Data-driven approaches to fragment merging

Stephanie Wills

Balliol College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Trinity term 2024

Acknowledgements

First and foremost, I would like to thank my supervisors, Charlotte and Frank. Their advice and guidance helped shape this thesis and I hope to have an ounce of the passion that they do for their work. I am eternally grateful to Ruben, who was a better mentor than I could have asked for, and who showed unwavering patience in the face of many stupid questions. I am also grateful for the feedback and wisdom from Steve, Rod and Andy, and for providing their industry perspectives.

Throughout the last few years, I have been fortunate to work alongside many wonderful researchers. I am grateful to the other members of OPIG and the Small Molecules group for their camaraderie and help, my office mates past and present (including Martin, 2.20 OG), the XChem 'geeks' (with special thanks to Fragmenstein wiz Matteo, Warren, Max and Kate), the XChem experimentalists, who I was very lucky to be able to collaborate with, and Alan and Tim, for all their database-related help.

Outside of academics, I am grateful to have made some wonderful friendships in Oxford, the Botley House alumni, who helped keep me sane and made Oxford a home, offering many laughs over taco Tuesdays and weekend brunch clubs. Thank you also to my friends in the Oxford jazz community. As upsettingly talented as they are, they gave me a creative outlet when I needed a break from my computer screen.

Thank you to Vanessa, Ethny and Janet, for their ceaseless support since fate brought us together on a Beit hall corridor 11 years ago. To Gabriel, my Canadian partner-in-crime, for always believing in me. Finally, thank you to my family: Richard, for all the big brother 'life skills', and my parents, who always pushed me to do hard things when I didn't think I was good enough.

Abstract

The COVID-19 pandemic provided stark evidence of the need for robust drug development pipelines that allow the rapid discovery and advancement of therapeutics such as antivirals. Increasing the pace with which we develop diverse compounds during the early stages of the pipeline is central to addressing problems of attrition and improving the efficiency of drug discovery.

Fragment-based drug discovery (FBDD) emerged as an alternative to the more established high-throughput screening, involving the screening of much smaller molecules, typically containing less than 20 heavy atoms, that exhibit higher hit rates and are able to probe the binding site and map out possible interaction opportunities. Fragment screening via X-ray crystallography has undergone rapid technological advancements, meaning finding fragment hits is no longer the main obstacle in the FBDD cascade. However, one of the longstanding bottlenecks is making use of the structural data from these screens to optimize fragments to become larger binders with on-scale affinities.

Automated approaches that exploit all the opportunities available from the fragment screen are lacking. Commonly employed approaches include manual design, which does not provide the scale needed and is subject to the biases of the medicinal chemist, and *de novo* and *in silico* methods, which can struggle to propose molecules that we can synthesize affordably and on rapid timescales. Moreover, when a set of follow-up compounds are proposed, there is no consensus method for prioritizing which are the most promising candidates.

The work described in this thesis focuses predominantly on the problem: how can we use the results of a crystallographic fragment screen to propose a set of diverse and synthetically accessible follow-up compounds that exploit all the interaction opportunities sampled by the original fragments? In Chapter 2, '*The Fragment Network as a novel source of fragment merges*', I describe the development of a pipeline that looks for fragment merges, referring to compounds that incorporate exact substructures of crystallographic fragment hits. The pipeline includes a set of filters that prioritize merges that adopt poses that recapitulate the orientations and interactions seen in the parent fragments. In Chapter 3, '*Expanding the scope of a catalogue search by finding bioisosteric merges*', I demonstrate how this tool can be extended to maximize the chemical space explored

within commercial catalogues, extending the search to what we term ‘bioisosteric merges’. This term denotes merges that incorporate substructures that recapitulate pharmacophoric properties of the original fragments without combining exact chemical structures. In Chapter 4, ‘*Usage of catalogue search tools in active XChem campaigns*’, I outline how active antiviral discovery campaigns provided a platform for method development, and describe the challenges encountered and lessons learnt in establishing the logistics for designing and ordering follow-up compounds based on fragment screens. Chapter 5, ‘*Developing a novel scoring function for evaluating compound elaboratability*’, outlines an approach that can be used to prioritize compounds according to their ‘elaboratability’, that is, how amenable they are to further synthesis and elaboration. The method considers both chemical and geometric aspects of assessing vector and compound elaboratability, and is contrasted against a deep learning-based approach trained on a synthetic dataset designed to contain realistic elaboration scenarios.

Contents

List of Figures

List of Tables

List of abbreviations

ADMET adsorption, distribution, metabolism, excretion and toxicity

AI artificial intelligence

AViDD Antiviral Drug Discovery

BCE binary cross-entropy

CADD computed-aided drug discovery

DL deep learning

DPP dipeptidyl peptidase

DSI Diamond–SGC–iNEXT

EC$_{50}$ half-maximal effective concentration

EGNN equivariant graph neural network

EV enterovirus

FBDD fragment-based drug discovery

FEP free energy perturbation

GA genetic algorithm

GNN graph neural networks

HBA hydrogen bond acceptor

HBD hydrogen bond donor

HSP heat shock protein

HTS high-throughput screening

IC$_{50}$ half-maximal inhibitory concentration

ITC isothermal titration calorimetry

$\boldsymbol{K}_{\mathbf{D}}$ dissociation constant

LE ligand efficiency

MCS maximum common substructure

ML machine learning

Mpro main protease

MST microscale thermophoresis

NIH National Institutes of Health

NMR nuclear magnetic resonance

nsp non-structural protein

OOD out-of-distribution

PARP poly(ADP-ribose) polymerase

PDB Protein Data Bank

PLIF protein-ligand interaction fingerprint

PLIP protein-ligand interaction profiler

PPI protein-protein interactions

QSPR quantitative structure–property relationship

RF-MS RapidFire mass spectrometry

RL reinforcement learning

SAR structure–activity relationship

SBDD structure-based drug discovery

SPR surface plasmon resonance

TEP target-enabling package

TSA thermal shift assay

t-SNE t-distributed stochastic neighbor embedding

USPTO US Patent and Trademark Office

USR ultrafast shape recognition

1. Introduction

Bringing a drug to market is an expensive and time-consuming process, with the mean costs estimated to be >$800 million and requiring decades of work (Leelananda et al., 2016; Sertkaya et al., 2024). The development pipeline is hampered by high rates of attrition, with the failure rate for drugs entering clinical trials estimated to be as high as 90% (Leelananda et al., 2016). The COVID-19 pandemic provided stark evidence for the need to develop therapeutics such as antivirals on rapid timescales (Delft et al., 2023). Improving the efficiency of the early stages of the pipeline, whereby hit compounds, found to bind to a target of interest, are developed into more potent candidates, will help to accelerate this process. The use of computer-aided drug discovery (CADD) in this step has proved invaluable in reducing the time and resource requirements by leveraging the power of existing datasets (Zhang, 2011). Below, we summarize the main steps involved in developing new therapeutics, in particular, outlining the role of hit discovery and optimization.

1.1 The drug discovery pipeline

Drugs are developed when there is an unmet clinical need for a therapeutic (Hughes et al., 2011a) (an overview of this process is shown in Figure 1.1). Drugs can generally be classed as small molecules, referring to synthetic medicinal chemicals that typically have a molecular weight of <1,000 Da, and biologics, which are comparatively larger, derived from cells or biological processes, and encompass therapeutics such as monoclonal antibodies and vaccines (Makurvet, 2021). To know how to treat a disease requires an understanding of its underlying biology and the identification of a relevant drug 'target' (Hughes et al., 2011a).

Most existing small-molecule drugs target a limited selection of proteins (Overington et al., 2006), referring to the macromolecules that undertake critical functions in our bodies to keep us alive, such as enzymes, hormones and structural proteins; although, drug targets may also be non-human such as those targeted by antimicrobials. The number of proteins targeted by existing therapeutics is relatively small in relation to the theoretical number of targets, with a 2017 analysis estimating there to be 667 targets modulated by existing drugs, despite there being >20,000 genes encoded by the human genome (Santos et al., 2017).

Most recent drugs have been developed using a target-based approach, with the goal of selectively modulating a target linked to the disease (Wagner, 2016). Several approaches are employed for target identification. For example, if there is an existing small molecule or drug implicated in the disease, direct or affinity-based approaches can be used to identify the interacting proteins (Tabana et al., 2023). Gene-targeting methods, such as use of the CRISPR–Cas9 system, can be used to selectively disrupt genes, isolating those that are associated with a specific disease phenotype (Rasul et al., 2022). In more recent years, omics-based approaches have proved valuable for the discovery of new biomarkers and targets, and can help in understanding mechanisms of action. Genomics approaches, such as genome-wide association studies, can be used to spot genetic variants, such as single nucleotide polymorphisms, helping to build a risk profile for a certain disease (Paananen et al., 2019). In contrast to target-based drug discovery, phenotypic screening can also be used for developing small-molecule therapeutics, by screening compounds to identify those that are able to modulate a biological process without explicit knowledge of the target (Wagner, 2016).

Validating a putative drug target is necessary to ensure that the protein plays a critical role in disease pathophysiology and is ‘druggable’, meaning that it is possible to modulate its activity using a drug molecule, leading to potential therapeutic benefit (Hughes et al., 2011a). This can be done through the use of chemical probes, which refer to highly characterized small molecules that can be used to interrogate biological processes by binding to and influencing protein function (Workman et al., 2010; Licciardello et al., 2022). Several other criteria need to be fulfilled to ensure it is possible to develop a safe therapeutic. For example, modifying the protein’s activity should lead to a measurable and assayable response via effects on biomarkers, referring to any molecule or quantifiable characteristic associated with a biological process or disease state. Additionally,

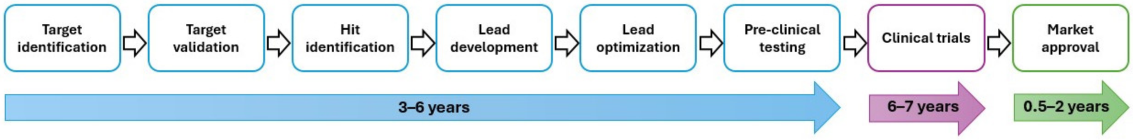


Figure 1.1: **Overview of the drug development pipeline.** The estimated timings for pre-clinical work, clinical trials (phases I–III) and processes for obtaining regulatory approval are shown.

modulation should be safe without leading to averse effects elsewhere. Fulfilling these criteria early has proved crucial for reducing attrition later on in the pipeline (Emmerich et al., 2021).

Given there is a validated target, hit compounds need to be identified; the term 'hit' is used here to denote molecules that bind to and demonstrate activity against a target. The major methods for hit discovery, high-throughput screening (HTS) and fragment-based drug discovery (FBDD), are discussed in Section 1.2. Screening involves assaying compound libraries against a target to find those that are able to bind and can provide starting points for further optimization to improve potency (Hughes et al., 2011a). It can also provide useful data regarding the accessible interactions within the binding site, subsequently informing structure-based drug discovery (SBDD, discussed in Section 1.3). While experimental screening approaches are highly optimized and well-established for hit discovery, computational methods, such as virtual screening (see Section 1.4.4.3), have shown promise in accelerating this process and reducing costs (Oliveira et al., 2023; Maia et al., 2020).

When a hit (or set of hits) is identified, it can enter a stage known as 'hit-to-lead optimization'. In contrast to hits, a lead compound refers to one that has undergone rigorous testing in several biological assays and demonstrates physiologically relevant binding affinity, typically improved by several orders of magnitude. The chemical features required for improved potency are identified using a structure–activity relationship (SAR) analysis, which correlates the addition of substituents with their effect on biochemical activity (Hoffer et al., 2018; Hevener et al., 2018).

During the lead optimization phase, not only is potency optimized, but the physicochemical profile is also monitored to ensure the compound is selective, lowering the risk of off-target effects, and demonstrates favourable pharmacokinetic properties (including absorption, distribution, metabolism, excretion and toxicity; ADMET) (Wunberg et al., 2006). This stage can be very expensive, involving several iterations of design and testing by medicinal chemists, with computational approaches sometimes used in parallel. What constitutes a lead-like drug can vary on a project-by-project basis, and may include several strategic and financial considerations (Hoffer et al., 2018). One illustration of this dilemma is Lipinski's rule-of-five, which was proposed based on observations of existing orally bioavailable drugs, and dictates that a lead-like drug should not violate more than one of the following: LogP $\leq$5, $\leq$5 hydrogen bond donors, $\leq$10 hydrogen bond acceptors and molecular weight $\leq 500\,Da$ (Lipinski et al., 1997); however, several exceptions exist for these rules (Zhang et al., 2007).

From the set of optimized leads, perhaps only a single compound series may be progressed into pre-clinical studies. These encompass *in vitro*, *in vivo*, *ex vivo* or *in silico* tests that are designed to mirror the disease system. The purpose of pre-clinical testing is to determine that the drug candidates are safe and efficacious enough to be administered to humans, after which they can enter clinical trials. Phase I trials involve a cohort of tens of individuals, with the main purpose being to collect data on safety, pharmacodynamics and pharmacokinetics, for example, tolerated dose, toxicity profile and adverse effects. Phase II trials are conducted in hundreds to thousands of patients over a longer period of months to years, and are divided into two stages: phase IIa mechanistic trials, which provide an initial evaluation of effectiveness and safety; and phase IIb efficacy trials, which collect data on drug efficacy compared with a placebo. Once the drug has been deemed to be safe and efficacious, phase III trials are initiated, involving tens of thousands of patients and conducted over a period of several years. These data are instrumental to applying for market approval, following which, if successful, the use of the drug for the public is continuously monitored in phase IV trials (Shegokar, 2020; Yuan et al., 2016). Phase IV trials are instrumental for evaluating how the drug performs in real-world settings, enabling researchers to study the effects of taking the drug over long periods of time, gathering ongoing safety data, such as the emergence of new side effects that were not previously observed (Suvarna, 2010).

1.2 Methods for early hit discovery

Hit discovery can range from screening libraries containing millions of compounds with HTS to screening fragment libraries containing perhaps only hundreds of compounds. *In silico* methods and virtual screening can also be used to reduce the experimental burden of hit discovery by prioritizing a smaller set of compounds for assaying (described in Section 1.4.4.3). In this Section, we discuss and contrast HTS against FBDD, summarizing the benefits and limitations of each.

1.2.1 High-throughput screening

HTS refers to the screening of large chemical libraries to identify biological or biochemical activity against a target or disease phenotype (Wigglesworth et al., 2015). The usage and scale of HTS began to explode in the late 1990s and 2000s, in parallel with the expansion of companies' internal screening catalogues, the growing number of potential drug targets and advances made in aspects such as assay miniaturization and automation (Mayr et al., 2009). However, a major limitation of HTS is that it typically results in low hit rates of between 0.01% and 0.14% (Zhu et al., 2013).

Much of the power of HTS depends on the quality and design of the screening library, for which there are several approaches. While library size is important, compounds are also selected for their druglikeness, physicochemical properties and absence of reactive groups. Chemical diversity may be maximized to increase the chances of finding a novel scaffold, especially when the target is not well-characterized, or libraries may be focused according to the target class or disease area. The latter approach is useful for well-characterized drug targets such as kinase enzymes, where several approved drugs already exist and related libraries are thus designed to contain analogues of these known compounds (Blay et al., 2020).

The success of HTS is also reliant on whether an assay is available that is sensitive, robust, has disease relevance and is economical enough to be performed at high-throughput. Binding-based assays focus solely on whether a compound binds the target, while biochemical assays additionally consider whether the compound exhibits the desired biological response. The response is usually competition-based, involving inhibition of the native substrate. Cell-based assays measure a specific phenotypic response, and thus are not limited to assessing the effects of a single target, instead measuring the outcome of a molecular pathway, while high-content imaging is a microscopy-based technique that uses advanced imaging analysis to extract changes in cellular features (Mayr et al., 2009; Blay et al., 2020).

While HTS has proved successful for multiple disease indications, particularly for well-characterized target classes, low hit rates and the occurrence of false positives can make its use difficult for less tractable targets that have not been extensively studied. Because of the existence of these false positives or promiscuous binders, orthogonal assays are often needed to deconvolute the results and separate artifacts from validated hits.

1.2.2 Fragment-based drug discovery

FBDD is an alternative approach for hit discovery that emerged in the late 1990s and involves screening of libraries containing low-molecular-weight compounds, usually no more than twenty heavy atoms, that can subsequently be optimized to become larger, more potent molecules (Erlanson et al., 2016).

The use of fragments presents many benefits over the molecules used in HTS. Firstly, fragments have a lower degree of molecular complexity, meaning they are more likely to bind to complementary sites on the protein and form favourable interactions without steric hindrance (Figure 1.2) (Hann et al., 2001; Leach et al., 2011). Fragments are usually designed to contain only one or two pharmacophore-binding motifs, meaning they are able to more easily probe the protein surface, resulting in higher rates (Erlanson et al., 2016). This makes them attractive for previously intractable targets, such as protein–protein interactions (PPIs), which are often flat and inaccessible to larger molecules (Murray et al., 2009). Fragments are also well-suited to detecting protein hotspots, which refer to areas of the protein that contribute disproportionately to binding free energy (Curran et al., 2020).

While the limited number of interactions means that fragments bind with weak affinity, in the high μM to low mM range, these interactions are considered to be high quality owing to the reduced chance of steric clashes and because fragment binding tends to be enthalpy-driven (Murray et al.,

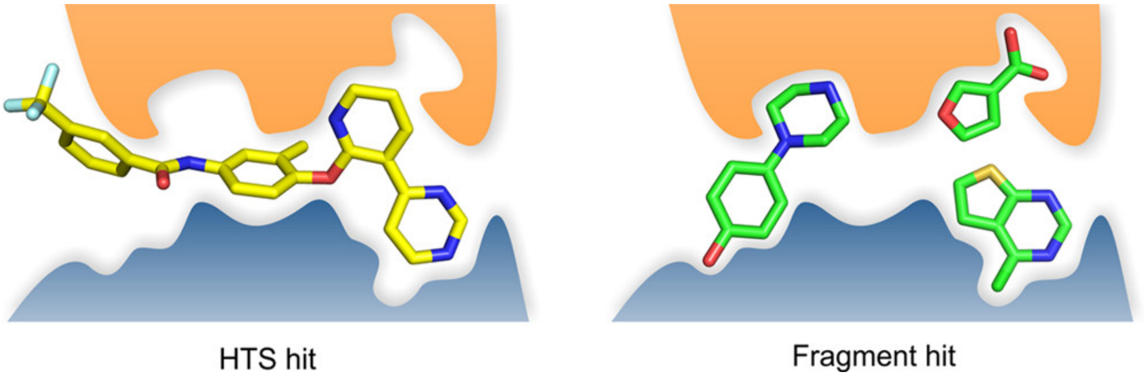


Figure 1.2: **High-throughput screening versus fragment-based drug discovery.** An example is shown of a hit observed using high-throughput screening (HTS) and with fragment screening. HTS screening involves libraries of larger compounds, typically up to 500 *Da* in molecular weight. Fragments are smaller, usually containing no more than 20 heavy atoms. Because of their small size and reduced molecular complexity, fragments are more likely to bind to complementary sites on the protein surface and face less steric hindrance. Reproduced with permission from Scott et al. (2012). Copyright 2012 American Chemical Society.

2009; Ferenczy et al., 2016; Keserű et al., 2016). To accurately assess interaction quality, the ligand efficiency (LE) metric was conceived, which normalizes the binding free energy of a ligand according to the number of heavy atoms (or an alternative measure of molecular size) (Hopkins et al., 2004). Fragments with an LE of at least 0.3 kcal mol^{-1}atom^{-1} are judged to represent promising molecular starting points (Erlanson et al., 2016).

Another key advantage of fragments is their ability to more efficiently sample chemical space. As the number of heavy atoms increases, so does the number of molecules that can be enumerated: for example, there are 10^8 estimated possible molecules that contain up to twelve heavy atoms, whereas this figure increases to at least 10^{30} for compounds with a molecular weight of 500 *Da* (around 36 heavy atoms) (Hall et al., 2014). In this way, a smaller fragment library containing only thousands of fragments samples a greater proportion of chemical space than an HTS library containing a million compounds (Murray et al., 2009).

The design of the fragment library is integral to a successful fragment screening campaign (discussed further in Section 1.3.2.2). Similar to Lipinski's rule of five (Lipinski et al., 1997), the rule of three emerged in 2003, based on retrospective observations of fragments used in a subset of screens. The rules stipulate that a fragment should not have a molecular weight greater than 300 *Da*, a cLogP of more than 3 and should contain no more than 3 hydrogen bond donors or acceptors (Jhoti et al., 2013). While the latter two rules are more subjective, size is a valid consideration as the fragment has to be small enough to enable the high hit rates integral to FBDD, but cannot be so small that the fragment binds promiscuously across the binding site. Other important factors include: high solubility, as high concentrations are required for detection; ensuring fragments do not contain reactive groups or substructures that may lead to aggregation; and that fragments have the synthetic handles to allow further elaboration (Erlanson et al., 2016; Keserű et al., 2016).

With regard to fragment optimization, a key advantage of fragments is that their small molecular size allows greater control over their properties during development. A high initial molecular weight and lipophilicity make molecules difficult to progress later on; thus, careful control over physicochemical properties can help counter these effects. Importantly, a fragment screen can provide a roadmap of the possible interactions within the binding site, and these data are central to optimizing potency, by maximizing the number of interactions made by a single ligand (typically using SBDD) (Murray et al., 2009).

1.3 Structure-based drug discovery using fragments

A key advantage of FBDD is that using small molecular starting points allows greater control over their subsequent optimization. Structural information regarding the fragment–protein binding pose can enable this process, though this is dependent on the screening method used. Common methods for fragment screening are described below, together with an overview of the workflow for the XChem facility, based at Diamond Light Source, UK, which provides the capabilities for performing high-throughput X-ray crystallographic fragment screening.

1.3.1 Methods for fragment screening

Due to the weakly binding nature of fragments, highly sensitive biophysical techniques are needed for their detection, with the most common being X-ray crystallography, nuclear magnetic resonance (NMR) and surface plasmon resonance (SPR) (Patel et al., 2014). Each method varies with regard

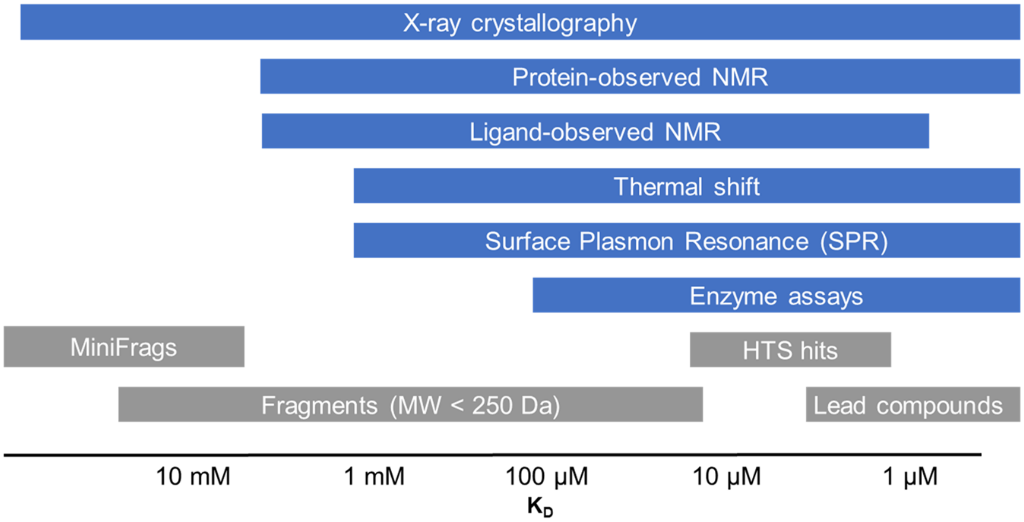


Figure 1.3: **Methods for fragment screening and the sensitivity of their detection.** Fragment screening techniques vary according to how sensitive they are to detecting ligands with different affinities (represented by the dissociation constant, K_{D}). Figure reproduced from Osborne et al. (2020). HTS, high-throughput screening; MW, molecular weight; NMR, nuclear magnetic resonance; SPR, surface plasmon resonance.

to factors such as whether they yield structural or affinity data, the amount of protein required for screening and how sensitive they are (summarized in Figure 1.3.)

SPR is a highly sensitive technique that involves immobilizing the target on a biosensor, washing the fragment over the sensor surface and recording whether fragment binding induces a change in the refractive index. The main advantages are that SPR is sensitive to weak binders, does not require large amounts of protein, allows calculation of the binding affinity (represented by the dissociation constant; K_{D}) and enables off-rate screening (Kirsch et al., 2019).

Thermal shift assays (TSAs) record changes in protein stability by determining the melting temperature, whereby the protein is heated until it unfolds, exposing a hydrophobic core to which fluorescent dye binds. Ligand binding induces a shift in the melting temperature, providing a relatively quick and easy method for fragment screening (Kirsch et al., 2019; Kranz et al., 2011).

Microscale thermophoresis (MST) measures changes in the movement of molecules in response to temperature gradients (referred to as thermophoresis) by capturing changes in fluorescence. The effect of fragment binding can be measured at different concentrations, allowing determination of the K_{D} (Kirsch et al., 2019; Kirkman et al., 2024).

Another technique, isothermal titration calorimetry (ITC), can be used to describe the thermodynamics of binding by measuring the difference in temperature between a reference cell and a protein-containing sample cell into which the ligand is injected. The results can be used to calculate the K_{D} and change in enthalpy (Kirsch et al., 2019).

The major limitation with these approaches however is that they do not result in structural information, meaning they need to be used orthogonally with another approach to enable SBDD (Kirsch et al., 2019).

NMR is a powerful technique for structural determination as it also allows determination of the binding affinity of the fragment and is sensitive to binders in the mM range (Osborne et al., 2020). Fragment binding can be detected by measuring changes in the NMR spectra for the ligand (ligand-observed NMR) or protein (protein-observed NMR), for example using chemical shifts and relaxation times. Ligand-observed NMR is more widely used as it does not require isotopic labelling of the protein, requires lower concentrations of fragment, is suitable for large proteins and provides better hit rates (Siegal et al., 2007; Harner et al., 2013; Mureddu et al., 2022).

Perhaps the gold standard for fragment screening, however, is X-ray crystallography, as it enables us to obtain high-resolution 3D structures of the fragment hits bound in the protein pocket and is sensitive enough to detect extremely weak binders. The structural information is highly valuable as it can help to map out all the possible interactions that can be made, assuming that the fragment hits adequately sample the pocket. These data can then be used to optimize the fragment hits and maximize the number of interactions made (Collins et al., 2018). Historically, X-ray crystallography has been limited by its throughput, owing to difficulties in finding the correct crystallization conditions (Collins et al., 2018; Li, 2020). However, waves of technological advancement and automation in recent years (Douangamath et al., 2021) mean that, in some settings, it is now possible to screen up to one thousand compounds within a week (Diamond Light Source., 2020). Some of the important considerations for crystallographic screening, including the method of crystallization and selection of the fragment library, are described below.

1.3.2 X-ray crystallography

1.3.2.1 Crystallization of fragment–ligand complexes

X-ray crystallography begins with the complex process of establishing the appropriate crystallization conditions for the target system. In ideal scenarios, the crystal structure needs to have a high resolution, at least 2.5Å, to allow the structure of the ligand to be accurately resolved (Patel et al., 2014). This can involve testing a matrix of different concentrations of solvents and optimizing the choice of precipitant, salt, buffer and additives. Certain conditions can ablate crystallization, such as the presence of solvents that affect fragment solubility, or those that compete with fragment binding in the active site. However, based on available data, there is no evidence of a correlation between certain conditions and whether the system will crystallize (Davies et al., 2011).

Screening can be performed by either crystal soaking or co-crystallization. The former involves soaking an existing protein crystal with solutions containing cocktails of fragments, whereby the fragments diffuse through solvent channels to reach and bind the pocket(Davies et al., 2011; Collins et al., 2018). There are challenges associated with fragment soaking, however. For example, the fragment solvent may affect stability or form interactions with the protein binding site (Patel et al., 2014); the conformation of the protein in the crystal may restrict ligand binding; or symmetry within the crystal lattice may block the binding site (Davies et al., 2011). Co-crystallization involves crystallizing the protein together with the fragment and is substantially more resource-intensive, as separate experiments are required for each ligand. The addition of small molecules in the crystallization process can result in false negatives if the protein fails to crystallize or the crystals contain unbound proteins. However, it has the benefit of circumventing some of the stability problems that may be observed with fragment soaking, and thus its use may be suitable for obtaining structures for specific ligands (Patel et al., 2014; Collins et al., 2018).

1.3.2.2 Fragment library selection

An additional important consideration in fragment screening is the choice of library. Beyond the limits on physicochemical properties described in Section 1.2.2, different fragment libraries have been optimized to fulfil specific purposes; some examples are provided below. Many of these libraries can be easily obtained from chemical suppliers such as Enamine (Enamine, 2024)

The use of halogenated fragment libraries has the benefit that halogens are well-suited to detection by X-ray crystallography owing to anomalous scattering. Additionally, halogens can form their own interactions with the protein and can easily be elaborated. An example halogenated fragment library is FragLites, where the fragments contain two hydrogen-bonding pharmacophoric features and a halogen substituent (Wood et al., 2019).

MiniFrags, described by Astex, is a library containing 81 highly soluble compounds with 'ultra-low' molecular weight, consisting of only 5–7 heavy atoms. These fragments have been designed to contain minimal pharmacophoric features, with the low molecular complexity conferring a way to exhaustively probe possible interactions on the protein and recognize hotspots (O'Reilly et al., 2019).

Fragment libraries can also be designed to contain covalent fragments, which contain electrophilic functional groups or 'warheads' that are able to covalently bond to nucleophilic amino acids (Keeley et al., 2020).

Another important consideration is 3D shape: fragments are often biased to having flat, 2D structures due to their small size and the presence of aromatic rings. While incorporating greater 3D character may increase molecular complexity and decrease hit rates, it enables us to access more chemical space and better sample the spatial complexities of the binding site. Efforts have been made to design libraries that represent more 3D characteristics, for example, by including more fragments with sp^3 hybridized carbons and aliphatic rings (Lovering et al., 2009; Hamilton et al., 2020). Downes et al. (2020) described a library containing 56 fragments that were selected by assessing the principal moment of inertia and conformational diversity.

Following from this, a common strategy in fragment library design is incorporating as much chemical and structural diversity as possible; however, analysis has shown that libraries designed in this way are not necessarily the most 'functionally diverse', whereby fragments are able to make diverse interactions across multiple protein targets. Analysis of ten protein targets found that functionally diverse sets of fragments are able to recover more information regarding the potential interaction opportunities in a fragment screen than those that are structurally diverse (Carbery et al., 2022).

A further important aspect in library design is whether fragment hits can be rapidly elaborated to identify follow-up opportunities during optimization. This requires the inclusion of synthetic handles that allow expansion with accessible chemistry. This has been referred to in the literature as 'fragment sociability', with the definition of a sociable fragment being one that can be readily elaborated or has purchasable close analogues (St. Denis et al., 2021). In a similar manner, the Diamond–SGC–iNEXT (DSI) Poised library (Cox et al., 2016) was designed to contain fragments that have been selected to allow rapid parallel chemistry through the incorporation of poised bonds, which allow the fragments to be decomposed into synthons. By purchasing analogues of these synthons, one can easily generate libraries of related compounds to bound fragment hits.

Below, we illustrate how these considerations have been applied within the XChem platform for high-throughput crystallographic screening.

1.3.2.3 XChem: achieving high-throughput crystallographic screening

As described above, progress made in automation has meant that X-ray crystallography is now a commonly used method for fragment screening and can be performed in high-throughput. XChem, based at Diamond Light Source, is a prime example of a crystallographic FBDD facility that has achieved this; since its conception in 2014, XChem's screening pipeline has been applied to a wide array of academic and industrial targets, totalling more than 300 experiments. The pipeline is highly streamlined, meaning up to 1,000 samples can be screened within a single week (Diamond Light Source., 2020; Douangamath et al., 2021; Pearce et al., 2022).

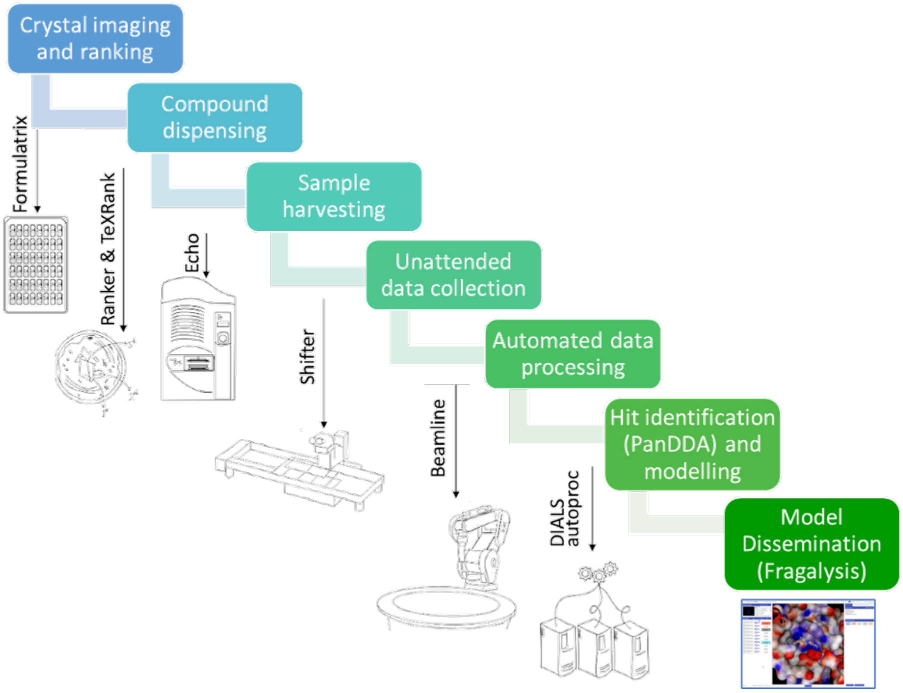

Crystal imaging and ranking
Compound dispensing
Sample harvesting
Unattended data collection
Automated data processing
Hit identification (PanDDA) and modelling
Model Dissemination (Fragalysis)
Formulatrix
Ranker & TeXRank
Echo
Shifter
Beamline
DIALS autoproc

Figure 1.4: **An overview of the XChem pipeline.** Figure reproduced from XChem (2024).

Here, we provide a high-level overview of the fragment screening process for XChem users (also shown in Figure 1.4).

- The soaking parameters (including solvent type and concentration) required for a reliable crystal system are established.
- The crystals are soaked with the fragment library.
  - The DSI-Poised library (Cox et al., 2016) is commonly used, designed to allow rapid parallel enumeration of analogues (described in Section 1.3.2.2).
  - TeXRank (Ng et al., 2014) is used to analyse the image of the crystal and select the drop dispensing location.
  - An ECHO acoustic liquid handler is used to dispense drops into the crystal (Collins et al., 2017).

  - The crystals are incubated with the fragment, allowing fragments to diffuse into the crystal and bind the target.
  - Crystals are manually harvested into a cryo-loop in liquid nitrogen with the assistance of a Shifter microscope (Wright et al., 2021).

- Crystals are screened at the I04-1 beamline, based at Diamond Light Source (Diamond Light Source., 2020).
  - Pre-screens are conducted for a subset of fragments to establish whether crystals can tolerate the compounds and soaking conditions.
  - A Pilatus detector is used to collect X-ray diffraction data unattended.

- The data are entered an automatic processing and analysis workflow, which is managed by XChemExplorer (Krojer et al., 2017).
  - Processed data are read in from Diamond software packages (Winter et al., 2013, 2018; Vonrhein et al., 2011, 2018).
  - One of the results is selected according to quality and similarity to a reference model.
  - Electron density maps are calculated using molecular replacement.
  - The PanDDA algorithm (Pearce et al., 2017b) is used to fit structures to the electron densities and identify partial-occupancy binders (Pearce et al., 2017a).
  - Models are refined and prepared for RCSB Protein Data Bank (PDB) (Berman et al., 2000) deposition (Krojer et al., 2017; Collins et al., 2018).

- Data are disseminated internally and can be used to inform follow-up design.

One initiative at XChem is using fragment screening data to design or select follow-up compounds for purchase or synthesis (a review of existing methods for fragment optimization is provided in Section 1.4). Several methods have been developed at Oxford to assist with this process (for example, those described by Cox et al. (2016), Hadfield et al. (2022), Scantlebury et al. (2023), Ferla et al. (2024)), including the work described in Chapters 2 and 3. The Fragalysis platform is a key part of this process and allows screening data and designed follow-up compounds to be made publicly available for viewing and download (found at `https://fragalysis.diamond.ac.uk/viewer/react/landing`).

A subsequent area of interest is the use of chemist-assisted robotics for the rapid in-house synthesis of follow-up compounds using readily accessible chemistry (Grosjean et al., 2024), enabling a fuller exploration of chemical space. The platform focuses, in particular, on enumerating possible elaborations of the follow-up compounds, enabling further SAR analysis. Many of the algorithmic design methods, including those described in this thesis, can thus be fed into this synthesis platform and serve as inspiration for further design iterations.

1.4 Methods for follow-up design

Following the identification of a set of crystallographic fragment hits, these data can be used to propose a set of follow-up compounds with the aim of improving, or detecting on-scale, potency. This refers to compounds that demonstrate a level of activity or inhibition that is detectable via the relevant assay.

1.4.1 Strategies for elaboration

There are three main strategies employed for elaborating fragments: fragment growing, linking and merging (Figure 1.5). Fragment growing involves adding additional atoms or groups to a fragment to access new regions within the binding pocket. Fragment linking involves designing a molecular linker to conjugate two fragments that bind to distinct, non-overlapping regions within the pocket. Fragment merging is used when fragments are in adjoining or partially overlapping space and involves designing compounds that incorporate substructural features from each (Souza Neto et al., 2020). While the latter two approaches represent efficient ways with which to maximize the number of interactions made by a single compound, fragment elaboration is the most commonly described strategy in the literature. This may be because, particularly for fragment linking, one of the challenges in elaborating compounds is ensuring that the ligand is able to retain the orientation

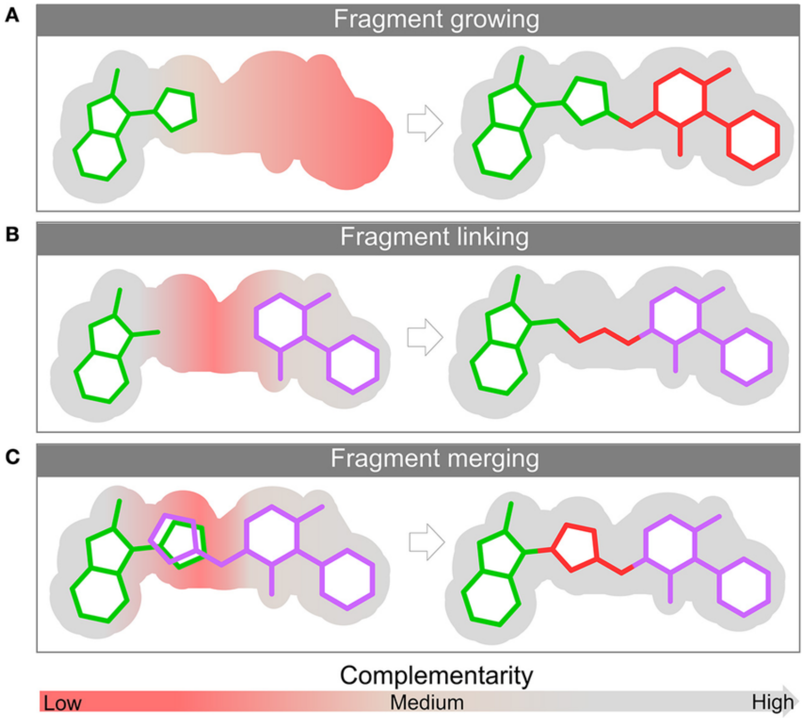


Figure 1.5: **Strategies for fragment elaboration.** The three main methods of fragment elaboration are shown: (**a**) fragment growing, whereby additional groups are added to a fragment to reach new interactions; (**b**) fragment linking, where two distinct fragments are joined via a molecular linker; and (**c**) fragment merging, whereby two partially overlapping or adjoining fragments are merged by identifying a compound that incorporates substructural features from each. Figure reproduced from Souza Neto et al. (2020).

of the parent fragment. An analysis by Malhotra et al. (2017) of 297 pairs of ligands bound to the same protein from the PDB, where one ligand represents an elaboration of the smaller ligand, found that 14% resulted in a change to the binding mode for the larger ligand. In this work, we focus predominantly on merging-based elaborations, as it is relatively under-represented in the literature but represents a powerful approach with which to increase potency.

1.4.2 Selecting fragments for elaboration

An important consideration when elaborating fragments is selecting which fragments from a screen represent the best candidates for elaboration. This can be based on simple parameters such as the B-factor, which is a measure of the atomic displacement, or occupancy, a correlated variable that refers to the proportion of time that an atom is present at a particular location. These metrics reflect the uncertainty regarding a given conformation (Cooper et al., 2011); it would follow that fragments that are more consistently observed at a particular location are more likely to maintain that conformation when elaborated. However, a fragment with lower occupancy may still provide evidence of useful interactions to be explored in compound design.

Fragment selection can also be based on specific interaction opportunities: given that, in a fragment screen, the fragments efficiently sample the possible interactions within the binding site, they can be used to map out hotspots, areas of the pocket that contribute disproportionately to the overall binding energy (Curran et al., 2020). This information can then be used to prioritize those fragments that bind in these regions for elaboration and to direct where a fragment should be elaborated by designing compounds that bridge multiple hotspots.

For fragment merging and linking, selection can be done by visual assessment of promising merging opportunities or using simple distance-based thresholds, ensuring the fragments are close enough to feasibly be combined. For example, researchers who performed a fragment screen against the SARS-CoV-2 macrodomain within the non-structural protein 3 (nsp3) identified several fragments that bound to a deep pocket in the binding site between the adenosine site and a symmetry mate, formed by crystal packing. They selected fragments for merging and linking by considering overlapping pairs that bridged the site and formed hydrogen bonds with the adenosine site, ruling out 24 fragments that formed hydrogen bonds with the symmetry partner alone (Schuller et al., 2021).

Designing elaborations based on fragment hits can be labour and computationally intensive. Thus, many of the above selection methods are used to limit the number of fragments that need to be considered. However, in scenarios where the aim is to maximize the chemical diversity of a set of designed follow-up compounds — to fully sample all the possible interaction opportunities within the binding site and increase the probability of identifying potential hits — a more exhaustive enumeration may be preferred. This requires the availability of automated, efficient approaches for follow-up design that can be scaled to designing elaborations based on the results of large-scale fragment screens. However, few approaches have been described for this purpose; this provides the motivation for much of the work described in Chapters 2 and 3.

1.4.3 *De novo* methods

1.4.3.1 Manual design

Perhaps the most widely used method for follow-up design is manual design by medicinal chemists, which has been shown to result in follow-up compounds with drastically improved potency across a range of targets (Woodhead et al. (2024) provides a summary of all fragment-to-lead papers published in 2022). For fragment merging, manual design is the most common method available. For example, Schade et al. (2020) identified highly selective inhibitors of Cathepsin S with picomolar affinities using fragment merging; and Ren et al. (2014) identified heat shock protein 90 (HSP90) inhibitors that exhibited nanomolar half-maximal inhibitory concentration (IC_{50}) values with a 200-fold increase in potency with respect to the initial fragment hits (see Miyake et al. (2019), Nikiforov et al. (2016), Hudson et al. (2012), Hughes et al. (2011b), Credille et al. (2016), Kaya et al. (2022), Prakash et al. (2021) for further examples).

The main challenges of manual design are that it requires considerable expertise and is not scalable when compared with *in silico* approaches, and thus is only used in scenarios when few select fragments are being optimized. Moreover, manual design is inherently biased to the experiences of the medicinal chemist, for example, towards compounds that can be easily synthesized by familiar chemical reactions, hence limiting the possible area of chemical space that can be explored (Brown et al., 2016). These factors also mean that, based on the results of a large-scale fragment screen in which there are tens of fragment hits, a chemist would likely only be able to consider a few merging opportunities from a very large possible chemical space.

1.4.3.2 Molecular hybridization

Molecular hybridization is a lead optimization strategy where several computational tools have been developed, which perform a related function to fragment merging. Typically, these tools rely on an overlapping substructure existing between a set of input ligands, both in terms of its chemistry and spatial organization. BREED (Pierce et al., 2004) is among the first of these techniques and relies on the existence of several protein–ligand structures for performing hybridization. This software identifies matching bonds between the set of input ligands and swaps fragments around these matched bonds to generate new hybrid structures. Matching bonds are identified by considering the bond order, distance between matched atoms and bond vectors, and whether the bonds belong to rings. The enumerated ligands can be crossed again in subsequent iterations; for example, after two iterations, novel ligands can be created that represent hybrids of four different input structures. LigMerge (Lindert et al., 2012) operates in a similar way but is receptor-independent, requiring only the 3D ligand structures to perform hybridization. LigMerge calculates the maximum common substructure (MCS) for a set of ligands and uses this as a template for alignment, maintaining MCS fragments that are geometrically equivalent (that is, the pairwise distances between atoms are approximately equal). Rotation and translation are performed to superimpose the ligands, allowing the identification of non-matching fragments that can be swapped to create new molecules. MolHyb (Wang et al., 2022a) also uses a similar approach but only takes a single ligand and protein as an input. MolHyb relies on a database of protein–ligand complexes, which combines both PDB structures and modelled complexes of ChEMBL molecules (Gaulton et al., 2012); ChEMBL molecules are placed into the binding pocket of their corresponding targets using either constrained embedding with known ligand complexes or docking.

Other methods of molecular hybridization exist that use an iterative approach, constructing compounds by assembling molecular building blocks. Fragment shuffling (Nisius et al., 2009) aligns several protein–ligand complexes and assigns scores to ligand atoms denoting their contribution to overall binding. The ligands are then fragmented and a selected seed fragment is iteratively grown using a tree search, whereby new fragments are added depending on their calculated fragment scores. Automatic Tailoring and Transplanting (AutoT&T) (Li et al., 2011, 2016) involves the optimization of a ligand using the results of a docking-based virtual screen. Matching bonds between the screening molecules and the reference molecule are identified, creating a set of fragments that can be transplanted onto the seed molecule if the fragment is predicted to have a better binding affinity.

Scaffold hopping approaches involve the replacement of the molecular core with that of a similar chemotype, for which there are various *in silico* tools. Recore (Maass et al., 2007), for example, creates a fragment database from a set of 3D structures. The molecule to be optimized is annotated with attachment points and pharmacophore atoms that are involved in forming interactions, and the database is then searched for fragments according to the pharmacophores and arrangement of their exit vectors. CReM (Polishchuk, 2020) is a tool for making ‘chemically reasonable mutations’, whereby interchangeable fragments are identified from a fragment database that can be used to perform either mutation, growing or linking of compounds; interchangeable fragments are selected according to whether they share the same chemical context. While these are not officially ‘merging’ techniques, as the cores are replaced with that from a chemical library rather than from a known ligand, these concepts can be applied to merge discovery.

Following these ideas, throughout this work, we define two types of merge: perfect (see Chapter 2) and bioisosteric merges (see Chapter 3). The term ‘bioisosteres’ refers to compounds where a functional group or fragment has been swapped with another that exhibits similar or enhanced biological activity. Thus, using this definition, we use perfect merges to refer to those that contain exact substructures of existing fragments, while bioisosteric merges refer to those that do not incorporate exact substructures of the parent fragments, but instead incorporate those with similar chemotypes or pharmacophoric properties.

1.4.3.3 Evolutionary algorithms

Other computer-aided approaches have been applied to ligand design, including the use of evolutionary algorithms. AutoGrow4 (Spiegel et al., 2020), for example, is a program for SBDD that relies on the use of a genetic algorithm (GA). Given an initial population of input compounds, optimized ligands are developed using the following GA principles: elitism, by selecting the compounds with the highest fitness score calculated using docking; mutation, by simulating chemical reactions to yield child compounds; and crossover, whereby compounds are merged based on a common substructure. This is similar to the concept of molecular hybridization, described above. GANDI (Dey et al., 2008) is another GA-based approach, whereby fragments are docked into a protein, linkers are generated between fragments using a table describing the connection vectors, and new child molecules are generated by crossing-over and mutating fragments. The fitness function consists of a weighted score that evaluates force field-based binding energy and 2D and 3D similarity to a target complex.

1.4.3.4 Generative deep learning models

De novo design using generative models has become increasingly common in the field of molecular optimization. Several deep learning (DL) models have been applied to this problem, including variational autoencoders, generative adversarial networks, recurrent neural networks, diffusion models and reinforcement learning (RL) algorithms, using different molecular representations such as SMILES (Weininger, 1988) and graphs.

In the FBDD space, the models developed thus far have largely been applied to fragment growing (for examples, see Lim et al., (2020), Green et al. (2021), Imrie et al. (2021), Hadfield et al. (2022), Arús-Pous et al. (2020), Li et al. (2020), Thomas et al. (2024), Xie et al. (2024), Powers et al. (2023)) and linking (for examples, see Tan et al. (2022), Imrie et al. (2020), Yang et al. (2020), Feng et al. (2022), Huang et al. (2022), Igashov et al. (2024)). The relative lack of *de novo* machine learning (ML) approaches for fragment merging may be due to the lack of data available for training such models and the difficulties of generating representative synthetic datasets. SILVR (Selective Iterative Latent Variable Refinement; Runcie et al. (2023)) is an example of an equivariant diffusion model that has been conditioned to generate new molecules by providing fragment hits; by providing superimposed fragments as input, the model can also be applied to fragment merging and linking problems.

1.4.3.5 Limitations

The main limitation with DL and *de novo* methods is that, while useful for allowing us to explore new areas of chemical space, many of the proposed molecules suffer from limited synthetic accessibility (Souza Neto et al., 2020) (methods for evaluating synthetic accessibility are described in Section 1.6.3). For example, hybridizing sets of ligands without regard for the synthesis schemes required to make them may result in infeasible molecules, or those that are very difficult and/or expensive to make. Efforts have been made to combat this; for example, LibINVENT (Fialková et al., 2022) uses RL to generate a focused library of elaborated compounds that share the same scaffold and molecular properties but are tailored using a set of defined chemical reactions, with the aim of enabling automated synthesis. However, techniques that efficiently exploit easily accessible areas of chemical space are still needed, as they may still offer the chemical diversity required to elucidate the SAR for a given target without the expense and time required for difficult synthesis. Section 1.4.4 below discusses the use of virtual libraries for exploring purchasable or synthesizable chemical space, and the various search methods that exist for this purpose.

1.4.4 Searching accessible chemical space

There are several compound databases commonly used for virtual screening purposes, such as ZINC (Irwin et al., 2020; Tingle et al., 2023), PubChem (Kim et al., 2023) and ChEMBL (Gaulton et al., 2012), designed to contain biologically relevant molecules. However, to find the most potent hits for a given target or to identify novel bioactive chemotypes, larger or more focused virtual libraries are often needed (Hoffmann et al., 2019). As a result, publicly accessible virtual libraries continue to grow in number and size. With this comes two main challenges: defining the chemical space of the library and having an appropriate method to search it (Coley, 2021).

1.4.4.1 The increasing size and availablility of virtual libraries

The number of theoretically possible molecules is estimated to be as high as 10^{60} for Lipinski-obeying compounds (Bohacek et al., 1996; Lyu et al., 2023; Reymond, 2015). Attempts have been made to represent the vastness of chemical space: using graph theoretical enumeration, the Reymond group designed GDB-17, a database containing 166.4 billion possible molecules containing up to 17 heavy atoms (Ruddigkeit et al., 2012). Given that HTS campaigns are only able to screen few millions of compounds, methods that allow us to navigate these vast virtual libraries are important for increasing the chemical space searched when trying to find the best candidates for a target. However, by not imposing any constraints on enumeration, it is likely that a large proportion of the molecules in theoretical libraries will not be synthesizable with our currently accessible chemistry.

Many groups have focused on creating chemical catalogues that are constrained to those molecules that are synthetically possible, using a 'make-on-demand' approach. These libraries are designed using sets of validated chemical reactions and applying these transformations to all possible combinations of available building blocks and reagents (Coley, 2021). By applying multiple reaction steps, recursive enumeration can build libraries containing billions of compounds. ZINC-22 (Tingle et al., 2023), for example, contains 37 billion 'tangible' molecules; Enamine's REadily AccessibLe (REAL) space (Grygorenko et al., 2020), generated by applying 2 or 3-component reactions to their in-house set of building blocks, contains a total of 69 billion molecules. While not all of the molecules stored in 'make-on-demand' databases are guaranteed as synthesizable, dependent on the reaction criteria used, these databases are typically quite robust; for example, the WuXI virtual database (containing 1.7 billion compounds) is estimated to have a success rate of 60–80%, while Enamine claim to have a synthesis success rate of >80% (Coley, 2021; Grygorenko et al., 2020).

1.4.4.2 Navigating chemical space

Given the size and widespread availability of these large compound libraries, approaches for efficiently navigating them are needed. It is generally infeasible to dock or evaluate every molecule in an ultra-large library, given the computational requirements, making it out of reach for even moderately sized research groups. There are, however, examples where this has been done successfully (Coley, 2021). For example, (Lyu et al., 2019) docked 99 million molecules against AmpC β-lactamase (AmpC) and 138 million against the dopamine D_4 receptor (though these figures are still several orders of magnitude smaller than ultra-large libraries); for AmpC, they optimized a hit to reach a potency of 77nM, and for the D_4 receptor, identified a novel chemotype demonstrating activity of 180pM. Additionally, VirtualFlow 2.0 (Gorgulla et al., 2023) is an example of a tool that has been developed for this purpose, which enables screening of the 69 billion molecules in Enamine REAL Space. The Adaptive Target-Guided Virtual Screening feature allows the user to screen an initial sparse representation of the library, followed by a more exhaustive screen of the compounds with molecular properties most likely to engage with the target. Moreover, VirtualFlow 2.0 scales linearly, enabling the use of up to 5.6 million virtual CPUs.

To circumvent the computational constraints, one approach is to use an active learning model, whereby the predictions of a quantitative structure–property relationship (QSPR) model are used to select a subset of compounds for screening. Bayesian optimization can be applied to balance the selection of uncertain candidates with those with positive model predictions. However, multiple iterations are typically needed to obtain good model performance, and models can suffer due to the lack of data, inability to generalize and dependence on the quality of the uncertainty prediction (Coley, 2021).

Several other methods exist for navigating libraries containing millions to billions of compounds, which have been comprehensively reviewed by Warr et al. (2022). If an initial hit exists for a target, performing an 'analogue-by-catalogue' search is a common method of identifying potential follow-up compounds, often done using substructure and similarity searching. However, these methods are often intractable for libraries containing several millions of compounds. Work has been done to develop faster search approaches to circumvent this problem (Warr et al., 2022): specific examples include the SmallWorld and Arthor tools, developed by NextMove, which are used for searching the ZINC catalogues (Irwin et al., 2020; Tingle et al., 2023; NextMove, 2024). SmallWorld enables sublinear searching using graph-edit distance and maximum common subgraphs. The database is pre-indexed according to anonymous graphs — graphs where atom identities and bond order have been removed — meaning, when performing a search using a query molecule, the database looks up the anonymous graph of the query then performs a more fine-grained search by using graph-edit modifications. Arthor, in contrast, scales linearly against database size, and allows substructure searching using SMARTS patterns. While less efficient than the former, Arthor exceeds alternative methods for substructure searching through its use of compact persistent binary molecule representations (Irwin et al., 2020).

Representing chemical space using a graph database is a relatively under-explored concept, yet there are several advantages for representing chemical space in this way. Most representations of chemical space envision it as a continuous space, where molecules are visualized as points associated with coordinates, calculated using molecular descriptors. Maggiora et al. (2014b) outline several limitations of such a representation, including the following: chemical space is finite and thus discrete rather than continuous; relationships within chemical space are dependent on the type of molecular representation used; chemical space is typically highly dimensional and usually requires some form of dimensionality reduction to be comprehensible; and continuous vector representations often have different units that require scaling.

Using a network, on the other hand, is much more conducive to depicting the discrete nature of chemical space. Networks, which consist of nodes, representing data points, connected by edges, representing relationships between nodes, are well-suited to describing relationships between molecules in chemical space. In a network where molecules are represented by nodes, similarity can easily be gauged by looking at the path distance between them. This can be done without the requirement for adding coordinates or performing dimensionality reduction. Additionally, networks can be annotated with node and edge properties, which are useful for improving the efficiency of a search (Maggiora et al., 2014b). One of the main points of variation between different chemical space networks is how edges are defined; for example, Scalfani et al. (2022) build two types of network by creating edges for pairs of molecules with similarity above a pre-defined threshold, calculated using either Tanimoto similarity (Bajusz et al., 2015) between molecular fingerprints or an MCS-based metric.

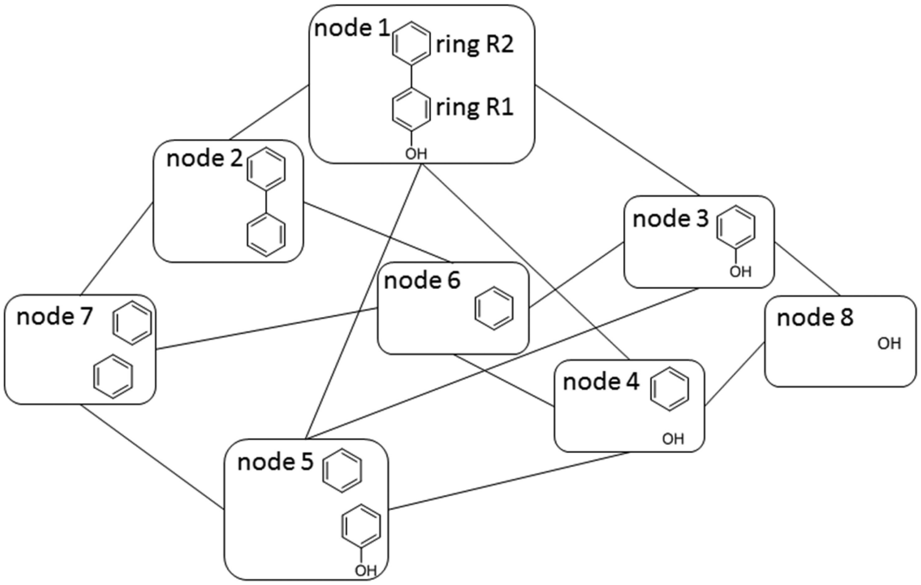


Figure 1.6: **Schematic of node enumeration using the Fragment Network.** An example is shown for how nodes, representing molecules, and edges, representing transformations, are generated for 4-hydroxy-biphenyl. Enumeration is done via the iterative removal of rings, linkers and substituents. Figure reproduced from Hall et al. 2017.

The Fragment Network, first described in a 2017 paper by Astex Pharmaceuticals (Hall et al., 2017), provides an alternative way to define relationships. It uses a graph database representation of chemical space, generated for a library of fragments, which can be used for identifying analogues to a given fragment hit. To populate the database, molecules are fragmented to generate a set of connected nodes; in this framework, nodes are generated via the iterative removal of rings, linkers and substituents (Figure 1.6). The edges joining molecule nodes represent transformations in which these groups are added or removed. By repeating this process for entire libraries of molecules, this results in a network where one can easily probe the chemical space around a molecule of interest by traversing these paths; paths involving several successive transformations can enable us to perform more complex transformations in an efficient way.

Hall et al. 2017 demonstrate the use of the Fragment Network for identifying 'medium changes' to a compound, referring to replacements of substituents and rings (representing a two-step query path), to probe chemical space around fragment hits. Using this methodology, they identified analogues against protein kinase B and hepatitis C virus protease–helicase, resulting in chemical modifications that would not have been retrievable using a substructure search alone.

One of the main advantages of this representation is that it is chemically intuitive, in that it mirrors the way that chemist's think about making molecule modifications. Furthermore, this allows retrieval of close neighbours (or analogues) within the database that, to the human eye, seem chemically similar, separated by short paths, but may not be deemed similar according to a traditional vector-based similarity metric. The benefits of this network are further explored in Chapters 2 and 3, in particular, in its application to more complex database queries by identifying follow-up merges based on fragment hits.

1.4.4.3 Virtual screening methods for identifying follow-up compounds

A well-established way of searching compound libraries is virtual screening, which computationally assesses which compounds in the library have the greatest probability of demonstrating biological activity against a target. While virtual screening can also be used as a method for early hit discovery, equivalent to HTS and fragment screening (discussed in Section 1.2), we provide an overview here to give context for how such methods can be applied to find follow-up compounds when there are known fragment hits.

Interest in virtual screening arose in the late 1990s due to the need to reduce the time and monetary costs associated with experimentally screening large numbers of molecules (Horvath, 1997; Maia et al., 2020). These *in silico* pipelines act as a funnel, reducing the number of compounds to a manageable amount that can be screened using more computationally intensive techniques (such as molecular dynamics simulations or quantum mechanical calculations) or *in vitro* assays (Vázquez et al., 2020; Oliveira et al., 2023). In addition, the properties that the molecules are being screened for do not need to be limited to bioactivity, and these techniques can be applied to other properties such as toxicity or pharmacokinetics (Oliveira et al., 2023). Virtual screening can be broadly classified into two main types: ligand-based and structure-based approaches. In this Section, we provide a brief overview of both, with a particular focus on pharmacophore-based methods.

Ligand-based virtual screening

Ligand-based approaches do not require an experimentally determined structure available for a target, and involve searching a database of molecules for those that exhibit similar properties to that of an existing reference ligand(s) (Vázquez et al., 2020; Oliveira et al., 2023). The rationale for these methods is governed by the similarity property principle, which posits that molecules with similar properties are likely to exhibit similar biological activity (Stumpfe et al., 2011).

The descriptors used to perform ligand-based screening can be classified according to their dimensionality. 1D descriptors include simple molecular properties such as molecular weight, rotatable bond count and lipophilicity. 2D descriptors encode information about the topology and connectivity of the molecule and are often represented using bit vectors (Oliveira et al., 2023). For example, MACCS key fingerprints record the presence or absence of 166 substructural features (Maggiora et al., 2014a; Durant et al., 2002); extended-connectivity fingerprints (ECFPs) encode substructure information by capturing the circular atomic environments of each atom, the size of which is controlled by the fingerprint radius (Rogers et al., 2010). 3D methods build on this by encoding information regarding the spatial properties of atoms or chemical features and molecular shape (Oliveira et al., 2023). Typically these methods require the generation of molecular conformers; however, this increases computational costs and adds an additional layer of complexity as one needs to choose the number of conformers that are needed to be representative of the conformational ensemble (Ebejer et al., 2012).

There are several advantages and disadvantages to the use of ligand-based virtual screening. In terms of the benefits, ligand-based methods can be used in absence of a target structure and perform particularly well if there is a large amount of existing experimental data. However, they may be biased to the properties of the reference molecules and not be able to account for large shifts in affinity observed for minor chemical modifications (termed activity cliffs) (Oliveira et al., 2023).

Structure-based virtual screening

Structure-based methods can be used when there is a structure available for the target protein, and involves using molecular docking to predict the binding mode of a set of compounds and scoring them by assessing the structural and chemical complementarity between the ligand the the receptor. This can be done using docking-associated or machine learning-based scoring functions (Vázquez et al., 2020). Commonly used programs include AutoDock VINA, GOLD and Glide (Trott et al., 2010; Jones et al., 1997; Friesner et al., 2004).

However, structure-based virtual screening methods are not a perfect solution and have several limitations: for example, they can struggle with generalizing to new domains and target classes; and many techniques are not able to account for protein flexibility or model the importance of water molecules in interactions (Vázquez et al., 2020; Oliveira et al., 2023). Most significantly, protein-ligand binding is a complex process and thus difficult to parameterize, meaning it can be difficult to obtain accurate results using docking algorithms alone (Maia et al., 2020), and machine learning-based docking algorithms can struggle to generate physically plausible poses (Buttenschoen et al., 2024). As a result, very large numbers of molecules need to be docked to obtain sufficient hit rates (Chen, 2015).

However, while both structure and ligand-based methods have their limitations, virtual screening still offers huge efficiency gains over large-scale *in vitro* screening, and ligand-based methods can be used to supplement docking approaches in cases where there is both a target structure available and sufficient experimentally determined affinity data (Maia et al., 2020; Oliveira et al., 2023). Below, we provide an overview of some of the pharmacophore-specific techniques available for screening.

1.4.4.4 Pharmacophore-based screening

Pharmacophore-based virtual screening involves searching for molecules that are able to replicate pharmacophoric features observed in a reference ligand or are complementary to pharmacophores observed in the target pocket. The term ‘pharmacophore’ can be defined as “the ensemble of steric and electronic features that is necessary to ensure the optimal supramolecular interactions with a specific biological target structure and to trigger (or to block) its biological response” (Wermuth et al., 1998). The pharmacophore features typically used in these search techniques include hydrogen bond donors and acceptors, hydrophobic groups, positive and negative ionizable groups and aromatic rings; additional pharmacophore features such as metal coordination and halogen bonds are also used but are less common. One of the main features of using a pharmacophore-based method is that they do not search for compounds that replicate exact substructural features of a reference ligand. This means one is able to access a larger portion of chemical space and identify potential binders with novel chemotypes, similar to scaffold hopping (Giordano et al., 2022).

Ligand-based methods for pharmacophore screening

There are several methods available for conducting pharmacophore-based screening using features that have been derived from an existing ligand. An example program that uses a traditional view of a molecule’s pharmacophore profile is Pharmit (Sunseri et al., 2016). Here, users provide as input a 3D ligand structure, from which the pharmacophores are extracted. Alternatively, a protein–ligand complex can be supplied, from which the interacting pharmacophores are derived. Using the query pharmacophores, large databases can be rapidly screened by performing an alignment of the database molecules with equivalent pharmacophore features and checking that they are within a defined tolerance radius.

Many pharmacophore and 3D shape-based descriptors and tools have been developed to facilitate easy and quick identification of structural analogues based on these properties, including vector and distribution-based representations of pharmacophoric features. The ultrafast shape recognition (USR) (Ballester et al., 2007) descriptor, for example, is a shape-based descriptor that is calculated from statistical moments using distances between reference atoms within the molecule. The USRCAT descriptor (Schreyer et al., 2012) builds on USR by considering the CREDO atom types (Schreyer et al., 2009), while ElectroShape (Armstrong et al., 2010), another USR extension, considers electrostatic properties. Other methods utilize distributions based on distances between pharmacophores, such as Ligity (Ebejer et al., 2019), which calculates distances between triangular and tetrahedral sets of three and four pharmacophoric interaction points (hotspots derived from ligands or protein–ligand complexes).

Pharmacophore-constrained molecular docking

In comparison with the methods described above, structure-based techniques provide additional context from the protein binding pocket to perform a pharmacophore search. Some of these methods rely on a structure in complex with a ligand, such as LigandScout (Wolber et al., 2005), which extracts the set of interactions based on the pharmacophores and can be used for virtual screening. Other methods construct pharmacophore models using the target structure alone. GRID, for instance, uses molecular probes to sample possible interactions in the binding site across a grid. Interaction energies can be used to construct a molecular interaction field, which can be used to identify hotspot regions of interest (Goodford, 1985; Schaller et al., 2020; Yang, 2010).

Beyond pharmacophore modelling methods, pharmacophores can be used as constraints within molecular docking (Hindle et al., 2002). In addition to their use as filters for a library to be docked (by ensuring that database ligands contain the required pharmacophores extracted from a query ligand), the identity and 3D position of the pharmacophores can be used to influence the scoring of the ligand pose; docking programs such as GOLD (Jones et al., 1997) and rDock (Ruiz-Carmona et al., 2014) can be used for this purpose. GOLD assigns higher scores to conformations whereby specific atom types are within a user-defined distance of a pharmacophoric constraint. rDock meanwhile allows the definition of a set of optional and compulsory pharmacophoric restraints; a penalty is applied for each restraint where the nearest ligand feature is beyond a specific tolerance radius (Jones et al., 1997; Ruiz-Carmona et al., 2014).

In this work, in Chapter 3, we focus on the use of pharmacophore-based methods to broaden the chemical space for follow-up compounds based on fragment hits, maximizing the possibility of finding new chemotypes and painting a fuller picture of the SAR for a target.

1.4.4.5 Catalogue methods for follow-up design

As mentioned, to rapidly progress fragment hits in drug discovery campaigns, commercial catalogues are often used as sources of follow-up compounds based on fragment hits. As these molecules are purchasable, they are guaranteed to be synthetically accessible and thus should be cheap and easy to acquire. However, there are only a handful of automated, formalized workflows described for this purpose, in particular those that exploit all possible interaction opportunities within the binding site. Here, we describe some examples where this has been explored in the literature, and how these methods offer promise for rapidly advancing the early stages of the FBDD pipeline.

Similarity and substructure searches are useful tools for identifying follow-up compounds from small-to-medium-sized libraries of compounds. Frag4Lead (Metz et al., 2021) describes a workflow used to perform an analogue-by-catalogue search within the MolPort database (MolPort, 2024) for five crystallographic fragment hits against aspartic protease endothiapepsin. This was done using three approaches: a similarity search using the MACCS fingerprint (Durant et al., 2002) (with a Tanimoto similarity threshold; Bajusz et al. (2015)); a superstructure search for molecules that contain the fragment as a substructure; and a search for compounds that are a substructure of the fragment. The superstructural analogues were docked using template-guided docking using a FlexX procedure (Rarey et al., 1996) to allow selection of a 28 compounds to be screened using crystallography and ITC. Of these, 10 of the compounds bound, the best of which resulted in a 266-fold improvement in affinity, reaching single-digit micromolar values.

The weakness of similarity and substructure searches is that the latter are often not computationally efficient, and fingerprint-based metrics can exhibit bias against smaller molecules. This bias occurs because fragments occupy fewer bits than larger molecules, and thus small changes in their structures can result in large drops in similarity according to a fingerprint-based metric. Substructure searching is not straightforward for identifying merging opportunities as this requires one to specify which pieces from each fragment to be included in the final molecule; specifying which combination and connectivity of substructures to include is a non-trivial task, and systematic enumeration of all possible combinations is likely intractable.

Andrianov et al. (2021) describe an approach in which they combined catalogue searching techniques with fragment growing and merging approaches to explore the kinase inhibitor space. Though not strictly starting from validated fragment hits, this work provides an example of a fragment merging-like approach applied to a well-characterized target class. Retrosynthesis was used to decompose kinase inhibitors, all featuring a hinge-binding core motif, into molecular building blocks that attach to the conserved core. The authors then searched commercial catalogues for analogues of the building blocks to add to the core, generating conformations using alignment based on the hinge-binding motif. Following this, they performed fragment merging by swapping building blocks between ligands with a low root mean square deviation (RMSD) between core motifs. In this way, the authors identified diverse analogues that are likely to be synthesizable using purchasable building blocks and via known chemical reactions. However, similar to the hybridization techniques described (Section 1.4.3.2), this approach is suited to ligands that share a conserved motif, which may not be applicable for all targets.

A few approaches have emerged recently that provide efficient ways of exploring make-on-demand chemical libraries. Sadybekov et al. (2022) describe a synthon-based method named V-SYNTHES; while this approach starts with virtual fragments (rather than validated crystallographic hits), it still provides a useful demonstration of the use of computational techniques to advance fragments using a make-on-demand library. In this work, the authors began with a library of 600,000 fragments, isolated from Enamine REAL space and representing common scaffolds. These fragments were docked, followed by iterative cycles whereby synthons were added to the most promising fragments and re-docked against the target. When this approach was applied to cannabinoid receptors 1 and 2, of the top 60 compounds identified, 21 demonstrated *in vitro* activity, with 6 showing inhibition at concentrations of $<10\mu$M (Sadybekov et al., 2022; Deane et al., 2022).

Meanwhile, Müller et al. (2022) detail a pipeline to progress four fragment hits against protein kinase A. This workflow involved placing fragments from Enamine REAL space (Grygorenko et al., 2020) based on the positions of the crystallographic fragments using FlexX (Rarey et al., 1996). The best scoring poses were selected, from which sub-libraries of larger molecules were enumerated using Enamine's in-house reactions. Repeating the template-guided docking and scoring of the enumerated molecules led to the selection and synthesis of 93 compounds, the best of which demonstrated a 13,500-fold gain in affinity.

In terms of expanding these approaches to identify more pharmacophorically similar follow-up compounds, there are a couple of examples that explore this. FRESCO (McCorkindale et al., 2022) takes as input an ensemble of 3D bound ligands, using unsupervised machine learning methods to learn 3D distributions of pharmacophores from a set of crystallographic fragment hits, which were in turn used to score and prospectively screen compounds from Enamine REAL space (Grygorenko et al., 2020). Moreover, a recent work by Correy et al. (2024) introduces their FrankenROCS pipeline, which uses fragment linking based on 3D shape-based search to explore the Enamine HTS collection, containing 2.1 million compounds. Based on several pairs of fragment hits against the SARS-CoV-2 nsp3 macrodomain, the Rapid Overlay of Chemical Structures (ROCS) (Hawkins et al., 2007) method is used to identify potential merges that replicate shape and pharmacophore-based properties of the fragments, identifying a follow-up compound (which did not contain a previously problematic carboxylic acid group) with an IC_{50} value of $130\mu M$. A larger-scale approach was implemented to search the Enamine REAL database, containing 22 billion molecules, using an active learning strategy, Thompson sampling (Klarich et al., 2024); Thompson sampling is applied in reagent space, enabling selection of the best-scoring building blocks using ROCS, which can be used to enumerate follow-up compounds using combinatorial expansion with Enamine's in-house reactions. From this, thirty-two compounds were selected for synthesis, leading to the identification of a compound with an IC_{50} value of $220\mu M$.

These examples show that catalogue search is an accessible yet powerful way to advance a FBDD campaign. Formalized workflows, involving *both* the identification and filtering of compounds, are integral for exploiting fragment screening data to rapidly identify compounds for further screening that can inform subsequent iterations of synthesis. A brief summary of how follow-up compounds are progressed to become more lead-like is provided below.

1.4.4.6 Optimization of follow-up compounds via the DMTA cycle

The design–make–test–analyse (DMTA) cycle describes the process by which candidate compounds are iteratively optimized and is particularly relevant in the stages immediately following the identification of promising hits or follow-up compounds.

DMTA begins with the **Design** stage, typically performed by medicinal chemists (in conjunction with computational chemistry techniques), whereby a set of compounds are proposed for synthesis. The compounds are usually selected to test a specific hypothesis, such as the chemical substructures related to improved potency or ADMET properties (Plowright et al., 2012). Computational methods may be used for prioritization, such as docking scores, machine learning models for affinity prediction or retrosynthesis predictions, ensuring the compounds can be synthesized using readily accessible chemistry (described in Section 1.6). It may be preferential to prioritize diversity at this stage, whether to boost the chances of finding a hit or to maximize the potential learnings for SAR, through the analysis of positive and negative data. Given the compounds can be successfully synthesized (**Make**), assays are conducted to yield experimental data to confirm (or reject) the hypothesis being tested and inform the next cycle of design (**Test**). Computational methods are particularly relevant, as they can be used, for example, to inform ML models that correlate molecular features with affinity, or to identify matched molecular pairs that result in activity cliffs (**Analyse**) (Plowright et al., 2012; Wesolowski et al., 2016).

Being able to perform this cycle in a timely manner is integral to improving the efficiency of drug discovery and is a major aim at XChem (see Section 1.3.2.3). Compound design campaigns often involve the purchase of commercially available (and thus synthetically accessible) compounds, or the use of robotic synthesis platforms to allow automated chemistry. These approaches allow the rapid generation of experimental data, which can then be used to inform the rapid progression of fragments to validated hits. A case study for how a crystallographic fragment screen was rapidly advanced to identify follow-up compounds against a COVID-19 target, also providing a model for how open-science collaboration can accelerate discovery, is provided below in Section 1.5.

1.5 Examples of fragment-based campaigns at XChem

Here, we provide an example case study for the design of follow-up compounds based on fragment screening results against a SARS-CoV-2 target at XChem and an overview of XChem's involvement in current antiviral campaigns as part of the AI-driven Structure-enabled Antiviral Platform (ASAP) Discovery Consortium. This case study provides context for the work performed in this thesis for performing exhaustive enumeration of follow-up compounds based on fragment merges.

1.5.1 The COVID Moonshot project: a collaborative campaign for inhibitors against SARS-CoV-2

The onset of the COVID-19 pandemic and the urgent need to develop antivirals let to the conception of the COVID Moonshot project (Figure 1.7) (Delft et al., 2021; Boby et al., 2023). Co-founded by XChem, the Moonshot was an open-science collaboration that advanced the results of a fragment screen against the SARS-CoV-2 main protease (Mpro; described in Section 2.3.2) to a potent lead compound. The project forged a new approach to open science for drug discovery, which relied on establishing an infrastructure for sharing data and scientific collaboration to drive the rapid progression of antiviral compounds, in aid of pandemic preparedness.

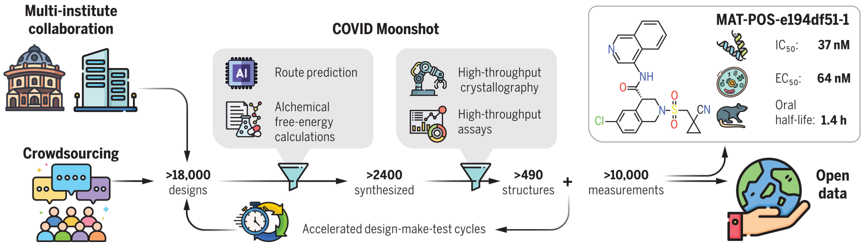


Figure 1.7: **An overview of the COVID Moonshot project.** Figure reproduced from Boby et al. (2023). EC$_{50}$, half-maximal effective concentration; IC$_{50}$, half-maximal inhibitory concentration.

The collaboration involved crowdsourcing designs from the community and working with institutions (both academic and industrial) that were able to provide synthesis, assaying, logistics and other capabilities. Following the fragment screen, the data were made available in March 2020 via an online platform, and scientists were able to submit follow-up designs based on these data (none of the fragments showed IC_{50} values of below $100\mu M$).

All data were immediately disclosed publicly, meaning the Moonshot bypassed any intellectual property-driven delays. Approximately 2,000 designs were uploaded within the first week from a variety of sources, including hypothesis-driven, computational and docking-based approaches. Retrospective analysis found that many designs appeared to use fragment merging and linking-like approaches based on overlapping fragments in the active site.

From the submissions, compounds were selected for synthesis (performed by Enamine) by expert medicinal chemists, assisted by computational tools. Alchemical free energy calculations, in particular, were found to be useful for identifying transformations that led to improved potency, such as swapping from a pyran group to a piperidine sulfonamide. The selected compounds were evaluated using further crystallographic screening and biochemical assays; these data were also deposited online to inform further design.

This collaboration led to the discovery of the first credible hit with an IC_{50} of 37nM (Boby et al., 2023). While the compound required further optimization in terms of its pharmacodynamic and pharmacokinetic properties (to improve bioavailability), this project provided a template for how collaboration can help to accelerate the development of new therapeutics. In particular, this work demonstrated how large-scale enumeration of potential fragment merges during follow-up design for screening campaigns could be a powerful technique. Chapters 2 and 3 describe pipelines that aim to contribute towards this goal.

1.5.2 ASAP: accelerating antiviral design for pandemic preparedness

The success of the Moonshot project led to the launch of the ASAP Consortium (Figure 1.8), one of the National Institutes of Health (NIH)-funded Antiviral Drug Discovery (AViDD) Centers for Pathogens of Pandemic Concern. Built on the template of the COVID Moonshot Project, ASAP focuses on using X-ray crystallographic fragment screens, coupled with artificial intelligence (AI) and computational-based methods, to accelerate the development of chemical leads against multiple viral targets, including coronaviruses, flaviviruses and picornaviruses.

XChem is a driver in the creation of target-enabling packages (TEPs), which create a foundation for rapid medicinal chemistry cycles to develop promising lead candidates later in the pipeline. The TEPs include, for a given target, a set of robust biochemical and biophysical assays, an exhaustive set of protein–ligand structures from the structural biology core and the design of an initial set of follow-up compounds (Figure 1.8).

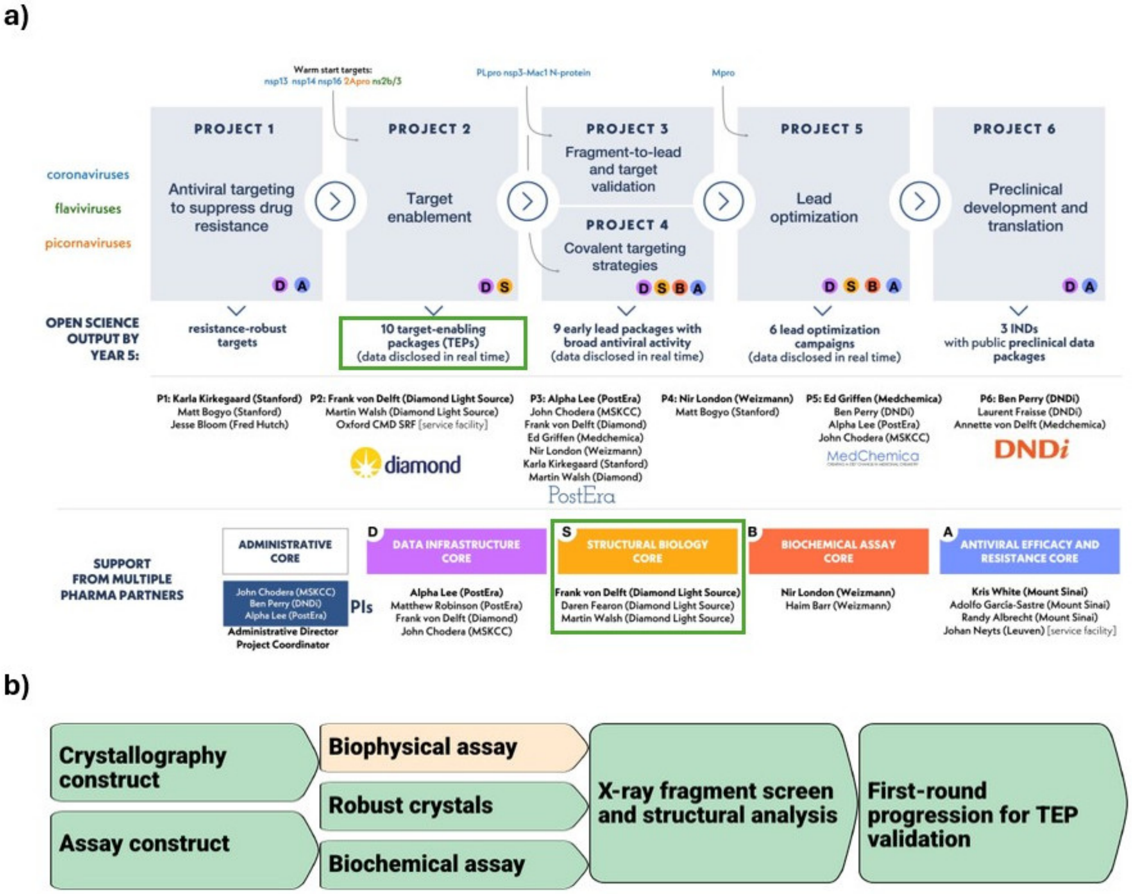

a)
Warm start targets:
nsp13 nsp14 nsp16 2Apro ns2b/3
PLpro nsp3-Mac1 N-protein
Mpro
coronaviruses
flaviviruses
picornaviruses
PROJECT 1
Antiviral targeting to suppress drug resistance
PROJECT 2
Target enablement
PROJECT 3
Fragment-to-lead and target validation
PROJECT 4
Covalent targeting strategies
PROJECT 5
Lead optimization
PROJECT 6
Preclinical development and translation
D A
D S
D S B A
D S B A
D A
OPEN SCIENCE OUTPUT BY YEAR 5:
resistance-robust targets
10 target-enabling packages (TEPs) (data disclosed in real time)
9 early lead packages with broad antiviral activity (data disclosed in real time)
6 lead optimization campaigns (data disclosed in real time)
3 INDs with public preclinical data packages
P1: Karla Kirkegaard (Stanford)
Matt Bogyo (Stanford)
Jesse Bloom (Fred Hutch)
P2: Frank von Delft (Diamond Light Source)
Martin Walsh (Diamond Light Source)
Oxford CMD SRF [service facility]
diamond
P3: Alpha Lee (PostEra)
John Chodera (MSKCC)
Frank von Delft (Diamond)
Ed Griffen (Medchemica)
Nir London (Weizmann)
Karla Kirkegaard (Stanford)
Martin Walsh (Diamond)
PostEra
P4: Nir London (Weizmann)
Matt Bogyo (Stanford)
P5: Ed Griffen (Medchemica)
Ben Perry (DNDi)
Alpha Lee (PostEra)
John Chodera (MSKCC)
MedChemica
P6: Ben Perry (DNDi)
Laurent Fraisse (DNDi)
Annette von Delft (Medchemica)
DNDi
SUPPORT FROM MULTIPLE PHARMA PARTNERS
ADMINISTRATIVE CORE
John Chodera (MSKCC)
Ben Perry (DNDi)
Alpha Lee (PostEra)
PIs
Administrative Director
Project Coordinator
D DATA INFRASTRUCTURE CORE
Alpha Lee (PostEra)
Matthew Robinson (PostEra)
Frank von Delft (Diamond)
John Chodera (MSKCC)
S STRUCTURAL BIOLOGY CORE
Frank von Delft (Diamond Light Source)
Daren Fearon (Diamond Light Source)
Martin Walsh (Diamond Light Source)
B BIOCHEMICAL ASSAY CORE
Nir London (Weizmann)
Haim Barr (Weizmann)
A ANTIVIRAL EFFICACY AND RESISTANCE CORE
Kris White (Mount Sinai)
Adolfo García-Sastre (Mount Sinai)
Randy Albrecht (Mount Sinai)
Johan Neyts (Leuven) [service facility]
b)
Crystallography construct
Assay construct
Biophysical assay
Robust crystals
Biochemical assay
X-ray fragment screen and structural analysis
First-round progression for TEP validation

Figure 1.8: **An overview of the ASAP project and target-enabled packages.** (**a**) A summary of the ASAP project pipeline is shown. Green boxes indicate components of the pipeline for which XChem is a key driver: the structural biology core, which produces an exhaustive set of protein–ligand crystal structures, necessary for elucidating protein hotspots and key interactions; and the creation of target-enabled packages (TEPs) during Project 2. (**b**) The different components involved in TEPs are shown. TEPs are data packages that provide the foundation for subsequent structure-based drug discovery and medicinal chemistry efforts. Figures reproduced from ASAP Discovery Consortium (2024).

Chapter 4 contains descriptions of my contributions to active ASAP campaigns against two enteroviral targets, involving collaborative efforts to design, select and assay purchasable follow-up fragment merges and elaborations.

1.6 Methods for prioritizing fragments and follow-up compounds

During the drug discovery process, it is repeatedly necessary to prioritize a subset of compounds from a larger selection due to a set of budget or resource constraints. This can happen at various stages of the pipeline: for example, when creating an initial library for screening, or in subsequent stages, when a list of compounds needs be narrowed down to select those to be purchased or synthesized. The evaluations are based on multiple factors, such as whether compounds fulfil a set of desired chemical properties, their synthetic accessibility, presence of reactive groups or predicted affinity based on scoring functions used in docking programs. Here, we describe some of the main areas of consideration for compound prioritization, with a particular focus on scoring follow-up compounds based on fragment hits.

1.6.1 Conservation of binding mode

One potential aim when designing follow-up compounds is ensuring that they will be able to recapitulate the pose and interactions made by the original fragments; this can be assessed by checking the overlap between the molecules, with a common metric being the RMSD, calculated based on the distances between corresponding atoms. The challenge of using RMSD is that this requires definition of the MCS between the fragment and larger ligand, which is not always evident, particularly if the follow-up compound represents a bioisostere or if there are multiple possible MCS mappings within either compound.

An alternative is to use a ‘shape and colour’ score; these metrics consider both the shape, defined as a 3D density, and chemical feature overlap between the elaborated compound and the fragment, providing a coarser-grained view that can be applied to compounds without an exact MCS match. Malhotra et al. (2017) used this method to check whether the binding mode was conserved between elaborations and their original ligands, applying a cut-off of between 0.4 and 0.55 (the score is scaled between 0 and 1). Importantly, this metric is asymmetric, ensuring that the larger ligand fully subsumes the volume of the smaller ligand, and not the other way round. Implementations of this metric using RDKit (Landrum, 2024) have been described since, such as the SuCOS (Leung et al., 2019) and $\mathrm{SC}_{\mathrm{RDKIT}}$ scores (Imrie et al., 2020).

Beyond considering only the structure of the ligand, an additional method to assess a docked ligand is to analyse the predicted interactions made with the protein using protein–ligand interaction fingerprints (PLIFs). PLIFs are useful as they enable us to capture 3D structural information in a binary vector, which can be used to assess whether the ligand has the potential to recapitulate the interactions observed within the original fragment or, additionally, the capacity to make new interactions not previously observed. There are several different programs available for this purpose (for examples, see Bouysset et al. (2021), Adasme et al. (2021), Jubb et al. (2017), Da et al. (2014)), which vary according to how interacting pharmacophores are identified (for example, using SMARTS patterns), the set of geometric constraints used as criteria, and whether the fingerprints operate on an atomic or residue level. These PLIFs can be used to probe the SAR for a set of compounds and aid optimization by SBDD.

1.6.2 Scoring functions for docked poses

One key method in SBDD is predicting binding affinity based on the docked protein–ligand complex. Classical scoring functions are used by docking software to predict the bound ligand pose, together with affinity. There are four main classes of scoring functions: force field-based, empirical, knowledge-based and machine learning-based scoring functions (Meli et al., 2022). One of the main limitations of traditional scoring functions is that, to be executed in a reasonable timeframe, they rely on approximations of the complex process of molecular recognition, leading to inaccuracies in their predictions (Velasquez-López et al., 2022). Several ML and DL-based methods, using graph neural networks (GNNs), have been developed (for a review, see Meli et al. (2022)), utilizing a wide array of architectures and descriptors, which have been shown to outperform classical scoring functions (Ghislat et al., 2021). However, recent work has shown that commonly used benchmarks allow ML-based scoring functions to exploit data biases, which has led to inflated performance;

baseline models trained to learn protein or ligand bias alone were competitive with state-of-the-art scoring functions, demonstrating that the models are not learning the underlying physics of the interactions (Boyles et al., 2019; Durant et al., 2023; Harren et al., 2024).

1.6.3 Synthetic accessibility

One important method of prioritization is to apply a synthetic accessibility score, which estimates how difficult it is to synthesize a given compound, and can be calculated in a number of ways. This is particularly useful for *de novo* methods of molecule generation or when searching theoretical compound libraries.

The first attempt to capture synthetic accessibility using a metric was the SAScore, proposed by Ertl et al. (2009), This score was devised based on the analysis of molecules present in PubChem (Kim et al., 2023), and comprises two parts: the 'fragmentScore', which compares substructures present in the molecule with their frequency of occurrence in the PubChem database; and the 'complexityPenalty', which applies a penalty for complex features such as macrocycles, stereocentres, ring systems and molecular size. This score is scaled between 1 and 10 and was validated by comparing the ranking obtained using the SAscore for a collection of 40 molecules against manual ranking by several medicinal chemists.

Several other metrics have been described since. The RAScore (Thakkar et al., 2021) is a machine learning classifier, which uses a feed-forward neural network that was trained to predict the outcomes of a retrosynthetic analysis using AiZynthFinder (Genheden et al., 2020); the output is a binary score, designed to mimic whether AiZynthFinder is able to solve for a synthetic route. The SCScore (Coley et al., 2018) was trained on Reaxys reaction data and is based on the assumption that the products of a reaction are more complex than the reactants. SYBA (Voršilák et al., 2020) is a Bernoulli naïve classifier model that aims to distinguish easy-to-synthesize from hard-to-synthesize molecules; easy-to-synthesize molecules were sourced from the ZINC database (Irwin et al., 2020), while hard-to-synthesize molecules were designed by Nonpher (Voršilák et al., 2017), which performs small chemical perturbations on starting molecules to generate greater complexity.

However, limitations of using such simple approximations have been described, for example, for

SAScore and SCScore. As these metrics do not consider the availability of reactants, the scores can be overestimated for complex molecules that can be synthesized from commercially available building blocks. Also, these scores may struggle to capture differences between structurally similar compounds that have vastly different synthetic routes (Makara et al., 2021; Gao et al., 2020b; Sanchez-Garcia et al., 2023).

1.6.4 Compound elaboratability

An under-explored but relevant concept in follow-up compound prioritization is that of 'elaboratability', which we define here as a compound's tractability for further synthesis given chemical and geometric considerations. To our knowledge, there are few studies in the literature that directly explore the concept of 'elaboratability'; however, one example is provided by Preston (2023), who developed a simple scoring function that determines if a molecule has the potential to undergo further reactions using comparisons against United States Patent and Trademark Office (USPTO) reaction data, a dataset comprising patent-based reactions (Schneider et al., 2016). For a given query molecule, the function determines if topologically equivalent reaction centre atoms exist within the dataset, and a score is generated based on how frequently an equivalent atom is observed and the diversity of reaction classes it participates in.

This is similar to the concept of 'fragment sociability' (St. Denis et al., 2021), which was introduced in Section 1.3.2.2, referring to whether fragments have synthetic handles and available purchasable close analogues to enable rapid elaboration. However, these methods have been applied to fragments in 2D, considering only the presence of elaboratable chemical features, rather than whether their position and the geometry of the pocket enable further growth. Developing methods that consider both the chemistry and the geometry of a ligand in evaluating its elaboratability is explored further in Chapter 5.

1.7 Project aims

While the throughput of X-ray crystallographic screening has dramatically improved, the optimization of crystallographic fragment hits still remains a bottleneck. Automated workflows for identifying and prioritizing the most promising follow-up compounds are needed that are able to efficiently search accessible chemical space, circumventing costly problems of low synthetic accessibility. Fragment merging techniques represent powerful methods for maximizing the number of interactions made by a single compound, and hence are a key focus of this work.

In this thesis, I will describe workflows developed for identifying fragment merges, using a chemically intuitive graph database representation of chemical space. The initial pipeline aims to identify perfect merges that are directly informed by the fragment hits, by incorporating exact substructures from both fragments (Chapter 2). The goal is to perform this task in a computationally efficient manner, circumventing the limitations associated with similarity and substructure searches, and providing a method that can be used for the exhaustive enumeration of merges based on all compatible pairs of fragments derived from a fragment screen. A procedure to expand the search space for fragment merges and increase chemical diversity among follow-up compounds will be introduced, by searching for what we term 'bioisosteric fragment merges' (Chapter 3).

Finally, we aim to describe a novel approach to prioritizing follow-up compounds by evaluating compound 'elaboratability', which aims to rank compounds and growth vectors according to their tractability for further synthesis to identify the most productive starting points for subsequent optimization (Chapter 5).

1.8 Thesis outline

The remainder of this thesis is organized as follows:

- **Chapter 2**, '*The Fragment Network as a novel source of fragment merges*', describes the initial pipeline for using the Fragment Network, a graph database representation of chemical space, for identifying purchasable fragment merges. The database searches for perfect merges, which incorporate exact substructures from the parent fragments. A filtering pipeline is implemented, which prioritizes compounds that have the potential to recapitulate the pose of the original fragments. This method of searching is contrasted against a more classical similarity search using molecular fingerprints, aiming to provide a way to improve the productivity of a catalogue search. Two retrospective case studies are provided, with the goal of evaluating whether the database search is able to recover known experimental binders.

- **Chapter 3**, '*Expanding the scope of a catalogue search by finding bioisosteric merges*' describes an updated version of the initial pipeline, which provides several feature enhancements, such as ensuring the substructures incorporated are spatially compatible and responsible for interactions with the protein. Importantly, this pipeline aims to expand the chemical space searched, by searching for bioisosteric merges, which recapitulate pharmacophoric properties of the original fragments. The purpose is to improve the possible hit rate of the catalogue search and enable initial exploration of the SAR in an automated manner. Comparisons against a pharmacophore-constrained docking workflow are made to show that the pipeline provides efficiency benefits that enable it to be used at scale.

- **Chapter 4**, '*Usage of catalogue search tools in active XChem campaigns*', summarizes work done to deploy the described pipelines in active campaigns as part of the ASAP project for two antiviral targets. These campaigns provide a platform for method development and aim to establish a workflow and logistics for advancing fragments in the XChem pipeline. Several design rounds are described, together with a description of some of the challenges encountered and lessons learned.

- **Chapter 5**, '*Developing a novel scoring function for evaluating compound elaboratability*', introduces a novel scoring function for evaluating compound elaboratability, which can be applied both to individual growth vectors, distinguishing which of them represent promising growth opportunities, and for ranking entire molecules based on their tractability for further synthesis. These methods are designed to consider whether compounds contain synthetic handles to enable elaboration and whether their spatial organization within the protein pocket enables growth and potential access to new interactions. A chemistry-informed and GNN approach are compared, and protocols for further validating these tools are outlined as goals for future work.

- **Chapter 6**, '*Conclusions and future work*', summarizes the main findings of this thesis and outlines possible avenues for future work.

2. The Fragment Network as a novel source of fragment merges

2.1 Chapter motivation

This Chapter describes a pipeline that I developed to provide a fully automated tool to perform catalogue-based fragment merging given a set of crystallographic fragment hits. The tool is based on the Fragment Network, a graph database approach for exploring chemical space, which was originally described in 2017 (Hall et al., 2017) by Astex Pharmaceuticals and re-implemented at XChem with the code made publicly accessible at `https://github.com/InformaticsMatters/fragmentor`. The work in this Chapter is largely described in a 2023 publication in the *Journal of Cheminformatics and Modelling* (Wills et al., 2023) and was carried out by myself unless stated.

2.2 Introduction

A key step in the fragment-based drug discovery (FBDD) pipeline is the development of fragment hits to become larger, more potent binders (Lamoree et al., 2017; Davis et al., 2017). As discussed in Section 1.3, while the throughput of X-ray crystallography for fragment screening has dramatically improved owing to advances in synchrotrons, automation and algorithms (Douangamath et al., 2021), a challenge remains in how to efficiently exploit the information contained in crystal complexes for the rapid structure-guided progression of fragments.

As described in Section 1.4, fragment merging is a powerful approach for improving potency, yet it remains poorly served by *in silico* approaches, and the published successful fragment merges have largely been manually designed (examples are provided in Section 1.4.3.1). Not only is this not transferable, but it does not allow the full exploitation of even modest numbers of hits, relying instead on the experience and biases of the medicinal chemist. Only *in silico* approaches will be able to more exhaustively sample the relevant chemical space, though the main challenge lies in recapitulating the orientation and interactions of the parent fragments. Furthermore, the main limitation of *de novo* approaches is that they tend to propose molecules with poor synthetic accessibility and thus are either impossible or difficult and expensive to make (Souza Neto et al., 2020).

Following from this, a key challenge in progressing fragments into an efficient drug discovery campaign is finding readily available compounds to allow elucidation of the SAR for a target. Commercial catalogues are generally used for this purpose as the compounds are guaranteed to be synthetically feasible and are cheap and easy to acquire. There are few efficient, fully automated workflows for the computational design of follow-up compounds using crystallographic fragment screens (discussed in Section 1.4.4.5). A traditional way to search these catalogues is using substructure and similarity searching (Metz et al., 2021), yet these methods can be computationally intensive and fingerprint-based similarity metrics can exhibit bias against fragments. Moreover, though such methods clearly hold great potential, many of these approaches are focused on finding analogues of single fragments and are thus more comparable with fragment growing strategies (Metz et al., 2021; Müller et al., 2022). The work described in this Chapter focuses on fragment merging and linking-like approaches, as incorporating features from multiple fragments represents an efficient way to maximize the number of interactions made by a single compound.

In this work, we demonstrate the use of a graph database as a method for finding fragment merges in commercial catalogues. As mentioned in Section 1.4.3.2, we differentiate between types of merges according to whether they maintain exact substructures of parent fragments (perfect merges), or whether they incorporate similar chemotypes or pharmacophoric properties (bioisosteric merges), the latter of which is similar to scaffold hopping techniques. In this work, as a starting point we focus on the identification of perfect merges to provide proof of concept for our approach (bioisosteric merging is explored further in Chapter 3).

To find fragment merges, we used an implementation of the Fragment Network (introduced in Section 1.4.4.2), a graph database approach that was first described by Astex Pharmaceuticals (Hall et al., 2017), which uses the RDKit cheminformatics toolbox (Landrum, 2024) and was loaded with 120 million commercial catalogue compounds. Using this database, we developed a pipeline to search for potential fragment merges, followed by several 2D and 3D filters that prioritize candidates by the likelihood of maintaining the binding pose and interactions of the parent fragments. This approach provides benefit beyond classical hybridization methods, as it guarantees synthetic accessibility by searching within catalogue space and can be applied to situations where there is no matching overlapping substructure between fragments (by linking together distinct substructures).

Our Fragment Network searches yield complementary results to classical similarity searches using molecular fingerprints, as each technique identifies filtered compounds for pairs of fragments for which the other technique fails, suggesting that the two techniques should be used in parallel. Specifically, the Fragment Network naturally lends itself to both identifying pure merges and limiting searches to local chemical space, leading to greater search efficiency. It is thus well-suited to fully exploiting all crystallographic data through large-scale enumeration of all possible pairs of fragment hits. Its suitability for fragment progression was confirmed in two retrospective analyses; comparison against experimental data from the COVID Moonshot (The COVID Moonshot Consortium et al., 2020) identified a known inhibitor and close analogues to other known inhibitors with IC_{50} values within the low-to-medium micromolar range. A known binder and close analogues to known inhibitors were also identified using experimental data against *Mycobacterium tuberculosis* transcriptional repressor protein EthR (Nikiforov et al., 2016).

2.3 Materials and methods

2.3.1 The Fragment Network database

The version of the Fragment Network used for this work was compiled in March 2022 and contains XChem fragment libraries and compounds from the Enamine, MolPort and Chemspace commercial catalogues, totalling >120 million compounds. The selection of compounds was conducted by previous members of XChem and was aimed at maximizing chemical diversity while ensuring compounds were closely related enough to ensure compounds could be connected within the database (explained further below). The code to generate the database is publicly available at `https://github.com/InformaticsMatters/fragmentor` and was written by a previous member of XChem, Anthony Bradley (and is maintained by InformaticsMatters).

The network was implemented as described in the original paper (Hall et al., 2017); generation of the network involves the decomposition of molecules — represented as nodes — into rings, linkers and substituents, with the iterative removal of these groups resulting in the enumeration of connected nodes. The network is populated with purchasable compounds that are connected via edges, which represent transformations between nodes. Transformations refer to the addition or removal of substructures, rather than reaction-based transformations. A schematic demonstrating the types of transformation that can be made is shown in Figure 2.1. The corresponding metadata for nodes and edges, describing features such as the substructure being removed or added when traversing an edge, or, optionally, molecular properties of the molecule node, can allow the tailoring of search queries. The XChem implementation uses neo4j (neo4j, 2022) as the graph database platform and queries are written in Cypher; this language can enable the creation of sophisticated search queries whereby a user can specify parameters such as the number of hops involved in the query (the number of hops refers to the path distance between nodes in the network), the type of hop to be used in the query path (that is, a contraction or expansion), the substructure(s) involved in the transformation and the filtering of results based on node properties (for example, heavy atom count).

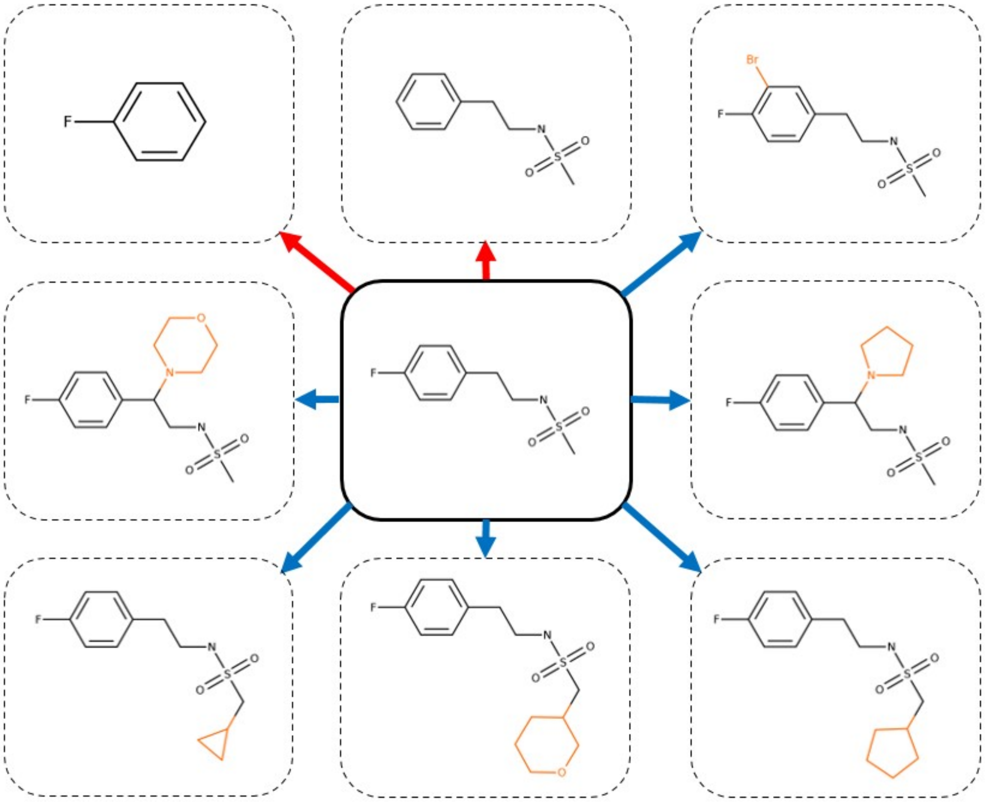


Figure 2.1: **Transformations between nodes in the Fragment Network.** Edges in the Fragment Network denote transformations in which a contraction (red arrows) or expansion (blue arrows) can be made, whereby a ring, linker or substituent is lost or gained. Example transformations are shown for a hit against non-structural protein 13 (nsp13; fragment x0276_0B). The substructure lost or gained (highlighted in orange) during the transformation is recorded in the edge label.

Table 2.1: Summary of the fragment screens used as test data

Target	Type of binding site	Number of fragments	Number of pairs for merging*
DPP11	Allosteric	11	55
PARP14	Active	13	75
nsp13	Active	9	35
Mpro	Active	19	134

*Value reflects the number of pairs after removing fragment pairs that are highly similar. DPP11, dipeptidyl peptidase 11; PARP14, poly(ADP-ribose) 14; Mpro, main protease; nsp13, non-structural protein 13.

2.3.2 XChem datasets

The four test cases are publicly available crystallographic fragment screens from the XChem facility, summarized in Table 2.1 and Appendix Table A.1. These data are available to download from the Fragalysis platform (`https://fragalysis.diamond.ac.uk/viewer/react/landing`). The specific experiments are described below.

Porphyromonas ginigivalis plays a causative role in the development of periodontal disease (Ohara-Nemoto et al., 2011). The dipeptidyl peptidase (DPP) enzymes are central to the energy metabolism of the bacterium (Ohara-Nemoto et al., 2011). Fragment screening found hits that bound to two sites: the active site, to which two fragments bound, and to a potential allosteric site. While the mode of action has yet to be established, 11 fragments were found to bind to the allosteric site (Figure 2.2a); the opportunity for inhibitor development means these 11 hits were selected for merging.

Human poly(ADP-ribose) polymerase 14 (PARP14) is part of a family of proteins involved in post-translational modification. Shared between the proteins is a highly conserved catalytic domain that binds to NAD+ and transfers negatively charged ADP-ribose to the target protein to be modified (Wahlberg et al., 2012). This study focuses on 13 fragment hits found to bind to the catalytic site (Figure 2.2b).

A helicase protein, SARS-CoV-2 non-structural protein 13 (nsp13) forms part of the replication

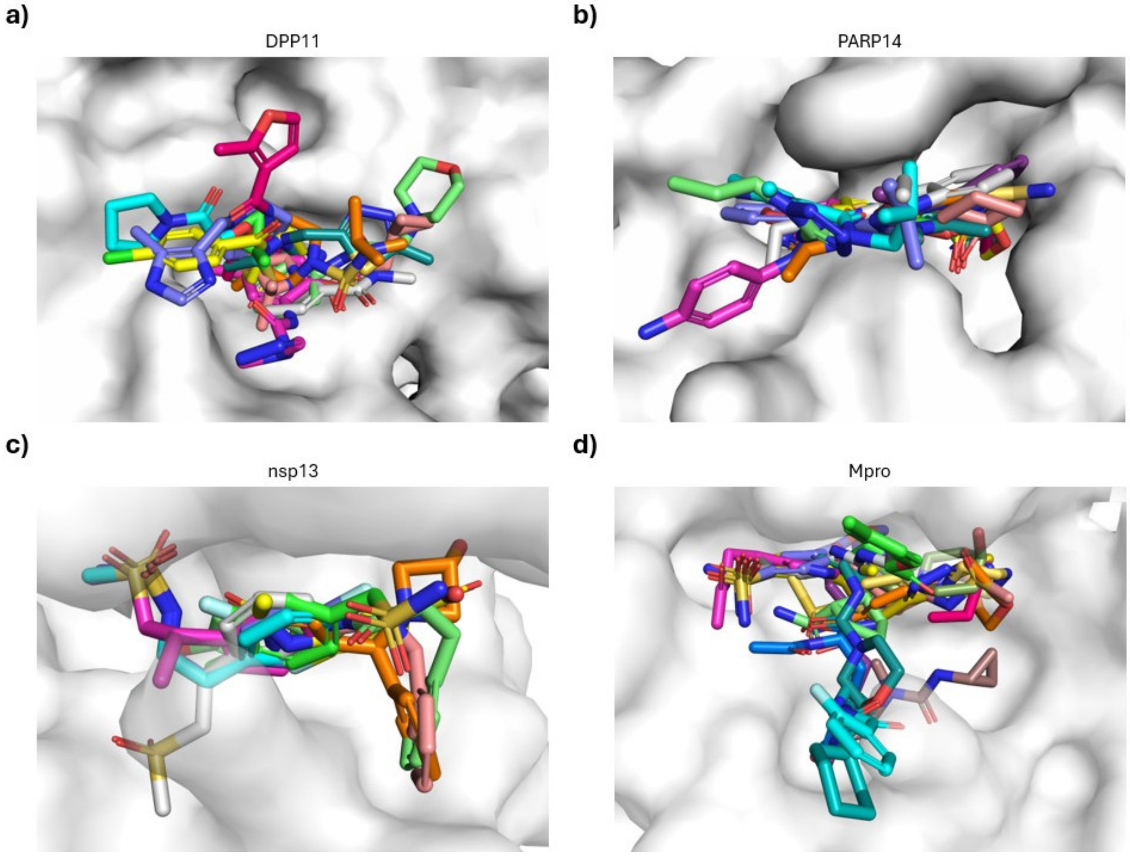


Figure 2.2: **Crystallographic fragment hits used for analysis.** The binding sites and crystallographic fragment hits used for testing the merging pipelines are shown for (**a**) *Porphyromonas gingivalis* dipeptidyl peptidase 11 (DPP11), (**b**) human poly(ADP-ribose) polymerase 14, (PARP14) (**c**) SARS-CoV-2 non-structural protein 13 (nsp13) and (**d**) SARS-CoV-2 main protease (Mpro). Data can be freely downloaded from `https://fragalysis.diamond.ac.uk/viewer/react/landing`. The fragment codes are provided in Appendix Table A.1.

and transcription machinery of SARS-CoV-2, together with 15 other non-structural proteins. The function of nsp13 is to catalyse the unwinding of genetic material using energy from nucleotide triphosphate hydrolysis. Two sites were found to have good ligandability in a fragment screen; for the purposes of this work, we focus on developing merges of fragments that bind to the nucleotide site, which is positioned between the 1A and 2A helicase subdomains (Newman et al., 2021). Nine overlapping fragment hits that offer good merging opportunities were chosen for testing the pipeline (Figure 2.2c).

The SARS-CoV-2 main protease (Mpro) is involved in processing polyproteins pp1a and pp1b of SARS-CoV-2, which are necessary for replication and transcription. Mpro is a homodimer of two polypeptides, protomers A and B. Each protomer consists of three domains and the substrate-binding site is positioned in a cleft between domains I and II, which comprise antiparallel β-barrel structures. The active site consists of multiple subsites, P1, P1′, P2, P3, P4 and P5 (Jin et al., 2020). In total, >1,250 fragments were soaked during screening, in which 71 fragments were found to populate the active site (Boby et al., 2023). In this work we focus on designing merges for 19 of these fragments (Figure 2.2d); an additional 5 fragments were used for an independent case study described below.

2.3.3 Computational workflow: database search

The overall pipeline for querying the database and filtering compounds is illustrated in Figure 2.3. For all compatible pairs of fragments, we query the Fragment Network as described in the following section. We define two fragments as compatible if they are close in space and are not highly similar to each other. Only pairs for which the distance between the closest pair of atoms is <5Å are considered. The chosen limit allows the identification of both merges and linkers; this distance threshold can be modified according to which type of elaborated compound to favour. We remove fragments that represent close analogues that bind in an equivalent position to ensure that hybrid molecules are substantially different from the parents. To do so, the MCS between the two fragments is calculated; if up to three heavy atoms in either fragment are not included in the MCS, the RMSD is calculated between the MCS atoms, and fragment pairs are removed, by default, if the RMSD is <2Å (these parameters can be modified by the user). These pairs typically do not represent useful merges as the resulting compounds do not incorporate unique substructures from each fragment. The total numbers of pairs to undergo querying are shown in Table 2.1.

2.3.3.1 Search by Fragment Network query

The Fragment Network query aims to identify close neighbours of one fragment, the seed fragment, that incorporate a substructure of the other fragment in the pair (all Fragment Network queries were written by myself). First, all possible substructures of one of the fragments in the pair are enumerated (this is done by traversing down all edges away from the fragment node that denote contractions and extracting the substructures that are removed). Substructures are filtered for those that contain at least three carbon atoms and do not exist in the other fragment in an overlapping position (considered as >50% overlap in volume). This ensures that the substructure makes a substantial and unique contribution to the final merge. Queries identify all nodes within a specified number of hops away from the seed fragment; the query then attempts to make an additional expansion hop from these nodes in which one of the substructures from the other fragment is incorporated (thus resulting in compounds that contain substructures of both original fragments).

To avoid redundancy in queries, for a single fragment, the substructures for all possible paired fragments are enumerated and pooled together (as the same substructure could be present in multiple fragments). The Fragment Network querying process is asymmetric (that is, there are two sets of queries for each pair of fragments, with each fragment undergoing expansion). The number of optional hops is a tuneable parameter; increasing the number of hops will result in a deeper search but will lead to an exponential increase in the time required and will result in molecules less similar to the original seed molecule. In this work, we limit the number of hops to a maximum of two.

All retrieved molecules are filtered for those that contain at least 15 heavy atoms to ensure the compounds represent true merges between the fragments. Query paths (that is, the path taken between the seed fragment node and the merge compounds) are limited to those where the node before expansion contains at least six carbon atoms and is not equivalent to the substructure used in the expansion (this prevents the query from retrieving all nodes connected to the substructure or

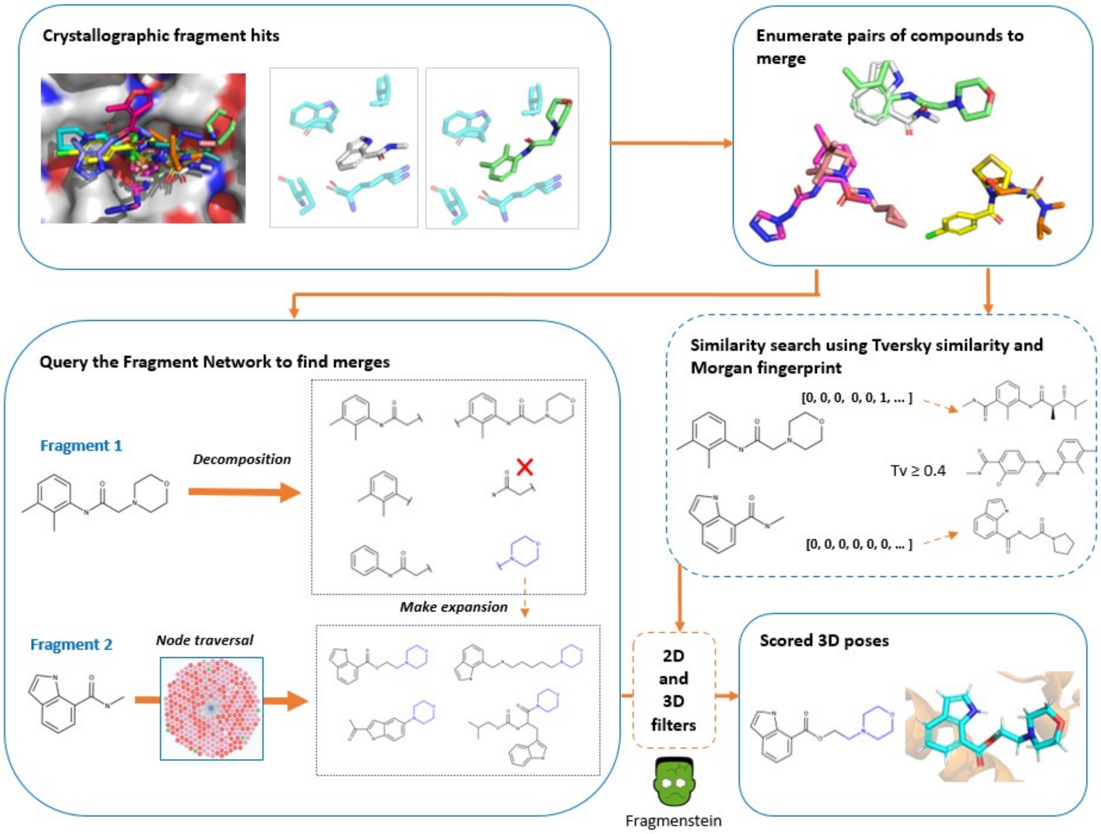

Crystallographic fragment hits
Enumerate pairs of compounds to merge
Query the Fragment Network to find merges
Fragment 1
Decomposition
Make expansion
Fragment 2
Node traversal
Similarity search using Tversky similarity and Morgan fingerprint
[0, 0, 0, 0, 0, 1, ...]
Tv ≥ 0.4
[0, 0, 0, 0, 0, 0, ...]
2D and 3D filters
Fragmenstein
Scored 3D poses

Figure 2.3: **Pipeline for identifying fragment merges.** Fragment hits from crystallographic fragment screens are used for finding fragment merges. All possible pairs of compounds are enumerated for merging (removing those with high similarity). Both the Fragment Network and similarity search are used to identify fragment merges. The Fragment Network enumerates all possible substructures of one of the fragments in the merge while the other fragment is regarded as the seed fragment. A series of optional hops are made away from the seed fragment (up to a maximum of two), after which an expansion is made by incorporating a substructure from the other fragment. The similarity search finds merges by calculating the Tversky (Tv) similarity against every compound in the database using the Morgan fingerprint (2,048 bits and radius 2). The Tversky calculation uses α and β values of 0.7 and 0.3, respectively. All compounds with a mean similarity $\geq$0.4 are retained. The merges pass through a series of 2D and 3D filters, including pose generation with Fragmenstein, to result in scored poses.

making expansions from small, ubiquitous nodes that make vast numbers of connections). A limit of 3,000 molecules per fragment pair was imposed before filtering to yield numbers comparable to that of the similarity search.

As a consequence of limiting the substructure for expansions to molecules containing at least three carbons, this can result in the Fragment Network not being able to incorporate substituents from the original fragment into the final merge. To combat this, we provide an option to perform R-group expansions of the final nodes. This can be performed as part of the initial query or as a post-refinement step of the most promising filtered molecules. To do so, a search is made for expansions of the merge compounds whereby a substituent, present in the original fragments, is added to the merge. Appendix Figure A.1 shows how the database query is able to retrieve the full scaffold incorporating key substructures from the original fragments that can then undergo R-group expansion in a subsequent step.

2.3.3.2 Similarity search

The similarity search was performed on the equivalent set of purchasable molecules available in the Fragment Network using molecular fingerprints (the code for the similarity search was written by Ruben Sanchez-Garcia). All molecular fingerprints were calculated using the RDKit implementation (Landrum, 2024) of the Morgan fingerprint (using 2,048 bits and a radius of 2; ECFPs are described in Section 1.4.4.3). All molecules in the database are represented without stereochemistry and thus the calculated Morgan fingerprints do not consider chirality. This fingerprint type is used in all similarity and clustering calculations.

For each pair of fragments, the Tversky similarity (an asymmetric metric; Senger (2009)) was calculated (using α and β values of 0.7 and 0.3, respectively) against every compound in the database. The formula is given by:

$$T(A, B) = \frac{|A \cap B|}{|A \cap B| + \alpha|A \setminus B| + \beta|B \setminus A|}$$

where A and B represent the fragment and database fingerprints, respectively. The Tversky index is an asymmetric metric and thus prioritizes whether the merges replicate the same bits as the fragments. The geometric mean was calculated between the two values and a mean value of 0.4 was used as a cut-off (the geometric mean was used to penalize compounds that are highly similar to only one of the fragments in the pair). The number of merges per pair was limited to the 3,000 with the highest similarity values (Appendix Figure A.2). Merges with <15 heavy atoms were removed to allow fair comparison with merges identified using the Fragment Network. While an alternative approach would be to perform a single query using the union of the fingerprints for the two fragments, we opted against using this approach, as this will result in smaller fragments contributing less weight to the similarity calculation.

2.3.4 Computational workflow: filtering molecules

The retrieved merges are passed through a set of 2D and 3D filters to retain the most promising compounds that are most likely to maintain the binding pose and interactions of the parent fragments (Figure 2.4). The same filters are applied to the merges from both the Fragment Network and similarity searches to allow fair comparison, with the exception of the expansion filter described below, which is only applied to Fragment Network-derived compounds. Because of the nature of the search, expansions are more likely to occur for Fragment Network-derived compounds (the similarity search uses the geometric mean of the similarity metric for both fragments). These filters are designed for large-scale fragment screens and are thus intentionally stringent. Our implementation of the pipeline was designed to enable flexibility: in cases where the filtering pipeline may result in few filtered compounds, it is straightforward to modify the filtering pipeline by either removing filtering steps or modifying thresholds, meaning that the pipeline can be used in all types of situations.

2.3.4.1 Filtering using calculated molecular descriptors

Compounds are filtered according to calculated molecular descriptors using Lipinski's rule of five (Lipinski et al., 1997) (see Section 1.1) and a maximum rotatable bond count of ten (Veber et al., 2002).

Furthermore, a limit on the number of consecutive non-ring bonds was imposed to remove 'stringy', flexible molecules. A threshold of eight bonds for 'linkers' (substructures connecting

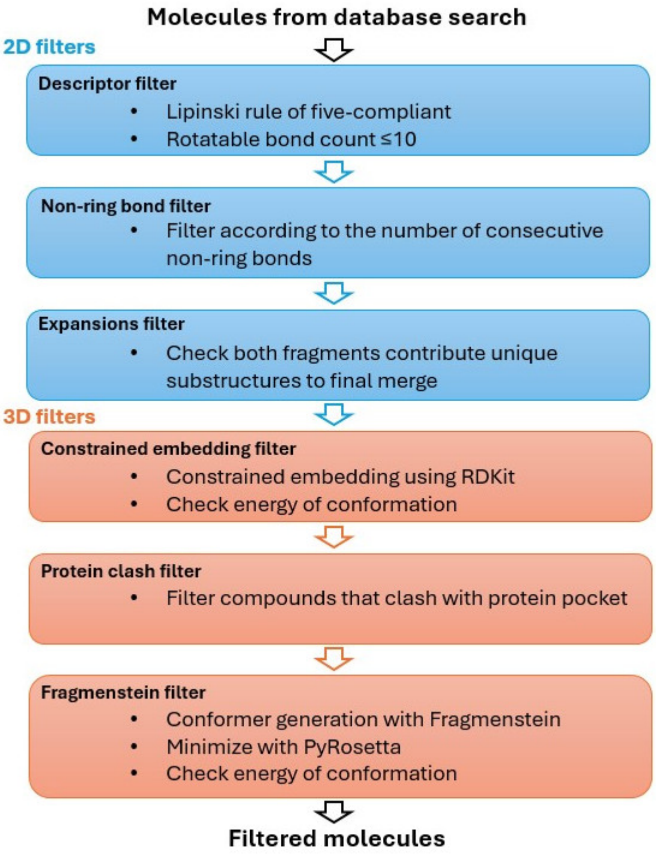

Molecules from database search
2D filters
Descriptor filter
Lipinski rule of five-compliant
Rotatable bond count ≤10
Non-ring bond filter
Filter according to the number of consecutive non-ring bonds
Expansions filter
Check both fragments contribute unique substructures to final merge
3D filters
Constrained embedding filter
Constrained embedding using RDKit
Check energy of conformation
Protein clash filter
Filter compounds that clash with protein pocket
Fragmenstein filter
Conformer generation with Fragmenstein
Minimize with PyRosetta
Check energy of conformation
Filtered molecules

Figure 2.4: **Filtering pipeline for Fragment Network merges.** Molecules retrieved from the Fragment Network database are passed through a series of 2D and 3D filters to retain the most promising compounds that are most likely to maintain the binding pose of the parent fragments. Molecules can subsequently be scored to aid selection for purchase.

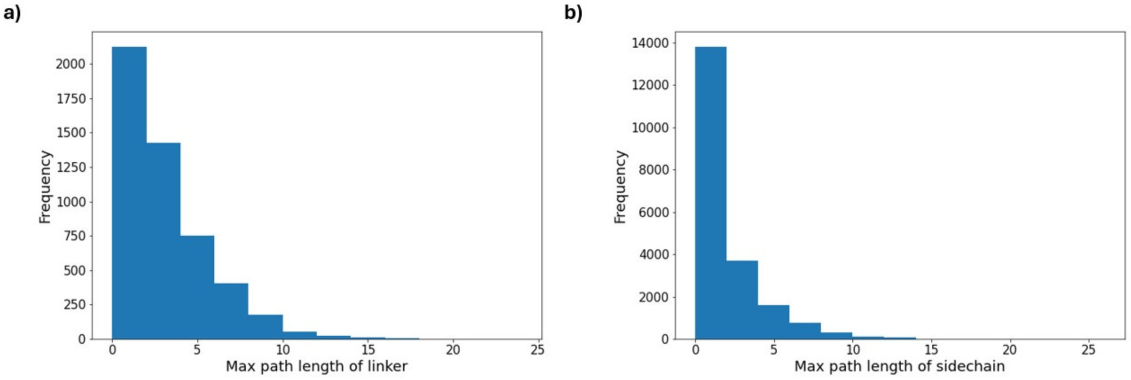


Figure 2.5: **The max linker and sidechain lengths found in ChEMBL drug molecules.** The max path length for (**a**) linkers, which join two rings, and (**b**) sidechains, which are attached to a single ring, was recorded for ChEMBL drug molecules (Gaulton et al., 2012) (CHEMBL29; after applying Lipinski filters and a maximum rotatable bond limit of 10). The 95th percentiles were used to set thresholds for the filter.

two rings) and six bonds for 'sidechains' (substructures connected to only one ring) was imposed. This was done by calculating the maximum path length of the non-ring substructures. Thresholds were set by analysing the number of consecutive non-ring bonds in compounds labelled as drugs in ChEMBL29 (Gaulton et al., 2012) (after applying the Lipinski and rotatable bond filters) and calculating the 95th percentile (Figure 2.5). These filters are optional and can be modified by the user; these thresholds may be too strict for certain scenarios, which is why the filters are applied as a filtering step rather than ensuring all compounds in the database meet these criteria.

2.3.4.2 Filtering compounds that resemble expansions of one fragment

A filter was implemented to remove compounds that resemble 'expansions' (that is, compounds that represent expansions of one fragment rather than true merges). This occurs when the fragment contributing the substructure for expansion does not contribute anything unique to the final merge. First, the MCS between the seed fragment and the merge is calculated and removed from the merge. The MCS is then calculated between the remainder of the molecule and the substructure used for expansion. If the MCS comprises at least three heavy atoms, the merge passes the filter. This filter is only applied to compounds derived from the Fragment Network approach as the search is found to result in a high proportion of these compounds when the substructure used for expansion is already present in the seed fragment (for example, the substructure may have been removed in the intermediate hops and then re-added in the final expansion step). This could be checked before the query; however, there may be circumstances in which we may want to maintain multiple copies of the same substructure in the final merge. This feature is addressed in a future iteration of the tool (Chapter 3).

2.3.4.3 Filtering compounds using constrained embedding

This filter assesses whether it is possible to generate physically reasonable conformations using distance-based constraints based on the poses of the parent fragments. The filter retrieves the atomic coordinates of substructures that were derived from the parent fragments by calculating the MCS or using the substructure that was used for expansion in the original query. The two sets of coordinates from both fragments are combined and used as a template for embedding the merge. As there may be multiple possible sets of atomic coordinates (for example, if there are several substructure matches between the MCS and the merge due to symmetries), every possible pair of coordinates is tried with every possible match with the merge. Merges for which it is possible to generate a physically reasonable structure using the RDKit constrained embedding implementation (Landrum, 2024) (that is, the bond distances fulfill the limits set by the distance bonds matrix) undergo an energy calculation using the conformation generated by the constrained embedding; the energy is calculated using RDKit using the UFF molecular force field (this force field is used for consistency with the RDKit constrained embedding implementation).

The energy of the constrained conformation is compared with the mean energy of 50 unconstrained conformations to rule out molecules with energetically infeasible poses. The conformations are randomly generated using RDKit and optimized using the UFF force field; the energy is calculated in the same way as the constrained conformations. A total of 50 conformations was deemed to be sufficient according to the rotatable bond count of the unfiltered molecules (the unfiltered datasets show an average number of rotatable bonds of less then 7) (Ebejer et al., 2012). If the ratio between the two energy calculations is >7, the molecule does not pass the filter. This threshold was selected based on an equivalent analysis of the PDBbind 2020 dataset (Su et al., 2019), selecting the

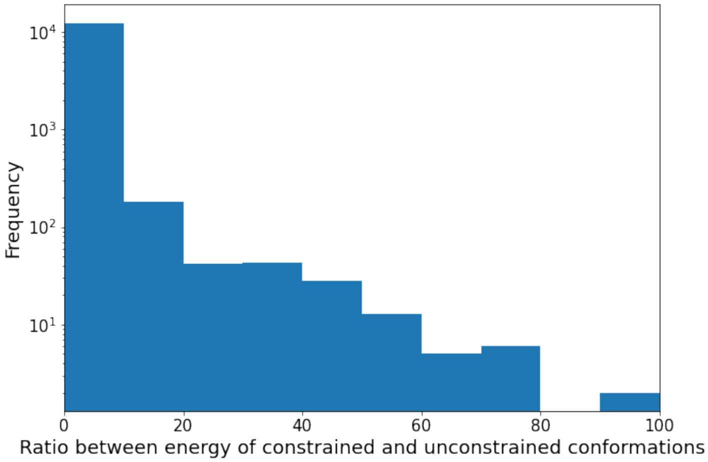


Figure 2.6: **Ratio between constrained and unconstrained conformations for PDBbind 2020 ligands.** The ratio between the energies of PDBbind ligands (Su et al., 2019) and 50 unconstrained conformers was calculated and plotted using a log scale.

value for the 95th percentile of the energy ratios (Figure 2.6). In cases where the filtering pipeline results in few filtered compounds, this filter can be removed or eased to increase the number of possible hits. Conformer generation is instead performed using only Fragmenstein (described in Section 2.3.4.5), which requires substantially greater compute time.

2.3.4.4 Filtering compounds that clash with the protein pocket

Compounds are filtered according to whether the merge fits the protein pocket. RDKit (Landrum, 2024) is used to calculate the protrusion volume between the merge and the protein (the proportion of the volume of the merge that protrudes from the protein). Molecules for which at least 16% of the volume clashes with the protein are removed. This threshold was set using analysis of Mpro crystallographic data; the clash distance between each compound and all protein structures was calculated and the value for the 95th percentile was used to set the threshold (Figure 2.7).

2.3.4.5 Filtering compounds following pose generation with Fragmenstein

Poses are generated using Fragmenstein (Ferla et al., 2024), which was developed as part of the COVID Moonshot project. Similar to the constrained embedding filter described above, Fragmen-

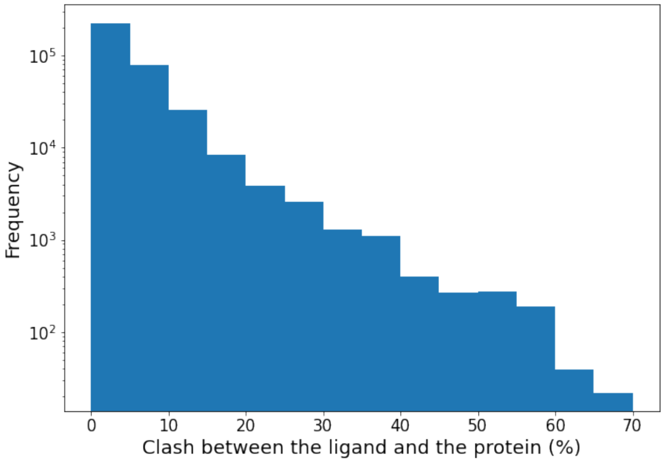


Figure 2.7: **Clash between Mpro ligands and all protein structures.** The protrusion between main protease (Mpro) ligands (available in Fragalysis) and all protein structures was calculated. The clash between all structures is shown.

stein attempts to generate poses using the coordinates of the parent fragments, but, due to its more advanced algorithm, it is more computationally expensive and thus is used at the final stage of filtering. Fragmenstein generates poses by calculating the MCS and the positional overlap between the merge and the fragments and uses the crystallographic atom coordinates for placing the merge into the protein structure. Following placement, the conformation undergoes energy minimization using PyRosetta (Chaudhury et al., 2010). Compounds are filtered for those for which Fragmenstein is able to generate a physically reasonable structure using the coordinates of both fragments. Other filters, including a combined RMSD with the fragments of <1Å, a negative $\Delta\Delta$G and the energy filter described above (to rule out unrealistic conformations), are also applied. A timeout was implemented of 10 minutes per molecule; molecules for which poses were not generated within this time limit were removed (this can occur for up to 10% of compounds entering the filter; this is likely due to the presence of multiple rotatable bonds).

2.3.5 Computational workflow: scoring and analysis

Filtered compounds are subsequently scored using various metrics to allow comparison between the two techniques. Packages used for scoring are described below.

2.3.5.1 Estimating possible protein–ligand interactions

The protein–ligand interaction profiler (PLIP) was used for estimating protein–ligand interactions present in the predicted poses. PLIP calculates five types of interaction: hydrophobic contacts, π-stacking interactions (face-to-face or edge-to-face), hydrogen bonds (with the protein as a donor or acceptor), salt bridges (with the protein positively or negatively charged and with metal ions) and halogen bonds (Adasme et al., 2021).

The predicted interactions were compared with those of the parent fragments to discern whether the merge is able to replicate unique interactions made by each of the fragments (representing useful merges). In addition to comparing chemical diversity of the compound sets, we also compared functional diversity by looking at the interactions made by each compound set. The analysis adapts the methodology described in (Carbery et al., 2022), which sought to propose functionally diverse fragment libraries that are able to recover more 'information' from fragment screens than those designed to be chemically diverse (referring to the ability of the set of fragment hits to recover a diverse set of protein interactions; see Section 1.3.2.2). Interaction diversity was analysed for each target by selecting the minimal subset of filtered compounds that represent all possible interactions (compounds are ranked according to the amount of 'information' recovered and are added until reaching a final set in which all possible interactions are recovered). This was repeated 100 times, shuffling each time (as multiple compounds will recover equivalent amounts of information). In each subset, the proportion of compounds that was proposed by either the Fragment Network or the similarity search was calculated to evaluate how diverse each compound set is with respect to diversity of interactions made.

2.3.5.2 Low-dimensional projections of chemical space

t-distributed stochastic neighbor embedding (t-SNE) plots were used to create low-dimensional representations of the compound data to allow easy visualization of chemical space (Maaten et al., 2008). The default parameters set in scikit-learn (version 1.0.1) (Pedregosa et al., 2011) were used to generate the images, with the maximum number of iterations set at 3,000.

2.3.6 Retrospective analysis using experimental data

We provide details of two separate case studies using inhibitors designed against Mpro (The COVID Moonshot Consortium et al., 2020) and *M. tuberculosis* transcriptional repressor protein EthR (Nikiforov et al., 2016).

2.3.6.1 Mpro

Existing experimental data consisting of IC_{50} values exist for Mpro and were used for retrospective analysis using a separate test set of fragments. We used data for compounds that were proposed and screened via the COVID Moonshot project and are available from the PostEra website (The COVID Moonshot Consortium et al., 2020; PostEra, 2022). We ran an independent analysis for five fragments that were used as inspiration for the manual design of 24 compounds (under submission name TRY-UNI-714a760b), 8 of which have recorded IC_{50} values in the low-to-medium micromolar range. This analysis was conducted independent of the main analysis and thus these five fragments are not included in the main test set described above. Furthermore, one of the inspirational fragments (fragment x1382-0A) is a covalent binder; however, the designed merges do not incorporate the chloroacetamide group and hence provide a useful test scenario for validating our method.

2.3.6.2 EthR

Nikiforov et al. (2016) describe a fragment-merging approach to designing *M. tuberculosis* EthR inhibitors using two fragment hits that were found to bind to the protein's hydrophobic cavity in two different locations using X-ray crystallography (denoted compounds 1 and 2 in the original paper; 2D structures of all described compounds are shown in Figure 2.8). The two fragments were used to propose three fragment merges based on the different arrangements in the cavity (compounds 3–5),

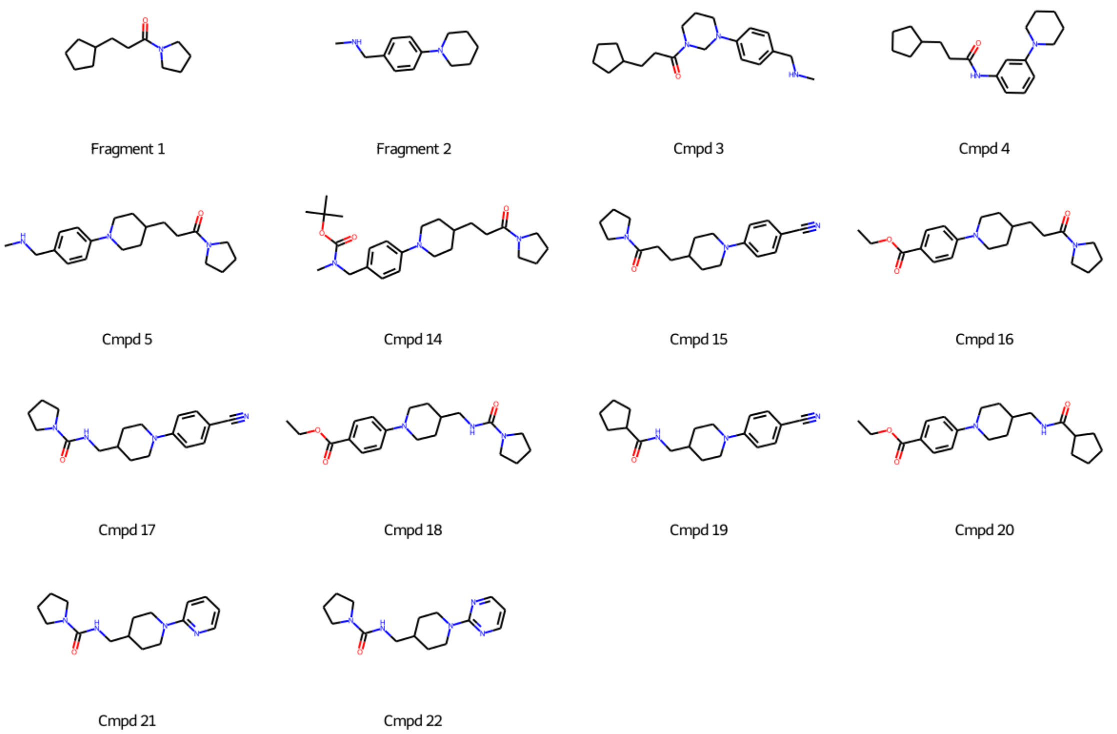


Figure 2.8: **Fragment hits and manually designed merges against *M. tuberculosis* EthR.** The 2D structures of two fragment hits (fragments 1 and 2) and manually designed merges are shown. Compounds 3–5 were initially designed, of which compounds 4 and 5 were found to recapitulate the poses of the original fragments. Compounds 14–22 were designed as analogues of compound 5.

of which compounds 4 and 5 were found to overlap with the volumes of their parent fragments. Several analogues of compound 5 were also synthesized (compounds 14–22) and were assayed using surface plasmon resonance to give IC_{50} values. We use compounds 4, 5 and 14–22 to determine whether the Fragment Network is able to known binders or close analogues to known binders.

2.4 Results

We tested the ability of the Fragment Network to identify merges using the results of crystallographic fragment screens against four targets: DPP11, PARP14, nsp13 and Mpro. Fragment hits against each target were selected and pairs of fragments were enumerated for merging. We contrasted the results with that of a more standard fingerprint-based similarity search against the equivalent database of compounds. The resulting merges identified from both techniques were passed through a filtering pipeline, comprising both 2D and 3D filters, to prioritize the most promising molecules based on properties such as molecular descriptors, the ability to fit the protein pocket and whether the compound is predicted to bind in such a way that maintains the orientation of the original fragments. Final poses were generated using Fragmenstein and the resulting molecules were scored to allow comparison of performance.

2.4.1 The Fragment Network readily identifies pure merges

For all targets, the Fragment Network yielded significant numbers of candidate merges that qualify as pure merges, as they could be placed in close agreement (<1Å RMSD) with their parent fragments (Table 2.2). We differentiate between four types of merging opportunity based on the overlap between the parent fragments (Appendix Figure A.3). Merging opportunities were categorized by visual inspection of the fragment pairs.

1. *Complete overlap merges* occur when most (at least ∼75%) of the volume of one fragment overlaps with the other. For these fragment pairs, a designed merge would not incorporate substructures that do not already overlap with the volume of the other fragment.

2. *Partial overlap by ring* occurs when only a ring overlaps between fragment pairs. These are the 'classical' merges described in the literature and are akin to the ligand overlap required for molecular hybridization. In these cases, the merging process is relatively straightforward as there is a clear hypothesis for the connectivity of the final molecule.

3. *Partial overlap without ring* occurs when the overlap is some other substructure. In these cases, the connectivity of the final compound is non-obvious.

Table 2.2: The numbers of filtered compounds identified using a Fragment Network versus similarity search.

	Number of hits	Search type	Before filtering	After filtering	% filtered	Number of intersect*	Complete overlap (%)	Partial overlap — ring (%)	Partial overlap — no ring (%)	No overlap (%)
DPP11	11	FN	22,903	198	0.9	8	0.0	29.4	70.6	NA
		SS	85,919	271	0.3		4.0	32.0	64.0	NA
PARP14	13	FN	48,320	70	0.1	0	0.0	18.3	45.1	36.6
		SS	78,116	56	0.1		16.1	16.1	37.5	30.4
nsp13	9	FN	36,239	503	1.4	1	0.4	99.6	0.0	NA
		SS	40,102	530	1.3		1.6	97.4	1.0	NA
Mpro	19	FN	109,012	952	0.9	4	2.5	6.8	49.6	41.2
		SS	169,424	918	0.5		71.0	7.1	12.5	9.4

DPP11, dipeptidyl peptidase 11; FN, Fragment Network; PARP14, poly(ADP-ribose) 14; Mpro, main protease; nsp13, non-structural protein 13; SS, similarity search. *Refers to the numbers of compounds present in both the FN and SS compound sets.

4. *Non-overlap of parent fragments* requires identifying linking opportunities, which was found in the PARP14 and Mpro datasets.

Categories 3 and 4 can be considered 'linker-like' merges, and for these it tends to not be immediately evident which pieces of the parent fragments to incorporate into the final merge, and how to do so in a way that will result in a physically reasonable structure that maintains the binding pose of the fragments. Nevertheless, for several such pairs, the Fragment Network identified how to join two distinct substructures of the parent fragments by a molecular linker, resulting in chemical series with diversity in the 'linker' region, created by the intermediate optional hops in the database query (Figures 2.9 and 2.10).

The Fragment Network enables searches not just for merges but also linkers, by including fragment pairs up to 5Å apart. Linking both fragments in their entirety is typically difficult, as it leads to compounds that are neither purchasable nor synthetically accessible. Instead, our search links partial structures of each parent, thereby generating many candidates; for Mpro, it identified 419 compounds that are as much linkers as merges.

2.4.2 The Fragment Network and similarity searches find comparable numbers of compounds

The outputs of the Fragment Network and similarity searches are compared in Table 2.2. The totals show the number of unique compounds found across all pairs of fragments, even though the same candidate compound can be identified across multiple fragment pairs. All instances of each compound are filtered, as the 3D structures of the parent fragments may affect their chances of successful filtering. The results show that the techniques produce comparable numbers of filtered compounds across all targets, while the filtering efficiency differed substantially between targets, from 0.1% for PARP14 compounds to 1.3–1.4% for nsp13 compounds.

Figure 2.11 shows the number of filtered compounds per pair using each technique, highlighting that this value is highly dependent on both the fragment pair and the technique used, with some pairs yielding no filtered merges from either technique. It was therefore not meaningful to compare scoring metrics between techniques, as there are few fragment pairs resulting in comparable numbers

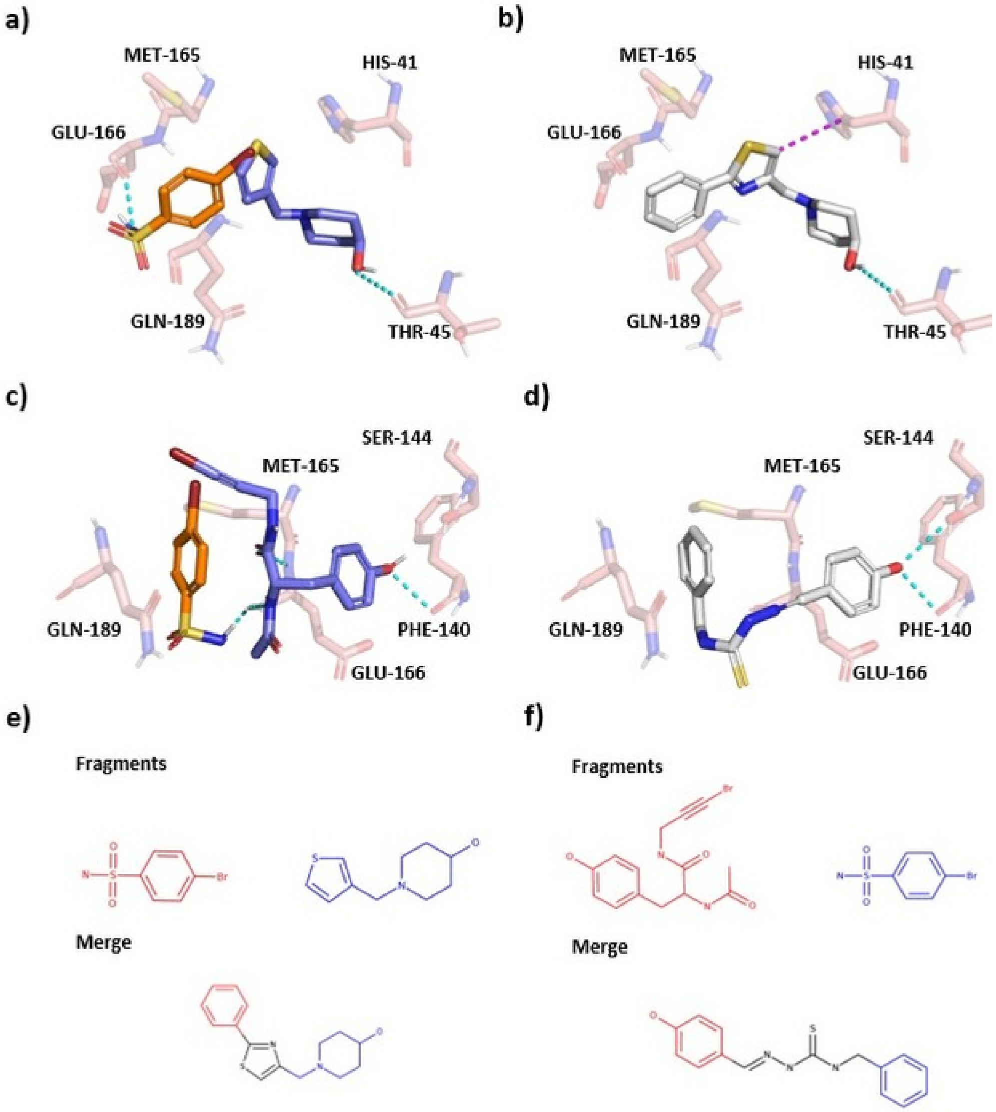


Figure 2.9: **The Fragment Network identifies pure merges.** (**a**) A fragment-merging opportunity for the main protease (Mpro) dataset. Interactions are predicted using the protein–ligand interaction profiler (PLIP). Hydrogen bonds and π-stacking interactions are shown by cyan and magenta dotted lines, respectively. (**b**) The linker-like merge (pose generated using Fragmenstein) joins substructures from partially overlapping fragments by a 'linker-like' region, maintaining the hydrogen bond with THR-45; a change in orientation of the thiazole ring (with respect to the thiophene ring in the fragment) enables an additional π-teeing interaction with HIS-41. (**c**) A fragment-linking opportunity for Mpro. (**d**) The proposed compound maintains a hydrogen bond with PHE-140 and makes an additional bond with SER-144. (**e**) The fragments and merge in **a,b** in 2D. (**f**) The fragments and merge in **c,d** in 2D.

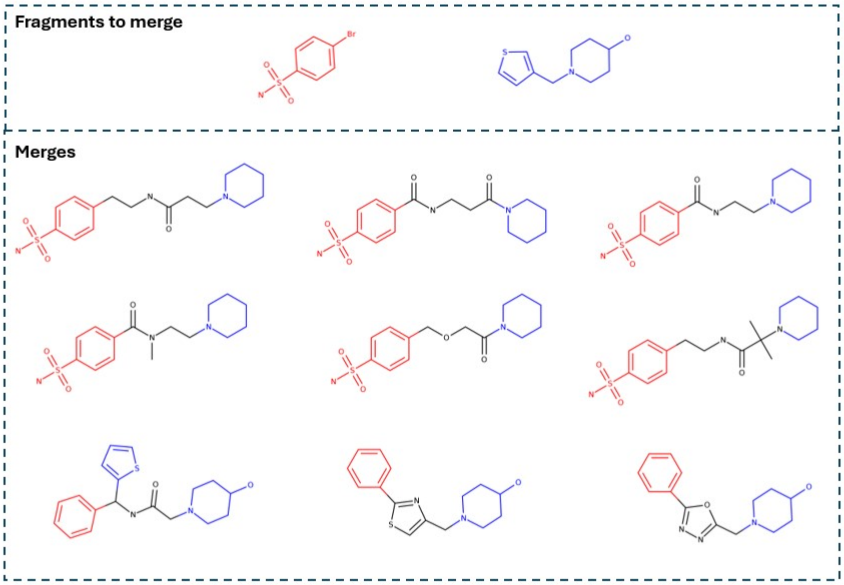


Figure 2.10: **Example linker-like merges.** Example 'linker-like' merges for two fragments found to bind to the main protease (Mpro). Diversity is generated in the 'linker' region joining the two substructures from the parent fragments. It is worth noting that the 2D compounds retrieved from the Fragment Network do not have stereochemistry assigned, such as for the bottom-left compound, which has a chiral centre.

of filtered compounds for both techniques.

The number of merges found by the Fragment Network depends on several factors. First, highly connected seed fragments (for example, molecules that decompose into simple, common substructures) will result in more possible query paths and thus greater likelihood of finding merges that incorporate the substructure of interest. Second, a query path that involves a ubiquitous, promiscuous node (for example, if an optional hop yields a benzene) will pull out more paths, as these nodes typically make hundreds of thousands of connections. Third, during filtering, if the fragments in the pair are well-aligned and share overlapping substructures, placement of the merge is more likely to be successful (this is relevant for both Fragment Network and similarity search-derived compounds). Fourth, if the query molecule is expanded by a common substructure, this will result in many paths. Specifically, the most common substructure used across all targets for expansion was a benzene ring, which is also a substructure for many of the catalogue compounds. For nsp13, 495 of the 510 (97%) filtered compounds (not accounting for redundancy) were identified by incorporating a benzene. However, the substructures incorporated for the other targets were more diverse, with benzene accounting for 37–52% of expansions (Appendix Table A.2).

The number of merges identified by the similarity search is affected by different factors: while it identifies comparable numbers of classical merges (category 2; Appendix Figure A.4) to the Fragment Network, it is much less effective at identifying linkers (category 4), as evident for Mpro, where the similarity searches identify only 90 linker compounds, compared with 419 from the Fragment Network. Completely overlapping (category 1) merges also present a peculiar difficulty, since it is inherently more difficult to filter out compounds that represent expansions rather than true merges, as it is harder to codify the rules for merges that are maintaining bits of the parent fragments rather than substructures. For example, fragments x0195-0A and x0946-0A of the Mpro dataset both share a phenylsulfonamide structure (Appendix Figure A.5), for which the similarity search identified 395 filtered compounds. However, most of the merges were analogues, as they maintained the phenylsulfonamide but had no unique contributions from the remainder of the fragment structures.

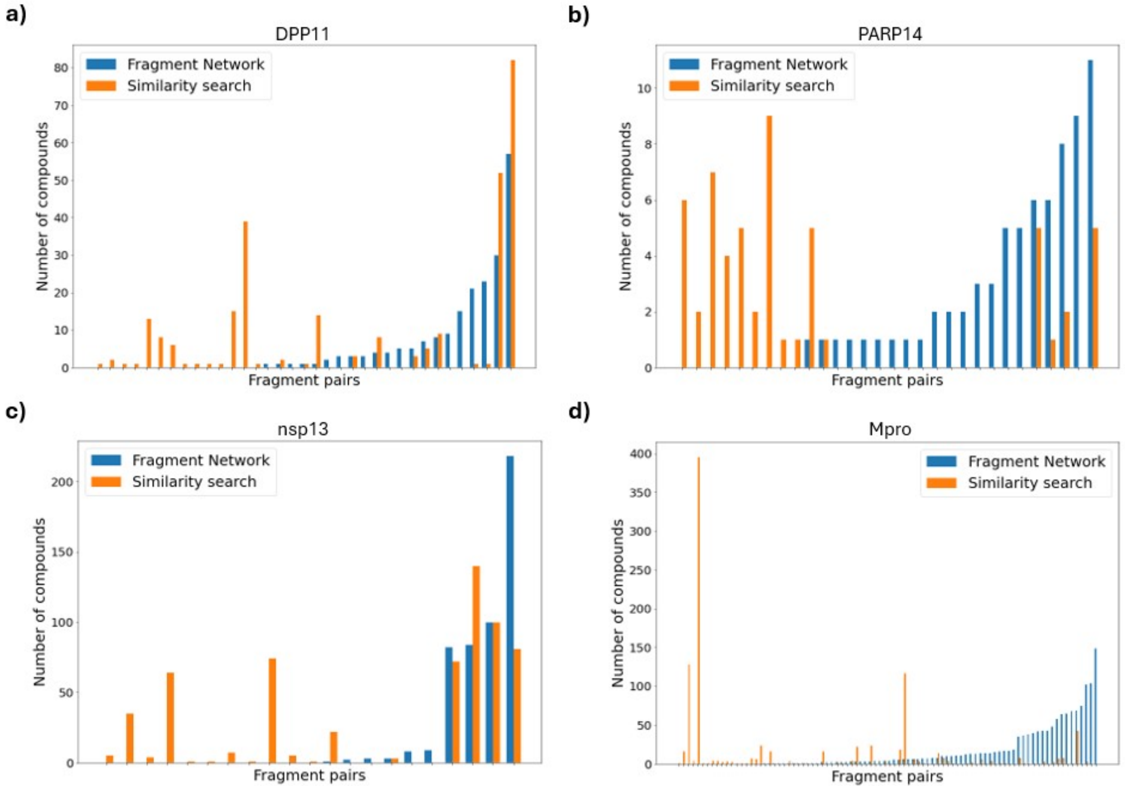


Figure 2.11: **The Fragment Network and similarity searches identify filtered compounds for different fragment pairs.** The numbers of filtered compounds for each fragment pair found using the Fragment Network (blue) or similarity search (orange) are shown across targets (**a**) dipeptidyl peptidase 11 (DPP11), (**b**) poly(ADP-ribose) polymerase 14, (PARP14) (**c**) non-structural protein 13 (nsp13) and (**d**) main protease (Mpro). Only pairs that resulted in filtered compounds are shown. The fragment pairs are ordered on the x-axis according to the number of Fragment Network compounds filtered for each pair (ascending from left to right). The data show that each search technique was able to identify filtered compounds for pairs where the other technique identified none.

Table 2.3: The number of pairs represented by Fragment Network and similarity search filtered compounds.

Target	Fragment Network-only	Similarity search-only	Both
DPP11	9	14	12
PARP14	15	9	6
nsp13	4	11	6
Mpro	29	18	36

DPP11, dipeptidyl peptidase 11; PARP14, poly(ADP-ribose) 14; Mpro, main protease; nsp13, non-structural protein 13.

2.4.3 The Fragment Network and similarity search identify compounds from distinct areas of chemical space

There are very few intersecting filtered compounds found from the two search techniques, with a maximum of eight compounds found to be in common for DPP11 (Table 2.2), which suggests they operate in different catalogue space. When we consider which fragment pairs result in successfully filtered merges, Figure 2.11 and Table 2.3 show that each technique is able to identify merges for merge pairs for which the other technique identifies none, supporting the notion that these techniques are complementary and could be used in parallel to increase the potential resultant hit rate of a catalogue search.

The T-SNE projections in Figure 2.12 further demonstrate that these techniques operate in distinct areas of chemical space, as both techniques seem to be clustered in different regions and show little overlap. This is particularly apparent for nsp13 and Mpro, owing to the greater availability of data. Figure 2.12 also shows that the Fragment Network-filtered compounds show greater overall chemical diversity. The compounds were clustered using the Taylor–Butina clustering algorithm (using a Tanimoto distance threshold of 0.3) (Butina, 1999), and the number of clusters containing only Fragment Network or similarity search compounds were counted; the Fragment Network was found to result in more clusters across all targets (Appendix Table A.3). The mean Tversky similarity between all filtered compounds and their parent fragments was also calculated (Appendix

Figure A.6); the Fragment Network compounds are far more dissimilar using these metrics and are able to access areas of chemical space not reached by the fingerprint-based similarity search, resulting in greater diversity in the compounds retrieved.

2.4.4 The Fragment Network and similarity searches identify compounds that form complementary interactions with different residues

Across both the Fragment Network and similarity searches, the filtered compound sets were found to interact with the same or greater number of residues than the parent fragments used to create the respective merges. Furthermore, the Fragment Network and similarity search filtered compound sets were each found to make interactions with residues that the other compound set was not able to reach, in particular for DPP11 and PARP14; for nsp13, the Fragment Network reached residues not reached by similarity search, while the opposite was shown for Mpro (Table 2.4). We also calculated the number of merges that represent true merges in terms of residue-level interactions, meaning merges that are able to replicate unique interactions made by each of the parent fragments (after discounting interactions that are made by both of the parent fragments). The Fragment Network was more efficient for identifying this type of 'true merge' for PARP14 and Mpro, while the similarity search performed better for nsp13 and DPP11 (Appendix Table A.4). This suggests that the two techniques may perform better for different targets.

Table 2.4: The number of interaction residues reached by Fragment Network and similarity search filtered compounds.

Target	Fragment Network-only	Similarity search-only	Both
DPP11	3	5	11
PARP14	1	5	9
nsp13	8	0	20
Mpro	0	7	25

DPP11, dipeptidyl peptidase 11; PARP14, poly(ADP-ribose) 14; Mpro, main protease; nsp13, non-structural protein 13.

In order to explore the concept of 'interaction diversity', whereby compounds are selected ac-

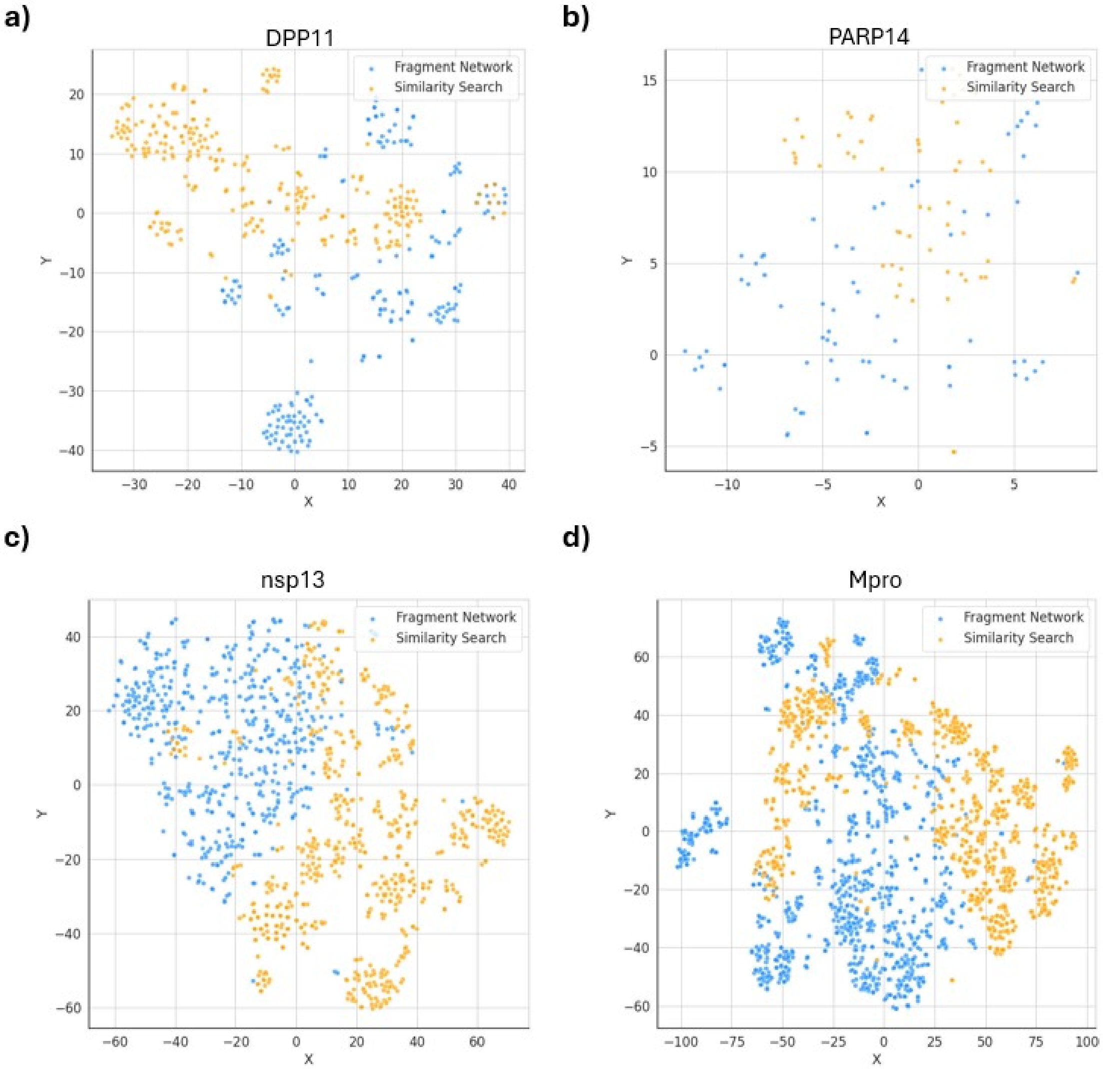


Figure 2.12: **Fragment Network and similarity search-derived compound sets populate different regions of chemical space.** The chemical space occupied by the filtered compound sets is projected into two dimensions using the T-SNE algorithm across targets (**a**) dipeptidyl peptidase 11 (DPP11), (**b**) poly(ADP-ribose) polymerase 14, (PARP14) (**c**) non-structural protein 13 (nsp13) and (**d**) main protease (Mpro). Fragment Network compounds are shown in blue and similarity search compounds are shown in orange. The two compound sets are shown to occupy distinct areas of chemical space.

Table 2.5: The composition of functionally diverse subsets identified from filtered compound sets.

Target	Fragment Network (%)	Similarity search (%)
DPP11	39.7 ± 5.0	60.3 ± 5.0
PARP14	40.3 ± 5.4	59.7 ± 5.4
nsp13	74.0 ± 2.8	26.0 ± 2.8
Mpro	30.5 ± 4.3	69.5 ± 4.3

Mean and standard deviation are provided. DPP11, dipeptidyl peptidase 11; PARP14, poly(ADP-ribose) 14; Mpro, main protease; nsp13, non-structural protein 13.

cording to the diversity in the interactions made (functional diversity) rather than their chemical diversity, we performed an analysis to identify the smallest compound sets for each target that make all possible interactions with the protein. This is done across both the Fragment Network and similarity search-identified compounds. In the final compound sets, we found that compounds from both search approaches were represented, again supporting the idea that the two techniques are complementary (Table 2.5). For the purposes of identifying compounds for purchase and further screening, we argue that these results support performing both techniques in parallel for catalogue search, expanding the set from which to identify the most useful purchase list for further screening and informing SAR.

2.4.5 Efficiency of the filtering pipeline

A series of 2D and 3D filters was implemented to remove compounds that lack desired molecular properties, that may not represent true merges, that do not fit the protein pocket and that are unable to replicate the binding pose of the original fragments. Appendix Table A.5 shows the percentages of compounds that are removed by each step in the filtering pipeline (as both the percentage of total compounds and the percentage of compounds entering the filter). The Fragment Network search typically identifies larger molecules and molecules that contain long linker regions containing many rotatable or consecutive non-ring bonds, while the similarity search compounds are typically more conservative with regard to their size. The constrained embedding filter removes a greater proportion of similarity search compounds compared with Fragment Network compounds. A possible explanation for this is that the Fragment Network is more likely to preserve exact substructures of the parent fragments and thus the MCS calculation used to extract the atomic coordinates of atoms to use as the embedding template is more likely to yield a greater number of atoms. Across both techniques, the step involving pose generation with Fragmenstein and filtering based on the resulting conformations is the most restrictive filter. We would expect this step to be the most stringent, as it is the most accurate and computationally intensive filter in our pipeline for determining whether the molecule can adopt a sensible binding pose that mirrors that of the parent fragments.

2.4.6 The Fragment Network has efficiency benefits over similarity search

The Fragment Network search method is more computationally efficient than the similarity search. The Fragment Network querying was run single-threaded and takes an average of 1.4–13.0 minutes (dependent on the target) per fragment pair. The similarity search was run on 16 CPUs and takes ∼2.5 minutes per fragment pair, requiring up to an estimated 40.0 minutes in total of CPU time (search timings are shown in Table 2.6).

Table 2.6: CPU time required for querying per fragment pair

Target	Query time (CPU minutes per fragment pair)	
	Fragment network	Similarity search
DPP11	2.1	40.0
PARP14	2.2	40.0
nsp13	13.0	40.0
Mpro	1.5	40.0

DPP11, dipeptidyl peptidase 11; PARP14, poly(ADP-ribose) 14; Mpro, main protease; nsp13, non-structural protein 13.

While the Fragment Network has potential efficiency benefits in terms of the search method, it required overall greater filtering time for 3 of the 4 targets (though different numbers of compounds were filtered through each). The time required for filtering all merges for a target ranged from 5.2–75.9 CPU days (this value varies considerably depending on the target and search technique). Timings for filtering and placement of molecules using Fragmenstein are provided in Table 2.7. Conformer generation using Fragmenstein represents the most computationally expensive step;

Table 2.7: CPU time required to run entire filtering pipeline (and Fragmenstein alone)

Target	Technique	Total filtering time (CPU days)	Total Fragmenstein time (CPU days)	Number of molecules filtered	Number of molecules through Fragmenstein
DPP11	Fragment Network	23.1	19.6	40,553	6,968
	Similarity search	53.0	46.0	116,819	16,451
PARP14	Fragment Network	13.0	11.6	116,084	13,336
	Similarity search	6.3	5.2	175,376	5,792
nsp13	Fragment Network	19.2	16.4	53,618	8,616
	Similarity search	16.4	13.0	88,539	7,203
Mpro	Fragment Network	75.9	64.7	175,024	23,093
	Similarity search	55.4	43.3	261,756	20,165

DPP11, dipeptidyl peptidase 11; PARP14, poly(ADP-ribose) 14; Mpro, main protease; nsp13, non-structural protein 13.

however, this step could easily be replaced with other placement tools such as constrained docking programs to speed up the pipeline if desired.

It is worth noting that either search method could be subjected to optimization; for example, a similarity search could be made faster by clustering molecules and running a hierarchical search (by calculating similarity against cluster centroids followed by similarity calculations against all members of the most similar clusters). However, the inherent efficiency benefit of the Fragment Network approach is that it restricts the search to a portion of the database within a specified number of hops from the seed fragments.

In comparison, building the Fragment Network database is a more computationally expensive procedure. To perform the fragmentation and processing of five million molecules for input into the Network requires on average 49 CPU days (this value is highly variable and dependent on molecular complexity). However, given the composition of the database and ability to represent entire catalogues, this is a rare operation to undertake as the overall chemical space of the catalogues will not change drastically over short timescales. This pre-computation also allows a more thorough exploration of chemical space than possible using a similarity search alone. Moreover, the addition of new molecules (as catalogues are updated) will not require the re-processing of all molecules and thus would be substantially less intensive. For example, we could perform the fragmentation of approximately 700,000 molecules within 7 CPU days, and thus should be able to perform updates at a rate that exceeds the rate of growth of any commercial catalogue.

2.4.7 Retrospective analysis using experimental data

2.4.7.1 Mpro

To test whether the Fragment Network is able to identify true binders, a retrospective analysis was performed using data from the public COVID Moonshot project (The COVID Moonshot Consortium et al., 2020). A set of five fragments were originally used to propose the manually designed TRY-UNI-714a760b series of fragment merges, which successfully achieved on-scale potency for several compounds. We queried the Fragment Network with the same fragments and looked for TRY-UNI-714a760b compounds in the results. Three of the five fragments are either highly similar or represent substructures of each other, and thus pairs between these fragments were removed, resulting in seven pairs of fragments for querying. We also compared our proposed merges against all other compounds with experimental data but without a covalent warhead. This resulted in a total of 1,247 compounds with an IC_{50} value of $<99\mu$M against Mpro recorded by either a fluorescence or RapidFire mass spectrometry (RF-MS) assay. The comparison involved checking whether there were experimental compounds with an identical SMILES to filtered compounds, or whether they were shown to be similar using Tanimoto metrics between Morgan fingerprints (calculated as described in Section 2.3.3.2).

After filtering, the Fragment Network did identify one known binder, along with compounds highly similar to several other known binders (Figure 2.13; Appendix Figure A.7). The direct match, LON-WEI-b2874fec-25, has an RF-MS IC_{50} value of 59.6μM; between the filtered compounds and the known binders, there were 13 pairs with Tanimoto similarity of >0.6, including TRY-UNI-714a760b-18 and TRY-UNI-714a760b-22 (Tanimoto similarities of 0.73 and 0.69, respectively; Figure 2.13c). Where available, the Fragmenstein-generated poses closely mimic the crystal poses. Compound TRY-UNI-714a760b-16 was also identified, which does not have an IC_{50} value.

Two additional known binders (JAN-GHE-83b26c96-22 and TRY-UNI-714a760b-18) were re-

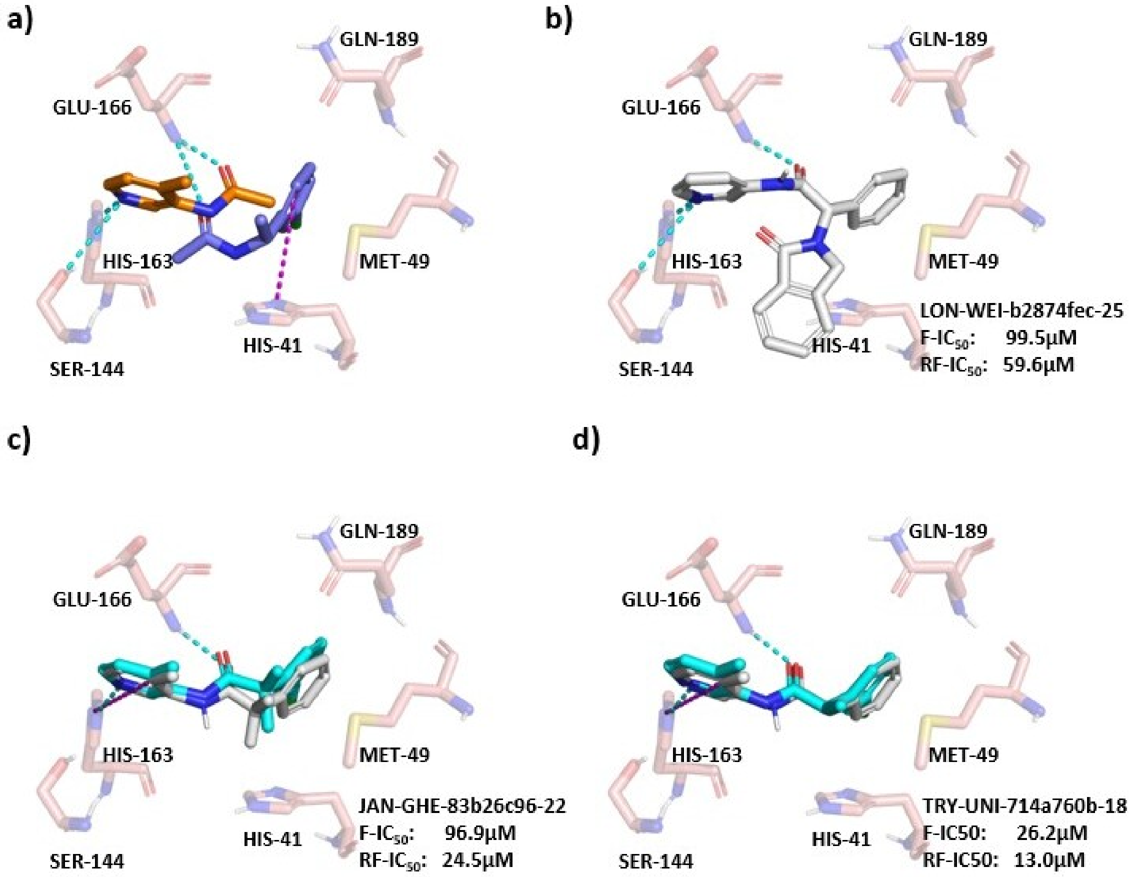


Figure 2.13: **The Fragment Network identifies a known binder against Mpro.** A Fragment Network search using (**a**) two fragment hits against the SARS-CoV-2 main protease (Mpro) identifies: (**b**) a known binder against Mpro (LON-WEI-b2874fec-25; RapidFire mass spectrometry (RF-MS) IC$_{50}$ value of 59.6μM); and similar compounds to known binders, (**c**) TRY-UNI-714a760b-18 (fluorescence and RF-MS IC$_{50}$ values of 26.2μM and 13.0μM) and (**d**) JAN-GHE-83b26c96-22 (RF-MS IC$_{50}$ value of 24.5μM). Fragmenstein-predicted merge poses are shown in white and crystal poses are in cyan. Interactions are predicted using the protein–ligand interaction profiler (PLIP) and key interaction residues are shown. Hydrogen bonds are shown in cyan and π-stacking interactions are shown in magenta.

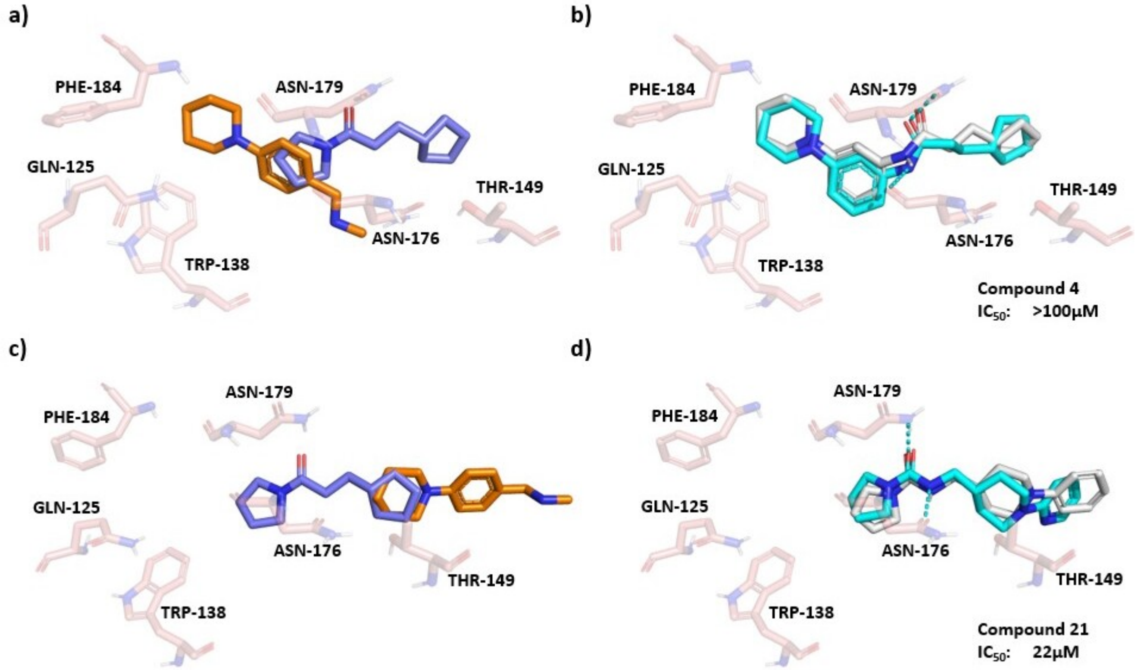


Figure 2.14: **The Fragment Network identifies a known binder against EthR.** A Fragment Network search using (**a**) two fragment hits against (which each bind in two different positions) *Mycobacterium tuberculosis* transcriptional repressor protein EthR identifies (**b**) a known binder (compound 4; IC_{50} value of $>100\mu M$). (**c**) An alternative crystallographic arrangement of the equivalent fragments identifies (**d**) a similar compound to a known binder, compound 21 (IC_{50} value of $22\mu M$). Fragmenstein-predicted merge poses are shown in white and crystal poses are in cyan. Hydrogen bonds are shown in cyan.

trieved by mitigating the limitation of the current implementation of the Fragment Network described in Section 2.3.3.1, which necessitates expansions to be made with substructures containing at least three carbons. When we performed R-group expansions on the retrieved analogues, using substituents from the original fragments, the additional binders were identified through the addition of a chlorine atom (Appendix Figures A.8 and A.9).

In comparison, the similarity search after filtering identified three of the retrieved analogues (JAN-GHE-83b26c96-22, TRY-UNI-714a760b-18 and TRY-UNI-714a760b-22). However, the similarity search did not identify Fragment Network match (LON-WEI-b2874fec-25), due to the low Tversky similarity between this compound and the parent fragments. This further supports the complementary nature of these two search techniques.

2.4.7.2 EthR

We perform an additional analysis using two fragments (compounds 1 and 2) found to bind to EthR (each fragment binds in two different conformations). As mentioned in Section 2.3.6.2, the original publication (Nikiforov et al., 2016) involved the design of two fragment merges that were found to recapitulate the original binding pose (compounds 4 and 5), plus the design of several analogues of compound 5 (compounds 14–22).

As compounds 1 and 2 already exist in the database, our Fragment Network search was run for these two fragments and the filtering pipeline was repeated using all four pairs of the different fragment conformations. This resulted in 10 filtered compounds (an example is shown in Appendix Figure A.10). While compound 4 was retrieved from the database, it was ruled out during filtering using our stringent default filtering parameters. As compound 4 was removed by the constrained embedding filter (a possible explanation for this is given below), we also tried relaxing the filtering pipeline by only using Fragmenstein for conformer generation, which increased the number of filtered compounds to 148.

With the new configuration, the Fragment Network search was able to identify compound 4 after filtering (Figure 2.14). This exemplifies the benefit of implementing a flexible filtering pipeline that can be tailored according to the features of the test system. While the compound has weak potency, with an IC_{50} value of $>100\mu$M, the Fragmenstein-predicted pose shows good overlap with the crystal pose. However, the change in orientation of the five-membered ring between the two structures (and with respect to the original fragment) is likely the reason for the initial failure of the constrained embedding filter, as RDKit was not able to generate a conformation that maintains the orientation of this ring in a physically reasonable way. On the contrary, Fragmenstein, that uses a smarter version of atom mapping for MCS is able to deal with this problem, thus leading to sensible poses. While the manually designed compound 5 is not present in the network, the Fragment Network identified several compounds that were highly similar to analogues of compound 5, shown in Appendix Figure A.11, with IC_{50} values ranging between 2–25μM. An example is shown in Figure 2.14d, for which the Fragmenstein-predicted pose again well reflects that of the original crystal pose. Moreover, the two compounds differ by a single nitrogen atom (which is not present in the original fragment and could not be expected to be identified by the Fragment Network pure merge search).

Collectively, these results provide validation for the use of the Fragment Network to progress crystallographic fragment hits to compounds with on-scale potencies in multiple systems.

2.5 Discussion

We have illustrated the use of the Fragment Network, a graph database containing compounds from commercial catalogues, for finding fragment merges, and show that our method can expand the scope of a traditional similarity-based catalogue search and increase the yield of potential follow-ups. We contrast the use of the Fragment Network and a fingerprint-based similarity search to find merges using the results of crystallographic fragment screens. The results support the idea that these two techniques are complementary and should be used in parallel, in that they are able to identify merges for pairs of fragments that are not represented by the other, they show little overlap in the compounds identified and chemical space in which they operate and they are able to identify compounds that form interactions not represented by the other. In particular, the Fragment Network is well-suited to identifying what we refer to as perfect merges, in other words, compounds that incorporate exact substructures of the parent fragments. The Fragment Network exploits a different type of similarity that is able to preserve substructures of the original fragments but that may appear dissimilar using a fingerprint-based metric. This also demonstrates usefulness beyond that of classical molecular hybridization techniques, which require overlapping substructures to exist between the set of input molecules.

The ability of the Fragment Network to identify known inhibitors of Mpro and EthR and highly similar compounds to inhibitors predicted to bind in the same orientation provides initial validation and supports the use of this approach in the FBDD pipeline. While the default thresholds have been carefully selected to yield sensible results in most scenarios, our flexible pipeline allows for customization to better adapt to different targets. Users are easily able to select which filters and scoring metrics to use and to set desired parameters and thresholds, and developers are able to incorporate new metrics and filters with ease.

The nature of the perfect merges identified has some repercussions; while the variation added to the molecule may not necessarily be better for maintaining all possible fragment interactions, we would hope that the preservation of exact substructures may result in the merges being more likely to mimic the original binding pose of the fragments. Future work (described in Chapter 3) aims to expand the pool from which we can select follow-up compounds by incorporating bioisosteric replacements into the merging process.

Regarding efficiency, one of the main advantages of the Fragment Network approach is the reduced compute time required to run a search. As the size of commercial catalogues is ever-expanding and new approaches are needed for managing the vast availability of compound data, efficient search techniques represent an important avenue of research. Search techniques such as these could become especially relevant for large virtual libraries with increasing coverage of chemical space.

In this work we also identified several potential areas for improvement. The efficiency of the search could be improved by limiting the substructures used in the expansion to those that are responsible for an interaction that we want to preserve in the final merge; similarly, an important step is ensuring that the seed fragment is not reduced down to a substructure that is not responsible for making an interaction.

A notable constraint of this pipeline was the inability to incorporate substituents from the parent fragments due to the limitation imposed on substructures for expansion of containing at least three carbon atoms. While we can easily tackle this by performing R-group searches and refining the filtering steps to maintain the most sensible R-group expansions, because of the lack of atom mapping between nodes, it is not feasible to track where on the molecule substituents are added *during* the database search, which would be more efficient. We describe this problem in more detail in Appendix Section A.1.

2.6 Conclusions

Currently, there are limited available *in silico* approaches for identifying fragment merges and, in particular, those that identify compounds that are synthetically accessible. We demonstrate the application of a graph database for identifying commercially available fragment merges that will allow rapid follow-up and progression of hits in FBDD campaigns.

Use of the Fragment Network has proved to be an effective way to improve the yield and potential hit rate of a catalogue search and has demonstrated itself to be complementary to a classical search using a fingerprint-based similarity metric. Our pipeline was able to successfully identify known inhibitors against Mpro and EthR, the former with micromolar activity. The results reported here support the use of the two techniques in parallel, with the aim of selecting the optimal compound subset for purchase from all enumerated compounds for further assaying and informing the next iteration of synthesis.

3. Expanding the scope of a catalogue search by finding bioisosteric merges

3.1 Chapter motivation

The pipeline described in Chapter 2 was able to find fragment merges across multiple targets and identify potential known binders in retrospective case studies. While the pipeline showed benefit beyond classical techniques for catalogue search, it was not able to find fragment merges for certain pairs of fragments for which the similarity search did, for example, if one of the fragment nodes makes few connections in the network and thus the size of the chemical space searched is limited. There are also possible avenues for improvement regarding the efficiency of the search, as described in Section 2.5, particularly with respect to the enumeration of the initial substructures for merging.

This Chapter explores how we can expand the chemical space covered by the Fragment Network search by pursuing what we have termed ‘bioisosteric merges’, referring to compounds that incorporate substructures with similar pharmacophoric properties to those seen in the parent fragments. This new search method finds follow-up compounds that are more chemically diverse yet directly informed by the substructures within the parent fragments, unlike virtual screening techniques, which involve descriptor calculations and/or docking experiments against either entire compound libraries or large subsets of libraries. The work in this Chapter is described in a 2024 *bioRXiv* preprint (Wills et al., 2024). I carried out all the work described within the Chapter unless explicity stated.

3.2 Introduction

As described in Section 1.4.4.5, existing pipelines for follow-up compound design based on the results of fragment screens have proved to be effective in reaching compounds with substantially improved affinities (low μM to nM), but have typically been restricted to scenarios with few (up to five) fragment hits (Wills et al., 2023; Metz et al., 2021; Müller et al., 2022). Many of these methods perform docking of virtual chemical libraries using template-guided docking based on the initial fragments. A major obstacle with docking tools is that they are computationally intensive, and thus docking entire compound libraries containing millions of compounds is not always practical. Consequently, appropriate pre-filtering steps to select promising compounds are needed. This can be achieved, for instance, through the use of molecular fingerprint-based similarity searches (Metz et al., 2021) or the use of workflows that iteratively dock compounds using reaction-based enumeration (Müller et al., 2022). Another issue is that docking results are not perfect in terms of accuracy (Chen, 2015), hence many candidate compounds have to be tested in order to achieve a sufficient hit rate.

The work in Chapter 2 and (Wills et al., 2023) provided an initial pipeline to address this problem. This pipeline used a graph database representation of chemical space, termed the Fragment Network (Hall et al., 2017), to find catalogue compounds that contain substructures from two parent fragments simultaneously. In this Chapter, I describe a new version of our algorithm, where we have addressed some of its previous limitations, such as merging substructures that may not be spatially compatible or important for binding (which wastes computational resources), or restricting our focus to 'perfect merges' — compounds that contain exact substructures of the parent fragments, which can limit productivity for certain pairs of fragments (Wills et al., 2023). The tool described in Chapter 2 enumerated all pairs of fragments for merging, removing only pairs that are near identical in terms of chemical structure and overlap, and that exceed a specified distance threshold. Our new pipeline aims to be more efficient by ruling out substructures that are unlikely to yield merges due to steric constraints and, optionally, make important interactions with the protein. Additionally, instead of limiting the search to exact substructures found in the initial fragment hits,

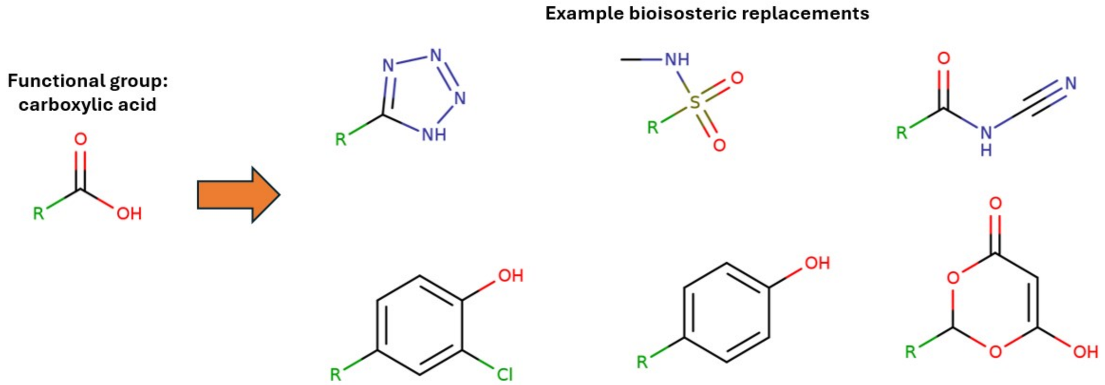


Figure 3.1: **Example bioisosteres for carboxylic functional groups.** Bioisosteres represent functional groups that are likely to exhibit similar biological properties. Example bioisosteric replacements for a carboxylic acid functional group are shown. ‘R’ denotes the attachment point. Figure adapted from Swain (2006).

it can find merges containing substructures that replicate pharmacophoric properties using simple pharmacophore fingerprints, a type of merge that we call a ‘bioisosteric merge’, combining both the power of similarity and substructure searches (with the Fragment Network representing a form of substructure search).

As discussed in Section 1.4.3.2, the term ‘bioisosteres’ refers to compounds where a functional group or fragment has been swapped with another that exhibits similar or enhanced biological activity (Figure 3.1). Tools for performing bioisosteric replacements have been developed to suggest substitutions for a given functional group based on various properties. For example, the Craig Plot (Ertl, 2020) suggests bioisosteric replacements on the basis of hydrophobicity or electron donating and accepting power. Additionally, there are recommenders for bioisosteric ring (Ertl et al., 2022) and linker replacements (Ertl et al., 2023), which are designed to increase bioactivity. Our tool provides the ability to find bioisosteres within the context of finding fragment merges and linkers, requiring the introduction of novelty in how to join the bioisosteric replacements to create a physically reasonable compound that maintains the orientations of the inspirational fragments.

There are many 3D shape and pharmacophore-based tools that have been developed to enable the identification of structural analogues for a given compound based on these properties, explored in Section 1.4.4.4. These tools are useful as they can enable us to perform virtual screening of large compound libraries by finding analogues for a known hit. To our knowledge, there are few examples in the literature of the use of such approaches within the context of fragment linking or merging. These descriptors are more commonly used for analogue searching, as linking and merging often require the generation of novelty, though two recent methods begin to explore this: FRESCO (McCorkindale et al., 2022) and FrankenROCS (Correy et al., 2024), which are discussed in detail in Section 1.4.4.5.

The potential benefits of expanding the search to bioisosteric merges include improved productivity and hit rates and maximizing the chemical and functional diversity of follow-up compound designs in the early stages of the FBDD pipeline (Bender et al., 2021). Maintaining this diversity is important as it increases the chances of developing promising leads further down the line (as chemically similar molecules are likely to exhibit similar biological properties (Johnson et al., 1990)). Moreover, as a candidate hit is developed into a larger molecule, it becomes increasingly difficult for the chemist to make changes to the core scaffold, thus there is an advantage to having several alternative scaffolds to work with. Thus, early exploration of chemical space is important, enabling a comprehensive understanding of the SAR and of which substructures are required for (or will ablate) binding (Schuffenhauer et al., 2006; Bender et al., 2021). The approach described in this Chapter effectively combines elements of fragment optimization, whereby the structure of a fragment hit is optimized within fragment space prior to elaboration, with the traditional fragment elaboration techniques of merging and linking.

We find that our approach of expanding the search to bioisosteric merges enables a broader search of the chemical space represented in the Fragment Network, resulting in compounds that incorporate more unique substructures than when performing perfect merging alone. As a consequence, we identify merges that represent a greater number of fragment pairs and are able to access potential interactions that are not seen for the set of perfect merges. Our bioisosteric merges are predicted to adopt poses that indicate comparable shape and pharmacophoric overlap with the crystallographic fragments and thus represent a promising way of increasing diversity and improving the potential hit rate of follow-up compound design.

Comparison against a more traditional, 'out-of-the-box' docking approach using pharmacophoric constraints shows that our tool provides substantial benefits with regard to computational efficiency. We find, when comparing follow-up compounds proposed for specific fragment pairs, the time required for conformer generation to identify a merging 'hit' is on average 9.7-fold ±11.4 greater for the docking protocol. This suggests our method offers a practical and formulaic approach for enumerating follow-up compounds at scale using data from an entire crystallographic screen (rather than a few individual fragment pairs). A case study using retrospective experimental data demonstrates that our bioisosteric merging method mimics substructure variations that have been manually designed by chemists when exploring SAR.

3.3 Methods

3.3.1 Crystallographic fragment screening data

To test our merging pipelines, we used the results of XChem crystallographic fragment screens against three antiviral targets, the data for which can be freely downloaded from the Fragalysis platform (`https://fragalysis.diamond.ac.uk/viewer/react/landing`). These test cases were chosen as they represented active targets within the ASAP project; see Section 1.5.2 for more details (ASAP Discovery Consortium, 2024). As XChem is a member of this project, these targets provided useful platforms for method development due to the potential to receive feedback from other ASAP members looking to progress compounds against these targets (this is discussed further in Chapter 4). Moreover, the use of three different test cases was useful for testing the robustness of the pipeline to find merges for diverse scenarios that vary with regard to the numbers of crystallographic fragment hits and distribution of hits within the binding site (Figure 3.2).

The first two targets are proteases from two non-polio enteroviruses (EVs), which are members of the Picornaviridae family, EV D68 and EV A71. EV D68 is most commonly associated with mild respiratory illnesses, although has been linked (in rare cases) to acute flaccid myelitis (Hu et al., 2020). EV A71 is linked to hand, foot and mouth disease, which affects children, but can also cause severe neurological diseases, including encephalitis (Wang et al., 2022b). We looked specifically at the 3C protease (of EV D68) and the 2A protease (of EV A71). These proteases are involved in self-cleavage of the viral polypeptide into individual viral proteins and are also involved in the cleavage of several host cell proteins (Hu et al., 2020). For EV D68 3C protease, a total of 1,231 fragments were screened, including fragments from the DSI-poised (Cox et al., 2016), MiniFrags, Fraglites, PepLite, York3D, SpotXplorer and Covalent MiniFrags libraries (Lithgo et al., 2024b). We searched for merges for 25 hits found to bind to the catalytic site (Figure 3.2a). For EV A71 2A protease, we examined 6 hits that were found to bind to the S1 region of the active site (Figure 3.2b). At the time we ran the pipeline (December 2023), only the results of a partial fragment screen were available, hence the low number of fragment hits used. However, this is a valuable test case to include as it reflects the realistic setting of drug development for a target, where there may be several iterations of data release and follow-up compound design. For the full set of crystallographic results, a total of 1,129 fragments were screened from the same libraries as above (Lithgo et al., 2024a).

The third target belongs to the Zika virus, which is a flavivirus that has been linked to neurological disorders, including Guillain–Barré syndrome, and the increased risk of microcephaly in fetuses and newborns owing to prenatal Zika infection (Lei et al., 2016). The virus consists of an ∼11kb RNA genome, which encodes a large polyprotein and is processed into several structural proteins and non-structural proteins. We focused on the NS2B–NS3 protease, which plays an essential role in processing the polyprotein to produce the virus' enzyamtic components and the replication complex (Hill et al., 2018; Hilgenfeld et al., 2018). A total of 1,076 fragments were screened against the NS2B co-factor from the DSI-poised (Cox et al., 2016), MiniFrags Probing, CovHetFrags and SpotXplorer libraries (Ni et al., 2024). The fragment screen shows a dense population of fragments in the S1 region of the active site for NS2B, with 38 fragments found to bind, but providing several linking opportunities to one fragment in the S1′ site and 3 fragments in the S2 site (Figure 3.2c). All fragments used in our analysis are listed in Appendix Table B.1.

3.3.2 Updates to the Fragment Network database

The Fragment Network version used in this Chapter contains >200 million compounds from vendors, including Enamine, MolPort and Chemspace. It is an expansion of the versions described in (Wills et al., 2023; Hall et al., 2017). As described in Chapter 2, the network comprises nodes, representing molecules or substructures, which are connected by edges that denote transformations in which rings, linkers and substituents are added or removed between nodes (neo4j, 2024). For this new pipeline, we provide the option to run the database using a Docker container to allow an easy-to-use implementation for new users.

3.3.2.1 Pharmacophore descriptors for substructures involved in transformations

Compared with the Fragment Network used in Chapter 2, we have made one significant change involving the addition of pharmacophore properties to enable the user to identify transformations involving substructures with desired pharmacophore properties. To do this, the edges of the Frag-

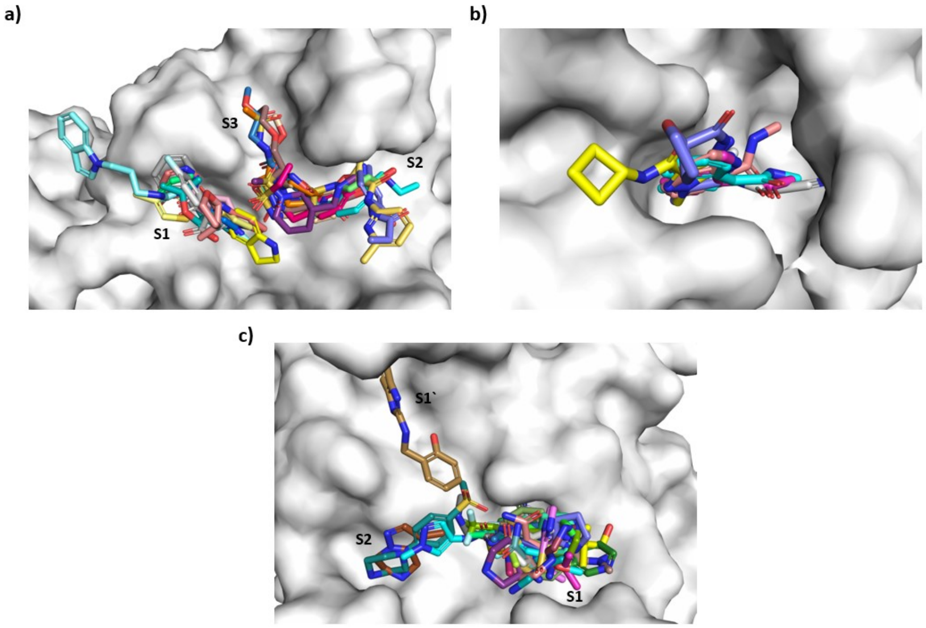


Figure 3.2: **Crystallographic fragment hits used for analysis.** The binding sites and crystallographic fragment hits used for testing the merging pipelines are shown for (**a**) enterovirus (EV) D68 3C protease, (**b**) EV A71 2A protease and (**c**) Zika NS2B. Individual subsites are labelled in **a** and **c**. Data can be freely downloaded from `https://fragalysis.diamond.ac.uk/viewer/react/landing`. The fragment codes are provided in the Appendix.

ment Network are now additionally labelled with 2D fingerprints describing the pharmacophore features of the substructure being added or removed in a transformation. This allows the creation of queries that search for merge compounds that incorporate substructures with similar pharmacophore features to a query substructure (the full querying process is described below). In the context of this Chapter, we focus on using 2D pharmacophore descriptors, as they have a reduced requirement for computational resources (as conformer generation is not needed) and we found that they were sufficient when searching for similar substructures (although the use of 3D fingerprints is still possible in graph database searches and may represent an interesting avenue for future research). Comparison of 2D versus 3D fingerprints has found that the former perform as well as state-of-the-art 3D methods for multiple tasks, including toxicity, solubility and simple binding affinity prediction (Gao et al., 2020a). For the updated algorithm, similarity metrics to determine appropriate replacement substructures are calculated on-the-fly as the database is queried.

The 2D pharmacophore fingerprint is calculated using the cheminformatics toolbox RDKit (Landrum, 2024). The fingerprint is generated by calculating topological distances between defined pharmacophore features, which are stored as bits in the fingerprint. We consider topological distances between pairs of pharmacophores, which are binned into three bins representing distances of 0–2, 2–5 and 5–8 bonds (to keep the length of the pharmacophore fingerprint short, resulting in fingerprints that were 84 bits in length). The pharmacophore features considered include hydrogen bond acceptors, hydrogen bond donors, negative ionizable, positive ionizable and aromatic groups. Two 'pseudo-pharmacophore' features were also added: aliphatic rings, as this was found to help maintain the shape of the substructure, and the attachment point atoms, which enables consideration of the location of the attachment point in relation to the rest of the substructure. Example substructure replacements are shown in Figure 3.3.

3.3.3 Enumeration of substructures for querying

In the previous iteration of the Fragment Network merge search (Chapter 2), the database was queried using pairs of fragments. In this new iteration, as discussed in Section 3.2, the search is performed for compatible pairs of substructures sourced from the fragments. Without controlling

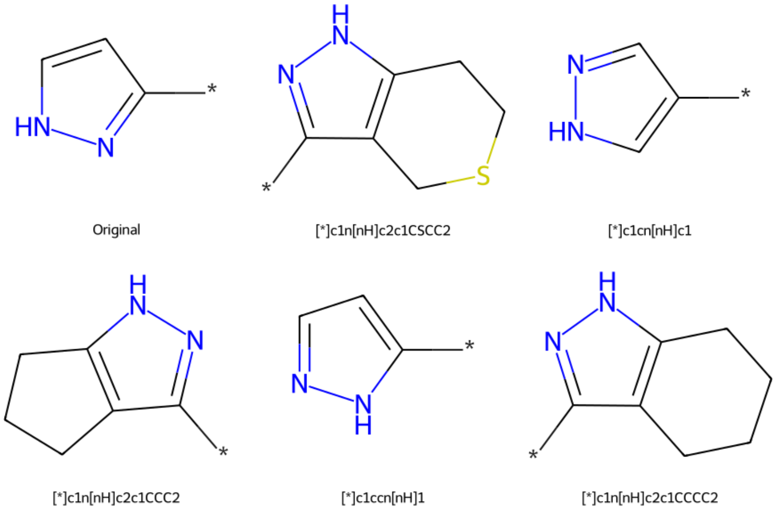


Figure 3.3: **Example substructure replacements used in bioisosteric merging.** An example substructure from a crystallographic fragment hit (top left) and five example replacement substructures found in the Fragment Network. The asterisk denotes the attachment point atom. The top-right replacement substructure for example is chemically identical to the original but differs in its attachment point.

which final substructures are incorporated into the merge, the merge may occur between substructures that are not spatially compatible, wasting computational time.

First, as in our previous version of the tool (Chapter 2), pairs of fragments are pre-filtered according to whether they share overlapping volume over a specified threshold, as these pairs do not represent good opportunities for merging (as they are unlikely to result in increased interactions). In this iteration of the tool, we removed pairs for which at least 56% of the volume of one fragment overlaps with the other. This threshold was determined by manually filtering pairs of fragments by visual inspection for an example fragment screen (and by calculating the threshold based on those that were filtered out). As in our previous iteration (Chapter 2), we additionally removed fragment pairs that show a minimum distance between the closest pair of atoms of $\geq$5Å to avoid the suggestion of long linkers. Both thresholds can be easily modified according to the needs of the user.

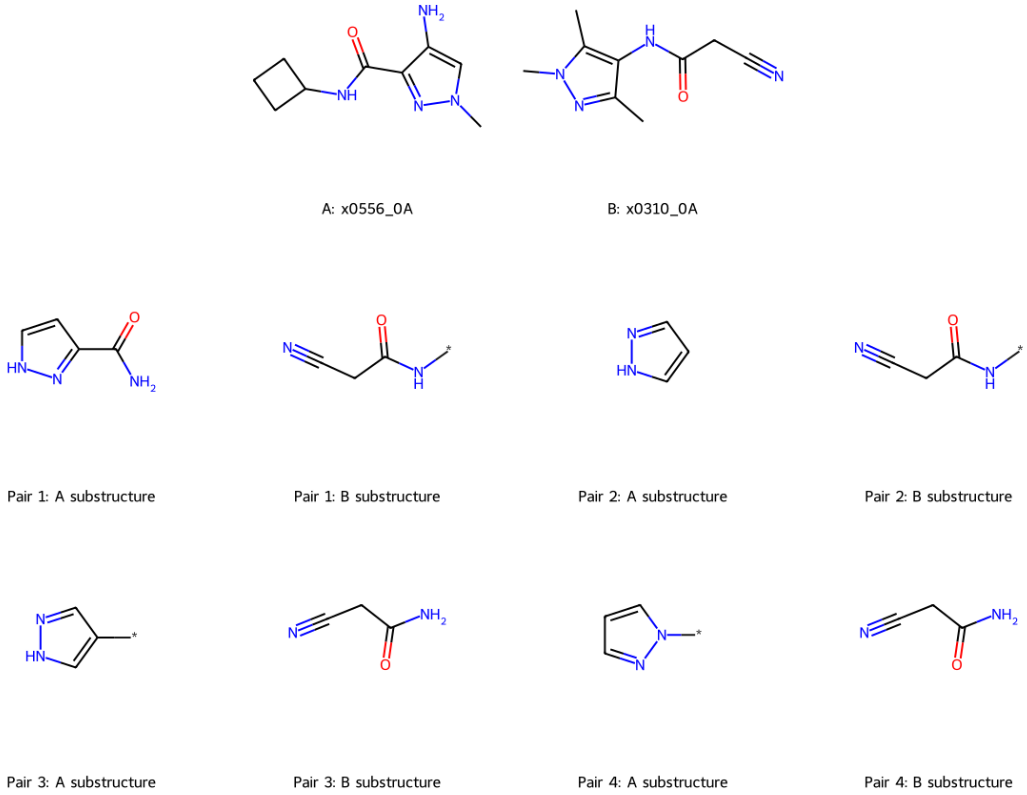
Following the enumeration of compatible fragment pairs, fragments are broken down by retrieving their constituent substructures from the network. Substructures are checked for whether they contain at least three carbon atoms (representing a substantial contribution to the final merge) and that they do not represent a single carbon-only ring (as these substructures are densely connected in the network and result in large numbers of molecules for filtering; this filter can be controlled by the user). To ensure the merges maximize binding, substructures are also filtered for those that make an interaction with the protein (calculated using ProLIF (Bouysset et al., 2021)). While PLIP (Adasme et al., 2021) was used for protein–ligand calculation in Chapter 2, we found that this tool could sometimes lead to errors in bond order during molecule preparation, which led to us using ProLIF instead. However, we recognize that interaction calculation is dependent on the software used, and each employs different thresholds for the detection of interactions.

We limited the number of compounds retrieved from the network, as some queries can result in large numbers of compounds for filtering. This is particularly common for 'terminal nodes' in the decomposition, meaning substructures that do not undergo any further contraction (such as single rings), which are extremely densely connected in the network. For each database query between substructures, results are limited to the first 500 compounds retrieved, although this limit is optional and can be easily modified. The parameters used in this work are described in Section 3.4.

Substructures are filtered to remove those for which the geometric mean of the overlapping volume for each substructure (meaning the percentage volume of each substructure that is overlapping with the other) is ≥30%. The set of non-redundant substructure pairs (as the equivalent substructures can be found in multiple pairs of fragments) is subsequently used for querying the database. Example pairs of substructures for querying are shown in Figure 3.4.

3.3.4 Querying the database

As the query is performed using substructures instead of fragments, the parameters of the search have been significantly modified compared with the previous iteration in (Chapter 2). The database query operates as follows: queries are performed for pairs of substructures (derived from two parent fragments) to be merged into a single compound. For clarity, we refer to the two substructures



A: x0556_0A

B: x0310_0A

Pair 1: A substructure

Pair 1: B substructure

Pair 2: A substructure

Pair 2: B substructure

Pair 3: A substructure

Pair 3: B substructure

Pair 4: A substructure

Pair 4: B substructure

Figure 3.4: **Example pairs of compatible substructures for merging.** Enumerated substructure pairs for querying are shown for fragment hits x0556-0A and x0310-0A against enterovirus A71 protease 2A. The substructures have been filtered according to their degree of overlap (<30% the volume of one fragment) and as to whether they make an interaction (calculated using ProLIF) (Bouysset et al., 2021).

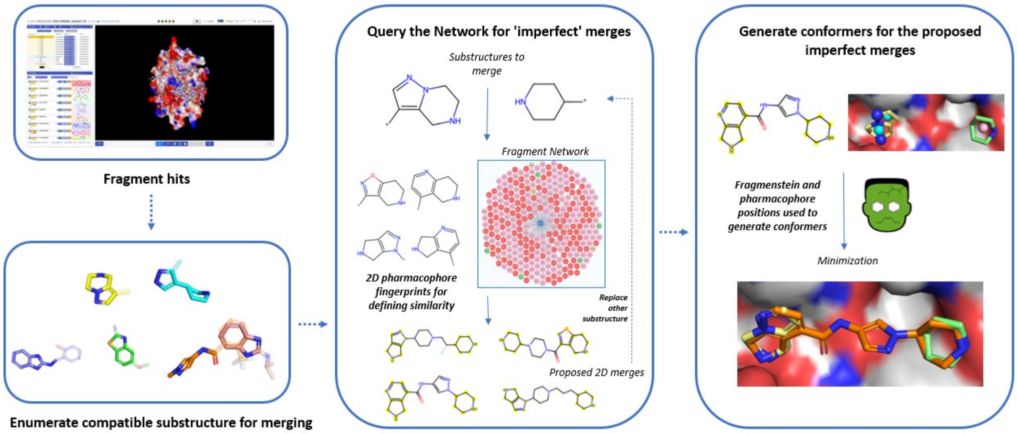

Fragment hits
Enumerate compatible substructure for merging
Query the Network for 'imperfect' merges
Substructures to merge
Fragment Network
2D pharmacophore fingerprints for defining similarity
Replace other substructure
Proposed 2D merges
Generate conformers for the proposed imperfect merges
Fragmenstein and pharmacophore positions used to generate conformers
Minimization

Figure 3.5: **Pipeline for identifying bioisosteric merges.** The overall pipeline for using the Fragment Network to find bioisosteric merges begins with the enumeration of compatible substructures from crystallographic fragment hits against a target. Pairs of substructures that are spatially compatible and make interactions with the protein (calculated using ProLIF (Bouysset et al., 2021)) are selected. The network is queried to find bioisosteric merges: given a pair of substructures, up to two expansion hops can be made away from the first substructure followed by incorporation of a pharmacophoric equivalent of the second substructure (calculated using similarity between pharmacophoric fingerprints implemented using RDKit (Landrum, 2024)). The first substructure can additionally be replaced in subsequent steps. Conformers are generated using an adapted version of Fragmenstein (Ferla et al., 2024); the replacement substructures are embedded using the pharmacophore coordinates from the original fragment substructures; these are used to generate custom atom-to-atom mappings that are fed into Fragmenstein for embedding the entire structure of the merge against the original fragments. Minimization is performed using PyRosetta (Chaudhury et al., 2010).

as substructure A and substructure B. Substructure A represents the ‘seed node’ in the query, and the query retrieves the database node representing this substructure to initiate the search. To identify perfect merges, from the seed node, up to two optional hops are made, in which an expansion is made, to generate diversity within the region linking the two substructures. A final hop is then made in which substructure B in the pair is incorporated into the compound (Cypher query examples can be found in Appendix Section B.1). A schematic demonstrating the key steps in the pipeline is shown in Figure 3.5.

3.3.4.1 Search for bioisosteric merges

To expand the search to identify ‘bioisosteric merges’, the query follows the same format, except the final expansion step is modified, instead specifying that the substructure to be incorporated need not be identical to the query substructure but should be similar in terms of its pharmacophoric features. As described above, the edges in the network are labelled with the pharmacophore fingerprint representing the substructure involved in the transformation. Thus, for a given query substructure, we can calculate similarity to substructures in the database by calculating similarity between their fingerprints, only accepting those with similarity above a specified threshold. The neo4j platform (neo4j, 2024) allows this calculation to be performed on-the-fly, enabling automatic filtering of the query results. For the purposes of our search, we used the Tanimoto metric to calculate similarity and retrieved all merges with similarity >0.9 (these query merges can be further prioritized at a later stage, but this threshold was found to result in sufficient numbers of compounds for filtering). This threshold can be adjusted according the requirements of the user.

To further increase diversity within the set of bioisosteric merges, promising compounds can be subjected to a further round of querying following initial conformer generation (Figure 3.6). To do this, merges for which an alignment has been successfully generated are selected (described below) and used to generate further bioisosteric merges by replacing the seed substructure in the query (also using pharmacophore fingerprint similarity). We provide options for a ‘strict’ and a ‘loose’ search. In the strict search, the seed substructure is removed, no modifications are made to the linker region in the merge and the seed substructure is replaced. In the loose search, an additional contraction and expansion are made within the linker region to generate further diversity and widen the search radius for merges (an example is shown in Appendix Figure B.1).

In the results presented here, the results of the perfect and bioisosteric merging pipelines are kept separate to enable comparison (meaning no perfect merges are included in the results of the bioisosteric pipeline).

As in Chapter 2, we also provide the option in the new pipeline to perform R-group expansion on the merges to recapitulate substituents seen in the original fragments (this is done in an automated

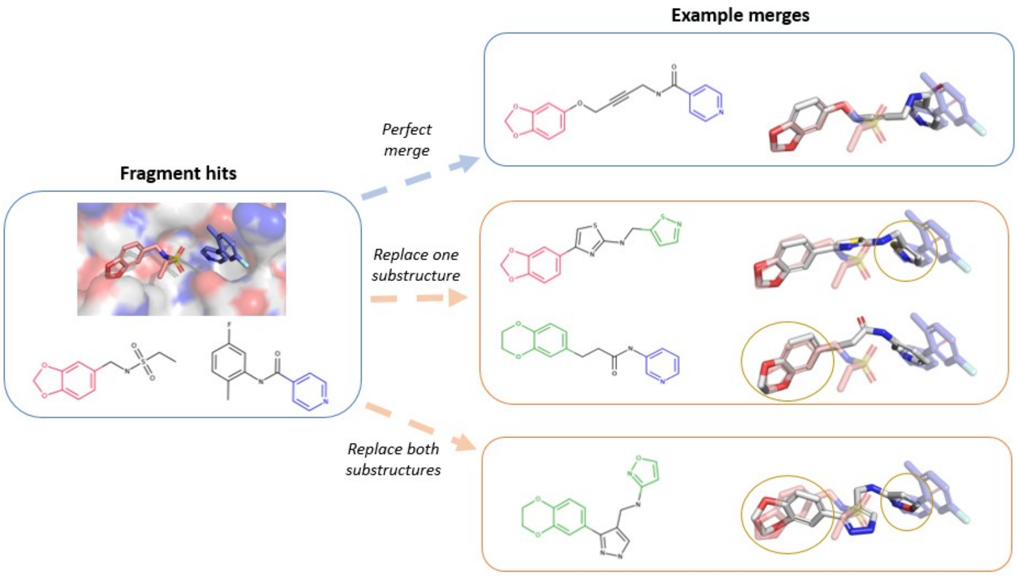


Figure 3.6: **Increasing diversity within fragment follow-up compounds.** Diversity within the follow-up compounds can be increased by replacing substructures with those that are 'pharmacophorically similar', determined by calculating similarity between pharmacophore fingerprints. Perfect merges maintain the exact substructures (shown in red and blue) seen within the parent fragments. Bioisosteric merges occur when we replace either one or both substructures from the parent fragments (shown in green), increasing chemical diversity within follow-up compounds.

manner requiring no manual curation of possible R-groups). In the results presented, R-group expansion has been additionally run for successfully placed merges (this can also be performed during the initial querying stage).

3.3.5 Prioritization of merges for filtering

Owing to the volume of potential merges retrieved from the database, we prioritized which compounds entered the computationally expensive step of conformer generation. For bioisosteric merges, we extracted all the replacement substructures that are used within the set of merges and generated embeddings using pharmacophores. Pharmacophores are extracted from both the original and replacement substructures and the replacement substructure is aligned to the original pharmacophores according to its 3D coordinates (a process that is relatively computationally cheap). This process results in several possible embeddings (based on which and how many pharmacophores are matched); we used the embeddings that demonstrate the most favourable shape and colour score (using an RDKit implementation, based on the $\mathrm{SC}_{\mathrm{RDKIT}}$ score described in Imrie et al. (2020)). The score aims to determine whether binding mode is conserved for elaborated compounds by calculating the overlap between the volume and pharmacophore features between the initial fragments and the elaborated compounds. Following this, possible interactions between the embedded replacement substructure and the protein were calculated using ProLIF (Bouysset et al., 2021) and compared against the interactions observed for the original substructure. We only maintained merges that incorporate substructures that are observed to make an interaction and ranked them according to the number of potential interactions that are maintained. According to the target and the size of the dataset, due to computational constraints, we can impose limits on the number of compounds that enter the conformer generation steps per fragment pair. In this work, we imposed a limit of 500 compounds per substructure pair.

3.3.6 Filtering generated merges

3.3.6.1 Conformer generation

The most important step of our filtering process is to evaluate whether the merges can adopt the conformations observed for the parent fragments. In Chapter 2, we used Fragmenstein (Ferla et al., 2024) to generate conformations, which predominantly relies on MCS matching and positional overlapping between the merge and fragments. Fragmenstein places the merge and can perform minimization within the protein with PyRosetta (Chaudhury et al., 2010). To place the merge, Fragmenstein computes an atomic correspondence between the merge and the fragments using a combination of the MCS and the atomic distance. This atomic mapping is then used to decide which atomic positions needs to be constrained for conformer generation and energy minimization. For our implementation, as the bioisosteric merges may no longer share exact substructures and may differ in shape (depending on the similarity threshold used), we have adapted the Fragmenstein protocol to perform placement of molecules using pharmacophores rather than molecules. In particular, a fast pharmacophore-based alignment is performed to identify the atomic mapping between the atoms of the fragment and the atoms of the molecule to be placed. This mapping is performed based on distances between atoms, mapping those that are within 1Å. The mappings can be provided to Fragmenstein, which subsequently attempts to generate conformations using these mappings; mappings for which Fragmenstein is unsuccessfully able to generate a conformation are discarded. We prioritized the mapped conformers (without minimization) with the highest $\mathrm{SC}_{\mathrm{RDKIT}}$ score to undergo subsequent minimization within the protein and selected the top 500 compounds to minimise per fragment pair).

Following conformer generation, we applied a series of filters to retain the most promising compounds. We only maintain merges that exhibit a negative ΔΔG value, a combined RMSD of $\leq 2\mathring{A}$ with the substructures used for placement, share overlapping volume with the substructures used for placement and are energetically favourable (according to the energy ratio between several unconstrained conformers and the constrained conformation generated by Fragmenstein, as calculated in the original pipeline (Wills et al., 2023)). These filters are less stringent than in the previous iteration of the pipeline, as we found in the case study described in Section 2.4.7.2 that strict filtering may result in the removal of favourable compounds. The pipeline has thus been tailored to enable an initial enumeration of sensible follow-up compounds that can be fed into the design–make–test cycle, and the parameters used have been made customizable according to the discretion of the user.

3.3.7 Comparison against pharmacophore-constrained docking

To compare with more ‘out-of-the-box’ standard techniques for computational follow-up of fragment screens, we also performed docking-based experiments for select fragment pairs. Owing to computational constraints, we did not perform docking of the entire equivalent set of >200 million compounds stored within the database (estimated to require >150,000 CPU days using our docking protocol). We instead extracted representative subsets to minimize the computational costs. Nine fragment pairs were used for comparison, consisting of three fragment pairs across each target that were well-represented amongst the results of the Fragment Network merging pipelines to provide sufficient numbers to enable comparison.

To extract a subset of compounds for docking, we performed a pharmacophore fingerprint-based search with the Tversky metric against every compound in the database, calculating the geometric mean of the similarity against each fragment in the pair. We used fingerprints calculated in a similar way as described above used in annotation of the network substructures using the same pharmacophore features (with the exception of the attachment point atoms). We then extracted the top 100,000 compounds for each pair. To provide a comparison search method that still considers the pharmacophores within the parent fragments, we used rDock (Ruiz-Carmona et al., 2014) to perform a virtual screen of these compound subsets, as rDock provides the ability to dock compounds using a set of specified mandatory and/or optional pharmacophore constraints.

To generate the set of pharmacophoric feature constraints to use for docking, for a given pair of fragments, the pharmacophoric features that are involved in interactions (calculated using ProLIF; Bouysset et al. (2021)) are extracted. The coordinates of the interacting atoms are recorded; for multi-atom pharmacophoric features, such as aromatic rings, the centroid of the coordinates is calculated. If a pair of pharmacophoric features are closer than 2Å, one of the features in the pair is removed, to ensure that pharmacophoric features are not too close to prevent finding sensible molecules. The interacting non-hydrophobic pharmacophoric features were used to generate a set of mandatory pharmacophoric feature constraints. Optional constraints were also imposed, where up to two of the hydrophobic pharmacophoric features are required to be fulfilled to successfully generate a conformation. A tolerance radius of 2Å was used for the pharmacophoric feature alignment. The parameters used for docking are provided in Appendix Table B.2. The full sets of pharmacophoric features used in our test cases as docking constraints can be found in Appendix Tables B.3–B.5.

3.3.8 Case study

To validate our method, we conducted a retrospective case study using hits that were screened against repeat-containing protein 5 (WDR5)–MYC (Chacón Simon et al., 2020). A merging opportunity was identified by overlaying a fragment-screening hit and a hit identified using HTS, for which a compound (compound 2i) was designed with a K_D value of 1.0μM. The authors also designed several analogues of the merged compound to optimize affinity by exploring several different

R-groups. To evaluate whether our bioisosteric pipeline is able to replicate some of the design ideas used in the optimization and thus could act as a useful tool for early SAR exploration, we ran our perfect and bioisosteric merging pipelines and provide comparison of the outputted compounds against the existing designed compounds with experimental data, consisting of K_{D} values from a fluorescence polarization-based assay.

3.4 Results

3.4.1 Enumeration of substructure pairs for merging

We tested our bioisosteric and new perfect merging pipelines using crystallographic fragment hits against three different antiviral targets: EV D68 3C protease, EV A71 2A protease and Zika NS2B.

For the EV D68 3C protease, we considered 25 fragment hits found to bind to the catalytic site of the protein. After running our enumeration pipeline (as described in Section 3.3.3), we found 152 pairs of fragments (querying is asymmetric; thus, the same fragment pair may be represented twice), leading to 1,128 pairs of substructures for querying. Totals for numbers of enumerated pairs are in Table 3.1. Timings for querying and conformer generation can be found in Appendix Tables B.6 & B.7, respectively.

In the case of the EV A71 2A protease, at the time we ran the pipeline (December 2023), only the results of a partial fragment screen were available, meaning we focused on only six compounds found to bind to the main active site of the target. The enumeration pipeline resulted in 11 fragment pairs and 88 pairs of substructures for querying.

For Zika NS2B, due to the dense population of fragments within the S1 site (with 37 fragments bound; Figure 3.2c), we focused only on merges that occur between subsites (that is, between S1 and S1′ or between S1 and S2) to find molecules that bridge across multiple sites. This enumeration resulted in 169 fragment pairs leading to 1,026 pairs of substructures for querying.

Table 3.1: Enumerated pairs of substructures for merging

Target	Number of fragment hits	Enumerated number of fragment pairs	Enumerated number of substructure pairs
EV D68 3C protease	25	152	1,128
EV A71 2A protease	6	11	88
Zika NS2B	41	169	1,026

EV, enterovirus.

3.4.2 Bioisosteric merging identifies potential follow-up compounds across all targets

Fragmenstein was used for generating conformations for our 2D proposed merges based on the parent fragments. Substantial redundancy was observed across the compound set as the same compound may be suggested for different pairs of fragments; however, the same compound was run through the filtering pipeline for each pair of fragments as the coordinates to where the design needs to be placed depend on the fragments, affecting the probability of success. Appendix Figure B.2 shows the numbers of unique filtered compounds that were successfully placed using Fragmenstein and passed our filters for each merging pipeline (all values for how many compounds pass through each step are shown in Table 3.2). Comparable numbers of compounds are identified when comparing the collective results from both rounds of bioisosteric merging with the results of perfect merging (across all targets). The efficiency of the search (defined as the percentage of compounds that enter minimization that are successfully placed with a $\mathrm{SC}_{\mathrm{RDKIT}}$ value of $\geq$0.55; see Section 3.3.5 for definition) ranged between 0.7–3.3%, with the lowest efficiency observed for the second round of bioisosteric merging for the EV A71 2A protease. The $\mathrm{SC}_{\mathrm{RDKIT}}$ threshold of 0.55 was chosen as this was found to correspond to an RMSD threshold of <2Å by Leung et al. (2019), a commonly used measure to determine whether docking was successful.

Upon visual inspection, the bioisosteric merges show a high degree of volume overlap with the fragments used for inspiration and many of the merges are observed to have high shape similarity (for example, overlapping rings) with the original fragments (particularly where replacement substructures maintain the rings seen in the original substructures), which indicates that bioisosteric merging still shows promise for recapitulating the original fragment poses. Examples of bioisosteric merges are shown in Figure 3.7, in which the computationally predicted structures of the merges mirror the structures of the crystal fragments and are predicted to maintain interactions seen in both of the parent fragments.

Table 3.2: Numbers of molecules passing through the merging pipeline

Target	Method	N molecules from database	N molecules entering minimization	N placed molecules	N filtered molecules	N filtered molecules (after R-group exp)
EV D68 3C protease	P	129,328 (65,753)	129,328 (65,753)	30,063 (18,387)	1,849 (1,103)	1,860 (1,114)
	B1	217,411 (99,245)	188,599 (91,139)	27,865 (15,532)	1,351 (750)	1,355 (754)
	B2	60,569 (14,577)	60,569 (14,577)	18,927 (4,709)	900 (282)	903 (284)
EV A71 2A protease	P	22,747 (13,351)	22,747 (13,351)	1,555 (1,360)	156 (137)	178 (153)
	B1	32,297 (20,106)	26,123 (16,245)	1,440 (1,104)	193 (161)	221 (179)
	B2	2,631 (1,836)	2,631 (1,836)	350 (248)	13 (12)	15 (14)
Zika NS2B	P	145,401 (74,263)	145,401 (74,263)	30,349 (17,150)	2,953 (1,564)	3,050 (1,640)
	B1	171,528 (71,431)	156,658 (61,450)	24,075 (11,601)	2,039 (1,159)	2,108 (1,211)
	B2	12,690 (4,898)	12,690 (4,898)	5,135 (1,702)	674 (161)	677 (163)

Brackets indicate the numbers of unique molecules. B1, first round of bioisosteric merging; B2, second round of bioisosteric merging; EV, enterovirus; P, perfect merging.

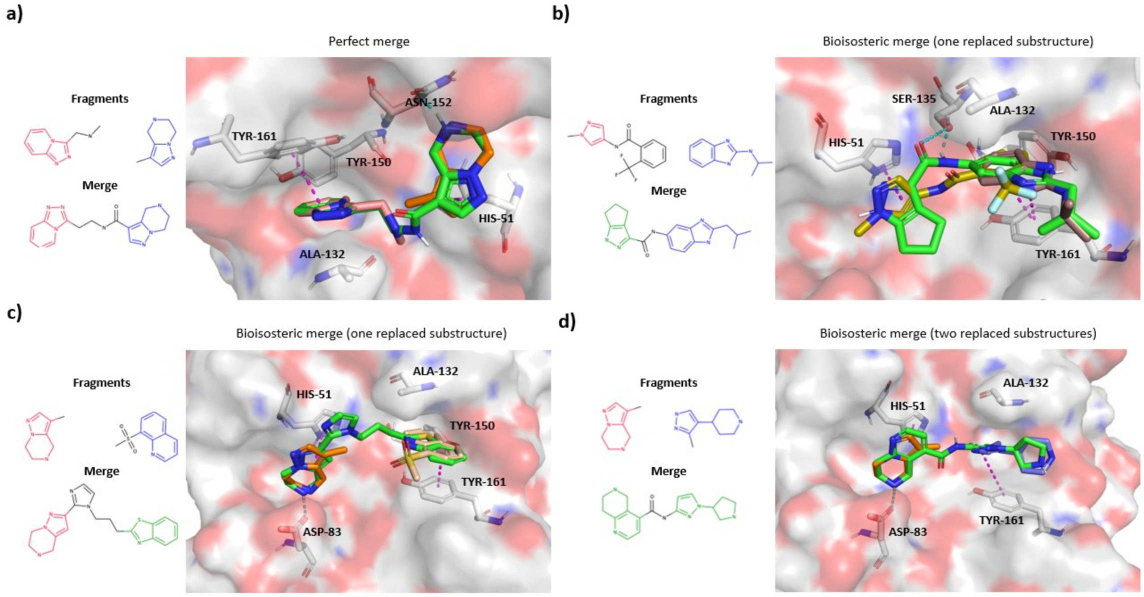


Figure 3.7: **Predicted merges of fragment hits against Zika NS2B.** Example merges against Zika NS2B are shown (Fragmenstein-predicted structures in green), overlaid with the crystallographic fragment hits used as inspiration for their design. The Fragment Network is used to identify (**a**) perfect merges, which contain exact substructures from two fragment hits, (**b,c**) bioisosteric merges that incorporate one replaced substructure and one exact substructure from the fragments and (**d**) bioisosteric merges that incorporate two replaced substructures. Residues predicted to be involved in interactions are shown. Hydrogen bonds are shown in cyan and π-stacking interactions are shown in magenta; hydrophobic interactions are not shown.

3.4.3 Bioisosteric merging expands the representation of fragment pairs

In the following sections, we compare various metrics across compounds found via bioisosteric merging (after both rounds of bioisosteric merging) versus perfect merging. To rule out compounds that do not represent favourable merges, the analysis is limited to compounds that have a favourable degree of volume and pharmacophore feature overlap using an $\mathrm{SC}_{\mathrm{RDKIT}}$ score threshold of $\geq$0.55 (Malhotra et al., 2017), and that are predicted to replicate an interaction observed for each original fragment (thus representing true merges). To remove redundant compounds (with the same SMILES), compounds are ordered according to their $\mathrm{SC}_{\mathrm{RDKIT}}$ and the SMILES with the highest $\mathrm{SC}_{\mathrm{RDKIT}}$ scores are maintained. The total numbers of compounds and their observed features are shown in Figure 3.8. Across all targets, bioisosteric merging is found to represent fragment pairs that were not represented by perfect merging and, for EV D68 3C protease and Zika NS2B, perfect merging found merges for pairs not represented by bioisosteric merging. When comparing the number of individual fragments represented, for EV D68 3C protease, both methods represent features from the same 16 fragments, while, for the EV A71 2A protease, bioisosteric merging identifies merges for 2 fragments not shown in the perfect merging set and both methods for Zika NS2B identify merges for a fragment not represented by the other. Bioisosteric merging is, therefore, valuable for expanding the search space of the Fragment Network search and identifies opportunities for fragments that are not well-represented by merging exact substructures alone.

3.4.4 Bioisosteric merging identifies potential interactions not represented by perfect merging

When we compare the unique interaction types that are predicted to be made by merges proposed by the bioisosteric versus perfect merging pipelines referring to the number of residue-level interactions made by each compound within the set, the results suggest that bioisosteric merging does not degrade the ability of the method to find merges that replicate the observed interactions. Bioisosteric merges perform similarly in terms of the mean total number of interactions made, dependent on target, being marginally higher (as observed for EV D68 2A protease and EV A71 2A protease) or slightly lower (for Zika NS2B) (Appendix Figure B.3). Similarly, when considering the propor-

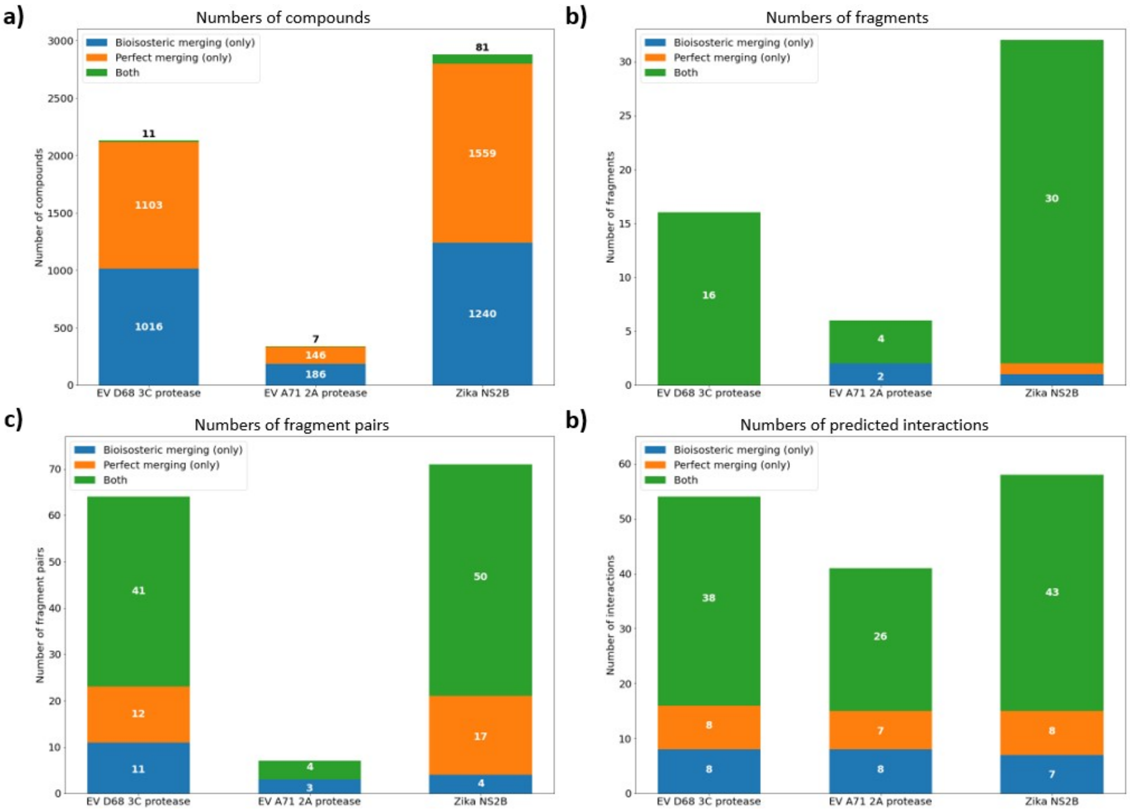


Figure 3.8: **Numbers of fragments, pairs of fragments and predicted interactions represented by merges found using perfect or bioisosteric merging.** For the top-performing merges (SC$_{\text{RDKIT}}$ score $\geq$0.55 and are predicted to maintain an interaction seen for both parent fragments), we show (**a**) the number of compounds identified, (**b**) the number of individual fragments used in the design of any of the compounds, (**c**) the number of fragment pairs for which it was possible to find at least one compound and (**d**) the total numbers of unique interactions predicted to be made by the entire compound sets. Colours indicate whether the compound, fragment, fragment pair or interaction was identified by compounds only retrieved via perfect merging, bioisosteric merging or both methods. Numbers represented by bioisosteric merging only are shown in blue, perfect merging only in orange and both techniques in green.

tion of preserved interactions, no technique performs consistently better than the other (Appendix Figure B.4).

Figure 3.8d shows, for each target, the number of interactions predicted to be made across the compound sets that are only observed for either the perfect or bioisosteric pipelines. Depending on the target, bioisosteric merges are found to identify between 7 and 8 interactions that are not represented with perfect merging alone, demonstrating that this method is useful for expanding the potential interaction space explored.

3.4.5 Bioisosteric merging represents more unique substructures than perfect merging

Figure 3.9 shows the number of unique substructures that are found among merges from both search techniques (that is, the number of substructures either taken directly from the fragments or that have been proposed as bioisosteric replacements, defined according to the SMILES string). Bioisosteric merging results in a substantial increase in the number of substructures represented, showing an average 3.4-fold ±1.3 increase across all targets. This demonstrates how bioisosteric merging can expand the search space and thus, potentially, the chemical diversity seen in the final compounds (Figure 3.10).

3.4.6 Bioisosteric merging increases the chemical space coverage of the search

One aim of bioisosteric merging is to access new areas of chemical space that are not explored when considering only perfect merges. Figure 3.11 shows low-dimensional representations of chemical space using t-SNE plots (Maaten et al., 2008) (calculated using Morgan fingerprints of length 1,024 bits and radius of 2; the dimensionality of the vectors was reduced to 50 using a principal component analysis before applying the t-SNE). For all targets, the plots show a separation between the regions of chemical space that are occupied by bioisosteric and perfect merges. There are clusters of compounds for bioisosteric merges that are not accessed by the perfect merge search (as well as some overlapping regions). To explore this further, we clustered the outputs of bioisosteric

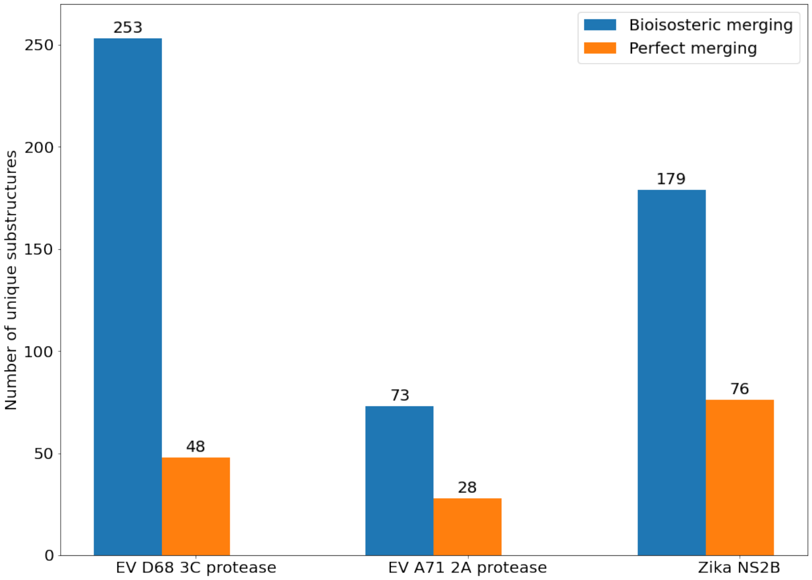


Figure 3.9: **Bioisosteric merging increases the numbers of unique substructures represented in follow-up compounds.** The numbers of unique substructures contributing to the final sets of merge compounds are shown across all targets. Bioisosteric merging increases the numbers of substructures observed compared with perfect merging. EV, enterovirus.

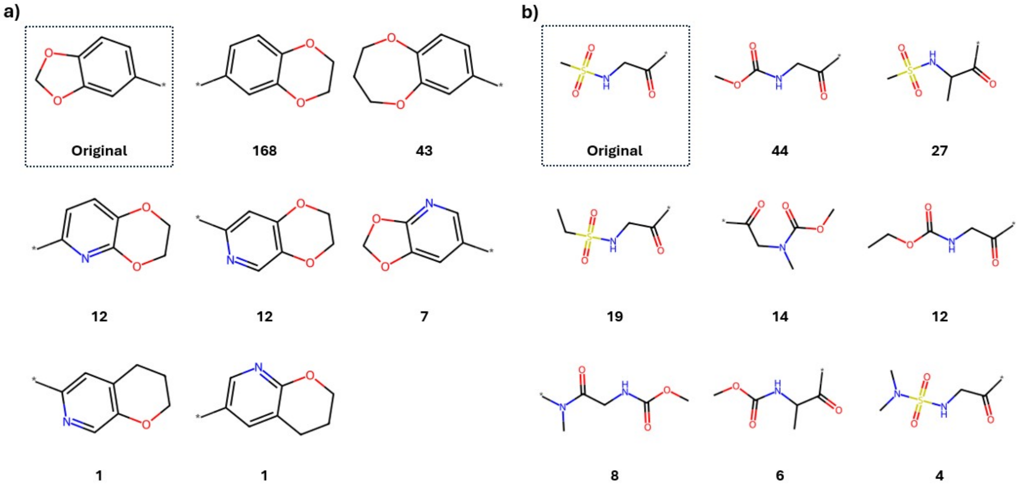


Figure 3.10: **Example substructure replacements from bioisosteric merging.** Example substructure replacements seen for two substructures used to form merges from the enterovirus D68 3C protease fragment screen. An example for a substructure containing (**a**) fused rings and (**b**) no rings are shown, together with the original substructures. Numbers indicate the number of compounds that contain the replacement substructure. The results are shown to include all compounds after R-group expansions.

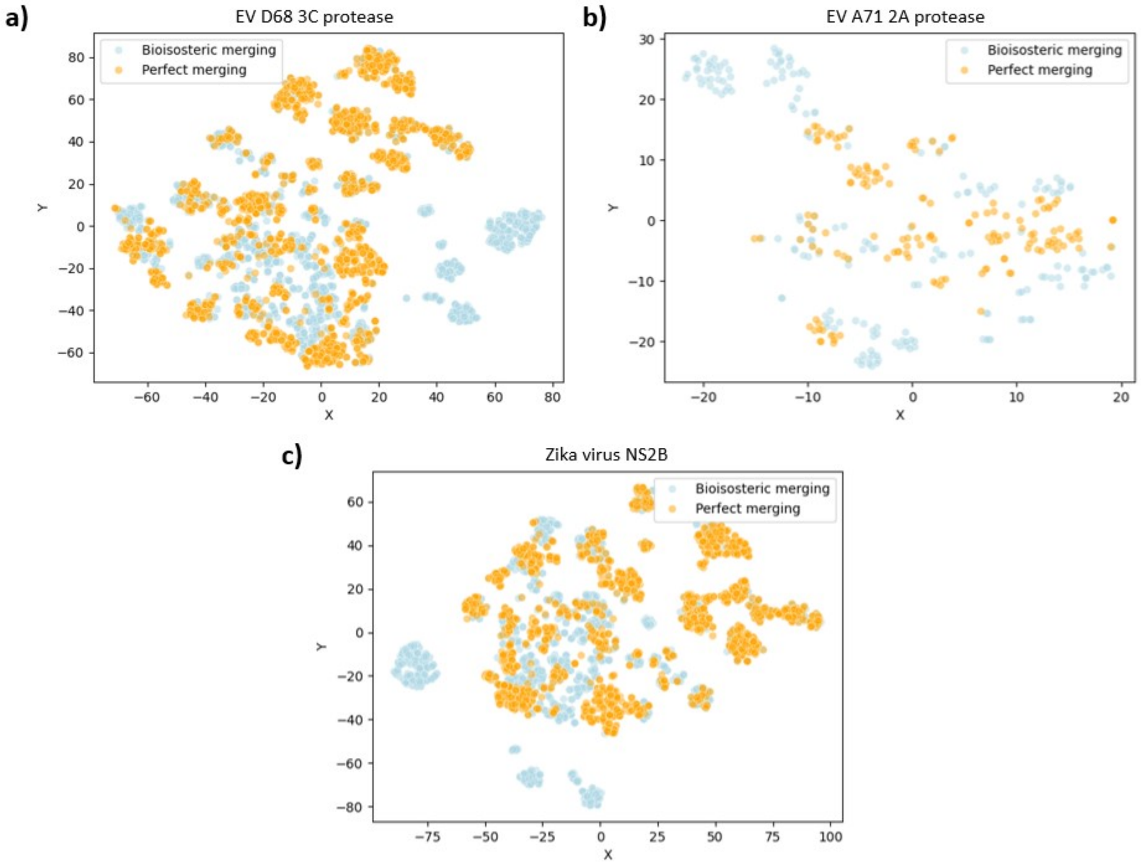


Figure 3.11: **Chemical space coverage by Fragment Network merges.** t-SNE plots depict low-dimensional representations of chemical space for filtered merge compounds identified for targets (**a**) enterovirus (EV) D68 3C protease, (**b**) EV A71 2A protease and (**c**) Zika virus NS2B. Molecules are represented using Morgan fingerprints of length 1,024 bits and a radius of 2.

and perfect merging for each target (Table 3.3), which shows the results of clustering using the Butina algorithm (Butina, 1999) with two different distance thresholds (calculated using Tanimoto similarity between Morgan fingerprints). Overall, this analysis showed that far fewer clusters contain compounds proposed by both techniques and bioisosteric merging was able to identify hundreds of clusters that contain no perfect merges, indicating that the method expands the chemical space searched.

Table 3.3: Clustering analysis for bioisosteric and perfect merging pipelines.

Distance threshold	Cluster composition	EV D68 3C protease	EV A71 2A protease	Zika NS2B
0.3	P + B	39	7	92
	P	801	122	1,226
	B	775	133	1,019
0.4	P + B	52	9	129
	P	547	85	735
	B	551	96	644

Numbers show the numbers of clusters containing compounds from perfect and/or bioisosteric merging pipelines after clustering all high-scoring filtered compounds. Compounds are clustered using Butina (Butina, 1999) clustering and Morgan fingerprint (1,024 bits; radius 2). B, bioisosteric merging; EV, enterovirus; P, perfect merging.

3.4.7 The Fragment Network merging pipeline offers potential efficiency benefits over pharmacophore-constrained docking

To enable comparison with an out-of-the-box docking approach, we performed pharmacophore-constrained docking using rDock for nine pairs of fragments, consisting of three fragment pairs against each protein target. The pharmacophore constraints were selected according to pharmacophore features that were responsible for making interactions (calculated using ProLIF; constraints are provided in Appendix Tables B.3–B.5). For each merge that passed through the rDock pipeline, the conformer with the best docking score was selected. The SC_{RDKIT} scores are consistently higher across all pairs for Fragment Network-identified compounds (Appendix Figure B.5), which can be attributed to several factors: the Fragment Network approach aims to maintain atoms (or similar substructures) directly inspired by the original fragments, but also the conformation generation method (using Fragmenstein) deliberately attempts to align the compounds with the parent fragments as closely as possible, whereas rDock conformers are ranked according to its scoring function that, although it considers pharmacophoric overlap as a term, also leverages energy-based terms. As a consequence, poses are not enforced to keep the pharmacophoric position of the parent fragments (example conformations for both methods are shown in Figure 3.12). Thus, to enable comparison between methods, we employed a less conservative SC_{RDKIT} threshold by filtering out compounds

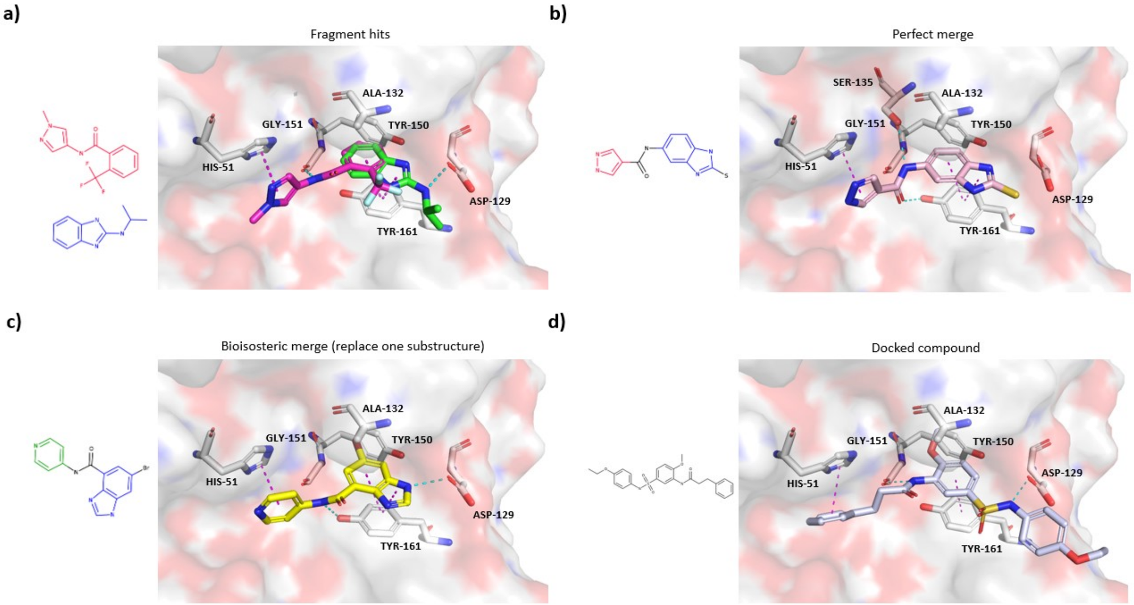


Figure 3.12: **Examples of high-scoring merges from Fragment Network and rDock pipelines.** Example merges of (**a**) two fragment hits that bind to Zika NS2B are shown. Merges have been proposed using different methods, including (**b**) a perfect merge, which maintains exact substructures seen in the fragments; (**c**) a bioisosteric merge, in which one substructure has been replaced; and (**d**) a merge from a pharmacophore-constrained docking protocol. All three merges are predicted to maintain interactions observed for both original fragments (hydrogen bonds are shown in cyan and π-stacking interactions are shown in magenta; hydrophobic interactions are not shown). 2D structures are provided; substructures that are maintained from the fragments in the merge are shown in red and blue while replacement substructures are shown in green.

with a $\mathrm{SC}_{\mathrm{RDKIT}}$ score of less than 0.4, to increase the number of rDock compounds to be used in our comparisons. Compounds are instead ranked according to the fraction of preserved interactions (meaning the proportion of unique residue-level interactions made by the parent fragments maintained by the merge); comparisons in Appendix Figures B.5-B.10 and Appendix Table B.8 are made between the top 100 compounds from both methods.

When comparing the chemical space coverage of the two techniques, it is noticeable across all pairs that the molecules identified by each technique occupy distinct areas of chemical space (Appendix Figure B.6). Clustering analysis also shows that the docking-identified compounds are more chemically diverse, occupying more clusters across all pairs of fragments than Fragment Network-identified compounds, and no clusters were found to include compounds from both techniques (Appendix Table B.8). Regarding the interactions that are predicted to be made by the follow-up compounds, neither technique identifies compounds that are predicted to make a greater number of interactions consistently across all pairs when normalized according to heavy atom count (Appendix Figure B.8).

To compare the efficiency of the two search approaches, we evaluated the ability of the two methods to identify merging 'hits'; we have defined this as compounds with a $\mathrm{SC}_{\mathrm{RDKIT}}$ score of at least 0.4 and that preserve at least 50% of unique residue-level interactions observed for the fragments (representing an increase from the number of interactions possible from a single parent fragment). We chose this less stringent $\mathrm{SC}_{\mathrm{RDKIT}}$ score for the reasons stated above (to enable comparison with the rDock technique). While true evaluation of whether compounds constitute a hit requires experimental validation in binding assays, this metric provides a general indication of the resources required to identify the most promising compounds. This definition of computational merging hits helps us discern which method may be more efficient for identifying high-scoring compounds that could contribute to a potential shortlist for purchase or synthesis. Table 3.4 shows, per fragment pair, the number of compounds (and associated CPU hours) required to yield the number of identified computational merging hits. While the docking protocol is efficient at identifying high-scoring compounds for certain pairs, in particular for those against the EV A71 2A protease (Figure 3.13), the Fragment Network pipeline requires less resources in terms of the CPU hours per hit across 8 of 9 pairs (an average of 9.7-fold ±11.4 reduction across all pairs). We hypothesize that the improved performance of the docking protocol for EV A71 2A protease may be due to the increase in overlapping volume between the fragments, meaning there are fewer (and closer) pharmacophores provided as docking constraints.

These results indicate that — while constrained docking can be a powerful method for follow-up development in certain contexts and does not have the same memory requirements for implementing the Fragment Network infrastructure — either larger numbers of molecules have to be screened, necessitating substantial compute, or more stringent pre-filtering methods are needed to reduce the library size to manageable levels. While the docking approach is adept at identifying highly diverse follow-up designs, the docking experiments conducted were not sufficient to consistently identify more compounds that maximize interactions across all of the targets when scaled by computational time. While both pipelines could be further optimized, and experimental validation is needed to fully assess the performance of each method, the ability of our tool to identify a comparable number of promising compounds with substantially less filtering time indicates it is well-suited to scenarios in which we want to quickly leverage all potential merging opportunities from a fragment screen.

3.4.8 Case study: our pipeline replicates design choices made during SAR analysis

To provide retrospective validation using existing experimental results for manually designed fragment merges, we compared designs proposed by our perfect and bioisosteric merging pipelines with the results of inhibitors designed against WDR5–MYC (Chacón Simon et al., 2020). In the original paper, the authors designed and optimized a merge of a fragment-screening hit (referred to as F2), and a hit identified by HTS (compound 1). The merge incorporates features from both hits and performs some additional optimization around the benzene ring of compound 1 through the addition of a hydroxyl group and a chlorine atom (Figure 3.14a,b). Several analogues were also designed that build a picture of the SAR by exploring different R-groups around the molecule.

In the original work by Chacón Simon et al. (2020), the analogue design process included the addition of alternative heterocycles to the imidazole ring from hit F2; there were two transformations whereby a pyrazole ring was added with different arrangements of nitrogens around the ring (referred to as substituents 3b and 3d).

We ran our pipeline to find catalogue merges incorporating substructures from F2 and compound 1. The initial parameters for filtering substructure pairs were loosened, such as removing substructures composed of single carbon rings from querying, due to their presence in the parent compounds and as we do not face the same computational limitations when only searching for merges for a single pair of fragments (these parameters can be freely modified by the user according to their requirements). We find that our pipelines proposed several merges that incorporate these pyrazole

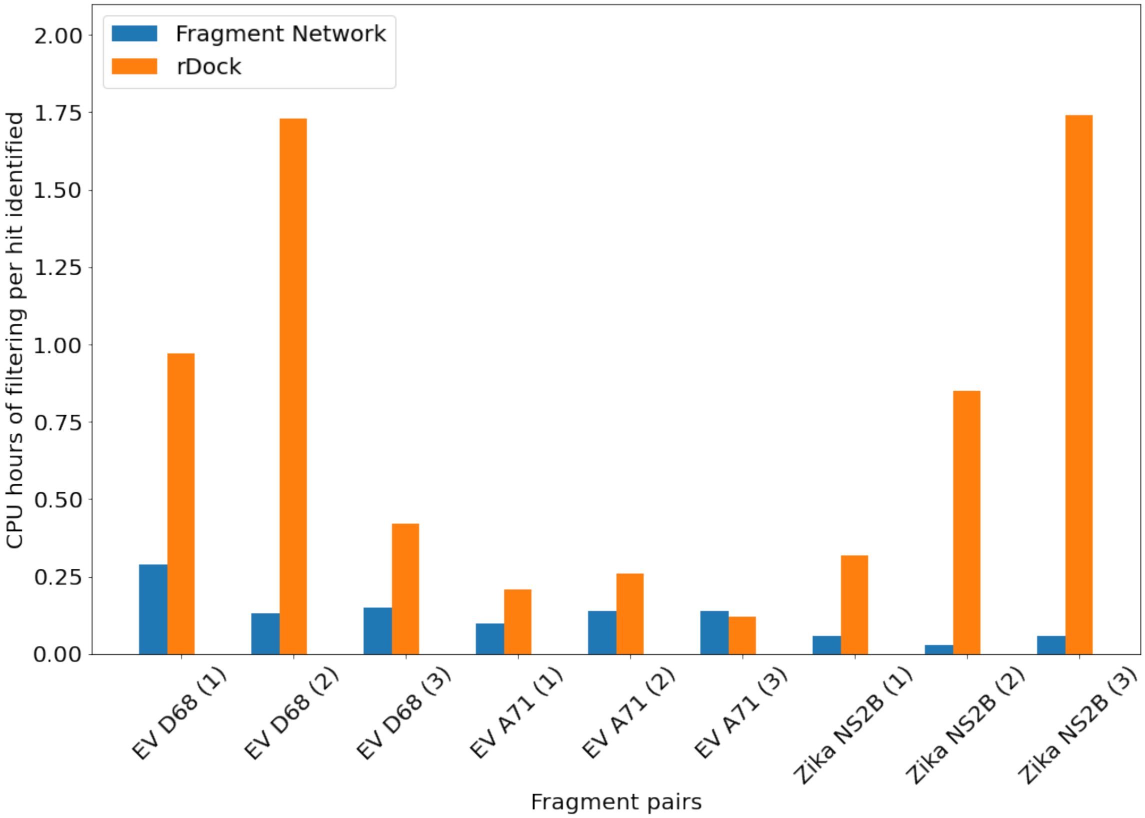


Figure 3.13: **CPU hours of filtering required per hit identified for pharmacophore-constrained docking versus Fragment Network merging.** The CPU hours required for conformer generation to identify a high-scoring merging 'hit' are shown using our Fragment Network merging pipelines (using 3D placement and minimization with Fragmenstein) versus pharmacophore-constrained docking following a similarity search using a pharmacophore fingerprint. The 'hits' are defined as compounds that preserve >50% interactions observed in the parent fragments and demonstrate a mean $\mathrm{SC}_{\mathrm{RDKIT}}$ of at least 0.4. EV, enterovirus.

Table 3.4: Pharmacophore-constrained docking versus Fragment Network merging

		Fragment Network pipelines				rDock pipeline			
Target	Fragment inspiration	Time (hours)	Number of compounds filtered	Number of 'hits'	Time per 'hit'	Time (hours)	Number of compounds filtered	Number of 'hits'	Time per 'hit'
EV D68 3C protease	x1071_0A x1498_0A	107	2,824	374	0.34	976	100,000	996	0.97
	x1083_0A x1498_0A	94	3,850	745	0.13	1,950	100,000	1,300	1.73
	x1140_0A x1498_0A	131	2,637	870	0.15	1,204	100,000	3,052	0.42
EV A71 2A protease	x0310_0A x0416_0A	41	9,595	406	0.10	1,532	100,000	7,981	0.21
	x0310_0A x0556_0A	59	9,674	407	0.14	1,013	100,000	4,143	0.26
	x0416_0A x0556_0A	39	8,985	279	0.14	1,183	100,000	10,765	0.12
Zika NS2B protease	x0089_0B x1098_0B	42	4,609	693	0.06	452	100,000	1,314	0.32
	x0429_0B x1098_0B	40	3,036	1,211	0.03	385	100,000	439	0.85
	x0687_0B x0969_1B	44	1,932	795	0.05	1,107	100,000	692	1.74

EV, enterovirus. Time refers to the total CPU hours required for conformer generation. 'Hits' refers to filtered compounds with a SC_{RDKIT} score of ≥0.4 and maintain ≥50% unique residue-level fragment interactions.

rings while maintaining the cyclopentane ring (found to have favourable effects for binding affinity in the original paper (Chacón Simon et al., 2020)). This included 12 compounds that incorporate the pyrazole ring arranged as in substituent 3b, and 6 compounds that incorporate the pyrazole ring arranged as in substituent 3d. Our pipelines also identify compounds that incorporate various substituents added to the benzene and pyrazole rings, which may provide further opportunities for elucidating SAR and suggesting potential optimizations. We include an R-group decomposition representing these proposed molecules and the various added substituents in Appendix Figures B.11 & B.12.

In addition to being able to recapitulate some of the manual design ideas, Figures 3.14c,d, show how the generated conformers demonstrate reasonable alignment with the crystallographic structures of the original fragments and the manually designed merge. This demonstrates how some of the substitutions incorporated in our bioisosteric merging pipeline can mirror some of the decisions made by chemists when exploring SAR for a given target.

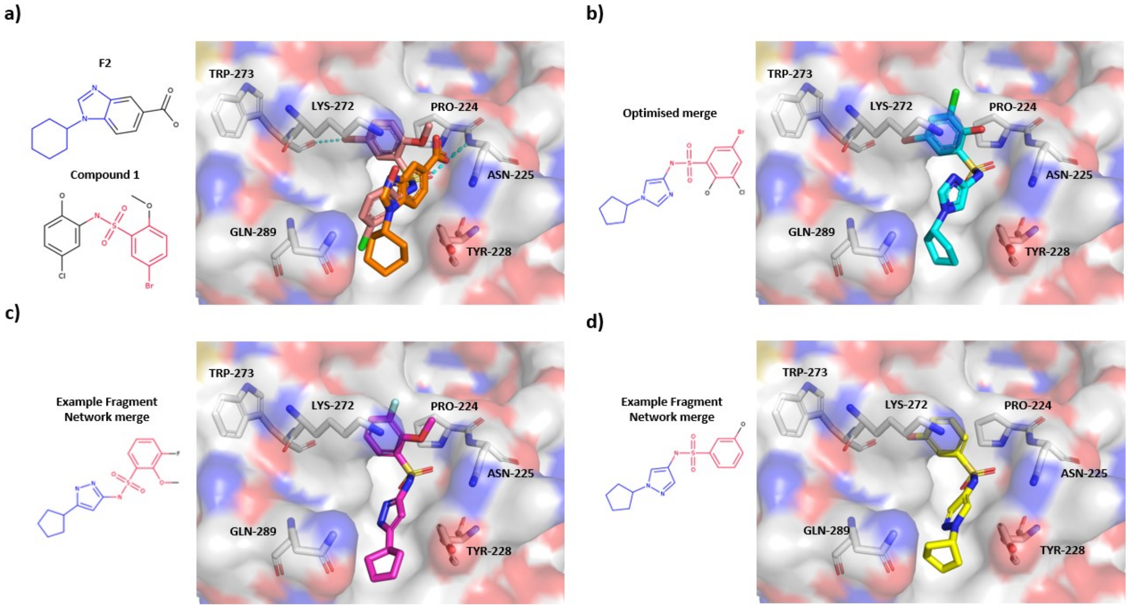


Figure 3.14: **Case study involving merges designed against WDR5–MYC.** (**a**) Crystal structures of two hits identified by fragment screening (compound F2; orange) and high-through screening (compound 1; pink) against WD repeat-containing protein 5 (WDR5)–MYC that were used as inspiration for merge design. (**b**) Crystal structure of the original manually designed merge; further optimization is performed through substituting in a hydroxyl group and addition of a chlorine atom. (**c,d**) Example merges proposed by the Fragment Network pipeline that replace the imidazole ring seen in the original fragment with pyrazole rings with different arrangements of nitrogen atoms. Conformations have been generated using Fragmenstein.

3.5 Discussion

In this work, we describe an update to our method for merging fragments using the Fragment Network that maximizes chemical diversity by incorporating substructures that recapitulate pharmacophoric features from the original fragments. Our bioisosteric merging method widens the search space for follow-up compounds, potentially improving the productivity of the catalogue search. Results are shown across multiple viral targets representing vastly different merging opportunities with regard to the numbers of fragment hits observed to bind, the size of the pockets and the distribution of fragment hits in pocket space, demonstrating the robustness of the pipeline and its suitability in multiple scenarios. Our tool aims to provide a more efficient search approach than traditional docking-based methods, requiring less computational resources and thus making it suitable for situations where we want to exploit all possible elaboration opportunities from a fragment screen.

When selecting compounds for purchase, a major consideration is maintaining as much diversity as possible and representing multiple core scaffolds in order to spread risk and increase the likelihood of finding a chemical series that can bind. Thus, it is promising that the bioisosteric merging pipeline is able to identify clusters of compounds in distinct areas of chemical space, representing up to a 5.3-fold increase in the number of unique substructures incorporated. Our retrospective analysis found that bioisosteric merging could help to replicate some of the SAR decisions made during analogue design by exploring different scaffolds and R-group expansions around those scaffolds. This is done without requiring expert chemistry knowledge and with limited user intervention.

Whilst delivering this increased diversity, the bioisosteric merging pipeline showed comparable performance with our original perfect merging in terms of the numbers of interactions predicted for a compound, the number of new unique interactions found that were not seen in the original fragments and the fraction of preserved interactions. Increased diversity is also reflected in terms of the observed interactions, as, across all targets, bioisosteric merges were able to identify potential new unique interaction types that were not observed using perfect merging alone. We also find that the generated conformers show favourable pharmacophoric overlap with the parent fragments, although many of the merges already share a high degree of shape similarity owing to the replacement substructures used.

We compared our new approach with a more standard out-of-the box approach based on pharmacophoric similarity search and constrained docking, which resulted in small differences in terms of the numbers of total potential interactions that can be exploited. To provide a method of comparison, we offer a definition of a merging 'hit', referring to compounds that demonstrate a $\mathrm{SC}_{\mathrm{RDKIT}}$ score of at least 0.4 and maintain at least 50% of the interactions observed within the parent fragments. The classical approach did sample a more diverse set of compounds but required substantially more compute when comparing the hours of conformer generation needed per merging 'hit' identified. These results suggest our method may be more convenient when scaling up to searching for merges and linkers of all possible fragment pairs. While the definition of a computational merging hit can be subjective and requires further experimental validation, our pipeline enables efficient identification of these favourably scoring compounds. Moreover, while the docking approach proved to be more diverse, we believe our method is still valuable as it is directly informed by the substructures observed within the parent fragments, reflected by the consistently higher $\mathrm{SC}_{\mathrm{RDKIT}}$ scores. In real-world settings, there is often a selection step involving expert chemists who may shortlist compounds according to budget constraints. In these settings, proposing compounds for which there is an obvious hypothesis for their design (informed by the original fragment hits) may give the user greater confidence in the potential for the compounds to bind in subsequent assays.

Overall, our tool provides increased chemical diversity compared with the previous iteration and provides a step towards the automation of SAR analysis for merges in a much more efficient way than previous techniques.

3.6 Conclusions

This work in this Chapter addresses several of the limitations of our first iteration of the Fragment Network merging tool (described in Chapter 2) and increases its potential productivity by expanding the search from 'perfect' to 'bioisosteric' merges. We demonstrate that this results in a favourable increase in chemical diversity and thus proposes compounds that exploit a greater number of fragments and interactions.

The pipeline has been shown to have favourable results in terms of efficiency, which we have defined as the computational resources required to identify 'computational merging hits', in comparison with a standard pharmacophore-constrained docking approach. This supports its suitability for use in larger-scale fragment screens with tens of fragment hits. Importantly, the limited requirement for user interaction means this pipeline should be able to replicate some of the design decisions made by expert chemists without pre-existing knowledge or understanding of the system. The work in Chapter 4 will describe how these tools are being deployed at XChem and how their active use has helped provide a useful platform for method development.

4. Usage of catalogue search tools in active XChem campaigns

This Chapter describes my work on progressing active XChem campaigns against targets for the ASAP project, introduced in Section 1.5.2. ASAP is an open-science collaboration aimed at progressing antivirals to aid pandemic preparedness. This work was both a useful platform for method development and contributed to efforts to identify potential follow-up compounds for live fragment experiments. I detail the work I carried out for these campaigns and give a description of the workflow employed for these projects, the compound selection process (involving multiple individuals) and the logistics involved in designing and ordering compounds. Most of the results presented were generated during the development of the pipeline described in Chapter 3 and so do not represent results from the final version of the pipeline. All the work involving the Fragment Network pipelines was performed by myself.

4.1 Introduction

Sections 1.3.2.3 and 1.5 provided an overview of the fragment screening pipeline at XChem and described its role as a co-founder of the COVID Moonshot Project, which played a key role in demonstrating how open-science discovery can accelerate the development of potent hits against antiviral targets, in this case, SARS-CoV-2 Mpro. This work helped launch the ASAP Consortium, a collaboration that focuses on using structural biology and computational techniques to develop leads against coronavirus, flavivirus and picornavirus targets. The design of initial follow-up compounds resulting from fragment screens against these targets is a key part of establishing

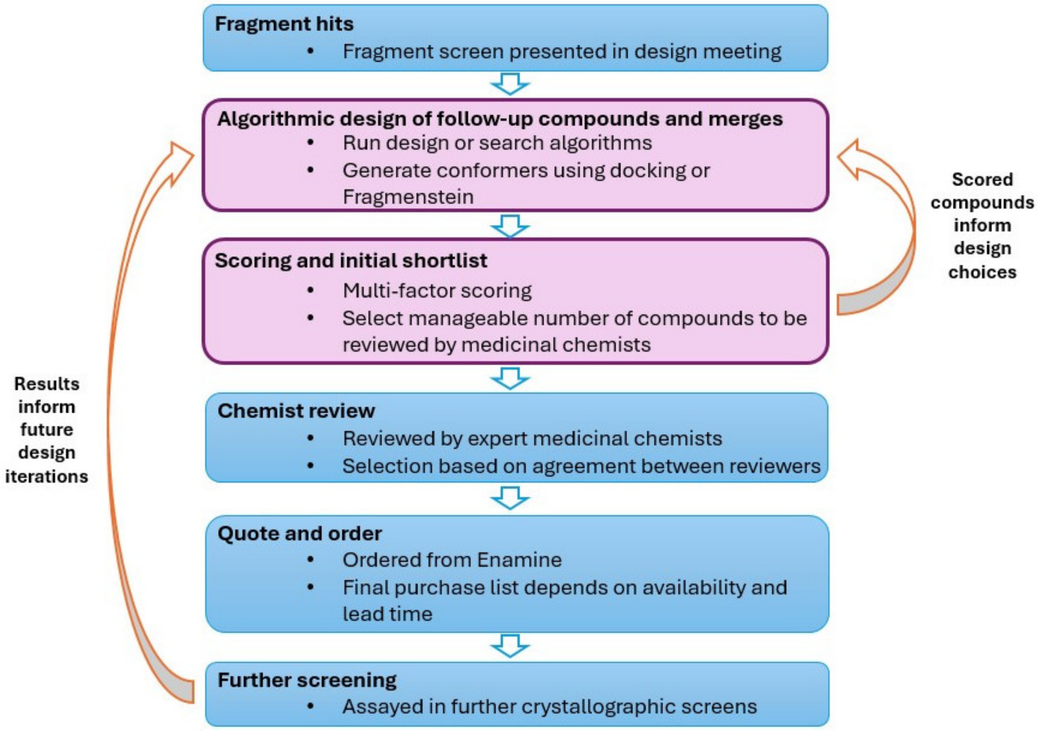

Fragment hits
Fragment screen presented in design meeting
Algorithmic design of follow-up compounds and merges
Run design or search algorithms
Generate conformers using docking or Fragmenstein
Scoring and initial shortlist
Multi-factor scoring
Select manageable number of compounds to be reviewed by medicinal chemists
Chemist review
Reviewed by expert medicinal chemists
Selection based on agreement between reviewers
Quote and order
Ordered from Enamine
Final purchase list depends on availability and lead time
Further screening
Assayed in further crystallographic screens
Scored compounds inform design choices
Results inform future design iterations

Figure 4.1: **XChem workflow for follow-up compound design.** The overall design workflow is shown. Opportunities for feedback loops are indicated by orange arrows, where visual analysis of scored compounds and results from future screens may inform algorithmic design in future iterations. Pink boxes indicate the relevant steps for compound designers.

target-enabled packages, which provide a foundation for subsequent SBDD (see Section 1.5.2).

Here, we describe the XChem compound design process and my contributions to two active ASAP campaigns and work done to establish logistics for this pipeline. The workflow continues to evolve, with the long-term goal being to commodify this process so that it is available to XChem users (as is part of XChem fragment screening experiments). The timelines reported here will therefore accelerate in future iterations.

Figure 4.1 shows the overall workflow for follow-up compound design as pursued for this Chapter. First, the crystallographer responsible for the screen analyses presents the results of the screen in a design meeting. This involves summarizing the protein sites of interest (for example, the active site or potential allosteric sites), the fragment hits and, optionally, potential interactions to focus on. This may also involve highlighting a list of fragments with unreliable densities to help inform the prioritization of follow-up compounds; this was found to be important during initial iterations.

The design tools are run by multiple individuals, resulting in a list of potential follow-up compounds based on the fragment screen. My contribution involved running early versions of the pipeline described in Chapter 3 and adapting it to the specific requests of each campaign. Other algorithms employed by the collaborators are listed below. Several of these tools are *de novo* methods, which, in general, do not yield purchasable compounds, so a similarity search must be performed for analogues, typically against the Enamine catalogue (Grygorenko et al., 2020). The algorithms used include the following:

1. SmallWorld Database search using manually defined SMARTS patterns (NextMove, 2024)
2. Fragment merging using Fragmenstein (Ferla et al., 2024)
3. Fragment elaboration with SILVR (Runcie et al., 2023), a diffusion-based generative model
4. Fragment elaboration using STRIFE (Hadfield et al., 2022), a deep learning model informed by hotspot mapping
5. Fragment elaboration using FEGrow (Bieniek et al., 2022), a user-controlled method that uses an R-group library together with a free energy perturbation (FEP) scoring function

Following the enumeration of potential follow-up compounds, scoring is used to reduce the number of compounds to a manageable amount that can be manually reviewed by medicinal chemists. Compounds were scored by each designer, typically using a combination of different metrics (for example, predicted interactions and shape and colour scores). In the initial submissions, the goal was to submit a maximum of 100–200 compounds per method for review; however, this was reduced in subsequent rounds to reduce the pressure on the reviewers. Following chemist triage of the compounds, an order list is compiled (the method of compilation varied depending on the iteration and budget) and sent to Enamine to obtain a quote, including information on whether compounds are in stock or require synthesis, and estimated lead times. Removing those that cannot be synthesized within realistic timeframes results in a final purchase list. Upon delivery, the compounds are then fed back into the XChem screening pipeline; however, the timings for results from further assays can vary depending on external factors such as when the beam is active and protein availability.

For the remainder of this Chapter, we provide case studies for design rounds conducted for two enteroviral ASAP targets.

4.2 Compound design rounds for two ASAP targets

This Section details the design of follow-up compounds against two targets, EV D68 3C protease and EV A71 2A protease. Descriptions of these targets and the relevant binding sites are provided in Section 3.3.1.

4.2.1 EV D68 3C protease

4.2.1.1 Design round 1: debugging deployment

For EV D68 3C protease, the aim of the first design attempt was predominantly to establish the logistics of the design and ordering process. Following release of the fragment–protein structures via Fragalysis (June 28th 2023), only a limited run of the pipeline described in Chapter 3 was performed. The fragments used are equivalent to those described in Chapter 3 and Appendix Table B.1. Compounds were not sent to medicinal chemists for review, as this represented an initial ‘sprint’ iteration.

Due to the short timelines, only 26 filtered compounds were proposed from the Fragment Network. In all iterations described, results were filtered for compounds present in Enamine (as the Fragment Network contains compounds from multiple vendors); however, there was limited selection based on other parameters during this iteration because of the limited number of results. Together with the other design methods used, including Fragmenstein (Ferla et al., 2024) and STRIFE (Hadfield et al., 2022), a total of 54 compounds were used to obtain a quote from Enamine, of which 30 were available (4 of which were off-the-shelf).

While we were not able to run full enumerations using the design algorithms, there were some lessons learnt during this process. One key example was gaining an understanding of the time required for the full design process; while the initial goal was to order compounds within 1–2 weeks, this does not accommodate for potential delays associated with running methods that are currently under development, such as the time required for debugging or installation on new servers. More realistic deadlines were thus set for future iterations. Furthermore, this iteration was also important for establishing the purchasing logistics for Enamine and helped forge the relevant communication channels.

4.2.1.2 Design round 2: obstacles with data quality

Following the first design round, a second round was initiated (July 17th 2023) using the equivalent set of fragments; throughout these campaigns, new design attempts were often initiated before receiving experimental results for follow-up compounds ordered during earlier iterations. In this design round, an early version of the pipeline described in Chapter 3 was run, expanding the search to potential bioisosteric merges. The selection criteria were fairly minimal at this stage, involving only Butina clustering (Butina, 1999), to select a chemically diverse subset of compounds, and manual selection. Full selection criteria for each round are provided in Appendix Table C.1. Throughout pipeline development, the scoring criteria evolved in response to the budget and results of previous rounds; the method of shortlisting compounds represents an area of active conversation at XChem (discussed further in Section 4.3).

While 139 Fragment Network compounds were uploaded to Fragalysis for medicinal chemist review (July 28th), all compounds proposed by this method were discounted. This was because all merges used fragment x0771_0A as one of the inspirations, which was found to be problematic owing to skepticism over the bond order and resulting 3D structure. Figure 4.2 shows that the 2D structure that was supplied to compound designers was chiral and aliphatic; however, the 3D crystal structure of the fragment appeared to be flat, suggesting that the fragment was aromatic and therefore indole. As the follow-up compounds were placed using this ambiguous fragment, they were discounted. Upon review of the original data, it was realized that the 2D structure supplied was incorrect, actually representing a dihydroindole. Furthermore, upon closer inspection from the crystallographer, it was concluded that the event maps seemed to indicate multiple possible conformations for the fragment; as the structure could not be accurately determined, it was advised to avoid using this fragment in future iterations.

While the results may have seemed initially discouraging, there were several valuable lessons from this iteration. First, similar to the previous design round, the occurrence of potential human or computational errors is to be expected when any workflow is being established and the timelines need to be flexible to accommodate for this. The example described shows the importance of performing rigorous quality checks and the value of manual inspection to spot issues that are not detected by design algorithms (and may not be initially evident to a non-chemist's eye). Second, this round showed that it may be preferable to select follow-up compounds that sample from (and spread risk across) a diverse set of fragments, rather than ranking the entire compound set or selecting a subset solely based on chemical diversity. A third important point is the need to maintain close communication with experimentalists and establish fragments have unclear densities. This information was subsequently used in future design rounds to de-prioritize certain follow-up compounds and proved particularly relevant given there is limited computational and/or monetary budget.

Other feedback from the medicinal chemists was more general: observations from other design methods included advice to use PAINS filters to rule out problematic substructures and to use torsion checks to remove unfavourable conformations (for example, using the Mogul software) (Baell et al., 2010; Bruno et al., 2004).

4.2.1.3 Design round 3: recognizing target-specific complexities

A third design round was initiated within the months following (September 4th 2023), which resulted in a larger set of 299 compounds from the Fragment Network that were sent for review. As fragment x0771_0A had been removed from the set of fragments used for enumeration, the only qualitative feedback from the chemists for this set was that five compounds needed to be removed due to a 'heterobicycle' that may not be stable, and a long PEG that would be difficult to replace (Figure 4.3). Across all design algorithms, a final purchase list was compiled containing 126 compounds, including 48 from the Fragment Network pipelines.

However, no promising crystal structures were identified for Fragment Network-derived compounds, and the few hits arising from other design methods did not maintain the pose predicted based on the parent fragments. While this shows that the 3C protease represents a difficult target, there may be ways to exploit this negative data in the future: as several compounds did not bind in the expected orientation, it could be possible that the interactions they were designed to

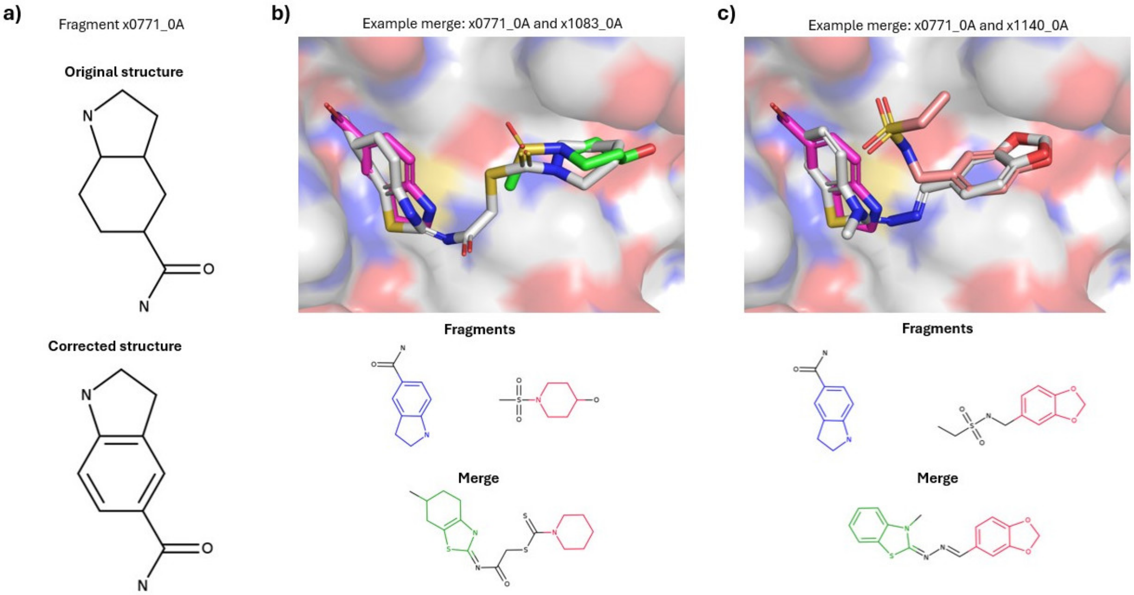


Figure 4.2: **Example merges proposed for EV D68 3C protease involving fragment x0771_0A.** (**a**) The original 2D structure supplied on Fraglaysis and the corrected structure are shown. (**b,c**) Example bioisosteric merges generated in the first design attempt using fragment x0771_0A for enterovirus (EV) D68 3C protease are shown, in combination with fragments (**b**) x1083_0A and (**c**) x1140_0A.

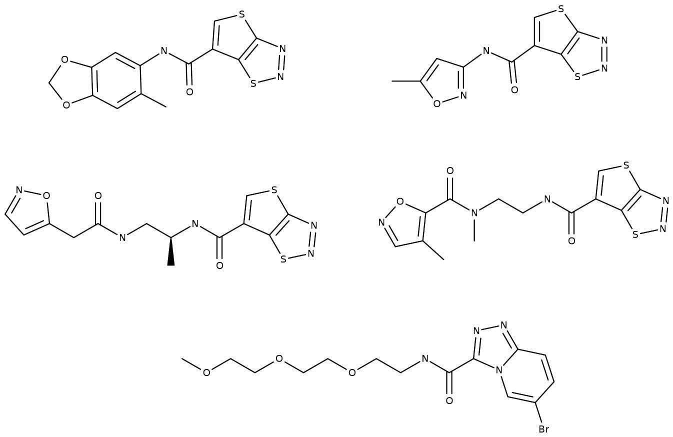


Figure 4.3: **Compounds removed for EV D68 3C protease.** Five compounds that were removed by the medicinal chemist for containing a heterobicycle (top four compounds) that may not be stable and a long PEG that would be difficult to replace (bottom compound).

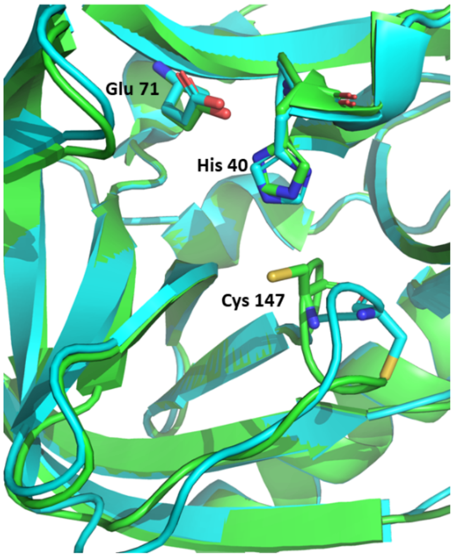


Figure 4.4: **Active site for the EV D68 3C protease.** Figure shows the overlaid crystal structure for the enterovirus (EV) D68 3C protease of monomers A and B, shown in green and cyan, respectively. CYS-147 is shown to adopt different conformations between the two.

recapitulate cannot be considered crucial for binding; for example, a water-coordinated interaction observed for an amide-containing fragment was not recapitulated. Also, compounds that were predicted to project towards the catalytic cysteine (Cys147) did not result in hits, which could suggest the cysteine is mobile and cannot be easily exploited (Figure 4.4). On the other hand, several of the follow-up compounds were able to replicate a hydrogen-bonding interaction with His161, suggesting it is a strong enthalpic driver (Lithgo et al., 2024b).

In such scenarios, it may be advantageous to proceed with a more focused approach to design rather than an enumeration exploiting all possible interactions and fragments. An interesting avenue of research may be to incorporate negative data into the design algorithms — for example, substructures or interactions from the original fragments that aren't recapitulated in crystal structures for follow-up compounds — to maximize the chances of finding compounds with improved potency.

4.2.2 EV A71 2A protease

4.2.2.1 Design round 1: the challenge of identifying productive merges

As mentioned in Section 3.3.1, the initial screen released for EV A71 2A protease only included six fragments within the active site. While this only represented a partial screen, initial design iterations were run to accelerate the development of follow-up compounds against this target. We focused on designing compounds against the P1 site and selected a protein structure for conformer generation that showed the protein in an open conformation (initiated on 13th November 2023). Between 1,300–1,400 compound designs were submitted, including 262 from the Fragment Network (24th November). Over a period of 5 days (27th November–1st December), the compounds were reviewed by three different medicinal chemists and 31 compounds from the Fragment Network were selected. Following an order to Enamine (7th December), 6 of these compounds were removed due to lack of availability, resulting in a final purchase list that included 25 Fragment Network compounds for ordering. Of these, 2 of the proposed follow-up compounds were found to result in crystallographic hits.

However, observation of these structures shows these hits are more akin to analogues of one fragment; while they contain, or are inspired by, substructures from both fragments, one of the substructures still overlaps with the volume of the other parent fragment and thus the merge does not maximize fragment hit overlap (Figure 4.5). Moreover, one of the hits results in a flipped conformation with respect to the conformer generated by Fragmenstein (this challenge is discussed in Malhotra et al. (2017)). Though this is not favourable, it still provides useful feedback on features that may not be critical for fragment binding. A key lesson was that, while it may be appropriate to search for analogues in certain applications, stricter thresholds can be imposed on follow-up compounds to find those that maximize the volume overlap with the original fragments, incorporating non-overlapping structures from both fragments.

4.2.2.2 Design round 2: scaling up to a larger fragment screen

Upon release of a larger dataset of 17 fragments, a further set of compounds was proposed using the Fragment Network, containing 35 compounds (3rd June 2024). More stringent scoring and

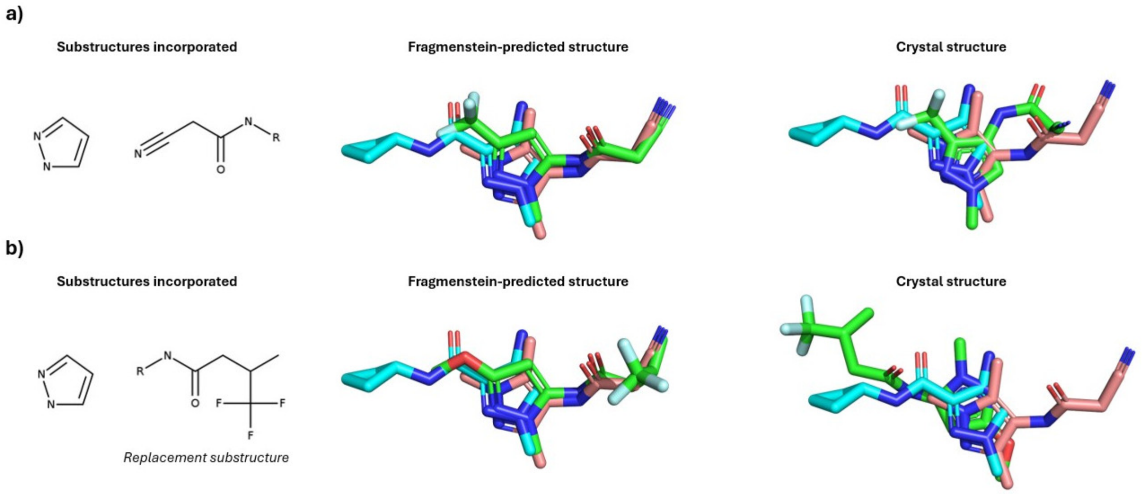

a)
Substructures incorporated
Fragmenstein-predicted structure
Crystal structure
N
N
N
N
R
O
b)
Substructures incorporated
Fragmenstein-predicted structure
Crystal structure
N
N
R
N
O
F
F
F
Replacement substructure

Figure 4.5: **Crystal structures for follow-up compounds against EV A71 2A protease.** Two follow-up compounds proposed by the Fragment Network are shown. Fragment hits are shown in pink and cyan; the structures of the follow-up compounds are shown in green (for both the predicted and crystallographic structures). The left-column indicates the substructures incorporated into the final merge. An example is shown for a (**a**) perfect and (**b**) bioisosteric merge. In particular, the follow-up compound shown in (**b**) results in a flipped conformation with respect to the computational prediction.

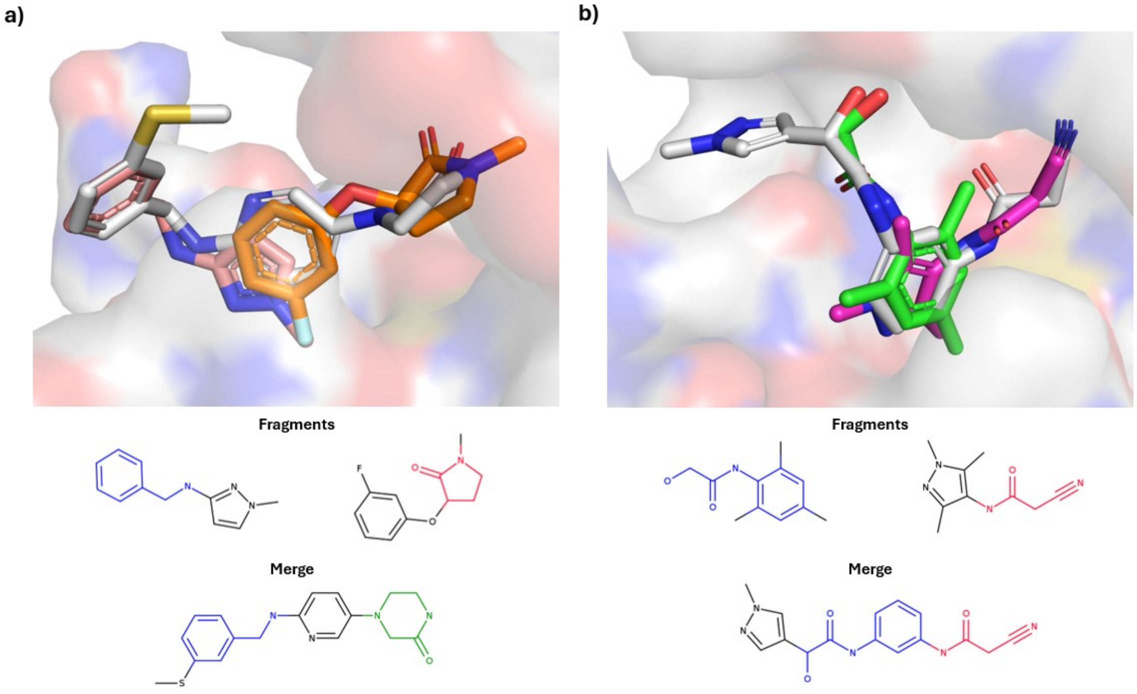


Figure 4.6: **Example merges proposed for ASAP target EV A71 2A protease.** Example merges, together with their parent fragment hits, proposed by the development version of the Fragment Network pipeline against enterovirus (EV) A71 2A protease. Fragmenstein-predicted structures of the merges are shown in white in 3D. Substructures maintained in from the original fragments are shown in blue and red in 2D; substructures that have been replaced with a bioisostere are shown in green.

filtering steps were employed, including an SC_{RDKIT} score $\geq$0.55, an RMSD of $\leq$ 1Å, maintaining $\geq$50% fragment interactions and a rotatable bond count $\leq$5. Another key focus during the selection process for this set was to select compounds representative of multiple fragment pairs. While these compounds were not subject to chemist review, informal feedback from other experimentalists at XChem recognized that the compounds sampled more of the pocket space and interactions (see examples in Figure 4.6). These compounds were entered into the ordering pipeline. At the time of writing, we are awaiting results for these compounds.

4.3 Discussion

XChem has been conducting fragment screens for nearly a decade, but it was not until the pandemic when the full potential of performing a large-scale enumeration of follow-up compounds was realized. The design rounds described in this Chapter served as a useful platform for method development that informed the design of the pipeline described in Chapter 3. This work was important in learning how to apply algorithmic tools in real-life campaigns and for establishing an optimized workflow and logistics for future XChem projects.

One of the challenges identified was the importance of establishing the compound design objectives at the start of the campaign. The algorithms have been optimized for different purposes and, consequently, they cannot perform well at both enumerating diverse compounds and proposing designs that remain close in similarity to the original fragments, which represent distinct goals. For example, a future application of the Fragment Network-derived compounds is to feed them into the synthesis platform described in 1.3.2.3; the purpose of the platform is to generate elaborations of the 'base compounds' using robotic chemistry that fully sample the possible interaction space within the pocket and build an exhaustive picture of the SAR. For this purpose, developing an initial set of compounds that represent perfect merges that do not deviate from the original fragments may be more desirable, given that chemical diversity will be generated later. For future campaigns, this could influence, for example, whether the perfect and/or bioisosteric merging pipelines are run.

Furthermore, in certain scenarios, a hypothesis-driven versus algorithmic enumeration may be more preferable, for example, if there is a particularly 'promising' fragment or interaction. The pipelines described in Chapters 2 and 3 have been designed to perform an exhaustive enumeration based on all the fragments supplied to the pipeline using a set of default filtering thresholds. However, in hypothesis-driven scenarios, greater user intervention is currently needed to tailor results, for example, to select specific substructures the user wishes to recapitulate or to filter for compounds that maintain a certain interaction. These features will need to be added to the pipeline so that they are evident to a user with less programming experience.

Similarly, we found that maintaining close communication with experimentalists, particularly regarding the reliability of the electron densities, was important to avoid prioritizing poor fragments that may be discounted during chemist curation. It is not yet clear how to communicate this information and to what extent a fragment should be de-prioritized. In situations where there is already a limited budget for review and/or purchase, inclusion of these fragments presents a potential waste of computational time and effort.

Regarding compound scoring, we found that choosing an appropriate method to enable short-listing was target-dependent on often involved performing an initial visual review of the filtered compounds. Moreover, the selection method was not consistent for the different molecule designers. Each used their own scoring method and most employed some form of manual inspection to select their favourite compounds, which is inherently subjective. Also, it may not be possible to determine the criteria used during chemist evaluation when selection is again partly based on intuition. Reaching a consensus on the best scoring methods could be improved by opening a dialogue with expert chemists as to what features they are looking for when triaging compounds, rather than selecting based on features such as predicted affinities, diversity, RMSD or interactions alone. However, as discussed above, scoring methods may need to be adjusted depending on the target and whether testing a specific hypothesis.

Throughout this work, we did identify some areas of the Fragment Network pipeline to optimize, for example, changing the way substructures are selected for merging, including prioritizing those that are responsible for interactions (which is addressed in Chapter 3) and filtering compounds that maintain ‘extraneous’ substructures (depending on the situation, allowing these elaborations may or may not be desirable). The pipeline in Chapter 3 was built to have a relatively sparse filtering step, primarily relying on conformer generation with Fragmenstein. However, additional filters may be added in the future if required by users and this work will continue to be progressed at XChem.

Importantly, we found these campaigns particularly productive for establishing the logistics of the design workflow, especially when data are passed between individuals. Important areas identified included having a way to document the design hypothesis employed by the compound designer and improving the tracking of compounds throughout the pipeline. As compounds are assigned new identifiers following delivery from the vendor, it can be difficult to track which compounds were proposed by each designer, which have been screened and which did/did not result in crystals. This will be particularly important for establishing the hit rates of the various techniques and providing negative data to inform future cycles.

4.4 Conclusions

This Chapter describes follow-up compound design work performed for two ASAP targets, EV D68 3C protease and EV A71 2A protease. These campaigns proved to be useful for improving various aspects of the Fragment Network pipeline, gaining experience of a real-life design campaign and establishing the logistics of the workflow for future XChem campaigns. One of the most important lessons of this work was how crucial having an experimental feedback loop was and how constant review by users can help the development of algorithmic approaches in drug discovery. This feedback helped hone the design of the pipeline and ensured that it can be targeted to the needs of the users.

5. Developing a novel scoring function for evaluating compound 'elaboratability'

5.1 Chapter motivation

The work in this Chapter covers the development of a tool to evaluate the 'elaboratability' of a compound, covering the general methodology and some initial results. We contrast a chemistry-informed, library-based approach for evaluating elaboratability with a deep learning approach using an equivariant GNN (EGNN), which is similar to methods described in the literature. The results show our tool has promise but requires further validation and refinement. All the work in this Chapter was performed by myself.

5.2 Introduction

During drug discovery, there are several stages that require the prioritization of a subset of compounds to purchase or synthesize from a larger selection so that they meet a set of budget constraints (discussed in Section 1.6). There is no universally established method for how to prioritize compounds, and this process may be highly subjective according to the individual and the metrics chosen for scoring (this challenge is discussed in Chapter 4.3). This Chapter outlines a novel way to assess molecules based on their 'elaboratability', which refers to their tractability for further synthesis given chemical and geometric considerations.

Rational drug discovery invariably consists of several cycles of DMTA (see Section 1.4.4.6), during which an initial hit scaffold is iteratively optimized to become a more potent binder. Medicinal chemistry and *in silico* techniques can be used to elaborate a hit scaffold and optimize its binding affinity by adding functional groups at different positions and exploring their effects on the SAR. However, one of the considerations is deciding which vector should be elaborated from, to best exploit potential new interaction opportunities within the pocket.

Additionally, some compounds may lack sufficient growth vectors to enable elaboration. As described in Section 1.2.2, one of the initial main aims during hit discovery and optimization is to identify compounds that demonstrate high LE (Hopkins et al., 2004, 2014) (a metric that is calculated by normalizing affinity according to molecular size). However, high LE alone does not necessarily mean that a ligand is a good candidate for further optimization, as it may indicate that the ligand is deeply 'buried' within the binding pocket, limiting further development (see Figure 5.1 for an example). This could occur because the molecule lacks the synthetic handles to allow it to be chemically modified, or, when the growth vectors exist, their location prevents elaboration because of steric constraints.

This leads us to the concept of 'elaboratability', where our aim is to develop an approach for scoring compounds based on whether they have tractable growth vectors. Our end goal is to compute a single overall score for each compound ('compound elaboratability'), depending on how elaboratable its vectors are ('vector elaboratability'), both of which represent useful metrics. Such tools could be useful for multiple applications. For example, assessing vector elaboratability could be useful for selecting which vectors to invest computational resources into with *de novo* and *in silico* scaffold decoration models. Additionally, evaluating compound elaboratability may aid in compound prioritization when selecting a set of follow-up compounds for purchase, by ensuring they represent valuable starting points for further optimization.

This concept has not been much explored in the literature. Preston (2023) developed a simple scoring function for elaboratability using comparisons against USPTO reaction data (Schneider et al., 2016), described in Section 1.6.4. To evaluate a given molecule, the function determines if any atoms in the compound have topologically equivalent atoms in the dataset, representing reaction centre atoms. A score is then generated based on how frequently such equivalent atoms are observed, and the diversity of reaction classes they participate in. While this was found useful for the purpose

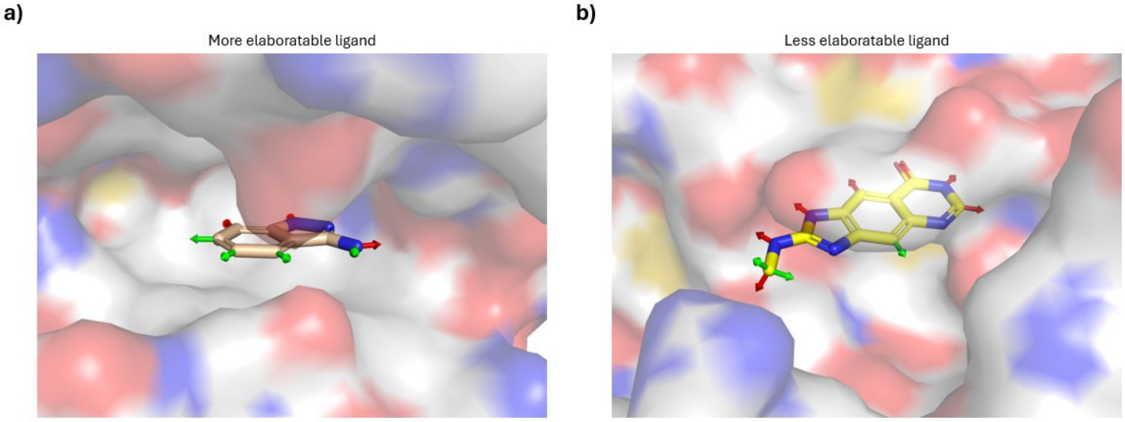


Figure 5.1: **Defining the concept of elaboratability.** We define two types of elaboratability: ‘vector elaboratability’, which assesses individual vectors according to how tractable they are for growth, and ‘compound elaboratability’, which assigns a score to the whole molecule according to the elaboratability of its individual vectors. Possible vectors are indicated by green and red arrows, where the latter shows vectors that are un-elaboratable (for example, due to lack of room in the pocket). Both an example of (**a**) an elaboratable compound, which has multiple growth vector opportunities, and (**b**) a less elaboratable compound, which is highly buried in the pocket and only has three accessible vectors, are shown.

of prioritizing fragments, the score depends on reaction classifications predicted by the proprietary NextMove software Pistachio (NextMove, 2017), making it unsuitable for implementation in an open-source format. Moreover, it does not account for steric considerations, ensuring that there is enough room in the pocket for growth.

A similar concept to compound elaboratability in the literature is ‘fragment sociability’, a concept that is applied in fragment library design by Astex Pharmaceuticals and seeks to ensure that fragment hits can be readily elaborated following screening. As discussed in Section 1.3.2.2, a sociable fragment is defined as one possessing multiple vectors that are synthetically tractable for elaboration — such as through heterocycle and scaffold modifications — and has a sufficient number of purchasable close analogues. A review by Astex (St. Denis et al., 2021) examined fragment sociability in past fragment screening campaigns, utilizing the Fragment Network (Hall et al., 2017) to determine whether growth vectors led to elaborations. A fragment was deemed sociable by calculating the ratio between the number of validated growth vectors and the number of theoretically possible vectors, meaning any atom bound to a hydrogen. An important observation was that apparently unsociable fragments could be socialized, for instance, by splitting an unsociable fragment into two sociable fragments, or by modifying a functional group.

The approaches described here offer promising ways to assess and enrich for synthetic tractability; however, the aim of this Chapter is to develop a tool that goes beyond these techniques by including geometric considerations in the evaluation of ligands when bound to a target.

The automated evaluation of elaboratability could be hugely beneficial for computational and *de novo* models. Many of the existing computational tools and DL generative models for fragment elaboration require scaffolds with explicitly defined attachment points as input (Fialková et al., 2022; Arús-Pous et al., 2020; Thomas et al., 2024; Bieniek et al., 2022). However, in practice, the optimal attachment point may not be known, or may require an expert to determine by eye.

There are few examples where attachment point selection has been integrated into the generative process. FRAME (Powers et al., 2023) is a geometric DL model for molecule generation, based on the iterative addition of fragments to a starting compound, which uses an EGNN to predict the appropriate attachment point for elaboration. This occurs in two stages: first, the model selects which attachment point should be elaborated from on the starting ligand; and second, a fragment is added, selected from a pre-defined library. For clarity, we refer to the smaller, un-elaborated ligand as the 'core ligand', and the ligand after elaboration as the 'full-sized ligand'. To train the attachment point prediction model, ligands from the PDBbind (Su et al., 2019) are fragmented into several pieces to generate core ligands. The fragmentation occurs as follows: any fragment within the full-sized ligand that exists in a pre-defined library is removed; of the remaining non-library structures, any fragment connected to a ring by a single bond is disconnected; these non-library fragments are then added to the full library of fragments. Thus, for a given fragment (or core ligand) in the pocket, an attachment point, defined as any hydrogen atom, is marked as positive if it is replaced by an elaboration in the full-sized ligand.

Another example, DiffDec (Xie et al., 2024), consists of a diffusion model conditioned with 3D information from the pocket using an EGNN, and is used for scaffold decoration. The dataset was generated using reaction-based rules to 'slice' ligands into a set of scaffolds and decorators, similar to the method described by Fialková et al. (2022); however, for DiffDec, this procedure is applied to ligands from the CrossDocked dataset (Francoeur et al., 2020), rather than using crystal structures alone. They subsequently tested the performance of the model for performing R-group decoration, with and without specifying an anchor atom, and found that the latter led to modest declines in chemical validity and the recovery of similar R-groups.

While these vector prediction models should be, ideally, learning an implicit definition of vector elaboratability, limitations in how the datasets are generated, using reaction-based slicing or fragmentation rules, may bias the definition so that it ignores aspects of whether the vectors are synthetically tractable. Additionally, these elaborations may not always reflect realistic optimization scenarios. In practice, the smaller ligand may not be entirely recapitulated within the larger elaborated ligand, and few-atom modifications may be required to enable elaboration; this is similar to the 'socialization' of unsociable fragments described above. We expand on these methods by developing an improved dataset for training and evaluating these tools that is more applicable to realistic elaboration scenarios. These tools could then be applied in multiple contexts beyond generative models.

In this Chapter, we describe the initial development of a tool designed to assess both vector and compound elaboratability, enabling prioritization of molecules for purchase. We outline a library-based method that considers both the chemistry of the ligand and the geometry of the protein pocket. We contrast this approach with an EGNN model trained on a synthetic dataset that simulates more realistic elaboration scenarios than previously published methods, where the core ligand may undergo small chemical modifications during elaboration. While further validation and refinement is required, this Chapter provides the first example for a comprehensive method for the assessment of elaboratability.

5.3 Methods

We developed two tools: a library-based and EGNN approach, the latter of which is similar to that described by Powers et al. (2023). Our library-based tool was designed to assess elaboratability in terms of both the geometry of the pocket, evaluating whether there is space to grow, and the chemistry, in terms of whether elaboration is chemically feasible.

5.3.1 A library-based method for assessing elaboratability

An exhaustive method for evaluating vector elaboratability would be to enumerate a set of all possible decorators and to determine if expansion would allow the creation of multiple elaborated compounds, while taking into account both spatial and chemical aspects. Since such a method would be intractable, we propose a proxy approach in which we generate constrained conformations for a library of candidate elaborations and calculate whether the decorators result in potential clashes, or new interactions, with the protein.

To make this proxy even more tractable, the set of putative decorators is used to generate a point cloud, representing the 'elaboration space'. This is created by: generating a set of conformers for a library of possible decorators, which are all aligned according to their attachment point atoms; clustering their atomic coordinates to generate the point cloud; and aligning this cloud to the query vector, which allows the calculation of whether the vector is tractable or not. The steps are described in detail below.

5.3.1.1 Library of elaborations

In this work, we used the dataset of decorators used for training LibINVENT (Fialková et al., 2022), a generative model that uses a recurrent neural network to add decorators to a scaffold at specified attachment points. LibINVENT uses a dataset generated from the ChEMBL database (Gaulton et al., 2012), where 37 curated reaction rules are used to 'slice' ChEMBL compounds into scaffolds and decorators. From the set of unique decorators (totalling 496,716), we selected all those that occurred at least 10,000 times, resulting in a total of 311. We tested decorator libraries of two different sizes, depending on the frequency of occurrence; however, we settled on a threshold of

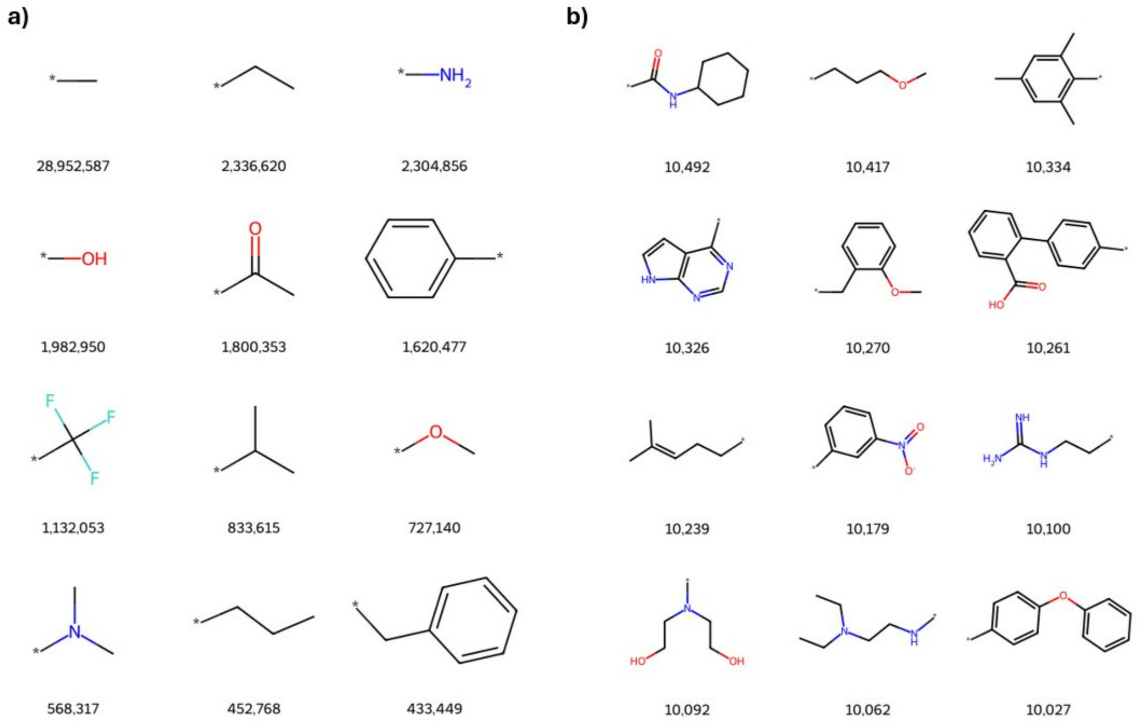


Figure 5.2: **Example decorators used in the library-based elaboratability evaluation.** The decorators used in the library-based elaboratability evaluation are sourced from the LibINVENT dataset of scaffold–decorator pairs, which was generated from ChEMBL using 37 curated reaction rules (Fialková et al., 2022). For the elaboratability tool, the decorators have been filtered for those that occur >10,000 times within the dataset. Examples of the (**a**) most common and (**b**) least common elaborations in our filtered set are shown; numbers indicate frequency of occurrence.

10,000 due to the substantially reduced computational requirements (see Section 5.4.4 for details). Figure 5.2 shows examples of the most versus least common decorators observed in the dataset.

5.3.1.2 Conformer generation for elaborations

We pre-generated conformers for the library of decorators to avoid the costly process of generating constrained conformers for full-sized molecules on-the-fly. For each decorator, several conformers were generated, using RDKit with distance geometry constraints (Landrum, 2024), the number of which was chosen according to the number of rotatable bonds present (Ebejer et al., 2012). All conformers were aligned so that the growth vectors, referring to the attachment points and their neighbouring atoms, have the same coordinates. To sample the full conformer space, each

Table 5.1: Number of clustered points generated from the decorator library

Threshold of occurrence	Number of decorators	Number of conformers	Number of clustered points
≥10,000	311	3,998	5,215
≥5,000	586	10,825	11,668

conformer was rotated around the growth vector at 30° intervals, resulting in 600 total conformers for each decorator. For each decorator, all its corresponding conformers were clustered using Butina clustering (Butina, 1999) and a distance threshold of 1.5Å RMSD to reduce redundancy. This resulted in a diverse set of aligned conformers that sample the full rotational space around the vector (see Appendix Figure D.1 for an example). Across all 311 decorators, this process resulted in a total of 3,998 conformers.

5.3.1.3 Point cloud generation

As the elaboration space covered by the conformers is very densely populated, to further improve computational efficiency, the geometry of the binding site is assessed by computing distances between a point cloud representation of the atoms in the conformers and in the protein. To do this, we used clustering operations to generate the point cloud based on atomic coordinates within the conformers.

First, the coordinates and elements from all atoms in each conformer were compiled; this also involved recording whether each atom represented a potential hydrogen bond donor (HBD) or acceptor (HBA) atom. For each element type, the coordinates of the corresponding atoms were clustered using agglomerative clustering with a distance threshold of 0.5Å. This resulted in a 'cloud' of clustered points. Each clustered point is associated with a specific atom type; a coordinate, based on the cluster centroid; a set of conformer and decorator identifiers that record which atoms can be attributed to the cluster; and whether the cluster point is associated with an atom that represents a potential HBD or HBA. The results are shown in Table 5.1.

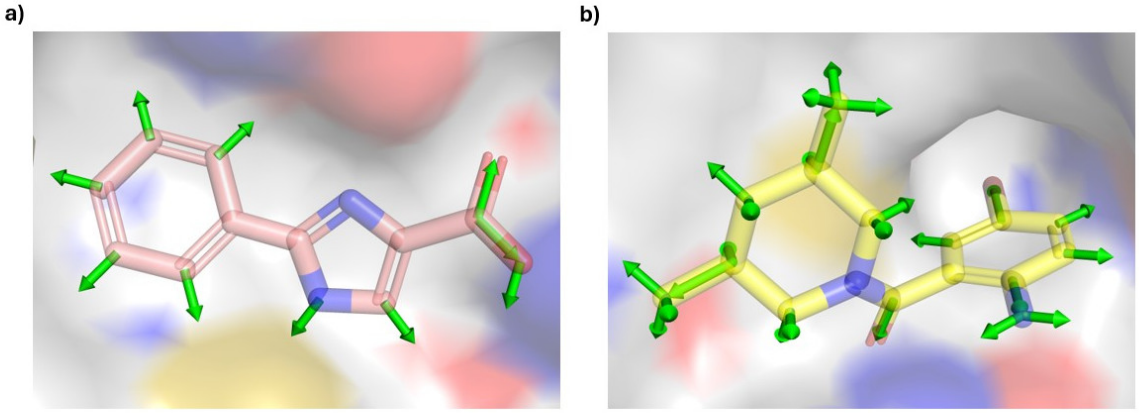


Figure 5.3: **Examples of enumerated vectors.** All hydrogen and ‘terminal’ heavy atoms (referring to any atom not bound to another heavy atom) are treated as possible growth vectors, whereby the hydrogen or terminal atom is replaced by the decorator during elaboration. The possible vectors are shown by green arrows.

5.3.1.4 Evaluation of a query growth vector

The aim of the library-based scoring function is to evaluate a ligand’s individual growth vectors when bound to the protein pocket. The scoring function takes as input: the point cloud, representing the possible elaboration space; the bound ligand, represented as an RDKit molecule; and the protein, represented by its PDB file. Hydrogens are added to the ligand using RDKit (Landrum, 2024) and all possible growth vectors are identified — we define two types of vectors: vector atoms bound to hydrogens, where elaboration occurs by replacing the hydrogen atom; and vector atoms bound to non-hydrogen, ‘terminal’ atoms, where the terminal atom is replaced by the decorator (see Figure 5.3 for an example). The protein is prepared by extracting the pocket atoms, defined as those within 8Å of a ligand atom, together with their atom type and radii, which are used in a later step for calculating atom clashes. We also record which are surface atoms that represent potential HBDs and HBAs, used later for calculating possible interactions.

Following this, the evaluation is run for each vector within the ligand. To do this, the point cloud is aligned to the vector based on the atomic coordinates of the vector atoms (the atom to be replaced and the neighbouring atom). The evaluation is done for all possible vector positions but we report a per vector atom score, aggregating the results of the all possible vector positions (for example, a vector atom may be attached to several hydrogens). Furthermore, to fully sample the possible coordinates for a hydrogen atom, for an atom bound to hydrogen atoms, the hydrogens are rotated so that six possible positions are sampled overall (this will likely be refined in future versions of the tool). Various metrics can then be calculated for scoring.

All distances are calculated between the point cloud and the pocket atoms and between the cloud and the ligand atoms using the 'cdist' function from SciPy (Virtanen et al., 2020); this is used to calculate potential clashes with the protein and with the ligand itself, considering the atomic radii. Due to the data structure employed, it is then straightforward to retrieve the conformers involved in a clash, and if all conformers for a specific decorator clashed.

To evaluate the *potential* to reach new interactions in the binding site, we use a coarse-grained representation of possible protein–ligand interactions by calculating if a cluster point associated with an HBA or HBD (from a non-clashing conformer) is within 3Å of a HBD or HBA within the protein. Refining the method of interaction prediction, and incorporating other types of interaction, would be an area to develop in future iterations of the tool.

5.3.1.5 Evaluation of chemical tractability

We also provide an additional feature that considers the chemical feasibility of the elaborations using the neural network policy of AiZynthFinder (Genheden et al., 2020), a tool for retrosynthesis prediction that uses a Monte Carlo tree search to propose a set or purchasable precursor molecules given an input molecule SMILES. AiZynthFinder, which was trained on USPTO data, uses a reaction filter network to shortlist retrosynthetic reaction templates, providing values for their probabilities.

We incorporate the AiZynthFinder policy network into our evaluation to filter out infeasible elaborations as follows. Specifically, reaction SMILES are created whereby the core ligand and the decorator represent the reactants and the full-sized ligand represents the product; a filter cutoff of 5% is applied to rule out unlikely reactions (see Appendix Figure D.2 for an example). This filter is applied to decorators that do not result in a clash and result in a potential interaction. The filter was implemented at this stage rather than before, based on the entire decorator library,

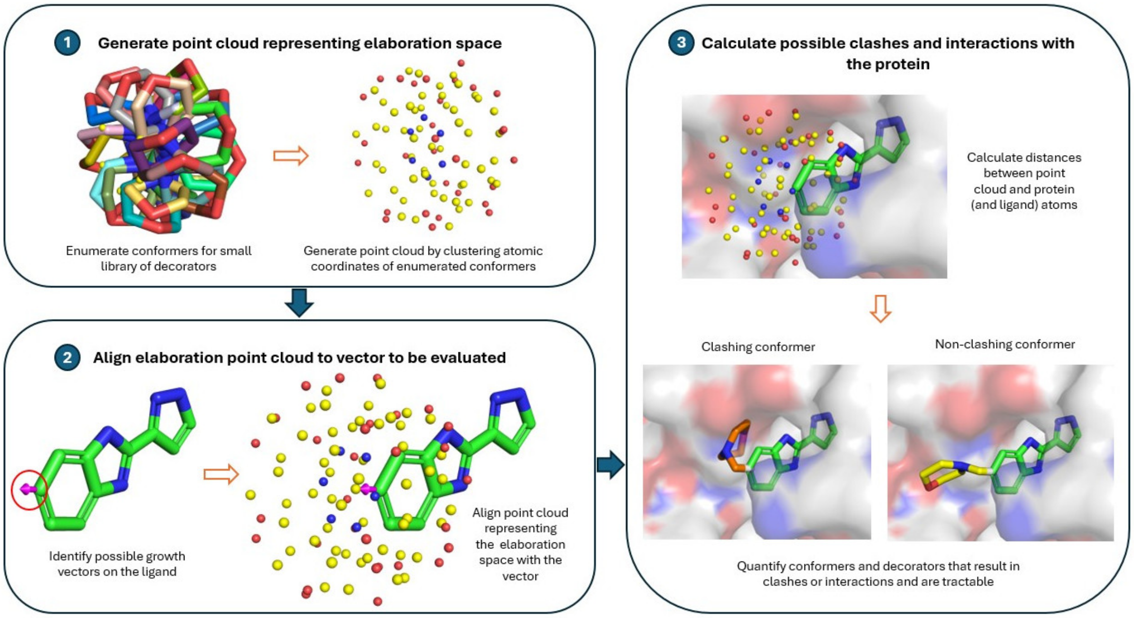

1 Generate point cloud representing elaboration space
Enumerate conformers for small library of decorators
Generate point cloud by clustering atomic coordinates of enumerated conformers
2 Align elaboration point cloud to vector to be evaluated
Identify possible growth vectors on the ligand
Align point cloud representing the elaboration space with the vector
3 Calculate possible clashes and interactions with the protein
Calculate distances between point cloud and protein (and ligand) atoms
Clashing conformer
Non-clashing conformer
Quantify conformers and decorators that result in clashes or interactions and are tractable

Figure 5.4: **Workflow for library-based method for scoring elaboratability.** To evaluate elaboratability, a point cloud representing the possible elaboration space is generated based on pre-enumerated conformers for a library of decorators. Given a vector for elaboration (shown by the pink arrow), the point cloud is aligned with the vector and all distances against nearby protein and ligand atoms are calculated. This allows the user to determine which conformers and decorators result in clashes and can potentially form interactions.

to save computational resources. We found this method sufficient for an initial estimation of the tractability of a growth vector, but exploring more accurate approaches for determining reaction feasibility represents an area for future development.

5.3.1.6 Selecting a scoring function

We explore multiple methods for assigning a score to evaluate the elaboratability of a specific vector. Several metrics are available based on the calculations described above, including the following:

- The mean number of clashing conformers or decorators (ruling out decorators where all associated conformers result in a clash)
    - The mean is calculated based on all possible vector positions: for each vector atom, the evaluation is calculated for all attached hydrogens or terminal atoms and their associated coordinates; the vector atom is determined to be a good vector based on the average of all these attached atoms
- The mean number of non-clashing conformers or decorators that result in a possible interaction
- The number of interaction residues reached by non-clashing conformers
- The mean number of non-clashing conformers or decorators that result in a possible interaction and are considered synthetically tractable

We provide one schema for a tiered scoring system based on the following criteria listed below, although there may be multiple ways to assess elaboratability using the metrics described above. A goal in a future iteration of this tool would be to aggregate these criteria into a single number that can be used for ranking vectors. Thresholds were selected for different components of the metric using the percentiles of the related value observed in our training PDBbind synthetic dataset (details of the datasets are described below). An upper and lower bound was used for the clash score threshold to reflect that we want to rule out vectors that result in few clashes, as they are likely solvent-exposed, while vectors resulting in high numbers of clashes will exhibit limited room for growth. Results using other clash score percentile thresholds are shown in Appendix Table D.2; to simplify our analysis, here, we focus on thresholds representing a 10–90 percentile range.

1. Clash evaluation
    - Between 4–80% of decorators within the decorator library can be placed at the vector without resulting in an atomic clash (meaning at least one of the conformers for that elaboration does not result in a clash)
    - These thresholds were chosen as they represent the 10th and 90th percentile when calculated across our synthetic training set
2. Clash + interaction evaluation
    - In addition to the criterion above, non-clashing conformers interact with at least one protein residue
3. Clash + interaction + tractability evaluation
    - In addition to the criteria above, considering only synthetically tractable elaborations, at least one protein residue is still reached

5.3.2 Training an EGNN for comparison

We train an EGNN model, similar to that described by Powers et al. (2023); however, our EGNN is trained on data designed to more accurately represent realistic elaboration scenarios, in which the core ligand may not be completely recapitulated by the full-sized ligand. We also use a more robust approach to evaluate generalizability, described below, which should be more robust against learning ligand or protein biases.

The model is implemented using PyTorch (Paszke et al., 2019) and PyTorch Geometric (Fey et al., 2019) and consists of three E(n)-equivariant graph convolutional layers (Satorras et al., 2022) with 30 hidden node features. The model takes as input a graph representation of the ligand–protein complex. Nodes are used to represent all ligand atoms and protein atoms within 8Å of any ligand atom (with atomic coordinates). Edges are constructed for all intermolecular distances that are within 10Å and intramolecular distances within 2Å. Node features consist of a one-hot encoded vector representing atom type and whether the atom belongs to the ligand or protein. Edge features consist of one-hot encoded vectors representing whether the edge exists between ligand atoms, protein atoms or between a ligand and protein atom.

The model was trained with a batch size of 8, together with the Adam optimizer and an initial learning rate of 0.001; a learning rate scheduler was used to reduce the learning rate on plateau, where the learning rate is reduced by a factor of 0.1 if performance had not improved after 5 epochs. The model was run for a maximum of 100 epochs, terminating early if performance does not improve after 10 epochs. Binary cross entropy (BCE) was used for the loss function; loss was averaged over all nodes in each ligand. The classes are highly imbalanced as there are few positively labelled vectors (described below) in the dataset; to compensate for this, the loss was weighted according to the ratio of positive to negative values within the training set. The full list of parameters tried and used in the final model are shown in Appendix Table D.2.

5.3.3 Evaluating compound elaboratability

As mentioned, we define compound elaboratability according to whether the compound has multiple elaboratable vectors. To provide a metric that can rank the compounds by elaboratability, and to enable comparison between the two approaches, we compute the ratio between the number of elaboratable vectors and the number of theoretically possible vectors: the former refers to the number of vectors predicted by the scoring function to be chemically and geometrically tractable; while the latter refers to the total number of possible vector atoms there are, meaning any atom bound to a hydrogen or terminal atom. This is similar to how fragment sociability is evaluated, as described by St. Denis et al. (2021).

5.3.4 Datasets for evaluation

5.3.4.1 Experimental design

To provide data to both train (or calibrate) and validate the scoring functions, we used two datasets. The first, larger dataset was generated from PDBbind 2020 (Wang et al., 2005). The second smaller set was sourced from the data compiled by Malhotra et al. (2017), described in detail below. The aim with the smaller dataset was, in contrast to the synthetic data, to perform manual annotation of the vectors, whereby all possible vectors within the ligand are labelled as positive or negative. The synthetic vectors, in comparison, contain positive and unlabelled data.

5.3.4.2 Synthetic PDBbind dataset

To generate a sufficiently large dataset containing both core ligands paired with elaborated, full-sized ligands against the same protein, we generated synthetic data using complexes in PDBbind v2020 (Wang et al., 2005). We wanted to generate a more realistic elaboration dataset, where the core ligand may not be entirely recapitulated by the full-sized ligand, unlike the data generated to train FRAME (Powers et al., 2023), which fragments PDBbind ligands by making disconnections within the compound (described above). This allows for few-atom changes to the core ligand, which are likely to occur in a medicinal chemistry optimization scenario.

To generate the dataset, we used the AiZynthFinder tree search (Genheden et al., 2020) using the USPTO filter policy to propose precursors that are available in ZINC (Irwin et al., 2020) for PDBbind ligands; these precursors represent the core ligand to be elaborated. Conformers were then generated for the precursors to provide a 3D structure than can be used for training. To place the precursors, we calculated the MCS between the original ligand and the precursor, which could be used as a template for placement. Filtering was performed to remove precursors where the MCS comprises less than three atoms and where the precursor contains at least six atoms that do not belong to the MCS. These filters are aimed at removing precursors that either represent a very small portion of the larger ligand or do not closely resemble a substructure of the ligand. The MCS was used to generate possible mappings between the precursor and the full-sized ligand, and the vector atoms were recorded.

Conformations were generated using RDKit's constrained embedding function (Landrum, 2024). To do this, the coordinates from the ligand are assigned to the MCS. The MCS is then used as a template for the constrained embedding function, which attempts to generate a physically reasonable conformation for the precursor using this template (Figure 5.5a,b).

The GT vectors on the core ligand that are elaborated from in the full-sized ligand are annotated as positive for training the model. Filtering was applied to these vectors to select only those that

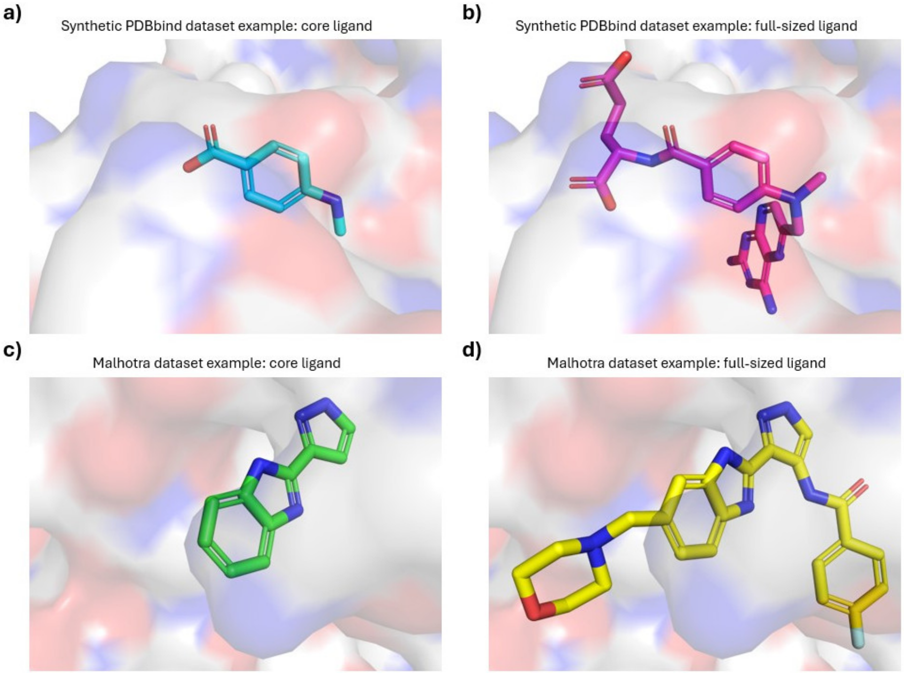

a)
Synthetic PDBbind dataset example: core ligand
b)
Synthetic PDBbind dataset example: full-sized ligand
c)
Malhotra dataset example: core ligand
d)
Malhotra dataset example: full-sized ligand

Figure 5.5: **Example data used for evaluation of elaboratability scoring tools.** (**a,b**) Synthetic data were generated using complexes in the PDBbind v2020 (Wang et al., 2005). Precursors (**a**) were generated using AiZynthFinder (Genheden et al., 2020) and placed using RDKit constrained embedding (Landrum, 2024) based on atomic coordinates from the full ligand (**b**). An example is shown for RCSB PDB ID: 1DHJ. (**c,d**) A small dataset curated from the Malhotra set (Malhotra et al., 2017), consisting of small and larger ligands that bind to the same protein, was also used for evaluation. Complexes were filtered for those where the smaller ligand (**c**) is fully recapitulated in the larger ligand (**d**) the $\mathrm{SC}_{\mathrm{RDKIT}}$ score between the two is >0.5.

represent productive elaborations. To annotate productive vectors, we retain only those that lead to an elaboration that is either responsible for an interaction or grows closer to the protein; this means that the minimum distance between an atom in the full-sized ligand and the protein is less than the minimum distance between any atom in the core ligand and the protein. This ensures that we ignore vectors that extend into the solvent or do not contribute meaningfully to compound optimization. Additionally, we ensure the elaboration only results in a maximum of a single-atom change to the core ligand. We refer to these vectors as positively labelled, ground-truth vectors (GT^+).

It is important to note, there may be other possible valid vectors within the dataset. While we treat all non-GT^+ as negative data together with using the BCE loss function, an avenue to explore is using strategies for positive and unlabelled data (see Bekker et al. (2020) for a survey).

We use the out-of-distribution (OOD) data split described in Warren et al. (2024) for training the EGNN. This data split was designed to minimize the similarity between proteins and ligands in the training set and the test set; this was calculated using both Tanimoto similarity and sequence identity (full details can be found in the original paper). Splitting the data in this way aims to provide a fair and unbiased benchmark for assessing the ability of a model to generalize to new complexes. We generated precursors for ligands within the data split rather than split the data based on the similarity of the precursors; this is because many of the precursors are repeated across different complexes. However, the data split ensures that similar elaboration opportunities are kept distinct as the full-sized ligand should be dissimilar between the training and test sets. We also removed complexes that are used in our separate curated Malhotra set (described below) and those for which the precursor is further than 8Å from a protein atom. The numbers of precursor–protein complexes are shown in Table 5.2; the corresponding numbers of PDB complexes that were used to generate the precursors and the total number of GT^+ vectors are also shown.

5.3.4.3 Small curated Malhotra set

We used the dataset described by (Malhotra et al., 2017), which was curated from complexes in the PDB (Berman et al., 2000) by identifying pairs of ligands, bound to the same protein, for which the

Table 5.2: The number of precursors generated for the PDBbind synthetic dataset

	Training set	OOD set
Number of PDBbind complexes	11,070	155
Number of precursors	23,288	297
Number of GT^{+} vectors	42,286	552

GT^{+}, positively labelled ground-truth; OOD, out-of-distribution.

smaller ligand represents a substructure of the larger ligand; this was calculated using a similarity metric based on molecular fingerprints. The authors ensured the difference in size is typical of that seen in typical elaborations, resulting in 297 crystal structures for pairs of ligands. We focus on a small 'perfect' subset of the data, for which there is an obvious hypothesis for elaboration (meaning the vector is easily identifiable) and overlap between the core and full-sized ligand. Specifically, we filtered the pairs for those where the core ligand is entirely recapitulated within the full-sized ligand, and where the SC_{RDKIT} score between the core ligand and the equivalent substructure within the full-sized ligand is ≥ 0.5 (the SC_{RDKIT} score is a measure of the volume and pharmacophoric overlap, described in Section 1.6.1). We also selected pairs where the core ligand had a maximum of twenty heavy atoms, resulting in a small subset that can easily be manually annotated. This resulted in a total of 64 ligand pairs (see Figure 5.5c,d for an example).

To allow annotation, a visualization tool was created (Appendix Figure D.3). Due to time constraints, thus far, I performed the annotation; however, this tool will allow expert chemists to annotate the data in future work. While an ideal experiment would involve having multiple experts label the dataset, to check concurrence between their manual annotations and with the ground-truth vectors, it provides an initial validation set to compare the results from our scoring tools. According to this annotation, across all possible vectors within the dataset, 62% were labelled as elaboratable, while 38% were labelled as non-elaboratable.

5.4 Results

To compare the two approaches for evaluating elaboratability and assess whether they are able to identify elaboratable vectors, we ran both scoring functions on the synthetic PDBbind dataset, in which the GT$^+$ vectors are the positive labels and all other vectors regarded as negative, and a small manually labelled dataset, where all vectors have been labelled as positive or negative vectors.

5.4.1 Results for the vectors included in the PDBbind synthetic calibration dataset

The PDBbind synthetic calibration dataset was used both to select thresholds for the library-based scoring function and to train the EGNN model to evaluate vector elaboratability.

With regard to the library-based scoring function, Figures 5.6 and 5.7 show the differences in distributions for the metrics recorded between the GT$^+$ versus non-GT$^+$ vectors, including the number of clashes, possible interactions and chemical tractability. The figures suggest the following: in general, the GT$^+$ vectors seem to result in conformers and decorators that form less clashes and make interactions with a greater number of residues, resulting in a greater proportion of elaborations that represent promising, tractable expansion opportunities. It is worth noting that the non-GT$^+$ vectors do not necessarily all represent bad vector opportunities and thus we would not expect there to be stark differences in the distributions for the two classes.

Using the distributions, we created a simple scoring function (described in Section 5.3.1.6), which provides an initial proof-of-concept for using these simple metrics to evaluate vector elaboratability. Conducting a thorough validation of the scoring function with greater experimentation and calibration to determine the optimal thresholds represents an avenue for future research.

5.4.2 Results for the vectors included in PDBbind synthetic OOD dataset

The PDBbind synthetic OOD dataset refers to a test set belonging to a data split that was designed to maximize ligand and protein dissimilarity to complexes in the training set. We evaluated whether the scoring functions were able to correctly identify GT$^+$ vectors for precursors generated from this OOD dataset.

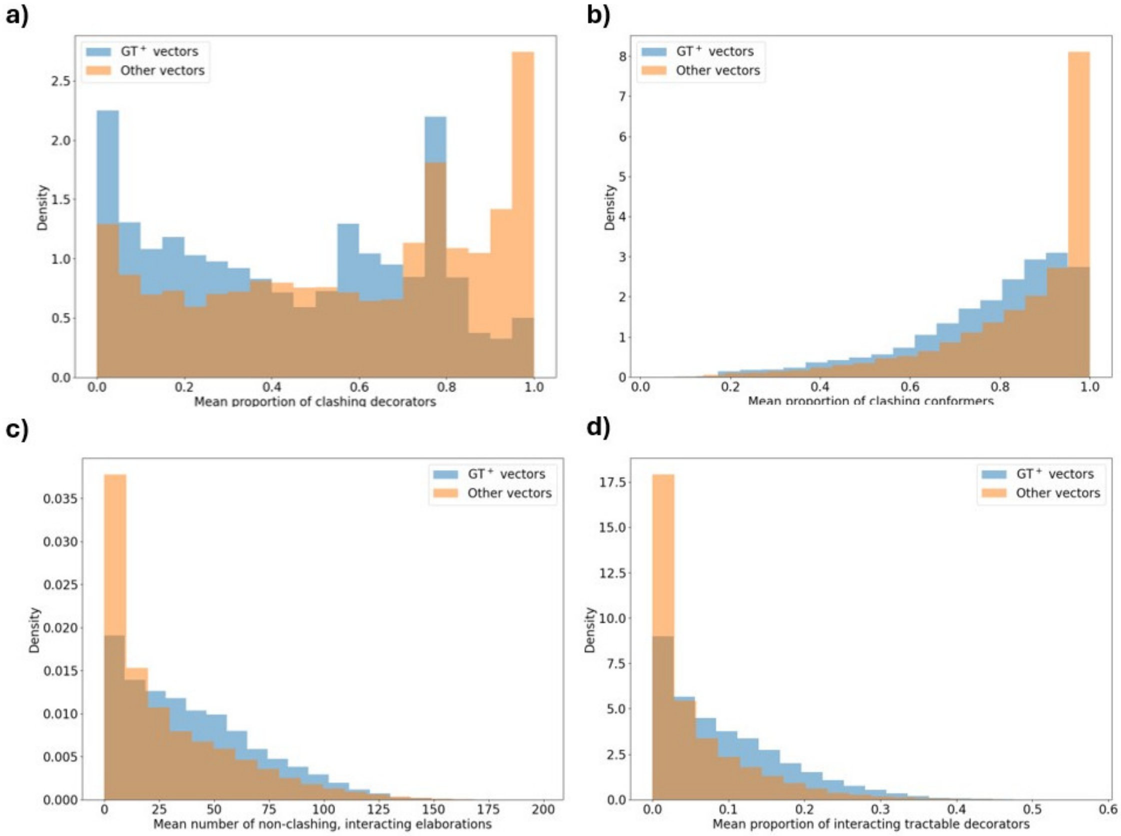


Figure 5.6: **Proportion of decorators or conformers added to vectors in the PDBbind synthetic calibration set that clash, form potential interactions and are tractable.** (**a**) The panel shows the mean proportions of decorators that clash with a protein or ligand atom. A non-clashing decorator is considered any for which a conformer has been successfully placed without a clash. (**b**) The mean proportions of conformers that clash with a protein or ligand atom are shown. (**c**) The mean proportion of decorators that do not clash and form a potential hydrogen bond interaction with a protein atom is shown. (**d**) The mean proportions of decorators that do not clash, form a potential hydrogen bond and are considered tractable according to AiZynthFinder are shown. Data are shown for the positively labelled ground-truth (GT^+) vectors versus all other possible vectors. The mean is calculated at each vector atom as the average across all possible vectors extending from this atom, meaning all positions of the attached hydrogen and terminal atoms.

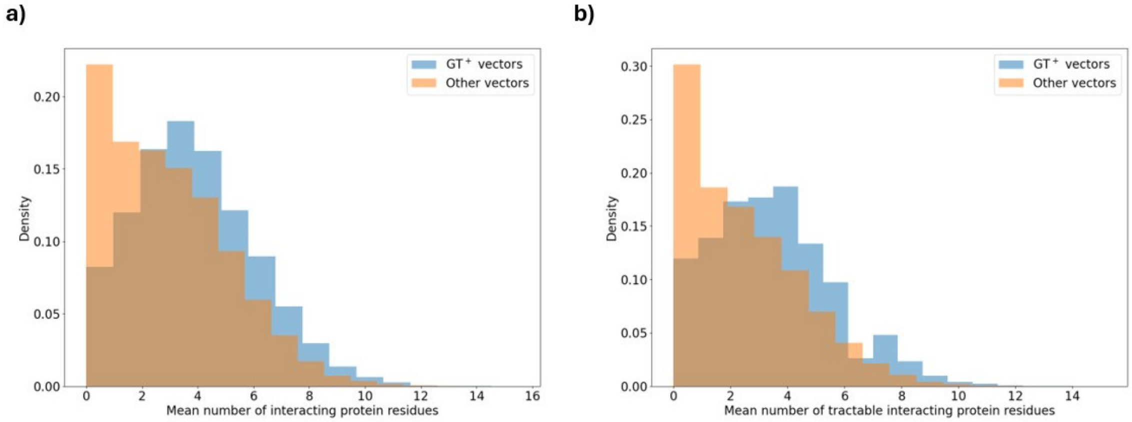


Figure 5.7: **The numbers of protein residues reached by non-clashing decorators for vectors in the PDBbind synthetic calibration set.** (**a**) The total number of protein residues that form a potential hydrogen bond with a non-clashing decorator in the elaboration 'cloud' for a given vector is shown. (**b**) The numbers of protein residues that form a potential hydrogen bond with a non-clashing decorator after filtering for decorators that are considered synthetically tractable for the vector are shown. Data are shown for the positively labelled ground-truth (GT$^+$) vectors versus all other possible vectors. The mean is calculated at each vector atom as the average across all possible vectors extending from this atom, meaning all positions of the attached hydrogen and terminal atoms.

Table 5.3 shows the percentages of vectors that were labelled as positive by both methods; the results are shown for the three tiers of the library-based scoring function and at different probability thresholds for the EGNN-based method. The EGNN score is able to more clearly distinguish between GT$^+$ and other vectors in the OOD set (see Figure 5.8); for example, at a probability threshold of 0.5, 82.4% of GT$^+$ vectors are labelled as positive, while 13.8% of other vectors are 'incorrectly' labelled as positive. The library-based scoring function, in comparison, is less precise and shows a less stark distinction in recall between the two vector classes; for example, when the scoring function takes into account the number of clashes, interactions and tractability, 74.2% of GT$^+$ vectors and 51.4% of other vectors are labelled positive. A similar pattern was observed when using different percentile thresholds based on the number of clashing decorators (Appendix Table D.1).

While not all of the GT$^+$ vectors will necessarily represent the best vector opportunity in the molecule, we would expect enrichment for elaboratable opportunities within this subset, which is reflected in the results across all three levels of scoring. Although we do not have a precise estimation of what is the expected percentage of other vectors to be positive, manual curation on our small Malhotra subset (described in Section 5.3.4.3) led us to estimate that ∼62% of vectors could be considered elaboratable; thus, in their present form, the library-based approach may seem to agree better with the data, and the EGNN-based approach may be overfitting to the training dataset.

Figure 5.9 shows two examples from the OOD test set and a visualization for the vector scoring for both methods. As we can see, the EGNN regards fewer vectors as elaboratable, and does not identify the correct vector in Figure 5.9a. The library-based method scores more vectors as elaboratable opportunities and visual inspection suggests that these vectors have greater room in the pocket for growth. While less precise, the results are perhaps more interpretable due to the nature of the scoring.

One observation for the OOD test set was that the molecular property filters, including a molecular weight threshold cut-off 1000 Da and no more than 20 rotatable bonds, were not strict enough as a subset of the ligands seem to be large and peptidic. These do not accurately reflect

Table 5.3: Percentage of vectors marked as elaboratable for the PDBbind synthetic dataset.

		Train		Test	
Method	Score level	% GT$^+$ vectors	% other vectors	% GT$^+$ vectors	% other vectors
Library-based	Clash score	80.0	63.0	79.0	59.3
	Clash + interaction score	76.1	58.7	76.5	55.4
	Clash + interaction + tractability score	74.3	54.5	74.2	51.4
EGNN-based	$P \geq 0.1$	98.4	32.9	97.5	35.9
	$P \geq 0.2$	96.3	23.5	94.2	26.9
	$P \geq 0.3$	93.6	18.0	91.9	21.3
	$P \geq 0.4$	90.5	13.8	87.5	16.9
	$P \geq 0.5$	86.3	10.5	82.4	13.8
	$P \geq 0.6$	80.8	7.6	76.1	10.0
	$P \geq 0.7$	73.5	5.0	71.2	6.5
	$P \geq 0.8$	62.7	2.8	59.1	4.0
	$P \geq 0.9$	45.1	1.0	36.8	1.4

EGNN, equivariant graph neural network; GT$^+$, positively labelled ground-truth.

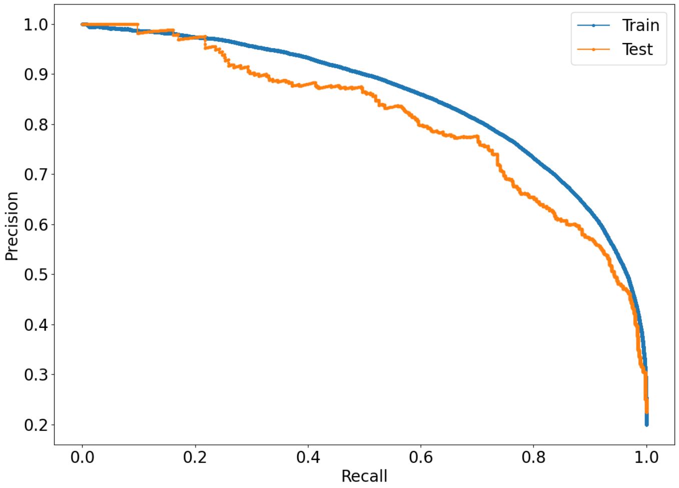


Figure 5.8: **Precision–recall curves for EGNN-based evaluation of vectors in the PDBbind synthetic dataset.** The curves show the precision and recall for predictions made by the equivariant graph neural network (EGNN)-based method for identifying ground-truth positively labelled vectors in the PDBbind synthetic calibration (training) and test sets. It is worth noting that the estimations for recall may not be realiable, as not all of the ground-truth vectors are labelled for the synthetic dataset.

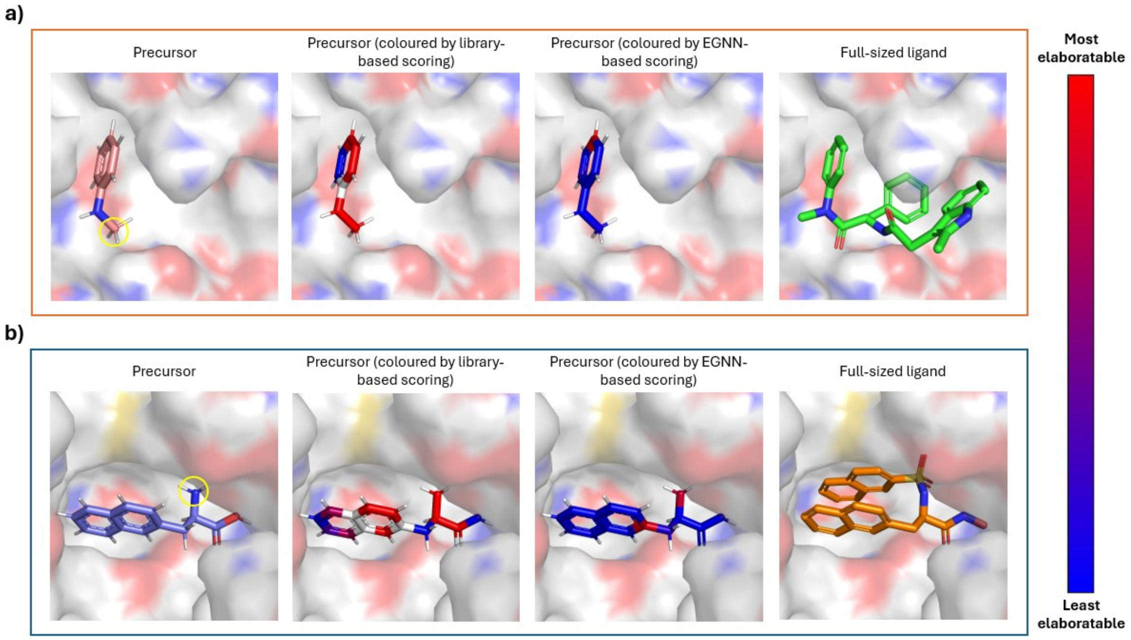

a)
Precursor
Precursor (coloured by library-based scoring)
Precursor (coloured by EGNN-based scoring)
Full-sized ligand
b)
Precursor
Precursor (coloured by library-based scoring)
Precursor (coloured by EGNN-based scoring)
Full-sized ligand
Most elaboratable
Least elaboratable

Figure 5.9: **Example visualization of vector scoring for compounds in the PDBbind synthetic OOD dataset.** Two examples of precursors plus the full-sized ligands in the PDBbind synthetic out-of-distribution (OOD) test set are shown, including PDB complexes (**a**) 2VES and (**b**) 2XDE. Vectors are scored using the library-based and equivariant graph neural network (EGNN)-based scoring approaches. Vectors are scored on a blue–red spectrum (colour bar is shown), where blue represent poor elaboration opportunities and red represents high-scoring vectors. White is used to indicate non-vector atoms, only differentiated by the library-based scoring method. The colours for the library-based scoring methods are based on the number of criteria obeyed by the vector. Precursors were generated using AiZynth (Genheden et al., 2020).

complexes we would expect from common elaboration scenarios and thus, in future work, these structures should be removed and the data should be filtered for pairs of complexes where the size increase reflects that expected in a typical elaboration scenario (similar to that described by Malhotra et al. (2017)).

5.4.3 Manually-labelled Malhotra dataset

The scoring functions were also run on the manually labelled Malhotra dataset to assess their ability to differentiate between positive and negative vectors, opposed to GT$^+$ and unlabelled vectors for the synthetic set.

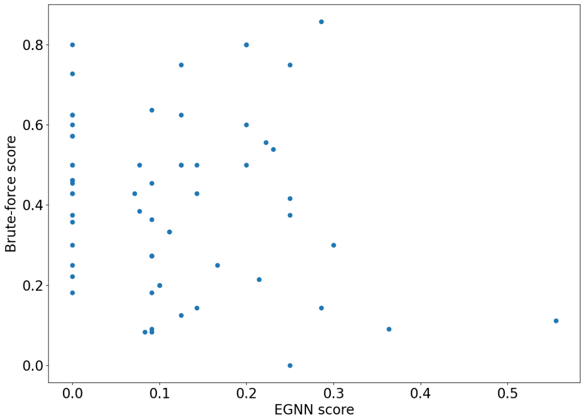


Figure 5.10: **Library-based versus EGNN-based scoring of compound elaboratability for molecules in the Malhotra dataset.** Compound elaboratability is scored by calculating the ratio of the number of elaboratable vectors against the total number of possible vectors within a molecule. The scores observed for the library-based scoring function and the equivariant graph neural network (EGNN)-based scoring function are shown.

Table 5.4 shows that the EGNN distinguished very few of the positive vectors as elaboratable opportunities (only 11.5% were marked as elaboratable with a probability threshold of ≥0.5) and there was little distinction between the scoring for positive versus negative vectors (9.2% for the latter). The library-based scoring function in comparison identified far fewer of the elaboratable positive vectors in comparison with in the synthetic set (47.2% for the most stringent tier of scoring) but still scored a higher proportion of the positive vectors as elaboratable in comparison to negative vectors (28.0%).

5.4.3.1 Ranking using compound elaboratablity

With regard to evaluating compound elaboratability, opposed to scoring vector elaboratability alone, whole molecules were evaluated by calculating the ratio between the number of elaboratable vectors and the number of theoretically possible vectors (described in 5.3.3); a probability threshold of 0.5 was applied to the EGNN predictions to distinguish elaboratable vectors. When using compound elaboratability scores to rank molecules, no statistically significant correlation was observed (Figure 5.10).

Table 5.4: Percentage of vectors marked as elaboratable for Malhotra dataset.

Method	Score level	% positive vectors	% negative vectors
Library-based	Clash score	49.1	31.3
	Clash + interaction score	48.5	30.8
	Clash + interaction + tractability score	47.2	28.0
EGNN-based	$P \geq 0.1$	33.5	23.8
	$P \geq 0.2$	24.5	17.5
	$P \geq 0.3$	17.6	15.4
	$P \geq 0.4$	14.1	12.1
	$P \geq 0.5$	11.5	9.2
	$P \geq 0.6$	9.1	5.8
	$P \geq 0.7$	6.7	4.6
	$P \geq 0.8$	4.0	3.8
	$P \geq 0.9$	2.4	0.4

EGNN, equivariant graph neural network.

Figures 5.11 and 5.12 show examples of the best and worst-scoring molecules according to both scoring functions. Figure 5.11 seems to show a noticeable distinction between highly elaboratable and non-elaboratable molecules with the library-based approach, as the former show ligands in wide pockets with clear elaboration oppotunities, while non-elaboratable molecules are visibly buried in the pocket, with few potential directions for growth available. This is less evident for the EGNN-based scoring function (Figure 5.12); for example, the third highest-scoring molecule appears to be in an internal pocket, and the poor-scoring molecules are in relatively open pockets. There may be several reasons for this drop in performance; for example, the EGNN may be learning some biases in relation to the precursor molecules, which may occur if the elaboratable vector atoms are often repeated across different protein complexes, or it may be a product of the threshold applied. This could also be a product of treating the unlabelled data as negative data using BCE as our loss function, rather than using a strategy for positive–unlabelled data. Further work will be needed to determine the generalizability of the EGNN approach.

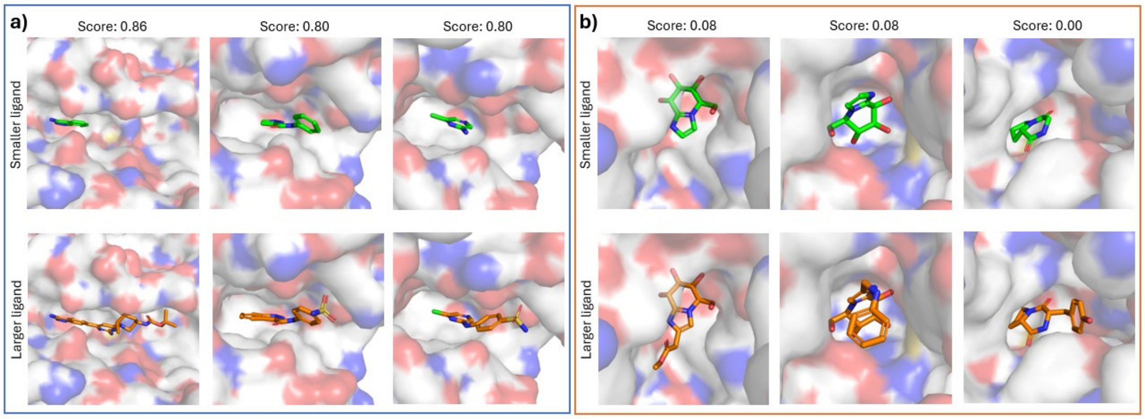


Figure 5.11: **Ranked Malhotra compounds according to their elaboratability score from the library-based scoring method.** Compounds in the Malhotra dataset were scored by calculating the ratio between the number of predicted elaboratable vectors versus the total number of theoretically possible vectors. The (**a**) best and (**b**) worst three compounds are shown (green; top row), together with the ground-truth full-sized ligands (orange; bottom row).

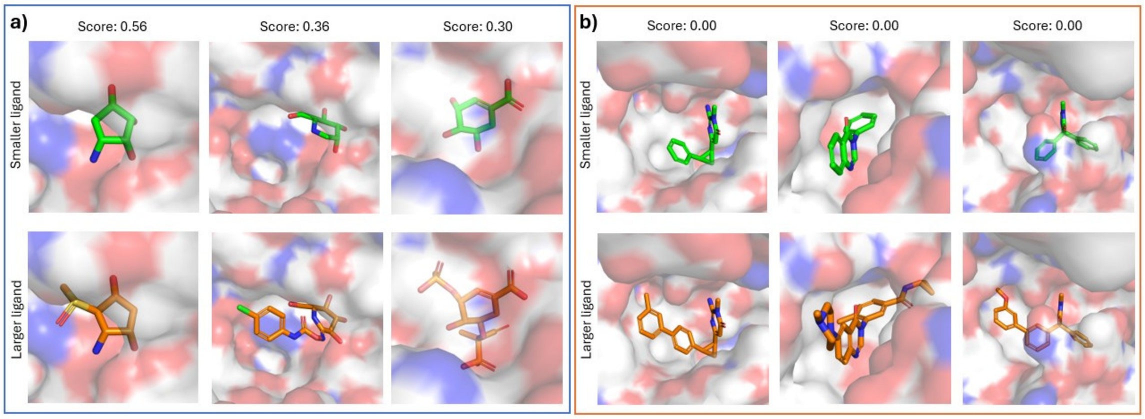


Figure 5.12: **Ranked Malhotra compounds according to their elaboratability score from the EGNN.** Compounds in the Malhotra data set were scored by calculating the ratio between the number of predicted elaboratable vectors (vectors with a probability of ≥ 0.5 predicted by the equivariant graph neural network (EGNN) model) versus the total number of theoretically possible vectors. The (**a**) best and (**b**) worst three compounds are shown (green; top row), together with the ground-truth full-sized ligands (orange; bottom row).

5.4.4 Computational efficiency for the library-based scoring method

As mentioned above, our library-based scoring method enabled us to evaluate elaboratability from both a geometric and chemical perspective, the latter of which required the use of AiZynthFinder's reaction filter network to rule out chemically infeasible decorators. When we analyse the time required to run the elaboratability evaluation, Figure 5.13 shows that the addition of AiZynthFinder to assess chemical tractability resulted in only a marginal increase in time required. The increase is much smaller than that observed when using a larger elaboration library (Figure 5.14). For this reason, we settled on the smaller elaboration library whereby elaborations occurred in the ChEMBL dataset as least 10,000 times. While the majority of molecules required less than 10 CPU seconds to be evaluated (99.8% and 99.2%, with and without AiZynthFinder, respectively), this value could be optimized in future iterations. In particular, the method we use to evaluate the position of the growth vectors could be improved. At present, hydrogen vectors are rotated around the vector to fully sample the elaboration space, but this could be condensed into a single calculation to improve efficiency. Moreover, the calculation could be run on multiple cores, to achieve better timings.

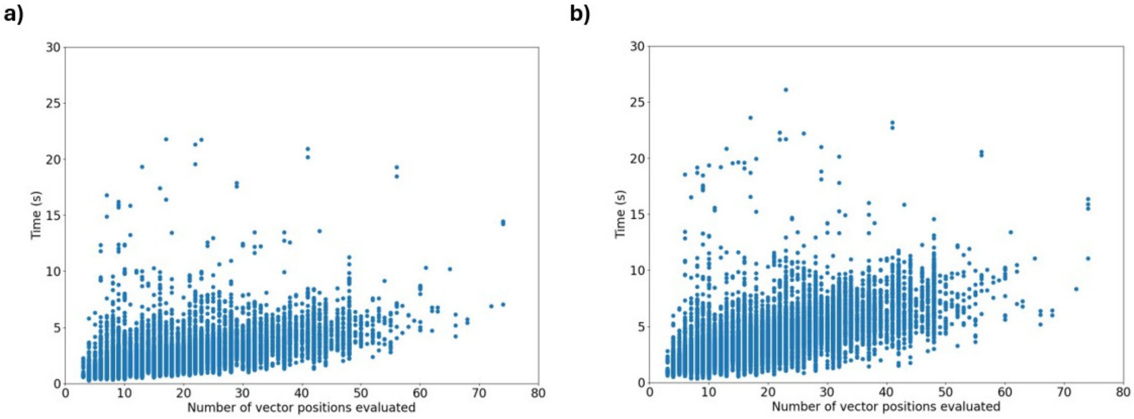


Figure 5.13: **Time required for evaluating elaboratability with the brute-force method for a given molecule in the PDBbind synthetic set.** Timings represent the total CPU time (in seconds), run on one CPU, and are shown for the scoring function (**a**) without AiZynth and (**b**) with AiZynth to evaluate synthetic tractability. The number of vector 'positions' are shown (this does not represent the number of unique vector atoms within the molecule, but the number of possible positions for across all vectors). These timings have the potential to be reduced in future iterations of the tool.

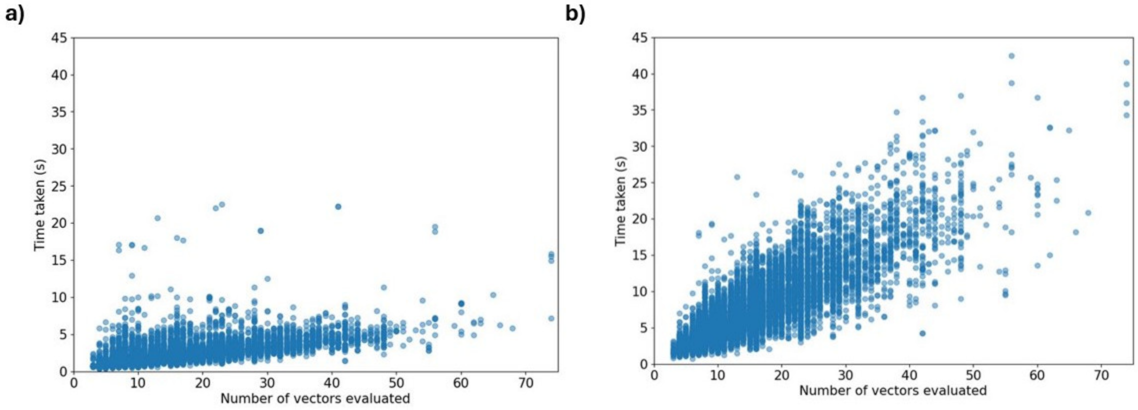


Figure 5.14: **Time required for evaluating elaboratability depending on elaboration library size for a given molecule in the PDBbind synthetic set.** Timings represent the total CPU time (in seconds), run on one CPU, and are shown for the scoring function without the use of AiZynth with an elaboration library where elaborations are filtered for those that occur (**a**) at least 10,000 (**b**) and at least 5,000 times. The number of vector 'positions' are shown (this does not represent the number of unique vector atoms within the molecule, but the number of possible positions for across all vectors).

5.5 Discussion

The work described in this Chapter provides a way to define elaboratability and outlines some initial approaches for how it can be evaluated. However, more validation and testing are needed to identify the optimal parameters for the scoring functions and to improve performance.

While, on the surface, the EGNN exhibited better recall for GT^+ vectors on the synthetic dataset than the library-based approach, this may be a result of it overfitting to the training data and thus ruling out too many of the unlabelled vectors as non-elaboratable. In their current form, it is difficult to draw exact conclusions on which method may be performing better and further work is needed to determine this. Moreover, it is worth noting that the EGNN underwent substantial parameter optimization; likely, more calibration will be needed to optimize the library-based method. However, it could provide a more interpretable way of distinguishing between the types of vector opportunity, such as whether the vector is tractable, solvent-exposed or faces steric impediments.

There are several potential avenues to explore in optimizing the tools described here. For both methods, it may be beneficial to generate a larger synthetic dataset for training or calibration by using docked ligands (for example, DiffDec used CrossDocked compounds for generating its training data) (Xie et al., 2024).

We could also explore other methods for generating possible precursors. While we chose the AiZynthFinder approach (Genheden et al., 2020) to provide more realistic elaboration scenarios — which allows small modifications to be made to the core ligand before elaboration — this may limit the number and diversity in the core ligands to be elaborated, as we observed some redundancy in the precursors across different protein complexes. Instead of using a reaction-based 'slicing' approach, as described in the literature, another option could be to place fragments from common fragment libraries by performing 'fuzzy' substructure matching, enabling few atom changes within the matched substructure. This may have the benefit of being more representative of fragment elaboration scenarios observed in the literature. In addition, incorporating negative data for the labelled vectors, or using positive versus unlabelled strategies for training, could be beneficial.

Another important filter to apply to the generated synthetic data is to limit the pairs to those where the size change reflects that typically observed in an elaboration scenario (as described by Malhotra et al. (2017)), to prevent incorporation of non-lead-like molecules and peptides.

Other changes could be made to the library-based scoring function itself. These include using other methods to assess whether a conformer in the generated decorator library represents a feasible addition to the molecule, by applying some measure of internal strain to determine whether the conformation is physically reasonable. Furthermore, we could expand the repertoire of possible interactions that are calculated between the elaboration cloud and the protein, beyond only hydrogen bonds. The geometry of the interaction could be assessed, for example, by comparing against data in the IsoStar library (Bruno et al., 1997), which contains CCDC data for non-bonded interactions observed in crystallographic structures. It is also important to add additional considerations as to whether the elaboration vectors are already involved in an interaction, and whether elaboration at a given vector will ablate an already important contributor to binding.

While we used a simple binary scoring method for labelling vectors as positive or negative, more sophisticated approaches to scoring could be used in future iterations of this tool. For example, it may be helpful to use some probability density estimation to score the probability of a query vector being elaboratable. Importantly, thorough validation of all scoring approaches will be needed, and providing a comparison against a manually annotated dataset by multiple expert chemists would be extremely valuable for determining the effectiveness of these methods.

5.6 Conclusions

This Chapter describes proof-of-concept work for evaluating vector and compound elaboratability using both a chemistry-informed, library-based approach and an EGNN-based method, similar to that described in the literature. Our library-based method relies on a pre-enumerated library of decorators to generate a point cloud representation of the elaboration space; this point cloud can subsequently be used to gauge whether elaboration at a given growth vector will be geometrically feasible — depending on whether there is volume to grow and accessible interactions — and chemically tractable, using a reaction filter assessment with AiZynthFinder. While the EGNN-based method seemed to exhibit better recall on a synthetically generated dataset of precursor ligands using complexes from PDBbind, it is unclear if this is due to the model learning dataset biases, as it did not appear to generalize well to an independent, manually labelled dataset. Fine-tuning these methods and providing comprehensive validation by comparing predictions against expert annotation will be valuable for improving the power of these tools.

6. Conclusions and future work

While crystallographic fragment screening has undergone rapid technological advancements in recent years, knowing how to effectively use the information gained from the resulting hits remains a challenge and a bottleneck within FBDD. A growing number of *de novo* and DL approaches have been developed to propose follow-up compounds for a fragment screen that maximize potency and explore new areas of chemical space. However, due to the costs and challenges associated with low synthetic accessibility of these compounds, catalogue-based search methods offer a way to accelerate this process. This thesis has largely focused on how to improve the efficiency with which we sample accessible chemical space to allow fragment progression, focusing on methods to merge fragment hits, and approaches to prioritize compounds and growth vectors based on their elaboratability. In this Chapter, we summarize the overall findings of this thesis and outline avenues for future work.

6.1 Catalogue-based approaches for fragment merging

In Chapter 2, we described the development of a pipeline for performing catalogue-based fragment merging using an implementation of the Fragment Network (Hall et al., 2014), a graph database that allows the user to probe the chemical space around fragment hits in a chemically intuitive manner.

The pipeline was developed to identify perfect merges that incorporate exact substructures of the parent fragments, identifying compounds that can be purchased from the vendor to enable rapid follow-up of fragment screening results. A filtering pipeline was implemented, taking into account 2D and 3D properties of the proposed compounds and, in particular, prioritizing compounds that have the potential to recapitulate the original fragment poses by generating conformers using

Fragmenstein (Ferla et al., 2024). We contrasted the results with a more traditional approach to analogue-by-catalogue searching, by performing a similarity search using molecular fingerprints. The results indicated that the two techniques were complementary and could improve the productivity of a catalogue search if used in parallel, as each was able to find merges for certain fragment pairs that were not represented by the other. Retrospective analyses using two case studies showed that the Fragment Network-based method was able to identify potential known binders with low double-digit micromolar affinity values.

In this work, we identified several limitations, some of which we addressed in Chapter 3. This Chapter built on the original pipeline by providing some new feature enhancements, and improved the power of the tool by expanding the chemical space searched, increasing the diversity of the resulting set of follow-up compounds. Some of the limitations addressed include querying the database at the substructure level and ensuring that the substructures to be incorporated into a single compound are spatially compatible and predicted to be responsible for making an interaction, thus improving the ability of the method to identify productive merges. Additionally, we expanded the search to identify what we term bioisosteric merges, as these compounds incorporate substructures from the original fragments that have pharmacophoric similarity, rather than representing an exact chemical match. Comparison against perfect merging showed bioisosteric merging was able to retrieve merges for fragment pairs not found by perfect merging alone. A case study involving compounds against WDR5-MYC showed that the new pipeline enables the network to begin to replicate some of the design decisions made during SAR analysis.

Importantly, we compared the results with an 'out-of-the-box' approach using a pharmacophore-constrained docking pipeline, which aimed to screen database compounds for those that replicate pharmacophore features observed in the fragments, maintaining a similar spatial arrangment. The time required to identify high-scoring follow-up compounds from the Fragment Network, which we term 'computational merging hits' based on their shape and colour scores and proportion of interactions recapitulated, was substantially reduced across several test cases involving nine fragment pairs across three targets. This indicates that the Fragment Network shows great promise in providing a highly automated method, requiring limited human intervention, for performing an exhaustive enumeration of fragment follow-up compounds based on the results of large-scale screens.

Chapter 4 provided an insight into how active antiviral campaigns as part of the ASAP project provided a platform for method development. Design rounds for two enteroviral targets helped to craft the logistics required for designing and ordering compounds, a collaborative process involving multiple individuals, and uncovered some key lessons and challenges, summarized here. First, an exhaustive enumeration approach may fail to generate productive follow-up compounds, particularly for difficult targets where flexibility in the binding site is observed, leading to compounds that either do not bind or bind in an unexpected orientation. A potential solution we identified is to incorporate filtering using multiple conformations of the target, if available. Additionally, the design rounds established the importance of setting explicit design objectives at the outset, as it provides clarity on whether an exhaustive or more focused approach is appropriate. Effective communication between experimentalists and designers was found to be crucial, especially when dealing with ambiguities in electron densities; establishing a clear process for managing these compounds will be key moving forward. Finally, a need for greater transparency in the scoring process was established, to ensure that design decisions are well-understood and can provide actionable feedback for further iterations.

Several other opportunities for improvement were identified for this work and for follow-up design tools more generally. First, incorporating negative data into the design process could be valuable for ensuring that labour and financial budgets are not wasted on particular substructures or interactions that do not contribute substantially to fragment binding, and so are less likely to be observed upon elaboration. Additionally, the Fragment Network approach, which has been optimized for exhaustive enumeration, could be adapted to allow for more hypothesis-driven design, for example, if a key interaction opportunity is to be prioritized. In certain circumstances, more control over the enumeration process may be necessary, particularly to prevent the addition of extraneous substructures that do not overlap with the fragment volume or fail to create new productive interactions.

There are several points to consider for future work, some of which will be addressed at XChem. While we outlined the challenges with incorporating complete atom mapping that considers molecular symmetry into the network, incorporating this into future databases would allow for finer control over query paths, what substructures are being merged together and the attachment points to be used for these transformations.

The composition of the network is currently being assessed, and new datasets are being designed so that the compounds are particularly suitable for synthesis, using the repertoire of reactions that are possible to execute on XChem's chemist-assisted robotics platform. An interesting perspective would be to evaluate the extent to which such datasets may expedite the DMTA process compared with the costs and timelines associated with ordering directly from vendors.

It may also be interesting to compare our approach with recent active learning methods used for rapid exploration of make-on-demand chemical spaces, described in Section 1.4.4.5, evaluating how they compare in terms of database coverage and efficiency. Another interesting aspect could be incorporating 3D pharmacophore or shape-based descriptors in the database, to determine if the benefit of using these properties in a database search outweighs the increased computational requirements.

Finally, prospective validation will be crucial for evaluating the efficacy of these methods. There are plans to use the tool in XChem's upcoming campaigns, which could provide valuable experimental validation and help further refine the method based on real-world results.

6.2 Elaboratability as a basis for compound prioritization

Chapter 5 focused on how to prioritize compounds and growth vectors for further elaboration, using a relatively new concept of elaboratability. While there are some similar concepts described in the literature — such as fragment sociability, or deep learning models, which learn an implicit definition of elaboratability by selecting the growth vector to expand from — there is still a lack of methods that provide a comprehensive assessment of both the chemical and geometric tractability for further synthesis. Such a tool could be useful for many applications in drug discovery.

The work in this Chapter outlined methodology for evaluating vector and compound elaboratability by pre-enumerating conformers for a library of decorators that could be used to construct a point cloud representing the 'elaboration space'; this could in turn be used to evaluate whether elaboration will result in a high volume of clashes or new potential interactions, in addition to considerations of chemical feasibility using a component of the AiZynthFinder tool (Genheden et al., 2020). We also constructed a new dataset containing pairs of core and elaborated ligands bound to the same protein, which was designed to more accurately represent realistic elaboration scenarios compared with existing synthetic datasets. While our tool seemed to exhibit worse recall for recovering positively labelled ground-truth vectors in our synthetic dataset, the EGNN method may be resulting in too many false negatives. Our tool could provide a more interpretable way of evaluating elaboratability and be less subject to the biases learnt by the EGNN model.

There are several interesting avenues to explore for this work. With regard to the dataset, possibilities include expanding the dataset by incorporating docked compounds, which may reduce bias to certain overrepresented precursors, or utilizing a different approach to synthetic data generation through the placement of fragments from common libraries using fuzzy substructure matching. Additionally, exploring more schemas for scoring the compounds could help improve score performance, for example, by conducting a more robust analysis of the calibration dataset to determine optimal thresholds, or by analysing probability densities for different scoring metrics. One of the most interesting pieces of validation would be to use a dataset where vectors have been annotated by multiple experts, together with a ranking of the compounds in terms of their elaboratability, to evaluate whether these methods replicate the chemist's decisions. This would be similar to the methodology described by Ertl et al. (2009), where compound rankings according to the SAScore were compared against rankings from nine chemists.

A key insight will be applying these scoring methods in real-world campaigns to assess whether they enhance the efficiency of the design and prioritization processes.

Bibliography

Adasme, Melissa F, Katja L Linnemann, Sarah Naomi Bolz, Florian Kaiser, Sebastian Salentin, V Joachim Haupt, and Michael Schroeder (2021). "PLIP 2021: expanding the scope of the protein–ligand interaction profiler to DNA and RNA". In: *Nucleic Acids Research* 49.W1, W530–W534.

Andrianov, Grigorii V., Wern Juin Gabriel Ong, Ilya Serebriiskii, and John Karanicolas (2021). "Efficient Hit-to-Lead Searching of Kinase Inhibitor Chemical Space via Computational Fragment Merging". In: *Journal of Chemical Information and Modeling* 61.12, pp. 5967–5987.

Armstrong, M. Stuart, Garrett M. Morris, Paul W. Finn, Raman Sharma, Loris Moretti, Richard I. Cooper, and W. Graham Richards (2010). "ElectroShape: fast molecular similarity calculations incorporating shape, chirality and electrostatics". In: *Journal of Computer-Aided Molecular Design* 24.9, pp. 789–801.

Arús-Pous, Josep, Atanas Patronov, Esben Jannik Bjerrum, Christian Tyrchan, Jean-Louis Reymond, Hongming Chen, and Ola Engkvist (2020). "SMILES-based deep generative scaffold decorator for *de novo* drug design". In: *Journal of Cheminformatics* 12.1, p. 38.

ASAP Discovery Consortium (2024). *AI-driven Structure-enabled Antiviral Platform (ASAP).* URL: `https://asapdiscovery.org/` (visited on 07/25/2024).

Astero, Maryam and Juho Rousu (2024). "Learning symmetry-aware atom mapping in chemical reactions through deep graph matching". In: *Journal of Cheminformatics* 16.1, p. 46.

Baell, Jonathan B. and Georgina A. Holloway (2010). "New Substructure Filters for Removal of Pan Assay Interference Compounds (PAINS) from Screening Libraries and for Their Exclusion in Bioassays". In: *Journal of Medicinal Chemistry* 53.7, pp. 2719–2740.

Bajusz, Dávid, Anita Rácz, and Károly Héberger (2015). "Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations?" In: *Journal of Cheminformatics* 7.1, p. 20.

Ballester, Pedro J. and W. Graham Richards (2007). "Ultrafast shape recognition to search compound databases for similar molecular shapes". In: *Journal of Computational Chemistry* 28.10, pp. 1711–1723.

Bekker, Jessa and Jesse Davis (2020). "Learning from positive and unlabeled data: a survey". In: *Machine Learning* 109.4, pp. 719–760.

Bender, Brian J., Stefan Gahbauer, Andreas Luttens, Jiankun Lyu, Chase M. Webb, Reed M. Stein, Elissa A. Fink, Trent E. Balius, Jens Carlsson, John J. Irwin, and Brian K. Shoichet (2021). "A practical guide to large-scale docking". In: *Nature Protocols* 16.10, pp. 4799–4832.

Berman, Helen M., John Westbrook, Zukang Feng, Gary Gilliland, T. N. Bhat, Helge Weissig, Ilya N. Shindyalov, and Philip E. Bourne (2000). "The Protein Data Bank". In: *Nucleic Acids Research* 28.1, pp. 235–242.

Bieniek, Mateusz K., Ben Cree, Rachael Pirie, Joshua T. Horton, Natalie J. Tatum, and Daniel J. Cole (2022). "An open-source molecular builder and free energy preparation workflow". In: *Communications Chemistry* 5.1, pp. 1–9.

Blay, Vincent, Bhairavi Tolani, Sunita P. Ho, and Michelle R. Arkin (2020). "High-Throughput Screening: today's biochemical and cell-based approaches". In: *Drug Discovery Today* 25.10, pp. 1807–1821.

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

Bohacek, R S, C McMartin, and W C Guida (1996). *The art and practice of structure-based drug design: a molecular modeling perspective.* United States.

Bouysset, Cédric and Sébastien Fiorucci (2021). "ProLIF: a library to encode molecular interactions as fingerprints". In: *Journal of Cheminformatics* 13.1, p. 72.

Boyles, Fergus, Charlotte M Deane, and Garrett M Morris (2019). "Learning from the ligand: using ligand-based features to improve binding affinity prediction". In: *Bioinformatics* 36.3, pp. 758–764.

Brown, Dean G. and Jonas Boström (2016). "Analysis of Past and Present Synthetic Methodologies on Medicinal Chemistry: Where Have All the New Reactions Gone?" In: *Journal of Medicinal Chemistry* 59.10, pp. 4443–4458.

Bruno, I. J., J. C. Cole, J. P. Lommerse, R. S. Rowland, R. Taylor, and M. L. Verdonk (1997). "IsoStar: a library of information about nonbonded interactions". In: *Journal of Computer-Aided Molecular Design* 11.6, pp. 525–537.

Bruno, Ian J., Jason C. Cole, Magnus Kessler, Jie Luo, W. D. Sam Motherwell, Lucy H. Purkis, Barry R. Smith, Robin Taylor, Richard I. Cooper, Stephanie E. Harris, and A. Guy Orpen (2004). "Retrieval of Crystallographically-Derived Molecular Geometry Information". In: *Journal of Chemical Information and Computer Sciences* 44.6, pp. 2133–2144.

Butina, Darko (1999). "Unsupervised data base clustering based on Daylight's fingerprint and Tanimoto similarity: a fast and automated way to cluster small and large data sets". In: *Journal of Chemical Information and Computer Sciences* 39.4, pp. 747–750.

Buttenschoen, Martin, Garrett M. Morris, and Charlotte M. Deane (2024). "PoseBusters: AI-based docking methods fail to generate physically valid poses or generalise to novel sequences". In: *Chemical Science* 15.9, pp. 3130–3139.

Carbery, Anna, Rachael Skyner, Frank von Delft, and Charlotte M. Deane (2022). "Fragment libraries designed to be functionally diverse recover protein binding information more efficiently than standard structurally diverse libraries". In: *Journal of Medicinal Chemistry* 65.16, pp. 11404–11413.

Chacón Simon, Selena, Feng Wang, Lance R. Thomas, Jason Phan, Bin Zhao, Edward T. Olejniczak, Jonathan D. Macdonald, J. Grace Shaw, Caden Schlund, William Payne, Joy Creighton, Shaun R. Stauffer, Alex G. Waterson, William P. Tansey, and Stephen W. Fesik (2020). "Discovery of WD Repeat-Containing Protein 5 (WDR5)–MYC Inhibitors Using Fragment-Based Methods and Structure-Based Design". In: *Journal of Medicinal Chemistry* 63.8, pp. 4315–4333.

Chaudhury, Sidhartha, Sergey Lyskov, and Jeffrey J. Gray (2010). "PyRosetta: a script-based interface for implementing molecular modeling algorithms using Rosetta". In: *Bioinformatics* 26.5, pp. 689–691.

Chen, Yu-Chian (2015). “Beware of docking!” In: *Trends in Pharmacological Sciences* 36.2, pp. 78–95.

Coley, Connor W. (2021). “Defining and Exploring Chemical Spaces”. In: *Trends in Chemistry* 3.2, pp. 133–145.

Coley, Connor W., Luke Rogers, William H. Green, and Klavs F. Jensen (2018). “SCScore: synthetic complexity learned from a reaction corpus”. In: *Journal of Chemical Information and Modeling* 58.2, pp. 252–261.

Collins, P. M., J. T. Ng, R. Talon, K. Nekrosiute, T. Krojer, A. Douangamath, J. Brandao-Neto, N. Wright, N. M. Pearce, and F. von Delft (2017). “Gentle, fast and effective crystal soaking by acoustic dispensing”. In: *Acta Crystallographica Section D: Structural Biology* 73.3, pp. 246–255.

Collins, Patrick M., Alice Douangamath, Romain Talon, Alexandre Dias, Jose Brandao-Neto, Tobias Krojer, and Frank von Delft (2018). “Achieving a good crystal system for crystallographic X-ray fragment screening”. In: *Methods in Enzymology* 610, pp. 251–264.

Cooper, David R., Przemyslaw J. Porebski, Maksymilian Chruszcz, and Wladek Minor (2011). “X-ray crystallography: Assessment and validation of protein-small molecule complexes for drug discovery”. In: *Expert Opinion on Drug Discovery* 6.8, pp. 771–782.

Correy, Galen J., Moira Rachman, Takaya Togo, Stefan Gahbauer, Yagmur U. Doruk, Maisie Stevens, Priyadarshini Jaishankar, Brian Kelley, Brian Goldman, Molly Schmidt, Trevor Kramer, Alan Ashworth, Patrick Riley, Brian K. Shoichet, Adam R. Renslo, W. Patrick Walters, and James S. Fraser (2024). “Extensive exploration of structure activity relationships for the SARS-CoV-2 macrodomain from shape-based fragment merging and active learning”. In: *bioRxiv* DOI: 10.1101/2024.08.25.609621.

Cox, Oakley B., Tobias Krojer, Patrick Collins, Octovia Monteiro, Romain Talon, Anthony Bradley, Oleg Fedorov, Jahangir Amin, Brian D. Marsden, John Spencer, Frank von Delft, and Paul E. Brennan (2016). “A poised fragment library enables rapid synthetic expansion yielding the first reported inhibitors of PHIP(2), an atypical bromodomain”. In: *Chemical Science* 7.3, pp. 2322–2330.

Credille, Cy V., Yao Chen, and Seth M. Cohen (2016). “Fragment-based identification of influenza endonuclease inhibitors”. In: *Journal of Medicinal Chemistry* 59.13, pp. 6444–6454.

Curran, Peter R., Chris J. Radoux, Mihaela D. Smilova, Richard A. Sykes, Alicia P. Higueruelo, Anthony R. Bradley, Brian D. Marsden, David R. Spring, Tom L. Blundell, Andrew R. Leach, William R. Pitt, and Jason C. Cole (2020). “Hotspots API: A Python Package for the Detection of Small Molecule Binding Hotspots and Application to Structure-Based Drug Design”. In: *Journal of Chemical Information and Modeling* 60.4, pp. 1911–1916.

Da, C. and D. Kireev (2014). “Structural Protein–Ligand Interaction Fingerprints (SPLIF) for Structure-Based Virtual Screening: Method and Benchmark Study”. In: *Journal of Chemical Information and Modeling* 54.9, pp. 2555–2561.

Davies, Douglas R., Darren W. Begley, Robert C. Hartley, Bart L. Staker, and Lance J. Stewart (2011). “Chapter four — Predicting the Success of Fragment Screening by X-Ray Crystallography”. In: *Methods in Enzymology*. Ed. by Lawrence C. Kuo. Vol. 493. Academic Press, pp. 91–114.

Davis, Ben J. and Stephen D. Roughley (2017). “Chapter Eleven - Fragment-Based Lead Discovery”. In: *Annual Reports in Medicinal Chemistry*. Ed. by Robert A. Goodnow. Vol. 50. Academic Press, pp. 371–439.

Deane, Charlotte and Maranga Mokaya (2022). “A virtual drug-screening approach to conquer huge chemical libraries”. In: *Nature* 601.7893, pp. 322–323.

Delft, Annette von, Matthew D. Hall, Ann D. Kwong, Lisa A. Purcell, Kumar Singh Saikatendu, Uli Schmitz, John A. Tallarico, and Alpha A. Lee (2023). “Accelerating antiviral drug discovery: lessons from COVID-19”. In: *Nature Reviews Drug Discovery* 22.7, pp. 585–603.

Delft, Frank von, Mark Calmiano, John Chodera, Ed Griffen, Alpha Lee, Nir London, Tatiana Matviuk, Ben Perry, Matt Robinson, and Annette von Delft (2021). “A white-knuckle ride of open COVID drug discovery”. In: *Nature* 594.7863, pp. 330–332.

Dey, Fabian and Amedeo Caflisch (2008). “Fragment-based de novo ligand design by multiobjective evolutionary optimization”. In: *Journal of Chemical Information and Modeling* 48.3, pp. 679–690.

Diamond Light Source. (2020). *Main protease structure and XChem fragment screen.* URL: `https://www.diamond.ac.uk/covid-19/for-scientists/Main-protease-structure-and-XChem.html` (visited on 07/08/2024).

Douangamath, Alice, Ailsa Powell, Daren Fearon, Patrick M. Collins, Romain Talon, Tobias Krojer, Rachael Skyner, Jose Brandao-Neto, Louise Dunnett, Alexandre Dias, Anthony Aimon, Nicholas M. Pearce, Conor Wild, Tyler Gorrie-Stone, and Frank von Delft (2021). "Achieving efficient fragment screening at XChem facility at Diamond Light Source". In: *JoVE*, e62414.

Downes, Thomas D., S. Paul Jones, Hanna F. Klein, Mary C. Wheldon, Masakazu Atobe, Paul S. Bond, James D. Firth, Ngai S. Chan, Laura Waddelove, Roderick E. Hubbard, David C. Blakemore, Claudia De Fusco, Stephen D. Roughley, Lewis R. Vidler, Maria Ann Whatton, Alison J.-A. Woolford, Gail L. Wrigley, and Peter O'Brien (2020). "Design and Synthesis of 56 Shape-Diverse 3D Fragments". In: *Chemistry – A European Journal* 26.41, pp. 8969–8975.

Durant, Guy, Fergus Boyles, Kristian Birchall, Brian Marsden, and Charlotte M. Deane (2023). "Robustly interrogating machine learning-based scoring functions: what are they learning?" In: *bioRxiv* DOI: 10.1101/2023.10.30.564251.

Durant, Joseph L., Burton A. Leland, Douglas R. Henry, and James G. Nourse (2002). "Reoptimization of MDL keys for use in drug discovery". In: *Journal of Chemical Information and Computer Sciences* 42.6, pp. 1273–1280.

Ebejer, Jean-Paul, Paul W. Finn, Wing Ki Wong, Charlotte M. Deane, and Garrett M. Morris (2019). "Ligity: a non-superpositional, knowledge-based approach to virtual screening". In: *Journal of Chemical Information and Modeling* 59.6, pp. 2600–2616.

Ebejer, Jean-Paul, Garrett M. Morris, and Charlotte M. Deane (2012). "Freely available conformer generation methods: how good are they?" In: *Journal of Chemical Information and Modeling* 52.5, pp. 1146–1158.

Emmerich, Christoph H., Lorena Martinez Gamboa, Martine C. J. Hofmann, Marc Bonin-Andresen, Olga Arbach, Pascal Schendel, Björn Gerlach, Katja Hempel, Anton Bespalov, Ulrich Dirnagl, and Michael J. Parnham (2021). "Improving target assessment in biomedical research: the GOT-IT recommendations". In: *Nature Reviews Drug Discovery* 20.1, pp. 64–81.

Enamine (2024). *Fragment Libraries.* URL: `https://enamine.net/compound-libraries/fragment-libraries` (visited on 10/03/2024).

Erlanson, Daniel A., Stephen W. Fesik, Roderick E. Hubbard, Wolfgang Jahnke, and Harren Jhoti (2016). "Twenty years on: the impact of fragments on drug discovery". In: *Nature Reviews Drug Discovery* 15.9, pp. 605–619.

Ertl, Peter (2020). "Craig plot 2.0: an interactive navigation in the substituent bioisosteric space". In: *Journal of Cheminformatics* 12.1, p. 8.

Ertl, Peter, Eva Altmann, and Sophie Racine (2023). "The most common linkers in bioactive molecules and their bioisosteric replacement network". In: *Bioorganic & Medicinal Chemistry* 81, p. 117194.

Ertl, Peter, Eva Altmann, Sophie Racine, and Richard Lewis (2022). "Ring replacement recommender: ring modifications for improving biological activity". In: *European Journal of Medicinal Chemistry* 238, p. 114483.

Ertl, Peter and Ansgar Schuffenhauer (2009). "Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions". In: *Journal of Cheminformatics* 1.1, p. 8.

Feng, Yu, Yuyao Yang, Wenbin Deng, Hongming Chen, and Ting Ran (2022). "SyntaLinker-Hybrid: a deep learning approach for target specific drug design". In: *Artificial Intelligence in the Life Sciences* 2, p. 100035.

Ferenczy, György G. and György M. Keserű (2016). "On the enthalpic preference of fragment binding". In: *MedChemComm* 7.2, pp. 332–337.

Ferla, Matteo P, Rubén Sánchez-García, Rachael E Skyner, Stefan Gahbauer, Jenny C Taylor, Frank von Delft, Brian D Marsden, and Charlotte M Deane (2024). "Fragmenstein: predicting protein-ligand structures of compounds derived from known crystallographic fragment hits using a strict conserved-binding–based methodology". In: *ChemRxiv* DOI: 10.26434/chemrxiv-2024-17w01.

Fey, Matthias and Jan Eric Lenssen (2019). "Fast Graph Representation Learning with PyTorch Geometric". In: *arXiv* DOI: 10.48550/arXiv.1903.02428.

Fialková, Vendy, Jiaxi Zhao, Kostas Papadopoulos, Ola Engkvist, Esben Jannik Bjerrum, Thierry Kogej, and Atanas Patronov (2022). "LibINVENT: reaction-based generative scaffold decoration for *in silico* library design". In: *Journal of Chemical Information and Modeling* 62.9, pp. 2046–2063.

Francoeur, Paul G., Tomohide Masuda, Jocelyn Sunseri, Andrew Jia, Richard B. Iovanisci, Ian Snyder, and David R. Koes (2020). "Three-Dimensional Convolutional Neural Networks and a Cross-Docked Data Set for Structure-Based Drug Design". In: *Journal of Chemical Information and Modeling* 60.9, pp. 4200–4215.

Friesner, Richard A., Jay L. Banks, Robert B. Murphy, Thomas A. Halgren, Jasna J. Klicic, Daniel T. Mainz, Matthew P. Repasky, Eric H. Knoll, Mee Shelley, Jason K. Perry, David E. Shaw, Perry Francis, and Peter S. Shenkin (2004). "Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy". In: *Journal of Medicinal Chemistry* 47.7, pp. 1739–1749.

Gao, Kaifu, Duc Duy Nguyen, Vishnu Sresht, Alan M Mathiowetz, Meihua Tu, and Guo-Wei Wei (2020a). "Are 2D fingerprints still valuable for drug discovery?" In: *Physical Chemistry Chemical Physics* 22.16, pp. 8373–8390.

Gao, Wenhao and Connor W. Coley (2020b). "The Synthesizability of Molecules Proposed by Generative Models". In: *Journal of Chemical Information and Modeling* 60.12, pp. 5714–5723.

Gaulton, Anna, Louisa J. Bellis, A. Patricia Bento, Jon Chambers, Mark Davies, Anne Hersey, Yvonne Light, Shaun McGlinchey, David Michalovich, Bissan Al-Lazikani, and John P. Overington (2012). "ChEMBL: a large-scale bioactivity database for drug discovery". In: *Nucleic Acids Research* 40, pp. D1100–D1107.

Genheden, Samuel, Amol Thakkar, Veronika Chadimová, Jean-Louis Reymond, Ola Engkvist, and Esben Bjerrum (2020). "AiZynthFinder: a fast, robust and flexible open-source software for retrosynthetic planning". In: *Journal of Cheminformatics* 12.1, p. 70.

Ghislat, Ghita, Taufiq Rahman, and Pedro J. Ballester (2021). "Recent progress on the prospective application of machine learning to structure-based virtual screening". In: *Current Opinion in Chemical Biology* 65, pp. 28–34.

Giordano, Deborah, Carmen Biancaniello, Maria Antonia Argenio, and Angelo Facchiano (2022). "Drug Design by Pharmacophore and Virtual Screening Approach". In: *Pharmaceuticals* 15.5, p. 646.

Goodford, P. J. (1985). "A computational procedure for determining energetically favorable binding sites on biologically important macromolecules". In: *Journal of Medicinal Chemistry* 28.7, pp. 849–857.

Gorgulla, Christoph, AkshatKumar Nigam, Matt Koop, Süleyman Selim Çınaroğlu, Christopher Secker, Mohammad Haddadnia, Abhishek Kumar, Yehor Malets, Alexander Hasson, Minkai Li, Ming Tang, Roni Levin-Konigsberg, Dmitry Radchenko, Aditya Kumar, Minko Gehev, Pierre-Yves Aquilanti, Henry Gabb, Amr Alhossary, Gerhard Wagner, Alán Aspuru-Guzik, Yurii S. Moroz, Konstantin Fackeldey, and Haribabu Arthanari (2023). "VirtualFlow 2.0 - The Next Generation Drug Discovery Platform Enabling Adaptive Screens of 69 Billion Molecules". In: *bioRxiv*.

Green, Harrison, David R. Koes, and Jacob D. Durrant (2021). "DeepFrag: a deep convolutional neural network for fragment-based lead optimization". In: *Chemical Science* 12.23, pp. 8036–8047.

Grosjean, Harold, Anthony Aimon, Storm Hassell-Hart, Warren Thompson, Lizbé Koekemoer, James Bennett, Cameron Anderson, Conor Wild, William Bradshaw, Edward A. FitzGerald, Tobias Krojer, Anthony Bradley, Oleg Fedorov, Philip C. Biggin, John Spencer, and Frank von Delft (2024). "Efficient large-scale exploration of fragment hit progression by exploiting binding-site purification of actives (B-SPA) through combining multi-step array synthesis and HT crystallography." In: *ChemRxiv* DOI: 10.26434/chemrxiv-2023-6m2s0-v2.

Grygorenko, Oleksandr O, Dmytro S Radchenko, Igor Dziuba, Alexander Chuprina, Kateryna E Gubina, and Yurii S Moroz (2020). "Generating multibillion chemical space of readily accessible screening compounds". In: *iScience* 23.11, p. 101681.

Hadfield, Thomas E., Fergus Imrie, Andy Merritt, Kristian Birchall, and Charlotte M. Deane (2022). "Incorporating target-specific pharmacophoric information into deep generative models for fragment elaboration". In: *Journal of Chemical Information and Modeling* 62.10, pp. 2280–2292.

Hall, Richard J., Paul N. Mortenson, and Christopher W. Murray (2014). “Efficient exploration of chemical space by fragment-based screening”. In: *Progress in Biophysics and Molecular Biology* 116.2, pp. 82–91.

Hall, Richard J., Christopher W. Murray, and Marcel L. Verdonk (2017). “The Fragment Network: a chemistry recommendation engine built using a graph database”. In: *Journal of Medicinal Chemistry* 60.14, pp. 6440–6450.

Hamilton, David J., Tom Dekker, Hanna F. Klein, Guido V. Janssen, Maikel Wijtmans, Peter O’Brien, and Iwan J. P. de Esch (2020). “Escape from planarity in fragment-based drug discovery: A physicochemical and 3D property analysis of synthetic 3D fragment libraries”. In: *Drug Discovery Today: Technologies* 38, pp. 77–90.

Hann, Michael M., Andrew R. Leach, and Gavin Harper (2001). “Molecular complexity and its impact on the probability of finding leads for drug discovery”. In: *Journal of Chemical Information and Modeling* 41.3, pp. 856–864.

Harner, Mary J., Andreas O. Frank, and Stephen W. Fesik (2013). “Fragment-Based Drug Discovery Using NMR Spectroscopy”. In: *Journal of Biomolecular NMR* 56.2, pp. 65–75.

Harren, Tobias, Torben Gutermuth, Christoph Grebner, Gerhard Hessler, and Matthias Rarey (2024). “Modern machine-learning for binding affinity estimation of protein–ligand complexes: Progress, opportunities, and challenges”. In: *WIREs Computational Molecular Science* 14.3, e1716.

Hawkins, Paul C. D., A. Geoffrey Skillman, and Anthony Nicholls (2007). “Comparison of Shape-Matching and Docking as Virtual Screening Tools”. In: *Journal of Medicinal Chemistry* 50.1, pp. 74–82.

Hevener, Kirk E., Russell Pesavento, JinHong Ren, Hyun Lee, Kiira Ratia, and Michael E. Johnson (2018). “Chapter Twelve - Hit-to-Lead: Hit Validation and Assessment”. In: *Methods in Enzymology.* Ed. by Charles A. Lesburg. Vol. 610. Academic Press, pp. 265–309.

Hilgenfeld, Rolf, Jian Lei, and Linlin Zhang (2018). “The structure of the Zika virus protease, NS2B/NS3pro”. In: *Dengue and Zika: Control and Antiviral Treatment Strategies.* Ed. by Rolf Hilgenfeld and Subhash G. Vasudevan. Singapore: Springer Singapore, pp. 131–145.

Hill, Maureen E., Anil Kumar, James A. Wells, Tom C. Hobman, Olivier Julien, and Jeanne A. Hardy (2018). “The unique cofactor region of Zika virus NS2B–NS3 protease facilitates cleavage of key host proteins”. In: *ACS Chemical Biology* 13.9, pp. 2398–2405.

Hindle, Sally A., Matthias Rarey, Christian Buning, and Thomas Lengaue (2002). “Flexible docking under pharmacophore type constraints”. In: *Journal of Computer-Aided Molecular Design* 16.2, pp. 129–149.

Hoffer, Laurent, Christophe Muller, Philippe Roche, and Xavier Morelli (2018). “Chemistry-driven Hit-to-lead Optimization Guided by Structure-based Approaches”. In: *Molecular Informatics* 37.9-10, p. 1800059.

Hoffmann, Torsten and Marcus Gastreich (2019). “The next level in chemical space navigation: going far beyond enumerable compound libraries”. In: *Drug Discovery Today* 24.5, pp. 1148–1156.

Hopkins, Andrew L., Colin R. Groom, and Alexander Alex (2004). “Ligand efficiency: a useful metric for lead selection”. In: *Drug Discovery Today* 9.10, pp. 430–431.

Hopkins, Andrew L., György M. Keserü, Paul D. Leeson, David C. Rees, and Charles H. Reynolds (2014). “The role of ligand efficiency metrics in drug discovery”. In: *Nature Reviews Drug Discovery* 13.2, pp. 105–121.

Horvath, Dragos (1997). “A Virtual Screening Approach Applied to the Search for Trypanothione Reductase Inhibitors”. In: *Journal of Medicinal Chemistry* 40.15, pp. 2412–2423.

Hu, Yanmei, Rami Musharrafieh, Madeleine Zheng, and Jun Wang (2020). “Enterovirus D68 antivirals: past, present, and future”. In: *ACS Infectious Diseases* 6.7, pp. 1572–1586.

Huang, Yinan, Xingang Peng, Jianzhu Ma, and Muhan Zhang (2022). “3DLinker: an E(3) equivariant variational autoencoder for molecular linker design”. In: *arXiv* DOI: 10.48550/arXiv.2205.07309, arXiv:2205.07309.

Hudson, Sean A., Kirsty J. McLean, Sachin Surade, Yong-Qing Yang, David Leys, Alessio Ciulli, Andrew W. Munro, and Chris Abell (2012). “Application of Fragment Screening and Merging to the Discovery of Inhibitors of the Mycobacterium tuberculosis CytochromeP450 CYP121”. In: *Angewandte Chemie International Edition* 51.37, pp. 9311–9316.

Hughes, JP, S Rees, SB Kalindjian, and KL Philpott (2011a). “Principles of early drug discovery”. In: *British Journal of Pharmacology* 162.6, pp. 1239–1249.

Hughes, Samantha J., David S. Millan, Iain C. Kilty, Russell A. Lewthwaite, John P. Mathias, Mark A. O’Reilly, Andrew Pannifer, Anne Phelan, Frank Stühmeier, Darren A. Baldock, and David G. Brown (2011b). “Fragment based discovery of a novel and selective PI3 kinase inhibitor”. In: *Bioorganic Medicinal Chemistry Letters* 21.21, pp. 6586–6590.

Igashov, Ilia, Hannes Stärk, Clément Vignac, Arne Schneuing, Victor Garcia Satorras, Pascal Frossard, Max Welling, Michael Bronstein, and Bruno Correia (2024). “Equivariant 3D-conditional diffusion model for molecular linker design”. In: *Nature Machine Intelligence* 6.4, pp. 417–427.

Imrie, Fergus, Anthony R. Bradley, Mihaela van der Schaar, and Charlotte M. Deane (2020). “Deep Generative Models for 3D Linker Design”. In: *Journal of Chemical Information and Modeling* 60.4, pp. 1983–1995.

Imrie, Fergus, Thomas E. Hadfield, Anthony R. Bradley, and Charlotte M. Deane (2021). “Deep generative design with 3D pharmacophoric constraints”. In: *Chemical Science* 12.43, pp. 14577–14589.

Irwin, John J., Khanh G. Tang, Jennifer Young, Chinzorig Dandarchuluun, Benjamin R. Wong, Munkhzul Khurelbaatar, Yurii S. Moroz, John Mayfield, and Roger A. Sayle (2020). “ZINC20—A Free Ultralarge-Scale Chemical Database for Ligand Discovery”. In: *Journal of Chemical Information and Modeling* 60.12, pp. 6065–6073.

Jhoti, Harren, Glyn Williams, David C. Rees, and Christopher W. Murray (2013). “The ‘rule of three’ for fragment-based drug discovery: where are we now?” In: *Nature Reviews Drug Discovery* 12.8, pp. 644–644.

Jin, Zhenming, Xiaoyu Du, Yechun Xu, Yongqiang Deng, Meiqin Liu, Yao Zhao, Bing Zhang, Xiaofeng Li, Leike Zhang, Chao Peng, Yinkai Duan, Jing Yu, Lin Wang, Kailin Yang, Fengjiang Liu, Rendi Jiang, Xinglou Yang, Tian You, Xiaoce Liu, Xiuna Yang, Fang Bai, Hong Liu, Xiang Liu, Luke W. Guddat, Wenqing Xu, Gengfu Xiao, Chengfeng Qin, Zhengli Shi, Hualiang Jiang, Zihe Rao, and Haitao Yang (2020). “Structure of Mpro from SARS-CoV-2 and discovery of its inhibitors”. In: *Nature* 582.7811, pp. 289–293.

Johnson, Marvin and Gerald M. Maggiora (1990). *Concepts and applications of molecular similarity.* URL: `https://api.semanticscholar.org/CorpusID:117506064` (visited on 07/25/2024).

Jones, Gareth, Peter Willett, Robert C Glen, Andrew R Leach, and Robin Taylor (1997). "Development and validation of a genetic algorithm for flexible docking". In: *Journal of Molecular Biology* 267.3, pp. 727–748.

Jubb, Harry C, Alicia P Higueruelo, Bernardo Ochoa-Montaño, Will R Pitt, David B Ascher, and Tom L Blundell (2017). "Arpeggio: A Web Server for Calculating and Visualising Interatomic Interactions in Protein Structures". In: *Journal of Molecular Biology* 429.3, pp. 365–371.

Kaya, Cansu, Isabell Walter, Samir Yahiaoui, Asfandyar Sikandar, Alaa Alhayek, Jelena Konstantinović, Andreas M. Kany, Jörg Haupenthal, Jesko Köhnke, Rolf W. Hartmann, and Anna K. H. Hirsch (2022). "Substrate-Inspired Fragment Merging and Growing Affords Efficacious LasB Inhibitors". In: *Angewandte Chemie International Edition* 61.5, e202112295.

Keeley, Aaron, László Petri, Péter Ábrányi-Balogh, and György M. Keserű (2020). "Covalent fragment libraries in drug discovery". In: *Drug Discovery Today* 25.6, pp. 983–996.

Keserű, György M., Daniel A. Erlanson, György G. Ferenczy, Michael M. Hann, Christopher W. Murray, and Stephen D. Pickett (2016). "Design principles for fragment libraries: maximizing the value of learnings from pharma fragment-based drug discovery (FBDD) programs for use in academia". In: *Journal of Medicinal Chemistry* 59.18, pp. 8189–8206.

Kim, Sunghwan, Jie Chen, Tiejun Cheng, Asta Gindulyte, Jia He, Siqian He, Qingliang Li, Benjamin A. Shoemaker, Paul A. Thiessen, Bo Yu, Leonid Zaslavsky, Jian Zhang, and Evan E. Bolton (2023). "PubChem 2023 update". In: *Nucleic Acids Research* 51.D1, pp. D1373–D1380.

Kirkman, Tim, Catharina dos Santos Silva, Manuela Tosin, and Marcio Vinicius Bertacine Dias (2024). "How to Find a Fragment: Methods for Screening and Validation in Fragment-Based Drug Dscovery". In: *ChemMedChem*, e202400342.

Kirsch, Philine, Alwin M Hartman, Anna K H Hirsch, and Martin Empting (2019). "Concepts and core principles of fragment-based drug design". In: *Molecules* 24.23.

Klarich, Kathryn, Brian Goldman, Trevor Kramer, Patrick Riley, and W. Patrick Walters (2024). "Thompson Sampling - An Efficient Method for Searching Ultralarge Synthesis on Demand Databases". In: *Journal of Chemical Information and Modeling* 64.4, pp. 1158–1171.

Kranz, James K. and Celine Schalk-Hihi (2011). "Chapter eleven - Protein Thermal Shifts to Identify Low Molecular Weight Fragments". In: *Methods in Enzymology.* Ed. by Lawrence C. Kuo. Vol. 493. Academic Press, pp. 277–298.

Krojer, T., R. Talon, N. Pearce, P. Collins, A. Douangamath, J. Brandao-Neto, A. Dias, B. Marsden, and F. von Delft (2017). "The XChemExplorer graphical workflow tool for routine or large-scale protein–ligand structure determination". In: *Acta Crystallographica Section D: Structural Biology* 73.3, pp. 267–278.

Lamoree, Bas and Roderick E. Hubbard (2017). "Current perspectives in fragment-based lead discovery (FBLD)". In: *Essays in Biochemistry* 61.5, pp. 453–464.

Landrum, Greg (2024). *RDKit: A software suite for cheminformatics, computational chemistry, and predictive modeling.* URL: `https://www.rdkit.org/RDKit_Overview.pdf` (visited on 08/05/2024).

Leach, Andrew R and Michael M Hann (2011). "Molecular complexity and fragment-based drug discovery: ten years on". In: *Current Opinion in Chemical Biology* 15.4, pp. 489–496.

Leelananda, Sumudu P. and Steffen Lindert (2016). "Computational methods in drug discovery". In: *Beilstein Journal of Organic Chemistry* 12.1, pp. 2694–2718.

Lei, Jian, Guido Hansen, Christoph Nitsche, Christian D. Klein, Linlin Zhang, and Rolf Hilgenfeld (2016). "Crystal structure of Zika virus NS2B–NS3 protease in complex with a boronate inhibitor". In: *Science* 353.6298, pp. 503–505.

Leung, Susan, Michael Bodkin, Frank von Delft, Paul Brennan, and Garrett Morris (2019). "SuCOS is better than RMSD for evaluating fragment elaboration and docking poses". In: *ChemRxiv* DOI: 10.26434/chemrxiv.8100203.v1.

Li, Qingxin (2020). "Application of fragment-based drug discovery to versatile targets". In: *Frontiers in Molecular Biosciences* 7, p. 180.

Li, Yan, Yuan Zhao, Zhihai Liu, and Renxiao Wang (2011). “Automatic Tailoring and Transplanting: a practical method that makes virtual screening more useful”. In: *Journal of Chemical Information and Modeling* 51.6, pp. 1474–1491.

Li, Yan, Zhixiong Zhao, Zhihai Liu, Minyi Su, and Renxiao Wang (2016). “AutoT&T v.2: an efficient and versatile tool for lead structure generation and optimization”. In: *Journal of Chemical Information and Modeling* 56.2, pp. 435–453.

Li, Yibo, Jianxing Hu, Yanxing Wang, Jielong Zhou, Liangren Zhang, and Zhenming Liu (2020). “DeepScaffold: a comprehensive tool for scaffold-based *de novo* drug discovery using deep learning”. In: *Journal of Chemical Information and Modeling* 60.1, pp. 77–91.

Licciardello, Marco P. and Paul Workman (2022). “The era of high-quality chemical probes”. In: *RSC Medicinal Chemistry* 13.12, pp. 1446–1459.

Lim, Jaechang, Sang-Yeon Hwang, Seokhyun Moon, Seungsu Kim, and Woo Youn Kim (2020). “Scaffold-based molecular design with a graph generative model”. In: *Chemical Science* 11.4, pp. 1153–1164.

Lindert, Steffen, Jacob D Durrant, and J Andrew McCammon (2012). “LigMerge: a fast algorithm to generate models of novel potential ligands from sets of known binders”. In: *Chemical Biology & Drug Design* 80.3, pp. 358–365.

Lipinski, Christopher A., Franco Lombardo, Beryl W. Dominy, and Paul J. Feeney (1997). “Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings”. In: *Advanced Drug Delivery Reviews* 23.1, pp. 3–25.

Lithgo, Ryan M., Charles W. E. Tomlinson, Michael Fairhead, Max Winokan, Warren Thompson, Conor Wild, Jasmin Cara Aschenbrenner, Blake H. Balcomb, Peter G. Marples, Anu V. Chandran, Mathew Golding, Lizbe Koekemoer, Eleanor P. Williams, SiYl Wang, Xiaomin Ni, Elizabeth MacLean, Charline Giroud, Andre Schutzer Godoy, Mary-Ann Xavier, Martin Walsh, Daren Fearon, and Frank von Delft (2024a). “Crystallographic Fragment Screen of Coxsackievirus A16 2A Protease identifies new opportunities for the development of broad-spectrum anti-enterovirals”. In: *bioRxiv* DOI: 10.1101/2024.04.29.591684.

Lithgo, Ryan M., Charles W. E. Tomlinson, Michael Fairhead, Max Winokan, Warren Thompson, Conor Wild, Jasmin Cara Aschenbrenner, Blake H. Balcomb, Peter G. Marples, Anu V. Chandran, Mathew Golding, Lizbe Koekemoer, Eleanor P. Williams, SiYl Wang, Xiaomin Ni, Elizabeth MacLean, Charline Giroud, Tryfon Zarganes-Tzitzikas, Andre Schutzer Godoy, Mary-Ann Xavier, Martin Walsh, Daren Fearon, and Frank von Delft (2024b). “Crystallographic fragment screen of Enterovirus D68 3C protease and iterative design of lead-like compounds using structure-guided expansions”. In: *bioRxiv* DOI: 10.1101/2024.04.29.591650.

Lovering, Frank, Jack Bikker, and Christine Humblet (2009). “Escape from Flatland: Increasing Saturation as an Approach to Improving Clinical Success”. In: *Journal of Medicinal Chemistry* 52.21, pp. 6752–6756.

Lyu, Jiankun, John J. Irwin, and Brian K. Shoichet (2023). “Modeling the expansion of virtual screening libraries”. In: *Nature Chemical Biology* 19.6, pp. 712–718.

Lyu, Jiankun, Sheng Wang, Trent E. Balius, Isha Singh, Anat Levit, Yurii S. Moroz, Matthew J. O’Meara, Tao Che, Enkhjargal Algaa, Kateryna Tolmachova, Andrey A. Tolmachev, Brian K. Shoichet, Bryan L. Roth, and John J. Irwin (2019). “Ultra-large library docking for discovering new chemotypes”. In: *Nature* 566.7743, pp. 224–229.

Maass, Patrick, Tanja Schulz-Gasch, Martin Stahl, and Matthias Rarey (2007). “Recore: a fast and versatile method for scaffold hopping based on small molecule crystal structure conformations”. In: *Journal of Chemical Information and Modeling* 47.2, pp. 390–399.

Maaten, Laurens van der and Geoffrey Hinton (2008). “Visualizing data using t-SNE”. In: *Journal of Machine Learning Research* 9.86, pp. 2579–2605.

Maggiora, Gerald, Martin Vogt, Dagmar Stumpfe, and Jürgen Bajorath (2014a). “Molecular Similarity in Medicinal Chemistry”. In: *Journal of Medicinal Chemistry* 57.8, pp. 3186–3204.

Maggiora, Gerald M. and Jürgen Bajorath (2014b). “Chemical space networks: a powerful new paradigm for the description of chemical space”. In: *Journal of Computer-Aided Molecular Design* 28.8, pp. 795–802.

Maia, Eduardo Habib Bechelane, Letícia Cristina Assis, Tiago Alves de Oliveira, Alisson Marques da Silva, and Alex Gutterres Taranto (2020). “Structure-Based Virtual Screening: From Classical to Artificial Intelligence”. In: *Frontiers in Chemistry* 8.

Makara, Gergely M., László Kovács, István Szabó, and Gábor Pőcze (2021). “Derivatization Design of Synthetically Accessible Space for Optimization: In Silico Synthesis vs Deep Generative Design”. In: *ACS Medicinal Chemistry Letters* 12.2, pp. 185–194.

Makurvet, Favour Danladi (2021). “Biologics vs. small molecules: Drug costs and patient access”. In: *Medicine in Drug Discovery* 9, p. 100075.

Malhotra, Shipra and John Karanicolas (2017). “When does chemical elaboration induce a ligand to change its binding mode?” In: *Journal of Medicinal Chemistry* 60.1, pp. 128–145.

Mayr, Lorenz M and Dejan Bojanic (2009). “Novel trends in high-throughput screening”. In: *Current Opinion in Pharmacology* 9.5, pp. 580–588.

McCorkindale, William, Ivan Ahel, Haim Barr, Galen J. Correy, James S. Fraser, Nir London, Marion Schuller, Khriesto Shurrush, and Alpha A. Lee (2022). “Fragment-based hit discovery via unsupervised learning of fragment–protein complexes”. In: *bioRxiv* DOI: 10.1101/2022.11.21.517375.

Meli, Rocco, Garrett M Morris, and Philip C Biggin (2022). “Scoring functions for protein-ligand binding affinity prediction using structure-based deep learning: A Review”. In: *Frontiers in Bioinformatics* 2.

Metz, A., J. Wollenhaupt, S. Glöckner, N. Messini, S. Huber, T. Barthel, A. Merabet, H.-D. Gerber, A. Heine, G. Klebe, and M. S. Weiss (2021). “Frag4Lead: growing crystallographic fragment hits by catalog using fragment-guided template docking”. In: *Acta Crystallographica Section D: Structural Biology* 77.9, pp. 1168–1182.

Miyake, Yuka, Yukihiro Itoh, Atsushi Hatanaka, Yoshinori Suzuma, Miki Suzuki, Hidehiko Kodama, Yoshinobu Arai, and Takayoshi Suzuki (2019). “Identification of novel lysine demethylase 5-selective inhibitors by inhibitor-based fragment merging strategy”. In: *Bioorganic & Medicinal Chemistry* 27.6, pp. 1119–1129.

MolPort (2024). *Compound Sourcing, Selling and Purchasing Platform.* URL: `https://www.molport.com/shop/index` (visited on 10/04/2024).

Müller, Janis, Raphael Klein, Olga Tarkhanova, Anastasiia Gryniukova, Petro Borysko, Stefan Merkl, Moritz Ruf, Alexander Neumann, Marcus Gastreich, Yurii S. Moroz, Gerhard Klebe, and Serghei Glinca (2022). "Magnet for the needle in haystack: "crystal structure first" fragment hits unlock active chemical matter using targeted exploration of vast chemical spaces". In: *Journal of Medicinal Chemistry* 65.23, pp. 15663–15678.

Mureddu, Luca G. and Geerten W. Vuister (2022). "Fragment-Based Drug Discovery by NMR. Where Are the Successes and Where can It Be Improved?" In: *Frontiers in Molecular Biosciences* 9.

Murray, Christopher W. and David C. Rees (2009). "The rise of fragment-based drug discovery". In: *Nature Chemistry* 1.3, pp. 187–192.

neo4j (2022). *Graph Modeling Guidelines. https://neo4j.com/developer/guide-data-modeling/.*

— (2024). *neo4j documentation.* URL: `https://neo4j.com/docs/` (visited on 09/10/2024).

Newman, Joseph A., Alice Douangamath, Setayesh Yadzani, Yuliana Yosaatmadja, Antony Aimon, José Brandão-Neto, Louise Dunnett, Tyler Gorrie-stone, Rachael Skyner, Daren Fearon, Matthieu Schapira, Frank von Delft, and Opher Gileadi (2021). "Structure, mechanism and crystallographic fragment screening of the SARS-CoV-2 NSP13 helicase". In: *Nature Communications* 12.1, p. 4848.

NextMove (2017). *Pistachio: Search and Faceting of Large Reaction Databases – NextMove Software.*

— (2024). *SmallWorld Search.* URL: `https://sw.docking.org/search.html` (visited on 08/09/2024).

Ng, J. T., C. Dekker, M. Kroemer, M. Osborne, and F. von Delft (2014). "Using textons to rank crystallization droplets by the likely presence of crystals". In: *Acta Crystallographica Section D: Biological Crystallography* 70.10, pp. 2702–2718.

Ni, Xiaomin, Andre Schutzer Godoy, Peter G. Marples, Michael Fairhead, Blake H. Balcomb, Matteo P. Ferla, Charles W. E. Tomlinson, Siyi Wang, Charline Giroud, Jasmin Cara Aschenbrenner, Ryan M. Lithgo, Max Winokan, Anu V. Chandran, Warren Thompson, Mary-Ann Xavier, Eleanor P. Williams, Martin Walsh, Daren Fearon, Lizbé Koekemoer, and Frank von Delft (2024). "Crystallographic fragment screening delivers diverse chemical scaffolds for Zika virus NS2B-NS3 protease inhibitor development". In: *bioRxiv* DOI: 10.1101/2024.04.29.591502.

Nikiforov, Petar O., Sachin Surade, Michal Blaszczyk, Vincent Delorme, Priscille Brodin, Alain R. Baulard, Tom L. Blundell, and Chris Abell (2016). "A fragment merging approach towards the development of small molecule inhibitors of *Mycobacterium tuberculosis* EthR for use as ethionamide boosters". In: *Organic & Biomolecular Chemistry* 14.7, pp. 2318–2326.

Nisius, Britta and Ulrich Rester (2009). "Fragment shuffling: an automated workflow for three-dimensional fragment-based ligand design". In: *Journal of Chemical Information and Modeling* 49.5, pp. 1211–1222.

O'Reilly, Marc, Anne Cleasby, Thomas G. Davies, Richard J. Hall, R. Frederick Ludlow, Christopher W. Murray, Dominic Tisi, and Harren Jhoti (2019). "Crystallographic screening using ultra-low-molecular-weight ligands to guide drug design". In: *Drug Discovery Today* 24.5, pp. 1081–1086.

Ohara-Nemoto, Yuko, Yu Shimoyama, Shigenobu Kimura, Asako Kon, Hiroshi Haraga, Toshio Ono, and Takayuki K. Nemoto (2011). "Asp- and Glu-specific novel dipeptidyl peptidase 11 of *Porphyromonas* gingivalis ensures utilization of proteinaceous energy sources". In: *Journal of Biological Chemistry* 286.44, pp. 38115–38127.

Oliveira, Tiago Alves de, Michel Pires da Silva, Eduardo Habib Bechelane Maia, Alisson Marques da Silva, and Alex Gutterres Taranto (2023). "Virtual Screening Algorithms in Drug Discovery: A Review Focused on Machine and Deep Learning Methods". In: *Drugs and Drug Candidates* 2.2, pp. 311–334.

Osborne, James, Stanislava Panova, Magdalini Rapti, Tatsuya Urushima, and Harren Jhoti (2020). "Fragments: where are we now?" In: *Biochemical Society Transactions* 48.1, pp. 271–280.

Overington, John P., Bissan Al-Lazikani, and Andrew L. Hopkins (2006). "How many drug targets are there?" In: *Nature Reviews Drug Discovery* 5.12, pp. 993–996.

Paananen, Jussi and Vittorio Fortino (2019). "An omics perspective on drug target discovery platforms". In: *Briefings in Bioinformatics* 21.6, pp. 1937–1953.

Paszke, Adam, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala (2019). "PyTorch: An Imperative Style, High-Performance Deep Learning Library". In: *arXiv* DOI: 10.48550/arXiv.1912.01703.

Patel, Disha, Joseph D. Bauman, and Eddy Arnold (2014). "Advantages of Crystallographic Fragment Screening: Functional and Mechanistic Insights from a Powerful Platform for Efficient Drug Discovery". In: *Progress in biophysics and molecular biology* 116.0, p. 92.

Pearce, Nicholas M., Anthony R. Bradley, Tobias Krojer, Brian D. Marsden, Charlotte M. Deane, and Frank von Delft (2017a). "Partial-occupancy binders identified by the Pan-Dataset Density Analysis method offer new chemical opportunities and reveal cryptic binding sites". In: *Structural Dynamics* 4.3, p. 032104.

Pearce, Nicholas M., Tobias Krojer, Anthony R. Bradley, Patrick Collins, Radosław P. Nowak, Romain Talon, Brian D. Marsden, Sebastian Kelm, Jiye Shi, Charlotte M. Deane, and Frank von Delft (2017b). "A multi-crystal method for extracting obscured crystallographic states from conventionally uninterpretable electron density". In: *Nature Communications* 8.1, p. 15123.

Pearce, Nicholas M., Rachael Skyner, and Tobias Krojer (2022). "Experiences From Developing Software for Large X-Ray Crystallography-Driven Protein-Ligand Studies". In: *Frontiers in Molecular Biosciences* 9.

Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay (2011). "Scikit-learn: machine learning in Python". In: *Journal of Machine Learning Research* 12.85, pp. 2825–2830.

Pierce, Albert C., Govinda Rao, and Guy W. Bemis (2004). "BREED: generating novel inhibitors through hybridization of known ligands. Application to CDK2, P38, and HIV protease". In: *Journal of Medicinal Chemistry* 47.11, pp. 2768–2775.

Plowright, Alleyn T., Craig Johnstone, Jan Kihlberg, Jonas Pettersson, Graeme Robb, and Richard A. Thompson (2012). "Hypothesis driven drug design: improving quality and effectiveness of the design-make-test-analyse cycle". In: *Drug Discovery Today* 17.1, pp. 56–62.

Polishchuk, Pavel (2020). "CReM: chemically reasonable mutations framework for structure generation". In: *Journal of Cheminformatics* 12.1, p. 28.

PostEra (2022). *Mpro activity data.* URL: `https://covid.postera.ai/covid/activity_data` (visited on 09/10/2023).

Powers, Alexander S., Helen H. Yu, Patricia Suriana, Rohan V. Koodli, Tianyu Lu, Joseph M. Paggi, and Ron O. Dror (2023). "Geometric Deep Learning for Structure-Based Ligand Design". In: *ACS Central Science* 9.12, pp. 2257–2267.

Prakash, Muthuraj, Yukihiro Itoh, Yoshie Fujiwara, Yukari Takahashi, Yuri Takada, Paolo Mellini, Elghareeb E. Elboray, Mitsuhiro Terao, Yasunobu Yamashita, Chika Yamamoto, Takao Yamaguchi, Masayuki Kotoku, Yuki Kitao, Ritesh Singh, Rohini Roy, Satoshi Obika, Makoto Oba, Dan Ohtan Wang, and Takayoshi Suzuki (2021). "Identification of Potent and Selective Inhibitors of Fat Mass Obesity-Associated Protein Using a Fragment-Merging Approach". In: *Journal of Medicinal Chemistry* 64.21, pp. 15810–15824.

Preston, Fergus (2023). "The identification and synthesis of novel fluorine containing heterocycles as starting-points of fragment-based drug discovery". PhD thesis. University of Dundee, Scotland, UK: University of Dundee.

Rarey, M., B. Kramer, T. Lengauer, and G. Klebe (1996). "A fast flexible docking method using an incremental construction algorithm". In: *Journal of Molecular Biology* 261.3, pp. 470–489.

Rasul, Azhar, Ammara Riaz, Iqra Sarfraz, Samreen Gul Khan, Ghulam Hussain, Rabia Zara, Ayesha Sadiqa, Gul Bushra, Saba Riaz, Muhammad Javid Iqbal, Mudassir Hassan, and Khatereh Khorsandi (2022). "Target Identification Approaches in Drug Discovery". In: *Drug Target Selection and Validation.* Ed. by Marcus T. Scotti and Carolina L. Bellera. Cham: Springer International Publishing, pp. 41–59.

Ren, Jing, Jian Li, Yueqin Wang, Wuyan Chen, Aijun Shen, Hongchun Liu, Danqi Chen, Danyan Cao, Yanlian Li, Naixia Zhang, Yechun Xu, Meiyu Geng, Jianhua He, Bing Xiong, and Jingkang Shen (2014). "Identification of a new series of potent diphenol HSP90 inhibitors by fragment merging and structure-based optimization". In: *Bioorganic Medicinal Chemistry Letters* 24.11, pp. 2525–2529.

Reymond, Jean-Louis (2015). "The Chemical Space Project". In: *Accounts of Chemical Research* 48.3, pp. 722–730.

Rogers, David and Mathew Hahn (2010). "Extended-Connectivity Fingerprints". In: *Journal of Chemical Information and Modeling* 50.5, pp. 742–754.

Ruddigkeit, Lars, Ruud van Deursen, Lorenz C. Blum, and Jean-Louis Reymond (2012). "Enumeration of 166 Billion Organic Small Molecules in the Chemical Universe Database GDB-17". In: *Journal of Chemical Information and Modeling* 52.11, pp. 2864–2875.

Ruiz-Carmona, Sergio, Daniel Alvarez-Garcia, Nicolas Foloppe, A. Beatriz Garmendia-Doval, Szilveszter Juhos, Peter Schmidtke, Xavier Barril, Roderick E. Hubbard, and S. David Morley (2014). "rDock: a fast, versatile and open source program for docking ligands to proteins and nucleic acids". In: *PLOS Computational Biology* 10.4, pp. 1–7.

Runcie, Nicholas T. and Antonia S.J.S. Mey (2023). "SILVR: guided diffusion for molecule generation". In: *Journal of Chemical Information and Modeling* 63.19, pp. 5996–6005.

Sadybekov, Arman A., Anastasiia V. Sadybekov, Yongfeng Liu, Christos Iliopoulos-Tsoutsouvas, Xi-Ping Huang, Julie Pickett, Blake Houser, Nilkanth Patel, Ngan K. Tran, Fei Tong, Nikolai Zvonok, Manish K. Jain, Olena Savych, Dmytro S. Radchenko, Spyros P. Nikas, Nicos A. Petasis, Yurii S. Moroz, Bryan L. Roth, Alexandros Makriyannis, and Vsevolod Katritch (2022). "Synthon-based ligand discovery in virtual libraries of over 11 billion compounds". In: *Nature* 601.7893, pp. 452–459.

Sanchez-Garcia, Ruben, Dávid Havasi, Gergely Takács, Matthew C. Robinson, Alpha Lee, Frank von Delft, and Charlotte M. Deane (2023). "CoPriNet: graph neural networks provide accurate and rapid compound price prediction for molecule prioritisation". In: *Digital Discovery* 2.1, pp. 103–111.

Santos, Rita, Oleg Ursu, Anna Gaulton, A. Patrícia Bento, Ramesh S. Donadi, Cristian G. Bologa, Anneli Karlsson, Bissan Al-Lazikani, Anne Hersey, Tudor I. Oprea, and John P. Overington (2017). "A comprehensive map of molecular drug targets". In: *Nature Reviews Drug Discovery* 16.1, pp. 19–34.

Satorras, Victor Garcia, Emiel Hoogeboom, and Max Welling (2022). “E(n) Equivariant Graph Neural Networks”. In: DOI: 10.48550/arXiv.2102.09844.

Scalfani, Vincent F., Vishank D. Patel, and Avery M. Fernandez (2022). “Visualizing chemical space networks with RDKit and NetworkX”. In: *Journal of Cheminformatics* 14.1, p. 87.

Scantlebury, Jack, Lucy Vost, Anna Carbery, Thomas E. Hadfield, Oliver M. Turnbull, Nathan Brown, Vijil Chenthamarakshan, Payel Das, Harold Grosjean, Frank von Delft, and Charlotte M. Deane (2023). “A Small Step Toward Generalizability: Training a Machine Learning Scoring Function for Structure-Based Virtual Screening”. In: *Journal of Chemical Information and Modeling* 63.10, pp. 2960–2974.

Schade, Markus, Beatrix Merla, Bernhard Lesch, Markus Wagener, Simone Timmermanns, Katrien Pletinckx, and Torsten Hertrampf (2020). “Highly selective sub-nanomolar cathepsin S inhibitors by merging fragment binders with nitrile inhibitors”. In: *Journal of Medicinal Chemistry* 63.20, pp. 11801–11808.

Schaller, David, Dora Šribar, Theresa Noonan, Lihua Deng, Trung Ngoc Nguyen, Szymon Pach, David Machalz, Marcel Bermudez, and Gerhard Wolber (2020). “Next generation 3D pharmacophore modeling”. In: *WIREs Computational Molecular Science* 10.4, e1468.

Schneider, Nadine, Nikolaus Stiefl, and Gregory A. Landrum (2016). “What’s What: The (Nearly) Definitive Guide to Reaction Role Assignment”. In: *Journal of Chemical Information and Modeling* 56.12, pp. 2336–2346.

Schreyer, Adrian and Tom Blundell (2009). “CREDO: A Protein–Ligand Interaction Database for Drug Discovery”. In: *Chemical Biology & Drug Design* 73.2, pp. 157–167.

Schreyer, Adrian M. and Tom Blundell (2012). “USRCAT: real-time ultrafast shape recognition with pharmacophoric constraints”. In: *Journal of Cheminformatics* 4.1, p. 27.

Schuffenhauer, Ansgar and Nathan Brown (2006). “Chemical diversity and biological activity”. In: *Drug Discovery Today: Technologies* 3.4, pp. 387–395.

Schuller, Marion, Galen J. Correy, Stefan Gahbauer, Daren Fearon, Taiasean Wu, Roberto Efraín Díaz, Iris D. Young, Luan Carvalho Martins, Dominique H. Smith, Ursula Schulze-Gahmen, Tristan W. Owens, Ishan Deshpande, Gregory E. Merz, Aye C. Thwin, Justin T. Biel, Jessica

K. Peters, Michelle Moritz, Nadia Herrera, Huong T. Kratochvil, QCRG Structural Biology Consortium, Anthony Aimon, James M. Bennett, Jose Brandao Neto, Aina E. Cohen, Alexandre Dias, Alice Douangamath, Louise Dunnett, Oleg Fedorov, Matteo P. Ferla, Martin R. Fuchs, Tyler J. Gorrie-Stone, James M. Holton, Michael G. Johnson, Tobias Krojer, George Meigs, Ailsa J. Powell, Johannes Gregor Matthias Rack, Victor L. Rangel, Silvia Russi, Rachael E. Skyner, Clyde A. Smith, Alexei S. Soares, Jennifer L. Wierman, Kang Zhu, Peter O'Brien, Natalia Jura, Alan Ashworth, John J. Irwin, Michael C. Thompson, Jason E. Gestwicki, Frank von Delft, Brian K. Shoichet, James S. Fraser, and Ivan Ahel (2021). "Fragment binding to the Nsp3 macrodomain of SARS-CoV-2 identified through crystallographic screening and computational docking". In: *Science Advances* 7.16, eabf8711.

Scott, Duncan E., Anthony G. Coyne, Sean A. Hudson, and Chris Abell (2012). "Fragment-Based Approaches in Drug Discovery and Chemical Biology". In: *Biochemistry* 51.25, pp. 4990–5003.

Senger, Stefan (2009). "Using Tversky Similarity Searches for Core Hopping: Finding the Needles in the Haystack". In: *Journal of Chemical Information and Modeling* 49.6, pp. 1514–1524.

Sertkaya, Aylin, Trinidad Beleche, Amber Jessup, and Benjamin D. Sommers (2024). "Costs of Drug Development and Research and Development Intensity in the US, 2000-2018". In: *JAMA Network Open* 7.6, e2415445–e2415445.

Shegokar, Ranjita (2020). "Chapter 2 - Preclinical testing—Understanding the basics first". In: *Drug Delivery Aspects*. Ed. by Ranjita Shegokar. Elsevier, pp. 19–32.

Siegal, Gregg, Eiso Ab, and Jan Schultz (2007). "Integration of fragment screening and library design". In: *Drug Discovery Today* 12.23, pp. 1032–1039.

Souza Neto, Lauro Ribeiro de, José Teófilo Moreira-Filho, Bruno Junior Neves, Rocío Lucía Beatriz Riveros Maidana, Ana Carolina Ramos Guimarães, Nicholas Furnham, Carolina Horta Andrade, and Floriano Paes Silva (2020). "*In* silico strategies to support fragment-to-lead optimization in drug discovery". In: *Frontiers in Chemistry* 8, p. 93.

Spiegel, Jacob O. and Jacob D. Durrant (2020). "AutoGrow4: an open-source genetic algorithm for de novo drug design and lead optimization". In: *Journal of Cheminformatics* 12.1, p. 25.

St. Denis, Jeffrey D., Richard J. Hall, Christopher W. Murray, Tom D. Heightman, and David C. Rees (2021). “Fragment-based drug discovery: opportunities for organic synthesis”. In: *RSC Medicinal Chemistry* 12.3, pp. 321–329.

Stumpfe, Dagmar and Jürgen Bajorath (2011). “Similarity searching”. In: *WIREs Computational Molecular Science* 1.2, pp. 260–282.

Su, Minyi, Qifan Yang, Yu Du, Guoqin Feng, Zhihai Liu, Yan Li, and Renxiao Wang (2019). “Comparative assessment of scoring functions: the CASF-2016 update”. In: *Journal of Chemical Information and Modeling* 59.2, pp. 895–913.

Sunseri, Jocelyn and David Ryan Koes (2016). “Pharmit: interactive exploration of chemical space”. In: *Nucleic Acids Research* 44.Web Server issue, W442–W448.

Suvarna, Viraj (2010). “Phase IV of Drug Development”. In: *Perspectives in Clinical Research* 1.2, pp. 57–60.

Swain, Chris (2006). *Acid Bioisosteres.* URL: `https://www.cambridgemedchemconsulting.com/DDResources/Bioisoteres/Acids_bioisosteres.html` (visited on 07/22/2024).

Tabana, Yasser, Dinesh Babu, Richard Fahlman, Arno G. Siraki, and Khaled Barakat (2023). “Target identification of small molecules: an overview of the current applications in drug discovery”. In: *BMC Biotechnology* 23.1, p. 44.

Tan, Youhai, Lingxue Dai, Weifeng Huang, Yinfeng Guo, Shuangjia Zheng, Jinping Lei, Hongming Chen, and Yuedong Yang (2022). “DRlinker: Deep Reinforcement Learning for Optimization in Fragment Linking Design”. In: *Journal of Chemical Information and Modeling* 62.23, pp. 5907–5917.

Thakkar, Amol, Veronika Chadimová, Esben Jannik Bjerrum, Ola Engkvist, and Jean-Louis Reymond (2021). “Retrosynthetic accessibility score (RAscore) — rapid machine learned synthesizability classification from AI driven retrosynthetic planning”. In: *Chemical Science* 12.9, pp. 3339–3349.

The COVID Moonshot Consortium et al. (2020). “COVID Moonshot: open science discovery of SARS-CoV-2 main protease inhibitors by combining crowdsourcing, high-throughput experiments, computational simulations, and machine learning”. In: *bioRxiv* DOI: 10.1101/2020.10.29.339317.

Thomas, Morgan, Mazen Ahmad, Gary Tresadern, and Gianni de Fabritiis (2024). “PromptSMILES: prompting for scaffold decoration and fragment linking in chemical language models”. In: *Journal of Cheminformatics* 16.1, p. 77.

Tingle, Benjamin I., Khanh G. Tang, Mar Castanon, John J. Gutierrez, Munkhzul Khurelbaatar, Chinzorig Dandarchuluun, Yurii S. Moroz, and John J. Irwin (2023). “ZINC-22A Free Multi-Billion-Scale Database of Tangible Compounds for Ligand Discovery”. In: *Journal of Chemical Information and Modeling* 63.4, pp. 1166–1176.

Trott, Oleg and Arthur J. Olson (2010). “AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization and multithreading”. In: *Journal of Computational Chemistry* 31.2, pp. 455–461.

Vázquez, Javier, Manel López, Enric Gibert, Enric Herrero, and F. Javier Luque (2020). “Merging Ligand-Based and Structure-Based Methods in Drug Discovery: An Overview of Combined Virtual Screening Approaches”. In: *Molecules* 25.20, p. 4723.

Veber, Daniel F., Stephen R. Johnson, Hung-Yuan Cheng, Brian R. Smith, Keith W. Ward, and Kenneth D. Kopple (2002). “Molecular properties that influence the oral bioavailability of drug candidates”. In: *Journal of Medicinal Chemistry* 45.12, pp. 2615–2623.

Velasquez-López, Yendrek, Eduardo Tejera, and Yunierkis Perez-Castillo (2022). “Chapter One - Can docking scoring functions guarantee success in virtual screening?” In: *Annual Reports in Medicinal Chemistry.* Ed. by Julio Caballero. Vol. 59. Academic Press, pp. 1–41.

Virtanen, Pauli, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C. J. Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, and Paul van Mulbregt (2020). “SciPy 1.0: fundamental algorithms for scientific computing in Python”. In: *Nature Methods* 17.3, pp. 261–272.

Vonrhein, Clemens, Claus Flensburg, Peter Keller, Andrew Sharff, Oliver Smart, Wlodek Paciorek, Thomas Womack, and Gérard Bricogne (2011). "Data processing and analysis with the autoPROC toolbox". In: *Acta Crystallographica. Section D, Biological Crystallography* 67.Pt 4, pp. 293–302.

Vonrhein, Clemens, Ian J. Tickle, Claus Flensburg, Peter Keller, Wlodek Paciorek, Andrew Sharff, and Gerard Bricogne (2018). "Advances in automated data analysis and processing within *autoPROC* , combined with improved characterisation, mitigation and visualisation of the anisotropy of diffraction limits using *STARANISO*". In: *Acta Crystallographica Section A Foundations and Advances* 74.a1, a360–a360.

Voršilák, Milan, Michal Kolář, Ivan Čmelo, and Daniel Svozil (2020). "SYBA: Bayesian estimation of synthetic accessibility of organic compounds". In: *Journal of Cheminformatics* 12.1, p. 35.

Voršilák, Milan and Daniel Svozil (2017). "Nonpher: computational method for design of hard-to-synthesize structures". In: *Journal of Cheminformatics* 9.1, p. 20.

Wagner, Bridget K. (2016). "The resurgence of phenotypic screening in drug discovery and development". In: *Expert Opinion on Drug Discovery* 11.2, pp. 121–125.

Wahlberg, Elisabet, Tobias Karlberg, Ekaterina Kouznetsova, Natalia Markova, Antonio Macchiarulo, Ann-Gerd Thorsell, Ewa Pol, Åsa Frostell, Torun Ekblad, Delal Öncü, Björn Kull, Graeme Michael Robertson, Roberto Pellicciari, Herwig Schüler, and Johan Weigelt (2012). "Family-wide chemical profiling and structural analysis of PARP and tankyrase inhibitors". In: *Nature Biotechnology* 30.3, pp. 283–288.

Wang, Hao, Xiaolin Pan, Yueqing Zhang, Xingyu Wang, Xudong Xiao, and Changge Ji (2022a). "MolHyb: a web server for structure-based drug design by molecular hybridization". In: *Journal of Chemical Information and Modeling* 62.12, pp. 2916–2922.

Wang, Jun, Yanmei Hu, and Madeleine Zheng (2022b). "Enterovirus A71 antivirals: past, present, and future". In: *Acta Pharmaceutica Sinica B* 12.4, pp. 1542–1566.

Wang, Renxiao, Xueliang Fang, Yipin Lu, Chao-Yie Yang, and Shaomeng Wang (2005). "The PDBbind database: methodologies and updates". In: *Journal of Medicinal Chemistry* 48.12, pp. 4111–4119.

Warr, Wendy A., Marc C. Nicklaus, Christos A. Nicolaou, and Matthias Rarey (2022). "Exploration of Ultralarge Compound Collections for Drug Discovery". In: *Journal of Chemical Information and Modeling* 62.9, pp. 2021–2034.

Warren, Matthew, Ísak Valsson, Charlotte Deane, Aniket Magarkar, Garrett Morris, and Philip Biggin (2024). "How to make machine learning scoring functions competitive with FEP". In: *ChemRxiv* DOI: 10.26434/chemrxiv-2024-bth5z.

Weininger, David (1988). "SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules". In: *Journal of Chemical Information and Computer Sciences* 28.1, pp. 31–36.

Wermuth, C. G., C. R. Ganellin, P. Lindberg, and L. A. Mitscher (1998). "Glossary of terms used in medicinal chemistry (IUPAC Recommendations 1998)". In: *Pure and Applied Chemistry* 70.5, pp. 1129–1143.

Wesolowski, Steven S. and Dean G. Brown (2016). "The Strategies and Politics of Successful Design, Make, Test, and Analyze (DMTA) Cycles in Lead Generation". In: *Lead Generation.* John Wiley & Sons, Ltd, pp. 487–512.

Wigglesworth, Mark J, David C Murray, Carolyn J Blackett, Michael Kossenjans, and J Willem M Nissink (2015). "Increasing the delivery of next generation therapeutics from high throughput screening libraries". In: *Current Opinion in Chemical Biology* 26, pp. 104–110.

Wills, Stephanie, Ruben Sanchez-Garcia, Tim Dudgeon, Stephen D. Roughley, Andy Merritt, Roderick E. Hubbard, James Davidson, Frank von Delft, and Charlotte M. Deane (2023). "Fragment merging using a graph database samples different catalogue space than similarity search". In: *Journal of Chemical Information and Modeling* 63.11, pp. 3423–3437.

Wills, Stephanie, Ruben Sanchez-Garcia, Stephen D. Roughley, Andy Merritt, Roderick E. Hubbard, Frank von Delft, and Charlotte M. Deane (2024). "Expanding the scope of a catalogue search to bioisosteric fragment merges using a graph database approach". In: *bioRxiv* DOI: 10.1101/2024.08.02.606367.

Winter, Graeme, Carina M. C. Lobley, and Stephen M. Prince (2013). "Decision making in xia2". In: *Acta Crystallographica. Section D, Biological Crystallography* 69.Pt 7, pp. 1260–1273.

Winter, Graeme, David G. Waterman, James M. Parkhurst, Aaron S. Brewster, Richard J. Gildea, Markus Gerstel, Luis Fuentes-Montero, Melanie Vollmar, Tara Michels-Clark, Iris D. Young, Nicholas K. Sauter, and Gwyndaf Evans (2018). "DIALS: implementation and evaluation of a new integration package". In: *Acta Crystallographica. Section D, Structural Biology* 74.Pt 2, pp. 85–97.

Wolber, Gerhard and Thierry Langer (2005). "LigandScout: 3-D Pharmacophores Derived from Protein-Bound Ligands and Their Use as Virtual Screening Filters". In: *Journal of Chemical Information and Modeling* 45.1, pp. 160–169.

Wood, Daniel J., J. Daniel Lopez-Fernandez, Leanne E. Knight, Islam Al-Khawaldeh, Conghao Gai, Shengying Lin, Mathew P. Martin, Duncan C. Miller, Céline Cano, Jane A. Endicott, Ian R. Hardcastle, Martin E. M. Noble, and Michael J. Waring (2019). "FragLites—Minimal, Halogenated Fragments Displaying Pharmacophore Doublets. An Efficient Approach to Druggability Assessment and Hit Generation". In: *Journal of Medicinal Chemistry* 62.7, pp. 3741–3752.

Woodhead, Andrew J., Daniel A. Erlanson, Iwan J. P. de Esch, Rhian S. Holvey, Wolfgang Jahnke, and Puja Pathuri (2024). "Fragment-to-Lead Medicinal Chemistry Publications in 2022". In: *Journal of Medicinal Chemistry* 67.4, pp. 2287–2304.

Workman, Paul and Ian Collins (2010). "Probing the Probes: Fitness Factors For Small Molecule Tools". In: *Chemistry & Biology* 17.6, pp. 561–577.

Wright, N. D., P. Collins, L. Koekemoer, T. Krojer, R. Talon, E. Nelson, M. Ye, R. Nowak, J. Newman, J. T. Ng, N. Mitrovich, H. Wiggers, and F. von Delft (2021). "The low-cost Shifter microscope stage transforms the speed and robustness of protein crystal harvesting". In: *Acta Crystallographica Section D: Structural Biology* 77.1, pp. 62–74.

Wunberg, Tobias, Martin Hendrix, Alexander Hillisch, Mario Lobell, Heinrich Meier, Carsten Schmeck, Hanno Wild, and Berthold Hinzen (2006). "Improving the hit-to-lead process: data-driven assessment of drug-like and lead-like screening hits". In: *Drug Discovery Today* 11.3, pp. 175–180.

XChem (2024). *The XChem Pipeline.* URL: `https://www.diamond.ac.uk/Instruments/Mx/Fragment-Screening/The-XChem-Pipeline.html` (visited on 08/12/2024).

Xie, Junjie, Sheng Chen, Jinping Lei, and Yuedong Yang (2024). “DiffDec: Structure-Aware Scaffold Decoration with an End-to-End Diffusion Model”. In: *Journal of Chemical Information and Modeling* 64.7, pp. 2554–2564.

Yang, Sheng-Yong (2010). “Pharmacophore modeling and applications in drug discovery: challenges and recent advances”. In: *Drug Discovery Today* 15.11, pp. 444–450.

Yang, Yuyao, Shuangjia Zheng, Shimin Su, Chao Zhao, Jun Xu, and Hongming Chen (2020). “SyntaLinker: automatic fragment linking with deep conditional transformer neural networks”. In: *Chemical Science* 11.31, pp. 8312–8322.

Yuan, Jiacheng, Herbert Pang, Tiejun Tong, Dong Xi, Wenzhao Guo, and Peter Mesenbrink (2016). “Seamless Phase IIa/IIb and Enhanced Dose Finding Adaptive Design”. In: *Journal of biopharmaceutical statistics* 26.5, pp. 912–923.

Zhang, Ming-Qiang and Barrie Wilkinson (2007). “Drug discovery beyond the ‘rule-of-five’”. In: *Current Opinion in Biotechnology* 18.6, pp. 478–488.

Zhang, Shuxing (2011). “Computer-Aided Drug Discovery and Development”. In: *Drug Design and Discovery: Methods and Protocols*. Ed. by Seetharama D. Satyanarayanajois. Totowa, NJ: Humana Press, pp. 23–38.

Zhu, Tian, Shuyi Cao, Pin-Chih Su, Ram Patel, Darshan Shah, Heta B. Chokshi, Richard Szukala, Michael E. Johnson, and Kirk E. Hevener (2013). “Hit identification and optimization in virtual screening: practical recommendations based upon a critical literature analysis”. In: *Journal of Medicinal Chemistry* 56.17, pp. 6560–6572.

A. Appendix items for Chapter 2

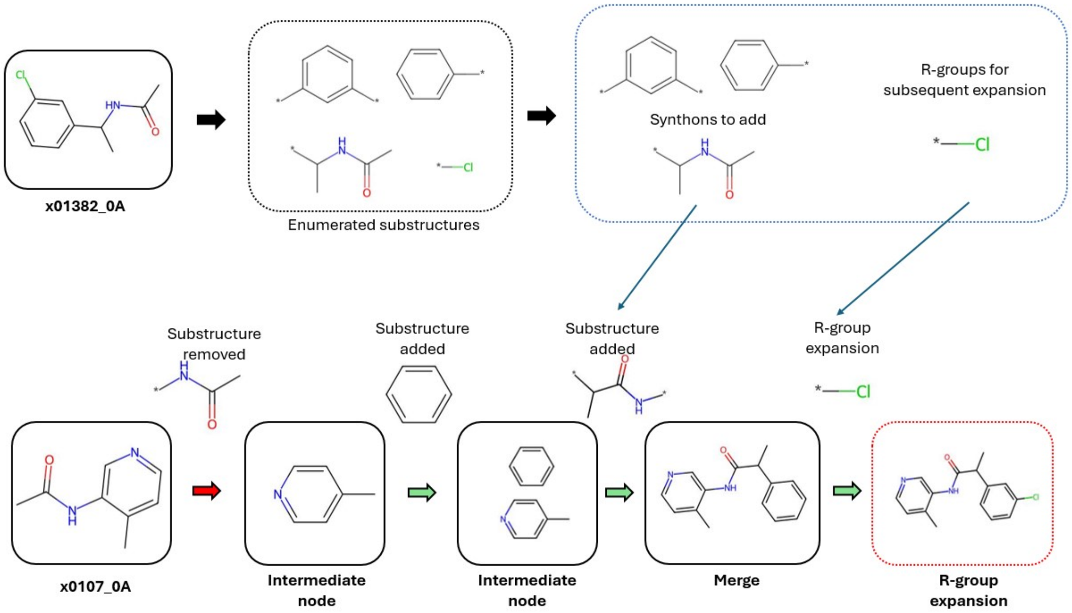


Figure A.1: **Example of the R-group expansion process.** The querying process for fragments x0107_0A and x1382_0A against the SARS-CoV2 main protease (Mpro) is shown, to demonstrate how merges retrieved from the database can subsequently undergo R-group expansion to add substituents. Fragment x1382_0A is broken down into substructures and only those with at least 3 carbon atoms are used in the database query (to ensure both fragments make a substantial contribution to the final merge). We provide an option to run a subsequent R-group expansion using the remaining substructures (here a chlorine atom) based on those merges. The bottom row shows the query path taken within the database.

Table A.1: Fragments used for analysis

Target	Fragments
DPP11	x0267_0A, x0230_1A, x0208_0A, x0228_0A, x0199_0A, x0346_0A, x0051_0A, x0115_0A x0083_0A, x0056_0A, x0087_0A
PARP14	x0161_1, x0238_1, x0266_1, x0315_1, x0324_1, x0334_1, x0412_1, x0457_1, x0473_1 x0505_1, x0590_1, x0637_1, x0712_1
nsp13	x0034_0B, x0176_0B, x0183_0B, x0208_0A, x0212_0B, x0246_0B, x0276_0B, x0311_0B x0438_0B
Mpro	x0354_0A, x0426_0A, x0104_0A, x0195_0A, x0072_0A, x0305_0A, x0161_0A x1077_0A, x0874_0A, x1249_0A, x0991_0A, x2193_0A, x1093_0A, x0946_0A, x0967_0A x0395_0A, x0387_0A, x0540_0A, x0397_0A
Mpro (case study)	x0434_0A, x0107_0A, x0678_0A, x0995_0A, x1382_0A

Crystal complexes are available to download from the Fragalysis platform: `https://fragalysis.diamond.ac.uk/viewer/react/landing`. DPP11, dipeptidyl peptidase 11; PARP14, poly(ADP-ribose) 14; Mpro, main protease; nsp13, non-structural protein 13.

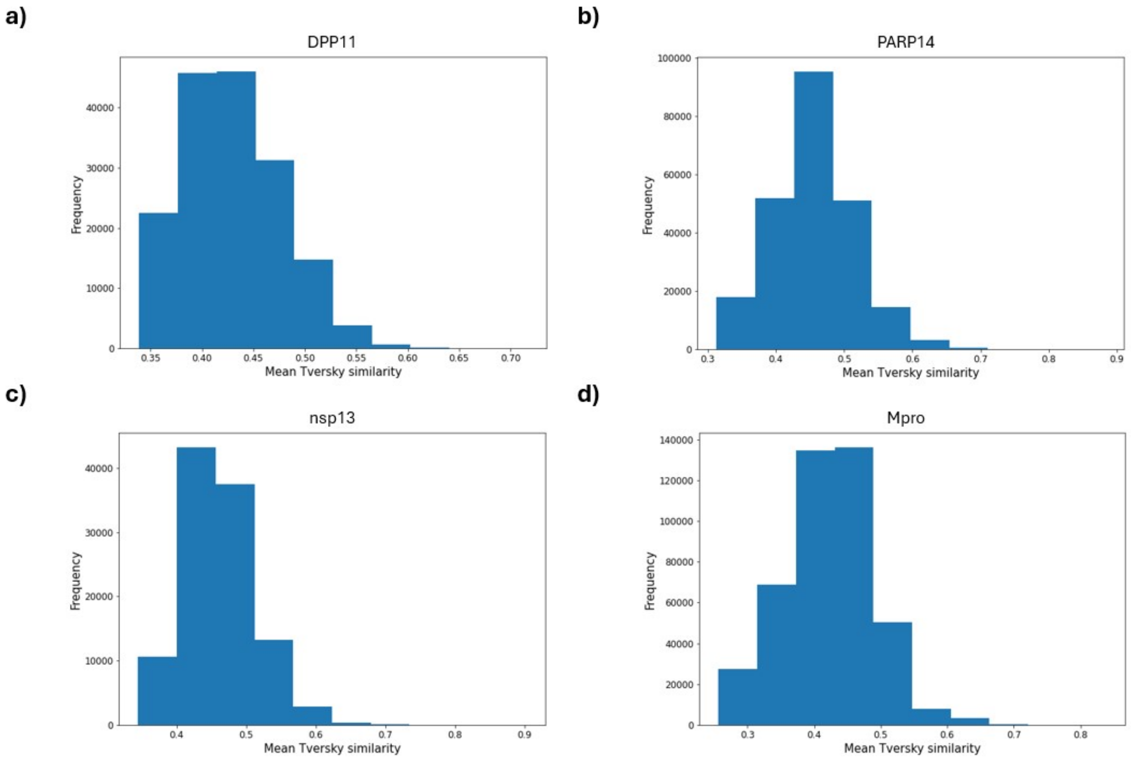


Figure A.2: **The mean Tversky similarities between unfiltered merges and their parent fragments found using similarity search.** The mean Tversky similarity between similarity search-identified merges (before removing those with Tversky <0.4 and before entering the filtering pipeline) and their parent fragments are shown for targets (**a**) dipeptidyl peptidase 11 (DPP11), (**b**) poly(ADP-ribose) polymerase 14 (PARP14), (**c**) non-structural protein 13 (nsp13) and (**d**) main protease (Mpro).

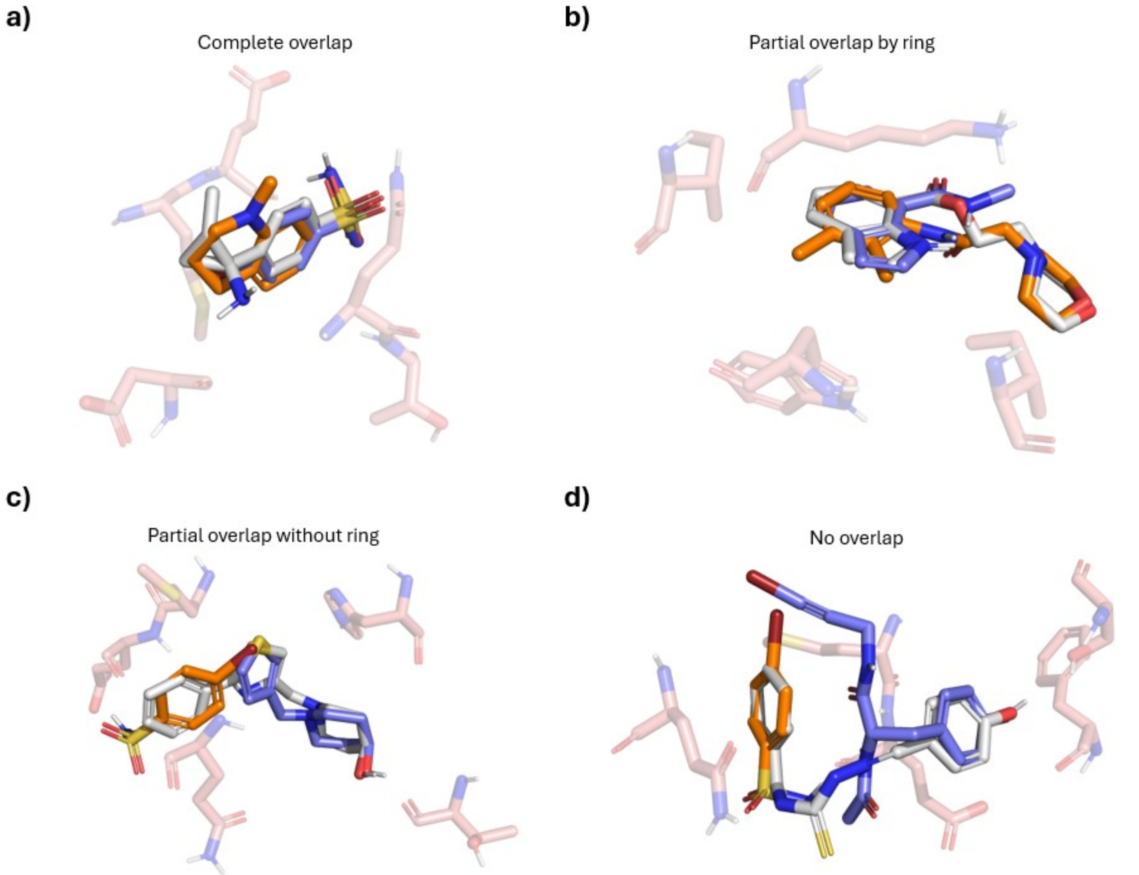


Figure A.3: **Types of merging opportunity.** We differentiate between four different types of merging opportunity depending on the overlap between the parent fragments (shown in purple and orange): (**a**) *complete overlap merges*, whereby the majority of the volume of one fragment overlaps with the other; (**b**) *partial overlap by ring*, which are fragments that share an overlapping ring structure and represent the classical merges seen in the literature; (**c**) *partial overlap without ring*, whereby the overlapping fragments do not share an overlapping ring structure; and (**d**) *non overlap of parent fragments*, which represent linking opportunities.

Table A.2: Substructures used in the expansion of Fragment Network-filtered compounds.

Target	Substructure SMILES	Frequency
DPP11	[*]c1ccccc1	83
	[*]N1CCOCC1	55
	[*]c1ccco1	26
	[*]C1CC1	22
	[*]N1CCCC1	12
	[*]c1ccncc1	3
	[*]c1ccoc1	2
	[*]N1CCc2ccccc2C1	1
PARP14A	[*]c1ccccc1	26
	CCC[*]	18
	[*]C1CCCCC1	9
	[*]n1cccn1	6
	[*]c1ccco1	5
	[*]c1ncccn1	3
	[*]c1ccc2ccccc2n1	2
	[*]c1nc2ccccc2[nH]1	1
	[*]c1cn2ccccc2n1	1
nsp13	[*]c1ccccc1	495
	[*]C1CNC1	11
	[*]N1CCC1	4
Mpro	[*]c1ccccc1	531
	[*]c1ccccn1	146
	[*]c1cccnc1	95
	[*]N1CCOCC1	85
	[*]N1CCCCC1	67
	[*]c1ccncc1	25
	[*]C1CC1	21
	[*]c1ccsc1	19
	[*]N1CCCc2ccccc21	15
	[*]C1CCNCC1	5
	[*]C1CCCCC1	4
	[*]N1CCCOCC1	3
	[*]c1c[nH]c2ncccc12	1
	[*]N1CCNCC1	1

[*] denotes the attachment point in the SMILES. DPP11, dipeptidyl peptidase 11; PARP14, poly(ADP-ribose) 14; Mpro, main protease; nsp13, non-structural protein 13.

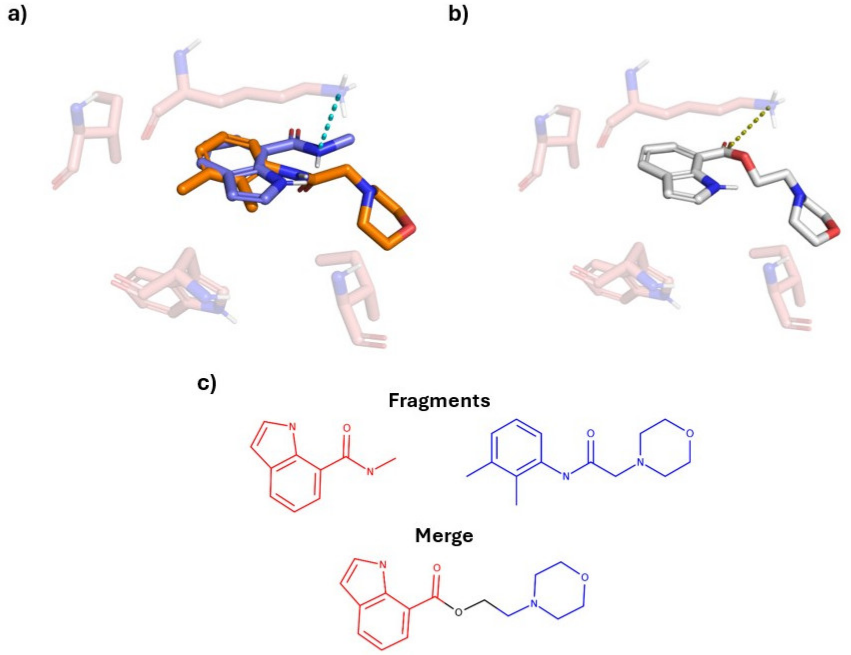


Figure A.4: **An example classical merge.** (**a**) The crystal structures of two parent fragments that represent a ‘classical merge’ for dipeptidyl peptidase 11 (DPP11), whereby the two fragments show an overlapping ring and the connectivity of the final compound is obvious. (**b**) The orientation of the merge (white) generated using Fragmenstein. Interactions are predicted using the protein–ligand interaction profiler (PLIP) and key interaction residues are shown. A salt bridge is depicted using a yellow dotted line. (**c**) The fragments and merge in 2D; colours indicate the substructures used in forming the merge.

Table A.3: Cluster composition after Butina clustering of filtered compounds

Target	Fragment Network-only	Similarity search-only	Both
DPP11	154	139	9
PARP14	63	48	0
nsp13	484	367	3
Mpro	791	646	10

DPP11, dipeptidyl peptidase 11; PARP14, poly(ADP-ribose) 14; Mpro, main protease; nsp13, non-structural protein 13.

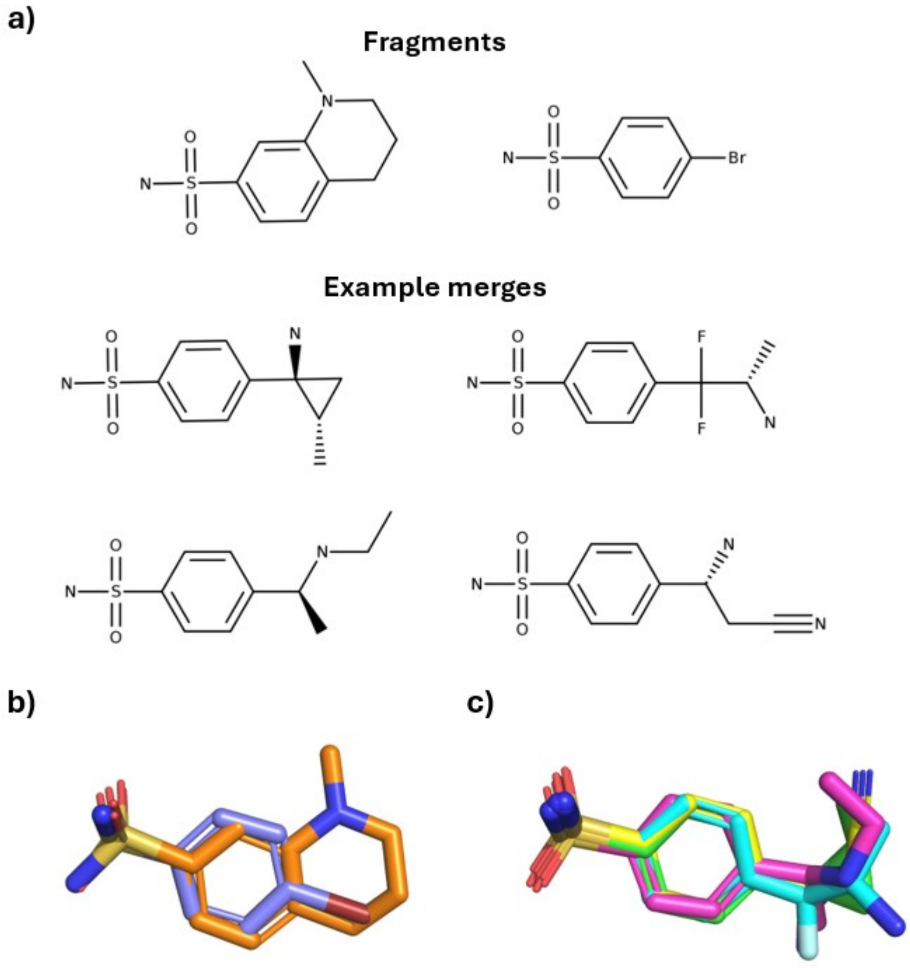


Figure A.5: **Non-useful merges identified using similarity search.** (**a**) Example compounds identified using similarity search for two fragments (x0195-0A and x0946-0A) against the main protease (Mpro). The compounds do not represent useful merges as they do not incorporate unique substructures from both fragments and thus are derivatives of one of the fragments. Crystal and Fragmenstein-predicted poses are shown of the fragments (**b**) and the merges (**c**), respectively.

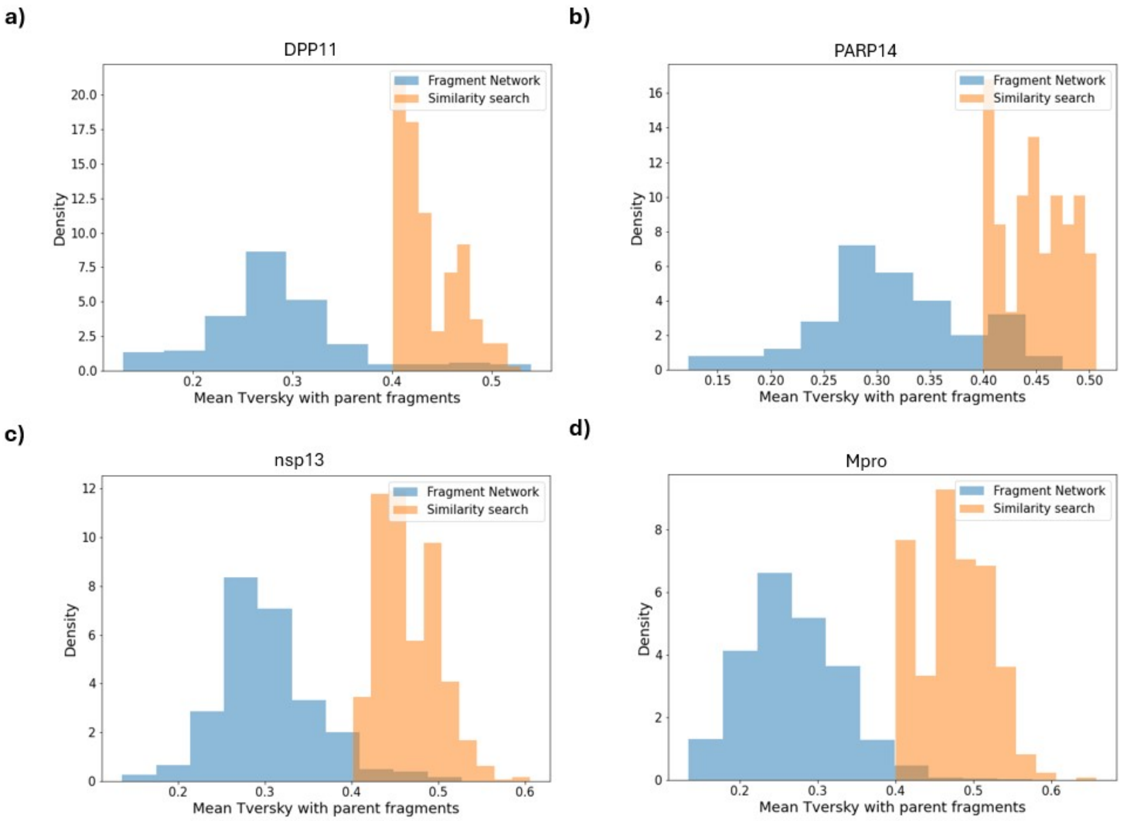


Figure A.6: **The mean Tversky similarities between filtered merges and their parent fragments found using similarity search.** The mean Tversky similarity between similarity search-identified merges (after filtering) and their parent fragments are shown for targets (**a**) dipeptidyl peptidase 11 (DPP11), (**b**) poly(ADP-ribose) polymerase 14, (PARP14) (**c**) non-structural protein 13 (nsp13) and (**d**) main protease (Mpro).

Table A.4: The number of 'true merges' found where both fragments contribute a unique interaction type with a specific residue to the final merge.

Target	Fragment Network			Similarity search		
	True merges*	Possible true merges**	Efficiency (%)	True merges*	Possible true merges**	Efficiency (%)
DPP11	32	203	15.8	88	272	32.4
PARP14	45	71	63.4	26	56	46.4
nsp13	82	509	16.1	142	616	23.1
Mpro	513	1,017	50.4	128	832	15.4

*This refers to the number of true merges that are found by the pipeline (meaning a unique interaction type from each fragment is present in the merge). **The number of possible true merges refers to the number of theoretically possible true merges there should be (meaning the parent fragments of these merges have non-intersecting interactions). DPP11, dipeptidyl peptidase 11; PARP14, poly(ADP-ribose) 14; Mpro, main protease; nsp13, non-structural protein 13.

Table A.5: The percentage of compounds removed by each filtering step.

		Fragment Network compounds		Similarity search compounds	
Target	Filter	% compounds removed*	% total compounds entering filter**	% compounds removed	% total compounds entering filter
DPP11	Descriptor	0.0	0.0	0.0	0.0
	Non-ring bond	17.6	17.6	3.1	3.1
	Expansion	1.0	1.2	–	–
	Embedding	60.8	74.7	79.6	82.1
	Overlap	3.6	17.7	2.0	11.7
	Fragmenstein	16.4	96.8	15.0	98.2
	Energy of pose	0.1	9.7	0.0	7.2
PARP14	Descriptor	0.1	0.1	0.1	0.1
	Non-ring bond	28.0	28.1	19.2	19.2
	Expansion	0.1	0.2	–	–
	Embedding	51.5	71.8	75.0	92.9
	Overlap	8.9	43.9	1.8	32.2
	Fragmenstein	11.2	98.9	3.8	98.7
	Energy of pose	0.1	50.7	0.0	24.3
nsp13	Descriptor	3.9	3.9	0.2	0.2
	Non-ring bond	51.0	53.0	24.0	24.1
	Expansion	0.0	0.0	–	–
	Embedding	21.3	47.2	62.0	81.8
	Overlap	7.6	31.8	3.5	25.1
	Fragmenstein	15.2	93.9	9.4	90.8
	Energy of pose	0.1	9.3	0.1	7.1
Mpro	Descriptor	0.8	0.8	0.0	0.0
	Non-ring bond	27.6	27.9	11.9	11.9
	Expansion	0.3	0.5	–	–
	Embedding	52.1	73.1	74.5	84.6
	Overlap	6.2	32.7	4.9	36.3
	Fragmenstein	12.1	93.8	8.1	93.3
	Energy of pose	0.2	28.7	0.2	33.9

*This refers to the percentage of all compounds retrieved from the database that are removed by each filter. **This refers to the percentage of compounds entering the filter that is then removed by that filter. DPP11, dipeptidyl peptidase 11; PARP14, poly(ADP-ribose) 14; Mpro, main protease; nsp13, non-structural protein 13.

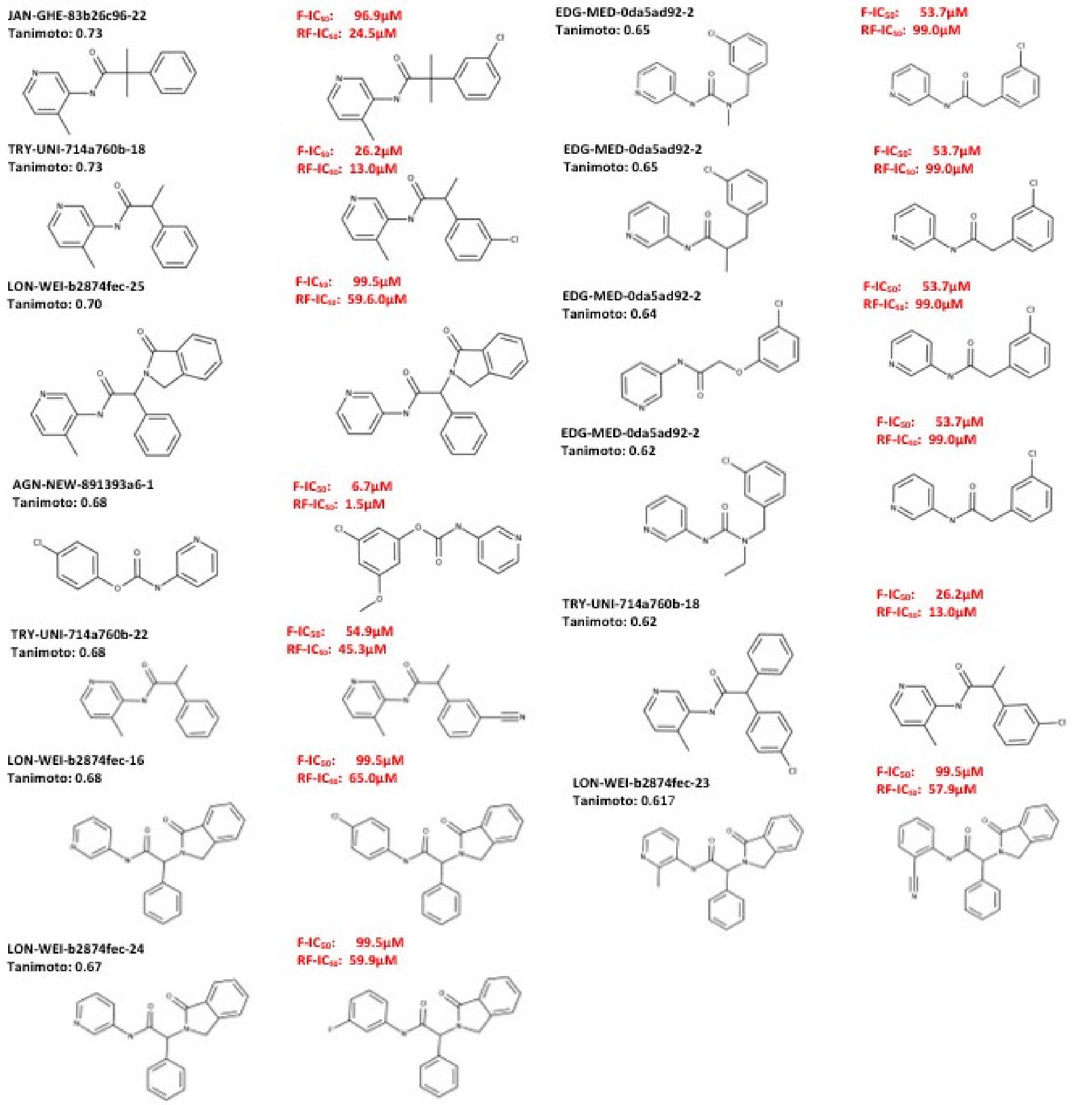


Figure A.7: **The Fragment Network identifies close analogues to known Mpro inhibitors.** Example compounds identified using the Fragment Network (left-hand side) and similar compounds with half-maximal inhibitory concentration (IC$_{50}$) values within the micromolar range, recorded using either a fluorescence assay (F-IC$_{50}$) or RapidFire mass spectrometry (RF-IC$_{50}$). Similar compounds were identified by calculating the Tanimoto similarity using Morgan fingerprints (radius 2; 2,048 bits).

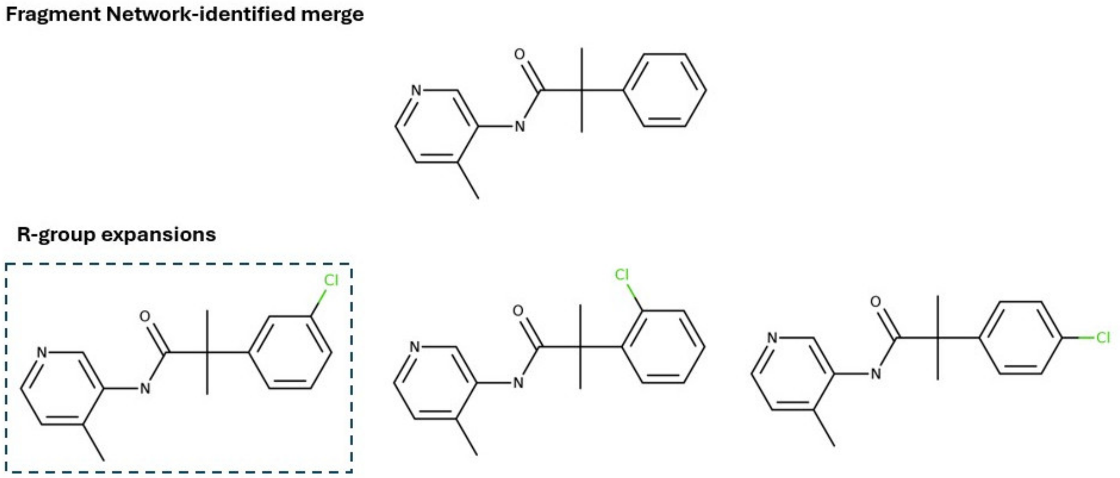


Figure A.8: **R-group expansion retrieves inhibitor JAN-GHE-83b26c96-22.** R-group expansion of a Fragment Network-derived merge identifies a known inhibitor with a RapidFire mass spectrometry half-maximal inhibitory concentration (IC_{50}) value of 24.5μM.

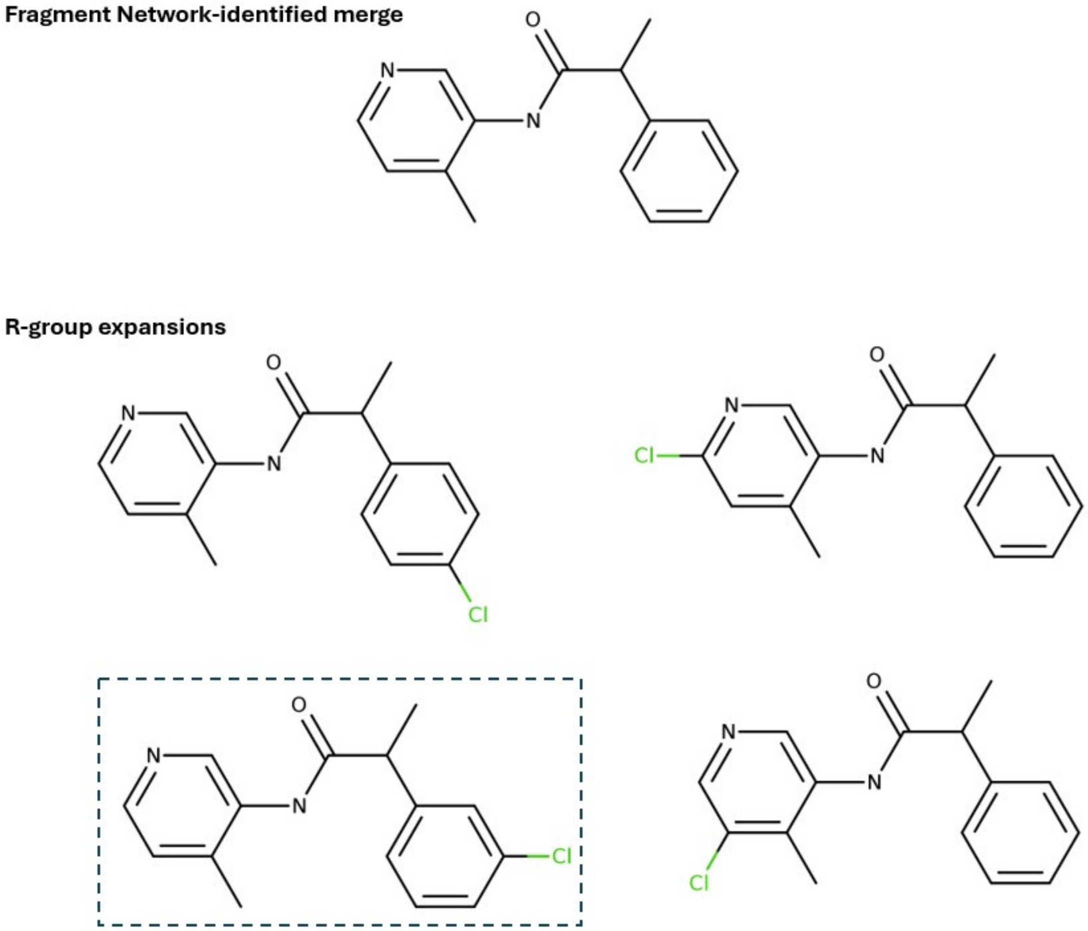


Figure A.9: **R-group expansion retrieves inhibitor TRY-UNI-714a760b-18.** R-group expansion of a Fragment Network-derived merge identifies a known inhibitor with a fluorescence assay half-maximal inhibitory concentration (IC_{50}) value of 25.2μM and a RapidFire mass spectrometry IC_{50} value of 13.0μM.

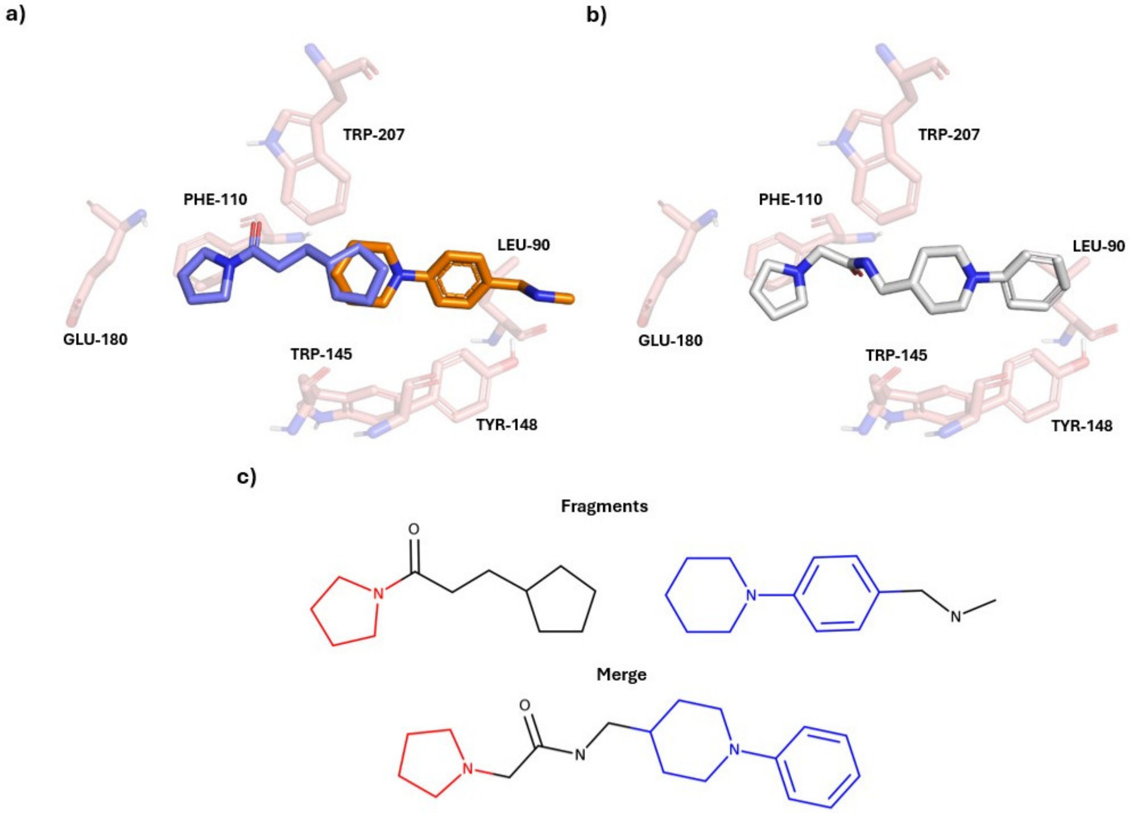


Figure A.10: **An example merge against EthR found using the entire filtering pipeline.** (**a**) The crystal structures of two parent fragments against EthR. (**b**) The orientation of the merge (white) generated using Fragmenstein. (**c**) The fragments and merge in 2D; colours indicate the substructures used in forming the merge.

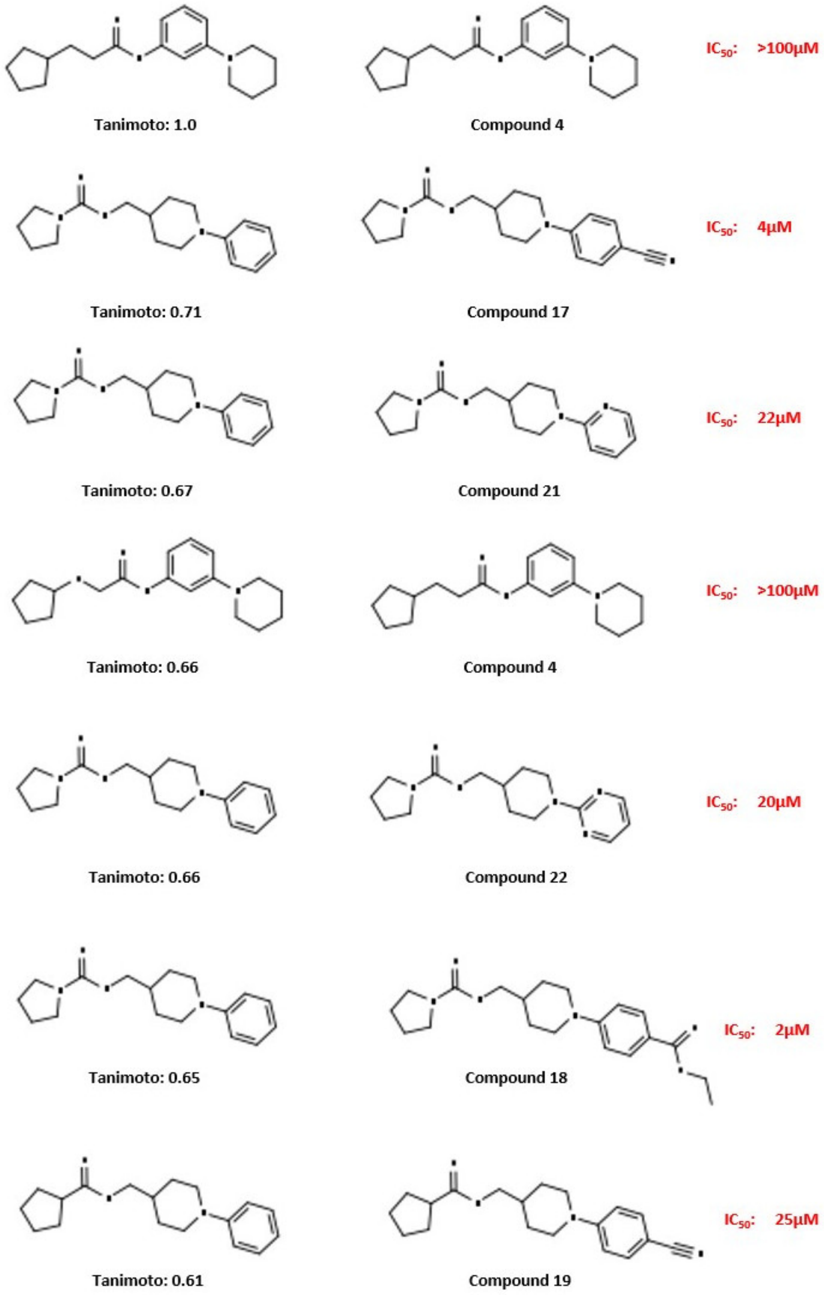


Figure A.11: **The Fragment Network identifies close analogues to known EthR inhibitors.** Example compounds identified using the Fragment Network (left-hand side) and similar compounds with IC_{50} values recorded using surface plasmon resonance. Similar compounds were identified by calculating the Tanimoto similarity using Morgan fingerprints (radius2; 2,048 bits).

A.1 Incorporating atom-to-atom mapping within the Fragment Network

As mentioned in Section 2.5, one limitation with the current Fragment Network architecture used at XChem is that it does not record complete atom-to-atom mappings (AAMs) for transformations between molecule nodes. The original implementation (Hall et al., 2017) describes a method to assign distinct labels for attachment point atoms on substructures undergoing transformations; this allows transformations at specific locations to be grouped, but does not describe a complete AAM method that would allow one to record atom correspondences between complex, multi-hop query paths.

While it is possible to correct for the lack of AAM in post-processing (as was done in the pipelines described in this thesis); addressing this would help to improve the preciseness of the database search. This was deemed to be beyond the scope of the work performed in Chapters 2 and 3, as the primary aim was to identify an initial set of merges that could recapitulate the scaffolds of the fragments, followed by a relatively computationally cheap R-group search on the filtered compounds.

One of the main challenges identified is the problem presented by molecular symmetry. This refers to the fact that for symmetric molecules, there may be multiple possible and valid mappings for a given transformation. For query paths involving several hops and transformation steps, it is important to be able to record all possible mappings throughout the query path, to prevent ruling out valid molecules. An example scenario for where this proves important is shown in Appendix Figure A.12. Incorporating AAM would enable the following functions:

1. Replace a substructure at the same position on a molecule
2. Decompose complex substructures into multi-step transformations
3. Rule out initial attachment points given the 3D arrangement of the molecules

For molecular symmetry in reaction data, a solution employed by the AMNet model was to identify topologically equivalent atoms (atoms with equivalent atomic environments) using the Weisfeiler–Lehman (WL) test (Astero et al., 2024), an algorithm for testing whether graphs are isomorphic (Astero et al., 2024).

A possible approach to handle molecular symmetry is to use the RDKit implementation of the Morgan fingerprint hashing procedure (Rogers et al., 2010) to detect topologically equivalent atoms (Landrum, 2024). To generate ECFPs, an identifier is assigned to each atom in the molecule; the identifier depends on certain atomic properties, such as the atom number, valence and charge. Following this, the identifier is updated over several iterations, whereby the identifier or the atoms and its neighbours (the size of the neighbourhood is increased in each successive iteration) are combined into an array and hashed to yield a new identifier. The number of iterations (and maximum size of the neighbourhood) is pre-determined by the user. In this way, we can generate a set of identifiers for each atom based on its atomic environment, meaning topologically equivalent atoms will be assigned identical identifiers.

To do this, given an edge in the database, via which a substructure is added, we start with SMILES representing the synthon (used here to refer to the substructure being added) and core (the substructure that is being added to). Following generating an AAM, the Morgan fingerprint can be calculated using RDKit (Landrum, 2024) (using neighbourhood size of 3). From this, we can record the identifiers for each atom at each iteration. This can be used both to extract the attachment point (as the identifier for the mapped atom at radius 0 will be different before and after the transformation), but we can also record what attachment points in the synthon and/or core are topologically equivalent. Using this, an atomic mapping can be generated and important equivalent atoms can be generated, as shown in Appendix Figure A.13.

Future work will require applying this to a small test database to finalize what the implementation should look like, the metadata that need to be recorded to allow new query types, and possible limitations, such as the increased memory required for storing the extra edge labels in the database.

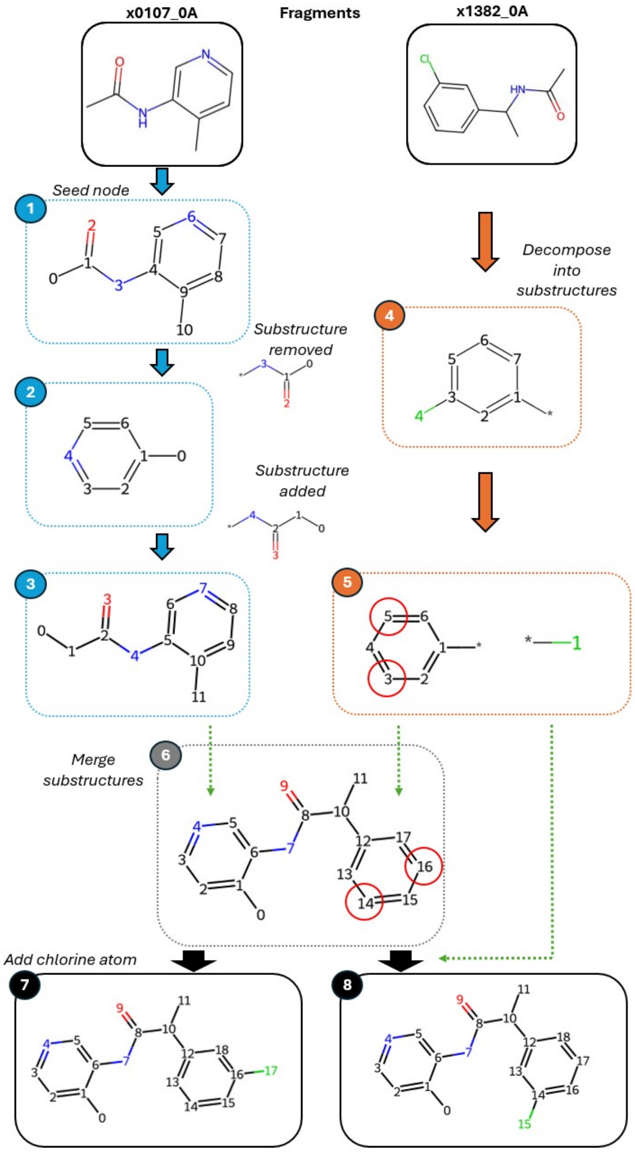
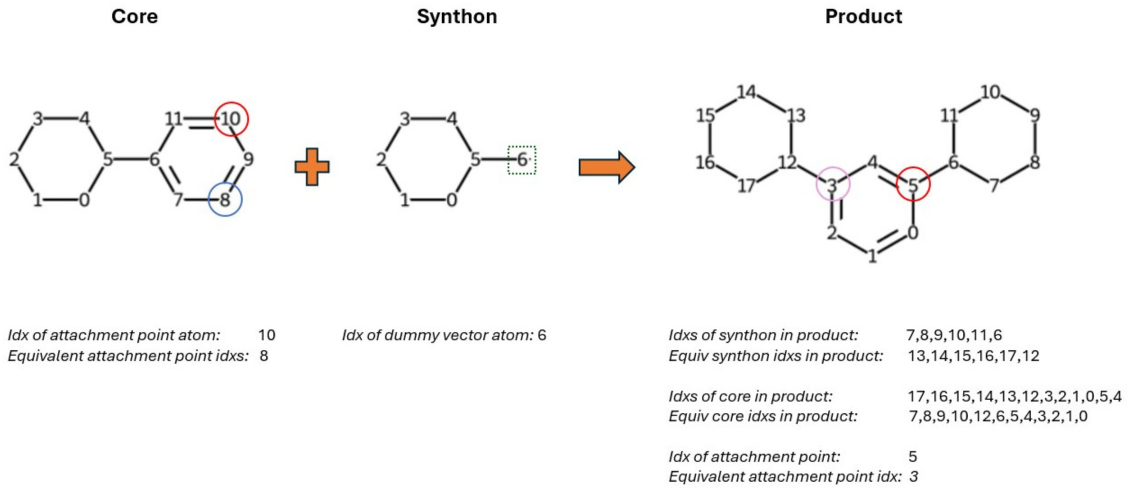


Figure A.12: **An example usage for atom mapping in the Fragment Network.** Atom mapping is required to track the atom correspondences between database hops and to record which are topologically equivalent. An example path is shown for merging fragments x0107_0A and x1382_0A. Hops are made away from fragment x0107_0A (the seed node in the network query) to generate diversity (boxes 1–3). Incorporating the chlorobenzene from fragment x1382_0A represents a two-hop expansion, as it can be decomposed into a benzene ring and chlorine atom (boxes 4 and 5). Due to the molecular symmetry within the chlorobenzene R-group, there are two equivalent attachment atoms for the chlorine atom (shown by the red circles; box 5). Following incorporation of the benzene, atom mapping would enable us to specify that we want to add the chlorine atom (box 6) at one of these topologically equivalent attachment points (represented by atoms 14 and 16 in box 6); the desired result is shown in boxes 7 and 8.



Idx of attachment point atom: 10
Equivalent attachment point idxs: 8

Idx of dummy vector atom: 6

Idxs of synthon in product: 7,8,9,10,11,6
Equiv synthon idxs in product: 13,14,15,16,17,12

Idxs of core in product: 17,16,15,14,13,12,3,2,1,0,5,4
Equiv core idxs in product: 7,8,9,10,12,6,5,4,3,2,1,0

Idx of attachment point: 5
Equivalent attachment point idx: *3*

Figure A.13: **An example of the information that can be recorded by using Morgan fingerprint identifiers in atom mapping.** Given SMILES using a core molecule, a synthon to be added and a product molecule (after addition of the synthon to the core), an atom mapping can be calculated using RDKit (Landrum, 2024). Morgan fingerprint (Rogers et al., 2010) identifiers can be used to identify topologically equivalent atoms (for example, an equivalent attachment point atom) in cases of molecular symmetry. Numberings depict the default numbering assigned by RDKit. Circles show described attachment points; circles within the same molecule are topologically equivalent.

B. Appendix items for Chapter 3

Table B.1: Crystallographic fragment screening hits used for merging.

Target	Fragments
EV D68 3C protease	**P1**: x0147_0A, x0771_0A, x0789_0A, x0980_0B, x1332_0A, x1498_0A, x1537_0A, x1594_0A, x1604_0A, x2099_0A, x2135_0A, x2148_0A, x2149_0A, x1163_0A **P2**: x0130_0A, x1020_0A, x1083_0A, x1084_0A, x1140_0A, x1247_0A **P2 & P3**: x1071_0A, x1329_0A, x1919_0A, x2021_0A, x1052_0A
EV A71 2A protease	x0451_0A x0554_0A x0556_0A x0566_0A x0310_0A x0416_0A
Zika NS2B	**S1**: x0051_0B x0089_0B x0101_0B x0182_0B x0227_0B x0229_0B x0382_0B x0386_0B x0422_0B x0425_0B x0429_0B x0435_0B x0443_0B x0455_0B x0465_0B x0472_0B x0490_0B x0553_0B x0559_0B x0589_0B x0605_0B x0680_0B x0687_0B x0693_0B x0727_0B x0773_0B x0788_0B x0800_0B x0803_0B x0852_0B x0884_0B x0904_0B x0917_0B x0945_0B x0951_0B x0969_0B x0990_0B **S1′**: x0846_0B **S2**: x0404_0B x0969_1B x1098_0B

EV, enterovirus.

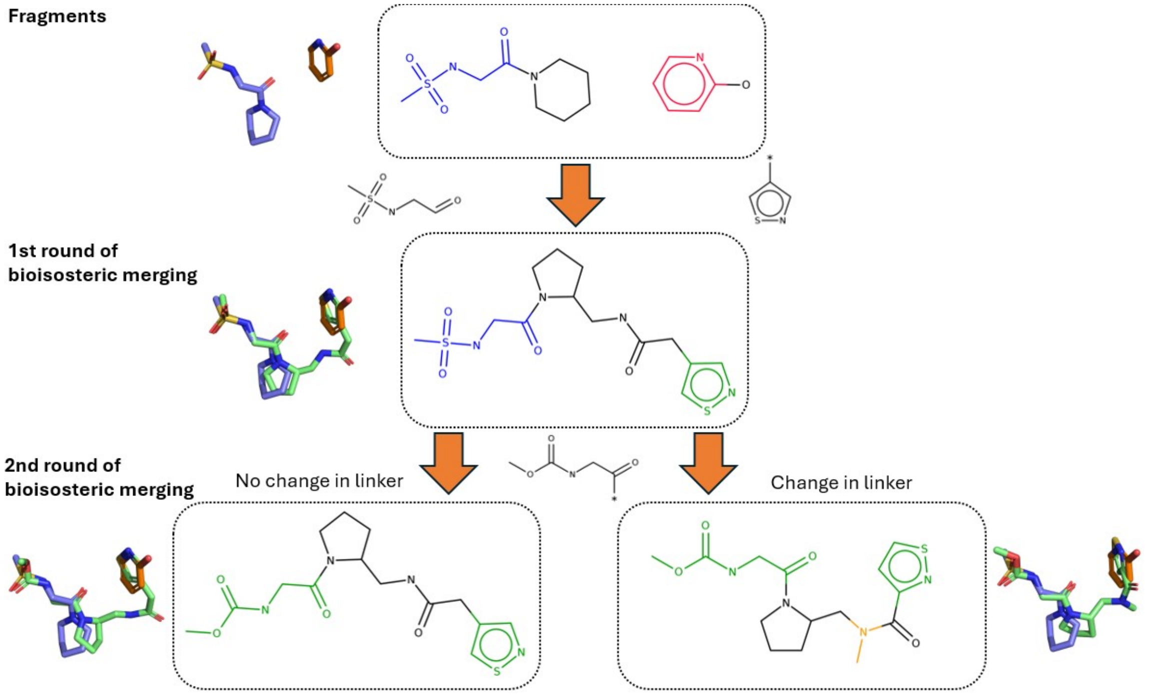


Figure B.1: **Bioisosteric merging process.** Given a pair of fragments (top box), a bioisosteric merge (middle box) is found containing one exact substructure (blue) from the original fragments and one substructure that is pharmacophorically similar (green). The database can be re-queried to replace the exact substructure so that two substructures are being incorporated that are pharmacophorically similar to those seen in the parent fragments (bottom boxes). This can be done using a strict search, where there is no change in the linker region (bottom-left box) or where changes can be made in the linker region (bottom-right box; change is shown in yellow).

B.1 Cypher query structure

Example Cypher queries are shown below for the different types of merging. The queries have been simplified for illustrative purposes. Values within inequality symbols (<VALUE>) indicate user defined parameters and values. Colons are used to represent labels (of specific node or edge types): :Mol represents nodes that are purchasable molecules; :F2 represent all molecule nodes (i.e., whether they represent purchasable molecules or intermediate transformations between purchasable molecules), and :FRAG represents edges whereby a transformation (a contraction or expansion) is being made.

Perfect merging

Below is an example Cypher query for finding a perfect merge. First, the node representing the seed substructure is identified (<SMILES>). Following this, up to two hops are made whereby expansions occur, which generates diversity ([:FRAG*0..2]), followed by an additional hop ([e:FRAG]), to identify a purchasable catalogue compound (c:Mol). The expansion specifies that the query substructure (<QUERY_SUBSTRUCTURE> , representing the other substructure in the pair) is to be incorporated, based on SMILES matching. The direction of the arrow representing the hop indicates whether a contraction or expansion is being made (for example, '(a:F2)-[:FRAG]->(b:F2)' indicates that a contraction has been made to molecule **a** to yield molecule **b**). The SMILES of the purchasable molecule (representing the merge) is returned on the final line (RETURN molecule_smiles).

```
MATCH (a:F2 {smiles: <SMILES> })<- [:FRAG*0..2] -(b:F2)<- [e:FRAG] -( c:Mol )
WHERE e.substructure = <QUERY_SUBSTRUCTURE>
WITH c.smiles as molecule_smiles
RETURN molecule_smiles
```

Bioisosteric merging: round 1

Below shows an example Cypher query for finding a bioisosteric merge. The node representing the seed substructure is identified (<SMILES>). As above, up to two hops are made whereby expansions occur ([:FRAG*0..2]), followed by an additional hop ([e:FRAG]) to identify a purchasable catalogue compound (c:Mol). Similarity is calculated between the pharmacophore fingerprint of the substructure incorporated in the final hop (e.pharmfp) and a query pharmacophore fingerprint, representing the other substructure in the pair (<QUERY_PHARMFP>), while ensuring merges are not retrieved with the exact original substructure (which is achieved by perfect merging). The SMILES of the final purchasable molecule is returned on the last line (RETURN molecule_smiles).

```
MATCH (a:F2 {smiles: <SMILES>})<-[:FRAG*0..2]-(b:F2)<-[e:FRAG]-(c:Mol)
WHERE EXISTS(e.pharmfp)
WITH similarity.tanimoto_similarity(e.pharmfp, <QUERY_PHARMFP>) as tanimoto,
c.smiles as molecule_smiles
WHERE tanimoto >= <THRESHOLD>
AND NOT e.substructure = <QUERY_SUBSTRUCTURE>
RETURN molecule_smiles
```

Bioisosteric merging: round 2

Below shows an example Cypher query for performing an additional round of bioisosteric merging. The SMILES of the first merge is matched (<SMILES>), which then undergoes a contraction to remove the original seed substructure in the query ([e1:FRAG]) followed by an additional contraction ([e2:FRAG]) and expansion ([e3:FRAG]) to generate diversity in the linker region linking the two substructure. (This step can be skipped to allow a stricter search.) Following this, an expansion is made in which a similar substructure is incorporated ([e4:FRAG]). The WHERE clauses set constraints on the query, ensuring that the same substructure isn't removed and re-added, and that the substructure added in the first round of bioisosteric merging isn't removed.

```
MATCH (a:F2 {smiles: <SMILES> })-[e1:FRAG]->(b:F2)-[e2:FRAG]->(c:F2)
<-[e3:FRAG]-(d:F2)<-[e4:FRAG]-( f:Mol )
WHERE e1.substructure = <QUERY_SUBSTRUCTURE>
AND EXISTS(e4.pharmfp)
AND NOT e2.substructure = <FIRST_SUBSTRUCTURE>
AND NOT e2.substructure = e3.substructure
AND NOT e3.substructure = e1.substructure
WITH similarity.tanimoto_similarity(e1.pharmfp, e4.pharmfp) as tanimoto,
f.smiles as molecule_smiles
WHERE tanimoto >= <THRESHOLD>
RETURN molecule_smiles
```

Table B.2: Parameters used for docking with rDock

Parameter	Value
Type of cavity method	Reference ligand
Cavity radius	8.0
Number of docks	30
Pharmacophore weight	1
Tolerance radius	2.0

Table B.3: Pharmacophore features for docking merges against EV D68 3C protease

Pair	Requirement	x	y	z	Type
x1071_0A-x1498_0A	Mandatory	-7.09	-4.283	-3.9	Acc
	Mandatory	-7.115	-1.971	-2.213	Don
	Mandatory	-5.876	-1.261	0.271	Acc
	Mandatory	-6.835	-3.99	-9.255	Acc
	Optional	-7.907	0.111	-0.552	Hyd
	Optional	-4.912	-7.449	-2.158	Hyd
	Optional	-4.912	-7.449	-2.158	Hyd
	Optional	-8.047	-6.301	-6.811	Hyd
x1083_0A-x1498_0A	Mandatory	-6.835	-3.99	-9.255	Acc
	Optional	-2.403	-7.764	1.32	Hyd
	Optional	-5.963	-8.561	-1.504	Hyd
	Optional	-8.047	-6.301	-6.811	Hyd
x1140_0A-x1498_0A	Mandatory	-6.741	-4.237	-3.743	Acc
	Mandatory	-6.835	-3.99	-9.255	Acc
	Optional	-6.412	-4.724	-0.83	Hyd
	Optional	-5.332	-8.174	-1.881	Hyd
	Optional	-8.047	-6.301	-6.811	Hyd

Constraints were generated using the procedure described in Section 3.3.7.

Table B.4: Pharmacophore features for docking merges against EV A71 2A protease

Pair	Requirement	x	y	z	Type
x0310_0A-x0416_0A	Mandatory	10.289	11.94	22.269	Don
	Optional	8.219	11.784	24.252	Hyd
	Optional	10.009	8.93	22.729	Hyd
	Optional	11.934	13.737	22.43	Hyd
	Optional	9.447	11.481	23.619	Hyd
	Optional	9.759	9.028	23.134	Hyd
x0310_0A-x0556_0A	Mandatory	10.289	11.94	22.269	Don
	Mandatory	6.244	9.398	26.831	Don
	Optional	8.219	11.784	24.252	Hyd
	Optional	10.009	8.93	22.729	Hyd
	Optional	11.934	13.737	22.43	Hyd
	Optional	8.695	11.508	23.717	Hyd
	Optional	4.995	8.454	28.88	Hyd
	Optional	4.995	8.454	28.88	Hyd
x0416_0A-x0556_0A	Mandatory	6.244	9.398	26.831	Don
	Optional	9.447	11.481	23.619	Hyd
	Optional	9.759	9.028	23.134	Hyd
	Optional	8.695	11.508	23.717	Hyd
	Optional	4.995	8.454	28.88	Hyd
	Optional	4.995	8.454	28.88	Hyd

Constraints were generated using the procedure described in Section 3.3.7.

Table B.5: Pharmacophore features for docking merges against Zika NS2B

Pair	Requirement	x	y	z	Type
x0089_0B-x1098_0B	Mandatory	-6.917	5.115	-18.61	Aro
	Mandatory	-4.785	5.51	-20.319	Don
	Mandatory	-9	4.657	-17.727	Aro
	Mandatory	-8.699	5.289	-11.548	Aro
	Mandatory	-9.161	4.117	-14.146	Don
	Optional	-2.549	4.447	-19.752	Hyd
	Optional	-9.427	4.563	-18.786	Hyd
	Optional	-2.549	4.447	-19.752	Hyd
	Optional	-7.384	5.028	-17.559	Hyd
	Optional	-9.345	4.386	-19.035	Hyd
	Optional	-7.729	5.164	-17.427	Hyd
	Optional	-8.825	4.899	-12.652	Hyd
	Optional	-7.729	5.164	-17.427	Hyd
x0429_0B-x1098_0B	Mandatory	-8.664	4.845	-17.588	Aro
	Mandatory	-9	4.657	-17.727	Aro
	Mandatory	-8.699	5.289	-11.548	Aro
	Mandatory	-9.161	4.117	-14.146	Don
	Optional	-9.99	4.459	-17.653	Hyd
	Optional	-8.059	4.976	-16.345	Hyd
	Optional	-9.345	4.386	-19.035	Hyd
	Optional	-7.729	5.164	-17.427	Hyd
	Optional	-8.825	4.899	-12.652	Hyd
	Optional	-7.729	5.164	-17.427	Hyd
x0687_0B-x0969_1B	Mandatory	-8.952	4.254	-15.218	Acc
	Mandatory	-6.504	5.241	-18.609	Aro
	Mandatory	-8.956	5.417	-9.301	Aro
	Mandatory	-9.077	1.514	-9.134	Don
	Optional	-7.121	5.464	-17.383	Hyd
	Optional	-9.356	4.727	-18.369	Hyd
	Optional	-7.121	5.464	-17.383	Hyd
	Optional	-8.708	5.605	-11.972	Hyd

Constraints were generated using the procedure described in Section 3.3.7.

Table B.6: Total hours for querying Fragment Network-derived compounds

Target	EV D68 3C protease	EV A71 2A protease	Zika NS2B protease
Bioisosteric (round 1)	186.31	1.96	419.85
Perfect	79.76	3.51	82.14
Bioisosteric (round 2: strict search)	0.41	0.39	0.23
Bioisosteric (round 2: loose search)	9.72	0.33	40.15
R-group: bioisosteric (round 1)	0.02	0.01	0.01
R-group: perfect	0.02	0	0.02
R-group: bioisosteric (round 2: strict search)	0	0	0
R-group: Bioisosteric (round 2: loose search)	0.01	0	0
Total time	276.25	6.2	542.4
Estimated average time per fragment pair	3.63	1.13	6.42

Figures show the total times (in hours) for running each individual query. Up to 8 queries were run in parallel on a VM with 120 CPUs. EV, enterovirus.

Table B.7: Total CPU hours for filtering Fragment Network-derived compounds

	EV D68 3C protease	EV A71 2A protease	Zika NS2B protease
Bioisosteric (round 1)	1754.81	87.8	791.9
Perfect	2492.74	138.74	1184.67
Bioisosteric (round 2)	1149.65	24.26	115.66
R-group: bioisosteric (round 1)	30.2	6.5	23.66
R-group: perfect	44.7	6.04	45.16
R-group: bioisosteric (round 2: loose search)	13.67	0.48	5.18
R-group: Bioisosteric (round 2: strict search)	9.82	0.98	3.37
Total	5495.59	264.8	2169.6

EV, enterovirus

Table B.8: The number of clusters containing the top-ranked Fragment Network or rDock-derived compounds.

Target	Pair	Number of FN clusters	Number of rDock clusters
EV D68 3C protease	x1071_0A-x1498_0A	56	100
	x1083_0A-x1498_0A	73	100
	x1140_0A-x1498_0A	70	100
EV A71 2A protease	x0310_0A-x0416_0A	62	100
	x0310_0A-x0556_0A	60	99
	x0416_0A-x0556_0A	65	99
Zika NS2B protease	x0089_0B-x1098_0B	58	98
	x0429_0B-x1098_0B	60	100
	x0687_0B-x0969_1B	54	100

Clustered using the Taylor-Butina algorithm using Tanimoto similarity (threshold of 0.4) and Morgan fingerprints (radius 2; 1,024 bits). EV, enterovirus; FN, Fragment Network.

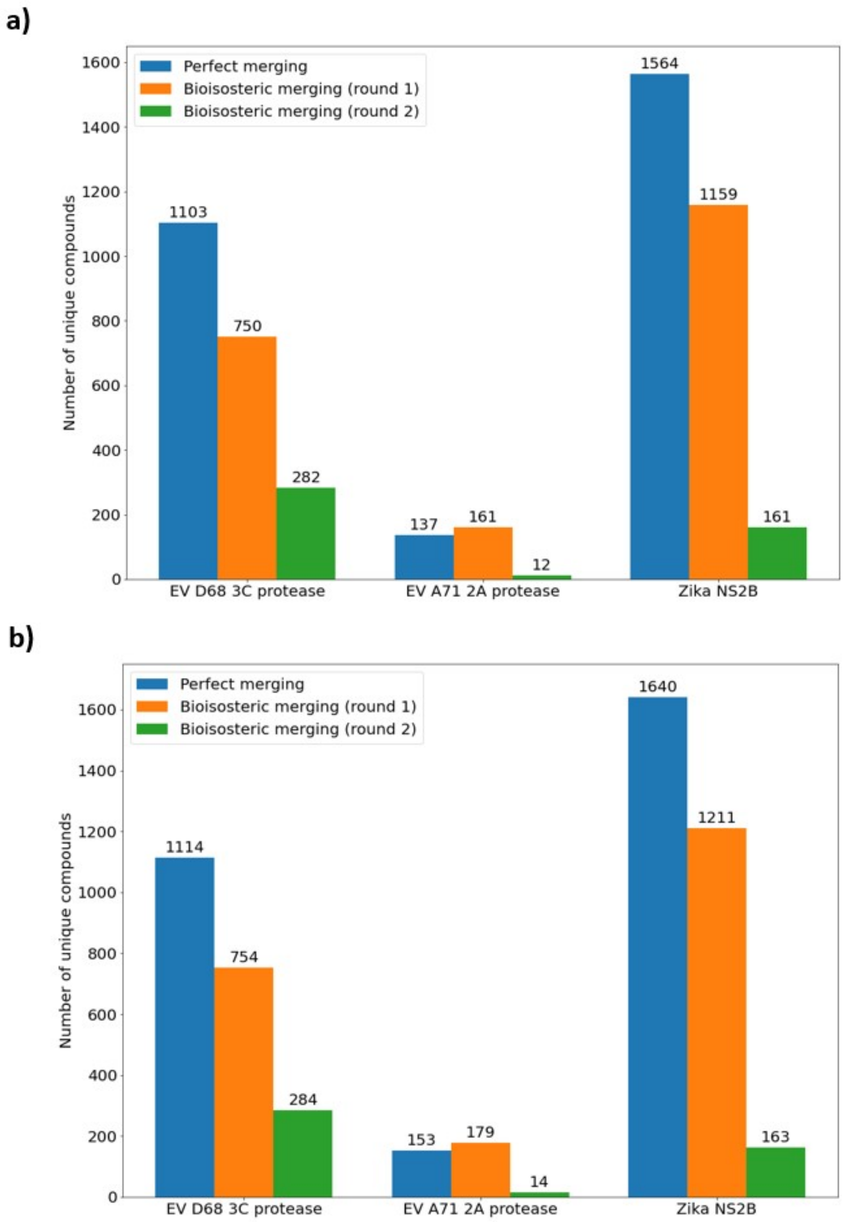


Figure B.2: **Numbers of successfully filtered compounds from Fragment Network merging pipelines.** Compounds retrieved from the database by either perfect merging or bioisosteric merging, replacing one (round 1) or two (round 2) substructures are placed using an adapted Fragmenstein protocol (Ferla et al., 2024) and minimized using PyRosetta (Chaudhury et al., 2010). The results are shown for compounds identified (**a**) before and (**b**) after performing an additional R-group search (to recapitulate substituents observed in the original fragments). Only the number of unique, successfully placed and filtered compounds with SC_{RDKIT} values ≥ 0.55 are shown. EV, enterovirus.

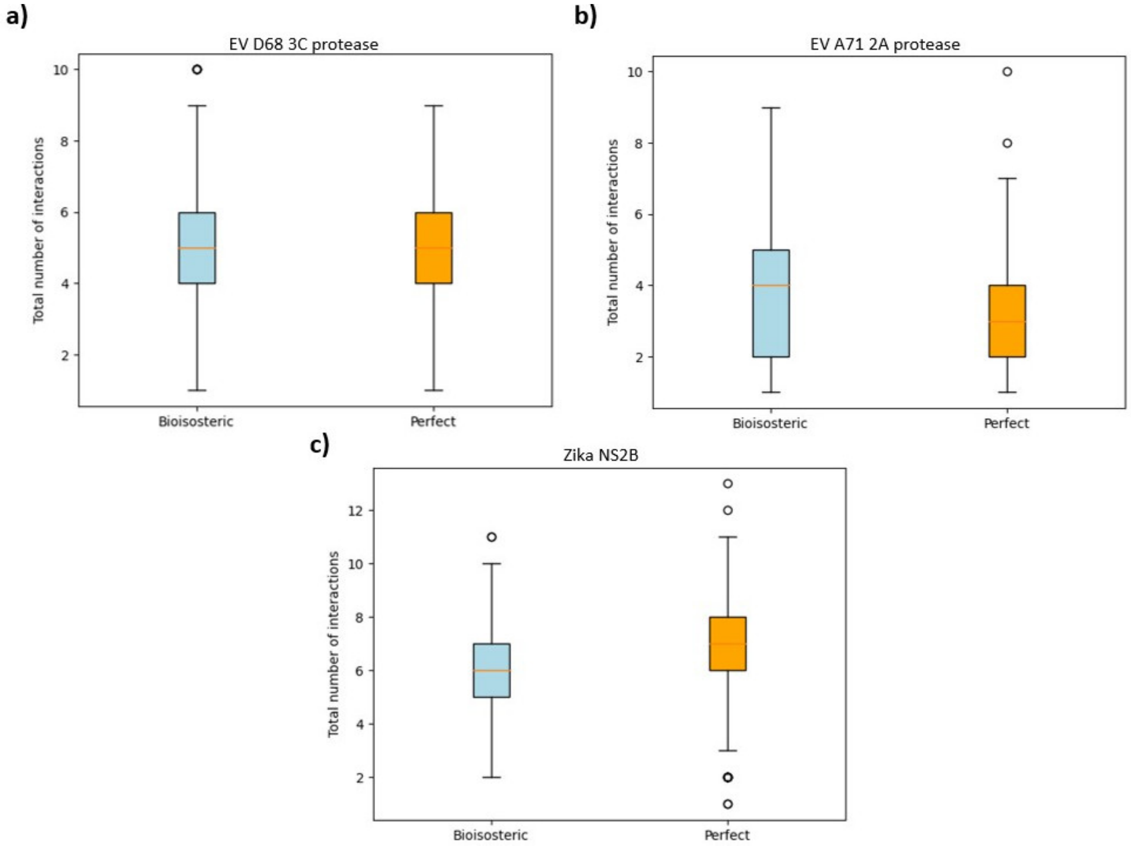


Figure B.3: **Total number of predicted interactions made by each merge.** The total number of predicted interactions for each compound proposed by bioisosteric merging and perfect merging for (**a**) enterovirus (EV) D68 3C protease, (**b**) EV A71 2A protease and (**c**) Zika NS2 protease. Interactions are calculated using ProLIF. Only compounds with a $\mathrm{SC}_{\mathrm{RDKIT}}$ score ≥ 0.55 and represent 'true merges' (potentially replicating an interaction from each fragment) are shown. The results are shown to include all compounds after R-group expansions.

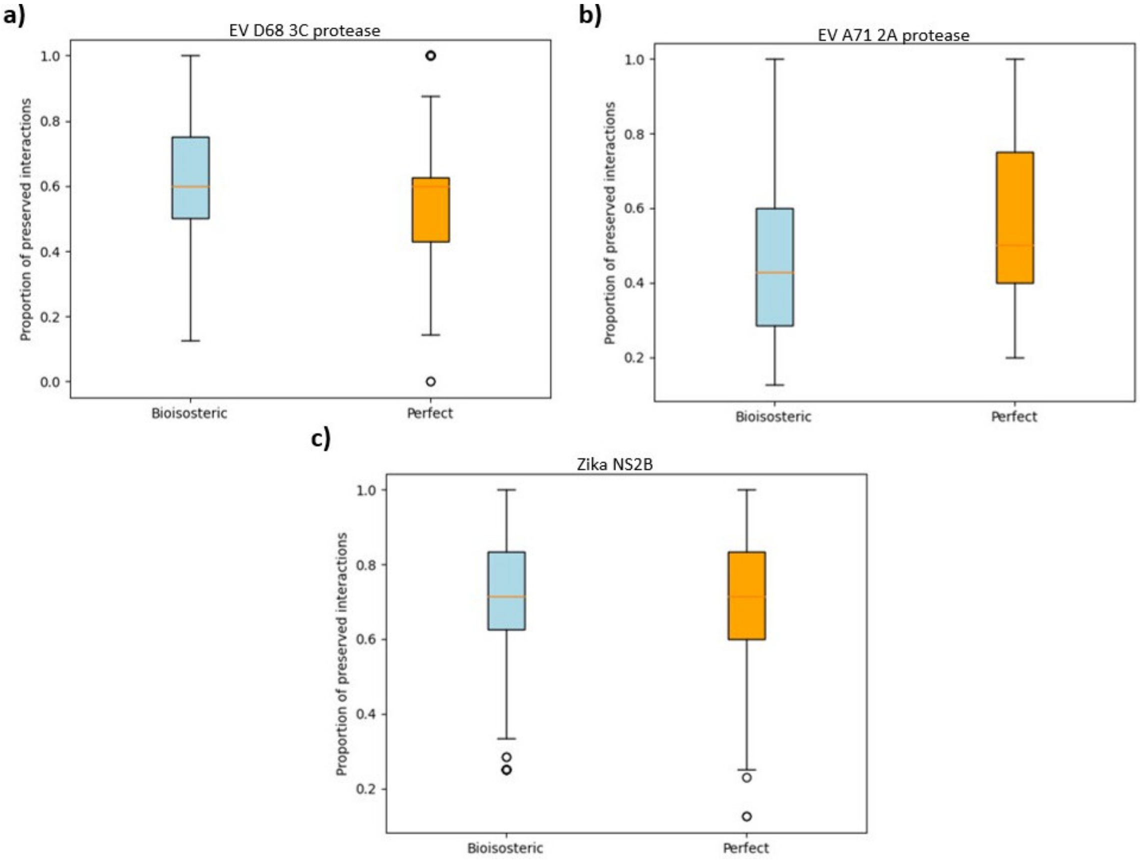


Figure B.4: **Fraction of preserved interactions.** The proportion of preserved predicted interactions observed at the residue-level (that is, unique interactions that are seen in the parent fragments that are replicated by the merge) for each compound proposed by bioisosteric merging and perfect merging for (**a**) enterovirus (EV) D68 3C protease, (**b**) EV A71 2A protease and (**c**) Zika NS2B protease. Interactions are calculated using ProLIF. Only compounds with a $\mathrm{SC}_{\mathrm{RDKIT}}$ score ≥ 0.55 and represent 'true merges' (replicating an interaction from each fragment) are shown. The results are shown to include all compounds after R-group expansions.

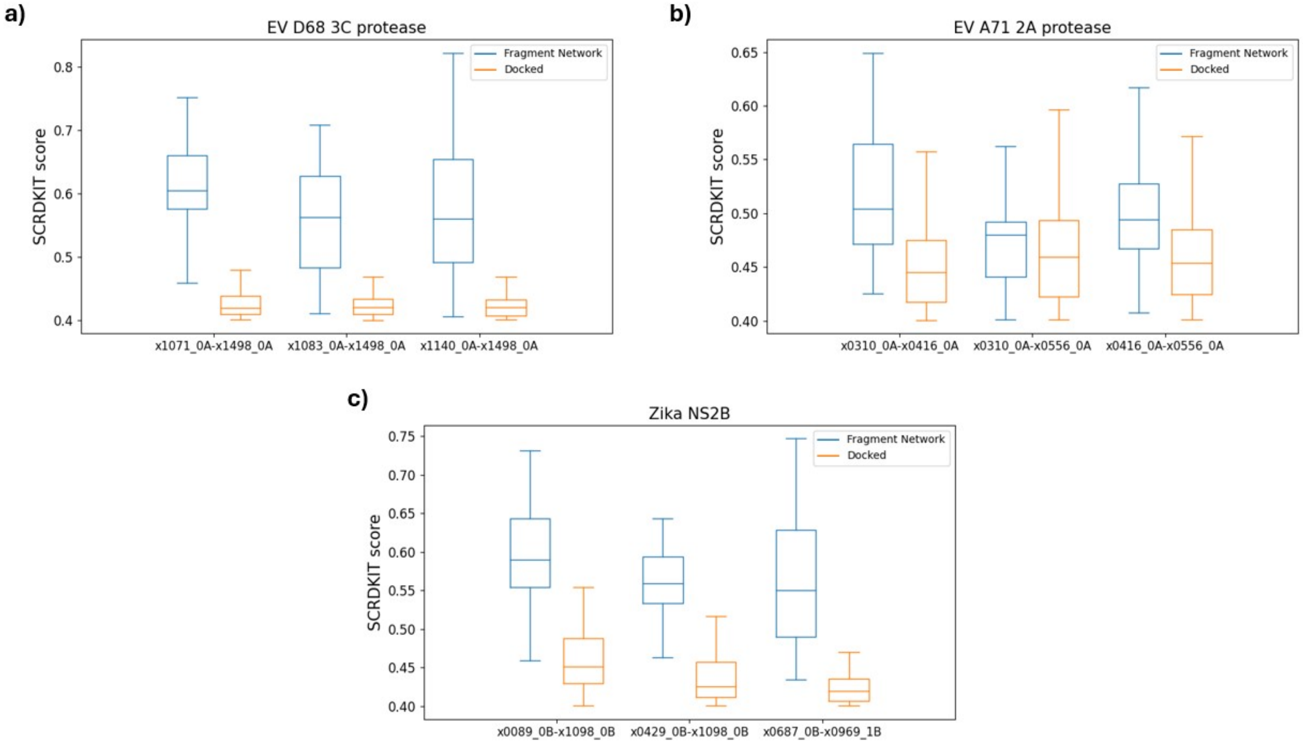


Figure B.5: **SC_{RDKIT} scores made by top 100 compounds identified using the Fragment Network merging pipelines versus pharmacophore-constrained docking.** EV, enterovirus.

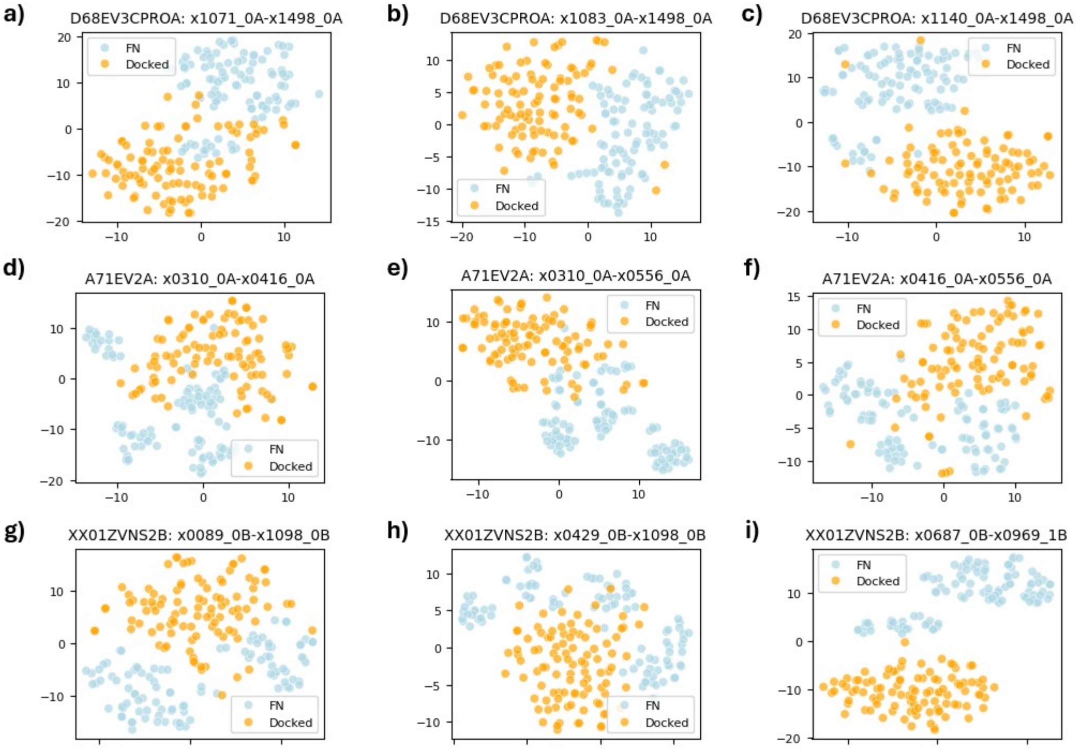


Figure B.6: **TSNEs for top 100 compounds identified using the Fragment Network merging pipelines versus pharmacophore-constrained docking.** EV, enterovirus.

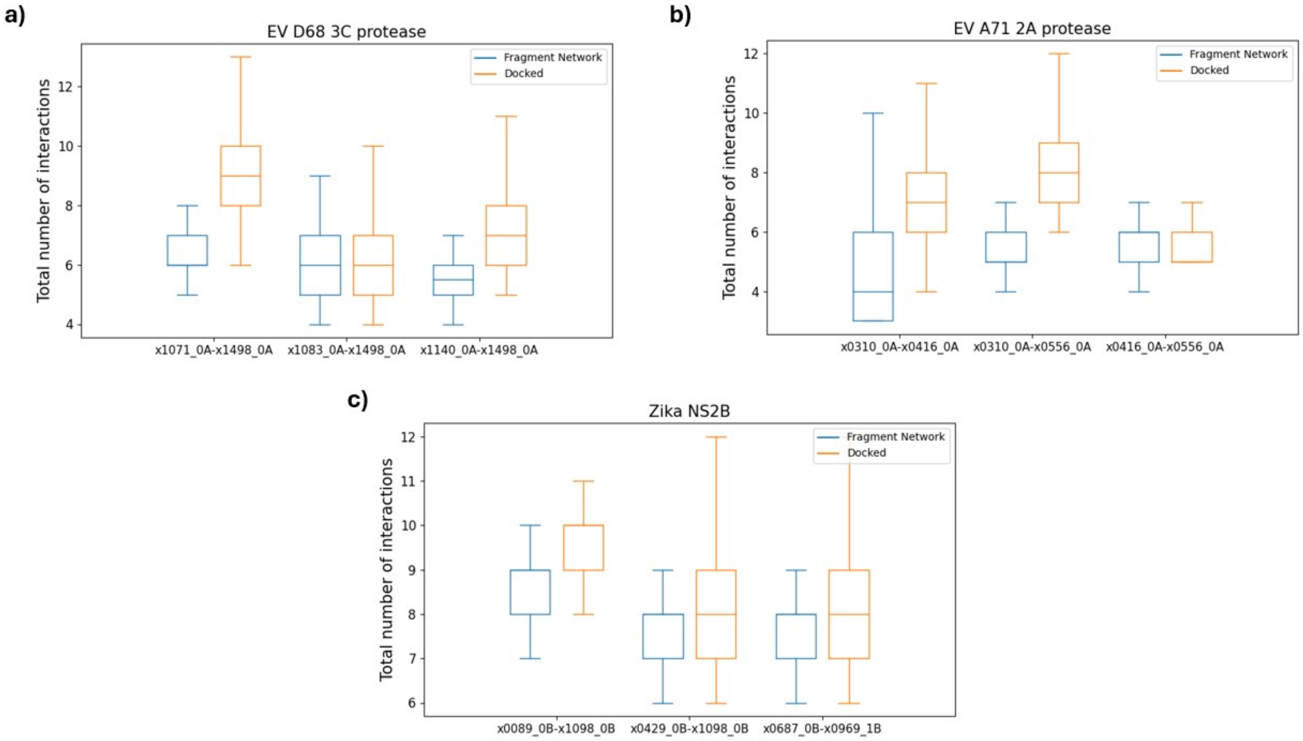


Figure B.7: **Total number of predicted interactions made by top 100 compounds identified using the Fragment Network merging pipelines versus pharmacophore-constrained docking.** EV, enterovirus.

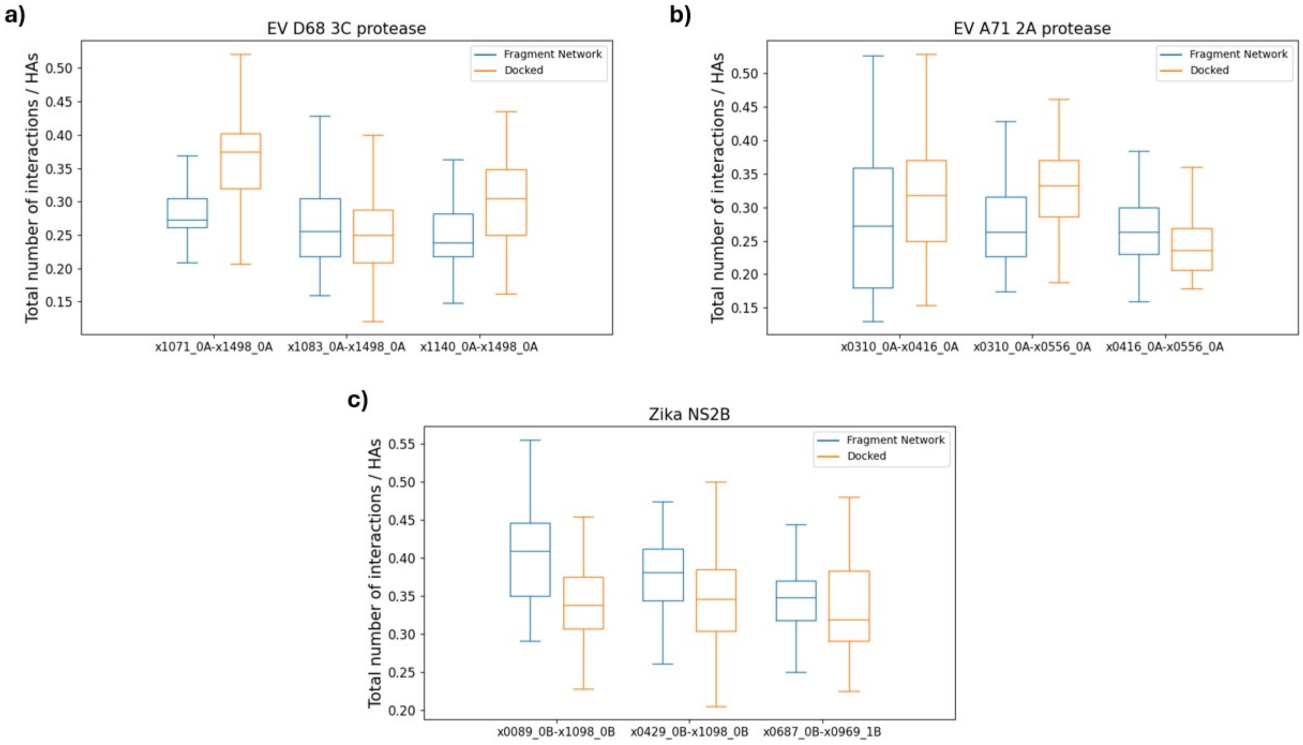


Figure B.8: **Number of predicted interactions normalized by heavy atom count made by top 100 compounds identified using the Fragment Network merging pipelines versus pharmacophore-constrained docking.** EV, enterovirus.

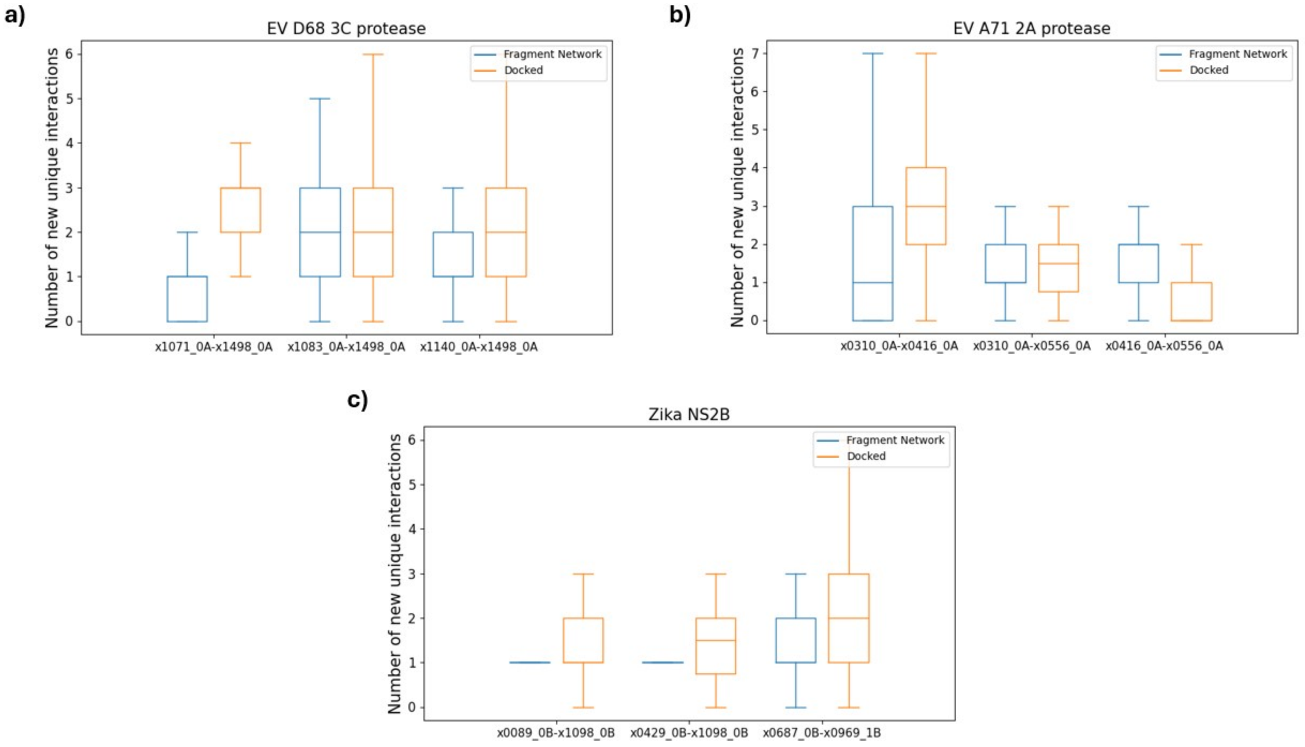


Figure B.9: **Number of new predicted interactions not seen in parent fragments made by top 100 compounds identified using the Fragment Network merging pipelines versus pharmacophore-constrained docking.** EV, enterovirus.

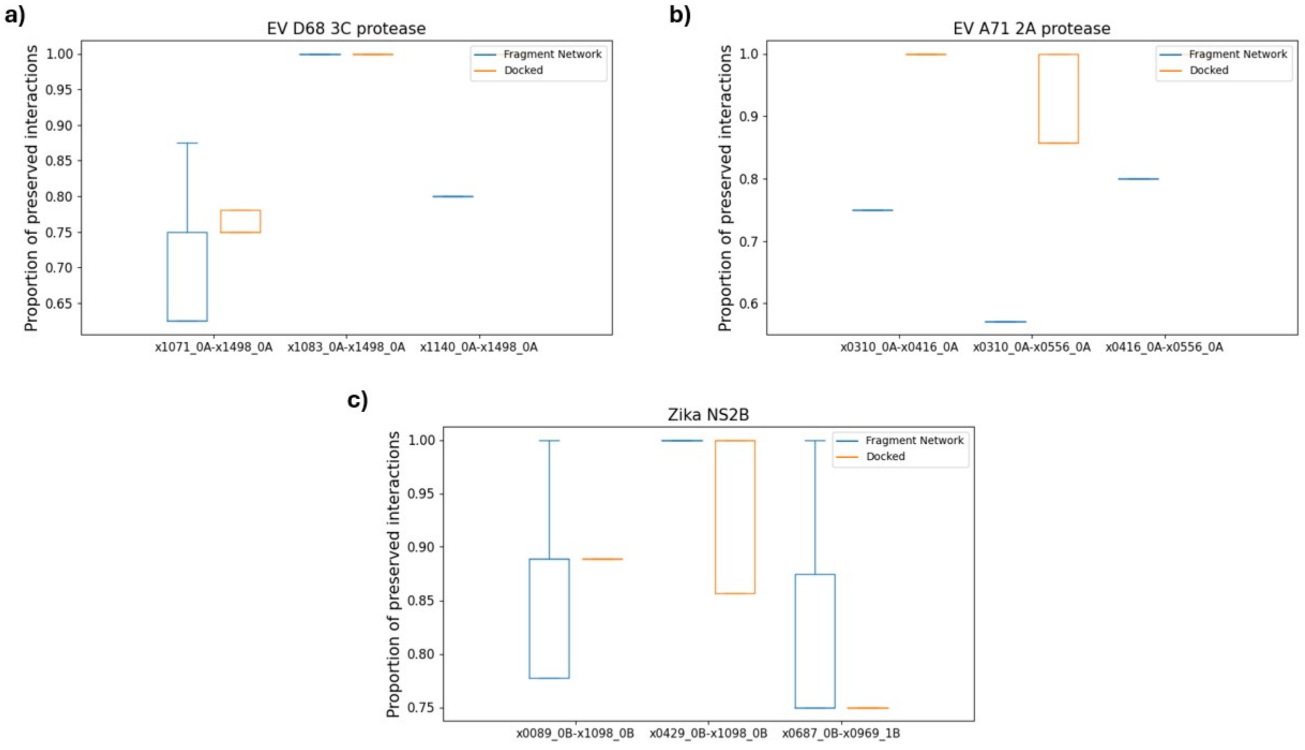


Figure B.10: **The fraction of preserved interactions from the parent fragments for top 100 compounds identified using the Fragment Network merging pipelines versus pharmacophore-constrained docking.** EV, enterovirus.

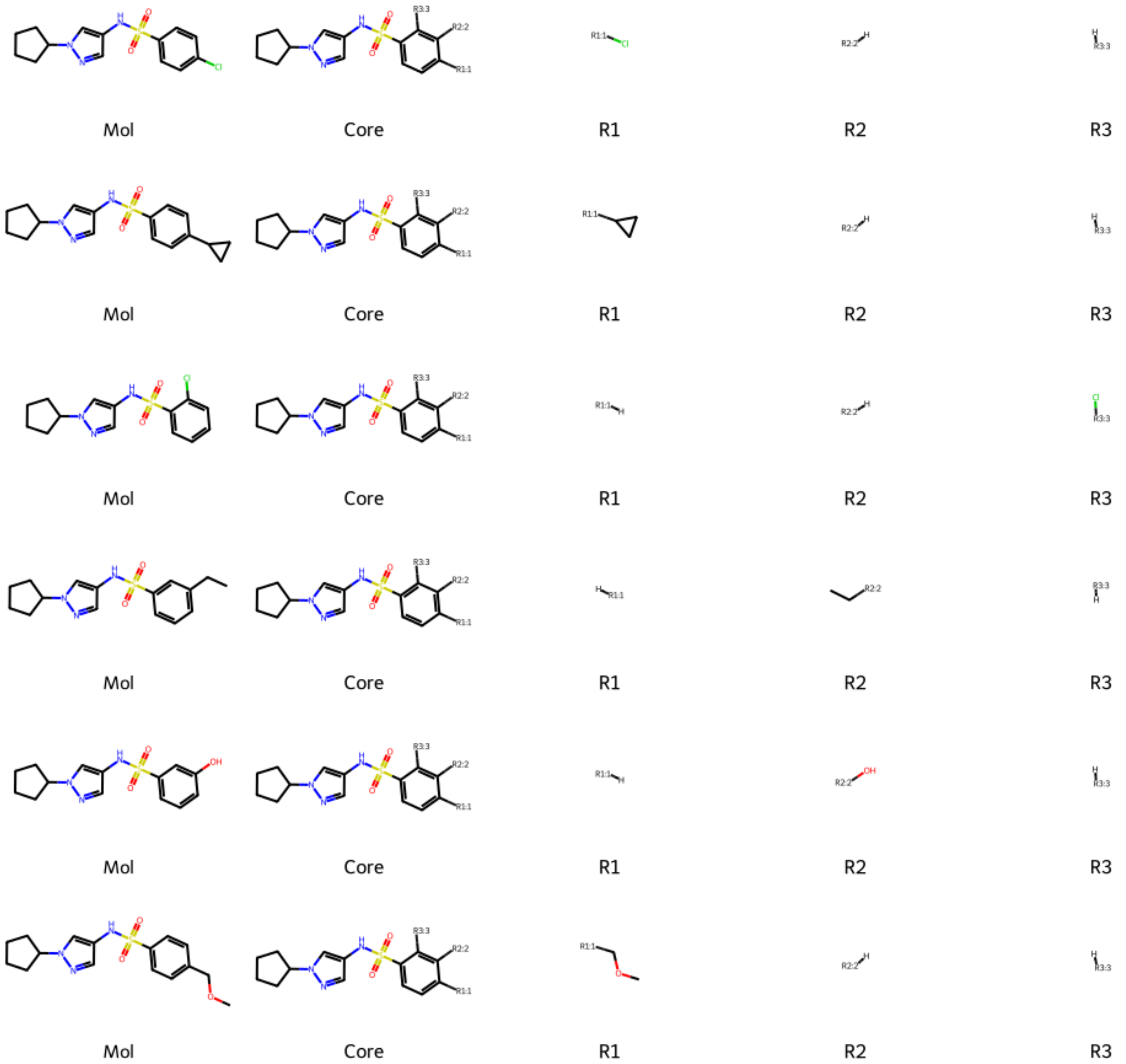
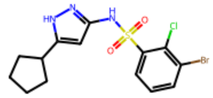
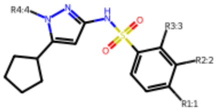


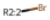
Figure B.11: **R-group decomposition for molecules similar to substituent 3d against WDR5–MYC.** Fragment Network-identified merges of fragments that bind to the WDR5–MYC complex. The identified merges incorporate the pyrazole ring identified for substituent 3d (manually designed in (Chacón Simon et al., 2020)) and also various substituents, which were outlined in the displayed R-group decomposition. Each row represents a different identified merge (1st column). The core molecule is shown in the 2nd column with the R-group attachment points labelled. The remaining columns show the R-groups added at each attachment point.


Mol



Core

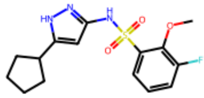
R1

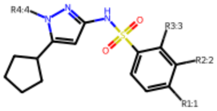


R2

R3

R4


Mol



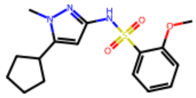
Core

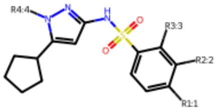
R1

R2

R3

R4


Mol



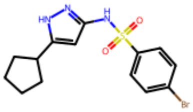
Core

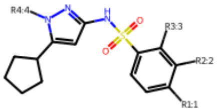
R1

R2

R3

R4


Mol



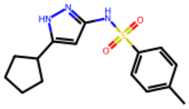
Core

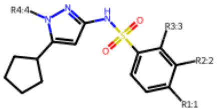
R1

R2

R3

R4


Mol



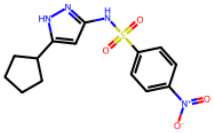
Core

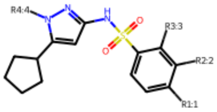
R1

R2

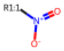
R3

R4


Mol



Core



R1

R2

R3

R4

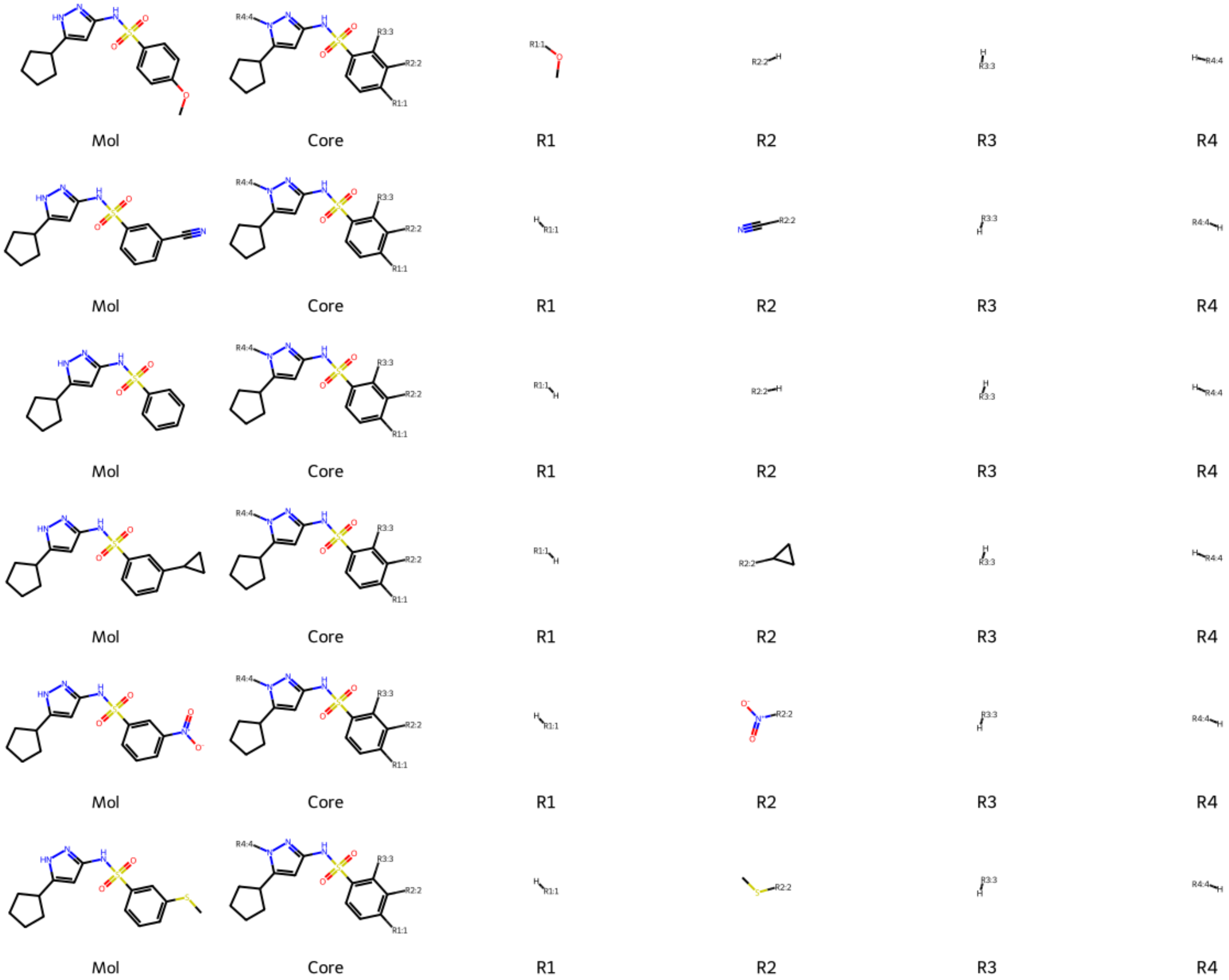


Figure B.12: **R-group decomposition for molecules similar to substituent 3b against WDR5–MYC.** Fragment Network-identified merges of fragments that bind to the WDR5–MYC complex. The identified merges incorporate the pyrazole ring identified for substituent 3b (manually designed in (Chacón Simon et al., 2020)) and also various substituents, which were outlined in the displayed R-group decomposition. Each row represents a different identified merge (1st column). The core molecule is shown in the 2nd column with the R-group attachment points labelled. The remaining columns show the R-groups added at each attachment point. The Figure is split across two pages.

C. Appendix items for Chapter 4

Table C.1: Selection criteria used during XChem campaigns

Target	Iteration	Selection criteria
EV D68 3C protease	1	—
	2	Butina clustering (Tanimoto threshold of 0.4) Manual selection
	3	Butina clustering (Tanimoto threshold of 0.4) $SC_{\mathrm{RDKIT}} \geq 0.4$ Manual selection
EV A71 2A protease	1	Butina clustering (Tanimoto threshold of 0.3) $SC_{\mathrm{RDKIT}} \geq 0.5$ RMSD $\leq$ 1.5A At least 2 interactions Mean overlap $\geq$ 55% Manual selection
	2	Butina clustering (Tanimoto threshold of 0.3) $SC_{\mathrm{RDKIT}} \geq 0.55$ RMSD $\leq$ 1.0A Maintain >50% interactions No more than 5 rotatable bonds Manual selection

EV, enterovirus.

D. Appendix items for Chapter 5

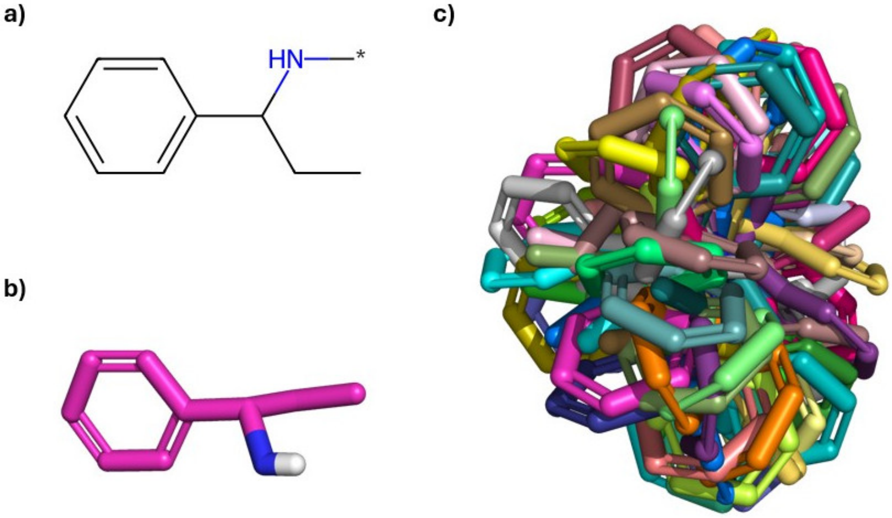


Figure D.1: **Example conformers generated for an elaboration.** (**a**) An example elaboration from the LibINVENT library (Fialková et al., 2022). (**b**) A single conformer generated for the elaboration. (**c**) A set of conformers that were rotated and clustered for the elaboration. All conformers are aligned to the same reference point (based on the coordinates of the growth vector) and rotated in 30° intervals. The resulting conformers were clustered using Butina clustering (Butina, 1999).

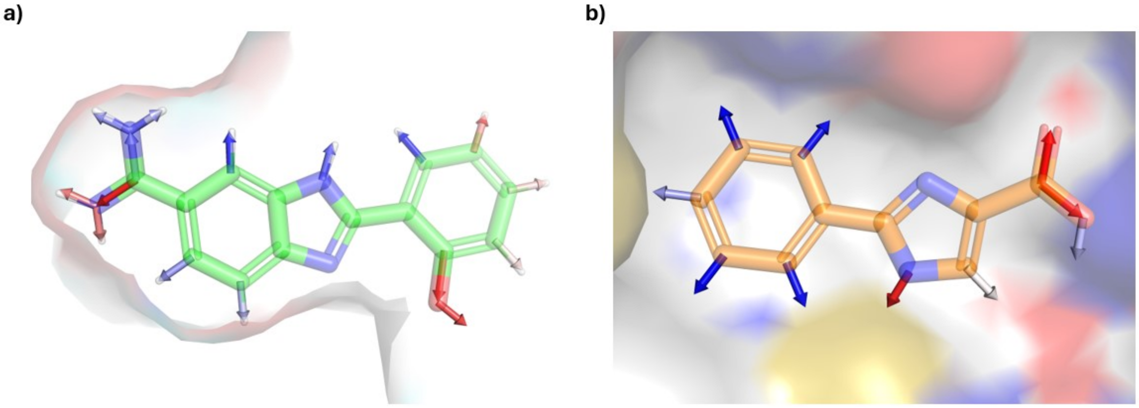
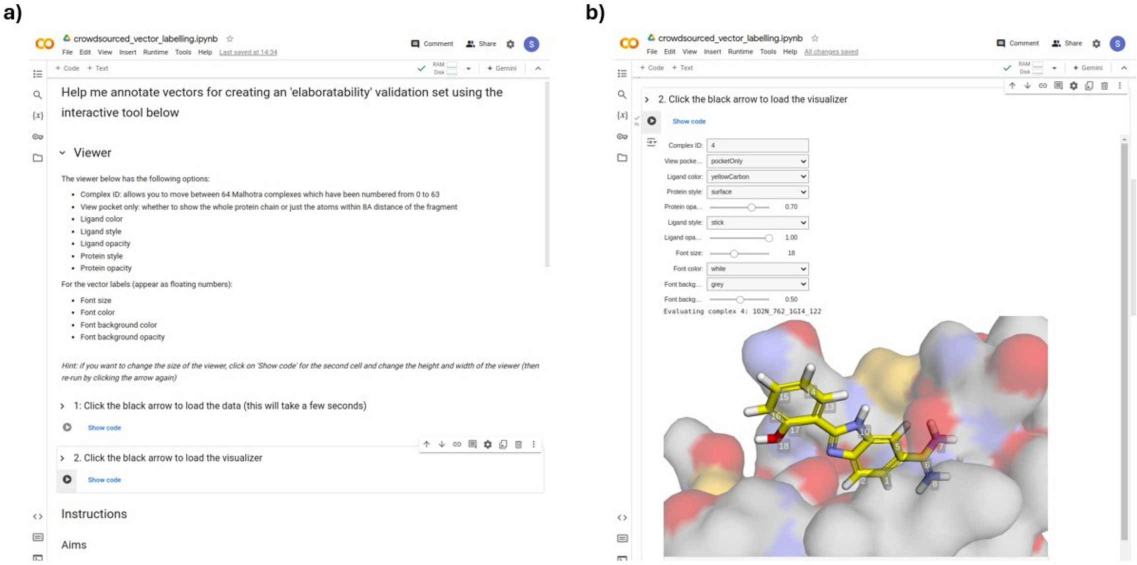


Figure D.2: **Example usage of AiZynthFinder for evaluating the tractability of growth vectors.** We used AiZynthFinder (Genheden et al., 2020) for evaluating the tractability of growth vectors for further synthesis. The examples show two molecules whereby all elaborations from our elaboration library (containing 311 elaborations) are added to the growth vector; the AiZynthFinder neural network is used to determine whether the reaction is feasible and the proportion of elaborations that can be successfully added is calculated. The arrows represent growth vectors are coloured using a blue–red spectrum from least to most 'reactive' (a greater proportion of elaborations can be added). The values have been normalized according to the most reactive vector.


a)
crowdsourced_vector_labelling.ipynb
File Edit View Insert Runtime Tools Help Last saved at 14:34
Comment
Share
S
+ Code + Text
RAM
Disk
Gemini
b)
crowdsourced_vector_labelling.ipynb
File Edit View Insert Runtime Tools Help All changes saved
Comment
Share
S
+ Code + Text
RAM
Disk
Gemini
15
14
13
16
17
18
10
5
7
6
2
1
8

a)

Help me annotate vectors for creating an 'elaboratability' validation set using the interactive tool below

Viewer

The viewer below has the following options:

- Complex ID: allows you to move between 64 Malhotra complexes which have been numbered from 0 to 63
- View pocket only: whether to show the whole protein chain or just the atoms within 8A distance of the fragment
- Ligand color
- Ligand style
- Ligand opacity
- Protein style
- Protein opacity

For the vector labels (appear as floating numbers):

- Font size
- Font color
- Font background color
- Font background opacity

Hint: if you want to change the size of the viewer, click on 'Show code' for the second cell and change the height and width of the viewer (then re-run by clicking the arrow again)

1: Click the black arrow to load the data (this will take a few seconds)

Show code

2. Click the black arrow to load the visualizer

Show code

Instructions

Aims

b)

2. Click the black arrow to load the visualizer

Show code

Complex ID: 4

View pocke...: pocketOnly

Ligand color: yellowCarbon

Protein style: surface

Protein opa...: 0.70

Ligand style: stick

Ligand opa...: 1.00

Font size: 18

Font color: white

Font backg...: grey

Font backg...: 0.50

```
Evaluating complex 4: 1O2N_762_1GI4_122
```

Figure D.3: **A molecule viewer to enable annotation of vector datasets.** A Google Colab notebook was created to aid annotation of vectors in a small dataset of protein–ligand complexes that could be used for validation of an elaboratability scoring tool. The tool was designed to be easy to use and accessible by non-computational scientists. (**a**) Instructions are provided on how to use the tool and load the viewer. (**b**) The molecule viewer, whereby vector atoms are labelled. The user inputs their annotations into a shared spreadsheet.

Table D.1: Percentage of vectors marked as elaboratable for the PDBbind synthetic dataset with different percentile thresholds.

Percentile thresholds	Score level	Train		Test	
		% GT$^+$ vectors	% other vectors	% GT$^+$ vectors	% other vectors
15–85	Clash score	70.0	56.0	69.7	52.1
	Clash + interaction score	66.8	52.4	67.6	48.9
	Clash + interaction + tractability score	65.2	48.9	65.4	45.2
10–90	Clash score	80.0	63.0	79.0	59.3
	Clash + interaction score	76.1	58.7	76.5	55.4
	Clash + interaction + tractability score	74.3	54.5	74.2	51.4
5–95	Clash score	89.5	73.5	91.2	71.6
	Clash + interaction score	84.9	67.4	88.1	66.1
	Clash + interaction + tractability score	82.7	62.1	85.0	61.3

Thresholds are varied for the clash score only. GT$^+$, positively labelled ground-truth.

Table D.2: Model parameters used for EGNN training

Parameter	Values
Batch size	**8**
Learning rate	0.01, **0.001**, 0.0001
Learning rate scheduler	None, Linear, **ReduceLROnPlateau**
Number of hidden node features	20, **30**, 40
Number of layers	2, **3**, 4
Loss calculation	Average over batch, average over complex, **average over molecule**
Loss function	**Binary cross entropy**
Activation function	**SiLU**

Bold indicates parameter used in final model.
EGNN, equivariant graph neural network.