Machine Learning Approaches for Computer Aided Drug Discovery

Marc Alexander Moesser

Green Templeton College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Trinity 2022

Statement of Originality

I, the undersigned, declare that this is my own work unless where otherwise acknowledged and referenced.

Candidate: Marc Alexander Moesser

Signed _____

Date 10.06.2022

Acknowledgements

Before I started my PhD, my expertise was in lab-based medicinal chemistry and organic synthesis. Although I had started to enjoy writing code in my free time before the start of my PhD, I was by no means well versed in data science or bioinformatics. Therefore, when I started my PhD, I set myself the goal to learn more computational drug discovery techniques to acquire a completely new set of skills, and was originally planned to only be a complementary part of my PhD. Three and a half years later, I now have a fully computational thesis. I've learned an immense amount and worked on methods and projects that I originally never thought I would, and I've really enjoyed the journey.

I therefore would like to thank my supervisors Garrett and Chris and my industrial supervisor Andrew, that enabled me to pursue projects I was passionate about, and gave me the freedom to explore completely new areas during my PhD. I believe this level of freedom is very rare and it has allowed me to stay positive and excited about my work throughout my entire PhD, I truly appreciate the trust and support you've all given me. Especially, I would like to say thank you to Garrett, who took a risk to accept a chemist with little coding experience into his group, and managed to turn him into a bioinformatician - although, as a mere mortal, I might never reach Garrett's level of command line magic proficiency.

Next, I would like to thank Charlotte and everyone in OPIG for providing a fun, welcoming and scientifically inspiring environment. In OPIG, I not only found great colleagues to talk about science, but many good friends that I already miss dearly. The Gloucester Green lunch crew (thanks Brennan for being the first person I met in OPIG that doesn't want to get miserable Tesco meal deals every day, you are my personal culinary hero), the Alpha Bar fanatics (cheese n' seeds enjoyers, I will not publicly shame you here, you know who you are), the Tuesday group meeting pub trips (Tom takes his social secretary responsibility serious, the chicken schnitzels at Royal Oak won't eat themselves), whiskey tastings at yours truly (Eve is a scotch-girl) and lots of other social events, it's been a fantastic and fun time. I hope someone maintains the meme wall, German engineering only works for so long.

Thanks to my friends at Oxford, back home, in Zurich or wherever you are in the world, I'm sorry I couldn't visit more often, I've got a long list of refunded plane tickets. I blame COVID, lethargy and the PhD, let's catch up soon!

Thank you to Susanne for always being there to talk about life, careers, friendship, everything. I use your Carbonara recipe weekly, thanks for keeping my cholesterol nice and high.

Especially, thank you to Kevin, who has gone through all the ups and downs of now over 10 years with me, thank you for always being there. Regardless of where

we are in life, I know we will always take time to chat, share stories, talk about life and build virtual logistics chains together, what more do you need?

Finally, thank you to my family for always supporting me. I would not be where I am today without your support. Thank you to my parents, Monika and Frank for always believing in me and allowing me to pursue my dreams - never compromising, always supportive. Thank you mum for pushing me to always do my best, whether at school, undergrad or PhD. Thank you dad for helping me to stay motivated with your trust and excitement for what I do. Thank you to Rüdiger for being a career role model, for being ambitious and for showing how much can be achieved with hard work and dedication. Thank you to my grandparents, Bernhard, Marianne and Elke. I cannot remember a time when you have not been supportive. Without you, I would not have been able to pursue my goals without worry. Thank you all, I'm extremely lucky to have you in my life.

Abstract

Pharmaceutical drug discovery is expensive, time consuming and scientifically challenging. In order to increase efficiency of the pre-clinical drug discovery pathway, computational drug discovery methods and most recently, machine learning-based methods are increasingly used as powerful tools to aid early stage drug discovery.

In this thesis, I present three complementary computer-aided drug discovery methods, with a focus on aiding hit discovery and hit-to-lead optimization. In addition, this thesis particularly focuses on exploring different molecular representations used to featurise machine learning models, in order explore how best to capture valuable information about protein, ligands and 3D protein-ligand complexes to build more robust, more interpretable and more accurate machine learning models.

First, I developed ligand-based models using a Gaussian Process (GP) as an easy-to-implement tool to guide exploration of chemical space for the optimization of protein-ligand binding affinity. I explored different topological fingerprint and autoencoder representations for Bayesian optimisation (BO) and showed that BO is a powerful tool to help medicinal chemists to prioritise which new compounds to make for single-target as well as multi-target optimisation. The algorithm achieved high enrichment of top compounds for both single target and multiobjective optimisation when tested on a well known benchmark dataset of the drug target matrix metalloproteinase-12 and a real, ongoing drug optimisation dataset targeting four bacterial metallo-β-lactamases.

Next, I present the development of a knowledge-based approach to drug design, combining new protein-ligand interaction fingerprints with a fragment-based drug discovery approach to understand SARS-CoV-2 M^{pro}-substrate specificity and to design novel small molecule inhibitors *in silico*. In combination with a fragment-based drug discovery approach, I show how this knowledge-based interaction fingerprint-driven approach can reveal fruitful fragment-growth design strategies.

Lastly, I expand on the knowledge-based contact fingerprints to create a ligand-shaped molecular graph representation (Protein Ligand Interaction Graphs, PLIGs) to develop novel graph-based deep learning protein-ligand binding affinity scoring functions. PLIGs encode all intermolecular interactions in a protein-ligand complex within the node features of the graph and are therefore simple and fully interpretable. I explore a variety of Graph Neural Network architectures in combination with PLIGs and found Graph Attention Networks to perform slightly better than other GNN architectures, performing amongst the best known protein-ligand binding affinity scoring functions.

Contents

List of Figures

List of Equations

Chapter 1

Introduction

1.1 Drug Discovery

Drug discovery is the interdisciplinary scientific process involving the identification, optimization and formulation of new pharmaceutical products. Modern drug discovery can typically be divided into six distinct steps, each with their own scientific challenges: 1) target discovery; 2) hit discovery; 3) hit-to-lead; 4) lead optimization; 5) *in vivo* activity, absorption, distribution, metabolism, and excretion (ADME) & toxicology optimization in animal models; and 6) human clinical trials (Figure 1.1).

During a typical target discovery project, new biological targets that could be used to treat a specific disease are identified. Often, understanding the role of the drug target in the underlying biological process of the disease is a key step in target discovery and validation. Once a suitable target has been identified, a hit compound needs to be identified to modulate the activity of the desired target. With the emergence of the field of biologics, therapies based on antibodies, enzyme therapies, and other proteins are increasing the scope of what constitutes a drug. However, for the purpose of this thesis, when discussing drugs and the drug discovery process more broadly, I am referring to small molecule drugs specifically.

The identification of a small-molecule hit compound is traditionally done through a high-throughput screen (HTS), where hundreds of thousands of small molecules

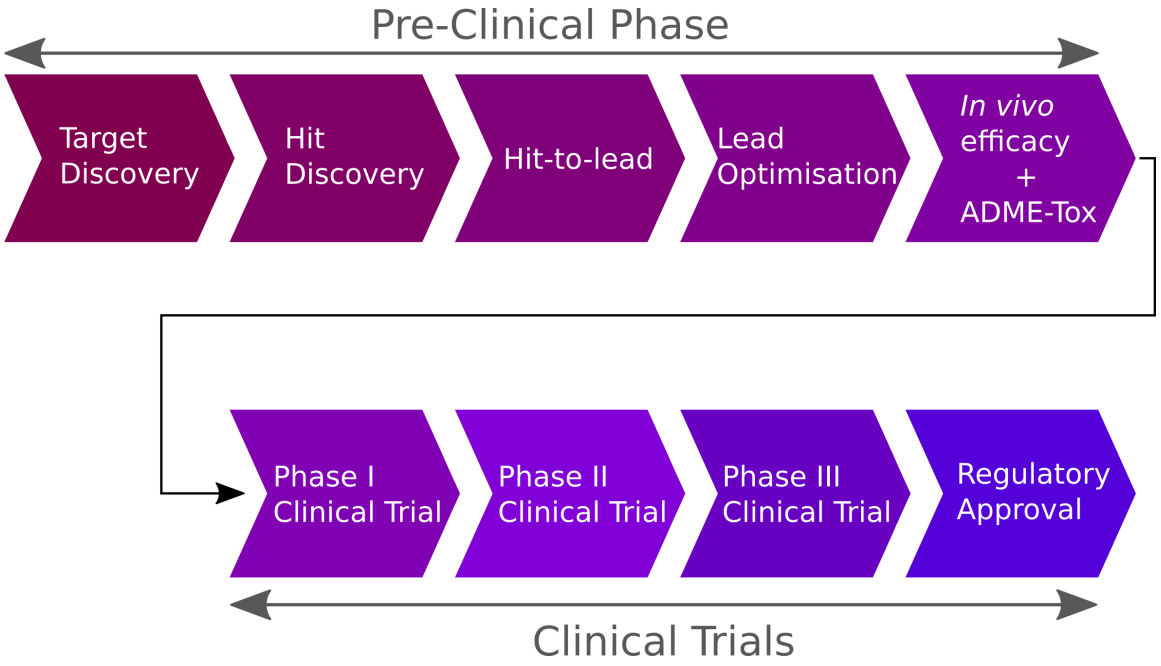

Figure 1.1: The steps of drug discovery: 1) target discovery, 2) hit discovery, 3) hit-to-lead, 4) lead optimization, 5) *in vivo* efficacy & ADME-toxicology studies, 6) human clinical trials, divided into 3 clinical trial phases. Steps 1-5 are considered "pre-clinical". This thesis focuses on the development of methods targeting steps 2-4.

are screened against the protein target with the goal of identifying compounds that modulate the activity of the target [Martis et al., 2011]. After identifying the initial hit compounds, their physical properties and biological activity needs to be optimised in a process called lead optimization, or Structure-Activity Relationship (SAR) optimization. As the last step in the pre-clinical process, lead compounds are further optimised for favourable absorption, distribution, metabolism, and excretion (ADME) properties as well as their toxicology profile *in vitro* and then later on, *in vivo*. This optimization is often done in parallel to the SAR optimization before lead compounds are able to be tested in clinical trials.

Each stage in this process is scientifically challenging, time consuming and expensive. Historically, pharmaceutical R&D costs for each drug brought to market have been increasing rapidly from around $100M in 1975 to $1.3B in 2005 [Roy, 2012] and even further in recent years to the current estimates of $1.3B - $2.8B [DiMasi et al.,

2016; Wouters et al., 2020]. Two of the main drivers for increased costs are the rising complexity of Phase III clinical trials, and the increased failure rate of new drug trials [Roy, 2012; Cook et al., 2014]. The high failure rate has been mainly attributed to the failure of drugs to show sufficient efficacy in the clinic, as well as failing safety and toxicity thresholds [Cook et al., 2014].

The increased cost and burden of proof required by regulators for clinical trials [Roy, 2012] in combination with the rising failure rate [Cook et al., 2014] underlines the necessity that only the best possible drug candidates should be taken forward into clinical trials. Traditionally, the hit identification, the hit-to-lead, and the lead optimization stages have been dominated by subjective decision making driven by the leading individual medicinal chemists' synthetic intuition and their biases. While medicinal chemists often agree on a high level about which features are desirable in lead compounds, the relative weighting and preferences placed on each feature varies between chemists, and some disagree completely on which properties are considered desirable [Kutchukian et al., 2012]. This leads to compounds reaching clinical trials that would be considered promising by one group of medicinal chemists, while being regarded as problematic by others. Standardising this subjective process through the introduction of computational and objective scoring functions is therefore crucial for the advancement of drug discovery. For this reason, the work described in this thesis focuses on the development of computational methods for the hit discovery and SAR optimization stages to improve efficiency and introduce novel computational approaches to aid medicinal chemists in making better, more objective decisions when choosing hit compounds and optimising them.

1.1.1 Hit Discovery & Hit-to-Lead

The objective during the hit discovery process is often the identification of small molecule inhibitors for the desired drug target from a large library of compounds (typically a corporate compound collection) in the shortest time possible. A widely-used method for this purpose is high-throughput screening (HTS). This single-target approach yields potential inhibitors that show activity either in an automated physical assay or more recently, by using high-throughput virtual screening (VS) methods such as protein-ligand docking [Rester, 2008] which at present are able to screen compounds *in silico* with greater speed, but lower accuracy than lab-based HTS methods. Furthremore, most recent ultra-large library screening methods can even enable the virtual screening of billions of compounds [Sadybekov et al., 2022; Gorgulla et al., 2020]. Additionally, thanks to the introduction of robots into biochemistry laboratories, tens to potentially even hundreds of thousands of compounds can be assayed in a single day, drastically increasing the number of compounds that can be physically screened in a single campaign [Martis et al., 2011]. An alternative approach to HTS is fragment-based drug discovery (FBDD), which is discussed in Section 1.1.4.

While HTS suffers from both type 1 (false positives) and type 2 (false negatives) errors, the major challenge for HTS and VS experiments is the separation of false positives from true positives [Martis et al., 2011]. Due to the extremely large libraries employed, screening campaigns usually suffer from an overabundance of hit compounds, where the elimination of type 2 errors would not add much value in comparison to the identification of type 1 errors. One method for the reduction of errors is the identification of compounds that are potential Pan Assay Interference Compounds (PAINS) through a substructure search algorithm [Baell and Holloway, 2010]. Compounds are usually subjected to PAINS filters before they enter a HTS database or before they are taken forward to the next stage of development. The

process of filtering from the initial hits of the screening campaign down to the true positives is referred to as the "hit-to-lead" process. A sense of the relative scale in terms of compound throughput of each step in the process is shown in Figure 1.2.

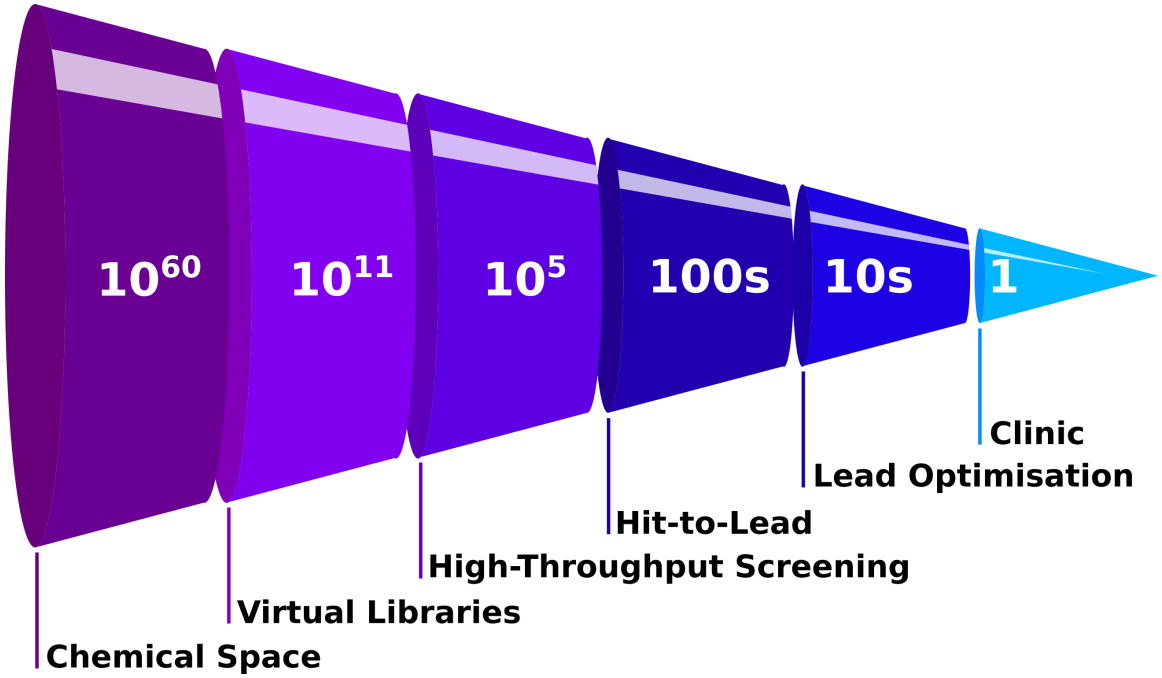

Figure 1.2: General overview of the number of compounds employed at each step of a typical modern drug discovery process. From the entire chemical space of drug-like compounds, huge virtual libraries are created and filtered to create libraries of hundreds of thousands of compounds during high-throughput screening, from which promising compounds are selected and derivatised to create hundreds of compounds. Tens of lead compounds are then selected from which one is chosen for clinical trials.

During the hit-to-lead process, medicinal chemists typically choose from the set of verified hits obtained from HTS to build a portfolio of lead compounds. At this stage, potential lead compounds are evaluated using a series of metrics such as their "drug-likeness", a loosely defined set of properties to compare how similar potential lead compounds are to existing drugs based on the assumption that higher similarity increases the likelihood of potential lead compounds to become approved drugs themselves. In addition, other factors such as toxicity, cell permeability and sometimes even performance during early *in vivo* mouse studies are also considered.

One well known set of properties to evaluate drug likeness of a molecule is the "Lipinski Rule of Five" [Lipinski et al., 2001], which states that orally active drugs should not violate any of the following rules: 1) no more than 5 hydrogen bond donors; 2) no more than 10 hydrogen bond acceptors; 3) a molecular mass of less than 500 Daltons and 4) a base 10-logarithm of the octanol-water partition coefficient below 5. However, studies have shown that many drugs do not follow the Lipinski Rule of Five and other metrics such as ligand efficiency might be more helpful in prioritising early leads [Hopkins et al., 2014]. Overall, this process suffers from a high reliance on the subjective biases of individual medicinal chemists and a lack of objective, widely applicable metrics.

1.1.2 Lead Optimization

Once a set of lead compounds has been identified in the hit-to-lead process, lead optimization begins. A lead compound needs to be optimised for several pharmacological properties including potency, solubility, stability, cost-effectiveness, ADME, and lack of toxicity. Lead optimization is therefore a complex multi-objective optimization problem, which is traditionally solved through highly iterative design-test-design cycles (a.k.a. the "design, make, test and analyse" framework or DMTA [Plowright et al., 2012] as seen in Figure 1.3).

For example, based on the lead compound(s), new derivatives may be designed to bind to the drug target more tightly, and a hypothesis about its effect is formulated. The chosen derivatives are then synthesised, purified and tested in the laboratory and the results inform future design decisions. Traditionally, this structure-activity relationship study is highly empirical and the compound is designed based on the experience (and subjective bias) of the lead medicinal chemist. Typically, a scaffold is identified as the core of the lead compound, and peripheral "R-groups" varied to

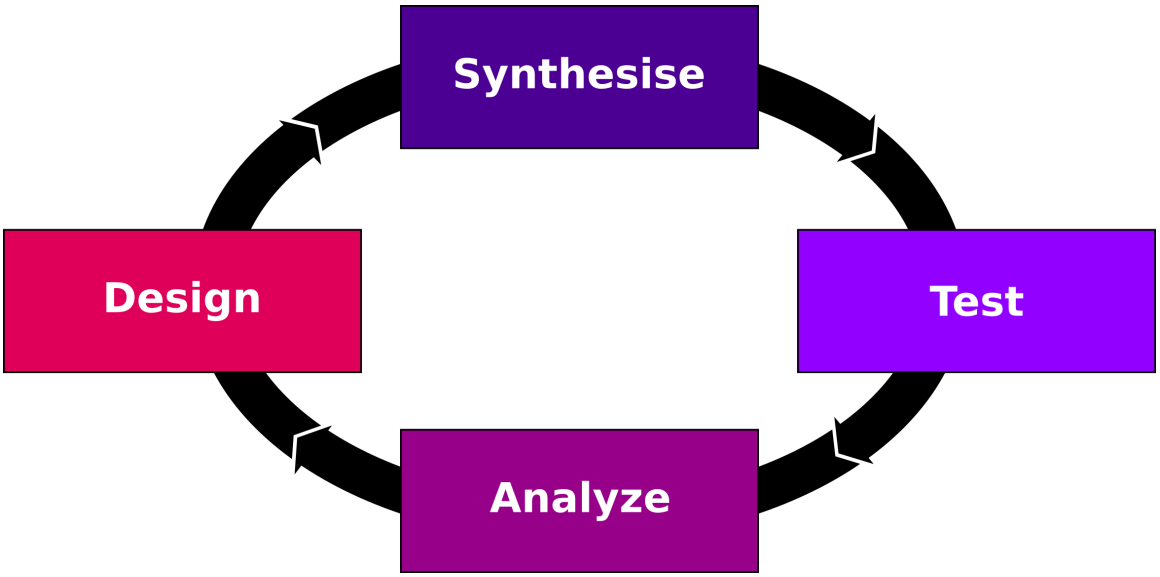

Figure 1.3: Overview of the design, make, test, analyse (DMTA) cycle. Drugs are designed based on a working hypothesis to improve its properties. It is then synthesized and tested in the lab. After analysis, the design hypothesis is adjusted based on the new information and the next generation of compounds designed accordingly.

explore an optimal combination. Even if optimization towards one of the desired properties, such as highly potent binding affinity, is straightforward, the balancing of the multi-objective optimization problem is much more challenging. For example, a protein binding pocket might be highly hydrophobic, requiring a more hydrophobic ligand which might result in less favourable solubility and ADME properties.

1.1.3 Selectivity versus Polypharmacology

Drug selectivity refers to the ability of a drug compound to bind primarily to the drug target of interest without significant binding affinity to other, undesired off-targets. One of the primary reasons for the optimization of protein-ligand binding affinity against a drug target is to reduce the dose of the drug required for biological activity, making the drug easier to take for the patient while also helping to avoid toxicity and unwanted side-effects.

Traditionally, in 20^{th} century drug discovery, phenotypic screening approaches

were used to identify promising drug candidates [Moffat et al., 2017], such as for example the discovery of Penicillin in 1928 by Alexander Fleming, who identified that certain compounds present in mold stop bacteria from growing, without knowing what the mechanism of action might be. However, phenotypic screening suffers from major challenges in hit validation and especially in target deconvolution. As a result, modern drug discovery approaches favor the single-target approach, which has now been widely adopted. However, as recent studies have shown, many drugs that were previously considered to be specific single-target drugs have subsequently been shown to derive their activity from polypharmacology, *i.e.* activity against more than one drug target by the same drug molecule [Paolini et al., 2006; Boran and Iyengar, 2010].

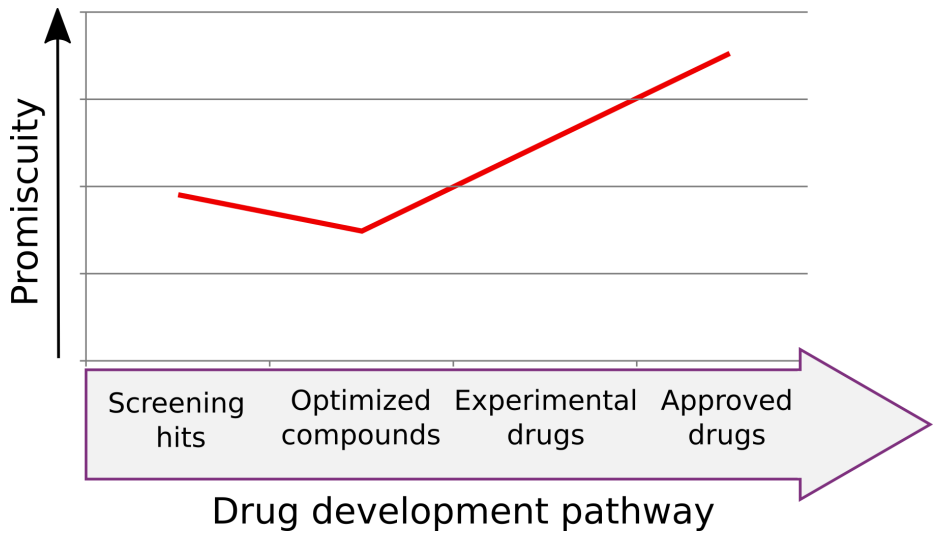

Figure 1.4: Overview of the relative promiscuity rates of small molecule drug compounds along the drug discovery pathway. Figure adapted from its original publication by Hu et al. [2014] under the open access CC BY 3.0 license.

As a result, the lead optimization process that has previously focused on selectivity and single-target activity has to be adapted to a new paradigm. One characteristic that has historically been considered to be a negative trait in lead compounds is binding promiscuity, which is reflected in the reduction of promiscuity of drugs at

the lead optimization stage (Figure 1.4, as published by Hu et al. [2014]). However, approved drugs were found to be more promiscuous, indicating that single-target approaches might not be as successful in the clinic as polypharmacology. Nonetheless, the addition of another dimension to the lead optimization problem increases the difficulty and the need for more objective evaluation methods that can simultaneously optimise desired properties as well as polypharmacological patterns.

1.1.4 Fragment-Based Drug Discovery

As an alternative to the classical small molecule drug discovery approach, where full sized drug-like molecules are identified as hits and derivatives designed to optimise their properties, fragment-based drug discovery (FBDD) has now become a mainstream method, with over 30 fragment-derived drug candidates in clinical trials and two fragment-derived drug approvals [Erlanson et al., 2016]. Instead of relying on huge HTS libraries with millions of drug-like compounds to generate hits, FBDD starts with significantly smaller libraries of several thousand drug fragments (typically compounds with fewer than 20 heavy atoms) that have been carefully chosen [Erlanson et al., 2016]. Just like in HTS campaigns, library design is very important for a successful FBDD campaign. One disadvantage of HTS libraries is that HTS is limited by its small coverage of chemical space. Although HTS libraries containing millions, or even hundreds of millions, of compounds might seem significant, the number of possible drug-like molecules has been estimated to be around 10^{63} [Erlanson et al., 2016], rendering even a large 10 million compound library insignificant by comparison. Since this explosion in complexity is driven by the number of atoms in a molecule, fragment libraries that contain molecules only half the size of drug molecules are able to cover a much greater percentage of chemical space with much smaller libraries. For example, according to [Ruddigkeit et al., 2012], the chemi-

cal space of fragment-sized molecules of around 17 heavy atoms encompasses only about 166 billion possible molecules, allowing a fragment library with thousands of compounds to cover a higher fraction of chemical space than the huge HTS libraries.

The hit-to-lead process in FBDD also differs from a traditional campaign. Instead of a typical SAR optimization where small changes are made to drug-like molecules, fragments alone are too small and their protein-ligand binding affinity often too weak (and too promiscuous) to be promising lead compounds by themselves [Erlanson et al., 2016]. Instead, two or more fragments with distinct binding modes in the protein binding pocket might be combined to form a full-sized inhibitor. Alternatively, the fragment might be grown considerably through synthetic addition of new functional groups. Both approaches require a detailed understanding of the 3D binding mode of the fragments in the binding pocket of the protein. FBDD has therefore profited tremendously from the recent advances in structural biology such as the development of cryo-EM pioneered by Jacques Dubochet, Joachim Frank and Richard Henderson (Nobel Prize 2017) and the development of high-throughput X-ray crystallography. For example, the high-throughput fragment-based screening campaign by the XChem facility at the Diamond Light Source in Oxfordshire was able to screen around 1500 fragments crystallographically in the span of just three weeks to identify fragment hits against SARS-CoV-2 Main Protease, a promising drug target against COVID-19 [Chodera et al., 2020]. In addition, structure-based computational methods such as protein-ligand binding affinity scoring functions and molecular docking (see Section 1.2.2) as well as other machine learning-based methods such as *de novo* fragment-growth methods [Imrie et al., 2021] have been instrumental in advancing the field of fragment-based drug discovery.

1.2 Computer-Aided Drug Design

In order to tackle the issue of subjective bias in drug design as discussed in Section 1.1, many different computer-aided drug design (CADD) methods have been developed and are now widely used throughout all stages of the drug discovery pathway [Yu and MacKerell Jr, 2017]. Broadly, CADD can be divided into structure-based drug design (SBDD) and ligand-based drug design (LBDD). SBDD methods include the analysis of 3D structural information such as X-ray or cryo-EM crystal structures of proteins and protein-ligand complexes to identify potential binding sites and to optimise the binding affinity of the ligand to the protein. LBDD focuses on the structure-activity relationship of the ligand itself, analysing which atoms in the ligand are more or less useful for binding and to fine-tune physical properties as well as pharmacokinetics. However, both approaches can be used in tandem, for example in modern machine learning models that use features derived from both active ligands and 3D structural data about the target.

1.2.1 Quantitative Structure-Activity Relationship (QSAR) Modeling

Today, Quantitative Structure-Activity Relationship (QSAR) models are firmly established as powerful predictive tools in pharmaceutical drug discovery. Many consider the founding of the field to be the publication of the SAR study of plant-growth regulators by Hansch et al. [1962]. Since then, QSAR models have significantly increased in complexity and are applied to a broader range of modeling tasks. The ultimate goal of QSAR models is to build a scoring function that can accurately predict the properties or biological activity of a molecule given its structure. While the first QSAR models have been limited to modeling linear relationships, for example between the base 10-logarithm of the water-octanol partition coefficient (log P) and

biological activity [Fujita et al., 1964], modern QSAR models employ sophisticated machine learning models using topological fingerprints [Rogers and Hahn, 2010] or even a combination of ligand and protein-derived descriptors for example in the field of Proteochemometrics (PCM) modeling [Cortés-Ciriano et al., 2015].

One major development in the field of QSAR modeling is the introduction of chemical descriptors as input features for QSAR models [Cherkasov et al., 2014]. Chemical descriptors can range from 1D descriptors (such as counts of different atom types), to the most popular 2D representation (the topological representation of molecules) to full 3D (molecular conformations) models [Cherkasov et al., 2014] and alignment-based comparative models such as Comparative Molecular Field Analysis (CoMFA) [Cramer et al., 1988].

The topological (2D) representation of molecules for QSAR models is the most commonly used chemical descriptor, since 1D descriptors often do not carry enough information to be useful, and 3D descriptors are limited by the accuracy of conformer generation software and choice of conformation. Descriptors derived from the topological representation of molecules can vary from molecular graph representations used in QSAR as early as 2000 [Ivanciuc, 2000] to lists of descriptors derived from the ligand structure such as the number of hydrogen bond donors and acceptors, the molecular weight or the topological polar surface area (TPSA) which are implemented for ease of use in QSAR models in popular cheminformatics toolkits such as RDKit [Landrum et al., 2006] which was used extensively in the work described in this thesis.

While the first, simple QSAR models were designed to accelerate drug discovery through automated compound evaluation, more recent advances have moved on beyond simple QSAR models. Classical QSAR models face a series of limitations [Cherkasov et al., 2014] such as the poor generalisability often driven by overfitting on the training data; the use of confounded descriptors in QSAR models, especially in

2D models where collinearity of descriptors is common; and finally the use of uninterpretable descriptors. There are thousands of different chemical descriptors that can be calculated with popular cheminformatics tools such as RDKit allowing easy access to hundreds of descriptors within a single module [Landrum et al., 2006], including descriptors that are unintuitive and where no clear physical-chemical interpretation available.

Alternatives to QSAR modeling exist, including three other major computational techniques that aim to resolve the limitations of classical QSAR models through different, complementary approaches: molecular docking, molecular dynamics (MD) simulations and modern machine learning-based scoring functions. While molecular docking software such as AutoDock 4 [Morris et al., 2009] employs semiempirical scoring functions to score large numbers of 3D protein-ligand complexes rapidly, molecular dynamics relies on more accurate but computationally expensive molecular force fields such as the modern AMBER force field [Weiner et al., 1984] to investigate molecular flexibility and stability of protein-ligand complexes. The most recent advances have been made in the field of machine learning scoring functions, where structure-based machine learning models are built with the goal of capturing non-linearity that linear classical scoring functions described above cannot. Ultimately, recent advances in ML-based scoring function design (Section 1.3) has focused on the development of generalisable models that can learn from the biophysics of interactions in protein-ligand complexes. For that purpose, a large number of different representations are being explored, ranging from 3D 3D protein-ligand interaction-derived fingerprints [Wójcikowski et al., 2018] to voxelised representations of the protein-ligand complex for convolutional neural networks (CNNs) [Jiménez et al., 2018] and molecular graph-based neural networks (GNNs) [Lim et al., 2019]. In all cases, 3D structural information of the protein-ligand complex is used to create the best performing models.

All three of these approaches (docking, MD and machine learning scoring functions) are featured extensively throughout this thesis (with novel methods being developed in each chapter) and each is described in more detail in Sections 1.2.2, 1.2.3 and 1.3, respectively.

1.2.2 Docking & Scoring

With advances in structural biology and the increased number of protein crystal structures available, the field of 3D protein-ligand docking emerged in the 1980s pioneered by Kuntz et al. [1982]. Instead of using the traditional 2D descriptor-based QSAR models as described above (Section 1.2.1), docking programs have been created to rapidly explore different geometrically feasible alignments between the ligand and protein in 3D and subsequently score the docked pose. Since the 1980s, major advances in protein-ligand docking have been made and a large range of different docking software is available today, ranging from commercial software such as GOLD [Jones et al., 1997] or Glide [Friesner et al., 2004] to the widely adopted open source docking tools of the AutoDock suite with the most recent versions of AutoDock 4 [Morris et al., 2009] and AutoDock Vina [Trott and Olson, 2010].

In general, modern docking software is composed of two major component: one or more search methods, and one or more scoring functions [Kitchen et al., 2004]. Since the AutoDock suite is currently one of the most popular software suite used for docking, a more detailed description of docking will be given using the example of AutoDock 4. The search method explores the translational, orientational, and conformational space of the protein-ligand complex, placing the ligand in different poses into the binding site of the target protein, and evaluating their score. For example, modern docking software such as AutoDock 4 use a Lamarckian Genetic Algorithm [Morris et al., 1998] to guide pose exploration (with the option to explore

protein sidechain slexibility as well).

The docking scoring function is used to approximate the free energy of binding, ΔG, of a given pose in complex with the receptor. Scoring function development in itself is a large field, including development of semiempirical, physics-based scoring functions [Kitchen et al., 2004] as well as more recent approaches exploring machine learning scoring functions such as GNINA [McNutt et al., 2021].

Currently, these novel machine learning scoring functions have not been incorporated into mainstream docking tools. The scoring function used by AutoDock4 [Huey et al., 2007] is a semiempririal force field that includes a pairwise evaluation of intramolecular interactions of the ligand ($V^{\text{L-L}}$) and the protein ($V^{\text{P-P}}$), intermolecular interactions between the protein and the ligand ($V^{P\text{-}L}$) and an estimation of the conformational entropy lost upon binding (ΔS_{conf}). The equation is shown below:

$$\Delta G = (V^{\text{L-L}}_{\text{bound}} - V^{\text{L-L}}_{\text{unbound}}) + (V^{\text{P-P}}_{\text{bound}} - V^{\text{P-P}}_{\text{unbound}}) + (V^{P-L}_{\text{bound}} - V^{\text{P-L}}_{\text{unbound}}) + \Delta S_{\text{conf}} \quad (1.1)$$

It is assumed that protein and ligand are sufficiently far apart in the unbound state that $V^{\text{P-L}}_{\text{unbound}}$ is zero. All pairwise atomic terms (V) are calculated as follows:

$$\begin{aligned} V = W_{\text{vdw}} \sum_{i,j} \left(\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} \right) + W_{Hb} \sum_{i,j} \left(\frac{C_{ij}}{r_{ij}^{12}} - \frac{D_{ij}}{r_{ij}^{10}} \right) \\ + W_{el} \sum_{i,j} \frac{q_i q_j}{\epsilon(r_{ij}) r_{ij}} + W_{sol} \sum_{i,j} (S_i V_j + S_j V_i) e^{(-r_{ij}^2/2\sigma^2)} \end{aligned} \quad (1.2)$$

Each term is calculated as the sum over all pairs of ligand atoms, i, and protein atoms, j. The equation overall includes four terms covering dispersion/repulsion, hydrogen bonding, electrostatics and desolvation, respectively. The first term is a typical 12-6 Lennard-Jones potential with parameters A_{ij} and B_{ij} from the AMBER

force field [Weiner et al., 1984]. The second term covers directional hydrogen bonding-based on a 12-10 potential [Goodford, 1985] with parameters C_{ij} and D_{ij} as well as directionality of the hydrogen bond inspired by the work of Wade et al. [1993] Electrostatic interactions are covered in the third term through a Coulombic potential between partial atomic charges q_i and q_j, and features a sigmoidal distance-dependent dielectric function $\epsilon(r_{ij})$ described by Mehler and Solmajer [Mehler and Solmajer, 1991]. AutoGrid4 uses the Coulombic potential to calculate electrostatic-interaction energy grid maps necessary for docking. Lastly, the desolvation potential is calculated in the fourth term using atomic fragmental volume (V) and solvation parameters (S), distance (r_{ij}) and a weighting factor σ.

Despite the advancements made in molecular docking and its widespread use throughout drug discovery, there are a series of outstanding challenges [Wang and Zhu, 2016]. Importantly, current docking protocols suffer from inaccurate scoring functions and the extremely large search space that arises when considering protein flexibility. Docking scoring functions are evaluated on the basis of four metrics [Li et al., 2018; Su et al., 2019]: (i) scoring power (accuracy in predicting the negative base-10 logarithm of the dissociation constant K_d (or inhibition constant K_i) of a protein-ligand complex); (ii) ranking power (accuracy in ranking known ligands for a single protein from best to worst binder); (iii) docking power (accuracy of finding the native ligand binding pose of a given protein-ligand complex); and (iv) screening power (ability to find true binders for a single protein target amongst a dataset of random molecules).

Docking scoring functions have been shown to struggle in the scoring power test (the Pearson correlation coefficient between the predicted and true binding affinity in the range of 0.21–0.63 for all docking software tested in CASF 2016 [Su et al., 2019]) and ranking power test. Conversely, the performance in predicting the native

binding pose (docking power) as well as in their screening power when tested against the CASF-2016 benchmark [Su et al., 2019] is high. The inaccuracy in predicting the absolute binding constant leads to downstream effects, weakening the other performance metrics. Since poses are ranked based on their predicted binding constant, inaccurate scoring functions could lead to inaccurate ranking. In addition, when evaluating screening power, poses have to be classified as binder *vs* non-binder based on an arbitrary threshold, often defined empirically for each protein target, rather than an absolute threshold. As a result, ongoing research into the development of more accurate scoring functions is crucial for the advancement of the field of molecular docking, with most recent advances in the field of machine learning, where deep learning-based scoring functions have been integrated into molecular docking directly [McNutt et al., 2021] and novel approaches for docking rescoring have been reported [Zhong et al., 2010].

An additional challenge is the modeling of water molecules that might be involved in ligand binding during the docking process. While the position of crystallographic water molecules is modelled by crystallographers and provided in protein-ligand co-crystal structures, the inclusion of water molecules into docking is challenging and most standard docking procedures remove water (and all other solvent molecules) before docking. In order to accurately score hydrated protein-ligand complexes, specialised force-fields such as the one developed by Forli and Olson [2012] are necessary that specifically take the entropic and enthalpic contributions of discrete water molecules to the overall binding energy of the complex into account.

Finally, a last major challenge is the treatment, or lack thereof, of protein conformational flexibility in docking methods. While there are methods that model flexible protein side chains during docking (such as in AutoDock 4 [Morris et al., 2009]), generally, high-throughput virtual screening experiments only allow conformational

flexibility on the ligand side, while considering the protein receptor to be rigid during docking [Wang and Zhu, 2016]. This reduces the search space of possible combinations of ligand-receptor conformations and allows docking to rapidly screen hundreds of thousands-, or in the case of recent GPU accelerated tools such as AutoDock GPU developed by the Forli group [Santos-Martins et al., 2021], millions of compounds. While throughput is increased, accuracy is decreased in comparison to detailed (and often knowledge-based), flexible docking-based approaches, presenting a challenging problem when considering the accuracy *vs* simplicity trade off. In addition, for novel systems, it is often unknown which amino acids in the binding site should be considered to be flexible, and allowing every residue to be flexible would not be feasible as it increases the possible search space explosively.

As one approach to restrict conformational degrees of freedom, modern docking tools such as AutoDock 4 [Morris et al., 2009], AutoDock Vina [Trott and Olson, 2010], GOLD [Jones et al., 1997] and Glide [Friesner et al., 2004] allow the user to constrain certain parts of the ligand, for example through covalent docking, which reduces the conformational degrees of freedom as well as eliminates the translational and orientational degrees of freedom of the ligand through covalent attachment to a protein residue, leading to more accurate poses. However, more complex constrained docking where the position of entire substructures of a ligand are constrained have only been directly implemented in commerical docking tools such as GOLD [Jones et al., 1997] and Glide [Friesner et al., 2004]. As a result, to increase availability and thus progress the field, there is a big need for the implementation of more sophisticated constrained docking protocols for the most popular open-source docking tools AutoDock 4 [Morris et al., 2009] and AutoDock Vina [Trott and Olson, 2010].

In Chapter 3 of this thesis, I present the development of a fragment-based *active-guided covalent docking* protocol implemented using AutoDock 4, with the goal of

utilising as much previous knowledge about the ligand and protein target as possible during docking, such as the induced-fit conformation of known protein-ligand complexes and covalent constraints to reduce docking inaccuracy without having to consider receptor flexibility explicitly. While this method currently does not include substructure-based constraints, it is a first step towards the implementation of a comprehensive open-source constrained docking methodology and future work on this project should include the expansion of the current covalent docking constraints to substructure-based constraints for the use in fragment-based drug discovery.

1.2.3 Molecular Dynamics

As an additional tool that is increasingly used in computational drug discovery, molecular dynamics (MD) is at the other end of the spectrum compared to QSAR modeling when it comes to the accuracy *vs* speed trade off. While classical QSAR models, and to some extent high-throughput virtual screening campaigns using docking, try to maximise throughput while sacrificing accuracy, MD is using computationally expensive and highly parameterized force fields to calculate the forces between atoms to calculate the overall energy of the system [De Vivo et al., 2016].

Instead of screening hundreds of thousands of compounds, MD simulations are carried out for tens of compounds at a later stage of lead optimization to give detailed information about the binding pose, energy contributions of each atom and potential induced fit effects to guide structure-based drug design [De Vivo et al., 2016]. This high accuracy approach is computationally expensive, processing a single protein-ligand complex in hours or days instead of minutes or even seconds for protein-ligand docking tools (depending on the computational resources available).

One example of the usefulness of MD in computational drug discovery are all-atom simulations that can be used to obtain more accurate estimates of the absolute

free energy of binding for drug molecules, outperforming classical scoring functions such as described above (Section 1.2.2) during docking when it comes to accuracy [Aldeghi et al., 2016]. In addition, MD can also be a powerful tool for evaluating the dynamic aspect of ligand binding, to allow detailed contact analysis for drug design [Chan et al., 2021a]. I describe one such application in Chapter 3 where MD was used to study substrate specificity and to create high quality substrate models to guide *in silico* design of SARS-CoV-2 Main protease inhibitors.

1.3 Machine Learning in Drug Discovery

The emergence of ML methods and data-science approaches throughout all scientific disciplines over the last decades has transformed the chemical and biological sciences and the drug discovery process specifically, with ML methods used throughout the entire drug discovery pipeline [Greener et al., 2022]. The work described in this thesis focuses on the development of ML methods targeting the improvement of pre-clinical small molecular drug discovery, with a particular focus on hit discovery and hit-to-lead optimisation. For these applications, ML methods focus primarily on generative methods for ML-guided *de novo* design [Popova et al., 2018; Gómez-Bombarelli et al., 2018] and the creation of scoring functions that are able to predict a range of different properties, such as protein-ligand binding affinity or absorption, distribution, metabolism, elimination and toxicology (ADMET) properties [Wenzel et al., 2019; Jiang et al., 2020a], as well as physical properties such as solubility [Boobier et al., 2020].

In this Section, I outline different machine learning approaches used for the development of such models and discuss the molecular representations currently used to featurise ligand-based as well as structure-based scoring models.

1.3.1 Molecular Representations

Many different molecular representations have been developed for use in machine learning scoring functions. While there is no single representation that is generally suitable for every task, different representations are created to tackle specific tasks. For example, molecular descriptors derived from ligand structures (see Section 1.2.1) are useful when building single-target protein-ligand affinity prediction models. Beyond classical QSAR descriptor-based models, current high-performing representations can be divided into four subcategories: (i) fingerprints; (ii) voxelised image representation; (iii) molecular graphs; and (iv) learned representations (such as autoencoders).

Fingerprints can come in many different shapes and sizes and can encode purely topological information about ligand structure such as Extended Connectivity Fingerprints (ECFP) [Rogers and Hahn, 2010] or include 3D protein-ligand contact information such as the Protein-Ligand Extended Connectivity fingerprint (PLEC) [Wójcikowski et al., 2018] and the Extended Connectivity Interaction Features (ECIF) [Sánchez-Cruz et al., 2020]. Generally, such fingerprints are developed to be used in protein-ligand affinity scoring functions based on random forests (RF) or gradient boosted decision tree models [Wójcikowski et al., 2018; Sánchez-Cruz et al., 2020; Boyles et al., 2019] but can also be used effectively in deep learning models such as a simple feed-forward neural network as I have shown in Chapter 4 [Moesser et al., 2022]. Fingerprint-based models, especially those used with RFs, are very popular due to their simplicity. For example, due to the straightforward implementation of RF models in data science packages such as scikit-learn [Pedregosa et al., 2011], a RF fingerprint model can be build in a day, the model trained on thousands of data points in minutes and a detailed relative feature importance analysis performed immediately using a single scikit-learn function (*feature_importance_*) from the *RandomForestRe-*

gressor module [Pedregosa et al., 2011]). Although fingerprint-based models have traditionally encoded 2D information [Landrum et al., 2006; Rogers and Hahn, 2010], modern 3D structure-based fingerprints have been shown to outperform 2D fingerprints when predicting protein-ligand binding affinity [Gao et al., 2020]. However, fingerprints are limited in how much information can be encoded in a given fingerprint size. By aiming to describe the biophysical interactions in 3D protein-ligand complexes as accurately as possible, 3D fingerprint vectors can become excessively long, creating models that use feature vectors with 10,000s or even 100,000s of features for a training dataset in the thousands, raising concerns about overfitting and generalisability [Wójcikowski et al., 2018; Boyles et al., 2019].

1.3.1.1 Convolutional Neural Networks

In order to address these concerns about how best to represent 3D protein-ligand complexes, advances in the field of image recognition using convolutional neural networks (CNN) were adopted for protein-ligand affinity prediction. To adjust the image recognition approach for protein-ligand complexes, the RBG (red, green, blue) channels that encode for the color values in a traditional CNN used for image recognition were re-purposed (typically with separate ligand and protein channels) to encode for different atomic properties such as hyrophobicity, aromaticity, hydrogen bond donor or acceptor properties, ionizability, etc. instead, and a 3D voxel grid created for the protein-ligand complex [Jiménez et al., 2018; McNutt et al., 2021; Ragoza et al., 2017]. A schematic representation of a CNN for protein-ligand binding affinity scoring function is shown in Figure 1.5.

Substantial progress has been made in the development of novel CNN-based scoring functions with high performing models used in virtual screening [Ragoza et al., 2017], protein-ligand binding affinity prediction [Jiménez et al., 2018] and even as a scoring function in protein-ligand docking [McNutt et al., 2021]. One advantage of

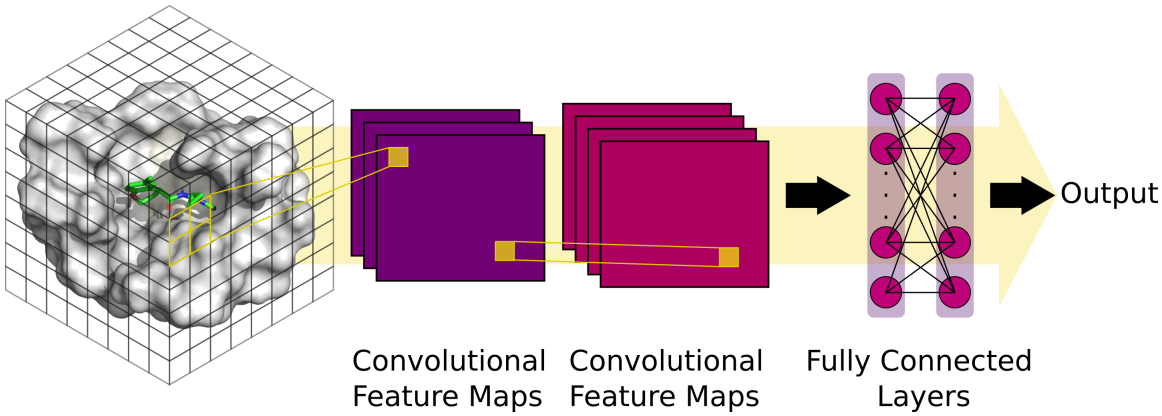

Figure 1.5: Schematic overview of the architecture of a 3D CNN model for protein-ligand binding affinity prediction (for example as reported by Ragoza et al. [2017]). A 3D voxelised grid is taken as input and N different convolutions are applied to create N convolutional maps, followed by a second convolutional layer and finally the fully connected layers to generate the prediction.

using convolutional layers in neural networks is the ability of CNNs to extract higher order features from the training set, given enough data and layer depth. Just like CNNs for image recognition have been shown to recognise and extract higher order features within images such as recognising a dog or a cat from an image of wildlife [Sermanet et al., 2014], CNN's have been applied to protein-binding affinity with the intent to extract meaningful interactions in the protein-ligand complex. However, while voxelised representations for CNNs are generally able to encode 3D information efficiently, in order for the model to extract chemically relevant features, a large training dataset is required. The need for substantially larger datasets than what is currently available for protein-ligand binding affinity data is therefore a limiting factor. This becomes evident when comparing one of the most popular databases that include 3D structural data and protein-ligand binding affinity data, the PDB-Bind dataset [Liu et al., 2015, 2017] (23496 entries) with the datasets used to train the image recognition algorithms such as AlexNet and GoogleNet which were trained on datasets with more than a million data points [Alzubaidi et al., 2021; Krizhevsky et al., 2012; Szegedy et al., 2015].

1.3.1.2 Graph Neural Networks

Alternatively, as one of the most recently emerging fields of research into molecular representation in deep learning, molecular graph-based models to solve these issues by respectively encoding for chemically-relevant bonds and atoms directly as edges and nodes in a mathematical graph. Many different GNN architectures are currently used such as Graph Isomorphism Networks (GIN, [Xu et al., 2019]), Graph SAGE [Hamilton et al., 2017], Graph Attention Networks (GATNet, [Veličković et al., 2018]), and Graph Convolutional Networks (GCN, [Kipf and Welling, 2017]). For example, GCNs use the same approach to convolution as CNN models, however, instead of applying a filter function over the voxelised grid representation of a 2D image, GCNs apply the filter function over the sub-graphs in the graph (Figure 1.6).

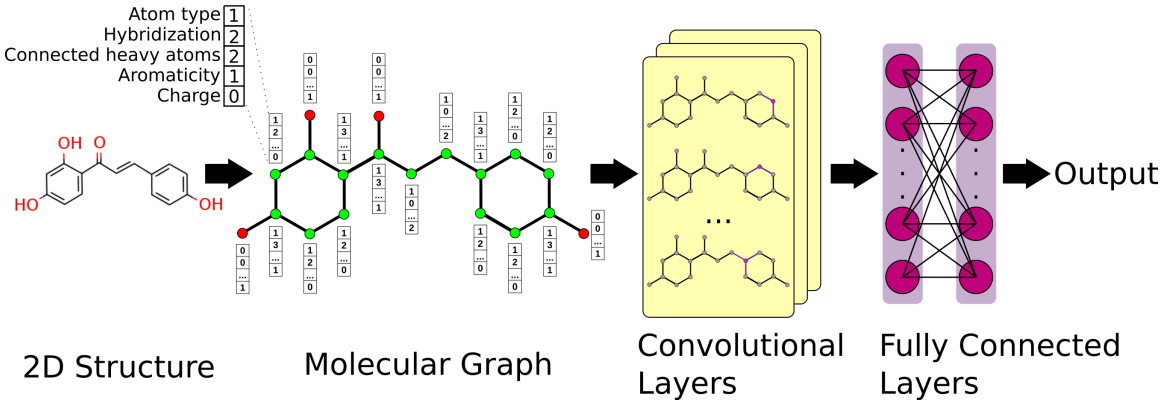

Figure 1.6: Schematic overview of the generation of a ligand-based graph and setup of a Graph Convolutional Network (GCN) for protein-ligand binding affinity prediction. The 2D structure of the ligand is encoded as a graph, with the one-hot encoded atomic features as the node features of the graph (example here shows 5 popular atomic features, but others are possible too depending on the setup). The GCN model applies the filter function over the nodes of the graph, generating N convolutional layers, depending on the setup. The convolutional layers connect into the fully connected layers to generate the prediction.

The first molecular GCNs only encoded ligand information, initialising a molecular graph-based on the connectivity of the ligand atoms, and by including ligand-based node features to describe each atom such as its element, connectivity, and valence [Wu et al., 2018]. However, over the last few years, different ways of including protein

information were developed such as the GraphDTA models [Nguyen et al., 2020] that encode the protein sequence in a separate 1D CNN. Most recently, new structure-based GCN models were developed that add protein atoms in proximity to the ligand into the molecular graph [Lim et al., 2019; Li et al., 2021] or by training 2 separate graphs for the ligand and protein, respectively in parallel [Jiang et al., 2020b]. Currently, GNN models are among the best performing models for protein-ligand binding affinity prediction together with CNN and fingerprint-based models [Sánchez-Cruz et al., 2020; Jiménez et al., 2018; Li et al., 2021; Lim et al., 2019]. However, no molecular representation and model architecture has been found to be superior, and the top performing models all perform very similarly (Pearson correlation coefficients between 0.82-0.87 [Sánchez-Cruz et al., 2020; Jiménez et al., 2018; Li et al., 2021; Lim et al., 2019; Zheng et al., 2019; Moesser et al., 2022]).

1.3.1.3 Molecular Autoencoders

As an alternative to the molecular representations described above that are primarily designed to be used as input for machine learning scoring functions, molecular autoencoders were developed to allow a continuous space representation of molecules. Autoencoders are unsupervised machine learning models where two networks are trained in tandem: an encoder that translates the input representation of a compound into (usually lower dimensional) latent space, and a decoder that reverses the translation to re-create the original compound representation from latent space (Figure 1.7). Autoencoders are trained to minimise the translation loss between encoder and decoder.

There are two major approaches to molecular representation in autoencoders: graph-based and SMILES-based methods. In graph-based autoencoders, molecules are first converted into a graph as described above for GNN methods, and the autoencoder tasked to encode and decode between graph and latens space. In SMILES-based

methods, the SMILES [Weininger, 1988] representation of a molecule is used instead. SMILES-based autoencoders take inspiration from language-based autoencoders such as word2vec [Mikolov et al., 2013] which was adapted into the popular molecular autoencoder mol2vec [Jaeger et al., 2018] as well as other SMILES-based autoencoders used widely for compound generation [Gómez-Bombarelli et al., 2018; Kadurin et al., 2017; Blaschke et al., 2018]. SMILES-based autoencoders are limited by the fact that SMILES strings of two extremely similar compounds can differ drastically, which would result in a significantly different location in latent space [Jin et al., 2018]. Additionally, the same compound can be written in different SMILES strings, therefore posing a challenge from a 1-to-1 translation perspective when training molecular autoencoders. However, this can be overcome through the usage of randomized SMILES during training [Arús-Pous et al., 2019]. Nonetheless, SMILES-based autoencoders have increasingly popular and are more widely used than their graph-based counterparts. A schematic example of a SMILES-based autonecoder setup is shown in Figure 1.7.

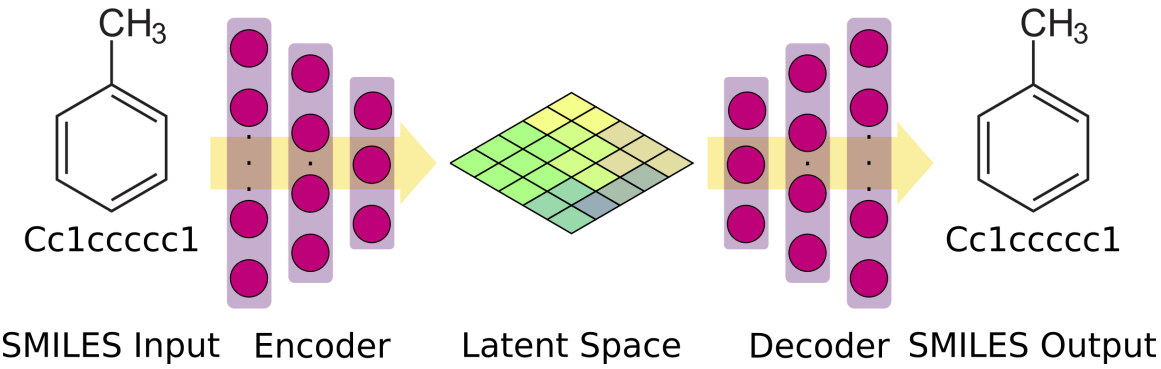

Figure 1.7: Schematic overview of the encoder and decoder setup of a generic SMILES-based autoencoder. A deep neural network is trained to encode the SMILES representation into latent space in tandem with the encoder, which takes the latent space representation and decodes it back into SMILES.

While there are autoencoders such as mol2vec [Jaeger et al., 2018] that were developed as an alternative to ECFP and other ligand-based descriptors, the most recent

application of molecular autoencoders is in the field of *de novo* compound generation. In order to generate a new compound from latent space, a random position in latent space or a specific location (for example close to a known molecule of interest) may be sampled and decoded into a new molecule. This process for example has been used in combination with Bayesian optimisation to generate molecules with increasing propensity to exhibit a specific property (such as hydrophobicity, structural features such as number of rings etc.) [Gómez-Bombarelli et al., 2018].

Currently, a large variety of different molecular representations and corresponding model architectures are used for protein-ligand binding affinity scoring functions. Nonetheless, no single representation has been found to be superior so far. All architectures and representations have at least some models among the top performing scoring functions [Sánchez-Cruz et al., 2020; Boyles et al., 2019; Jiménez et al., 2018; Moon et al., 2022; Moesser et al., 2022]. However, models can differ drastically when it comes to secondary considerations such as interpretability, with deep learning models being notoriously difficult to interpret, and are often referred to as "black-box" models. Especially for the emerging field of molecular graph-based neural networks, it is important that future graph-based representations are created with interpretability and generalisability in mind.

1.3.2 Ligand-Based Models

As discussed in Section 1.2.1 & 1.3.1, ligand-based approaches have been widely used in single-target QSAR models. However, recent studies have shown that ligand-based machine learning models also perform well on mutli-target datasets such as the widely-used Comparative Assessment of Scoring Functions (CASF) 2016 benchmark [Boyles et al., 2019, 2021]. The CASF-2016 dataset which is a commonly used benchmark consisting of 285 protein-ligand complexes and their corresponding experimentally

determined binding affinity chosen to evaluate protein-ligand binding affinity scoring functions [Su et al., 2018] (see Section 1.3.3 for a more detailed discussion of CASF-2016). While some correlation is to be expected when evaluating the binding-affinity of a given ligand against any protein simply based on ligand information alone (since there are generally favourable properties a ligand can have to be considered a good drug), a truly generalisable model that is able to distinguish between different proteins should not be possible with only ligand-based features. The high performance of ligand-based models on multi-protein datasets therefore suggests a strong ligand bias in the training and test sets used throughout the field and has been highlighted as an area of concern [Boyles et al., 2019]. As a result, as described by Boyles et al. [2019] and highlighted in the work described in Chapter 4 of this thesis, future studies into the development of structure-based, generalisable models should focus on interpretable models in order to analyse if models are truly learning the biophysics of protein-ligand interactions and molecular recognition, or are just regurgitating ligand biases while neglecting structure-based features [Boyles et al., 2019, 2021; Moesser et al., 2022].

1.3.3 Structure-Based Scoring Functions

Structure-based models have been developed to overcome the limitations of ligand-based models and to build a generalisable model that is able to score the binding affinity of any given protein-ligand complex accurately. As previously mentioned (see Section 1.3.1), common examples of structure-based machine learning methods include CNN [Jiménez et al., 2018; Zheng et al., 2019; McNutt et al., 2021], GNN [Li et al., 2021; Lim et al., 2019; Moesser et al., 2022], 3D fingerprint [Sánchez-Cruz et al., 2020; Wójcikowski et al., 2018] and 3D descriptor-based [Durrant and McCammon, 2011; Ballester and Mitchell, 2010] machine learning models and appear to outperform

classical semiempirical scoring functions used in molecular docking (see Section 1.2.2) [Morris et al., 2009; Trott and Olson, 2010; Friesner et al., 2004; Jones et al., 1997].

Although the highest performing structure-based scoring functions achieve high Pearson correlation coefficients (between 0.82-0.87 depending on the model, [Jiménez et al., 2018; Zheng et al., 2019; Sánchez-Cruz et al., 2020; Wójcikowski et al., 2018; Boyles et al., 2019; Moesser et al., 2022; Li et al., 2021]) when trained on the PDBbind database and tested on the popular CASF-2016 benchmark [Su et al., 2018], consideration of generalisability has only recently emerged. The Comparative Assessment of Scoring Functions (CASF) benchmark is a dataset of 285 protein-ligand complexes sourced from the PDBbind 2016 refined set [Liu et al., 2015, 2017] which is a collection of high quality crystal structure obtained from the Protein Data Bank (PDB) [Berman et al., 2000]. The CASF-2016 benchmark has been commonly accepted as one of the primary scoring power benchmarks that almost every scoring function described in this thesis has been tested against. However, as Boyles [2020] points out, since the CASF-2016 dataset is chosen to only include protein-ligand pairs of proteins that are already present in the PDBbind training set, the CASF-2016 set is not a suitable benchmark to assess model generalisability. This has now been recognized more broadly, and more recently reported scoring functions are including an additional generalisability test, such as the elimination of protein-ligand complexes from the training set if that protein is within a certain sequence identity threshold to any protein in the CASF-2016 test set [Boyles et al., 2019; Moon et al., 2022; Moesser et al., 2022]. In all cases [Boyles et al., 2019; Moon et al., 2022; Moesser et al., 2022], model performance drops strongly when controlling for training and test set similarity, highlighting that the creation of truly generalisable protein-ligand binding affinity scoring functions is still an open problem.

1.3.4 Bayesian Optimisation in Drug Discovery

Chemical space is vast and the experimental exploration of it is expensive, time consuming, and often prone to subjective biases in compound prioritisation (see Section 1.1). As a data-driven alternative to the optimisation approaches used in traditional medicinal chemistry, active search methods such as Bayesian optimisation (BO) have recently been applied to drug discovery to parse chemical space more efficiently and optimise molecular properties [Pyzer-Knapp, 2018; Gómez-Bombarelli et al., 2018; Griffiths and Hernández-Lobato, 2020].

Bayesian optimisation is an optimisation strategy that was originally pioneered by Jonas Mockus in the 1980s to optimise expensive-to-evaluate functions [Mockus, 1989]. The goal of Bayesian optimisation is to find the optimal value of an objective function in the minimum number of steps. Since the objective function is unknown to start with, BO uses a surrogate function, often a Gaussian Process (GP), to model both the objective function and assign a measure of uncertainty to each point in the function. To choose a new data point most likely to improve the optimisation, an acquisition function is used that evaluates each point in the GP surrogate function and chooses the next point to sample based on an exploration *vs* exploitation trade-off (Figure 1.8).

Exploration describes the process of evaluating the gain in information about the surrogate function that could be achieved by sampling data points that the GP assigns a high uncertainty to, and that could therefore alter the surrogate function drastically. Most often, high uncertainty is associated to parts of the function where little information is currently available. Alternatively, exploitation describes the process of sampling data points close to the current optimum value to find a new optimum. This search strategy does not increase the information in the system as much, but might yield incremental improvements to find the global optimum. Different acqui-

sition functions assign different weights to the exploration *vs* exploitation trade-off and it is therefore possible to adjust the acquisition function based on the goal of the optimisation.

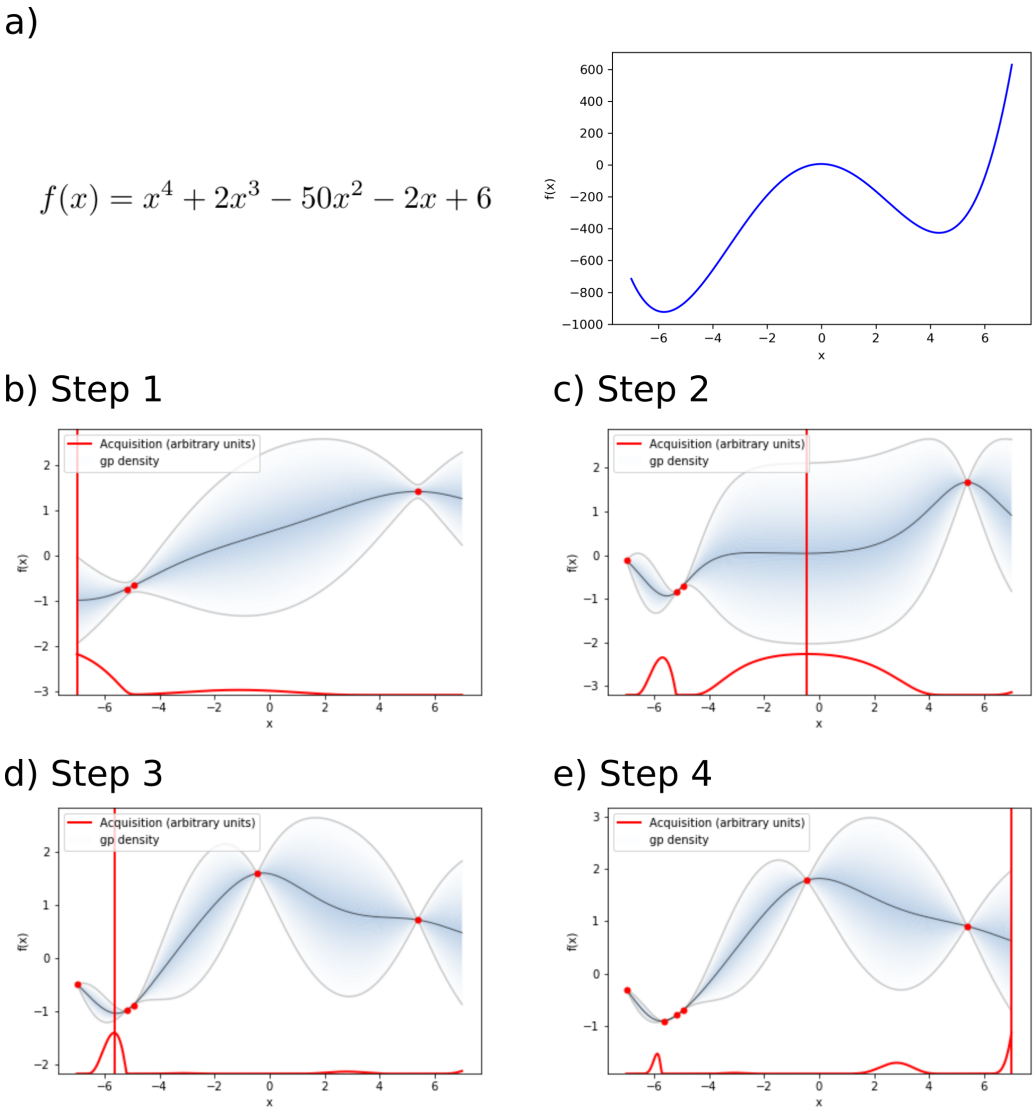

Figure 1.8: Overview of a simple Bayesian optimisation process using a GP surrogate function and the Expected Improvement (EI) acquisition function to find the minimum of the objective function. The acquisition function is shown in red, with the next data point to be sampled (the maximum of the acquisition function) indicated by the red vertical line. The uncertainty of the GP model is shown by blue shading. a) The objective function. b) The initial model of the function based on three random data points. c) The first point is sampled and the GP adjusts its estimation. At each step the acquisition function is updated. d) The next point is sampled. e) Another point is sampled, and the Bayesian optimization algorithm has identified the minimum of the objective function.

This sequential evaluation of the exploration *vs* exploitation trade-off closely resembles the traditional design, make, test, analyse (DMTA, [Plowright et al., 2012]) cycle in drug discovery (Figure 1.3), where medicinal chemists re-evaluate which compounds to make next every cycle based on the information gained in the previous cycle. BO is therefore highly suitable as a tool to guide the decision making for medicinal chemists during hit-to-lead and lead optimisation.

While most studies that have focused on the application of Bayesian optimisation for drug discovery have used continuous BO using a latent space representation (see Section 1.3.1) prospectively to find and generate new compounds that excel in a specific property such as solubility [Gómez-Bombarelli et al., 2018; Griffiths and Hernández-Lobato, 2020], Pyzer-Knapp described a retrospective multi-armed bandit approach using a discrete ECFP representation of compounds [Pyzer-Knapp, 2018] to find the best inhibitor in a large dataset of compounds with known binding affinities. This multi-armed bandit approach has not been widely explored in the context of drug discovery, and open questions about the optimal compound representation and the usefulness in a real drug discovery project still remain.

1.4 Project Aims

As outlined in this Chapter, the costs, both in time and money, of pharmaceutical drug discovery has been steadily increasing over many years. Drug targets considered as "low hanging fruits" are slowly running out, forcing scientists to target more challenging diseases. In combination with the increasing strictness of regulatory authorities when evaluating drug approvals, this has lead to an increase in failure rate in the clinic. Improvement in efficiency as well as in the quality of drug candidates is therefore crucial for the sustainability of pharmaceutical drug discovery in the future. In this thesis, I present three complementary computer-aided drug discovery methods

with the aim of increasing efficiency in the pre-clinical drug discovery pipeline, with a particular focus on hit discovery and hit-to-lead optimization.

The overarching theme of this thesis focuses on exploration of different representations of molecular structure, intermolecular interactions and 3D protein-ligand complexes. In Chapter 2, I present the development of ligand-based models using a Gaussian Process (GP) as an easy-to-implement tool to guide exploration of chemical space for the optimization of binding affinity against one or more protein targets. I explore different topological fingerprint and autoencoder representations for Bayesian optimisation (BO) and show that BO is a powerful tool to help medicinal chemists prioritise new compounds to make for single-target as well as multi-target optimisation.

In Chapter 3, I present the development of a knowledge-based approach to drug design, combining quantitative contact fingerprint-based similarity calculations with a fragment-based drug discovery approach to understand SARS-CoV-2 M^{pro}-substrate specificity and to design novel small molecule inhibitors *in silico*. I show that the M^{pro} protein-ligand interaction fingerprints can be powerful tools for knowledge-based design, allowing the identification of protein-ligand interactions at several levels of detail and its direct use in drug design. In addition, by filtering virtual screening (VS) results for contacts of high interest, the identification of high quality VS hits was enabled. In combination with a fragment-based drug discovery approach, I showcase how this knowledge-based interaction fingerprint-driven approach can reveal fruitful fragment-growth design strategies. The work in this chapter has been published in Chemical Science [Chan et al., 2021a].

Finally, in Chapter 4, I expand on the knowledge-based contact fingerprints in Chapter 3 to create a ligand-shaped molecular graph representation (Protein Ligand Interaction Graphs, PLIGs) for graph-based deep learning that are able to encode

all intermolecular interactions in a protein-ligand complex within the node features of the graph. I explore a variety of Graph Neural Network (GNN) architectures in combination with PLIGs to create several high-performing protein-ligand affinity prediction models with comparable performance, highlighting that careful design of molecular representations outweighs small gains that can be made by optimizing deep learning architectures. However, overall I found Graph Attention Networks (GAT-Net, Veličković et al. [2018]) to perform slightly better than other GNN architectures. PLIGs were designed with the goal of advancing the field of scoring function development to find generalisable models that are able to encode the biophysics of molecular recognition while retaining simplicity and most importantly, full interpretability. The work in this chapter is published on biorxiv [Moesser et al., 2022] and will be send for peer review in an appropriate journal.

Chapter 2

Exploration of Bayesian Optimization for Structure-Activity Relationship Modeling

2.1 Introduction

As one of the more recently emerging fields in cheminformatics, active learning strategies coupled with *de-novo* design have proven to be promising tools for the generation and optimization of molecules towards a certain objective. Examples include the use of recurrent neural networks (RNNs) [Ar´us-Pous et al., 2019; Popova et al., 2018], autoencoders [Gómez-Bombarelli et al., 2018], generative adversarial neural networks (GAN) [Méndez-Lucio et al., 2020], and synthesis-based methods [Hartenfeller et al., 2012; Vinkers et al., 2003]. However, the fine tuning of these generative methods towards the design of compounds that have desired drug-like properties remains challenging. First steps towards tackling this challenge include methods that employ reinforcement learning (RL), where a predictive model takes the previously generated compound as input and evaluates its performance against a certain property (*e.g.* hydrophilicity or number of H-bond donors). Subsequently, the generator is rewarded for proposing molecules with the desired property (or properties, in multi-objective

optimization) and over many cycles learns to generate new molecules optimised in that desired regime [Popova et al., 2018]. However, the usefulness of RL methods as well as autoencoders for the generation of new molecules with the desired properties is limited by the accuracy of predictive models, especially for binding affinity prediction. Previous approaches have chosen easily calculated physicochemical properties such as the water-octanol partition coefficient, or structural features such as the number of aromatic rings [Popova et al., 2018; Gómez-Bombarelli et al., 2018]. Nonetheless, for protein families where more accurate models for the binding affinity prediction exist, RL-based methods have been applied to generate new molecules that are predicted to be high affinity binders [Popova et al., 2018; Olivecrona et al., 2017].

This highlights a big problem in the cheminformatics field: the need for more diverse datasets with high quality data that cover proteins and ligands beyond the commonly observed protein families such as kinases, proteases, transferases or G-protein coupled receptors (GPCRs) that collectively cover over 32% of all protein targets represented in the ChEMBL database [Davies et al., 2015; Mendez et al., 2019]. In order to address these issues, this project aimed to test the performance of active learning strategies for the optimization of protein-ligand binding affinity on novel, real-world data of an ongoing drug discovery project. I used high quality experimental binding data from two sources: i) the matrix metalloproteinase-12 (MMP-12) dataset [Pickett et al., 2011] as a validation dataset since it had been previously used as a benchmark in similar studies [Pyzer-Knapp, 2018]; and ii) a novel metallo-β-lactamase (MBL) inhibition dataset obtained from the Schofield group at the University of Oxford.

MBLs were chosen as targets in part with the aim of contributing to the global challenge of antibiotic resistance. Today, the most important class of antibiotics are β-lactam derivatives which correspond to over half of the global antibacterial market

[Elander, 2003]. One of the largest ongoing global health threats is growing antibiotic resistance, which renders antibacterial agents useless through the evolution of defense mechanisms within bacteria. β-Lactam resistance poses a significant threat, since it puts the viability of the most widely used class of antibacterial agents at risk. β-Lactam antibiotics target penicillin-binding proteins (PBPs), which are transpeptidases involved in peptidoglycan synthesis, an essential step in bacterial cell wall construction. The β-lactam ring of the inhibitors is cleaved by PBPs and the inhibitor becomes covalently attached to the active site, rendering the PBP inactive and ultimately resulting in bacterial cell lysis [Yocum et al., 1980]. Evolutionary adjustment of bacteria to this threat has lead to the emergence of bacterial β-lactamases as a defense mechanism. Resistance is obtained by cleavage of the active β-lactam ring of the compound by β-lactamases. As an enzyme class, β-lactamases are divided into two subclasses, the serine-β-lactamases (SBL) and the metallo-β-lactamases (MBL). Although SBL inhibitors such as Sulbactam and Tazobactam (Figure 2.1) are widely used drugs (in combination with a β-lactam antibiotic), no MBL inhibitors have yet been brought to market. One major challenge in the development of MBL inhibitors

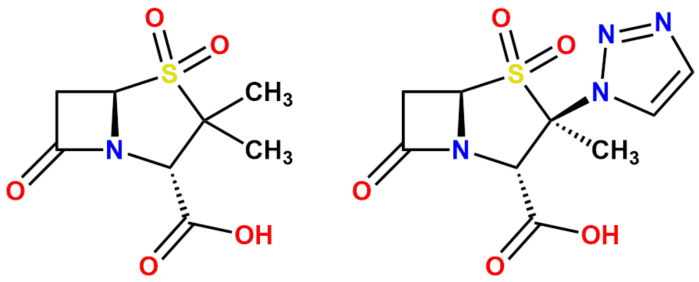

Figure 2.1: Structures of Sulbactam (left) and Tazobactam (right).

is their required polypharmacological profile. In order to be clinically viable against a broad range of bacterial infections, MBL inhibitors should be active against as many of the four most clinically important bacterial MBL enzymes as possible: the Verona integron-encoded metallo-β-lactamases (VIM-1 and VIM-2), the Imipenemase

(IMP-1) and the New Delhi metallo-β-lactamase (NDM-1) [Walsh et al., 2005]. Unfortunately, the development of multi-target MBL inhibitors without sacrificing other desired properties such as solubility or cell-penetration has not yet been achieved.

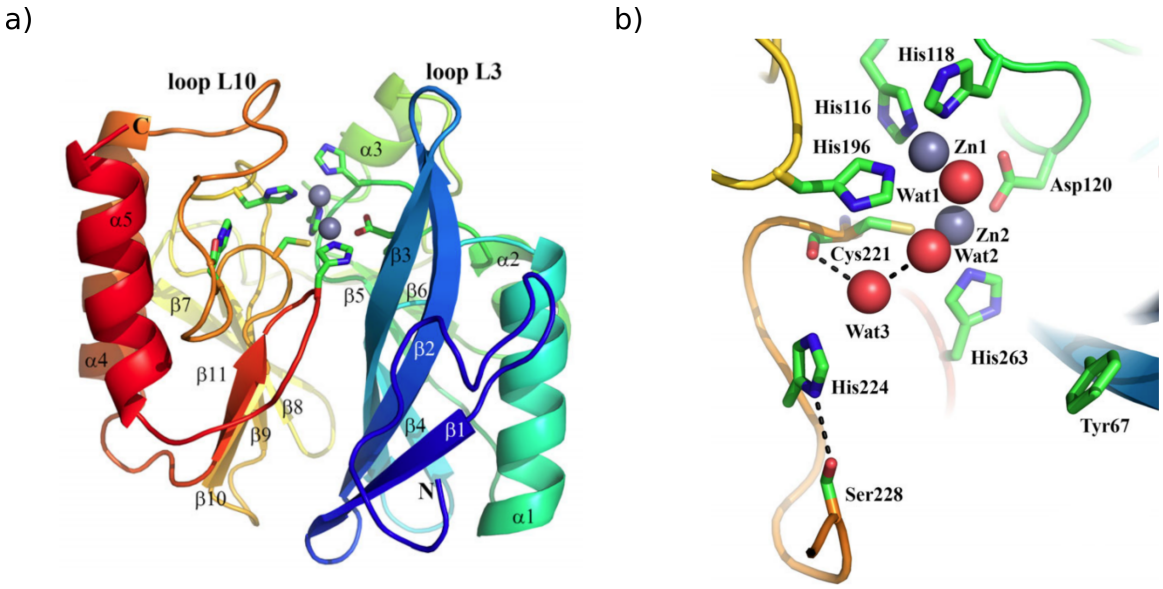

Figure 2.2: Views from a crystal structure of VIM-1 is shown as published by Salimraj et al. [2019] (CC BY 4.0 license). (a) The VIM-1 tertiary structure showing the overall fold and active site residues, color coded from blue (N terminus) to red (C terminus). Zinc ions are shown as grey spheres and the side chains of important active site residues shown in green. (b) The VIM-1 active site, with they key active site amino acids labelled, and water oxygens shown as red spheres.

The MBL family is thus an excellent model system to use for the prospective validation of new computational methods for the design of selective/promiscuous compounds. MBLs have been extensively researched in the Schofield lab at the University of Oxford as part of the global European Gram Negative Antibacterial Engine (ENABLE) project, where a series of indole-2-carboxylate-based small molecule inhibitors have been identified as broad spectrum MBL inhibitors. These compounds are non-covalent inhibitors that bind in the active site of the MBLs and coordinate to the zinc ions (Figure 2.2). The structures of the four major MBL targets are known and their active site motifs are conserved between them [Salimraj et al., 2019]. As an example, the structure of VIM-1 and a close-up of its active site is shown in Figure

2.2. The drug discovery efforts of the ENABLE project gave rise to a confidential MBL dataset containing 558 compounds with biological activity data against the four MBL targets VIM-1, VIM-2, IMP-1, and NDM-1. The limited information available as well as the fact that optimization work is still ongoing therefore poses a more realistic challenge than the retrospective analyses of well studied families such as kinases, and the development of new multi-target MBL inhibitors directly serves an unmet need for the development of new antibacterial agents.

The work by Pyzer-Knapp [2018] has shown that active learning techniques such as multi-armed bandit Bayesian Optimization (BO) can be used to optimize "black-box" functions such as experimental protein-ligand binding affinity, potentially speeding up the discovery process by helping medicinal chemists prioritize compounds for synthesis and testing more efficiently. BO is a black-box optimization technique that is particularly suitable for expensive objective functions. For protein-ligand binding affinity prediction, the objective function is the relationship between the structure and properties of the protein and ligand. Evaluation of this function means expensive and time-consuming lab experiments would have to be conducted to synthesize and then test the compounds in order to gain the affinity of a given compound. Since the values of the objective function is unknown for all compounds at the start of the optimization (such as at the start of an inhibitor optimization project where less than 10 high quality hits might have been found in a high-throughput screen), the Bayesian strategy is to create a surrogate function to model the relationship between the inhibitor's structure and its binding affinity.

In this project, I used a Gaussian Process as implemented in the open-source toolkit GPyOpt [Consortium, 2020b] as the surrogate function. The GP uses the data it has seen so far to estimate the binding affinity as well as the associated predicted uncertainty for all remaining compounds in the dataset. The goal of this

method is to use the BO to help chemists choose which compound to make next from a set of possible compounds based on the likelihood it will improve the existing "best compound" (exploitation) or the likelihood that it will open up new chemical space by sampling compounds with high uncertainty (exploration). A more detailed overview of Bayesian Optimization in general can be found in Chapter 1, Section 1.3.4. This active-learning strategy could enable chemists to apply resources more efficiently than in traditional structure-activity relationship optimization campaigns, reducing the amount of compounds synthesized and tested and ultimately increasing speed and improving the cost-efficiency of drug discovery. A detailed description of the implemented BO algorithm can be found in Section 2.2.3.

The work by Gómez-Bombarelli et al. [2018] has been foundational to the field of VAE-based molecular generation coupled with BO, showing that variational autoencoders (VAEs) are capable of encoding compounds in a continuous space based on their chemical structure as well as their corresponding physical properties such as the base-10 logarithm of their water-octanol partition coefficient. They also showed VAEs could be used in combination with Gaussian Processes to sample from the encoded latent space and optimize for a combination of drug likeness and synthetic accessibility. As a result, their method was able to output new compound designs. As discussed in Chapter 1, Section 1.3.1, the representation of compounds in latent space is one of many currently employed representation techniques. Overall, the search for the most optimal molecular representation to tackle protein-ligand binding affinity prediction is still an active area of research and composes one of the overarching themes in this thesis, with each chapter approaching the problem from a different angle and with different methods.

Ligand-based models, particularly in QSAR modeling, have used classical topological fingerprints such as ECFP [Rogers and Hahn, 2010] or computed molec-

ular descriptors such as counts of functional groups and physical properties, exemplified by the Descriptor module in RDKit [Landrum et al., 2006]. More recent models, especially deep learning models have utilized more creative representations such as text-based autoencoders [Gómez-Bombarelli et al., 2018], molecular graphs [Nguyen et al., 2020] or 3D voxels for convolutional neural networks [Jiménez et al., 2018; McNutt et al., 2021]. In the work published by Pyzer-Knapp [Pyzer-Knapp, 2018], only extended connectivity fingerprints (ECFP) [Rogers and Hahn, 2010] were used as input features for a single objective optimization. In order to expand upon the technique outlined by Pyzer-Knapp and to optimize its potential, I focused first on determining the most optimal compound representation for BO. Herein, I describe my investigations into Bayesian multi-objective optimization, along with the exploration of three different compound representations and their effect on BO performance: extended connectivity fingerprints (ECFP) [Rogers and Hahn, 2010], connected subgraph fingerprints (CSFP) [Bellmann et al., 2019] and the autoencoder Mol2vec [Jaeger et al., 2018]. In this work, I show that ECFP perform best overall. In addition, I implement a multi-objective optimization approach that combines simultaneous optimization against two MBL targets. My optimization approach outperforms previously described methods on the same benchmark (MMP-12 dataset) [Pyzer-Knapp, 2018] and was further validated on an independent, real-life lead-optimization dataset (MBL dataset) as well as in a multi-objective optimization scenario. This chapter is fully my own work. My implementation of the Bayesian optimisation algorithm can be found on Github `https://github.com/MarcMoesser/Bayesian-Optimization-For-Drug-Discovery`.

2.2 Materials & Methods

2.2.1 Data Sets

2.2.1.1 Matrix Metalloproteinase-12 Dataset

The dataset used to benchmark the performance of my BO methods against the results obtained by Pyzer-Knapp [Pyzer-Knapp, 2018] was the matrix metalloproteinase-12 (MMP-12) dataset [Pickett et al., 2011]. MMP-12 is a metalloprotease primarily expressed by macrophages and has been linked to several pathological conditions such as rheumatoid arthritis and the formation of aneurysms [Dean et al., 2008]. The dataset was originally published by Pickett et al. [2011] as an attempt to synthesize and test a complete 50×50 matrix of different R_1 and R_2 group combinations while keeping a constant biaryl sulfonamide core (Figure 2.3).

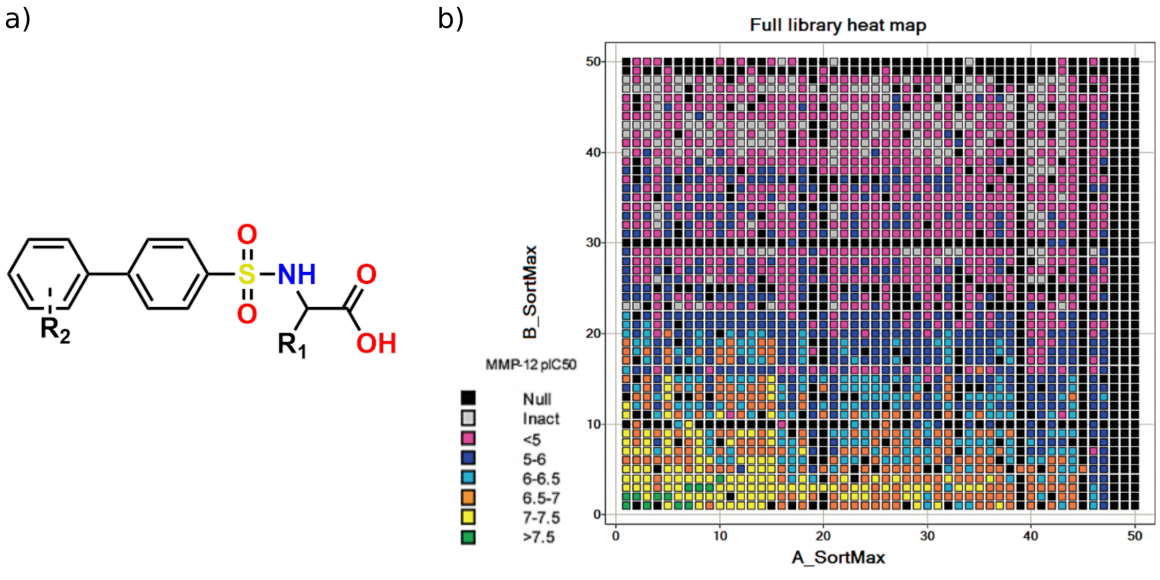

Figure 2.3: (a) The biaryl sulfonamide core present in all compounds in the MMP-12 dataset. (b) Heatmap of the 50×50 biaryl sulfonamide array. Reprinted (adapted) with permission from [Pickett et al., 2011] Copyright 2011 American Chemical Society. *A_SortMax* corresponds to modifications in R_1 and *B_SortMax* to R_2. Values in black are marked as "Null" and correspond to compounds where the compound was not synthesized or the assay failed. Values in grey were found to be inactive in the assay.

The dataset contains pIC_{50} data for 1880 out of the 2500 possible compounds. The

absent compounds were either not made, not assayed, or failed the assay. I manually prepared the dataset, setting the pIC_{50} values of compounds labelled *"inactive"* to 0 and discarding the compounds labelled *"assay failed"*, *"not assayed"* and *"not made"*. This dataset was used as a control to compare directly to the original results obtained by the BO method developed by Pyzer-Knapp [Pyzer-Knapp, 2018] on the same dataset.

2.2.1.2 Metallo-β-Lactamase Dataset

The metallo-β-lactamase (MBL) dataset was obtained from an ongoing pre-clinical drug discovery project in the Schofield group at the University of Oxford. The dataset is currently confidential and has not yet been published. However, once more significant progress on the discovery project is made, the data will be publically released. It currently consists of a total of 558 compounds that have been tested against four different MBL targets: VIM-1, VIM-2, NDM-1 and IMP-1. Not all compounds have been tested against every protein target. The dataset includes the compound's chemical structure and the experimentally measured IC_{50} value for each compound-protein combination. IC_{50} values against each MBL were determined using a fluorogenic assay as described by van Berkel et al. [2013]. All compounds in the dataset originate from the same strucure-activity relationship study and all share a common indole-2-carboxylate core shown in Figure 2.4 with two substituent groups (R_1 and R_2) that are varied. However, the dataset does not follow a matrix search approach where all possible combinations of the R_1 and R_2 groups were tested, but rather follows a more typical approach adopted by medicinal chemists, where a broader spectrum of R groups were explored and only certain, high performing ones kept constant.

The dataset was split for the BO method into separate subsets for each target: VIM-1, VIM-2, IMP1, and NDM1; with 495, 527, 533, and 550 compounds, respectively. In addition, one dataset for every protein-protein pair was also created for

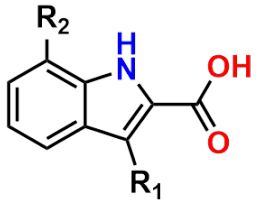

Figure 2.4: Common indole-2-carboxylate core present in every molecule of the MBL dataset. The substituent groups R_1 and R_2 are modified throughout the dataset to explore the SAR.

multi-objective optimization. For the combined datasets, all compounds with IC_{50} values measured against each target pair were selected, the geometric mean of their respective IC_{50} values against both targets calculated, and then combined into a new dataset. This resulted in the following six datasets: IMP-1+VIM-1, IMP-1+VIM-2, NDM-1+IMP-1, NDM-1+VIM-1, NDM-1+VIM-2 and VIM-2+VIM-1; with 494, 525, 540, 502, 533, and 494 compounds respectively.

2.2.2 Implementation of Compound Representations

The performance of the BO method was evaluated using three different compound representations: Extended-Connectivity Fingerprints [Rogers and Hahn, 2010] (ECFP), Connected Subgraph Fingerprints [Bellmann et al., 2019] (CSFP) and mol2vec [Jaeger et al., 2018].

ECFPs are well established topological fingerprints specifically designed for molecular similarity calculations and structure-activity modeling [Rogers and Hahn, 2010]. ECFPs are generated by giving each atom an initial identifier and generating a bit string of predefined length that contains the chemical features of the atom and its neighboring atoms up to a defined radius (or a number of bonds). This process is repeated a specified number of times and duplicate identifiers removed. Afterwards, the updated identifier of each atom is hashed to determine which bits should be set in the output fingerprint. The remaining unique set of atom identifiers after hashing constitutes the ECFP.

CSFPs have been developed more recently and are a modified version of the ECFP procedure where unique identifiers are placed on substructures instead of atoms. Each possible substructure of specified size (minimum and maximum number of atoms per substructure) is identified and the chemical environment of the substructure such as atom type and bond type is recorded as a numeric identifier for each atom. A depth-first search algorithm is applied to each substructure which combines atom identifiers and bond identifiers in the order that they are seen by the algorithm into a single identifier. Lastly, duplicates are identified and summarized and the final fingerprint compressed into a bit vector of predetermined length.

By contrast, mol2vec is an autoencoder inspired by the unsupervised machine-learning algorithm word2vec [Mikolov et al., 2013]. Like ECFP and CSFP, which define a molecule as a sum of its substructures, mol2vec adopts the natural language processing approach of word2vec which embeds all the words in a corpus (perhaps the 250,000 words of the English language) in a much lower dimensional space, typically around 200. First, all substructures of radius 0 and 1 are generated the same way as described above for ECFPs. All resulting identifiers are ordered into a *"molecular sentence"* using the respective canonical SMILES atom order. The network has been trained by Jaeger et al. [2018] on 19.9 million compounds to create a 300-dimensional vector representation for each compound. Just like word2vec, which encodes words with similar meaning to be close together in latent space, mol2vec aims to create a latent space representation where similar molecules are close together. Although, since SMILES strings can differ drastically between two molecules with only small structural differences it is not known how this representation would compare to the topological fingerprints such as CSFP and ECFP which are a more direct representation and can encode chemical similarity extremely well.

All fingerprints were generated as follows. The implemented RDKit function *Get-*

MorganFingerprintAsBitVect of RDKit [Landrum et al., 2006] (v2019.09.1.0) was used to generate ECFPs with a radius of 2 and either 1024 bits or 512 bits vector. CSFPs were generated using the CSFPy script version 0.9.0 [Bellmann et al., 2019] with a lower bound of 2 and an upper bound of 4 and a 1024 bit vector. Mol2vec representations were generated using the pre-trained model supplied by Jaeger et al. [2018] to output a 300-dimensional vector for each compound.

2.2.3 Bayesian Optimization

The Bayesian optimization (BO) algorithm used for this work uses a Gaussian Process (GP) to model the Bayesian response surface. GPs are a *stochastic process* where every finite linear combination of random variables sampled from the process is normally distributed. This allows GP methods, in contrast to classical machine-learning point predictions, to predict a normal distribution with a defined standard deviation and mean. As a result, the prediction also incorporates a measure of uncertainty. Bayesian optimization leverages this fact by defining the *predited improvement* Z :

$$Z = \frac{\mu(x) - f(x^+)}{\sigma(x)} \tag{2.1}$$

Here, $f(x^+)$ is defined as the best target value found so far, $\mu(x)$ is the predicted mean, x its location in the search space, and $\sigma(x)$ the corresponding standard deviation. Next, an acquisition function is defined that provides agency to the BO method and determines which point in the search space to sample next. There are several different acquisition functions that try to balance exploitation (searching close to the current best value) and exploration (searching at a point of highest uncertainty). In this work, the *Expected Improvement* (EI) acquisition function was employed, which is defined as follows:

$$EI(x) = (\mu(x) - f(x^+))\Phi(Z) + \sigma(x)\phi(Z) \tag{2.2}$$

Here, $\Phi(Z)$ and $\phi(Z)$ are the cumulative distribution function and the probability density function, respectively. Equation (2.2) is made up of two terms, with the first term responsible for the exploitation, controlled by the predicted mean; and the second term relating to exploration, which is controlled by the predicted standard deviation. With each iteration, the method then seeks to maximise the value of *EI(x)*. In the context of chemical space, a further distinction between between optimization of continuous versus discrete data has to be made.

Since bioactivity data is often made up of individual data points of compounds tested in the lab, it is discrete by nature. Although efforts have been made to describe chemical space in a continuous matter, for example through the use of variational autoencoders [Gómez-Bombarelli et al., 2018], in this work the search space will be considered to be discrete. As a result, it is more accurate to describe the efforts of the BO method as a prioritization problem, rather than an optimization of a continuous function. This method is known as the *"Multiarmed Bandit Problem"* and rephrases the objective to: "Out of this set of compounds, in what order should they be synthesized and tested so as to discover the best candidates in the fewest iterations". This question directly mimics the traditional medicinal chemistry approach to lead optimization where chemists design a set of compounds based on their prior knowledge of chemistry and the system, then prioritise and test the compounds. For this work, the Bayesian optimization algorithm was implemented using GPyOpt [Consortium, 2020b], an open source Python library based on GPy [Consortium, 2020a], a framework for GP modeling.

Another important consideration in the design of the BO model is the kernel function. The kernel determines the smoothness of the model since it is expected that two points close to each other in the search space, x, also have similar observed values, $f(x)$. In this work, the Radial Basis Function (RBF) kernel is used, which is one of the most commonly used kernels, and is defined as follows:

$$RBF(p, p^*) = a_0 e \frac{-|p - p^*|^2}{2l^2} \tag{2.3}$$

where a_0 and l are the variance and length scale respectively and p and p^* are two points in search space. The *GPy.kern.RBF* function of GPyOpt was used for all compound representations with a lengthscale of 5 and a variance of 1. My implementation of the Bayesian optimisation algorithm can be found on Github `https://github.com/MarcMoesser/Bayesian-Optimization-For-Drug-Discovery`.

2.2.3.1 Performance Evaluation of Bayesian Optimization

Performance of the Bayesian optimization algorithm was evaluated through two different metrics. First, the performance of the optimization task is evaluated by counting the number of compounds explored before the algorithm found the best possible compound in the dataset. As a control, a random sampling algorithm was used, which samples a random compound at each iteration from the dataset. The second evaluation method is to determine the number of compounds found at each iteration that fall within the top 10% of pIC_{50} values in the dataset. Through this approach, it is possible to gauge the enrichment of actives in the dataset, which is a technique used in real-world virtual screening campaigns and structure-activity relationship studies where medicinal chemists are not only interested in the compound with the highest potency, but instead are looking for several compounds with high potency but differ-

ent chemical and physical properties. All Bayesian Optimization and random search experiments were repeated 10 times each with different random number generator seeds.

2.2.4 Tanimoto Similarity and Activity Cliffs

For the evaluation of compound similarity, 1024 bit radius 2 ECFP fingerprints were generated and the RDKit (v2019.09.1.0) TanimotoSimilarity function used to calculate a similarity score between 0 (dissimilar) and 1 (identical). Activity cliffs were calculated using Structure-Activity Similarity (SAS) maps as described by Shanmugasundaram and Maggiora [2001]; Guha [2012]. The activity similarity score $\mathrm{Sim}_{\mathrm{Act}}$ for two molecules, A and B, is calculated as follows:

$$\mathrm{Sim}_{\mathrm{Act}}(A, B) = 1 - \frac{|\mathrm{Act}(A) - \mathrm{Act}(B)|}{|\mathrm{Act}_{\mathrm{max}} - \mathrm{Act}_{\mathrm{min}}|} \tag{2.4}$$

The resulting activity similarity is plotted against the corresponding Tanimoto similarity to create SAS maps such as Figure 2.5.

2.3 Results and Discussion

2.3.1 Analysis of the MBL Dataset

Each MBL subset containing compounds and pIC_{50} values against a single MBL protein were prepared as described in Section 2.2.1. The distribution of pIC_{50} values is shown in Appendix A, Figure A.1. The VIM-2 (Appendix A, Figure A.1 b) and NDM-1 (Appendix A, Figure A.1 d) datasets are visibly skewed towards the higher end of the pIC_{50} spectrum with more high potency ligands than the VIM-1 and IMP-1 datasets, indicating a more challenging SAR for VIM-1 and IMP-1.

Next, the compound similarity in each dataset was analyzed and the set tested for activity cliffs. The Tanimoto similarity between all compound combinations in

each MBL dataset was calculated. The distribution of similarity scores follows approximately a normal distribution in all four datasets (Appendix A, Figure A.2). In order to identity activity cliffs, Structure-Activity-Similarity (SAS) maps [Shanmugasundaram and Maggiora, 2001] were created. SAS maps divide the structure-activity landscape into four distinct zones, as shown in Figure 2.5. The four zones are defined by Shanmugasundaram and Maggiora [2001] as: (I) rough regions (activity cliffs), (II) nondescript regions, (III) scaffold hops and (IV) smooth region. The SAS maps depict each pair of compounds in the dataset, its corresponding SAS score, and their Similarity score. Most of the dataset is distributed between quadrants II, III and IV with the majority of compounds in the "scaffold hops" quadrant (III, i.e. dissimilar compounds but similar activity). The number of compound pairs in zone I (activity cliffs) for VIM-1, VIM-2, IMP-1 and NDM-1 are 115 (0.0009%), 126 (0.0010%), 530 (0.0043%) and 322 (0.0026%), respectively. Thus, none of the datasets show a significant number of activity cliffs. The GP that the Bayesian optimization algorithm is built on can only be used if the underlying function that the GP is simulating is reasonably smooth and does not contain major discontinuities or "spikes". In the case of biological activity of chemical compounds, this would correspond to large changes in biological activity when there are small changes in the structure of the compound. The activity cliff analysis shows that the BO algorithm is appropriate for the MBL dataset as it does not contain a large number of activity cliffs that could disrupt the GP.

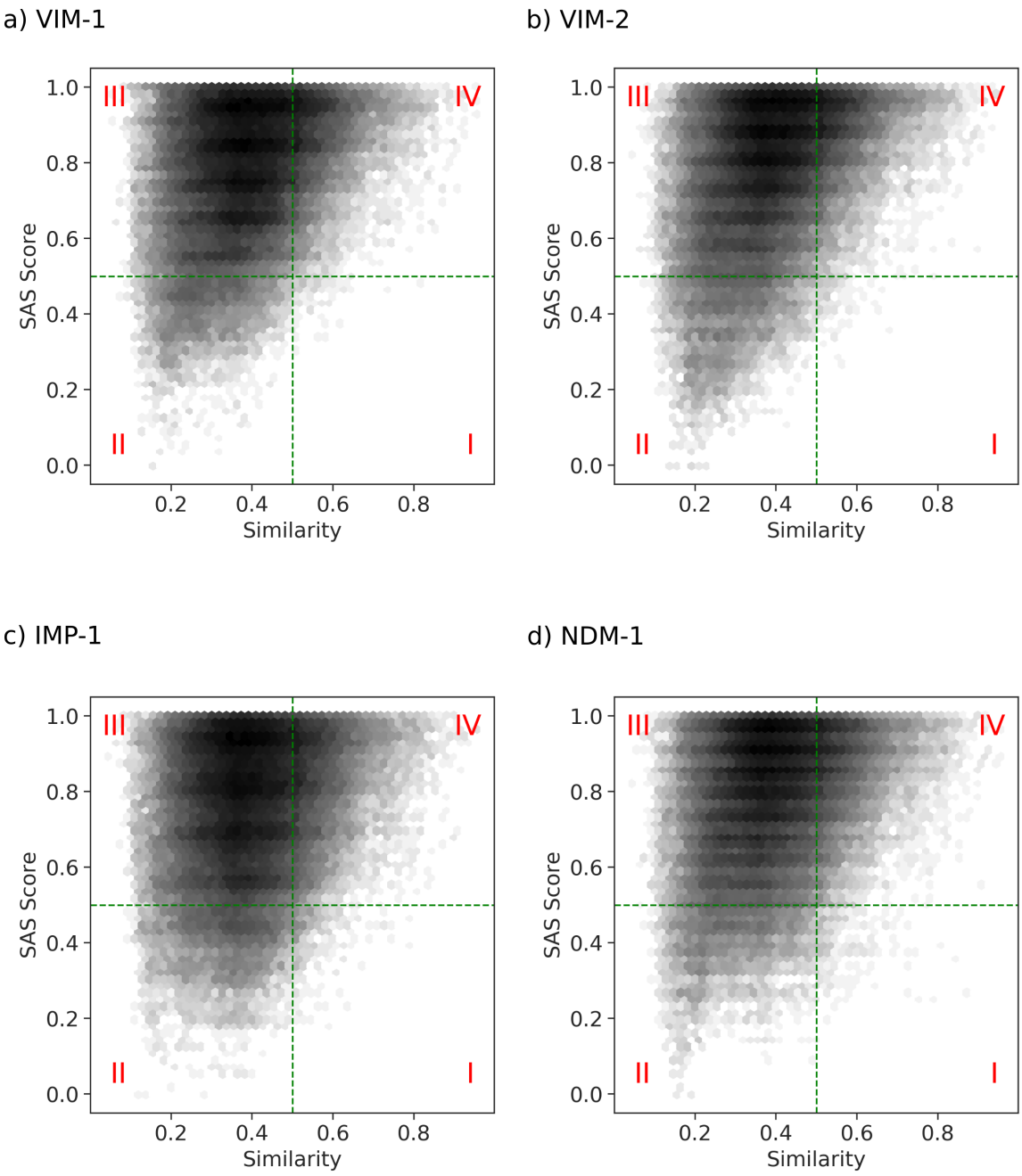

Figure 2.5: SAS maps of each MBL inhibition dataset using logarithmic scaling on the bin density, with darker bins corresponding to more compound pairs. The quadrants I, II, III and IV correspond to activity cliffs, non-descript region, scaffold hops and smooth regions, respectively (as defined by Shanmugasundaram and Maggiora [2001]). The number of compound pairs in the activity cliff zone I for VIM-1, VIM-2, IMP-1 and NDM-1 are (a) 115 (0.09% of total 122265 combinations), (b) 126 (0.09% of total 138601 combinations), (c) 530 (0.37% of total 141778 combinations) and (d) 322 (0.21% of total 150975 combinations), respectively.

2.3.2 Bayesian Optimization Model Performance

In order to analyse the performance of Bayesian optimization (BO), two plots were generated for each experiment to show the raw optimization power (number of iterations required to find the best compound in the dataset by pIC_{50}, Figure 2.6 a) and the ability of the algorithm to identify the most potent compounds (number of compounds that fall into the top 10% of pIC_{50} values that had been found at each iteration, Figure 2.6 b). The results are plotted as follows: the results for the first 200 iterations are shown and plots are labelled to identify the protein target, feature, kernel and acquisition function. The RBF kernel and EI acquisition function were used for all experiments. All experiments were repeated 10 times, the mean plotted in bold and the 95% confidence intervals as shaded regions. The different molecular representations are color coded: ECFP 1024 bits radius 2: blue; ECFP 512 bits radius 2: dark green; CSFP: orange; mol2vec: light green; and random sampling: red.

First, in order to validate my BO implementation and to compare it to previous results on the MMP-12 dataset obtained by Pyzer-Knapp [Pyzer-Knapp, 2018], the BO method was run on the same dataset. As shown in Figure 2.6, the BO algorithm is able to discover the best molecule in the dataset after fewer than 200 iterations using ECFP 1024 bits, which significantly outperforms Pyzer-Knapp's method, where the best molecule was found after over 800 iterations [Pyzer-Knapp, 2018] despite Pyzer-Knapp using the EI acquisition function, RBF kernel and ECFP 512 bit radius 2. However, it is not clear from the paper how exactly the method was implemented and no code was supplied. In comparison, the mol2vec representation seems to be the best performing method in this scenario, although not by a large margin.

Since Pyzer-Knapp only used ECFP 512 bits representations, the effect of varying the fingerprint bit vector length of ECFPs was investigated. However, no significant difference between larger 1024 bit radius 2 ECFP fingerprints and smaller 512 bit

radius 2 fingerprints were observed when comparing the general performance of the method, although 1024 bit radius 2 ECFPs slightly outperformed the shorter variant when discovering top compounds from the dataset (Appendix A, Figures A.3 - A.6). For all future experiments, 1024 bit radius 2 ECFPs were used in the method and labelled just as "ECFP".

2.3.3 Bayesian Optimization against MBL Targets

2.3.3.1 Performance Against a Single Protein Target

After establishing the performance of my implementation of the BO algorithm against past benchmarks in the MMP-12 dataset, the method was applied to the MBL dataset. All variations of the Bayesian optimization method are able to find the best compound in fewer than 175 iterations, with no large difference between the various compound representations when looking at the general performance, although ECFPs are able to identify the best compound fastest in all cases (Figures 2.7-2.10 a). A substantial gap in performance between molecular representations is revealed when looking at the second performance indicator: the number of "good" compounds obtained at each iteration (Figures 2.7-2.10 b). While optimization against VIM-1 and NDM-1 (Figure 2.7 b and Figure 2.10 b) show no significant difference, optimization of VIM-2, and especially, IMP-1 activity is significantly more efficient using 1024 bit ECFPs. Interestingly, while mol2vec and CSFP perform almost identically for 3 of the 4 MBL targets, performance of CSFP drops significantly after the initial 60 iterations in comparison to ECFP and mol2vec when optimizing against NDM-1 (Figure 2.10). In all cases, Bayesian optimization outperforms random sampling, which is not able to find the best compound in the dataset within 200 iterations. A larger difference in performance between Bayesian optimization and random sampling can be seen when looking at the enrichment of top compounds, where the ECFP method, for example,

is able to find around double the number of "good" compounds after 200 iterations (Figures 2.7-2.10 b).

2.3.3.2 Simultaneous Optimization Against Two Protein Targets

After testing Bayesian optimization against single MBL targets, multi-task optimization was explored by calculating the geometric mean of the corresponding pIC_{50} values of each compound between two MBL targets and using new datasets for each pairwise combination of proteins (6 total: IMP-1+VIM-1, IMP-1+VIM-2, NDM-1+IMP-1, NDM-1+VIM-1, NDM-1+VIM-2, VIM-2+VIM-1). In comparison to the results for the single target optimization, a larger difference in general optimization performance between the compound representations can be observed in the multi-objective approach (Figures 2.11 - 2.16). Overall, ECFPs perform best, especially for the IMP-1+VIM-2 combination (Figure 2.12) where the ECFP method reaches the maximum in fewer than 50 iterations, as opposed to over 100 for the other methods. The overall performance of BO against the multi-objective datasets is increased in comparison to the single target optimizations, with all multi-objective experiments reaching the maximum in fewer iterations than single target optimization. For the enrichment of "good" compounds in the case of the multi-objective optimization, the same trend in performance can be seen with ECFPs performing best, followed by mol2vec and CSFPs performing similarly.

In summary, the investigation of the impact of different compound representations on the performance of Bayesian optimization shows that although the performance of all three representations are very similar, 1024 bit radius 2 ECFPs outperform the other two representations. In general, Bayesian optimization is able to find the best compound in the dataset quickly, and most importantly, to greatly enrich the number of "good" compounds in comparison to random exploration.

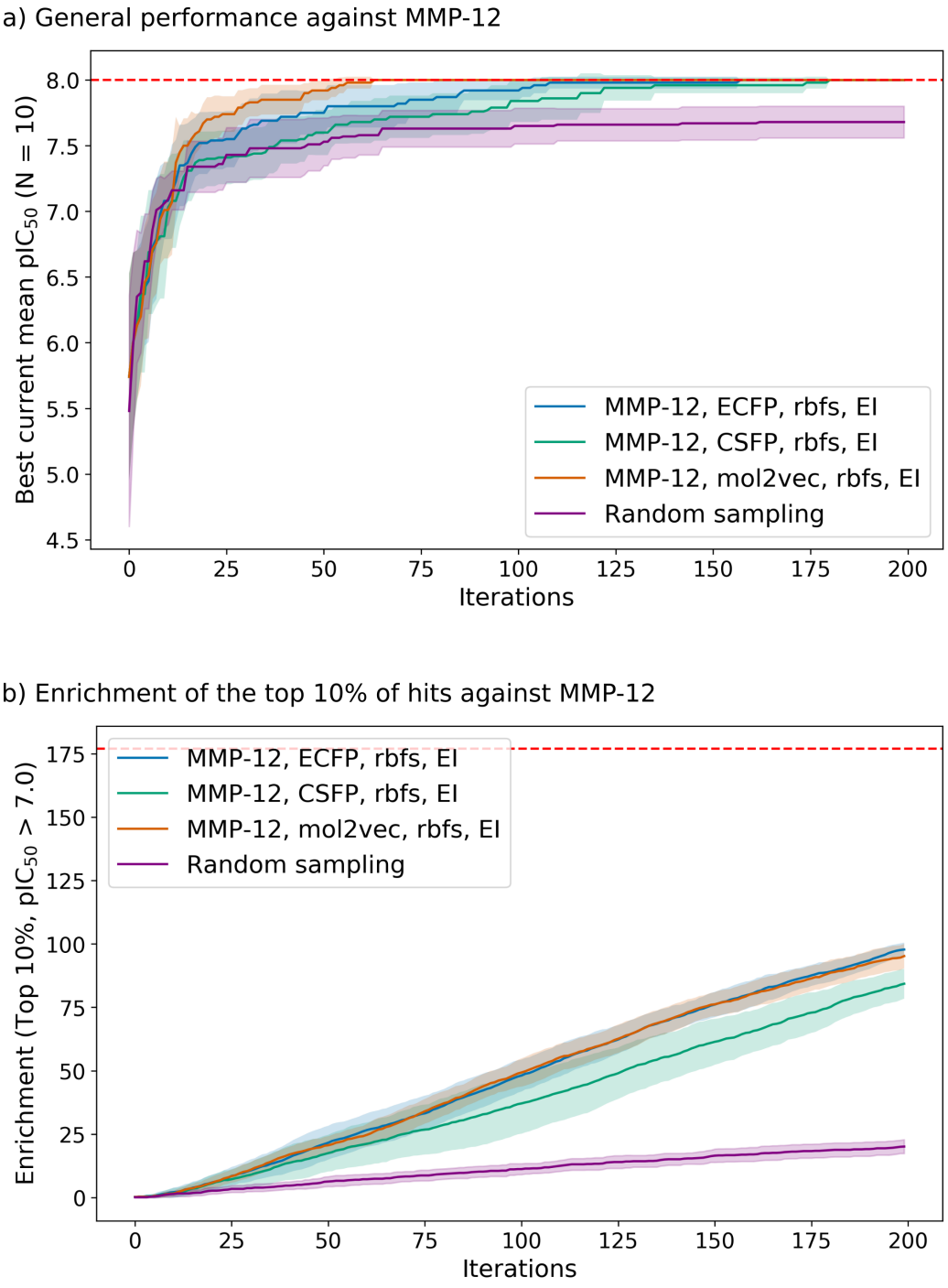

Figure 2.6: Performance of Bayesian optimization with the MMP-12 dataset using the ECFP, CSFP and mol2vec representations, versus random sampling. The dashed red line indicates the maximum possible value for both: the maximum pIC_{50} and the maximum count of top compounds; bold lines indicate the mean of 10 experiments and the shaded area is the 95% confidence interval (CI). a) General performance plot showing the best current mean pIC_{50} found over 200 iterations of BO or random sampling. b) Enrichment plot showing the total count of top compounds found so far at each iteration. The best performing methods are the ECFP and mol2vec implementation.

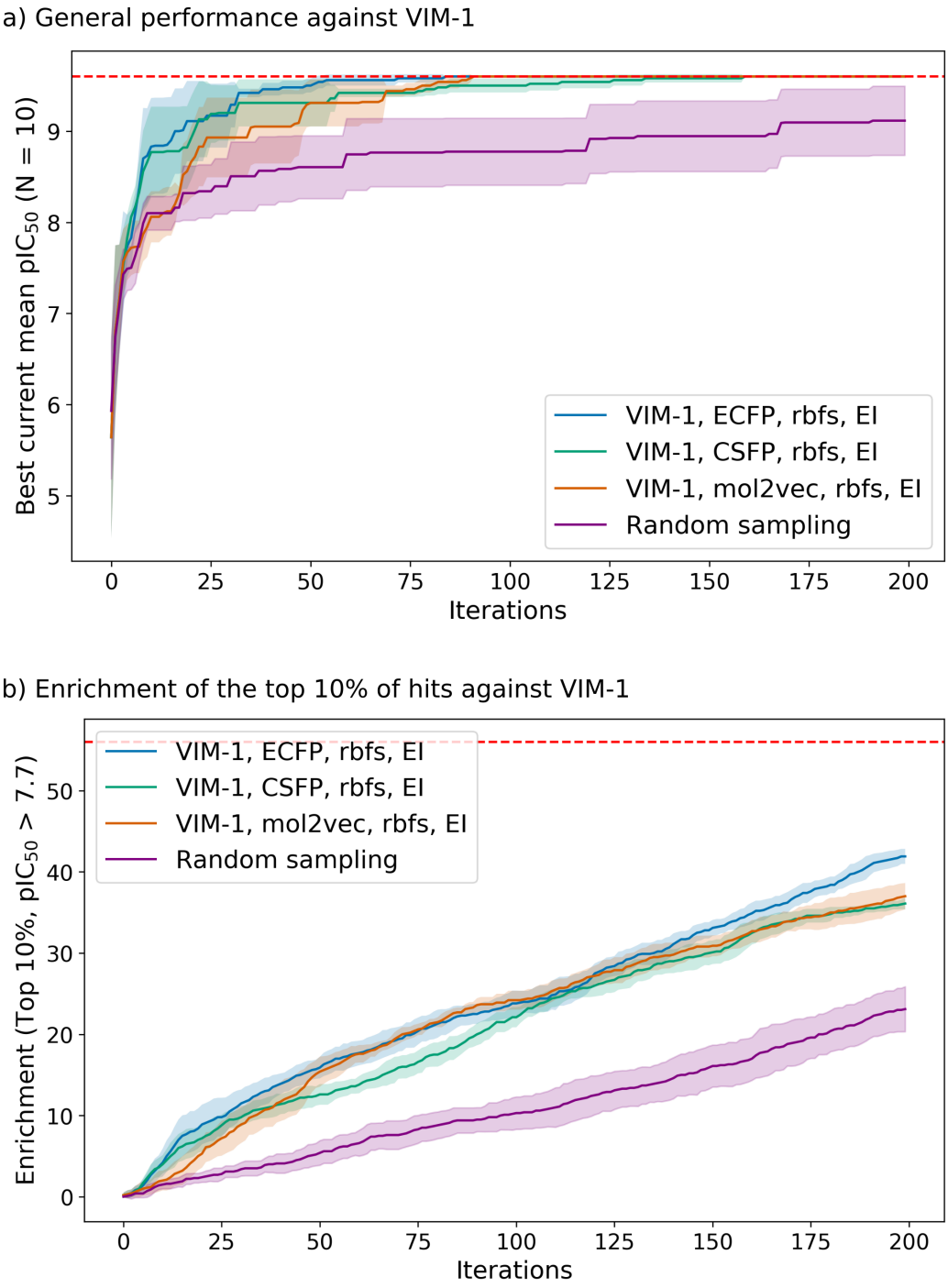

Figure 2.7: Performance of Bayesian optimization using the VIM-1 dataset. The dashed red line indicates the maximum possible value for both: the maximum pIC_{50} and the maximum count of top compounds; bold lines indicate the mean of 10 experiments and shaded area the 95% CI. a) General performance plot showing the best current mean pIC_{50} found over 200 iterations of BO or random sampling. b) Enrichment plot showing the total count of top compounds found so far at each iteration. The ECFP implementation performs best, but all BO methods perform significantly better than random search.

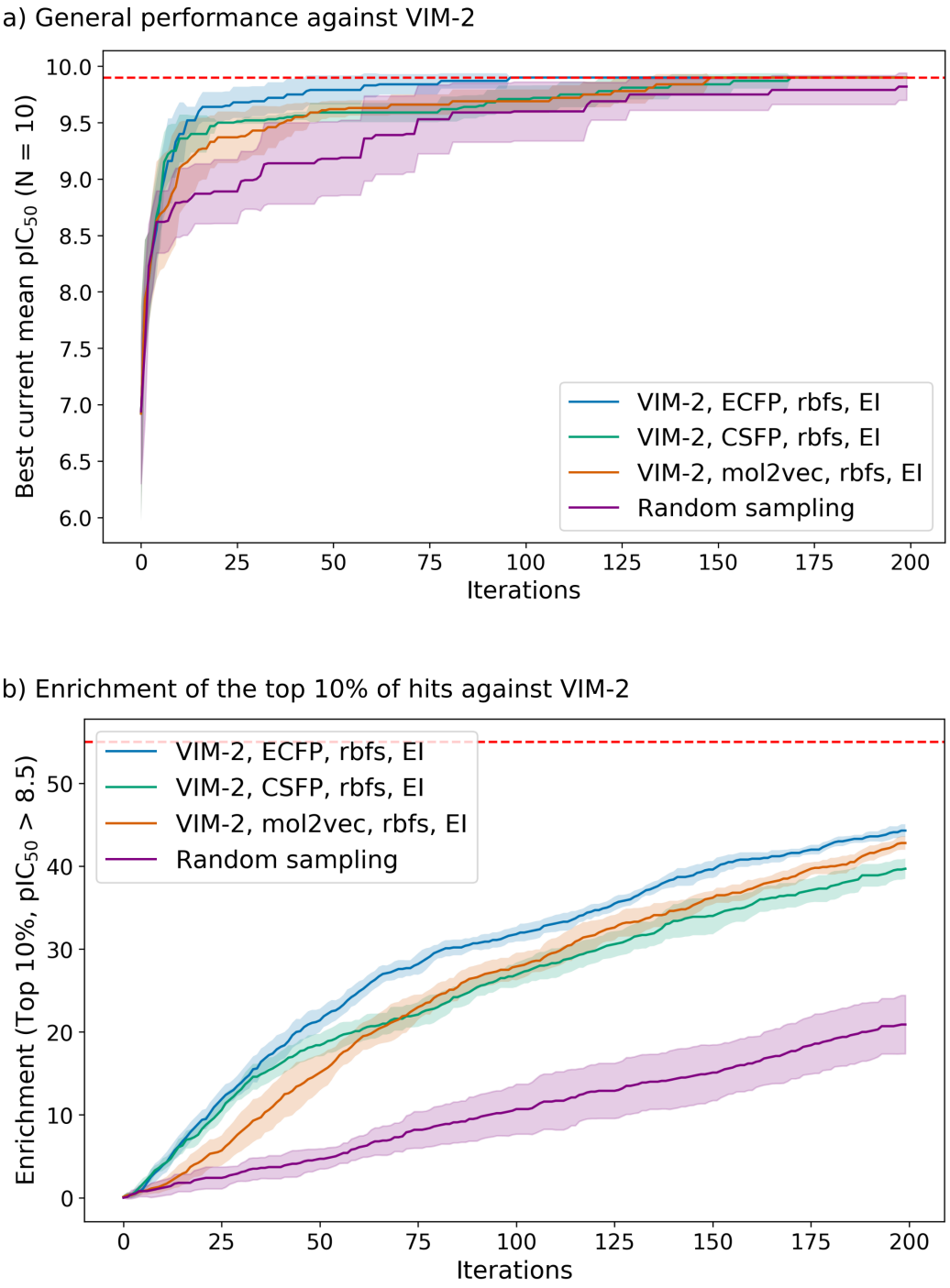

Figure 2.8: Performance of Bayesian optimization using the VIM-2 dataset. The dashed red line indicates the maximum possible value for both: the maximum pIC_{50} and the maximum count of top compounds; bold lines indicate the mean of 10 experiments and shaded area the 95% CI. a) General performance plot showing the best current mean pIC_{50} found over 200 iterations of BO or random sampling. b) Enrichment plot showing the total count of top compounds found so far at each iteration. The ECFP implementation performs best, but all BO methods perform significantly better than random search.

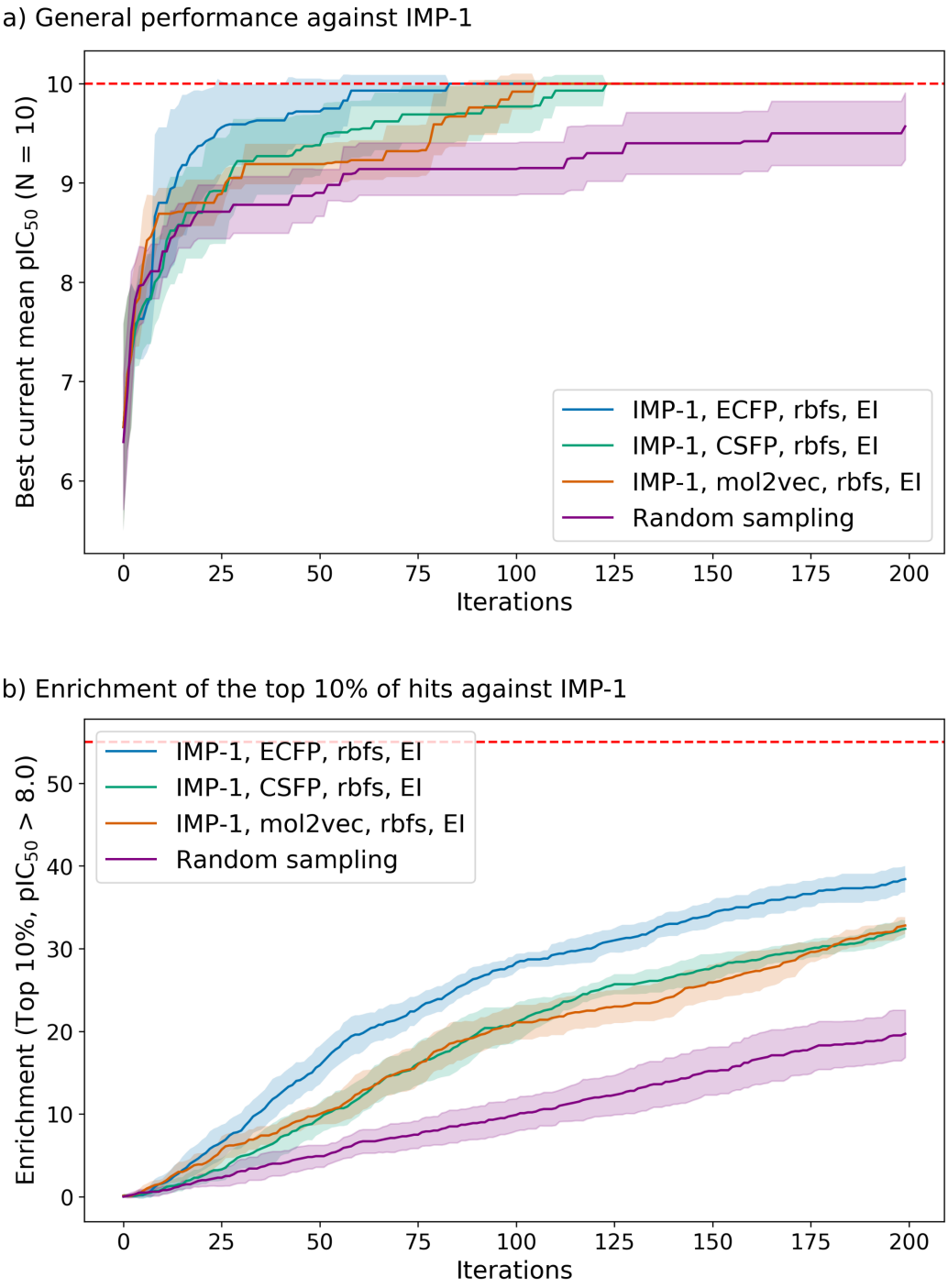

Figure 2.9: Performance of Bayesian optimization using the IMP-1 dataset. The dashed red line indicates the maximum possible value for both: the maximum pIC_{50} and the maximum count of top compounds; bold lines indicate the mean of 10 experiments and shaded area the 95% CI. a) General performance plot showing the best current mean pIC_{50} found over 200 iterations of BO or random sampling. b) Enrichment plot showing the total count of top compounds found so far at each iteration. The ECFP implementation performs best, but all BO methods perform significantly better than random search.

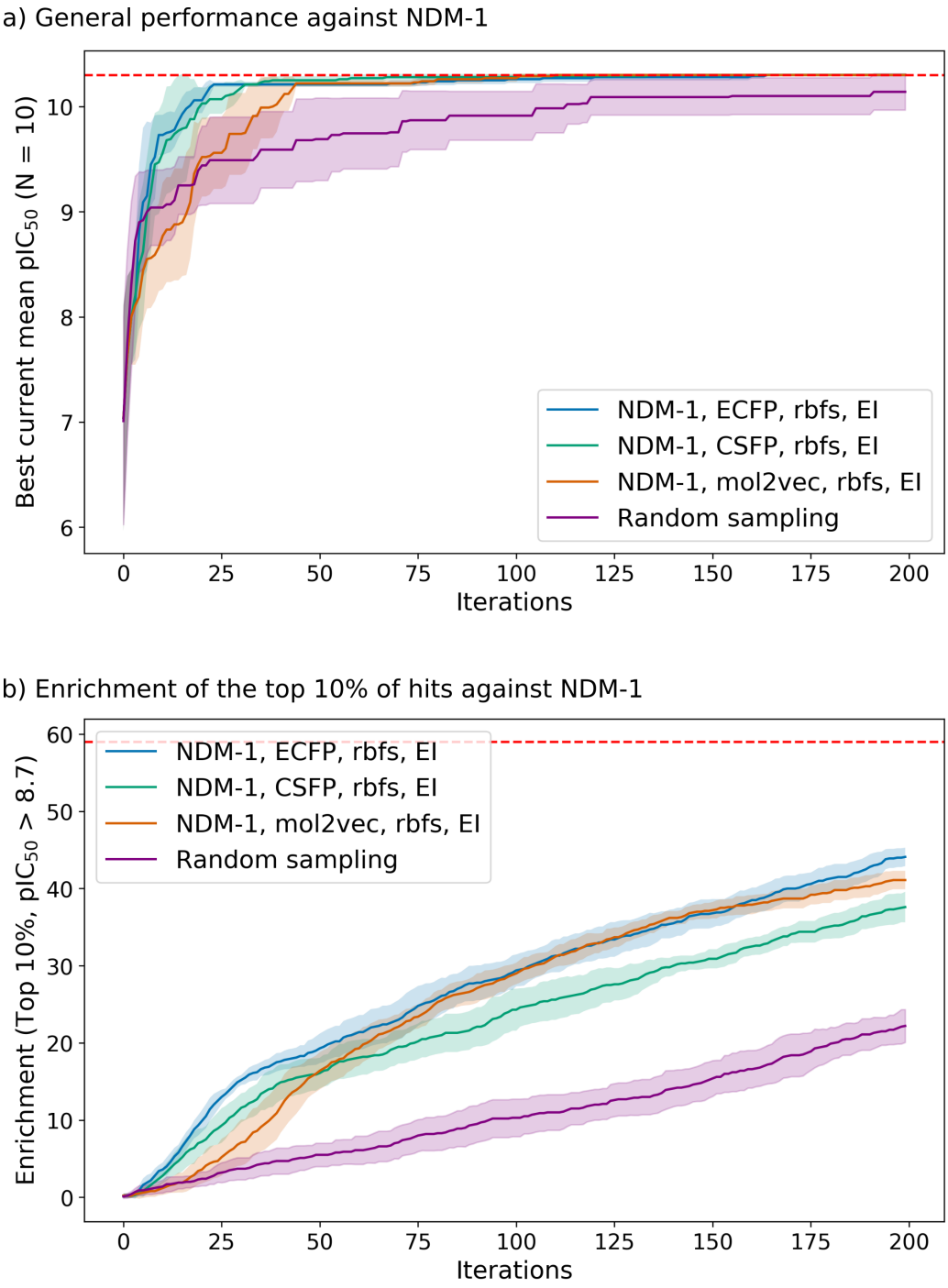

Figure 2.10: Performance of Bayesian optimization using the NDM-1 dataset. The dashed red line indicates the maximum possible value for both: the maximum pIC_{50} and the maximum count of top compounds; bold lines indicate the mean of 10 experiments and shaded area the 95% CI. a) General performance plot showing the best current mean pIC_{50} found over 200 iterations of BO or random sampling. b) Enrichment plot showing the total count of top compounds found so far at each iteration. The ECFP implementation performs best, but all BO methods perform significantly better than random search.

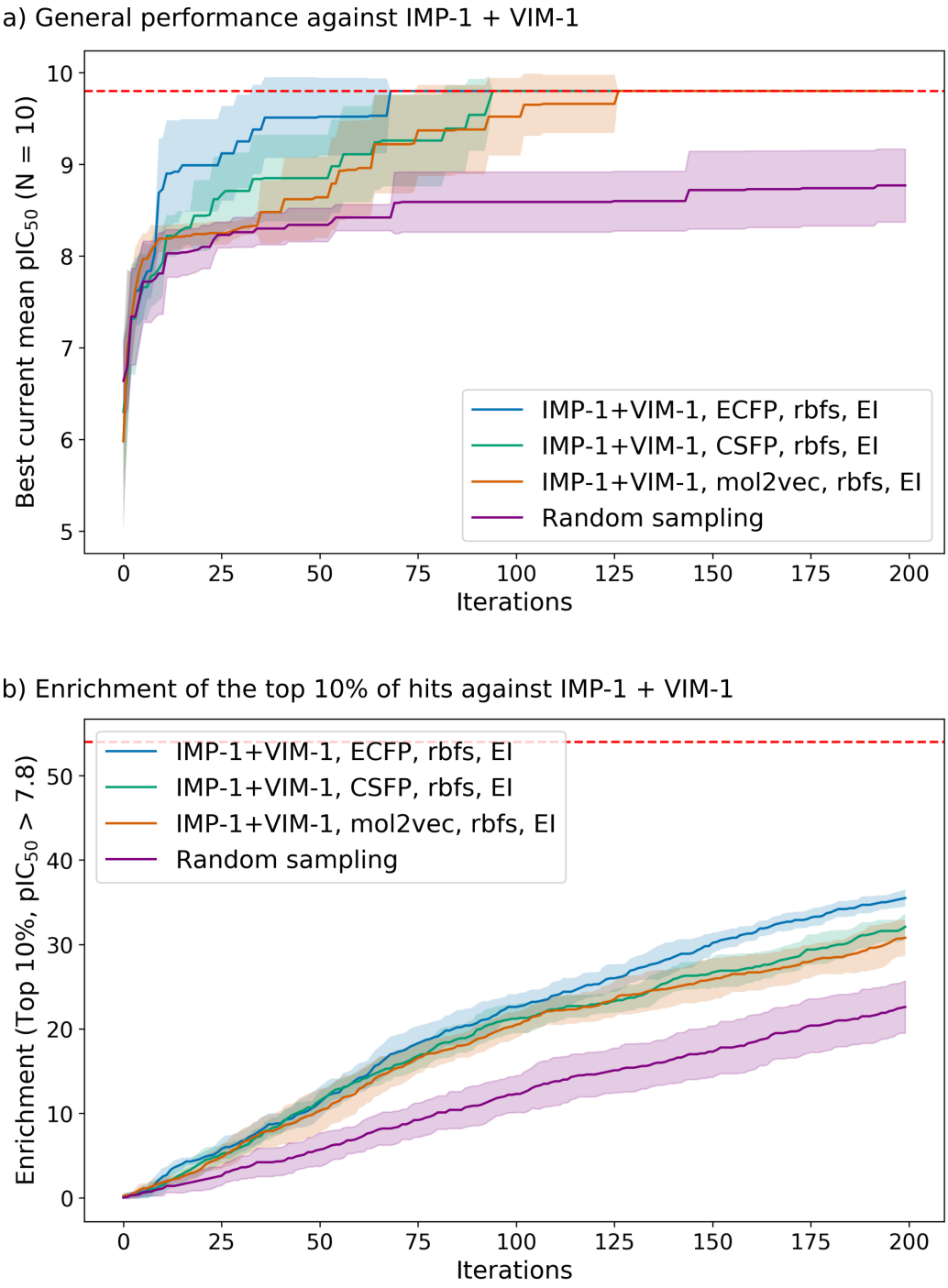

Figure 2.11: Multi-objective optimization performance of Bayesian optimization using the combined IMP-1 + VIM-1 dataset. The dashed red line indicates the maximum possible value for both: the maximum pIC_{50} and the maximum count of top compounds; bold lines indicate the mean of 10 experiments and shaded area the 95% CI. a) General performance plot showing the best current mean pIC_{50} found over 200 iterations of BO or random sampling. b) Enrichment plot showing the total count of top compounds found so far at each iteration. The ECFP implementation performs best, but all BO methods perform significantly better than random search.

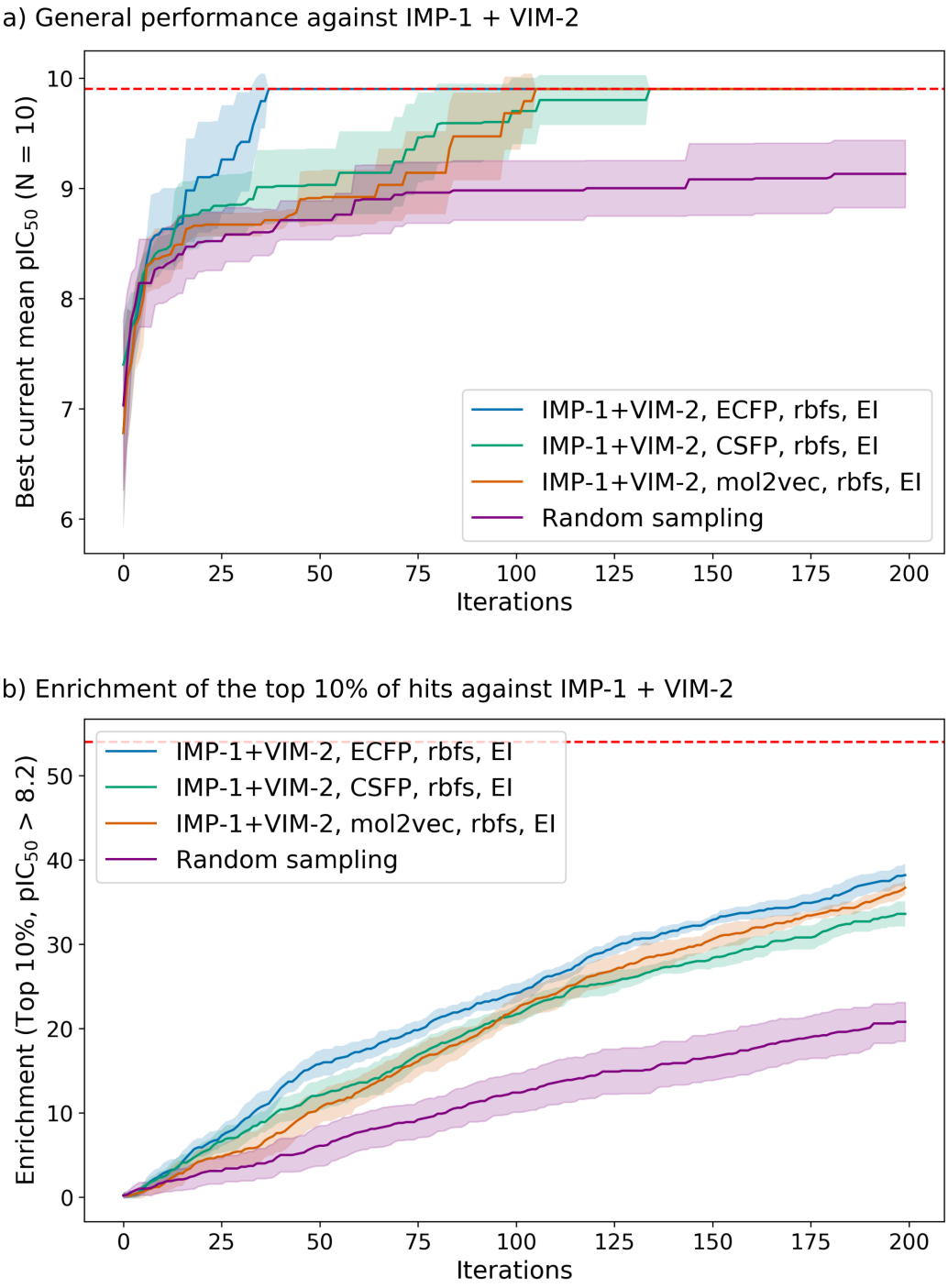

Figure 2.12: Multi-objective optimization performance of Bayesian optimization using the combined IMP-1 + VIM-2 dataset. The dashed red line indicates the maximum possible value for both: the maximum pIC_{50} and the maximum count of top compounds; bold lines indicate the mean of 10 experiments and shaded area the 95% CI. a) General performance plot showing the best current mean pIC_{50} found over 200 iterations of BO or random sampling. b) Enrichment plot showing the total count of top compounds found so far at each iteration. The ECFP implementation performs best, but all BO methods perform significantly better than random search.

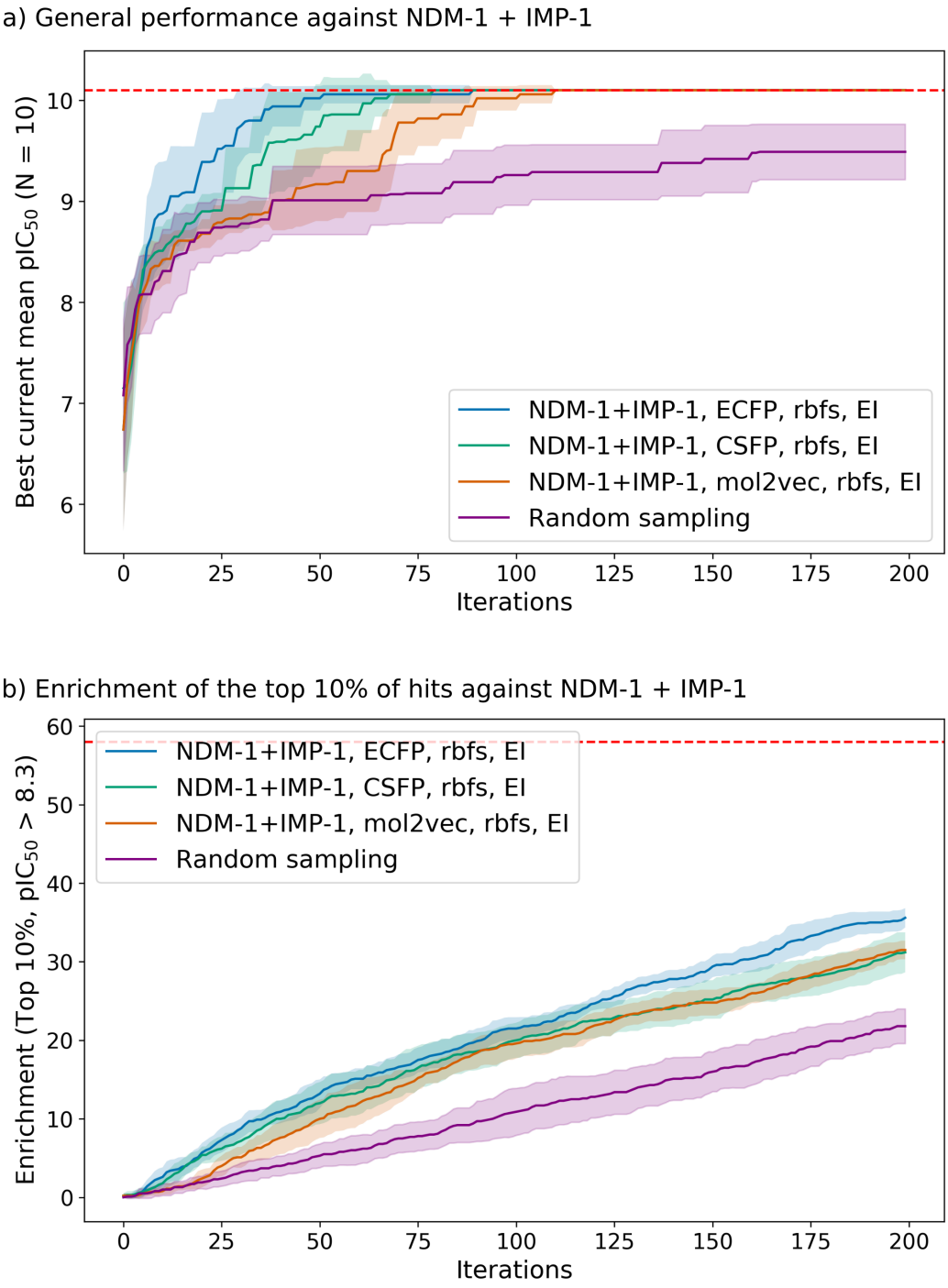

Figure 2.13: Multi-objective optimization performance of Bayesian optimization using the combined NDM-1 + IMP-1 dataset. The dashed red line indicates the maximum possible value for both: the maximum pIC_{50} and the maximum count of top compounds; bold lines indicate the mean of 10 experiments and shaded area the 95% CI. a) General performance plot showing the best current mean pIC_{50} found over 200 iterations of BO or random sampling. b) Enrichment plot showing the total count of top compounds found so far at each iteration. The ECFP implementation performs best, but all BO methods perform significantly better than random search.

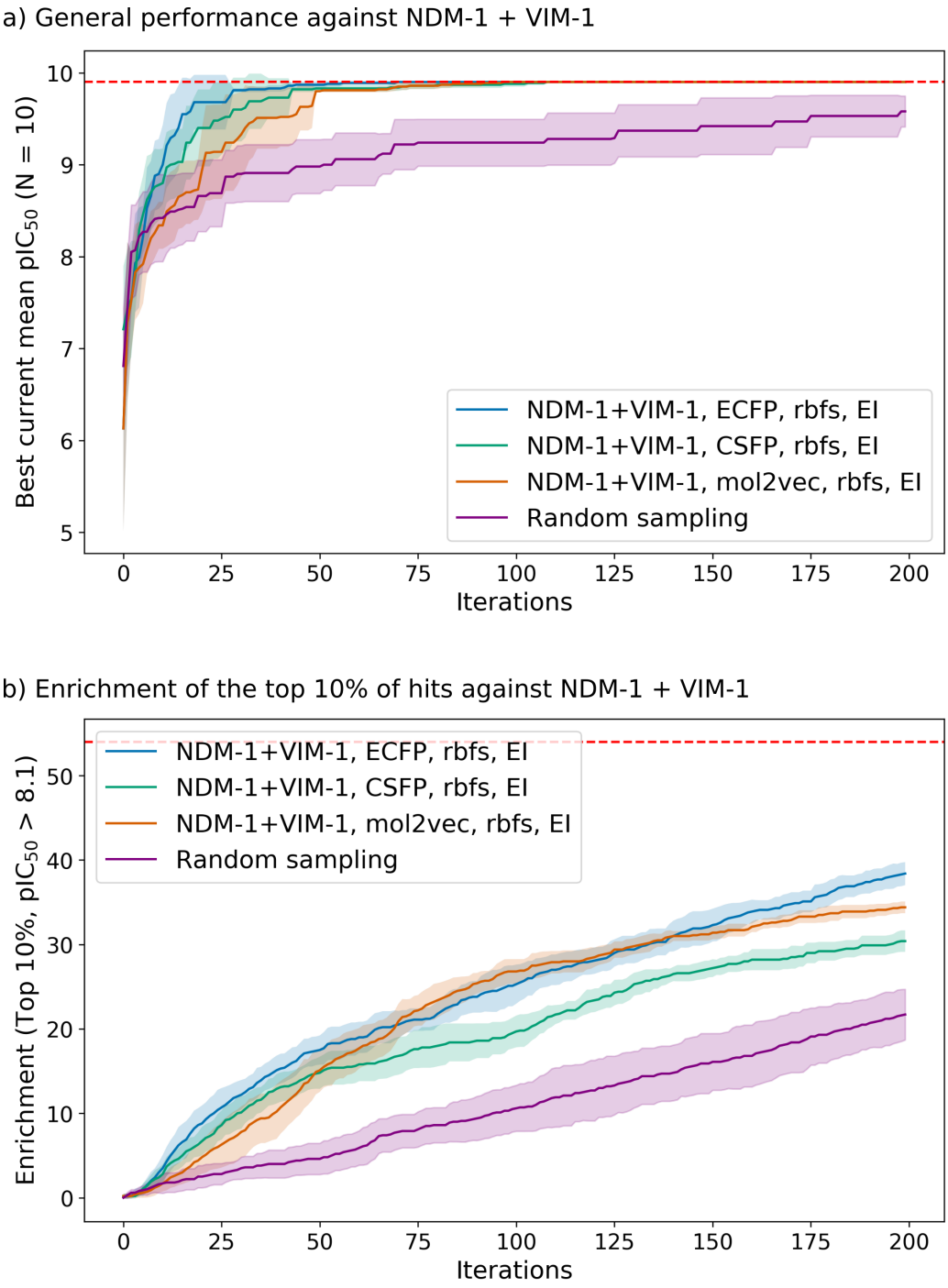

Figure 2.14: Multi-objective optimization performance of Bayesian optimization using the combined NDM-1 + VIM-1 dataset. The dashed red line indicates the maximum possible value for both: the maximum pIC_{50} and the maximum count of top compounds; bold lines indicate the mean of 10 experiments and shaded area the 95% CI. a) General performance plot showing the best current mean pIC_{50} found over 200 iterations of BO or random sampling. b) Enrichment plot showing the total count of top compounds found so far at each iteration. The ECFP implementation performs best, but all BO methods perform significantly better than random search.

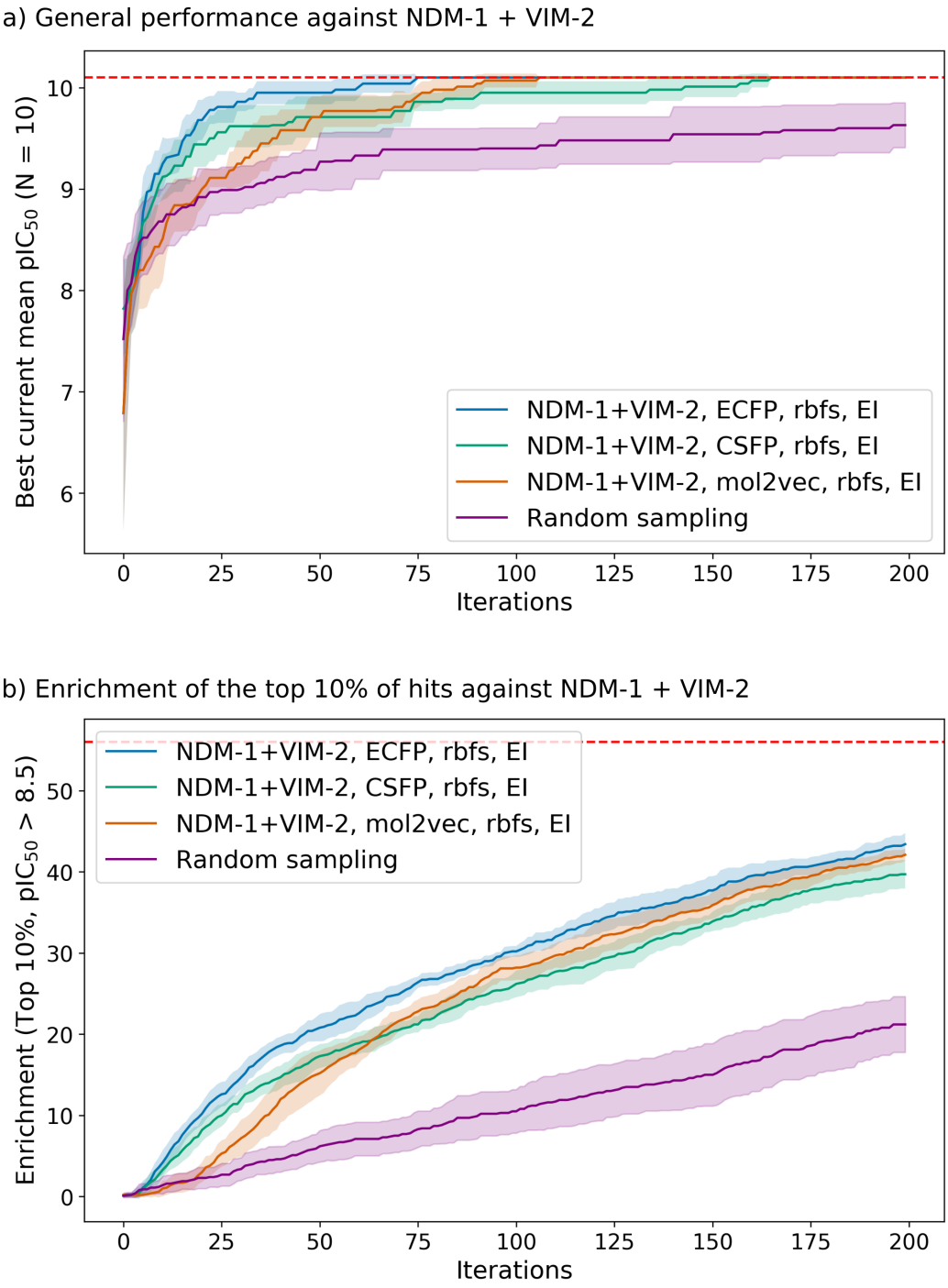

Figure 2.15: Multi-objective optimization performance of Bayesian optimization using the combined NDM-1 + VIM-2 dataset. The dashed red line indicates the maximum possible value for both: the maximum pIC_{50} and the maximum count of top compounds; bold lines indicate the mean of 10 experiments and shaded area the 95% CI. a) General performance plot showing the best current mean pIC_{50} found over 200 iterations of BO or random sampling. b) Enrichment plot showing the total count of top compounds found so far at each iteration. The ECFP implementation performs best, but all BO methods perform significantly better than random search.

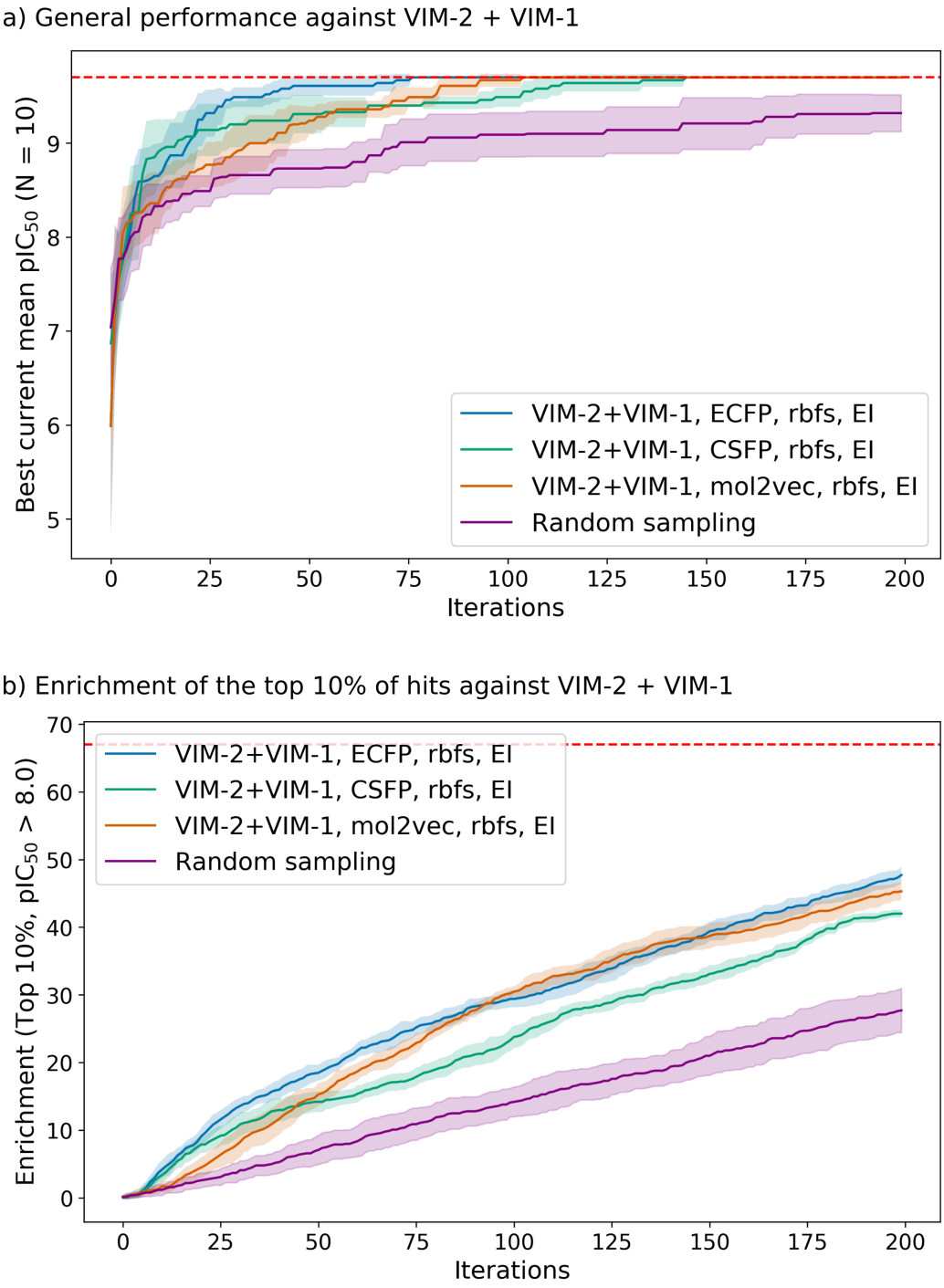

Figure 2.16: Multi-objective optimization performance of Bayesian optimization using the combined VIM-2 + VIM-1 dataset. The dashed red line indicates the maximum possible value for both: the maximum pIC_{50} and the maximum count of top compounds; bold lines indicate the mean of 10 experiments and shaded area the 95% CI. a) General performance plot showing the best current mean pIC_{50} found over 200 iterations of BO or random sampling. b) Enrichment plot showing the total count of top compounds found so far at each iteration. The ECFP implementation performs best, but all BO methods perform significantly better than random search.

2.4 Discussion

The primary goal of the experiments was first to validate the BO method against a known dataset (MMP-12); second, to apply the method to a novel, real-life drug optimisation dataset, the MBL dataset obtained from the Schofield lab at the University of Oxford which is the result of an ongoing effort to discover new antibiotics; and third, to investigate the impact of different molecular representations of chemical space on the performance of BO. The BO method was analyzed using two different performance metrics: the general power of the optimization algorithm as measured by the number of iterations needed to find the most potent compound (*i.e.* maximum pIC_{50}) in the dataset; and second, the enrichment of potent compounds, as measured by the total number of potent compounds (top 10% of pIC_{50} values in the dataset) found at each iteration.

In general, the difference between random sampling and BO in all cases was found to be larger when looking at the second performance metric (the identification of "good" compounds) than the first metric (raw optimization to find the best compound in the dataset). This second metric is also arguably more important in an early stage lead discovery or lead optimization project, where the discovery of a single high affinity compound is less important than the discovery of a series of compounds, all with good activity, which opens up a more diverse base to optimize desirable properties such as ADME properties, solubility, toxicology, and cell permeability.

However, on a methodological level, the current version of the BO model has the limitation that it chooses only one molecule per iteration, which is less applicable in a traditional lab setting. Normally, compounds are synthesized and assayed in batches for efficiency. In order to make the BO method as realistic as possible, future work should include the improvement of the algorithm to enable batches of molecules to be chosen at each iteration.

Since the BO surrogate function is able to update fewer times and is therefore forced to choose new compounds with less information when choosing molecules in batches, the resulting loss of information might lead to decreased performance when compared to single molecule iterations. Nonetheless there are several ways to implement the batch sampling process beyond just variation in batch sizes. For example, exploration could be prioritized early in the optimization process to collect as much information as possible before prioritizing exploitation to find the best compound. However, the performance of the current version of BO is still a good indication of its ability to create a realistic surrogate function to model the correlation between multi dimensional fingerprints and biological activity.

The MMP-12 dataset was primarily used in this work to serve as a benchmark for the BO algorithm and to compare it to previously described results by Pyzer-Knapp [Pyzer-Knapp, 2018]. The results obtained from the BO method on the MMP-12 dataset is interesting (Figure 2.6, given that Pyzer-Knapp reported a very similar method with significantly worse performance. [Pyzer-Knapp, 2018] However, since Pyzer-Knapp has not made the details of the implementation of his BO method public, it is unknown what could have caused the difference in performance. One aspect I found to greatly influence the performance of the BO algorithm while building the GPyOpt implementation is the RBF kernel function. I found the length-scale restriction could drastically diminish sampling performance to levels worse than random sampling. In summary, I have shown that my method has passed the benchmark and surpasses Pyzer-Knapp's method. [Pyzer-Knapp, 2018]

Next, the BO method was applied to the MBL dataset which was split into four single target datasets (VIM-1, VIM-2, IMP-1, NDM-1) and 6 dual-objective datasets (IMP-1+VIM-1, IMP-1+VIM-2, NDM-1+IMP-1, NDM-1+VIM-1, NDM-1+VIM-2

and VIM-2+VIM-1). An interesting feature of the single target datasets is the underlying distribution of pIC_{50} values in each set (Appendix A, Figure A.1). The distribution of VIM-2 and NDM-1 is shifted towards more potent pIC_{50} values in comparison to VIM-1 and IMP-1 (Figure A.1. One reason for the shift could be a more challenging SAR for VIM-1 and IMP-1 (medicinal chemists on the project were able to discover less high affinity ligands), which highlights the need for additional tools to support medicinal chemists in the design of new high affinity ligands against challenging targets. The shift in the pIC_{50} distribution could also directly affect the performance evaluation of the BO method, since the second performance metric relies on classifying good compounds as the top 10% of pIC_{50} values in each dataset (*i.e.* for VIM-1: $pIC_{50} > 7.7$, for IMP-1: $pIC_{50} > 8.0$, for VIM-2: $pIC_{50} > 8.5$ and for NDM-1: $pIC_{50} > 8.7$). Since the top 10% threshold is applied locally on each dataset rather than globally across all MBL datasets, this leads to a higher threshold in the case of the VIM-2 and NDM-1 datasets due to the greater number of high affinity ligands in the dataset. A dynamic adjustment of the performance metric through local thresholds as described above is advantageous for the purpose of benchmarking, since it sets a higher bar to datasets where it is statistically more likely to find high affinity compounds. Without the dynamic top 10% metric, a comparison between the performance of BO against two datasets with a different pIC_{50} distribution would be heavily biased. The BO algorithm would be more likely (purely statistically) to find high affinity ligands in a dataset with a shifted $pIC50_{50}$ distribution towards the top end of pIC_{50} scores, making direct comparisons between datasets convoluted.

The performance of the BO algorithm was similar between all four single target datasets (Figure 2.7-2.10) as well as the six multi-objective approaches (Figure 2.11-2.16) showing that the BO algorithm is applicable to different proteins and dataset distributions.

Since the method is purely ligand based and no information of the protein target is included except the corresponding pIC_{50} values, it might be expected that the method will perform equally well against all four targets or combinations thereof, since the compounds included in each subset are identical and no large activity cliffs were found in any of the sets. However, while this is observed in most MBl datasets (and combinations thereof), as was discussed above, differences in SAR for some protein targets might produce biases in the underlying dataset, resulting in different performance on different datasets. One such exception was observed when optimizing against the protein pair IMP-1+VIM-2, where the ECFP-based method was able to identify the best compounds in under 50 iterations (Figure 2.12 a), significantly faster than any other method or dataset. However, the difference in performance disappears when looking at the more important second metric (Figure 2.12 b) where not only the other compound representations perform comparably, but also no significant performance difference can be observed when compared to the other dual protein datasets. This observation highlights the importance of analyzing the optimization results from a different angle than purely greedy optimization. In general, the 1024 bit radius 2 ECFPs did substantially better in the second performance metric in almost all single and multi-objective datasets than the CSFP and Mol2vec representation. Although in isolated instances, such as in the case of NDM-1, (Figure 2.10) and VIM-2+VIM-1, (Figure 2.16) the Mol2vec and ECFP representations perform comparably and stay within the 95% confidence intervals of each other. The more similar performance of ECFP and Mol2vec could stem from the fact that Mol2vec used ECFPs as molecular "words" to train the Mol2vec model. The poorer performance of CSFPs could potentially be due to information loss inherent to the CSFP creation process, when the extremely high dimensional fingerprint is compressed into 1024 bit vectors. While this is true for ECFPs as well, the way CSFPs are designed increases the severity

of the effect. The main feature of CSFPs is the ability to exhaustively enumerate every possible subgraph of a given size within a molecule, instead of just each atom and its surrounding area as is done with ECFPs. A 1024 bit vector might not be sufficient to capture the increased information density as compared to ECFPs. Even for ECFPs, a slight increase in performance was observed when comparing 512 bit and 1024 ECFPs (Appendix A, Figures A.3 - A.6).

A potential avenue for further exploration of these results would be to increase the bit size for CSFPs, although a larger fingerprint would also increase the computational power required for the BO algorithm, so the trade-off would need to be considered carefully. In addition, the subgraph size of the CSFPs is a hyperparameter that has to be fine tuned since, it was originally reported that CSFP subgraph size does have an impact on machine learning model performance when using CSFP [Bellmann et al., 2019]. For this work, a subgraph with a lower bound of 2 and an upper bound of 4 was used to capture subgraphs with between 2 and 4 atoms. Although this is the recommended range, fine tuning could be done to optimize the these values. Overall, my results whow that BO is able to be used with different ligand-based representations including fingerprints as well as molecular autoencoders with ECFP performing best.

There are two high-level areas for future improvement. First, a variational autoencoder similar to the VAE published by Gómez-Bombarelli et al. [2018] that captures both the chemical structure of a molecule and its properties in a continuous latent space representation would enable the algorithm to generate novel compounds by adding Gaussian noise to the latent space representation of an identified top compound, or by interpolating between two potent compounds. A similar vector in the latent space could be decoded by the autoencoder to create a new molecule that is similar to the original compound and potentially close in activity as well, in ac-

cordance with the molecular similarity principle [Duran-Frigola et al., 2020]. This process could be used to generate a diverse set of active ligands based on the current best known compound. Since I have demonstrated that the mol2vec autoencoder is a viable molecular representation for BO, using latent space compound representations in a continuous BO instead of the multi-armed bandit described herein should be feasible.

A second area of improvement would be to continue the exploration of different compound representations. For example, an autoencoder could be trained to create a vector representation that is specifically designed to optimize the performance of the Bayesian optimization algorithm by using a modified outcome of the optimization as the loss function for training. Additionally, it is possible to replace the GP as the surrogate function for the BO algorithm. Since the current implementation uses a simple ligand-based scoring function for binding affinity based on a GP alone, replacing the surrogate function with a more accurate Bayesian neural network, or a modified neural network that is able to output point predictions as well as the associated uncertainty could increase the performance of the overall optimization. First steps have been made in this direction by Dominik Klein, who conducted research for his Master thesis under my supervision on the implementation of final-layer GP models into graph-based neural networks for affinity prediction.

As the work described here was coming to a conclusion, a global pandemic was declared by the World Health Organization. With the agreement of my supervisors, it was decided to change focus to a new target and sub-project: the SARS-COV-2 main protease M^{pro} and the discovery of novel peptide inhibitors as well as the development of a *active guided docking* method described in Chapter 3.

Chapter 3

In silico Design and Validation of SARS-CoV-2 M^{pro} Inhibitors from Modelling Substrate and Ligand Binding

3.1 Preamble

Emerging at the end of 2019, and hence colloquially referred to as "COVID-19", the Severe Acute Respiratory Syndrome coronavirus 2 (SARS-CoV-2) outbreak was declared to be a global pandemic in March 2020 by the World Health Organization. Very quickly, the scientific community rallied to tackle COVID-19 and to improve the understanding of SARS-CoV-2 structure and function in an effort to develop vaccines and drugs more quickly than ever before. Among the first major contributions, a crystal structure of the SARS-CoV-2 main cysteine protease, M^{pro}, in complex with a peptidomimetic inhibitor, N3, was solved and published by Jin et al. [2020]. Quickly thereafter, a wealth of fragment-bound co-crystal structures of M^{pro} was released by the XChem facility at UK's Diamond Light Source which screened over 600 fragments against M^{pro}, revealing an initial 80 fragment hits (later additions add up to 91 fragment hits total) [Douangamath et al., 2020]. This spawned the COVID Moonshot Project led by PostEra.ai and the XChem facility at Oxford [Chodera et al.,

2020; Achdout et al., 2020]. The project crowd-sourced the design of inhibitors of SARS-CoV-2 M^{pro} inspired by the original fragments and quickly created a database of known actives (fragments and crowd-sourced fragment elaborations) with experimentally determined inhibitory activity (IC_{50}) with the goal of developing a small molecule drug against COVID-19. The assayed bioactivity data of fragments and designed compounds was made publicly available on the COVID Moonshot Project Github [COVID-19 Moonshot project, 2020] and the PostEra.ai website [PostEra.Ai, 2020], and the crystal structures were hosted on the Fragalysis website provided by the Diamond Light Source [Diamond, 2020].

Based on the rapid response of structural biologists to the pandemic, structure-based computer aided drug design methods were now possible to be used effectively to aid the global effort against COVID-19. In March of 2020, under the leadership of Prof. Garrett Morris, a large collaboration of 28 computational and laboratory-based scientists from Europe and Japan formed around the question: "How does SARS-CoV2 M^{pro} bind and cleave its substrates and how can we use that information to design inhibitors?". This work resulted in a large collaborative publication that I co-first authored in the journal of Chemical Science as an Edge Article [Chan et al., 2021a].

My contribution to the collaboration was the in-depth intermolecular interaction analysis of the 11 substrate models obtained from molecular dynamics simulations carried out by my collaborator Henry Chan, followed by a characterization of the binding sub-sites of M^{pro} and the extrapolation of the observed binding details to guide future small molecule ligand design. In addition, I developed a 3D interaction fingerprint for SARS-CoV-2 M^{pro} hat is able to characterize ligands by their binding pose. I also developed a constrained small molecule alignment method for the implementation of a covalent protein-ligand docking workflow named *Active-Guided*

Covalent Docking (AGCD). This work ties into the previous project surrounding the Bayesian optimization of binding affinity (Chapter 2) in two ways. First, as one of the overarching goals of the work described in this thesis is the development of methods to help medicinal chemists make better decisions and increase efficiency during hit discovery and hit-to-lead optimization, the development of the AGCD methodology adds an additional method to the computational toolkit of a medicinal chemists. Furthermore, AGCD could be used in tandem with the other methods described in this thesis and is therefore a synergistic addition. Second, the theme of exploring diverse molecular representations to find novel ways of representing molecules in medicinal chemistry challenges is continued here. In this project, I create a 3D interaction fingerprint that represents all interactions between protein and ligand (or peptide substrate) on a residue or atom level. The fingerprint is able to distinguish between different interaction types (*e.g.* hydrogen bonds vs, hydrophobic interactions) and was used in this project to determine favourable interactions in different parts of the M^{pro} binding site to guide inhibitor design.

The work of our collaboration has been published in Chemical Science [Chan et al., 2021a]; due to my major contributions I was named as a co-first author. As a result of the tightly interlinked nature of the work of each co-first author and in order to present the story of the overall project clearly, I include some figures and paragraphs as published in the original paper, which was published under an open source CC BY 3.0 license [Chan et al., 2021a]. I have indicated in each section, result and for each method, which part of the work was performed by my collaborators. In addition, the Appendix features a list of the methods used by my collaborators (and marked as such) for the results shown in this thesis that were not directly created by myself. I have included the relevant methodology for clarity and to understand the results my collaborators contributed which I have built upon in this work.

3.2 Introduction

The family of coronaviruses has long been known to infect humans and other animals. When they jump from one species to another (zoonosis) they can lead to well documented and devastating outbreaks. This occurred twice before, with the outbreaks of severe acute respiratory syndrome (SARS) in 2003 and Middle East respiratory syndrome (MERS) in 2012 [De Wit et al., 2016]. In both these cases, the outbreak was halted (although there are still occasional cases of MERS today), but the more recent outbreak in late 2019 of COVID-19, which is caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) was not contained, and has resulted in a global pandemic that remains with us today [Zhu et al., 2020]. It prompted a rapid response of the academic, non-profit, and commercial biomedical community to find targeted treatments for COVID-19. In April 2020, [Jin et al., 2020] reported a crystal structure of the SARS-CoV-2 main cysteine protease, M^{pro}, in complex with a peptidomimetic inhibitor, N3, as well as an approach to identify potential antiviral agents against the virus.

The main protease (M^{pro}), also known as 3-chymotrypsin-like protease, or $3CL^{pro}$, has been previously identified as a key enzyme in mediating viral maturation in SARS-CoV. [Anand et al., 2002; Yang et al., 2003] More specifically, the SARS-CoV-2 genome encodes for two essential polyproteins: pp1a and pp1ab, which are required for viral replication and transcription. [Zhou et al., 2020; Wu et al., 2020] In order to release the functional proteins from the polyproteins, extensive proteolytic processing by M^{pro} (and another protease, papain-like protease, or PL^{pro}) is vital (Figure 3.1 c). As a result, M^{pro} is crucial for the correct processing or pp1a and pp2b after translation and is directly responsible for the release of many key proteins (including itself) from pp1a and pp2b (Figure 3.1 c).

M^{pro} exists predominantly as a homodimer and fulfills its proteolytic function in a similar way to classic cysteine proteases, through its catalytic cysteine (Cys-145), which ultimately cleaves the peptide bond between key sites in the polypeptides. The location of the catalytic dyad [Jin et al., 2020] in the active site of M^{pro} consisting of Cys-145 and His-41 is located close to the dimer interface. Dimerisation has been proposed to be required for M^{pro} catalytic activity, most likely since the active site is located so close to the dimer interface and amino acids from one monomer make up part of the active site of the other [Zhang et al., 2020]. In addition, monomeric SARS-CoV-2 M^{pro} has been found to be inactive by itself [Xia and Kang, 2011], and experiments using non-denaturing mass spectrometry (MS)-based assays suggest that the monomeric form binds truncated versions (11-mer) of the natural substrates with significantly lower affinity [El-Baba et al., 2020].

The substrate specificity of SARS-CoV-2 M^{pro} is fairly narrow, recognizing all 11 natural substrates with the following motif: [P4:Small] [P3:X] [P2:Leu/Phe/Val/Met] [P1:Gln] ↓ [P1′:Gly/Ala/Ser/Asn], with the scissile amide indicated by the "↓" symbol [Rut et al., 2021; Zhu et al., 2011]. In this case, the following amino acids are classified as "Small": Ala, Val, Pro or Thr. In addition, position "X" does not have a specific motif and can stand for any amino acid. Throughout this work, the Berger & Schechter notation is used to refer to protein and peptide residues in relation to the scissile amide [Schechter and Berger, 1967]. The truncated 11-mer sequences of the 11 natural substrates are shown in Figure 3.1. A visualization of the catalytic cleavage of substrate s01, with the unique amino acid positions highlighted, is shown in Figure 3.2. Overall, the importance of M^{pro} for viral replication in SARS-CoV-2, its unique sequence specificity as well as its established importance in the broader coronavirus family and classical cysteine protease activity profile indicate that M^{pro} is an attractive antiviral drug target.

a)

	P6	P5	P4	P3	P2	P1	P1′	P2′	P3′	P4′	P5′
s01	T	S	A	V	L	Q	S	G	F	R	K
s02	S	G	V	T	F	Q	S	A	V	K	R
s03	K	V	A	T	V	Q	S	K	M	S	D
s04	N	R	A	T	L	Q	A	I	A	S	E
s05	S	A	V	K	L	Q	N	N	E	L	S
s06	A	T	V	R	L	Q	A	G	N	A	T
s07	R	E	P	M	L	Q	S	A	D	A	Q
s08	P	H	T	V	L	Q	A	V	G	A	C
s09	N	V	A	T	L	Q	A	E	N	V	T
s10	T	F	T	R	L	Q	S	L	E	N	V
s11	F	Y	P	K	L	Q	S	S	Q	A	W

b)

c)

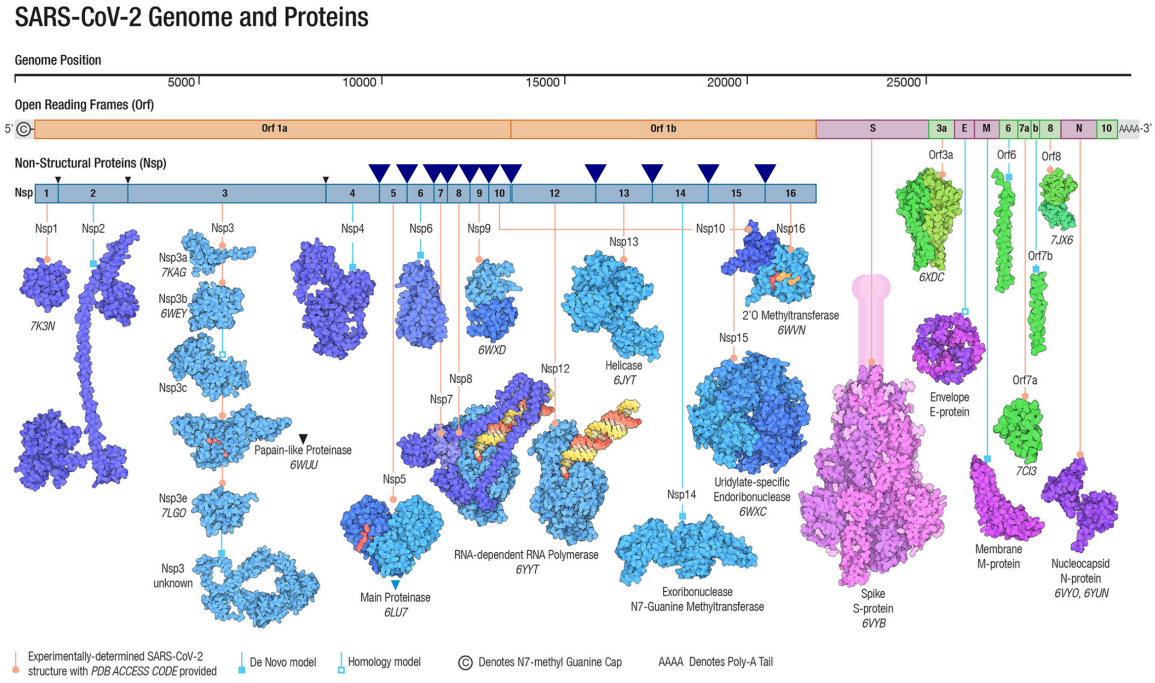

Figure 3.1: Figure a) and b) are adapted from the original publication of this work [Chan et al., 2021a]. a) Sequences of the 11 SARS-CoV-2 M^{pro} cleavage sites as 11-residue peptides (s01-s11). Positively and negatively charged amino acids are colored blue and red respectively. Polar amino acids are colored green, cysteine yellow and histidine purple. b) Analysis of the relative abundance of residues at each position generated by WebLogo [Crooks et al., 2004]. Glutamine is conserved at P1 and the residue at P2 always has a hydrophobic side chain. c) Figure reprinted (adapted) with permission from John Wiley and Sons (Copyright 2021 John Wiley and Sons) as published in [Lubin et al., 2022]. Overview of the SARS-CoV-2 genome and proteome. The viral proteome is derived from the non-structural polyproteins pp1a and pp1ab (shades of blue), the virion structural proteins (pink/purple) and the open reading frame proteins (Orfs, shades of green). The pp1a and pp1ab cleavage sites are indicated by inverted triangles (black for PL^{pro}, blue/bold for M^{pro}). Substrates s01-s11 (shown in a) are the sequences at the M^{pro} cleavage sites separating the Nsps in order of appearance.

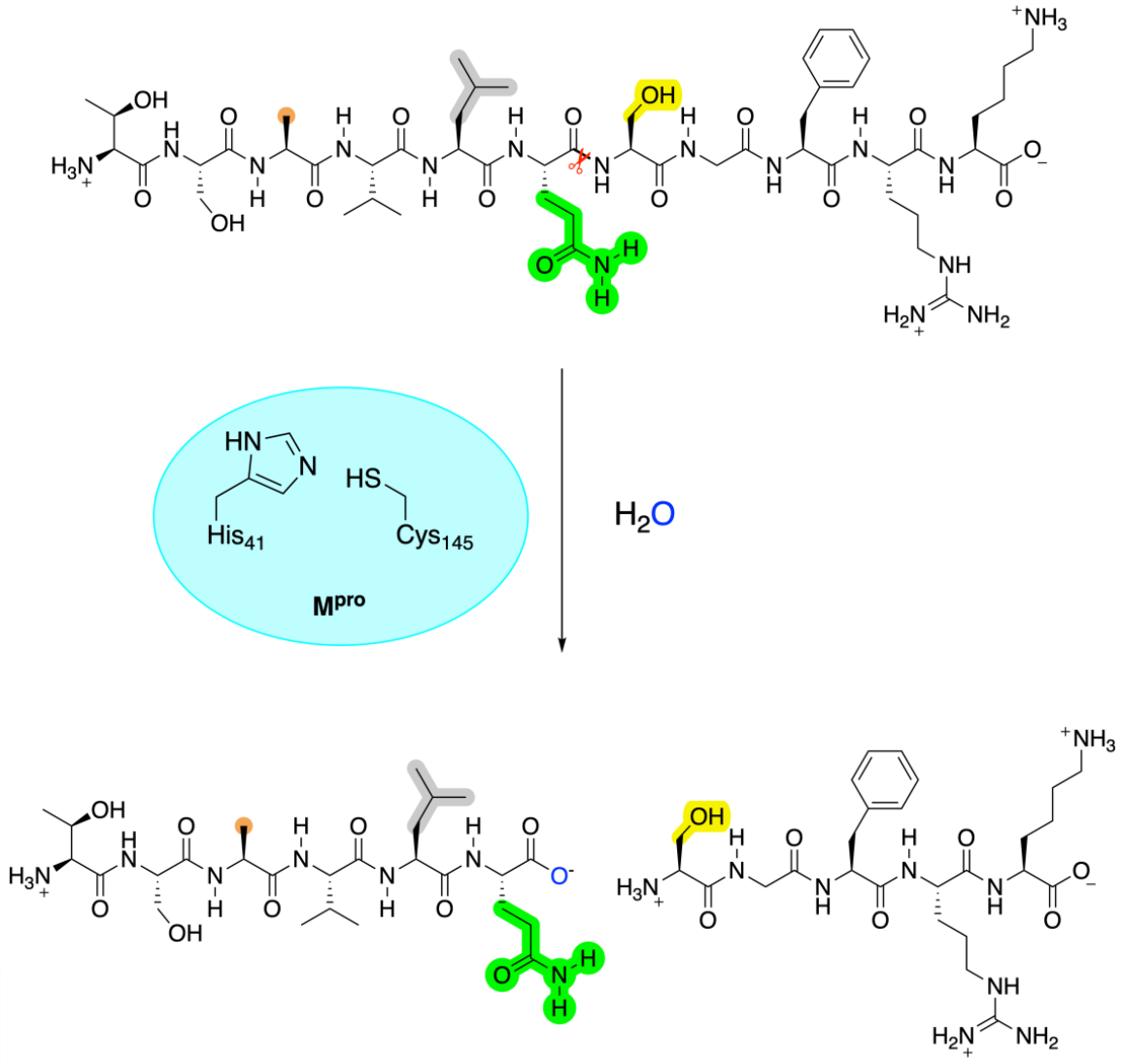

Figure 3.2: Substrate cleavage reaction catalysed by the M^{pro} catalytic dyad on the truncated 11-mer sequence of the natural substrate s01. Figure adapted from the original publication of this work [Chan et al., 2021a]. The cleavage site is indicated as red scissors and the following positions necessary for substrate recognition marked in color: orange (P4, "Small"), grey (P2, Leu/Phe/Val/Met), green (P1, Glu), yellow (P1′, Gly/Ala/Ser/Asn).

Previous studies on the closely related SARS-CoV had also identified M^{pro} as a viable drug target for SARS [Pillaiyar et al., 2016]. Furthermore, M^{pro}'s substrate specificity described above is unlike any known human protease, therefore indicating a potentially favourable side-effect profile since specific SARS-CoV-2 M^{pro} inhibitors that mimic the natural substrate binding mode would not be expected to be recog-

nised by human proteases [Zhang et al., 2020].

When this work was performed, no clinically approved drugs targeting M^{pro} were available. However, several peptidomimetics as well as small molecule inhibitors were known to target SARS-CoV M^{pro} and through the quick response of medicinal chemists around the world, also against SARS-CoV-2 M^{pro} [Pillaiyar et al., 2016; Mengist et al., 2021; Chodera et al., 2020]. Indeed, during the course of my work, the covalent M^{pro} inhibitor PF-07321332 developed by Pfizer had just entered clinical trials with limited information available to the public and to our collaboration [Owen, 2021; Halford, 2021]. As is now known, PF-07321332 (now known as "Nirmatrelvir" [Owen et al., 2021]) was approved by the FDA on the 22^{nd} of December 2021 as an oral M^{pro} inhibitor for the treatment of SARS-CoV-2 in combination with the protease inhibitor "Ritonavir" under the brand name "Paxlovid" [Tantibanchachai, 2021]. In addition, as one of the largest academic SARS-CoV-2 drug discovery projects, the COVID Moonshot Project [Chodera et al., 2020; Achdout et al., 2020] was created with the goal of discovering a novel drug against SARS-CoV-2 M^{pro}. The COVID Moonshot Project brought together many researchers from around the globe with contributors from University of Oxford, University of Cambridge, Diamond Light Source, Weizmann Institute of Science in Rehovot, Temple University, Memorial Sloan Kettering Cancer Center, PostEra, University of Johannesburg, and the Drugs for Neglected Diseases initiative (DNDi) in Switzerland. Soon after the collaboration started, the Diamond Light Source conducted a high-throughput X-ray crystallographic screen of over 600 fragments against M^{pro} that resulted in initial 81 fragment hits (later additions add up to 91 fragment hits total) [Douangamath et al., 2020]. Figure 3.3 shows an example of one of these fragment co-crystal structures of the SARS-CoV-2 M^{pro} homo dimer covalently bound to the ligand "x0830" obtained from Fragalysis [Diamond, 2020].

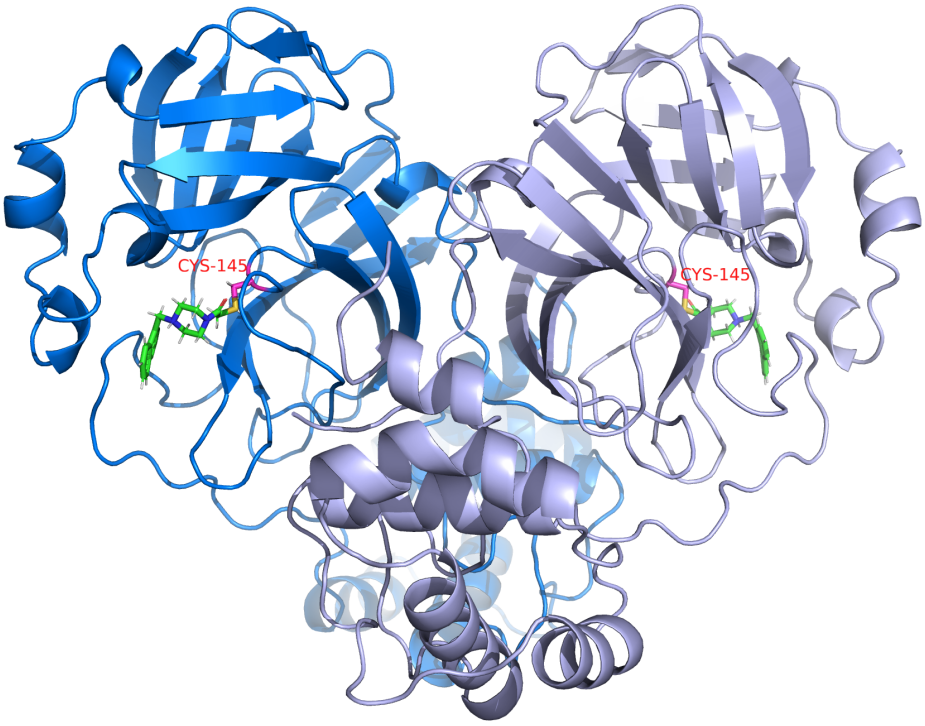

Figure 3.3: Crystal structure of the M^{pro} homo-dimer co-crystallized with the fragment x0830. [Chodera et al., 2020] The two subunits of the dimer are shown in blue and lilac, and the ligands' carbon, nitrogen, oxygen and hydrogen atoms shown in green, blue, red and white, respectively. The fragment x0830 is covalently bound to the active site cysteine-145 shown in magenta and with sulfur in yellow. The structure was obtained from the COVID Moonshot Project as part of a large scale crystallographic fragment screen. This structure and all other structures created by the COVID Moonshot Project can be found on Fragalysis [Diamond, 2020].

The COVID Moonshot Project followed a unique approach to drug design by crowd sourcing the design process. Based on the initial fragment screen, the COVID Moonshot Project urged scientists from around the globe to submit designed fragment elaborations that use one or more of the original fragments as inspiration. All submissions are available online on the PostEra website and the COVID Moonshot Project Github [PostEra.Ai, 2020; COVID-19 Moonshot project, 2020]. In this work, I use the COVID Moonshot Project submission database as the source of potential covalent inhibitors for the AGCD methodology and will henceforth refer to the crowd-sourced compounds as "designs" and the original fragment they are based on as "inspiration fragment".

In addition to the drug discovery projects mentioned above, multiple studies utilizing crystallographic and computational modelling techniques were conducted focused on elucidating the catalytic mechanism [Świderek and Moliner, 2020; Arafet et al., 2021; Ramos-Guzmán et al., 2020; Mondal and Warshel, 2020] and mechanism of inhibition [Acharya et al., 2020; Chodera et al., 2020; Loschwitz et al., 2021; Abel et al., 2020; Ghahremanpour et al., 2020; Zhang et al., 2021] of SARS-CoV-2 M^{pro}, many of which were deposited in the public COVID-19 Open Research Dataset (CORD-19), set up by a consortium of researchers and the United States White House to collect and organise research output about SARS-CoV-2 and other coronaviruses [Wang et al., 2020]. These studies proposed that the Cys-145 thiol is deprotonated by His-41 during catalysis, increasing its nucleophilicity thus allowing the thiol to react with the carbonyl of the scissile amide, forming the acyl-enzyme intermediate. A network of hydrogen bonds between the enzyme and the peptide then stabilises the intermediate, with important hydrogen bonds forming between the scissile amide carbonyl and the "oxyanion hole" which is comprised of a series of backbone amide nitrogens (Gly-143, Ser-144, Cys-145) at the active site. Regeneration of the active M^{pro} catalytic thiol *via* hydrolysis of the acyl-enzyme intermediate releases the N-terminal product and completes the catalytic cycle, enabling M^{pro} to cleave the next peptide.

While the current understanding of M^{pro} catalysis has been greatly advanced, important questions still remain regarding the details of substrate binding and recognition, the influence of induced fit on active site conformation, the role of water, and catalysis as well as how insights of substrate binding can be used for inhibitor design.

In order to answer these questions, a collaboration was formed in April 2020 that united 28 experts in a variety of computational and laboratory-based methods such as molecular mechanics (MM) and quantum mechanical (QM) techniques, non-covalent and covalent automated docking, molecular dynamics (MD) simulations,

density functional theory (DFT), combined quantum mechanics/molecular mechanics (QM/MM) modelling, and interactive MD in virtual reality (iMD-VR). The result of this collaboration was the publication titled "Discovery of SARS-CoV-2 M^{pro} Peptide Inhibitors from Modelling Substrate and Ligand Binding" published in Chemical Science in September 2021 [Chan et al., 2021a] of which I was a co-first author. In this work, my co-authors and I provide detailed, atomic-level insights into the interactions between SARS-CoV-2 M^{pro} (henceforth just referred to as M_{pro}) and the 11-residue substrate models of the 11 natural cleavage sites (named "s01" to "s11", in order of appearance in the sequence of the viral polyproteins pp1a and pp1ab) and utilize this information to design and test peptide inhibitors *in vitro* as well as conduct an *in silico* design approach for small molecule M^{pro} inhibitors. An overview of the methods used and results obtained is shown in Figure 3.4.

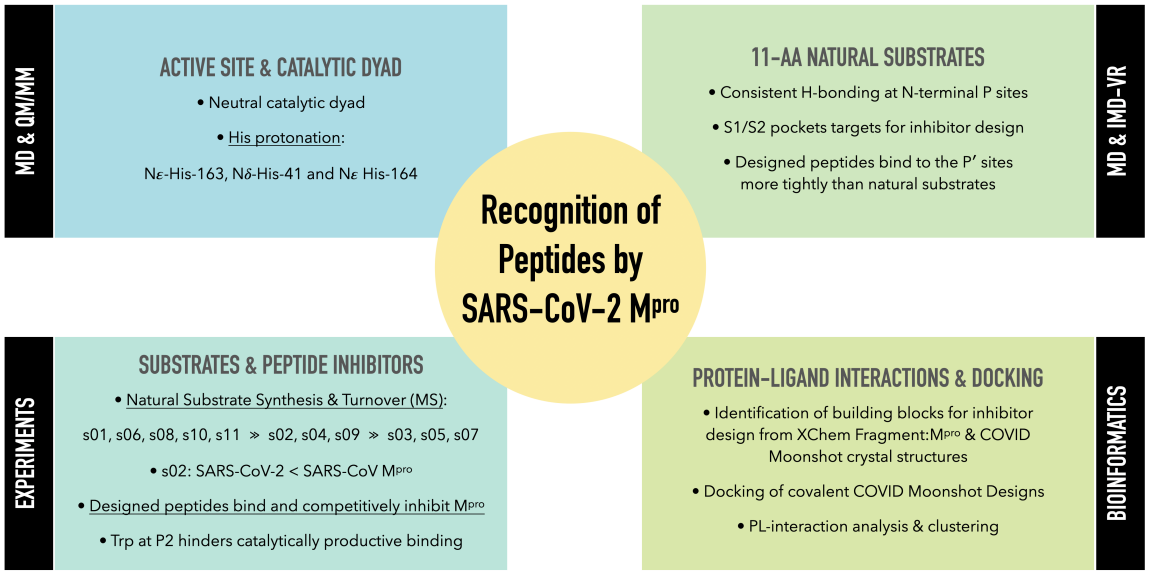

Figure 3.4: Figure adapted from the original publication summarising the methods, topics and insights gained into how SARS-CoV-2 M^{pro} recognises its substrates and how peptide and small molecule inhibitor design can be improved [Chan et al., 2021a].

A detailed summary of which part of the project was undertaken by myself can be found in the Preamble to this thesis above (Section 3.1). In this chapter, I describe

primarily my contribution to this large collaborative project, focusing on the detailed interaction analysis of substrates and peptide inhibitors with M^{pro} as well as the small molecule screening and design campaign. However, for a full description of the entire work performed by the collaboration, please refer to the original publication [Chan et al., 2021a]. In addition, all results are freely available *via* GitHub (`https://github.com/gmm/SARS-CoV-2-Modelling`).

3.3 Materials & Methods

All of the following methods are my own work except otherwise stated. Some figures and paragraphs have been directly adapted from the original publication with minor adjustments as covered by the CC BY 3.0 license [Chan et al., 2021a]. Methods for results generated by my collaborators that are mentioned (and labelled as such) are described in the Appendix B, Section B.1.

3.3.1 Experimental Studies on M^{pro} Activity and Inhibition

Experimental studies were performed by Tika R. Malla, Tobias John, Eidarus Salah, Petra Lukacik, Claire Strain-Damerell, C. David Owen, Martin A. Walsh and Victor Mikhailov. Details about peptide synthesis, substrate and peptide inhibitor turnover assay, substrate binding assay, dose response curve analysis and LCMS analysis are given in the Appendix B, Section B.1.

3.3.2 Molecular Dynamics Simulations

Details about the methodology performed by H.T. Henry Chan are given in the Appendix B, Section B.1.

3.3.3 *In silico* Design of SARS-CoV-2 M^{pro} Peptide Inhibitors

Details about the methodology performed by Debbie K. Shoemark are given in the Appendix B, Section B.3.

3.3.4 Active-Guided Covalent Docking

The idea behind Active-Guided Covalent Docking (AGCD) is to match each compound design proposed by a crowd-sourced medicinal chemist to the corresponding covalent origin fragment for which we have an X-ray crystal structure and use the binding pose information of the fragment for the docking process. The justification for this approach is inspired by fragment-based drug discovery and in particular the findings of [Malhotra and Karanicolas, 2017] who found that for elaborated ligands, in 86% of the 297 paired ligands, the larger elaborated ligand did not change its binding mode relative to the smaller ligand.

In the COVID Moonshot Project, the inspiration fragment was cited by the crowd-sourced designer of the compound [Chodera et al., 2020; Achdout et al., 2020]. I matched each of the designed compounds to their corresponding inspiration fragments, and calculated the maximum common substructure (MCS) between them. Next, an initial conformation of the design was aligned to the fragment before docking. A full overview of the AGCD workflow is described in Section 3.4.4 and Figure 3.18. Since the alignment method in RDKit [Landrum et al., 2006] (v2020.03.1) was insufficiently constrained, I created a new, stricter constrained alignment method (named *MCS-align*, see Section 3.4.4.1) that uses the maximum common substructure between two molecules as the basis for the alignment. The method uses atom constraints to force the corresponding atom positions of the MCS into the same conformation, followed by a constrained energy minimisation, thus keeping the con-

formation of the MCS constant. Docking was performed using AutoDock4 (AD4), which by default treats ring conformations to be rigid when sampling ligand conformations before docking [Morris et al., 2009]. Although it is possible to use specialised docking methods such as the implementation of "glue dummy atoms" developed by Forli and Botta [2007] to overcome this issue and create flexible ring docking, this limitation is useful for AGCD where the goal is to incorporate as much information of known actives into the docking process as possible. As a result of the alignment process, all rings present in the MCS for the designed compounds are aligned to the crystallographically observed binding pose of the rings in the inspiration fragment. Each design was then docked to the corresponding M^{pro} crystal structure of the origin fragment, after generation of the homodimer, and the protonation and charge optimization using Protonate3D in MOE [Chemical Computing Group ULC., 2019].

The FlexRes method in AD4 for covalent docking [Morris et al., 2009] was used. It was assumed that each design with a cysteine-targeting covalent warhead would react with the Cys-145 of M^{pro}. The covalent adduct of the COVID Moonshot design after reacting with the active site Cys-145 was selected as a flexible residue and a water molecule was included as the "dummy" ligand. Docking and grid parameter files were generated for each Cys-145-inhibitor adduct individually with the rest of the corresponding co-crystallised dimeric M^{pro} structure treated as the rigid receptor molecule. The PDBQT files for docking were generated using the *prepare_receptor4.py* and *prepare_ligand4.py* scripts implemented in MGLTools v1.5.7 [Morris et al., 2009] for the protein and ligand, respectively. Docking with AD4 (v4.2.6) [Morris et al., 2009] was performed using the Lamarckian Genetic Algorithm (LGA) and the following AD4 hyperparameters: population size 300; maximum number of energy evaluations 250,000; maximum number of generations 27,000; number of dockings 100. The remaining parameters were kept at default levels.

The scoring function used by AD4 includes pairwise evaluation of intermolecular interactions of the ligand and the protein, intramolecular interactions within the ligand and between residues of the protein and the covalent adduct, and an estimation of the conformational entropy lost upon binding (for details about the scoring function see Chapter 1 Section 1.2.2). For the evaluation of the covalent docking procedure, only the intramolecular terms of the covalent adduct are relevant, since they correspond to the changes in the energy of the ligand bonded to the covalently-bound Cys-145 residue. Since AD4 automatically clusters docking results by the total estimated free energy of binding, covalent docking results must be re-clustered using the "Final Total Internal Energy" instead (as reported in the DLG docking log file output by AD4). For clustering the docked poses of the covalent adducts, I implemented a hierarchical clustering procedure (similar to the one used in the native AD4 method). A new cluster was seeded with the lowest energy pose, and all remaining poses within a threshold (< 2 Å RMSD) are added to that cluster. The procedure was repeated for the next lowest energy pose, until all docked poses have been assigned to a cluster. RMSD values between poses were calculated using the Open Drug Discovery Toolkit (ODDT [Wójcikowski et al., 2015a]), to account for symmetry, such as rotations of equivalent methyls in tertiary butyl groups.

Docked poses were compared to the original inspiration fragment crystal structure using SuCOS [Leung et al., 2019]. SuCOS produces scores to a value between 0 and 1, where 1 indicates perfect overlap and identical molecules. SuCOS computes the shape and pharmacophoric feature overlaps between two molecules, and both scores are weighted equally in the final SuCOS score. Based on work by Leung et al. [2019], a SuCOS score of 0.55 between two molecules was found to be equivalent to a conformational RMSD of 2 Å. The original default parameters for SuCOS do not produce normalized scores which might result in SuCOS scores of larger than 1 for

some molecules. As a result, I created a new, normalized version of SuCOS, which is freely available for use on GitHub (`github.com/MarcMoesser/SuCOS` [Moesser, 2021]). This adapted and normalized SuCOS method was used for this work.

Finally, covalent docking for Nirmatrelvir was performed identically to that for the other covalent Moonshot designs with the exception that no pre-alignment of the ligand was performed prior to docking. Instead, a random conformation of the ligand was used to seed the docking process. The azanide nitrogen resulting from the covalent attachment of the nitrile warhead to Cys-145 was assigned a negative charge prior to docking (Structure and docked pose of Nirmatrelvir is shown in Section 3.4.5 Figure 3.20.

3.3.5 Analysis of SARS-CoV-2 M^{pro} Active Site Interactions

Snapshots from MD models (see Appendix B, Section B.1) as well as the XChem derived crystal structures and covalent docking poses were analysed using the interaction analysis tool Arpeggio [Jubb et al., 2017]. Arpeggio was developed to identify and classify interactions within and between proteins and protein, DNA, or small-molecule ligands. Arpeggio uses a radial distance cutoff of 5 Å between interacting atoms and the expanded definition of interaction types implemented in Interaction Fingerprints (IFPs) published by Marcou and Rognan [Marcou and Rognan, 2007] which were based on previous interaction type definitions used in Structural Interaction Fingerprints (SIFt, [Singh et al., 2006]). This approach combines matching complementary atom types (*e.g.* hydrogen bond donor and acceptor) with geometric rules (*e.g.* a donor-acceptor distance of less than 3.5 Å and an angle constraint on the D-H-A angle for hydrogen bonds) to identify interactions. Beyond interactions present in the original SIFt, Arpeggio is able to identify van der Waals', ionic, carbonyl, metal, hydrophobic, halogen bond, hydrogen bonds and specific atom–aromatic

ring (cation–π, donor–π, halogen–π, and carbon–π) and aromatic ring–aromatic ring (π–π) interactions.

For Arpeggio analysis of the MD snapshots of the substrate and designed peptides (Section 3.4.1.1 & Section 3.4.2), a representative snapshot for each complex was chosen by selecting the snapshot within the highest populated cluster that had the lowest RMSD compared to all other snapshots in that cluster. From the docked covalent Moonshot submission compounds, the lowest energy pose of the highest populated cluster obtained from AD4 dockgin (see Section 3.3.4) was chosen. Analysis of the XChem fragments and Moonshot designs was done using the reported crystallographically observed conformations obtained from Fragalysis where available [Diamond, 2020; Chodera et al., 2020; Achdout et al., 2020; Douangamath et al., 2020]. Before Arpeggio analysis, the ligand-M^{pro} complex was processed by cleaning the PDB file using PDBtools and running Arpeggio on all ligand-M^{pro} interactions as described by Jubb [2020]; Jubb et al. [2017]. Then, Arpeggio was used to identify all intermolecular atom-atom interactions in a given protein-ligand or protein-peptide complex and to classify them as "Clash", "Covalent", "VdW Clash", "VdW", "Proximal", "Hydrogen Bond", "Weak Hydrogen Bond", "Halogen Bond", "Ionic", "Metal Complex", "Aromatic", "Hydrophobic", "Carbonyl", "Polar" or "Weak Polar" [Jubb et al., 2017]. Based on this atom-level description of the environment of the peptide or ligand in the active site, three levels of analysis were conducted, describing the interaction in different levels of detail (see Figure 3.5).

First, inspired by previous approaches to interaction fingerprints such as IFPs [Marcou and Rognan, 2007] and SIFts [Singh et al., 2006] a custom interaction fingerprint was created that denotes the presence or absence of a general "interaction" (1) *vs* "no interaction" (0) in the fingerprint bit-vector for each active-site residue in the protein-ligand or protein-peptide complex (Figure 3.5 left branch).

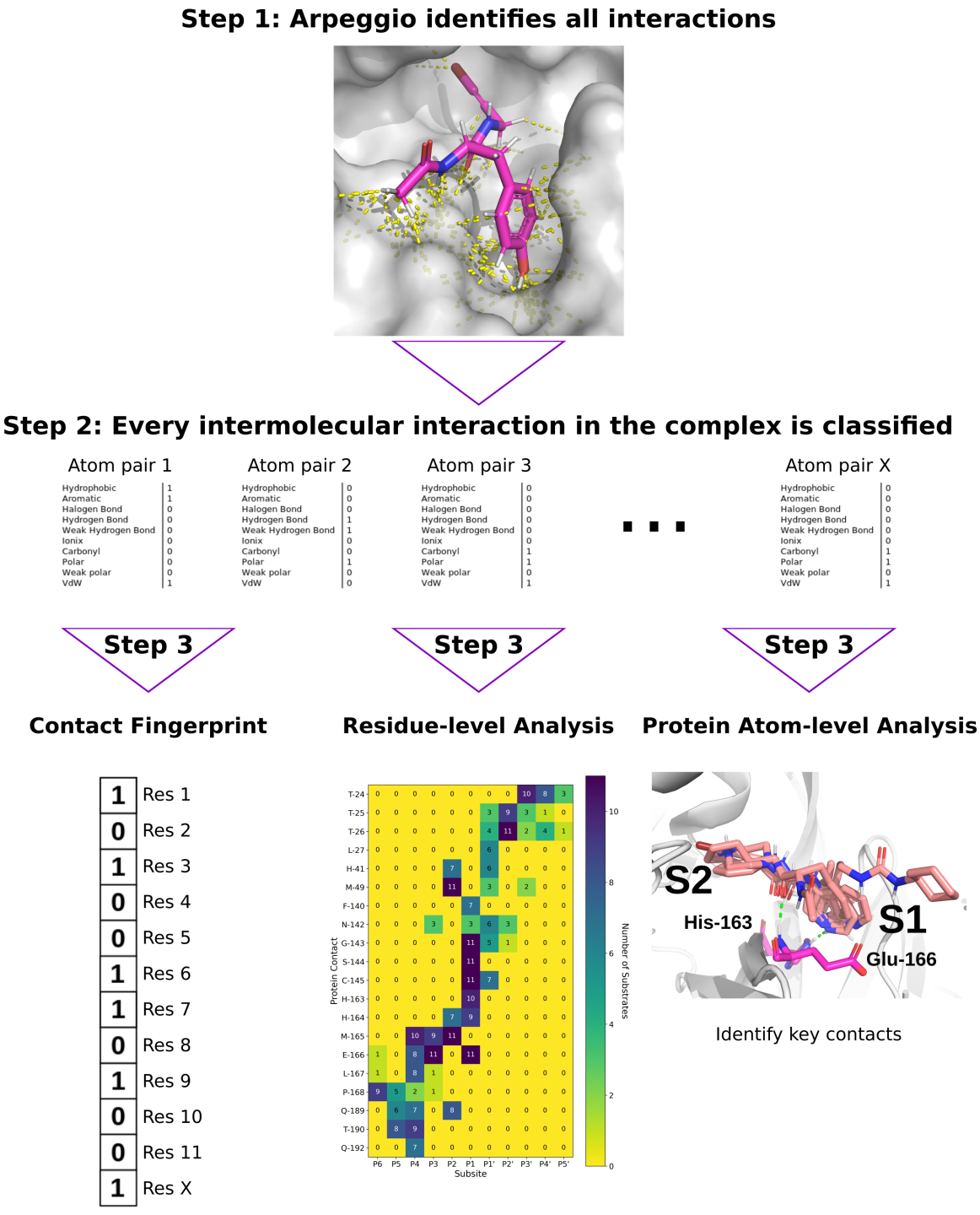

Figure 3.5: Arpeggio-derived intermolecular interactions between M^{pro} and the ligand or peptide are analyzed at three levels of detail. High level fingerprints are created (left branch) based on the existence of any interactions between ligand and protein residues. Residue level interaction analysis (middle branch) was performed to obtain active-site subsites and the common type of interaction observed at each site. Atom-level analysis (right branch) reveals specific interactions such as hydrogen bonds and hydrophobic pockets to guide inhibitor design.

An "interaction" was classified as any of the following Arpeggio interaction types: "VdW", "Hydrogen Bond", "Weak Hydrogen Bond", "Ionic", "Halogen Bond", "Aromatic", "Hydrophobic", "Carbonyl", "Polar" or "Weak Polar".

As has been shown by Rácz et al. [2018], binary interaction fingerprints such as IFPs can be compared using the Tanimoto distance. I applied the same principle to the Arpeggio-derived interaction fingerprints, calculating the Jaccard distance [Jaccard, 1912] between different fingerprint bit vectors. In addition, to identify and analyse similar ligands, I created a clustering algorithm to cluster the ligands by their calculated Jaccard distance using two different thresholds of 0.5 and 0.7. I used a hierarchical clustering approach (same as described above for clustering of docked poses) that identifies all molecules with pairwise similarities below the threshold, creates a cluster for them and then moves on to the rest of the ligands not currently part of a cluster to repeat the process.

The residue level interaction matrix of the substrates and peptide inhibitors (Figure 3.5 middle branch) was created by identifying the residues in each binding subsite in the protein active site by their interaction with the peptide substrate residues. Highly conserved residues between all peptide substrates were noted on a residue and protein-atom level to guide inhibitor design.

Finally, on an atomic interaction level (Figure 3.5 right branch), the pairwise atomic interactions between the ligand and M^{pro} atoms in the fragment crystal structures, as well as between the atoms of the peptide substrates and M^{pro}, were used as a baseline to guide potential fragment elaboration. To determine if the docked poses of the COVID Moonshot designs exhibit the same binding profile as one of the identified fragment clusters, a standardised cluster profile was created for each fragment cluster which records the presence of a residue-level interaction if it was classified by Arpeggio as one of the following "major" interactions: Aromatic, Hydrophobic, Halo-

gen Bond, Polar, Hydrogen Bond, Ionic, Carbonyl. If more than 70% of all recorded major protein atom interactions (*e.g.* a strong polar interaction with an aspartic acid carboxylic oxygen) of a particular cluster are occupied for an individual ligand, the ligand was classified as a member of that cluster.

3.3.6 Hydrophilicity Maps

To calculate whether a given M^{pro} subsite corresponded to a hydrophilic or hydrophobic pocket, I developed an interaction-based hydrophilicity score. All identified (using Arpeggio) intermolecular atom-atom interactions with a given residue in the substrate were classified as either hydrophobic (hydrophobic, aromatic or halogen bond Arpeggio interaction types) or hydrophilic (hydrogen bond, weak hydrogen bond, ionic, carbonyl, or polar Arpeggio interaction types). I excluded the "VdW" and "Weak Polar" interaction types since they were deemed less informative and in some cases even redundant as individual atom-atom interactions. The hydrophilicity score, Z_{hydro}, of a given subsite over all pairwise atom-atom interactions was then calculated as follows:

$$Z_{hydro} = \sum_{i,j} C^{i,j}_{hydrophilic} - \sum_{k,l} C^{k,l}_{hydrophobic} \tag{3.1}$$

The first term is calculated over the sum of all pairs of ligand atoms, i, and protein atoms, j that form hydrophilic interactions and the second term is calculated over the sum of all pairs of ligand atoms, k, and protein atoms, l that form hydrophobic interactions. The sum of all hydrophobic atom-atom interactions, $\sum_{k,l} C^{k,l}_{hydrophobic}$, was then subtracted from the sum of all hydrophilic atom-atom interactions, $\sum_{i,j} C^{i,j}_{hydrophilic}$, to create the hydrophilicity score, Z_{hydro} for each subsite. A higher hydrophilicity score correlates with more hydrophilic interactions and *vice versa*.

3.3.7 Plasticity Analysis

I analysed the effect of ligand binding on the conformational plasticity of M^{pro} by calculating the per-residue heavy atom root-mean square deviation (RMSD) between all pairs of 333 different M^{pro} co-crystal structures obtained from Fragalysis [Diamond, 2020] using MDAnalysis v. 1.1.1 [Richard J. Gowers et al., 2016; Michaud-Agrawal et al., 2011].

3.4 Results and Discussion

3.4.1 SARS-CoV-2 M^{pro}-Substrate Interaction Analysis

3.4.1.1 Models of SARS-CoV-2 M^{pro}-Substrate Peptide Complexes

The initial comparative modeling and MD simulation was performed by Garrett M. Morris, with subsequent MD calculations being carried out by H.T. Henry Chan.

In order to obtain high-quality 3D models of the protein-substrate complex, an 11-mer peptide was created as a surrogate for each of the 11 natural cleavage sites ranging from position P6 to P5′ (Figure 3.1) [Wu et al., 2020]. These 11-mer peptides are referred to as "substrates" since all of them have been experimentally confirmed to be cleaved by M^{pro}, though with different efficiencies (by Tika R. Malla, see Appendix B, Section B.2). Following the comparative modelling procedure outlined in Appendix B, Section B.1.1, the substrates were modelled into the active site of chain A of the M^{pro} dimer obtained from the crystal structure 6yb7 (high resolution structure with unliganded active site, deposited by Owen et al. [2020]) and subjected to three independent explicit-solvent MD simulations of 200 ns each. In the following sections, all M^{pro} residue numbers and residue names unless otherwise stated, refer to chain A of the dimer.

During explicit-solvent MD, all substrates remained tightly bound in the active site and substrate backbone stability was maintained. However, some local sidechain

fluctuations were observed for individual models, specifically around P3. Overall, C-terminal P′ residues were observed to be more conformationally flexible than N-terimal P-side residues.

3.4.1.2 Hydrogen Bond Interaction Network

In order to alayse the role ff hydrogen bonding in M^{pro} substrate recognition and binding, the persistence of individual hydrogen bonds (HBs) was monitored during MD simulations (performed by H.T. Henry Chan), while I analysed the MD-derived snapshots using the bioinformatics tool Arpeggio [Jubb et al., 2017], as outlined in Section 3.3.5.

During MD and Arpeggio analysis, 12 HBs were consistently identified (Figure 3.6). Backbone-backbone HBs at Glu-166 (position P3, HB 2 & 3) and Thr-26 (position P2′, HB 10 & 11) were the most persistent during MD (Figure 3.6). Hydrogen bonds 5-9 are all formed between the highly conserved Gln residue at P1 and the protein, divided between side chain (HBs 6 and 7) and backbone (HBs 5,8 and 9). Despite the presence of a Gln in all 11 substrates at position P1, the hydrogen bonds formed by Gln at P1 are not as constant as the stabilizing backbone hydrogens at P3 and P2′. However, HBs 6 and 8 are significantly more stable than the other HBs at P1 and correspond to the oxyanion hole (which has been reported to play an important role in the stabilisation of the acyl-enzyme intermediate as described above). It is comprised of the backbone HBs between the carbonyl oxygen at P1 and the backbone N-H on residue Cys-145 and Gly-143.

A visualization of the binding pose of substrate s01 in the active site of M^{pro} with hydrogen bonds HB1-3 and HB5-12 is shown in Figure 3.7. Hydrogen bond 4 between the backbone nitrogen of P2 and residue Gln-189 in the M^{pro} active site was not observed in this snapshot.

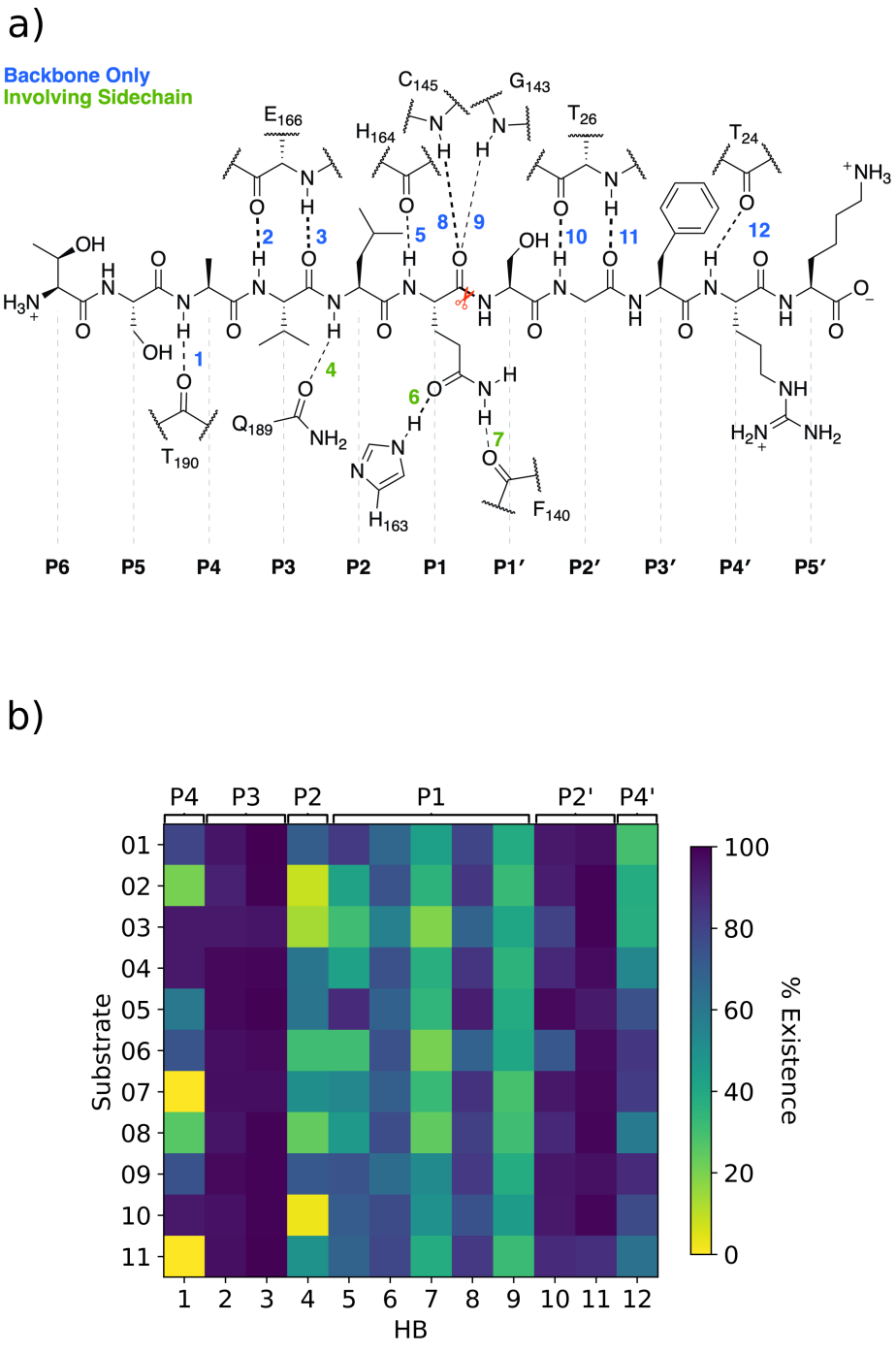

Figure 3.6: Figure created by H.T. Henry Chan and adapted from the original publication of this work [Chan et al., 2021a]. a) Overview of the 12 major identified hydrogen bonds exemplified by substrate s01. The scissile amide is indicated by a red pair of scissors. b) The observed percentage prevalence of hydrogen bonds 1-12 for each substrate. Snapshots were taken every nanosecond from 600 ns of explicit-solvent MD per substrate. HBs 2-3 and 10-11 are observed in nearly 100% of snapshots, strongly stabilizing the pose of the substrate. HBs 6-9 form the Gln recognition motif.

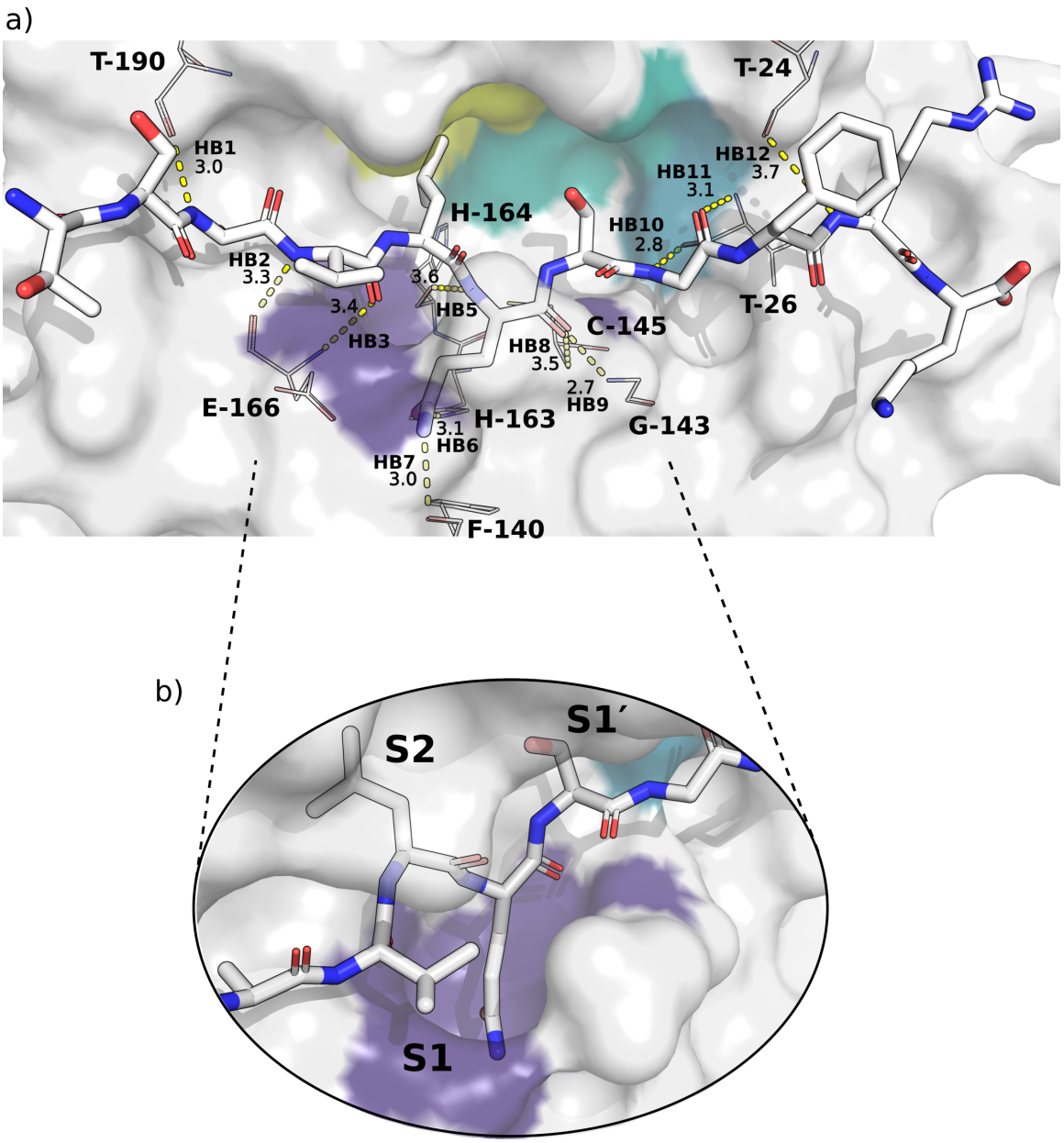

Figure 3.7: a) Overview of the binding pose of substrate s01 in the active site of M^{pro}. The most representative conformation obtained from MD used for interaction analysis as described in Section 3.3.5 is shown. Key M^{pro} amino acids involved in binding are labelled and the hydrogen bonds 1-4 and 6-12 that are involved in s01 recognition (see Section 3.4.1.2) are indicated by yellow dotted lines and their distance labelled. Subsite S2-S2′ are colored in yellow, purple, light blue and turquoise for S2, S1, S1′ and S2′, respectively. b) Rotated close-up of the binding site at the scissile amide bond showing the deeply buried Gln residue at S1. The hydrophobic S2 pocket accomodating the s01 Leu residue at P2 buries the Leu side chain deeply into the pocket, where it interacts primarily with Met-49 and Met-165.

However, the key hydrogen bonds for stability (backbone HBs 2 & 3 and 10 &

11 and the Gln recognition motif at the oxyanion hole and the Gln side chain HBs 6 & 7 are clearly visible. This interaction pattern is present in all 11 substrates. In addition, Gln at P1 is clearly deeply buried in the active site pocket, fixed in place by several hydrogen bonds as well as the lack of space confined in the pocket itself.

3.4.1.3 Non-Covalent Interaction Analysis

In order to identify all key interactions (beyond just HBs as described above) between the modelled substrates and M^{pro} as well as to classify the binding sub-pockets, I conducted a full interaction analysis covering all types of non-covalent interactions using the bioinformatics tool Arpeggio [Jubb et al., 2017] on snapshots extracted from the explicit-solvent MD simulations as described in Section 3.3.5 [Jubb et al., 2017]. A residue level analysis was conducted that summarized all interactions between the substrates and M^{pro} in order to identify the sub-pockets that each residue on the substrate binds to (Figure 3.8 and Figure 3.7). At position P1, six of the eight most common interactions with M^{pro} residues are present in most (*i.e.*, $\geq 9/11$) substrates. This indicates a rigid binding mode at position P1 driven by the specific recognition of the conserved Gln residue present in every substrate at P1. Specifically, M^{pro} residues Gly-143, Ser-144, Cys-145, His-163 and Glu-166 interact with at least 10 of the 11 substrates. Interactions at P2 (His-41, Met-49, His-164 and Met-165) are also conserved between substrates, although to a lesser extent than at P1. Overall, residue-level interactions appear to be less conserved with increasing distance to the scissile amide bond between P1 and P1′, indicating a higher degree of movement during MD and therefore weaker binding. Nonetheless, some M^{pro}-substrate interactions further away from the scissile amide bond, such as with M^{pro} residue Thr-24, Thr-26 and Glu-166 at P3′, P2′ and P3, respectively, are highly conserved. In addition, interactions on the P′ side are overall less conserved and less frequent than interactions on the

P-side, indicating a tighter, more well defined binding groove for substrates on the P-side.

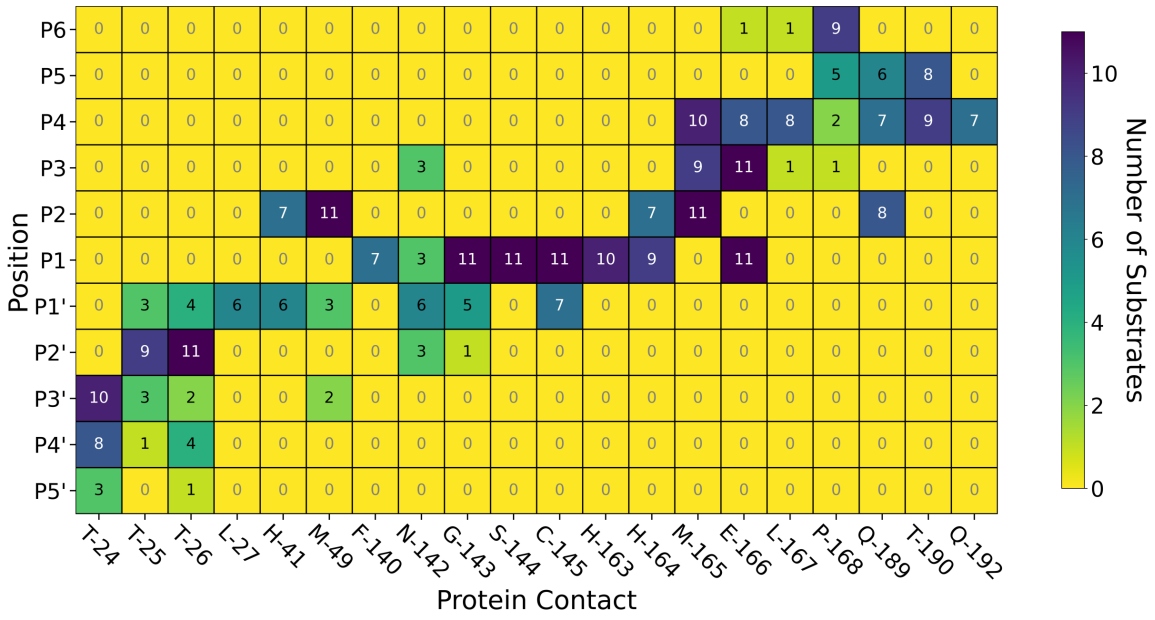

Figure 3.8: Map of non-covalent interactions between the 11 substrates and M^{pro} derived from interactions identified by Arpeggio [Jubb et al., 2017]. The most representative pose for each substrate was obtained from explicit-solvent MD as described in Section 3.3.5. Yellow indicates that no substrate forms this interaction at a given subsite, while dark blue indicates the interaction is formed by most/all substrates as labelled. The P1 subsite has highly conserved interactions between all 11 substrates, mostly interacting with the same amino acids in the protein, indicating a conserved, highly specific binding mode for recognition of the Gln residue at P1.

3.4.1.4 Hydrophilicity Analysis

After identifying key residue-level contacts between M^{pro} and the substrate, I analyzed the degree to which M^{pro} subsites can be characterized as either hydrophilic or hydrophobic pockets. For that purpose, I created an Arpeggio-derived hydrophilicity score as described in Section 3.3.6. To analyze hydrophilicity on a subsite-level, I identified every interaction between the protein and substrate at every subsite and calculated the corresponding hydrophilicity score. The hydrophilicity analysis revealed that the S1 subsite is the most hydrophilic site of all the 11 subsites in the substrate–M^{pro} complexes, while the S2 site was consistently identified as a hydropho-

bic pocket (Figure 3.9). In addition, with increasing distance from the cleavage site on both the prime and non-prime sites, the hydrophilicty score varies more strongly between substrates and tends to be amphiphilic. This analysis is in agreement with the MD-based results described above where no consistent HBs were identified towards the N- and C-termini of the substrate 11-mers and higher degrees of movement during MD was observed. Nonetheless, subsites S3 and S2′ are slightly biased towards hydrophilic interactions.

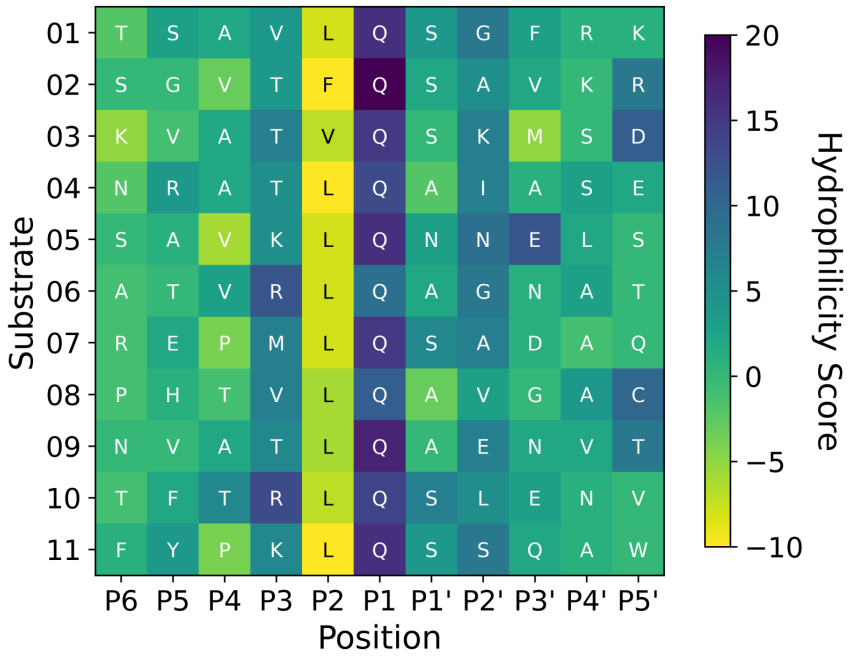

Figure 3.9: Figure showing the hydrophilicity or hydrophobicity of each M^{pro} subsite. The hydrophilicty score was calculated for each substrate as described in the Methods Section 3.3.6 using Arpeggio to identify interactions [Jubb et al., 2017]. A higher score correlates to a more hydrophilic pocket. Subsite P1 is shown to be highly hydrophilic, recognizing the conserved Gln residue mainly through polar interactions. Pocket P2 was observed to be a hydrophobic pocket. Other subsites did not show a major bias towards hydrophilic or hyrophobic interactions.

3.4.1.5 Conformational plasticity in M^{pro} crystal structures

Previous studies have compared the dynamics of ligand binding sites across SARS-CoV-2, SARS-CoV and MERS-CoV M^{pro} [Cho et al., 2021]. Here, I investigated

the conformational plasticity of the SARS-CoV-2 M^{pro} active site upon binding by comparing 333 M^{pro}-ligand co-crystal structures obtained from Fragalysis [Diamond, 2020] to a reference *apo* structure of M^{pro}, PDB entry 6yb7 [Owen et al., 2020] by computing per-residue RMSD values (Figure 3.10 a). A high degree of plasticity was observed at residues Gln-19, Thr-24, Thr-25, His-41, Thr-45, Ser-46, Met-49, Asn-119, Asn-142, Met-165, Glu-166, Arg-188, Gln-189 and Ala-191. A visualization of the plasticity as an overlay of all 333 structures is shown in Figure 3.10 b. The S1 subsite is particularly rigid, with almost no movement across all 333 crystal structures. Since S1 is primarily responsible for Gln recognition, which is a driving force of substrate specificity, a high degree of rigidity is to be expected. However, high plasticity was observed at the S2 site, changing drastically upon ligand binding, especially at residues Thr-45, Ser-46 and Met-49 (Figure 3.10 b). In all 11 natural substrates, the S2 site accommodates a hydrophobic amino acid. The high degree of plasticity at S2 therefore suggests that larger hydrophobic groups could be accommodated in the S2 pocket and should be considered when designing peptide and small molecule ligands. It also has implications for structure-based virtual screening: to capture the M^{pro} plasticity when using rigid-protein docking, multiple M^{pro} conformations should be used, ideally based on ligand similarity between known structures and the screening compounds. Potential ligands with high similarity to known binders (such as fragment elaborations of known fragment binders) should be docked into the *holo* M^{pro} structure of the highly similar known binder to take advantage of the induced-fit conformation of M^{pro} around the active site.

3.4.1.6 Summary of Key Insights into M^{pro}-Substrate Binding

The following trends emerge from the above described study on SARS-CoV-2 M^{pro} in complex with models of all 11 of its substrates: (i) binding stability is partly conferred by a series of HBs from P4 to P4′, in particular between the backbones of

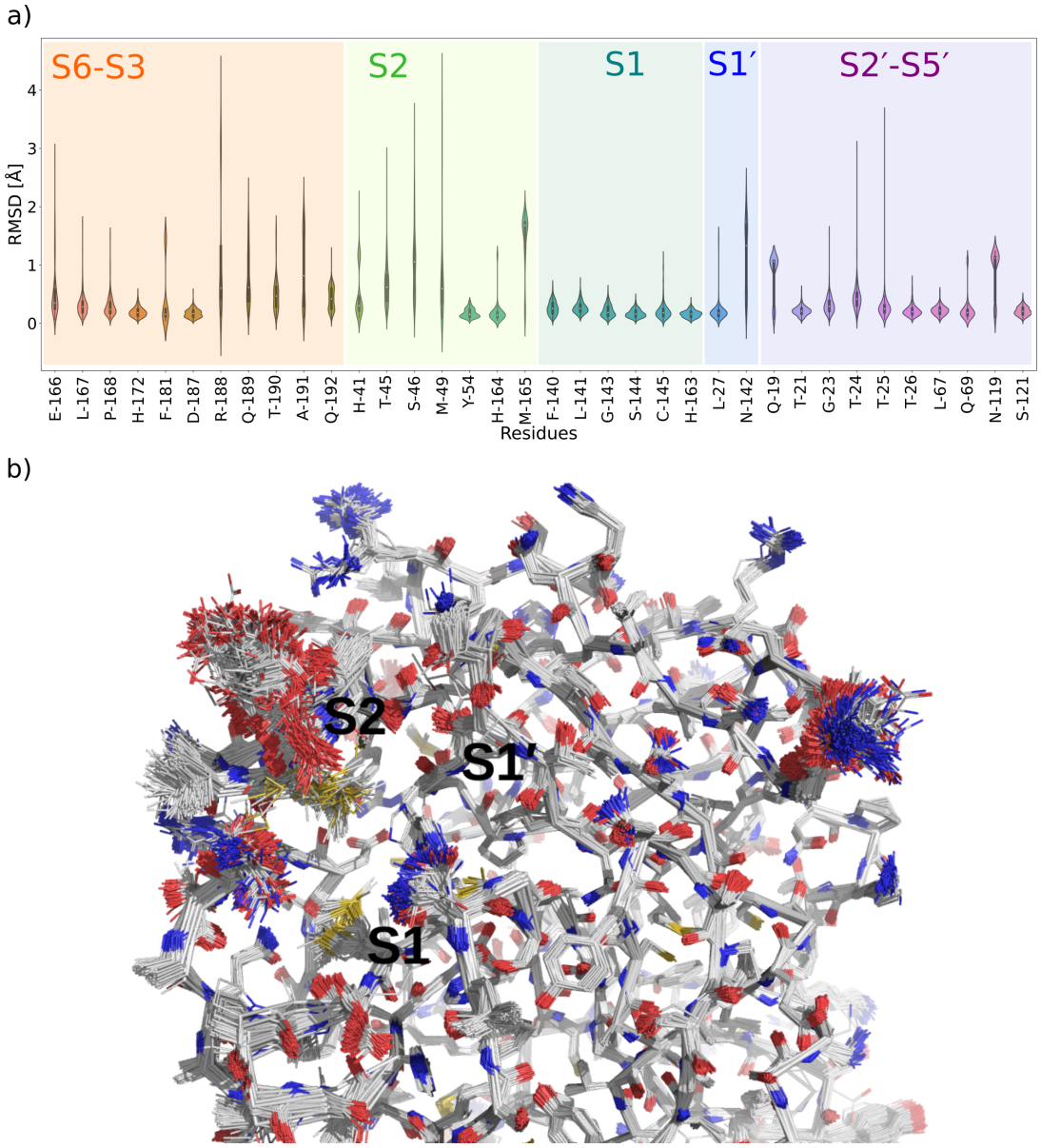

Figure 3.10: a) Analysis of the active site plasticity of 333 M^{pro} co-crystal structures. Active site residues (res. 19, 21, 23–27, 41, 45, 46, 49, 54, 67, 69, 119, 121, 140–145, 163–168, 172, 181, 187–192) were chosen based on the MD analysis of the 11 substrate–M^{pro} models and correspond to all M^{pro} residues that interaction any substrate. The violin plots show the distributions of per-residue heavy atom RMSD values between the 333 M^{pro}–ligand co-crystal structures53 and a reference uncomplexed structure (PDB 6yb7 [Owen et al., 2020]). Each M^{pro} subsite is colour-coded. b) Overlay of the lines representation of all 333 M^{pro} co-crystal structures. The subsites S2, S1 and S1′ are labelled. Subsites S1′ and especially S1 show low plasticity, barely moving between co-crystal structures. S2 however shows a large degree of flexibility, especially at residues Ser-46 and Met-49 where large movements of the side chain occur to alter the size of the S2 pocket.

M^{pro} Glu-166 and Thr-26 at substrate positions P3 and P2′ respectively, as well as HBs involving the conserved P1 Gln sidechain; (ii) substrate residues on the N-terminal side of the cleavage site (P-side) form more, and more consistent, interactions with M^{pro} compared to the P′ side, with interactions at Met-49, Gly-143, Ser-144, Cys-145, His-163, His-164, Met-165 and Glu-166 being most conserved. This observation is in accordance with results published by Shaqra et al. [2022], who were able to obtain crystal structures for nine out of the 11 natural substrates in complex with M^{pro} and report that the binding mode of non-prime residues is more conserved between substrates, while the binding mode of prime-side residues is more varied and even lack full electron density in many cases, particularly for residues beyond the P3′ position. These results suggest that the S1 and S2 pockets are therefore prime targets for active site substrate-competing inhibitor design due to their well-defined and consistent character, large energy contributions to substrate binding, high degree of flexibility in S2 and vital conserved hydrogen bonds in S1 to compete with substrate recognition.

3.4.2 *In silico* Design and Experimental Validation of Peptide Inhibitors

3.4.2.1 *In silico* Mutational Analysis of Substrate Peptides Enables Inhibitor Design

As part of our collaboration, an *in silico* mutational analysis of our 11-mer substrate models was performed by Dr. Deborah K. Shoemark at the University of Bristol, with the aim of designing peptide inhibitors that would bind more tightly to M^{pro}, out-competing the natural substrates without being turned over themselves. A more detailed elaboration of the method provided by Deborah Shoemark is given in Appendix B, Section B.3.

Substitution of residues at each position in the substrate via an *in silico* alanine scanning experiment was conducted using the interactive web application BAlaS which uses the BudeAlaScan [Wood et al., 2020] and the BUDE_SM algorithm [Sessions, 2021] for Predictive Saturation Variation Scanning (PreSaVs) [Hetherington et al., 2021]. By "mutating" to the 19 other amino acids at each position, the algorithm computes the $\Delta\Delta G$ value at each position between the original substrate residue and the substituted residue to find the best possible 11-residue sequence that would maximize the binding affinity at each subsite, resulting in five designed peptides (Figure 3.11). Interestingly, all of the designed peptide inhibitors, with the exception of p16, had a bulky, aromatic amino acid at position P2. This is in agreement with our previous interaction and plasticity analysis of the natural substrates (Section 3.4.1) that showed that the hydrophobic binding pocket at S2 has a high degree of flexibility, potentially enabling M^{pro} to accommodate larger hydrophobic side chains such as the Trp or Phe side chains of peptide p12-15.

To test if the designed peptides are indeed inhibitors of M^{pro}, the inhibitory activity of peptides 12, 13, 15 and 16 was determined by dose-response analysis using a mass spectrometry-based assay which monitors both substrate s01 depletion and corresponding N-terminally cleaved product formation with different peptide inhibitor concentrations (this experiment was performed by Tika R. Malla). Details of the experimental procedure and results can be found in the Appendix B, Section B.4. All four tested peptides showed moderate inhibitory activity: p12 (IC_{50} = 5.36 ± 2.17 μM), p13 (IC_{50} = 3.11 ± 1.80 μM), p15 (IC_{50} = 5.31 ± 1.08 μM), and p16 (IC_{50} = 3.76 ± 1.19 μM) with p13 being the most potent inhibitor among the set (Appendix B, Figure B.4).

	P6	P5	P4	P3	P2	P1	P1′	P2′	P3′	P4′	P5′
p12	K	Y	T	F	W	Q	Y	S	Q	F	Y
p13	K	Y	L	T	W	Q	N	S	Q	I	N
p14	N	S	V	T	F	W	I	S	Q	F	Q
p15	L	T	I	N	W	Q	K	Y	F	N	T
p16	W	F	T	L	K	Q	Y	W	Q	T	N

Figure 3.11: Sequences of designed peptides p12–p16. Positively charged amino acids are colored blue. Polar amino acids are colored green.. No negatively charged amino acids were part of the designed peptides. Strikingly, except for p15, the P2 position was occupied by a bulky amino acid with a large aromatic side chain (Trp or Phe).

3.4.2.2 Understanding the Basis of SARS-CoV-2 M^{pro} Inhibition by the Designed Peptides

Explicit-solvent molecular dynamics simulations of the designed peptides were performed by H.T. Henry Chan (details about the method can be found in the Appendix B, Section B.1.2). Using the conformation of the most representative pose (pose with the lowest RMSD to all other poses in the highest populated cluster) obtained from MD, I performed the Arpeggio-based interaction analysis to identify key interactions and compared the insights gained to the original substrates.

The residue-level analysis of the peptide inhibitors revealed a similar pattern to the substrates with highly conserved interactions between all inhibitors at S1, S2 and the key stabilizing backbone HBs with Thr-26 and Glu-166 (Figure 3.12 a & Figure 3.13). All the peptide inhibitors were found to bind in the oxyanion hole, making contact with Cys-145, Gly-143 as well as Ser-144 and His-163 in the S1 binding site (Figure 3.12 a & Figure 3.13). Interestingly, peptide p14 does not contain a glutamine residue at P1 (Figure 3.11) but was still found to form all the key interactions at P1 (Figure 3.12 a), indicating that nitrogen containing heterocycles could be a possible alternative to the glutamine side chain when binding into the S1 pocket.

The main difference in substrate and peptide binding was observed at S2. BUDE predicted that the substitution of Leu or Val present in 10 out of the 11 substrates to a bulky amino acid with an aromatic side chain such as Trp or Phe would increase

binding affinity significantly (see Appendix B, Section B.3).

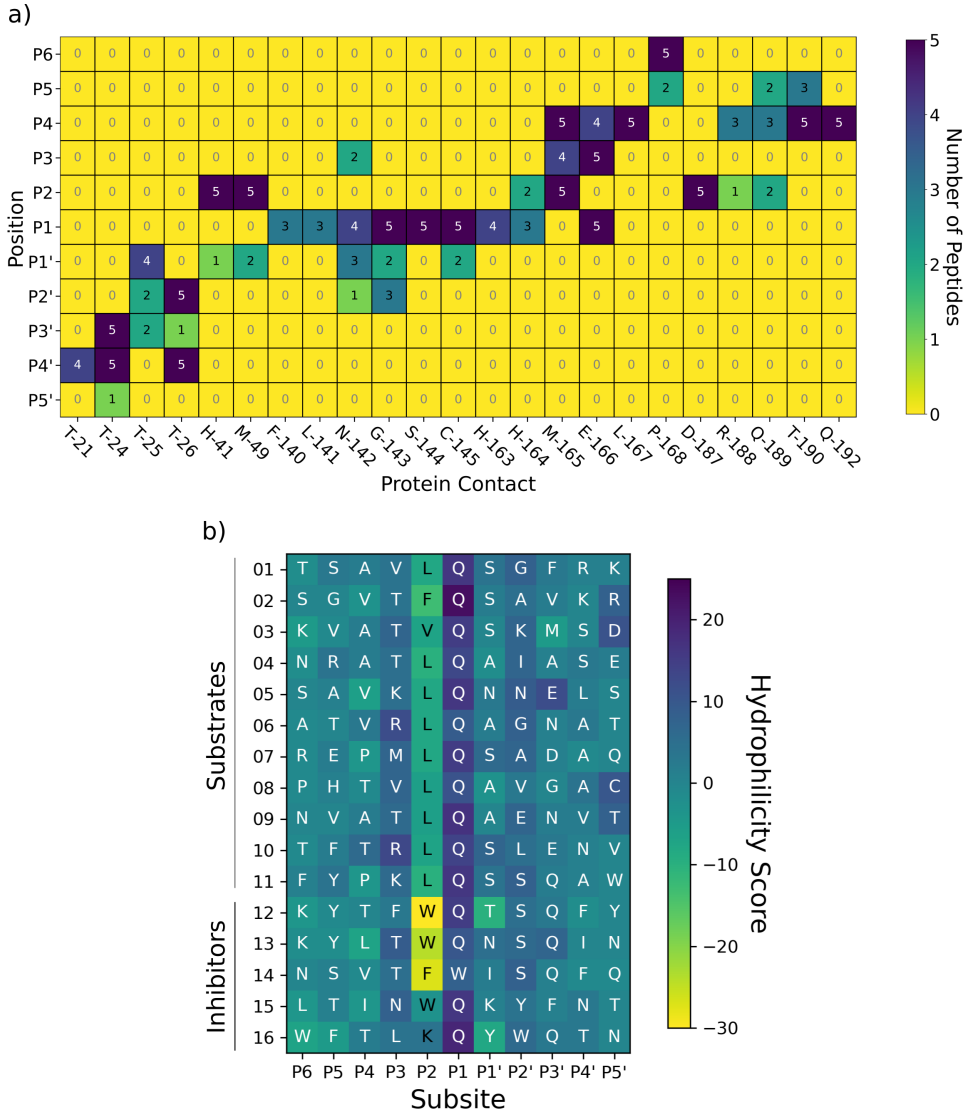

Figure 3.12: a) Map of non-covalent interactions between the MD-derived peptide models and M^{pro} identified by Arpeggio [Jubb et al., 2017]. The most representative pose for each peptide was obtained from explicit-solvent MD as described in Section 3.3.5. Yellow color indicates that no peptide forms this interaction at a given subsite, while dark blue hue indicates the interaction is formed by most/all peptides as labelled. b) Hydrophilicity map for both, the peptide inhibitors and the substrates for comparison. The hydrophilicty score was calculated as previously described (Section 3.3.6). A higher score correlates to a more hydrophilic pocket. Similar trends emerge as for the substrate analysis (Figure 3.8 & 3.9). Subsite P1 is highly hydrophilic while the P2 pocket is highly hydrophobic. Other subsites did not show a major bias towards hydrophilic or hyrophobic interactions. Strikingly, p12, p13 and p14 form almost double the number of hydrophobic interactions in P2 than the natural substrates, indicating that Trp and Phe might be able to bury more deeply into S2.

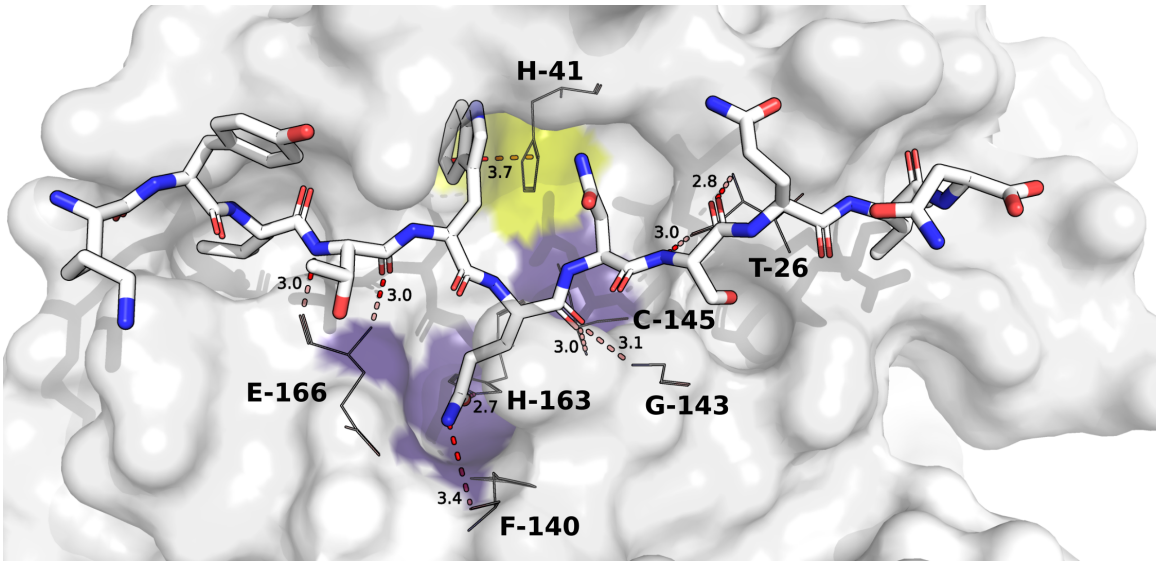

Figure 3.13: Overview of the binding mode of the peptide inhibitor p13 modelled in the active site of M^{pro}. Shown is the most representative pose obtained from MD used for interaction analysis as described in Section 3.3.5. Key M^{pro} amino acids involved in binding are labelled and the key hydrogen bonds and the π-π stacking with His-41 are highlighted as red dotted lines. Subsites S2 and S1 are colored in yellow and purple, respectively. Peptide inhibitor p13 occupies all key hydrogen bonds previously identified to be curcial for tight binding (Thr-26, Glu-166) as well as glutamine recognition (oxyanion hole Cys-145, Gly-143 and HBs with Phe-140 and His-163. In addition, the Trp residue at P2 forms π-π stacking interactions with His-41 blocking the catalytic dyad.

One reason for this high increase in predicted affinity could be the ability of the S2 pocket to encompass relatively large groups (as described in Section 3.4.1.5, allowing for a larger hydrophobic surface area for interaction and therefore increasing the binding affinity. Indeed, the hydrophilicity score analysis (Figure 3.12 b) of all non-covalent interactions between the P2 residue in peptides p12, p13 and p14 and the S2 site in M^{pro} revealed that they form more than double the number of hydrophobic interactions than any of the substrates, indicating that the Trp and Phe residues might be buried more deeply in the hydrophobic pocket. In addition, the Trp residue at p13 was observed to form additional π-π stacking interactions with the His-41 residue of the catalytic dyad in the active site (Figure 3.13). The close contact to His-41 might partially explain the inability of M^{pro} to catalyse the cleavage of the

P2-Trp-containing peptide inhibitors, however, no fully conclusive evidence has been found yet to explain why the peptide inhibitors are indeed inhibitors and not turned over by M^{pro} (for turnover experiments see Appendix B, Section B.4).

3.4.3 Fragment-based *In silico* Design of Small Molecule Inhibitors

Having elucidated how M^{pro} recognises its substrates and the designed peptide inhibitors, the next step was to transfer the lessons learned to the design of small molecule inhibitors. Specifically, I explored whether ligands sharing the same interactions as the natural substrates and peptide inhibitors could lead to better inhibitory activity. To answer this question, I started the analysis with the initial 91 X-ray structures of small molecule fragments in complex with M^{pro} obtained by high-throughput crystallographic screening at Diamond's XChem facility [Douangamath et al., 2020], as well as the dataset of 798 designed inhibitors and 245 crystal structures obtained from the COVID Moonshot Project [Chodera et al., 2020; Achdout et al., 2020; Douangamath et al., 2020] available at the time of this work (accessed January 2021). In addition, I developed an Active-Guided Covalent Docking (AGCD) procedure and selected promising candidates to inform inhibitor design and fragment elaboration. A full overview of the small molecule analysis pipeline is shown in Figure 3.14.

Similarly to the natural substrate and peptide inhibitor interaction analysis described above, I used the bioinformatics toolkit Arpeggio [Jubb et al., 2017] to identify all protein-ligand interactions. Using all interactions, I created a protein-ligand interaction fingerprint, denoting the presence or absence of interactions to specific protein residues in order to cluster and compare known binders to the high-quality substrate models (for methods see Section 3.3.5). Finally, I created the AGCD method for the alignment and covalent docking of 540 ligand designs obtained from the Moonshot

project. Using the detailed interaction analysis and the protein-ligand fingerprint approach, I identified promising candidates from the virtual screening campaign and suggested future directions for fragment elaboration in order to cover key protein interactions derived from my previous substrate analysis.

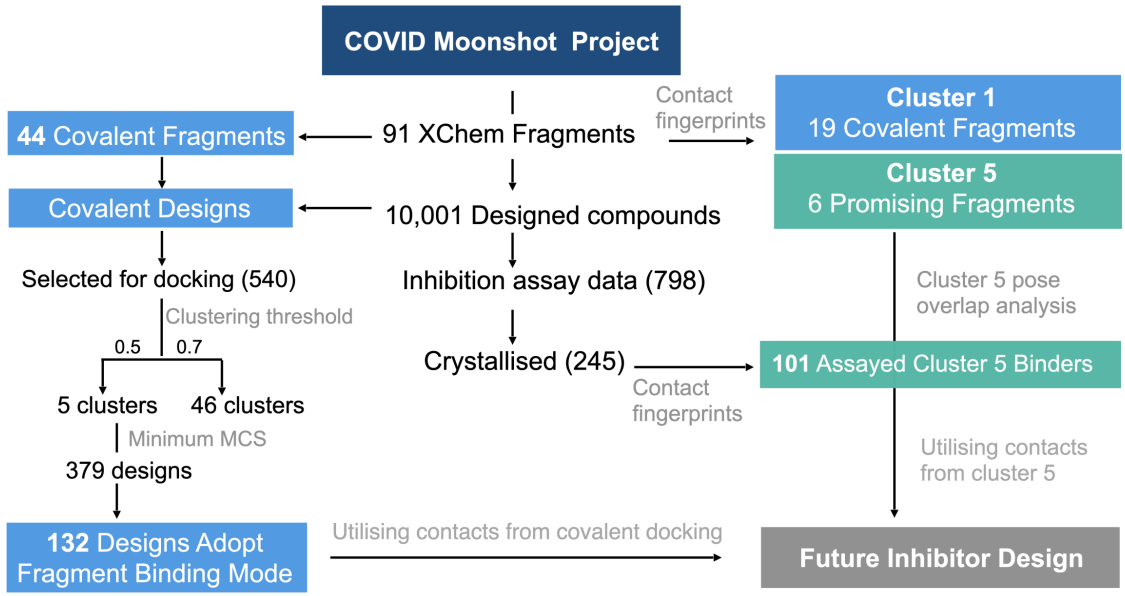

Figure 3.14: Analysis of fragment and designed compounds from the Moonshot project and XChem fragment screen [Chodera et al., 2020; Achdout et al., 2020; Douangamath et al., 2020] adapted from the original publication of this work [Chan et al., 2021a]. Workflow used to identify promising fragments and guide novel designs. The COVID Moonshot Project database was used to select the 44 covalent fragment co-crystal structures and all covalent compound designs present at the time of the analysis (10,001 compounds). All covalent designs based on specific covalent fragments were selected for AGCD and the resulting docked poses filtered to select 132 compounds with high quality poses that adopt the original fragment binding mode. The designs were then analysed in the context of known, key interactions obtained from the 11 substrate models and 91 XChem fragment structures. All 91 XChem fragments were analysed using Arpeggio and an interaction fingerprint created to cluster fragments by binding mode. The most important cluster (cluster 5) was found to occupy key interactions responsible for substrate recognition and was used as a template to identify known cluster 5 binders in the COVID Moonshot Project database and to guide fragment elaboration.

3.4.3.1 Interaction Analysis of XChem Fragments

As part of the COVID Moonshot Project, the XChem facility at UK's Diamond Light Source released a set of 91 fragment-bound crystal structures of SARS-CoV-2 M^{pro}

[Douangamath et al., 2020]. I separated the the XChem fragments found to bind to M^{pro} into non-active-site binders (25 fragments) and active-site binding/likely-substrate competing molecules (66 fragments; Figure 3.15). Since the goal was to leverage the extensive interaction analysis of the substrates (Section 3.4.1.3) to understand how best to create substrate-competing, active-site binding inhibitors, only the active-site binding fragments were considered in the analysis.

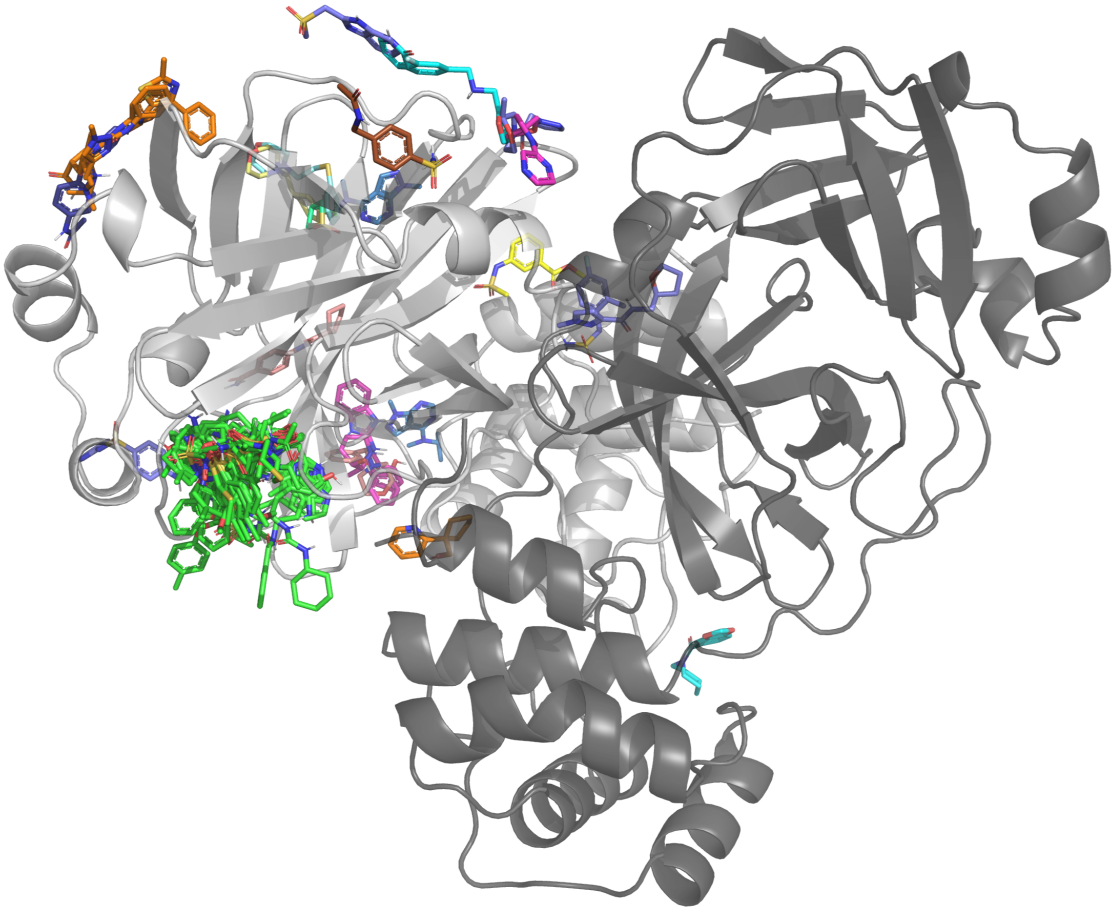

Figure 3.15: Overlay of all 91 XChem fragments bound to the M^{pro} dimer [Douangamath et al., 2020]. All of the binding sites on chain A (white) are shown. As a representative structure, the M^{pro} crystal structure of the x0830 co-crystal structure was used [Douangamath et al., 2020]. There are 66 fragments that bind into the active site (green fragments) and 25 fragments binding in remote pockets (fragment 1101 binds in two different remote sites). The non-active site binding fragments are broadly colored to distinguish different binding sites.

Next, an interaction fingerprint bit-vector was constructed for every active-site

binding fragment, with each bit denoting the presence or absence of any interaction with any M^{pro} residues (as described in Section 3.3.5). This binary interaction fingerprint was used to cluster fragments by their interaction fingerprint Tanimoto similarity [Jaccard, 1912; Rácz et al., 2018], with 1 corresponding to identical interaction fingerprints, and 0 to the absence of any shared interactions. Note that the Tanimoto index was computed between Arpeggio-derived protein-ligand interaction fingerprints rather than, for example, the widely used ligand-based extended-connectivity fingerprints (ECFPs) [Rogers and Hahn, 2010]. A series of different Tanimoto similarity thresholds for the clustering was employed in steps of 0.1 between 0.1 and 0.9 (see Appendix B, Table B.2) using the hierarchical clustering approach described in Section 3.3.5. For this analysis, two clustering thresholds were chosen: a broader (0.5) and a tighter (0.7) threshold (for details about the choice of clustering thresholds see Appendix B, Section B.6.1). While the tighter clustering threshold produced 29 clusters total, the number of clusters with more than one fragment remained the same in both thresholds (nine clusters, see Appendix B, Table B.2), indicating that tighter clustering above 0.5 does not lead to the addition of meaningful clusters. In addition, despite the lower threshold, the broader (0.5) clustering method was able to create distinct clusters (Top 5 most populated clusters for both thresholds are shown in Appendix B, Figure B.7) and both thresholds identify a series of primarily covalent fragments (binding to Cys-145) as the highest populated cluster (all interactions in each cluster shown in Appendix B, Figure B.8). Nonetheless, cluster 5 of the broader (0.5 threshold) clustering was found to be the most unique cluster with great potential for inhibitor design (Figure 3.16), as it was the only one that targets the key HBs 3 and 6 identified to be crucial for both stability and M^{pro} substrate recognition (see Section 3.4.1.2). Subsequent mentions of cluster numbers therefore refer to clusters obtained from the clustering threshold of 0.5.

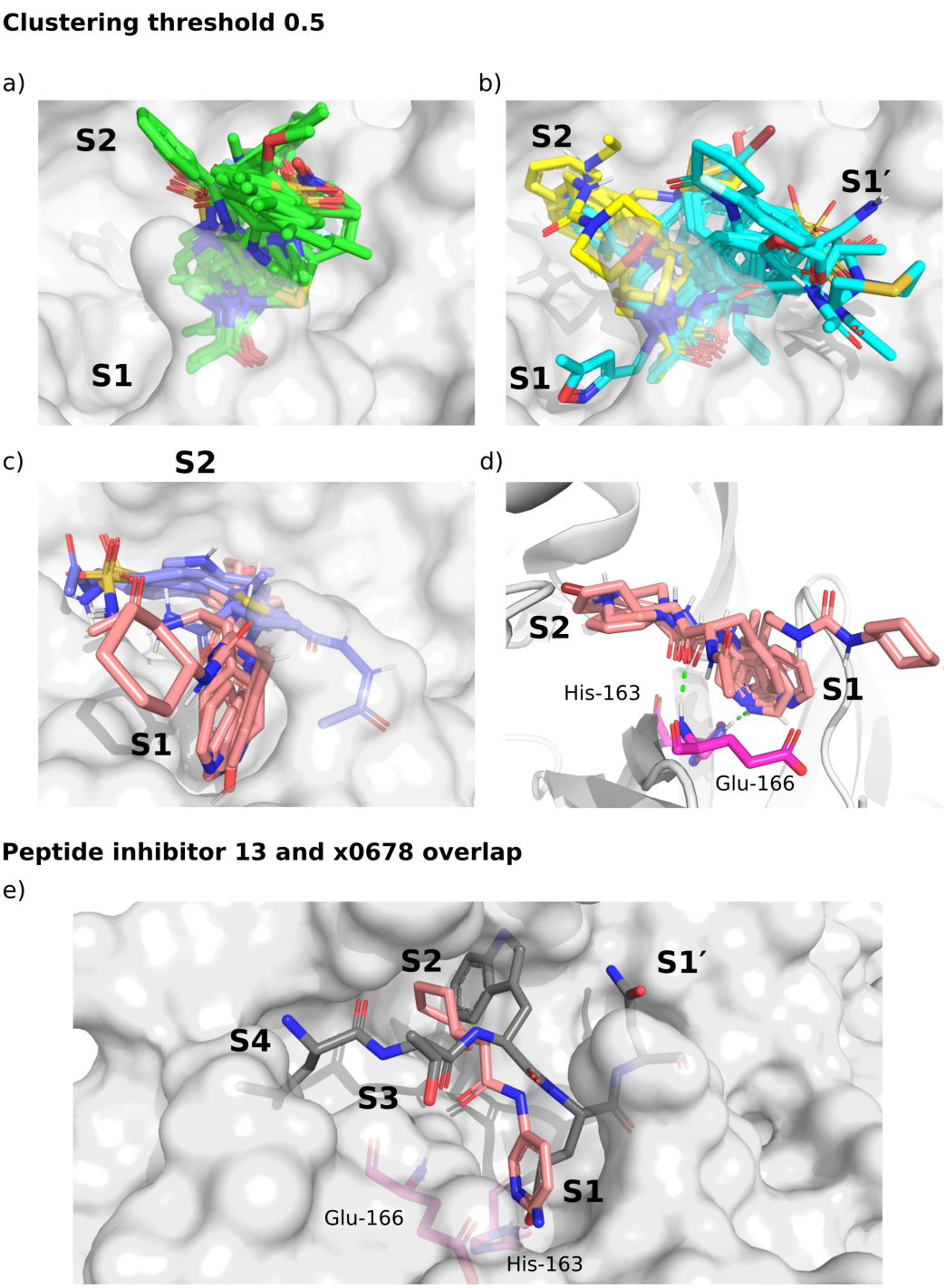

Figure 3.16: Molecular surface of the x0830-bound M^{pro} structure (white surface) and the top 5 most populated fragment clusters using a clustering threshold of 0.5. a) Cluster 1 fragments (green); b) clusters 2 (cyan) and 3 (yellow); c) clusters 4 (lilac) and 5 (pink); d) close-up of cluster 5. Green dotted lines indicate the two key HBs between the fragment carbonyl oxygen and the backbone nitrogen of Glu-166 (HB 3, Figure 3.6), and between the His-163 Nϵ and the heterocyclic nitrogen of the fragment (HB 6, Figure 3.6). e) Overlay of the P4–P1′-truncated structure of p13 (grey) and cluster 5 binder x0678 (pink), with the x0678 co-crystal M_{pro} structure (white surface). Cluster 5 inhibitors are well suited to mimic Gln recognition at S1 as well to extend into the hydrophobic S2 pocket.

All the fragments and ligands in clusters 1 and 2 (except x0397, x0978 and x0981) are covalently bound to Cys-145. A highly conserved binding mode was observed for the carbonyl-containing covalent warheads (e.g., chloroacetamides), where the carbonyl oxygen binds into the oxyanion hole between residues Gly-143 and Cys-145, mimicking substrate HBs 8 and 9 (Figure 3.16).

Cluster 5 stands out as the only major cluster with fragments that bind deeply in the S1 pocket, one of the most conserved interactions identified in all M^{pro} substrates. Cluster 5 shows a distinct binding motif primarily driven by: (i) hydrogen bonding between a carbonyl oxygen on the fragment and the Glu-166 backbone NH-group (HB 3) ; and (ii) a HB/strong polar interaction between His-163 and the fragment (HB 6). In addition, it appears that the protonation state of the imidazole of His-163 depends on the fragment hydrogen bonding polarity. Based on the presence or absence of either a HB donor or acceptor on the fragment, the protonation state of His-163 could be inferred. This suggests that for x0107, x0434, x0540, x0678 and x0967, the His-163 ϵ-nitrogen is protonated, forming a hydrogen bond to the pyridine nitrogen (x0107, x0434, x0540 and x0678) or phenol oxygen (x0967). For x1093, the δ-nitrogen is protonated, leaving the ϵ-nitrogen free to form a hydrogen bond with the indole -NH of x1093, reversing the hydrogen bond polarity compared to the other fragments in the cluster. Nonetheless, the same binding geometry is observed in both cases and the clustering algorithm correctly assigns the molecules into the same cluster.

Overall, the primary functionality that facilitates interaction with His-163 is the nitrogen-containing heterocycle present in almost all ligands in cluster 5 (Figure 3.17); the exception is x0967, which forms the His-163 HB via its phenol oxygen. Such heterocycles were found to be well suited to replace the substrate Gln sidechain at P1 by mimicking its HB donor/acceptor abilities while also containing potential conformational entropic advantages. This trend was also observed during the BUDE_SM

PreSaVS analysis which predicted that Trp at P1 might be better suited to replace Gln for peptide p14 (Figure 3.11 & Appendix B, Figure B.3).

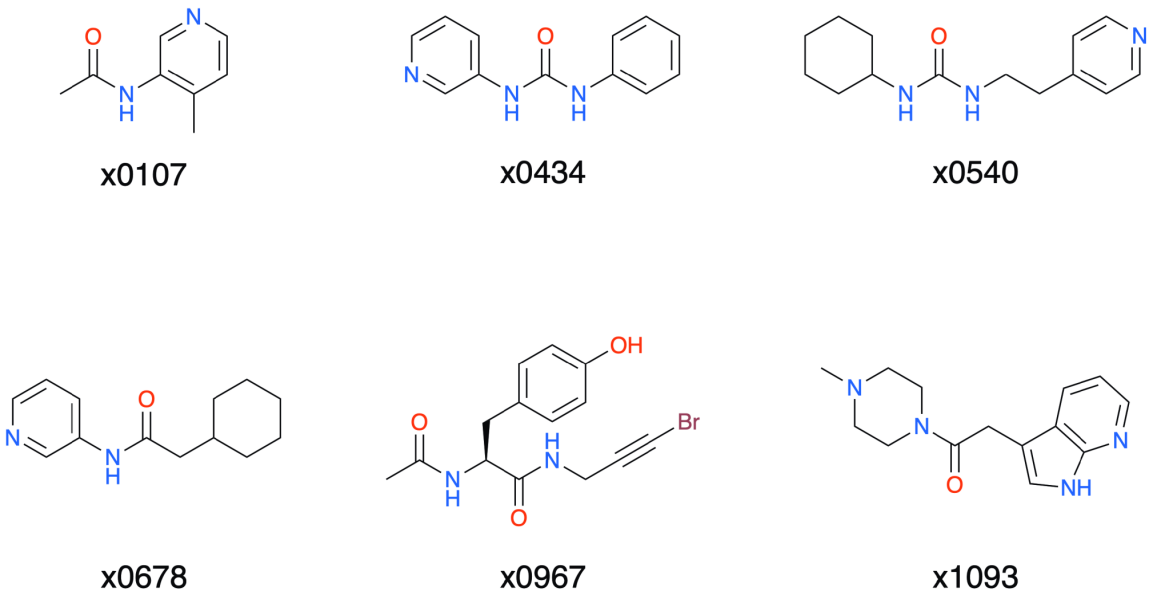

Figure 3.17: Chemical structures of the cluster 5 XChem fragments. Note the prevalence of nitrogen-containing heterocycles or in the case of x0967, the phenol containing tyrosine derivative structure, responsible for the hydrogen bonding in the S1 pocket.

In addition, most cluster 5 binders also extend into the hydrophobic S2 pocket, although there is no clear preference in functional group at S2. This agrees with our plasticity analysis, which shows that S2 can accommodate a large variety of functional groups (Figure 3.10). As seen in the overlap of peptide inhibitor p13 and cluster 5 representative x0678 (Figure 3.16 e), the binding modes of both inhibitors in the S1 and S2 subsites are very similar, with both forming HBs to His-163 (HB 6) and Glu-166 (HBs 2 & 3) and binding deelpy in the S2 pocket. In addition, all cluster 5 ligands (Figure 3.17) contain an amide or urea linker between the P1 and P2 binding groups, making them interesting building blocks for the development of peptidomimetics.

The structure of all analysed XChem fragments, the clustering code and the resulting clusters are available on Github (`https://github.com/gmm/SARS-CoV-2-Modelling/tree/main/Interaction_Clustering`).

3.4.4 Active-Guided Covalent Docking of COVID Moonshot Designs

Next, I utilized the large amount of structural data provided by the COVID Moonshot Project in the form of "Moonshot designs" to guide inhibitor design. "Moonshot designs" are structures of inhibitors submitted by crowd-sourced medicinal chemists and members of the COVID Moonshot consortium that use the structures of one or more of the previously described XChem fragments as inspriation [Chodera et al., 2020; Achdout et al., 2020]. At the time of this analysis (January 2021), 10,001 Moonshot designs had been submitted and were publicly available on the COVID Moonshot Project Github and the PostEra.ai website [COVID-19 Moonshot project, 2020; PostEra.Ai, 2020]. In order to use the existing structural information of the XChem fragments, I created the Active-Guided Covalent Docking (AGCD) approach based on AutoDock4 [Morris et al., 2009]. The docking workflow is shown in Figure 3.18. Since the majority of XChem fragments bound in the active site were covalent inhibitors and with the goal of maximising the amount of information used during docking, a covalent docking procedure was developed. Although AGCD uses constraints in the form of a covalent attachment point, it is not yet a constrained docking protocol. While there are fully implemented methods to do constrained docking in proprietary docking software such as Gold [Jones et al., 1997] or Glide [Friesner et al., 2004], there is not yet a straightforward way to do constrained docking in the most widely used open-source docking tools Autodock4 [Morris et al., 2009] and AutoDock Vina [Trott and Olson, 2010]. The development of AGCD is the first step towards the implementation of a dedicated constrained docking workflow for an open-source docking tool.

To accommodate induced fit and create high-quality poses of covalent inhibitors for future optimisation, I selected 540 compounds with covalent warheads from the

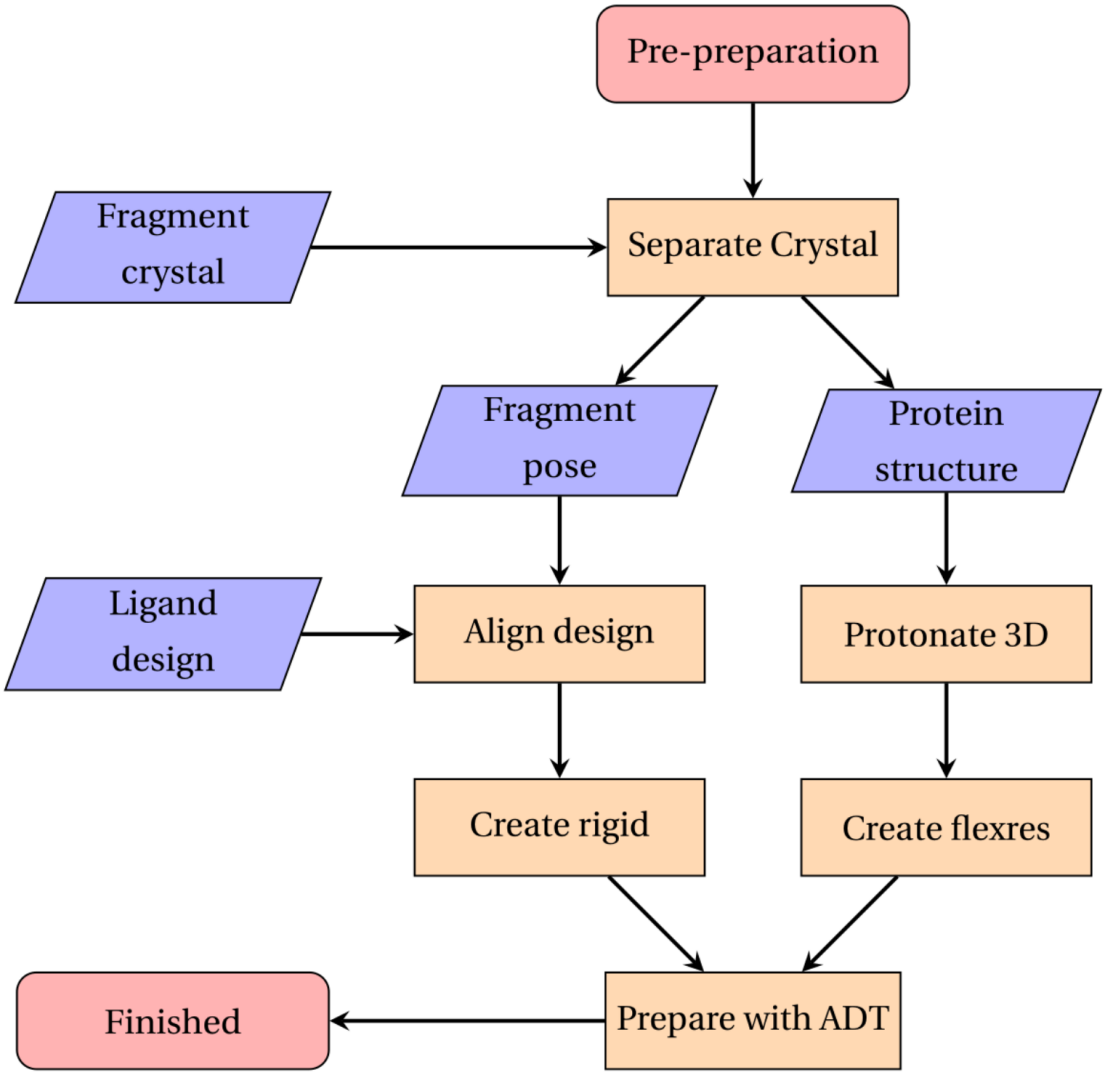

Figure 3.18: Flowchart showing the steps and data used during active-guided covalent docking.

10,001 Moonshot-designed compounds and docked them using the knowledge-based AGCD method using AutoDock4 [Morris et al., 2009]. The structure and docked pose for all 540 docked compounds is available on GitHub [Chan et al., 2021b]. Only compounds with a matching covalent warhead to the inspiration fragment that also cite a single covalent fragment as their inspiration were selected to form the dataset of 540 compounds. To take advantage of the many diverse induced-fit conformations of M^{pro},

each designed compound was docked into the M^{pro} structure of the corresponding covalent "inspiration fragment" (fragment referenced by chemist as the inspiration for a design). Lastly, to leverage the wealth of structural information about M^{pro} crystal structures, I developed an alignment algorithm named *MCS-Align* to expand on the *Constraint Embedding* method in RDKit [Landrum et al., 2006] (v2020.03.1) to pre-align each covalent Moonshot design to the corresponding inspiration fragment before covalent docking.

3.4.4.1 Constrained Alignment using *MCS-Align*

With the goal of ultimately building a knowledge-based constrained docking protocol, I developed a pre-alignment workflow based on RDKit (v2020.03.1) to force the conformation of two molecules to align, given a sufficiently large maximum common substructure (MCS). The standard RDKit [Landrum et al., 2006] *AlignMol* function only roughly aligns atoms in a molecule, rather than fixing atom positions exactly. I therefore created a stricter constrained alignment workflow using RDKit called MCS-Align. It uses pairwise atom constraints for all pairs of atoms in the maximum common substructure (MCS) between the designed molecule and the crystallographic fragment to force the position of the atom based on the crystallographic coordinates. This aligns the MCS between fragment and design exactly, forcing the common substructure into the crystallographic binding mode. This is followed by a constrained energy minimisation, while keeping the atomic coordinates of the MCS constant. An implementation of the *MCS-Align* method can be found on GitHub (`https://github.com/MarcMoesser/SARS-CoV-2-Modelling/tree/main/Covalent_Docking/ligand_alignment`).

In the first step, the algorithm calculates the maximum common substructure between the target molecule and the template using the RDKit *FindMCS* function. In

the next step, the position of each atom in the MCS from the crystallographic fragment is retrieved and the target atoms adjusted, using the RDKit functions *GetAtomPosition* and *SetAtomPosition*, respectively. Finally, a random 3D conformer is initialized and a constrained energy minimization performed using the Universal Force Field (UFF) [Rappe et al., 1992] implementation in RDKit with a strict distance constraint using the RDKit *AddDistanceConstraint* function and a relative force constant setting of 100. The resulting alignment method is able to align substructures even in extremely complex molecules that the native RDKit *ConstrainedEmbed* function either is unable to align at all or only roughly aligns.

3.4.4.2 Active-Guided Covalent Docking Results

In order to evaluate the docked poses, I analysed to what extent the docking procedure was able recapitulate the binding pose of the parent fragment when docking the fragment-based Moonshot designs. As previously shown by Malhotra and Karanicolas [2017], a fragment elaborated small molecule ligand and the original fragment it is derived from are highly likely to have the same binding mode. I therefore compared the shape and pharmacophoric overlap using SuCOS [Leung et al., 2019; Moesser, 2021] of the lowest energy pose of the highest populated cluster for each Moonshot compound with the inspiration covalent XChem fragment referenced by the designers (Appenidx Figure B.9). A SuCOS score of 0.55 or higher is generally considered sufficient to consider the binding poses of the crystallographic fragment and docked design as conserved [Leung et al., 2019], being equivalent to having an RMSD of less than 2 Å. Likely in part due to the creative freedom in the design process, some of the designed compounds do not overlap significantly with the inspiration fragments and in some extreme cases only have the covalent warhead in common. When controlling for the smallest maximum common substructure (MCS) that encompasses at

least the covalent warhead and one additional atom in the compound, 379 docked designs remain, from which 132 (34.8%) recovered the binding mode of the inspiration fragment. Given the high similarity between the fragments and the docked designed compounds, it is likely that these binding modes are more representative of the actual binding mode of the ligand.

At the time of this analysis (January 2021), only 6 of the 540 docked covalent compounds had been crystallised by the COVID Moonshot Consortium and deposited in Fragalysis [Diamond, 2020]. These structures were used as a limited benchmark for the docking method. For comparison between the binding modes of the docked designs and crystal fragments, an overlay of the crystallographic conformation of the Moonshot design, the lowest energy pose of the highest populated cluster of the design from docking, and the corresponding crystallographic structure of the inspiration XChem fragment is shown in Figure 3.19.

Fragment x10899 (Figure B.10) was excluded from further analysis since it binds via a crystal interaction to a third symmetry-related M^{pro} molecule, rather than the biologically relevant dimeric state. The binding modes of two compounds, x3077 and x10306 (Figure 3.19 a & e) respectively), were reproduced almost identically, which is reflected in their SuCOS score between inspiration fragment and docked pose (0.88 and 0.82 for x3077 and x10306, respectively). In the case of fragment x3324 (Figure 3.19 b) , docking places the aromatic sidechain correctly into the S2 pocket of M^{pro} but varies on placement of the linker when compared to the crystal structure. However, the provided inspiration fragment has minimal overlap with the designed compound (SuCOS = 0.51). Since the induced fit AGCD approach aims to leverage protein and ligand conformations of known actives to improve docking, in cases with low overlap (x3324), it is to be expected that docking pose quality would be lower than for high overlap cases (x3077 and x10306).

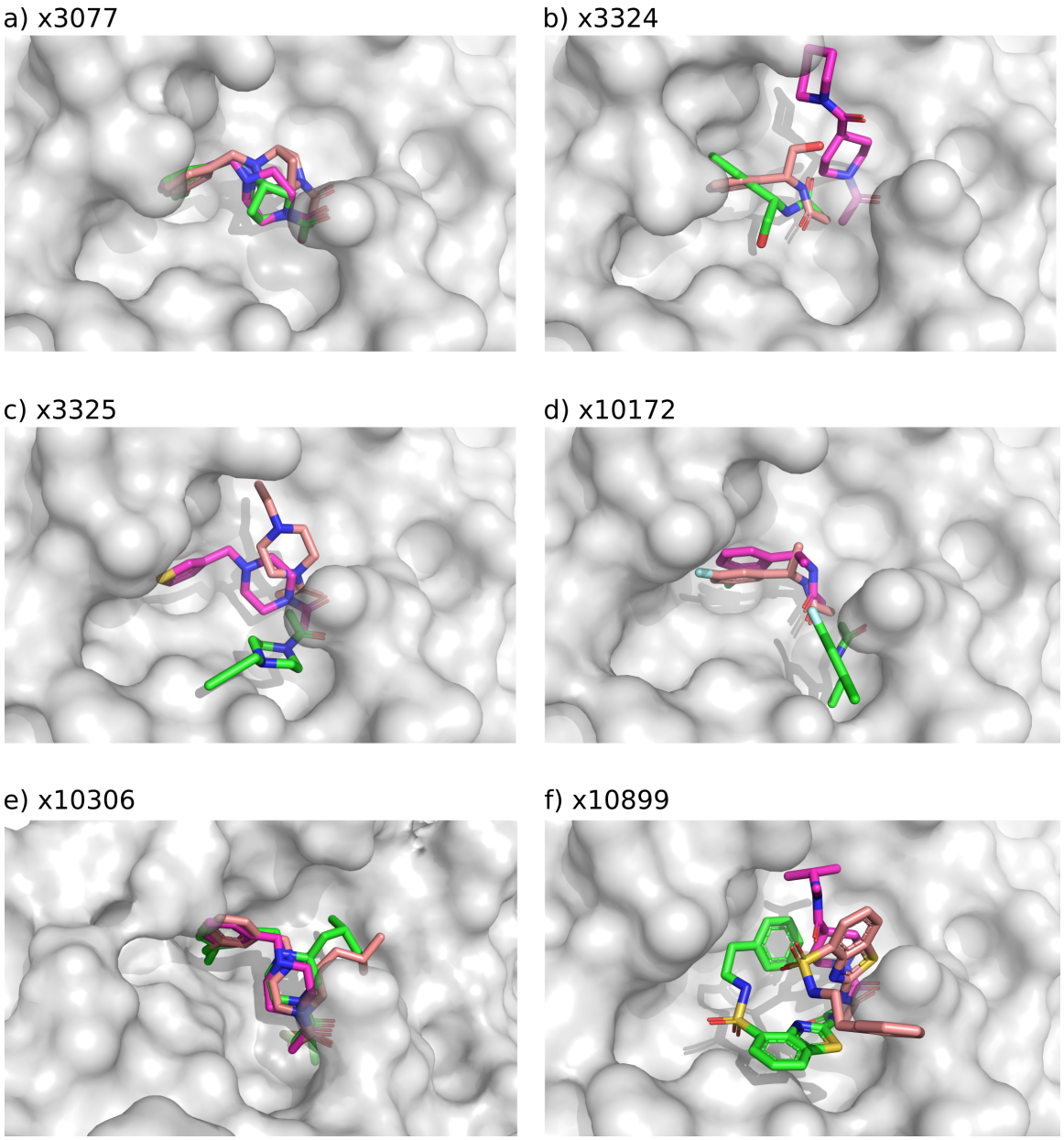

Figure 3.19: Overlay of the lowest energy pose in the highest populated cluster of the AD4 covalent docking procedure for the design (green) with the crystal structure of the original inspiration fragment (pink) and the crystal structure of the design (salmon). For every docking, the M^{pro} protein structure of the corresponding inspiration fragment co-crystal structure was used. a) Moonshot design X3077 (salmon) with inspiration fragment X0770 (pink) and the docked pose of X3077 (green). b) Moonshot designed compound X3324 (salmon) with inspiration fragment X1380 (pink) and the docked pose of X3324 (green). c) Moonshot design X3325 (salmon) with inspiration fragment X1386 (pink) and the docked pose of X3325 (green). d) Moonshot design X10172 (salmon) with inspiration fragment X1382 (pink) and the docked pose of X10172 (green). e) Moonshot design X10306 (salmon) with inspiration fragment X0770 (pink) and the docked pose of X10306 (green). f) Moonshot design X10899 (salmon) with inspiration fragment X1458 (pink) and the docked pose of X10899 (green). Poses of the designs x3077 and x3324 almost perfectly resemble the original crystal pose.

Finally, for x3325 and x10172 (Figure 3.19 c) & d), respectively), the selected lowest-energy pose of the highest populated cluster did not match the binding pose of the crystal structure and the corresponding SuCOS between design and inspiration fragment low in both cases (0.09 and 0.03 for x3325 and x10172, respectively). Docking did not produce high quality poses in these cases.

The structure of all docked COVID Moonshot designs, all docked poses and the selected, low energy pose for each compound is provided on Github (`https://github.com/gmm/SARS-CoV-2-Modelling/tree/main/Covalent_Docking`).

3.4.4.3 Interaction Fingerprint Clustering of the Docked Poses

The lowest energy docked pose in the highest populated cluster of each docking of the Moonshot designs was used to identify interactions using Arpeggio and generate the interaction Tanimoto distance matrix. Tanimoto similarity thresholds of 0.5 and 0.7 was employed for pose clustering as described above for the XChem fragments. The broader clustering threshold of 0.5 leads to a total of five clusters, with the first cluster containing 477 of the 540 poses (88%) and no single-pose clusters; while the stricter threshold of 0.7 results in 46 clusters with 12 single molecule clusters. As expected, the diversity of the binding modes for these compounds is much lower than in the original XChem fragment set, due to the limited number of fragments (44) and the reduced structural diversity of the designs, all being covalent S1/S1′ binders. As a result, I developed the following *in silico* design approach (Section 3.4.5) to filter the docked Moonshot designs further in order to identify promising elaboration pathways for known M^{pro} fragment-based inhibitors.

3.4.5 Implications for Future Inhibitor Design

Finally, an *in silico* design approach was developed to bring together different insights gained throughout this project. For that purpose, I compared the interactions of

the cluster 5 fragments with those in the substrates, peptide inhibitors, crystallised Moonshot designs and docked Moonshot designs.

First, I filtered the known crystal structures of 245 Moonshot designs obtained from Fragalysis [Diamond, 2020] by their overlap to the key interactions identified by Arpeggio in the fragment cluster 5. A compound was considered a cluster 5-type binder if at least 70 % of the cluster 5 interactions are present in the complex. Interestingly, unlike the peptides, almost none of the identified Moonshot compounds that exhibit cluster 5-like binding interact with the oxyanion hole of M^{pro}. The only crystallised cluster 5 Moonshot compounds where this interaction is made are a series of covalent inhibitors, none of which showed promising potency (Appendix B, Figure B.11). An exhaustive search of resolved Moonshot crystal structures showed that at the time of the analysis, no non-covalent inhibitor had been tested that includes both the typical cluster 5 binding mode while also being able to interact with the oxyanion hole, which presents a potential route for the elaboration of cluster 5 binders.

In order to identify compounds that could cover both, cluster 5 binding sites and the oxyanion site, the covalently docked Moonshot designs were also analysed for cluster 5 overlap. Out of all 540 docked designs, only 3 compounds were identified to cover key cluster 5 interactions: FOC-CAS-e3a94da8-1, MIH-UNI-e573136b-3 and NIR-THE-ed286faa-1. For further study, FOC-CAS-e3a94da8-1 and MIH-UNI-e573136b-3 were selected based on their high normalized SuCOS overlap (0.73 and 0.63, respectively) with their inspiration fragments, strongly suggesting that their docked binding modes reflect the actual poses [Malhotra and Karanicolas, 2017] as opposed to NIR-THE-ed286faa-1 (SuCOS of 0.11). Both, FOC-CAS-e3a94da8-1 and MIH-UNI-e573136b-3 were also found to bind into the oxyanion hole as well as into S1 and S2, providing a clear opportunity for extension of the cluster 5 binders (Figure 3.20 a & Appendix B, Figure B.12).

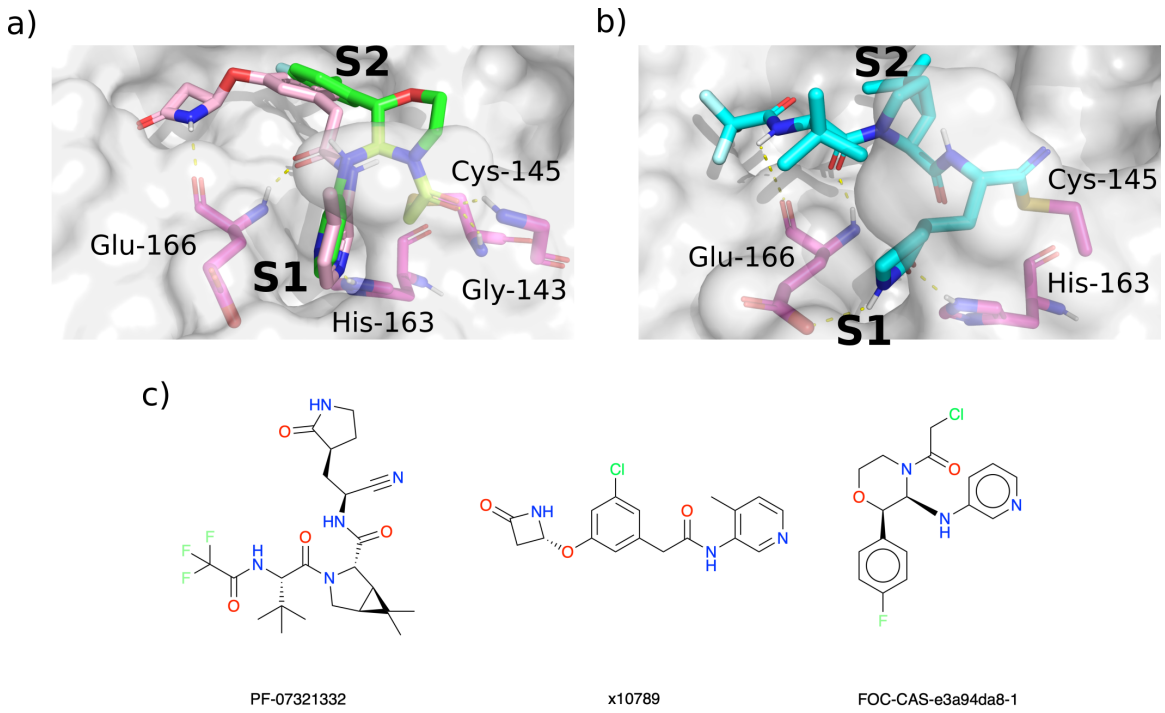

Figure 3.20: Docking informs novel inhibitor design. HBs between M^{pro} residues (magenta) and the ligands are shown as dotted yellow lines. a) Overlay of the docked pose of FOC-CAS-e3a94da8-1 (green and greenish-yellow) with the crystal structure of x10789 (pink) on the M^{pro} surface (PDB entry 5RER; 1.88 Å resolution). Derivatisation of x10789 into the oxyanion hole could be achieved by attaching a methylene amide group present in x0830 (highlighted greenish-yellow). b) Docked pose of Pfizer's FDA approved drug compound PF-07321332 (Nirmatrelvir), covalently docked into M^{pro} (6XHM; 1.41 Å resolution) [Hoffman et al., 2020]. Nirmatrelvir (cyan) is covalently attached to Cys-145. The docked Nirmatrelvir adopts the same major interactions as the "combination" of x10789 and x0830, namely the double HB to the backbone of Glu-166, the HB to His-163 in the S1 subsite, and a series of hydrophobic interactions in the S2 subsite. c) Structures of Moonshot designed compound FOC-CAS-e3a94da8-1, crystallographic fragment x10789, and inhibitor Nirmatrelvir. Overall, the expansion into the oxyanion hole for the Moonshot compound series would enable binding to exactly the same key interactions as the approved drug Nirmatrelvir.

For the XChem fragments, most cluster 5 binders place the aromatic heterocycle into the S1 site and the carbonyl oxygen of the amide linker bonds to Glu-166 (Figure 3.16). In the case of FOC-CAS-e3a94da8-1, the position of this amide nitrogen linker on x10789 overlays perfectly with the ring amine present in the docked compound FOC-CAS-e3a94da8-1. Thus, extension of cluster 5 binders into the oxyanion hole could be achieved by adding a substituent at the amide nitrogen. A promis-

ing candidate for extension is x10789 (IC_{50} = 3.6 μM, determined by fluorescence assay [Achdout et al., 2020; COVID-19 Moonshot project, 2020; PostEra.Ai, 2020]), which makes a HB with the backbone oxygen of Glu-166 (Figure 3.20 a) and mimics the non-prime side binding mode of peptide inhibitor p13 (Figure 3.16 e), even binding into the S4 site via its β-lactam ring (Figure 3.20 a). Additional expansion into the oxyanion hole and S1′ through the amide linker could yield a powerful peptidomimetic inhibitor, combining protein interactions observed for the substrates, peptide inhibitors and small molecule fragments.

At the time of analysis, no crystal structure of the then clinical drug candidate PF-07321332 (Nirmatrelvir) was known. Therefore, in order to compare the fragment-based inhibitor design approach with Nirmatrelvir, I covalently docked Nirmatrelvir into the active site of M^{pro} structure of 6XHM [Hoffman et al., 2020], which is a crystal structure in complex with ligand PF-00835231, a precursor in the design of Nirmatrelvir [Hoffman et al., 2020; Owen et al., 2021]. Then, in November 2021, the team at Pfizer deposited a high resolution crystal structure of Nirmatrelvir in complex with M_{pro} in the PDB (7RFS, 1.91 Å) [Owen et al., 2021]. In order to quantify the quality of the docked pose of Nirmatrelvir, I calculated the symmetry-aware RMSD between the docked pose of Nirmatrelvir and the crystal pose (2.08 Å) as well as the the SuCOS score (0.48), indicating that the docked pose almost resembles the crystal pose. An overlay of the crystal and docked pose can be found in the Appendix B, Figure B.13. While the docked pose closely resembles the crystal structure in the S1 site and around the covalent attachment point to Cys-145, it deviates around the S2 and S3 site (Appendix B, Figure B.13). This result is in accordance with the plasticity analysis (Section 3.4.1.5), which indicates that the S2 and S3 sites are highly plastic and can change strongly upon ligand binding, making rigid docking in this pocket challenging. However, note that for docked Nirmatrelvir, AutoDock4

was unable to place the negatively charged azanide nitrogen in the oxyanion hole, which is the expected position given its similarity to related warheads previously docked. In addition, while it was assumed before docking that covalent reaction of Nirmatrelvir with M^{pro} would result in a negatively charged azanide, the authors of 6XHM chose to model a neutral, protonated nitrogen at that position [Owen et al., 2021]. Nonetheless, the key interactions between Nirmatrelvir and M^{pro} were retained between the crystal and docked pose, with key hydrogen bonds formed with Glu-166 and His-163, and hydrophobic interactions deep in the S2 pocket.

Finally, when comparing interactions exhibited by cluster 5 binders (Glu-166, His-163) or covalent fragments (Gly-143, Cys-145) with the interactions present in the docked structure of Nirmatrelvir (Figure 3.20), a nearly identical interaction pattern to the cluster 5 binding motif is observed, validating the *in silico* design approach outlined in this work.

3.5 Conclusions

The methods that I developed in this project and the results I generated were all part of an ongoing international collaboration to understand better understand SARS-CoV-2 on a molecular level in order to contribute to the ongoing drug discovery effort to tackle the COVID-19 pandemic. Our collaboration was formed around understanding how SARS-CoV-2 M^{pro} binds to its 11 natural substrates on a molecular level and how this detailed understanding of its substrate binding might be used for the design of peptide- and small molecule inhibitors. This chapter details my contribution to this collaboration, focusing on the detailed interaction analysis of the M^{pro} substrates, peptide inhibitors, and all fragments and covalent inhibitors in the COVID Moonshot Project database available at the time of this study (accessed January 2021) [Chodera et al., 2020; Achdout et al., 2020; Douangamath et al., 2020].

However, this work would not have been possible without the effort of many collaborators, especially my co-first authors on the original publication: H.T. Henry Chan, Tika R. Malla and Rebecca K. Walter [Chan et al., 2021a]. Throughout this chapter and the corresponding Appendix B, I have marked their contributions to this work as appropriate. Together, this collaborative work combines insights from a large variety of computational and laboratory-based drug discovery techniques, spanning explicit-solvent molecular dynamics, protein-ligand docking, cheminformatics, as well as mass spectrometry and peptide synthesis to create a comprehensive and exhaustive analysis of the SARS-CoV-2 M^{pro} activity on a molecular level, creating a case study on how *in silico* driven insights from enzymology can be used effectively in inhibitor design.

Understanding the structural and dynamic features responsible for catalysis and substrate selectivity is challenging in the case of M^{pro} due to its dimeric nature and multiple substrate sequences. While a detailed understanding of a protease on this level might not be absolutely necessary for the discovery of potential drugs, the drug quality and the efficiency with which they are developed may be improved. Indeed, the peptide inhibitors created during this work were designed completely *in silico* with a 100% hit-rate, as all synthesized and tested peptides were competitive inhibitors (Appendix B, Figure B.4). In addition, due to the high propensity of SARS-CoV-2 to mutate, as has been observed since the outbreak in the formation of hundreds of different lineages and several SARS-CoV-2 variances of concern such as most recently, the Delta variant [Rambaut et al., 2020; O'Toole et al., 2021], concerns are high that SARS-CoV-2 mutations might affect drug resistance (including resistance against M^{pro} inhibition), similar to the global effort against the HIV pandemic. Having identified the key interactions that drive substrate binding and selectivity/recognition, inhibitors could be designed that specifically target those interactions, under the

assumption that key amino acids driving the substrate selectivity of the key protease in the viral life cycle are unlikely to mutate, therefore potentially creating a mutation resistant drug.

The 11 natural substrate models that were developed using comparative modelling followed by explicit solvent MD were used as a basis for the analysis of key interactions involved in substrate binding and recognition. Using the structural biology tool Arpeggio [Jubb et al., 2017], I conduced a detailed interaction analysis of the 11 substrate-Mp_{pro} complexes. This methodology expands upon the known interaction fingerprint methods such as SIFt [Singh et al., 2006], IFP [Marcou and Rognan, 2007] and pSIFt [Chuaqui et al., 2005] which have primarily been developed for virtual screening and to calculate the similarity of ligand poses. The usage of the expanded interaction identification tool Arpeggio [Jubb et al., 2017] to cover more significant interactions as well as a more in-depth analysis of the fingerprints through clustering and key interaction identifications has enabled Arpeggio-derived fingerprints to be useful beyond similarity calculations and initial hit-compound identification, enabling them to be directly used to guide inhibitor design and formulate a detailed binding hypothesis of the peptide substrates of M^{pro}. In addition, I have shown that Arpeggio-derived interaction fingerprints can be used on several levels of detail beyond traditional interaction fingerprints, enabling a detailed, atomic level analysis, a summarized, residue-level analysis and on the highest level a description of the protein subsites and their character.

The analysis revealed a series of important trends. First, the P′ (C-terminal) side of the substrates was found to bind much less tightly than the P (N-terminal) side (Figure 3.8 & 3.6). Hydrogen bonds as well as other non-covalent interactions found on the P-side showed remarkable consistency across all substrates (Figure 3.8). One explanation of this binding behaviour might be the need for the P′ side of the substrate

to leave after acyl-enzyme formation prior to hydrolysis. It is therefore possible that P-side residues are more important for initial substrate recognition and suggest that inhibitors targeting S-side binding sites could be more effective as substrate competing inhibitors. Indeed, other known potent inhibitors such as the peptidometic inhibitor N3 [Jin et al., 2020] and the Pfizer drug PF-07321332 (Nirmatrelvir) [Owen, 2021] (Figure 3.20) both bind almost exclusively in the S subsites. However, the fragment-based clustering analysis combined with molecular docking shows that there is considerable scope for the growth of known S1-S3 binding inhibitors into the oxyanion hole and towards the S1′ site and beyond (Figure 3.20). Although, more effort might be required to optimize S′ binders towards high affinity.

Next, beyond the overall subsite analysis, I focused on the analysis of the P1 Gln recognition motif present at S1 in M^{pro} and the critical role that the Gln residue plays in productive substrate binding. The interaction analysis reveals that HBs 6 & 7 are directly involved in Gln recognition, binding to the amide carbonyl and amide nitrogen, holding the Gln side chain deep within the S1 pocket (Figure 3.7 & 3.6). In addition, strong hydrogen bonds are formed between the backbone nitrogens of M^{pro} residues Glu-143 and Cys-145 (the oxyanion hole) to the backbone carbonyl oxygen of Gln. While this interaction is not necessarily unique to Gln, in aggregate, the P1 Gln residue forms a total of 4-5 HBs (HB 5 not always formed), the most out of any substrate residue. As identified by computational alanine scanning mutagenesis in peptide p14 (Appendix B, Figure B.3) and the XChem fragment clustering experiment (Figure 3.16), a potential replacement for the Gln binding motif could be a nitrogen containing heterocycle such as the indole ring on the P1 Trp in peptide p14, or the pyridine rings present in cluster 5 binders (Figure 3.17). This case study is an important example in how structural insight into substrate binding can directly inform inhibitor design and functional group choice.

While the P1 residue is completely conserved in all M^{pro} substrates, the residues at P2 were found to be highly conserved in regards to their hydrophobic character (Figure 3.9). In 9 out of the 11 substrates, Leu was present at P2, with Phe and Val present at P2 in s02 and s03, respectively (Figure 3.1). Although no specific hydrogen bonds were formed, the interaction analysis indicates that the deep hydrophobic pocket is nonetheless crucial for recognition and inhibition. Due to the high plasticity at the S2 site (Figure 3.10), M^{pro} is able to accommodate a range of different hydrophobic residues, including significantly larger groups than the Leu residue present in most substrates, although not always in a productive conformation. BAlaS-guided design of peptide inhibitors predicts a large gain in binding affinity upon exchange of the P2 residue to large aromatic amino acids in 4 out of 5 designed peptides (Trp: p12, p13 and p15; Phe: p14, Appendix B, Figure B.3). As shown by the interaction analysis of the MD-derived pose of peptide p13 (Figure 3.13) , the loss of M^{pro} activity upon binding might be due to the π-π stacking of the Trp indole to His-41 in the S2-site, hindering His-41 to adopt a productive conformation as part of the catalytic dyad. Alternatively, studies with other nucleophilic proteases such as elastase revealed that substitution of the scissile residue can by itself already cause inhibition [Wright et al., 2001; Wilmouth et al., 1997]. However, computational and laboratory-based examination into why the designed peptides are competitive inhibitors instead of substrates has not been concluded yet, and the reason not fully known. In addition, further SAR exploration should be done to convert the peptide inhibitor to a peptidomimetic inhibitor, possible starting at the derivatisation of Trp at P2 to maximize binding affinity or the cyclisation *via* insertion of a methylene group linking position 2 at the indole ring to the backbone nitrogen of Trp [Castelli et al., 2015]. Furthermore, the flexibility of the S2 site can also be exploited by small molecule inhibitors. Some cluster 5 inhibitors are able to connect from the S1 site into S2 with

a variety of functional groups (Figure 3.16). Since the hydrophobic effect of binding can be increased by increasing the hydrophobic surface area involved in binding, a plastic hydrophobic pocket could be exploited by maximizing the ligand-M^{pro} contact area at S2, increasing ligand binding affinity. Nonetheless, since even substrates with the same residue at P2 (Leu) are turned over at different rates suggest that more interactions beyond the P1 and P2 sites are important (Figure B.1). Therefore, interactions beyond the immediate active site are likely to contribute significantly to selectivity of both: substrate binding as well as modulating the rate of reaction of the enzyme–substrate complex.

Beyond the substrates and peptide inhibitors, I conducted an exhaustive interaction analysis of all existing fragments and Moonshot design in the COVID Moonshot Project database [Chodera et al., 2020; Achdout et al., 2020; Douangamath et al., 2020] (Figure 3.14). First, I analysed the binding mode and interactions between all 91 XChem fragments [Douangamath et al., 2020] obtained from Fragalysis [Diamond, 2020] and compared them to the previously identified substrate interactions. I created a new Arpeggio-derived fingerprint specific towards M^{pro}-ligand binding that encodes for the absence or presence of residue-level interactions for each ligand (Figure 3.5). Using this fingerprint representation, I clustered all active site binding fragments, identifying the promising binding cluster 5 (Figure 3.16). Fragments in this cluster occupy several of the key interactions described above (Figure 3.16): i) a heterocyclic moiety occupies S1 and forms a HB with His-163, blocking the key Gln recognition HB 6; ii) the amide (or urea) linker between the heterocycle in S1 and the rest of the fragment forms a HB with the backbone nitrogen of Glu-166, blocking the key stabilising HB 3 which is observed in every substrates; iii) most cluster 5 binders feature a hydrophobic ring on the opposite end of the linker, occupying the hydrophobic S2 pocket. Cluster 5 binders stretch from S1-S2, while occupying key interactions

and binding motifs identified as crucial for substrate binding and recognition and are therefore interesting starting points for fragment elaboration. In addition, a large covalent cluster (cluster 1, Figure 3.16) was identified which overwhelmingly formed HBs 8 and 9 with the oxyanion hole (Gly-143, Cys-145). This interaction was not found in cluster 5 and could thus be used as a promising elaboration path.

Furthermore, by combining all the above mentioned insights with the covalent docking campaign conducted on covalent COVID Moonshot designs, I identified a direction for future inhibitor design (Figure 3.20). When combining a known small fragment elaboration from Fragalysis (x10789 [Diamond, 2020]), that was found to have a high Tanimoto similarity with cluster 5 binders, with the docked pose of inhibitor design FOC-CAS-e3a94da8-1, a path for elaboration of x10789 into the oxyanion hole was discovered. At the time of the creation of this inhibitor design strategy, the structure and existence of Pfizer drug PF-07321332 was not known. However, as is known now, compound PF-07321332 exploits the same "privileged" interactions that I designed future inhibitors to bind and the binding pose of PF-07321332 occupies the same space in the binding pocket as the most promising moonshot compounds and the elaboration designs identified through my interaction analysis approach (Figure 3.20).

Lastly, the development of AGCD represents the first step towards the implementation of a dedicated constrained docking workflow. The ultimate goal of the AGCD method is to create a knowledge-based constrained docking approach for fragment-based drug discovery where new compounds are docked by constraining their MCS with a known fragment crystal structures during docking, while allowing flexibility elsewhere on the ligand. This approach is inspired by the findings of [Malhotra and Karanicolas, 2017] who found that for elaborated ligands, in 86% of the 297 paired ligands, the larger elaborated ligand did not change its binding mode relative to

the smaller ligand. While there are fully implemented methods to do constrained docking in proprietary docking software such as Gold [Jones et al., 1997] or Glide [Friesner et al., 2004], there is not yet a straightforward way to do constrained docking on this scale in the most widely used open-source docking tools Autodock4 [Morris et al., 2009] and AutoDock Vina [Trott and Olson, 2010]. Future work should therefore focus on the implementation of the needed functionalities into either AD4 or AutoDock Vina to create a dedicated open-source constrained docking method to advance fragment-based drug discovery.

This chapter continues the overall topic of this thesis about finding the best possible representation of a given drug discovery problem (in this case the question of protein-ligand complex binding modes). I created a practical, easy to compute and highly effective protein-ligand interaction fingerprint that was used as a vector representation of the 3D interactions in M^{pro}-ligand complexes. During this work on protein-ligand interaction fingerprints, I searched the literature for other forms of protein-ligand interaction representations and found methods such as the widely known machine learning-based protein-ligand affinity scoring function PLEC[Wójcikowski et al., 2018] that encode interactions as Extended Connectivity Fingerprints [Rogers and Hahn, 2010] between the protein and ligand atoms and a new atom-atom interaction fingerprint called ECIF [Sánchez-Cruz et al., 2020]. ECIF encode protein-ligand atom-atom interactions by pre-computing all possible combinations of protein atoms and ligand atoms in a dataset and then counting the occurrence of each unique atom-atom interaction in 3D. Inspired by ECIF and my work on Arpeggio-based interaction fingerprints, I wondered if there is a way to encode interactions similar to ECIF without the need to pre-compute the feature space and having to re-train the model every time a new, unknown data point is added. The next chapter details by work on creating Protein-Ligand Interaction Graphs which

follows on from my previous work on affinity prediction and combines it with the interaction analysis done in this chapter to create a novel protein-ligand binding affinity scoring function.

Chapter 4

Protein-Ligand Interaction Graphs: Learning from Ligand-Shaped 3D Interaction Graphs to Improve Binding Affinity Prediction

4.1 Preamble

As my work described in Chapter 3 on SARS-CoV-2 M^{pro} came to a conclusion with the publication of our collaborative effort [Chan et al., 2021a], I continued working on the idea of creating an interaction-based representation of protein-ligand complexes that simultaneously account for 3D interactions and ligand structure. The Arpeggio-based fingerprints described in Chapter 3 were useful in distinguishing between the binding modes of ligands *via* clustering and in conducting direct contact-contact mapping, but they do not allow for a more detailed encoding of the biophysics of interactions and give no further detail about the shape of the ligand. I therefore explored the idea further, searching for a way to improve the detail with which the interaction is encoded, such as classifying the interaction on an atom-atom level, rather than the residue level used during clustering in Chapter 3, without compromising the simplicity of the model. Overall, a variety of different protein-ligand interaction fingerprints are know, although in all cases, the encoding of interactions was compressed into

a feature vector to be used in machine learning models, most often random forests or gradient boosted trees [Wójcikowski et al., 2018; Singh et al., 2006; Gao et al., 2020; Sánchez-Cruz et al., 2020]. While working on this issue, Sánchez-Cruz et al. [2020] published a new interaction fingerprint named Extended Connectivity Interaction Features (ECIF), that uses similar concepts as my work on Arpeggio-fingerprints, but classifies interactions as the combination of two unique atom types in proximity to each other. Insprired by the work of Sánchez-Cruz et al. [2020], I combined the ECIF approach to classification of protein-ligand interactions with the molecular graph representation of the emerging field of molecular Graph Neural Networks (GNNs) that has been increasingly used for protein-ligand affinity prediction [Nguyen et al., 2020; Lim et al., 2019; Li et al., 2021] creating a novel molecular graph representation called "Protein-Ligand Interaction Graphs" or PLIGs for short.

ECIF fingerprints need to re-define the dimensionality of the fingerprint every time a new ECIF atom type is added (*e.g.* when screening a new compound) and thus need to be re-trained from scratch. Most recent graph-based methods such as SIGN [Li et al., 2021] or the work of Lim et al. [2019] encode 3D protein-ligand complexes either into massive graphs (*e.g* by incorporating protein nodes into the graph), reducing the advantage graphs have when representing the actual chemical structure of the underlying ligand. PLIGs are able to overcome the limitations of both ECIF fingerprints and current graph-based protein-ligand affinity scoring functions.

The following chapter details my work on the development and benchmarking of PLIGs with different GNN architectures on the CASF-2016 benchmark set [Su et al., 2018]. The work in this chapter was published (on March 7th, 2022) as a preprint on BioRxiv (`https://doi.org/10.1101/2022.03.04.483012`, [Moesser et al., 2022]) where it has already gathered significant attention by the community, with over 906 full text views (PDF and HTML) and over 1484 abstract views.

This project includes contributions from Dominik Klein who conducted his Master thesis under my and Prof. Garrett Morris' supervision. Dominik Klein contributed to this project by first re-implementing the models originally published in the GraphDTA method by Nguyen et al. [2020] to reproduce the results therein, and by conducting the hyperparameter tuning of the models used in this work (see Appendix C, Section C.1). Contributions by Dominik Klein are indicated when applicable, but all other results and methods were produced by myself.

4.2 Introduction

In early stage pre-clinical drug discovery, one of the most important properties of a small molecule drug is its binding affinity for the correct protein target. High binding affinity is crucial for the overall efficacy of a drug while the careful design of multi-target affinity profiles is crucial to avoid toxicity and side-effects. In addition, higher binding affinity allows drugs to be administered at lower doses to generate the desired efficacy, which reduces overall toxicity and increases practicality. Computer-aided drug design (CADD) has been firmly established as a powerful technique in the drug discovery pipeline, and increasingly machine-learning-based methods [Vamathevan et al., 2019].

In particular, interest in applying machine learning (ML) to the development of more accurate scoring functions that can predict protein-ligand binding affinity has grown in the last decade. Classical scoring functions use physics-based methods (force fields), linear combinations of (semi-)empirical terms, or knowledge-based potentials. They are used in popular docking software such as AutoDock4 [Morris et al., 2009], AutoDock Vina [Trott and Olson, 2010] or GOLD [Jones et al., 1997] to score 3D ligand poses. While these classical scoring functions, and espcially the scoring functions employed in docking software, perform well in docking and virtual screening

tasks, they struggle with binding affinity prediction and ranking tasks [Li et al., 2018; Su et al., 2019]. Especially during early stage drug discovery, namely the hit-to-lead and lead optimization processes, it is highly desirable to be able to accurate predict the protein-ligand binding affinity of potential ligands in order to prioritize promising candidates and obtain highly potent ligands.

For this purpose, more recently, ML-based scoring functions that can outperform classical scoring functions for binding affinity prediction have emerged. These models employ a diverse set of ML architectures and features, from classical ML techniques such as random forests and gradient boosted trees [Sánchez-Cruz et al., 2020; Boyles et al., 2019; Durrant and McCammon, 2011] to deep learning (DL) models [Jiménez et al., 2018; Li et al., 2021; Nguyen et al., 2020]. As the most accurate scoring functions use many different ML architectures, it is not yet clear how protein-ligand complexes should be featurised for model training. Older quantitative structure activity relationship (QSAR) models, mostly used against a single protein target, use ligand-based features such as ECFP fingerprints [Rogers and Hahn, 2010] or computed molecular descriptors [Cherkasov et al., 2014]. Recent scoring functions that aim to be able to generalise across multiple proteins incorporate additional 3D-structural information with the main goal of creating models that can learn the biophysics of protein-ligand interactions, rather than regurgitate biases in their training sets.

In general, since the binding pockets of proteins can differ dramatically between protein families, the structure of a matching ligand will also be subtly different. Models learning purely ligand-based features should therefore be unsuited to learn how to evaluate ligands against two or more protein targets (unless they are very similar) and should only be used in isolation when building single protein target models. Existing models utilize different approaches to featurisation based on the ML architecture employed. Models using classical ML models such as random forests

use interaction based fingerprints (*e.g.* PLEC [Wójcikowski et al., 2018] and ECIF [Sánchez-Cruz et al., 2020]) or include feature vectors based on 2D and 3D descriptors [Boyles et al., 2019, 2021]. DL-based models such as 3D convolutional neural networks (CNN) use voxelized representations of the protein-ligand complex [Jiménez et al., 2018]. Another very successful approach in building scoring functions that can learn from interactions is the use of Atomic Environment Vectors (AEV) as described by Meli et al. [2021] which combines a deep learning approach with features derived from atom-centred symmetry functions using the ANI neural-network potential [Smith et al., 2017].

Graph-based neural networks (GNN) have recently emerged as a powerful method for protein-ligand binding affinity prediction [Nguyen et al., 2020; Li et al., 2021; Karlov et al., 2020]. While different methods usually use differently engineer graph representations as well as different GNN architectures, the search for the optimal architectures as well as the best graph representation for protein-ligand binding affinity scoring functions is still ongoing. Although simple ligand-based graphs alone are already useful for affinity prediction [Nguyen et al., 2020], higher performance for affinity prediction and virtual screening tasks on multi-target datasets has been achieved in graphs that can incorporate 3D-structural information [Li et al., 2021; Lim et al., 2019].

In order to encode 3D information, previous studies have chosen to expand the graph itself, by including nodes corresponding to protein atoms that are close to the ligand [Li et al., 2021; Lim et al., 2019]. This expands the graph's complexity considerably while obfuscating the topology of the ligand, reducing the advantage that GNNs might have when encoding molecular structures, especially when it comes to full transparency and interpretability. In this work, I present Protein-Ligand Interaction Graphs (PLIGs) to solve this issue of structural data incorporation into

GNNs and to create a fully interpretable graph representation. PLIGs can incorporate 3D protein-ligand interactions directly into the ligand atom nodes of molecular graphs, encoding all intermolecular interactions made by each ligand atom (within a pre-defined threshold) in the 3D protein-ligand complex. This enables graphs to retain the shape of the ligand, by only altering node features. In addition, to explore potential performance difference between different GNN architectures, I compared a large variety of modern GNN architectures that are currently popular: (i) Graph Convolutional Neural networks (GCN, [Kipf and Welling, 2017]); (ii) Graph Attention Networks (GATNet, [Veličković et al., 2018]); (iii) Graph Isomorphism Networks (GIN, [Xu et al., 2019]); (iv) a combined GAT-GCN network [Nguyen et al., 2020]; (v) Graph SAGE [Hamilton et al., 2017]; and (vi) Simple Graph Convolutional Networks (SGC, [Wu et al., 2019]). When tested on the CASF-2016 benchmark and featurised using PLIGS, most GNN architectures perform similarly, with GATNet + PLIG performing best, outperforming other well known scoring functions tested on CASF-2016, such as PLEC-based random forests [Wójcikowski et al., 2018], K_{DEEP} [Jiménez et al., 2018], graph-based SIGN [Li et al., 2021] and performs within range of the currently best performing GNN model graphDelta [Karlov et al., 2020], however no standard deviation, confidence interval, model error or robustness was reported by graphDelta, making a direct comparison challenging.

4.3 Materials & Methods

4.3.1 Training and Test Sets

I investigated the ability of PLIGs with different graph-based neural networks (GNN) to predict the protein-ligand binding affinity. For training and testing, I used the PDBbind database [Liu et al., 2015, 2017], a curated set of 3D-structures of protein-ligand complexes and their corresponding experimentally determined binding affinity,

obtained from the Protein Data Bank (PDB, [Berman et al., 2000]). In order to utilize the most up-to-date data, I used the PDBBind 2020 General Set and supplemented it with data from the PDBBind 2016 Refined Set for a total of 19451 protein-ligand complexes. Additionally, for evaluation against docked poses rather than crystal poses, I subjected the combined dataset to the pre-processing and docking procedure described in the Methods Section 4.3.2, resulting in 14981 valid, docked protein-ligand complexes with the corresponding crystal poses. The dataset was randomly split into training (14254 data points) and validation (455 data points) sets. I used the PDBbind 2016 "Core Set", also referred to as the "CASF-2016 set", as our test set since it was used as the "scoring power" benchmark in the CASF exercise in 2016 [Su et al., 2018]. Our test dataset excluded 13 protein-ligand complexes unable to go through the docking procedure (for a breakdown of the processing see Methods Section 4.3.2), resulting in 272 valid data points (less than the full 285 normally present in CASF-2016). The CASF-2016 test set has been widely used in the community as a test set for evaluating scoring function performance [Boyles et al., 2019; Sánchez-Cruz et al., 2020; Li et al., 2021; Karlov et al., 2020], and notwithstanding the small exclusion of 13 data points, it serves as a good comparison of our models to previously published scoring function benchmarks.

Each model was trained to predict the inhibition constant K_i, the dissociation constant K_d, or the half-maximal inhibitory concentration IC_{50}, depending on which was provided by PDBBind for each complex. For the purpose of this study, these values were considered as interchangeable and are henceforth referred to as "the binding constant, K". This practice of combining two or all three different affinity values (IC_{50}, K_i and K_d) has been previously used for similar studies [Boyles et al., 2019; Sánchez-Cruz et al., 2020]. While the conversion between K_d or K_i and IC_{50} is possible using the Cheng-Prusoff equation [Yung-Chi and Prusoff, 1973], it requires

the substrate concentration which is not reported in PDBbind and often not known for individual complexes in the dataset. For training and performance evaluation, the negative base-10 logarithm of K, commonly denoted as "pK" was used:

$$\mathrm{p}K = -\log_{10} K \tag{4.1}$$

For every model, I evaluated its performance by computing the Pearson correlation coefficient (ρ) as well as the root-mean-square error (RMSE in pK units) between the predicted pK value and the corresponding, experimentally determined pK value for every complex in the test set. Cross validation was performed using a 5-fold split of the training set excluding the validation and test sets. Details about the cross validation can be found in Appendix C, Section C.3. Performance on the CASF-2016 benchmark was evaluated for every model by training and testing 10 times using different random seeds on the same dataset split. The ρ and RMSE were calculated and reported as the average over 10 runs with the corresponding standard deviation. Model runs were also combined into ensemble models, by averaging each protein-ligand prediction between each of the 10 models in the test set and calculating the ρ or RMSE between the average prediction and the true value.

4.3.2 Protein-Ligand Docking Methods

First, the ligand files provided by PDBbind were processed using the cheminformatics toolkit RDKit [Landrum et al., 2006] (accessed October 2021, v2021.03.5). To avoid the starting conformation of the ligand biasing the docking process, new conformers were generated using the ETKDG method [Riniker and Landrum, 2015] followed by energy minimization with the Merck Molecular Force Field (MMFF, [Halgren, 1999]) as implemented in RDKit. Out of the 19451 original ligand files, some could not be processed by RDKit, and the remaining ligands were subjected to additional filtering criteria (removal of metal containing ligands, removal of molecules with more than

20 rotatable bonds and a molecular weight of more than 1000 Da) resulting in 15317 compounds. This step was performed to filter out non-drug-like ligands. The resulting 15317 ligands were processed into PDBQT-formatted files using Open Babel v3.1.0 [O'Boyle et al., 2011] and the corresponding protein PDB file processed into PDBQT-formatted files using the *prepare_receptor4.py* function as implemented in MGLTools v1.5.7 [Morris et al., 2009], resulting in one PDBQT file which could not be generated by the software (PDB code 4BPS). The ligand and protein PDBQT files were then docked using Smina [Koes et al., 2013], a user friendly fork of AutoDock Vina [Trott and Olson, 2010] using default parameters except for the following: *exhaustiveness* = 20; *autobox_add* = 8; and *num_modes* = 20. The grid box for each docking run was determined using Smina's "autobox ligand" feature by passing the original crystal pose of the ligand into Smina and calculating the grid box from its location. For each ligand, 20 diverse poses were generated and the best scoring pose was used for featurization in our models. After docking, the resulting PDBQT ligand files (which lack bond order information) were parsed by RDKit to assign the correct bond orders to the docked pose using the original compound SMILES string. This resulted in the exclusion of 336 ligands as they failed to be parsed or processed using RDKit, resulting in a final set of 14981 valid docked protein-ligand complexes. In order to assess the quality of the docked poses, I computed the root-mean-square deviation (RMSD) between the coordinates of each atom in the docked pose and its corresponding atom in the original crystal pose using the symmetry-aware RMSD method implemented in the Open Drug Discovery Toolkit (ODDT, [Wójcikowski et al., 2015b]).

4.3.3 Architecture of Machine-Learning Models

I explored a wide variety of graph-based neural networks (GNNs) as well as a multi-layer perceptron neural network (MLPNet) in a two-branch setup (Figure 4.1), with

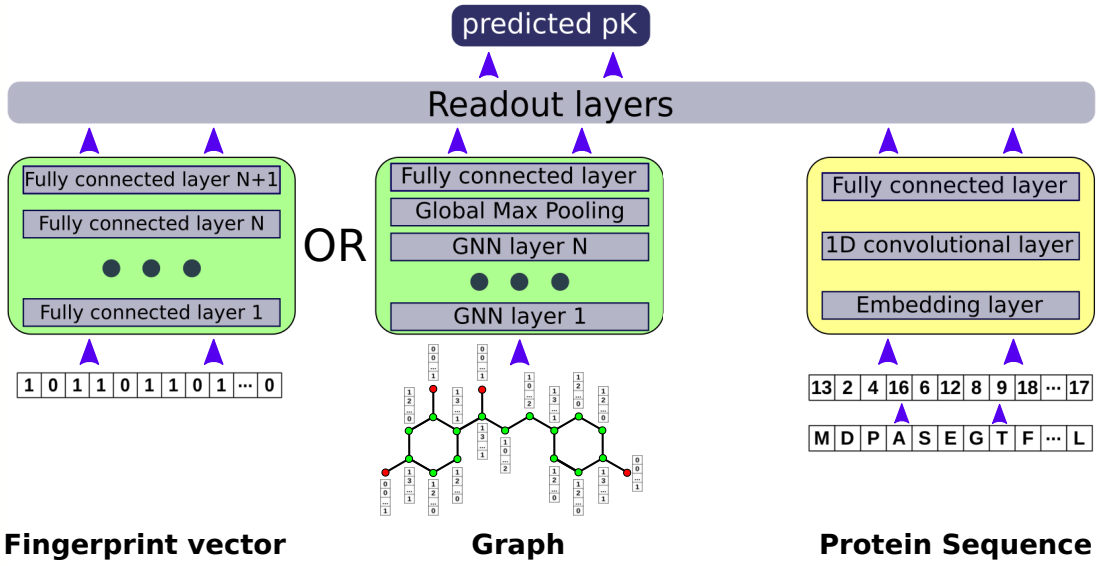

Figure 4.1: General architecture of the models with the ligand branch (green) and the sequence branch (yellow). The ligand is either embedded as a fingerprint vector in the MLPNet branch (leftmost branch), or a molecular graph is created for the GNN branch (middle branch). On the right branch, the protein embedding is shown, where a 1D-convolutional layer is fed with the embedded protein amino acid sequence. The outputs from the protein and ligand branches are concatenated and fed into the three fully-connected readout layers and a single pK prediction is made.

either the GNN or MLPNet embedding the ligand with or without the protein structure in the first branch and a second branch encoding the protein sequence in a 1D convolutional neural network. Both branches combine into the readout layers, consisting of three fully-connected layers. This two-branch architecture is adapted from the GraphDTA architecture described by Nguyen et al. [2020]. All models were implemented using PyTorch v1.9.0 [Paszke et al., 2019] and PyTorch Geometric v2.0.0 [Fey and Lenssen, 2019]. The implementation of the graph convolutional neural network (GCN, [Kipf and Welling, 2017]), graph attention network (GATNet, [Veličković et al., 2018]), graph isomorphism network (GIN, [Xu et al., 2019]), and the combined GAT-GCN were directly adopted from [Nguyen et al., 2020] with the only changes being the selection of optimal hyperparameters after hyperparameter tuning (see Appendix C, Section C.1, implemented in collaboration with Dominik Klein). In

addition, GraphSAGE [Hamilton et al., 2017] as well as Simple Graph Convolutional Networks (SGC, [Wu et al., 2019]) were incorporated into the GraphDTA architecture using the PyTorch Geometric SAGEConv and SGCConv implementation respectively (in collaboration with Dominik Klein).

The differences in GNN design between the different GNNs used in this work are often subtle and abstract and therefore well suited to answer the question: "how much do design changes in attention, convolution, aggregation, or layer activation approaches affect the final performance of GNN models". A brief, high-level summary of the differences between the GNN implementations is shown below.

GNNs in general, and GCNs in particular were already discussed in detail in Chapter 1 Section 1.3.1. Briefly, GCNs [Kipf and Welling, 2017] (just like classical convolutional neural networks) are comprised of layers of learned spectral filters which are typically activated by a nonlinear activation function. At each layer, the filter is applied to a node and its neighboring nodes that are one hop away to obtain a new learned feature representation. Therefore, after k layers, a node now contains feature information from all nodes that were k-hops away in the original graph, creating higher order feature representations. SGCs [Wu et al., 2019] alter the GCN concept by collapsing all GCN layers into a single layer and by removing the nonlinear activation function, only keeping the softmax function.

GATNets [Veličković et al., 2018] are also based on GCNs and alter the architecture by changing the process of how feature information from neighboring nodes is aggregated on a given node. Instead of calculating the normalized sum of node features during a graph convolution, GATNets introduce an attention mechanism between neighboring nodes, calculating an un-normalized attention score between neighboring nodes which implicitly assigns higher weights to neighboring nodes of higher importance during training. The combined GAT-GCN model incorporates

both, GATNet and normal GCN layers into the model. The GraphSAGE [Hamilton et al., 2017] architecture is very similar to the GATNet approach since it changes the way feature aggregation works between neighboring nodes. While GCNs produce the normalized sum over the node features of neighboring nodes, GraphSAGE simply averages over neighboring node features instead. This also enables GraphSAGE to be used inductively instead of transductively, for example on evolving graphs with unseen nodes. However for the purpose of this work, all models are employed transductively. Finally GINs [Xu et al., 2019] are based on the Weisfeiler-Leman test (WL test), which is able to tell if two graphs are anti-isomorphic (graphs with non identical structure). GIN replace the aggregator that is used to update the node features with an aggregator inspired by the WL test to maximize the representational (or discriminative) power of the GIN.

The MLPNet model follows a simple feed-forward neural network architecture implemented using PyTorch with the dimension (width) of each layer and the number of fully-connected layers (depth) determined by hyperparameter optimization as described in the Appendix C, Section C.1. The embedding and setup of the protein branch was adopted from GraphDTA [Nguyen et al., 2020], in collaboration with Dominik Klein, with only a minor change: reducing the number of convolutional layers from three layers to only a single layer, simplifying the model. The convolutional kernel size and the number of filters are hyperparameters which were tuned individually for every GNN/MLPNet and protein branch combination (see Appendix C, Section C.1). The implementation and code for all models can be found at `http://github.com/MarcMoesser/Protein-Ligand-Interaction-Graphs`.

4.3.4 Ligand-Based Graphs

Ligand-based graphs were generated based on the bonds and atoms in the small molecule. Each atom (node) is represented as a one-hot encoded 40-dimensional feature vector, and each covalent bond (regardless of the bond type) as an edge in the graph. The atom feature set used by GraphDTA [Nguyen et al., 2020] consisted of the following one-hot encoded features: atomic symbol; number of adjacent heavy atoms; number of adjacent hydrogens; implicit valence; and whether the atom is in an aromatic ring. In order to accurately reflect the ligand and add relevant additional ligand information, this was expanded by adding additional features: a boolean variable describing whether the atom is in a ring; the one-hot encoded formal charge; and the one-hot encoded hybridization type; and by changing the encoding of the implicit valence to the explicit valence. Since PDBBind [Liu et al., 2015, 2017] carefully preprocesses ligands and assigns physiologically relevant atom charges, the inclusion of features such as the explicit valence and the formal charge gives a more complete description of the molecule. All features were calculated using RDKit [Landrum et al., 2006] (v2021.03.5).

4.3.5 Small Molecule Fingerprints

The molecular fingerprints used for the MLPNet models were calculated using RDKit (v2021.03.5) [Landrum et al., 2006]. I investigated the effect of using Morgan fingerprints with the RDKit implementation of Extended-Connectivity Fingerprints and Functional-Class Fingerprints (ECFP and FCFP respectively, [Rogers and Hahn, 2010]) with a radius of 2, and either 512 or 1024-bit vectors. They are referred to as "ECFP512" or "ECFP1024" respectively.

4.3.6 Protein-Ligand Interaction Graphs (PLIGs)

Protein-Ligand Interaction Graphs (or "PLIGs") were inspired by the work of [Sánchez-Cruz et al., 2020] who introduced Extended Connectivity Interaction Features (ECIF) as high performing features for random forests and gradient boosted tree networks for binding affinity prediction.

PLIGs combine ligand-based graphs with 3D protein-ligand interaction features (Figure 4.2). Initially, a ligand-based graph is generated (see Section (4.3.4) with only five RDKit-derived atom node features: number of adjacent heavy atoms; number of adjacent hydrogens; explicit valence; aromaticity; and ring membership. Then, all possible, unique protein atom types with the following criteria: atom symbol; explicit valence; number of attached heavy atoms; number of attached hydrogens; aromaticity; and ring membership, are identified. This resulted in 22 unique protein atom types based on the 20 proteinogenic amino acids, excluding selenocysteine (for full list see Appendix C, Section C.2). To identify the interaction features, the existence of every protein atom in the proximity of a given ligand atom within a defined distance threshold in the 3D-protein-ligand complex is recorded. A 22-dimensional integer vector is created, with each element in the vector corresponding to the count of the protein atoms of that unique protein atom type in the vicinity of the ligand atom. Each position in the ligand atom node feature vector is therefore interpretable as it (i) encodes for a specific, identifyable protein atom type (*e.g.* an aromatic carbon in phenylalanine), and (ii) since it encodes for the number of interactions made with this type for the given ligand atom. Previous work on ECIF [Sánchez-Cruz et al., 2020] determined 6 Å to be the optimal proximity threshold, however for this work, I investigated the performance of different thresholds (between 4-8 Å in 1 Å-intervals). Finally, the 5-dimensional ligand-derived atom feature vector is concatenated with the 22-dimensional ligand-protein interaction features, to generate

the final 27-dimensional node feature vector (Figure 4.2). This procedure is repeated for every atom in the ligand to generate the final PLIG.

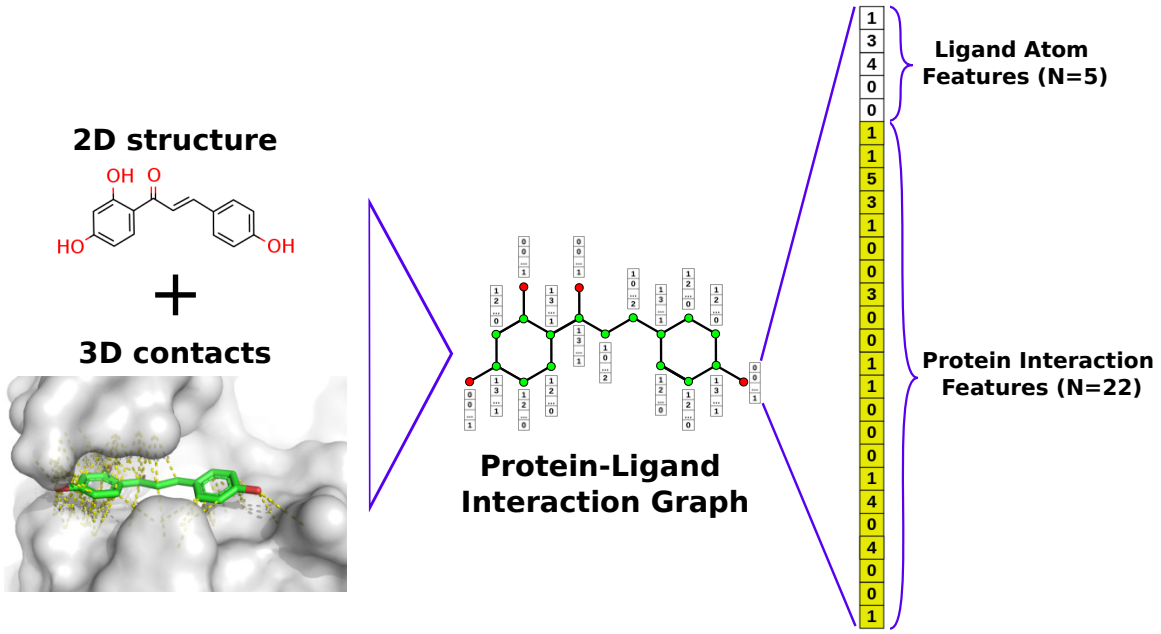

Figure 4.2: Stylised representation of the construction of a Protein-Ligand Interaction Graph. On the left, the chemical structure of the ligand of 6ajv and 3D-structure of the binding site are shown, with every intermolecular interaction in the radius of 4 Å around each ligand atom marked in yellow. PLIGs are created by assigning an integer feature vector to each node consisting of the RDKit-derived atom descriptors (number of adjacent heavy atoms; number of adjacent hydrogens; explicit valence; aromaticity; and ring membership) as well as the number of interactions made with each protein atom type (one for each of the 22 unique atom types) that are within a pre-defined distance threshold to create the final 27-dimensional feature vector for every atom node in the PLIG.

One major disadvantage of the ECIF fingerprints is that the dimensionality of the vector depends on the dataset. Since the ECIF fingerprint is comprised of all possible combinations of protein atom types and ligand atom types, if a novel ligand is to be scored using an ECIF model, it cannot contain a new ligand atom type. An ECIF-featurized model therefore needs to be retrained from scratch every time the model is asked to evaluate a ligand with a new "ECIF atom type" that was not represented in the original dataset. Given the diversity of chemical space and the uniformity of known drug-like molecules, this is especially problematic when dealing

with novel molecule design, especially for early stage drug discovery where a diverse set of molecules might be screened or a set of novel scaffolds needs to be explored for candidate optimization. PLIGs overcome this limitation by defining atom types on the protein side, which limits the possible options to the 20 proteinogenic amino acids (inclusion of more amino acids is also possible if needed), thus eliminating the problems of unknown atom types. Therefore, any new protein-ligand complex can be scored with a pre-trained PLIG model, regardless of the ligand structure. This is especially useful for building prospective models on targets where little structural information is known and docking is needed for generating ligand poses, and a diverse set of new ligands needs to be screened. Additionally, the PLIG architecture is able to represent 3D protein-ligand complexes in a ligand-based graph, without needing additional edges and nodes, only changing the node feature vector.

4.4 Results and Discussion

4.4.1 Quality of Docked Poses

The quality of all 14981 docked poses was estimated by calculating the symmetry-aware RMSD using ODDt [Wójcikowski et al., 2015b] between all heavy atoms of the ligand's original crystal pose and the highest scoring docked pose for each protein-ligand complex. A docked pose is considered to be of high quality if the calculated crystallographic RMSD is 2 Å or less. Overall, 39 % of the docked poses were high quality. However, when splitting the dataset by the quality of the underlying crystal structures into the Refined Set and General Set, a subtle difference in pose quality was observed (42 % of docked poses from the Refined Set and 38 % of poses from the General Set were high quality, Figure 4.3). This might be due to the lower resolution of structures in the General Set where crystal structures have resolutions that worse than the 2 Å cutoff used above for pose quality estimation. In order to simulate a

more realistic docking campaign where lower quality poses can be expected, I used all available docked poses as input for the scoring functions, regardless of pose accuracy.

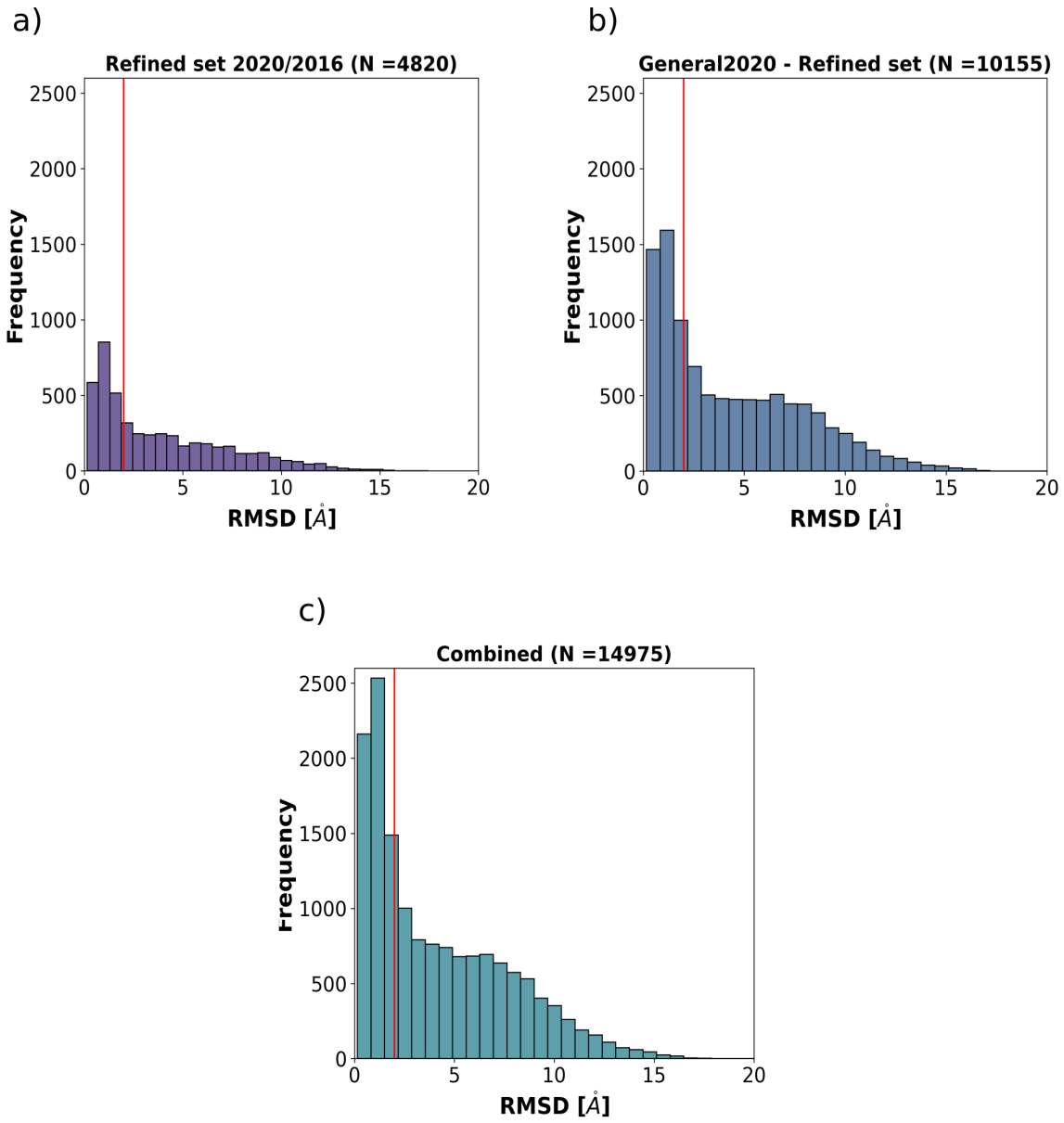

Figure 4.3: Overview of the docking pose quality for the PDBbind dataset. The red line marks the evaluation cutoff of 2 Å. All RMSD values were calculated as symmetry aware using OODT [Wójcikowski et al., 2015b]. a) Distribution of RMSD values (reported in Å) for the subset of the dataset that is part of the Refined Set 2020 and Refined Set 2016. b) Distribution of RMSD values (reported in Å) for the complexes in the General Set 2020 without complexes that are part of Refined Set 2020 or Refined Set 2016. c) Distribution of RMSD values (reported in Å) for the entire dataset (Refined Set 2016 and General Set 2020 including Refined Set 2020).

4.4.2 Model Combinations

For this study, six different GNN and one MLPNet architecture were employed. The GNN models were either featurized using ligand-based graphs or PLIGs. The MLPNet model was featurized with structure-based fingerprints (ECIF [Sánchez-Cruz et al., 2020]) or ligand-based fingerprints (ECFP512, ECFP1024, FCFP512, or FCFP1024). All resulting models were trained and tested either in a two-branch architecture with an additional protein sequence encoding branch (see Figure 4.1), or by themselves as a standalone model. This resulted in the 34 models listed in Table 4.1. Each of the models were trained and tested separately on crystallographic and docked poses.

In addition to the models listed in Table 4.1, six different multi-model ensembles were created as described in Section 4.3.1. The list of created multi-model ensembles is given in Table 4.2. The architecture of each model (*e.g.* number of layers and other parameters) were determined during hyperparameter tuning an can be found in Appendix C, Section C.1.

4.4.3 Model Stability and Ensemble Model Performance

Since the model predictions are stochastic between different training and test runs on the same dataset split, model performance against the withheld test set (CASF-2016) for all trained models was evaluated: (1) using the average and standard deviation (SD) of the Pearson correlation coefficient (ρ) as well as the root-mean-square error (RMSE in pK units) over 10 runs; and (2) as the ensemble model over all 10 models by averaging the individual predictions for each protein-ligand complex in the test set from all 10 models. Model stability between the 10 training runs was high for all models, with a maximum ρ standard deviation of 0.023 for the model with highest variability (SGCNet-LB) and a maximum RMSE standard deviation of 0.08 pK (SGCNet+PLIG model and the SGCNet+PLIG + PB model trained and tested on

Model Architecture	Features	Sequence
GATNet + PLIG + PB	3D-based (PLIG)	Yes
GATNet + PLIG	3D-based (PLIG)	No
GATNet-LB + PB	ligand-based	Yes
GATNet-LB	ligand-based	No
GCNNet + PLIG + PB	3D-based (PLIG)	Yes
GCNNet + PLIG	3D-based (PLIG)	No
GCNNet-LB + PB	ligand-based	Yes
GCNNet-LB	ligand-based	No
GIN + PLIG + PB	3D-based (PLIG)	Yes
GIN + PLIG	3D-based (PLIG)	No
GIN-LB + PB	ligand-based	Yes
GIN-LB	ligand-based	No
GAT/GCN + PLIG + PB	3D-based (PLIG)	Yes
GAT/GCN + PLIG	3D-based (PLIG)	No
GAT/GCN-LB + PB	ligand-based	Yes
GAT/GCN-LB	ligand-based	No
SGCNet + PLIG + PB	3D-based (PLIG)	Yes
SGCNet + PLIG	3D-based (PLIG)	No
SGCNet-LB + PB	ligand-based	Yes
SGCNet-LB	ligand-based	No
SageNet + PLIG + PB	3D-based (PLIG)	Yes
SageNet + PLIG	3D-based (PLIG)	No
SageNet-LB + PB	ligand-based	Yes
SageNet-LB	ligand-based	No
MLPNet + ECIF + PB	3D-based (ECIF)	Yes
MLPNet + ECIF	3D-based (ECIF)	No
MLPNet + ECFP512 + PB	ligand-based	Yes
MLPNet + ECFP512	ligand-based	No
MLPNet + ECFP1024 + PB	ligand-based	Yes
MLPNet + ECFP1024	ligand-based	No
MLPNet + FCFP512 + PB	ligand-based	Yes
MLPNet + FCFP512	ligand-based	No
MLPNet + FCFP1024 + PB	ligand-based	Yes
MLPNet + FCFP1024	ligand-based	No

Table 4.1: All model architecture and feature combinations (34 total). All PLIGS were generated using a proximity threshold of 6 Å. The presence or absence of the two-branch architecture featuring the additional protein branch (PB) is denoted in the "Sequence" column and the architecture name. Ligand-based GNN models have the "LB" label.

docked poses). For a detailed overview of model stability and standard deviations between different runs, see Appendix C, Section C.4. Performance improved for all

Ensemble Model Architectures	Features	Protein Branch?
All PLIG models	3D	No
GATNet (PLIG)+MLPNet (ECIF)	3D	No
GATNet (PLIG)+MLPNet (ECFP512)	3D & LB	No
MLPNet (ECIF)+MLPNet (ECFP512)	3D & LB	No

Table 4.2: All multi-model ensembles. All models were trained without the sequence embedding. There are four multi-model ensembles total. All PLIGs were generated using a proximity threshold of 6 Å.

models when using the ensemble predictions in comparison to the averaged ρ score (Appendix C, Figure C.11 & C.12; Main text Figure 4.5) and therefore all further results are reported for the ensemble models.

4.4.4 Model Performance on Crystal Poses

Model performance when trained and tested on crystal poses is shown in Figure 4.5, with the GATNet+PLIG model (no sequence; ρ=0.84, RMSE=1.22 pK) and the MLPNet ECIF model (with sequence; ρ=0.84, RMSE=1.19 pK) performing best overall, although differences in performance between GATNet+PLIG (ρ=0.84, RMSE=1.22 pK) and GAT-GCN PLIG (ρ=0.82, RMSE=1.24 pK) were small, and all other graph architectures only marginally worse (Figure 4.5). ECIF fingerprints were previously only tested using random forest or gradient boosted tree models [Sánchez-Cruz et al., 2020], showing that ECIF fingerprints can be used efficiently in a deep learning framework as well as in classical machine learning models. In addition, although performance is strong across all GNN architectures, the GATNet PLIG model outperforms all other graphs (scatter plot of GATNet+PLIG trained and tested on crystal poses is shown in Figure 4.4 a). Overall, I show that PLIGs are a high-performing graph representation regardless of GNN architecture, and performance difference between ligand-based and PLIG-based GNNs is larger than differences between GNN architectures, highlighting that good molecular representation design

is more important than subtle differences in GNN architecture design. Furthermore, while performance of PLIG-based GNN models slightly lacks behind the reported performance of graphDelta (ρ=0.87 on CASF-2016, [Karlov et al., 2020]), no standard deviation, confidence interval, model error or robustness was reported by graphDelta, making a direct comparison challenging, especially since the standard deviation of the Pearson correlation coefficient between individual runs of GATNet+PLIG for example is 0.02, indicating that graphDelta might be almost in range. Nonetheless, PLIG-based models are able to outperform some of the best and most widely used protein-ligand binding affinity prediction methods such as random forest-based PLEC (ρ=0.817, [Wójcikowski et al., 2018]), OnionNet (ρ=0.816, [Zheng et al., 2019]) and K_{DEEP} (ρ=0.82, [Jiménez et al., 2018]) as well as other recent GNN-based scoring functions such as SIGN (ρ=0.797, [Li et al., 2021]) and PIGNet (ρ=0.761, [Moon et al., 2022]).

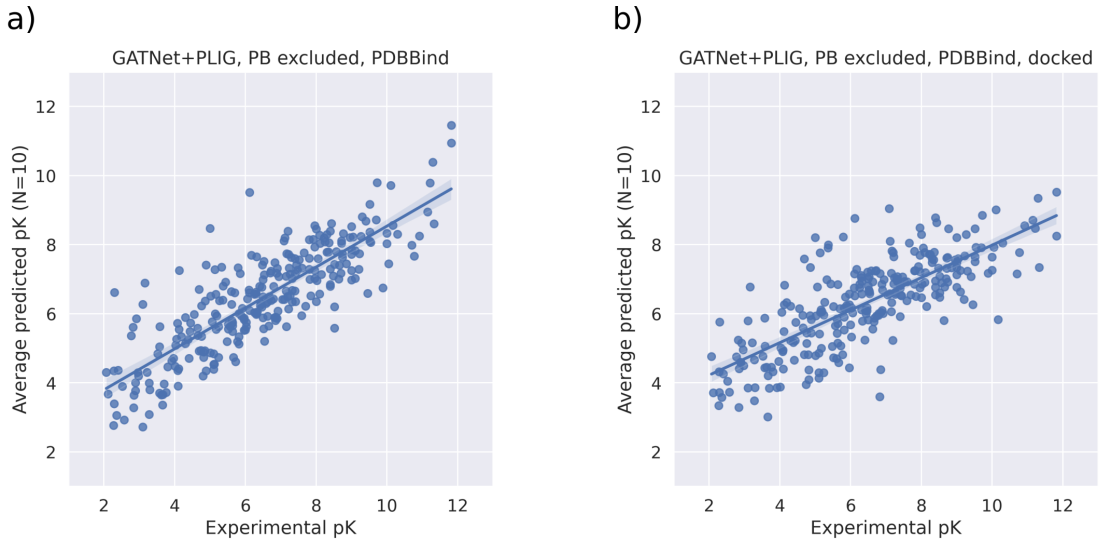

Figure 4.4: Scatter plots of the average prediction versus the experimentally determined pK value for each protein-ligand complex between the 10 model runs for: a) GATNet+PLIG, no protein sequence embedding, trained and tested on crystal structures; b) PLIG, no protein sequence embedding, trained and tested on docked structures.

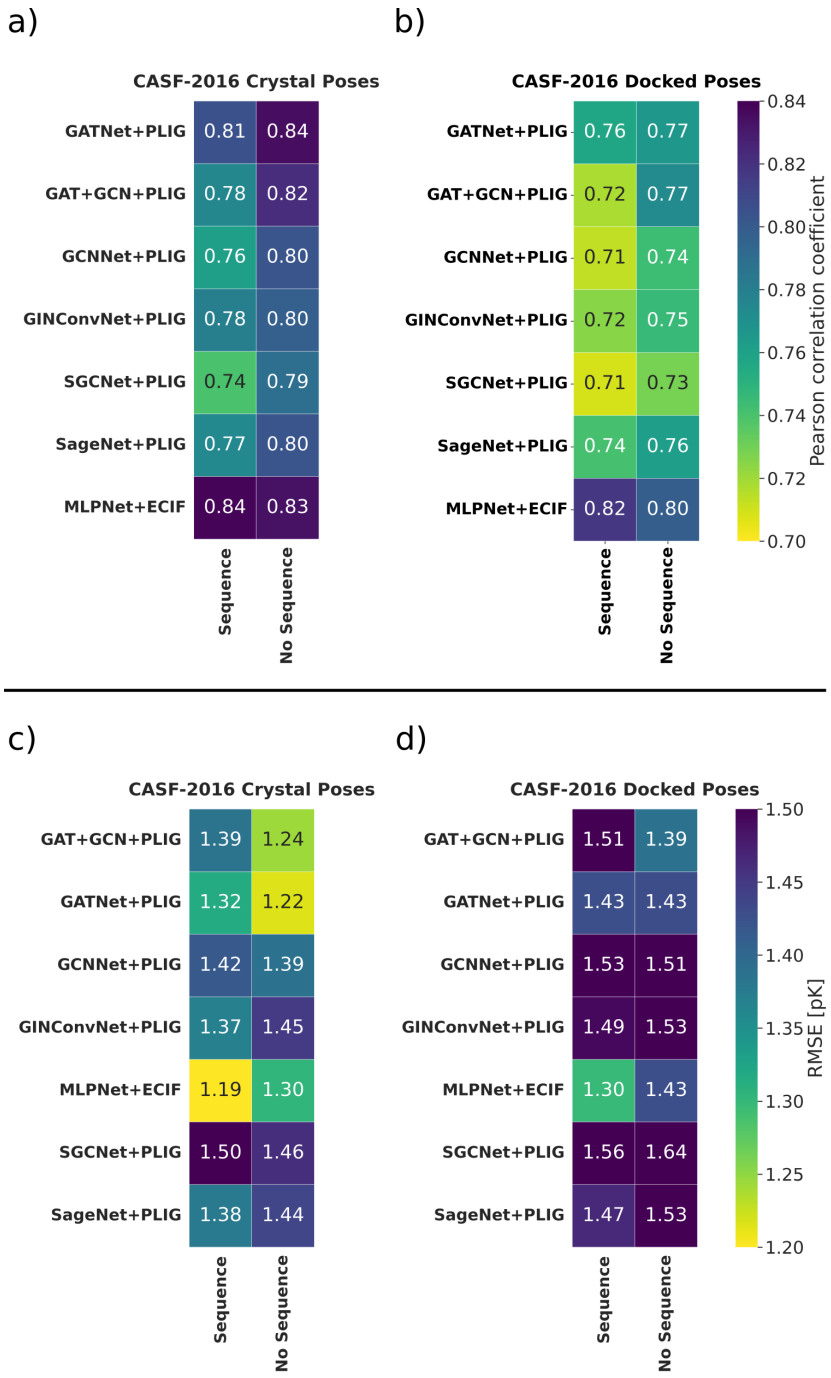

Figure 4.5: The reported Pearson correlation coefficient (ρ) and root-mean-square error (RMSE) of the structure-based ensemble models. All PLIGs are generated using a proximity threshold of 6 Å. For both, ρ and RMSE results, *"Sequence"* and *"No Sequence"* denote the presence or absence of the protein sequence-encoding branch in the model architecture, respectively. (a and c) Performance of the structure-based GNN and MLPNet models when trained and tested on crystal structures. (b and d) Performance of the structure-based GNN and MLPNet models when trained and tested on docked poses. The best performing model is the GATNet+PLIG (ρ=0.84, RMSE=1.24 pK) model.

4.4.5 Protein Sequence Embedding

For all PLIG models, the inclusion of the protein sequence branch in the model decreases performance. Since PLIGs already implicitly encode the structure of the protein (at least the part that interacts with the ligand), inclusion of 1D-protein sequence information might be introducing confounding information or noise, hindering PLIG model performance. In addition, all protein-branch containing models are more unstable during cross validation, reaching maximum performance quickly, after a small number of epochs, therefore raising concerns of overfitting (Appendix C, Section C.3). As a result, I recommend using the single-branch GNN or MLPNet implementations of any of the models implemented in this work without the protein sequence branch in future studies and will be focused on those models in the discussion of further results. In addition, caution is advised in general when dealing with protein-ligand binding affinity scoring functions that utilize sequence information directly in 1D convolutional networks as described in this work because of potential overfitting.

4.4.6 Model Performance on Docked Poses

Overall, the performance of all models decreased when trained and tested on docked poses in comparison to training and testing on crystal poses (Figure 4.5). As with models trained and tested on crystal poses, the performance difference between different GNN architectures in combination with PLIG was low (ρ=0.73-077 between all GNN+PLIG models) when trained and tested on docked poses. As one of the top performing graph-based PLIG models, the scatter plot for GATNet+PLIG when trained and tested on docked poses is shown in Figure 4.4 (b).

The MLPNet+ECIF model performed better than any graph-based methods when comparing the Pearson correlation coefficient (MLPNet+ECIF ρ=0.80), but drops in

performance in comparison with the best GNN-based method when comparing RMSE (GATNet+PLIG: RMSE=1.39 pK; MLPNet+ECIF: RMSE=1.43 pK, see Figure 4.5 d). However, all differences are fairly minimal in total. As mentioned in the previous section, models that included the protein sequence embedding were less robust during cross validation (Appendix C, Section C.3) and models without the sequence embedding are preferred and more stable. The models were also trained on crystal structures and tested on docked poses (Appendix C, Figure C.13) however no noteworthy difference to models trained and tested on docked poses was observed. Overall, the drop in performance from training on crystal to docked poses for structure-based methods is to be expected, as docked poses sometimes diverge from the original crystal structures, and the identified interactions might not reflect the actual interactions in the crystal.

4.4.7 Model Performance with Ligand-Based Features

In contrast to the structure-based methods where incorporation of the sequence embedding did not improve performance for PLIG GNN, ligand-based GNN models are either improved (GAT-GCN, GCNNet, GIN, SGC) or are not negatively affected (GATNet, SageNet), as can be seen in Figure 4.6. When using purely ligand-derived graphs, no notable difference in performance between the different graph architectures was observed. In addition, as described for the PLIG-based GNNs, inclusion of the sequence branch is not recommended for robust model design. Overall, purely ligand-based models do not perform as well as the structure-based models trained and tested on crystal structures, but perform comparably to the structure-based models trained and tested on docked poses (Figure 4.5 & 4.6), showing that any additional structural noise introduced to the dataset by docking leads to 3D-based models performing as well as ligand-based models that have no 3D data at all. This same trend was ob-

served by Boyles et al. [2021] for random forest-based models. This also highlights that the ligand-based models benefit from the inherent bias in the benchmark sets, and are memorizing ligand information rather than interactions between the ligand and the protein. Their ability to extrapolate to unseen ligands and generalize will therefore be compromised.

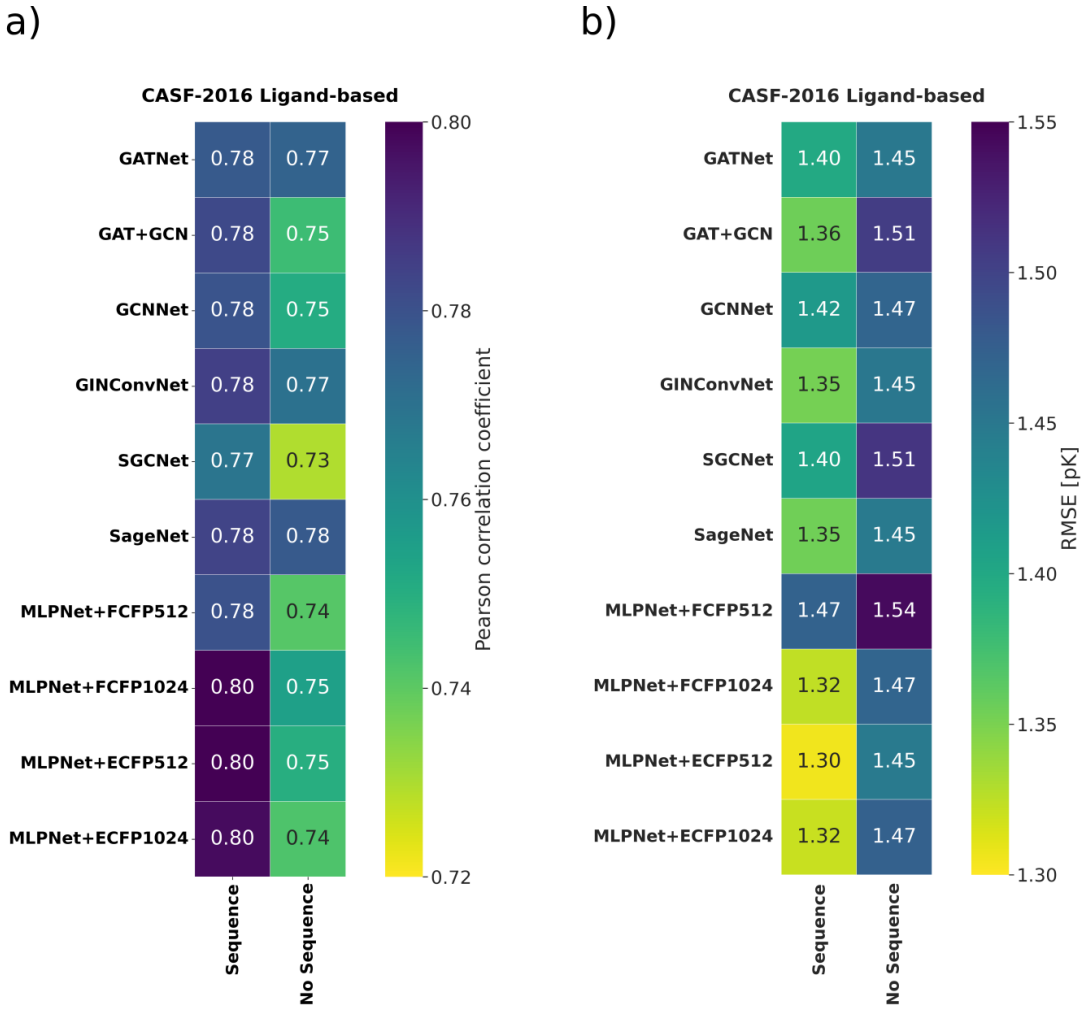

Figure 4.6: (a) The reported Pearson correlation coefficient (ρ); and (b) root-mean-square error (RMSE) of the ligand-based ensemble models. For both, ρ and RMSE results, *"Sequence"* and *"No Sequence"* denote the presence or absence of the protein sequence-encoding branch in the model architecture, respectively. Outside of some outliers that performed especially poorly (ligand-based SGCNet and ligand-based MLPNet+FCFP512), no strong difference in performance is observed between all ligand-based models, however, models that include the protein sequence branch performed better than without, but made the models prone to overfitting (see Section 4.4.5.

4.4.8 Model Performance of Multi-Model Ensembles

Since intra-model ensembles performed better than individual models, I tested the impact of multi-model ensembles. Overall, no notable improvement over the GATNet+PLIG model was observed for any model combination (Figure 4.7) when trained and tested on crystal poses. However, combining GATNet+PLIG models with ligand-based ECFP512 fingerprint features yielded a synergistic improvement in comparison to the individual models trained and tested on docked poses (ensemble: ρ=0.81, RMSE=1.35 pK versus GATNet+PLIG: ρ=0.77, RMSE=1.43; and ECFP512: ρ=0.75, RMSE=1.45 pK). However, ensembling ligand-based models with MLPNet+ECIF models did not increase performance when trained and tested on docked poses.

The addition of ligand-based features has been previously shown to recover most of the lost performance when replacing crystal with docked poses (Random Forest models by Boyles et al. [2021]). I observed the same effect when using 3D structure-based graph neural networks and ligand-based feed forward neural networks (MLPNet) in an ensemble model framework, recovering lost performance between crystal structures and docked poses. Being able to apply model ensembles has the advantage over the work reported by Boyles et al. [2019] that models do not have to be retrained once the decision has been made to include additional ligand features to rescue performance. Rather, in the case of ensemble models, a single new ligand-based model can be created and combined post-prediction with the structure-based models to rescue performance. Overall, the GATNet+PLIG models perform best, outperforming all other models on crystal poses and when combined with ligand-based models in a multi-model ensemble regain top performance on docked poses.

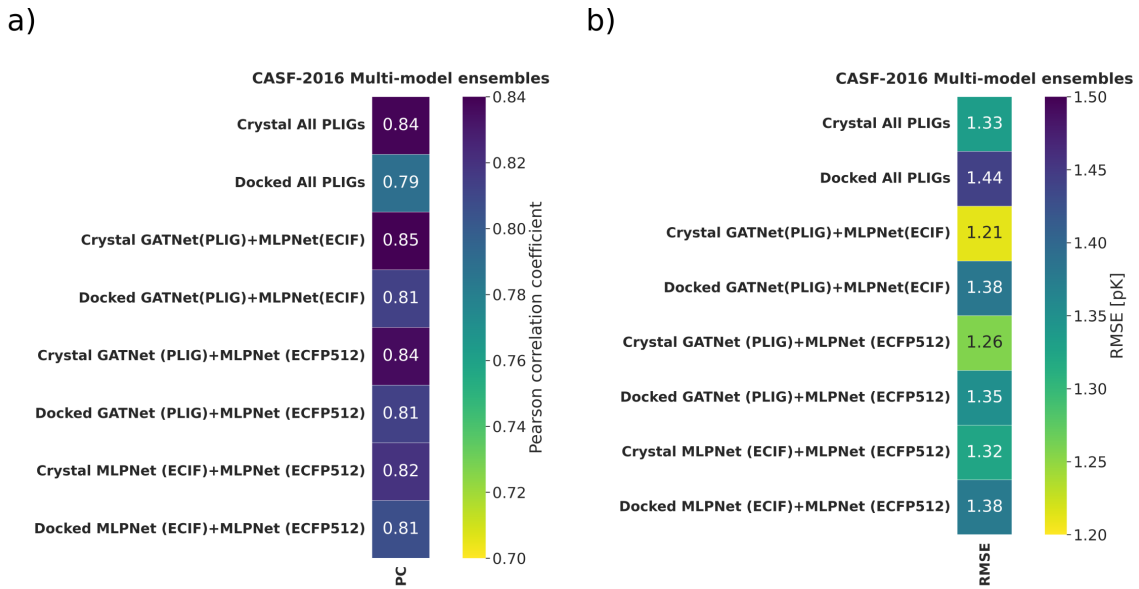

Figure 4.7: Performance of the multi-model ensembles as measured by ρ (a) and RMSE (b) for docked and crystal poses. All PLIGs are generated using a proximity threshold of 6 Å. The "All PLIGs" model is an ensemble of all PLIG GNN models. PLIG, ECIF and ECFP models were used without the sequence embedding for all ensembles. The best performing single model, the GATNet PLIG without protein sequence embedding (Figure 4.5 ρ=0.84, RMSE=1.24 pK), was not improved further through multi-model ensembles, but performance of the GATNet+PLIG model (ρ=0.77, RMSE = 1.39 pK, see Figure 4.5) was improved against docked poses by ensembling with the MLPNet ECFP512 model (ρ=0.81, RMSE=1.35 pK).

4.4.9 Influence of Proximity Threshold on Performance

I further investigated the effect of varying the protein-ligand interaction proximity thresholds during PLIG generation on the performance of the GATNet PLIG model. I generated PLIGs using thresholds of 4-8 Å in 1 Å steps. These particular thresholds were chosen since values smaller than 4 Å might exclude some intermolecular interactions and therefore give an incomplete picture of the surrounding atom environment. In addition, proximity values beyond 8 Å would include protein atoms too far away to form meaningful interactions. The GATNet model without the sequence embedding was chosen as the model architecture for this experiment and 5-fold cross validation as well as training and testing on the CASF-2016 benchmark was performed as described above (Methods Section 4.3.1).

Model Architecture	Dataset	Ensemble ρ	Ensemble RMSE [pK]
GATNet (PLIG) 4 Å	Crystal	0.81	1.32
GATNet (PLIG) 5 Å	**Crystal**	**0.84**	**1.21**
GATNet (PLIG) 6 Å	**Crystal**	**0.84**	**1.22**
GATNet (PLIG) 7 Å	Crystal	0.81	1.29
GATNet (PLIG) 8 Å	Crystal	0.80	1.33
GATNet (PLIG) 4 Å	Docked	0.76	1.46
GATNet (PLIG) 5 Å	Docked	0.76	1.44
GATNet (PLIG) 6 Å	Docked	0.77	1.43
GATNet (PLIG) 7 Å	Docked	0.77	1.39
GATNet (PLIG) 8 Å	Docked	0.76	1.43
All proximities ensemble	Crystal	0.84	1.22
All proximities ensemble	Docked	0.79	1.38

Table 4.3: Performance of the GATNet PLIG model on the CASF-2016 benchmark using different proximity thresholds. Reported Pearson correlation coefficient (ρ) and Root mean square error (RMSE) are calculated as the ensemble between 10 train and test model iterations. The best performance on crystal structures was achieved with PLIGs using the 5 and 6 Å thresholds (marked bold, ρ=0.84 / 0.84 and RMSE=1.21 / 1.22 for 5 / 6 Å threshold respectively) with no notable difference in performance between both. Multi-model ensembles between all proximity thresholds does not lead to increased performance.

For a discussion of the performance during cross validation as well as the average performance and standard deviation before ensembling see Appendix C, Section C.5. The ensemble performance between 10 train and test model iterations is shown in Table 4.3. The best performance when trained and tested on crystal structures was observed for PLIGs with thresholds of 5 Å and 6 Å (ρ=0.84 / 0.84 and RMSE=1.21 / 1.22 for the 5 / 6 Å thresholds, respectively) with no notable advantage of one over the other. Furthermore, performance dropped for all thresholds when training and testing on docked poses slightly, although varying proximity thresholds do not seem to strongly affect performance in this instance (Table 4.3). This observation is in line with our expectations since a majority of docked poses were found to have a RMSD between the crystal and docked pose of larger than 2 Å (Section 4.4.1). Changing the PLIG threshold within the docking error rate (more than 2 Å for many docked poses) should therefore not alter results drastically.

Finally, multi-model ensembles using all thresholds does not improve performance. Overall, I found the 6 Å threshold to perform best between docked and crystal poses and recommend usage of 6 Å PLIGs for further studies.

4.4.10 Model Generalizability

In order to assess the ability of the best performing GATNet PLIG model to generalize between different protein families, protein-ligand pairs in the training dataset were eliminated based on their sequence identity to proteins represented in the CASF-2016 benchmark using five sequence identity threshold levels ranging between 50 % and 100 % identity. The GATNet+PLIG model was trained on the reduced dataset and tested against the full CASF-2016 dataset a total of 10 times, and the ensemble model obtained as described above (a description of the model stability is given in the Appendix C, Section C.6). As expected, performance decreases as increasingly dissimilar proteins are eliminated with stricter identity thresholds (Figure 4.8), with the largest drop in performance observed between the original full dataset and the elimination of 100% identical proteins from the training set (meaning that only proteins in the training set with identical sequence to proteins in the test set are removed). This observation is in line with similar experiments such as reported by Boyles et al. [2019], where model performance decreased regardless of random forest model featurization when eliminating test set-similar protein structures from the training set. However, part of the decrease in performance could be due to the reduced dataset size when eliminating data points in the training set (Table 4.4). In addition, a decrease in performance with decreasing sequence similarity could also suggest that the model might be susceptible to ligand-specific bias, as similar ligands are expected to bind similar proteins. This has previously been observed for random forest-based models as reported by Boyles et al. [2019].

Nonetheless, as CASF-2016 is deliberately chosen to be representative of the proteins present in the PDBbind refined set, eliminating this bias through techniques such as the sequence identity elimination described herein, a fairer evaluation of generalizability would be achieved and should therefore be utilized in future evaluations of other protein-ligand affinity scoring functions.

Threshold	Training Set Size
Full training set	14254
100 % threshold	11917
95 % threshold	11033
90 % threshold	10969
70 % threshold	10855
50 % threshold	10408

Table 4.4: Size of the training set based on the sequence identity threshold.

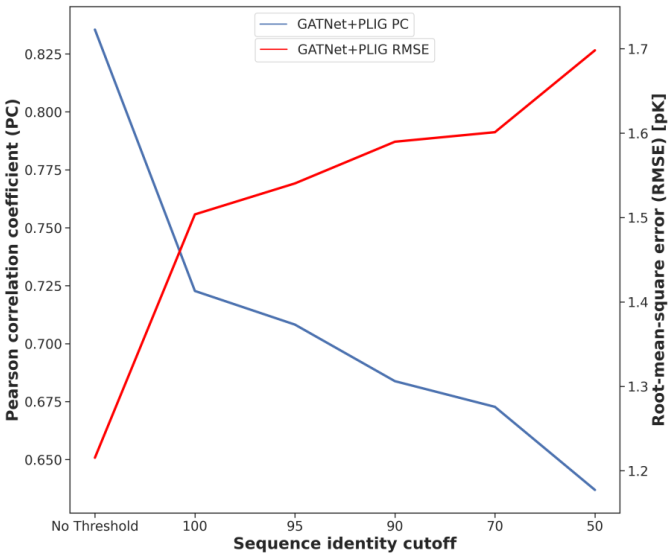

Figure 4.8: Pearson correlation coefficient (ρ) and Root mean square error in pK units (RMSE) of the predicted versus the experimental binding affinity for the GATNet PLIG model (no sequence, ensemble over 10 runs) when trained and tested on crystal poses. Protein-ligand complexes in the training set with a sequence identity at or above the cut-off value to proteins in the CASF-2016 test set were excluded, resulting in a smaller dataset at every step.

4.5 Discussion

The recent interest in GNN-based scoring functions for protein-ligand binding affinity prediction has led to something of an arms race in complexity. New models that try to incorporate structural features into GNNs add protein nodes to the graph [Li et al., 2021; Lim et al., 2019] which changes the shape of the graph and introduces unnecessary complexity of the model. In this study, I have introduced Protein-Ligand Interaction Graphs (PLIGs), simple molecular graphs that retain the chemical topology of the ligand but which also encode all of the 3D protein-ligand interactions made by the ligand, by incorporating proximity-based interactions into each ligand atom's node features. Despite their simplicity, PLIGs perform among the top scoring functions when tested against the CASF-2016 benchmark. Using the novel PLIG-based featurization, I conducted a comprehensive analysis of six different GNN architectures with PLIGS and found only small differences in performance between architectures with Graph Attention Networks (GATNet) performing best for predicting binding affinity, indicating that subtle differences in feature aggregation, attention mechanism or activation function are less important than good molecular representation design. In addition, simplicity in model design seems to be desirable for PLIG models specifically, since the hybrid branch architecture that combines the protein sequence and the GNN using PLIG performs worse than models with just PLIG-based GNNs.

Despite their simplicity, GATNet PLIGs (ρ=0.84, RMSE=1.22) perform comparably to some of the best current scoring functions against the CASF-2016 benchmark, approaching methods such as ECIF fingerprints (ρ=0.867, [Sánchez-Cruz et al., 2020]) and graphDelta (ρ=0.87, [Karlov et al., 2020]). In addition, PLIGs outperform other well-established structure-based models such as random forest-based PLEC (ρ=0.817, [Wójcikowski et al., 2018]), OnionNet (ρ=0.816, [Zheng et al., 2019]) and K_{DEEP}

(ρ=0.82, [Jiménez et al., 2018]) as well as other recent GNN-based scoring functions such as SIGN (ρ=0.797, [Li et al., 2021]) and PIGNet (ρ=0.761, [Moon et al., 2022]).

Nonetheless, performance between different graph architectures are fairly similar, and although GATNet performs best, other architectures such as GCN and GAT-GCN models perform almost as well (Figure 4.5). If chosen by random, any herein described GNN architecture in combination with PLIGs would have performed close to, or among the best current scoring functions. This highlights the crucial importance of choosing the right molecular representation rather than fine tuning or developing new and complex deep neural network architectures. While the current top performing graph-based protein-ligand binding affinity scoring functions aim to create more complex architectures to increase performance [Li et al., 2021; Nguyen et al., 2020; Jiang et al., 2020b; Lim et al., 2019], simplicity in architecture and careful featurisation such as through PLIGS can yield similar or better performance while retaining interpretability and easy customizability.

Next, I have shown that model ensembles can be a powerful tool in overcoming the performance drop of binding affinity scoring functions when going from training on crystal to training on docked poses. Rather than creating a hybrid model by adding ligand-based features such as ECFP fingerprints directly into the feature space of a structure-based model (as described by Boyles et al. [2019]), a simple ensemble of a structure-based model and a separate ligand-based fingerprint model can increase performance to levels higher than the individual models by themselves. However, ensembling different structure-based methods does not seem to improve performance, at least for the GNN PLIG and MLPNet ECIF models described here. Additional investigation would be required to test if a more diverse set of model ensembles could improve performance further.

This study shows that simply including proximity-based protein-ligand interac-

tions into the atomic nodes of molecular graphs of the ligand boosts performance of graph neural networks when predicting protein-ligand binding affinity. As a result, it opens up a large space for exploration of which other node features might be included in molecular graphs. Rather than increasing the size of the graph, adding nodes or increasing the complexity, this method of incorporating 3D-structural information into ligand graphs is simple, powerful and compatible with a large variety of GNN architectures.

Furthermore, PLIGs are designed to be easily modified, opening up many possibilities for further improvement of the graphs. The PLIG architecture can therefore serve as an extensible framework to enable researchers to investigate which other kinds of protein-ligand complex representations might be valuable in GNNs. Potential areas of improvement could be to integrate more details about the intermolecular interactions themselves for example through computed biophysical quantities of the interaction. For example, docking-derived atomic scores such as Van der Waals, electrostatics, or desolvation terms could be extracted by scoring the crystal poses of the ligands, and directly incorporated into the atom nodes of PLIGs. Alternatively, as Meli et al. [2021] have shown, atomic-environment vectors (AEVs) are powerful features for the creation of protein-ligand scoring functions and could also be added into atom nodes of PLIGs. Indeed, first steps into the implementation of PLIG/AEV crossover models have already been started in collaboration with Rocco Meli. In theory, virtually any molecular representation that extract features on a ligand atom level could be used as potential node features in PLIGs, opening up many different paths for exploration and potential improvement.

Finally, as the protein-ligand interaction features present in PLIGs are fully interpretable since every dimension in the node feature vector of a PLIG corresponds to a specific, known atom type present in the proteinogenic amino acids of every protein,

further study into the feature importance and interpretability of PLIGs could provide great insights. Understanding the importance that the model assigns to interactions made with every unique atom type in a protein could not only lead to improved models with better performance and generalizability, but could also enable a deeper understanding of which ligand atoms and protein atoms drive high-affinity binding. PLIGs might therefore not only be used as a tool to screen compounds during lead optimisation to prioritise compounds for synthesis, but also as a tool to gain insight into the details of ligand recognition on an atomic level. The first steps towards the implementation of an interpretability analysis pipeline have already been made by Samuel Homberg using GNNExplainer [Ying et al., 2019] under my and Prof. Garrett Morris' supervision, confirming that PLIGs are indeed interpretable on an atom-atom interaction level and can be used to gain insights into interaction importance and might be used in the future to create comprehensive binding hypothesis for binding pockets of protein targets.

Chapter 5

Conclusions

As outlined in Chapter 1, pharmaceutical drug discovery is becoming more and more time consuming and expensive. It is possible that many "low hanging fruits" have already been found, forcing scientists to target more challenging diseases. In addition, the regulatory environment around drug approvals is become stricter, which, in combination with the increased difficulty in targeting challenging diseases, is leading to a higher failure rate in the clinic. The pharmaceutical industry therefore needs to either increase efficiency in pre-clinical drug discovery in order to fail more quickly (and more cheaply) or to otherwise improve the quality of lead compounds that enter the clinical trials. In this thesis, I have outlined three distinct but complementary computer aided drug discovery methods to increase efficiency for the pre-clinical drug discovery pipeline, mainly focusing on hit discovery and hit-to-lead optimization.

One of the main themes of this thesis is the exploration of different representations of molecular structures and non-covalent interactions present in protein-ligand complexes. Since computer-aided drug discovery borrows from the wide field of computer science, CADD does not suffer from a shortage of different methods, models or approaches. Given the variety of choices, the decision about which exact method and representation one needs for a given method or a given task is therefore not trivial. In this thesis, I described three different approaches to molecular representa-

tion, highlighting that to answer slightly different questions within the drug discovery pipeline, completely different molecular representations and machine learning models can be necessary. In Chapter 2, I described the construction of ligand-based models using a Gaussian Process and explore the effect of representing the search space with different ligand-based fingerprints and a molecular autoencoder on the search power of Bayesian optimisation. In Chapter 3, I employed a knowledge-based approach to drug design, comparing protein-ligand interaction fingerprints derived from crystal structures and high-quality models of protein-substrate complexes to docked poses of potential inhibitors to inform fragment-based drug design. Finally, in Chapter 4, I explored a wide range of Graph Neural Network Architectures for protein-ligand binding affinity prediction and expanded upon the interaction fingerprints to create a ligand-shaped graph representation (Protein Ligand Interaction Graph, PLIG) of small molecule ligands that encodes for all protein-ligand interactions made by a given ligand atom within the node features of the graph. GNN-based models using PLIGs are high-performing protein-ligand affinity prediction models.

5.1 Bayesian Optimization for Drug Discovery

In Chapter 2, I focused on the search for the best molecular representation for Bayesian optimisation (BO) used for the hit-to-lead optimisation of small molecule ligands and on the development of a multi-objective optimisation method for the design of compounds with polypharmacological profiles. I used a discrete multi-armed bandit implementation of BO on two datasets, a previously described MMP-12 dataset [Pickett et al., 2011] as well as a novel β-lactamase inhibitor dataset obtained from the Schofield group at the University of Oxford. I showed that BO has the potential to be a powerful tool to aid medicinal chemists in prioritising new molecules to test during structure-activity-relationship (SAR) optimisation. In addition, I found that

fingerprint-based as well as autoencoder-based representations are able to perform well in Bayesian optimization, with ECFP performing best overall. Additionally, I highlighted the current limitations of BO in inhibitor discovery, such as the fairly simple GP surrogate function, and the inability to generate new compounds *de novo* rather than choosing from a list of given compounds.

However, the first steps have already been taken to create follow-on projects to be undertaken in the Oxford Protein Informatics Group to improve the BO implementation described herein and to solve these issues. For example, under the supervision of myself and Prof. Garrett Morris, Dominik Klein worked on the implementation of a Bayesian variant of the PLIG GATNet models described in Chapter 4 by replacing the final layer of the GATNet model with a GP. This would allow the currently used ligand-based fingerprint GP to be replaced with a Bayesian PLIG model, which is a more generalisable surrogate function that encodes for protein-ligand interactions in 3D. While this will most likely not increase performance against a single protein target, the ability to differentiate between the active sites of different targets, *e.g.* in designing selectivity for metallo-β-lactamases, will be highly beneficial for the creation of multi-objective, multi-target optimisation approaches where structural information is key to fine-tune selectivity *vs* promiscuity between proteins of the same family. In order to tackle the issue of compound generation, the first steps towards the implementation of a latent-space representation using variational autoencoders have been made by myself and Ruoyang Feng. Using a latent space representation as the search space would allow the BO to be used as a generative method, sampling from the latent-space to propose new molecules rather than the current implementation, which samples from a discrete set of pre-determined compounds. Implementation of an autoencoder-based compound representation for the purpose of Bayesian optimisation has already been reported and should therefore be a straightforward process

[Gómez-Bombarelli et al., 2018; Griffiths and Hernández-Lobato, 2020].

Finally, additional lab-based validation of the BO method is needed to fully validate its usefulness for medicinal chemists. For that, a batch selection version of the optimisation would need to be implemented that enables the BO algorithm to choose a batch of several molecules at once instead of the single molecule cycles currently used. This would be necessary for practical reasons, as synthesizing and testing molecules one by one would not be efficient. However, a drop in performance would be expected for the BO algorithm when choosing molecules in batches since the model has to select several molecules with the same level of information, rather than being able to update the surrogate function after every selection. Nonetheless, my *in silico* study of BO indicates that BO is applicable and highly useful for early stage hit-to-lead optimisation, especially when optimising compounds with a common scaffold.

5.2 Insights into SARS-CoV-2 M^{pro} Molecular Recognition

In Chapter 3, I outlined my work on the detailed interaction analysis of SARS-CoV-2 M^{pro}-substrate and peptide inhibitor complexes and its insights into the design of novel fragment-based small molecule inhibitors. This work was part of a large collaborative project to tackle the ongoing COVID-19 pandemic by answering the question: "How does SARS-CoV-2-M^{pro} bind and cleave its substrates and how can we use this information to design inhibitors?" and was published in Chemical Science [Chan et al., 2021a].

Using the structural bioinformatics tool Arpeggio [Jubb et al., 2017], I identified all protein-ligand interactions in a large number of different substrate-, peptide inhibitor-, and small molecule ligand-M^{pro} complexes and created a SARS-CoV-2 M^{pro} intermolecular interaction fingerprint representation. This fingerprint was generated

with varying levels of detail ranging from residue-level to atom-level interactions. Using this fingerprint, I found that filtering known fragment binders for key interactions made by the natural substrates revealed highly promising fragment elaboration starting points. Using the Active-Guided Covalent Docking (AGCD) method, I was able to identify potential growth opportunities for the fragments, culminating in the *in silico* design of novel inhibitors. My design approach was validated indirectly by the release of Nirmatrelvir, the M^{pro} inhibitor in the Pfizer COVID-19 drug Paxlovid, which was designed independently and occupies the same key interactions in the M^{pro} active site as the predicted pose of my designs.

There are several avenues for future work. First, to fully validate the small molecule inhibitor design methodology, the designs need to be synthesized and validated experimentally. Second, the *in silico* design and interaction analysis of the peptide inhibitors did not sufficiently answer why the peptide inhibitors are not turned over as substrates. The current results suggest that the peptide inhibitors are able to bind to His-41 (which is part of the catalytic dyad) in a way that forces His-41 into an unproductive conformation, chiefly through π-π stacking interaction with the larger aromatic groups present at the P2 position in many of the peptide inhibitors. However this evidence is currently inconclusive and does not explain why the inhibitors that do not have bulky group at P2 are not turned over as well. Further experiments and computational study of the peptide-M^{pro} binding mode as well as a crystal structures of the inhibitory peptide-M^{pro} complexes are needed to resolve this question.

In addition, I developed AGCD as the first step towards creating a dedicated knowledge-based constrained docking workflow for fragment-based drug discovery. Ultimately, the goal is to create an advanced docking method where new compounds are docked by constraining their MCS with a known fragment crystal structures during docking, while allowing flexibility elsewhere on the ligand. This approach

takes inspiration from Malhotra and Karanicolas ([Malhotra and Karanicolas, 2017]) who identified that for elaborated ligands, in 86% of the 297 paired ligands, the larger elaborated ligand did not change its binding mode relative to the smaller ligand. Although fully implemented methods to do constrained docking already exist in proprietary docking software such as Gold [Jones et al., 1997] or Glide [Friesner et al., 2004], no straightforward implementation to do constrained docking on this scale exists in the most popular open-source docking tools Autodock4 ([Morris et al., 2009]) and AutoDock Vina ([Trott and Olson, 2010]). Therefore, future work should be focussed on the implementation of the needed functionalities into either AD4 or AutoDock Vina to create a dedicated open-source constrained docking method to advance fragment-based drug discovery.

Overall, in this chapter I have shown the effectiveness of classical structural bioinformatics and cheminformatics techniques for the efficient *in silico* design of inhibitors. Using the knowledge-based protein-ligand interaction fingerprint approach in combination with a sophisticated molecular docking pipeline, I was able to identify and design promising small molecule ligands to inhibit SARS-CoV-2 M^{pro}. Building on this work, with the aim of using these interaction fingerprints for machine learning to score protein-ligand complexes, I explored different ways of encoding interactions for deep learning techniques. This ultimately resulted in the Protein Ligand Interaction Graphs (PLIGs) described in Chapter 4.

5.3 Development of Protein Ligand Interaction Graphs

Chapter 4 outlined my work on exploration of GNNs and the development of a novel graph-based representation of protein-ligand complexes for protein-ligand affinity prediction. Inspired by my previous work on interaction fingerprints for SARS-CoV-2

M^{pro} described in Chapter 3, I explored how protein-ligand interactions could be best encoded for use in machine learning models. In order to address both the limitations of interaction-based fingerprints as well as known structure-based graph neural networks, I created Protein Ligand Interaction Graphs (PLIGs). PLIGs retain the topology of the ligand but also encode for all protein-ligand interactions made in 3D by incorporating counts of intermolecular atom-atom interactions into the node features of the graph. PLIGs perform among the best known protein-ligand affinity scoring functions when tested on the CASF-2016 benchmark, despite their simplicity, especially in comparison to other high performing, but much more complicated graph-based models.

First and foremost, I created PLIGs to be simple and easily adaptable, opening up many possibilities for further improvement. For example, the first steps have been made by my collaborator Rocco Meli, to explore which other atomic features could be incorporated into PLIG node features to describe the interactions more accurately. The overarching goal of this follow-up study is to find ways to represent protein-ligand complexes to allow the model to learn the biophysics of molecular recognition in order to build a generalisable model that is able to score complexes accurately, based on their predicted or actual intermolecular interactions. The most straightforward choices for potential feature additions would be the intermolecular atomic contributions of the docking score which can be extracted from docking software such as AutoDock4 [Morris et al., 2009] or AutoDock Vina [Trott and Olson, 2010], or features from other high performing scoring functions such as Atomic Environment Vectors [Meli et al., 2021].

In addition to studies to improve the PLIG model itself, an investigation into the interpretability of PLIGs has been started by Samuel Homberg under the supervision of myself and Prof. Garrett Morris. PLIGs are designed to be fully interpretable,

since every position in the node feature vectors of a given ligand directly corresponds to a unique, identifiable protein atom in proximity to the ligand atom. Preliminary results from Samuel Homberg indicate that it is indeed possible to gain an atom-level insight into the importance of any protein atom within 6 Å which is involved in binding, as well as the importance of the ligand atom-derived features in the node feature vector. This could not only help in guiding further work to improve PLIGs, for example by eliminating protein atom types from the representation that are less important, but could enable the usage of PLIGs by medicinal chemists to guide SAR optimisation directly. Once trained on a given dataset, PLIGs could be analysed to reveal important sets of atoms in the binding site of the protein as well as in the ligand itself. This could help guide medicinal chemists to prioritise interactions with important protein atoms as well as reveal potential functional groups on the ligand that are more suitable for a given protein target.

Overall, the high performance and simplicity of PLIGs are designed in light of a current trend in the field: "Increased complexity of a model does not necessarily correlate with more accurate predictions". Improvement in predictive performance (for example on the CASF-2016 benchmark) has been minimal between models, and incremental at best, over the last few years as indicated by the list of models mentioned in Chapter 4. Model complexity however, especially for GNN-based models, has increased significantly. PLIGs therefore highlight that instead of aiming for high complexity and a small improvement in performance, better interpretability should be a more desirable design goal when creating novel machine-learning based scoring functions.

5.4 Concluding Remarks

In this thesis I have focused on the development of practical, easy to implement, yet powerful computer-aided drug discovery tools. In a world where computational resources are increasingly abundant, and the choice of different, highly complex machine learning algorithms seems almost endless, it is crucial for the future development of computer-aided drug discovery tools to be focused on interpretability, generalisability and practicality. Models need to be interpretable to convince medicinal chemists to use them with confidence when designing new molecules. Additionally, good models also need to be robust and generalisable to ensure that dataset biases like similar ligand structures or the overabundance of one protein family does not create machine-driven optimisation pipelines where instead of exploring unknown chemical space, models generate similar drug-like molecules, reinforcing previous biases. Finally, the practicality of a model is important in regards to its usage in a real-life drug discovery project. Before selecting a potential new tool to use, drug discovery researchers need to answer: Does the model need to be retrained/adjusted for every new dataset? Can it integrate into the existing tools in the pipeline? Is the implementation of the code easy to setup and the model results reproducible? How easy is it to add new features? How computationally expensive is the model?

Overall, this thesis presents three approaches to improve the efficiency of early stage pre-clinical drug discovery. All implementations of my methods and models developed in each chapter are available on GitHub and easily adjustable and executable. Finally, the projects described herein have already been fundamental for the creation of a series of ongoing follow-on projects to improve and expand upon the discoveries made in this thesis to drive forward the improvement of pre-clinical drug discovery.

Appendix A

Exploration of Bayesian Optimization for Structure-Activity Relationship Modeling

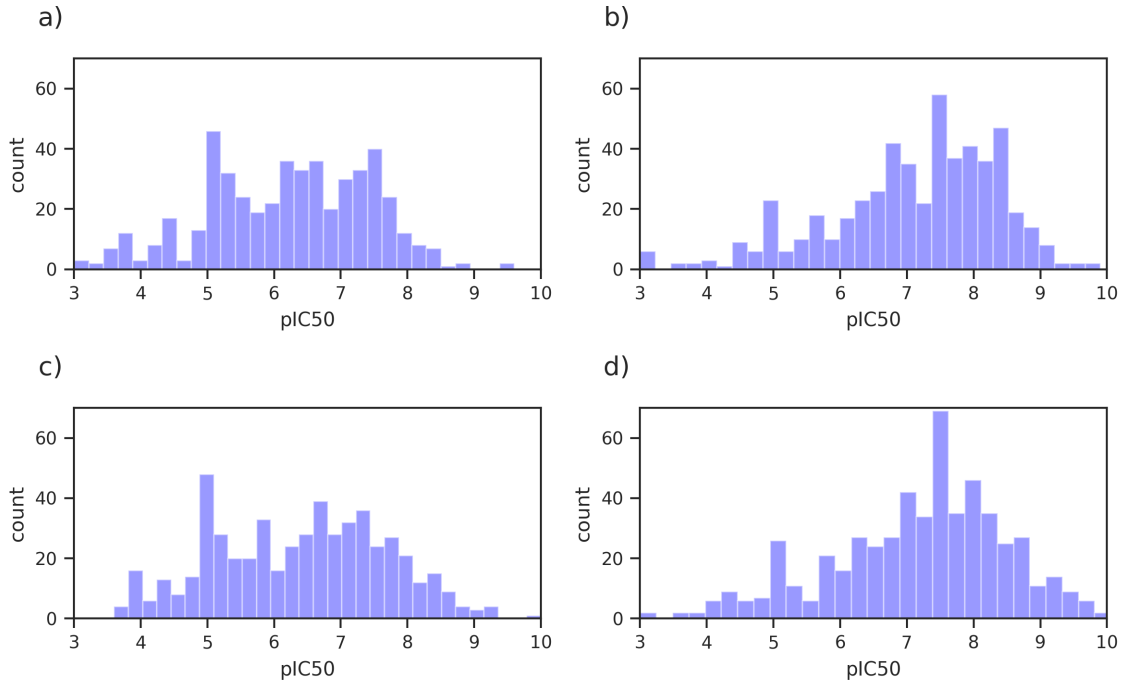

Figure A.1: Distribution of pIC_{50} values for all inhibitors against each target. VIM-1, VIM-2, IMP-1, NDM-1 are shown in (a), (b), (c) and (d), respectively. The VIM-2 b) and NDM-1 d) datasets are skewed towards more high affinity ligands in comparison to VIM-1 a) and IMP-1 c).

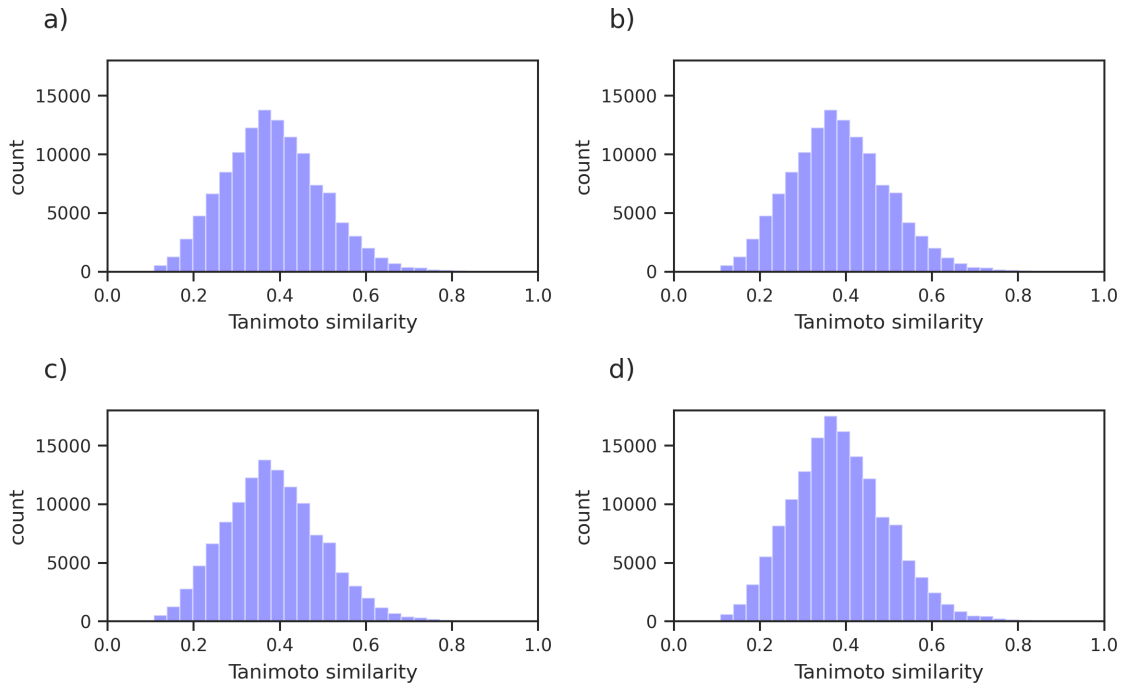

Figure A.2: Distribution of Tanimoto similarity scores between all compounds in each respective dataset. VIM-1, VIM-2, IMP-1, NDM-1 are shown in (a), (b), (c) and (d), respectively. No large differences in ligand similarity between datasets since the compounds in each dataset are mostly the same. Overall, the dataset is relatively diverse given that every ligand has the same scaffold, with a mean Tanimoto similarity of around 0.4 for all four datasets.

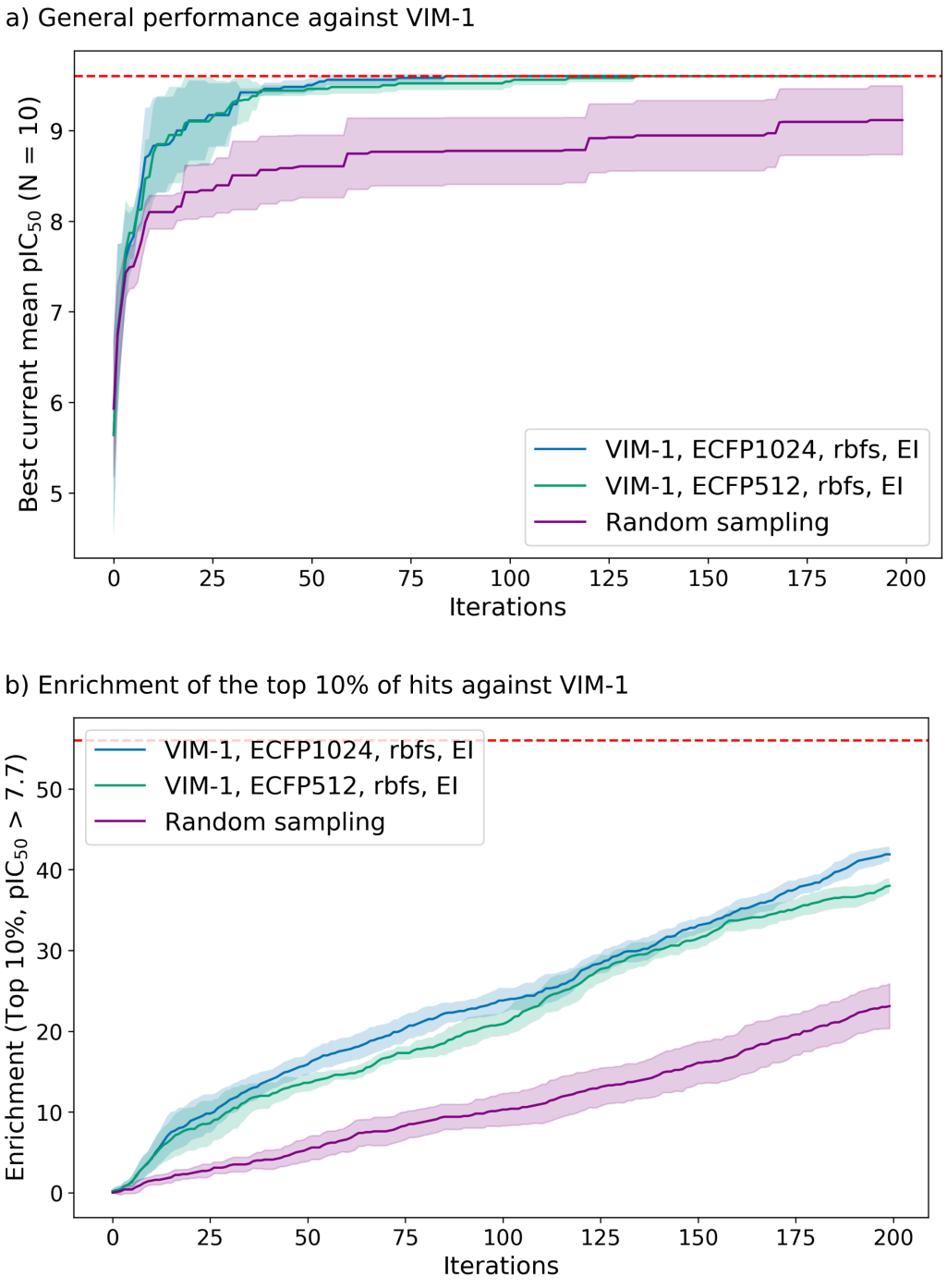

Figure A.3: Performance of Bayesian optimization using the VIM-1 dataset and varying ECFP bit length. The dashed red line indicates the maximum possible value for both: maximum pIC_{50} and maximum count of top compounds; bold lines indicate the mean of 10 experiments and shaded area the 95% CI. a) General performance plot showing the best current mean pIC_{50} found over 200 iterations of BO or random sampling. b) Enrichment plot showing the total count of top compounds found so far at each iteration. Both representations perform equally well on metric a), but ECFP1024 outperforms slightly on metric b).

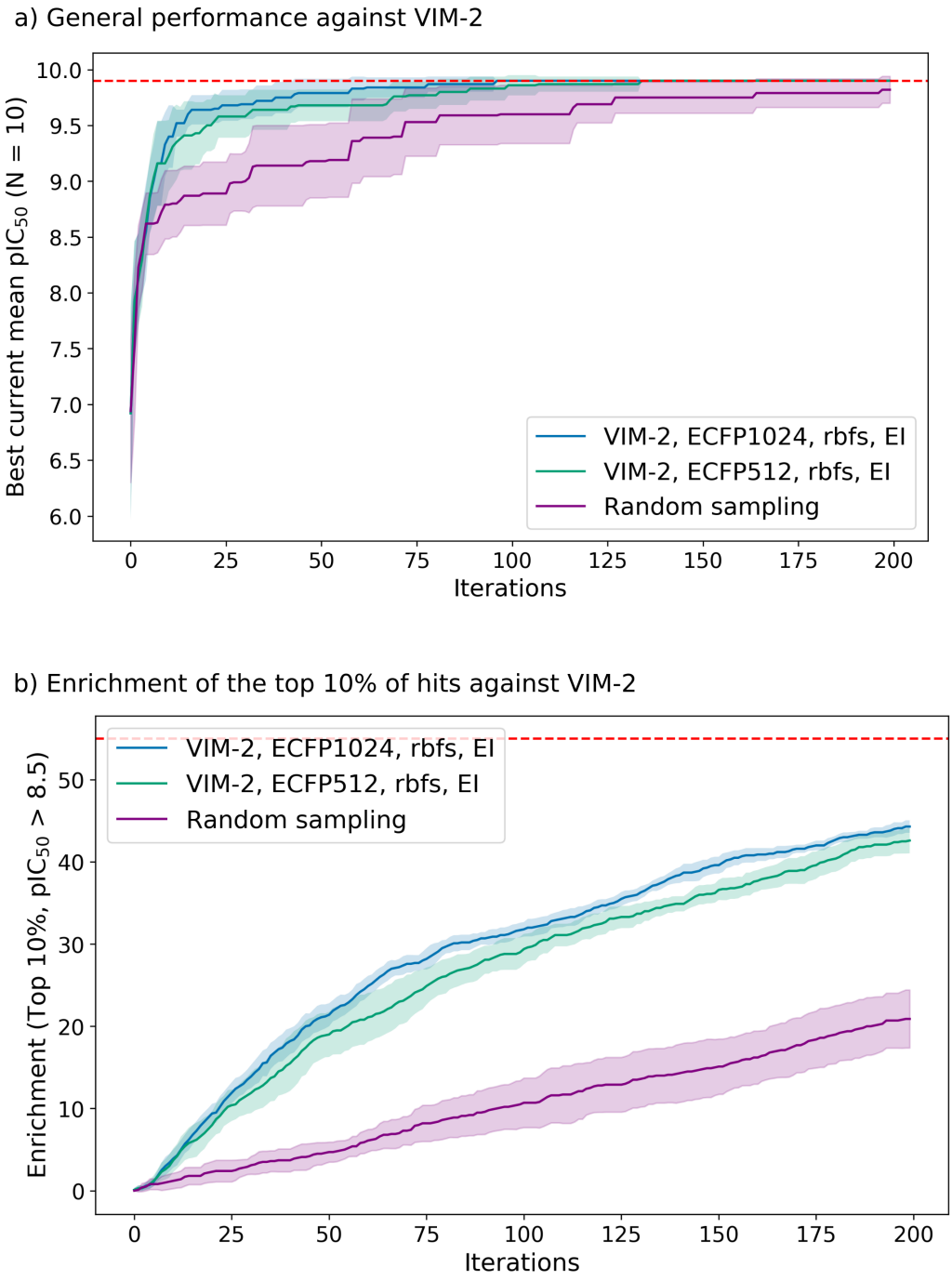

Figure A.4: Performance of Bayesian optimization using the VIM-2 dataset and varying ECFP bit length. The dashed red line indicates the maximum possible value for both: maximum pIC_{50} and maximum count of top compounds; bold lines indicate the mean of 10 experiments and shaded area the 95% CI. a) General performance plot showing the best current mean pIC_{50} found over 200 iterations of BO or random sampling. b) Enrichment plot showing the total count of top compounds found so far at each iteration. Both representations perform equally well on metric a), but ECFP1024 outperforms slightly on metric b).

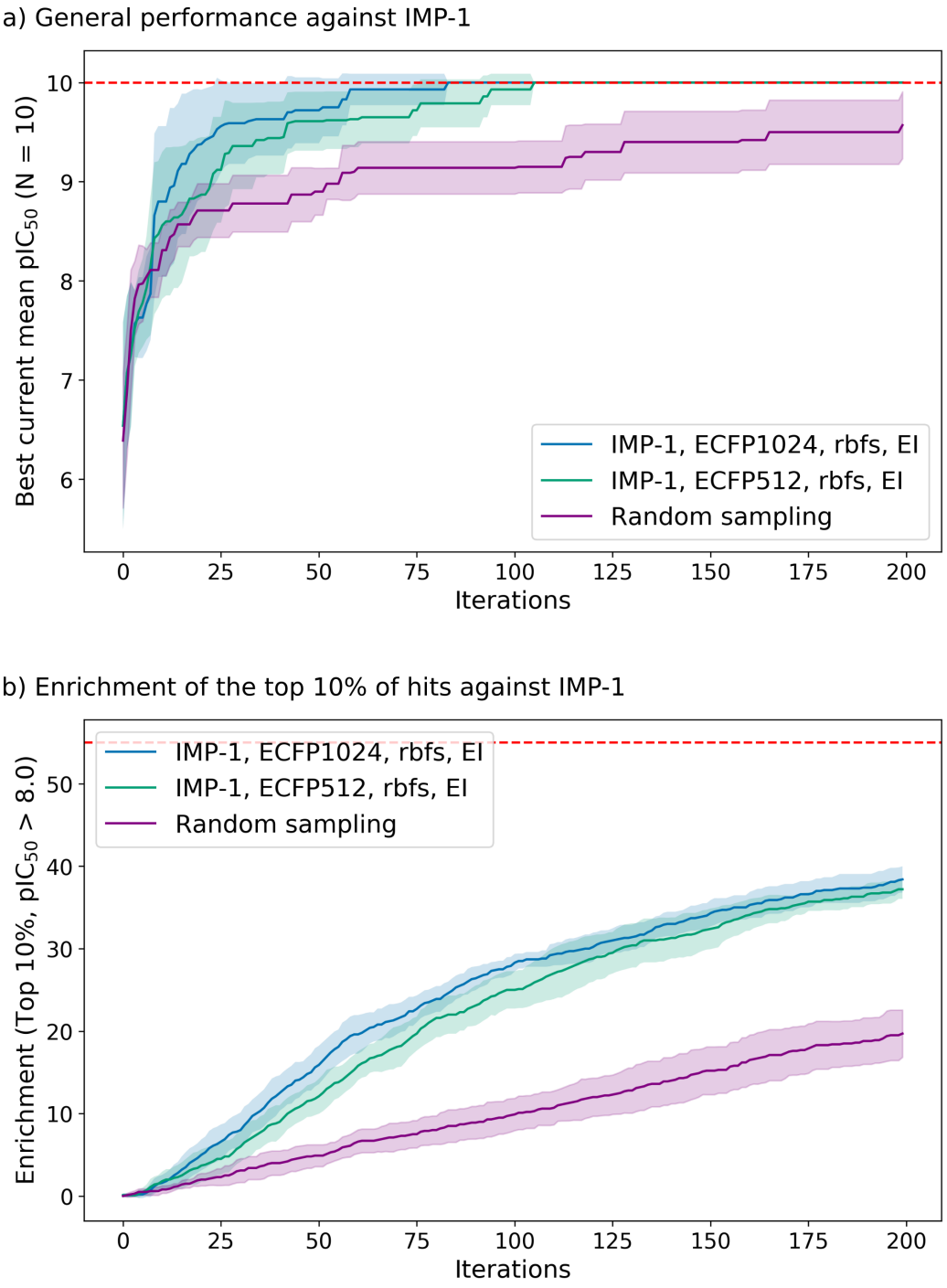

Figure A.5: Performance of Bayesian optimization using the IMP-1 dataset and varying ECFP bit length. The dashed red line indicates the maximum possible value for both: maximum pIC_{50} and maximum count of top compounds; bold lines indicate the mean of 10 experiments and shaded area the 95% CI. a) General performance plot showing the best current mean pIC_{50} found over 200 iterations of BO or random sampling. b) Enrichment plot showing the total count of top compounds found so far at each iteration. Both representations perform equally well on metric a), but ECFP1024 outperforms slightly on metric b).

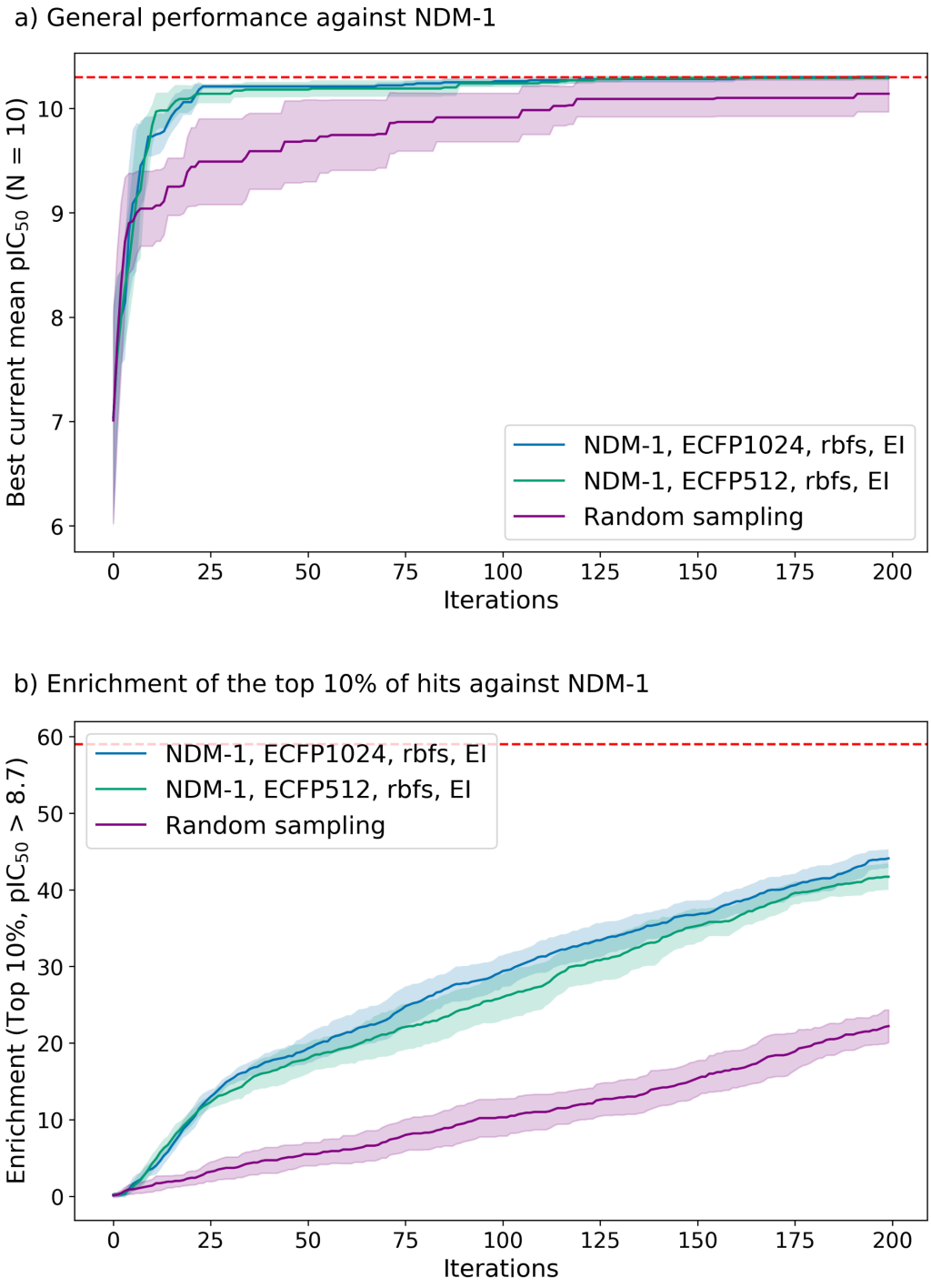

Figure A.6: Performance of Bayesian optimization using the NDM-1 dataset and varying ECFP bit length. The dashed red line indicates the maximum possible value for both: maximum pIC_{50} and maximum count of top compounds; bold lines indicate the mean of 10 experiments and shaded area the 95% CI. a) General performance plot showing the best current mean pIC_{50} found over 200 iterations of BO or random sampling. b) Enrichment plot showing the total count of top compounds found so far at each iteration. Both representations perform equally well on metric a), but ECFP1024 outperforms slightly on metric b).

Appendix B

In silico Design and Validation of SARS-CoV-2 M^{pro} Inhibitors from Modelling Substrate and Ligand Binding

B.1 Supplementary Methods Provided by Collaborators

The following methods were developed and performed by collaborators and are not my own work. I have included the following procedures to allow convenient access for the reader since results obtained from these methods were either directly used in follow-on experiments conducted by myself or are directly relevant to insights gained by my work. The descriptions of the results and methods from co-authors were directly adapted from the original collaborative publication with minor adjustments as covered by the CC BY 3.0 license [Chan et al., 2021a].

B.1.1 Comparative Modelling of the SARS-CoV-2 M^{pro}-Peptide Complexes

This method was developed and carried out by Garrett M. Morris. I have included this method description provided by Garrett M. Morris as published under the CC BY 3.0 license [Chan et al., 2021a].

"A crystal structure of the TSAVLQ↓SGFRK 11-mer peptide substrate bound to the H41A mutant of dimeric SARS-CoV M^{pro} (PDB entry 2q6g [Xue et al., 2008]) was superimposed with one of unmodified dimeric SARS-CoV-2 M^{pro} (PDB entry 6yb7; 1.25 Å resolution [Owen et al., 2020]). The substrate was transferred over to the chain A active site of the catalytically-competent SARS-CoV-2 M^{pro} structure. The sequences of the 11 native cleavage sites processed by SARS-CoV-2 M^{pro} (s01-s11) were identified by aligning the sequences of the ORF1ab polyproteins of both SARS-CoV (GenBank accession code NC_004718.3 [He et al., 2004]) and SARS-CoV-2 isolate Wuhan-Hu-1 (accession code MN908947.3 [Wu et al., 2020]) using MUSCLE [Edgar, 2004]. For each of the 11 cleavage sites, atomic models of an 11-mer peptide matching positions P6 to P5′ and charged N- and C-termini were constructed using the mutagenesis tool of the open source version of PyMOL (v. 2.3.0) [Schrödinger LLC., 2020]. For every sidechain from positions P6 to P5′, apart from Gly and Ala, the highest-probability backbone-dependent conformer with the least steric clash and the most chemical complementarity was selected [Dunbrack Jr. and Cohen, 1997]. Using CCG MOE version 2019.0104 [Chemical Computing Group ULC., 2019], each of the resulting 11 models of the SARS-CoV-2 M^{pro} dimer complexed with each 11-mer substrate in the A-chain active site underwent structure preparation protonation using Protonate 3D. Each model was then solvated using 0.1 M NaCl and explicit water and subjected to energy minimization using the $AMBER_{10:EHT}$ force field [Cornell et al., 1995; Gerber and Müller, 1995] and periodic boundary conditions, until convergence with an RMS of 0.4184 kcal mol^{-1} per iteration was reached."

B.1.2 Explicit-Solvent Molecular Dynamics

This method was developed and carried out by H.T. Henry Chan. I have included this method description provided by H.T. Henry Chan for clarity as published under

the CC BY 3.0 license [Chan et al., 2021a].

"Pre-solvation models of the dimeric M^{pro}-peptide complexes constructed as described above (Method Section B.1.1) were used as starting points for MD simulations. All additives and crystallographic water molecules were removed from PDB entry 6yb7, except HOH 644 which provides bridges between His-41, His-164 and Asp-187. Protonation and rotameric states of histidines and other titratable residues were assigned at pH 7.4 based on a combination of Reduce (MolProbity, Duke University [Williams et al., 2018]), H++ (Virginia Tech [Anandakrishnan et al., 2012]), PROPKA3 (PDB2PQR [Olsson et al., 2011]), and visual inspection, with a final M^{pro} monomeric charge of -4. His-41 (protonated on its δ-nitrogen) and Cys-145 were assigned neutral.

MD simulations were performed using GROMACS (v. 2019.2 [Abraham et al., 2015]) employing the AMBER99SB-ILDN force field [Lindorff-Larsen et al., 2010]. Each of the constructed complexes was solvated (TIP3P water model [Jorgensen et al., 1983]) in a rhombic dodecahedral box (1.0 nm buffer), neutralised, and minimised using the steepest descent algorithm until the maximum force was below 1000 $kJ\,mol^{-1}\,nm^{-1}$. For each peptide sequence, three independent simulations were initiated by random velocities at 298.15 K. In each case, the system was equilibrated under NVT (200 ps; 1 fs timestep) and NPT (200 ps; 1 fs timestep) conditions, before being subjected to 200 ns MD simulation (2 fs timestep) at 298.15 K and 1 bar, during which protein-peptide interactions were monitored. All simulations were performed with three-dimensional periodic boundary conditions. Long-range electrostatics was calculated using the smooth particle mesh Ewald method [Essmann et al., 1995]. All bond lengths involving hydrogen atoms were constrained with the LINCS algorithm [Hess et al., 1997]. Hydrogen bonds between M^{pro} and the peptides were monitored over the course of the simulations, defined using a combined criteria on the donor-

acceptor distance (dD-A $\leq$ 3.5 Å) and the proton-donor-acceptor angle ((H-D-A) $\leq$ 30°).

Models of the designed sequences complexed with SARS-CoV-2 M^{pro} (PDB entry 6yb7) were built using a comparative modelling approach similar to that described above (Methods Section B.1.1), starting from the previously constructed model of the M^{pro}-s02 complex. Each constructed complex was then solvated, minimised, equilibrated, and subjected to 3 × 200 ns MD as described above, except the retention of a backbone restraint during NPT equilibration to allow longer relaxation of the non-native peptide side chains and the M^{pro} binding pockets.

To generate representative structures of M^{pro}-peptide complexes for interaction analysis, frames extracted every ns from the concatenated 3 × 200 ns MD trajectories were fitted using the M^{pro} backbone, before performing clustering based on the heavy-atom RMSD of the peptide, using the gromos algorithm as implemented in GROMACS (v. 2019.2) [Daura et al., 1999]. A cut-off of 2.0 Å (for native substrates) or 2.5 Å (for p12 and p13, due to heavier residues in their terminal regions) was used."

B.1.3 Protein Production and Purification

Recombinant M^{pro} protein for the *in-vitro* assays was obtained from Claire Strain-Damerell, Eidarus Salah and Petra Lukacik and was produced and purified as previously reported [Malla et al., 2021].

B.1.4 Peptide Synthesis

This method was developed and carried out by Tika R. Malla and Tobias John. I have included this method description provided by Tika R. Malla for clarity as published under the CC BY 3.0 license [Chan et al., 2021a].

"Peptide synthesis was performed as previsouly reported [Malla et al., 2021]. s01, s01-LP2W , s01-QP1W, p12, p13, p13-WP2L, p15 and p16 were synthesised on a 0.1-0.25 mmol scale from C- to N-terminus on Rink amide-MBHA resin (100–200 mesh, 0.6–0.8 $\mathrm{mmol\,g^{-1}}$ loading, AGTC Bioproducts) using a muwave assisted LibertyBlue peptide synthesizer (CEM) and N-Fmoc protected α-amino acids (CS Bio, Novabiochem, Sigma-Aldrich, TCI, Alfa Aesar, Merck or AGTC Bioproducts). N,N'-diisopropylcarbodiimide (TCI Europe) and Oxyma Pure (Merck) in DMF and 20 % (v/v) piperidine in DMF (peptide synthesis grade, AGTC Bioproducts) were used for iterative cycles of coupling and deprotection respectively under the manufacturer's standard protocol. Following the terminal Fmoc-deprotection step, the resin was washed with CH2Cl2, dried in air, then treated with 5-10 mL of a deprotection solution (2.5:2.5:2.5:92.5 (v/v) 1,3-dimethoxybenzene, triisopropylsilane, MilliQ water and trifluoroacetic acid) for 3 h at ambient temperature. The resulting mixture was filtered and the filtrate was diluted with cooled Et2O (3×45 mL) to precipitate the peptide. Et2O was decanted, peptide dried on air and lyophilised overnight.

Peptides apart from P1 mutant of s01 were dissolved in DMSO and quantified by spiking the sample with 3 $\mathrm{mmol\,g^{-1}}$ of an internal standard 3-(Trimethylsilyl)propionic-2,2,3,3-d4 acid sodium salt. The eleven substrate peptides (s01-s11) were purchased from GLBioChem (Shanghai)."

B.1.5 Substrate Turnover Analysis Under Denaturing Conditions

This method was developed and carried out by Tika R. Malla. I have included this method description provided by Tika R. Malla for clarity as published under the CC BY 3.0 license [Chan et al., 2021a].

"20 μM stock of all 11 native substrate peptide sequences were prepared in the assay buffer (20 mM HEPES, pH 7.5, 50 mM NaCl). E1-ClipTip™ Bluetooth™ Elec-

tronic multichannel pipette (ThermoFisher) was used to dispense 5 μL/well (x24) of each peptide in a single row of a 384 well plate. The first column was treated with a final concentration of 1 % (v/v) aqueous formic acid to obtain a 0 min time point. M^{pro} was dispensed using Multidrop to obtain a final concentration of 0.15 μM M^{pro} with 2 μM peptides in all wells. Each column was sequentially quenched (every minute) with 1 % (v/v) aqueous formic acid. Samples were analysed by solid-phase extraction (SPE) coupled to mass spectrometry (MS) using a RapidFire Mass Spectrometer. The operating parameters in the positive ion mode were: capillary voltage (4000 V), nozzle voltage (1000 V), fragmentor voltage (365 V), drying gas temperature (280 °C), gas flow (13 $\mathrm{L\,min^{-1}}$), sheath gas temperature (350 °C) and sheath gas flow (12 $\mathrm{L\,min^{-1}}$). Samples were loaded onto a SPE C4-cartridge, which was then washed with 0.1 % (v/v) aqueous formic acid to remove non-volatile buffer salts (5.5 s, 1.5 $\mathrm{mL\,min^{-1}}$) followed by elution with aqueous 85 % (v/v) acetonitrile in 0.1 % (v/v) formic acid (5.5 s, 1.25 $\mathrm{mL\,min^{-1}}$). The cartridge was equilibrated with 0.1 % (v/v) aqueous formic acid (0.5 s, 1.25 $\mathrm{mL\,min^{-1}}$) prior to every sample injection. Data were exported in a plate list mode and processed in Excel to calculate percentage product turnover."

B.1.6 Substrate Binding and Turnover Analysis Under Non-Denaturing Conditions

This method was developed and carried out by Victor Mikhailov. I have included this method description provided by Victor Mikhailov for clarity as published under the CC BY 3.0 license [Chan et al., 2021a].

"Non-denaturing mass spectra were obtained using a Waters Synapt HDMS Q-TOF mass spectrometer coupled with an automated chip-based nano-electrospray ion source (TriVersa Nanomate, Advion). A larger concentration of M^{pro} than the one used in the denaturing MS assays was used to provide sufficient sensitivity. 5

μM of M^{pro} was mixed with 13-fold molar excess of a substrate (s01-s11) in 200 mM of ammonium acetate (pH 6.9) at room temperature and electrosprayed (1.77 kV spray voltage, 0.55 psi spray backing gas pressure and 4.3 mbar inlet pressure). The sample and extractor cone voltages were maintained at 180 V and 1 V, respectively; no in-source dissociation of M^{pro} dimers was observed at these voltages. Mass spectra were recorded after 1, 3, 6, 9 and 12 min incubation. Measurements were taken in duplicate for each substrate. Data collection and analysis were carried out using Waters MassLynx software. Integrated peak areas of the substrate ions and cleavage product ions were compared at different time points: the sum of substrate and product ions intensities was set at 100 % for each measurement, and the level of depletion of the substrate ions was used as a measure of the turnover efficiency."

B.1.7 Dose Response Curve Analysis

This method was developed and carried out by Tika R. Malla. I have included this method description provided by Tika R. Malla for clarity as published under the CC BY 3.0 license [Chan et al., 2021a].

"Methods for SPE coupled RapidFire MS-based inhibition assay were as previously reported [Malla et al., 2021]. In brief, in a 384 polypropylene well plate, 100 μL of 2.5 mM stocks of the designed peptides were transferred. 11 point 3 fold serial dilutions of the peptides were performed in 60 μL using E1-ClipTip™ Bluetooth™ Electronic multichannel pipette (ThermoFisher) in DMSO with 5 mix cycles of 30 muL volume for mixing. 10 μL was drawn from each well and 5 μL was transferred to two wells of a new destination 384 well polypropylene plate. 5 μL of DMSO (positive turnover control) and 5 μL of 10 % (v/v) aqueous formic acid (negative turnover control) were added to 16 wells each on every destination plate. 25 μL/well of x2 stock of enzyme in assay buffer (20 mM HEPES, pH 7.5, 50 mM NaCl) was dispensed

using a Multidrop Combi machine; incubation for 15 minutes was followed by dispensation of x2 stock of substrate in each well to obtain 0.15 μM M^{pro} and 2 μM s01 concentration. The reaction was allowed to progress for 10 min ($\sim$50 % turnover in DMSO control), then quenched with 5 μL of 10 % (v/v) aqueous formic acid. The plates were centrifuged for $\sim$15 s after addition of each reagent at 2500 rpm (Star lab) to ensure all dispensed solutions were pooled at the bottom of the plate. The plates were analysed by SPE coupled MS under the conditions specified in Section B.1.5. RapidFire integrator software was used to extract and integrate abundance peaks of the +1 charge states of the substrate (1191.68 Da) and N-terminal cleaved product (617.34 Da). Data was exported in a plate list mode and processed in Excel to calculate percentage product turnover, normalisation of percentage activity followed by deduction of percentage inhibition. Normalised percentage inhibition data were exported to GraphPad Prism 8 and non-linear regression analysis was performed to obtain IC_{50} values. Top and bottom constraints of 100 % and 0 % were applied respectively for the analysis of reported IC_{50} values curves. Z' of the assay was always $\leq$0.8."

B.1.8 Dose Response Curve Analysis with Varying Substrate Concentrations

This method was developed and carried out by Tika R. Malla. I have included this method description provided by Tika R. Malla for clarity as published under the CC BY 3.0 license [Chan et al., 2021a].

"The designed peptides were dispensed using an Echo 550 acoustic liquid handling robot. Samples were prepared as described above (Section B.1.7) with final substrate concentrations of 2 μM, 10 μM, 20 μM and 40 μM TSAVLQ/SGFRK-NH_2 (s01) with 10, 10, 15 and 20 minutes of incubation with substrates, respectively."

B.1.9 Designed Peptide Turnover Analysis Under Denaturing Conditions

This method was developed and carried out by Tika R. Malla. I have included this method description provided by Tika R. Malla for clarity as published under the CC BY 3.0 license [Chan et al., 2021a].

"100 μM stocks of p12, p13, p15, p16, p13-WP2L, s01-LP2W and s01-QP1W were prepared. 0.15 μM of enzyme was dispensed and incubated with 2 μM peptide (see Section B.1.5); the reaction was allowed to proceed overnight at 37°C, 300 rpm in a thermomixer. Samples were analysed by SPE coupled MS. After integration using RapidFire Integrator, the data was analysed in Excel and presented using GraphPad Prism 8. For a summary of the m/z charge state see Table B.1."

Peptides	Sequence	Substrate m/z (Da)	Product m/z (Da)
s01	TSAVLQ↓SGFRK	1191.68 (+1)	617.34 (+1)
s02	SGVTFQ↓SAVKR	1177.65 (+1)	637.30 (+1)
s03	KVATVQ↓SKMSD	1191.62 (+1)	
s04	NRATLQ↓AIASE	1171.55 (+1)	
s05	SAVKLQ↓NNELS	1200.57 (+1)	644.37 (+1)
s06	ATVRLQ↓AGNAT	1099.53 (+1)	686.36 (+1)
s07	REPMLQ↓SADAQ	1243.51 (+1)	772.39 (+1)
s08	PHTVLQ↓AVGAC	2185.98 (+2)	
s09	NVATLQ↓AENVT	1157.60 (+1)	644.35 (+1)
s10	TFTRLQ↓SLENV	1305.62 (+1)	764.37 (+1)
s11	FYPKLQ↓SSQAW	1352.59 (+1)	794.37 (+1)
p12	KYTFWQYSQFY	1558.75 (+1)	
p13	KYLTWQNSQIN	1392.70 (+1)	
p15	LTINWQKYFNT	1427.62 (+1)	
p16	WFTLKQYWQTN	1514.70 (+1)	
p13-WP2L	KYLTLQNSQIN	1319.71 (+1)	
s01-LP2W	TSAVWQ↓SGFRK	1264.65 (+1)	690.33 (+1)
s01-QP1W	TSAVLWSGFRK	1249.68 (+1)	

Table B.1: Table adapted from the original publication of this work under the CC BY 3.0 license [Chan et al., 2021a]. Observed mass (Da) and (m/z) charge states of the peptides that were extracted using RapidFire Integrator for peak integration.

B.1.10 LCMS Analysis for Designed Peptides

This method was developed and carried out by Tika R. Malla and Victor Mikhailov. I have included this method description provided by Tika R. Malla and Victor Mikhailov for clarity as published under the CC BY 3.0 license [Chan et al., 2021a].

"LCMS experiments were performed using an Agilent Infinity Series II System attached to QTOF 6650 using an Agilent Zorbax C-18 Extend column. Solvent A: LCMS grade water with 0.1 % formic acid, and solvent B: 100 % acetonitrile in 0.1 % (v/v) formic acid was used at 0.2 $\mathrm{mL\,min^{-1}}$ flow rate to elute the peptides over a gradient of 22-55 % of solvent B over 8 minutes. The operating parameters for the LCMS were the same as above (Section B.1.5). In a 96 well plate, samples consisting of 0.15 μM M^{pro} were prepared. p12, p13, p15, p16 and s01 were transferred from source wells to destination wells with M^{pro} using the multi injector program and samples injected immediately after mixing. 30 min, 3 h, 6 h, 1 day and 2 days time points were obtained for peptides. Samples were covered with a polypropylene cover to limit evaporation."

B.1.11 Designed Peptide Binding and Turnover Analysis Under Non-Denaturing Conditions

This method was developed and carried out by Victor Mikhailov. I have included this method description provided by Victor Mikhailov for clarity as published under the CC BY 3.0 license [Chan et al., 2021a].

"The binding of designed peptides p12, p13, p15 and p16 to M^{pro} dimers and their effects on substrate turnover were investigated using non-denaturing mass spectrometry (Section B.1.6). 5 μM of M^{pro} was mixed with designed peptides at different levels of peptide excess in 200 mM of ammonium acetate (pH 6.9) at room temperature. Non-denaturing mass spectra were recorded for different protein-peptide molar

concentration ratios (up to 16-fold excess of peptide relative to the protein). At the final step, the native s01 substrate was added to the protein-peptide mixture at 4-fold excess over the protein, and its turnover recorded after 3- and 6-min incubation."

B.2 Supplementary Results - Monitoring of Substrate Sequence Hydrolysis by Mass Spectrometry

The following description of the results was directly adapted from the original publication with minor adjustments as covered by the CC BY 3.0 license [Chan et al., 2021a]. This work was done by Tika R. Malla and Victor Mikhailov.

"To rank the SARS-CoV-2 M^{pro} preferences for hydrolysis of the 11 cleavage sites, we monitored turnover of 11-mer peptides under non-denaturing MS conditions using ammonium acetate buffer (Figure B.1 a). Peptides s01, s06, s08, s10 and s11 evidenced fast turnover. The level of substrate ion depletion was $>$ 70 % after 1 min and $>$ 90 % after 6 min incubation. Peptides s02, s04 and s09 showed substrate ion depletion from 35 to 45 % after 1 min incubation, $>$ 70 % depletion after 6 min, and $>$ 90 % depletion after 12 min. Peptides s03, s05 and s07 demonstrated slow turnover that was below 50 % after 12 min incubation.

Under non-denaturing conditions (Figure B.1) turnover was: fast (s01, s11, s06, s10, and s08), medium (s04, s02, and s09), and slow (s05, s03, and s07). Substrates s01, s11 and s06 turned over fastest; while s07, s05 and s03 were slow as measured by both methods.

The observed turnover of all 11 SARS-CoV-2 cleavage-site-derived peptides by M^{pro} is consistent with our atomistic models (Main Section 3.4.1, where the peptides remain bound in the active site during MD simulations and where the scissile amide carbonyl remains well-positioned in the oxyanion hole for reaction initiation.

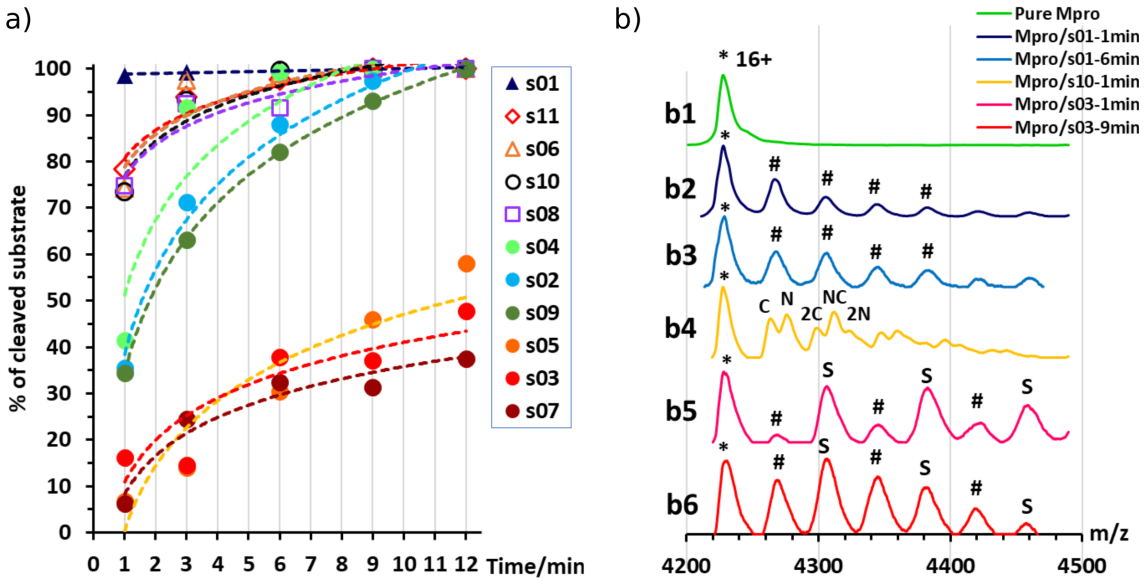

Figure B.1: Figure adapted from the original publication of this work under the CC BY 3.0 license [Chan et al., 2021a]. Non-denaturing mass spectrometry of M^{pro} substrate turnover as published in the original publication of this work [Chan et al., 2021a]. a) Substrate turnover versus incubation time as measured by non-denaturing MS. Trend lines are given for visual guidance only. b) Examples of mass spectra showing normalized intensity in the m/z region around the 16+ charge state of M^{pro} dimer (asterisk, *): b1) pure M^{pro} solution (5 μM); b2) M^{pro} and s01 solution after 1 min incubation, hashes (#) indicate the mass peaks corresponding to the s01 cleaved fragments sequentially attached to the M^{pro} dimer; note: the resolution is not sufficient to distinguish between the N- or C-terminal fragments (mass shifts of 617 and 593 Da, respectively); b3) same solution as (b2) after 6 min incubation; (b4) M^{pro} and s10 solution after 1 min incubation, 'N' labels N-terminal fragment(s) attached (765 Da), 'C' labels C-terminal fragment(s) attached (560 Da); (b5) M^{pro} and s03 solution after 3 min incubation, 'S' labels intact substrate(s) attached, hashes (#) label attached substrate fragments, but the N- and C-terminal fragments cannot be distinguished (mass shift 644 and 566 Da, respectively); b6) same solution as b5) after 9 min incubation. All substrates are turned over during non-denaturing mass spectrometry, validating the substrate models.

The stability of the M^{pro}–peptide interactions involving the S2 and S1 subsites, as well as backbone–backbone HBs 2, 3, 10 and 11, could explain the observation using non-denaturing MS of complexes of M^{pro} with products — because of slow product dissociation. Nevertheless, we envisage that the order of substrate turnover rates is likely determined by various factors, including peptide conformations, the influence of the P2 and P1′ residues on the catalytic dyad, entropic effects, and rates of product dissociation, all of which prompt ongoing experimental and computational

investigations."

B.3 Supplementary Results - *In silico* Mutational Analysis of Substrate Peptides Enables Peptide Inhibitor Design

The following description of the results was adapted from the original publication with minor adjustments as covered by the CC BY 3.0 license [Chan et al., 2021a]. This work was done by Debbie K. Shoemark.

"We used the interactive web application BAlaS to perform Computational Alanine-Scanning mutagenesis (CAS) using BudeAlaScan [Wood et al., 2020] and the BUDE_SM algorithm [Sessions, 2021] for Predictive Saturation Variation Scanning (PreSaVS) [Hetherington et al., 2021]. Both are built on the docking algorithm BUDE, [McIntosh-Smith et al., 2015] which uses a semi-empirical free energy forcefield to calculate binding energies [McIntosh-Smith et al., 2012]. To identify key binding interactions of the natural substrate peptides to M^{pro}, the 11 substrate:M^{pro} complexes were first subjected to CAS using BAlaS. By sequentially substituting for alanine, the energetic contribution of each substrate residue to the overall interaction energy between the singly mutated peptide and M^{pro} is calculated using:

$$\Delta\Delta G = \Delta G_{Ala} - \Delta G_{wt} \tag{B.1}$$

where ΔGwt is the interaction energy between the peptide and M^{pro}, and ΔG_{Ala} is the interaction energy for the peptide with a single alanine mutation at a given position. The more positive the value for each residue, the greater the contribution from that substrate residue to binding. This method was used later to evaluate potential inhibitor peptides.

Having identified residues contributing most to the binding energy of the natural M^{pro} substrates, each of the sequences was subjected to PreSaVS using the BUDE_SM algorithm. This sequentially substitutes each substrate residue with a range of residues (D, E, F, H, I, K, L, M, N, Q, R, S, T, V, W and Y). BUDE_SM calculates the $\Delta\Delta G$ for the binding interaction of each, entire, singly mutated peptide with M^{pro}. Substitutions predicted to improve binding over wildtype sequences have a positive $\Delta\Delta G$. Figure B.2 shows an example of the BUDE_SM PreSaVS results for all the P2 substitutions for the 11 substrate peptides (normally Leu, Phe, or Val in the 11 substrates). The most positive results suggest that Phe, Trp and Tyr favour increased predicted affinity at the P2 position (Figure B.2). However, although Tyr generally increased the predicted binding affinity ($\Delta\Delta G_{sum} = 68.8\ kJ.mol^{-1}$, it was not considered for substitution at P2 due to its negative effect at this position in s11 (scoring -18.9, Figure B.2). Candidate residues for each position, from P6 to P5′, were shortlisted similarly based on those with the best total, and the fewest unfavourable, scores.

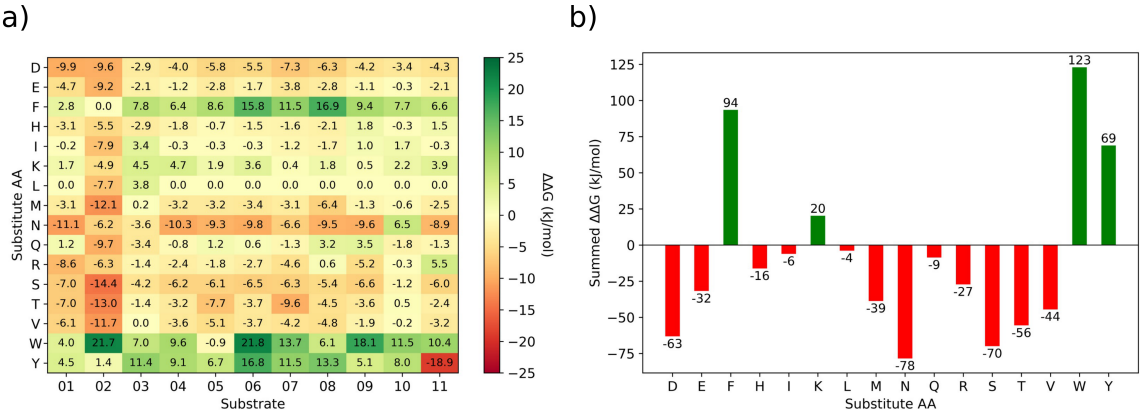

Figure B.2: Figure adapted from the original publication of this work under the CC BY 3.0 license [Chan et al., 2021a]. BUDE_SM PreSaVS results for the P2 position as published in the original publication of this work [Chan et al., 2021a]. a) Heat-map for BUDE_SM PreSaVS saturation mutagenesis at P2, showing the $\Delta\Delta G$ value calculated for each substitution and each M^{pro} substrate. Mutations predicted to improve peptide binding have a positive $\Delta\Delta G$ and are greener; those disfavouring binding are in red. b) The summed $\Delta\Delta G$ values for each residue type substituted at P2. The most positive results suggest that Phe, Trp and Tyr favour increased predicted affinity at the P2 position.

In addition to the computed $\Delta\Delta G$ values, we considered the propensity of each residue to promote an extended conformation. All bound substrates are largely extended, so entropic penalties may be avoided if inherently extended conformations could be favoured in the designed peptide. Thus, the best β-forming (and therefore least α-forming) residues from the first triage were selected (Figure B.3) [Pace and Scholtz, 1998]. We also considered solubility. This was achieved by limiting the number of hydrophobic residues in each designed peptide and ensuring a net positive charge (except p14, which was neutral).

Employing the criteria described above, five new peptides, p12–p16, were designed (Figure B.3 b). Comparison of the computed $\Delta\Delta G$ values for s01–s11 (Figure B.3 c) and p12–p16 (Figure B.3 d) reveals that substitutions at the P sites provide only occasional, moderate improvements to binding energy over the corresponding substrate P sites, with the notable exception of P2, which can accommodate Trp, Phe or Lys. These results agree with the HB analysis, which predicts that the sidechains of residues that are on the N-terminal side of the cleavage site (P sites) contribute more to binding than C-terminal, P$'$ sites. The most striking difference between substrates and designed peptides is in this P$'$ region, where the predicted binding energy contributions for the designed peptides exceed those of the substrates, an advantage that is distributed over most of the designed P$'$ positions.

The final step in design was to assess the relative binding affinities of the substrates and designed peptides. Hence the summed $\Delta\Delta G$s (Figure B.3 e) provide a proxy for the binding energies (BAlaS) [Ibarra et al., 2019] for the substrates and designed peptides with M^{pro}.

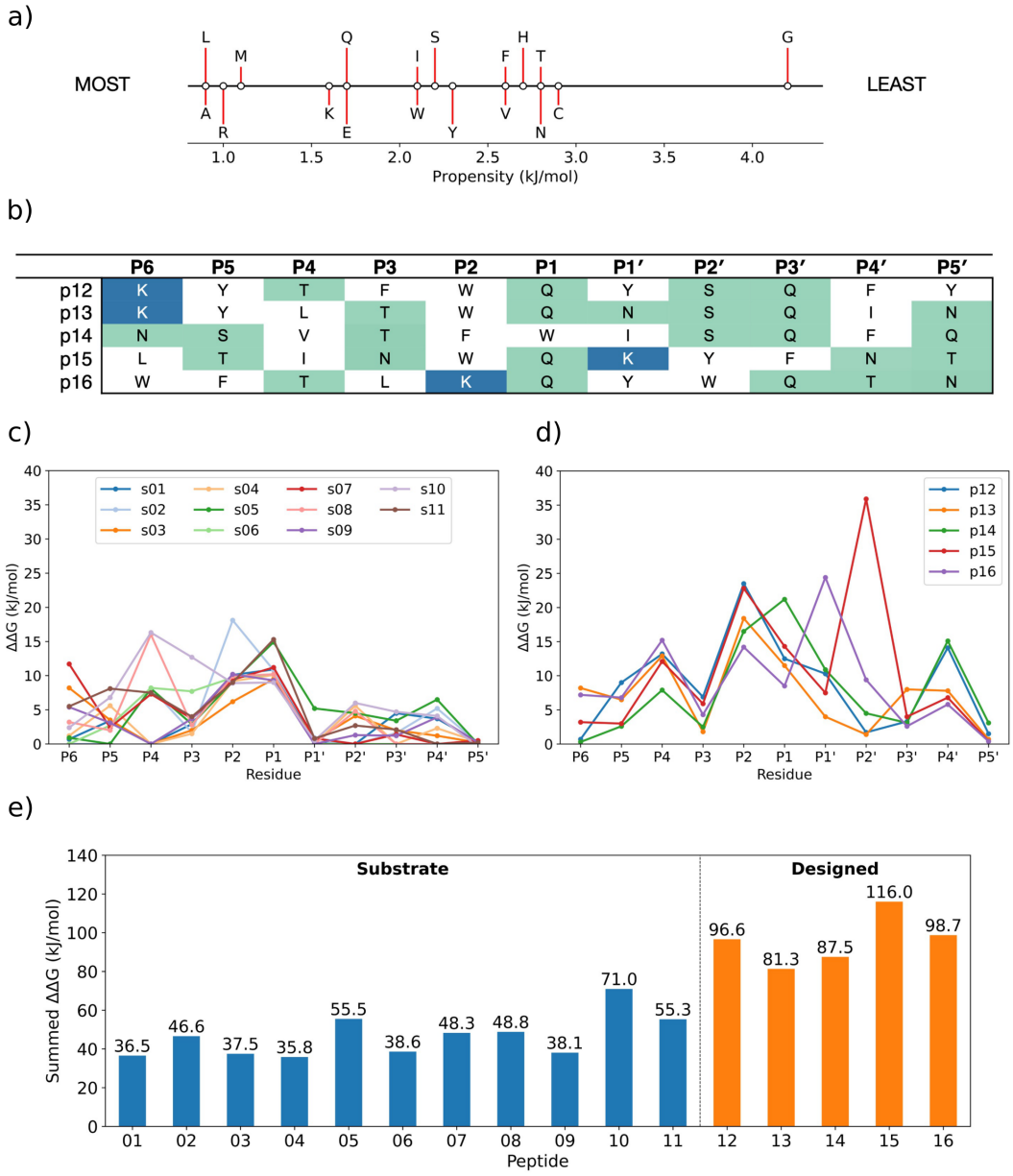

Figure B.3: Figure adapted from the original publication of this work under the CC BY 3.0 license [Chan et al., 2021a]. BAlaS-guided design of tight-binding peptides as published in the original publication of this work [Chan et al., 2021a]. a) Propensity scale of each amino acid to form an α-helical peptide conformation. b) Sequences of designed peptides p12–p16. c,d) Scatter plots with predicted BAlaS $\Delta\Delta G$ values on substitution to alanine for each residue of the 11 M^{pro} natural substrates and the designed peptides, respectively. The more positive the value, the greater the contribution made by the sidechain to the overall binding energy. e) The BAlaS $\Delta\Delta G_{sum}$ comparing values between complexes of M^{pro} with substrate and designed peptides as a proxy for predicting relative binding affinity (larger score = tighter binder). All designed peptides p12-16 were predicted to be tighter binders to M^{pro} than the natural substrates.

The substrate:M^{pro} complexes are stabilised by an average of 46.5 $kJ.mol^{-1}$, whereas the designed-peptide:M^{pro} complexes are predicted to have, in some cases, double the interaction stability of the substrates, with an average of 96.0 $kJ.mol^{-1}$. The full analysis can be found on the GitHub repository of this work (`https://github.com/gmm/SARS-CoV-2-Modelling` file SI_BAlaS_BUDE_SM_12-04-2021.xlsx)."

B.4 Supplementary Results - Synthesis and Experimental Analysis of Designed Peptides

The following description of the results was directly adapted from the original publication with minor adjustments as covered by the CC BY 3.0 license [Chan et al., 2021a]. This work was done by Tika R. Malla and Victor Mikhailov.

"To test the designed sequences, p12, p13, p15 and p16 were synthesised with a carboxyl-amide C-terminus by solid phase synthesis. Their M^{pro} inhibitory activity was determined by dose–response analysis using SPE MS, monitoring both substrate s01 (1191.68 Da) depletion and N-terminally cleaved product (617 Da) formation. Ebselen which reacts multiple times with M^{pro} was used as a standard ($IC_{50} = 0.14 \pm 0.04$ μM; Fig. 10). All four designed pepetides manifested similar potentcy with IC^{50} values ranging from 3.11 μM to 5.36 μM (Figure B.4) with p13 as the most potent inhibitor ($IC_{50} = 3.11 \pm 1.80$ μM).

We probed the inhibition mode of the designed peptides by monitoring changes in IC_{50} while varying the substrate concentration (2 μM, 10 μM, 20 μM and 40 μM TSAVLQ↓SGFRK-NH2 s01; $K_m \approx 14.4$ μM) [Malla et al., 2021]. The results indicated a linear dependency between substrate concentration and IC_{50} values (Figure B.4 a–d). This was not observed with a control 15-mer peptide or ebselen (Figure B.4 e and f).

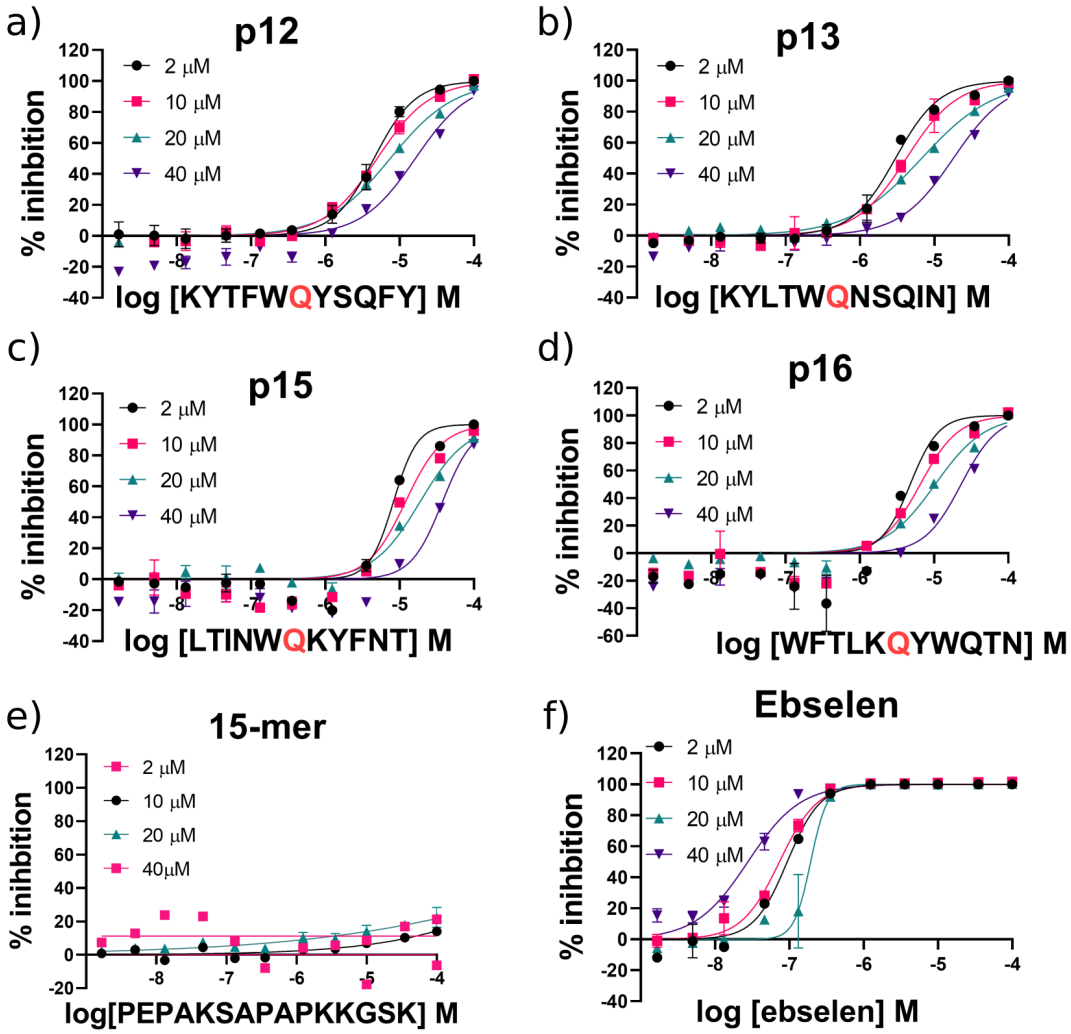

Figure B.4: Figure adapted from the original publication of this work under the CC BY 3.0 license [Chan et al., 2021a]. IC_{50} of designed peptide against M^{pro} with varied substrate concentrations as published in the original publication of this work [Chan et al., 2021a]. IC_{50}s for a) p12 ($IC_{50} = 5.36 \pm 2.17$ μM), b) p13 ($IC_{50} = 3.11 \pm 1.80$ μM), c) p15 ($IC_{50} = 5.31 \pm 1.08$ μM), d) p16 ($IC_{50} = 3.76 \pm 1.19$ μM), e) 15-mer control peptide and f) ebselen ($IC_{50} = 0.14 \pm 0.04$ μM); with 2 μM, 10 μM, 20 μM and 40 μM of substrate peptide s01. IC_{50} values were calculated from technical duplicates. See Appendix Section B.1.5 for assay details. Peptide p13 is the most potent inhibitor ($IC_{50} = 3.11 \pm 1.80$ μM).

Analysis of the data by the procedure of [Wei et al., 2007] implies competitive inhibition. By contrast, the same analysis for ebselen did not support competitive inhibition, consistent with MS studies showing it has a complex mode of inhibition [Malla et al., 2021].

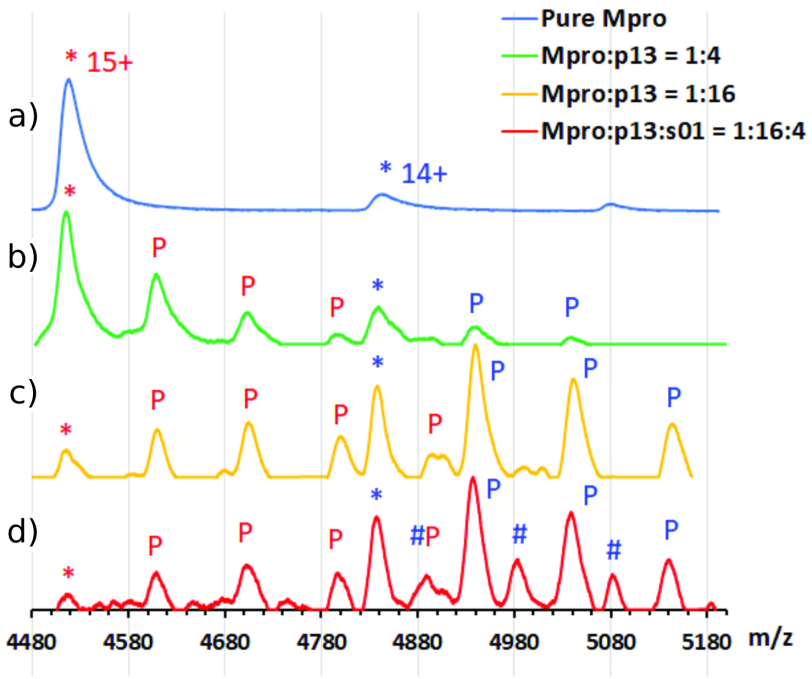

Figure B.5: Figure adapted from the original publication of this work under the CC BY 3.0 license [Chan et al., 2021a]. Non-denaturing MS analysis of designed peptides binding to the M^{pro} dimer as published in the original publication of this work [Chan et al., 2021a]. Inhibitor binding from non-denaturing MS showing normalized intensity in the m/z region around the 14+ and 15+ charge states of the M^{pro} dimer. (*) indicates unbound M^{pro} dimer. (a) 5 μM M^{pro} solution; (b) 4-fold excess of p13 relative to the M^{pro} dimer; 'P' indicates sequential binding of p13 peptides to M^{pro} in the 15+ charge state (red) and 14+ state (blue); (c) 16-fold excess of p13; (d) 16-fold excess of p13 and 4-fold excess of s01; hash (#) indicates sequential binding of s01-cleavage products (note: the resolution is not sufficient to distinguish between the N- and C-terminal fragments; some non-specific binding of p13 is also observed in (c) and (d) due to the high concentration of the peptide).

We then used non-denaturing protein MS to study enzyme–substrate/product/inhibitor complexes simultaneously with turnover. Complexes between M^{pro} dimer and p12 and p13 were observed, together with the uncomplexed M^{pro} dimer in the protein region of the mass spectra. No binding was observed for p15 and p16, due to relatively high noise in that m/z region. None of the designed peptides were cleaved by M^{pro}, as recorded in the peptide region. As a control, s01 was added to the protein/inhibitor mixtures; for all the inhibitors, turnover of s01 was observed after 3 min incubation. Depletion of s01 was 95 %, 91

%, 70 % and 78 % in the presence of p12, p13, p15 and p16, respectively, with an 8-fold excess of inhibitor over M^{pro}, versus > 98 % depletion for the M^{pro}/s01 mixture without the inhibitor. In the protein region of the mass spectra, complexes between M^{pro} dimers and the s01-cleavage products were observed in the presence of p13, but the abundance of these complexes was lower than the abundance of M^{pro}/p13 complexes (Figure B.5). These results validate the above-described evidence that the peptide inhibitors both bind and competitively inhibit M^{pro}."

B.5 Supplementary Results - Substrate and Peptide Inhibitor Contact Analysis

The following results are my own work and have been directly adapted from the original publication with minor adjustments as covered by the CC BY 3.0 license [Chan et al., 2021a].

Using the 3D contact identification tool Arpeggio [Jubb et al., 2017], all 11 natural substrate models and the 5 peptide inhibitor models were analyzed as described in Section 3.3.5. All atomic contacts identified to interact with each residue on the substrates and peptides are shown in Figure B.6 a) and b) respectively. Trends identified on an atomic level as described in Section 3.4.1.3 are visible on an atomic level. Contacts at P1 and P2 are highly conserved between all substrates and to a slightly lesser extent between peptide inhibitors (Figure B.6). However, key contacts such as backbone HBs to Thr-26, Glu-166 or the oxyanion hole at Gly-143 and Cys-145 are conserved on an atomic level.

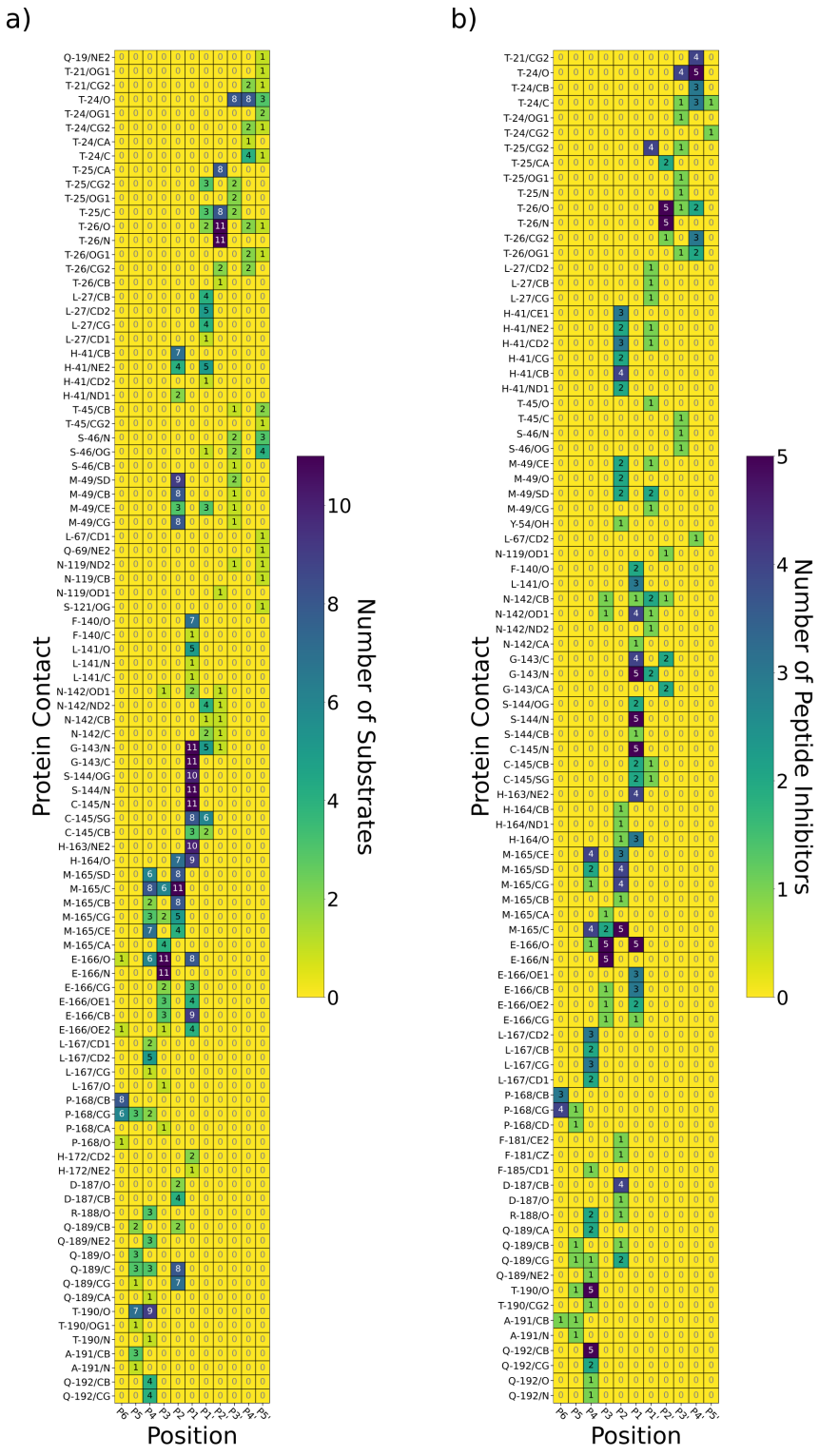

Figure B.6: Figure adapted from the original publication of this work under the CC BY 3.0 license [Chan et al., 2021a]. All contacts made by every residue in a) the 11 natural substrates (s01-s11) and b) the 5 peptide inhibitors (p12-p16) on an atomic level. Even on an atomic level, P1 and P2 are highly conserved.

B.6 Supplementary Results - Fragment-based *In silico* Design of Small Molecule Inhibitors

The following results are my own work and have been directly adapted from the original publication with minor adjustments as covered by the CC BY 3.0 license [Chan et al., 2021a].

B.6.1 Fragment Clustering by Interaction Fingerprints

After contact fingerprint creation using the 91 XChem fragments, the fragments were clustered using different Tanimoto similarity thresholds in steps of 0.1 between 0.1-0.9. The resulting cluster sizes are shown in table B.2.

Threshold	Clusters	Single molecule clusters	Average cluster size
0.1	2	0	33.0
0.2	2	0	33.0
0.3	5	1	13.2
0.4	9	0	7.3
0.5	11	2	6.0
0.6	16	4	4.1
0.7	29	20	2.3
0.8	37	27	1.8
0.9	54	46	1.2

Table B.2: Table adapted from the original publication of this work under the CC BY 3.0 license [Chan et al., 2021a]. Relationship between Tanimoto similarity threshold and cluster sizes for the XChem fragment crystal structures. Only active-site binders were considered.

For this analysis, two clustering thresholds were chosen: a broader (0.5) and a tighter (0.7) threshold. Despite the tighter clustering method producing 29 clusters total, the number of clusters with more than one fragment stays the same between both methods (nine clusters, Table B.2), indicating that tighter clustering above 0.5 does not lead to the addition of meaningful clusters. In addition, despite the lower threshold, the broader (0.5) clustering method is able to create distinct clusters (Top 5 most populated clusters for both thresholds are shown in Figure B.7) and both

methods identify a series of primarily covalent fragments (binding to Cys-145) as the highest populated cluster (all contacts in each cluster shown in Figure B.8).

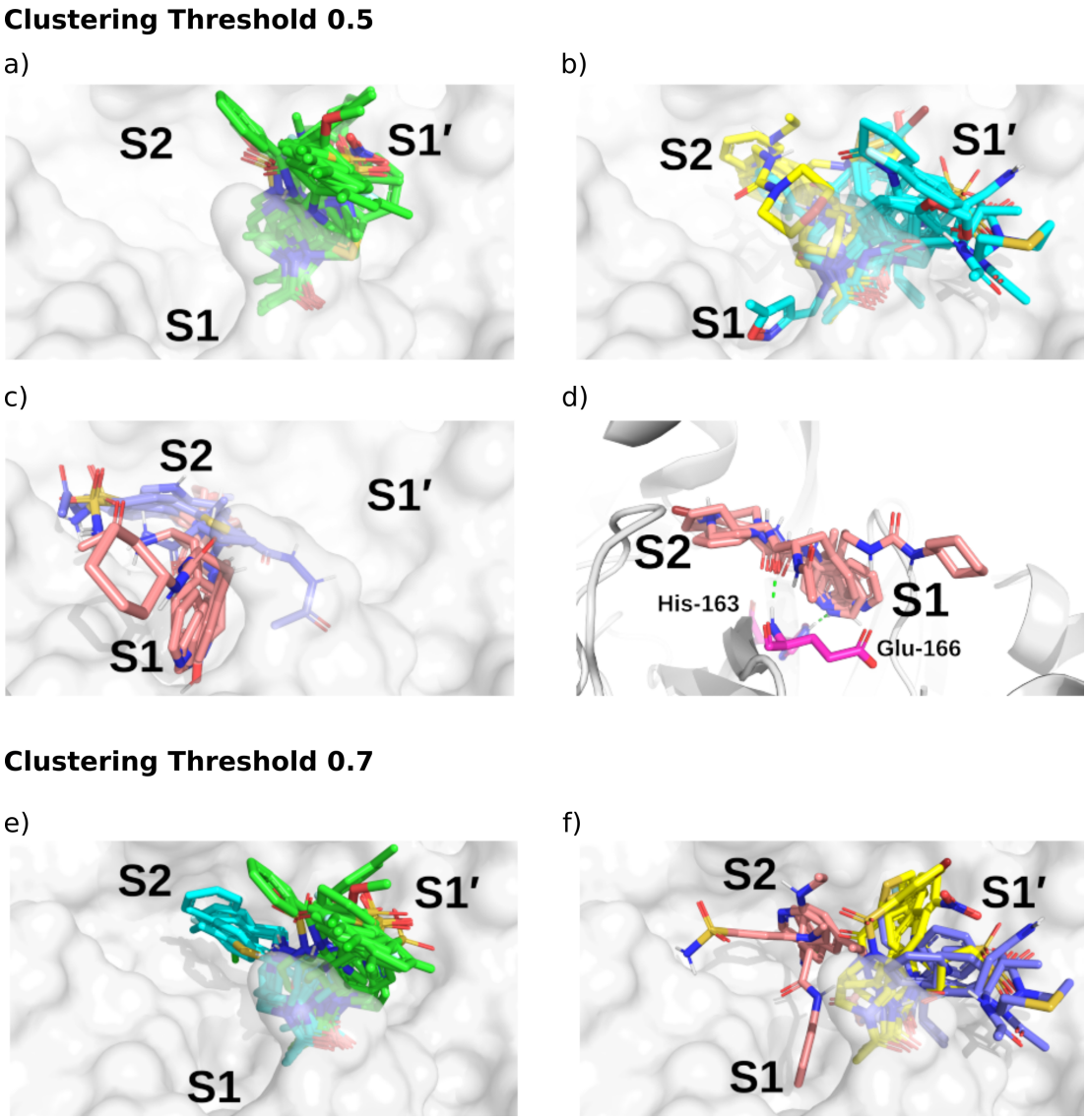

Figure B.7: Figure adapted from the original publication of this work under the CC BY 3.0 license [Chan et al., 2021a]. M^{pro} crystal structure (x0830) in complex with the top 5 most populated clusters using a clustering threshold of 0.5: a) cluster 1 (green); b) clusters 2 (cyan) and 3 (yellow); c) clusters 4 (blue) and 5 (salmon). d) Close-up on the binding pose of cluster 5. Shown in green are the two key HBs between the fragment carbonyl oxygen and the backbone nitrogen of Glu-166 (HB 3 as identified in main text Section 3.4.1.2), and between the His-163 Nϵ and the nitrogen heterocycle of the fragment (HB 6 as identified in main text Section 3.4.1.2). Also shown are the top 5 most populated clusters using a threshold of 0.7: e) clusters 1 (green) and 2 (cyan); f) clusters 3 (yellow), 4 (blue) and 5 (salmon). The clustering threshold of 0.5 produces more meaningful clusters.

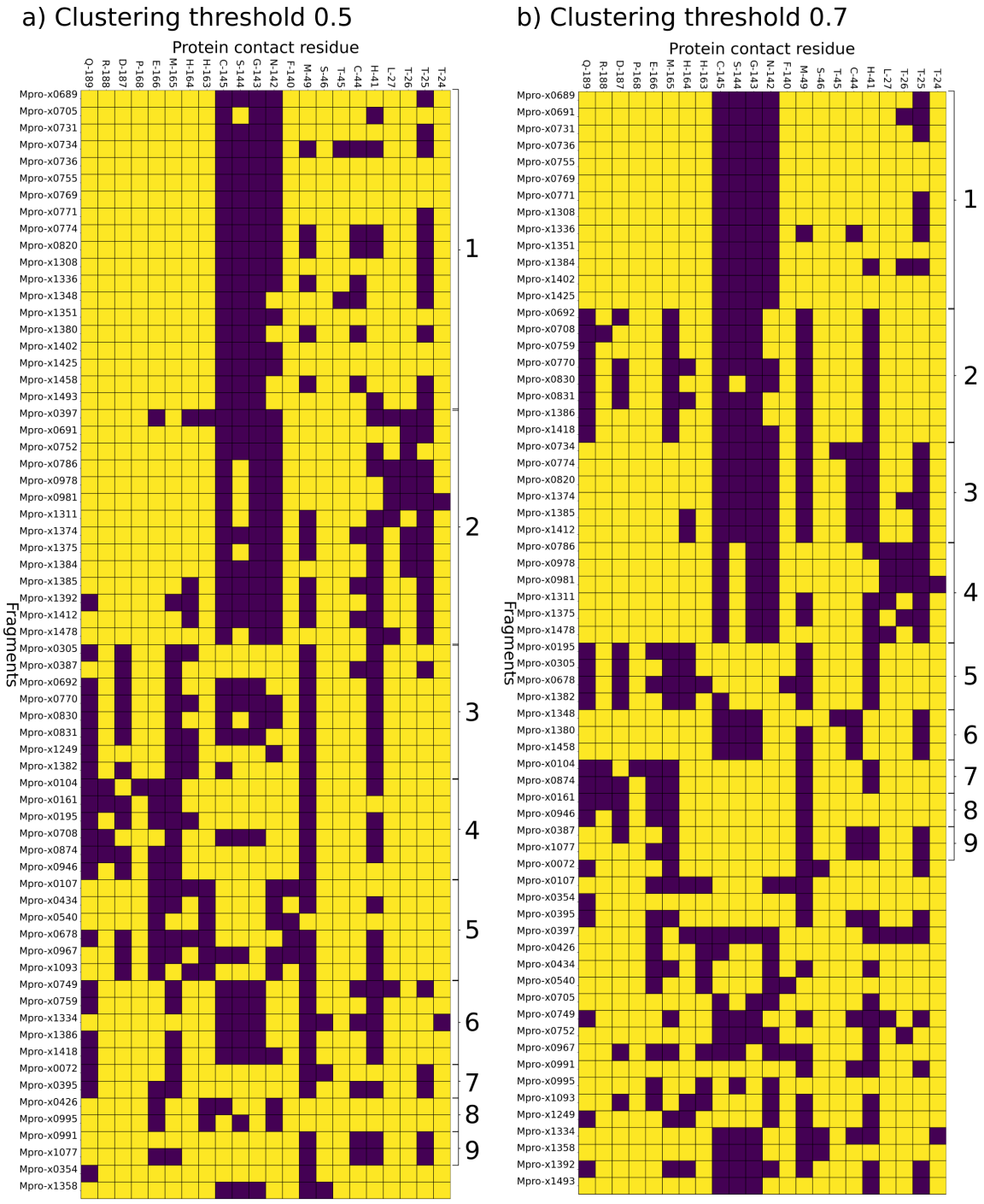

Figure B.8: Figure adapted from the original publication of this work under the CC BY 3.0 license [Chan et al., 2021a]. Contact matrix for the 66 active site XChem fragments sorted by their assigned cluster for a) the broader clustering threshold 0.5 and b) the tighter clustering threshold 0.7. Contacts are shown in purple and no contact in yellow. Clusters with more than 1 molecule are marked by brackets (except x0354 and x1358).

Nonetheless, cluster 5 of the broader (0.5 threshold) clustering method was found

to be the most unique cluster with great potential for inhibitor design, as it was the only one that targets the key HBs 3 and 6 identified to be crucial for both, stability and substrate recognition (see main text Section 3.4.1.2).

B.6.2 Active-Guided Covalent Docking Results

Active-guided covalent docking (AGCD) was performed as described in the main text Section 3.3.4. I compared the shape and pharmacophoric (SuCOS) overlap of the lowest energy pose of the highest populated cluster for each Moonshot compound with the inspiration covalent XChem fragment referenced by the designers (Figure B.9). A SuCOS score of 0.55 and higher is generally considered sufficient to consider the binding poses of the crystallographic fragment and docked design as conserved [Leung et al., 2019]. Due to creative freedom in the design process, some of the designed compounds do not overlap significantly with the inspiration fragments and in some extreme cases only have the covalent warhead in common. When controlling for the smallest maximum common substructure (MCS) that encompasses at least the covalent warhead and one additional atom in the compound, 379 docked designs remain, from which 132 (34.8 %) recovered the binding mode of the inspiration fragment. Given the high similarity between the fragments and the docked designed compounds, it is likely that these binding modes are more representative of the actual binding mode of the ligand. Furthermore, by selecting docked compounds that did not have significant modifications to the original fragment (less than 10 atoms difference to the MCS between fragment and docked compounds), the number of compounds regaining the binding pose was 87 (54.7 %) of the 159 compounds. Interestingly, the distribution of SuCOS over the data subsets resembles a bimodal distribution in all cases, with one peak around a SuCOS of 0.25 and a second peak between 0.7-0.8 depending on the subset (Figure B.9). This suggests that the AD4

docking process performs generally well in identifying binding poses similar to the original fragment (especially in cases where few modifications were made), but fails completely in some cases, resulting in almost no overlap between the pose and the fragment and an extremely low SuCOS (< 0.3). This observation is in line with the expected changes in binding mode of a molecule when subjected to large changes in structure and does not necessarily correspond to incorrect docked poses. As a result, when filtering for compounds where only minor changes were made to the inspiration fragment and the docking pose generated regained the same binding pose as the fragment, the probability of gaining relevant poses is increased. However, thorough validation of this hypothesis is yet to be done.

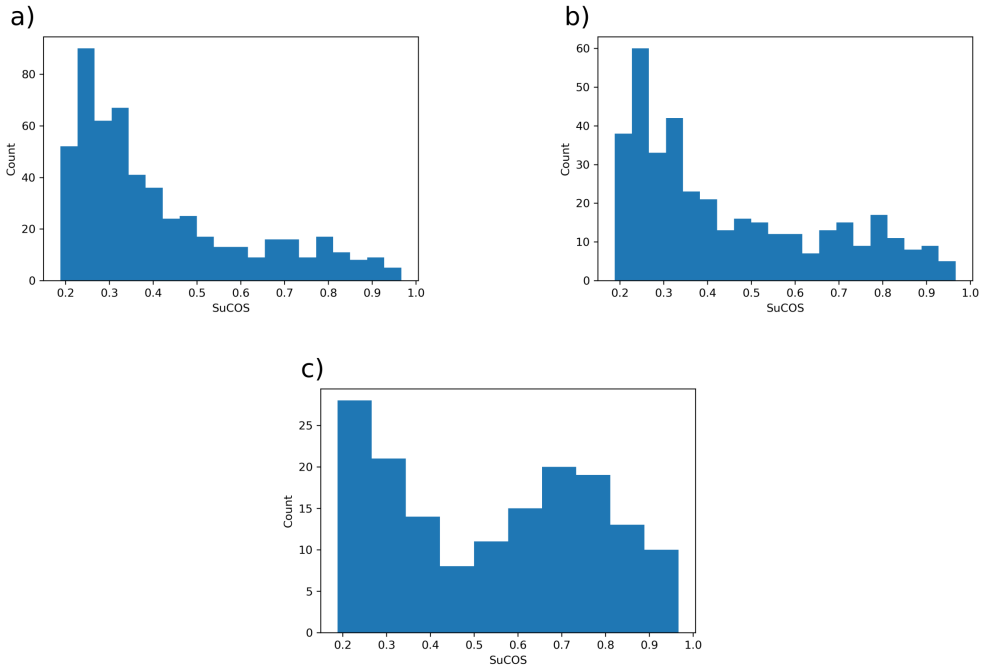

Figure B.9: Figure adapted from the original publication of this work under the CC BY 3.0 license [Chan et al., 2021a]. Distribution of SuCOS between the docked pose of the covalent Moonshot design and the original covalent fragment used as a basis for design. a) SuCOS scores of the 540 docked compounds; b) SuCOS scores of the 379 docked compounds with significant MCS overlap (i.e. more than 8 atoms MCS match) to the fragment; c) SuCOS scores of the 159 docked compounds with significant MCS overlap to the fragment and only small changes to its structure (< 10 atoms difference between the compound and its MCS with the inspiration fragment)

B.6.3 Docking Pose of Moonshot Design x10899

The docked pose of Fragment x10899 found to bind via a crystal contact to a third symmetry-related Mpro molecule, rather than the biologically relevant dimeric state (Figure B.10).

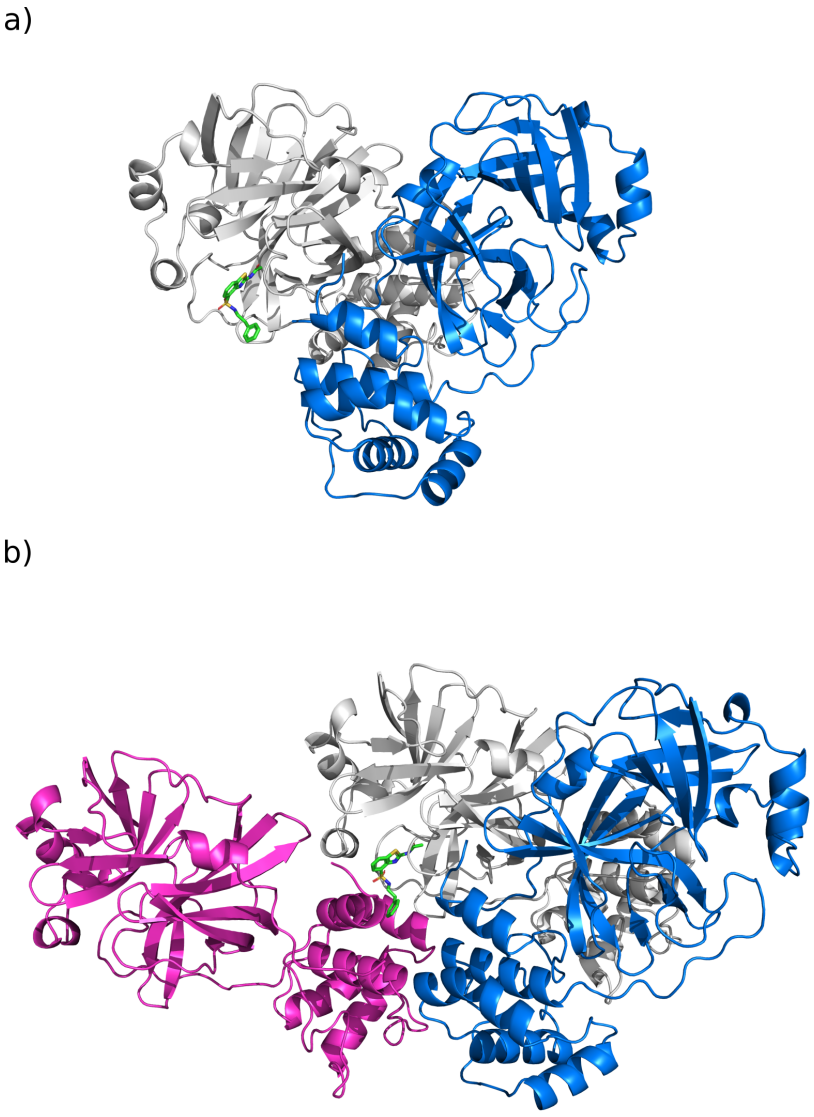

Figure B.10: Figure adapted from the original publication of this work under the CC BY 3.0 license [Chan et al., 2021a]. a) View from the crystal structure of x10899 (ligand in green) from the perspective of the biologically relevant dimer. b) View of the crystal structure of x10899 with one additional crystal packing symmetry mate. Chains A, B (dimer) and C (crystal-contacting chain) are coloured white, blue and pink, respectively. The small molecule X10899 is bound in the active site of chain A (white); but its terminal aromatic sidechain interacts with the symmetry-related chain C (magenta), leading to what would appear to be an unusual binding mode if only the dimer were considered (top).

B.6.4 Implications for Future Inhibitor Design

B.6.4.1 Potency of Known Cluster 5 Moonshot Designs

Moonshot designs that are confirmed to bind into the oxyanion hole (Figure B.11. Data collected from the COVID MOonshot Project GitHub, accessed January 2021 [COVID-19 Moonshot project, 2020].

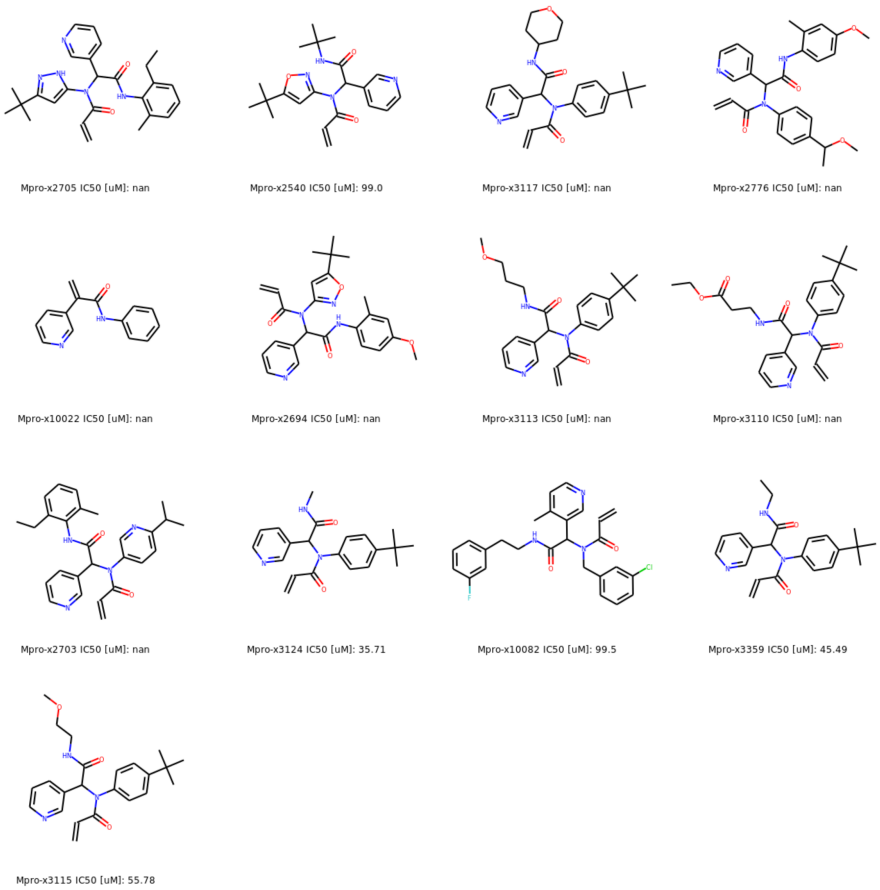

Figure B.11: Figure adapted from the original publication of this work under the CC BY 3.0 license [Chan et al., 2021a]. Selection of all Moonshot compounds in cluster 5 that bind in the M^{pro} oxyanion hole. All compounds are covalent inhibitors, reacting with Cys-145 via the acrylamide warhead. IC_{50} values are obtained from the postera.ai GitHub page [COVID-19 Moonshot project, 2020]. Note: "nan" indicates that the compound has not been assayed.

B.6.4.2 Compound Elaboration for MIH-UNI-e573136b-3

Compound MIH-UNI-e573136b-3 could be grown into the oxyanion hole to increase its potency (Figure B.12).

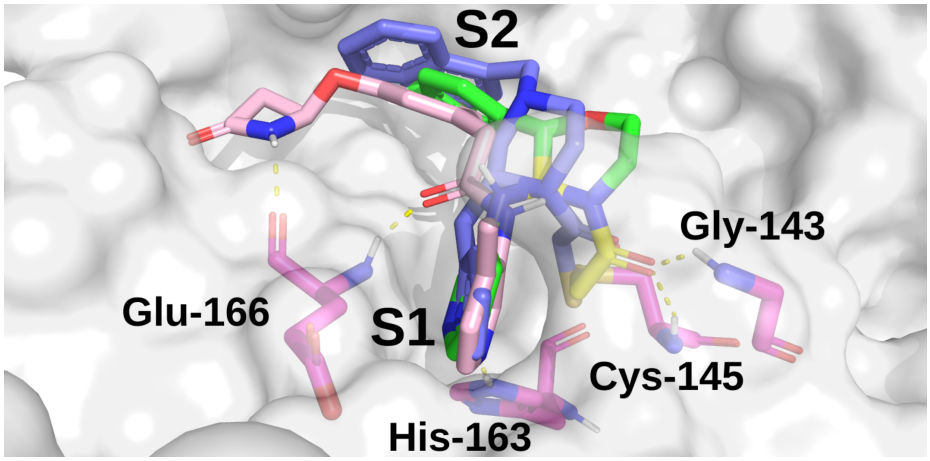

Figure B.12: Figure adapted from the original publication of this work under the CC BY 3.0 license [Chan et al., 2021a]. Overlay of the docked pose of FOC-CAS-e3a94da8-1 (green), the docked pose of MIH-UNI-e573136b-3 (blue), and a crystallographically observed binding mode of x10789 (salmon) with Mpro (PDB: 5RER) [Douangamath et al., 2020]. The proposed expansion of x10789 into the oxyanion hole is shown in yellow on compound FOC-CAS-e3a94da8-1.

B.6.4.3 Docked and Crystal Pose of Nirmatrelvir

I covalently docked Nirmatrelvir into the active site of M^{pro} structure of 6XHM [Hoffman et al., 2020], which is a crystal structure in complex with ligand PF-00835231, a precursor in the design of Nirmatrelvir [Hoffman et al., 2020; Owen et al., 2021]. The team at Pfizer deposited a high resolution crystal structure of Nirmatrelvir in complex with M^{pro} in the PDB in November 2021 (7RFS, 1.91 Å) [Owen et al., 2021]. An overlay of the crystal and docked pose can be found in Figure B.13 b. Around the S1 site and the covelent attachment point to M^{pro} Cys-145, the docked pose closely resembles the crystal structure. However, it deviates around the S2 and S3 site (Figure B.13 b). The plasticity analysis performed on 333 M^{pro} holo structures (Section 3.4.1.5), supports this results, since it shows that the S2 and S3 sites are highly plastic

and can change strongly upon ligand binding, making rigid docking in this pocket challenging. Although the binding pose is slightly shifted, the key interactions between Nirmatrelvir and M^{pro} were retained between the crystal and docked pose (HB with Glu-166 and His-163, and hydrophobic interactions in S2, Figure B.13 & Figure 3.20).

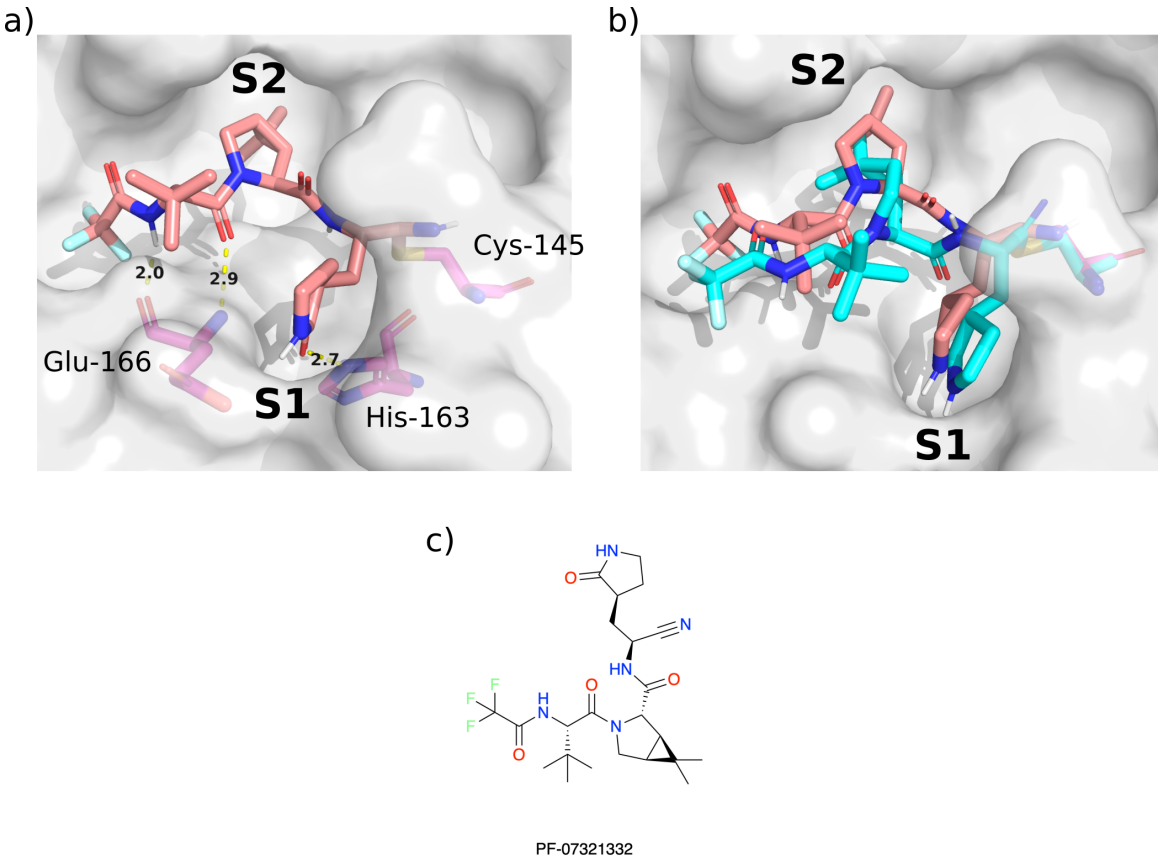

Figure B.13: a) Crystal pose of Nirmatrelvir (salmon), covalently attached to M^{pro} (7RFS; 1.91 Å resolution) [Owen et al., 2021]. Key M^{pro} residues are shown in purple and key hydrogen bonds marked in yellow dotted lines with their distance labelled in Å. The crystal structure of Nirmatrelvir adopts the same major interactions as previously identified in cluster 5 (Figure 3.16) and for the natural substrate models (Section 3.4.1, namely the double HB to the backbone of Glu-166, the HB to His-163 in the S1 subsite, and a series of hydrophobic interactions in the S2 subsite. b) Overlay of the docked (cyan) and crystal (salmon) pose of Nirmatrelvir on the surface of M^{pro} (7RFS) [Owen et al., 2021]. c) 2D structure of Nirmatrelvir.

Appendix C

Protein-Ligand Interaction Graphs: Learning from Ligand-Shaped 3D Interaction Graphs to Improve Binding Affinity Prediction

C.1 Model Hyperparameter Tuning

C.1.1 Hyperparameter Optimization Setup

The hyperparameters were optimized using Optuna 2.8.0 [Akiba et al., 2019]. Optimization was performed on the PDBbind refined set 2016 [Liu et al., 2015, 2017], with a random train/validation split resulting in a training dataset of 3229 and a validation dataset of 359 compounds. The CASF-2016 test set that was used as a benchmark set was removed from the PDBBind refined set 2016 prior to hyperparameter optimization. Since each specific architecture has its own hyperparameters, some changes had to be made between models for optimization. Some hyperparameters appear in every model:

- The learning rate was optimized based on a loguniform prior on $[1e-4, 1e-2]$.
- The activation function in the MLP/GNN layers was optimized with categorical values
  $\{ReLU, LeakyReLU, Sigmoid\}$.

- The dropout rate for the MLP/GNN layers was optimized based on a uniform prior on $[0, 0.9]$.
- The dropout rate for the readout layers was optimized based on a uniform prior on $[0, 0.9]$.
- The number of MLP/GNN layers was optimized with possible values $\{1, ..., 10\}$.

The readout layers were set to a fixed size as they only serve to combine the features extracted in previous layers. On the protein branch side (if included in the model) the following parameters were optimized:

- The number of filters in the 1D-convolutional layer was optimized with possible values $\{2^2, ..., 2^5\}$.
- The size of the filters in the 1D-convolutional layer was optimized with possible values $\{2^2, ..., 2^5\}$.

The idea behind optimizing these hyperparameters is that the protein branch should not be a bottleneck for the model performance and hence some flexibility should be allowed. In the GNN/MLP branch hyperparameters which were optimized are:

- The number of MLP/GNN layers (except for SGCs) was optimized with possible values $\{1, ..., 10\}$ for GNN models and $\{0, ..., 10\}$ for MLP layers. As the N+1$^{\text{st}}$ layer is a fully connected layer $N = 0$ is possible for fingerprint-based models.
- The dimension of each MLP (width of layer) / GNN layer (length of node features) as a multiple of the input dimension was optimized with possible values values $\{F^{(0)} \cdot 2^0, F^{(0)} \cdot 2^1, F^{(0)} \cdot 2^2\}$ where $F^{(0)}$ is the input feature dimension, e.g. $F^{(0)} = 40$ for LBGs and $F^{(0)} = 512$ for the ECFP-512 model.

SGCs are restricted to one GNN layer as their idea is to replace N GCN layers by one SGC layer [Wu et al., 2019]. Model-specific hyperparameters which are optimized are:

- the power of the adjacency matrix K of SGCs with possible values $\{1, ..., 10\}$,
- the N possibly different dimensions of the GAT heads for GATs with possible values $\{1, ..., 10\}$,
- the N possibly different dimensions of the GAT heads for GAT-GCNs with possible values $\{1, ..., 10\}$,
- the initial ϵ-value (uniform on $[0, 2]$) for GINs and the MLP activation function of the MLP which is part of a GIN layer among $\{ReLU, LeakyReLU, Sigmoid\}$.

Each hyperparameter optimization comprised 1000 trials with values suggested by the TPE (Tree-structured Parzen Estimator) algorithm which is based on Gaussian Mixture Models [Bergstra et al., 2011]. The results for each model can be found below.

C.1.2 Hyperparameter Optimization Results

PB included corresponds to the combined 2 branch setup where the protein sequence branch is included. PB Excluded corresponds to the GNN or MLPNet model without protein sequence embedding. The MLPNet ECIF models were not tuned independently and instead utilized the MLPNet ECFP 1024 bit model parameters for the MLPNet ECIF models. The MLPNet ECIF implementation was never intended to get the best possible performance (as [Sánchez-Cruz et al., 2020] have already optimized the performance using random forests) but rather serve as a quick comparison to the new PLIG models as they use similar ECIF style atom representations.

Hyperparameter	**PB Included**	**PB Excluded**
Learning rate	0.0010	0.0006
Activation function	Leaky ReLU	Leaky ReLU
Number of filters protein convolution	2^4	–
Kernel size of protein convolution	2^3	–
Dropout rate GNN layers	0.033	0.000
Dropout rate readout layers	0.028	0.065
Number of GNN layers	4	3
Dimension of GNN layer 1	$40 \cdot 2^2$	$40 \cdot 2^2$
Dimension of GNN layer 2	$40 \cdot 2^2$	$40 \cdot 2^2$
Dimension of GNN layer 3	$40 \cdot 2^2$	$40 \cdot 2^2$
Dimension of GNN layer 4	$40 \cdot 2^2$	–
Pearson correlation on validation set	0.759	0.729

Table C.1: Results of hyperparameter tuning of GCN with ligand-based graphs.

Hyperparameter	**PB Included**	**PB Excluded**
Learning rate	0.0008	0.0006
Activation function	Leaky ReLU	ReLU
Number of filters protein convolution	2^4	–
Kernel size of protein convolution	2^3	–
Dropout rate GNN layers	0.021	0.110
Dropout rate readout layers	0.021	0.080
Dimension of GNN layer 1	$40 \cdot 2^0$	$40 \cdot 2^2$
K	10	10
Pearson correlation on validation set	0.755	0.735

Table C.2: Results of hyperparameter tuning of SGC with ligand-based graphs.

Hyperparameter	PB Included	PB Excluded
Learning rate	0.0008	0.0007
Activation function	Leaky ReLU	Leaky ReLU
Number of filters protein convolution	2^4	–
Kernel size of protein convolution	2^4	–
Dropout rate GNN layers	0.016	0.029
Dropout rate readout layers	0.013	0.013
Number of GNN layers	2	3
Dimension of GNN layer 1	$40 \cdot 2^2$	$40 \cdot 2^0$
Dimension of GAT head 1	9	10
Dimension of GNN layer 2	$40 \cdot 2^1$	$40 \cdot 2^2$
Dimension of GAT head 2	1	5
Dimension of GNN layer 3	–	$40 \cdot 2^2$
Dimension of GAT head 3	–	6
Pearson correlation on validation set	0.753	0.745

Table C.3: Results of hyperparameter tuning of GAT with ligand-based graphs.

Hyperparameter	PB Included	PB Excluded
Learning rate	0.0003	0.0009
Activation function	Leaky ReLU	ReLU
Number of filters protein convolution	2^4	–
Kernel size of protein convolution	2^5	–
Dropout rate GNN layers	0.000	0.000
Dropout rate readout layers	0.001	0.090
Number of GNN layers	2	1
Dimension of GNN layer 1	$40 \cdot 2^1$	$40 \cdot 2^1$
Dimension of GAT head 1	5	2
Dimension of GNN layer 1	$40 \cdot 2^1$	–
Dimension of GAT head 1	10	–
Pearson correlation on validation set	0.761	0.732

Table C.4: Results of hyperparameter tuning of GAT-GCN with ligand-based graphs.

Hyperparameter	**PB Included**	**PB Excluded**
Learning rate	0.0005	0.0003
Activation function	ReLU	ReLU
Number of filters protein convolution	2^4	–
Kernel size of protein convolution	2^3	–
Dropout rate GNN layers	0.001	0.017
Dropout rate readout layers	0.100	0.050
Number of GNN layers	2	3
Dimension of GNN layer 1	$40 \cdot 2^1$	$40 \cdot 2^2$
Dimension of GNN layer 2	$40 \cdot 2^2$	$40 \cdot 2^1$
Dimension of GNN layer 3	–	$40 \cdot 2^1$
Initial ϵ	0.369	0.070
GIN MLP activation function	Leaky ReLU	Leaky ReLU
Pearson correlation on validation set	0.754	0.721

Table C.5: Results of hyperparameter tuning of GIN with ligand-based graphs.

Hyperparameter	**PB Included**	**PB Excluded**
Learning rate	0.0003	0.0010
Activation function	Leaky ReLU	Leaky ReLU
Number of filters protein convolution	2^5	–
Kernel size of protein convolution	2^5	–
Dropout rate GNN layers	0.000	0.019
Dropout rate readout layers	0.016	0.147
Number of GNN layers	6	2
Dimension of GNN layer 1	$40 \cdot 2^1$	$40 \cdot 2^2$
Dimension of GNN layer 2	$40 \cdot 2^0$	$40 \cdot 2^0$
Dimension of GNN layer 3	$40 \cdot 2^1$	–
Dimension of GNN layer 4	$40 \cdot 2^0$	–
Dimension of GNN layer 5	$40 \cdot 2^2$	–
Dimension of GNN layer 6	$40 \cdot 2^2$	–
Pearson correlation on validation set	0.755	0.734

Table C.6: Results of hyperparameter tuning of SAGE with ligand-based graphs.

Hyperparameter	PB Included	PB Excluded
Learning rate	0.0003	0.0001
Activation function	ReLU	Leaky ReLU
Number of filters protein convolution	2^3	–
Kernel size of protein convolution	2^5	–
Dropout rate MLP layers	0.021	0.000
Dropout rate readout layers	0.106	0.169
Number of MLP layers	1	6
Dimension of MLP layer 1	$512 \cdot 2^1$	$512 \cdot 2^2$
Dimension of MLP layer 2	–	$512 \cdot 2^0$
Dimension of MLP layer 3	–	$512 \cdot 2^1$
Dimension of MLP layer 4	–	$512 \cdot 2^0$
Dimension of MLP layer 5	–	$512 \cdot 2^2$
Dimension of MLP layer 6	–	$512 \cdot 2^2$
Pearson correlation on validation set	0.781	0.767

Table C.7: Results of hyperparameter tuning of MLP with ECFP-512 fingerprints.

Hyperparameter	PB Included	PB Excluded
Learning rate	0.0001	0.0002
Activation function	Leaky ReLU	ReLU
Number of filters protein convolution	2^5	–
Kernel size of protein convolution	2^5	–
Dropout rate MLP layers	0.019	0.099
Dropout rate readout layers	0.719	0.232
Number of MLP layers	8	4
Dimension of MLP layer 1	$1024 \cdot 2^1$	$1024 \cdot 2^0$
Dimension of MLP layer 2	$1024 \cdot 2^0$	$1024 \cdot 2^0$
Dimension of MLP layer 3	$1024 \cdot 2^1$	$1024 \cdot 2^0$
Dimension of MLP layer 4	$1024 \cdot 2^2$	$1024 \cdot 2^1$
Dimension of MLP layer 5	$1024 \cdot 2^1$	–
Dimension of MLP layer 6	$1024 \cdot 2^1$	–
Dimension of MLP layer 7	$1024 \cdot 2^1$	–
Dimension of MLP layer 8	$1024 \cdot 2^2$	–
Pearson correlation on validation set	0.787	0.767

Table C.8: Results of hyperparameter tuning of MLP with ECFP-1024 fingerprints.

Hyperparameter	PB Included	PB Excluded
Learning rate	0.0002	0.0001
Activation function	ReLU	Leaky ReLU
Number of filters protein convolution	2^4	–
Kernel size of protein convolution	2^5	–
Dropout rate MLP layers	0.075	0.068
Dropout rate readout layers	0.726	0.458
Number of MLP layers	7	8
Dimension of MLP layer 1	$512 \cdot 2^0$	$512 \cdot 2^0$
Dimension of MLP layer 2	$512 \cdot 2^2$	$512 \cdot 2^1$
Dimension of MLP layer 3	$512 \cdot 2^1$	$512 \cdot 2^2$
Dimension of MLP layer 4	$512 \cdot 2^2$	$512 \cdot 2^1$
Dimension of MLP layer 5	$512 \cdot 2^0$	$512 \cdot 2^0$
Dimension of MLP layer 6	$512 \cdot 2^2$	$512 \cdot 2^0$
Dimension of MLP layer 7	$512 \cdot 2^1$	$512 \cdot 2^1$
Dimension of MLP layer 8	–	$512 \cdot 2^1$
Pearson correlation on validation set	0.775	0.761

Table C.9: Results of hyperparameter tuning of MLP with FCFP-512 fingerprints.

Hyperparameter	PB Included	PB Excluded
Learning rate	0.0001	0.0001
Activation function	Leaky ReLU	ReLU
Number of filters protein convolution	2^5	–
Kernel size of protein convolution	2^4	–
Dropout rate MLP layers	0.134	0.034
Dropout rate readout layers	0.200[1]	0.401
Number of MLP layers	4	6
Dimension of MLP layer 1	$1024 \cdot 2^1$	$1024 \cdot 2^1$
Dimension of MLP layer 2	$1024 \cdot 2^1$	$1024 \cdot 2^1$
Dimension of MLP layer 3	$1024 \cdot 2^2$	$1024 \cdot 2^2$
Dimension of MLP layer 4	$1024 \cdot 2^2$	$1024 \cdot 2^0$
Dimension of MLP layer 5	–	$1024 \cdot 2^0$
Dimension of MLP layer 6	–	$1024 \cdot 2^0$
Pearson correlation on validation set	0.774	0.755

Table C.10: Results of hyperparameter tuning of MLP with FCFP-1024 fingerprints.
[1]Result of hyperparameter optimization is 0.842 but since this results in bad performances (especially in combination with early stopping of 5) the value was set to 0.200.

Hyperparameter	PB Included	PB Excluded
Learning rate	0.0010	0.0007
Activation function	Leaky ReLU	Leaky ReLU
Number of filters protein convolution	2^5	–
Kernel size of protein convolution	2^3	–
Dropout rate GNN layers	0.001	0.038
Dropout rate readout layers	0.017	0.058
Number of GNN layers	5	3
Dimension of GNN layer 1	$27 \cdot 2^1$	$27 \cdot 2^2$
Dimension of GNN layer 2	$27 \cdot 2^1$	$27 \cdot 2^2$
Dimension of GNN layer 3	$27 \cdot 2^2$	$27 \cdot 2^0$
Dimension of GNN layer 4	$27 \cdot 2^2$	–
Dimension of GNN layer 5	$27 \cdot 2^0$	–
Pearson correlation on validation set	0.748	0.772

Table C.11: Results of hyperparameter tuning of GCN with PLIGs.

Hyperparameter	PB Included	PB Excluded
Learning rate	0.0006	0.0005
Activation function	Leaky ReLU	ReLU
Number of filters protein convolution	2^3	–
Kernel size of protein convolution	2^3	–
Dropout rate GNN layers	0.080	0.058
Dropout rate readout layers	0.001	0.066
Dimension of GNN layer 1	$27 \cdot 2^2$	$27 \cdot 2^2$
K	2	8
Pearson correlation on validation set	0.761	0.765

Table C.12: Results of hyperparameter tuning of SGC with PLIGs.

Hyperparameter	PB Included	PB Excluded
Learning rate	0.0001	0.0008
Activation function	ReLU	ReLU
Number of filters protein convolution	2^2	–
Kernel size of protein convolution	2^2	–
Dropout rate GNN layers	0.001	0.012
Dropout rate readout layers	0.034	0.020
Number of GNN layers	3	3
Dimension of GNN layer 1	$27 \cdot 2^1$	$27 \cdot 2^1$
Dimension of GAT head 1	5	6
Dimension of GNN layer 2	$27 \cdot 2^2$	$27 \cdot 2^2$
Dimension of GAT head 2	6	5
Dimension of GNN layer 3	$27 \cdot 2^1$	$27 \cdot 2^2$
Dimension of GAT head 3	9	8
Pearson correlation on validation set	0.786	0.787

Table C.13: Results of hyperparameter tuning of GAT with PLIGs.

Hyperparameter	PB Included	PB Excluded
Learning rate	0.0008	0.0002
Activation function	ReLU	Leaky ReLU
Number of filters protein convolution	2^4	–
Kernel size of protein convolution	2^5	–
Dropout rate GNN layers	0.004	0.000
Dropout rate readout layers	0.023	0.073
Number of GNN layers	1	2
Dimension of GNN layer 1	$27 \cdot 2^2$	$27 \cdot 2^2$
Dimension of GAT head 1	6	8
Dimension of GNN layer 2	–	$27 \cdot 2^2$
Dimension of GAT head 2	–	9
Pearson correlation on validation set	0.761	0.776

Table C.14: Results of hyperparameter tuning of GAT-GCN with PLIGs.

Hyperparameter	PB Included	PB Excluded
Learning rate	0.0007	0.0012
Activation function	ReLU	Leaky ReLU
Number of filters protein convolution	2^2	–
Kernel size of protein convolution	2^5	–
Dropout rate GNN layers	0.000	0.000
Dropout rate readout layers	0.014	0.307
Number of GNN layers	1	1
Dimension of GNN layer 1	$27 \cdot 2^2$	$27 \cdot 2^2$
Initial ϵ	1.583	1.509
GIN MLP activation function	sigmoid	sigmoid
Pearson correlation on validation set	0.761	0.769

Table C.15: Results of hyperparameter tuning of GIN with PLIGs.

Hyperparameter	PB Included	PB Excluded
Learning rate	0.0007	0.0003
Activation function	ReLU	Leaky ReLU
Number of filters protein convolution	2^5	–
Kernel size of protein convolution	2^3	–
Dropout rate GNN layers	0.000	0.020
Dropout rate readout layers	0.018	0.253
Number of GNN layers	3	5
Dimension of GNN layer 1	$27 \cdot 2^2$	$27 \cdot 2^1$
Dimension of GNN layer 2	$27 \cdot 2^2$	$27 \cdot 2^1$
Dimension of GNN layer 3	$27 \cdot 2^2$	$27 \cdot 2^2$
Dimension of GNN layer 4	–	$27 \cdot 2^2$
Dimension of GNN layer 5	–	$27 \cdot 2^2$
Pearson correlation on validation set	0.768	0.762

Table C.16: Results of hyperparameter tuning of SAGE with PLIGs.

C.2 Protein Atom Types

The 22 ECIF protein atom types were identified based on the identifiers outlined by Sánchez-Cruz et al. [2020] from the PDBbind 2020+2016 combined dataset. An atom is defined through the following parameters: atom symbol, explicit valence, number of attached heavy atoms, number of attached hydrogens, aromaticity and ring membership. The possible protein atom types based on the naturally occurring amino acids utilized in PLIG identified based on those parameters is shown in Figure

C.1. Out of all possible atom types based on the amino acids only 22 unique atom types are identified. The unique atom types are shown in Table C.17. Those atom types form the dimensionality of the PLIG contact vector of the node features.

	Protein Atom	ECIF Atom Type
1	ALA-C	C;4;3;0;0;0
2	ALA-CA	C;4;3;1;0;0
3	ALA-CB	C;4;1;3;0;0
4	ALA-N	N;3;2;1;0;0
5	ALA-O	O;2;1;0;0;0
6	ALA-OXT	O;2;1;0;0;0
7	ARG-C	C;4;3;0;0;0
8	ARG-CA	C;4;3;1;0;0
9	ARG-CB	C;4;2;2;0;0
10	ARG-CD	C;4;2;2;0;0
11	ARG-CG	C;4;2;2;0;0
12	ARG-CZ	C;6;3;0;0;0
13	ARG-N	N;3;2;1;0;0
14	ARG-NE	N;4;2;1;0;0
15	ARG-NH1	N;4;1;2;0;0
16	ARG-NH2	N;4;1;2;0;0
17	ARG-O	O;2;1;0;0;0
18	ARG-OXT	O;2;1;0;0;0
19	ASN-C	C;4;3;0;0;0
20	ASN-CA	C;4;3;1;0;0
21	ASN-CB	C;4;2;2;0;0
22	ASN-CG	C;4;3;0;0;0
23	ASN-N	N;3;2;1;0;0
24	ASN-ND2	N;3;1;2;0;0
25	ASN-O	O;2;1;0;0;0
26	ASN-OD1	O;2;1;0;0;0
27	ASN-OXT	O;2;1;0;0;0
28	ASP-C	C;4;3;0;0;0
29	ASP-CA	C;4;3;1;0;0
30	ASP-CB	C;4;2;2;0;0
31	ASP-CG	C;5;3;0;0;0
32	ASP-N	N;3;2;1;0;0
33	ASP-O	O;2;1;0;0;0
34	ASP-OD1	O;2;1;0;0;0
35	ASP-OD2	O;2;1;0;0;0
36	ASP-OXT	O;2;1;0;0;0
37	CYS-C	C;4;3;0;0;0
38	CYS-CA	C;4;3;1;0;0
39	CYS-CB	C;4;2;2;0;0
40	CYS-N	N;3;2;1;0;0
41	CYS-O	O;2;1;0;0;0
42	CYS-OXT	O;2;1;0;0;0
43	CYS-SG	S;2;1;1;0;0
44	GLN-C	C;4;3;0;0;0
45	GLN-CA	C;4;3;1;0;0
46	GLN-CB	C;4;2;2;0;0
47	GLN-CD	C;4;3;0;0;0
48	GLN-CG	C;4;2;2;0;0
49	GLN-N	N;3;2;1;0;0
50	GLN-NE2	N;3;1;2;0;0
51	GLN-O	O;2;1;0;0;0
52	GLN-OE1	O;2;1;0;0;0
53	GLN-OXT	O;2;1;0;0;0
54	GLU-C	C;4;3;0;0;0
55	GLU-CA	C;4;3;1;0;0
56	GLU-CB	C;4;2;2;0;0
57	GLU-CD	C;5;3;0;0;0
58	GLU-CG	C;4;2;2;0;0
59	GLU-N	N;3;2;1;0;0
60	GLU-O	O;2;1;0;0;0
61	GLU-OE1	O;2;1;0;0;0
62	GLU-OE2	O;2;1;0;0;0
63	GLU-OXT	O;2;1;0;0;0
64	GLY-C	C;4;3;0;0;0
65	GLY-CA	C;4;2;2;0;0
66	GLY-N	N;3;2;1;0;0
67	GLY-O	O;2;1;0;0;0
68	GLY-OXT	O;2;1;0;0;0
69	HIS-C	C;4;3;0;0;0
70	HIS-CA	C;4;3;1;0;0
71	HIS-CB	C;4;2;2;0;0
72	HIS-CD2	C;4;2;1;1;1
73	HIS-CE1	C;4;2;1;1;1
74	HIS-CG	C;4;3;0;1;1
75	HIS-N	N;3;2;1;0;0
76	HIS-ND1	N;3;2;0;1;1
77	HIS-NE2	N;3;2;1;1;1
78	HIS-O	O;2;1;0;0;0
79	HIS-OXT	O;2;1;0;0;0
80	ILE-C	C;4;3;0;0;0
81	ILE-CA	C;4;3;1;0;0
82	ILE-CB	C;4;3;1;0;0
83	ILE-CD1	C;4;1;3;0;0
84	ILE-CG1	C;4;2;2;0;0
85	ILE-CG2	C;4;1;3;0;0
86	ILE-N	N;3;2;1;0;0
87	ILE-O	O;2;1;0;0;0
88	ILE-OXT	O;2;1;0;0;0
89	LEU-C	C;4;3;0;0;0
90	LEU-CA	C;4;3;1;0;0
91	LEU-CB	C;4;2;2;0;0
92	LEU-CD1	C;4;1;3;0;0
93	LEU-CD2	C;4;1;3;0;0
94	LEU-CG	C;4;3;1;0;0
95	LEU-N	N;3;2;1;0;0
96	LEU-O	O;2;1;0;0;0
97	LEU-OXT	O;2;1;0;0;0
98	LYS-C	C;4;3;0;0;0
99	LYS-CA	C;4;3;1;0;0
100	LYS-CB	C;4;2;2;0;0
101	LYS-CD	C;4;2;2;0;0
102	LYS-CE	C;4;2;2;0;0
103	LYS-CG	C;4;2;2;0;0
104	LYS-N	N;3;2;1;0;0
105	LYS-NZ	N;4;1;3;0;0
106	LYS-O	O;2;1;0;0;0
107	LYS-OXT	O;2;1;0;0;0
108	MET-C	C;4;3;0;0;0
109	MET-CA	C;4;3;1;0;0
110	MET-CB	C;4;2;2;0;0
111	MET-CE	C;4;1;3;0;0
112	MET-CG	C;4;2;2;0;0
113	MET-N	N;3;2;1;0;0
114	MET-O	O;2;1;0;0;0
115	MET-OXT	O;2;1;0;0;0
116	MET-SD	S;2;2;0;0;0
117	PHE-C	C;4;3;0;0;0
118	PHE-CA	C;4;3;1;0;0
119	PHE-CB	C;4;2;2;0;0
120	PHE-CD1	C;4;2;1;1;1
121	PHE-CD2	C;4;2;1;1;1
122	PHE-CE1	C;4;2;1;1;1
123	PHE-CE2	C;4;2;1;1;1
124	PHE-CG	C;4;3;0;1;1
125	PHE-CZ	C;4;2;1;1;1
126	PHE-N	N;3;2;1;0;0
127	PHE-O	O;2;1;0;0;0
128	PHE-OXT	O;2;1;0;0;0
129	PRO-C	C;4;3;0;0;0
130	PRO-CA	C;4;3;1;0;1
131	PRO-CB	C;4;2;2;0;1
132	PRO-CD	C;4;2;2;0;1
133	PRO-CG	C;4;2;2;0;1
134	PRO-N	N;3;3;0;0;1
135	PRO-O	O;2;1;0;0;0
136	PRO-OXT	O;2;1;0;0;0
137	SER-C	C;4;3;0;0;0
138	SER-CA	C;4;3;1;0;0
139	SER-CB	C;4;2;2;0;0
140	SER-N	N;3;2;1;0;0
141	SER-O	O;2;1;0;0;0
142	SER-OG	O;2;1;1;0;0
143	SER-OXT	O;2;1;0;0;0
144	THR-C	C;4;3;0;0;0
145	THR-CA	C;4;3;1;0;0
146	THR-CB	C;4;3;1;0;0
147	THR-CG2	C;4;1;3;0;0
148	THR-N	N;3;2;1;0;0
149	THR-O	O;2;1;0;0;0
150	THR-OG1	O;2;1;1;0;0
151	THR-OXT	O;2;1;0;0;0
152	TRP-C	C;4;3;0;0;0
153	TRP-CA	C;4;3;1;0;0
154	TRP-CB	C;4;2;2;0;0
155	TRP-CD1	C;4;2;1;1;1
156	TRP-CD2	C;4;3;0;1;1
157	TRP-CE2	C;4;3;0;1;1
158	TRP-CE3	C;4;2;1;1;1
159	TRP-CG	C;4;3;0;1;1
160	TRP-CH2	C;4;2;1;1;1
161	TRP-CZ2	C;4;2;1;1;1
162	TRP-CZ3	C;4;2;1;1;1
163	TRP-N	N;3;2;1;0;0
164	TRP-NE1	N;3;2;1;1;1
165	TRP-O	O;2;1;0;0;0
166	TRP-OXT	O;2;1;0;0;0
167	TYR-C	C;4;3;0;0;0
168	TYR-CA	C;4;3;1;0;0
169	TYR-CB	C;4;2;2;0;0
170	TYR-CD1	C;4;2;1;1;1
171	TYR-CD2	C;4;2;1;1;1
172	TYR-CE1	C;4;2;1;1;1
173	TYR-CE2	C;4;2;1;1;1
174	TYR-CG	C;4;3;0;1;1
175	TYR-CZ	C;4;3;0;1;1
176	TYR-N	N;3;2;1;0;0
177	TYR-O	O;2;1;0;0;0
178	TYR-OH	O;2;1;1;0;0
179	TYR-OXT	O;2;1;0;0;0
180	VAL-C	C;4;3;0;0;0
181	VAL-CA	C;4;3;1;0;0
182	VAL-CB	C;4;3;1;0;0
183	VAL-CG1	C;4;1;3;0;0
184	VAL-CG2	C;4;1;3;0;0
185	VAL-N	N;3;2;1;0;0
186	VAL-O	O;2;1;0;0;0
187	VAL-OXT	O;2;1;0;0;0

Figure C.1: All heavy atoms in every amino acid in the PDBbind dataset and their corresponding ECIF atom type.

C.3 Cross Validation

5-fold cross validation was performed on the combined PDBBind general 2020 and PDBBind refined 2016 set (dataset details in the main text Section 4.3.1). The validation and test set (CASF-2016) used in the main study was removed from the cross

C;4;2;2;0;0
N;4;1;2;0;0
S;2;2;0;0;0
C;4;3;0;1;1
N;3;1;2;0;0
C;4;3;0;0;0
C;4;1;3;0;0
N;3;2;1;1;1
S;2;1;1;0;0
C;4;3;1;0;1
N;3;2;0;1;1
C;4;3;1;0;0
O;2;1;1;0;0
N;4;2;1;0;0
N;3;2;1;0;0
N;4;1;3;0;0
C;6;3;0;0;0
C;4;2;2;0;1
C;4;2;1;1;1
O;2;1;0;0;0
N;3;3;0;0;1
C;5;3;0;0;0

Table C.17: Unique 22 ECIF atom types present in the 20 proteinogenic amino acids in the PDBbind 2020/2016 dataset.

validation set, to leave the training set of 14254 compounds. Cross validation was done separately for the docked and crystal-based datasets to ensure model stability for both, crystal derived structures and docked poses. This set was split into 5 random folds, using 20 % of the dataset as validation in each fold. Models were run across all 5 folds until model performance converged. The performance of all models during cross validation is recorded for every epoch. The number of epochs to train each model for training and test on the CASF-2016 benchmark was determined as the point where no significant performance increase (difference of less than 0.01 pearson correlation coefficient between epoch n and $n-1$). The determined optimal number of epochs for each model is given in Tables C.18, C.19 and C.20.

Model Architecture	PB Included	PB Excluded
GATNet + PLIG	9	16
GCNNet + PLIG	8	18
GIN + PLIG	6	18
GAT/GCN + PLIG	9	19
SGCNet + PLIG	4	14
SageNet + PLIG	6	21
MLPNet + ECIF	5	15

Table C.18: Optimal number of epochs for models trained on crystal structures.

Model Architecture	PB Included	PB Excluded
GATNet + PLIG	9	11
GCNNet + PLIG	5	12
GIN + PLIG	6	17
GAT/GCN + PLIG	3	18
SGCNet + PLIG	4	14
SageNet + PLIG	8	21
MLPNet + ECIF	5	6

Table C.19: Optimal number of epochs for models trained on docked poses.

Model Architecture	PB Included	PB Excluded
GATNet	4	13
GCNNet	5	15
GIN	7	22
GAT/GCN	7	15
SGCNet	5	22
SageNet	7	21
MLPNet + ECFP512	5	6
MLPNet + ECFP1024	7	6
MLPNet + FCFP512	5	8
MLPNet + FCFP1024	5	4

Table C.20: Optimal number of epochs for ligand-based models.

It is noteworthy, that the model architecture that includes the protein branch (PB Included) consistently reaches peak performance significantly faster than the models without protein sequence embedding (all models except the ligand-based ECFP and FCFP fingerprints models using the MLPNet architecture). On average, GNNs using PLIGs and do not include the protein branch take 17 epochs to converge while

the same models with the protein branch take only 6 epochs. This could be due to the protein branch models over fitting during training, reaching peak performance quickly, while pure GNN PLIG models need several epochs to slowly learn more meaningful information about the system. The following pages include the performance evaluation of all models during cross validation for the reader to inspect.

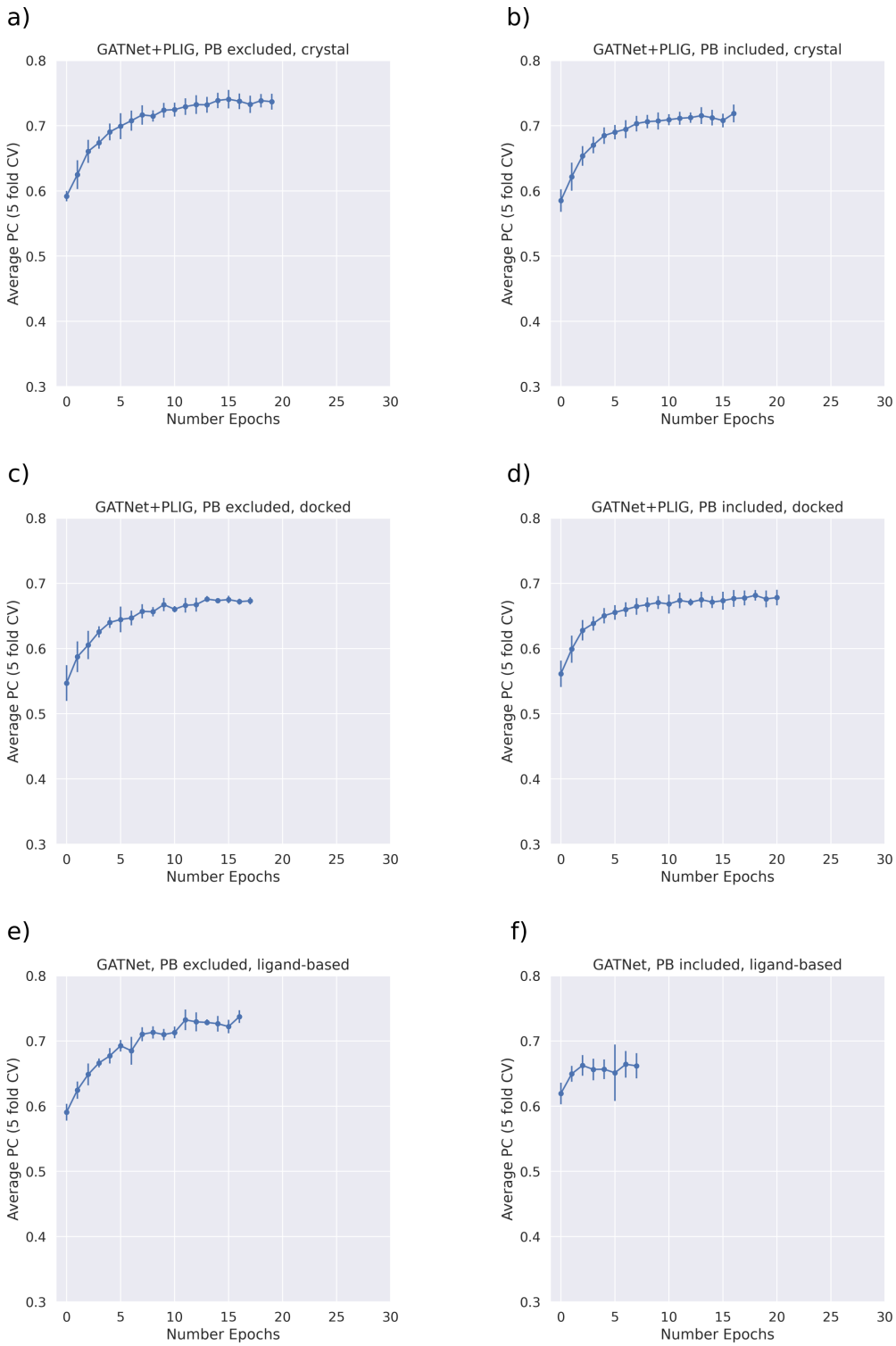

Figure C.2: 5-fold cross validation of the GATNet models reported as the average calculated pearson correlation coefficient and its corresponding standard deviation (error bars) at each epoch. a) PB excluded, crystal structures; b) PB included, crystal structures; c) PB excluded, docked poses; d) PB included, docked poses, e) PB excluded, ligand-based, f) PB included, ligand-based. Models including the protein branch reach peak performance extremely fast, indicating potential overfitting.

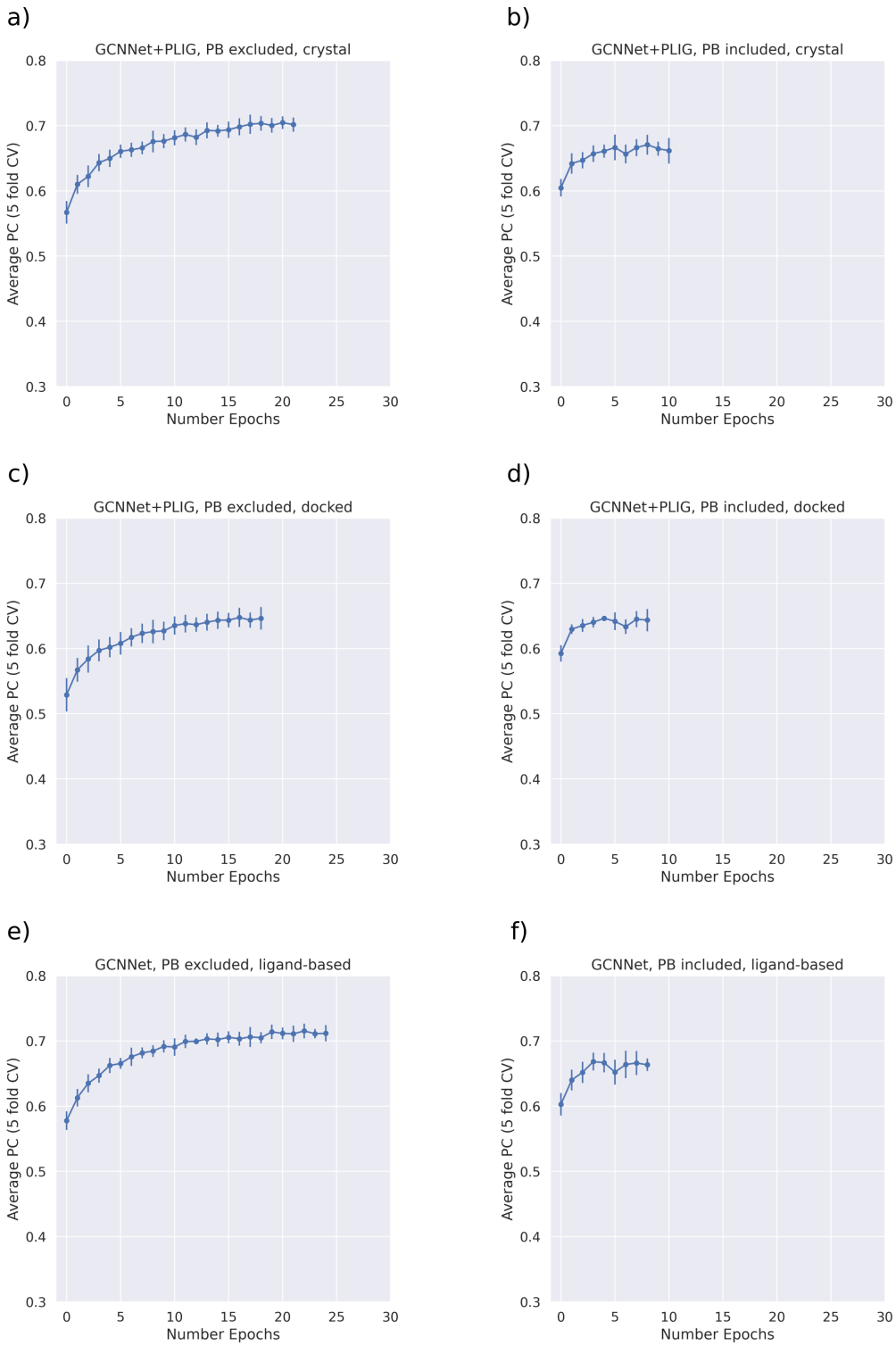

Figure C.3: 5-fold cross validation of the GCNNet models reported as the average calculated pearson correlation coefficient and its corresponding standard deviation (error bars) at each epoch. a) PB excluded, crystal structures; b) PB included, crystal structures; c) PB excluded, docked poses; d) PB included, docked poses, e) PB excluded, ligand-based, f) PB included, ligand-based. Models including the protein branch reach peak performance extremely fast, indicating potential overfitting.

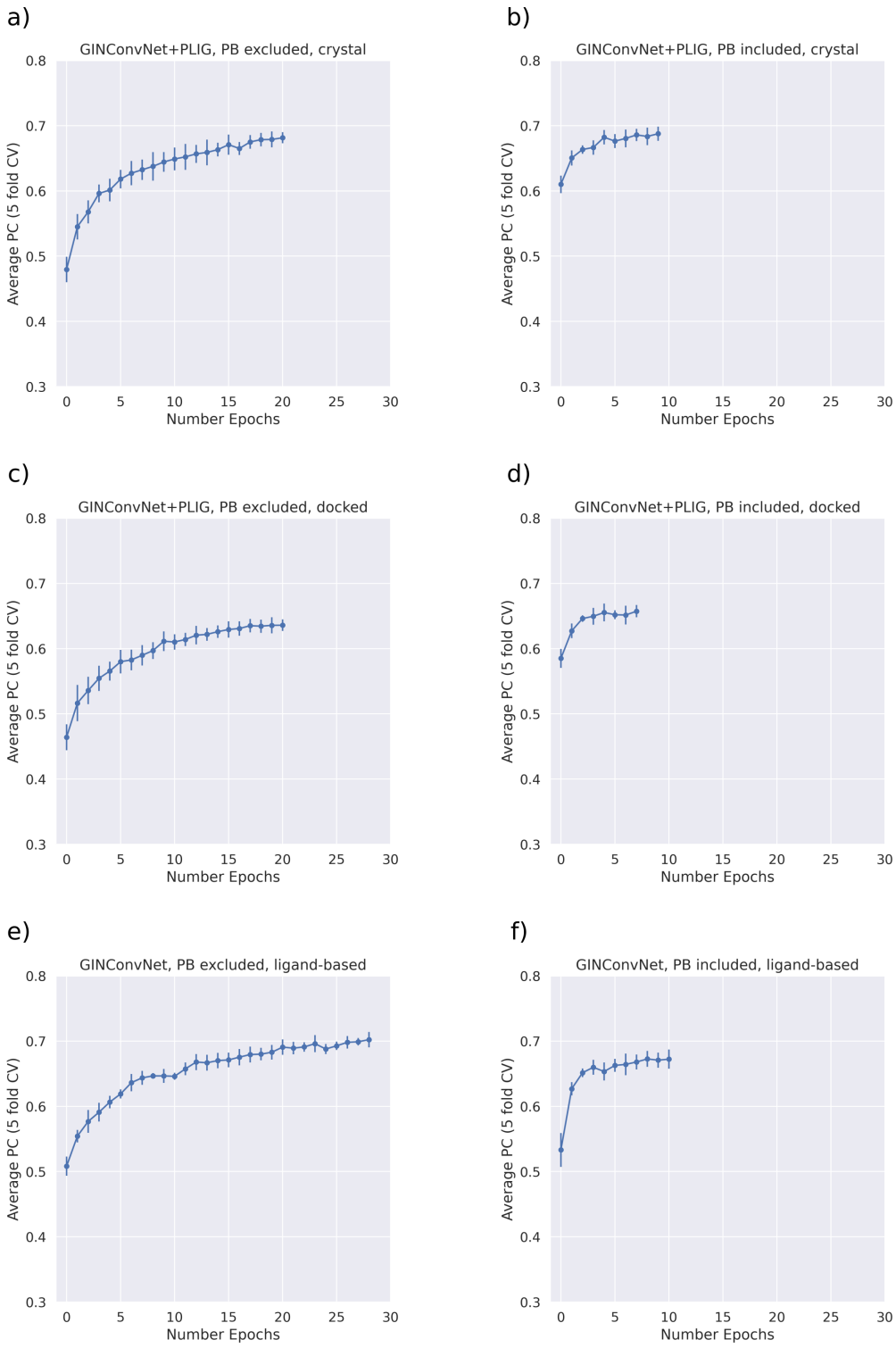

Figure C.4: 5-fold cross validation of the GIN models reported as the average calculated pearson correlation coefficient and its corresponding standard deviation (error bars) at each epoch. a) PB excluded, crystal structures; b) PB included, crystal structures; c) PB excluded, docked poses; d) PB included, docked poses, e) PB excluded, ligand-based, f) PB included, ligand-based. Models including the protein branch reach peak performance extremely fast, indicating potential overfitting.

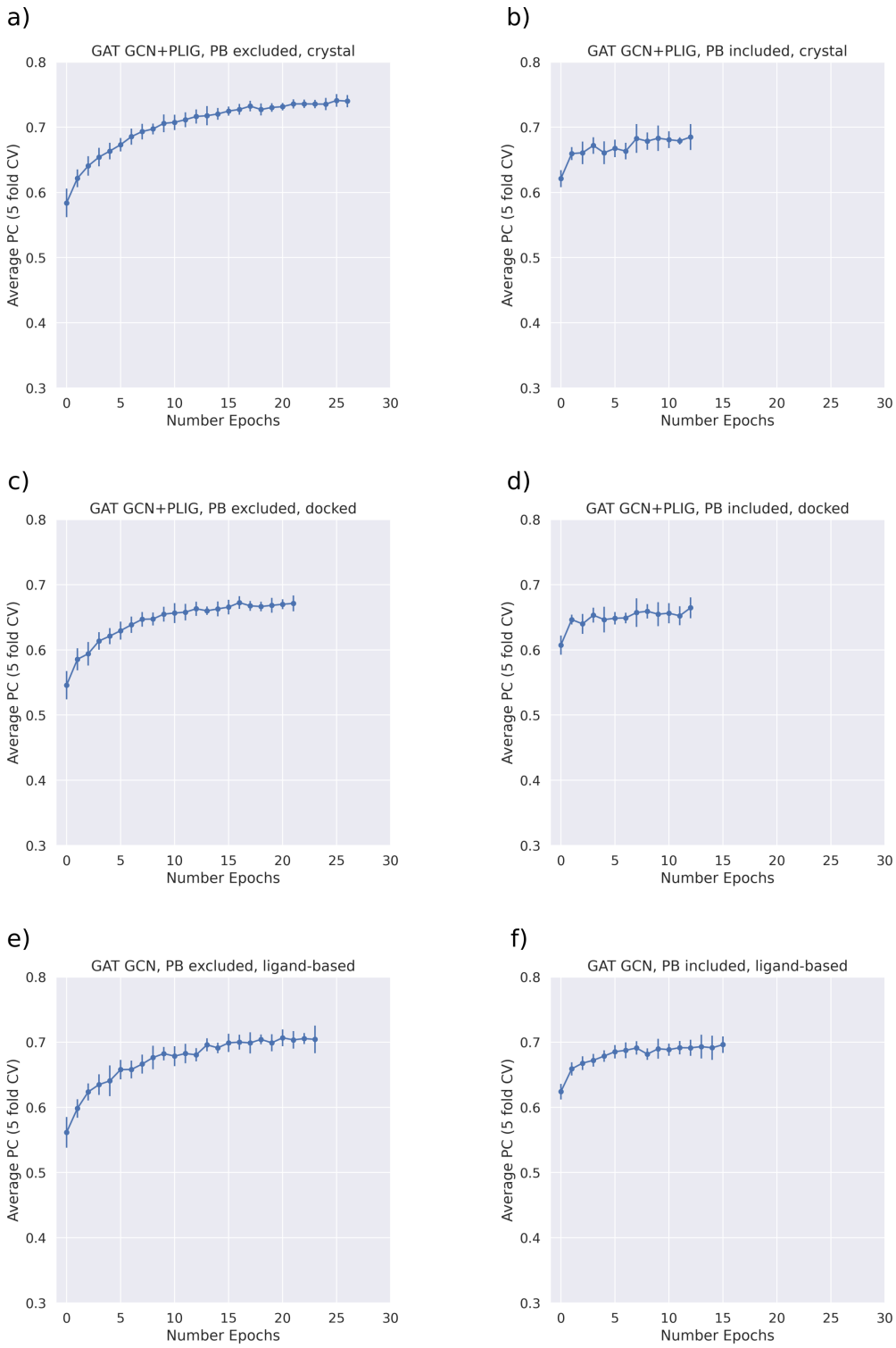

Figure C.5: 5-fold cross validation of the GAT/GCN models reported as the average calculated pearson correlation coefficient and its corresponding standard deviation (error bars) at each epoch. a) PB excluded, crystal structures; b) PB included, crystal structures; c) PB excluded, docked poses; d) PB included, docked poses, e) PB excluded, ligand-based, f) PB included, ligand-based. Models including the protein branch reach peak performance extremely fast, indicating potential overfitting.

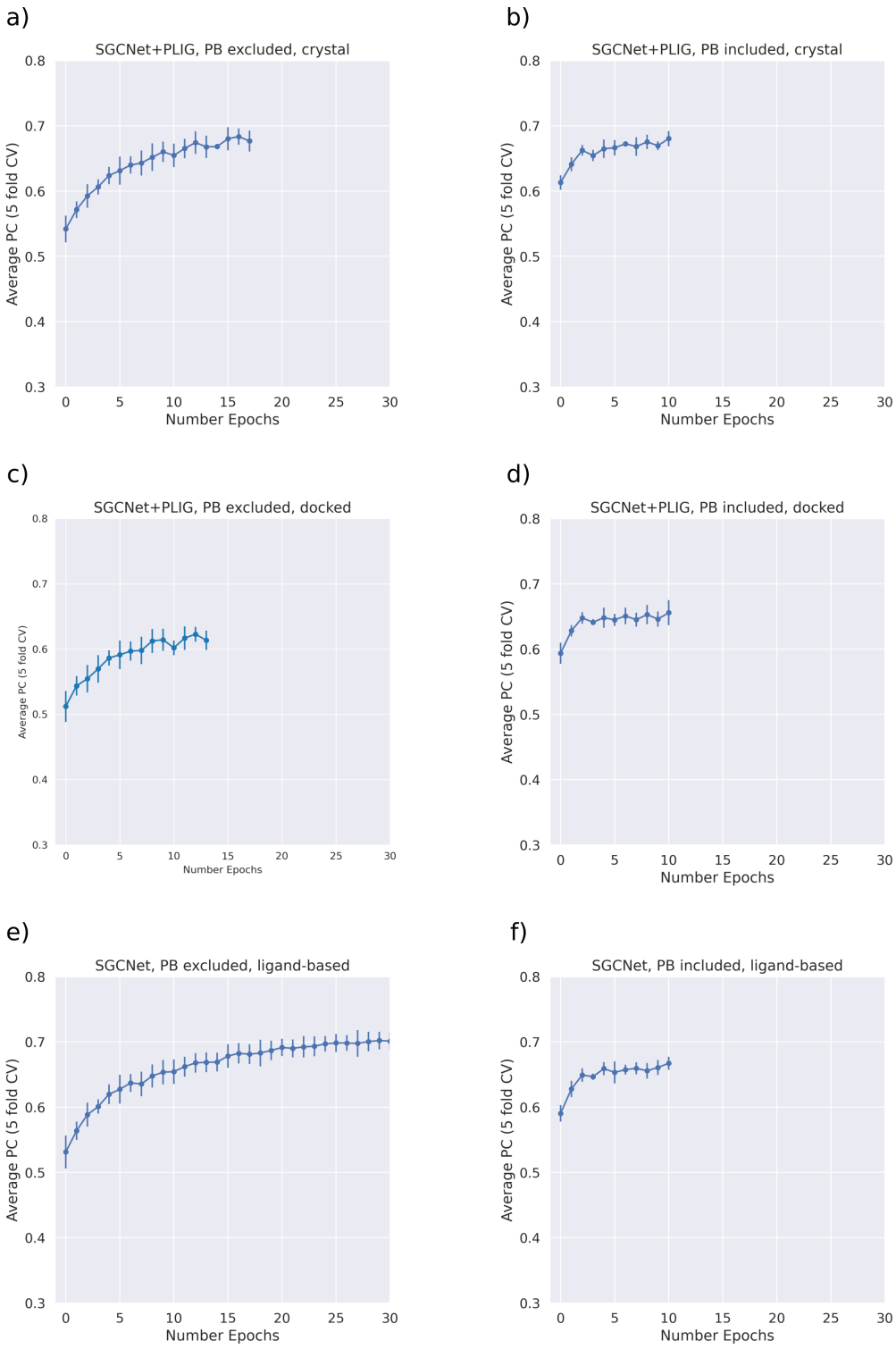

Figure C.6: 5-fold cross validation of the SGCNet models reported as the average calculated pearson correlation coefficient and its corresponding standard deviation (error bars) at each epoch. a) PB excluded, crystal structures; b) PB included, crystal structures; c) PB excluded, docked poses; d) PB included, docked poses, e) PB excluded, ligand-based, f) PB included, ligand-based. Models including the protein branch reach peak performance extremely fast, indicating potential overfitting.

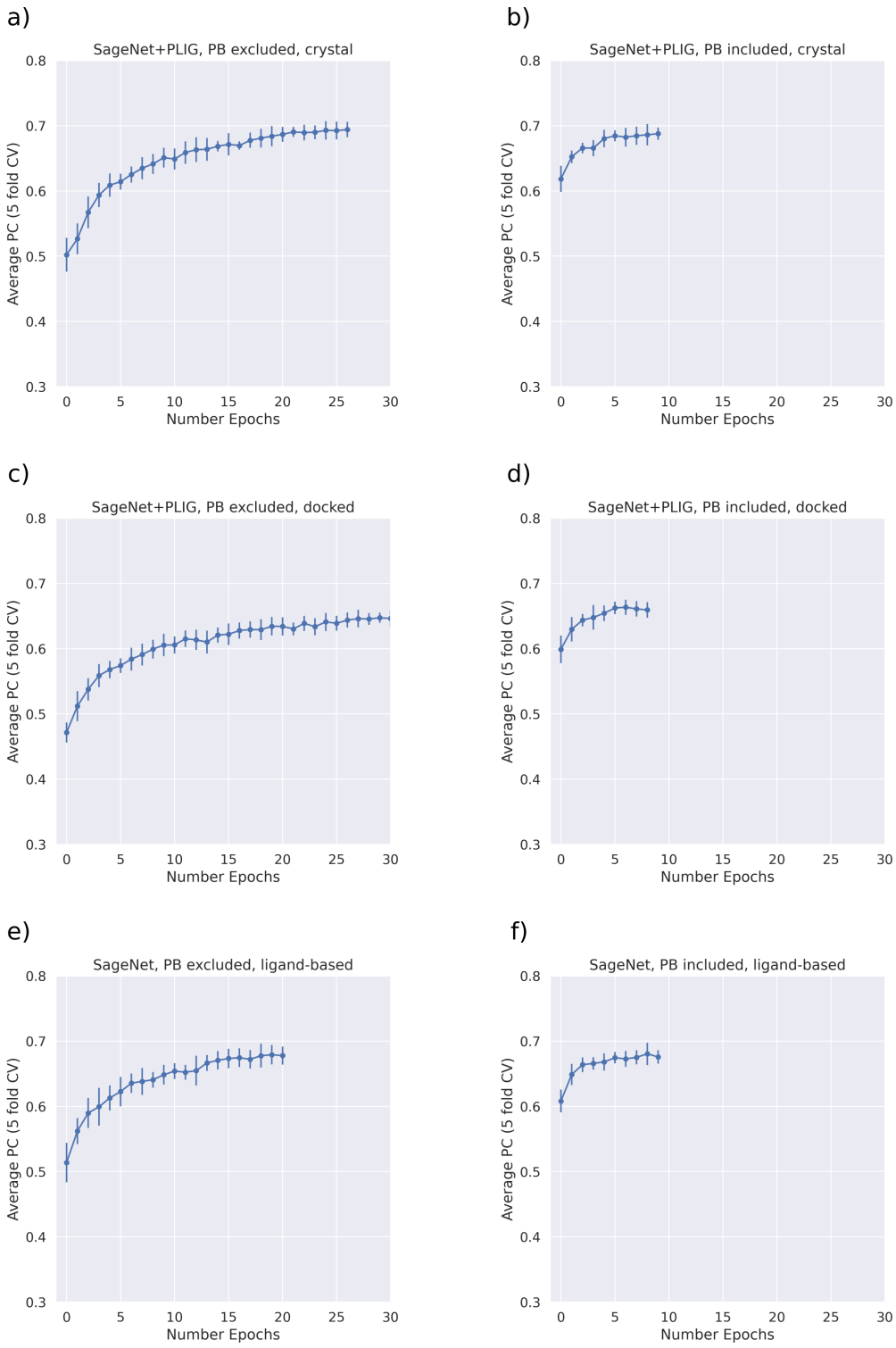

Figure C.7: 5-fold cross validation of the Sage models reported as the average calculated pearson correlation coefficient and its corresponding standard deviation (error bars) at each epoch. a) PB excluded, crystal structures; b) PB included, crystal structures; c) PB excluded, docked poses; d) PB included, docked poses, e) PB excluded, ligand-based, f) PB included, ligand-based. Models including the protein branch reach peak performance extremely fast, indicating potential overfitting.

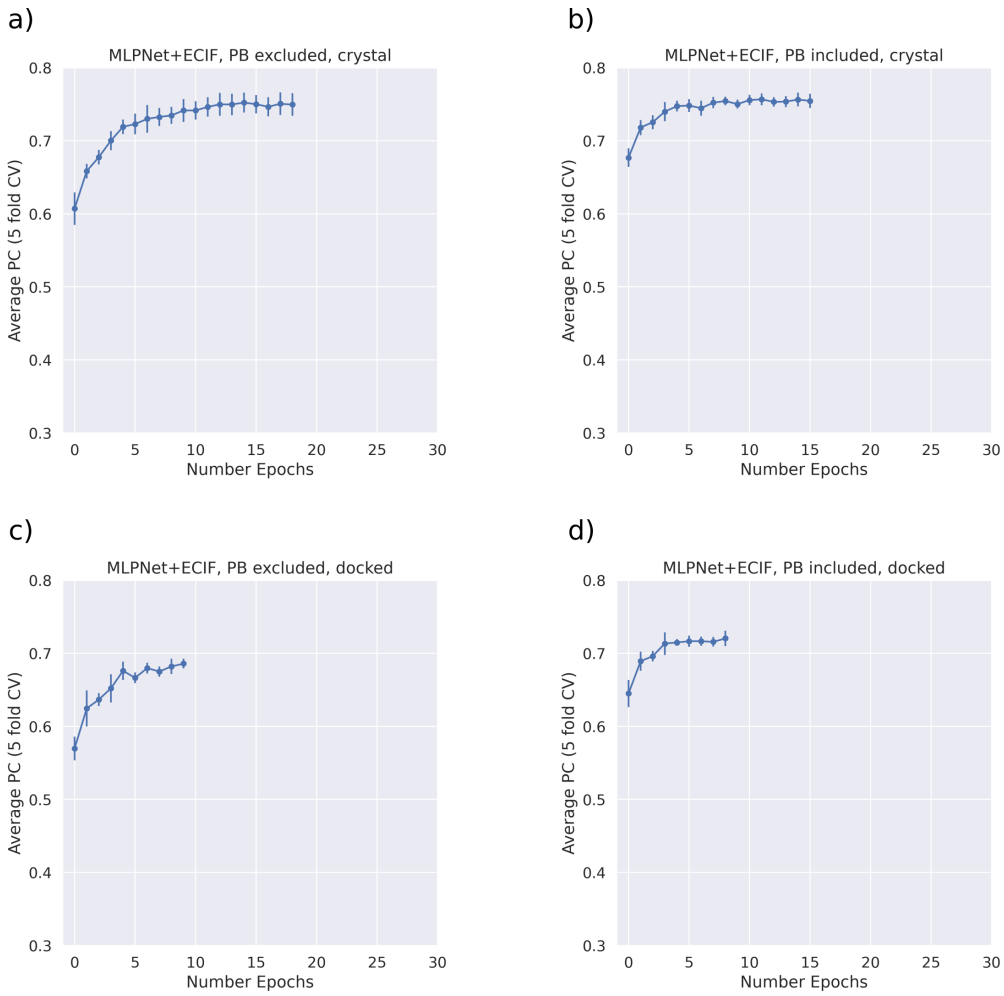

Figure C.8: 5-fold cross validation of the MLPNet + ECIF models reported as the average calculated pearson correlation coefficient and its corresponding standard deviation (error bars) at each epoch. a) PB excluded, crystal structures; b) PB included, crystal structures; c) PB excluded, docked poses; d) PB included, docked poses. Models including the protein branch reach peak performance extremely fast, indicating potential overfitting.

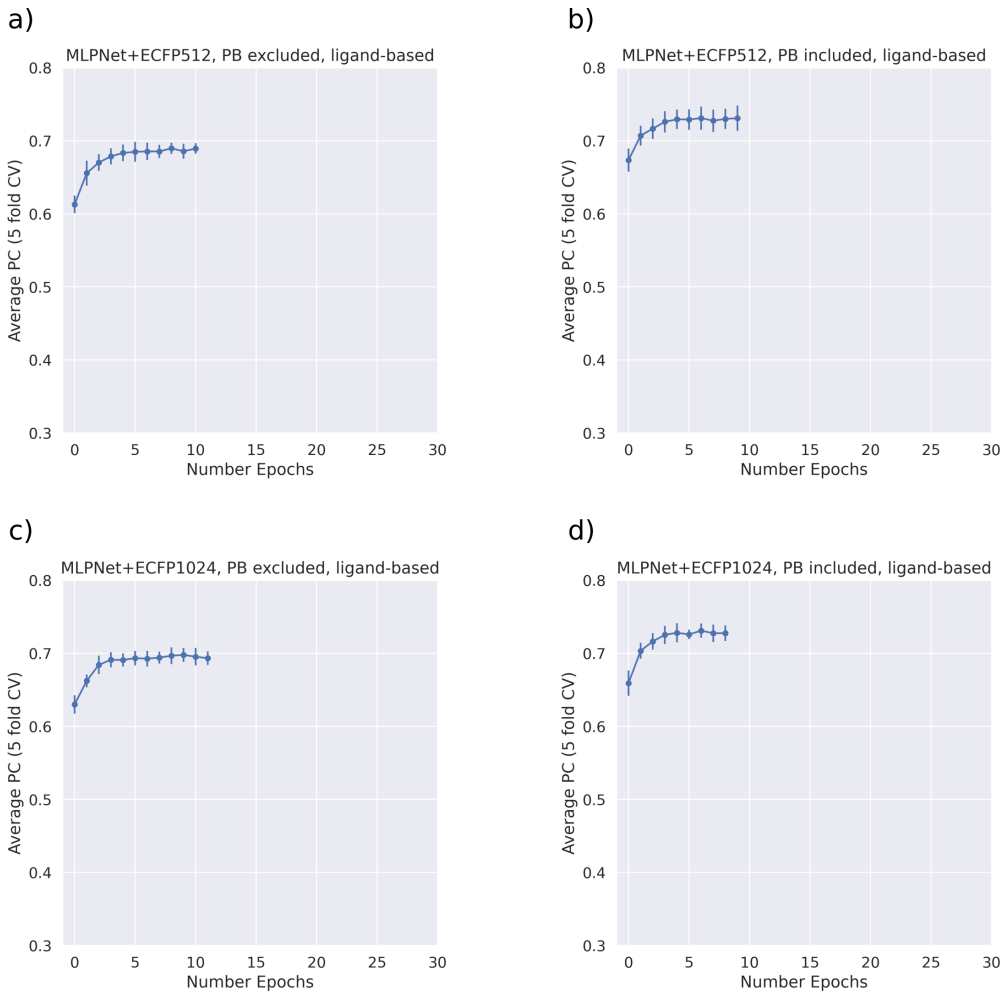

Figure C.9: 5-fold cross validation of the MLPNet + ECFP models reported as the average calculated pearson correlation coefficient and its corresponding standard deviation (error bars) at each epoch. a) ECFP512, PB excluded; b) ECFP512, PB included; c) ECFP1024, PB excluded; d) ECFP1024, PB included. Models including the protein branch reach peak performance extremely fast, indicating potential overfitting.

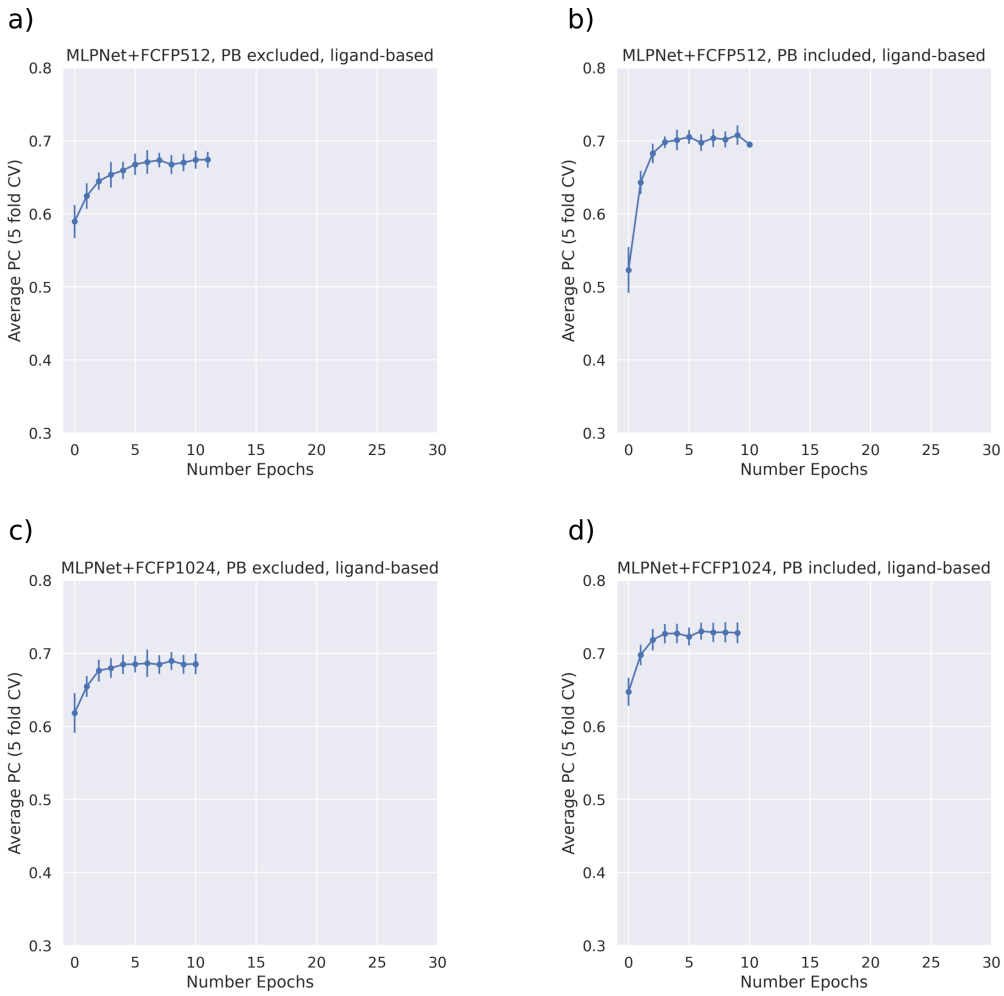

Figure C.10: 5-fold cross validation of the MLPNet + FCFP models reported as the average calculated pearson correlation coefficient and its corresponding standard deviation (error bars) at each epoch. a) FCFP512, PB excluded; b) FCFP512, PB included; c) FCFP1024, PB excluded; d) FCFP1024, PB included. Models including the protein branch reach peak performance extremely fast, indicating potential overfitting.

C.4 Performance and Stability of All Models

All models were trained on the PDBBind dataset (description see main text Section 4.3.1) and tested on the CASF-2016 benchmark set (crystal case is trained and tested on crystal structures, docked case is trained and tested on docked poses and the ligand case does not use 3D information). Since the models' predictions are somewhat stochastic, model performance and stability against the withheld test set (CASF-

2016) for all trained models was evaluated using the average and standard deviation (SD) of the pearson correlation coefficient (ρ) as well as the root-mean-square error (RMSE) over 10 runs. ρ and RMSE and their corresponding standard deviations are shown in Figure C.11 and FigureC.12. The scatter plots of the average prediction versus the experimentally determined pK value for each protein-ligand complex between the 10 model runs for each model and feature combination are shown in Figure C.14-C.22. The best performing PLIG model was the GATNet PLIG without protein sequence embedding ($\rho = 0.80$) and the best performing model overall was the MLPNet ECIF model with protein sequence embedding ($\rho = 0.82$). However, since the standard deviation of the Pearson correlation coefficient is as large as the difference between the GATNet PLIG and MLPNet ECIF model (Figure C.11, standard deviation of 0.018 versus 0.009 for GATNet PLIG and MLPNet ECIF, respectively), the difference is not significant and both models should be considered to be of similar performance. Overall, model stability as measured by ρ standard deviation varied between 0.007 (GCNNet PLIG, no sequence trained and tested on crystal poses) and 0.023 (SGCNet ligand-based, no sequence)

In addition, the case where models were trained on crystal poses and tested on docked poses was tested as well to investigate if performance improves. The results are shown in Figure C.13. There is no significant difference when training on crystal poses and testing on docked poses as opposed to the standard case (trained and tested on docked poses).

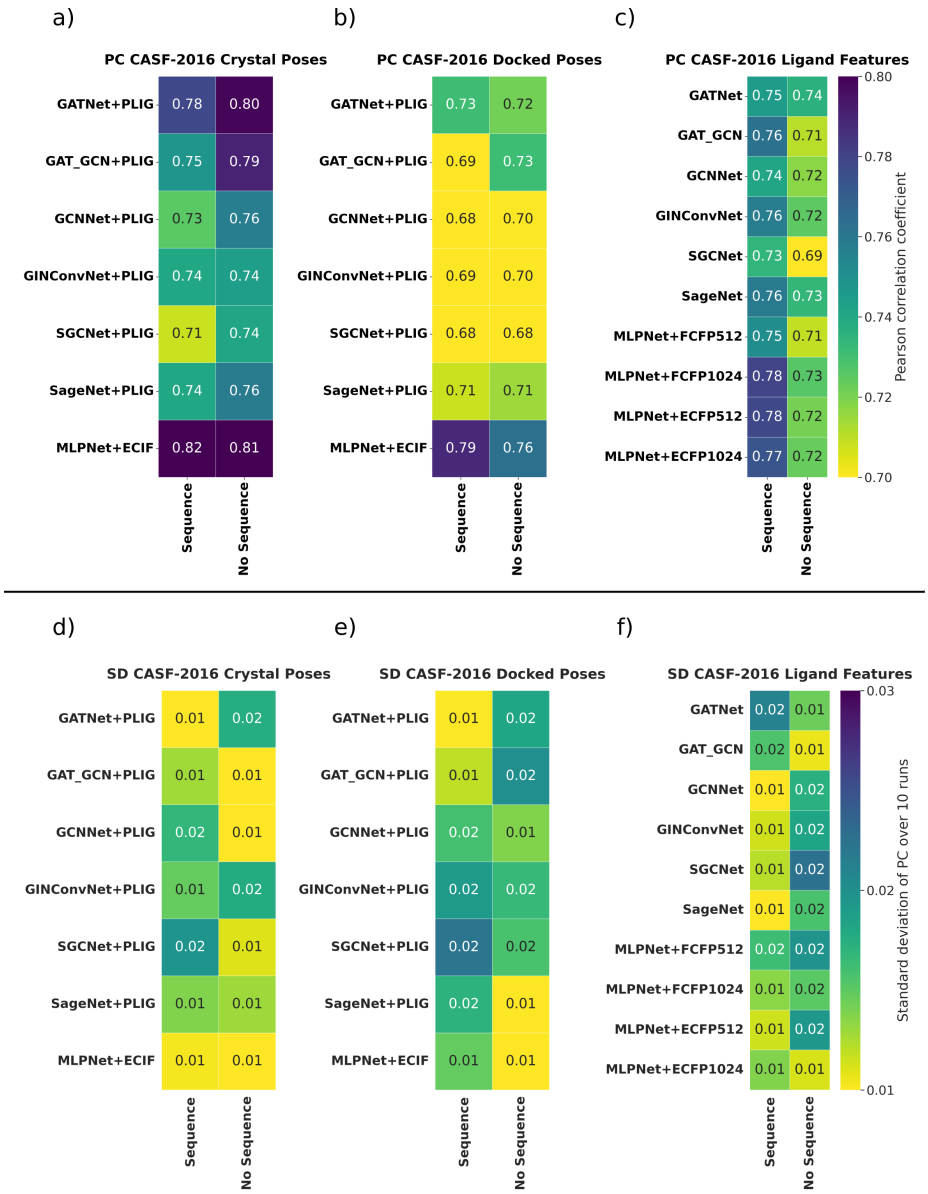

Figure C.11: Model stability of all models measured by Pearson correlation coefficient (*rho*) over 10 runs. *Sequence* and *No Sequence* denote the presence or absence of the protein sequence encoding branch in the model architecture, respectively. (a and d) ρ and SD of the structure-based GNN and MLPNet models when trained and tested on crystal structures. (b and e) ρ and SD of the structure-based GNN and MLPNet models when trained and tested on docked poses. (c and f) ρ and SD of all ligand-based GNN and MLPNet models. ECFP and FCFP fingerprint radius was 2 in all cases. Model stability as measured by ρ standard deviation varied between 0.007 (GCNNet PLIG no sequence trained/tested on crystal poses) and 0.023 (SGCNet ligand-based no sequence).

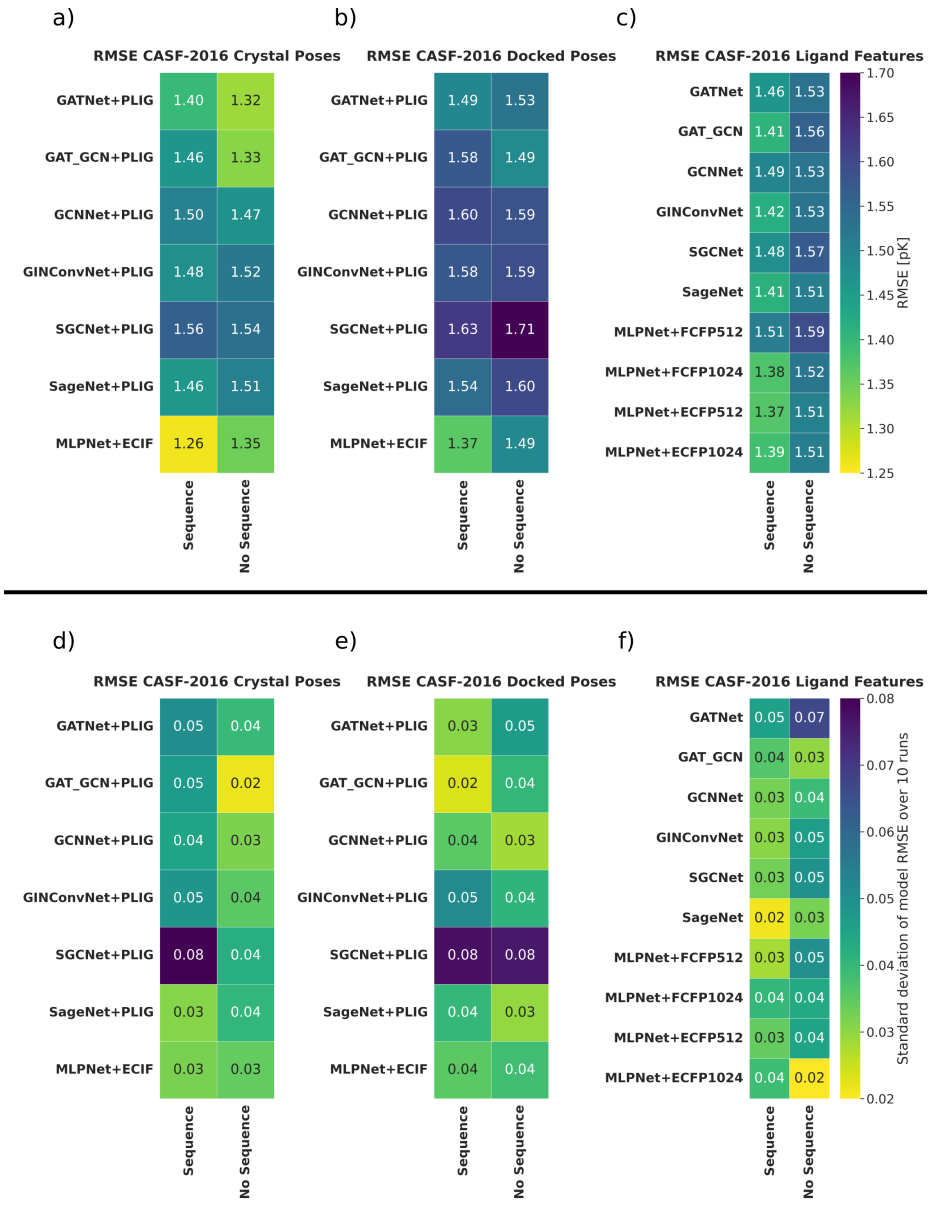

Figure C.12: Model stability of all models measured by RMSE over 10 runs. *Sequence* and *No Sequence* denote the presence or absence of the protein sequence encoding branch in the model architecture, respectively. (a and d) RMSE and SD of the structure-based GNN and MLPNet models when trained and tested on crystal structures. (b and e) RMSE and SD of the structure-based GNN and MLPNet models when trained and tested on docked poses. (c and f) RMSE and SD of all ligand-based GNN and MLPNet models. ECFP and FCFP fingerprint radius was 2 in all cases. Model stability as measured by RMSE standard deviation varied between 0.02 (several models) and 0.08 (several SGCNet based models).

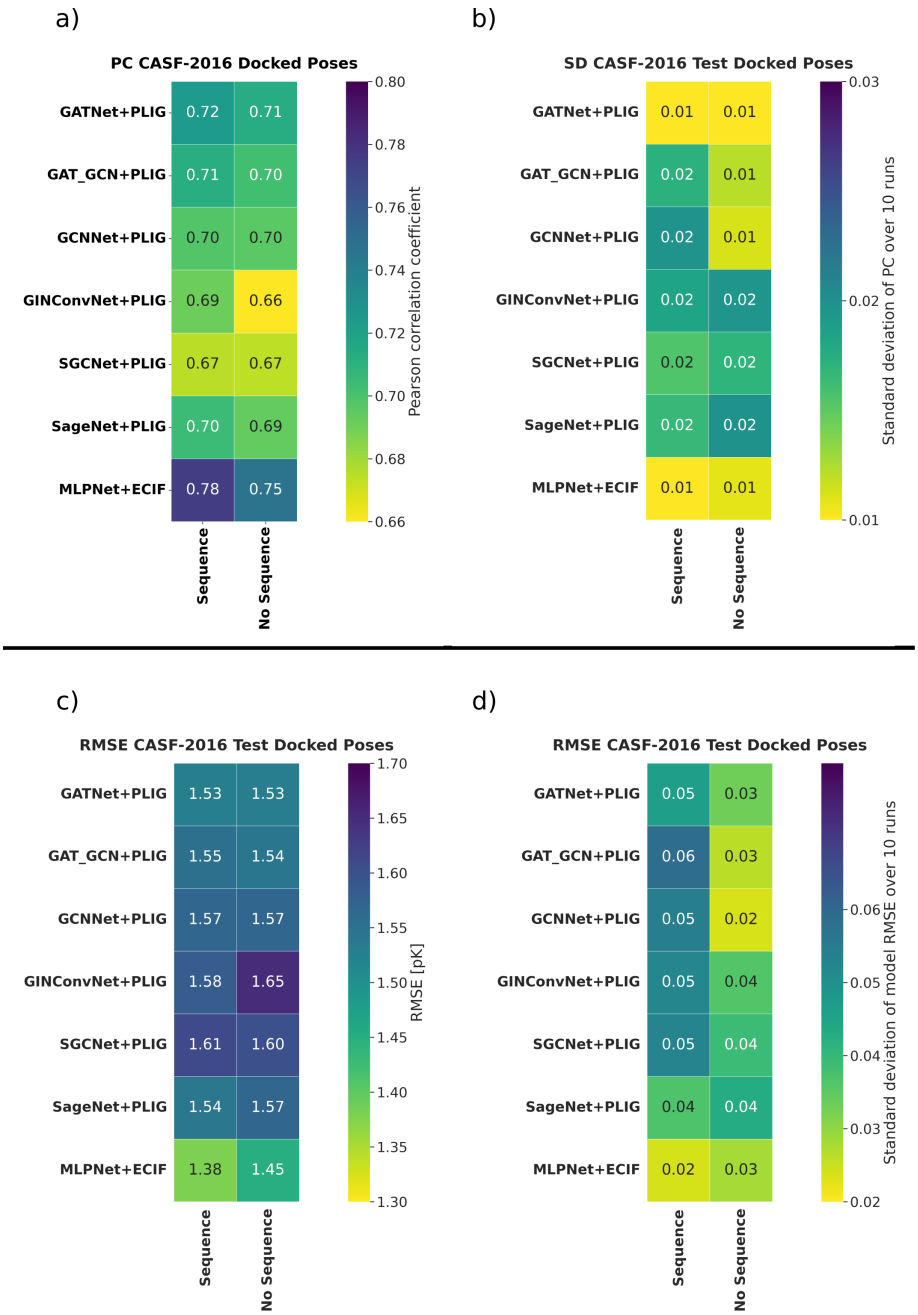

Figure C.13: Model performance when trained on crystal structures and tested on docked poses. *Sequence* and *No Sequence* denote the presence or absence of the protein sequence encoding branch in the model architecture, respectively. (a and b) ρ and SD of the structure-based GNN and MLPNet models when trained on crystal poses and tested on docked poses. (c and d) RMSE and SD of the structure-based GNN and MLPNet models when trained on crystal poses and tested on docked poses. Best performing model overall was the MLPNet ECIF model with protein sequence embedding ($\rho = 0.78$, RMSE $= 1.38$). There is no significant performance difference to models trained and tested on docked poses as shown in Figure C.11 and C.12

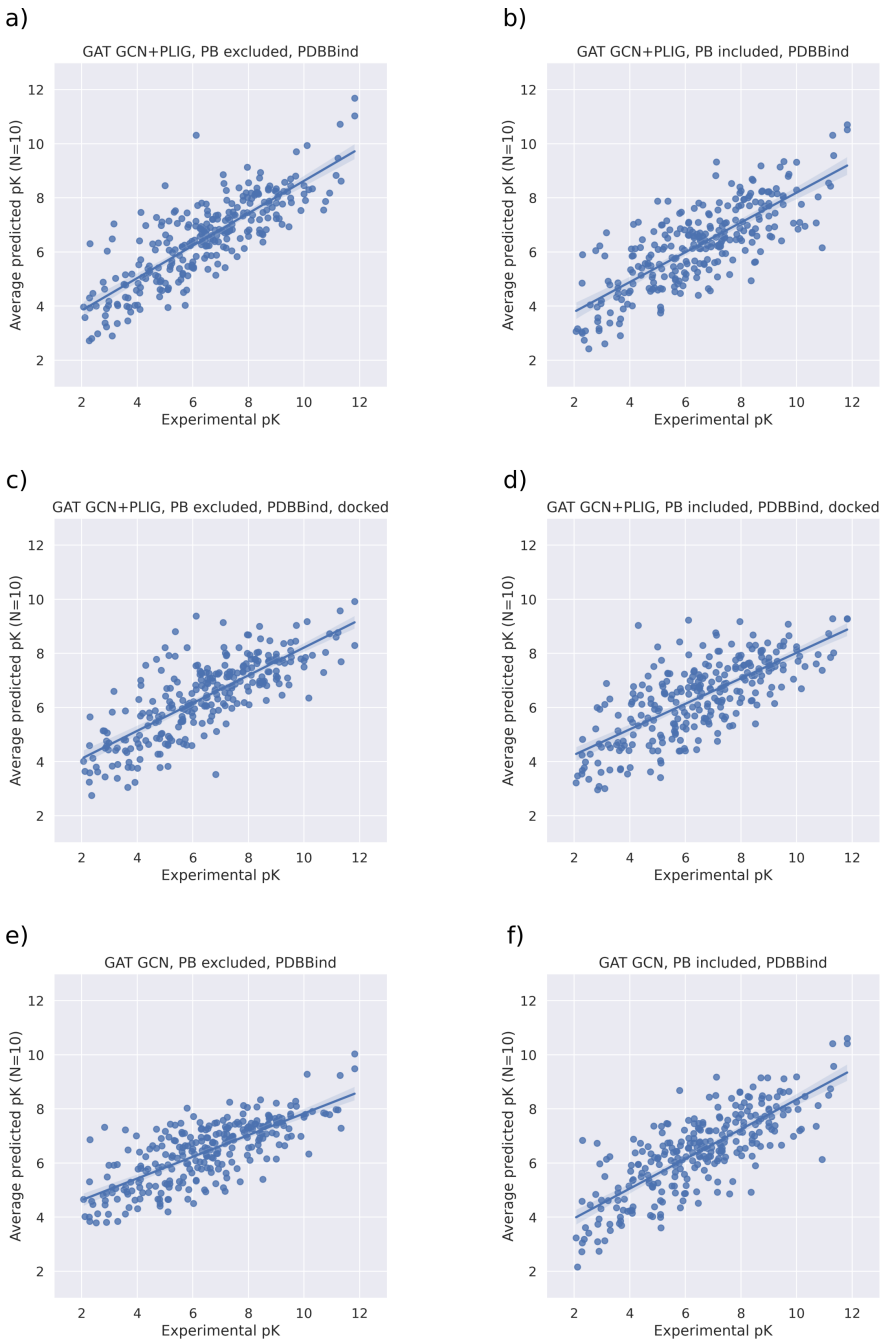

Figure C.14: Scatter plots of the average prediction versus the experimentally determined pK value for each protein-ligand complex between the 10 model runs for all GAT+GCN models. a) PLIG, no protein sequence embedding, trained and tested on crystal structures; b) PLIG, with sequence embedding, trained and tested on crystal structures; c) PLIG, no protein sequence embedding, trained and tested on docked structures; d) PLIG, with protein sequence embedding, trained and tested on docked structures; e) ligand-based graph, no protein sequence embedding; f) ligand-based graph, with protein embedding.

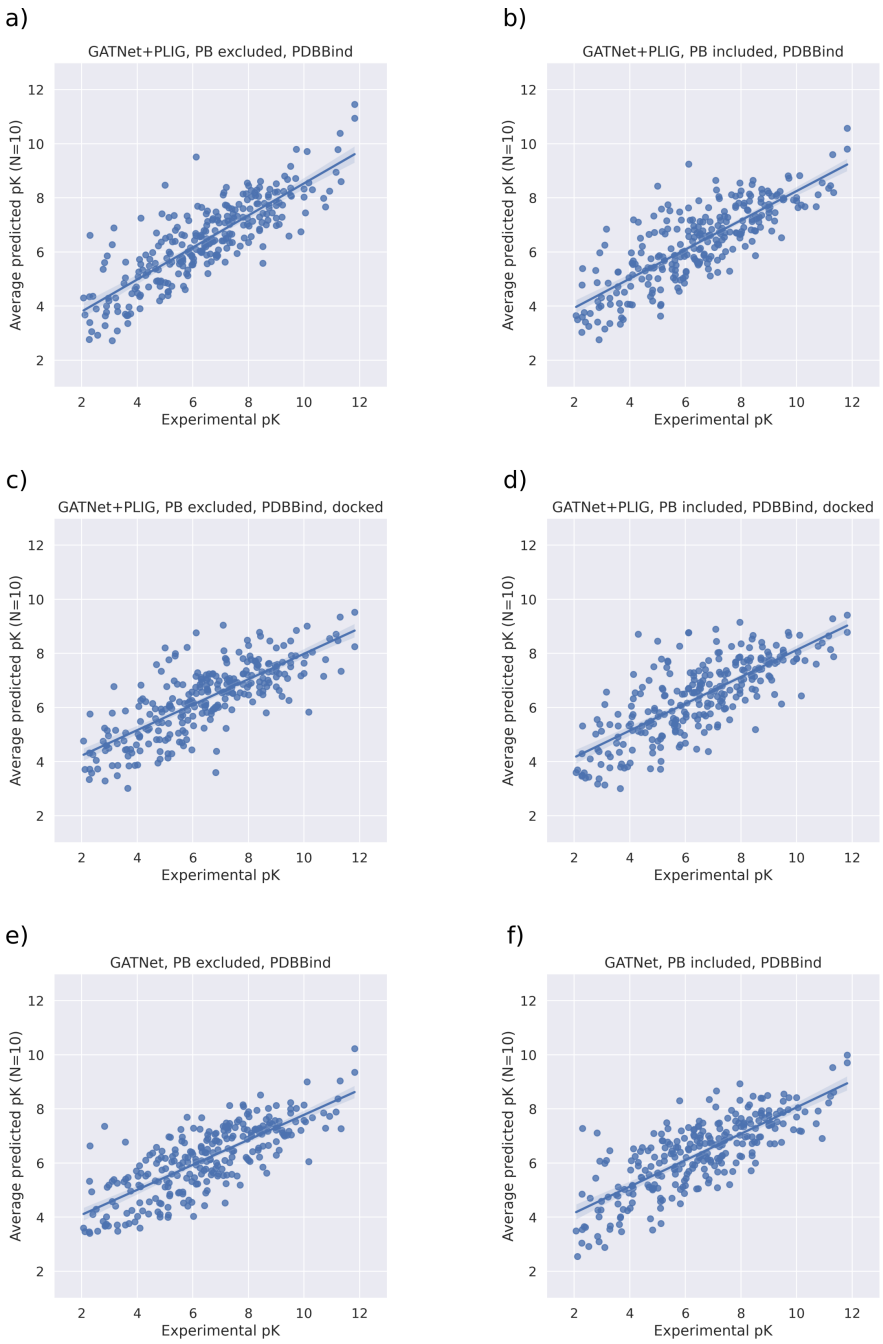

Figure C.15: Scatter plots of the average prediction versus the experimentally determined pK value for each protein-ligand complex between the 10 model runs for all GATNet models. a) PLIG, no protein sequence embedding, trained and tested on crystal structures; b) PLIG, with sequence embedding, trained and tested on crystal structures; c) PLIG, no protein sequence embedding, trained and tested on docked structures; d) PLIG, with protein sequence embedding, trained and tested on docked structures; e) ligand-based graph, no protein sequence embedding; f) ligand-based graph, with protein embedding.

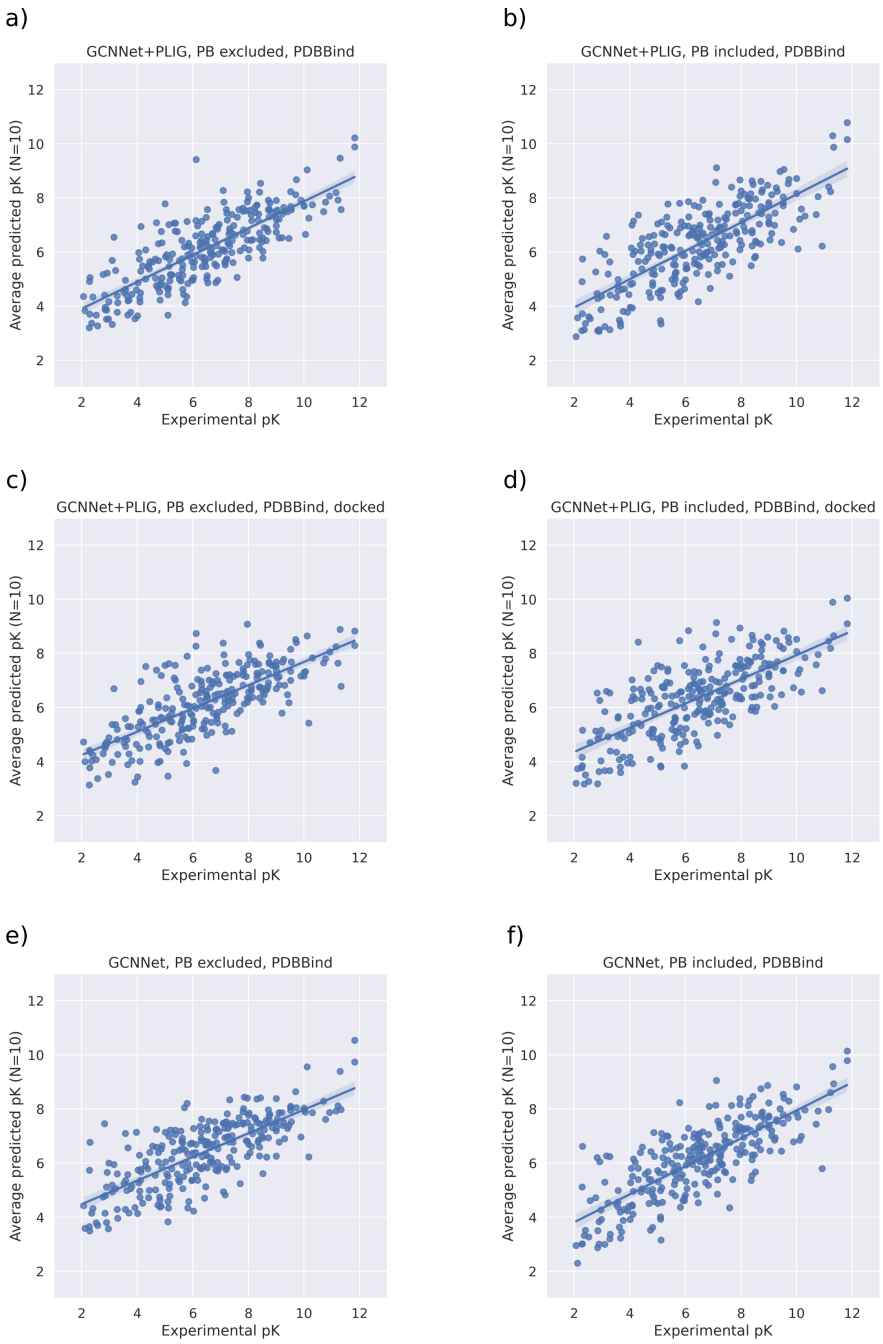

Figure C.16: Scatter plots of the average prediction versus the experimentally determined pK value for each protein-ligand complex between the 10 model runs for all GCNNet models. a) PLIG, no protein sequence embedding, trained and tested on crystal structures; b) PLIG, with sequence embedding, trained and tested on crystal structures; c) PLIG, no protein sequence embedding, trained and tested on docked structures; d) PLIG, with protein sequence embedding, trained and tested on docked structures; e) ligand-based graph, no protein sequence embedding; f) ligand-based graph, with protein embedding.

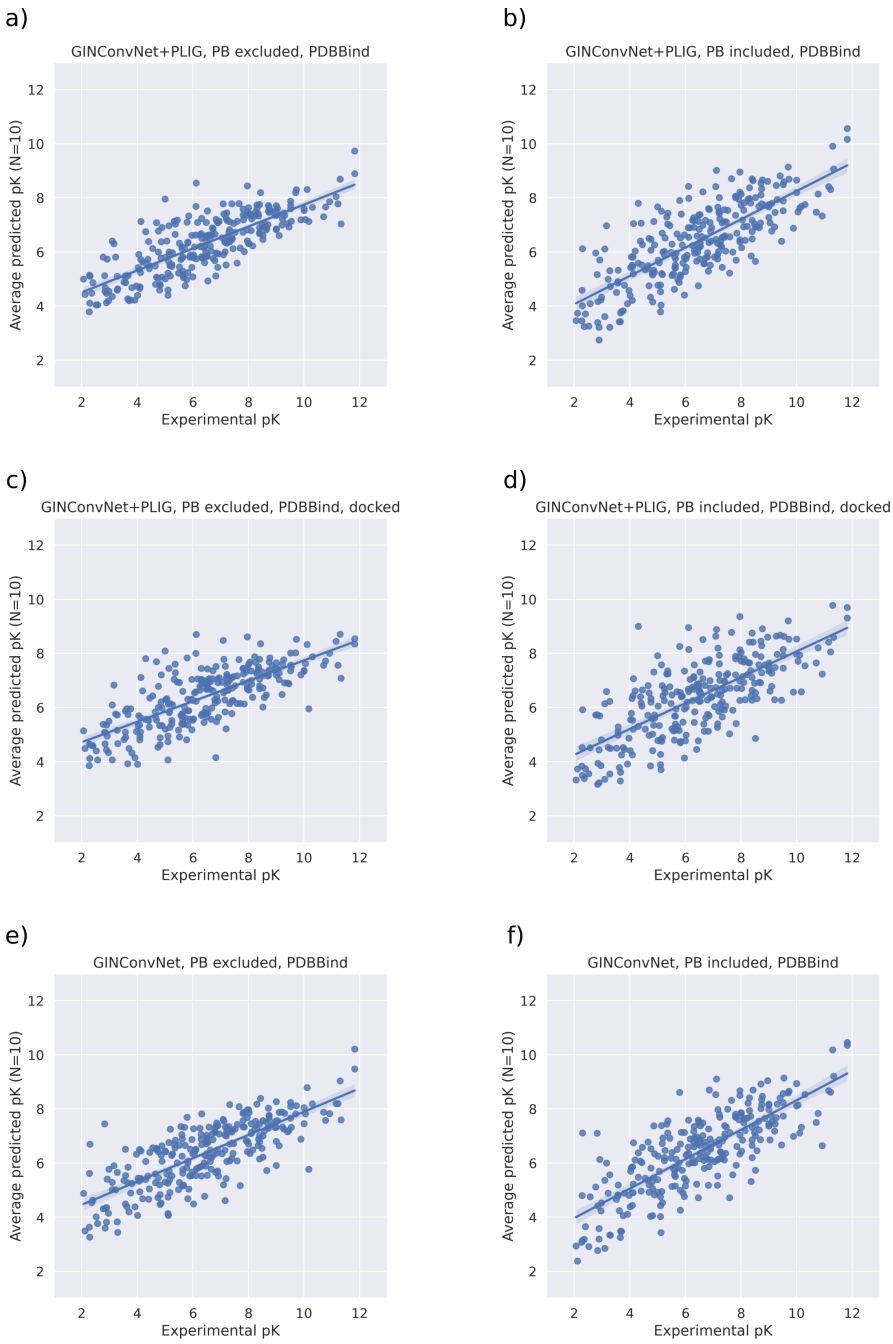

Figure C.17: Scatter plots of the average prediction versus the experimentally determined pK value for each protein-ligand complex between the 10 model runs for all GIN models. a) PLIG, no protein sequence embedding, trained and tested on crystal structures; b) PLIG, with sequence embedding, trained and tested on crystal structures; c) PLIG, no protein sequence embedding, trained and tested on docked structures; d) PLIG, with protein sequence embedding, trained and tested on docked structures; e) ligand-based graph, no protein sequence embedding; f) ligand-based graph, with protein embedding.

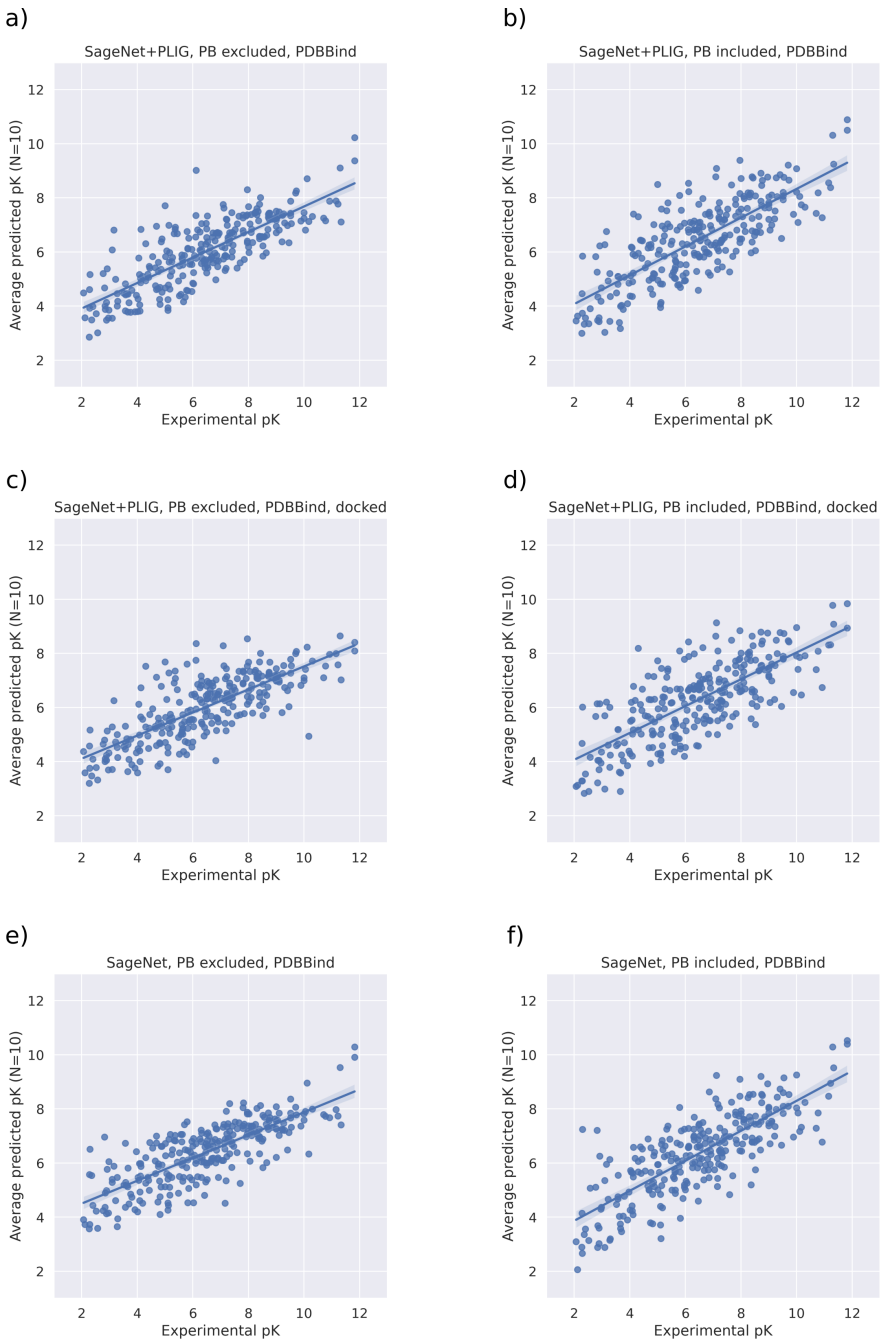

Figure C.18: Scatter plots of the average prediction versus the experimentally determined pK value for each protein-ligand complex between the 10 model runs for all SageNet models. a) PLIG, no protein sequence embedding, trained and tested on crystal structures; b) PLIG, with sequence embedding, trained and tested on crystal structures; c) PLIG, no protein sequence embedding, trained and tested on docked structures; d) PLIG, with protein sequence embedding, trained and tested on docked structures; e) ligand-based graph, no protein sequence embedding; f) ligand-based graph, with protein embedding.

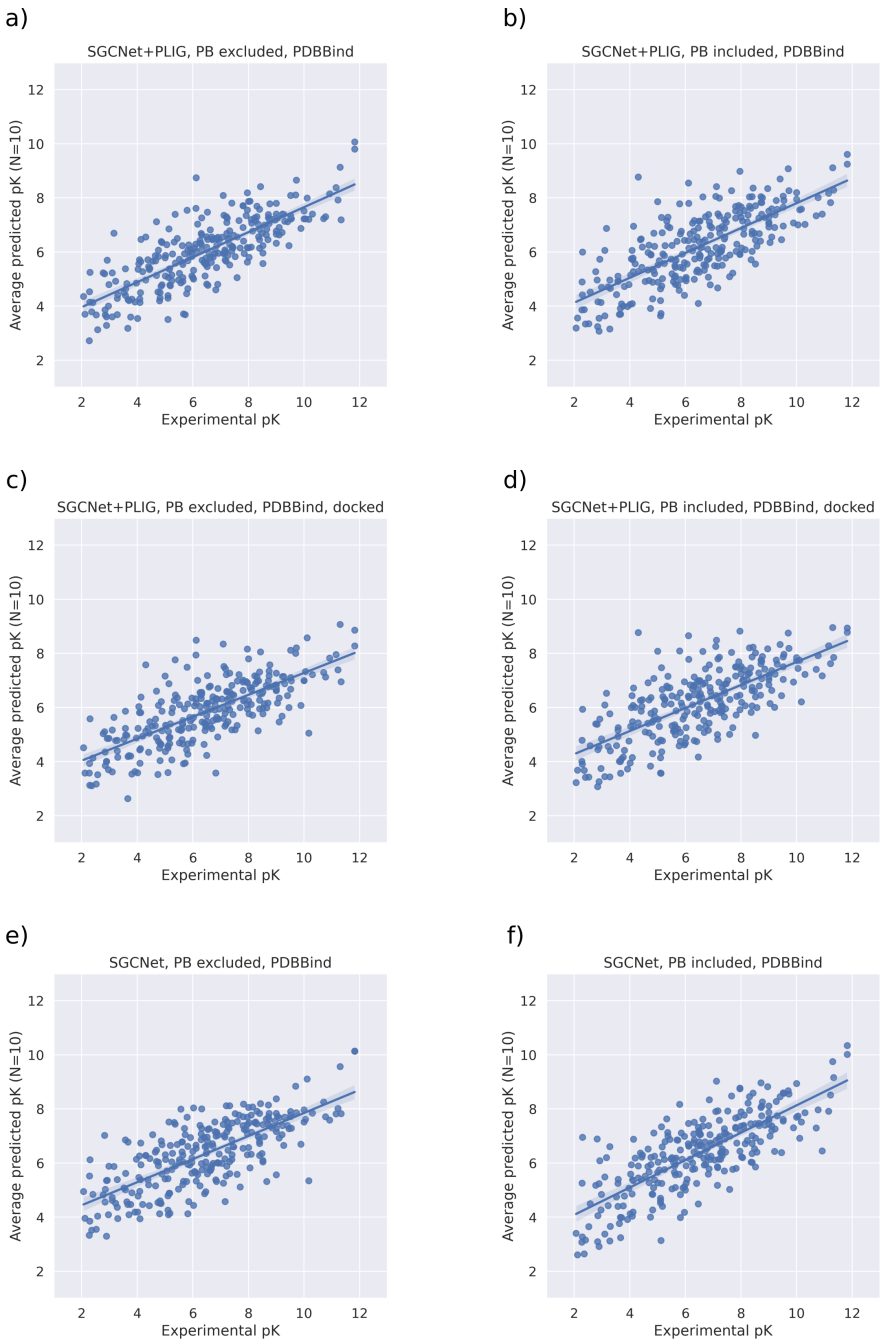

Figure C.19: Scatter plots of the average prediction versus the experimentally determined pK value for each protein-ligand complex between the 10 model runs for all SGCNet models. a) PLIG, no protein sequence embedding, trained and tested on crystal structures; b) PLIG, with sequence embedding, trained and tested on crystal structures; c) PLIG, no protein sequence embedding, trained and tested on docked structures; d) PLIG, with protein sequence embedding, trained and tested on docked structures; e) ligand-based graph, no protein sequence embedding; f) ligand-based graph, with protein embedding.

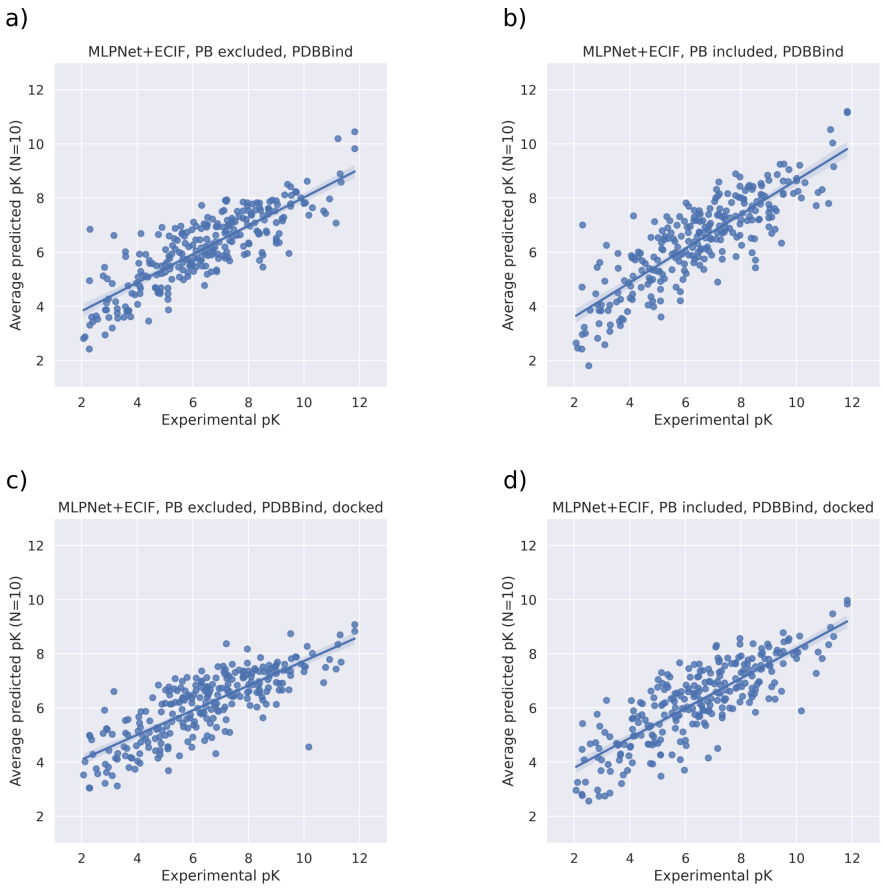

Figure C.20: Scatter plots of the average prediction versus the experimentally determined pK value for each protein-ligand complex between the 10 model runs for all MLPNet + ECIF models. a) no protein sequence embedding, trained and tested on crystal structures; b) including sequence embedding, trained and tested on crystal structures; c) no protein sequence embedding, trained and tested on docked structures; d) including protein sequence embedding, trained and tested on docked structures.

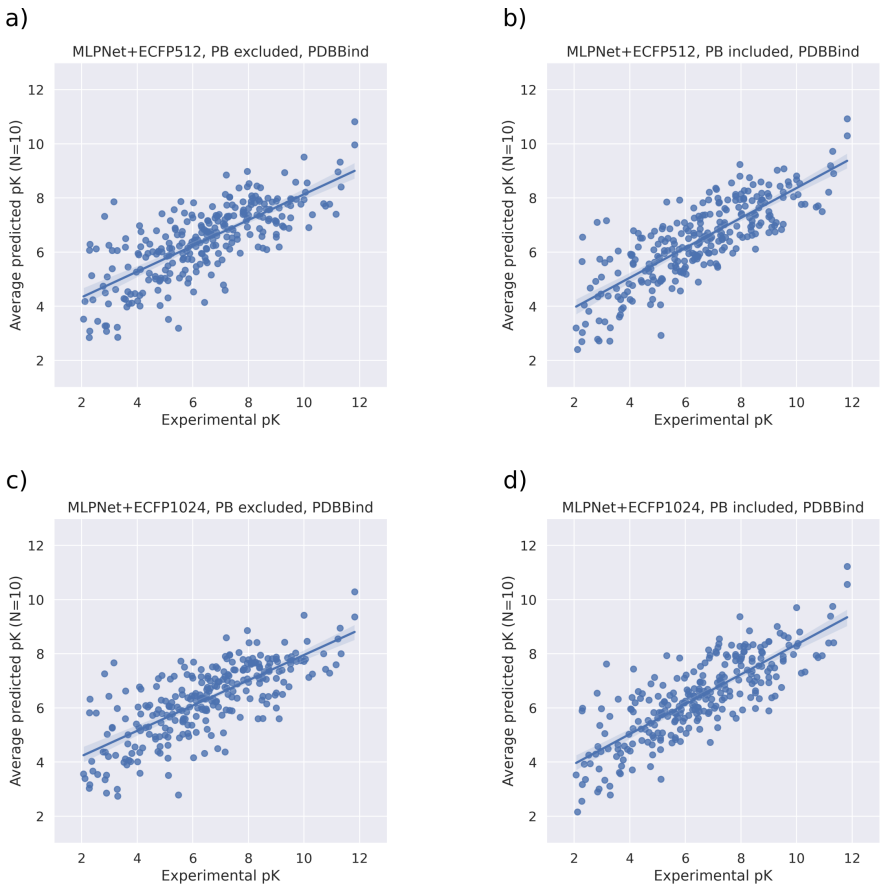

Figure C.21: Scatter plots of the average prediction versus the experimentally determined pK value for each protein-ligand complex between the 10 model runs for all MLPNet + ECFP models. a) ECFP512, no protein sequence embedding, trained and tested on ligand information; b) ECFP512, including sequence embedding, trained and tested on ligand information; c) ECFP1024, no protein sequence embedding, trained and tested on ligand information; d) ECFP1024, including protein sequence embedding, trained and tested on ligand information

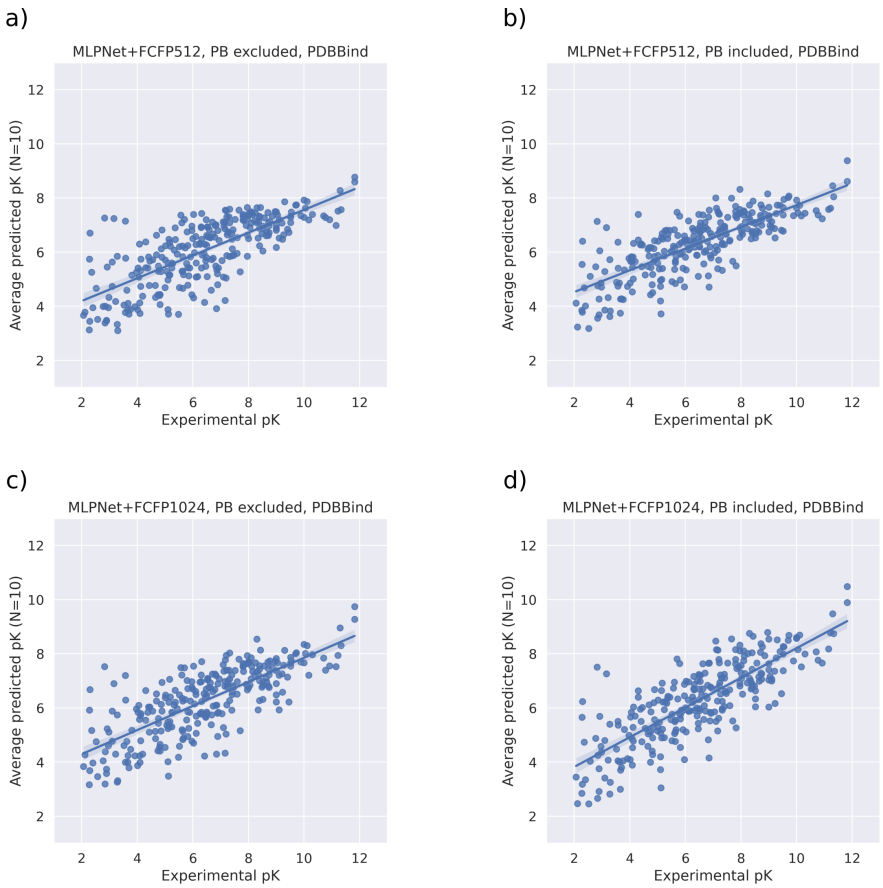

Figure C.22: Scatter plots of the average prediction versus the experimentally determined pK value for each protein-ligand complex between the 10 model runs for all MLPNet FCFP models. a) FCFP512, no protein sequence embedding, trained and tested on ligand information; b) FCFP512, including sequence embedding, trained and tested on ligand information; c) FCFP1024, no protein sequence embedding, trained and tested on ligand information; d) FCFP1024, including protein sequence embedding, trained and tested on ligand information

C.5 Proximity Analysis

C.5.1 Cross Validation

5-fold cross validation was performed on the combined PDBBind general 2020 and PDBBind refined 2016 set (dataset details in the main text Section 4.3.1). The validation and test set (CASF-2016) used in the main study was removed from the cross validation set, to leave the training set of 14254 compounds. Cross validation was done separately for the docked and crystal-based datasets to ensure model stability

for both, crystal derived structures and docked poses. This set was split into 5 random folds, using 20 % of the dataset as validation in each fold. Models were run across all 5 folds until model performance converged. The performance of all models during cross validation is recorded for every epoch. The number of epochs to train each model for training and test on the CASF-2016 benchmark was determined as the point where no significant performance increase (difference of less than 0.01 Pearson correlation coefficient between epoch n and $n-1$). The GATNet model was chosen without the protein sequence embedding branch as the architecture to try out different PLIG proximity thresholds since it was the best performing model in the main study. The optimal number of epochs as determined by cross validation is shown in Table C.21.

Proximity threshold	**Epochs (Crystal)**	**Epochs (Docked)**
PLIG 4 Å	12	11
PLIG 5 Å	14	9
PLIG 6 Å	16	11
PLIG 7 Å	14	15
PLIG 8 Å	15	13

Table C.21: Optimal number of epochs for different PLIG thresholds when trained and tested on docked and crystal poses using the GATNet PLIG architecture without the protein sequence embedding.

The performance of the 4,5,7 and 8 Å GATNet PLIG models during cross validation is shown in Figure C.23 & C.24. Details on cross validation for the 6 Å PLIG model can be found in the general cross validations Section C.3 above.

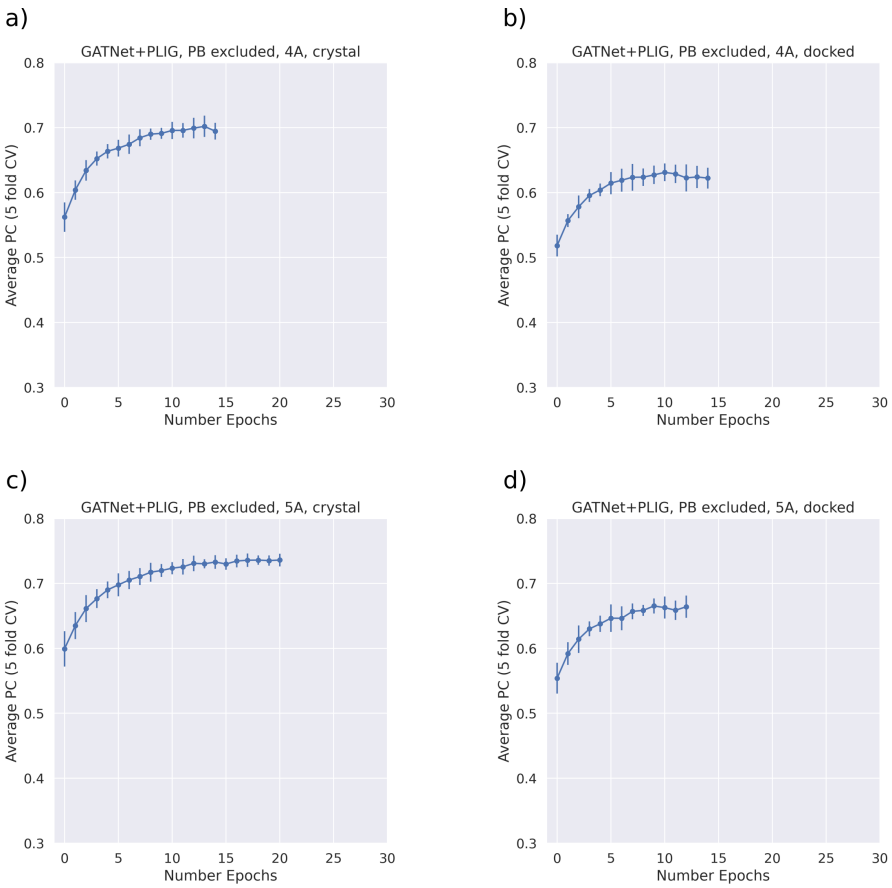

Figure C.23: 5-fold cross validation of the GATNet PLIG models (no sequence) with 4 and 5 Å proximity thresholds reported as the average calculated pearson correlation coefficient and its corresponding standard deviation (error bars) at each epoch. a) 4 Å threshold, trained and tested on crystal poses; b) 4 Å threshold, trained and tested on docked poses; c) 5 Å threshold, trained and tested on crystal poses ; d) 5 Å threshold, trained and tested on docked poses.

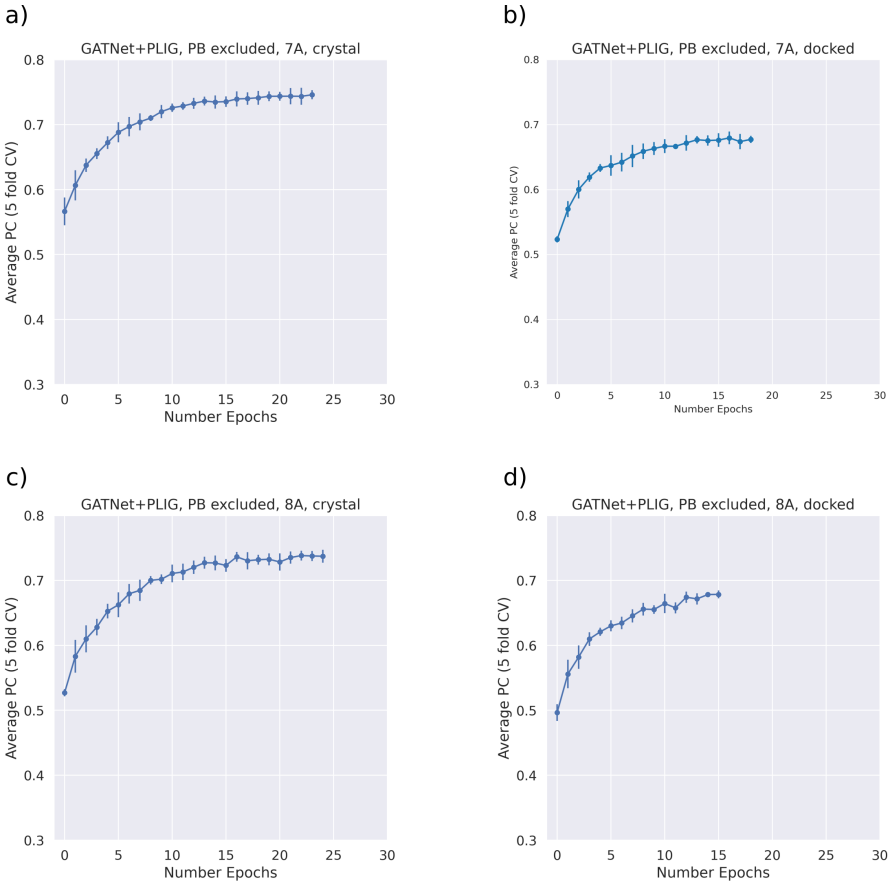

Figure C.24: 5-fold cross validation performance of the GATNet PLIG models (no sequence) with 7 and 8 Å proximity thresholds reported as the average calculated pearson correlation coefficient and its corresponding standard deviation (error bars) at each epoch. a) 7 Å threshold, trained and tested on crystal poses; b) 7 Å threshold, trained and tested on docked poses; c) 8 Å threshold, trained and tested on crystal poses ; d) 8 Å threshold, trained and tested on docked poses.

C.5.2 Performance and Stability over 10 Stochastic Runs

All models were trained on the PDBBind dataset (description see main text Section 4.3.1) and tested on the CASF-2016 benchmark set (crystal case is trained and tested on crystal structures, docked case is trained and tested on docked poses and the ligand case does not use 3D information). Since the models' predictions are somewhat stochastic, model performance and stability against the withheld test set (CASF-2016) for all trained models was evaluated using the average and standard deviation (SD) of the pearson correlation coefficient (ρ) as well as the root-mean-square error

(RMSE) over 10 independent runs. ρ and RMSE and their corresponding standard deviations are shown in figure C.25 and figure C.26. Overall, the 5 and 6 Å representations reach similar performance metrics (ρ = 0.80 / 0.80, RMSE = 1.33 / 1.32 for 5 / 6 Å respectively) however the 6 Å model is slightly more stable with a lower standard deviation for the RMSE (SD RMSE = 0.06 / 0.04 for 5 / 6 Å respectively). The difference in ρ and standard deviation of the ρ is insignificant between the two models. In addition, performance of all models against docked poses is extremely similar and overall lower than the crystal poses. This observation is in line with our expectations since a majority of docked poses were found to have a RMSD between the crystal and docked pose of larger than 2 Å (Main text Section 3.1). A threshold change of the same magnitude should therefore not alter results drastically, as most poses have an equally large inaccuracy in their pose.

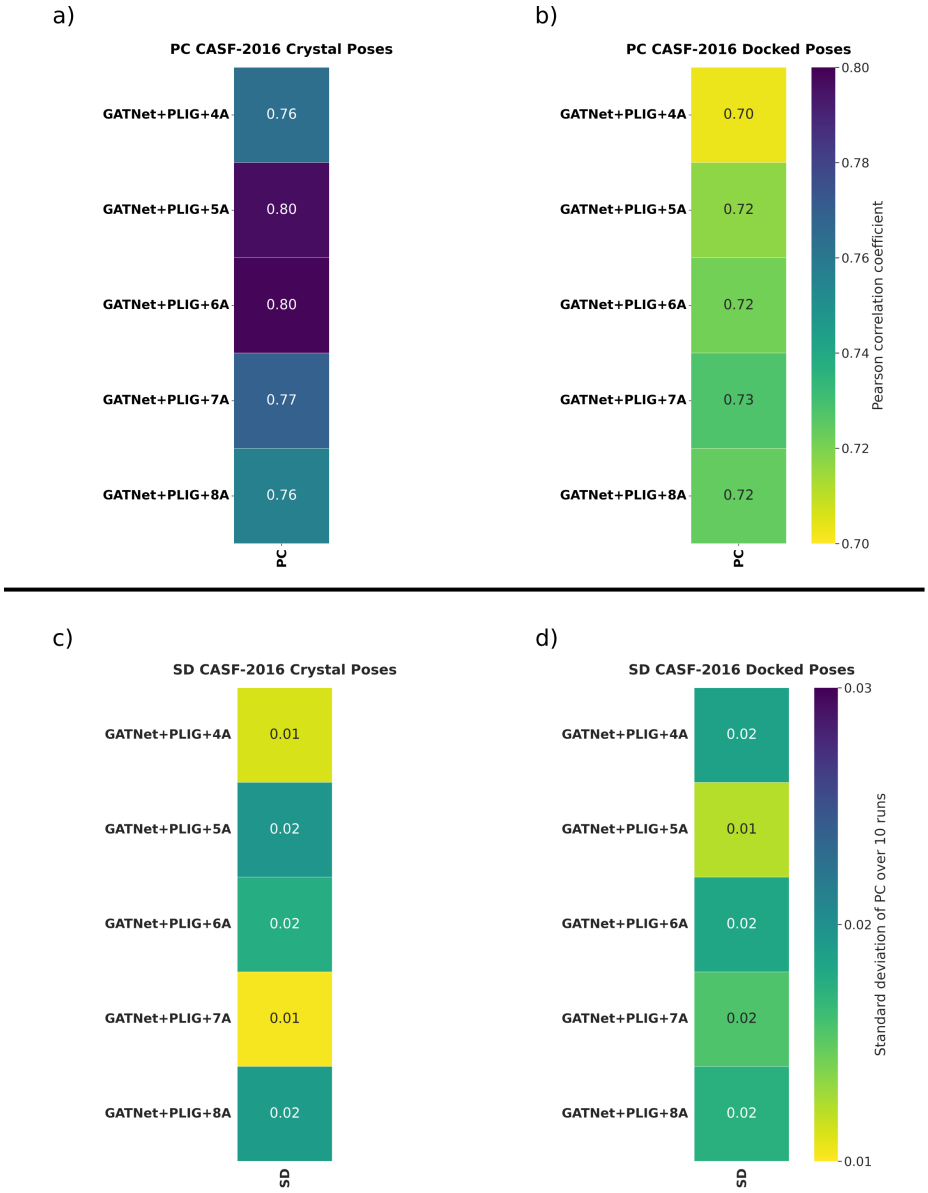

Figure C.25: Model stability of different threshold GATNet PLIG models measured by *rho* and its corresponding SD. All GATNet models were used without the second protein sequence embedding branch implementation. (a and c) ρ and SD of the GATNet PLIG models with different proximity thresholds when trained and tested on crystal structures. (b and d) ρ and SD of the GATNet PLIG models with different proximity thresholds when trained and tested on docked poses. Model stability as measured by ρ standard deviation varied between 0.01 (GATNet PLIG 7 Å trained/tested on crystal poses) and 0.02 (GATNet PLIG 5 Å, trained/tested on crystal poses).

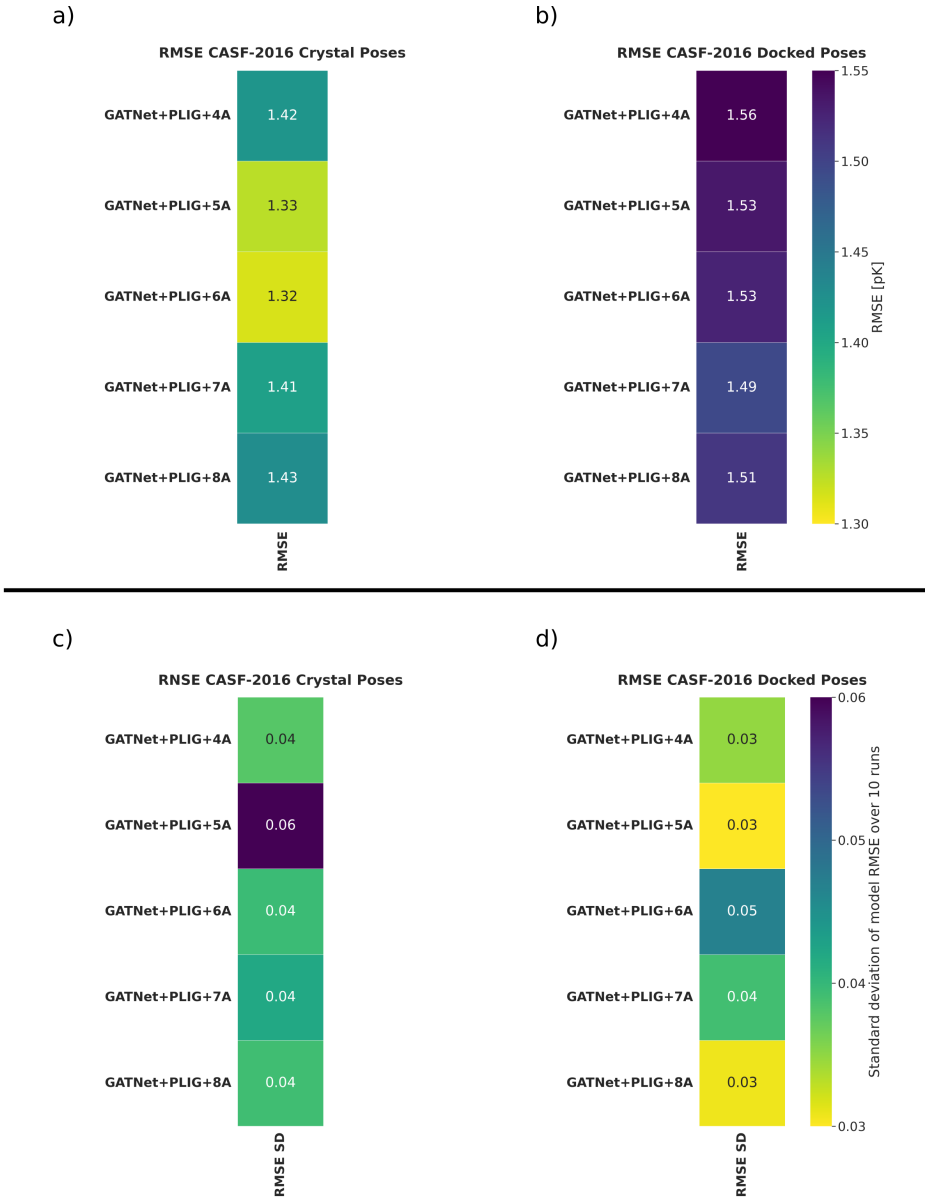

Figure C.26: Model stability of different threshold GATNet PLIG models measured by RMSE and its corresponding SD. All GATNet models were used without the second protein sequence embedding branch implementation. (a and c) RMSE and SD of the GATNet PLIG models with different proximity thresholds when trained and tested on crystal structures. (b and d) RMSE and SD of the GATNet PLIG models with different proximity thresholds when trained and tested on docked poses. Model stability as measured by ρ standard deviation varied between 0.03 (GATNet PLIG 4,5 and 8 Å trained/tested on docked poses) and 0.06 (GATNet PLIG 5 Å trained/tested on crystal poses).

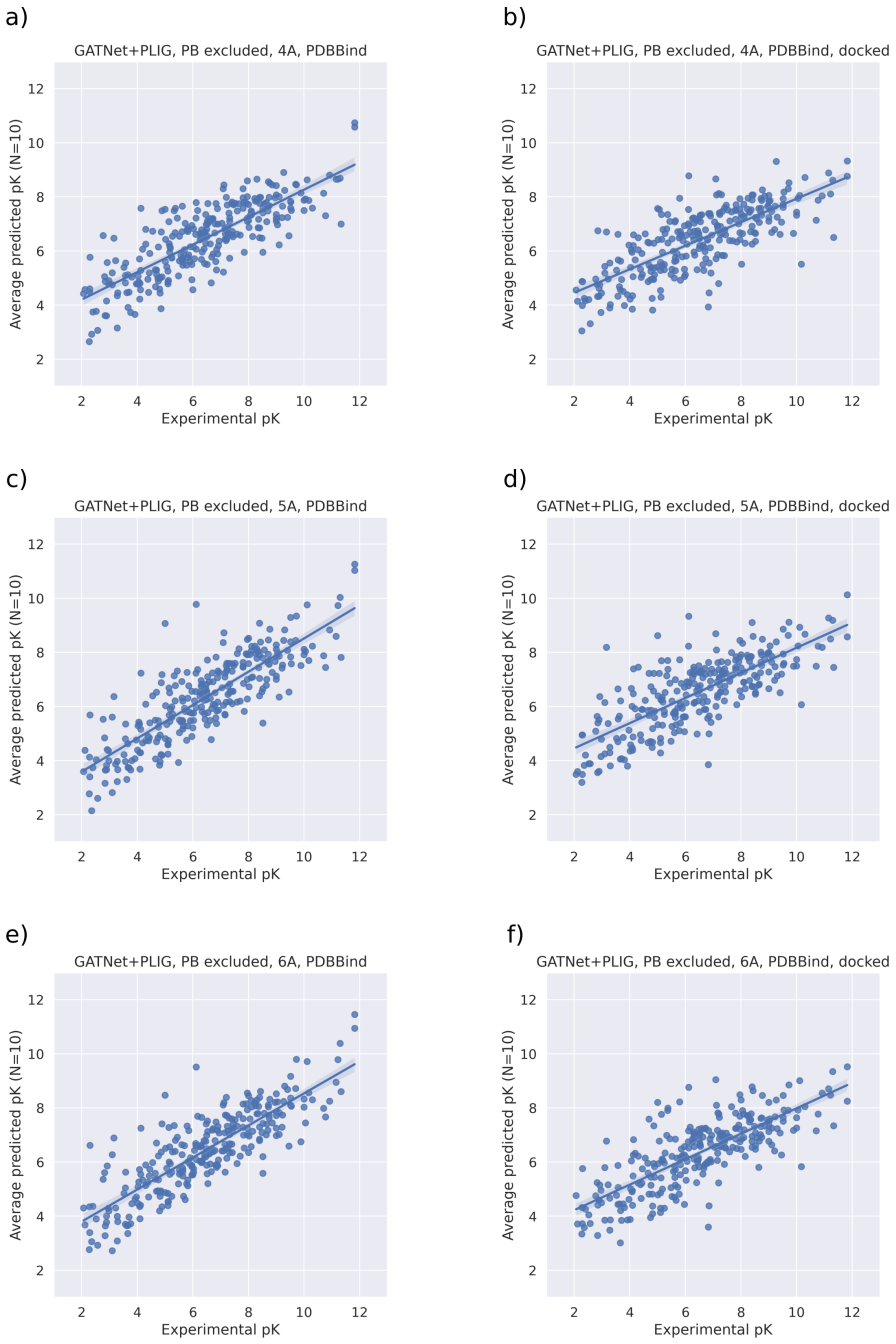

Figure C.27: Scatter plots of the average prediction versus the experimentally determined pK value for each protein-ligand complex between the 10 model runs for the GATNet PLIG model (no sequence) at different proximity thresholds. a) GATNet PLIG, 4 Å threshold, trained and tested on crystal structures; b) GATNet PLIG, 4 Å threshold, trained and tested on docked poses; c) GATNet PLIG, 5 Å threshold, trained and tested on crystal structures; d) GATNet PLIG, 5 Å threshold, trained and tested on docked poses; e) GATNet PLIG, 6 Å threshold, trained and tested on crystal structures (same as Figure C.15 a); f) GATNet PLIG, 4 Å threshold, trained and tested on docked poses (same data as Figure C.15 b).

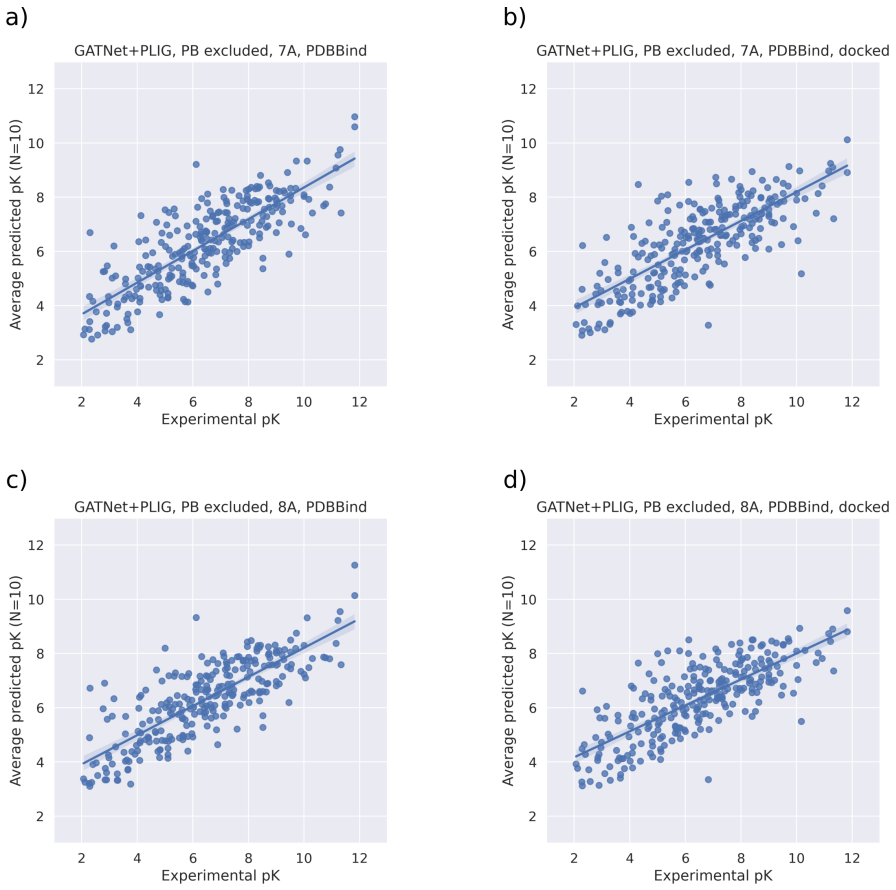

Figure C.28: Scatter plots of the average prediction versus the experimentally determined pK value for each protein-ligand complex between the 10 model runs for the GATNet PLIG model (no sequence) at different proximity thresholds. a) GATNet PLIG, 7 Å threshold, trained and tested on crystal structures; b) GATNet PLIG, 7 Å threshold, trained and tested on docked poses; c) GATNet PLIG, 8 Å threshold, trained and tested on crystal structures; d) GATNet PLIG, 8 Å threshold, trained and tested on docked poses.

C.6 Sequence Similarity Threshold Experiments

In order to assess the ability of the best performing GATNet PLIG model to generalize between different protein families, protein-ligand pairs in the training dataset were eliminated based on their sequence identity to proteins represented in the CASF-2016 benchmark using 5 threshold levels between 50-100 % identity. The GATNet PLIG (no sequence) model was trained on the reduced dataset and tested against the full CASF-2016 dataset. The average PC and RMSE over 10 model runs as well as the corresponding standard deviation is shown in Figure C.29. The scatter plots of the average prediction versus the experimentally determined pK value for each protein-ligand complex between the 10 model runs for each sequence similarity threshold are shown in Figure C.30. In addition to a decrease in performance with increasing strictness of the threshold, the standard deviation between model runs is increasing as well, indicating that models are more unstable as dataset size decreases and similarity between training and test set decreases.

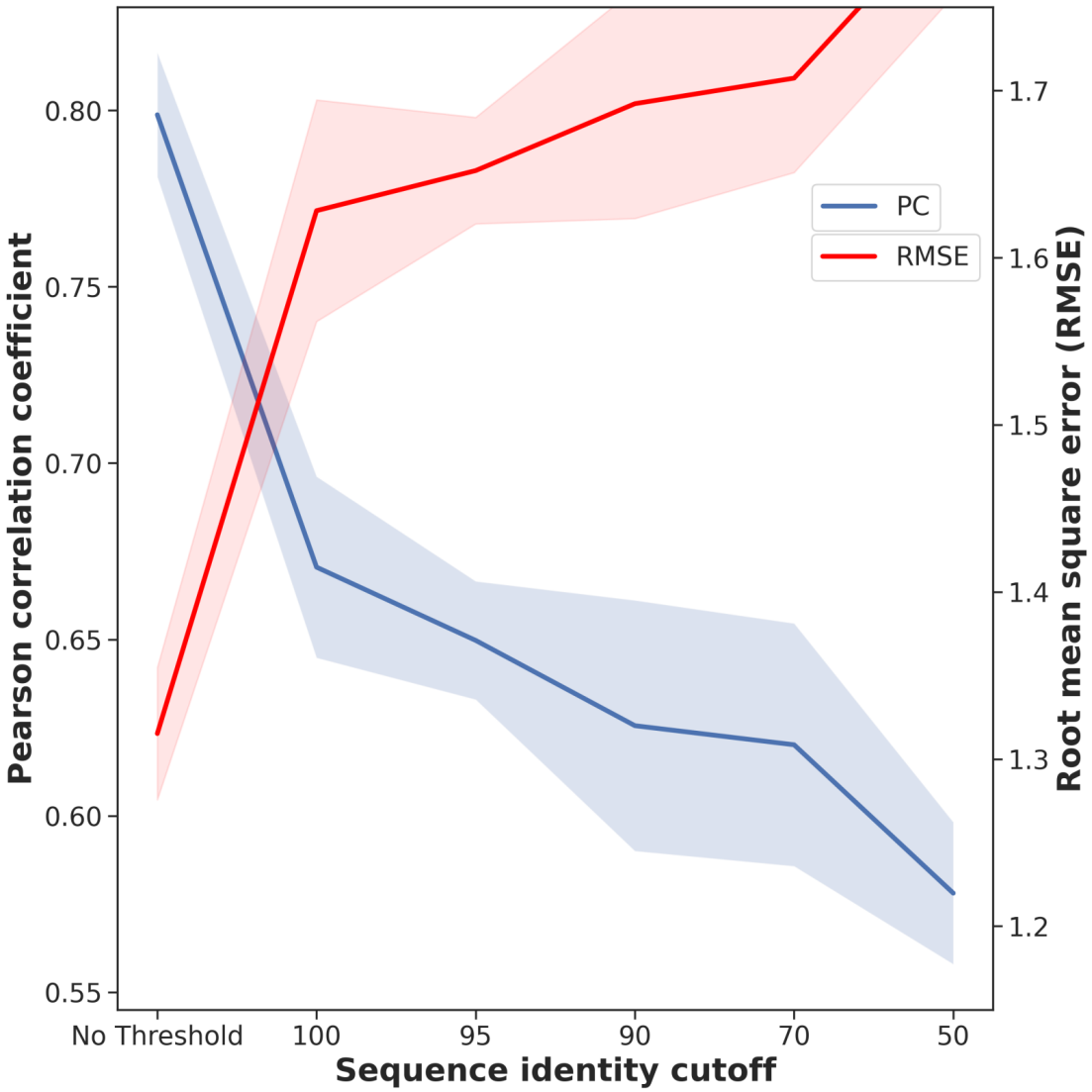

Figure C.29: Pearson correlation coefficient (ρ) and Root-mean-square error (RMSE) of predicted versus experimental binding affinity for the GATNet PLIG model (no sequence, average and SD between 10 runs) when trained and tested on crystal poses. Protein-ligand complexes in the training set with a sequence identity at or above the cut-off value to proteins in the CASF-2016 test set were excluded resulting in a smaller dataset at every step.

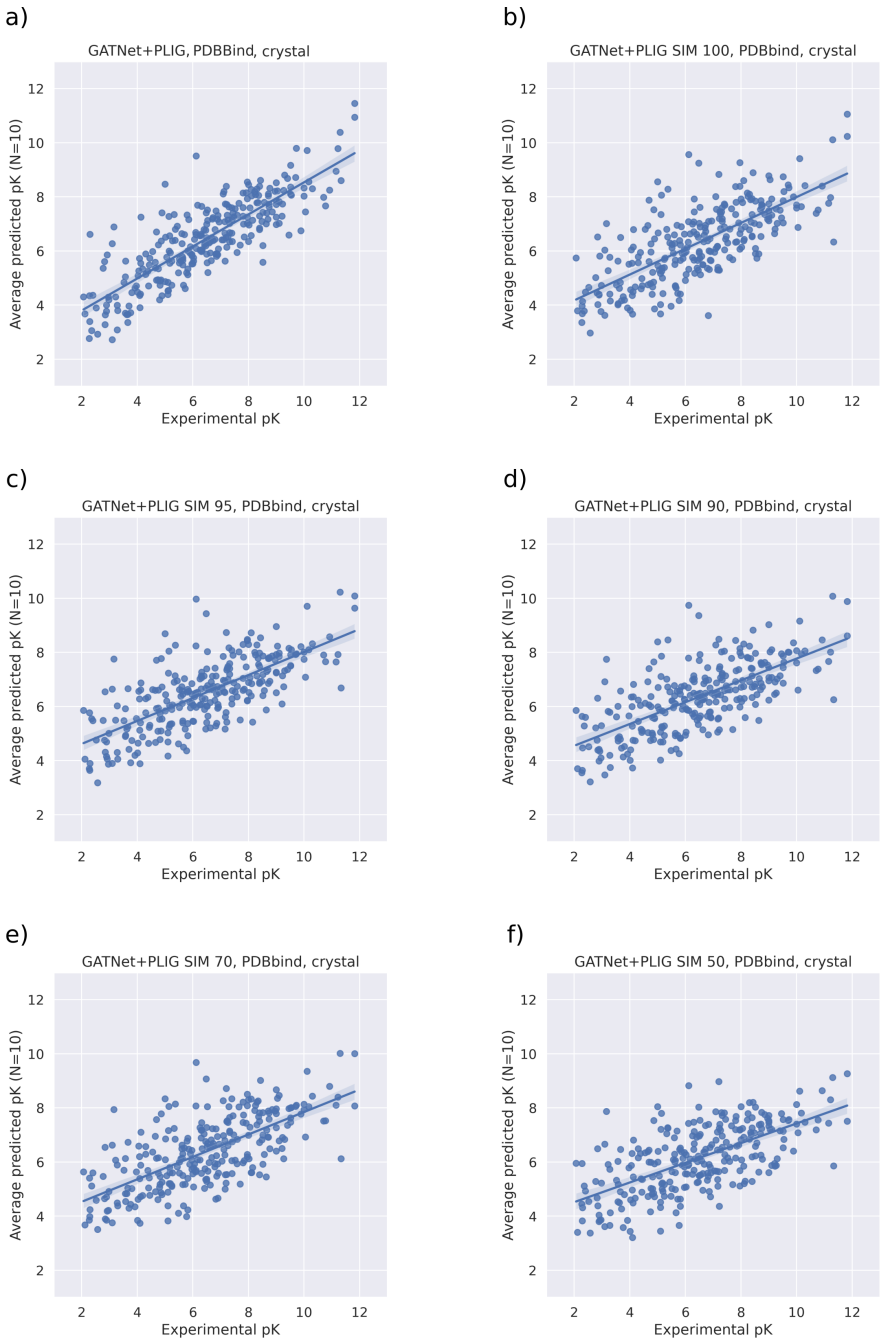

Figure C.30: Scatter plots of the average prediction versus the experimentally determined pK value for each protein-ligand complex between the 10 model runs for each sequence similarity threshold. All models are GATNet PLIG models with no protein sequence embedding. a) No threshold, full training set; b) similarity threshold of 100%; c) similarity threshold of 95%; d) similarity threshold of 90%; e) similarity threshold of 70%; f) similarity threshold of 50%

Bibliography

Abel, R., Paredes Ramos, M., Chen, Q., Pérez-Sánchez, H., Coluzzi, F., Rocco, M., Marchetti, P., Mura, C., Simmaco, M., Bourne, P. E., Preissner, R., and Banerjee, P. (2020). Computational Prediction of Potential Inhibitors of the Main Protease of SARS-CoV-2. *Frontiers in Chemistry*, 8(1162).

Abraham, M. J., Murtola, T., Schulz, R., Páll, S., Smith, J. C., Hess, B., and Lindahl, E. (2015). GROMACS: High Performance Molecular Simulations Through Multi-Level Parallelism from Laptops to Supercomputers. *SoftwareX*, 1-2:19–25.

Acharya, A., Agarwal, R., Baker, M. B., Baudry, J., Bhowmik, D., Boehm, S., Byler, K. G., Chen, S. Y., Coates, L., Cooper, C. J., Demerdash, O., Daidone, I., Eblen, J. D., Ellingson, S., Forli, S., Glaser, J., Gumbart, J. C., Gunnels, J., Hernandez, O., Irle, S., Kneller, D. W., Kovalevsky, A., Larkin, J., Lawrence, T. J., LeGrand, S., Liu, S. H., Mitchell, J. C., Park, G., Parks, J. M., Pavlova, A., Petridis, L., Poole, D., Pouchard, L., Ramanathan, A., Rogers, D. M., Santos-Martins, D., Scheinberg, A., Sedova, A., Shen, Y., Smith, J. C., Smith, M. D., Soto, C., Tsaris, A., Thavappiragasam, M., Tillack, A. F., Vermaas, J. V., Vuong, V. Q., Yin, J., Yoo, S., Zahran, M., and Zanetti-Polzi, L. (2020). Supercomputer-Based Ensemble Docking Drug Discovery Pipeline with Application to Covid-19. *Journal of Chemical Information and Modeling*, 60(12):5832–5852.

Achdout, H., Aimon, A., Bar-David, E., Barr, H., Ben-Shmuel, A., Bennett, J.,

Bobby, M. L., Brun, J., Sarma, B., Calmiano, M., Carbery, A., Cattermole, E., Chodera, J. D., Clyde, A., Coffland, J. E., Cohen, G., Cole, J., Contini, A., Cox, L., Cvitkovic, M., Dias, A., Douangamath, A., Duberstein, S., Dudgeon, T., Dunnett, L., Eastman, P. K., Erez, N., Fairhead, M., Fearon, D., Fedorov, O., Ferla, M., Foster, H., Foster, R., Gabizon, R., Gehrtz, P., Gileadi, C., Giroud, C., Glass, W. G., Glen, R., Glinert, I., Gorichko, M., Gorrie-Stone, T., Griffen, E. J., Heer, J., Hill, M., Horrell, S., Hurley, M. F. D., Israely, T., Jajack, A., Jnoff, E., John, T., Kantsadi, A. L., Kenny, P. W., Kiappes, J. L., Koekemoer, L., Kovar, B., Krojer, T., Lee, A. A., Lefker, B. A., Levy, H., London, N., Lukacik, P., Macdonald, H. B., MacLean, B., Malla, T. R., Matviiuk, T., McCorkindale, W., Melamed, S., Michurin, O., Mikolajek, H., Morris, A., Morris, G. M., Morwitzer, M. J., Moustakas, D., Neto, J. B., Oleinikovas, V., Overheul, G. J., Owen, D., Pai, R., Pan, J., Paran, N., Perry, B., Pingle, M., Pinjari, J., Politi, B., Powell, A., Psenak, V., Puni, R., Rangel, V. L., Reddi, R. N., Reid, S. P., Resnick, E., Robinson, M. C., Robinson, R. P., Rufa, D., Schofield, C., Shaikh, A., Shi, J., Shurrush, K., Sittner, A., Skyner, R., Smalley, A., Smilova, M. D., Spencer, J., Strain-Damerell, C., Swamy, V., Tamir, H., Tennant, R., Thompson, A., Thompson, W., Tomasio, S., Tumber, A., Vakonakis, I., van Rij, R. P., Varghese, F. S., Vaschetto, M., Vitner, E. B., Voelz, V., von Delft, A., von Delft, F., Walsh, M., Ward, W., Weatherall, C., Weiss, S., Wild, C. F., Wittmann, M., Wright, N., Yahalom-Ronen, Y., Zaidmann, D., Zidane, H., and Zitzmann, N. (2020). COVID Moonshot: Open Science Discovery of SARS-CoV-2 Main Protease Inhibitors by Combining Crowdsourcing, High-Throughput Experiments, Computational Simulations, and Machine Learning. *bioRxiv*, page 2020.10.29.339317.

Akiba, T., Sano, S., Yanase, T., Ohta, T., and Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th*

ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pages 2623–2631.

Aldeghi, M., Heifetz, A., Bodkin, M. J., Knapp, S., and Biggin, P. C. (2016). Accurate calculation of the absolute free energy of binding for drug molecules. *Chemical Science*, 7:207–218.

Alzubaidi, L., Zhang, J., Humaidi, A. J., Al-Dujaili, A., Duan, Y., Al-Shamma, O., Santamaría, J., Fadhel, M. A., Al-Amidie, M., and Farhan, L. (2021). Review of Deep Learning: Concepts, CNN Architectures, Challenges, Applications, Future Directions. *Journal of Big Data*, 8(1):53.

Anand, K., Palm, G. J., Mesters, J. R., Siddell, S. G., Ziebuhr, J., and Hilgenfeld, R. (2002). Structure of Coronavirus Main Proteinase Reveals Combination of a Chymotrypsin Fold with an Extra α-Helical Domain. *The EMBO Journal*, 21(13):3213–3224.

Anandakrishnan, R., Aguilar, B., and Onufriev, A. V. (2012). H++ 3.0: Automating pK Prediction and the Preparation of Biomolecular Structures for Atomistic Molecular Modeling and Simulations. *Nucleic Acids Research*, 40(W1):W537–W541.

Arafet, K., Serrano-Aparicio, N., Lodola, A., Mulholland, A. J., González, F. V., Świderek, K., and Moliner, V. (2021). Mechanism of Inhibition of SARS-CoV-2 M^{pro} by N3 Peptidyl Michael Acceptor Explained by QM/MM Simulations and Design of New Derivatives with Tunable Chemical Reactivity. *Chemical Science*, 12(4):1433–1444.

Ar´us-Pous, J., Blaschke, T., Ulander, S., Reymond, J.-L., Chen, H., and Engkvist, O. (2019). Exploring the GDB-13 Chemical Space Using Deep Generative Models. *Journal of Cheminformatics*, 11(1).

Arús-Pous, J., Johansson, S. V., Prykhodko, O., Bjerrum, E. J., Tyrchan, C., Reymond, J.-L., Chen, H., and Engkvist, O. (2019). Randomized SMILES Strings Improve the Quality of Molecular Generative Models. *Journal of Cheminformatics*, 11(1):71.

Baell, J. B. and Holloway, G. A. (2010). New Substructure Filters for Removal of Pan Assay Interference Compounds (PAINS) from Screening Libraries and for Their Exclusion in Bioassays. *Journal of Medicinal Chemistry*, 53(7):2719–2740.

Ballester, P. J. and Mitchell, J. B. O. (2010). A Machine Learning Approach to Predicting Protein-Ligand Binding Affinity with Applications to Molecular Docking. *Bioinformatics*, 26(9):1169–1175.

Bellmann, L., Penner, P., and Rarey, M. (2019). Connected Subgraph Fingerprints: Representing Molecules Using Exhaustive Subgraph Enumeration. *Journal of Chemical Information and Modeling*, 59(11):4625–4635.

Bergstra, J., Bardenet, R., Bengio, Y., and Kégl, B. (2011). Algorithms for hyper-parameter optimization. *Advances in Neural Information Processing Systems*, 24.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000). The Protein Data Bank. *Nucleic Acids Research*, 28(1):235–242.

Blaschke, T., Olivecrona, M., Engkvist, O., Bajorath, J., and Chen, H. (2018). Application of Generative Autoencoder in *de novo* Molecular Design. *Molecular Informatics*, 37(1-2):1700123.

Boobier, S., Hose, D. R. J., Blacker, A. J., and Nguyen, B. N. (2020). Machine Learning with Physicochemical Relationships: Solubility Prediction in Organic Solvents and Water. *Nature Communications*, 11(1):5753.

Boran, A. D. W. and Iyengar, R. (2010). Systems Approaches to Polypharmacology and Drug Discovery. *Current Opinion in Drug Discovery & Development*, 13(3):297–309.

Boyles, F. (University of Oxford, Department of Statistics, 2020). *Developing Novel Scoring Functions for Protein-Ligand Docking Using Machine Learning*. PhD thesis, University of Oxford.

Boyles, F., Deane, C. M., and Morris, G. M. (2019). Learning from the Ligand: Using Ligand-Based Features to Improve Binding Affinity Prediction. *Bioinformatics*, 36(3):758–764.

Boyles, F., Deane, C. M., and Morris, G. M. (2021). Learning from Docked Ligands: Ligand-Based Features Rescue Structure-Based Scoring Functions when Trained on Docked Poses. *Journal of Chemical Information and Modeling*.

Castelli, R., Tognolini, M., Vacondio, F., Incerti, M., Pala, D., Callegari, D., Bertoni, S., Giorgio, C., Hassan-Mohamed, I., Zanotti, I., Bugatti, A., Rusnati, M., Festuccia, C., Rivara, S., Barocelli, E., Mor, M., and Lodola, A. (2015). $\Delta(5)$-Cholenoyl-Amino Acids as Selective and Orally Available Antagonists of the Eph-Ephrin System. *European Journal of Medical Chemistry*, 103:312–324.

Chan, H. T. H., Moesser, M. A., Walters, R. K., Malla, T. R., Twidale, R. M., John, T., Deeks, H. M., Johnston-Wood, T., Mikhailov, V., Sessions, R. B., Dawson, W., Salah, E., Lukacik, P., Strain-Damerell, C., Owen, C. D., Nakajima, T., Świderek, K., Lodola, A., Moliner, V., Glowacki, D. R., Spencer, J., Walsh, M. A., Schofield, C. J., Genovese, L., Shoemark, D. K., Mulholland, A. J., Duarte, F., and Morris, G. M. (2021a). Discovery of SARS-CoV-2 M^{pro} Peptide Inhibitors from Modelling Substrate and Ligand Binding. *Chemical Science*, 12(41):13686–13703.

Chan, H. T. H., Moesser, M. A., Walters, R. K., Malla, T. R., Twidale, R. M., John, T., Deeks, H. M., Johnston-Wood, T., Mikhailov, V., Sessions, R. B., Dawson, W., Salah, E., Lukacik, P., Strain-Damerell, C., Owen, C. D., Nakajima, T., Świderek, K., Lodola, A., Moliner, V., Glowacki, D. R., Spencer, J., Walsh, M. A., Schofield, C. J., Genovese, L., Shoemark, D. K., Mulholland, A. J., Duarte, F., and Morris, G. M. (2021b). Discovery of SARS-CoV-2 M^{pro} Peptide Inhibitors from Modelling Substrate and Ligand Binding. `https://github.com/gmm/SARS-CoV-2-Modelling`.

Chemical Computing Group ULC. (2019). Molecular Operating Environment (MOE) version 2019.0104.

Cherkasov, A., Muratov, E. N., Fourches, D., Varnek, A., Baskin, I. I., Cronin, M., Dearden, J., Gramatica, P., Martin, Y. C., Todeschini, R., Consonni, V., Kuz'min, V. E., Cramer, R., Benigni, R., Yang, C., Rathman, J., Terfloth, L., Gasteiger, J., Richard, A., and Tropsha, A. (2014). QSAR Modeling: Where Have You Been? Where Are You Going To? *Journal of Medicinal Chemistry*, 57(12):4977–5010.

Cho, E., Rosa, M., Anjum, R., Mehmood, S., Soban, M., Mujtaba, M., Bux, K., Moin, S. T., Tanweer, M., Dantu, S., Pandini, A., Yin, J., Ma, H., Ramanathan, A., Islam, B., Mey, A. S. J. S., Bhowmik, D., and Haider, S. (2021). Dynamic Profiling of β-Coronavirus 3CL M^{pro} Protease Ligand-Binding Sites. *Journal of Chemical Information and Modeling*.

Chodera, J., Lee, A. A., London, N., and von Delft, F. (2020). Crowdsourcing Drug Discovery for Pandemics. *Nature Chemistry*, 12(7):581.

Chuaqui, C., Deng, Z., and Singh, J. (2005). Interaction Profiles of Protein Kinase-

Inhibitor Complexes and Their Application to Virtual Screening. *Journal of Medicinal Chemistry*, 48(1):121–133.

Consortium, T. G. (2012, accessed: Jan 2020a). GPy: A Gaussian Process Framework in Python. `http://github.com/SheffieldML/GPy`.

Consortium, T. G. (2016, accessed: Jan 2020b). GPyOpt: A Bayesian Optimization Framework in Python. `http://github.com/SheffieldML/GPyOpt`.

Cook, D., Brown, D., Alexander, R., March, R., Morgan, P., Satterthwaite, G., and Pangalos, M. N. (2014). Lessons Learned From the Fate of AstraZeneca's Drug Pipeline: a Five-Dimensional Framework. *Nature Reviews Drug Discovery*, 13(6):419–431.

Cornell, W. D., Cieplak, P., Bayly, C. I., Gould, I. R., Merz, K. M., Ferguson, D. M., Spellmeyer, D. C., Fox, T., Caldwell, J. W., and Kollman, P. A. (1995). A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *Journal of the American Chemical Society*, 117(19):5179–5197.

Cortés-Ciriano, I., Ain, Q. U., Subramanian, V., Lenselink, E. B., Méndez-Lucio, O., IJzerman, A. P., Wohlfahrt, G., Prusis, P., Malliavin, T. E., van Westen, G. J. P., and Bender, A. (2015). Polypharmacology Modelling Using Proteochemometrics (PCM): Recent Methodological Developments, Applications to Target Families, and Future Prospects. *Medicinal Chemistry Communications*, 6(1):24–50.

COVID-19 Moonshot project (2020). (accessed: July 2021), XChem Facility at UK's Diamond Light Source and PostEra.Ai. `https://github.com/postera-ai/COVID_moonshot_submissions`.

Cramer, R. D., Patterson, D. E., and Bunce, J. D. (1988). Comparative Molecular Field Analysis (CoMFA). 1. Effect of Shape on Binding of Steroids to Carrier Proteins. *Journal of the American Chemical Society*, 110(18):5959–5967.

Crooks, G. E., Hon, G., Chandonia, J. M., and Brenner, S. E. (2004). WebLogo: A Sequence Logo Generator. *Genome Research*, 14(6):1188–1190.

Daura, X., Gademann, K., Jaun, B., Seebach, D., van Gunsteren, W. F., and Mark, A. E. (1999). Peptide Folding: When Simulation Meets Experiment. *Angewandte Chemie International Edition*, 38(1-2):236–240.

Davies, M., Nowotka, M., Papadatos, G., Dedman, N., Gaulton, A., Atkinson, F., Bellis, L., and Overington, J. P. (2015). Chembl web services: Streamlining access to drug discovery data and utilities. *Nucleic Acids Research*, 43:W612–W620.

De Vivo, M., Masetti, M., Bottegoni, G., and Cavalli, A. (2016). Role of Molecular Dynamics and Related Methods in Drug Discovery. *Journal of Medicinal Chemistry*, 59(9):4035–4061.

De Wit, E., Van Doremalen, N., Falzarano, D., and Munster, V. J. (2016). SARS and MERS: Recent Insights into Emerging Coronaviruses. *Nature Reviews Microbiology*, 14(8).

Dean, R. A., Cox, J. H., Bellac, C. L., Doucet, A., Starr, A. E., and Overall, C. M. (2008). Macrophage-Specific Metalloelastase (MMP-12) Truncates and Inactivates ELR+ CXC Chemokines and Generates CCL2, -7, -8, and -13 Antagonists: Potential Role of the Macrophage in Terminating Polymorphonuclear Leukocyte Influx. *Blood*, 112(8):3455–3464.

Diamond (2020). Fragalysis (accessed January 2021). https://fragalysis.diamond.ac.uk/.

DiMasi, J. A., Grabowski, H. G., and Hansen, R. W. (2016). Innovation in the Pharmaceutical Industry: New Estimates of R&D Costs. *Journal of Health Economics*, 47:20–33.

Douangamath, A., Fearon, D., Gehrtz, P., Krojer, T., Lukacik, P., Owen, C. D., Resnick, E., Strain-Damerell, C., Aimon, A., Ábrányi-Balogh, P., Brandão-Neto, J., Carbery, A., Davison, G., Dias, A., Downes, T. D., Dunnett, L., Fairhead, M., Firth, J. D., Jones, S. P., Keeley, A., Keserü, G. M., Klein, H. F., Martin, M. P., Noble, M. E. M., O'Brien, P., Powell, A., Reddi, R. N., Skyner, R., Snee, M., Waring, M. J., Wild, C., London, N., von Delft, F., and Walsh, M. A. (2020). Crystallographic and Electrophilic Fragment Screening of the SARS-CoV-2 Main Protease. *Nature Communications*, 11(1):5047.

Dunbrack Jr., R. L. and Cohen, F. E. (1997). Bayesian Statistical Analysis of Protein Side-Chain Rotamer Preferences. *Protein Science*, 6(8):1661–1681.

Duran-Frigola, M., Pauls, E., Guitart-Pla, O., Bertoni, M., Alcalde, V., Amat, D., Juan-Blanco, T., and Aloy, P. (2020). Extending the Small-Molecule Similarity Principle to All Levels of Biology with the Chemical Checker. *Nature Biotechnology*, 38(9):1087–1096.

Durrant, J. D. and McCammon, J. A. (2011). NNScore 2.0: A Neural-Network Receptor–Ligand Scoring Function. *Journal of Chemical Information and Modeling*, 51(11):2897–2903.

Edgar, R. C. (2004). MUSCLE: Multiple Sequence Alignment with High Accuracy and High Throughput. *Nucleic Acids Research*, 32(5):1792–1797.

El-Baba, T. J., Lutomski, C. A., Kantsadi, A. L., Malla, T. R., John, T., Mikhailov, V., Bolla, J. R., Schofield, C. J., Zitzmann, N., Vakonakis, I., and Robinson,

C. V. (2020). Allosteric Inhibition of the SARS-CoV-2 Main Protease: Insights from Mass Spectrometry Based Assays. *Angewandte Chemie International Edition*, 59(52):23544–23548.

Elander, R. P. (2003). Industrial Production of β-Lactam Antibiotics. *Applied Microbiology and Biotechnology*, 61(5):385–392.

Erlanson, D. A., Fesik, S. W., Hubbard, R. E., Jahnke, W., and Jhoti, H. (2016). Twenty Years On: The Impact of Fragments on Drug Discovery. *Nature Reviews Drug Discovery*, 15(9):605–619.

Essmann, U., Perera, L., Berkowitz, M. L., Darden, T., Lee, H., and Pedersen, L. G. (1995). A Smooth Particle Mesh Ewald Method. *The Journal of Chemical Physics*, 103(19):8577–8593.

Fey, M. and Lenssen, J. E. (2019). Fast graph representation learning with PyTorch Geometric. In *ICLR Workshop on Representation Learning on Graphs and Manifolds*.

Forli, S. and Botta, M. (2007). Lennard-Jones Potential and Dummy Atom Settings to Overcome the AUTODOCK Limitation in Treating Flexible Ring Systems. *Journal of Chemical Information and Modeling*, 47(4):1481–1492.

Forli, S. and Olson, A. J. (2012). A Force Field with Discrete Displaceable Waters and Desolvation Entropy for Hydrated Ligand Docking. *Journal of Medicinal Chemistry*, 55(2):623–638.

Friesner, R. A., Banks, J. L., Murphy, R. B., Halgren, T. A., Klicic, J. J., Mainz, D. T., Repasky, M. P., Knoll, E. H., Shelley, M., Perry, J. K., Shaw, D. E., Francis, P., and Shenkin, P. S. (2004). Glide: A New Approach for Rapid, Accurate Docking

and Scoring. 1. Method and Assessment of Docking Accuracy. *Journal of Medicinal Chemistry*, 47(7):1739–1749.

Fujita, T., Iwasa, J., and Hansch, C. (1964). A new substituent constant, π, derived from partition coefficients. *Journal of the American Chemical Society*, 86(23):5175–5180.

Gao, K., Nguyen, D. D., Sresht, V., Mathiowetz, A. M., Tu, M., and Wei, G.-W. (2020). Are 2d fingerprints still valuable for drug discovery? *Physical Chemistry Chemical Physics*, 22:8373–8390.

Gerber, P. R. and Müller, K. (1995). MAB, a Generally Applicable Molecular Force Field for Structure Modelling in Medicinal Chemistry. *Journal of Computer-Aided Molecular Design*, 9(3):251–268.

Ghahremanpour, M. M., Tirado-Rives, J., Deshmukh, M., Ippolito, J. A., Zhang, C.-H., Cabeza de Vaca, I., Liosi, M.-E., Anderson, K. S., and Jorgensen, W. L. (2020). Identification of 14 Known Drugs as Inhibitors of the Main Protease of SARS-CoV-2. *ACS Medicinal Chemistry Letters*, 11(12):2526–2533.

Gómez-Bombarelli, R., Wei, J. N., Duvenaud, D., Hernández-Lobato, J. M., Sánchez-Lengeling, B., Sheberla, D., Aguilera-Iparraguirre, J., Hirzel, T. D., Adams, R. P., and Aspuru-Guzik, A. (2018). Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. *ACS Central Science*, 4(2):268–276.

Goodford, P. J. (1985). A Computational Procedure for Determining Energetically Favorable Binding Sites on Biologically Important Macromolecules. *Journal of Medicinal Chemistry*, 28(7):849–857.

Gorgulla, C., Boeszoermenyi, A., Wang, Z.-F., Fischer, P. D., Coote, P. W., Padmanabha Das, K. M., Malets, Y. S., Radchenko, D. S., Moroz, Y. S., Scott, D. A.,

Fackeldey, K., Hoffmann, M., Iavniuk, I., Wagner, G., and Arthanari, H. (2020). An Open-Source Drug Discovery Platform Enables Ultra-Large Virtual Screens. *Nature*, 580(7805):663–668.

Greener, J. G., Kandathil, S. M., Moffat, L., and Jones, D. T. (2022). A Guide to Machine Learning for Biologists. *Nature Reviews Molecular Cell Biology*, 23(1):40–55.

Griffiths, R.-R. and Hernández-Lobato, J. M. (2020). Constrained bayesian optimization for automatic chemical design using variational autoencoders. *Chemical Science*, 11:577–586.

Guha, R. (2012). Exploring Structure-Activity Data Using the Landscape Paradigm. *Wiley Interdisciplinary Reviews. Computational Molecular Science*, 2(6):10.1002/wcms.1087.

Halford, B. (2021). C&EN: Pfizer Unveils its Oral SARS-CoV-2 Inhibitor.

Halgren, T. A. (1999). MMFF VI. MMFF94s Option for Energy Minimization Studies. *Journal of Computational Chemistry*, 20(7):720–729.

Hamilton, W. L., Ying, R., and Leskovec, J. (2017). Inductive Representation Learning on Large Graphs. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*.

Hansch, C., Maloney, P. P., Fujita, T., and Muir, R. M. (1962). Correlation of biological activity of phenoxyacetic acids with hammett substituent constants and partition coefficients. *Nature*, 194(4824):178–180.

Hartenfeller, M., Zettl, H., Walter, M., Rupp, M., Reisen, F., Proschak, E., Weggen, S., Stark, H., and Schneider, G. (2012). DOGS: Reaction-Driven *de novo* Design of Bioactive Compounds. *PLOS Computational Biology*, 8(2):1–12.

He, R., Dobie, F., Ballantine, M., Leeson, A., Li, Y., Bastien, N., Cutts, T., Andonov, A., Cao, J., Booth, T. F., Plummer, F. A., Tyler, S., Baker, L., and Li, X. (2004). Analysis of Multimerization of the SARS Coronavirus Nucleocapsid Protein. *Biochemical and Biophysical Research Communications*, 316(2):476–483.

Hess, B., Bekker, H., Berendsen, H. J. C., and Fraaije, J. G. E. M. (1997). LINCS: A Linear Constraint Solver for Molecular Simulations. *Journal of Computational Chemistry*, 18(12):1463–1472.

Hetherington, K., Dutt, S., Ibarra, A. A., Cawood, E. E., Hobor, F., Woolfson, D. N., Edwards, T. A., Nelson, A., Sessions, R. B., and Wilson, A. J. (2021). Towards Optimizing Peptide-Based Inhibitors of Protein–Protein Interactions: Predictive Saturation Variation Scanning (PreSaVS). *RSC Chemical Biology*.

Hoffman, R. L., Kania, R. S., Brothers, M. A., Davies, J. F., Ferre, R. A., Gajiwala, K. S., He, M., Hogan, R. J., Kozminski, K., Li, L. Y., Lockner, J. W., Lou, J., Marra, M. T., Mitchell, L. J., Murray, B. W., Nieman, J. A., Noell, S., Planken, S. P., Rowe, T., Ryan, K., Smith, G. J., Solowiej, J. E., Steppan, C. M., and Taggart, B. (2020). Discovery of Ketone-Based Covalent Inhibitors of Coronavirus 3CL Proteases for the Potential Therapeutic Treatment of COVID-19. *Journal of Medicinal Chemistry*, 63(21):12725–12747.

Hopkins, A. L., Keserü, G. M., Leeson, P. D., Rees, D. C., and Reynolds, C. H. (2014). The Role of Ligand Efficiency Metrics in Drug Discovery. *Nature Reviews Drug Discovery*, 13(2):105–121.

Hu, Y., Gupta-Ostermann, D., and Bajorath, J. (2014). Exploring Compound Promiscuity Patterns and Multi-Target Activity Spaces. *Computational and Structural Biotechnology Journal*, 9(13):e201401003.

Huey, R., Morris, G. M., Olson, A. J., and Goodsell, D. S. (2007). A Semiempirical Free Energy Force Field with Charge-Based Desolvation. *Journal of Computational Chemistry*, 28(6):1145–1152.

Ibarra, A. A., Bartlett, G. J., Hegedüs, Z., Dutt, S., Hobor, F., Horner, K. A., Hetherington, K., Spence, K., Nelson, A., Edwards, T. A., Woolfson, D. N., Sessions, R. B., and Wilson, A. J. (2019). Predicting and Experimentally Validating Hot-Spot Residues at Protein–Protein Interfaces. *ACS Chemical Biology*, 14(10):2252–2263.

Imrie, F., Hadfield, T. E., Bradley, A. R., and Deane, C. M. (2021). Deep Generative Design with 3D Pharmacophoric Constraints. *Chemical Science*, 12(43):14577–14589.

Ivanciuc, O. (2000). QSAR Comparative Study of Wiener Descriptors for Weighted Molecular Graphs. *Journal of Chemical Information and Computer Sciences*, 40(6):1412–1422.

Jaccard, P. (1912). The Distribution of the Flora in the Alpine Zone. *New Phytologist*, 11(2):37–50.

Jaeger, S., Fulle, S., and Turk, S. (2018). Mol2vec: Unsupervised Machine Learning Approach with Chemical Intuition. *Journal of Chemical Information and Modeling*, 58(1):27–35.

Jiang, D., Lei, T., Wang, Z., Shen, C., Cao, D., and Hou, T. (2020a). ADMET Evaluation in Drug Discovery. Prediction of Breast Cancer Resistance Protein Inhibition Through Machine Learning. *Journal of Cheminformatics*, 12(1):16.

Jiang, M., Li, Z., Zhang, S., Wang, S., Wang, X., Yuan, Q., and Wei, Z. (2020b). Drug–target affinity prediction using graph neural network and contact maps. *RSC Advances*, 10:20701–20712.

Jiménez, J., Škalič, M., Martínez-Rosell, G., and De Fabritiis, G. (2018). K_{DEEP}: Protein–Ligand Absolute Binding Affinity Prediction via 3D-Convolutional Neural Networks. *Journal of Chemical Information and Modeling*, 58(2):287–296.

Jin, W., Barzilay, R., and Jaakkola, T. (2018). Junction tree variational autoencoder for molecular graph generation. In *International Conference on Machine Learning*, pages 2323–2332. PMLR.

Jin, Z., Du, X., Xu, Y., Deng, Y., Liu, M., Zhao, Y., Zhang, B., Li, X., Zhang, L., Peng, C., Duan, Y., Yu, J., Wang, L., Yang, K., Liu, F., Jiang, R., Yang, X., You, T., Liu, X., Yang, X., Bai, F., Liu, H., Liu, X., Guddat, L. W., Xu, W., Xiao, G., Qin, C., Shi, Z., Jiang, H., Rao, Z., and Yang, H. (2020). Structure of M^{pro} from SARS-CoV-2 and Discovery of Its Inhibitors. *Nature*, 582(7811):289–293.

Jones, G., Willett, P., Glen, R. C., Leach, A. R., and Taylor, R. (1997). Development and Validation of a Genetic Algorithm for Flexible Docking. *Journal of Molecular Biology*, 267(3):727–748.

Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W., and Klein, M. L. (1983). Comparison of Simple Potential Functions for Simulating Liquid Water. *The Journal of Chemical Physics*, 79(2):926–935.

Jubb, H. C. (accessed: April 2020). PDBTools.

Jubb, H. C., Higueruelo, A. P., Ochoa-Montaño, B., Pitt, W. R., Ascher, D. B., and Blundell, T. L. (2017). Arpeggio: A Web Server for Calculating and Visualising Interatomic Interactions in Protein Structures. *Journal of molecular biology*, 429(3):365–371.

Kadurin, A., Nikolenko, S., Khrabrov, K., Aliper, A., and Zhavoronkov, A. (2017). druGAN: An Advanced Generative Adversarial Autoencoder Model for *de dovo* Generation of New Molecules with Desired Molecular Properties in Silico. *Molecular Pharmaceutics*, 14(9):3098–3104.

Karlov, D. S., Sosnin, S., Fedorov, M. V., and Popov, P. (2020). graphDelta: MPNN Scoring Function for the Affinity Prediction of Protein–Ligand Complexes. *ACS Omega*, 5(10):5150–5159.

Kipf, T. N. and Welling, M. (2017). Semi-Supervised Classification with Graph Convolutional Networks. In *Proceedings of the International Conference on Learning Representations*.

Kitchen, D. B., Decornez, H., Furr, J. R., and Bajorath, J. (2004). Docking and Scoring in Virtual Screening for Drug Discovery: Methods and Applications. *Nature Reviews Drug Discovery*, 3(11):935–949.

Koes, D. R., Baumgartner, M. P., and Camacho, C. J. (2013). Lessons Learned in Empirical Scoring with Smina from the CSAR 2011 Benchmarking Exercise. *Journal of Chemical Information and Modeling*, 53(8):1893–1904.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'12, page 1097–1105, Red Hook, NY, USA. Curran Associates Inc.

Kuntz, I. D., Blaney, J. M., Oatley, S. J., Langridge, R., and Ferrin, T. E. (1982). A geometric approach to macromolecule-ligand interactions. *Journal of Molecular Biology*, 161(2):269–288.

Kutchukian, P. S., Vasilyeva, N. Y., Xu, J., Lindvall, M. K., Dillon, M. P., Glick, M., Coley, J. D., and Brooijmans, N. (2012). Inside the Mind of a Medicinal Chemist: The Role of Human Bias in Compound Prioritization during Drug Discovery. *PLOS One*, 7(11):e48476.

Landrum, G. et al. (2006). RDKit: Open-source Cheminformatics. `https://www.rdkit.org/docs/index.html`.

Leung, S., Bodkin, M., von Delft, F., Brennan, P., and Morris, G. (2019). SuCOS is Better than RMSD for Evaluating Fragment Elaboration and Docking Poses. *ChemRxiv*.

Li, S., Zhou, J., Xu, T., Huang, L., Wang, F., Xiong, H., Huang, W., Dou, D., and Xiong, H. (2021). Structure-Aware Interactive Graph Neural Networks for the Prediction of Protein-Ligand Binding Affinity. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*.

Li, Y., Su, M., Liu, Z., Li, J., Liu, J., Han, L., and Wang, R. (2018). Assessing Protein–Ligand Interaction Scoring Functions with the CASF-2013 Benchmark. *Nature Protocols*, 13(4):666–680.

Lim, J., Ryu, S., Park, K., Choe, Y. J., Ham, J., and Kim, W. Y. (2019). Predicting Drug–Target Interaction Using a Novel Graph Neural Network with 3D Structure-Embedded Graph Representation. *Journal of Chemical Information and Modeling*, 59(9):3981–3988.

Lindorff-Larsen, K., Piana, S., Palmo, K., Maragakis, P., Klepeis, J. L., Dror, R. O., and Shaw, D. E. (2010). Improved Side-Chain Torsion Potentials for the Amber ff99SB Protein Force Field. *Proteins: Structure, Function, and Bioinformatics*, 78(8):1950–1958.

Lipinski, C. A., Lombardo, F., Dominy, B. W., and Feeney, P. J. (2001). Experimental and Computational Approaches to Estimate Solubility and Permeability in Drug Discovery and Development Settings. *Advanced Drug Delivery Reviews*, 46(1):3–26.

Liu, Z., Li, Y., Han, L., Li, J., Liu, J., Zhao, Z., Nie, W., Liu, Y., and Wang, R. (2015). PDB-Wide Collection of Binding Data: Current Status of the PDBbind Database. *Bioinformatics*, 31(3):405–412.

Liu, Z., Su, M., Han, L., Liu, J., Yang, Q., Li, Y., and Wang, R. (2017). Forging the Basis for Developing Protein-Ligand Interaction Scoring Functions. *Accounts of Chemical Research*, 50(2):302–309.

Loschwitz, J., Jäckering, A., Keutmann, M., Olagunju, M., Eberle, R. J., Coronado, M. A., Olubiyi, O. O., and Strodel, B. (2021). Novel Inhibitors of the Main Protease Enzyme of SARS-CoV-2 Identified via Molecular Dynamics Simulation-Guided in Vitro Assay. *Bioorganic Chemistry*, page 104862.

Lubin, J. H., Zardecki, C., Dolan, E. M., Lu, C., Shen, Z., Dutta, S., Westbrook, J. D., Hudson, B. P., Goodsell, D. S., Williams, J. K., Voigt, M., Sarma, V., Xie, L., Venkatachalam, T., Arnold, S., Alfaro Alvarado, L. H., Catalfano, K., Khan, A., McCarthy, E., Staggers, S., Tinsley, B., Trudeau, A., Singh, J., Whitmore, L., Zheng, H., Benedek, M., Currier, J., Dresel, M., Duvvuru, A., Dyszel, B., Fingar, E., Hennen, E. M., Kirsch, M., Khan, A. A., Labrie-Cleary, C., Laporte, S., Lenkeit, E., Martin, K., Orellana, M., Ortiz-Alvarez de la Campa, M., Paredes, I., Wheeler,

B., Rupert, A., Sam, A., See, K., Soto Zapata, S., Craig, P. A., Hall, B. L., Jiang, J., Koeppe, J. R., Mills, S. A., Pikaart, M. J., Roberts, R., Bromberg, Y., Hoyer, J. S., Duffy, S., Tischfield, J., Ruiz, F. X., Arnold, E., Baum, J., Sandberg, J., Brannigan, G., Khare, S. D., and Burley, S. K. (2022). Evolution of the sars-cov-2 proteome in three dimensions (3d) during the first 6 months of the covid-19 pandemic. *Proteins: Structure, Function, and Bioinformatics*, 90(5):1054–1080.

Malhotra, S. and Karanicolas, J. (2017). When Does Chemical Elaboration Induce a Ligand To Change Its Binding Mode? *Journal of Medicinal Chemistry*, 60(1):128–145.

Malla, T. R., Tumber, A., John, T., Brewitz, L., Strain-Damerell, C., Owen, C. D., Lukacik, P., Chan, H. T. H., Maheswaran, P., Salah, E., Duarte, F., Yang, H., Rao, Z., Walsh, M. A., and Schofield, C. J. (2021). Mass Spectrometry Reveals Potential of β-Lactams as SARS-CoV-2 M^{pro} Inhibitors. *Chemical Communications*, 57(12):1430–1433.

Marcou, G. and Rognan, D. (2007). Optimizing Fragment and Scaffold Docking by Use of Molecular Interaction Fingerprints. *Journal of Chemical Information and Modeling*, 47(1):195–207.

Martis, E. A., Radhakrishnan, R., and Badve, R. R. (2011). High-Throughput Screening: The Hits and Leads of Drug Discovery - an Overview. *Journal of Applied Pharmaceutical Science*, 1(1):2–10.

McIntosh-Smith, S., Price, J., Sessions, R. B., and Ibarra, A. A. (2015). High Performance In-silico Virtual Drug Screening on Many-Core Processors. *The International Journal of High Performance Computing Applications*, 29(2):119–134.

McIntosh-Smith, S., Wilson, T., Ibarra, A. Á., Crisp, J., and Sessions, R. B. (2012). Benchmarking Energy Efficiency, Power Costs and Carbon Emissions on Heterogeneous Systems. *The Computer Journal*, 55(2):192–205.

McNutt, A. T., Francoeur, P., Aggarwal, R., Masuda, T., Meli, R., Ragoza, M., Sunseri, J., and Koes, D. R. (2021). GNINA 1.0: Molecular Docking with Deep Learning. *Journal of Cheminformatics*, 13(1):43.

Mehler, E. L. and Solmajer, T. (1991). Electrostatic Effects in Proteins: Comparison of Dielectric and Charge Models. *Protein Engineering, Design and Selection*, 4(8):903–910.

Meli, R., Anighoro, A., Bodkin, M. J., Morris, G. M., and Biggin, P. C. (2021). Learning Protein-Ligand Binding Affinity with Atomic Environment Vectors. *Journal of Cheminformatics*, 13(1).

Mendez, D., Gaulton, A., Bento, A. P., Chambers, J., De Veij, M., Félix, E., Magariños, M., Mosquera, J., Mutowo, P., Nowotka, M., Gordillo-Marañón, M., Hunter, F., Junco, L., Mugumbate, G., Rodriguez-Lopez, M., Atkinson, F., Bosc, N., Radoux, C., Segura-Cabrera, A., Hersey, A., and Leach, A. (2019). Chembl: Towards direct deposition of bioassay data. *Nucleic Acids Research*, 47:D930–D940.

Méndez-Lucio, O., Baillif, B., Clevert, D.-A., Rouquié, D., and Wichard, J. (2020). *De novo* Generation of Hit-Like Molecules From Gene Expression Signatures Using Artificial Intelligence. *Nature Communications*, 11(1).

Mengist, H. M., Dilnessa, T., and Jin, T. (2021). Structural Basis of Potential Inhibitors Targeting SARS-CoV-2 Main Protease. *Frontiers in Chemistry*, 9(7).

Michaud-Agrawal, N., Denning, E. J., Woolf, T. B., and Beckstein, O. (2011). Mdanalysis: A toolkit for the analysis of molecular dynamics simulations. *Journal of Computational Chemistry*, 32(10):2319–2327.

Mikolov, T., Corrado, G. S., Chen, K., and Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space.

Mockus, J. (1989). Bayesian Approach to Global Optimization.

Moesser, M. (2021). Normalized SuCOS. `https://github.com/MarcMoesser/SuCOS`.

Moesser, M. A., Klein, D., Boyles, F., Deane, C. M., Baxter, A., and Morris, G. M. (2022). Protein-Ligand Interaction Graphs: Learning from Ligand-Shaped 3D Interaction Graphs to Improve Binding Affinity Prediction. *bioRxiv*, page 2022.03.04.483012.

Moffat, J. G., Vincent, F., Lee, J. A., Eder, J., and Prunotto, M. (2017). Opportunities and Challenges in Phenotypic Drug Discovery: An Industry Perspective. *Nature Reviews Drug Discovery*, 16(8):531–543.

Mondal, D. and Warshel, A. (2020). Exploring the Mechanism of Covalent Inhibition: Simulating the Binding Free Energy of α-Ketoamide Inhibitors of the Main Protease of SARS-CoV-2. *Biochemistry*, 59(48):4601–4608.

Moon, S., Zhung, W., Yang, S., Lim, J., and Kim, W. Y. (2022). Pignet: A physics-informed deep learning model toward generalized drug–target interaction predictions. *Chemical Science*, 13:3661–3673.

Morris, G. M., Goodsell, D. S., Halliday, R. S., Huey, R., Hart, W. E., Belew, R. K., and Olson, A. J. (1998). Automated docking using a lamarckian genetic algorithm

and an empirical binding free energy function. *Journal of Computational Chemistry*, 19(14):1639–1662.

Morris, G. M., Huey, R., Lindstrom, W., Sanner, M. F., Belew, R. K., Goodsell, D. S., and Olson, A. J. (2009). AutoDock4 and AutoDockTools4: Automated Docking with Selective Receptor Flexibility. *Journal of computational chemistry*, 30(16):2785–2791.

Nguyen, T., Le, H., Quinn, T. P., Nguyen, T., Le, T. D., and Venkatesh, S. (2020). GraphDTA: Predicting Drug–Target Binding Affinity with Graph Neural Networks. *Bioinformatics*, 37(8):1140–1147.

O'Boyle, N. M., Banck, M., James, C. A., Morley, C., Vandermeersch, T., and Hutchison, G. R. (2011). Open Babel: An Open Chemical Toolbox. *Journal of Cheminformatics*, 3(1):33.

Olivecrona, M., Blaschke, T., Engkvist, O., and Chen, H. (2017). Molecular *de-novo* Design Through Deep Reinforcement Learning. *Journal of Cheminformatics*, 9(1).

Olsson, M. H. M., Søndergaard, C. R., Rostkowski, M., and Jensen, J. H. (2011). PROPKA3: Consistent Treatment of Internal and Surface Residues in Empirical pKa Predictions. *Journal of Chemical Theory and Computation*, 7(2):525–537.

O'Toole, Á., Scher, E., Underwood, A., Jackson, B., Hill, V., McCrone, J. T., Colquhoun, R., Ruis, C., Abu-Dahab, K., Taylor, B., Yeats, C., du Plessis, L., Maloney, D., Medd, N., Attwood, S. W., Aanensen, D. M., Holmes, E. C., Pybus, O. G., and Rambaut, A. (2021). Assignment of Epidemiological Lineages in an Emerging Pandemic using the Pangolin Tool. *Virus Evolution*, 7(2). veab064.

Owen, C. D., Lukacik, P., Strain-Damerell, C. M., Douangamath, A., Powell, A. J., Fearon, D., Brandao-Neto, J., Crawshaw, A. D., Aragao, D., Williams, M., Flaig,

R., Hall, D., McAauley, K., Stuart, D. I., von Delft, F., and Walsh, M. A. (2020). COVID-19 Main Protease with Unliganded Active Site. *PDB 6YB7*.

Owen, D. (2021). Pfizer Oral COVID-19 Antiviral Clinical Candidate.

Owen, D. R., Allerton, C. M. N., Anderson, A. S., Aschenbrenner, L., Avery, M., Berritt, S., Boras, B., Cardin, R. D., Carlo, A., Coffman, K. J., Dantonio, A., Di, L., Eng, H., Ferre, R., Gajiwala, K. S., Gibson, S. A., Greasley, S. E., Hurst, B. L., Kadar, E. P., Kalgutkar, A. S., Lee, J. C., Lee, J., Liu, W., Mason, S. W., Noell, S., Novak, J. J., Obach, R. S., Ogilvie, K., Patel, N. C., Pettersson, M., Rai, D. K., Reese, M. R., Sammons, M. F., Sathish, J. G., Singh, R. S. P., Steppan, C. M., Stewart, A. E., Tuttle, J. B., Updyke, L., Verhoest, P. R., Wei, L., Yang, Q., and Zhu, Y. (2021). An oral sars-cov-2 m^{pro} inhibitor clinical candidate for the treatment of covid-19. *Science*, 374(6575):1586–1593.

Pace, C. N. and Scholtz, J. M. (1998). A Helix Propensity Scale Based on Experimental Studies of Peptides and Proteins. *Biophysical Journal*, 75(1):422–427.

Paolini, G. V., Shapland, R. H. B., van Hoorn, W. P., Mason, J. S., and Hopkins, A. L. (2006). Global Mapping of Pharmacological Space. *Nature Biotechnology*, 24(7):805–815.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. (2019). PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems 32*.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A.,

Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Pickett, S. D., Green, D. V. S., Hunt, D. L., Pardoe, D. A., and Hughes, I. (2011). Automated Lead Optimization of MMP-12 Inhibitors Using a Genetic Algorithm. *ACS Medicinal Chemistry Letters*, 2(1):28–33.

Pillaiyar, T., Manickam, M., Namasivayam, V., Hayashi, Y., and Jung, S.-H. (2016). An Overview of Severe Acute Respiratory Syndrome - Coronavirus (SARS-CoV) 3CL Protease Inhibitors: Peptidomimetics and Small Molecule Chemotherapy. *Journal of Medicinal Chemistry*, 59(14):6595–6628.

Plowright, A. T., Johnstone, C., Kihlberg, J., Pettersson, J., Robb, G., and Thompson, R. A. (2012). Hypothesis driven drug design: improving quality and effectiveness of the design-make-test-analyse cycle. *Drug Discovery Today*, 17(1):56–62.

Popova, M., Isayev, O., and Tropsha, A. (2018). Deep Reinforcement Learning for *de novo* Drug Design. *Science Advances*, 4(7).

PostEra.Ai (2020). COVID-19 Moonshot Project. `https://postera.ai/covid`.

Pyzer-Knapp, E. O. (2018). Bayesian Optimization for Accelerated Drug Discovery. *IBM Journal of Research and Development*, 62(6):2:1–2:7.

Rácz, A., Bajusz, D., and Héberger, K. (2018). Life Beyond the Tanimoto Coefficient: Similarity Measures for Interaction Fingerprints. *Journal of Cheminformatics*, 10(1):48.

Ragoza, M., Hochuli, J., Idrobo, E., Sunseri, J., and Koes, D. R. (2017). Protein–Ligand Scoring with Convolutional Neural Networks. *Journal of Chemical Information and Modeling*, 57(4):942–957.

Rambaut, A., Holmes, E. C., O'Toole, Á., Hill, V., McCrone, J. T., Ruis, C., du Plessis, L., and Pybus, O. G. (2020). A Dynamic Nomenclature Proposal for SARS-CoV-2 Lineages to Assist Genomic Epidemiology. *Nature Microbiology*, 5(11):1403–1407.

Ramos-Guzmán, C. A., Ruiz-Pernía, J. J., and Tuñón, I. (2020). Unraveling the SARS-CoV-2 Main Protease Mechanism Using Multiscale Methods. *ACS Catalysis*, 10(21):12544–12554.

Rappe, A. K., Casewit, C. J., Colwell, K. S., Goddard, W. A., and Skiff, W. M. (1992). UFF, a Full Periodic Table Force Field for Molecular Mechanics and Molecular Dynamics Simulations. *Journal of the American Chemical Society*, 114(25):10024–10035.

Rester, U. (2008). From Virtuality to Reality - Virtual Screening in Lead Discovery and Lead Optimization: A Medicinal Chemistry Perspective. *Current Opinion in Drug Discovery & Development*, 11(4):559–568.

Richard J. Gowers, Max Linke, Jonathan Barnoud, Tyler J. E. Reddy, Manuel N. Melo, Sean L. Seyler, Jan Domański, David L. Dotson, Sébastien Buchoux, Ian M. Kenney, and Oliver Beckstein (2016). MDAnalysis: A Python Package for the Rapid Analysis of Molecular Dynamics Simulations. In Sebastian Benthall and Scott Rostrup, editors, *Proceedings of the 15th Python in Science Conference*, pages 98–105.

Riniker, S. and Landrum, G. A. (2015). Better Informed Distance Geometry: Using What We Know to Improve Conformation Generation. *Journal of Chemical Information and Modeling*, 55(12):2562–2574.

Rogers, D. and Hahn, M. (2010). Extended-Connectivity Fingerprints. *Journal of Chemical Information and Modeling*, 50(5):742–754.

Roy, A. S. A. (2012). Stifling New Cures: The True Cost of Lengthy Clinical Drug Trials. Technical report, Manhattan Institute for Policy Research.

Ruddigkeit, L., van Deursen, R., Blum, L. C., and Reymond, J.-L. (2012). Enumeration of 166 Billion Organic Small Molecules in the Chemical Universe Database GDB-17. *Journal of Chemical Information and Modeling*, 52(11):2864–2875.

Rut, W., Groborz, K., Zhang, L., Sun, X., Zmudzinski, M., Pawlik, B., Wang, X., Jochmans, D., Neyts, J., Młynarski, W., Hilgenfeld, R., and Drag, M. (2021). SARS-CoV-2 M^{pro} Inhibitors and Activity-Based Probes for Patient-Sample Imaging. *Nature Chemical Biology*, 17(2):222–228.

Sadybekov, A. A., Sadybekov, A. V., Liu, Y., Iliopoulos-Tsoutsouvas, C., Huang, X.-P., Pickett, J., Houser, B., Patel, N., Tran, N. K., Tong, F., Zvonok, N., Jain, M. K., Savych, O., Radchenko, D. S., Nikas, S. P., Petasis, N. A., Moroz, Y. S., Roth, B. L., Makriyannis, A., and Katritch, V. (2022). Synthon-Based Ligand Discovery in Virtual Libraries of over 11 Billion Compounds. *Nature*, 601(7893):452–459.

Salimraj, R., Hinchliffe, P., Kosmopoulou, M., Tyrrell, J. M., Brem, J., Berkel, S. S., Verma, A., Owens, R. J., McDonough, M. A., Walsh, T. R., Schofield, C. J., and Spencer, J. (2019). Crystal Structures of VIM-1 Complexes Explain Active Site Heterogeneity in VIM-class Metallo-β-Lactamases. *The FEBS Journal*, 286(1):169–183.

Santos-Martins, D., Solis-Vasquez, L., Tillack, A. F., Sanner, M. F., Koch, A., and Forli, S. (2021). Accelerating AutoDock4 with GPUs and Gradient-Based Local Search. *Journal of Chemical Theory and Computation*, 17(2):1060–1073.

Schechter, I. and Berger, A. (1967). On the Size of the Active Site in Proteases. I. Papain. *Biochemical and Biophysical Research Communications*, 27(2):157–162.

Schrödinger LLC. (accessed: 2020). The PyMOL Molecular Graphics System, Version 2.3.0.

Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R., and LeCun, Y. (2014). Overfeat: Integrated recognition, localization and detection using convolutional networks. In *2nd International Conference on Learning Representations, ICLR 2014*.

Sessions, R. B. (accessed: April 2021). BUDE_SM. `https://github.com/richardbsessions/BUDE_SM`.

Shanmugasundaram, V. and Maggiora, G. (2001). Characterizing Property and Activity Landscapes Using an Information-Theoretic Approach. *CINF-032. 222nd ACS National Meeting, Chicago, IL, United States.*

Shaqra, A. M., Zvornicanin, S., Huang, Q. Y., Lockbaum, G. J., Knapp, M., Tandeske, L., Barkan, D. T., Flynn, J., Bolon, D. N. A., Moquin, S., Dovala, D., Kurt Yilmaz, N., and Schiffer, C. A. (2022). Defining the Substrate Envelope of SARS-CoV-2 Main Protease to Predict and Avoid Drug Resistance. *bioRxiv*, page 2022.01.25.477757.

Singh, J., Deng, Z., Narale, G., and Chuaqui, C. (2006). Structural interaction fingerprints: A new approach to organizing, mining, analyzing, and designing protein-small molecule complexes. *Chemical Biology & Drug Design*, 67(1):5–12.

Smith, J. S., Isayev, O., and Roitberg, A. E. (2017). ANI-1: An Extensible Neural Network Potential with DFT Accuracy at Force Field Computational Cost. *Chemical Science*, 8(4):3192–3203.

Su, M., Yang, Q., Du, Y., Feng, G., Liu, Z., Li, Y., and Wang, R. (2018). Comparative Assessment of Scoring Functions: The CASF-2016 Update. *Journal of Chemical Information and Modeling*, 59(2):895–913.

Su, M., Yang, Q., Du, Y., Feng, G., Liu, Z., Li, Y., and Wang, R. (2019). Comparative Assessment of Scoring Functions: the CASF-2016 Update. *Journal of Chemical Information and Modeling*, 59(2):895–913.

Świderek, K. and Moliner, V. (2020). Revealing the Molecular Mechanisms of Proteolysis of SARS-CoV-2 M^{pro} by QM/MM Computational Methods. *Chemical Science*, 11(39):10626–10630.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2015). Going deeper with convolutions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–9.

Sánchez-Cruz, N., Medina-Franco, J. L., Mestres, J., and Barril, X. (2020). Extended connectivity interaction features: improving binding affinity prediction through chemical description. *Bioinformatics*, 37(10):1376–1382.

Tantibanchachai, C. (2021). Coronavirus (COVID-19) Update: FDA Authorizes First Oral Antiviral for Treatment of COVID-19.

Trott, O. and Olson, A. J. (2010). AutoDock Vina: Improving the Speed and Accuracy of Docking with a New Scoring Function, Efficient Optimization, and Multithreading. *Journal of Computational Chemistry*, 31(2):455–461.

Vamathevan, J., Clark, D., Czodrowski, P., Dunham, I., Ferran, E., Lee, G., Li, B., Madabhushi, A., Shah, P., Spitzer, M., and Zhao, S. (2019). Applications of Machine Learning in Drug Discovery and Development. *Nature Reviews Drug Discovery*, 18(6):463–477.

van Berkel, S. S., Brem, J., Rydzik, A. M., Salimraj, R., Cain, R., Verma, A., Owens, R. J., Fishwick, C. W. G., Spencer, J., and Schofield, C. J. (2013). Assay Platform for Clinically Relevant Metallo-β-Lactamases. *Journal of Medicinal Chemistry*, 56(17):6945–6953.

Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., and Bengio, Y. (2018). Graph Attention Networks. In *Proceedings of the International Conference on Learning Representations*.

Vinkers, H. M., de Jonge, M. R., Daeyaert, F. F. D., Heeres, J., Koymans, L. M. H., van Lenthe, J. H., Lewi, P. J., Timmerman, H., Van Aken, K., and Janssen, P. A. J. (2003). SYNOPSIS: SYNthesize and OPtimize System in Silico. *Journal of Medicinal Chemistry*, 46(13):2765–2773.

Wade, R. C., Clark, K. J., and Goodford, P. J. (1993). Further Development of Hydrogen Bond Functions for Use in Determining Energetically Favorable Binding Sites on Molecules of Known Structure. 1. Ligand Probe Groups with the Ability to Form Two Hydrogen Bonds. *Journal of Medicinal Chemistry*, 36(1):140–147.

Walsh, T. R., Toleman, M. A., Poirel, L., and Nordmann, P. (2005). Metallo-β-Lactamases: the Quiet before the Storm? *Clinical Microbiology Reviews*, 18(2):306–325.

Wang, G. and Zhu, W. (2016). Molecular Docking for Drug Discovery and Development: a Widely Used Approach But Far From Perfect. *Future Medicinal Chemistry*, 8(14):1707–1710.

Wang, L. L., Lo, K., Chandrasekhar, Y., Reas, R., Yang, J., Eide, D., Funk, K., Kinney, R., Liu, Z., Merrill, W., Mooney, P., Murdick, D., Rishi, D., Sheehan, J., Shen, Z., Stilson, B., Wade, A. D., Wang, K., Wilhelm, C., Xie, B., Raymond, D.,

Weld, D. S., Etzioni, O., and Kohlmeier, S. (2020). CORD-19: The COVID-19 Open Research Dataset. *ArXiv*, page arXiv:2004.10706v2.

Wei, M., Wynn, R., Hollis, G., Liao, B., Margulis, A., Reid, B. G., Klabe, R., Liu, P. C., Becker-Pasha, M., Rupar, M., Burn, T. C., McCall, D. E., and Li, Y. (2007). High-Throughput Determination of Mode of Inhibition in Lead Identification and Optimization. *J. Biomol. Screen.*, 12(2):220–228.

Weiner, S. J., Kollman, P. A., Case, D. A., Singh, U. C., Ghio, C., Alagona, G., Profeta, S., and Weiner, P. (1984). A New Force Field for Molecular Mechanical Simulation of Nucleic Acids and Proteins. *Journal of the American Chemical Society*, 106(3):765–784.

Weininger, D. (1988). Smiles, a chemical language and information system. introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences*, 28(1):31–36.

Wenzel, J., Matter, H., and Schmidt, F. (2019). Predictive Multitask Deep Neural Network Models for ADME-Tox Properties: Learning from Large Data Sets. *Journal of Chemical Information and Modeling*, 59(3):1253–1268.

Williams, C. J., Headd, J. J., Moriarty, N. W., Prisant, M. G., Videau, L. L., Deis, L. N., Verma, V., Keedy, D. A., Hintze, B. J., Chen, V. B., Jain, S., Lewis, S. M., Arendall III, W. B., Snoeyink, J., Adams, P. D., Lovell, S. C., Richardson, J. S., and Richardson, D. C. (2018). MolProbity: More and Better Reference Data for Improved All-Atom Structure Validation. *Protein Science*, 27(1):293–315.

Wilmouth, R. C., Clifton, I. J., Robinson, C. V., Roach, P. L., Aplin, R. T., Westwood, N. J., Hajdu, J., and Schofield, C. J. (1997). Structure of a Specific Acyl-

Enzyme Complex Formed Between β-Casomorphin-7 and Porcine Pancreatic Elastase. *Nature Structural Biology*, 4(6):456–462.

Wójcikowski, M., Kukiełka, M., Stepniewska-Dziubinska, M. M., and Siedlecki, P. (2018). Development of a protein–ligand extended connectivity (PLEC) fingerprint and its application for binding affinity predictions. *Bioinformatics*, 35(8):1334–1341.

Wójcikowski, M., Zielenkiewicz, P., and Siedlecki, P. (2015a). Open Drug Discovery Toolkit (ODDT): A New Open-Source Player in the Drug Discovery Field. *Journal of Cheminformatics*, 7(1):26.

Wójcikowski, M., Zielenkiewicz, P., and Siedlecki, P. (2015b). Open Drug Discovery Toolkit (ODDT): A New Open-Source Player in the Drug Discovery Field. *Journal of Cheminformatics*, 7(1):26.

Wood, C. W., Ibarra, A. A., Bartlett, G. J., Wilson, A. J., Woolfson, D. N., and Sessions, R. B. (2020). BAlaS: Fast, Interactive and Accessible Computational Alanine-Scanning Using BudeAlaScan. *Bioinformatics*, 36(9):2917–2919.

Wouters, O. J., McKee, M., and Luyten, J. (2020). Estimated Research and Development Investment Needed to Bring a New Medicine to Market, 2009-2018. *Journal of the American Medical Association*, 323(9):844–853.

Wright, P. A., Wilmouth, R. C., Clifton, I. J., and Schofield, C. J. (2001). Kinetic and Crystallographic Analysis of Complexes Formed Between Elastase and Peptides From Beta-Casein. *European Journal of Biochemistry*, 268(10):2969–2974.

Wu, F., Souza, A., Zhang, T., Fifty, C., Yu, T., and Weinberger, K. (2019). Simplifying Graph Convolutional Networks. In *International Conference on Machine Learning*.

Wu, F., Zhao, S., Yu, B., Chen, Y.-M., Wang, W., Song, Z.-G., Hu, Y., Tao, Z.-W., Tian, J.-H., Pei, Y.-Y., et al. (2020). A New Coronavirus Associated with Human Respiratory Disease in China. *Nature*, 579(7798):265–269.

Wu, Z., Ramsundar, B., Feinberg, E., Gomes, J., Geniesse, C., Pappu, A. S., Leswing, K., and Pande, V. (2018). Moleculenet: a benchmark for molecular machine learning. *Chemical Sciience*, 9:513–530.

Xia, B. and Kang, X. (2011). Activation and Maturation of SARS-CoV Main Protease. *Protein & Cell*, 2(4):282–290.

Xu, K., Hu, W., Leskovec, J., and Jegelka, S. (2019). How Powerful are Graph Neural Networks? In *Proceedings of the International Conference on Learning Representations*.

Xue, X., Yu, H., Yang, H., Xue, F., Wu, Z., Shen, W., Li, J., Zhou, Z., Ding, Y., Zhao, Q., Zhang, X. C., Liao, M., Bartlam, M., and Rao, Z. (2008). Structures of Two Coronavirus Main Proteases: Implications for Substrate Binding and Antiviral Drug Design. *J. Virol.*, 82(5):2515–2527.

Yang, H., Yang, M., Ding, Y., Liu, Y., Lou, Z., Zhou, Z., Sun, L., Mo, L., Ye, S., Pang, H., and Others (2003). The Crystal Structures of Severe Acute Respiratory Syndrome Virus Main Protease and its Complex with an Inhibitor. *Proceedings of the National Academy of Sciences*, 100(23):13190–13195.

Ying, R., Bourgeois, D., You, J., Zitnik, M., and Leskovec, J. (2019). GNNExplainer: Generating Explanations for Graph Neural Networks. *Advances in Neural Information Processing Systems*, 32:9240–9251.

Yocum, R. R., Rasmussen, J. R., and Strominger, J. L. (1980). The Mechanism of Action of Penicillin. Penicillin Acylates the Active Site of *Bacillus Stearothermophilus* D-Alanine Carboxypeptidase. *Journal of Biological Chemistry*, 255(9):3977–3986.

Yu, W. and MacKerell Jr, A. D. (2017). Computer-Aided Drug Design Methods. *Methods in Molecular Biology (Clifton, N.J.)*, 1520:85–106.

Yung-Chi, C. and Prusoff, W. H. (1973). Relationship Between the Inhibition Constant (K_I) and the Concentration of Inhibitor Which Causes 50 per cent Inhibition (IC_{50}) of an Enzymatic Reaction. *Biochemical Pharmacology*, 22(23):3099–3108.

Zhang, C.-H., Stone, E. A., Deshmukh, M., Ippolito, J. A., Ghahremanpour, M. M., Tirado-Rives, J., Spasov, K. A., Zhang, S., Takeo, Y., Kudalkar, S. N., Liang, Z., Isaacs, F., Lindenbach, B., Miller, S. J., Anderson, K. S., and Jorgensen, W. L. (2021). Potent Noncovalent Inhibitors of the Main Protease of SARS-CoV-2 from Molecular Sculpting of the Drug Perampanel Guided by Free Energy Perturbation Calculations. *ACS Central Science*, 7(3):467–475.

Zhang, L., Lin, D., Sun, X., Curth, U., Drosten, C., Sauerhering, L., Becker, S., Rox, K., and Hilgenfeld, R. (2020). Crystal Structure of SARS-CoV-2 Main Protease Provides a Basis for Design of Improved α-Ketoamide Inhibitors. *Science*, 368(6489):409–412.

Zheng, L., Fan, J., and Mu, Y. (2019). OnionNet: A Multiple-Layer Intermolecular-Contact-Based Convolutional Neural Network for Protein–Ligand Binding Affinity Prediction. *ACS Omega*, 4(14):15956–15965.

Zhong, S., Zhang, Y., and Xiu, Z. (2010). Rescoring Ligand Docking Poses. *Current Opinion in Drug Discovery & Development*, 13(3):326–334.

Zhou, P., Yang, X.-L., Wang, X.-G., Hu, B., Zhang, L., Zhang, W., Si, H.-R., Zhu, Y., Li, B., Huang, C.-L., Chen, H.-D., Chen, J., Luo, Y., Guo, H., Jiang, R.-D., Liu, M.-Q., Chen, Y., Shen, X.-R., Wang, X., Zheng, X.-S., Zhao, K., Chen, Q.-J., Deng, F., Liu, L.-L., Yan, B., Zhan, F.-X., Wang, Y.-Y., Xiao, G.-F., and Shi, Z.-L. (2020). A Pneumonia Outbreak Associated with a New Coronavirus of Probable Bat Origin. *Nature*, 579(7798):270–273.

Zhu, L., George, S., Schmidt, M. F., Al-Gharabli, S. I., Rademann, J., and Hilgenfeld, R. (2011). Peptide Aldehyde Inhibitors Challenge the Substrate Specificity of the SARS-Coronavirus Main Protease. *Antiviral Research*, 92(2):204–212.

Zhu, N., Zhang, D., Wang, W., Li, X., Yang, B., Song, J., Zhao, X., Huang, B., Shi, W., Lu, R., Niu, P., Zhan, F., Ma, X., Wang, D., Xu, W., Wu, G., Gao, G. F., and Tan, W. (2020). A Novel Coronavirus from Patients with Pneumonia in China, 2019. *New England Journal of Medicine*, 382(8):727–733.