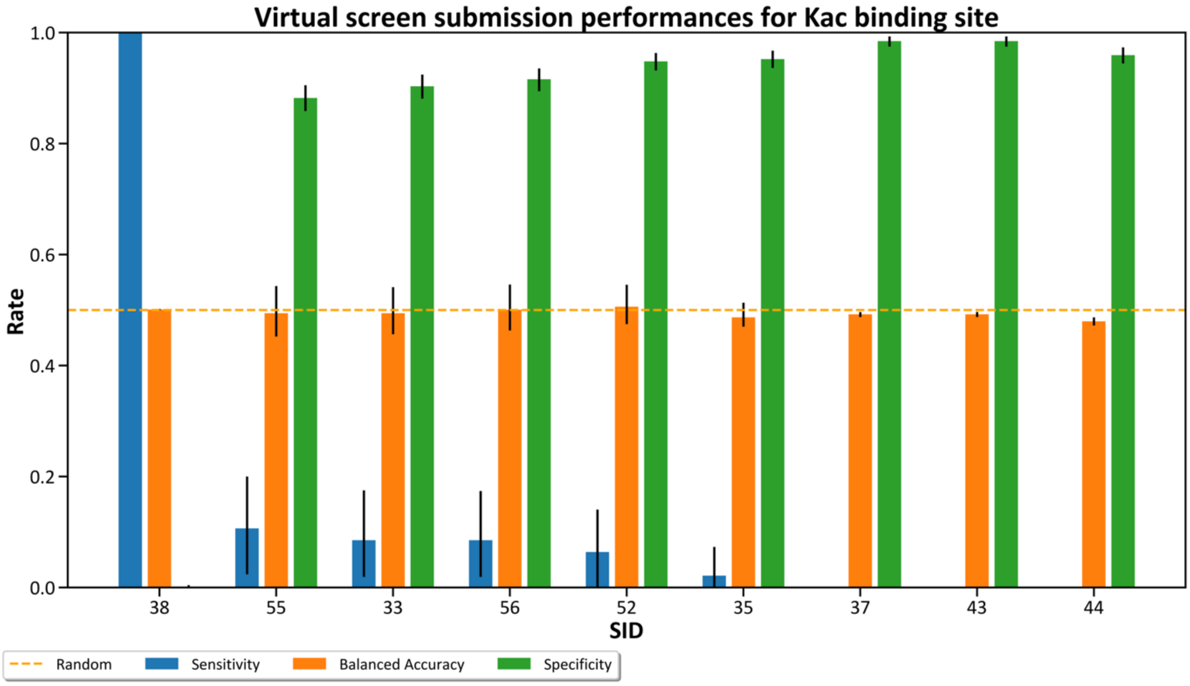


Figure 3.2: Performance summary of submissions to Stage 1 of the challenge. The Sensitivity, Balanced accuracy and Specificity rates for each submission are shown in blue, orange and green, respectively. The error bars associated which each metric was estimated via bootstrapping of the data and the orange horizontal line shows a random perfection based on balanced accuracy values.

This stage turned out to be extremely challenging with all submissions performing approximately random with an average balanced accuracy across all submissions of 0.49 ± 0.01 (**Fig. 3.3**). Overall, all methods, but SID 38, tended to correctly classify the majority of the fragments as non-binders as indicated by their high specificity rates. However, these failed to identify most binders with submissions 37, 43 and 44 being unable to identify a single binding event and submissions 55, 56, 52 and 35 only correctly identifying 5 (F95, F199, F558, F579, F740), 4 (F13, F14, F96, F488), 3 (F275, F529, F760) and 1 (F362) binding events, respectively (**Table 3.1**). SID 38, however, showed an opposite outcome where all fragments were categorised as binders, therefore, leading to a Sensitivity rate of 1.

3.4.3. Stage 2: Binding pose predictions for fragment binders

The objective of the second stage was to correctly predict the crystallographically resolved binding pose for the 47 protein-fragment complexes resolved at the acetylated-lysine binding site. The participants were provided with the same C2 apo state structure (as in Stage 1) and binders SMILES strings. It was also specified that compounds were purchased as racemic mixtures and that higher affinity conformers should be revealed in the electron density or that both stereo-isomers could bind, thus resulting in an average electron density.[108]

The entrants were tasked to submit at least one and no more than 5 poses for each of the 47 fragments and given about 2 weeks to submit predictions. It was also asked the poses to be ranked (from best to worst) if more than one was submitted. Finally, participants were reminded that predictions should be done considering the C2 space group as crystallographic screening against different crystal forms often leads to differential fragment-hit identification.[109] It is known that this is true for PHIP(2) as a previous screen against a $P2_12_12$ form revealed different binders than against the C2 form.[98] It was decided to hide the information about differential crystal binding to strengthen the blinded aspect of the challenge and evaluate how the participant select their system. Only 5 submissions were received for this stage of the challenge for publication, summarised in **Table 3.5**.

Table 3.5: Overview of the Stage 2 binding pose prediction submissions. 5 sets of predictions were received for Stage 2. The first, second, third and fourth columns correspond to the submission identification number (SID), the affiliation of the participants, the method's name and the list of software used.

SID	Participant Affiliation	Method Name	Method category	Software used
77	Molecular Modeling Section Lab (Prof. Stefano Moro) University of Padova, Italy	HT-SuMD	Other, (MD)	MOE 2019.01 Acemd3 VMD Python 3.6 scikit-learn 0.21.3 AmberTools 2016
75	Acellera	rDock-rDeep	Docking, ML	rdkit 2018.03.4 HTMD Playmolecule proteinPrepare rDock rDeep v0
64	Not provided	2d feature model	Docking	Smina
79	The University of Tokyo, Japan SHaLX Inc.	Template docking	Docking	Molegro Virtual Docker (7.0.0)
80	Institut de Chimie des Substances Naturelles, CNRS, Gif-sur-Yvette, France	ranking_stage2	Docking	Schrodinger LigPrep v48012 CACTVS Chemoinformatics Toolkit V3.4.6.26 CORINA v4.2.0 CCDC GOLD v5.7.1 (CSDS-2019-1)

Only SID 77 was totally agnostic of docking. Instead, they used Supervised Molecular Dynamic Simulations (SuMD).[110] SuMD is a relatively novel Molecular Dynamics-based method that aims to sample binding and unbinding paths between a ligand-receptor pair. The algorithm positions the ligand in an MD box, monitors the translation of the ligand towards the binding site and selects frames that are closer to the target binding site. Once the ligand is sufficiently close to the binding site (around 5 Å), the engine switches to a classical unbiased molecular dynamic simulation of the protein-ligand complex. The final unbiased production runs are analysed and scored. Triplicate SuMD runs were carried out and the relevant conformations were extracted from production simulations using a DBSCAN clustering. The different clusters were ranked by a consensus approach that accounts for the cluster sizes, MMGBSA energies and MOE-Hyd estimation of the hydrophobic contributions to binding. Finally, the centroid structures for the top 5 scoring clusters were selected.

SID 75 kept structural water molecules and prepared their system at pH 6.0 with playmolecules.com.[111] 100 poses per fragment were generated with rDock[106] which were rescored with a seemingly *in house*, convolutional neural network function and the top 5 non-redundant poses were submitted.

SID 64 used a simple docking protocol where they used AutockVina[82] to dock the fragments into acetylated-lysine binding site of the provided apo structure. The participants gave little additional information about their protocol.

SID 79 selected 3 ligand-bound PHIP(2) PDB structures: 3MB3, 5ENF and 5ENI. The provided binders were classified by template matching which resulted in 3MB3, 5ENF and 5ENI being associated with 19, 15 and 12 binders, respectively. Molegro Virtual Docker[112] was used for docking of the binders into the receptors. The provided apo structure was only used in one case after alignment of 3MB3.

SID 80 used a similar method to their Stage 1 submission. The same validation set of 1499 compounds which contained 45 actives was used. The systems were prepared as mentioned in Stage 1 and the validation compounds were each docked onto 3 different ligand conformations with 50 poses generated for each. Again, ChemScore on the second conformation performed best. The provided fragment-hits were prepared and docked using the workflow described in Stage 1 and the top scoring pose submitted.

The performance of each submission was assessed by calculating the fragment root mean square deviation (RMSD) between the docked and the crystallographic poses. The top scoring (here referred as “Best” or “Top 1”) and lowest RMSD poses (here referred as “Best in all” or “Top 5”) amongst submitted poses for a given fragment, were recorded for success rate evaluation. Calculation of the success rate as a function of RMSD shows how a particular submission performs as the RMSD cut-off defining success rate is changed. An RMSD cut-off of ≤ 2Å was chosen to categorised docking pose predictions as successful (**Fig. 3.4**).

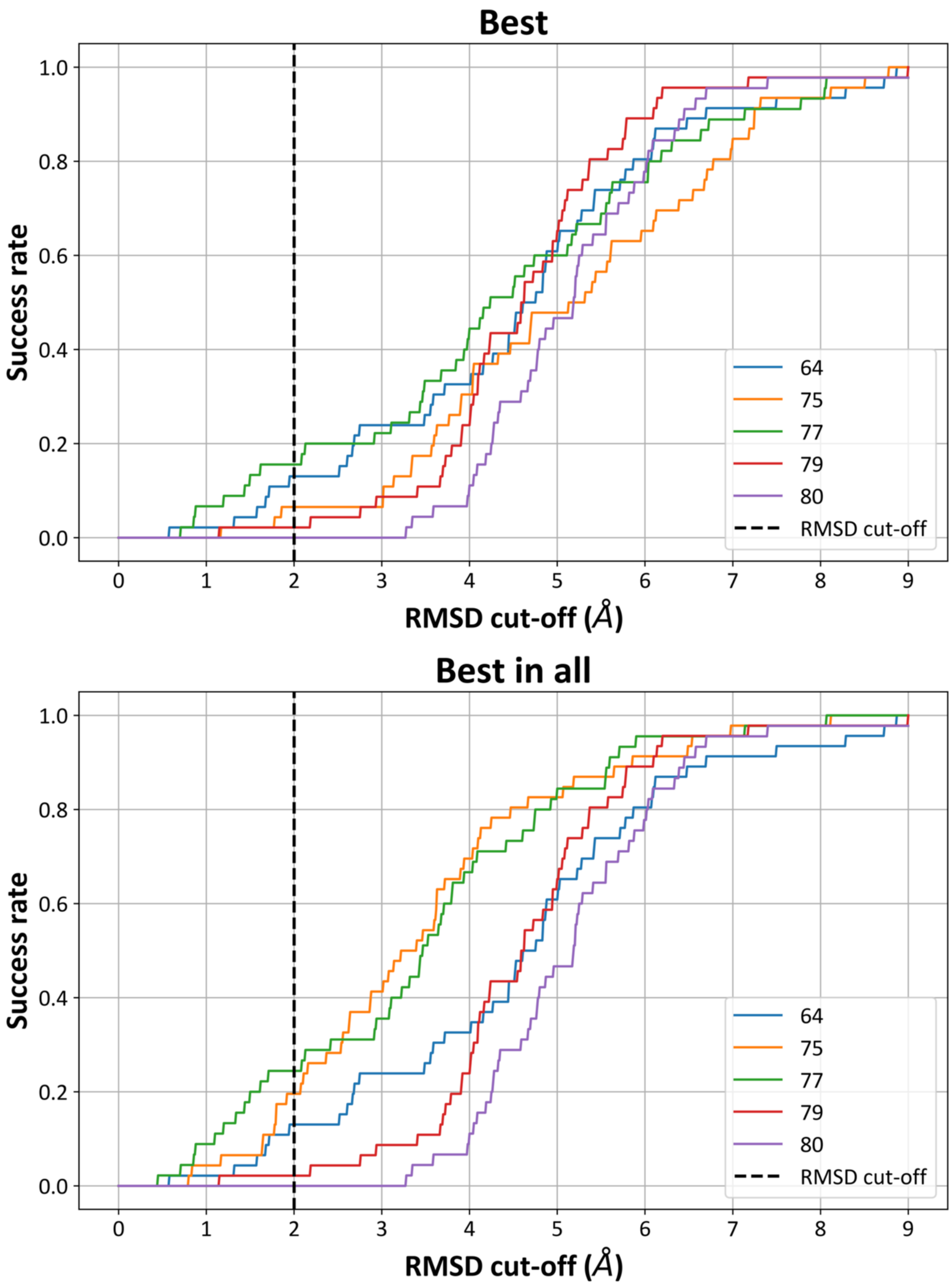


Figure 3.3: Success rate variation as a function of RMSD cut-off value. The top (Best) and bottom (Best in all) panels show the variation in success rate for the best scoring and lowest RMSD poses.

There was no difference in success rates between the Top 1 (Best) and Top 5 (Best in all) for SIDs 64, 79 and 80 as these participants only submitted one pose per fragment. This resulted in a better performance of methods 75 and 77 when the lowest RMSD pose is considered over the best scoring pose (**Fig. 3.5A**).

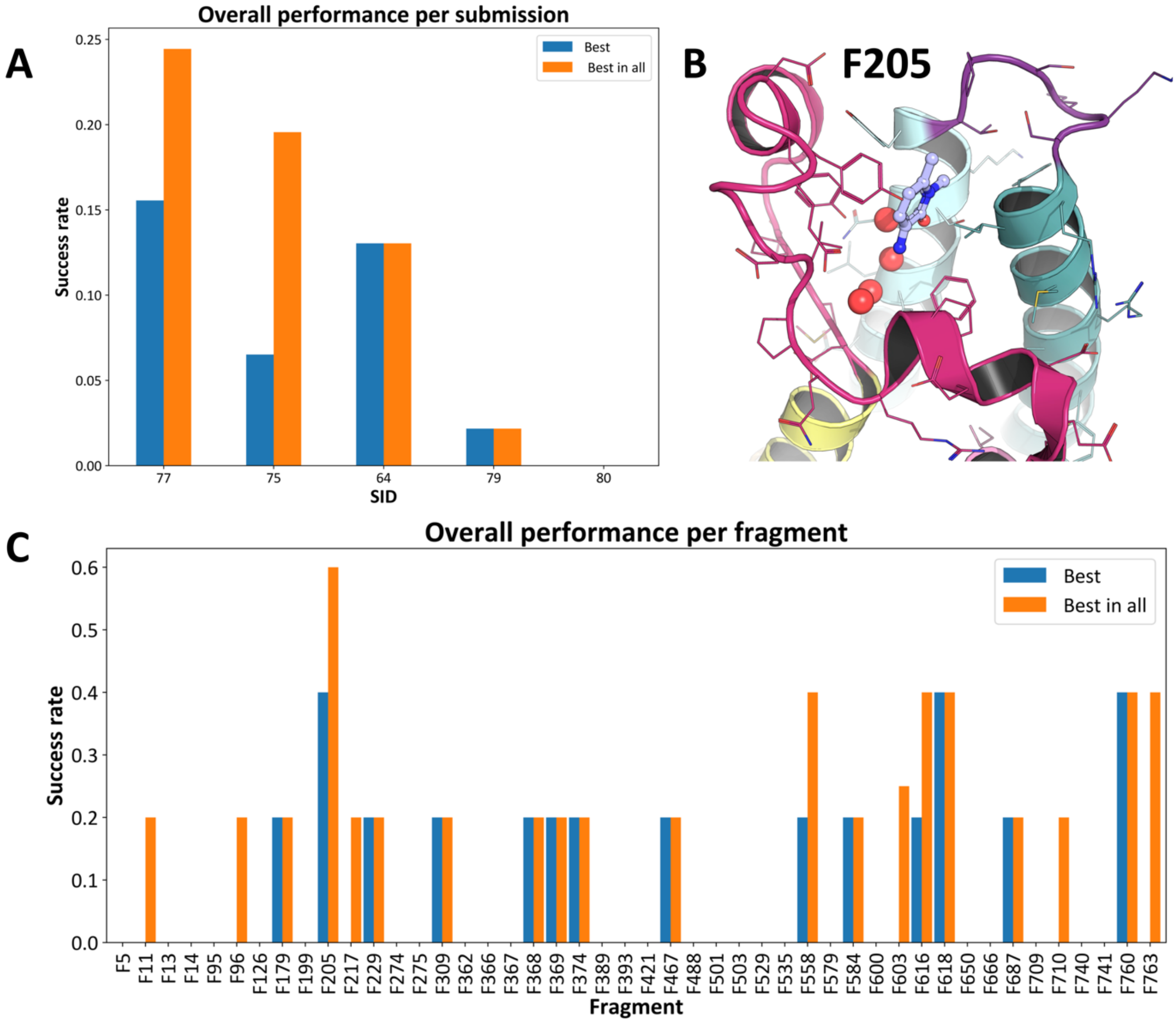


Figure 3.4: Submission's success rate at RMSD cut-off of 2Å. Submissions' success rate at a RMSD cut-off of 2Å. The top left panel (A) compares the success rate across all fragments for the different submissions. The top left panel (B) show the pose for the best predicted fragment pose, F205. The bottom panel (C) compares the success rate per fragment across all submissions. The top right panel shows the binding pose of the fragment with the highest success rate.

Overall, this stage also appeared to be challenging (**Fig. 3.5A**). The method that was the most capable to reproduce binding poses was SID 77, which performed best for both rankings. SID 77 predicted correctly 7 and 11 fragment poses out of 45 for the top 1 and top 5, respectively. SID 77 treated the systems dynamically, which resulted in binding site conformations different to the one observed in the provided C2 crystal form. However, this did not prevent them from correctly identifying the binding pose of 16% and 24% of the binders for the Top 1 and Top 5, respectively. The success of submissions 75 and 77 were significantly enhanced when looking at the Top 5 over the Top 1 which was 9 and 13 % higher, respectively. Thus, the SID 75 success experienced the most dramatic drop between the Top 5 and Top 1. For the Top 5, SIDs 75 and 77, however, performed better than the others that did not submit 5 poses implying that, in this case, the Top 1 pose is not necessarily the one with the lower RMSD (**Fig. 3.5A**). Submissions 79 and 80 performed relatively poorly with success rates of 2% and 0%, respectively.

These poor results might reflect the fact that the provided C2 apo state was not used in their workflow. Instead, other ligand-bound structures were used as receptors, which have different binding site conformations. Evidently, the scoring scheme used in submissions 79 and 80 was targeted toward different static binding site states that prevented the identification of the correct binding poses. Contrarily, submissions 64 and 75 employed the provided C2 structures which led to an improvement in the results when compared to submissions 79 and 80.

When looking at the success rates for individual fragments (**Fig. 3.5C**), no clear correlation could be robustly established given the small number of submissions, which translated into an even smaller number of predictions per ligand. Qualitative observations could, however, be made. F205 was the fragment predicted with the highest success rate for the Top 1 (**Fig. 3.5B**). F205 was predicted correctly in 3 and 2 submissions out of 5 for the top 1 and top 5, respectively. This relatively polar and flat fragment hit fills the central void and makes an H-bond with serine 1392. F558, F616, F618, F760 and F763 were correctly predicted twice for the Top 5. F763 has a similar chemotype and binding pose to F205. F558 one of the largest of the fragments that binds to 3 out of 4 subsites (the BC-interface, the central void and the ZA-channel) and fills most of the available binding site volume. Interestingly, F616, F618 and F760 all displace the 4th water molecule from the water network and their binding pose were relatively well predicted (**Fig. 3.2**). This highlights how correctly accounting for water molecules' dynamics and/ or positioning can lead to better results in binding pose predictions for fragments.

3.4.4. Stage 3: Selection of fragment follow-up compounds from a database

The goal of the third and final stage of this SAMPL challenge was to enumerate fragment follow-up compounds from a provided database. These follow-ups should target the PHIP(2) acetylated-lysine binding site and aim to be potent. The participants were provided with the co-crystal structures of the Round 2 fragment hits to support their selection. They were also given a library of more than 40 million molecules, which was a combination of the Molport "AllStockCompounds" (dating from July 2019) and a subset of the Enamine Real© library.[113]

In addition, library subsets generated using the fragment network[114] (by Dr Tim Tim Dudgeon) were also provided in case participants preferred to work with smaller numbers of compounds. The fragment network provides a convenient way to identify similar compounds. Molecules are fragmented by the removal of substituents, rings and linkers and a graph database assembled with nodes corresponding to molecules and their sub-fragments and edges corresponding to the parent-child relationships between those nodes. The search algorithm requires an input molecule and 3 parameters: 1- the number of edges to traverse from the query molecule, 2- number of changes in heavy atom count, 3- number of changes in ring atoms counts. The fragment hits were used in queries with number of edges to traverse, changes in heavy atom count and number of changes in ring atoms counts ranging between 1 and 4, 3 and 5 and 1 and 2, respectively. The resulting molecules typically correspond to variants of the query with substituents, rings and linkers added and, deleted and/or replaced.

Participants were asked to suggest between 10 and 100 follow-up compounds from the provided library, along with a confidence score, for each follow-up, between 0 and 10. The participants were encouraged to submit poses when using Molecular Dynamics- or docking-based methods. The entrants were given 1 month to submit their follow-ups.

Initially, the submitted follow-up compounds would have been purchased and validated against the C2 crystal form at the XChem by X-ray crystallography and also potentially with biochemical and/ or biophysical assays. The follow-ups and generation methods would have been examined by medicinal and computational chemists to assess whether the molecules were worthy of being bought. A total of 50 to 100 compounds were expected to be purchased given the available budget at the time of setting up stage 3.

Unfortunately, the COVID-19 pandemic resulted in a diversion of funds before validation and the challenge terminated at this point. Here, instead, the top 10 submitted follow-ups will be criticised and comment on their value as likely molecules for purchase. Only 4 submissions were received in total and summarised in **Table 3.6**.

Table 3.6: Overview of the Stage 3 follow-ups enumeration submissions. 4 sets of predictions were submitted for stage 3. The first, second, third and fourth columns correspond to the submission identification number (SID), the affiliation of the participants, the method's name and the list of software used.

SID	Participant Affiliation	Method Name	Method category	Software used
82	Peking University	XGboost_gnina_Yu	Docking, ML	rdkit xgboost gnina
85	Acellera	SkeleDock-rDock	Docking	SkeleDock RDKit HTMD rDock
86	Institut de Chimie des Substances Naturelles, CNRS, Gif-sur-Yvette, France	ranking_stage3	Docking	Schrodinger LigPrep v48012 CACTVS Chemoinformatics Toolkit V3.4.6.26 CORINA v4.2.0 CCDC GOLD v5.7.1 (CSDS-2019-1)
87	The University of Tokyo, Japan SHaLX Inc.	Template filtering then template docking	Docking	RDKit (2018.09.1) Molegro Virtual Docker (7.0.0)

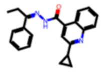
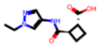
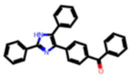
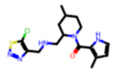
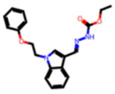
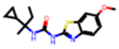
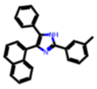
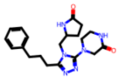
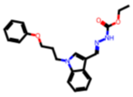
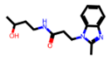
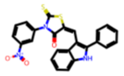
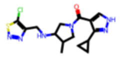
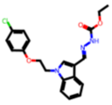
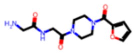
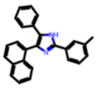
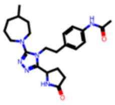
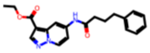
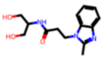
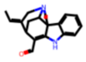
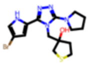
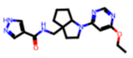
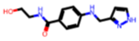
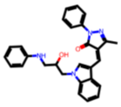
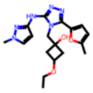
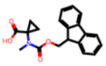
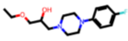
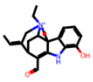
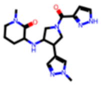
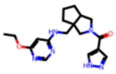
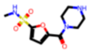
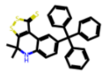
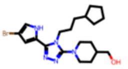
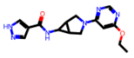
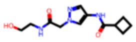
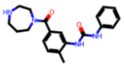
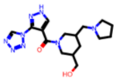
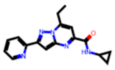
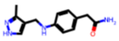
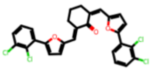
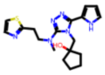
SID 82 failed to provide a complete description of the method employed for follow-up identification. However, the method appears to have filtered the library using an Xboost classifier built as part of Stage 1. The database molecules were represented as Molecular ACCess System (MACCS) fingerprints and the classifier achieved a mean area under the curve of 0.73 with 10-fold cross-validation. The top 200 scoring molecules were taken forward for docking with GNINA[84] into all the provided crystal structures. The molecules were ranked by their final mean predicted affinities.

SID 85 used a fragment-network generated library subset corresponding to 2, 5 and 2 numbers of edges to traverse, changes in heavy atom count and number of changes in ring atoms counts, respectively. They justified that those parameters would provide a good balance between similarity to the co-crystallised compounds and increased molecular size of the follow-ups which is necessary to increase the number of favourable contacts. They removed compounds that have a Dice Similarity over Morgan Fingerprints of the co-crystallised compounds smaller than 0.7. Finally, they removed compounds lacking large enough common substructures which resulted in a final set of >8000 molecules. They then used SkeleDock[111] to generate follow-up poses. This software uses maximum common substructure constrained docking over provided protein-ligand structure which were, here, the co-crystallised fragments. The poses were scored with rDock[106] and the top 100 compounds were selected and manually inspected.

SID 86 used did not filter the database in their workflow. All compounds were generated from the provided SMILES with CORINA[100] and protonated at physiological pH with Schrodinger LigPrep. Single docking poses were generated in GOLD and scored with ChemScore using a ligand efficiency of 30% (against 200% in Stage 1) and top 100 scoring compounds submitted.

SID 87 applied filtering to the whole database based on chemical descriptors. Selected compounds should have between 10 and 34 heavy atoms, more than 2 hydrogen bond donors, between 50 and 100% of the heavy atoms in a SP3 hybridisation form, between 3 to 10 rotatable bonds, between 2 and 3 aromatic rings, 1 or more aromatic heterocycle, between 1 and 2 aliphatic rings, no more than 4 rings and a molecular weight between 360 and 440 Da. They further filtered out compounds containing amide bonds except tertiary amides and lactams. They built a pharmacophore in Molegro Virtual Docker[112] by merging the poses of F367, F389, F558 and F584 which cover all areas of the acetylated-lysine binding site. The pharmacophore was used to apply template filtering to the selected compounds and the top 10% of those were selected. Finally, the poses were scored against the C2 apo crystal structure and the top 16 structures were submitted.

Table 3.7: Chemical structures of the top 10 compounds submitted in Stage 3 follow-ups enumeration submissions. The column and row labels indicate the SID and ranks, respectively. The Molport or Enamine Real identifiers are shown below each 2D molecular representation.

Rank:	SID: 82	85	86	87
1	 MOLPORT:000-796-582	 REAL:Z2897472896	 MOLPORT:000-800-360	 REAL:PV-000076313713
2	 MOLPORT:007-564-137	 REAL:PV-001941530881	 MOLPORT:001-732-911	 REAL:PV-002045872457
3	 MOLPORT:002-568-963	 REAL:Z1727122569	 MOLPORT:002-231-584	 REAL:PV-000062484624
4	 MOLPORT:000-827-347	 REAL:Z2433636166	 MOLPORT:001-732-911	 REAL:PV-002002900770
5	 REAL:PV-001818177585	 REAL:Z1727485589	 MOLPORT:000-702-788	 REAL:PV-002045938857
6	 REAL:PV-000010447219	 REAL:Z1975353425	 MOLPORT:000-815-332	 REAL:PV-002030117876
7	 MOLPORT:044-730-763	 REAL:Z1195969280	 MOLPORT:000-701-509	 REAL:PV-000075137018
8	 REAL:PV-000011799512	 REAL:Z927289330	 MOLPORT:000-784-522	 REAL:PV-002045955111
9	 REAL:PV-000019326486	 REAL:Z1614521200	 REAL:Z2439864311	 MOLPORT:027-823-837
10	 MOLPORT:010-790-580	 REAL:Z751931838	 MOLPORT:008-322-386	 REAL:PV-002034310941

Expectedly, all submissions showed an increase in all molecular descriptor counts (**Table 3.8**), with respect to the provided binding fragments, which is paired with an increase in molecular weight. Similarly, the Tanimoto coefficient distributions are shifted toward 1 for each submission indicating that follow-ups have a higher degree of chemical similarity than the native fragments. The intrinsic molecular size bias of the Tanimoto coefficient may also participate in the shift. The larger the molecules the larger number of features present in the binary array and is reflected by a higher similarity.[115] The top 10 molecules for each submission are displayed in **Table 3.7**.

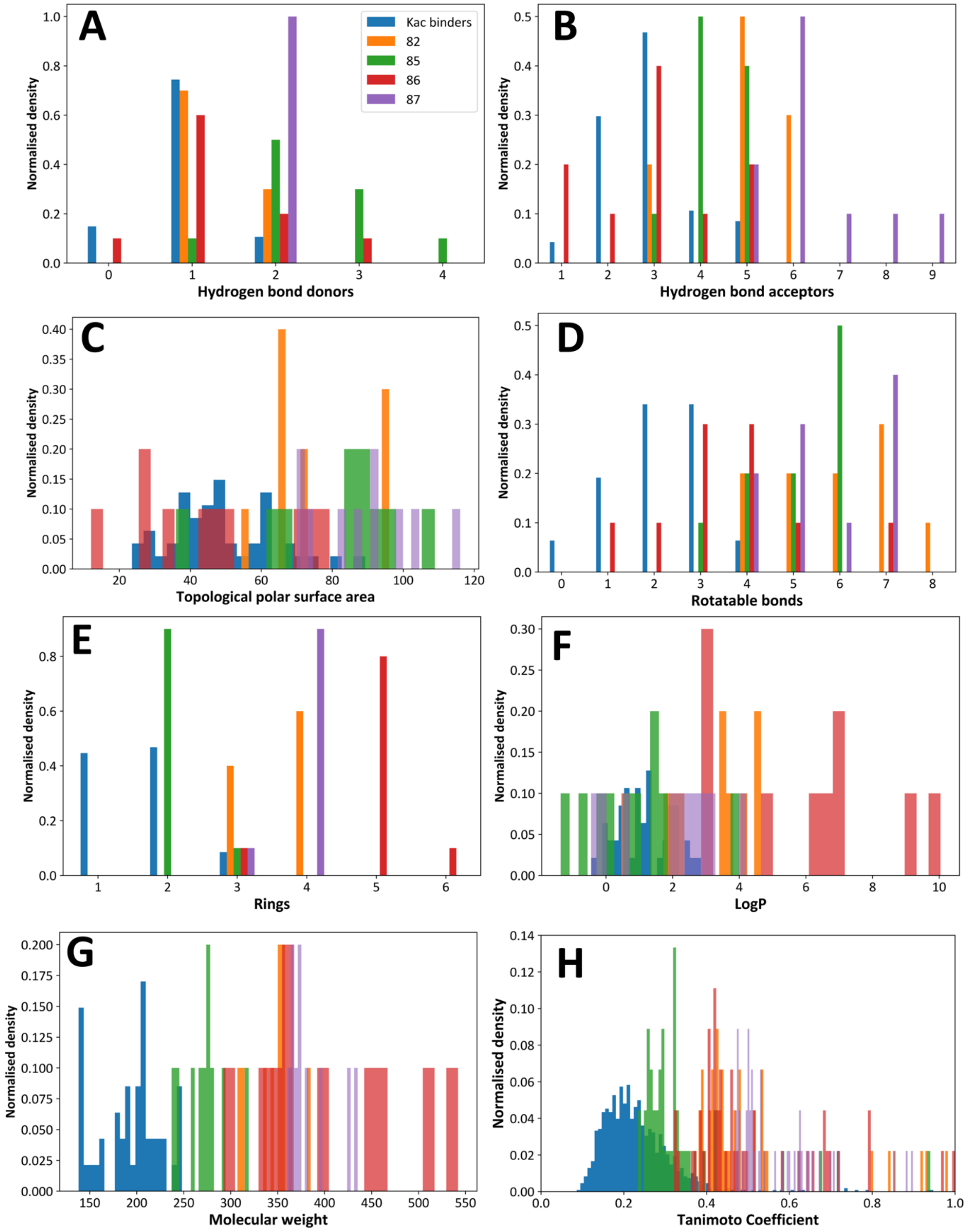
SID 82 increased the number of rotatable bonds and molecular weight. However, the other descriptors were moderately raised. This implies that these molecules would be relatively more flexible for a low number of H-bond donor and acceptors implying a potentially high entropic penalty of binding for a low enthalpic compensation. Some molecules also showed a high degree of chemical similarity to each other with a Tanimoto coefficient higher than 0.8 (**Table 3.8**). Molecules ranked 2 to 9 are too flexible to be selected for validation. Molecules ranked 1 and 10 could potentially have good scaffold sampling all areas of the binding site while maintaining reduced flexibility relatively to other compounds in that submission (**Table 3.7**).

SID 85 increased relatively largely the H-bond donors and acceptors whilst maintaining the number of rings and molecular weight relatively low which had a positive and negative impact on the topological surface area and clogP, respectively (**Table 3.7**). Reducing hydrophobicity in that fashion may be a suitable strategy, however, PHIP(2) and bromodomain in general have a relatively hydrophobic binding site implying that more hydrophobic compounds are likely to have an increased affinity. In addition, some compounds have low feature counts keeping them in a fragment-like category. The compounds are also relatively diverse when compared to other submissions. No molecules from that submission would have been selected for biophysical validation and they are too small and/ or too similar to the fragment hits and would likely not significantly increase binding affinity. For example, molecules 2, 3, 4, 5 and 7 are almost identical to fragments 579, 217, 709, 217 and 710 (**Table 3.8**).

SID 86 increased the most the number of rings without a similar increase in polar features in the follow-ups which resulted in large clogP and molecular weight against a low topological polar surface area indicative of greasy compounds (**Table 3.7**). Only 4 out of the top 10 were rule of 5 compliant. Although those compounds would probably sample well the predominantly hydrophobic binding site, they would lead to aggregation problems in solution. Only compounds 5, 7 and 9 have an acceptable predicted clogP. 5 and 7 are extremely similar so one of the two would be selected along with compound 7 for further testing (**Table 3.8**).

Finally, the molecules submitted by SID 87 are relatively similar and have relatively large and varied numbers of H-bond donors and rotatable bonds while the count of ring and H-bond acceptors remain low (**Table 3.8**). They are almost all based on a central ring that expands into 3 different directions to potentially sample the whole binding site (**Table 3.1**). This appears to be suitable strategy, but care should be given to keeping molecules flexibility to a minimum. For example, molecules 2, 8 and 10 display a high number of rotatable bonds. Molecules 1 and 3 would not be able to sample the water cavity. Thus molecules 4, 5, 6 and 7 would have been selected for biophysical validation (**Table 3.7**).

Table 3.8: Molecular descriptors for acetylated-lysine site fragment elaborations and top 10 for each submission. The plots show the normalised density of a molecule having a descriptor value for a given set (top 10 follow-ups for submission or Kac fragment binders). A, B, C, D, E F, G show the distributions of H-bond donors, H-bond acceptors, rings, and rotatable bonds counts, respectively. Panel H shows the distribution of Tanimoto coefficients for all against all molecules within the sets.



3.5. Discussion

The application of computational methods is now an essential part of almost all fragment-based drug discovery pipelines. These tools are, however, better suited for larger molecules and there is a lack of prospective evaluation in the fragment space. To that end, a unique opportunity arising from the PHIP(2) work presented itself. The crystallographic fragment screening resulted in the identification of 52 binders across 4 sites (**Fig. 3.1**), 47 of which were bound to the pharmacologically relevant acetylated-lysine binding site (**Table 3.1**) translating to hit rates of 5.88% and 6.50%, respectively. Crude analysis of historical Xchem data shows that hit rates vary between approximately 2% and 15% with an average of 7%, per protein target.

Stage 1 assessed the ability of submitted methods to discriminate crystallographic binders and non-binders. A variety of protocols were submitted with techniques involved ranging in computational expense and complexity (**Table 3.4**). Some methods applied the correct bias strategy. All submissions but 1 predicted a large excess of non-binders which is in accordance with the experimental binder to non-binder ratio (**Fig. 3.3**). Few studies have prospectively assessed virtual screening of both fragment-like and larger molecules. Notably, SAMPL3 focused on trypsin binders and built a dataset of 544 fragments against which some bound structures were resolved, and affinity measurements were collected by ITC and SPR.[116] 20 of these 544 fragments were considered as true binders. SAMPL3 virtual screening performance was better than with the PHIP(2) fragments with submissions achieving an areas under the curve of up to 0.8 with the best submission employing a similar docking protocol to SID 52.[117]

SAMPL4 focused on fragment follow-ups that bind to the HIV integrase. They built an elaboration library around 4 fragments in an attempt to design a more potent inhibitor implying that the subsequent follow-ups share a similar scaffold and therefore have some degree of chemical similarity.[118] Compounds resolved with X-ray crystallography and did not bind better than 2 mM, based on SPR measurements, were classified as inactives. A total of 305 compounds, including 56 actives, were put forward for the challenge. 5 submissions out of 26 achieved an area under the curve equal or better than 0.6 which is, again better than the predictions presented above (**Fig. 3.5**).[119] It is hard to precisely rationalise the poor results of the Stage 1 submissions (**Fig. 3.3**).

One notable difference between this task and SAMPL3 and SAMPL4 is that the classification was solely based on crystallographic binding. This was emphasised when setting up the stage, by suggesting that participants should aim to reproduce the crystallographic screening results. This may be a more important consideration than first assumed as there is the possibility that the PHIP(2) fragments do not necessarily bind in solution, whereas scoring functions are mostly calibrated and validated against affinity and structural data.[120] Furthermore, the fragments also have relatively more similar physico-chemical characteristics than larger compounds (**Table 3.2**) which may further hinder predictions due to a relatively reduced number of discriminatory features.

Stage 2 was a binding pose prediction exercise - a task that is more frequently assessed than virtual screenings. Only 5 ranked submissions were received for this stage of the challenge (**Table 3.5**). This low number of participations was likely due to the short timeframe imposed; only 2 weeks (due to internal pressures on Stage 3). Although this stage was shorter than initially planned, the small number of submissions suggests a lack of flexible, fast and automated workflows available. The 5 submissions received did not perform well, stressing the difficulty and potential scope for improvement regarding fast 3D binding pose prediction for fragments (**Fig. 3.5**). The best performance was SID 77 and employed a biased MD-based protocol (**Fig. 3.5**) providing an example where a protocol treating protein and water as flexible can outperform rigid dockings when trying to reproduce static crystallographic data. The second and third best performing methods docked the compounds into the provided apo structure. The two worst performing methods picked different fragment-bound PHIP(2) structures from the PDB. This show and example where binding pose predictions for fragments by docking-based methods, can be sensitive to the chosen receptor conformation. Additionally, the availability of alternative ligand-bound structures does not consistently enhance prediction quality (**Fig. 3.5**).

Another challenging aspect of docking to PHIP(2), and bromodomains in general, is that only some fragments retain the conserved water network. This implies that the choice of retaining or removing these waters will have an adverse impact on some predictions.[121] For example, SID 75, kept all structural waters in their protocol, which made impossible the correct prediction of water-displacing fragments such as F584 (**Fig. 3.2**).

SAMPL4 also assessed binding pose prediction of a similar number of ligands. Only one third (3 out 15) of the submissions performed better than the predictions presented in this work.[119] Here most ligands were incorrectly predicted in all submissions as opposed to SAMPL3 where most ligands were correctly predicted in at least 1 submission, but this is likely due to the relatively higher number of submissions in SAMPL3.[117,119] D3R grand challenges also assess blinded predictions of binding pose as well as other quantities meaningful to drug discovery. D3R grand challenge 4 submissions achieved excellent results for binding pose predictions of macrocycles against BACE1.[122] The good performances were partially attributed to the presence of many ligand-bound crystal structures in the PDB which could guide docking results.

The most obvious difference with the data presented here is the ligand size, thus suggesting, that docking of larger ligand may be an easier task than smaller fragment-like molecules although correct sampling of the large number of ligand conformations may present issues. Fragments typically have relatively low potencies, in the µM to mM ranges, suggesting that their free energy surfaces are relatively flat with shallow minima potentially associated with alternative binding poses and therefore, rendering discrimination tasks more difficult.

Stage 3 of this challenge was atypical. It consisted in selecting compounds from a large database. To the best of available information, this was the first challenge of the kind and compounds would have been initially screened by X-ray crystallography at the XChem with potential additional assays. Unfortunately, this stage did not go forward because of the COVID-19 pandemic. The participants were given 1 month to submit up to 100 molecules, and only 4 submissions were received (**Table 3.6**) which, like stage 2, indicates a lack of responsive workflow despite this potentially unique opportunity to test follow-up enumeration methods in a truly prospective way. One striking submission from the top 10 compounds was within SID 86 where the library was not filtered, and all compounds docked (**Table 3.6**). This resulted in large molecules with larger clogP values that in most cases violated the rule of 5 (**Table 3.8**). Such molecules would not be viable candidates for screening and would likely not have been tested experimentally which illustrates that appropriate library or pose filtering is necessary. The workflows employed in this stage were computationally cheap compared to what is available and more routinely applied such MD-based screening or free energy calculations.

3.6. Conclusion and future directions

To the best of available information, no inhibitor has been discovered against PHIP(2), despite its implicating in cancers[55] thus legitimising the crystallographic effort. This dataset presented an ideal scenario to set up a SAMPL challenge to explore the capability of computational tools for exploring fragment binding. The challenge was initially divided into 3 stages: 1) virtual screening, 2) binding pose prediction and 3) enumeration of fragment follow-up compounds from a database. All stages of this SAMPL edition were clearly challenging. What was learned from this?

Most submissions did not directly aim to reproduce experimental results by mimicking crystallographic conditions and binding although this was strongly emphasised during the challenge. Chemical factors were not taken into consideration despite the detailed experimental protocol being provided. For example, some participants protonated the protein and ligands assuming a physiological pH of 7.4 although it was highlighted that PHIP(2) crystals were grown around pH 4.6.

No ligand-bound structural information with this C2 crystal form was available prior to the start of the challenge. Thus, participants employed fragment-bound PHIP(2) structures crystalised in other space groups to guide their predictions. Given that using various ligand-bound structures showed improved predictions in previous challenges this was a sensible move. However, previous challenges considered larger, non-fragment molecules. In this challenge, using other fragment-bound structures appeared to have had a negative effect on binding pose predictions (**Fig. 3.5**). Indeed, small conformational variations between binding sites may significantly change the energy landscape sampled by the docking algorithms. For example, the C2 and $P2_12_12$ crystal forms show conformational variation at the BC- and ZA-loops (**Fig. 3.1**), where amino acids important in binding, such as tyrosine 1350 or threonine 1396 have different orientations. These structural differences at the binding site would explain the impact on fragment docking performances as they have a direct impact on binding site volume and available interactions. Thus, it is perceived that, in this case, using different fragment-bound crystals was a mistake and computational identification of crystal binders should be preferably performed against the targeted crystal form.

Another important factor may be that scoring methods are normally calibrated based on existing structures paired with affinity data[123], whereas the PHIP(2) fragments benchmark was only based on crystallographic results, and it are presently unknown if these bind into solution. Thus, commonly used methods may not be able to detect nor reproduce such weak crystallographically observed events. Other target specific considerations, such as the important water networks, likely also participate in making this SAMPL7 challenge difficult. Overall, performance of those that did enter illustrates that this is difficult to apply computational predictions on fragments in a prospective way at least in the case of this target.

Despite the relatively poor performance of the predictions, this work illustrates that there is plenty of scope of improvement in this area. This also implies that experimental procedures are still heavily needed to identify crystallographic fragment binders. The small number of submissions received, particularly for Stage 2 and 3 (**Tables 3.5 and 3.6**) also highlights the lack of responsive workflows although the short timescales were designed to reflect real-world conditions.

The potential avenues to improve such predictions may include better treatment of water molecules. This could be done by selecting strongly interacting water by using MD-based methods or Grand Canonical Monte Carlo analysis[124]. Alternatively, creating an ensemble of receptors that include all possible combinations of water networks within and surrounding the binding site may also be useful. Identification of relevant conformational changes may also be achieved via MD-based methods or flexible docking. Including crystal symmetry mates around the binding site may also improve prediction by mimicking crystal conditions more precisely.

Fragments make few but usually “high-quality” interactions and thus an explicit stage that captures those might well improve confidence. This could be done via hotspot or pharmacophore rescoring.[29,125] Creating tailored scoring functions trained on fragments and crystallographic data may yield better results.[126] Those improvements may also be applied to the selection of fragment follow-ups, which would benefit from better library filtering tools.

Future challenges may also benefit from participants submitting the scores associated with different molecules and poses for virtual screening and binding pose prediction, respectively. This would allow the investigation of the energy surface, for example by plotting screening rank or RMSD against docking score. An additional method for scoring binding pose predictions could be to compare the protein-ligand interaction fingerprint of the docking model with the experimentally determined ligand. This approach is often employed in assessments of protein-protein docking studies[127] and may be effective in identifying significant binding site interactions. The evaluation could be quantified through similarity calculations, such as the Tanimoto coefficient, comparing the modelled and experimental fingerprints.

Overall, this work emphasises the need for more prospective evaluations and the development of improved computational workflows for computational fragment identification.

3.7. Methods

3.7.1. Crystallographic screening

Proteins were expressed, purified and crystalised as described in Chapter 2, Section 2.2.1. Crystallographic screening was also performed as described in Chapter 2, Section 2.2.2. The crystals were soaked for 2 hours at a final fragment concentration of 20 mM. The fragments used for screening were purchased from Enamine stored in ethylene glycol at a concentration of 100 mM. Structural models and crystallographic statistics can be retrieved from the PDB deposition ID: G_1002162.

3.7.2. Small molecule featurisation and distance calculations

Small molecules SMILES strings were used as input in featurisation and fingerprinting after removal of salt and ion atoms. Standard RDKit fingerprints were generated as bit vectors using default setup.[128] RDKit was used to extract the number of hydrogen bond donor, hydrogen bond acceptors, rings and rotatable bonds. RDKit was also used to estimate topological surface areas, molecular weights, and octanol–water partition coefficients. Chemical similarity calculations of molecule pairs were performed using RDKit and were based on the Tanimoto coefficient between RDKit fingerprint bit vectors.[79]

3.7.3. RMSD calculations

Symmetry-corrected RMSD calculations for submitted binding pose prediction heavy atoms were performed with spyrmsd[129] against the experimentally resolved fragments. Predictions were considered successful if the RMSD was equal to or lower than 2 Å.

3.7.4. Statistical analysis

All statistical values were calculated with scikit-learn. Due to large class imbalance in favouring the negative (non-binding) class, the balanced accuracy was employed to evaluate virtual screening performances. The balanced accuracy is mean of the sensitivity and specificity which correspond to the true positive and true negatives rates (TPR & TNR) respectively.

$$\textbf{Sensitivity} \qquad \mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \qquad \textbf{Equation 3.1}$$

Specificity $$\mathrm{TNR} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}$$ **Equation 3.2**

Balanced accuracy $$\mathrm{BA} = \frac{\mathrm{TPR} + \mathrm{TNR}}{2}$$ **Equation 3.3**

Where TN, TP, FN and FP correspond to the number of true negatives (correctly classified negative cases), true positives (correctly classified positive cases), false negatives (incorrectly classified negative cases), and false positives recovered by the classifiers (correctly classified positive cases), respectively. Balanced accuracy values increase with model quality.

3.8. Data availability

All the experimental data, participant's predictions, code for analysis and resulting output is available on GitHub.

https://github.com/samplchallenges/SAMPL7/tree/master/protein_ligand

Chapter 4. High-throughput X-ray crystallography for rapid fragment growth from crude compound arrays by low-cost robotics

4.1. Credits

This chapter is the outcome of a multidisciplinary and collaborative endeavour. Harold Grosjean took responsibility for protein expression, purification, and crystallisation, as well as structure refinement, electron density analysis, and cheminformatic analysis. Furthermore, Harold Grosjean drafted the manuscript and assembled the figures. Follow-up synthesis, crude reaction mixtures preparation and crystallographic screening were conducted by Dr Anthony Aimon and Dr Storm Hassell-Hart under the supervision of Prof John Spencer and Prof Frank von Delft. Dr Tobias Krojer, Dr Anthony Aimon and Dr William Bradshaw helped with model building and refinement. Dr Warren Thompson developed and benchmarked the automated quality control analysis software. Dr Storm Hassell-Hart carried out estimations of solvent and time gains. The grating-coupled interferometry assay was carried out by Dr Lizbé Koekemoer and Dr Edward A. FitzGerald, while the Alpha-screening assay was performed by Dr James Bennett. Cameron Anderson contributed to the cheminformatic analysis. Prof Philip C. Biggin, Prof John Spencer and Prof Frank von Delft reviewed and provided guidance on the manuscript. All involved participated in the redaction of the manuscript.

4.2. Preprint

The research conducted in this chapter led to a preprint, which can be accessed using the following reference:

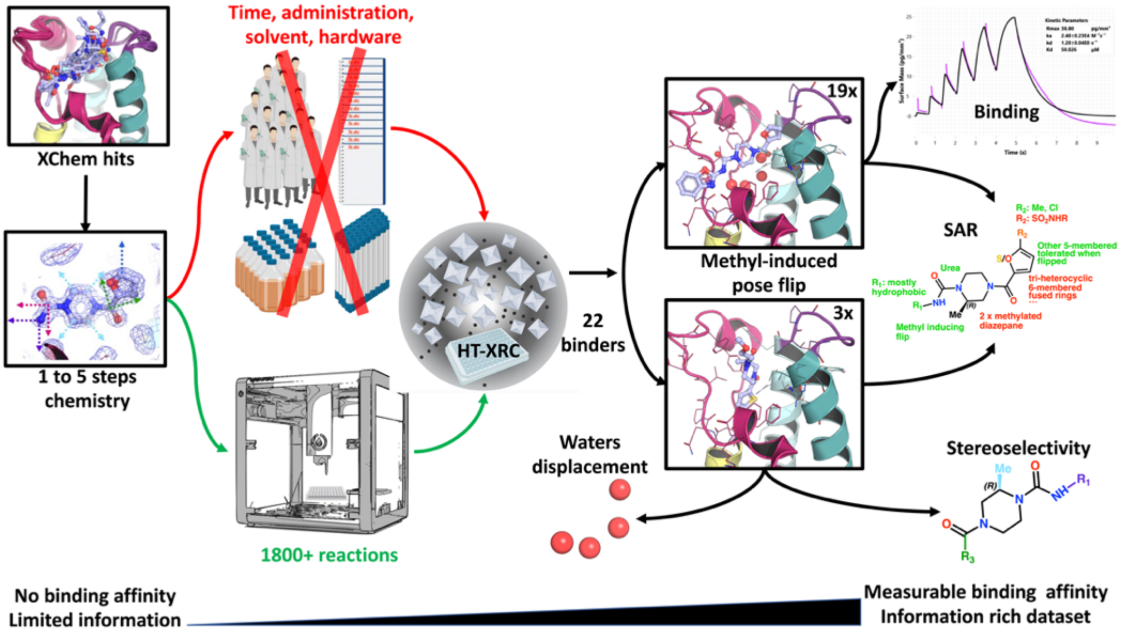
Grosjean, H. *et al.* (2023) “High-throughput crystallography for rapid fragment growth from crude arrays by low-cost robotics,” *ChemRxiv* [Preprint].

Available at: https://doi.org/10.26434/chemrxiv-2023-6m2s0.

4.3. Introduction

In Chapter 1, novel fragment binders were identified against PHIP(2) using X-ray crystallographic screening. It was also demonstrated that computational prediction of those fragment is a challenging task thus suggesting that efforts should instead be focused on generating follow-up compounds. Chemical synthesis is a costly and slow process presenting a bottleneck in fragment elaboration to potency.

In this chapter, hundreds of follow-up compounds were synthesised by low-cost robotics and crude reaction mixtures resolved by X-ray crystallographic and analytic techniques. This provides an optimised way of generating extensive structural Structure-Activity data bypassing small molecule purification requirements therefore saving the use of polluting solvent and time. Starting from a fragment, F709, resolved in Chapter 1, a collection of more than 1850 compounds was enumerated, and thousands of crude reaction mixtures were synthesised, at the success rate of about 60%, on one OpenTrons robotic liquid handler. Crystallographically, 969 usable X-ray diffraction datasets were acquired. This resolved 22 unique reaction products binding to the protein, 19 with conserved poses relative to the original fragment and 3 with a new, unexpected binding pose with 1 structural hit also resolved in assays. Crystallographic and cheminformatic analysis showed that a single methyl addition at a chiral centre caused the change in binding pose. This Chapter address bottlenecks in fast follow-up generation and validation.



Graphical introduction 4.1: Array synthesis optimises follow-up generation by lower time and solvent requirement and yield information rich dataset.

4.4. Results

4.4.1. Fragment F709 offers opportunities for robotic elaborations

Previous X-ray crystallographic fragment screening efforts presented in chapter 3 yielded multiple hits against PHIP(2). These experiments were used in the assembly of the DSI-poised and FragLites fragment libraries[12,13], and the fragment hits used as the starting point for the SAMPL7 challenge and for this automated chemistry study.[130] One fragment, F709 (**Fig. 4.1)**, displayed interesting characteristics including clear electron density at the binding site and expansion vectors displaying chemical moieties enabling parallel robotic chemistry. This would allow sampling of the chemical landscape around those vectors to potentially extract SAR from different elaboration series. Overall, these vectors enabled fragment growth around 3 different regions of the binding site defined in Chapter 3: i) across the ZA-channel, ii) within the central hydrophobic cavity and iii) opposite the water cavity (**Fig. 4.1**).

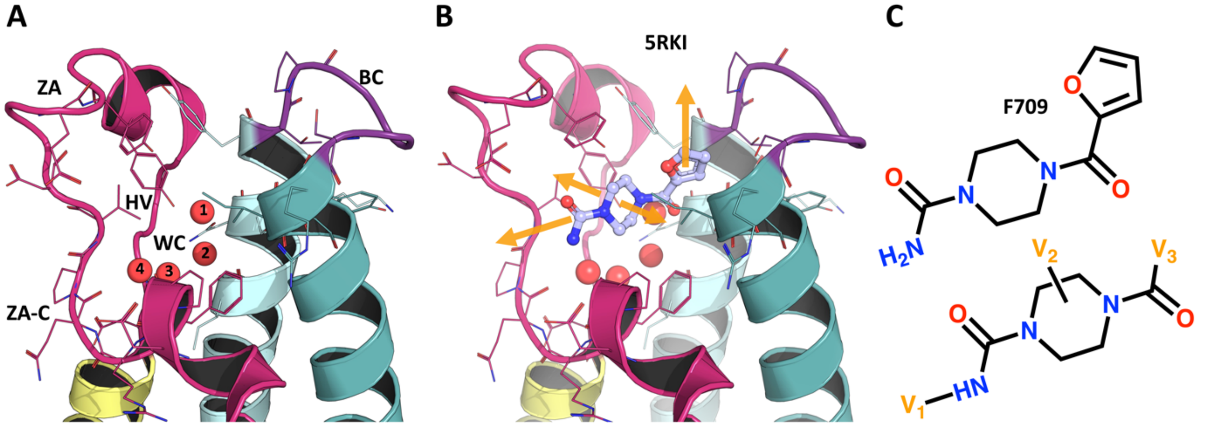


Figure 4.1: Crystallographic screening of PHIP(2) identified a fragment with good vectors for automated chemistry elaborations. Panel A shows the structure of PHIP(2) acetylated-lysine binding site with the bromodomain waters displayed in red spheres. Helices Z, A, B and C are shown in pink, yellow, cyan and teal, respectively. ZA (in fuchsia) and BC (in purple) indicate the connecting loops between the corresponding α-helices. ZA-C, WC and HV indicate the ZA-channel, water cavity and hydrophobic void, respectively. Panel B shows the F709-bound co-crystal structure (PDB ID: 5RKI) with elaboration vectors in orange. Panel C shows the chemical structure of F709 and elaboration vectors in orange.

F709 interacts with the protein notably via H-bonds with the backbone oxygen of proline 1340 and the side chain oxygen of serine 1392. The 5-membered furan ring is also positioned perpendicularly to tyrosine 1395 thus promoting a π-π interaction and the piperazine ring occupies the central hydrophobic cavity (**Fig. 4.1)**.

4.4.2. Performing complex reactions robotically with minimal resources

To determine feasibility, reactions were initially tested and developed (by Dr Anthony Aimon and Dr Storm Hassell-Hart) using standard bench chemistry with a specific focus on solvent selection and solubility. Dimethylacetamide (DMA) proved to be a general solvent and was used in all the chemistry developed and executed on the OpenTrons. Thereafter, the reaction conditions were translated to the OpenTrons as a liquid transfer-only process (done by Dr Anthony Aimon and Dr Storm Hassell-Hart).

The manual Quality Control (QC) analysis of targets required 3240 samples analysed, with a time of 7 minutes required per sample for LCMS analysis (**Fig. 4.3**). This manual process was a bottleneck in assessing high-throughput synthesis success, so an open-source mass spectrum peak finding tool, MSCheck (developed by Dr Warren Thompson). Briefly, MSCheck searches for different parent ion matches, analyses mass spectrum patterns, and generates a report. MSCheck had an 82.9% recall when benchmarked against manual processing in the retrospective analysis of post-work-up runs. MSCheck took five days to complete (**Fig. 4.3**), but the time can be optimised to within a day.[131]

To test the validity of using the OpenTrons for chemistry, three iterations (Iterations 1.0, 1.1 and 2.0) of single and two-step chemistry for the formation of urea and amides were explored (**Fig. 4.2**) (done by Dr Anthony Aimon and Dr Storm Hassell-Hart). 180 single-step reactions were executed, with a combined success rate of 83%, as measured by the identification of the expected molecular ion peak by MSCheck. Iteration 1.1, exploring a “deletion” strategy, by effectively replacing the piperazine central core by a morpholine and removing the furan-2-carbaldehyde, and Iteration 2.0, exploring an amide linkage against the urea in Iteration 1.0, yielded no structural hits (**Fig. 4.2**).

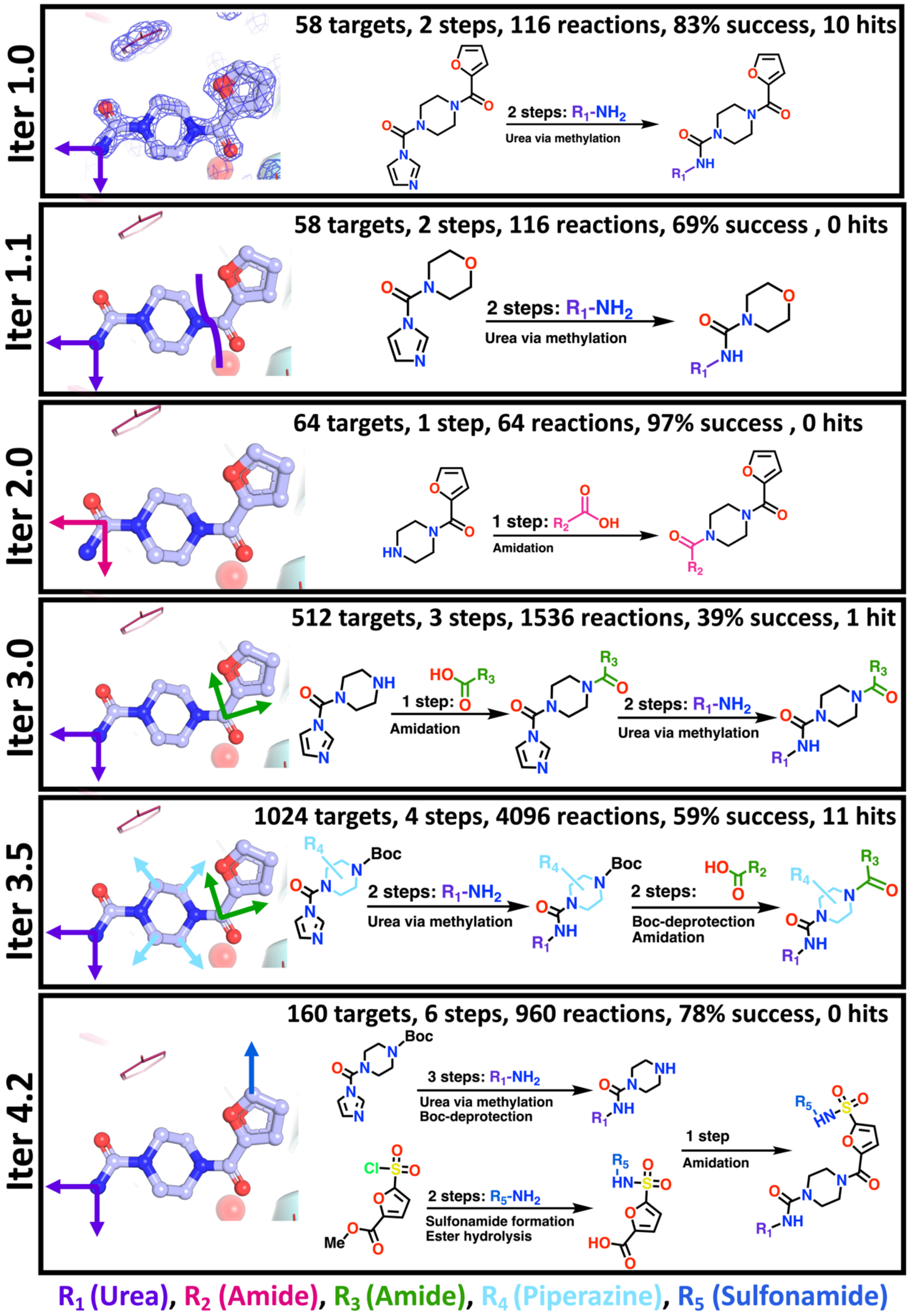

Iter 1.0
58 targets, 2 steps, 116 reactions, 83% success, 10 hits
2 steps: R1-NH2
Urea via methylation
Iter 1.1
58 targets, 2 steps, 116 reactions, 69% success , 0 hits
2 steps: R1-NH2
Urea via methylation
Iter 2.0
64 targets, 1 step, 64 reactions, 97% success , 0 hits
1 step: R2
Amidation
Iter 3.0
512 targets, 3 steps, 1536 reactions, 39% success, 1 hit
1 step:
Amidation
2 steps: R1-NH2
Urea via methylation
Iter 3.5
1024 targets, 4 steps, 4096 reactions, 59% success, 11 hits
2 steps: R1-NH2
Urea via methylation
2 steps:
Boc-deprotection
Amidation
Iter 4.2
160 targets, 6 steps, 960 reactions, 78% success, 0 hits
3 steps: R1-NH2
Urea via methylation
Boc-deprotection
1 step
Amidation
2 steps: R5-NH2
Sulfonamide formation
Ester hydrolysis
R1 (Urea), R2 (Amide), R3 (Amide), R4 (Piperazine), R5 (Sulfonamide)

Figure 4.2: OpenTrons synthesis allows rapid and reliable chemistry for initial SAR scoping via fragment growth. Summary of the different synthetic routes (iterations, Iters) and are shown in each panel. The number of fragment growth targets, steps performed on the robot, total number of reactions performed for each step, success rates for those reactions and crystallographic hits are displayed at the top of each box. The success rates are defined by whether a product with expected mass can be detected by the quality control pipeline. The functional groups exploited at the various vectors are colour coded with labels bellow the figures and displayed with respect with the original fragment on the left.

Building on the single-step iteration, three follow-up iterations were achieved (done by Dr Anthony Aimon and Dr Storm Hassell-Hart) with a total of 1696 multi-step targets, via a combined total of 6592 reactions, and a combined success rate of 54.37% (**Fig. 4.2**). Iteration 4.2 was the most complicated synthesis attempted on the OpenTrons with 160 attempts of two by two-step routes being combined towards the final product with an overall 78% success rate (**Fig. 4.2**). Iteration 4.2 showed that it is possible to use relatively complex multi-step chemistry to yield significant fragment elaborations.

Table 4.1: The automated chemistry workflow reduces time and solvent usage compared to human operations. Comparison of the estimated time, columns/workups, and litres of solvent, between the automated and manual synthetic approaches to the target libraries. (Red = Manual synthesis, Green = Automated Synthesis).

Iteration	Number of targets	Time per Analogue (days)	Total Time (days)	Manual Columns and workups per target	Total columns and workups	Solvent per Analogue	Total Solvent Volume
1	58	3	174	1	58	1 L	58 L
1	58	-	3	-	1	5 mL	290 mL
1.1	58	3	174	1	58	1 L	58 L
1.1	58	-	3	-	1	5 mL	290 mL
2	64	1	64	1	64	1 L	64 L
2	64	-	2	-	1	5 mL	320 mL
3	512	5	2560	2	1024	2.5 L	1280 L
3	512	-	4	-	2	10 mL	5.12 L
3.5	1024	5	5120	2	2048	2.5 L	2560 L
3.5	1024	-	4	-	2	10 mL	10.28 L
4.2	160	8	1280	3	480	4.5 L	720 L
4.2	160	-	4	-	3	15 mL	2.40 L
Total number of targets	1880	**Total working years**	25.68/ 0.05	**Total columns and workups**	3732/ 10	**Total litres of solvent**	4740/ 18.70
Total Saving		**Time**	**25.63 years**	**Columns and workups**	**3722**	**Litres of Solvent**	**4721.3**

Based on the manual synthesis and purification of control compounds (done by Dr Storm Hassell-Hart), it was estimated that the human time required to manually prepare an analogue varies between 1-8 working days, depending on the number of synthetic steps and purifications (**Table 4.1**). If the synthesis of the complete library was conducted in a linear fashion this would take about 25 working years to complete.

The use of automated synthesis also avoids the requirement for large-scale aqueous workups and purifications. Synthesis of the library under classical conditions would require up to 3740 chromatography stages, which when combined with the work-up stages results in a total solvent usage of 4740 L. In contrast, the automated protocol avoids the requirement for purification, with an estimated total solvent usage of 5-15 mL (**Table 4.1**) per analogue. This would require a total of 18.7 L of solvent, an over 250-fold reduction in costly and potentially polluting solvent usage.

4.4.3. High-throughput X-ray crystallography of crude arrays

High throughput X-ray crystallography requires that protein crystals must diffract at a good resolution and resist fracture during soaking.[97] Previous efforts demonstrated that crude reaction mixtures can be prepared and soaked onto protein crystals and that reaction products can be resolved.[132] Here, crude reaction mixture preparation was also included in the robotic framework (done by Dr Anthony Aimon and Dr Storm Hassell-Hart) where the products were concentrated in the organic phase and solvent exchanged. Due to the intrinsic affinity of bromodomains for dimethyl sulfoxide (DMSO) and/ or dimethylacetamide (DMA) these solvents were replaced with ethylene glycol (EG).

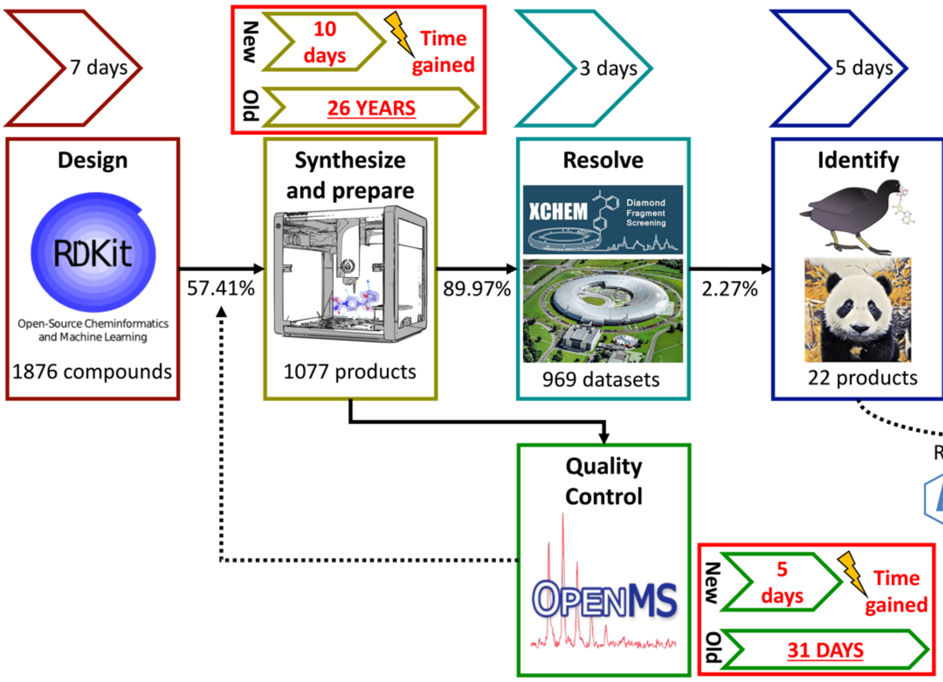
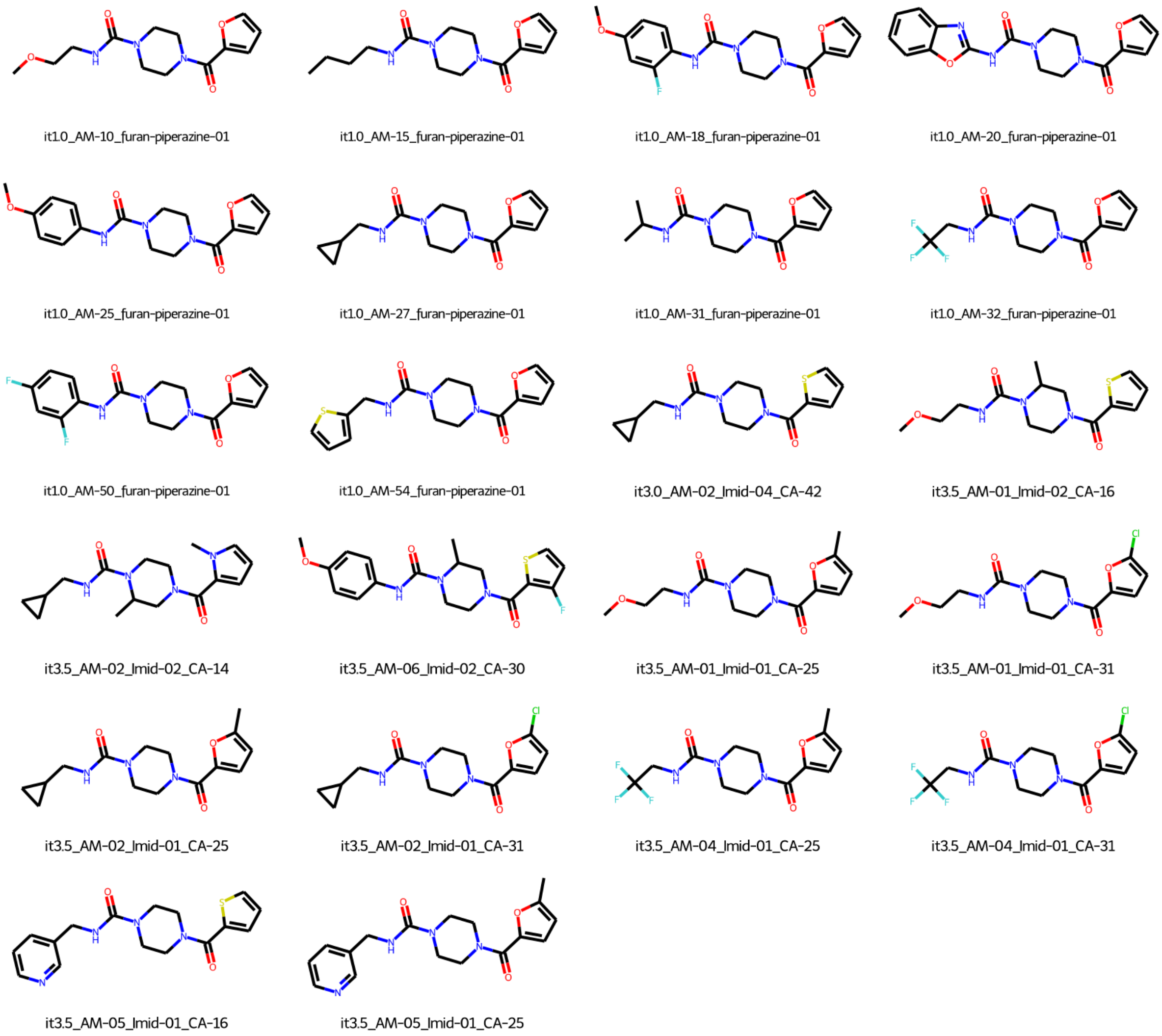


Figure 4.3: Automated chemistry and quality control fits with the XChem workflow, reduces human labour. Each step of the process is represented in a box with a picture illustrating the dominant piece of infrastructure at this step. The molecular and reaction designs were made with the Python package RDKit; preparation of CRMs on an OpenTrons; CRM QC via LC-MS and analysed with MSCheck; the crystals were prepared and resolved at the XChem, the product hits were identified from electron density maps with Coot and PanDDa; and the hits were confirmed via GCI and an alpha-screening assay. The quality control step is parallel to the main workflow because crystals were soaked with all crude reaction mixtures while the LC-MS outcome determined the number of successful reactions, as indicated by the dotted arrow. The new and old timings are shown as arrow boxes with estimated values within them. New and old refer to the timings achieved by this protocol versus the estimated time it would take a human to process an equal number of compounds. The validation required ordering the pure compounds from Enamine. Success rates are shown bellow the arrow connecting the steps and are indicative of the number of successful outcomes for a given step over the number of successful outcomes of the step before.

Reaction products were measured in 1077 out of 1876 syntheses by MSCheck, the *in house* quality control protocol,[131] representing a 57.41% success rate across all iterations. Crude reaction mixtures were soaked onto protein crystals (done by Dr Anthony Aimon and Dr Storm Hassell-Hart) at XChem (**Fig. 4.3**). This yielded 969 usable X-ray diffraction datasets for the crude reaction mixtures with successful syntheses, translating into a crystal deterioration rate upon soaking of crude reaction mixtures of only 8.73% (**Fig. 4.3**). The crystallographic experiments resolved a total of 29 unique bound structures, which included 7 starting materials and 22 reaction products (**Table 4.2**). A hit rate of 2.27% (22/969) was achieved when looking at target-bound reaction products from successful syntheses and usable X-ray diffraction data after soaking (**Fig. 4.4**). This is a lower hit rate than the previous crystallographic XChem fragment screening against this crystal form, which yielded a 6.51% hit rate with similar soaking conditions.[130]

Table 4.4: Crystallographic hits have conserved scaffold. The 2D structure for all crystallographic hits are displayed with its unique identifier as legend. The first field of the compounds' legend indicate the iterations from which the compound originated with subsequent identifiers for reaction products.


it1.0_AM-10_furan-piperazine-01
it1.0_AM-15_furan-piperazine-01
it1.0_AM-18_furan-piperazine-01
it1.0_AM-20_furan-piperazine-01
it1.0_AM-25_furan-piperazine-01
it1.0_AM-27_furan-piperazine-01
it1.0_AM-31_furan-piperazine-01
it1.0_AM-32_furan-piperazine-01
it1.0_AM-50_furan-piperazine-01
it1.0_AM-54_furan-piperazine-01
it3.0_AM-02_Imid-04_CA-42
it3.5_AM-01_Imid-02_CA-16
it3.5_AM-02_Imid-02_CA-14
it3.5_AM-06_Imid-02_CA-30
it3.5_AM-01_Imid-01_CA-25
it3.5_AM-01_Imid-01_CA-31
it3.5_AM-02_Imid-01_CA-25
it3.5_AM-02_Imid-01_CA-31
it3.5_AM-04_Imid-01_CA-25
it3.5_AM-04_Imid-01_CA-31
it3.5_AM-05_Imid-01_CA-16
it3.5_AM-05_Imid-01_CA-25

4.4.4. Resolving the binding landscape through structural analysis

All starting materials bind at the same location between helices B and C (**Fig. 4.1**) and display a conserved binding mode where they interact with the protein via 2 π-π stackings and hydrogen bonding (**Fig. 4.1**). All products bound maintain a 4-formylpiperazine-1-carboxamide scaffold (**Table 4.2**), where the piperazine moiety occupies the central hydrophobic cavity. All products also retained the original furan, analogous thiophene or pyrrole rings, with some hits having either a 2-chlorine or 2-methyl furan.

19 unique products observed in crystals (**Fig. 4.4**) have the same binding mode defined by the starting fragment (**Fig. 4.1**). Iteration 1.0 and the exploration for the urea vector was the only single-step iteration to yield ten follow-up crystal hits on the XChem platform (**Fig. 4.2**). Three of the binding events were from overlapping compounds from iteration 1.0. The importance of the furan ring for binding was highlighted when no structural hits were found for Iteration 1.1's urea vector exploration (**Fig. 4.2**) where the furan ring was not included. Iteration 2.0 (**Fig. 4.2**) highlighted the importance of retaining the urea group of the initial fragment as no hits were found unless the urea group from the original fragment was retained. Structurally, the urea group forms a hydrogen bond with the proline 1340 backbone oxygen (**Fig. 4.4**).

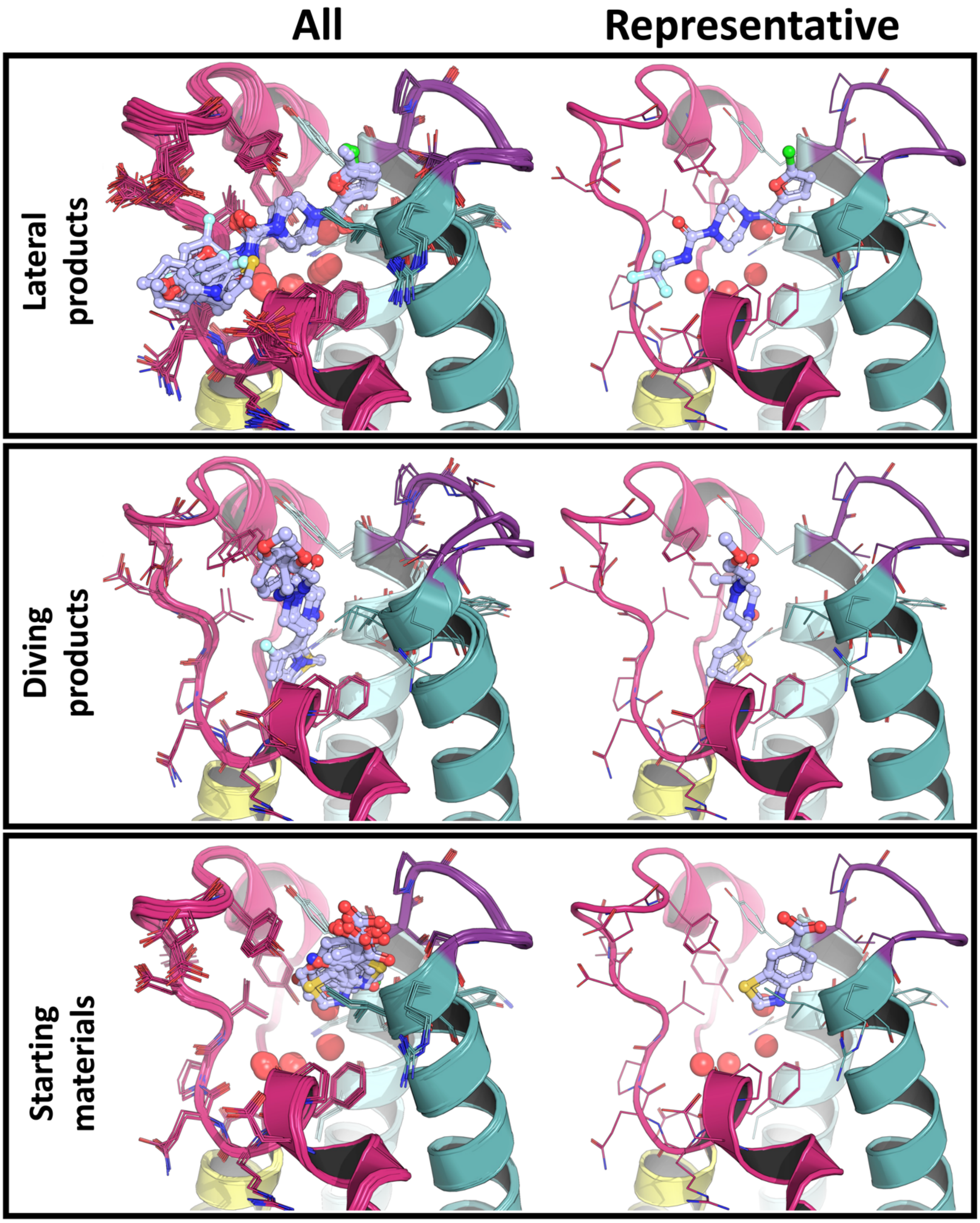


Figure 4.4: XChem screening of CRMs yielded reaction product-bound structures, with conserved and non-conserved binding poses, and starting materials. The first and second panels show structures of the laterally and diving bound products, respectively. The third panel shows structures of the reaction starting material-bound proteins. Each panel is subdivided into 2 columns. The first (All) aggregates all bound structures while the second (representative) show a single binder that has a binding mode representative of the others. The representative binders were arbitrarily selected to illustrate the corresponding binding mode.

Unexpectedly, 3 products were resolved in an alternative pose relative to the original fragment F709 (**Fig. 4.1**) as shown in **Figure 4.5**, demonstrating the effect of modifying the piperazine ring in Iteration 3.5 (**Fig. 4.2**). Those products rotated by about 90° around the piperazine core (with respect with the original pose) and so that their 5-membered ring displaces waters 2 to 4 while the nearby amide displaces water 1 of the network (**Fig. 4.1**). There, the 5-membered rings seem to bind mostly via hydrophobic interactions with the amino acids comprising the water cavity.

In addition, the cis conformation of the two amides was observed in the divers, while the lateral products displayed a trans conformation. A relatively large protein conformational change is paired with this novel diving binding pose. The ZA-loop adopts a generally more relaxed conformation, resulting in a more voluminous binding site that accommodates the flipped products, thus illustrating that some level of conformational motion is allowed by the C2 crystal system (**Fig. 4.4**).

A striking feature of the diving products is that they all have methyl-substituted piperazine ring. When looking at their closest lateral neighbours based on chemical similarity distance, it seems that this alkyl substitution alone is required to change binding orientation (**Fig. 4.5**). The relatively low incidence of diving relative to lateral binders also indicates that a high number of experiments are needed to resolve this unexpected pose.

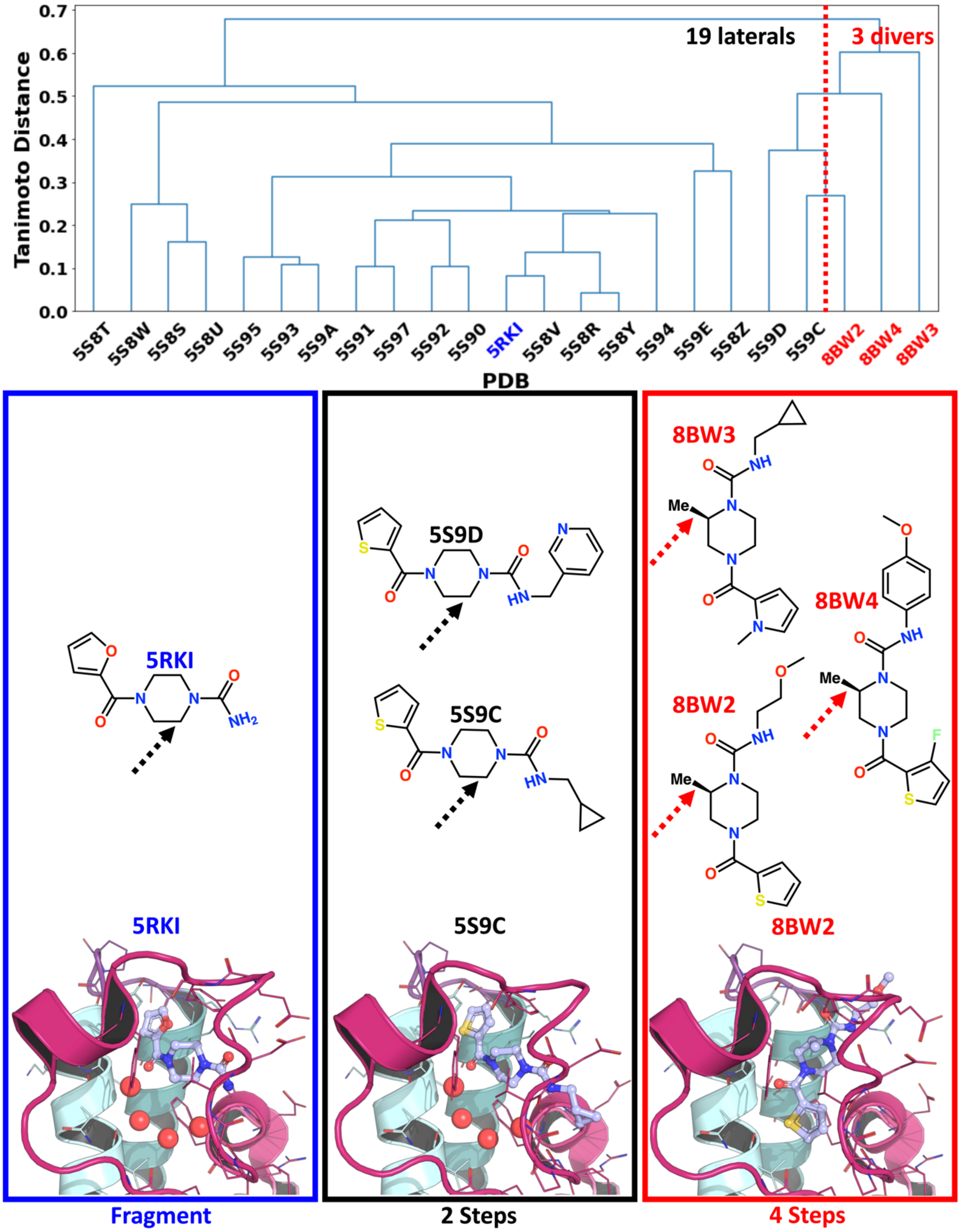


Figure 4.5: Unexpected diving binding pose appears to be triggered by the addition of a methyl to the core piperazine ring. A dendrogram showing the chemical fingerprint (Tanimoto) distance between bound reaction products is shown on top. The compounds are labelled by PDB accession IDs. Information related to starting fragment, lateral and diving compounds is highlighted in blue, black and red respectively. The 2D structures for the fragment, the divers' closest neighbour (laterals) and divers with a representative 3D structure are shown in the first, second and third columns, respectively. The number of synthetic steps required to obtain those compounds is shown under the columns for lateral and diving bound reaction products.

The synthesis of the 3 diving binder molecules made use of racemic building blocks (**Fig. 4.2**), resulting in crude reaction mixtures containing a mix of enantiomers and other by-products in unknown ratios. The quality control protocol, however, identified that products with the expected mass were present in the mixture. PanDDA event maps were used instead of traditional electron density maps to fit the compounds due to weak density in standard electron density maps.[133] The position of the methyl group around the piperazine ring was ambiguous and required further inspection (**Fig. 4.6**). Each compound has two stereo-isomers, each having two possible location with one on each side of the piperazine.

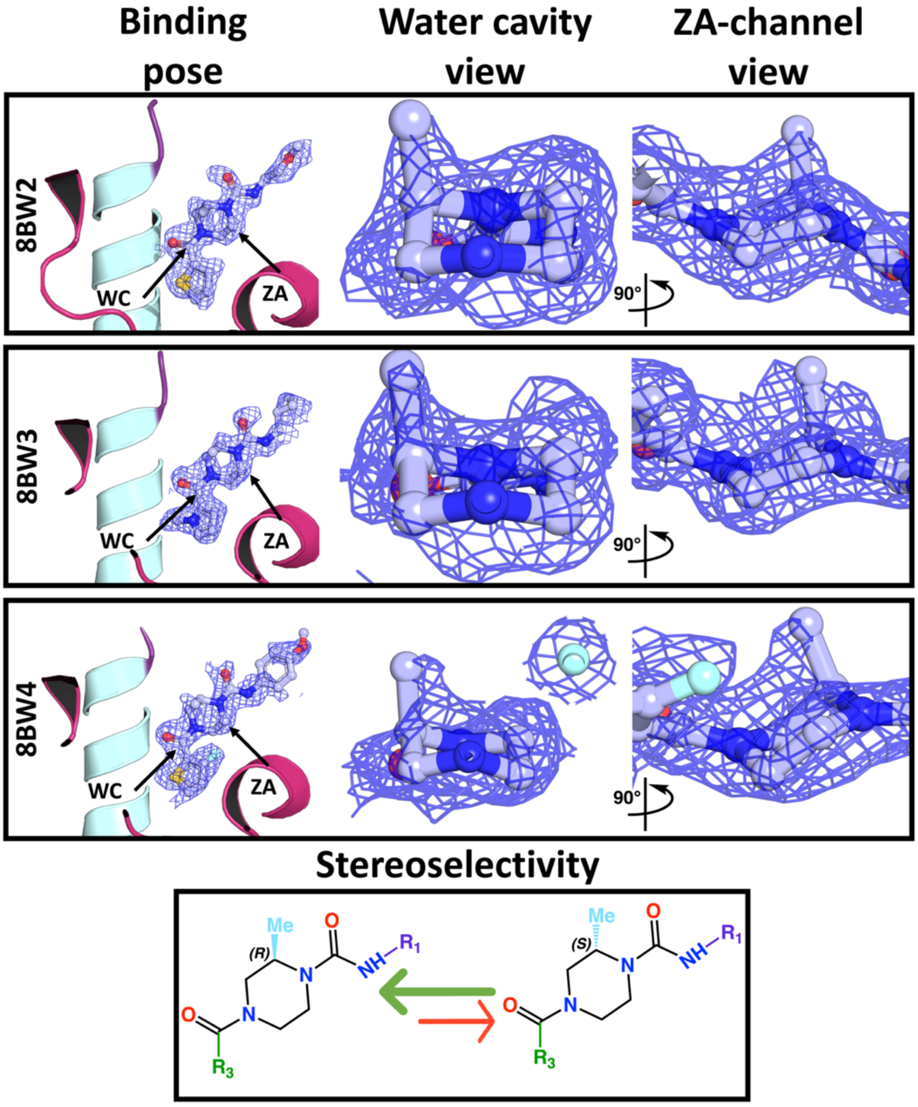


Figure 4.6: PanDDA event maps of racemate crude reaction mixtures suggest stereo-selectivity of the PHIP(2) binding for products containing an R-methylpiperazine moiety. The first three rows show the PanDDA event maps for the divers. The first, second, and third columns show the event density for the overall binding pose and the methylpiperazine moiety density viewed from the water cavity (WC) and the ZA-channel (ZA), respectively. For clarity, the point of views are indicated with arrows in the binding pose column. The last row shows the two possible chiralities, with the green arrow indicating which one was modelled in the binding site.

These maps showed a consistent protrusion at the same location, suggesting the presence of the additional methyl group (**Fig. 4.6**). Other possible locations do not have a similar protrusion and positioning the group there would result in a clash with the protein and/or the compound itself. At the identified position, the methyl group interact with the binding site through hydrophobic interactions (**Fig. 4.4**). Overall, these results indicate that the protein binding site defined by this crystal system is stereo-selective for the methyl group oriented in an "up" (R) fashion and located on the same side as the compounds' cis-oriented amide group oxygens (**Fig. 4.6**).

The interchangeability of chemical groups in crystallographic hits resulted from the combinatorial approach and maintained essential fragment features (**Fig. 4.7**). These features preserve pre-existing interactions with the binding site (**Fig. 1**), while more diverse changes at the R_1 urea (**Fig. 4.7**) extending across the ZA-channel into the solvent, where no novel interactions are acquired (**Fig. 4.4**). Several modifications, including replacing the piperazine a with diazepane moiety or dimethyl-piperazine, replacing the furan with a 6-membered ring or tri-heterocyclic ring, and adding a sulphonamide to the furan, in Iteration 4.2 (**Fig. 4.2**), consistently resulted in non-binding events despite quality control confirming that their synthesis has been successful.

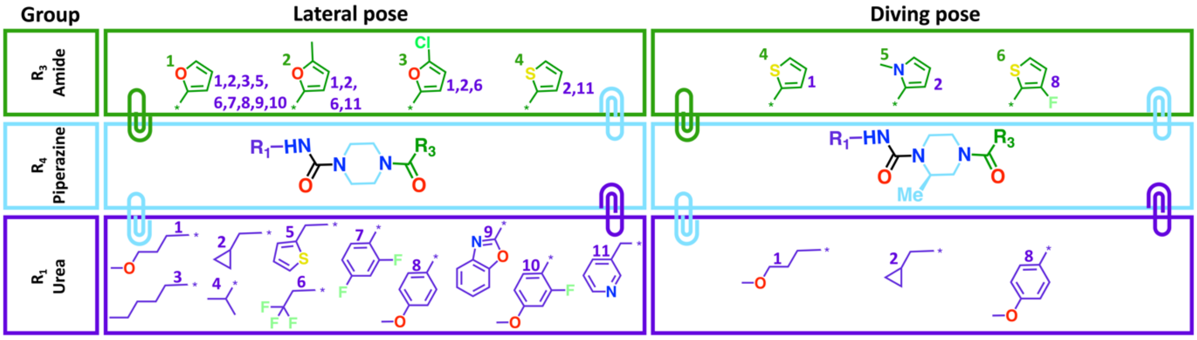


Figure 4.7: Combinatorial R-groups leading to crystallographic binding and binding pose change. The different groups causing crystal binding for each pose are showed in the panels. The numbers next to the group for the amide and urea expansions highlight the observed combinations between those groups.

4.4.5. Structural hit exhibits assay binding

The staring template fragment, F709, did not yield any detectable signal in solution assays and elaborating fragments aims at increasing the compound's affinity for the target. Hence, once identified, crystallographic hits must be confirmed via orthogonal methods. For this purpose, the pure compounds were independently synthesised from Enamine Ltd. These were then evaluated with Amplified Luminescence Proximity Homogeneous Assay (AlphaScreen™)[134] and grating-coupled interferometry analysis using the Creoptix repeated analyte pulses of increasing duration (waveRAPID)[68] instrument (done by Dr James Bennet and Dr Lizbé Koekemoer) thus testing biochemical and kinetics activities, respectively (**Fig. 4.8**).

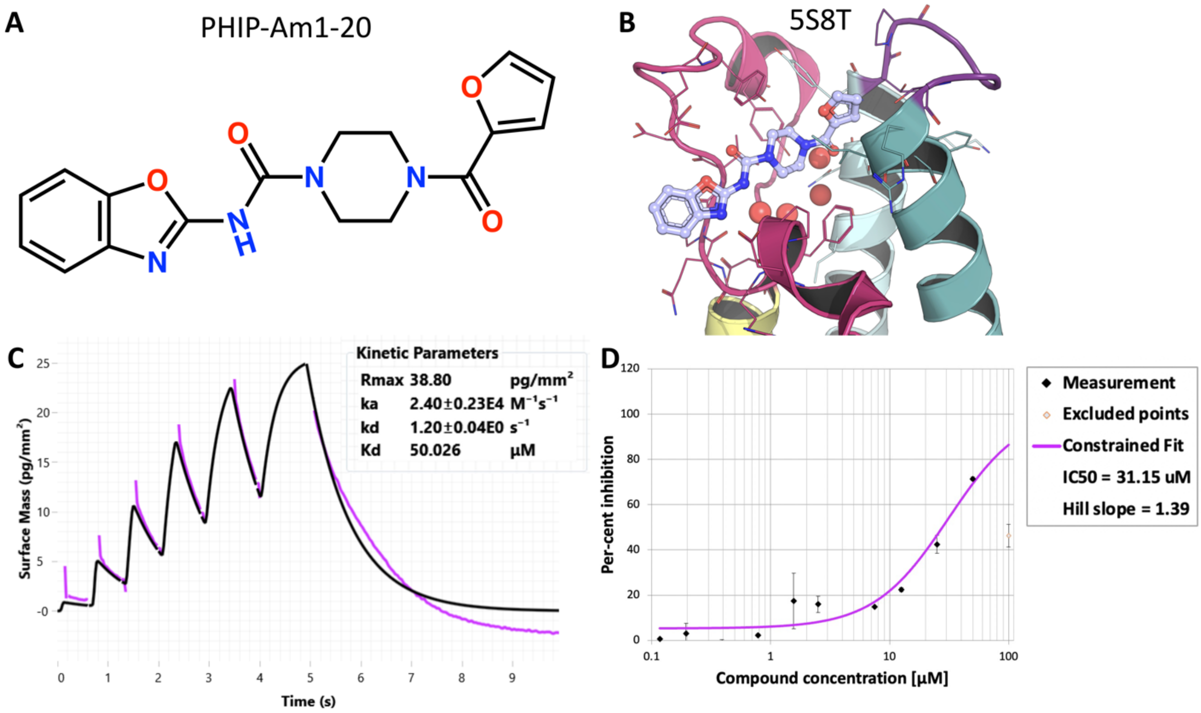


Figure 4.8: Structural and assay binding of reaction product PHIP-AM1-20. Panels A and B show the chemical structure and crystallographic binding pose of PHIP-AM1-20, respectively. The PDB code of the bound complex is written above B. Panels C and D showing kinetic rate constants and affinities from the Grating-Coupled Interferometry (GCI) and AlphaScreen assays, respectively, with the raw data in black and fitted binding curve in purple. ka and kd are indicative of the association and dissociation rates with Kd, the dissociation constant, equalling to association over the dissociation rates. IC50 corresponds to the half maximal inhibitory concentration.

Compound PHIP-AM1-20, product of Iteration 1 (**Fig. 4.2**), had a measurable effect in both assays with a K_d and a IC_{50} of 50.03 µM and 31.15 µM corresponding to ligand efficiencies of 2.00 and 1.25 µM per heavy atom for the kinetic analysis and alpha-screen, respectively (**Fig. 4.8**).

4.5. Discussion

Optimising potency rapidly and cheaply is a long-standing medicinal chemistry goal although fragment-based drug discovery (FBDD) approaches often struggle to generate potent molecules from protein structure-derived hits.[135] In recent years, progress has been made in developing low-cost high-throughput platforms[136] for the synthesis, purification and analysis of compounds.[137–139] These are, however, rarely directly paired with downstream biophysical and biochemical assays. This is due to the need for high-purity compounds to ensure accurate and convincing assay readout, which increases overhead costs, for example, due to higher solvent consumption through extraction, purification, and analysis. In turn, extensive solvent use can have detrimental environmental effects due to their toxic, flammable, and volatile nature, as well as their potential to contribute to climate change through their greenhouse gas emissions.[140]

The OpenTrons platform offers a convenient solution for automated synthesis paired with high-throughput validation by crystallography (**Fig. 4.3**). The platform's ability to complete the synthesis in a matter of weeks with just one trained chemist is a significant improvement over traditional synthetic approaches, which could take up years to complete (**Table 4.1**). Traditional synthetic approaches can also be fastened by the use of parallel synthesis and bulk preparation of common intermediates, as well as an increase in the number of chemists involved.

Another advantage of the OpenTrons is its compact size. The unit can be easily contained within a single fume hood and automated protocols can be run for every stage of the synthesis, from stock solution preparation to final sample solution preparation and quality control sample preparation (**Table 4.1**). This contrasts with traditional synthetic approaches that require substantial amounts of equipment and space, as well as almost constant operation by the chemist or project team. The design of larger arrays would also be possible,[141] but the size of the iterations were limited to levels that are manageable for downstream processing at the XChem, which involves human interventions at certain steps, such as crystal fishing.[22]

This integration of the automated synthesis workflow with the existing high-throughput crystallographic pipeline can generate a large volume of usable data quickly, while maintaining crystal integrity (**Fig. 4.3**). The quality control step also provides traceability of reaction outcomes, allowing for the tracking of electron density maps that may contain bound reaction products.[131] In comparison to previous studies, such as the work of Baker et al., (2020),[132] the use of the OpenTrons leads a substantial increase in throughput at the XChem by generating thousands of crude reaction mixtures. In their study, 83 reaction mixtures were processed in triplicate, while only a single experiment was performed per compound, here. While the approach, assembled here, may have led to a lower hit rate, it prioritised data quantity over quality, resulting in a 22-fold gain in processing compounds or a 7-fold gain in performing experiments compared to the work of Baker et al., (2020).[132]

The hit rate at the acetylated-lysine binding site from the screening of crude reaction mixtures (**Fig. 4.4**) was about 75% lower than the fragment screen presented in Chapter 3.[130] It is important to note that the low hit rate observed in this study may have also been due to inadequate compound designs, as the exercise was largely driven by the available chemistry surrounding the starting fragment. This also highlights the importance of maintaining interactions resolved in the original fragment (**Fig. 4.1**) as modifications that perturbs those led to the absence of crystallographic binding. It is also hard to manually process the copious amounts of non-binding data and it is expected that several false negatives may be hiding in the data. For example, simple rescreening of fragments in Chapter 3 led to identification of a relatively large number of hits, an operation that was not performed here.

However, the thorough sampling of the chemical space (**Fig. 4.5**) identified changes in binding pose and small protein conformational changes (**Fig. 4.4**) that would not have been observed through commonly used docking methods. PanDDa event maps[133] were used to guide the ambiguous positioning of a single methyl at a stereo-centre (**Fig. 4.6**). The position of the methyl at the stereo-centre was however deduced from visual inspection of the PanDDA event maps and the electron density and/ or event may be resulting from overlapping crystallographic state.[132] The consistent protrusion corresponding to the R (up) isomer and binding site context guided the final choice.

It is hard to explain the underlying energetics causative of such pose change but contributing factors may include entropic gain via water displacement.[142] Previous computational study of bromodomains' water network stability has suggested that the PHIP(2) network is unstable with a positive Gibbs binding free energy.[143] This would imply that the water network does not make strong interactions with the neighbouring binding site and is easily displaceable, potentially resulting in an entropic reward to Gibbs binding free energy of ligand binding. The change in binding pose may also be driven by acquisition of more favourable interactions in the new pose or steric clashes introduced by the chemical modification in the originating lateral pose.[144] The addition of the methyl may also change the ligand's conformation associated with its overall energy minima. In the diving pose, the two amides were resolved in a trans conformation which may be more stable than the cis conformation resolved in the lateral pose (**Fig. 4.4**). Previous observations of so called "magic methyls" have also highlighted the ability of such modification to drastically alter binding affinities.[145]

The ability to perform such exploration highlights the importance of multi-step iterations which are more laborious but necessary (**Figs. 4.2 and 4.5**). It is anticipated that better follow-up designs will result in increased hit rates. This may also extend to downstream hit validation in solution assays. In addition, all compounds except 1 (**Fig. 4.8**) did not yield readable signals in assays suggesting that these crystal binding events are weak and may not happen in solution.[146] The only assay binder (**Fig. 4.8**), however, exhibited a 2- to 3- fold increase in binding affinity when compared to the best binder obtained by Cox et al., (2016) that were elaborations around the first DSI-poised fragment binders identified against PHIP(2).[13]

The method has proven to be effective at generating large amounts of crystallographic data but lacks equally automated means of validating those hits via assays. Purchase of pure compounds (**Fig. 4.3**), or resynthesis *in house* using conventional chemistry is still required for validation, but efforts have been made in measuring dissociation rates from crude reaction mixtures, which may combine efficiently with the novel pulsed injection schemes implemented for the time-resolved kinetic analysis using grating-coupled interferometry.[44,68]

The strategy of prioritising data quantity over quality has pros and cons. On one hand, the upshot of this choice lies in the broad coverage of chemical space. This increases the knowledge and certainty of modifications leading to crystallographic binding (and non-binding) as well as the chances of uncovering unexpected findings, such as alternative binding poses (**Fig. 4.5**). On the other hand, the downside of this approach is the potential compromise on data integrity. The large volume of data generated can present considerable challenges in terms of interpreting the information accurately. In addition, the adoption of high-throughput methodologies, especially soaking of crude reaction mixtures, is expected to lead to false negatives due to inconsistencies in the quality control readout, varying concentrations, and binding affinities.

4.6. Conclusion and future direction

Overall, the conceptual fragment to drug-like molecule growth exercise (**Fig. 4.1**) has demonstrated that it is possible to save time, solvent usage, and many synthetic, extraction, work-up and purification bottlenecks (**Table 4.1**) by combining crude array synthesis and X-ray crystallographic structural determination of the resulting molecules soaked onto PHIP(2) crystals (**Fig. 4.3**). The approach is also more environmentally friendly than traditional synthesis as it prevents the use of polluting solvent (**Table 4.1**).[147]

The XChem screen, against PHIP(2), presented in Chapter 3, led to the piperazinyl hit F709 (**Fig. 4.1**). Consequently, several thousand multi-step, robotic-driven reactions were designed reactions with minimum automated workups (**Fig. 4.2**). Crude reaction mixtures were submitted to XChem for X-ray analysis at high-throughput levels and without compromising crystal integrity thus, bypassing purification steps (**Fig. 4.3**). Furthermore, the quality control step (MSCheck) supplements the workflow by allowing rapid tractability of reaction outcomes thus enabling tracking of which electron density maps could have a bound product.[131]

This resulted in the identification of crystal binders (**Table 4.2 and Fig. 4.4**) which enabled to map which vectors can be elaborated, with what type of chemical group and to what extent. Together these findings provide an estimate of the crystallographic SAR landscape around fragment F709 defined by the synthetic iterations (**Fig. 4.7**). In addition, a novel diving pose was recovered.

Cheminformatic (**Fig. 4.5**) and electron density inspection (**Fig. 4.6**) highlighted its role and orientation, respectively. To confirm the position of the methyl, solutions containing pure R or S enantiomers could be soaked on the protein crystals and resolved by X-ray crystallography. The aim here was also to rapidly generating crystallographic information from crude reaction mixtures and thus, further experiments enantiomers were beyond the scope of this study. Time, budget an expertise constrains were further prohibitive of further analysis.

Overall, this illustrates how features can rapidly be scoped and "fail quickly and cheaply" which could be used to quickly redefine synthetic strategies.[148] Even negative results, obtained quickly, could have insightful impact in these labour-intensive processes (**Fig. 4.2**). The structural information acquired from this experiment such multiple bound structures, ligand conformations and flipped pose may be useful in future iteration design, for example by being leveraged in conjecture with constrained and/ or ensemble docking.[149]

However, automated methods are also needed to systematically analyse and rationalise the data resulting from these increasingly large high-throughput crystallographic screenings. This is especially true for the non-binding data which is less tractable by visual inspection. A single crystallographic binder was also found to be active in biophysical and biochemical assays (**Fig. 4.8**) and of better potency than previously reported.[13] This demonstrates that at scale follow-up potency can be achieved, starting from an undetectable fragment, using this protocol. Future work may also include mass dissociation rate screening using crude reaction mixtures by grating-coupled interferometry assay. This would remove resynthesis or purification needs to increase throughput and gain essential potency information.

This case study serves as an exemplar on how to generate structural data quickly with minimal bottlenecks, caveated by the fact that improved binding interactions are not necessarily commensurate with increased activity nor are they a substitute for biological activity derived from an assay. Overall, this work minimised the risk of human errors and optimised resource usage whilst also highlighting important needs for downstream assays and data analysis tools.

4.7. Methods

4.7.1. Protein expression, purification, and crystallisation

Proteins were expressed, purified and crystalised as described in Chapter 2, Section 2.2.1. Model building and refinement was also performed as detailed in Chapter 2, Section 2.2.2. Structural models and crystallographic statistics can be retrieved from the PDB deposition ID: G_1002190.

4.7.2. Chemical similarity calculations and clustering

Chemical similarity calculations for small molecule pairs were performed using RDKit and were based on the Tanimoto coefficient between molecular fingerprint bit vectors.[79] The Tanimoto distance for each pair yields a distance matrix enabling to perform hierarchical clustering using complete linkage method in Euclidean space using SciPy.

4.8. Data availability

The directory (curated by Dr Warren Thomson) was deposited on Zenodo, containing X-ray and LCMS results for the reactions executed on the OpenTrons, output reports and summaries from MSCheck (semi-automated LCMS analyzer tool) and the Python scripts used to execute single and multistep chemistry on the OpenTrons.

The accession URL is https://zenodo.org/record/7586212#.ZCLC6ezMLn5

Chapter 5. Algorithmic recovery of fragment elaboration binding features from large scale crystallographic readouts

5.1. Credits

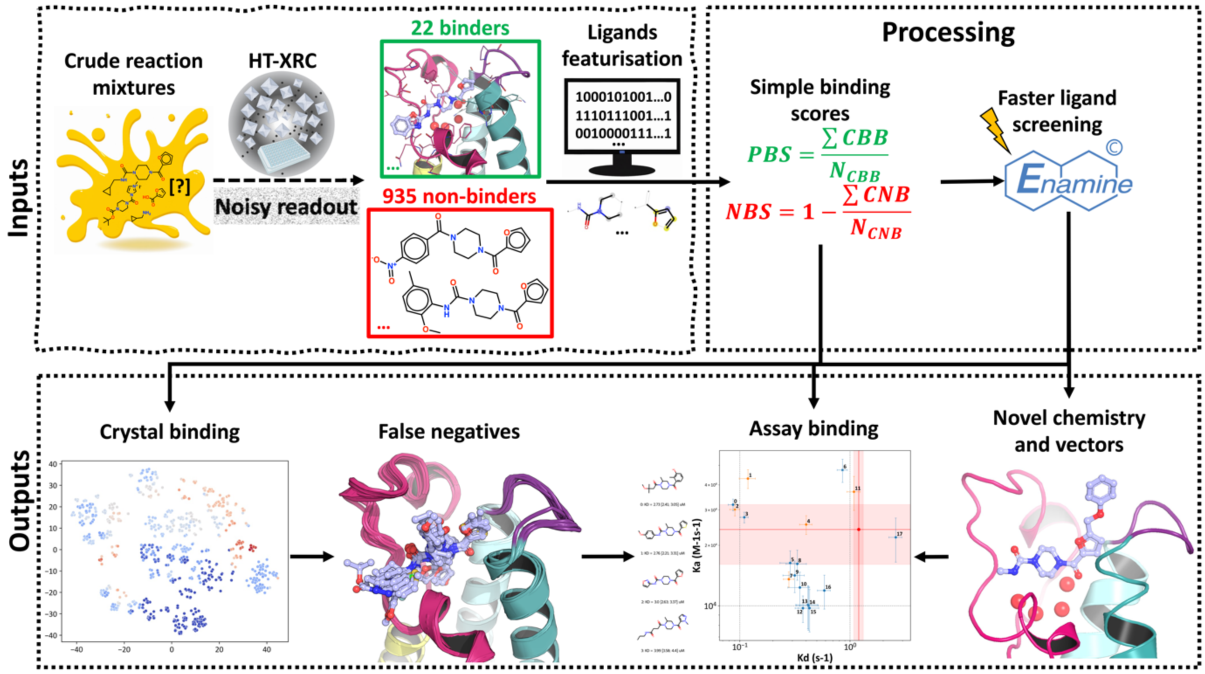
The research and manuscript presented in this chapter were conducted and authored by Harold Grosjean. Harold Grosjean conceived, coded, validated and benchmarked the method with input from Dr Ruben Sanchez-Garcia, Dr Rocco Meli, and Cameron Anderson. Harold Grosjean carried out protein expression, purification, and crystallisation, as well as crystallographic screening, structure refinement, electron density analysis, and grating-coupled interferometry assays. Additionally, Harold Grosjean wrote the chapter and created the figures. Dr Charlie Tomlinson, Dr Lizbé Koekemoer, Dr Tobias Krojer, Dr William, Dr Daren Fearon, and Dr Edward A. FitzGerald contributed to experimental approaches and the processing of resultant data. Dr Ruben Sanchez-Garcia conducted the catalogue ligand- and structure-based screening. Prof Philip C. Biggin and Prof Frank von Delft reviewed and offered guidance on both the manuscript and the overall project.

5.2. Introduction

Chapter 4 illustrated how robotic-assisted chemistry can be paired with X-ray crystallography at high-throughput levels to generate vast amounts of crystallographic structure-activity data. The understanding of such wealth of information remained however challenging due to the lack of data analysis methods and the expected degree of uncertainty associated with the use of combined high-throughput approaches. Indeed, the variable purity in the crude reaction mixtures paired with imperfect quality control tools may degrade data quality. Manual analysis of the non-binding data also identified several compounds recapitulating binding features suggesting false negatives.

Thus, this chapter aimed to develop a data analysis approach that unbiasedly identifies molecular features associated with binding, or lack thereof, which are then used to map false negatives and thus deconvoluting the binding landscape. Chemical connectivity fingerprints were calculated and molecular features associated with binding and non-binding events extracted. The rates of binding and non-binding features were combined into the positive and negative binding scores, respectively. Those metrics were used to identify potential false negatives from the chemical space defined by robotic synthesis, in Chapter 4, which were purchased in pure form and rescreened with X-ray crystallography. Overall, 23 and 3 false negatives were identified in the lateral and diving poses, respectively, thus doubling the initial hit rates. Enrichment analysis for the rescreening experiment showed the positive binding score to significantly enrich binders versus non-binders. The performance of the approach was compared with other commonly used ligand-based classifier including an optimised random forest. Features extraction also presented a natural opportunity in virtual screening. The binding scores were used to select follow-ups from a catalogue of commercially available compounds. Validation of those compounds revealed crystallographic binders with novel chemistry and enhanced affinity in assays.

Altogether, this work demonstrates how binding and non-binding information can be recovered from a noisy crystallographic binding landscape and used in a prospective virtual screening effort.



Graphical introduction 5.1: Ligand features extraction from a noisy crystallographic dataset enables false negative identification and virtual screening of catalogue compounds.

5.3. Results

5.3.1. A ligand-based method for analysing large scale crystallographic readouts

Extracting meaningful information from large datasets often requires specialised computational analysis. In Chapter 4, a crystallographic screening of PHIP(2) robotic fragment elaborations generated a large dataset of binding and non-binding events (**Fig. 5.1**).[150] However, manual inspection revealed that some non-binding reaction products were structurally similar to previously resolved compounds, making it difficult to rationalise their lack of crystallographic binding. False negatives may explain this observation due to experimental factors such as defective quality control,[43] relative crystal tolerance to ligands,[151] compound solubility-concentration,[152] inaccurate dispensing,[21] as well as other crystal pathologies.[153] Some elaboration series systematically were never observed in crystals indicating that theses truly prevent follow-up binding.

To recover information from this seemingly intractable data, a simple cheminformatic approach that's extracts ligand features associated with binding and non-binding events was developed. A data subset of 957 crystallographic ligand binding readouts was curated from approximately 1850 entries based on (1) successful synthesis outcomes, as measured by automated by Liquid chromatography–Mass Spectrometry (LCMS) analysis of crude reaction mixtures, and (2) usable X-ray diffraction data.[150] A "hit" or “binder” was defined as a compound manually resolved in the electron density, while a "non-hit" or “non-binder” lacked crystallographic evidence of binding.

This refined dataset was divided into three data subsets based to include pose information. The "lateral" and "diving" subsets correspond to hits that bind in the lateral and diving pose previously identified in chapter 4, respectively. The "all" data subset aggregates hits from both poses. (**Fig. 5.1**).

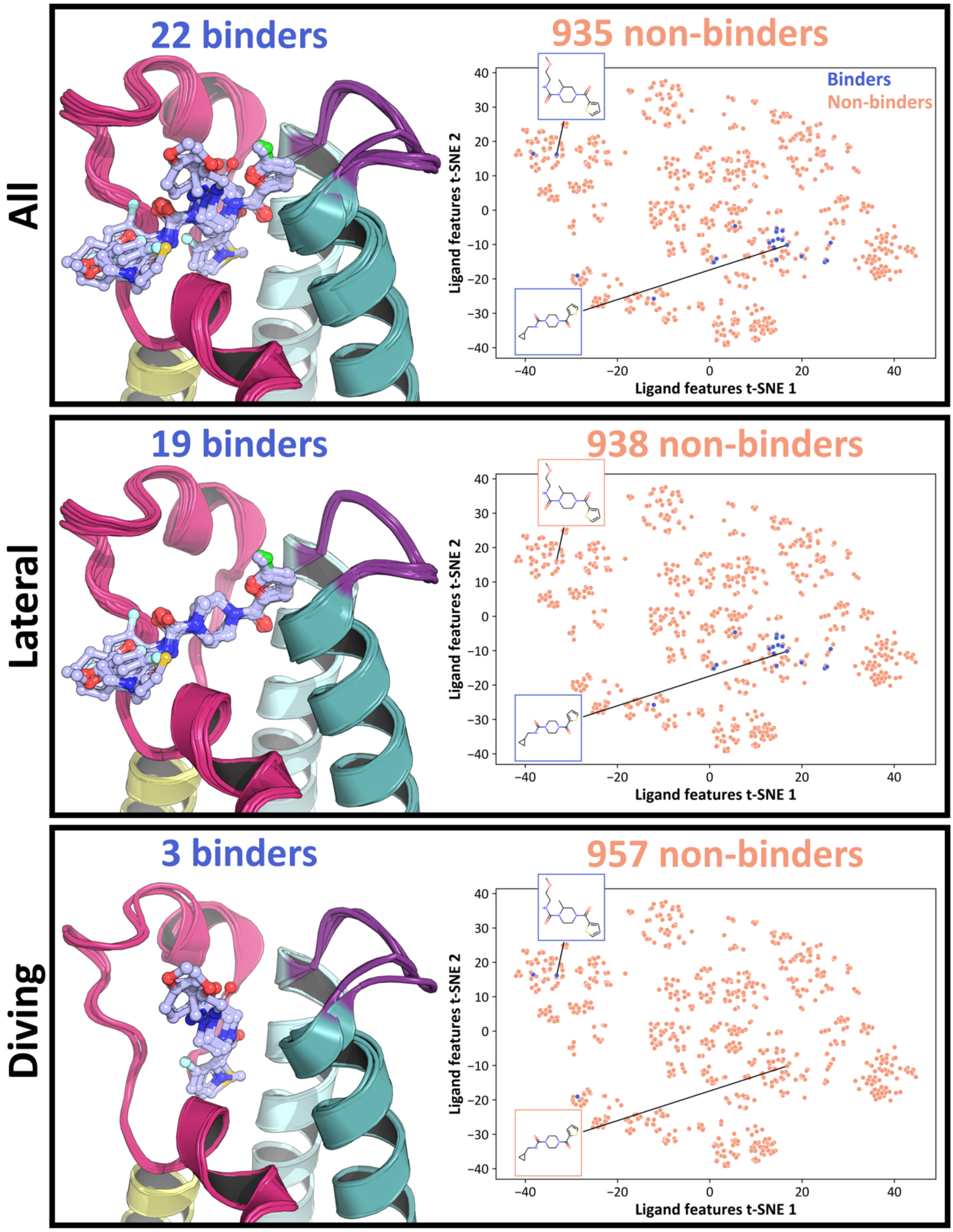


Figure 5.1: Chemical space defined by robotic elaboration showing binders and non-binders identified from crystallographic screening of crude reaction mixture. The left column displays previously resolved reaction products-bound crystal structures. Data subsets of combined (All) and individual poses (Lateral and Diving) as showed on each row. The chemical landscapes-based t-Distributed Stochastic Neighbourhood Embedding of underlying connectivity fingerprints folded it in two dimensions are showed on the right column.

For the extraction of important ligand features, it is essential to break down the related compounds into formats that machines can process. Connectivity fingerprints of ligands offer a convenient avenue as they encode molecular structures as fixed-length bit strings with each bit representing the presence (“activated-bit” or “on-bit”) or absence (“deactivated-bit” or “off-bit”) of a particular chemical feature.[76] Furthermore, these fingerprints can be used to visualise the chemical landscape defined by group of compounds. t-Distributed Stochastic Neighbourhood Embedding (t-SNE) was used to reduce the highly-dimensional molecular binary representations to a 2-dimensional space for visualisation.[80] Overall, the lateral binders are located together around the middle of the chemical space whereas the diving compounds are more distantly located from each other (**Fig. 5.1**).

This scoring scheme developed here focuses first on information pertaining to hit compounds as these can be characterised with a higher degree of certainty compared to non-binding compounds. Two simple schemes (**Fig. 5.2**) were designed to extract binding and non-binding information from the high-throughput crystallographic screening of crude reaction mixtures.

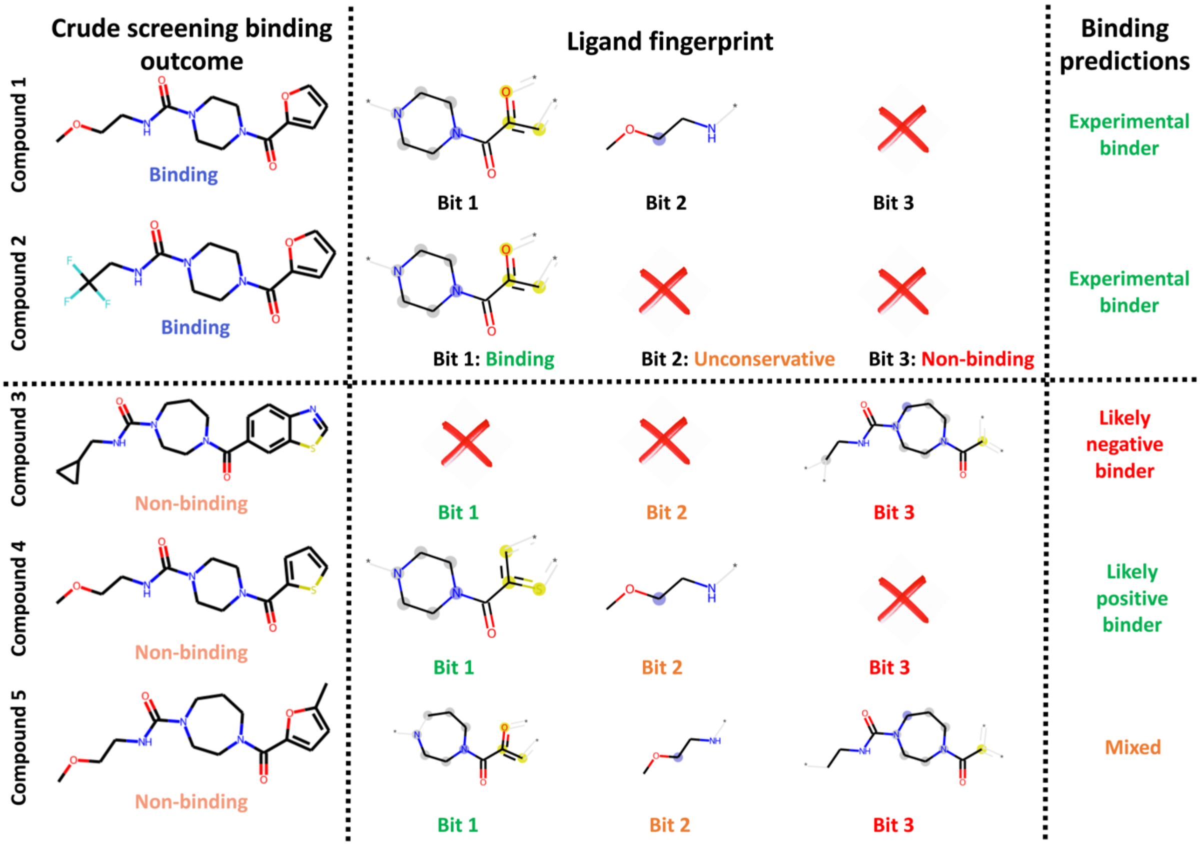


Figure 5.2: Identification of conserved ligand features and compound scoring for simple binding predictions amongst non-binders. This shows a simplified example of how non-binders can be predicted for binding using this method. Binding and non-binding features are identified with respect with binding hits and then mapped onto non-binders for prediction to likely or unlikely binders.

Recognising the preserved features of both binding and non-binding compounds facilitates compound scoring (refer to the methods section 5.6.2 for details). The Bit Conservation Score (BCS) is calculated for individual bits and across hits. This enables to classify each bit into Conserved Binding-Bits (CBB) and Conserved Non-binding Bits (CNB) which correspond to features associated with crystallographic binding or lack thereof (**Fig. 5.2**), respectively.

The Positive and Negative Binding Scores (PBS and NBS) are then calculated by counting the conserved binding and non-binding bits in non-hits. This allows predict for these non-hits weather they are likely or unlikely binders (**Figs. 5.2 and 5.11**). A positive and negative binding score of 1 indicates that a compound has features strongly and weakly associated with binding and non-binding, respectively (**Suppl fig. 5.1**).

5.3.2. Feature identification rationalises crystallographic binding

This scoring scheme was applied to all three data subsets (All, Lateral, and Diving Pose), with bit classification performed prior to scoring (**Fig. 5.3**). This enables visualisation of the conserved binding bits to further rational pose-specific binding.

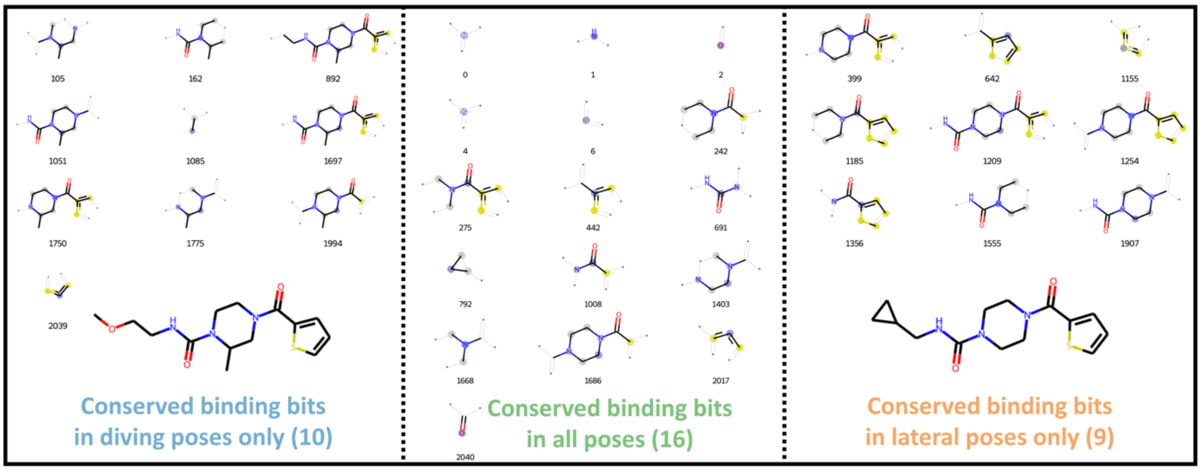

105
162
892
1051
1085
1697
1750
1775
1994
2039
Conserved binding bits in diving poses only (10)
0
1
2
4
6
242
275
442
691
792
1008
1403
1668
1686
2017
2040
Conserved binding bits in all poses (16)
399
642
1155
1185
1209
1254
1356
1555
1907
Conserved binding bits in lateral poses only (9)

Figure 5.3: Identification of conserved binding-bits enables comparison of pose-specific binding features. The conserved binding-bits for all, lateral and diving binders are compared highlighting pose-specific binding features. The dotted line illustrate that the binding bits present in all binders are also shared with lateral and diving binders.

The comparison of overlapping and diverging bit classification between data subsets provides insights into underlying features responsible for binding and pose change. The presence of novel chemical features in the binders was associated with an increased and decreased number of unconservative and conserved non-binding bits, respectively (**Suppl fig. 5.2**). Despite a 6.3-fold difference in hit count between the lateral and diving poses, both exhibited a similar number of conserved binding bits, suggesting that the accumulation of binders quickly reveals the scaffold required for pose-specific binding (**Fig. 5.3**). The number of conserved binding bits was manageable for visual inspection. This reinforced the association between the methylpiperazine modification and the diving pose, as this chemical group is present in 90% of the pose-specific conserved binding bits.[150]

Identification of conserved binding and non-binding bits (**Fig. 5.3**) enabled the scoring of each compound based on positive and negative binding scores (**Fig. 5.2**), where higher values indicate the presence and absence of binding and non-binding features, respectively. Compounds with positive and negative binding scores of 1 or less than 1 were classified as likely or unlikely binders (**Fig. 5.3**). Positive and negative binding scores were calculated for each pose, resulting in a total of six landscapes, and each compound was classified accordingly based on its calculated scores (**Suppl fig. 5.3**).

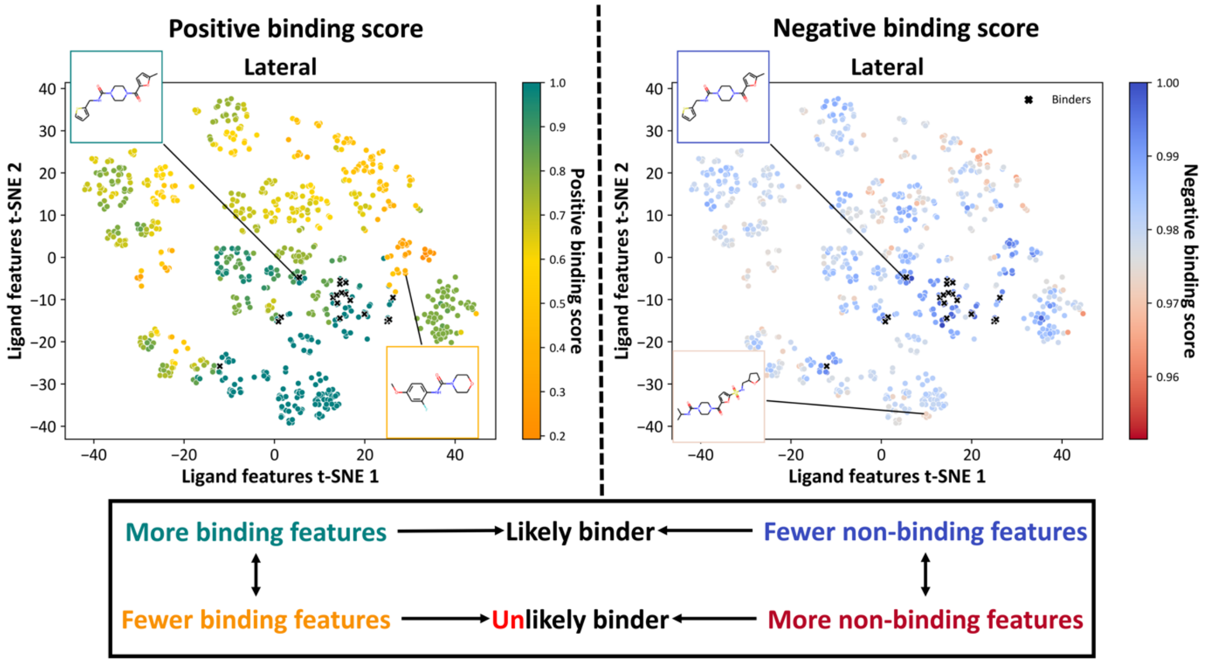


Figure 5.4: Positive and negative binding scores contextualise crystallographic binding landscapes in term of ligand features associated with binding or lack thereof. The top row displays the binding landscape colour-coded by positive or negative binding scores. The binders are showed on each landscape with black crosses. The bottom panel illustrates how these scores should be interpreted.

The number of underlying bits accounted for by the score affects the smoothness for the binding surface (**Fig. 5.4**), with the negative binding score showing less pronounced gradients due to the greater number of conserved non-binding bits (**Suppl fig. 5.2**). Overall, both landscapes were in good agreement with each other. However, some clusters can have conflicting interpretation, such as the one around landscape coordinates of (Ligand features t-SNE 1) 30 and (Ligand features t-SNE 2) 0, have unfavourable positive binding scores and favourable negative binding scores, respectively. This due to the absence of the amide and 5-membered ring essential for binding whilst not bearing any groups associated with non-binding. Reversibly, compounds with a sulphonamide extension, which repeatedly led to non-binding, exhibited a low negative binding score in the cluster around landscape coordinates of (Ligand features t-SNE 1) -30 and (Ligand features t-SNE 2) 0. For those compounds, the conserved binding features were maintained, resulting in large positive binding scores (**Fig. 5.4**).

Compounds that did not bind in the initial X-ray crystallographic experiment of crude reaction mixtures were subjected to predictive classification (**Suppl fig. 5.3**). Likely and unlikely binders were predicted based on positive and negative binding scores, respectively (**Fig. 5.2**). A total of 189 and 34 compounds were predicted to be likely binders for the lateral and diving pose, respectively, with the differential chemical series exploration sizes being the reason for the 5.56-fold difference (**Suppl fig. 5.4**). The classification based on the negative binding score was more permissive, with 915 and 951 compounds classified as unlikely binders for the lateral and diving pose, respectively, due to the presence of at least one unobserved feature in the compound (**Suppl fig. 5.4**). Compounds with a negative binding score of 1 are a chimera of existing binders because all activated bits are either binding or unconservative, meaning that they were observed in at least one binder (**Fig. 5.4**).

5.3.3. Identification of false negatives in rescreening experiment

This study explored two hypotheses: (1) there are crystallographic false negatives among non-binders, and (2) the scoring scheme presented above can predict crystallographic ligand binding. In order to determine whether the two hypotheses hold, an experiment was designed using the data presented in Chapter 4.

The proposed scores were computed for all non-binders with the aim of selecting potential false negatives for experimental validation. To maximise chances of resolving false negatives, a logarithmic sampling strategy for compound selection was applied to both positive and negative binding scores with bin densities biased towards higher score values (**Fig. 5.5**). The cheapest compound was selected within each bin, and if compounds had the same price, the one with the score closest to the average score for that bin was chosen. A total of 97 non-binders were purchased from Enamine in pure form (**Suppl. table 5.1**), and these were tested with X-ray crystallography against the previously used PHIP(2) crystals in a C2 space group.[130] Triplicate 24-hour soaks were performed to increase the likelihood of identifying binding events.[43]

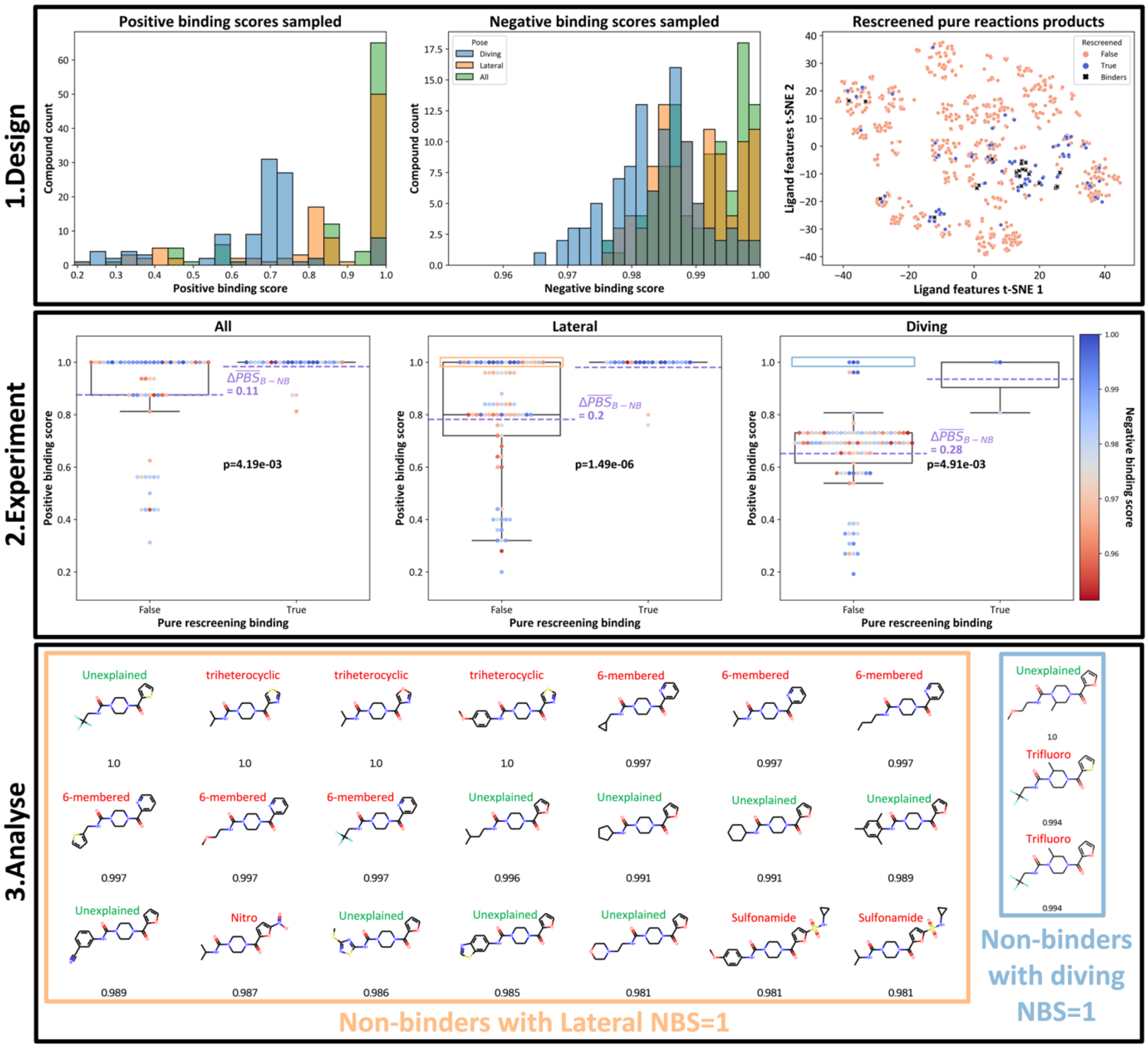


Figure 5.5: Crystallographic rescreening of pure compounds covers binding landscape with the positive binding score significantly enriching false negatives identification. The chemical space sampled with associated positive and negative binding scores for each pose is shown in the first row (1). The second row presents the enrichment analysis, with True and False categories representing the pure crystallographic rescreening binding outcome on the x-axis (2). The spread for each class generated from the positive binding score is showed along the y axis. Purple dotted lines show the average positive binding scores for each class and the resulting difference (ΔPBS) between binders (B) and non-binders (NB). The p-value, calculated using a The Mann-Whitney U test, indicate the significance of the positive binding score in discriminating binders from non-binders. The associated negative binding score values are colour-coded for all datapoints. The lower row displays non-binders with a positive binding score of 1. Each compound is labelled with features hypothesised to cause non-binding and negative binding scores, respectively.

The selection sampled well the chemical space with the positive binding score significantly identifying false negatives for all 3 datasets (**Fig. 5.5**). The chemical space was well covered by the rescreening strategy, except for the top right cluster. This region had low positive and negative binding scores due to the replacement of the essential 5-membered diheterocyclic ring with various groups . The compound selection was biased towards higher positive and negative binding score regions for the lateral pose. This implies that selected compounds recapitulated most conserved binding bits and lacked conserved non-binding bits for that pose (**Fig. 5.2**).

Crystallographic rescreening of the pure compounds identified 26 novel binders, doubling the initial hit rates for both poses (**Fig. 5.1**). Enrichment analysis based on positive binding scores of binders versus non-binders proved to be significant in all three data subsets, with the lateral and diving datasets being more significant and better enriched than the "All" data subset. The difference in average positive binding score for binders versus non-binders was the largest in the diving pose. However, only 3 false negatives were recovered for this data subset and statistics may therefore suffer from a lack of sampling.

The disparity in the number of recovered diving binders in comparison to lateral ones is due to the reduced number of rescreened compounds harbouring a methylpiperazine group(**Fig. 5.5**). 65 contained a piperazine moiety, 11 contained a methylpiperazine, and 21 had neither of these functionalities (**Suppl table 5.1**). Considering the number of compounds binding in the lateral and diving conformations, there were 21 piperazine and 3 methylpiperazine variants, respectively (**Fig. 5.6**). As such, the resultant specific hit rates for these two groups are 32.31% and 27.27%.

The experimental results demonstrated that the positive binding score significantly enriched binders over non-binders across all datasets and performs better than the negative binding score (**Suppl fig. 5.6**). However, there were 21 and 3 non-binders with maximal positive binding scores for the lateral and diving poses, respectively. Features such as 5-membered triheterocyclic rings, 6-membered rings, nitro and sulphonamide groups for the lateral pose, and trifluoromethyl for the diving pose, repeatedly led to non-binding events. Additionally, 9 and 1 non-binders had no obvious features that could explain the lack of crystallographic binding for the lateral and diving poses, respectively (**Fig. 5.5**).

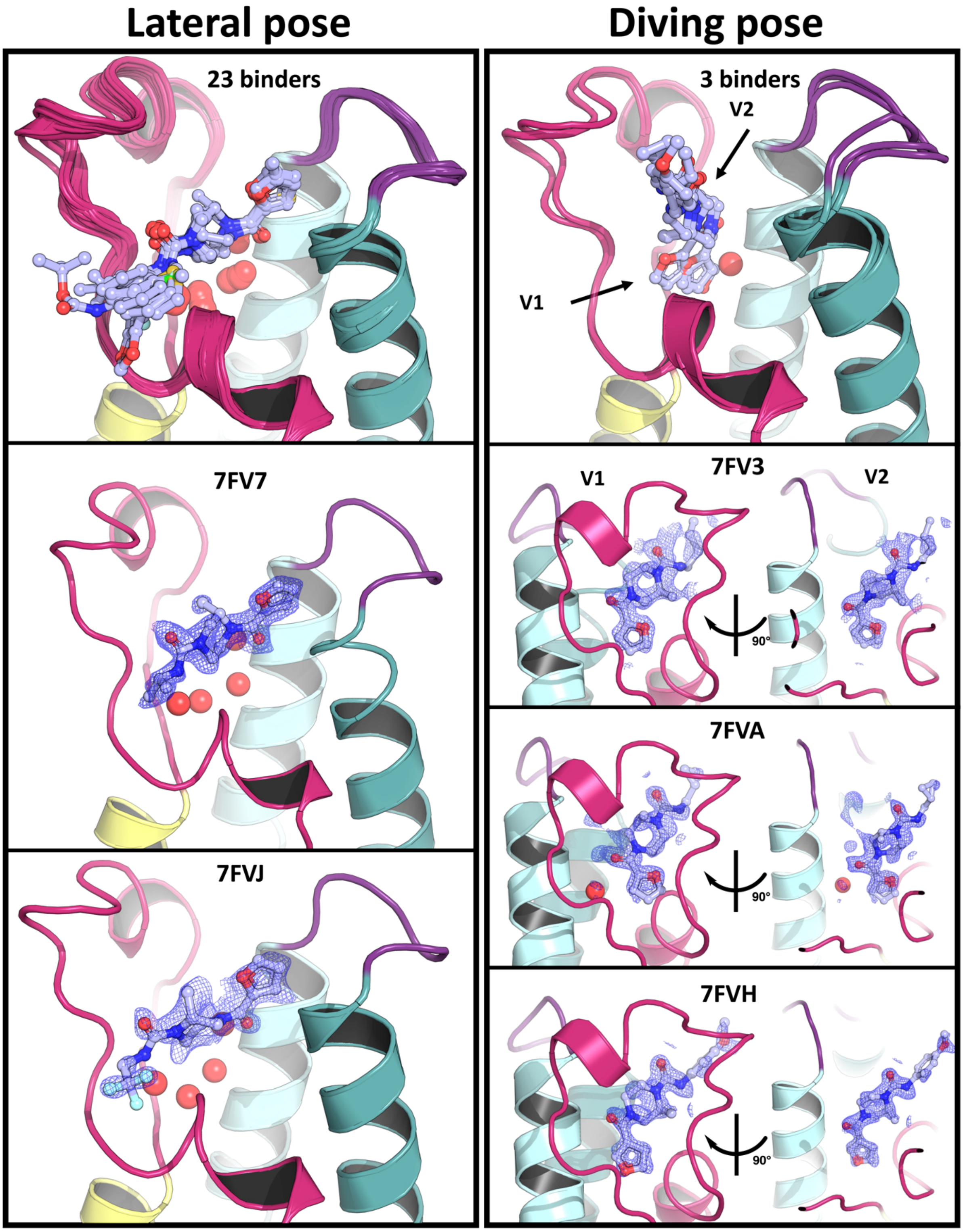


Figure 5.6: Rescreening of pure compounds resolves false negatives and reveals outliers poses and chemotypes. The top of each column shows lateral and diving false negatives. The left column shows the structures and associated PanDDa event maps for lateral false negative outliers bearing a methylpiperazine. The right column also shows structures and associated PanDDa event maps for all 3 diving false negatives.

Structural and electron density inspection revealed interesting features for binding outliers. To fit binding events, visual inspection of the PanDDa event maps was required due to their weak natures (**Fig. 5.6**).[24] The screening of pure compounds revealed 23 lateral binders with a similar pose to the previous experiment presented in Chapter 4, where elaborations extend across the ZA-channel (**Fig. 5.5**).[150] However, two lateral binders (7FV7 and 7FVJ) had an unexpected single methyl addition to the conserved piperazine core, pointing towards the inside of the binding site (**Fig. 5.6**).

All diving binders had a furan binding in the water cavity, which was not previously resolved.[150] These binders also exhibited structural differences compared to the initial screening of crude reaction mixtures presented in Chapter 4. In two cases (7FV3 and 7FVH), the methyl group was resolved on the other side of the piperazine's nitrogen-nitrogen axis, than previously resolved.[150] In one of these cases (7FV3), a furan-3-carbonyl was resolved instead of the furan-2-carbonyl found in the other two binders. One furan-2-carbonyl binder (7FVA) had a piperazine methyl in an R conformation similar to that of the previously resolved diving binders in Chapter 4. For this compound, its pose was shifted towards the ZA-channel which was paired with the re-insertion of a water-network molecule (**Fig. 5.6**).

5.3.4. Benchmarking binding scores against ligand-based approaches

To test the performance of the positive and negative binding scores, these were benchmarked against other commonly used ligand-based approaches. A training set was created by excluding the 97 rescreened compounds from the initial 957 compound large dataset, thus totalling to 860 compounds and initial screening outcomes. The test set consisted only of those 97 compounds purchased in pure with new rescreening binding outcomes (**Fig. 5.6**). The conserved binding and non-binding bits were extracted from the training set only and used to score testing set compounds.

The benchmark included a random forest classifier with hyperparameters optimised against the test set (**Fig. 5.7**). It is important to note that random forest hyperparameters were tuned on the test set to obtain the best possible results for comparison purposes, even though this may lead to overoptimistic results. Random forest classifiers are commonly used in ligand-based predictions such as for bio-availability, bioactivity and toxicity.[154–156] Two Tanimoto chemical similarity distance-based scores were also included in the benchmark. The average

Tanimoto coefficient value of a target compound against a set of experimental binders is commonly used in catalogue screening experiments[157–159]. The maximum Tanimoto coefficient of a target compound aginst a set of experimental binders is expected to efficiently identify false negatives, as it rewards compounds highly similar to known binders.[160] The precision recall are under the curve was selected as comparison metric as it focuses on positive data and better handles data imbalance.[161]

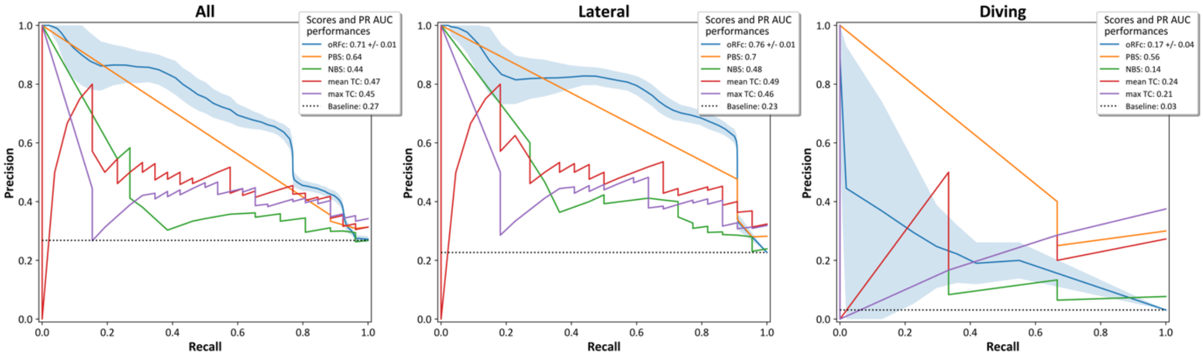


Figure 5.7: Benchmarking ligand-based methods for reproducing crystallographic rescreening results. Tested methods include an optimised random forest classifier (oRFc), positive and negative binding scores (PBS and NBS), and mean and maximum Tanimoto coefficients (mean TC and max TC), shown in blue, orange, green, red, and purple, respectively. The baseline, shown as a dotted black horizontal line, represents the hit rate associated with each data subset. A PR-AUC equal to the hit rate indicates random prediction. The error plotted and recorded for the optimised random forest were related to precision and overall PR-AUC values, respectively.

The positive binding score was the second-best performing metric for the “all” and “lateral” datasets where it performed similarly to the random forest in the low to mid recall regions (**Fig. 5.7**). For the diving pose data subset, the positive binding score was the best classifier with other metrics, including the optimised random forest, performing poorly. Tanimoto-based metrics and the negative binding score performed poorly across all data subsets (**Fig. 5.7**). The optimised random forest had the best performance for the “all” and “lateral” data subsets despite some associated errors. Again, the random forest hyper-parameters were optimised against the task thus likely generating overoptimistic predictions over a blind scenario.

5.3.5. Virtual screening identifies novel informative series

The binding scores presented above were used to prospectively screen a catalogue of commercially available compounds to select further follow-up candidates (**Fig. 5.8**). Additionally, a constrained docking step was added, using Fragmenstein, to offer additional structure-based filtering opportunities based on affinity predictions.[109] The pipeline below was applied twice, once for the lateral and once for the diving pose data subsets.

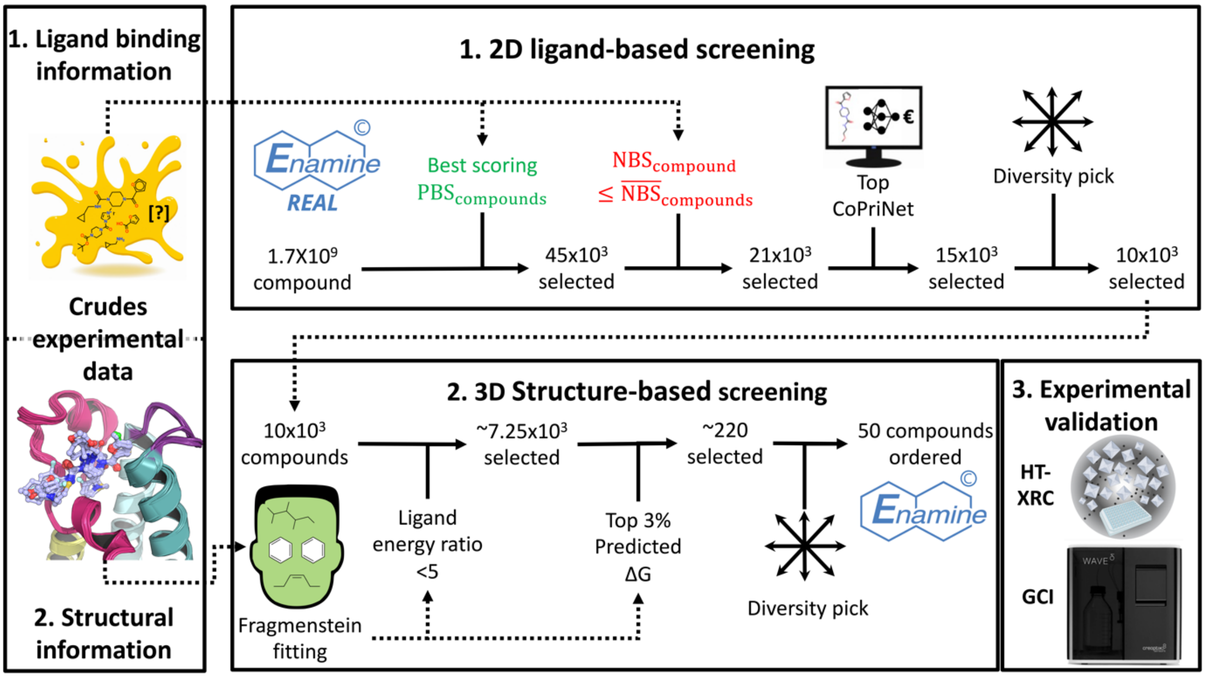


Figure5.8: Enamine Real subset virtual screening selects 50 compounds per pose for experimental validation. The crystallographic screening of crude reaction mixtures was used in two different scales for an integrative approach. The positive and negative binding scores were used to extract compounds from the ligand catalogue, and then binding predictions using bound structures were made using Fragmenstein. 50 compounds per pose were selected for further analysis based on energy scores.

Initially, a 2D ligand-based virtual screening (done by Dr Ruben Sanchez-Garcia) on a fraction of the Enamine Real library, which totals to approximately 1.7 billion compounds (**Fig. 5.8**).[162] The catalogue was filtered to retrieve the top 45,000 compounds based on a positive binding score. Compounds with negative binding scores greater than or equal to the average, amongst those selected 45.000 compounds, were removed (**Fig. 5.8**). CoPriNet was used to removed compounds predicted as expensive, and a chemical diversity selection was applied to select for a more diverse pool of compounds.[163] This process resulted in the selection of 10,000 compounds per pose for the ligand-based screening thus totalling to 20,000 compounds.

Subsequently, Fragmenstein[86] was used to fit the selected compounds into product-bound crystal structures, therefore also utilising the structural information obtained from the initial crystallographic screening of crude reaction mixtures (**Fig. 5.1**).[150] Unphysical poses were eliminated using the ligand energy ratio, which accounted for approximately 26% of the fitted compounds. The top 3% scoring compounds based on the Rosetta All-Atom Energy Function[164] were then selected for a final chemical diversity filtering.

The final selection included 47 (**Suppl table 5.2**) and 46 (**Suppl table 5.3**) compounds for the lateral and diving pose, respectively, totalling to 93 compounds for crystallographic validation at the XChem (**Fig. 5.8**).

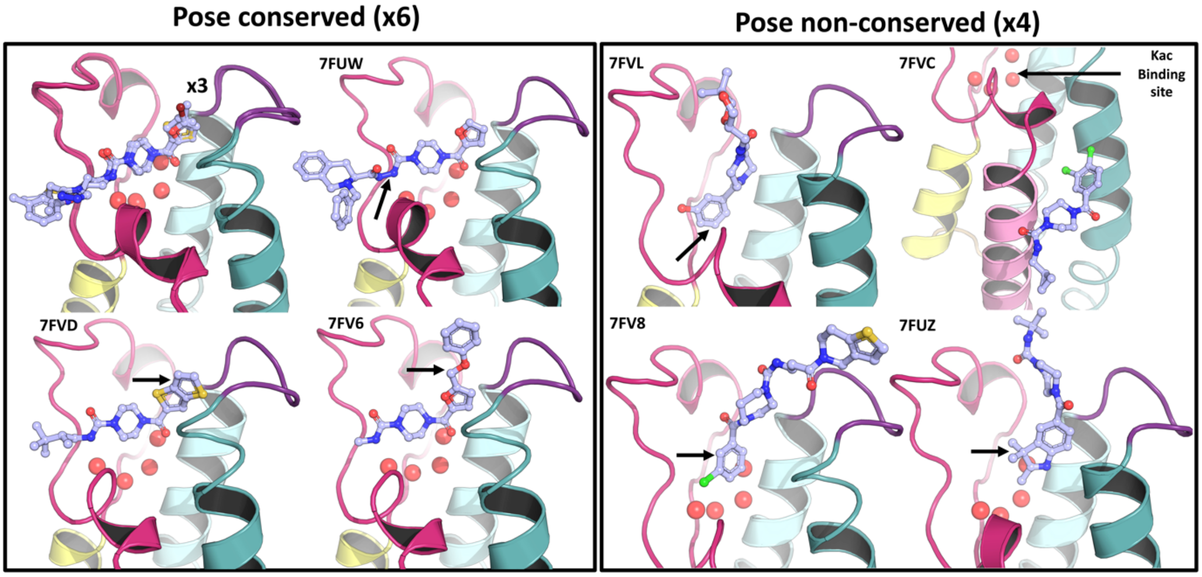


Figure 5.9: Crystallographic validation of selected compounds reveals novel binding poses, chemistries, and elaboration vectors. The identification of 10 compounds resulted in 6 maintaining the initial lateral pose (left panel). Among the lateral binders, 3 displayed novel features, while 4 exhibited modified binding (right panel) compared to the initial pose.

Crystallographic evaluation of pure catalogue compounds yielded 10 compounds, achieving a global hit rate of 10.75% (10/93) (**Fig. 5.9**). All bound compounds resulted from virtual screening applied to the lateral pose, implying pose-specific hit rates of 21.28% and 0.0% for the lateral and diving poses, respectively. Among the 6 binders with a conserved lateral pose, 3 had a similar molecular structure to previously resolved compounds, while the other 3 presented novel features: one had a 1,2-diformylhydrazine bond forming new hydrogen bonds (7FUW), another had an unobserved bicyclic thieno[3,2-b]thiophene ring system that made amplified hydrophobic contacts with neighbouring amino acids (7FVD), and the third had an anisole extension to the furane that bound in a previously unsampled vector (7FV6).

The remaining 4 binders exhibited modified conformations with respect to the lateral pose used for template fitting (**Fig. 5.9**), including a phenol ring that displaced water molecules with an unaltered piperazine ring occupying the central cavity (7FV8), 2 compounds that flipped and translated without displacing the water network and had long extensions towards the solvent (7FV8 and 7FUZ), and one instance where the follow-up was bound at a different site (7FVC) located between α-helices Z and C (**Fig. 5.9**).

Additional, validation necessitates the evaluation the interacting strength between the follow-up and the target, PHIP(2). A grating-coupled interferometry (GCI) assay[165] was used to test the binding kinetics for the 194 compounds, thus including both the compounds purchased (**Suppl table 5.1**) for retrospective rescreening and prospective virtual screening (**Suppl tables 5.2 and 5.3**). This allowed for the evaluation of differential binding affinities between the two sets of compounds.

Binding events were further considered if the k_a and k_d errors were lower than 25% and paired with an R_{max} value greater than 10. Although no universally accepted hit-calling criteria exist, the thresholds mentioned above are relatively stringent and applied to potentially minimise the likelihood of false positives in comparison to more lenient values.

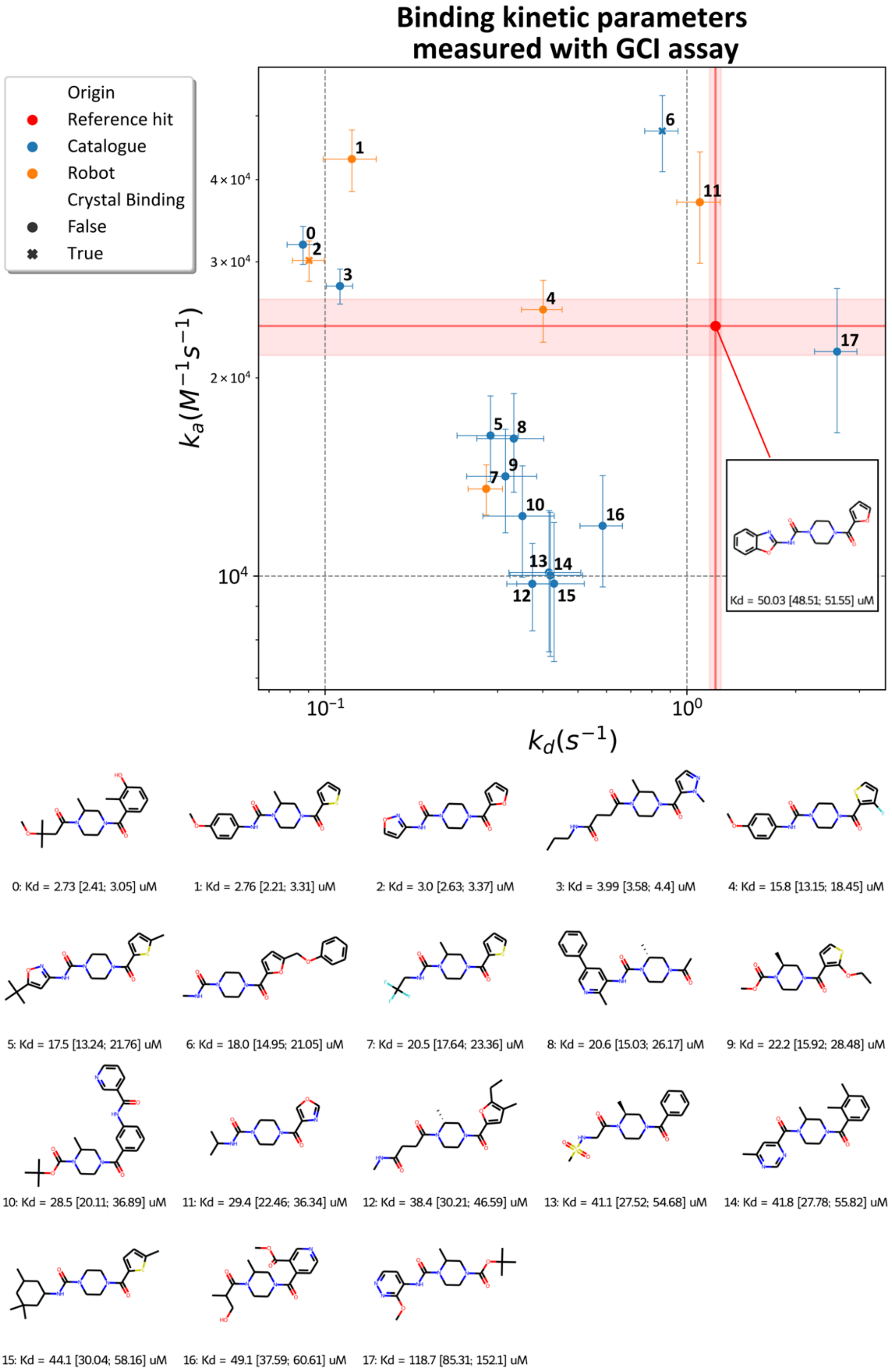


Figure 5.10: Selected GCI assay binders exhibit increased affinity against the initial robotic chemistry hit and chemotypes unresolved in crystal structures. k_d and k_a values are presented in the top panel, with associated errors, in second^{-1} and, molar^{-1} per second^{-1}, respectively. The reference binder, resolved from crude reaction mixtures, is represented by a red dot and shaded regions indicating k_a and k_d values and errors. Hits resolved or unresolved in crystals are marked by points or crosses, respectively. 2D ligand structures for each GCI assay hit are shown on the bottom panel, with identifiers corresponding to the affinity scatterplot. Final K_d values, expressed in micromolar (µM), with upper and lower ranges calculated from k_a and k_d errors, are displayed below corresponding 2D structure. ka and kd are indicative of the association and dissociation rates with Kd, the dissociation constant, equalling to association over the dissociation rates.

GCI kinetic measurements identified 18 high-quality binding events, with 5 from the robot chemistry library and 13 from virtual screening. Four compounds showed an approximate 10-fold increase in binding affinity compared to the assay binder, previously identified in Chapter 5 (**Fig. 5.10**). Ten compounds contained a methylpiperazine ring, initially associated with the diving pose. Only two compounds were resolved in crystal structures: binder 2 was a crystallographic false negative (**Fig. 5.6**), and binder 6 was the virtual screening compound with the anisole vector extension (**Fig. 5.9**, 7FV6).

Compounds with larger 6-membered rings or modified 5-membered rings contained a methylpiperazine. Binders 0, 10, 13, 14, and 16 had a benzene ring replacing the original 5-membered ring (**Fig. 5.10**). Binders 1 and 7 exhibited high scoring positive and negative binding score values in the diving pose (**Fig. 5.5**), while binders 2, 5, and 15 had features associated with lateral binders (**Fig. 5.3**). Binder 4 was almost identical to a diving crystallographic binder but lacked the methyl on the piperazine. Binder 11 had a triheterocyclic ring that was systematically not resolved in crystals, although its positive binding score was high in the lateral pose. Binders 12 and 13 had long flexible extensions lacking a urea, and binders 8 and 17 lacked the conserved amide aromatic ring (**Fig. 5.10**).

5.4. Discussion

The study initially aimed to rationalise the binding landscape of fragment F709 elaborations, designed against PHIP(2)[150], by extracting ligand features relevant for crystallographic binding and non-binding. The initial chemical landscape was generated by a robotic synthesis and resolved using X-ray crystallography in Chapter 4 (**Fig. 5.1**). The congeneric nature of the elaboration and limited dataset size favoured the development of a simple heuristic (**Fig. 5.2**) over the use of machine learning, which is commonly used for feature extraction or classification.[166] Molecular features associated with binding and lack thereof were extracted and compared between each pose, highlighting the significant role of methyl in inducing the diving pose (**Fig. 5.3**).

Aggregation of these features into the positive and negative binding scores (**Fig. 5.2**) allowed the identification chemical landscape regions associated with crystallographic follow-up binding and non-binding, respectively (**Fig. 5.4**). Certain chemical space regions exhibited conflicting scores due to follow-ups recapitulating binding features but also containing elaborations linked to non-binding events, such as sulphonamides. The binding scores guided the identification of experimental false negatives (**Fig. 5.4**), which was validated by crystallographic rescreening of pure compounds (**Fig. 5.5**). This led to the identification of 23 and 3 false negatives for the lateral and diving pose, respectively (**Fig. 5.6**).

The SAMPL7 challenge presented in Chapter 3 and Baker et al., (2020) also identified crystallographic false negatives through repeated screening, highlighting the occurrence of crystallographic false negatives.[43,130] However, studies experimentally addressing data misclassification in drug discovery are scarce, although false negatives exacerbate data imbalance, which can confound predictive technologies (**Suppl fig. 5.7**) or lead to poor generalisation performance.[167,168]

Identifying outliers, such as the piperazine methyl in the lateral pose, with previously unseen features, expands the knowledge of pose-specific binding features (**Fig. 5.6**). This suggests that features may act cooperatively or destructively to drive pose-specific binding, highlighting the possibility of examining feature combinations for a better understanding of ligand crystallographic binding, rather than considering them as a static ensemble. A single modification, such as methyl addition, is unlikely to fully explain changes in binding pose. Instead, multiple factors should be considered. Malhorta et al., (2017) suggested that ligand flexibility and other factors contribute to binding pose changes, with weak ligands being more susceptible to alterations.[169] Protein flexibility[170] and water dynamics[171] were also found to participate in ligand rearrangement, which may apply here given the dynamic[124] and hydration of bromodomain binding sites.[91] One diving compound showed some evidence of alternative lateral binding in the event map, but insufficient for proper fitting while another had a shifted pose with the re-insertion of a water molecule (**Fig. 5.6**).

The piperazine ring's methyl position were resolved at different locations, suggesting that event maps may reflect a superposition of states (**Fig. 5.6**).[43] Virtual screening structural hits demonstrated further pose changes with one binding in the diving orientation, but lacking previously identified conserved diving features, while another changed the binding site (**Fig. 5.9**).

Increasing hit diversity reduces the non-binding bits ensemble (**Suppl fig. 5.1**), which is approximately 60 times more populated than the conserved binding bits in both pose-specific data subsets (**Fig. 5.3**). Benchmarking based on the negative binder score (**Suppl fig. 5. 6**) performed worse than the positive binding score. Categorising non-binding bits is challenging as they address an unobserved signal, which has proven to be noisy and can also be more influenced by data biases (**Fig. 5.5**). Non-binding events are however essential to consider as they are populated, here, the majority of crystallographic readouts (**Fig. 5.1**). Focusing solely on binders by disregarding non-binding events creates a case of survival bias that has proven to be detrimental in various fields[172] and may be relevant for virtual screening, where negative data is rarely included in follow-up selection. Recent research has shown that incorporating decoys into the training set improves predictions, emphasising the importance of considering non-binding events.[173]

The detected associations and correlations between features and binding outcomes might stem from the original compound designs, adding a layer of complexity to the accurate feature identification process. Indeed, the different iterations did not systematically consider R group combinations. Consequently, there could be features that display correlation, in that they consistently appear together or are mutually absent, which results in their association with both binding and non-binding events. However, it is important to emphasise that the presence of correlation does not unequivocally infer causation. A single feature from a correlated set may be the actual driver for a binding or non-binding event. This underscores the importance of careful experimental design that aims to systematically consider feature combinations and diversity, thereby enhancing the efficacy of subsequent analyses and enabling credible causal inferences.

Non-binding compounds with high positive and negative scores also raised questions about the reliability of crystallographic readouts (**Fig. 5.5**). In some cases, the reason for their lack of binding remained unexplained. Non-binders were either chimeras of existing binders or had seemingly suitable extensions that would extend across the ZA-channel, some previously resolved in SAMPL7 fragments such as F368 and F369 (PDB: 5RKW and 5RKJ, respectively).[130] Other modifications, such as replacing the diheterocyclic by a triheterocyclic ring or adding a sulphonamide/nitro group, led to non-binding and are more difficult to rationalise (**Fig. 5.5**). These modifications can increase compound volume, polarity or solubility at a hydrophobic location which may promote the ligand to go into solvent.[174] However, replacing a 5-membered with a 6-membered ring modifies piperazine and linking amine conformation, disturbing the hydrogen bond with the nearby serine thus explaining non-binding (**Fig. 5.5**).

The validation analysis of rescreening robotically generated compounds in pure form showed significant enrichment between binders and non-binders despite apparent irregularities in crystallographic readout (**Fig. 5.5**). Virtual screening non-binders align with the crude reaction mixture and validation results, as they exhibit recurring non-binding features, such as 6-membered rings, tri-heterocyclic rings, and diazepane (**Suppl tables 5.2 and 5.3**). This underscores the significance of properly accounting for non-binding characteristics. The absence of virtual screening crystal binders, for the diving predictions, may also result from an incomplete categorisation of conserved bits due to the small number of initial binders. It is important to note that virtual screening was performed with no prior knowledge of the rescreening results of potential false negatives (**Fig. 5.8**). The validations by crystallography and GCI assays (**Fig. 5.10**) were simultaneous and identical for both compound sets, enabling comparison of the results. Thus, it expected that the results gathered, here, will improve future screening campaigns.

Benchmarking against other ligand-based methods was encouraging for the positive binding score, outperforming negative binding score and other Tanimoto-based metrics (**Fig. 5.7**). The positive binding score performed better than the random forest for the diving pose, operating at unbalanced data regimes inappropriate for random forest classification. The random forest classifier outperformed positive binding score for "all" and "lateral" data subsets but had similar performance between low and mid recall regions (**Fig. 5.7**). In resource-limited compound discovery, high precision is more important than high recall.[166] Thus, methods with similar behaviours at the beginning of the curve should be considered equally effective (**Fig. 5.5**). The random forest was also optimised against the test set and likely performs better than a blinded scenario. Validation and benchmarking were less conclusive for the “all” versus the “lateral,” showing that mixing pose information can confounds classification.

The positive and negative binding scores were about 20-fold faster than the random forest (**Suppl fig. 5.8**), potentially enabling deeper catalogue search whilst incorporating target-specific information. In comparison, Amendola et al. (2021)[175] developed a target-specific ligand screening approach which can process up to 272,000 compounds in 1 hour using multiple CPUs whereas scoring with both positive and negative binding scores can process about 145,000 compounds per hour on only a single CPU (**Suppl fig. 5.8**).

Robotic and virtual compounds were evaluated for affinity using grating-coupled interferometry assays (GCI) which measured low micromolar binding events (**Fig. 5.10**). Four compounds had up to 10-fold increase in affinity and ligand efficiency compared to the previously identified ligand in Chapter 4.[150] These assay binders displayed various chemotypes, including the piperazine methyl ring (**Fig. 5.10**), which favours the diving pose in crystal structure. Non-systematic R-group substitution make simple structure-activity-relationships deduction harder whereas the limited data points restricted the use of machine learning for quantitative structure-activity-relationships.

Most GCI binders were not resolved by crystallography (**Fig. 5.10**), which presents a challenge in relating binding affinities to binding site context. The difficulty in resolving assay binders from soaked crystals is a known limitation that may result from the inability of the lattice to adapt to the conformation required for ligand binding.[27] These findings emphasise the difference between crystallographic and assay structure-activity-relationships, which are both essential to structure-based drug discovert efforts.

5.5. Conclusion and future direction

A simple heuristic was developed to extract fragment follow-up binding and non-binding features from noisy datasets (**Fig. 5.2**), enabling crystallographic structure-activity rationalisation. The methyl's role in pose flipping was highlighted (**Fig. 5.3**), and the binding landscape was contextualised in terms of existing and missing molecular features in F709 follow-up elaborations (**Fig. 5.4**). Future research could explore refining feature identification, using bit frequency thresholding or weighting,[176] and enriched by adding 3D context[177] using machine learning[178] or alchemical binding free energy calculations.[179] Using 2D pharmacophoric fingerprints could increase feature generalisability across more chemically diverse ligand series.[180,181]

The scoring schemes developed, here, were useful for identifying crystallographic false negatives, as rescreening revealed 26 binders, thus doubling the initial hit rate (**Fig. 5.5**).[150] Enrichment analysis based on positive binding scores was significant in all 3 data subsets, but some non-binders had high positive binding scores, thus questioning the reliability of crystallographic readouts (**Fig. 5.5**).

Lateral outliers with methylpiperazine were also observed, suggesting combinatorial factors driving pose changes. The position of the methyl in the diving pose was different than previously observed (**Fig. 5.6**). To understand these dynamics, an integrative approach may be needed and may include room temperature crystallography or Nuclear Magnetic Resonance (NMR) spectroscopy.[182,183] Isothermal Titration Calorimetry (ITC) may help resolve the thermodynamics and entropic changes linked to water displacement.[184] Absolute binding free energies of different conformers may help identifying the lower energy stereo-isomer and pose.[185]

The positive binding score outperformed other Tanimoto-based metrics and the negative binding score in benchmarking (**Fig. 5.7**). Notably, the positive binding score outperformed all other metrics for the diving pose, indicating that it can be applied to degraded datasets with few positives (**Fig. 5.7**). To further improve performance, a consensus approach combining positive and negative binding scores could be explored.

The positive and negative binding score facilitated fast catalogue searches in prospective virtual screenings, thus incorporating site-specific information (**Fig. 5.8**). The virtual screening identified 10 crystal binders with informative chemistry and elaboration vectors, and previously unresolved poses, including a case where a compound changed binding sites (**Fig. 5.9**). GCI assays revealed hits with up to 10-fold increase in affinity and ligand efficiency than the hit previously identified in Chapter 5 (**Fig. 5.10**).[150]

The wealth of acquired structural information could be combined with template docking of the assay binders, not resolved in crystals, and binding free energy calculations to rationalise the assay results from a structural perspective.[186] While crystallographic structure-activity relationships was the primary focus in the initial dataset, it is also important to account for differential binding between structural and assay-based methods. In addition to soaking, co-crystallisation of assay binders may also provide bound structures and should be considered as an alternative approach.[27]

Overall, this work presents a novel framework for interpreting noisy crystallographic binding landscape. This feature extracted from these can be used for incorporating target-specific information in virtual screening. This framework is expected to be transferable to different targets and fingerprints and will hopefully support the analysis of future robotically assisted fragment elaboration campaigns. Future research possibilities were also highlighted to potentially improve the performance of the method and rationalise the wealth of data generated from the computational, crystallographic, and kinetic experiments.

5.6. Methods

5.6.1. Small molecule fingerprinting and dimensionality reduction.

Feature-based invariant Morgan fingerprints were generated using a length of 2048 and a radius of 6 with RDKit and used as input for feature extraction (detailed bellow) and t-Distributed Stochastic Neighbourhood Embedding using scikit-learn.[80] The fingerprint bit vectors were embedded in 2 dimensions allowing for visualisation using the Jaccard distance metric with a perplexity of 30. Chemical similarity calculations of molecule pairs were performed using RDKit.[79]

5.6.2. Bit and compound scoring methodology

The small molecules SMILES strings were first translated into bit vectors using Morgan fingerprints (as detailed in above) and partitioned into two subsets: hits and non-hits, corresponding to their crystallographic outcomes. Initially, the unsampled bits, which are never observed in either hits or non-hits, were eliminated from the array. Subsequently, the Bit Conservation Score (BCS) was computed for each bit, representing the activation rate of a specific bit across binders.

$$\textbf{Bit Conservation Score} \qquad \mathrm{BCS} = \frac{\sum \mathrm{bit_{on}}}{\mathrm{N_{hits}}} \qquad \textbf{Equation 5.1}$$

Where $\sum \mathrm{bit_{on}}$ is the sum of activated bits across crystallographic hits for single bit and $\mathrm{N_{hits}}$ is the number of crystallographic hits. Each bit must then be classified based on the conserved binding and non-binding bit thresholds which define the Bit Conservation Score at which a bit is classified, respectively.

$$\textbf{Conserved binding bit} \qquad \mathrm{BCS} \geq \theta_{\mathrm{CBB}} \Rightarrow \mathrm{CBB} \qquad \textbf{Equation 5.2}$$

$$\textbf{Conserved non-binding bit} \qquad \mathrm{BCS} \leq \theta_{\mathrm{CNB}} \Rightarrow \mathrm{CNB} \qquad \textbf{Equation 5.3}$$

Where θ_{CBB} and θ_{CNB} are the Bit Conservation Score thresholds defining the conserved binding bit and conserved non-binding bit labels for individual bits. If those criterions are not met, then the bit is labelled unconservative. This enables calculating the positive and negative binding scores for individual compounds. The positive binding score is number of activated conserved binding bit across one compound over the total number of conserved binding bit identified. The negative binding score is one minus the number of activated conserved non-binding bit accross one compound over the total number of conserved non-binding bit identified.

Positive binding score $$\mathrm{PBS} = \frac{\sum \mathrm{CBB_{on}}}{\mathrm{N_{CBB}}}$$ **Equation 5.4**

Negative binding score $$\mathrm{NBS} = 1 - \frac{\sum \mathrm{CNB_{on}}}{\mathrm{N_{CNB}}}$$ **Equation 5.5**

Where $\sum\mathrm{CBB_{on}}$ and $\sum\mathrm{CNB_{on}}$ are the sum of activated conserved binding and non-binding bits within a compound, respectively. $\mathrm{N_{CBB}}$ and $\mathrm{N_{CNB}}$ are the number of activated conserved binding and non-binding bits identified, respectively. Compounds are then predicted as likely (LB) or unlikely binders (UB) using the positive and negative binding score thresholds.

Likely binder $$\mathrm{PBS} \geq \theta_{\mathrm{PBS}} \Longrightarrow \mathrm{LB}$$ **Equation 5.6**

Unlikely binder $$\mathrm{NBS} \geq \theta_{\mathrm{NBS}} \Longrightarrow \mathrm{UB}$$ **Equation 5.7**

Where θ_{PBS} and θ_{NBS} are the positive and negative binding score thresholds defining the score values at which a compound is labelled likely or unlikely binder, respectively.

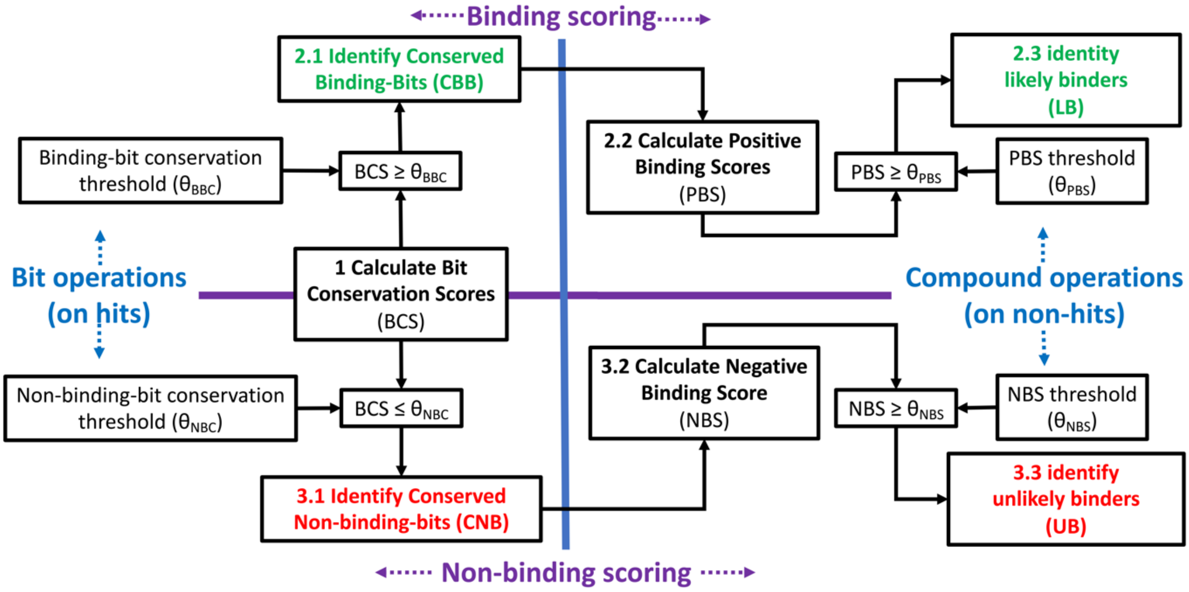

Binding scoring
2.1 Identify Conserved Binding-Bits (CBB)
2.3 identity likely binders (LB)
Binding-bit conservation threshold (θBBC)
BCS ≥ θBBC
2.2 Calculate Positive Binding Scores (PBS)
PBS ≥ θPBS
PBS threshold (θPBS)
Bit operations (on hits)
1 Calculate Bit Conservation Scores (BCS)
Compound operations (on non-hits)
3.2 Calculate Negative Binding Score (NBS)
Non-binding-bit conservation threshold (θNBC)
BCS ≤ θNBC
NBS ≥ θNBS
NBS threshold (θNBS)
3.1 Identify Conserved Non-binding-bits (CNB)
3.3 identify unlikely binders (UB)
Non-binding scoring

Figure 5.11: Algorithmic workflow for feature extraction and compound scoring. Fingerprint bits are first processed by calculating the bit conservation score across binders. The conserved binding and non-binding bits define the positive and negative binding scores on non-binders, respectively. Those scores are then used to define whether a non-binding compound is a likely or unlikely binder.

5.6.3. Random forest hyper-parameters optimisation and classification

The hyperparameters for the random forest model were optimised across all 3 data subsets employing the Optuna library and its Tree-structured Parzen Estimator algorithm.[187] The training set consisted of the initial 957 compound large dataset minus the 97 rescreened compounds from, thus totalling to 860 compounds. The test set consisted only of those 97 compounds rescreened in pure form. For both Feature-based invariant Morgan fingerprints served as inputs and crystallographic outcomes as ground truth.

The optimisation process comprised 1,000 initial trials followed by a total of 2,500 trials. Four hyperparameters were optimised: maximum tree depth, maximum number of features used at each split, class weight, and number of trees in the forest. The maximum depth was evaluated within a range of 1 to 40 units with a step size of 1, while the maximum features values were assessed between "None" and various options, including "sqrt", "log2", and float values from 0.6 to 1.0 with a step size of 0.1. The class weight was examined among four distinct options: None, "balanced", {0:1, 1:10}, and {0:1, 1:100}. Class 0 symbolises negatives or non-binders, and class 1 represents positives or binders. The number of trees in the forest was explored from 10 to 1,000 with a step size of 50. The objective criterion aimed to maximise the average precision-recall area under the curve across 20 repeats calculated with scikit-learn.

Random forest classifications were applied to the same validation and test sets using scikit-learn, employing the best-identified hyperparameters. The average precision-recall area under the curve across 200 repeats was recorded to evaluate performance. To compare model performances between raw and corrected datasets, leave-one-out cross-validations were executed in triplicate for all datasets and average precision-recall AUC recorded and utilised to assess performance.

5.6.4. Statistical analysis

The Mann-Whitney U test, implemented using the SciPy library, was chosen for testing how significantly the binding scores enrich the identification of binders over non-binder during the rescreening experiment.[188] Due to the limited sample size, independent nature of the data points (different ligands were screened) and the non-normal binding score distributions, a non-parametric statistical test was employed. The test was conducted with a two-sided alternative hypothesis to assess the significance of differences between the two groups.

The predictions generated by the different classifiers were ranked thus allowing for calculation of the precision-recall AUC. The precision-recall area under the curve accounts for class imbalance and is therefore superior to receiver operating characteristic curve.[161] The precision and recall correspond to the positive predictive value (PPV) and true positive rate, respectively.

Precision

$$\mathrm{PPV} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$$

Equation 5.8

Recall

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

Equation 5.9

Where TP, FP and FN correspond to the number of true positives (correctly classified positive cases), false positives (incorrectly classified positive cases), and false negatives (incorrectly classified negative cases), respectively. The precision-recall area under the curve value can be estimated by calculating the precision and recall values across possible threshold values.

Precision-recall area under the curve

$$\mathrm{PR\ AUC} = \int_0^1 \mathrm{PPV}(\mathrm{TPR})\mathrm{dTPR}$$

Equation 5.10

Where PPV(TPR) is the precision function at recall value TPR. This equation states that the PR-AUC can be obtained by integrating the precision function with respect to recall over the entire range from 0 to 1. Here this value was estimated using the trapezoidal rule implemented in scikit-learn.

5.6.5. Virtual screening

The initial ligand-based screening of a subset of Enamine REAL 1.7 (done by Dr Ruben Sanchez-Garcia). The database subset was assembled in April 2021 and was composed of about 1.7 billion compounds with heavy atom counts ranging between 6 and 38. The top 45,000 molecules were retrieved based on the positive binding score for the two-pose specific-datasets. Then compounds of higher or equal negative binding score than the average negative binding score across all 45,000 molecules were removed. CoPriNet[163] was then used for price prediction and the 15,000 cheapest scoring compounds in both datasets were selected for further processing. A diversity selection was performed to select the 10,000 most divers compounds, based on the tanimoto distance, for each datasets using the lazy MaxMin diversity picker implemented in RDKit. This resulted in a combined library of 20,000 molecules.

Prior to binding pose predictions, the crystalised ligands were separated from the protein and waters removed with MDAnalysis[189] providing template material for structure-based steps. Binding pose predictions for chosen follow-up compounds (done Dr Ruben Sanchez-Garcia) with Fragmenstein where the follow-up compounds were placed into the dehydrated co-crystal structure of the most chemically similar reaction product based on the tanimoto distance,. The template reaction product crystallographic conformers employed to fit the follow-up were the three most chemically similar ones. Five fits were performed for each follow-up compound. Fitted follow-ups were not processed further if the minimum ligand energy ratio exceeded 5. Subsequently, the top 3% of compounds were chosen based on their affinity scores. An additional diversity pick was applied to this selection, resulting in the choice of 50 ligands.

5.6.6. Crystallographic screening

Proteins were expressed, purified and crystalised as described in Chapter 2, Section 2.2.1. Crystallographic screening was also performed as described in Chapter 2, Section 2.2.2. The crystals were soaked for 24 hours at a final compound concentration of 8 mM. The compounds used for screening were purchased from Enamine stored in ethylene glycol at a concentration of 40 mM. Structural models and crystallographic statistics can be retrieved from the PDB deposition ID: G_1002265.

5.6.7. Grating-coupled interferometry assay

Pulsed single-concentration surface-based biophysical measurements of binding kinetics were performed as detailed in Chapter 2, section 2.2.3. Specifically, PHIP(2) was diluted to in to 5 µg/mL and immobilised to a final level of 6554 surface mass (pg/mm^2) corresponding to an injection time of 420 s. The running buffer was composed of 20mM HEPES, pH 7.5, 50 mM NaCl, 0.5% (v/v) Tween-20, and 0.5% (v/v) ethylene glycol.

The WaveRapid[68] GCI analysis method quantifies the error associated with both k_a and k_d values, expressing them as a percentage relative to the measured values. Consequently, these must be manually propagating to estimate K$_d$ error value to properly estimate the affinity range of a particular compound for the target. The errors were propagated using root-sum-square method.

Equilibrium constant error $$\mathrm{K_{d_{err}}} = \mathrm{K_d} \times \sqrt[2]{\left(\frac{\mathrm{k_a} \times \delta_{\mathrm{k_a}}}{\mathrm{k_a}}\right)^2 + \left(\frac{\mathrm{k_d} \times \delta_{\mathrm{k_d}}}{\mathrm{k_d}}\right)^2}$$ **Equation 5.11**

Where K$_d$ and $\mathrm{K_{d_{err}}}$ are the equilibrium dissociation constant and associated error, respectively. k$_a$ and δ_{ka} are the association rate and, associated error expressed in percent, respectively. k$_d$ and δ_{kd} are the dissociation rate and, associated error expressed in percent, respectively.

Chapter 6. Integrating structure-based methods for PHIP(2) fragment elaboration

6.1. Credits

The research and manuscript presented in this chapter were conducted and authored by Harold Grosjean. Harold Grosjean carried out hotspot mapping, virtual screening library generation, binding pose predictions and rescoring, molecular dynamic simulations, and free energy calculations with assistance from Dr Irfan Alibay, Dr William (Zhiyi) Wu, Dr Rocco Meli, and Dr Mihaela Smilova. Compound synthesis was kindly performed by Thomas Grimes under the supervision of Paul Brennan. Furthermore, Harold Grosjean performed co-crystallisation screening, crystallographic analysis, and grating-coupled interferometry assays with the help of Dr Lizbé Koekemoer and Dr William Bradshaw. Prof Philip C. Biggin reviewed and provided guidance on both the manuscript and the overall project.

6.2. Introduction

In Chapter 5, a simple approach was developed to extract important ligand features from large and noisy crystallographic screening of crude reaction mixture. The approach retrospectively focused on one fragment, F709, and its robotically generated elaborations before prospectively identifying assay binders. However, these did not surpass low micromolar potencies.

In this chapter, a larger gain in affinity was attempted by employing more conventional but intensive structure-based methods, applied to all 47 fragments resolved at the acetylated-lysine binding site. The discovery of an anti-PHIP(2) inhibitor could provide a pathway to treat various lethal and previously untargetable cancers. Here, crystallographic fragments were used to query databases and construct a library of chemically accessible fragment elaborations. The follow-ups were subsequently docked into their originating co-crystal structures. The resulting docks were filtered based on overlays against the fragment pose, selecting more favourable docking poses that reproduce crystallographic features. Molecular dynamic simulations of the best follow-up poses enabled free energy scoring via Molecular Mechanics Generalised Born Surface Area and Interaction Entropy methods, leading to the selection of six compounds for experimental validation with Grating-Coupled Interferometry assay and co-crystallisation. Experimental evidence suggested two compounds binding weakly in the Grating-Coupled Interferometry assay but lacking crystallographic evidence.

Overall, this chapter emphasises the complexity of early-stage drug discovery and underscores the importance of considering appropriate vector elaborations, ligand physico-chemical characteristics, binding site interactions, and conformational dynamics, along with target-specific experimental validation constraints.

6.3. Results

6.3.1. Investigating binding site interactions and dynamic fragment binding

To gain an initial understanding of the binding site and available interactions within it, ensemble hotspot maps were generated (**Fig. 6.1**).[29] Ensemble maps are a clustering of individual hotspot maps, then are utilised to discern pharmacophoric trends and fluctuations across multiple structures. This method allows for the identification of recurrently occurring features and the combination of more widely scattered pharmacophoric points. However, it is important to note that ensemble maps may overlook infrequent interactions that could still hold relevance. All 47 co-crystals structures were included in those calculations thus accounting for the conformational heterogeneity captured by the initial crystallographic fragment screening experiment. 3 types of maps were calculated. Apolar, acceptor and donor maps measure the inclination of a binding site location to favourably interact with a hydrophobic, hydrogen bond donor and hydrogen bond acceptor probe, respectively.

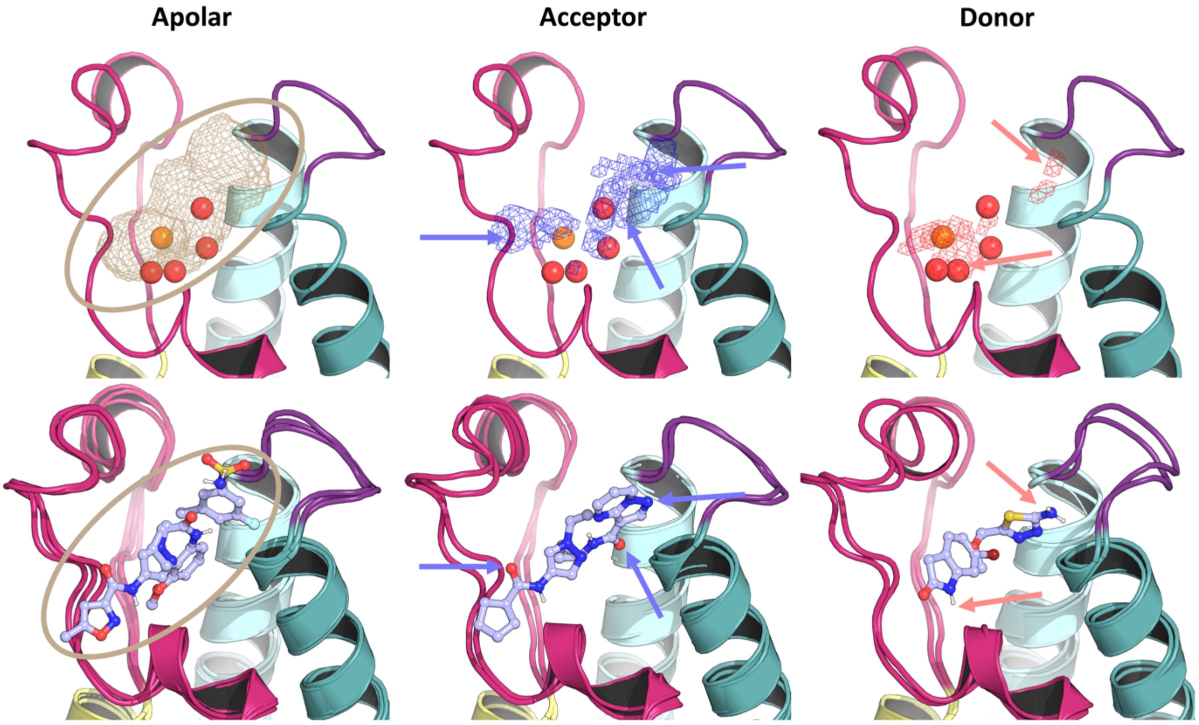


Figure 6.1: Ensemble hotspot mapping highlights primarily hydrophobic binding alongside polar interaction located around loop regions and a concealed water molecule at a donor site. The upper panel displays three distinct hotspot maps, with apolar, acceptor, and donor probes represented in beige, blue, and red, respectively. Apolar, acceptor, and donor correspond to probe types. The conserved bromodomain waters are depicted in red, while the hidden water at an acceptor site is illustrated in orange. The lower panel presents examples of fragment binding at hotspot regions.

The apolar map indicates a large density spanning the entire binding site and also engulfing the acceptor and donor points (**Fig. 6.1**). The acceptor hotspots (donor protein for acceptor ligand) were primarily located around α-helix B and the BC-loop, with some density also observed at the ZA-channel. The donor hotspots (acceptor protein for donor ligand) showed the opposite pattern, with greater opportunities at the ZA-channel, followed by the α-helix B and BC-loop regions. The fragment screening hits sampled the majority of the identified hotspots, except for one donor point corresponding to a water coordination site buried deep within the water cavity (**Fig. 6.1**).

Short molecular dynamic (MD) simulations were then conducted to evaluate the stability of the binding poses (**Fig. 6.2**) thus enhancing the static crystal structures information. A root mean square deviation (RMSD) cutoff of 2.5 Å was used to distinguish stable from unstable crystal poses, and a binding mode was considered stable if the median RMSD of the fragment heavy atoms over the trajectory was lower than the aforementioned cutoff.

Out of the 47 binders, 22 fragments were found to have an unstable binding pose, as defined by the RMSD cutoff (**Fig. 6.2**). These fragments tended to be relatively small, more solvent-exposed, and had their predominant moieties in the hydrophobic void, leading to fewer polar interactions with the binding site.

25 fragments had a more stable binding mode and tended to make polar and/or more interactions with the binding site (**Fig. 6.2**). For instance, F558 and F709 spanned the entire binding site from the ZA-channel to the BC-loop. F421 was relatively small in size but had a polar 5-membered ring that interacted with the BC-loop. F584 and F618 displaced 4 and 1 water molecules from the conserved network, respectively, with the former being the most stable and buried fragment.

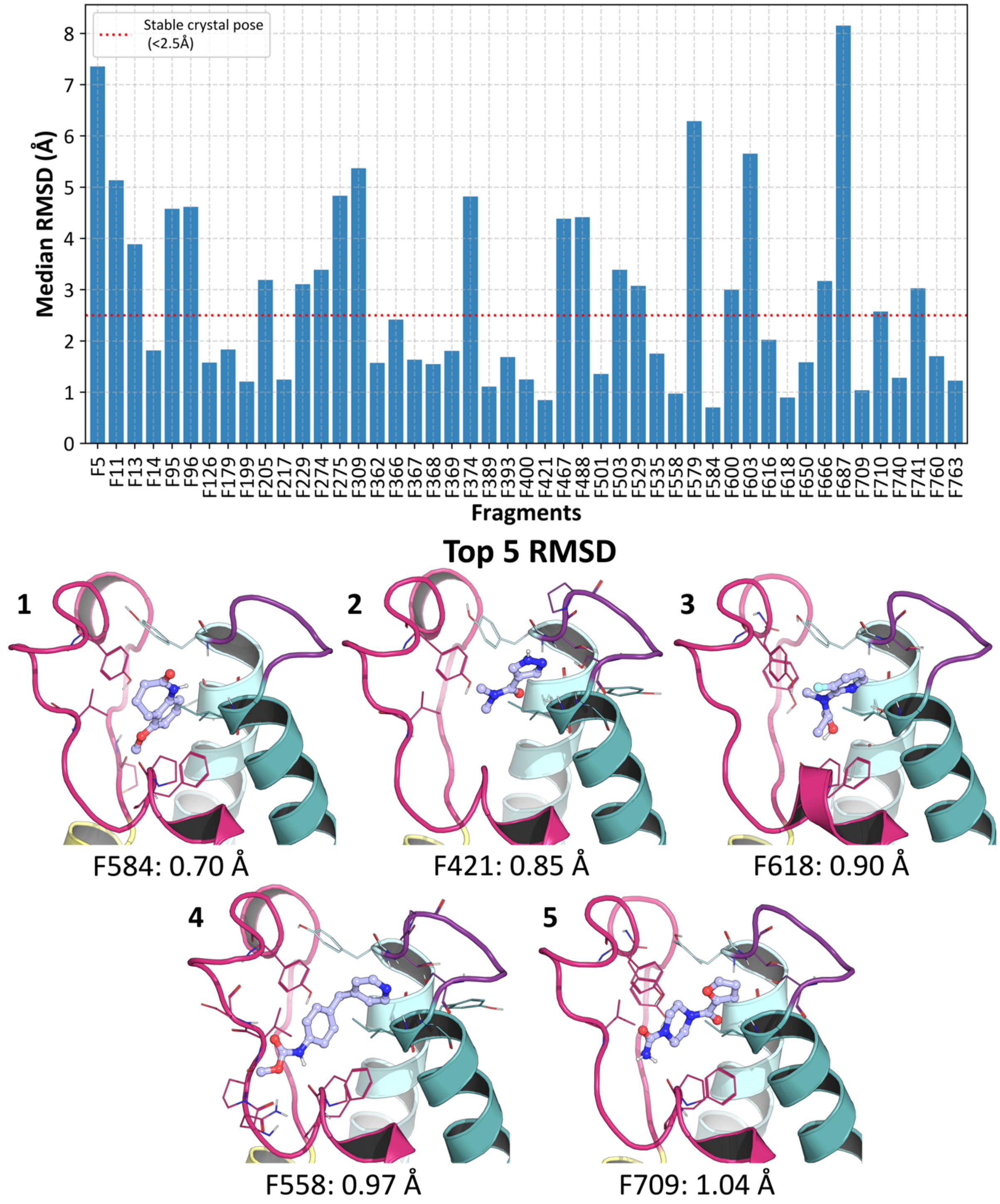


Figure 6.2: Crystallographic fragment poses demonstrate varying stability in molecular dynamic simulations. The upper panel presents the median RMSD for fragment crystal poses sampled throughout molecular dynamic simulation trajectories. A pose is deemed stable if its median RMSD is below 2.5Å, denoted by the dotted red line. The lower panel displays the top 5 most stable fragment crystal poses, with the fragment ID and corresponding median RMSD value provided below each crystallographic pose.

Free energy calculations provide a means to estimate binding affinity, which is crucial for understanding the thermodynamics of fragment binding and interaction strength. Here the enthalpy and entropy were estimated via Molecular Mechanics Generalised Born Surface Area (MMBSA)[190] and Interaction Entropy (IE)[93] calculations from MD simulation trajectories leading to an estimate of the Gibbs binding free energies for individual fragments (**Fig. 6.3**).

All but 9 fragments had negative (favourable) predicted binding free energies (**Fig. 6.3**). The unfavourable compounds tended to bear charged groups near polar protein regions (**Suppl fig. 6.1**). More stable poses (**Fig. 6.2**) tended to have more favourable predicted binding free energies indicative of better interaction with the binding site (**Suppl fig. 6.1**). For favourable fragment poses, the free energy ranking was relatively smooth with no clear best binder(s) (**Fig. 6.3**).

Inspection of top 5 predicted most favourable binders revealed a diversity of interactions and of chemical scaffold. F558 and F709 spanned the entire binding site from the ZA-channel to the BC-loop. F616, displaced a conserved water and interacted via a relatively rare (for this set of fragments) halogen bond α-helix B while F362 has a double hetrocyclic ring that interacts with nearby tyrosines (**Fig. 6.3**).

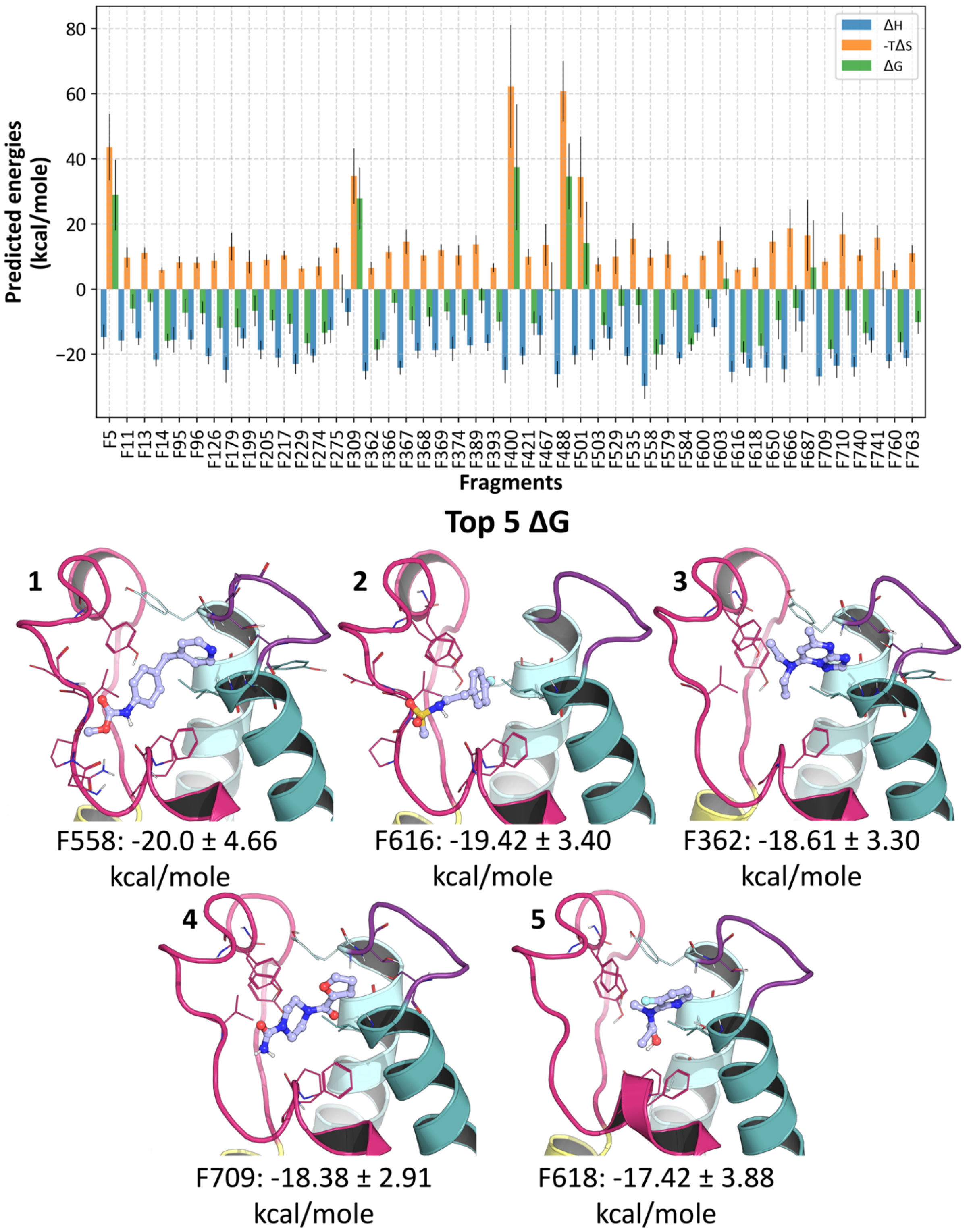


Figure 6.3: Crystallographic fragment poses exhibit a smooth predicted binding free energy landscape. The upper panel displays the predicted binding free energy terms for each fragment, with average enthalpy, entropy, and Gibbs binding free energy terms represented in blue, orange, and green, respectively. Enthalpy (ΔH) and entropy (-TΔS) are derived from MMGBSA and Interaction Entropy analysis of MD simulation trajectories, respectively. The Gibbs binding free energy (ΔG) is the sum of these two terms. The vertical black lines indicate errors associated with the calculations. The lower panel presents the fragment crystal poses with the most favourable predicted binding free energies.

6.3.2. Generating a library of synthetically accessible fragment elaborations

To increase fragment affinity, larger follow-up compounds can be selected so that they engage with more available interactions within the binding site (**Fig. 6.1**). To achieve this, the fragment SMILES strings were used to create a library of follow-up compounds using three different search tools (**Fig. 6.4**).

The substructure search is restricted to follow-up compounds that replicate the entire structure of the initial queried fragment. On the other hand, the similarity and fragment network[114] searches allow for some degree of variation of the starting fragment in the follow-up compounds. The follow-up compounds must also be chemically tractable to facilitate downstream synthesis and experimental validation compounds (**Fig. 6.4**).[191]

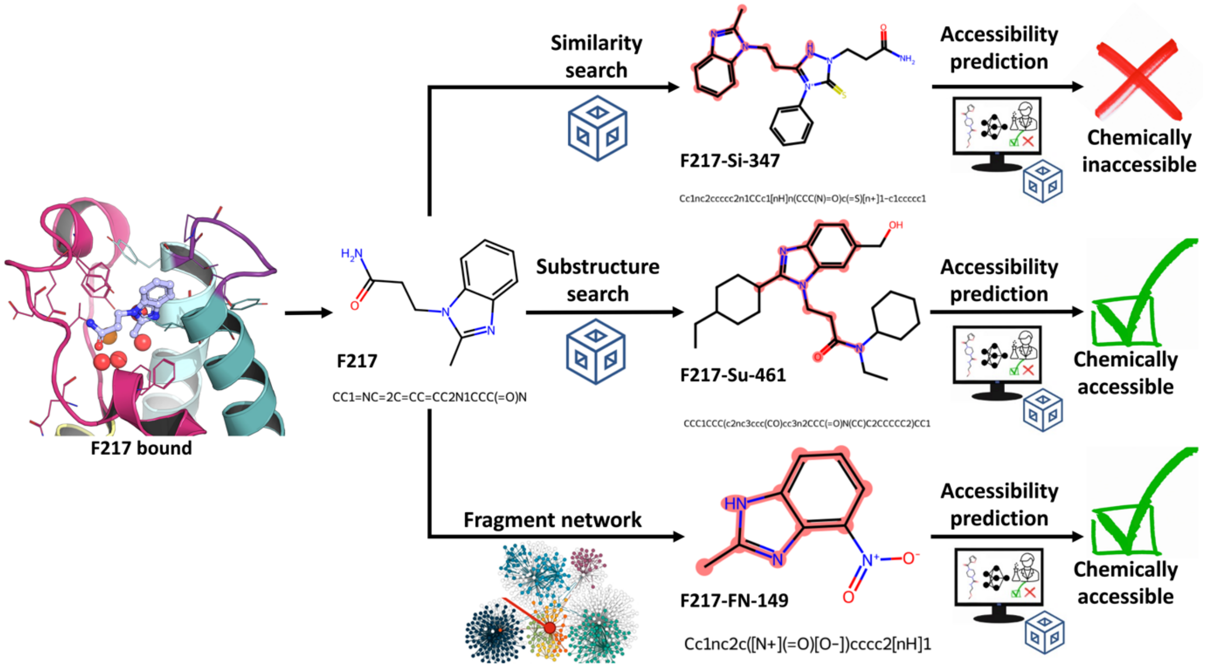


Figure 6.4: Crystallographic fragments enable database searches to construct a library of follow-up elaborations deemed chemically accessible. Initially, fragments were resolved crystallographically and inputed as 2D structures in SMILES format to seed database searches. Fragment network, manifold similarity, and substructure searches are applied to all fragments, generating follow-up compounds. The pink overlay represents the maximum common substructure relative to the originating fragment. All follow-ups are assessed for synthetic accessibility, and those deemed inaccessible are rejected from the final library.

The queries returned a total of 77,378 follow-up compounds. These compounds were then scored for chemical accessibility, and only those predicted to be accessible were retained for further analysis. This removed 23.11% of the initial library, leaving 59,507 follow-up compounds selected for further filtering, averaging to 1,266.11 follow-ups per queried fragment. To better understand the variability associated with each query, the number of follow-up compounds obtained per method per fragment was plotted (**Fig. 6.5A**). To visualise the chemical landscape defined by the final virtual library, a two-dimensional representation was generated using t-Distributed Stochastic Neighbourhood Embedding[80] (t-SNE) using Feature-based invariant Morgan fingerprints as inputs (**Fig. 6.5B**).[76]

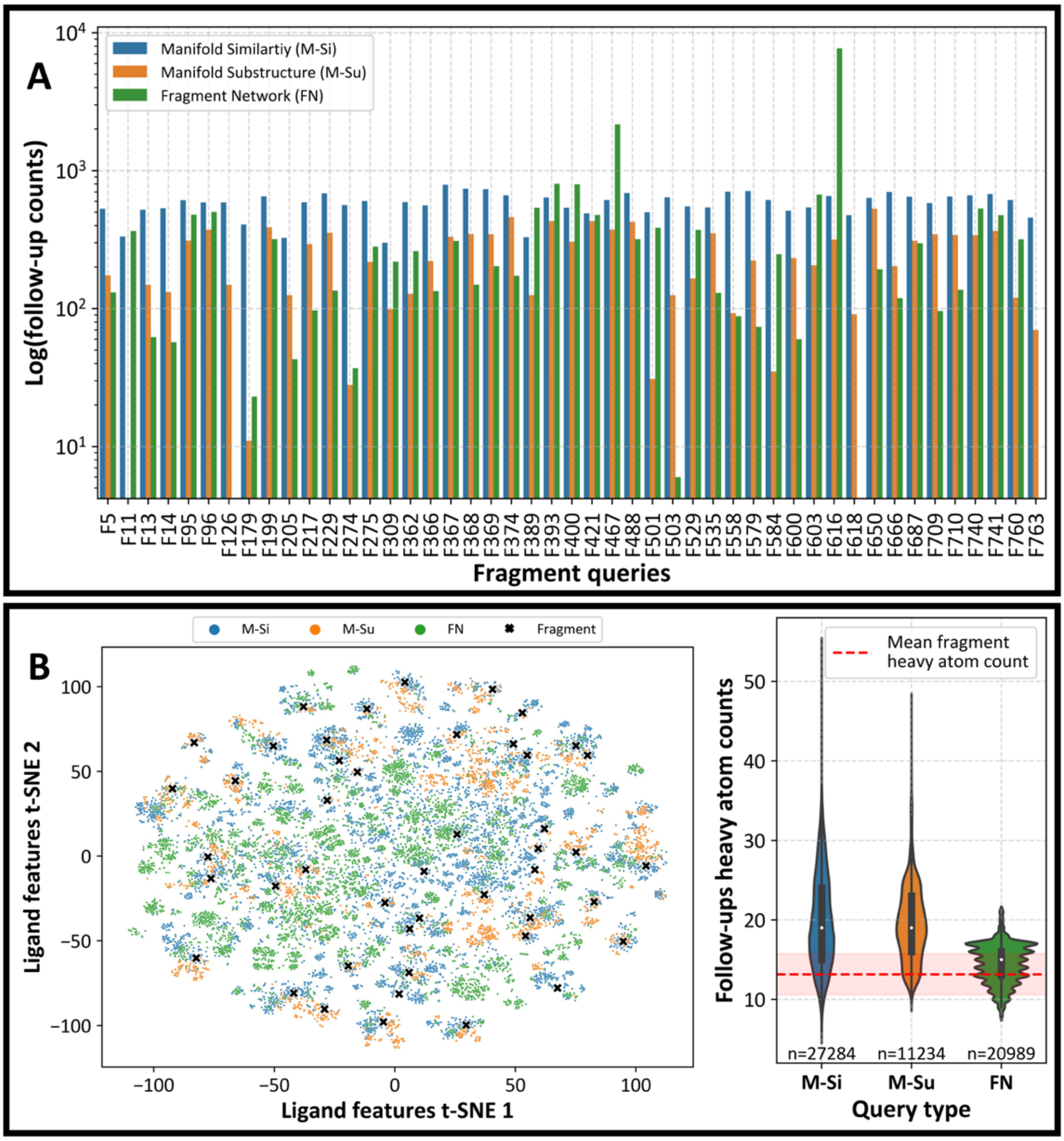


Figure 6.5: Library building reveals fragment-to-fragment variability in follow-up numbers returned and small size of fragment network compounds. The upper panel illustrates (A) the number of chemically accessible follow-ups per fragment per query method, using a log Y scale to mitigate visual effects of variable follow-up numbers. The bottom (B) left panel presents a 2D representation of the chemical landscape defined by the final library with query fragments showed as black crosses. The bottom right panel displays distributions of follow-ups' heavy atom counts within all three query types, with the number of compounds returned for queries shown below the distributions. Manifold similarity (M-Si), manifold substructure (M-Su), and fragment network (FN) searches are depicted in blue, orange, and green, respectively.

The three different types of fragment-based queries yielded varying numbers of follow-up compounds (**Fig. 6.5A**). The similarity search returned the highest number of compounds and showed little variation between fragment queries. In contrast, the substructure and fragment network searches were more variable and yielded the most extreme changes in the number of follow-up compounds returned between fragment queries. For instance, the fragment network search returned 0 and 7,703 follow-up compounds for fragments F503 and F616, respectively. The substructure search produced the fewest follow-up compounds, followed by the fragment network and similarity searches (**Fig. 6.5B**).

The fragment network compounds covered a larger portion of the chemical space compared to the substructure and similarity searches, which tended to cluster closer to the original fragment queries. Furthermore, the fragment network compounds had a lower heavy atom count than the substructure and similarity searches (**Fig. 6.5C**).

6.3.3. Binding pose prediction and filtering for fragment elaborations

X-ray crystallographic screening produced high-resolution fragment-bound structures, in Chapter 3, that can support structure-based inhibitor development (**Fig. 6.1**).[130] To this end, the follow-up compounds, stored as linear SMILES strings, must undergo conformer generation and protonation.[99] The protein (receptor) structure must also be processed to remove cofactors, such as ligand or solvent molecules, and hydrogen atoms must be added (**Fig. 6.6**).[192] Here, the bromodomain waters were removed as they were found to be displaceable, and the additional water was also eliminated as targeting this portion of the binding site may result in high-quality interactions (**Fig. 6.1**).

The follow-ups were then docked, using GNINA,[193] into PHIP(2)'s binding pocket which generated a variety of conformers each associated with a binding score (**Fig. 6.6**). This step is, however, agnostic of fragment crystal pose information and rescoring can used to select for overlapping docked poses[37] thus making use of both, protein and ligand crystal structure information (**Fig. 6.6**).

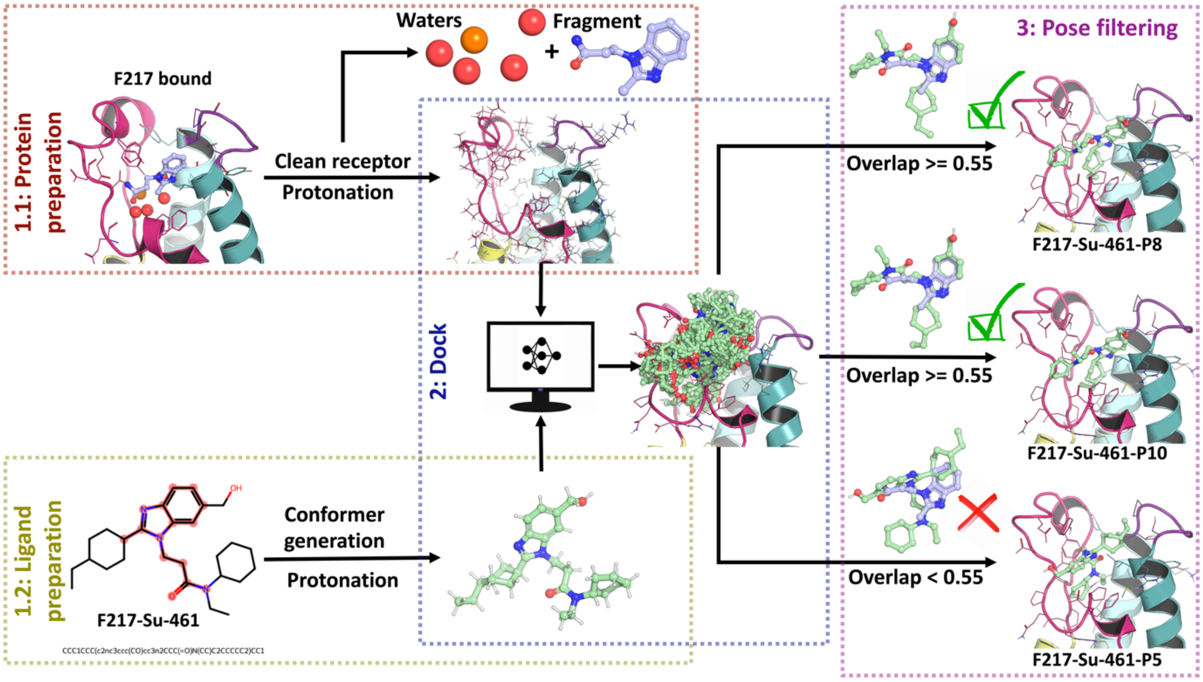


Figure 6.6: Docking workflow facilitates binding pose prediction, scoring, and filtering for follow-up compounds. First, the originating receptor protein structure is cleared of all waters and cofactors (1.1), and conformers are generated for follow-up compounds (1.2). Both receptor and follow-ups are protonated prior to binding pose prediction and scoring through docking (2). Resulting poses for each follow-up are filtered out if they don't overlap with the original fragment crystallographic pose (3).

Follow-up compounds were docked back into the co-crystal structure initially bound to their originating fragment, generating 100 poses per follow-up (**Fig. 6.6**). The docked poses did not always replicate experimentally resolved fragment poses, considered here as ground truth.

Consequently, SuCos[37] was employed to filter out docked poses that did not align with the crystallographic fragment pose. In brief, SuCos calculates volume and features overlap, such as hydrogen bond donor and acceptor, between docked and crystal poses; higher scores indicate more similar poses (**Fig. 6.6**).

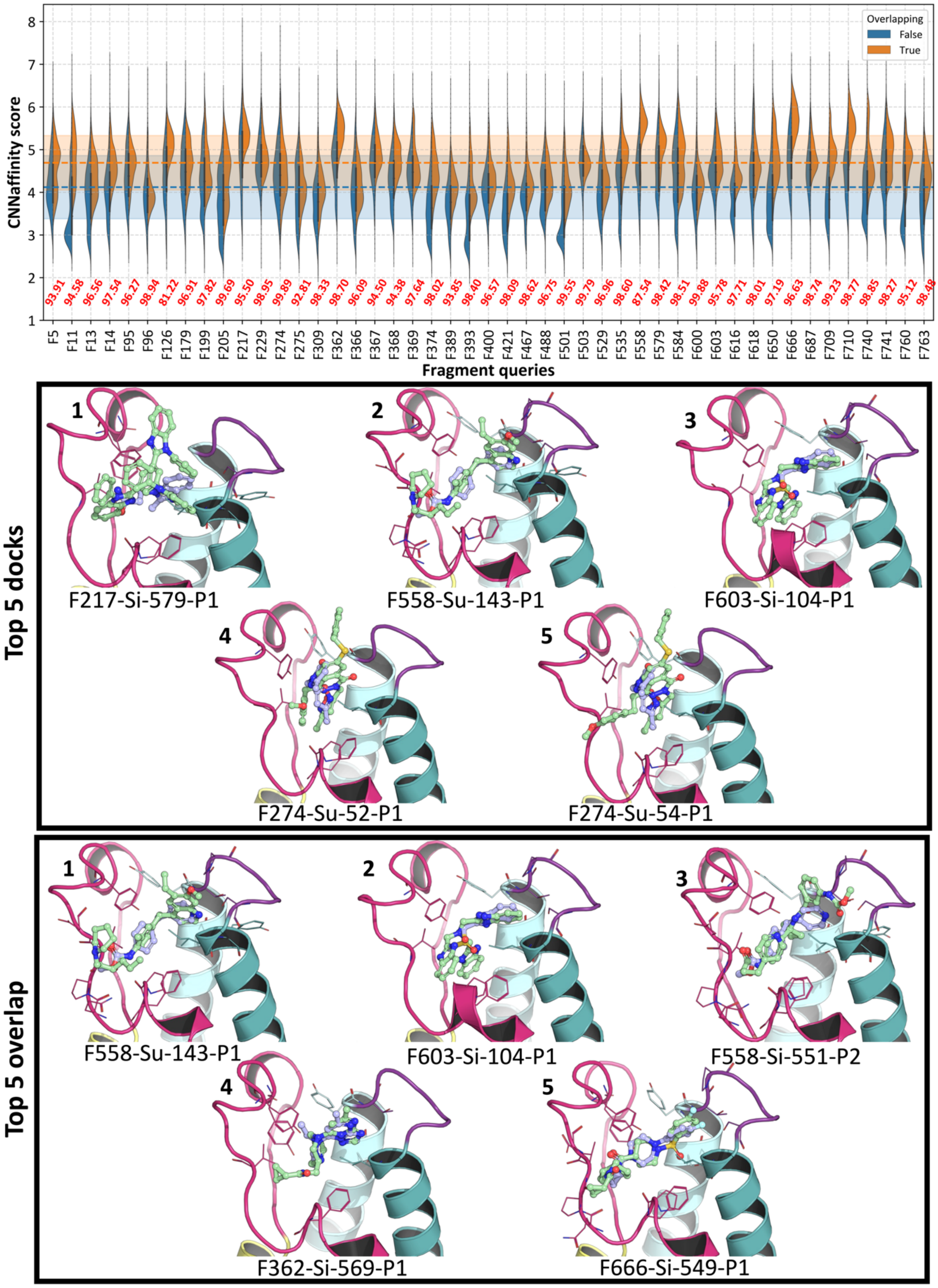


Figure 6.7: SuCos effectively eliminates unfavourable docked poses not overlapping with the crystallographic fragment. The top panel displays docking score distributions for all follow-up poses within fragment queries. The docking score distributions for overlapping and non-overlapping follow-up poses are shown in blue and orange, respectively. The percentage of poses removed by overlay filtering is indicated in red below the distributions for each fragment query. The top 5 scoring overlapping, and non-overlapping follow-up poses are presented in the middle and bottom panels, respectively. Light green and light blue ligands represent the crystallographic fragment and follow-up poses, respectively.

The docking procedure generated numerous poses with a score distribution for follow-ups derived from fragment-based queries (**Fig. 6.7**). Implementing SuCos for pose filtering eliminated the majority of docked poses, with removal rates varying from 81.22% to 99.89% for specific fragment queries and averaging 96.86% across all queries.

Overlapping poses tended to exhibit more favourable docking scores than non-overlapping poses. Therefore, rescoring procedure generally extracted the upper, more favourable, tails of the docking score distributions, with some variation between fragment queries (**Fig. 6.7**).

For instance, F217 follow-ups showed relatively good scores, while F393 follow-ups had poor scores in both overlapping and non-overlapping ensembles. In contrast, the rescoring procedure effectively selected the best-scoring docked poses for F11 follow-ups, while F96 and F503 follow-ups exhibited similar score distributions between the two ensembles.

It is important to note that top-scoring poses did not always recapitulate the crystallographic fragment poses (**Fig. 6.7**). For example, F217-Si-59, F603-Si-104, F274-Si-52, and F274-Su-54 comprised the unfiltered top 5 scoring poses, while only F558-Su-143 appeared in both the unfiltered and filtered top 5s (**Fig. 6.7**).

6.3.4. Free energy ranking of docked poses for follow-up selection

To better assess the energetics of the docking poses, MD simulation followed by free energy scoring by MMGBSA,[94] and Interaction Entropy[194] were applied (**Fig. 6.8**). This enables to account for protein and ligand dynamics and energetics of binding in the follow-up selection process.

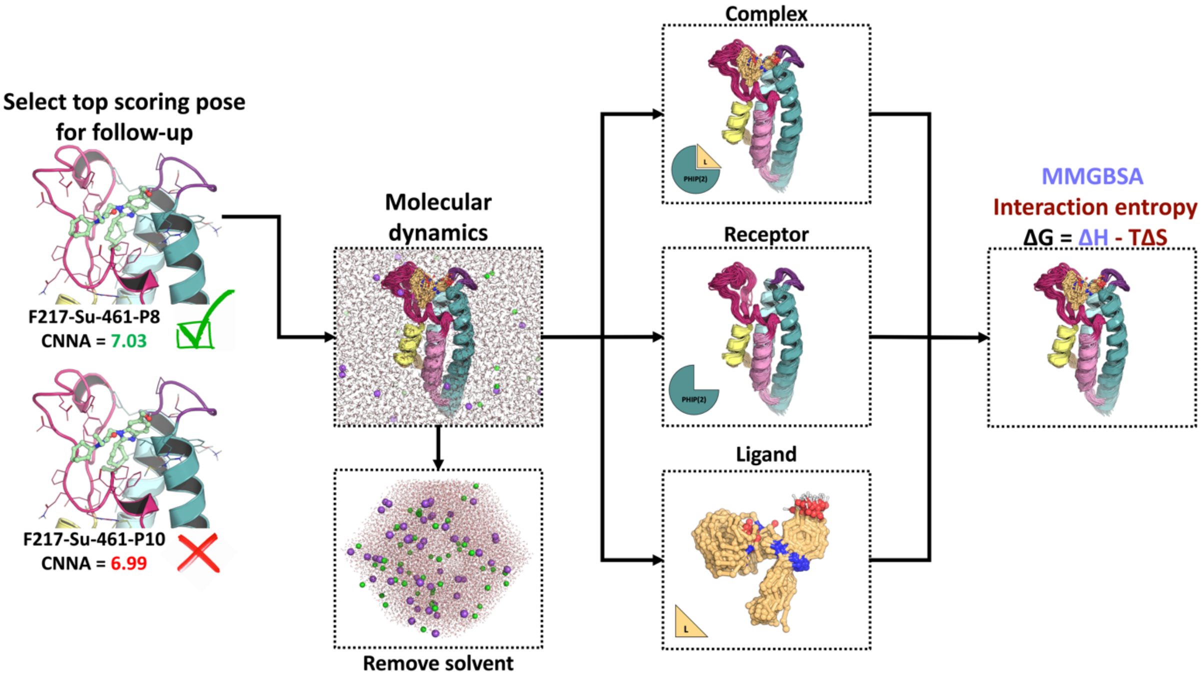


Figure 6.8: Best-scoring docking poses for follow-ups initiate molecular dynamic simulations enabling free energy scoring via MMGBSA and Interaction Entropy. To prevent redundant follow-ups, top-scoring poses are selected based on CNN affinity (CNNA). Follow-up bound structures seed molecular dynamic simulation trajectories in explicit solvent. Holo, apo, and ligand state trajectories are extracted from an initial MD trajectory and used for MMGBSA and Interaction Entropy calculations.

The best scoring pose for each individual follow-up was selected, and the top 15 best scoring follow-ups were chosen for each fragment, resulting in a total of 705 unique follow-ups selected for molecular dynamic simulation-based evaluation (**Fig. 6.8**). The coordinates of these 705 docked poses and receptors were used to initiate 15.2 nanosecond MD simulation trajectories in explicit solvent, resulting in a combined simulation time of 10.72 microseconds.

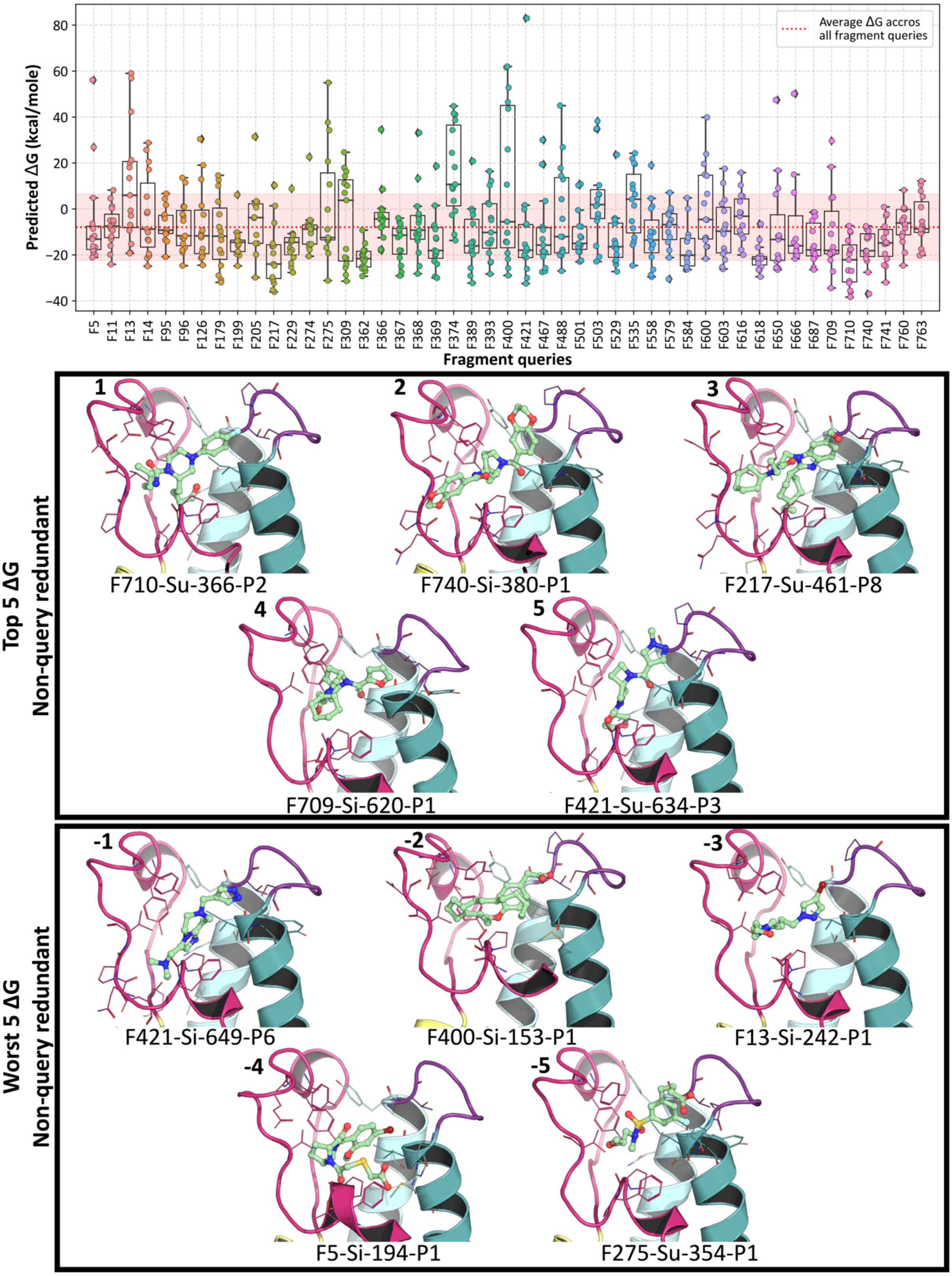


Figure 6.9: Free energy calculations uncover diverse chemotypes related to (un)favourable binding. The top panel displays average predicted Gibbs binding free energies for follow-ups within fragment queries, with the overall average shown by the dotted red line. The middle and lower panels present the top and worst non-query redundant follow-up poses, respectively. The docked follow-up poses are shown in pale green.

The average calculated binding free energy across all queries was -7.92 kcal/mol (**Fig. 6.9**). There was a considerable range in the predicted binding free energy distributions for the different fragment queries. Generally, local densities of similar predicted binding free energies within fragment queries correspond to chemically alike follow-ups adopting a comparable pose (**Fig. 6.9**).

The outliers also tended to be in regions with high positive binding free energies. Some queries exhibited a bimodal distribution, such as F309 and F367, which represent two dominating chemotypes within those follow-up selections. When examining the top five free energy scoring follow-ups, these tended to interact with the BC-loop and possessed large apolar groups either in the central or water cavity. Interactions at the ZA-channel were sampled in a smaller proportion (**Fig. 6.9**).

For follow-up poses predicted to be the least favourable, they all featured some degree of charged or highly polar groups (such as halogen atoms) and had more rotatable bonds than the top-scoring compounds. All follow-up compounds were subsequently ranked based on their predicted binding free energies (**Fig. 6.10**).

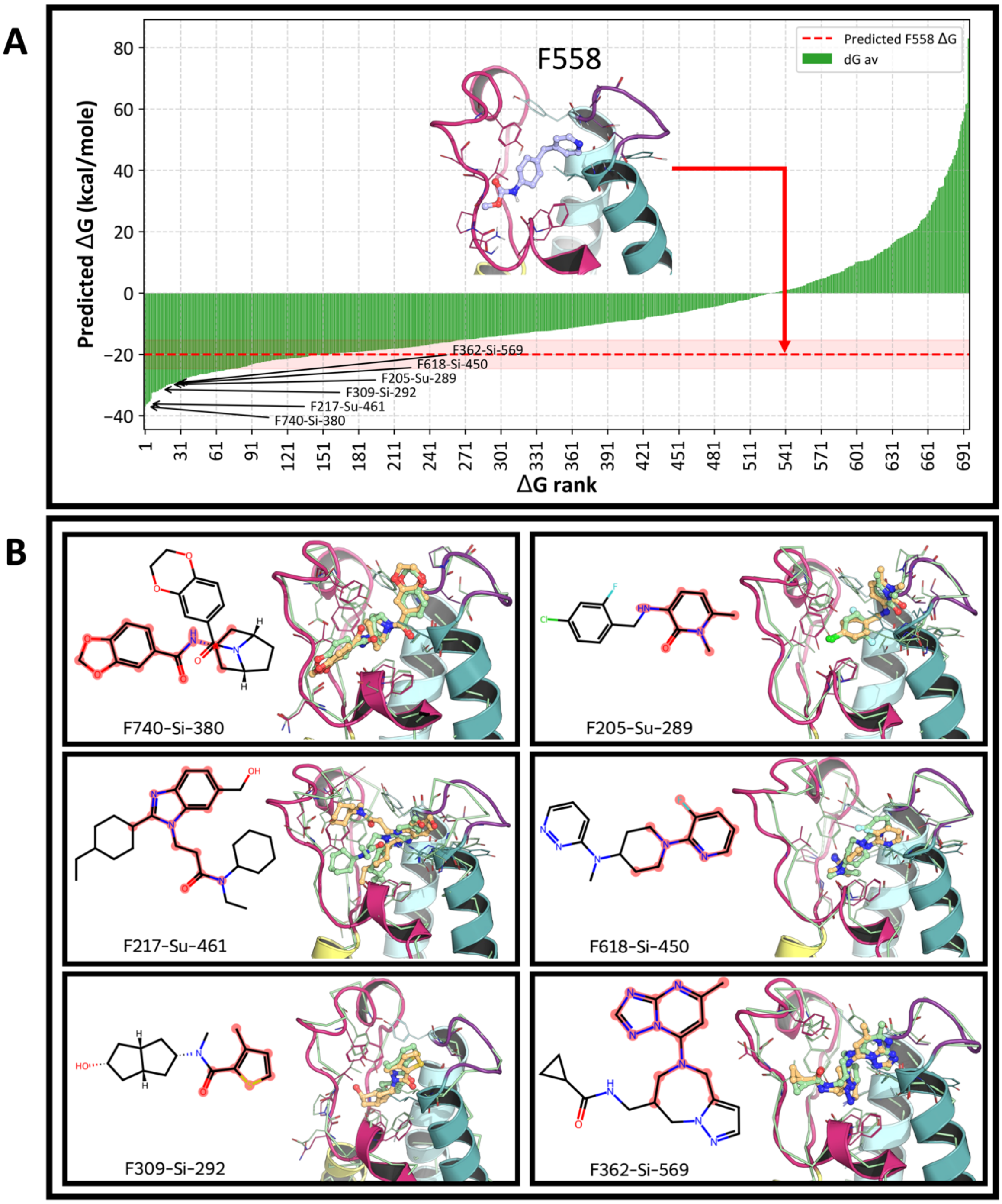


Figure 6.10: Free energy scoring ranks follow-up poses, enabling manual selection of top compounds for synthesis and experimental validation. The top panel (A) displays the ranked free energy profile accross all follow-ups, with the dotted red line indicating the predicted free energy value of the most favourable crystal fragment, F558. The bottom left of panel A shows the free energy ranking location of the follow-ups selected for validation using black arrows. The 2D chemical structures of selected follow-ups are displayed below (B), with the maximum common substructure relative to their originating fragment highlighted in pink. The 3D docked and final molecular dynamic simulation poses are shown in light green and beige, respectively, alongside initial and final protein conformations (extracted from MD trajectories) as pale green ribbons and coloured cartoons, respectively.

The analysis of the free energy profile for the follow-up compounds revealed that most poses were either similarly or less favourable than the most favourable fragment, F558 (**Fig. 6.2**). Specifically, out of the 705 selected follow-ups, 114 and 552 poses were predicted similarly or less favourable than F558 (**Fig. 6.10A**). In addition, 167 follow-up poses had a positive average predicted binding free energy implying that the bound state is estimated to be less stable than the apo state.

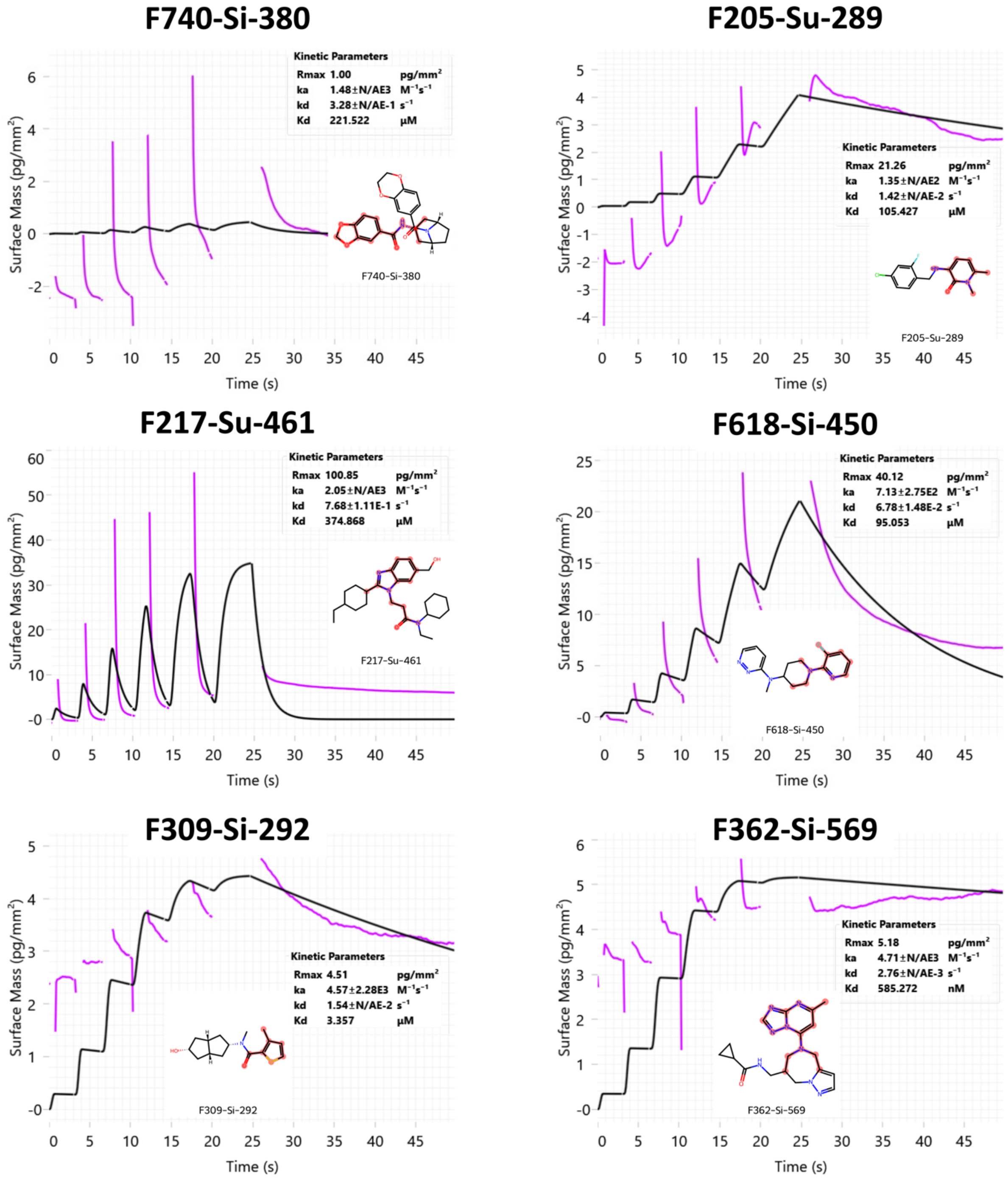
Only 30 compounds were predicted to be significantly more energetically favourable than F558 (**Suppl. table 6.1**). A team consisting of biochemists, synthetic, medicinal, and computational chemists collaborated in the selection process of the follow-up compounds. The team inspected all 30 poses and manually selected 6 for synthesis of them based on specific criteria, including pose stability (with respect to the initial fragment and docked pose), synthetic feasibility, and quality of the interaction formed with the binding site. Those 6 compounds were then synthetised by Thomas Grimes.

Overall, the docked poses for selected compounds were stable over the course of the MD trajectories (**Fig. 6.10B**) and maintained the initial fragment and docked pose including interactions with the binding site (**Fig. 6.1**). F217-Su-461 was the selected compound that experienced the most important conformational change with its aliphatic ring initially located at the ZA-channel moving opposite to the water cavity therefore inducing a relaxation of the ZA-loop. The F618-Si-450 system also experience conformational motions of the ZA-loop in the MD trajectory (**Fig. 6.10B**).

6.3.5. Validation identifies weak assay binders lacking crystal structure

The selected compounds (**Fig. 6.10B**) were synthesised by Thomas Grimes (**Suppl fig. 6.2**) and the final products dissolved in Dimethyl sulfoxide (DMSO) which is a polar aprotic solvent commonly used to dissolve organic molecules.[195] The selected follow-up compounds were tested for affinity against PHIP(2) using a grating-coupled interferometry (GCI) assay (**Table 6.1**).[137,165]

Table 6.1: Grating-coupled interferometry assay identifies two weakly binding follow-ups. The GCI sensograms for follow-up compounds are shown in purple, with fitted binding models in black. Kinetic values extracted from the models are displayed in the black box. The name and 2D structures of the tested follow-ups are presented alongside the sensograms. ka and kd are indicative of the association and dissociation rates with Kd, the dissociation constant, equalling to association over the dissociation rates.



The sensograms for follow-ups F740-Si-380, F205-Su-289, F309-Si-292, and F362-Si-569 did not provide conclusive results due to poor raw data. Therefore, the software was unable to fit the binding model and extract kinetic values. However, the last three follow-up compounds showed some weak signals of mass accumulating on the chip, which may be indicative of some level of binding. On the other hand, F618-Si-450 and F217-Su-461 exhibited weak binding signals with estimated K_d values of 374.87 and 95.06 µM, respectively (**Table 6.1**).

Structural validation is an essential step in iterative structure-based drug discovery as it enables to understand how a compound binds to the target and this information can be used in future rounds of optimisation. In addition, co-crystallisation can be more effective at resolving larger compounds that induce protein conformational motions.[27] Thus all 6 synthesised ligands were screened for co-crystallisation instead of the previously used crystal soaking protocol (**Fig. 6.11**).

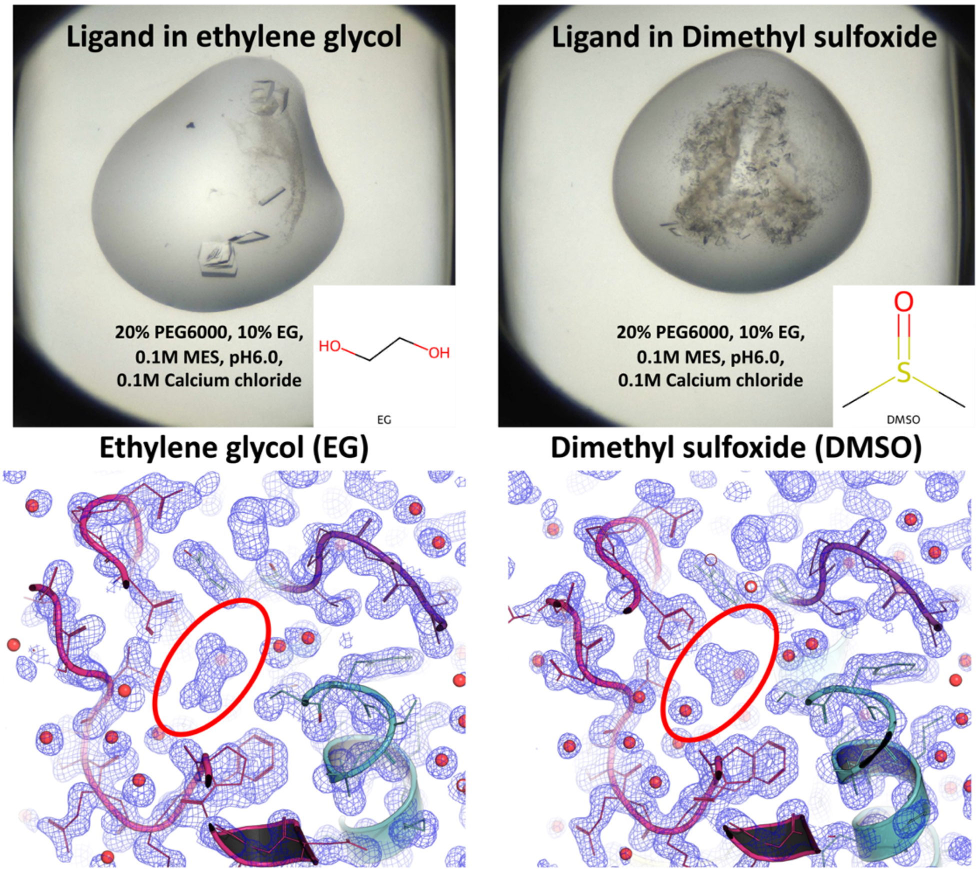


Figure 6.11: Dimethyl sulfoxide degrades PHIP(2) crystal morphology compared to ethylene glycol and binds to the acetylated lysine binding site. The top panel displays crystal drops with identical composition, co-crystallised with ligands dissolved in either ethylene glycol or dimethyl sulfoxide. The corresponding solvent 2D structures are shown in the bottom right. The bottom panel presents typical electron density for the PHIP(2) acetylated lysine binding site when bound to either ethylene glycol or dimethyl sulfoxide.

Unfortunately, no crystals suitable for x-ray diffraction experiments were obtained for F740-Si-380 and F205-Su-298. However, F309-Si-292 produced three crystals that yielded electron diffraction data, while F217-Su-461, F362-Si-569, and F618-Si-450 produced usable electron diffraction data from one, two, and two crystals, respectively (**Suppl. table 6.2**). Interestingly, all the usable crystals were formed under conditions that included some levels of ethylene glycol (**Suppl. table 6.2**). Furthermore, co-crystallisation with ligand in DMSO appeared to degrade crystal morphology compared to ethylene glycol (**Fig. 6.11**). Notably, electron density analysis showed the presence of DMSO in the binding site for each crystal.

6.4. Discussion

A virtual library of 59,507 follow-up compounds (**Fig. 6.5**), scored chemically accessible,[191] was assembled using the fragment network,[114] similarity, and substructure searches[25] using individual acetylated-lysing binding site fragments as seeds (**Fig. 6.4**). The fragment network produced the second-highest number of follow-ups, but with significantly lower heavy atom counts than the other methods (**Fig. 6.5**). Their fragment-like sizes resulting in low docking scores[196] implied none being selected for evaluation via MD-based free energy scoring (**Fig. 6.10B**). The number of follow-ups retrieved varied between fragment queries, implying that certain that scaffolds are better at seeding searches than others (**Fig. 6.5A**). However, this isn't always associated with chemical expansion at appropriate vectors.[197] For instance, F584, the most stable and water displacing fragment (**Fig. 6.2**),[198] retrieved follow-ups expanding at vectors causing clashes with the binding site or breaking its stabilising 7-membered lactam ring. Similarly, F616, the second most favourable fragment, had elaborations extending towards the solvent, leading to less favourable docking scores (**Fig. 6.7**).

Overall, the elaborations tended not to capture the variety of available interactions within the binding site (**Fig. 6.1**), suggesting single-fragment queries may limit interactions recovery from collective crystallographic fragment hits. Indeed, fragment growing (**Fig. 6.4**), employed here, only considers one scaffold during elaboration therefore reducing the use of experimental information compared to linking or merging.[199] 5 follow-ups were synthesised and 1 ordered from catalogues (**Suppl fig. 6.2**). In some cases, their synthesis was non-trivial and required multiple steps, highlighting that synthetic accessibility scores can be misleading (**Suppl fig. 6.2**).

The follow-ups were docked in their originating co-crystal structures, generating various poses (**Fig. 6.6**). These poses were then selected based on their overlay with the crystallographic fragment binding pose (**Fig. 6.6**), therefore making use of experimental information[37] before more computationally expansive MD-based free energy calculations.[94] The initial filtering step successfully extracted more favourable poses based on docking scores for fragment queries (**Fig. 6.7**). The workflow ensured equal numbers of follow-ups for each fragment query in the final selection for evaluation (**Fig. 6.8**). However, some fragment queries, such as F96 or F393, scored worse on average, indicating relatively unfavourable docked poses (**Fig. 6.7**). Follow-ups from these queries never achieved more favourable binding free energies than the best fragment, F558.

Large errors in, or unfavourable predicted free energies were observed for F96 and F393 queries (**Fig. 6.9**). Thus, considering all fragment queries equally before more intensive calculations may detrimentally impact the overall outcome, as computational resources are allocated to undesirable poses. Additionally, 36.64% of follow-ups selected for MD simulations had a predicted binding free energy significantly greater than 0 (**Fig. 6.10A**), caused by large entropic penalties, suggesting unfavourable poses. In general, such cases had many rotatable bonds and, to a lesser extent, highly polar groups (**Fig. 6.9**). While the Interaction Entropy method filters out most flexible compounds, it also unnecessarily penalised charged compounds[194] potentially forming favourable interactions (**Suppl Fig. 6.1**), an issue also observed for some fragments (**Fig. 6.3**).

Six follow-up compounds were chosen for synthesis and experimental validation (**Fig. 6.10B**). Interestingly, these follow-ups recovered features from other fragments. For example, F740-Si-380 and F618-Si-450 had aliphatic rings in the central binding cavity similar to F709 and F362-Si-569 had a 7-membered ring like water-displacing fragment F584 (**Fig. 6.10**). This implies that those top scoring follow-ups are coincidental fragment merges.

Some follow-ups were highly hydrophobic, necessitating DMSO over ethylene glycol as solvent.[195] This negatively affected validation crystallographic efforts, degrading crystal morphology and binding to the target site with higher affinity than ethylene glycol (**Fig. 6.11**).[200] Only F217-Su-461 and F618-Si-450 had low and medium micromolar binding affinities, respectively (**Table 6.1**). The use of DMSO was however, also found to interfere with bromodomain biosensor readouts.[201]

Comprehending the target binding site is vital for rationalising ligand binding. Key characteristics in protein-ligand interactions include water stabilities ,[121] pharmacophoric features ,[202] and protein motions .[203] All waters were removed here (**Fig. 6.6**), even the one deeply buried in the water cavity (**Fig. 6.1**). The energetics of this previously undisplaced water is unknown, and is potentially being highly stable and therefore undisplaceable.[124] F217-Su-461 and F309-Si-292 were initially docked at this location but later in the trajectory moved, allowing the buried water back (**Fig. 6.10B**). Retaining this water during docking could alter the docking free energy landscape and therefore, alter follow-up selection.[204]

Polar interactions were located around more flexible loop regions, and identified by docking and maintained in MD simulations (**Fig. 6.1**). However, trajectories might not be long enough to sample representative loop conformations in the solution state, possibly differing from crystal and short simulations.[91] This effect could be relevant as selected follow-ups originated from BC-loop fragment binding (**Fig. 6.10B**).[205] Targeting flexible regions may reduce, or even prohibit, follow-up binding due to an high binding entropic barrier associated with the loss of protein conformational motions.[206] Intriguingly, the two follow-up GCI binding events were paired with ZA-loop relaxation compared to original crystal conformation (**Fig. 6.10B**). This loop conformation may better represent global and energetically favourable binding site minima. However, no structural information was recovered to validate this.

A trade-off exists between running longer simulations for better system characterisation and increased MMGBSA errors, sometimes hindering predictions.[207] Optimising the dielectric constant for a specific binding site or ligand set may improve performance, but the absence of known actives prohibited prior optimisation.[207] Interaction Entropy correction's inclusion, although computationally inexpensive, can also adversely affect some predictions and requires prior knowledge of ligand binding for benchmarking purposes.[208] The virtual screening here was guided by a prior study of bromodomain inhibitors, which informed the parameters used, including simulation time, dielectric constant value, and Interaction Entropy correction.[209]

6.5. Conclusion and future direction

This study aimed to enhance the potency of crystallographic fragments against the cancer target PHIP(2)[210] using computational methods. Insights into binding site interactions and fragment stabilities were acquired through hotspot mapping (**Fig. 6.1**),[29] MD simulations (**Fig. 6.2**), and free energy calculations[94] of co-crystal structures (**Fig. 6.3**).

The initial library building, using Fragment Network, similarity, and substructure searches, resulted in 77,378 follow-up compounds, later reduced to 59,507 through chemical accessibility scoring (**Fig. 6.4**). The Fragment Network exhibited the most variation in compounds returned, all small in size (**Fig. 6.5A**).[114] These follow-ups were docked[193] into co-crystals, and SuCos overlay[37] was employed to filter 96.86% of poses (**Fig. 6.6**). This approach extracted poses with better docking scores and overlap with the experimentally resolved pose (**Fig. 6.7**).

The top 15 scoring non-redundant follow-up docks per fragment query, amounting to 705 compounds, were chosen for molecular dynamic simulations and free energy calculations using MMGBSA[190] and Interaction Entropy (**Fig. 6.8**).[194] A total of 30 compounds demonstrated more favourable predicted binding affinities than the most favourable crystallographic fragment, F558 (**Fig. 6.11** and **Suppl. table 6.1**). Out of these, six were chosen for synthesis (**Fig. 6.10**).

Some compounds required complex multistep chemistry for material generation. Experimental validation using crystallography (**Suppl. table 6.2**) and GCI assay[137] (**Table 6.1**) proved challenging; no crystallographic evidence for binding was found (**Fig. 6.11**). However, two weak binding events out of six tested compounds were identified via GCI assay (**Table 6.1**), both associated with loop motions in MD trajectories (**Fig. 6.10B**), potentially indicating the significance of accounting for protein motions.

Several areas for potential improvement were highlighted. These include focusing on fragment selection by opting for fragments with suitable and chemically accessible elaboration vectors,[197] stable binding poses, and energetics to maximise efficient computational resource usage. The elaboration process could be enhanced by conducting additional iteration cycles of the Fragment Network to obtain larger compounds, or by aiming for direct fragment merges or linkers to optimise the use of crystallographic information.[211]

It is also crucial to consider certain physico-chemical characteristics, such as rotatable bonds or lipophilicity, in advance of expansive calculations and experimental validation. Docking and rescoring optimisation could involve implementing ensemble docking against MD-generated structures of fragment-bound conformations to account for conformational motions.[203] Including water molecules during binding pose predictions may also help improving the results.[209] However, selecting appropriate waters before docking is not a straight forward problem.

Applying diversity selection of best poses across all fragment queries to improve coverage of the chemical landscape may be more judicious than selecting a constant number of compounds per fragment query. Increasing MD and MMGBSA sampling by extending simulations to eliminate unstable poses and performing free energy scoring only for stable ones also seems advisable. More rigorous alchemical binding free energy calculations may also improve the predictions,[209] but their use was impractical, here, due to limited access to computational resources. Understanding binding site conformational motions may enable the reprioritisation of interaction hotspots compared to the crystal structure. Finally, addressing the experimental validation challenges entails investigating ways to counteract the detrimental effects of DMSO on the target, repeating and optimising experimental validation steps.

This attempt was a prospective endeavour, as no prior information about active compounds related to the C2 crystal form was available. This hindered optimisation and calibration of computational methods, illustrating the challenges of identifying potent compounds without assay actives. Overall, this study highlights the complexity of early-stage drug discovery, emphasising the need to account for appropriate vector elaborations, ligand physico-chemical characteristics, binding site interactions, and conformational dynamics while considering target-specific experimental validation constraints.

6.6. Methods

6.6.1. Hotspot mapping

Ensemble hotspot maps[29] were generated using all 47 acetylated-lysine binding site fragment bound structures after NPT minimisation (see section 6.6.2). Protein hydrogens were removed prior to calculations. Individual maps were first generated using with standard donor, acceptor, apolar, positive, and negative probes with 6000 rotations each for sampling step. For ensemble map calculation, the apolar and polar frequency threshold were set to None and 10, respectively and maps were combined based on median values.

6.6.2. Molecular dynamic simulations

Crystallographically resolved conformers and water molecules were extracted from PDB files with MDAnalaysis.[189] Fragment and follow-up conformer protonation states were assigned with Quacpac Toolkit (OpenEye).[99] For all fragment-bound co-crystal structures used, missing protein termini were added with Modeller[212] assuming a static conformation for atoms present in the initial PDB structure. Missing heavy atoms and hydrogens were added with PDB2PQR[213] using PARSE as force field, PROPKA[192] for protonation an assuming physiological pH of 7.4.

Ligands were parametrised with GAFF2[214] using AM1-BCC charges.[215] Ligand parameters were converted to Gromacs format with ParmEd.[216] Protein residues were parametrised with the AMBER99SB-ILDN force field.[89] All protein-ligand complexes were solvated with TIP3P waters in dodecahedral box with periodic boundary conditions. The minimum distance between the edge and the closest atom of the complex was set to 12 Å. Water molecules were replaced with Na^+ and Cl^- ions to reach a concentration of 150 mM and a neutral charge with the simulation box. The systems were energy minimised using the steepest descent method with target energy of 1000 kJ/mol/nm. 500 ps equilibrations in the canonical (NVT) ensemble were achieved at 298.15 K using a Berendsen thermostat[217] with temperature coupling within protein and non-protein atoms. 1 ns NPT equilibration were carried out at 1 bar with a Berendsen barostat.

For all system, 15.2 ns production runs were started upon removal of the position restraints on the heavy atoms of the proteins and ligand. Van der Waals and short-range electrostatic interactions distance cut-off were set to 1.2 Å while long-range electrostatics were computed with the particle mesh Ewald method.[218] Hydrogens to heavy atom bonds were constrained with the LINCS allowing a 2 fs time step using.[219] The stochastic leap-frog integrator was used for numerical integration with periodic boundary conditions and Parrinello-Rahman barostat. Trajectory frames were saved every 2 ps. To allow for further processing steps, the protein was centred in the protein in the simulation box, rotation and translation of the protein C-alpha atoms removed, and the first frame of the aligned trajectory extracted. All minimisation, trajectories and fittings were generated using Gromacs 2020.[220]

6.6.3. RMSD calculations

Ligand RMSD calculation for heavy atoms compared the first frame of the production run against all subsequent frames with MDanalysis[189] without additional centring or superposition. The median RMSD was then reported.

6.6.4. Free energy calculations

MMGBSA and Interaction Entropy for fragment and follow-up trajectories were calculated with the gmx_MMPBSA suite.[190] The whole trajectories were process with an interval of 30 frames corresponding to 60 ps of simulation timesteps for both MMGBSA and Interaction Entropy. The OBC1 model was used for Generalised Born estimations along with an implicit salt concentration of 150 mM. The values reported were the average MMGBSA energies and interaction entropies along with standard deviations. Binding enthalpy and entropy were estimated from MMGBSA and Interaction Entropy, respectively and summed thus providing an estimate of Gibbs binding free energy.

Gibbs free energy estimation

$$\Delta G_{est} = \Delta H_{mmgbsa} - T\Delta S_{IE}$$

Equation 6.1

In order to estimate for errors around the average Gibbs binding free energy values calculated, both binding enthalpy and entropy standard deviations were also propagated using the root-sum-square method.

Gibbs binding free energy standard deviation

$$\Delta G_{std} = \sqrt[2]{\Delta H_{std}^{2} + \Delta S_{std}^{2}}$$

Equation 6.2

Where ΔH_{std} and ΔS_{std} correspond to binding enthalpy and entropy standard deviations calculated from MMBSA and Interaction Entropy analysis of the MD trajectories, respectively.

6.6.5. Library construction

The library was built using the Fragment network and Manifold similarity and substructure searches using fragment hit SMILES strings to seed the searches. Fragment network follow-up compounds were enumerated via the graph-network made available through the fragalysis API.[221] The Manifold similarity and substructure searches were performed by requesting corresponding services to https://postera.ai/ via HTTP posting.[222] On occasion, multiple molecules were returned within one SMILES string. Only the larger molecules were retained for those cases. Each follow-up was then scored for chemical accessibility by querying to Postera for Synthetic Accessibility / Fast Score and retained for further processing if their Fast Score was none or low. All queries to Postera were made around the 13st of June 2021.

6.6.6. Follow-up fingerprinting and dimensionality reduction

Feature-based invariant Morgan fingerprints were generated using a length of 2048 and a radius of 6 with RDKit and used as input for and t-Distributed Stochastic Neighbourhood Embedding using scikit-learn.[80] The fingerprint bit vectors were embedded in 2 dimensions allowing for visualisation using the Jaccard distance metric with a perplexity of 50.

6.6.7. Docking and pose filtering

Follow-up conformers were generated with RDKit.[128] First, all hydrogens were added before attempting a maximum of 15.000 embeddings with enforced chirality. For unsuccessful embeddings, minimisation of the hydrogenated molecule was performed instead using the MMFF94 force field.[223] Follow-up conformer protonation states were then assigned with Quacpac Toolkit (OpenEye).[99]

Protonated follow-up conformers were docked with GNINA[84] back into their original fragment-bound co-crystal structures refined from NPT minimisation (see section 6.6.2). A docking search box with dimensions of 18, 19, and 16 for X, Y, and Z, respectively, was defined, encompassing the entire binding site. For each follow-up compound, 100 conformers were generated using 200 Monte Carlo steps, an RMSD filter of 0.25 Å, and an exhaustiveness of 118. Resulting poses were excluded if their SuCos[37] score against the experimentally resolved fragment was less than or equal to 0.55. The top 15 non-redundant poses for each follow-up were then chosen for further analysis using molecular dynamic simulations.

6.6.8. Co-crystallisation screening

Proteins were expressed and purified as described in Chapter 2, Section 2.2.1. The 6 selected follow-up compounds were kindly synthesised or purchased by Thomas Grimes, under the supervision of Prof Paul Brennan, and stored in dimethyl sulfoxide at a concentration of 40 mM. Co-crystallisation condition screening for each of the follow-up compounds were then conducted. First, protein aliquots were incubated with the dissolved compound at a final concentration of 199 µM on ice for 1 hour.

Subsequently, vapor diffusion experiments were set up using sitting drop plates and a Ligand Friendly Screen (Molecular Dimensions)[26] for each compound. Drops of 100 nL of mother liquor were combined with 100, 75, or 50 nL of protein solution and 30 nL of seed stock derived from the C2 crystal form.[198] The plates were incubated at 4°C for one week prior to crystal harvesting.

All harvested crystals were subjected to X-ray diffraction analysis at the i04-1 beamline located at the Diamond Light Source (Harwell, UK). Suitable MTZ files were extracted from ISPyB[224] and then combined and truncated with AIMLESS.[225] Molecular replacement was performed using PHASER-MR[226] with the crystal structure of 7AV9 as the search model and a sequence conservation value of 0.7.

6.6.9. Grating-coupled interferometry assay

Pulsed single-concentration surface-based biophysical measurements of binding kinetics were performed as detailed in Chatper 2, section 2.2.3. Specifically, PHIP(2) was diluted to 15 µg/mL and immobilised to a final level of 15951 surface mass (pg/mm^2) corresponding to an injection time of 600 s. The running buffer was composed of 20mM HEPES, pH 7.5, 50 mM NaCl, 0.5% (v/v) Tween-20, and 0.5% (v/v) DMSO.

Chapter 7. Lessons and outlooks

7.1. Lessons

This work examined various aspects of fragment-based drug discovery, encompassing both experimental and computational methods for an integrative approach. The study focused on fragments identified crystallographically against the pharmacologically relevant Second Bromodomain of the Pleckstrin-Homology Domain Interacting Protein (PHIP(2)). An initial XChem crystallographic fragment screening, presented in Chapter 3, laid the groundwork for this research. Particular attention was given to fragment F709 and its follow-up derivatives in Chapters 4 and 5, while all acetylated-lysine binding site fragments were examined in Chapters 3 and 6.

Chapter 3 assessed the capability of computational methods to recover fragment binding and crystallographic poses through a community-led exercise, as well as evaluating participants' ability to enumerate sensible follow-ups. Overall, performances across all three stages were modest, likely due to a variety of factors. For instance, predictions not applied to the provided apo structure in the C2 space group fared poorly for binding pose predictions. Other factors contributing to poor predictions may be more difficult to untangle, such as a potential lack of affinity of the fragments against PHIP(2) in solution.

These findings underscore the challenges in applying computational methods to replicate crystallographic screening of small molecule fragments. The study therefore implies that concentrating computational resources on fragment predictions might be suboptimal, recommending instead that efforts focus on designing follow-up compounds to "escape" the low potency fragment space, a task that also proved to be difficult. The poor predictions also highlight that experimental approaches remain required to reliably identify fragments against a protein target.

Chapter 4 organised the data obtained from a large crystallographic screening of robotically generated crude reaction mixtures, carried out by collaborators. This chapter's primary feature was the multidisciplinary, collaborative, and extensive nature of the project, which necessitated a comprehensive understanding of all methods involved to collate the data productively. In total, nearly 2,000 follow-up compounds were enumerated, with over half successfully synthesised robotically. Soaking crude reaction mixtures without causing significant crystal deterioration allowed for the acquisition of nearly 1000 diffraction datasets. Crystallographic analysis revealed 22 reaction product binders, with three exhibiting a non-conserved pose due to methyl insertion and one showing affinity in assays.

The findings demonstrated how high-throughput chemistry can be coupled with X-ray crystallographic screening of crude reaction mixtures, bypassing small molecule purification, thus reducing time, solvent, and hardware requirements. Therefore, this showcases how follow-up compounds can be synthetised and validated in faster, cheaper, and more environmentally friendly fashion. The study also highlighted the need for improved design and analysis methods to increase hit rates and extract meaningful information from large crystallographic dataset. Finally, this chapter also indicates that numerous compounds and experiments are needed to explore synthetically accessible chemical landscapes and identify binders and binding pose changes.

Chapter 5 developed an algorithmic approach aimed at extracting features associated with crystallographic binding or the lack thereof, using the data generated in Chapter 4. Binding and non-binding features were initially extracted, then aggregated into positive and negative binding scores. This method further emphasised the importance of methyl addition in altering the binding pose. Crystallographic rescreening of selected follow-ups in pure form revealed 26 false negatives, effectively doubling the initial hit rate, with respect with the experiment performed in Chapter 4, where the positive binding score significantly enriched the identification of binders over non-binders. The positive and negative binding scores were then employed prospectively in a virtual screening exercise, incorporating both ligand- and structure-based information. This revealed a collection of crystallographic binders displaying novel chemical groups, vectors, and poses. A series of assay binders with low micromolar potencies were also identified.

Overall, this work demonstrated how features relevant to binding or the lack thereof can be extracted to enhance the understanding of large crystallographic screening data for congeneric follow-up elaboration series. Crucially, these features can be exploited in subsequent designs to integrate site- and elaboration-specific binding and non-binding information.

In Chapter 6, more traditional and computationally intensive methods were combined to develop the acetylated-lysine binding site fragments into more potent follow-up compounds via a structure-based virtual screening. An initial virtual library of follow-up compounds deemed chemically accessible was generated, using the original fragments to seed database searches. Library follow-up compounds were subsequently docked back into their originating co-crystal structures. Docked poses not overlapping with their native crystallographic fragment conformations were removed, leading to the selection of a relatively small number of high-scoring follow-up poses. Molecular dynamic simulation trajectories were generated for the best-scoring non-redundant poses, from which binding free energy rankings were performed using Molecular Mechanics Generalised Born Surface Area with Interaction Entropy methods. Six follow-ups from the top 30 were chosen for experimental validation. X-ray crystallography did not resolve any binders, while grating-coupled interferometry kinetic assay identified two weak binding events. Both weak binding events were associated with binding site motions in molecular dynamic simulations.

Overall, this work emphasised the complexity of elaborating fragments to sub-micromolar potencies and highlights how existing methodologies can be limiting. Thus, better method integration is required for more efficient design. This work underscored the importance of library construction and efficient allocation of computational resources to worthwhile series. Additionally, it implied the importance of establishing early and cheap structure-activity-relationship for follow-up series and the potential significance of accounting for protein and water dynamics, while the absence of active data was prohibitive in benchmarking computational methods.

7.2. Outlooks

Numerous lessons and perspectives can be drawn from the work undertaken in this thesis. Effective virtual library design and compound selection are crucial prior to utilising more extensive structure-based approaches. Furthermore, fragment growing may be chemically more tractable, but suboptimal compared to linking and merging strategies. Recent machine learning-based generative approaches holds promising potential for initial follow-up enumeration. These generative algorithms could be constrained to chemically or robotically accessible chemical space for reduced lead time whilst intensive physics-based methods could be used for more robust compounds selection.

Although some challenges may be specific to PHIP(2), lessons learned here may apply to other protein families. Protein and ligand dynamics are integral aspects of their interactions, but often neglected in high-throughput efforts. Consequently, successful elaborations depend on the target binding site representing a low-energy conformer in solution, which may differ from crystal structures. This is particularly relevant in areas with flexibility that might be artificially stabilised by crystal packing effects. Therefore, comparing experimental temperature factors with binding site motions extracted from molecular dynamic simulations may inform on whether ligand binding is artifactually stabilised by lattice effects.

Molecular dynamics simulations may also be useful in identifying low-energy protein conformations from crystal structures, and therefore help choosing crystal systems when different options are available. Before conducting crystallographic screening, space group selection criteria could be adopted to identify a conformation resembling the low-energy conformer generated through molecular dynamics simulations. However, practical considerations such as crystal reproducibility, solvent resistance, and binding site resolution may remain crucial decision drivers.

Crystallographic and assay structure-activity relationships do not necessarily corroborate. Therefore, ligands resolved in crystal structures may not bind in assays and vice versa. This discrepancy might also result from differences in crystal and solution conformational state(s). Therefore, establishing an early, tractable structure-activity relationship for elaboration series that can be rationalised structurally is essential. Selecting elaboration series should also be guided by the ability of the fragment to be modified into compounds that bind in both crystals and assays.

The acquisition of negative data may be informative for the elaboration process, but such information tends to be less accessible due to factors such as initial misclassification or the impossibility to obtain structural data for these compounds. One approach to consider may be to weight chemical fingerprint bit elements to account for the frequency at which they were observed in non-binding events thus refining the initial approach outlined in Chapter 5.

Crystallographic analysis of robotically generated crude reaction mixtures soaked onto protein crystals can rapidly and inexpensively map the binding landscape, demonstrating potential for informing early structure-activity relationships across multiple compounds. This process could benefit from better compound design and diversity to enhance the chances of identifying more potent elaborations. A major limitation in rapidly acquiring structure-activity relationships data is that assays are performed using pure compounds, meaning robotic compounds must be independently resynthesised or purified. This can be prohibitive for reasons related to time, reagent, hardware, or finances. Using crude reaction mixtures might be applicable to high-throughput dissociation rate screening with grating-coupled interferometry. However, the resulting data is expected to be highly noisy, and control measurements of pure starting materials will be required to deconvolute the signal. Molecular dynamics simulations may also supplement the analysis by simulating ligand residence time within the binding site, which can then be compared with experimentally measured dissociation rates, potentially enabling the removal of outliers thus generating a cleaner dataset.

PHIP and its second bromodomain, PHIP(2), remain undrugged targets despite significant efforts. Inhibiting this bromodomain may offer a potential avenue for treating lethal cancers. The design of proteolysis targeting chimeras (PROTAC) may represent a suitable alternative, where lower potency but specific binders can be used. In the context of PHIP overexpression in cancerous cells, this strategy may be particularly interesting. Multidomain targeting may also be a viable alternative, potentially achieved by dual inhibition of PHIP(1) and PHIP(2) with separate warheads against each bromodomain, connected by a linker region. Additionally, the identification of potent compounds may help uncover PHIP's elusive biology via chemical biology methods.

In summary, this research showcases the application, integration, and limitations of contemporary fragment-based drug discovery techniques. Key lessons and insights related to fragment identification, follow-up design, synthesis, and validation have been emphasised. Moreover, the insights and materials produced in this research may contribute to future structure- and fragment-based efforts aimed to develop a novel anti-PHIP(2) inhibitor with potential applications in treating lethal cancers.

Literature and references

[1] F. D. Makurvet, *Medicine in Drug Discovery* **2021**, *9*, 100075.

[2] T. S. Maurer, M. Edwards, D. Hepworth, P. Verhoest, C. M. N. Allerton, *Drug Discovery Today* **2022**, *27*, 538–546.

[3] J. Hughes, S. Rees, S. Kalindjian, K. Philpott, *British Journal of Pharmacology* **2011**, *162*, 1239–1249.

[4] A. Rezabakhsh, A. Mahmoodpoor, H. Soleimanpour, *J Cardiovasc Thorac Res* **2021**, *13*, 179–180.

[5] A. S. Pina, A. Hussain, A. C. A. Roque, in *Ligand-Macromolecular Interactions in Drug Discovery* (Ed.: A.C.A. Roque), Humana Press, Totowa, NJ, **2010**, pp. 3–12.

[6] L. M. Mayr, D. Bojanic, *Current Opinion in Pharmacology* **2009**, *9*, 580–588.

[7] K. L. M. Drew, H. Baiman, P. Khwaounjoo, B. Yu, J. Reynisson, *Journal of Pharmacy and Pharmacology* **2012**, *64*, 490–495.

[8] C. A. Lipinski, *Drug Discovery Today: Technologies* **2004**, *1*, 337–341.

[9] C. W. Murray, D. C. Rees, *Nature Chemistry* **2009**, *1*, 187–192.

[10] G. G. Ferenczy, G. M. Keserű, *Med. Chem. Commun.* **2016**, *7*, 332–337.

[11] A. Carbery, R. Skyner, F. Von Delft, C. M. Deane, *J. Med. Chem.* **2022**, *65*, 11404–11413.

[12] D. J. Wood, J. D. Lopez-Fernandez, L. E. Knight, I. Al-Khawaldeh, C. Gai, S. Lin, M. P. Martin, D. C. Miller, C. Cano, J. A. Endicott, I. R. Hardcastle, M. E. M. Noble, M. J. Waring, *J. Med. Chem.* **2019**, *62*, 3741–3752.

[13] O. B. Cox, T. Krojer, P. Collins, O. Monteiro, R. Talon, A. Bradley, O. Fedorov, J. Amin, B. D. Marsden, J. Spencer, F. von Delft, P. E. Brennan, *Chem. Sci.* **2016**, *7*, 2322–2330.

[14] A. L. Carvalho, J. Trincão, M. J. Romão, in *Ligand-Macromolecular Interactions in Drug Discovery* (Ed.: A.C.A. Roque), Humana Press, Totowa, NJ, **2010**, pp. 31–56.

[15] J. C. Kendrew, G. Bodo, H. M. Dintzis, R. G. Parrish, H. Wyckoff, D. C. Phillips, *Nature* **1958**, *181*, 662–666.

[16] M. F. Perutz, Giulio. Fermi, D. J. Abraham, Claude. Poyart, E. Bursaux, *J. Am. Chem. Soc.* **1986**, *108*, 1064–1078.

[17] J. Blaney, *J Comput Aided Mol Des* **2012**, *26*, 13–14.

[18] A. Douangamath, A. Powell, D. Fearon, P. M. Collins, R. Talon, T. Krojer, R. Skyner, J. Brandao-Neto, L. Dunnett, A. Dias, A. Aimon, N. M. Pearce, C. Wild, T. Gorrie-Stone, F. von Delft, *JoVE* **2021**, 62414.

[19] J. W. Kaminski, L. Vera, D. P. Stegmann, J. Vering, D. Eris, K. M. L. Smith, C.-Y. Huang, N. Meier, J. Steuber, M. Wang, G. Fritz, J. A. Wojdyla, M. E. Sharpe, *Acta Crystallogr D Struct Biol* **2022**, *78*, 328–336.

[20] J. T. Ng, C. Dekker, M. Kroemer, M. Osborne, F. von Delft, *Acta Crystallogr D Biol Crystallogr* **2014**, *70*, 2702–2718.

[21] P. M. Collins, J. T. Ng, R. Talon, K. Nekrosiute, T. Krojer, A. Douangamath, J. Brandao-Neto, N. Wright, N. M. Pearce, F. von Delft, *Acta Crystallogr D Struct Biol* **2017**, *73*, 246–255.

[22] N. D. Wright, P. Collins, L. Koekemoer, T. Krojer, R. Talon, E. Nelson, M. Ye, R. Nowak, J. Newman, J. T. Ng, N. Mitrovich, H. Wiggers, F. von Delft, *Acta Crystallogr D Struct Biol* **2021**, *77*, 62–74.

[23] T. Krojer, R. Talon, N. Pearce, P. Collins, A. Douangamath, J. Brandao-Neto, A. Dias, B. Marsden, F. von Delft, *Acta Crystallogr D Struct Biol* **2017**, *73*, 267–278.

[24] N. M. Pearce, T. Krojer, A. R. Bradley, P. Collins, R. P. Nowak, R. Talon, B. D. Marsden, S. Kelm, J. Shi, C. M. Deane, F. von Delft, *Nature Comms.* **2017**, *8*, 15123.

[25] A. Morris, W. McCorkindale, T. C. M. Consortium, N. Drayman, J. D. Chodera, S. Tay, N. London, A. A. Lee, *Chem. Commun.* **2021**, *57*, 5909–5912.

[26] J. T. Ng, C. Dekker, P. Reardon, F. von Delft, *Acta Crystallogr D Struct Biol* **2016**, *72*, 224–235.

[27] B. Wienen-Schmidt, M. Oebbeke, K. Ngo, A. Heine, G. Klebe, *ChemMedChem* **2021**, *16*, 292–300.

[28] P. Kirsch, A. M. Hartman, A. K. H. Hirsch, M. Empting, *Molecules* **2019**, *24*, 4309.

[29] D. Bajusz, W. S. Wade, G. Satała, A. J. Bojarski, J. Ilaš, J. Ebner, F. Grebien, H. Papp, F. Jakab, A. Douangamath, D. Fearon, F. von Delft, M. Schuller, I. Ahel, A. Wakefield, S. Vajda, J. Gerencsér, P. Pallai, G. M. Keserű, *Nat Commun* **2021**, *12*, 3201.

[30] H.-W. Wang, J.-W. Wang, *Protein Science* **2017**, *26*, 32–39.

[31] M. Saur, M. J. Hartshorn, J. Dong, J. Reeks, G. Bunkoczi, H. Jhoti, P. A. Williams, *Drug Discovery Today* **2020**, *25*, 485–490.

[32] L. Wang, J. Gao, R. Ma, Y. Liu, M. Liu, F. Zhong, J. Hu, S. Li, J. Wu, H. Jiang, J. Zhang, K. Ruan, *Magnetic Resonance Letters* **2022**, *2*, 107–118.

[33] D. A. Erlanson, S. W. Fesik, R. E. Hubbard, W. Jahnke, H. Jhoti, *Nat Rev Drug Discov* **2016**, *15*, 605–619.

[34] L. R. de Souza Neto, J. T. Moreira-Filho, B. J. Neves, R. L. B. R. Maidana, A. C. R. Guimarães, N. Furnham, C. H. Andrade, F. P. Silva, *Front. Chem.* **2020**, *8*, 93.

[35] A. Kumar, K. Y. J. Zhang, *Methods* **2015**, *71*, 26–37.

[36] A.-J. Banegas-Luna, J. P. Cerón-Carrasco, H. Pérez-Sánchez, *Future Medicinal Chemistry* **2018**, *10*, 2641–2658.

[37] S. Leung, M. Bodkin, F. von Delft, P. Brennan, G. Morris, *SuCOS Is Better than RMSD for Evaluating Fragment Elaboration and Docking Poses*, Chemistry, **2019**.

[38] M. Fischer, R. G. Coleman, J. S. Fraser, B. K. Shoichet, *Nature Chem* **2014**, *6*, 575–583.

[39] M. Bissaro, M. Sturlese, S. Moro, *Drug Discovery Today* **2020**, *25*, 1693–1701.

[40] S. Genheden, U. Ryde, *Expert Opinion on Drug Discovery* **2015**, *10*, 449–461.

[41] J. D. St. Denis, R. J. Hall, C. W. Murray, T. D. Heightman, D. C. Rees, *RSC Med. Chem.* **2021**, *12*, 321–329.

[42] S. Chow, S. Liver, A. Nelson, *Nat Rev Chem* **2018**, *2*, 174–183.

[43] L. M. Baker, A. Aimon, J. B. Murray, A. E. Surgenor, N. Matassova, S. D. Roughley, P. M. Collins, T. Krojer, F. von Delft, R. E. Hubbard, *Commun Chem* **2020**, *3*, 122.

[44] J. B. Murray, S. D. Roughley, N. Matassova, P. A. Brough, *J. Med. Chem.* **2014**, *57*, 2845–2850.

[45] D. E. Scott, A. G. Coyne, S. A. Hudson, C. Abell, *Biochemistry* **2012**, *51*, 4990–5003.

[46] G. Holdgate, K. Embrey, A. Milbradt, G. Davies, *ADMET DMPK* **2019**, *7*, 222–241.

[47] A. Yasgar, A. Jadhav, A. Simeonov, N. P. Coussens, in *High Throughput Screening* (Ed.: W.P. Janzen), Springer New York, New York, NY, **2016**, pp. 77–98.

[48] A. Roy, *High-Throughput* **2018**, *7*, 4.

[49] M. S. Salahudeen, P. S. Nishtala, *Saudi Pharmaceutical Journal* **2017**, *25*, 165–175.

[50] E. N. Muratov, J. Bajorath, R. P. Sheridan, I. V. Tetko, D. Filimonov, V. Poroikov, T. I. Oprea, I. I. Baskin, A. Varnek, A. Roitberg, O. Isayev, S. Curtalolo, D. Fourches, Y. Cohen, A. Aspuru-Guzik, D. A. Winkler, D. Agrafiotis, A. Cherkasov, A. Tropsha, *Chem. Soc. Rev.* **2020**, *49*, 3525–3564.

[51] J. Farhang-Fallah, X. Yin, G. Trentin, A. M. Cheng, M. Rozakis-Adcock, *J. Biol. Chem.* **2000**, *275*, 40492–40497.

[52] A. Podcheko, P. Northcott, G. Bikopoulos, A. Lee, S. R. Bommareddi, J. A. Kushner, J. Farhang-Fallah, M. Rozakis-Adcock, *Mol Cell Biol* **2007**, *27*, 6484–6496.

[53] S. Li, A. B. Francisco, C. Han, S. Pattabiraman, M. R. Foote, S. L. Giesy, C. Wang, J. C. Schimenti, Y. R. Boisclair, Q. Long, *FEBS Letts* **2010**, *584*, 4121–4127.

[54] D. De Semir, M. Nosrati, V. Bezrookove, A. A. Dar, S. Federman, G. Bienvenu, S. Venna, J. Rangel, J. Climent, T. M. Meyer Tamgüney, S. Thummala, S. Tong, S. P. L. Leong, C. Haqq, P. Billings, J. R. Miller, R. W. Sagebiel, R. Debs, M. Kashani-Sabet, *Proc. Natl. Acad. Sci.* **2012**, *109*, 7067.

[55] D. de Semir, V. Bezrookove, M. Nosrati, A. A. Dar, C. Wu, J. Shen, C. Rieken, M. Venkatasubramanian, J. R. Miller, P.-Y. Desprez, S. McAllister, L. Soroceanu, R. J. Debs, N. Salomonis, D. Schadendorf, J. E. Cleaver, M. Kashani-Sabet, *Proc. Natl. Acad. Sci. USA* **2018**, *115*, E5766.

[56] D. de Semir, V. Bezrookove, M. Nosrati, K. R. Scanlon, E. Singer, J. Judkins, C. Rieken, C. Wu, J. Shen, C. Schmudermayer, A. A. Dar, J. R. Miller, C. Cobbs, G. Yount, P.-Y. Desprez, R. J. Debs, N. Salomonis, S. McAllister, J. E. Cleaver, L. Soroceanu, M. Kashani-Sabet, *Proc. Natl. Acad. Sci.* **2020**, *117*, 9064.

[57] C. Xu, J. Min, *Protein Cell* **2011**, *2*, 202–214.

[58] N. Zaware, M.-M. Zhou, *Nat Struct Mol Biol* **2019**, *26*, 870–879.

[59] P. Filippakopoulos, S. Picaud, M. Mangos, T. Keates, J.-P. Lambert, D. Barsyte-Lovejoy, I. Felletar, R. Volkmer, S. Müller, T. Pawson, A.-C. Gingras, C. H. Arrowsmith, S. Knapp, *Cell* **2012**, *149*, 214–231.

[60] G. Klebe, in *Drug Design* (Ed.: G. Klebe), Springer Berlin Heidelberg, Berlin, Heidelberg, **2013**, pp. 61–88.

[61] X. Du, Y. Li, Y.-L. Xia, S.-M. Ai, J. Liang, P. Sang, X.-L. Ji, S.-Q. Liu, *IJMS* **2016**, *17*, 144.

[62] P. T. Wingfield, *CP Protein Science* **2015**, *80*, DOI 10.1002/0471140864.ps0601s80.

[63] P. Evans, A. McCoy, *Acta Crystallogr D Biol Crystallogr* **2008**, *64*, 1–10.

[64] N. M. Pearce, T. Krojer, F. Von Delft, *Acta Crystallogr D Struct Biol* **2017**, *73*, 256–266.

[65] N. M. Pearce, A. R. Bradley, T. Krojer, B. D. Marsden, C. M. Deane, F. Von Delft, *Structural Dynamics* **2017**, *4*, 032104.

[66] D. Patko, K. Cottier, A. Hamori, R. Horvath, *Opt. Express* **2012**, *20*, 23162.

[67] M. J. E. Fischer, in *Surface Plasmon Resonance* (Eds.: N.J. Mol, M.J.E. Fischer), Humana Press, Totowa, NJ, **2010**, pp. 55–73.

[68] Ö. Kartal, F. Andres, M. P. Lai, R. Nehme, K. Cottier, *SLAS Discovery* **2021**, *26*, 995–1003.

[69] M. Wojdyr, R. Keegan, G. Winter, A. Ashton, *Acta Crystallogr A Found Crystallogr* **2013**, *69*, s299–s299.

[70] P. Emsley, K. Cowtan, *Acta Crystallogr D Biol Crystallogr* **2004**, *60*, 2126–2132.

[71] G. Bricogne, E. Blanc, M. Brandl, C. Flensburg, P. Keller, W. Paciorek, P. Roversi, A. Sharff, O. S. Smart, C. Vonrhein, T. O. Womack, **2017**.

[72] R. A. Nicholls, F. Long, G. N. Murshudov, in *Advancing Methods for Biomolecular Crystallography* (Eds.: R. Read, A.G. Urzhumtsev, V.Y. Lunin), Springer Netherlands, Dordrecht, **2013**, pp. 231–258.

[73] D. Weininger, *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31–36.

[74] M. Mokaya, F. Imrie, W. P. Van Hoorn, A. Kalisz, A. R. Bradley, C. M. Deane, *Nat Mach Intell* **2023**, *5*, 386–394.

[75] N. M. O'Boyle, *J Cheminform* **2012**, *4*, 22.

[76] H. L. Morgan, *J. Chem. Doc.* **1965**, *5*, 107–113.

[77] D. Rogers, M. Hahn, *J. Chem. Inf. Model.* **2010**, *50*, 742–754.

[78] A. Capecchi, D. Probst, J.-L. Reymond, *J Cheminform* **2020**, *12*, 43.

[79] D. Bajusz, A. Rácz, K. Héberger, *J Cheminform* **2015**, *7*, 20.

[80] L. van der Maaten, G. Hinton, *Journal of Machine Learning Research* **2008**, *9*, 2579–2605.

[81] L. Breiman, *Mach. Learn.* **2001**, *45*, 5–32.

[82] O. Trott, A. J. Olson, *J. Comput. Chem.* **2010**, *31*, 455–61.

[83] G. M. Morris, D. S. Goodsell, R. S. Halliday, R. Huey, W. E. Hart, R. K. Belew, A. J. Olson, *J. Comput. Chem.* **1998**, *19*, 1639–1662.

[84] A. T. McNutt, P. Francoeur, R. Aggarwal, T. Masuda, R. Meli, M. Ragoza, J. Sunseri, D. R. Koes, *J. Cheminf.* **2021**, *13*, 43.

[85] S.-Y. Huang, X. Zou, *IJMS* **2010**, *11*, 3016–3034.

[86] M. Ferla, **2023**.

[87] E. Braun, J. Gilmer, H. B. Mayes, D. L. Mobley, J. I. Monroe, S. Prasad, D. M. Zuckerman, *LiveCoMS* **2019**, *1*, DOI 10.33011/livecoms.1.1.5957.

[88] V. Hornak, R. Abel, A. Okur, B. Strockbine, A. Roitberg, C. Simmerling, *Proteins* **2006**, *65*, 712–725.

[89] K. Lindorff-Larsen, S. Piana, K. Palmo, P. Maragakis, J. L. Klepeis, R. O. Dror, D. E. Shaw, *Proteins* **2010**, *78*, 1950–1958.

[90] J. A. Maier, C. Martinez, K. Kasavajhala, L. Wickstrom, K. E. Hauser, C. Simmerling, *J. Chem. Theory Comput.* **2015**, *11*, 3696–3713.

[91] L. Raich, K. Meier, J. Günther, C. D. Christ, F. Noé, S. Olsson, *Proc. Natl. Acad. Sci. U.S.A.* **2021**, *118*, e2017427118.

[92] V. S. Inakollu, D. P. Geerke, C. N. Rowley, H. Yu, *Current Opinion in Structural Biology* **2020**, *61*, 182–190.

[93] L. Duan, X. Liu, J. Z. H. Zhang, *J. Am. Chem. Soc.* **2016**, *138*, 5722–5728.

[94] T. Tuccinardi, *Expert Opinion on Drug Discovery* **2021**, *16*, 1233–1237.

[95] H. Sahakyan, *J Comput Aided Mol Des* **2021**, *35*, 731–736.

[96] P. A. Greenidge, C. Kramer, J.-C. Mozziconacci, R. M. Wolf, *J. Chem. Inf. Model.* **2013**, *53*, 201–209.

[97] P. M. Collins, A. Douangamath, R. Talon, A. Dias, J. Brandao-Neto, T. Krojer, F. von Delft, in *Methods Enzymol.* (Ed.: C.A. Lesburg), Academic Press, **2018**, pp. 251–264.

[98] O. B. Cox, T. Krojer, P. Collins, O. Monteiro, R. Talon, A. Bradley, O. Fedorov, J. Amin, B. D. Marsden, J. Spencer, F. von Delft, P. E. Brennan, *Chem. Sci.* **2016**, *7*, 2322–2330.

[99] P. C. D. Hawkins, A. G. Skillman, G. L. Warren, B. A. Ellingson, M. T. Stahl, *J Chem Inf Model* **2010**, *50*, DOI 10.1021/ci100031x.

[100] J. Sadowski, J. Gasteiger, *Chem. Rev.* **1993**, *93*, 2567–2581.

[101] G. Jones, P. Willett, R. C. Glen, Leach A. R, R. Taylor, *J. Mol. Biol.* **1997**, *267*, 727–748.

[102] P. T. Lang, S. R. Brozell, S. Mukherjee, E. F. Pettersen, E. C. Meng, V. Thomas, R. C. Rizzo, D. A. Case, T. L. James, I. D. Kuntz, *RNA* **2009**, *15*, 1219–1230.

[103] N. M. O'Boyle, M. Banck, C. A. James, C. Morley, T. Vandermeersch, G. R. Hutchison, *J. Cheminf.* **2011**, *3*, 33.

[104] O. Korb, T. Stützle, T. E. Exner, in (Eds.: M. Dorigo, L.M. Gambardella, M. Birattari, A. Martinoli, R. Poli, T. Stützle), Springer Berlin Heidelberg, **2006**, pp. 247–258.

[105] D. Alvarez-Garcia, X. Barril, *J. Med. Chem.* **2014**, *57*, 8530–8539.

[106] S. Ruiz-Carmona, D. Alvarez-Garcia, N. Foloppe, A. B. Garmendia-Doval, S. Juhos, P. Schmidtke, X. Barril, R. E. Hubbard, S. D. Morley, *PLOS Comp. Biol.* **2014**, *10*, e1003571.

[107] S. Ruiz-Carmona, P. Schmidtke, F. J. Luque, L. Baker, N. Matassova, B. Davis, S. Roughley, J. Murray, R. Hubbard, X. Barril, *Nature Chemistry* **2017**, *9*, 201–206.

[108] M. Wandhammer, E. Carletti, M. Van der Schans, E. Gillon, Y. Nicolet, P. Masson, M. Goeldner, D. Noort, F. Nachon, *J. Biol. Chem.* **2011**, *286*, 16783–16789.

[109] M. Schuller, G. J. Correy, S. Gahbauer, D. Fearon, T. Wu, R. E. Díaz, I. D. Young, L. Carvalho Martins, D. H. Smith, U. Schulze-Gahmen, T. W. Owens, I. Deshpande, G. E. Merz, A. C. Thwin, J. T. Biel, J. K. Peters, M. Moritz, N. Herrera, H. T. Kratochvil, Qcrg Structural Biology Consortium, A. Aimon, J. M. Bennett, J. Brandao Neto, A. E. Cohen, A. Dias, A. Douangamath, L. Dunnett, O. Fedorov, M. P. Ferla, M. R. Fuchs, T. J. Gorrie-Stone, J. M. Holton, M. G. Johnson, T. Krojer, G. Meigs, A. J. Powell, J. G. M. Rack, V. L. Rangel, S. Russi, R. E. Skyner, C. A. Smith, A. S. Soares, J. L. Wierman, K. Zhu, P. O'Brien, N. Jura, A. Ashworth, J. J. Irwin, M. C. Thompson, J. E. Gestwicki, F. von Delft, B. K. Shoichet, J. S. Fraser, I. Ahel, *Sci. Adv.* **2021**, *7*, eabf8711.

[110] D. Sabbadin, S. Moro, *J. Chem. Inf. Model.* **2014**, *54*, 372–376.

[111] G. Martínez-Rosell, T. Giorgino, G. De Fabritiis, *J. Chem. Inf. Model.* **2017**, *57*, 1511–1516.

[112] G. Bitencourt-Ferreira, W. F. Azevedo, *Methods Mol. Biol.* **2019**, *2053*, 149–167.

[113] O. O. Grygorenko, D. S. Radchenko, I. Dziuba, A. Chuprina, K. E. Gubina, Y. S. Moroz, *iScience* **2020**, *23*, 101681.

[114] R. J. Hall, C. W. Murray, M. L. Verdonk, *J. Med. Chem.* **2017**, *60*, 6440–6450.

[115] M. A. Fligner, J. S. Verducci, P. E. Blower, *Technometrics* **2002**, *44*, 110–119.

[116] J. Newman, O. Dolezal, V. Fazio, T. Caradoc-Davies, T. S. Peat, *J Comput Aided Mol Des* **2012**, *26*, 497–503.

[117] J. L. Kulp, S. N. Blumenthal, Q. Wang, R. L. Bryan, F. Guarnieri, *J. Comput-Aided Mol. Des.* **2012**, *26*, 583–594.

[118] T. S. Peat, O. Dolezal, J. Newman, D. Mobley, J. J. Deadman, *J. Comput-Aided Mol. Des.* **2014**, *28*, 347–362.

[119] D. L. Mobley, S. Liu, N. M. Lim, K. L. Wymer, A. L. Perryman, S. Forli, N. Deng, J. Su, K. Branson, A. J. Olson, *J Comput Aided Mol Des* **2014**, *28*, 327–345.

[120] J. Schiebel, S. G. Krimmer, K. Röwer, A. Knörlein, X. Wang, A. Y. Park, M. Stieler, F. R. Ehrmann, K. Fu, N. Radeva, M. Krug, F. U. Huschmann, S. Glöckner, M. S. Weiss, U. Mueller, G. Klebe, A. Heine, *Structure* **2016**, *24*, 1398–1409.

[121] H. Zhong, Z. Wang, X. Wang, H. Liu, D. Li, H. Liu, X. Yao, T. Hou, *Phys. Chem. Chem. Phys.* **2019**, *21*, 25276–25289.

[122] C. D. Parks, Z. Gaieb, M. Chiu, H. Yang, C. Shao, W. P. Walters, J. M. Jansen, G. McGaughey, R. A. Lewis, S. D. Bembenek, M. K. Ameriks, T. Mirzadegan, S. K. Burley, R. E. Amaro, M. K. Gilson, *J. Comput-Aided Mol. Des.* **2020**, *34*, 99–119.

[123] R. Wang, X. Fang, Y. Lu, C.-Y. Yang, S. Wang, *J. Med. Chem.* **2005**, *48*, 4111–4119.

[124] M. Aldeghi, G. A. Ross, M. J. Bodkin, J. W. Essex, S. Knapp, P. C. Biggin, *Commun. Chem.* **2018**, *1*, 19.

[125] A. E. Wakefield, C. Yueh, D. Beglov, M. S. Castilho, D. Kozakov, G. M. Keserű, A. Whitty, S. Vajda, *J. Chem. Inf. Model.* **2020**, *60*, 6612–6623.

[126] O. Méndez-Lucio, M. Ahmad, E. A. del Rio-Chanona, J. K. Wegner, *Nat Mach Intell* **2021**, *3*, 1033–1039.

[127] J. Janin, *Protein Science* **2005**, *14*, 278–283.

[128] G. Landrum, **2016**.

[129] R. Meli, P. C. Biggin, *J. Cheminf.* **2020**, *12*, 49.

[130] H. Grosjean, M. Işık, A. Aimon, D. Mobley, J. Chodera, F. von Delft, P. C. Biggin, *J Comput Aided Mol Des* **2022**, *36*, 291–311.

[131] W. Thompson, **2022**, DOI 10.5281/ZENODO.7018551.

[132] L. M. Baker, A. Aimon, J. B. Murray, A. E. Surgenor, N. Matassova, S. D. Roughley, P. M. Collins, T. Krojer, F. von Delft, R. E. Hubbard, *Commun Chem* **2020**, *3*, 1–11.

[133] N. M. Pearce, T. Krojer, A. R. Bradley, P. Collins, R. P. Nowak, R. Talon, B. D. Marsden, S. Kelm, J. Shi, C. M. Deane, F. von Delft, *Nat Commun* **2017**, *8*, 15123.

[134] M. Philpott, J. Yang, T. Tumber, O. Fedorov, S. Uttarkar, P. Filippakopoulos, S. Picaud, T. Keates, I. Felletar, A. Ciulli, S. Knapp, T. D. Heightman, *Mol Biosyst* **2011**, *7*, 2899–2908.

[135] S. D. Roughley, R. E. Hubbard, *J. Med. Chem.* **2011**, *54*, 3989–4005.

[136] C. N. Johnson, D. A. Erlanson, W. Jahnke, P. N. Mortenson, D. C. Rees, *J. Med. Chem.* **2018**, *61*, 1774–1784.

[137] S. Chow, S. Liver, A. Nelson, *Nat Rev Chem* **2018**, *2*, 174–183.

[138] M. R. Bentley, O. V. Ilyichova, G. Wang, M. L. Williams, G. Sharma, W. S. Alwan, R. L. Whitehouse, B. Mohanty, P. J. Scammells, B. Heras, J. L. Martin, M. Totsika, B. Capuano, B. C. Doak, M. J. Scanlon, *J. Med. Chem.* **2020**, *63*, 6863–6875.

[139] R. P. Thomas, R. E. Heap, F. Zappacosta, E. K. Grant, P. Pogány, S. Besley, D. J. Fallon, M. M. Hann, D. House, N. C. O. Tomkinson, J. T. Bush, *Chem. Sci.* **2021**, *12*, 12098–12106.

[140] C.-J. Li, B. M. Trost, *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 13197–13202.

[141] L. Gao, S. Shaabani, A. Reyes Romero, R. Xu, M. Ahmadianmoghaddam, A. Dömling, *Green Chem.* **2023**, 10.1039.D2GC04312B.

[142] M. Schauperl, P. Czodrowski, J. E. Fuchs, R. G. Huber, B. J. Waldner, M. Podewitz, C. Kramer, K. R. Liedl, *J. Chem. Inf. Model.* **2017**, *57*, 345–354.

[143] M. Aldeghi, G. A. Ross, M. J. Bodkin, J. W. Essex, S. Knapp, P. C. Biggin, *Commun. Chem.* **2018**, *1*, 19.

[144] S. Malhotra, J. Karanicolas, *J. Med. Chem.* **2017**, *60*, 128–145.

[145] E. J. Barreiro, A. E. Kümmerle, C. A. M. Fraga, *Chem. Rev.* **2011**, *111*, 5215–5246.

[146] J. Bajorath, *Expert Opinion on Drug Discovery* **2021**, *16*, 719–721.

[147] S. Kar, H. Sanderson, K. Roy, E. Benfenati, J. Leszczynski, *Chem. Rev.* **2022**, *122*, 3637–3710.

[148] J. T. DiPiro, M. A. Chisholm-Burns, *AJPE* **2013**, *77*, 159.

[149] T. Cheng, Q. Li, Z. Zhou, Y. Wang, S. H. Bryant, *AAPS J* **2012**, *14*, 133–141.

[150] H. Grosjean, A. Aimon, S. Hassell-Hart, W. Thompson, L. Koekemoer, J. Bennett, C. Anderson, E. A. FitzGerald, T. Krojer, A. Bradley, O. Fedorov, P. C. Biggin, J. Spencer, F. von Delft, *High-Throughput Crystallography for Rapid Fragment Growth from Crude Arrays by Low-Cost Robotics*, Diamond Light Source, **2023**.

[151] P. M. Collins, A. Douangamath, R. Talon, A. Dias, J. Brandao-Neto, T. Krojer, F. von Delft, in *Methods in Enzymology*, Elsevier, **2018**, pp. 251–264.

[152] L. Ciccone, L. Vera, L. Tepshi, L. Rosalia, A. Rossello, E. A. Stura, *Biotechnology Reports* **2015**, *7*, 120–127.

[153] C. X. Weichenberger, P. V. Afonine, K. Kantardjieff, B. Rupp, *Acta Crystallogr D Biol Crystallogr* **2015**, *71*, 1023–1038.

[154] S. Tian, Y. Li, J. Wang, J. Zhang, T. Hou, *Mol. Pharmaceutics* **2011**, *8*, 841–851.

[155] J. Wu, Q. Zhang, W. Wu, T. Pang, H. Hu, W. K. B. Chan, X. Ke, Y. Zhang, *Bioinformatics* **2018**, *34*, 2271–2282.

[156] Y. Low, T. Uehara, Y. Minowa, H. Yamada, Y. Ohno, T. Urushidani, A. Sedykh, E. Muratov, V. Kuz'min, D. Fourches, H. Zhu, I. Rusyn, A. Tropsha, *Chem. Res. Toxicol.* **2011**, *24*, 1251–1262.

[157] Y. Gilad, K. Nadassy, H. Senderowitz, *J Cheminform* **2015**, *7*, 61.

[158] S. Riniker, G. A. Landrum, *J Cheminform* **2013**, *5*, 26.

[159] R. Todeschini, V. Consonni, H. Xiang, J. Holliday, M. Buscema, P. Willett, *J. Chem. Inf. Model.* **2012**, *52*, 2884–2901.

[160] P. Willett, J. M. Barnard, G. M. Downs, *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983–996.

[161] T. Saito, M. Rehmsmeier, *PLoS ONE* **2015**, *10*, e0118432.

[162] "REAL Database - Enamine," can be found under https://enamine.net/compound-collections/real-compounds/real-database, **n.d.**

[163] R. Sanchez-Garcia, D. Havasi, G. Takács, M. C. Robinson, A. Lee, F. von Delft, C. M. Deane, *Digital Discovery* **2023**, *2*, 103–111.

[164] R. F. Alford, A. Leaver-Fay, J. R. Jeliazkov, M. J. O'Meara, F. P. DiMaio, H. Park, M. V. Shapovalov, P. D. Renfrew, V. K. Mulligan, K. Kappel, J. W. Labonte, M. S. Pacella, R. Bonneau, P. Bradley, R. L. Dunbrack, R. Das, D. Baker, B. Kuhlman, T. Kortemme, J. J. Gray, *J. Chem. Theory Comput.* **2017**, *13*, 3031–3048.

[165] Ö. Kartal, F. Andres, M. P. Lai, R. Nehme, K. Cottier, *SLAS Discovery* **2021**, *26*, 995–1003.

[166] S. Dara, S. Dhamercherla, S. S. Jadav, C. M. Babu, M. J. Ahsan, *Artif Intell Rev* **2022**, *55*, 1947–1999.

[167] S. Korkmaz, *J. Chem. Inf. Model.* **2020**, *60*, 4180–4190.

[168] A. V. Zakharov, M. L. Peach, M. Sitzmann, M. C. Nicklaus, *J. Chem. Inf. Model.* **2014**, *54*, 705–712.

[169] S. Malhotra, J. Karanicolas, *J. Med. Chem.* **2017**, *60*, 128–145.

[170] M. N. Drwal, G. Bret, C. Perez, C. Jacquemard, J. Desaphy, E. Kellenberger, *J. Med. Chem.* **2018**, *61*, 5963–5973.

[171] M. Schauperl, P. Czodrowski, J. E. Fuchs, R. G. Huber, B. J. Waldner, M. Podewitz, C. Kramer, K. R. Liedl, *J. Chem. Inf. Model.* **2017**, *57*, 345–354.

[172] J. P. A. Ioannidis, *PLoS Med* **2005**, *2*, e124.

[173] Y. O. Adeshina, E. J. Deeds, J. Karanicolas, *Proc. Natl. Acad. Sci. U.S.A.* **2020**, *117*, 18477–18488.

[174] Y. Kawasaki, E. Freire, *Drug Discovery Today* **2011**, *16*, 985–990.

[175] G. Amendola, S. Cosconati, *J. Chem. Inf. Model.* **2021**, *61*, 3835–3845.

[176] S. M. Arif, J. D. Holliday, P. Willett, *J. Chem. Inf. Model.* **2010**, *50*, 1340–1349.

[177] J. B. Jasper, L. Humbeck, T. Brinkjost, O. Koch, *J Cheminform* **2018**, *10*, 15.

[178] J. Scantlebury, L. Vost, A. Carbery, T. E. Hadfield, O. M. Turnbull, N. Brown, V. Chenthamarakshan, P. Das, H. Grosjean, F. von Delft, C. M. Deane, *PointVS: A Machine Learning Scoring Function That Identifies Important Binding Interactions*, Bioinformatics, **2022**.

[179] Y. Khalak, G. Tresadern, D. F. Hahn, B. L. de Groot, V. Gapsys, *J. Chem. Theory Comput.* **2022**, *18*, 6259–6270.

[180] N. Stiefl, I. A. Watson, K. Baumann, A. Zaliani, *J. Chem. Inf. Model.* **2006**, *46*, 208–220.

[181] Q. Zhang, I. Muegge, *J. Med. Chem.* **2006**, *49*, 1536–1548.

[182] M. Fischer, *Quart. Rev. Biophys.* **2021**, *54*, e1.

[183] C. D. Blundell, M. J. Packer, A. Almond, *Bioorganic & Medicinal Chemistry* **2013**, *21*, 4976–4987.

[184] J. E. Ladbury, *Chemistry & Biology* **1996**, *3*, 973–980.

[185] M. Aldeghi, A. Heifetz, M. J. Bodkin, S. Knapp, P. C. Biggin, *Chem. Sci.* **2016**, *7*, 207–218.

[186] E. Kronenberg, F. Weber, S. Brune, D. Schepmann, C. Almansa, K. Friedland, E. Laurini, S. Pricl, B. Wünsch, *J. Med. Chem.* **2019**, *62*, 4204–4217.

[187] T. Akiba, S. Sano, T. Yanase, T. Ohta, M. Koyama, in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, ACM, Anchorage AK USA, **2019**, pp. 2623–2631.

[188] H. B. Mann, D. R. Whitney, *Ann. Math. Statist.* **1947**, *18*, 50–60.

[189] R. J. Gowers, M. Linke, J. Barnoud, T. J. E. Reddy, M. N. Melo, S. L. Seyler, D. L. Dotson, J. Domanski, S. Buchoux, I. M. Kenney, O. Beckstein, in *Proceedings of the 15th Python in Science Conference* (Eds.: S. Benthall, S. Rostrup), **2016**.

[190] M. S. Valdés-Tresanco, M. E. Valdés-Tresanco, P. A. Valiente, E. Moreno, *J. Chem. Theory Comput.* **2021**, *17*, 6281–6291.

[191] P. Ertl, A. Schuffenhauer, *J Cheminform* **2009**, *1*, 8.

[192] M. Rostkowski, M. H. Olsson, C. R. Søndergaard, J. H. Jensen, *BMC Struct Biol* **2011**, *11*, 6.

[193] A. T. McNutt, P. Francoeur, R. Aggarwal, T. Masuda, R. Meli, M. Ragoza, J. Sunseri, D. R. Koes, *J Cheminform* **2021**, *13*, 43.

[194] V. Ekberg, U. Ryde, *J. Chem. Theory Comput.* **2021**, *17*, 5379–5391.

[195] M. A. Kalam, A. Alshamsan, M. Alkholief, I. A. Alsarra, R. Ali, N. Haq, M. K. Anwer, F. Shakeel, *ACS Omega* **2020**, *5*, 1708–1716.

[196] R. G. Govindaraj, M. Brylinski, *BMC Bioinformatics* **2018**, *19*, 91.

[197] M. Bon, A. Bilsland, J. Bower, K. McAulay, *Molecular Oncology* **2022**, *16*, 3761–3777.

[198] H. Grosjean, M. Işık, A. Aimon, D. Mobley, J. Chodera, F. von Delft, P. C. Biggin, *J Comput Aided Mol Des* **2022**, *36*, 291–311.

[199] B. Lamoree, R. E. Hubbard, *Essays in Biochemistry* **2017**, *61*, 453–464.

[200] P. Bamborough, C. Chung, *Med. Chem. Commun.* **2015**, *6*, 1587–1604.

[201] I. Navratilova, T. Aristotelous, S. Picaud, A. Chaikuad, S. Knapp, P. Filappakopoulos, A. L. Hopkins, *ACS Med. Chem. Lett.* **2016**, *7*, 1213–1218.

[202] M. L. Peach, M. C. Nicklaus, *J Cheminform* **2009**, *1*, 6.

[203] K. W. Lexa, H. A. Carlson, *Quart. Rev. Biophys.* **2012**, *45*, 301–343.

[204] B. C. Roberts, R. L. Mancera, *J Chem Inf Model* **2008**, *48*, 397–408.

[205] S. Steiner, A. Magno, D. Huang, A. Caflisch, *FEBS Letters* **2013**, *587*, 2158–2163.

[206] J. Wallerstein, V. Ekberg, M. M. Ignjatović, R. Kumar, O. Caldararu, K. Peterson, S. Wernersson, U. Brath, H. Leffler, E. Oksanen, D. T. Logan, U. J. Nilsson, U. Ryde, M. Akke, *JACS Au* **2021**, *1*, 484–500.

[207] T. Hou, J. Wang, Y. Li, W. Wang, *J. Chem. Inf. Model.* **2011**, *51*, 69–82.

[208] H. Sun, L. Duan, F. Chen, H. Liu, Z. Wang, P. Pan, F. Zhu, J. Z. H. Zhang, T. Hou, *Phys. Chem. Chem. Phys.* **2018**, *20*, 14450–14460.

[209] M. Aldeghi, M. J. Bodkin, S. Knapp, P. C. Biggin, *J. Chem. Inf. Model.* **2017**, *57*, 2203–2221.

[210] D. de Semir, V. Bezrookove, M. Nosrati, A. A. Dar, C. Wu, J. Shen, C. Rieken, M. Venkatasubramanian, J. R. Miller, P.-Y. Desprez, S. McAllister, L. Soroceanu, R. J. Debs, N. Salomonis, D. Schadendorf, J. E. Cleaver, M. Kashani-Sabet, *Proc. Natl. Acad. Sci. U.S.A.* **2018**, *115*, E5766–E5775.

[211] S. Wills, R. Sanchez-Garcia, S. D. Roughley, A. Merritt, R. E. Hubbard, T. Dudgeon, J. Davidson, F. von Delft, C. M. Deane, *The Use of a Graph Database Is a Complementary Approach to a Classical Similarity Search for Identifying Commercially Available Fragment Merges*, Bioinformatics, **2022**.

[212] N. Eswar, B. Webb, M. A. Marti-Renom, M. S. Madhusudhan, D. Eramian, M. Shen, U. Pieper, A. Sali, *CP in Bioinformatics* **2006**, *15*, DOI 10.1002/0471250953.bi0506s15.

[213] T. J. Dolinsky, J. E. Nielsen, J. A. McCammon, N. A. Baker, *Nucleic Acids Research* **2004**, *32*, W665–W667.

[214] J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman, D. A. Case, *J. Comput. Chem.* **2004**, *25*, 1157–1174.

[215] A. Jakalian, D. B. Jack, C. I. Bayly, *J. Comput. Chem.* **2002**, *23*, 1623–1641.

[216] M. R. Shirts, C. Klein, J. M. Swails, J. Yin, M. K. Gilson, D. L. Mobley, D. A. Case, E. D. Zhong, *J Comput Aided Mol Des* **2017**, *31*, 147–161.

[217] H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, A. DiNola, J. R. Haak, *The Journal of Chemical Physics* **1984**, *81*, 3684–3690.

[218] U. Essmann, L. Perera, M. L. Berkowitz, T. Darden, H. Lee, L. G. Pedersen, *The Journal of Chemical Physics* **1995**, *103*, 8577–8593.

[219] B. Hess, H. Bekker, H. J. C. Berendsen, J. G. E. M. Fraaije, *J. Comput. Chem.* **1997**, *18*, 1463–1472.

[220] M. J. Abraham, T. Murtola, R. Schulz, S. Páll, J. C. Smith, B. Hess, E. Lindahl, *SoftwareX* **2015**, *1–2*, 19–25.

[221] **2022**.

[222] “Manifold API,” can be found under https://api.postera.ai/api/v1/docs/, **n.d.**

[223] P. Tosco, N. Stiefl, G. Landrum, *J Cheminform* **2014**, *6*, 37.

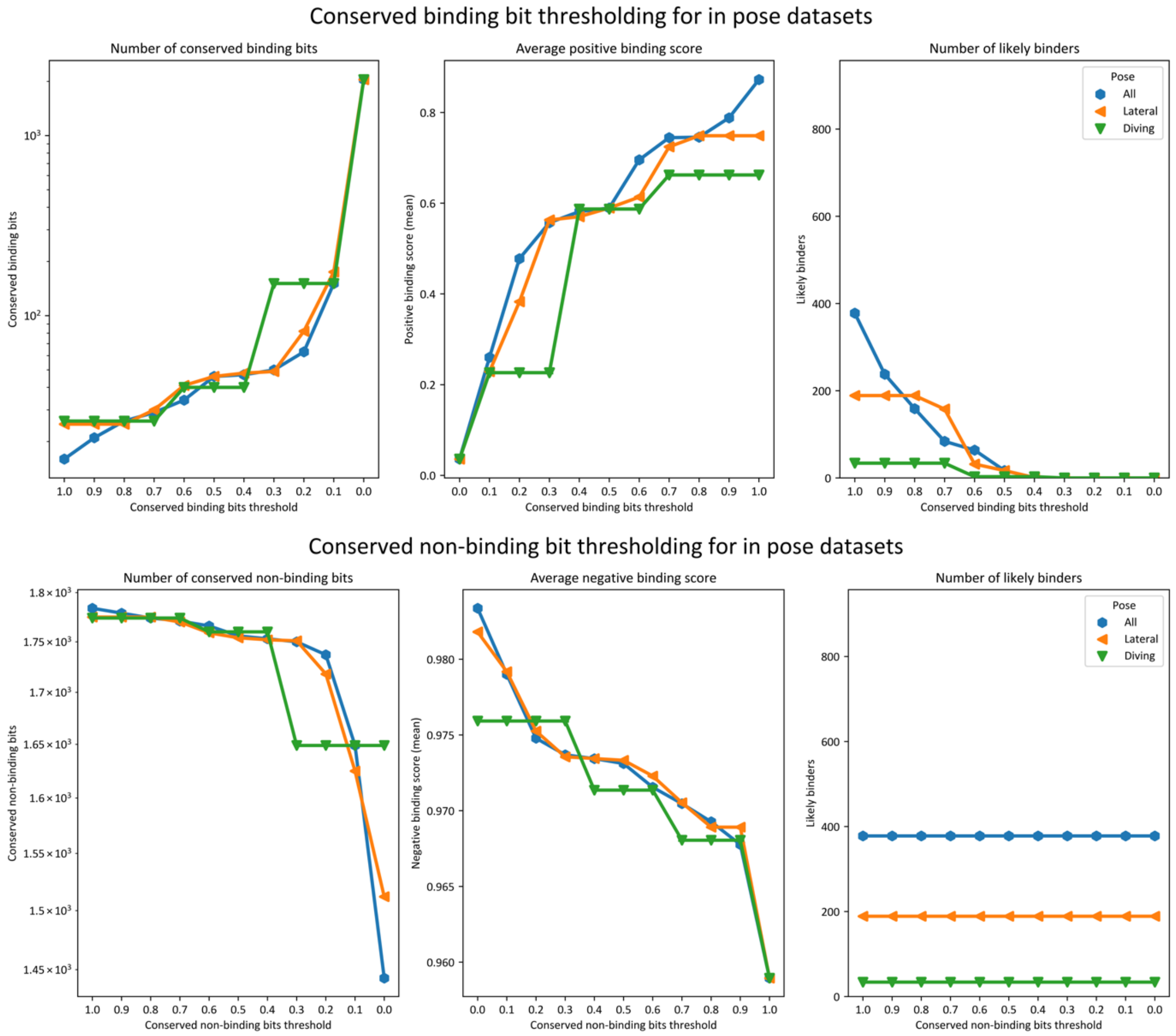
[224] S. Delagenière, P. Brenchereau, L. Launer, A. W. Ashton, R. Leal, S. Veyrier, J. Gabadinho, E. J. Gordon, S. D. Jones, K. E. Levik, S. M. McSweeney, S. Monaco, M. Nanao, D. Spruce, O. Svensson, M. A. Walsh, G. A. Leonard, *Bioinformatics* **2011**, *27*, 3186–3192.

[225] P. R. Evans, G. N. Murshudov, *Acta Crystallogr D Biol Crystallogr* **2013**, *69*, 1204–1214.

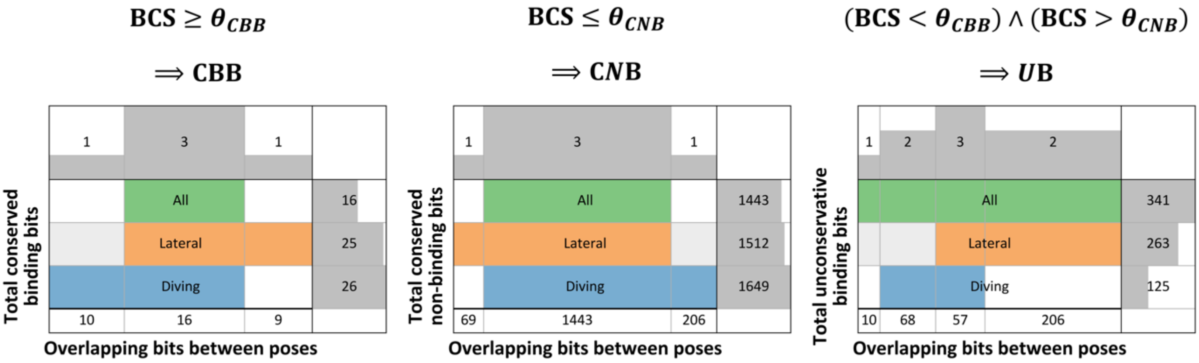
[226] A. J. McCoy, R. W. Grosse-Kunstleve, P. D. Adams, M. D. Winn, L. C. Storoni, R. J. Read, *J Appl Crystallogr* **2007**, *40*, 658–674.

Supplementary information

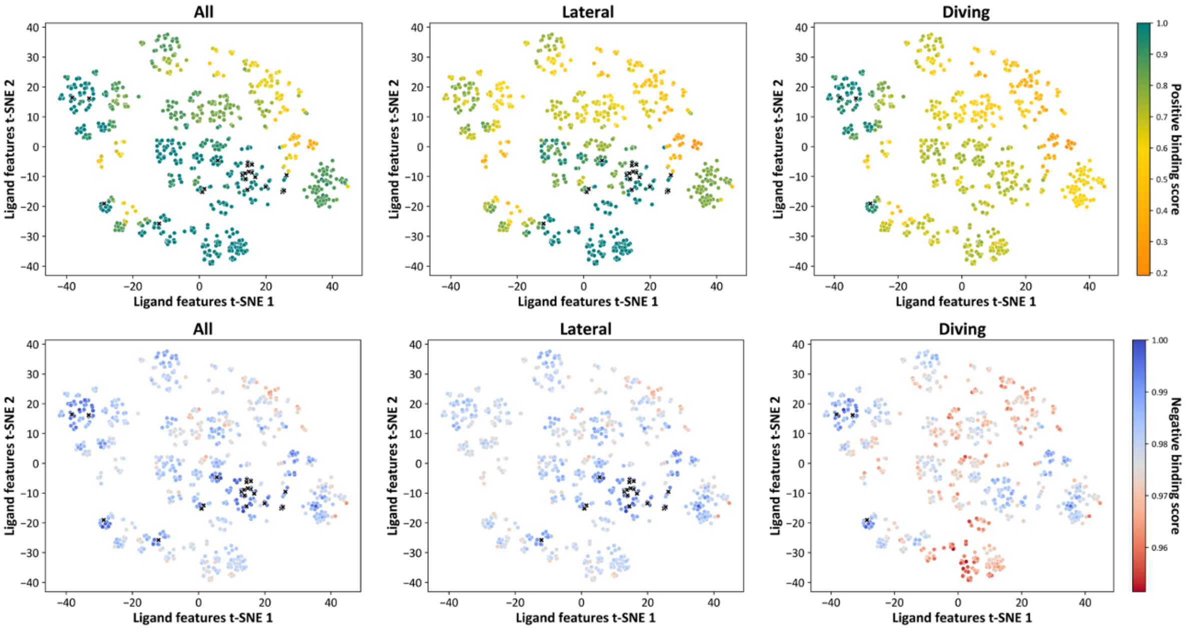
Supplementary information to Chapter 5



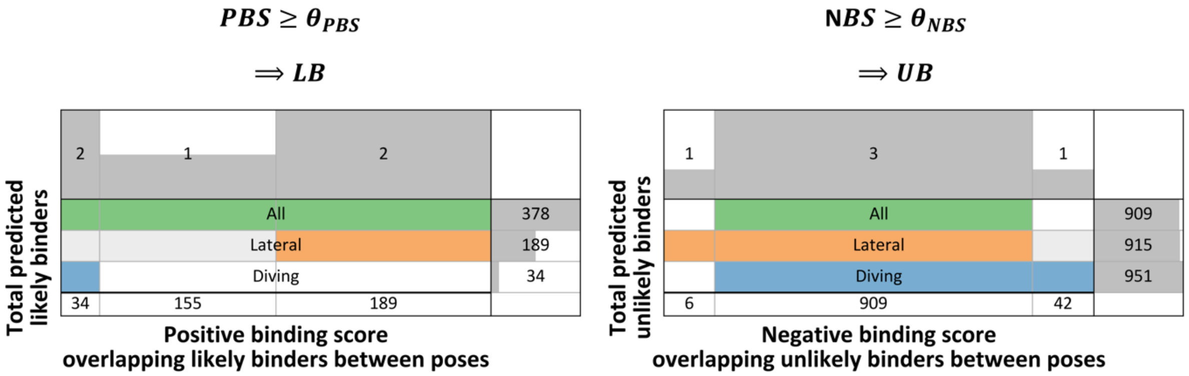
Supplementary figure 5.1: Binding bit conservation score thresholding and resulting effect on the binding scores and binding labels. The binding and non-binding bit conservation thresholds values were sampled between 0 and 1, and effect on the number of conserved bits, average positive and negative binding scores, and, prediction of likely and unlikely binders showed for each pose.



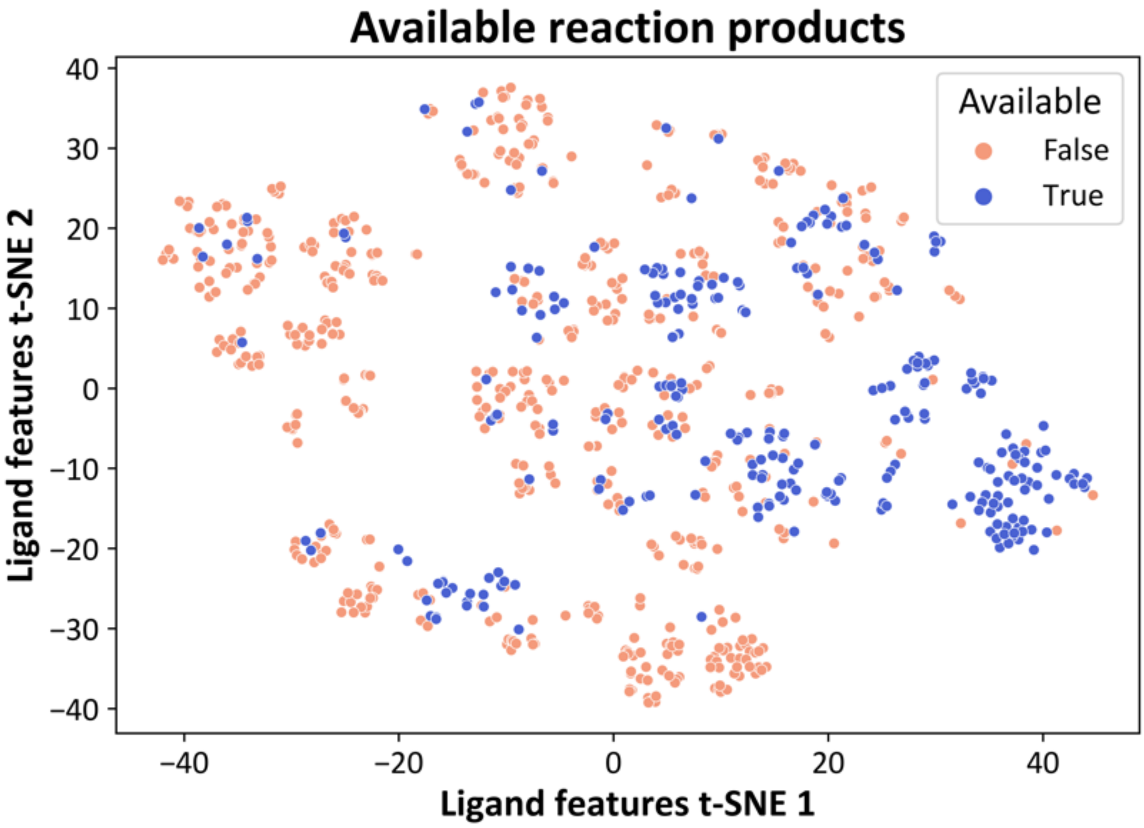
Supplementary figure 5.2: Bit ensemble comparison for different poses. 2D Venn diagrams comparing poses within each bit classification label: conserved binding, conserved non-binding, and unconservative bits. Each Venn diagram is separated into 5 rows and 4 columns, showing the number of sub-ensembles the bits are present in, the total number of bits composing the intersection, and the bit ensemble for and intersections between the different poses. The first 3 columns relate to the different ensembles and subsequent overlaps, while the last column shows the total number of bits for a given pose. The condition defining the classification is showed above each Venn diagram.



Supplementary figure 5.3: Pose specific binding landscapes. The top and bottom rows show pose-specific positive and negative binding score across all follow-up compounds, respectively. Initial binders resolved from the crystallographic screening of crude reaction mixtures are showed with black crosses.

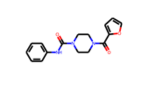


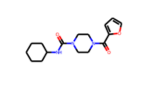
Supplementary figure 5.4: Compound binding predictions comparison for different poses. 2D Venn diagrams comparing poses within each compound classification label: likely and unlikely binders. Each Venn diagram is separated into 5 rows and 4 columns, showing the number of sub-ensembles the compounds are present in, the total number of compounds composing the intersection, and the compounds ensemble for and intersections between the different poses. The first 3 columns relate to the different ensembles and subsequent overlaps, while the last column shows the total number of compounds for a given pose. The condition defining the classification is showed above each Venn diagram.

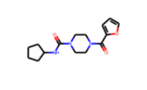


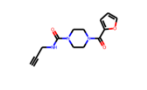
Supplementary figure 5.5: Available chemical space regions for rescreening of pure robotic compounds. The compounds initially enumerate from the automated chemistry experiments were enquired to Enamine for purchase in pure form. The available and unavailable compounds are showed in purple and pink, respectively.

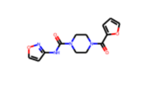
Supplementary table 5.1: 2D structures of robotic compounds purchased and rescreened in pure form. 97 robotically enumerated and screened compounds were purchased in pure form and rescreened at the by X-ray crystallography and by GCI assay.

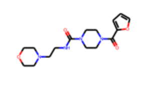

it1.0_AM-02_furan-piperazine-01

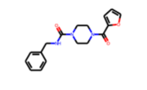

it1.0_AM-03_furan-piperazine-01

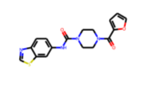

it1.0_AM-06_furan-piperazine-01

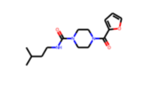

it1.0_AM-11_furan-piperazine-01

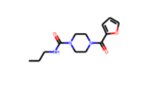

it1.0_AM-12_furan-piperazine-01

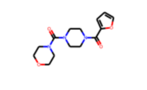

it1.0_AM-13_furan-piperazine-01

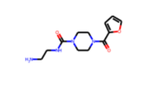

it1.0_AM-14_furan-piperazine-01

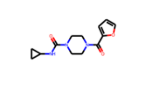

it1.0_AM-17_furan-piperazine-01

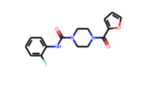

it1.0_AM-19_furan-piperazine-01

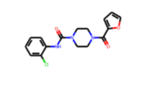

it1.0_AM-24_furan-piperazine-01

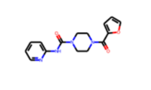

it1.0_AM-26_furan-piperazine-01

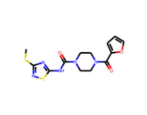

it1.0_AM-28_furan-piperazine-01

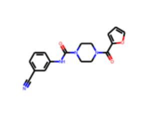

it1.0_AM-30_furan-piperazine-01

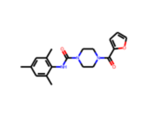

it1.0_AM-35_furan-piperazine-01

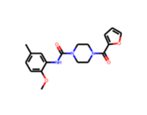

it1.0_AM-36_furan-piperazine-01

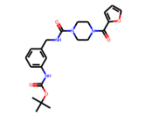

it1.0_AM-37_furan-piperazine-01

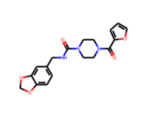

it1.0_AM-39_furan-piperazine-01

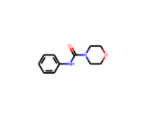

it1.0_AM-45_furan-piperazine-01

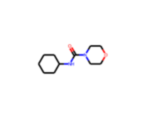

it1.0_AM-48_furan-piperazine-01

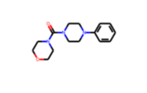

it1.0_AM-49_furan-piperazine-01

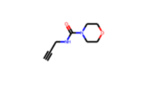

it1.0_AM-53_furan-piperazine-01

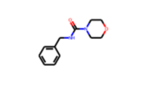

it1.0_AM-56_furan-piperazine-01

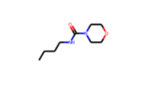

it1.0_AM-02_morpholine-01

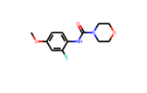

it1.0_AM-03_morpholine-01

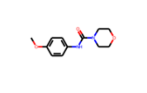

it1.0_AM-07_morpholine-01

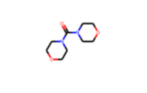

it1.0_AM-11_morpholine-01

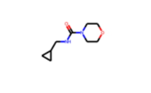

it1.0_AM-14_morpholine-01

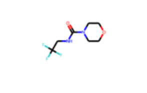

it1.0_AM-15_morpholine-01

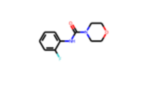

it1.0_AM-18_morpholine-01

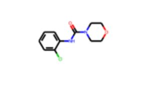

it1.0_AM-25_morpholine-01

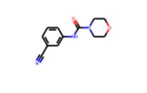

it1.0_AM-26_morpholine-01

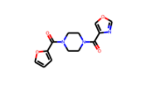

it1.0_AM-27_morpholine-01

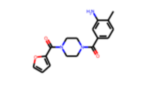

it1.0_AM-32_morpholine-01

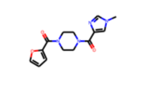

it1.0_AM-35_morpholine-01

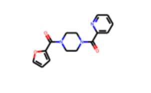

it1.0_AM-36_morpholine-01

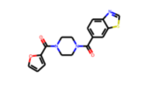

it1.0_AM-45_morpholine-01

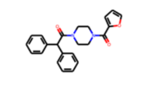

it2.0_CA-04_piperazine-01

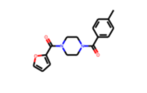

it2.0_CA-08_piperazine-01

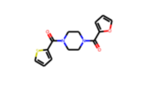

it2.0_CA-12_piperazine-01

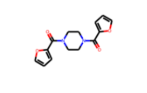

it2.0_CA-14_piperazine-01

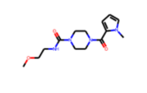

it2.0_CA-18_piperazine-01

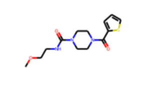

it2.0_CA-27_piperazine-01

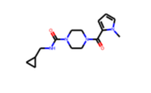

it2.0_CA-34_piperazine-01

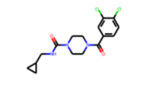

it2.0_CA-42_piperazine-01

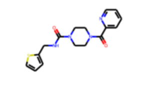

it2.0_CA-53_piperazine-01

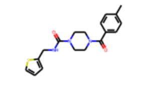

it3.0_AM-01_Imid-04_CA-35

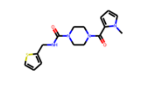

it3.0_AM-01_Imid-04_CA-42

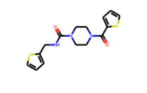

it3.0_AM-02_Imid-04_CA-35

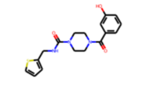

it3.0_AM-02_Imid-04_CA-56

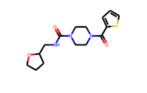

it3.0_AM-03_Imid-04_CA-14

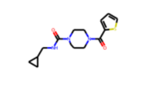

it3.0_AM-03_Imid-04_CA-34

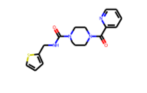

it3.0_AM-03_Imid-04_CA-35

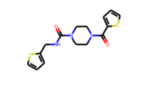

it3.0_AM-03_Imid-04_CA-42

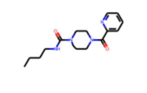

it3.0_AM-03_Imid-04_CA-58

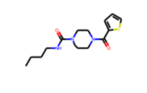

it3.0_AM-06_Imid-04_CA-42

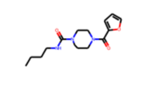

it3.5_AM-02_Imid-04_CA-16

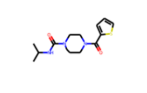

it3.5_AM-03_Imid-04_CA-07

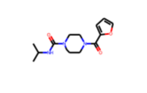

it3.5_AM-03_Imid-04_CA-16

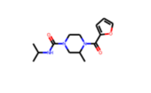

it3.5_AM-07_Imid-04_CA-07

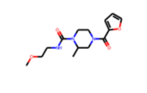

it3.5_AM-07_Imid-04_CA-16

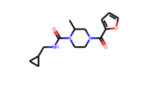

it3.5_AM-07_Imid-04_CA-20

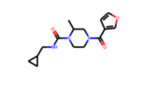

it3.5_AM-08_Imid-04_CA-16

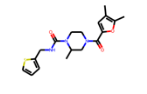

it3.5_AM-08_Imid-04_CA-20

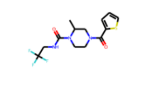

it3.5_AM-08_Imid-03_CA-20

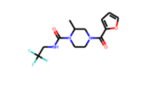

it3.5_AM-01_Imid-02_CA-20

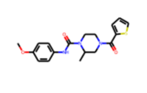

it3.5_AM-02_Imid-02_CA-20

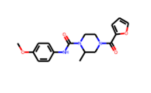

it3.5_AM-02_Imid-02_CA-22

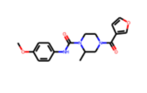

it3.5_AM-03_Imid-02_CA-29

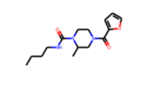

it3.5_AM-04_Imid-02_CA-16

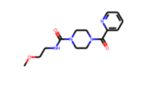

it3.5_AM-04_Imid-02_CA-20

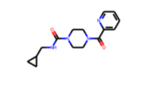

it3.5_AM-06_Imid-02_CA-16

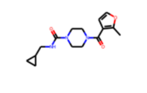

it3.5_AM-06_Imid-02_CA-20

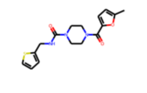

it3.5_AM-06_Imid-02_CA-22

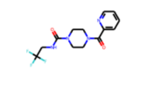

it3.5_AM-07_Imid-02_CA-20

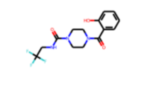

it3.5_AM-01_Imid-01_CA-07

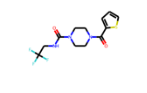

it3.5_AM-02_Imid-01_CA-07

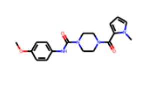

it3.5_AM-02_Imid-01_CA-27

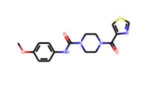

it3.5_AM-03_Imid-01_CA-25

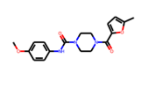

it3.5_AM-04_Imid-01_CA-07

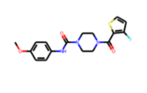

it3.5_AM-04_Imid-01_CA-09

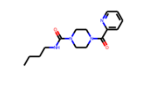

it3.5_AM-04_Imid-01_CA-16

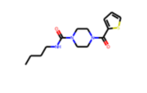

it3.5_AM-06_Imid-01_CA-14

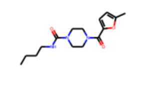

it3.5_AM-06_Imid-01_CA-23

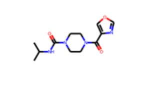

it3.5_AM-06_Imid-01_CA-25

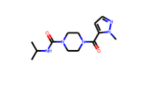

it3.5_AM-06_Imid-01_CA-30

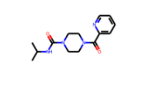

it3.5_AM-07_Imid-01_CA-07

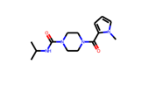

it3.5_AM-07_Imid-01_CA-16

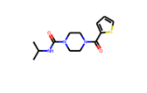

it3.5_AM-07_Imid-01_CA-25

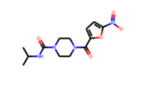

it3.5_AM-08_Imid-01_CA-02

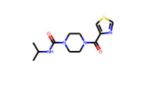

it3.5_AM-08_Imid-01_CA-04

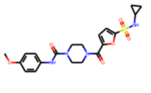

it3.5_AM-08_Imid-01_CA-07

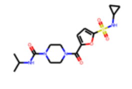

it3.5_AM-08_Imid-01_CA-14

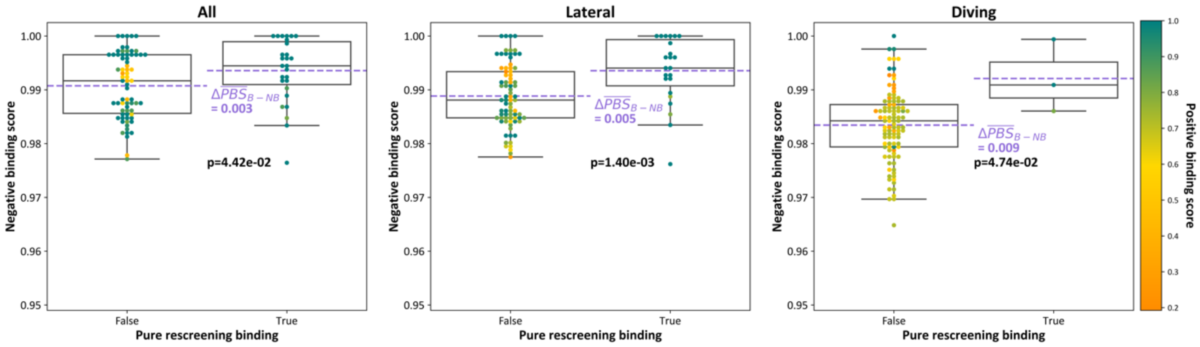

it3.5_AM-08_Imid-01_CA-16


it3.5_AM-08_Imid-01_CA-18

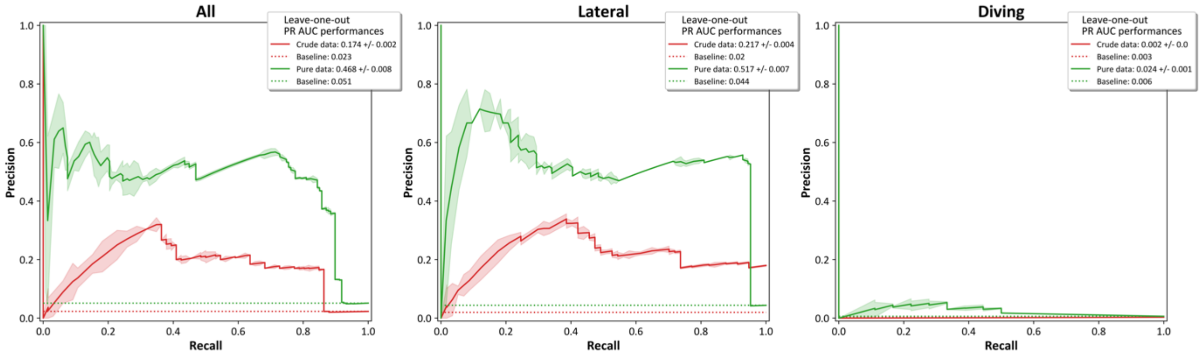

it3.5_AM-08_Imid-01_CA-23


it4.2_AM-06_Imid-01_SO2Cl-01_AM-SO2Cl-12

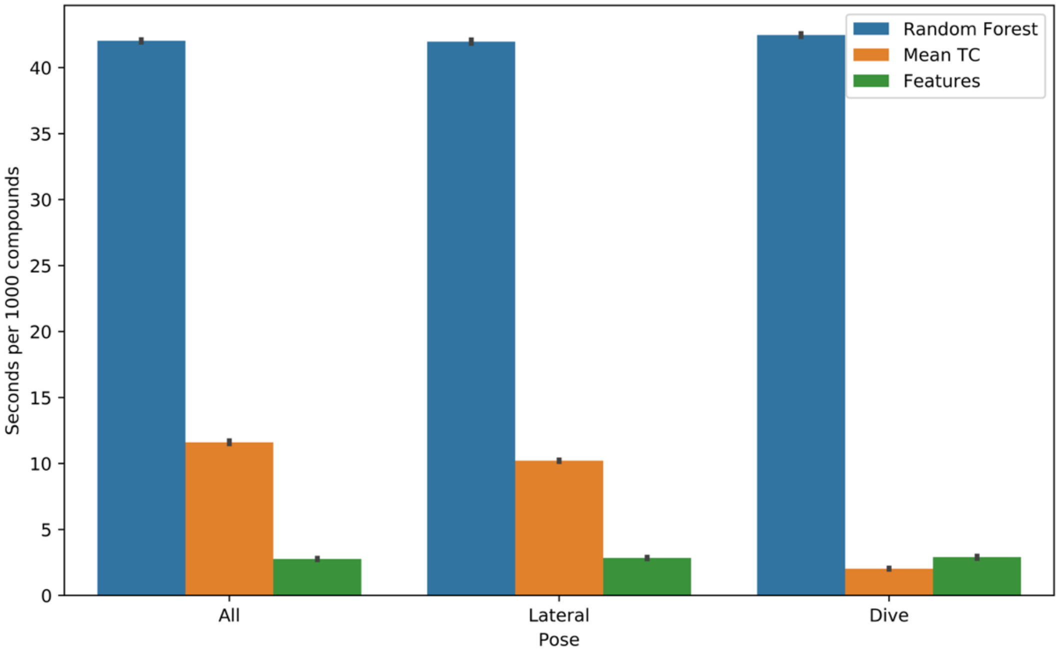

it4.2_AM-08_Imid-01_SO2Cl-01_AM-SO2Cl-12



Supplementary figure 5.6: Rescreening analysis based on the negative binding score. This shows the enrichment analysis, with True and False categories representing the pure crystallographic rescreening binding outcome on the x-axis and the distribution for each class generated from the positive binding score. Purple dotted lines show the average positive binding scores for each class and the resulting difference (ΔPBS) between binders (B) and non-binders (NB). p-value, calculated from a Mann-Whitney U test, indicate the significance of the metric in discriminating binders from non-binders. The associated positive binding score values are colour-coded for all datapoints.

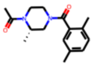


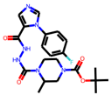
Supplementary figure 5.7: Leave-one-out validation of random forest classifier performance before and after rescreening experiment and correction of experimental false negative labels. A leave-one-out validation of the random forest classifier was applied to the dataset coming out of the crystallographic screening of crude reaction mixtures (in red) and the same dataset where binding labels of the 26 false negatives are updated. The operation was done once for each dataset. The errors showed are related to the precision and the ones reported to the overall PR-AUC.

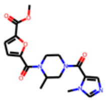


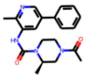
Supplementary figure 5.8: Features-based scoring is faster than random forest classification and is hit-rate independent. A speed benchmark was carried out to assess which method performs faster ligand-based virtual screening. The random forest was slower than retrieving both positive and negative binding scores for screened molecule. The mean Tanimoto coefficient was slower than the feature method for the “All” and “Lateral” data subsets as it is dependent on the number of positive. It was faster than the feature method for the “Dive” dataset has it only contains 3 binders.

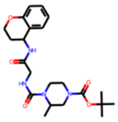
Supplementary table 5.2: Virtual screening compounds for the diving pose purchased and screened. A total of 46 compounds were purchased following the virtual screening workflow applied to the diving binders.

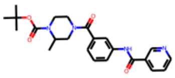

DIV0

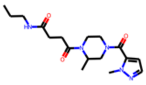

DIV1

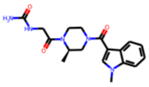

DIV2

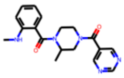

DIV3

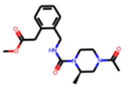

DIV4

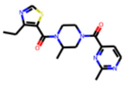

DIV5

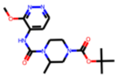

DIV6

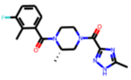

DIV7

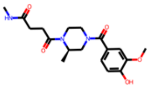

DIV8

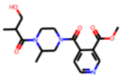

DIV9

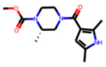

DIV10

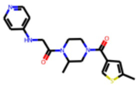

DIV11

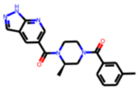

DIV12

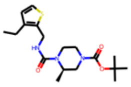

DIV13

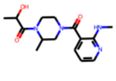

DIV15

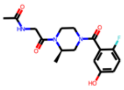

DIV16

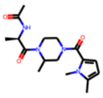

DIV17

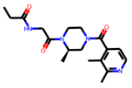

DIV18

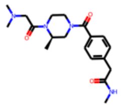

DIV19

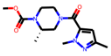

DIV20

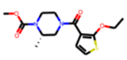

DIV21

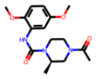

DIV22

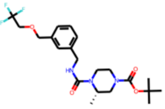

DIV24

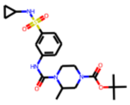

DIV25

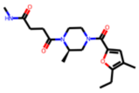

DIV26

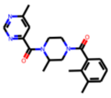

DIV27

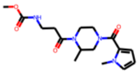

DIV28

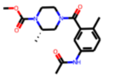

DIV29

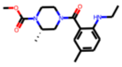

DIV30

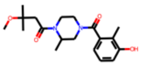

DIV31

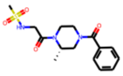

DIV34

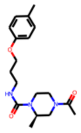

DIV35

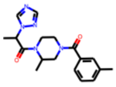

DIV36

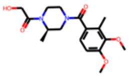

DIV37

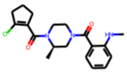

DIV38

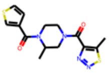

DIV39

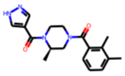

DIV40

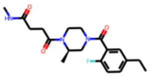

DIV41

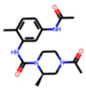

DIV42

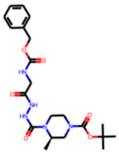

DIV43

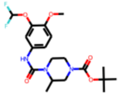

DIV44


DIV45

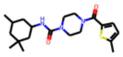

DIV46

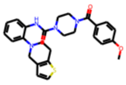

DIV47

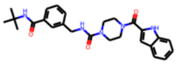

DIV48

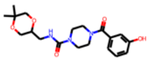

DIV49

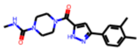
Supplementary table 5.3: Virtual screening compounds for the lateral pose purchased and screened. A total of 47 compounds were purchased following the virtual screening workflow applied to the lateral binders.

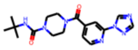

LAT0

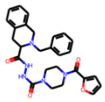

LAT1

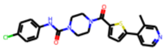

LAT2

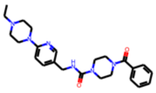

LAT3

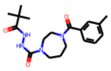

LAT4

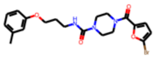

LAT5

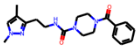

LAT6

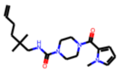

LAT7

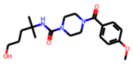

LAT8

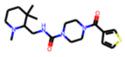

LAT9

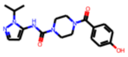

LAT10

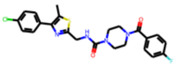

LAT11

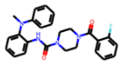

LAT12

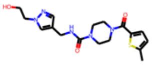

LAT13

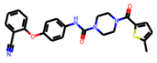

LAT14

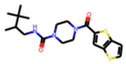

LAT15

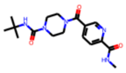

LAT18

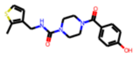

LAT19

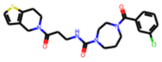

LAT20

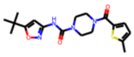

LAT21

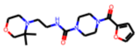

LAT22

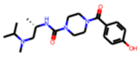

LAT23

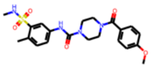

LAT24

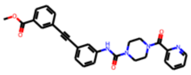

LAT25

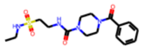

LAT26

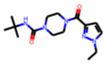

LAT27

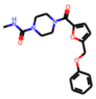

LAT28

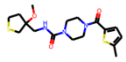

LAT29

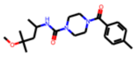

LAT30

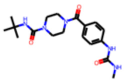

LAT31

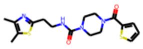

LAT32

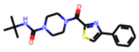

LAT33

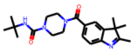

LAT34

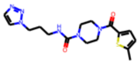

LAT35

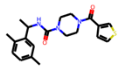

LAT36

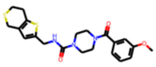

LAT37

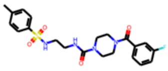

LAT38

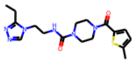

LAT39

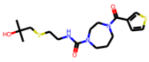

LAT41

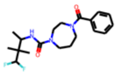

LAT42

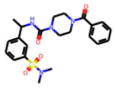

LAT43

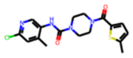

LAT44


LAT45

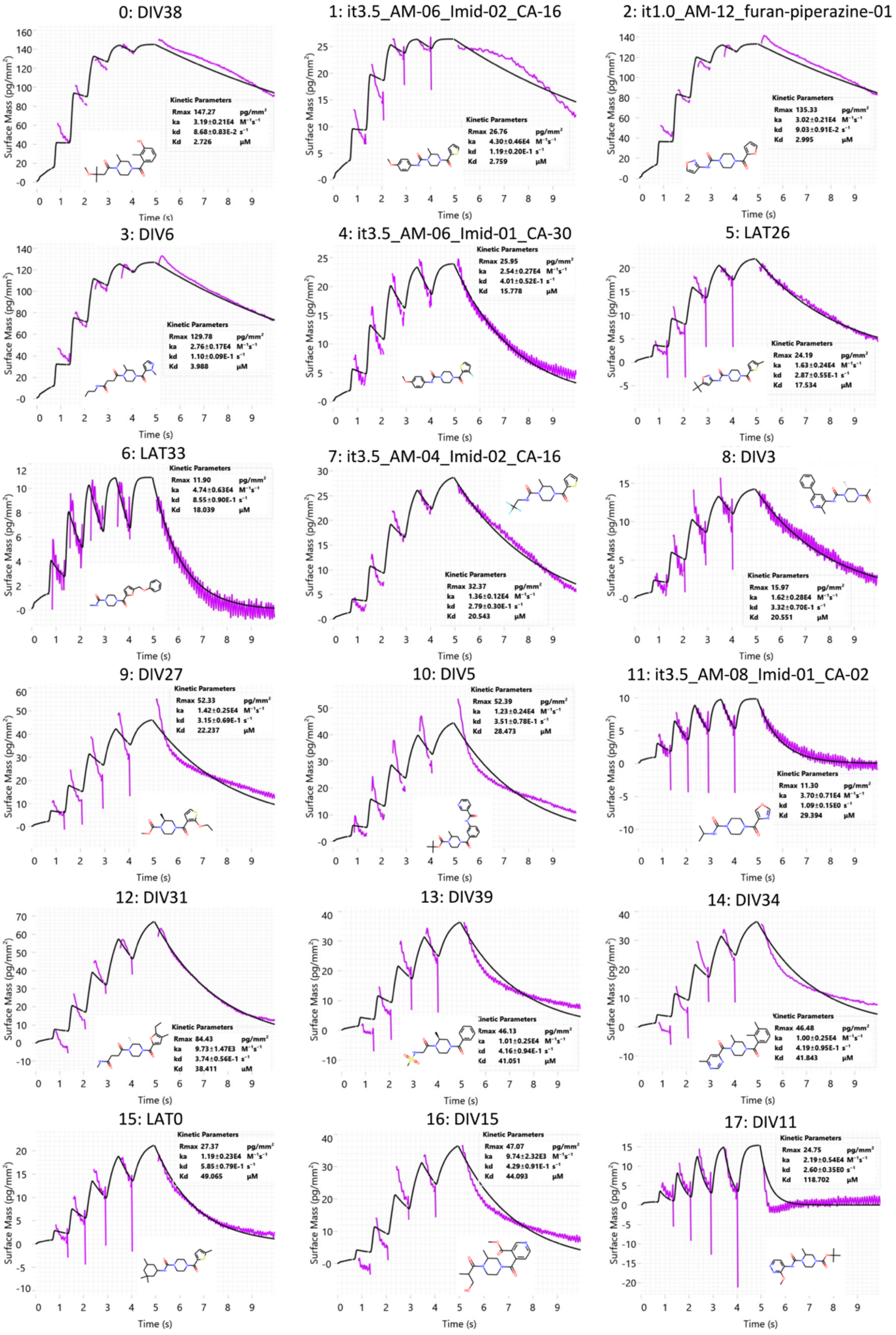

LAT46


LAT47

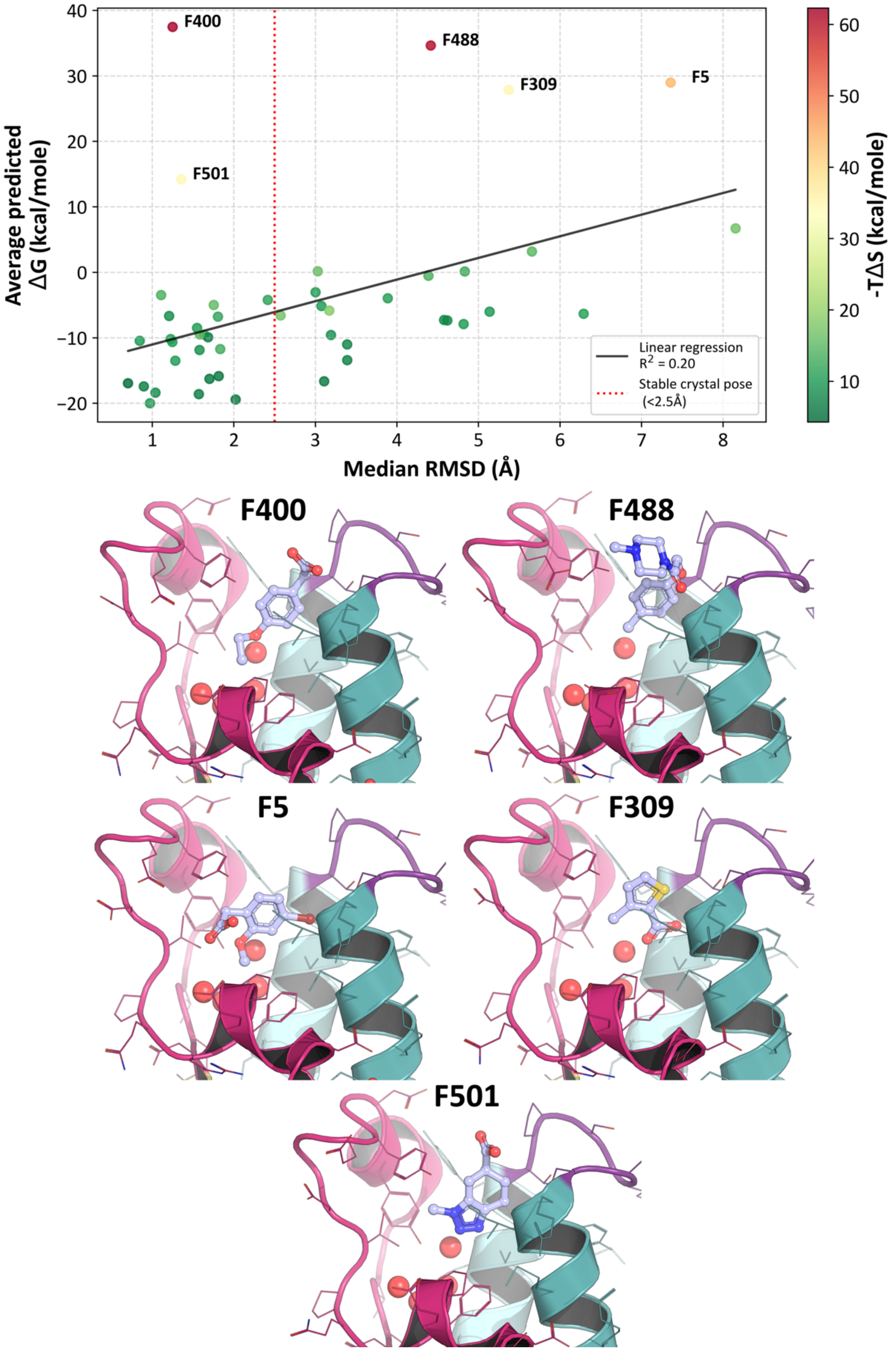

LAT48


LAT49

Supplementary table 5.4: GCI binding curves for selected binders. This table presents the binding curves obtained from GCI assays, with the black line representing the raw data and the fitted model shown in purple. The kinetic parameters, including association rate (k_a), dissociation rate (k_d), and equilibrium dissociation constant (Kd), were deduced from the fitted model, and are displayed in a separate box. The table also includes 2D molecule structures to illustrate the chemical structures of the molecules assayed. ka and kd are indicative of the association and dissociation rates with Kd, the dissociation constant, equalling to association over the dissociation rates.

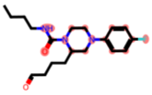


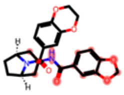
Supplementary information to Chapter 6

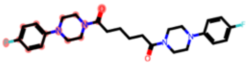


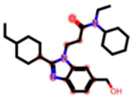
Supplementary figure 6.1: MMGBSA with Interaction Entropy free energy calculations correlated with pose stability and creates outliers in compounds with charged groups. The top panel show the median RMSD for fragment molecular dynamic simulations against corresponding calculated binding free energies. The lower panel shows the crystal structures for outliers with high Interaction entropy values.

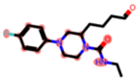
Supplementary table 6.1: 2D structures of follow-up compounds with significantly better binding free energy scores than F558. The maximum common substructure relative to the originating fragment is highlighted in pink.

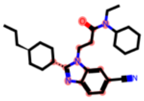

F710-Su-366

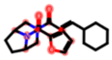

F740-Si-380

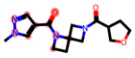

F710-Si-91

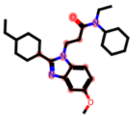

F217-Su-461

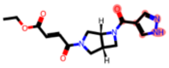

F710-Su-289

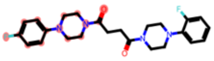

F217-Su-472

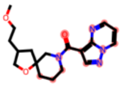

F709-Si-620

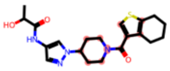

F421-Su-634

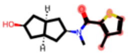

F217-Su-463

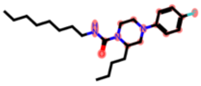

F389-Si-421

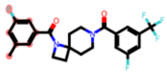

F710-Si-364

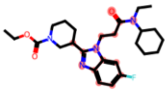

F741-Si-597

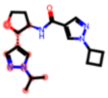

F179-Si-555

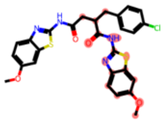

F309-Si-292

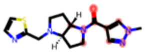

F710-Su-420

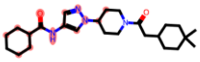

F275-Si-565

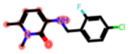

F217-Su-482

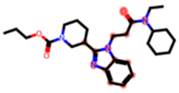

F179-Si-413

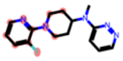

F579-Si-487

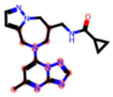

F421-Su-636

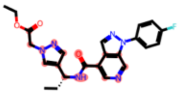

F369-Si-568

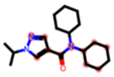

F205-Su-289

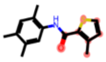

F217-Su-483

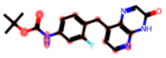

F618-Si-450

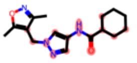

F362-Si-569

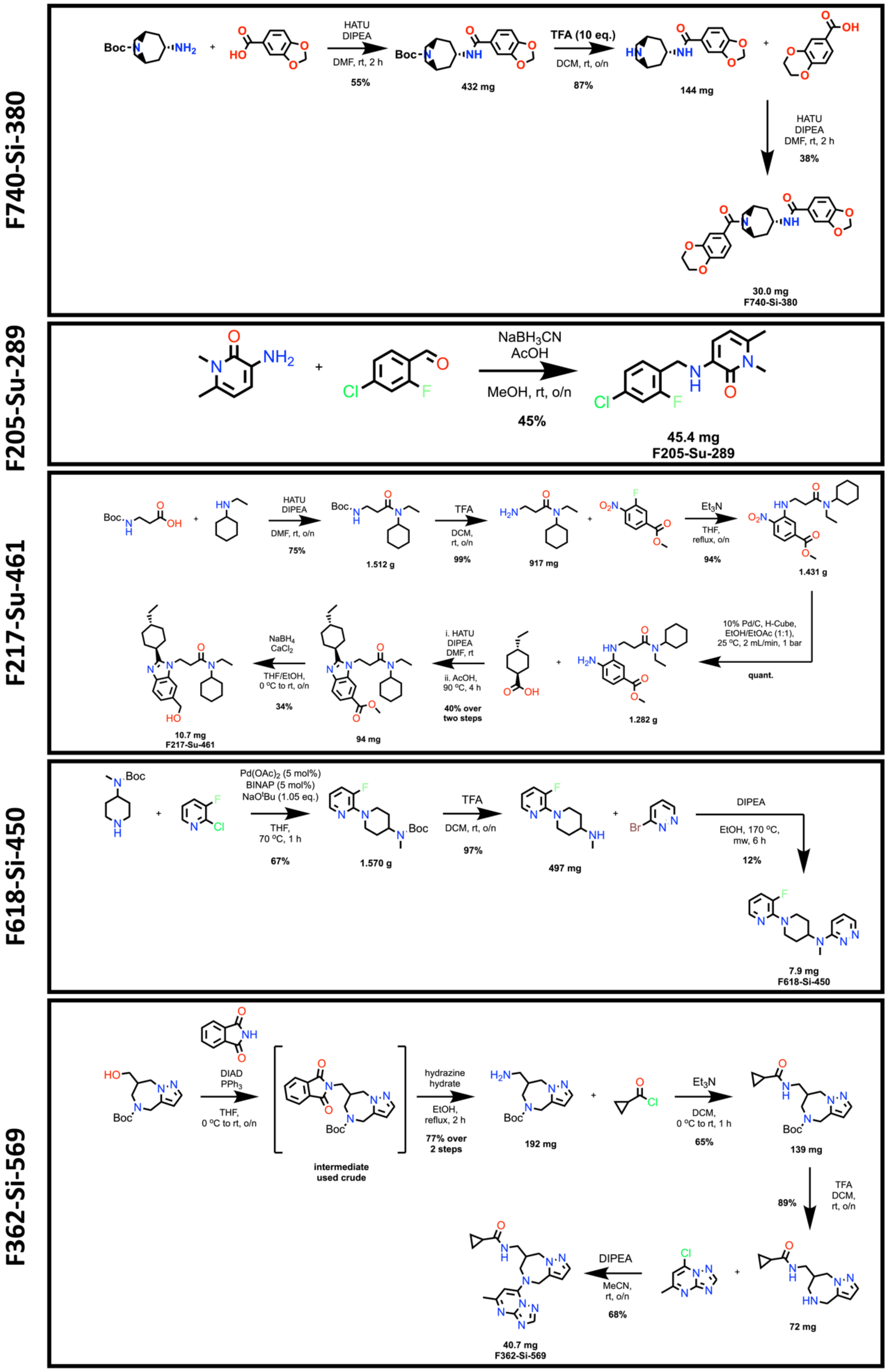

F367-Su-404


F179-Si-416


F309-Si-276

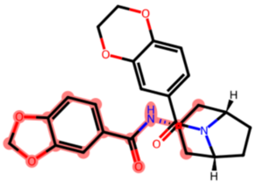

F558-Su-130


F368-Si-292



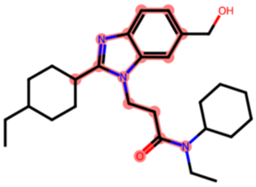
Supplementary figure 6.2: Synthetic routes for selected follow-up compounds. Molecular structures for building block and synthetic intermediates are showed along with reaction condition and yields. The synthesis and figure were made by Thomas Grimes under the supervision of Prof Paul Brennan and are showed here for the record. F309-Si-202 was directly purchased from catalogue.

Supplementary table 6.2: Co-crystallisation drop conditions for selected follow-up compounds. The chosen follow-ups were screened for co-crystallisation conditions, and all mountable crystals underwent X-ray diffraction experiments. The best diffracting crystals were analysed using molecular replacement. No ligand binding was observed; instead, DMSO occupied the binding site. The table below records selected co-crystal conditions yielding suitable electron densities. A red cross indicates no suitable crystals were identified or resolved for specific ligands.

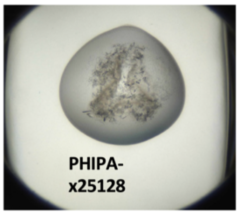


F740-Si-380

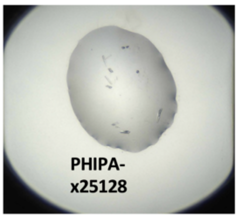
❌



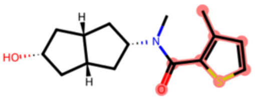
F217-Su-461



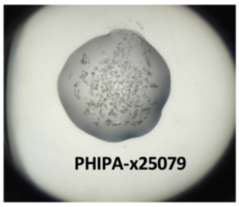
20% PEG6000, 10% EG,
0.1M MES, pH6.0,
0.1M $CaCl_2$



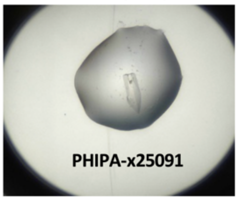
20% PEG3350, 10% EG,
0.1M bis-tris-propane, pH8.5,
0.2M Na/K phosphate



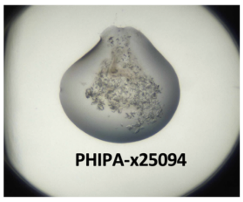
F309-Si-292



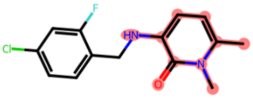
20% PEG6000, 10%EG,
0.2M NH_4Cl



20% PEG6000, 10%EG,
0.1M Tris, pH 7.5, 0.2M NH_4Cl

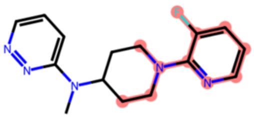


20% PEG6000, 10%EG,
0.1M Tris, pH 7.5, 0.2M $CaCl_2$

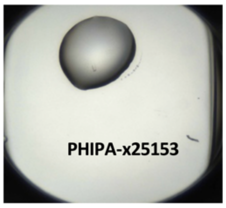


F205-Su-289

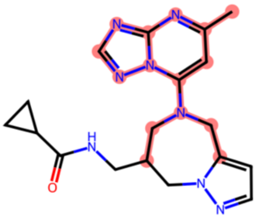
❌



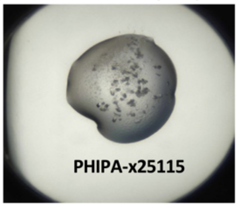
F618-Si-450



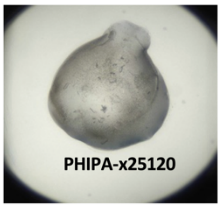
20% PEG3350, 10% EG,
0.1M bis-tris-propane,
pH 6.5, 0.02M Na/K phosphate



F362-Si-569



30% PEG1000, 0.1M PCB,
pH 8.0



20%PEG, 10% EG,
0.1M tris, pH 7.5, 0.2M $CaCl_2$