The Effects of Sequence Variations in the Structural Dynamics and Ligand Interactions of Main Proteases in Betacoronaviruses

by

Shubham Devesh Ramgoolam

A thesis

presented to the University of Waterloo

in fulfillment of the

thesis requirement for the degree of

Master of Science

in

Pharmacy

Waterloo, Ontario, Canada, 2025

© Shubham Devesh Ramgoolam 2025

Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners. I understand that my thesis may be made electronically available to the public.

Abstract

The main protease (Mpro) of betacoronaviruses is an essential enzyme for viral replication and a premier target for antiviral drug discovery. The high conservation of its active site across the genus makes it an ideal candidate for developing pan-coronavirus therapeutics to combat future outbreaks. In response to the Coronavirus disease 2019 (COVID-19) pandemic, numerous large-scale screening campaigns were initiated to identify Mpro inhibitors. However, these efforts were frequently hindered by the fundamental pose classification problem, which is the inability to distinguish between correct and incorrect binding poses. Compounding this issue, the research response generated an unprecedented volume of Mpro structural data, yet existing bioinformatics platforms lack the integrated tools for its systematic, comparative analysis. These combined challenges hinder the rational design of next-generation inhibitors.

In this work, we address these challenges by developing a novel, integrated computational toolkit. We first present CoviProDigy, a web-based platform designed for the comprehensive and comparative analysis of Mpro-ligand interactions, featuring specialized tools for subpocket occupancy analysis and elucidating common and unique interactions across different ligands. This functionality is designed to support scaffold hopping and medicinal chemistry optimization based on the molecular-level insights gained from known structures. To overcome the major issue of inaccurate pose selection in virtual screening, we then developed and validated a fine-tuned machine learning model that demonstrates enhanced accuracy in classifying ligand binding poses for the *Betacoronavirus* Mpro family. Together, these contributions provide a robust, end-to-end framework that accelerates the discovery of potent inhibitors, presenting a valuable resource for pandemic preparedness and the ongoing search for broad-spectrum antiviral agents.

Acknowledgements

I am immensely grateful to my primary supervisor Dr. Ganesan, and co-supervisor Dr. Kalyaanamoorthy, for accepting me as their graduate student. This journey would not have been possible without their guidance and patience as I honed my skills as a scientist through trial and error. I also want to thank my committee members, Dr. Nekkar and Dr. Wong, for their valuable feedback, and all my colleagues from ArGan'sLab for their constant support. I gratefully acknowledge the funding from MITACS, which allowed me to focus on my studies without financial struggles.

On a more personal note, I have been lucky to be surrounded by a wonderful support system. My thanks go to Kru Shannon, PC Kru Veronica, and my Muay Thai friends for their positivity and mentorship. To Cheryl, Darren, Everett, and Kimberly, thank you for treating me like family. To Jennifer, thank you for never giving up on me.

Finally, to my parents: I owe everything to you. Thank you for your unwavering love and support since day 1.

Table of Contents

Author's Declaration ii

Abstract iii

Acknowledgements iv

List of Abbreviations xii

Chapter 1: Introduction 1

1.1 COVID-19 Pandemic Brings the Light back on Coronaviruses 2

1.2 SARS-CoV-2: The Etiological Agent of COVID-19 Pandemic 5

1.3 Main Protease: A Promising Target for COVID-19 Therapeutics 8

1.4 Coronavirus Mpro: Targets for Pan-Coronavirus Therapeutics 12

1.5 Hypothesis and Objectives 16

Hypothesis 16

Objective 1: Building a New Database to Facilitate Comparative SARS-CoV-2 Mpro-Ligand Exploration 16

Objective 2: Benchmarking a Machine Learning Model for Ligand Pose Selection in Betacoronaviruses Mpro Complexes 17

Objective 3: Integrating a Fine-tuned, Mpro Specific Pose Classification Model to Support Antiviral Search 18

Chapter 2: Methodology 19

2.1 Frontend Web Development 20

2.1.1 Overview and Rationale 20

2.1.2 Core Programming Languages 20

2.1.3 React as a Framework for Frontend Web Development 21

2.1.4 Graphics and Visualization 21

2.2 Backend Web Development 21

2.2.1 Overview and Rationale 21

2.2.2 Flask as a Framework for Backend Web Development 22

2.2.3 Cheminformatics Toolkits 22

2.3 Web Deployment 23

2.3.1 Deployment Overview 23

2.3.2 Heroku for Hosting Web Applications 23

2.3.3 Docker for Containerization 23

2.4 Interaction Fingerprinting 24
2.4.1 Overview 24
2.4.2 Predicting Non-Bonded Interactions 24
2.4.3 Algorithms for Generating IFPs 25
2.5 Tanimoto Coefficient 25
2.5.1 Overview 25
2.5.2 Morgan Fingerprints 26
2.5.3 Tanimoto Coefficient Formula 26
2.6 Lipinski Properties 27
2.6.1 Overview 27
2.6.2 Rule of Five 27
2.7 Molecular Docking 28
2.7.1 Overview 28
2.7.2 Systematic Search 28
2.7.3 Stochastic Search 28
2.7.4 Scoring Function 29
2.8 Machine Learning in Structure-Based Drug Design 30
2.8.1 Overview 30
2.8.2 Ligand Pose Classification with Machine Learning 30
2.8.3 Training and Benchmarking ML Models 31
2.9 Multivariate Data Analysis 31
2.9.1 Overview 31
2.9.2 PCA & Clustering 32
Workflow 33
Chapter 3: A New Database to Facilitate Comparative SARS-CoV-2 Mpro-Ligand Exploration 34
3.1 Introduction 35
3.2 Methodology 40
3.2.1 CoviProDigy Design Strategy 40
3.2.2 Data Input and Validation 40
3.2.3 Data Pre-processing 41
3.2.4 Data Mining and Interaction Profiling 41
3.2.5 Data Clustering 41
3.2.6 Structural Alignment 42

3.2.7 Data Curation and Management 42
3.3 Results 42
3.3.1 Optimized Algorithm for Characterizing Subpocket Occupancy 43
3.3.2. Crystal Ligand Dataset 46
3.3.3 Data Exploration 52
3.3.4 3D Structure Mining and Interaction Profiling 54
3.4 Discussion 58
3.5 Conclusion 59
Chapter 4: Machine Learning for Ligand Pose Selection in Betacoronaviruses Mpro Complexes 61
4.1 Introduction 62
4.2 Methodology 65
4.2.1 *Betacoronavirus* Mpro Dataset Curation 65
4.2.2 *Betacoronavirus* Mpro Redocking 65
4.2.3 Adapting Docking Results for Supervised Learning 66
4.2.4 Training Set-Test Set Partition 66
4.2.5 Training and Benchmarking the ML model 66
4.3 Results 67
4.3.1 Redocking of PanCoronavirus Mpro Ligands 68
4.3.2 Fingerprinting Redocked Pan Coronavirus Ligand Poses for ML Classification 70
4.3.3 Benchmarking ClassyPose using Redocked PanCoronavirus Ligands 72
4.3.4 Optimizing ClassyPose for Pan Coronavirus Mpro Ligand Poses Classification 74
4.3.5 Benchmarking ClassyPose and the Optimized SVM model on Testing Dataset 79
4.3.6 Benchmarking ClassyPose and Optimized SVM model on Hard Dataset 82
4.3.7 Integration of the Optimized SVM Model in CoviProDigy 85
4.4 Discussion 86
4.5 Conclusion 87
Chapter 5: Summary and Conclusion 88
5.1 Overview 89
5.2 The Gap in Current Drug Discovery Pipeline 89
5.3 Key Achievements of this Thesis 90
5.4 Overall Significance 91
5.5 Case Study: Investigating the Binding Mode of a Novel Molecule Docked against SARS-CoV-2 Mpro 91

5.6 Current Limitations 94
5.7 Future Directions 95
5.8 Concluding Remarks 96
References 97
Appendix 112

List of Figures

Chapter 1

Fig. 1.1: Relationship between Vaccination, Death Rate, and Strain Emergence for COVID-19 Pandemic 4

Fig. 1.2: Schematic of the SARS-CoV-2 Virion and its Replicative Cycle 7

Fig. 1.3: Subsite Architecture of SARS-CoV-2 Mpro 9

Fig. 1.4: Binding Modes and Chemical Scaffolds of SARS-CoV-2 Mpro Inhibitors 11

Fig. 1.5: Coronavirus Mpro as a Conserved Target for Broad-Spectrum Antiviral Drug Design 14

Chapter 2

Fig. 2.1: Methodology Workflow Diagram 33

Chapter 3

Fig. 3.1: Active Site Architecture of SARS-CoV-2 Mpro 36

Fig. 3.2: Binding Modes of Selected Peptidomimetic and Non-Peptidomimetic SARS-CoV-2 Mpro Inhibitors 38

Fig. 3.3: Annual Deposition of Ligand-Bound SARS-CoV-2 Mpro in the PDB 43

Fig. 3.4: Characterizing Subpocket Occupancies of SARS-CoV-2 Mpro Ligands 45

Fig. 3.5: Physicochemical Properties Characterization of SARS-CoV-2 Mpro Crystal Ligands 48

Fig. 3.6: Identifying Structurally Related Ligands Across SARS-CoV-2 Crystal Ligands Dataset via Morgan Fingerprint-based Tanimoto Similarity Clustering 50

Fig. 3.7: Identifying Ligands of Interest across Summary Plots 51

Fig. 3.8: Xplorer Tab Interface 53

Fig. 3.9: Architecture of Mine Tab and SARS-CoV-2 Mpro Dedicated Toolkits 55

Fig. 3.10: Comparison of Binding Modes for the X77 and Related Analogues 57

Chapter 4

Fig. 4.1: Conservation of *Betacoronavirus* Active Site Residues 63

Fig. 4.2: Benchmarking the Scoring Function of Vina-GPU 2.0194 via Redocking of Pan Coronavirus Mpro Ligands 69

Fig. 4.3: PLEC Generation of SARS-CoV-2 Ligand X77 71

Fig. 4.4: Outcome of Pan Coronavirus Redocked Poses Classification by ClassyPose 73

Fig. 4.5: Supervised Learning Protocol of SVM ML Model for Pan Coronavirus Mpro Optimization . 76

Fig. 4.6: Outcome of Benchmarking ClassyPose and the Optimized SVM Model on the Testing Dataset 81

Fig. 4.7: Outcome of Benchmarking ClassyPose and the Optimized SVM Model on the Hard Dataset 84

Fig. 4.8: Pose Quality Assessment in CoviProDigy 85

Chapter 5

Fig. 5.1: Characterizing the Binding Mode of the Docked Poses of the Novel Fused Molecule and Novel Fused Molecule Analogues 93

List of Tables

Chapter 3

Table 3.1: Optimized Subpocket Definitions and Distance Cutoff for Assessing Subpocket Occupancies 44

Chapter 4

Table 4.1: Retrieval Pan Coronavirus Crystal Structures 68
Table 4.2: Characteristics of the Training, Testing, and Hard Datasets..... 78

List of Abbreviations

cLogP	octanol-water partition coefficient
COVID-19	Coronavirus disease 2019
ECFP	Extended Connectivity Fingerprint
FDA	Food and Drug Administration
HBA	Hydrogen Bond Acceptor
HBD	Hydrogen Bond Donor
HKU1	Human coronavirus HKU1
HKU4	Tylonycteris bat coronavirus HKU4
IFP	Interaction Fingerprint
MERS	Middle East Respiratory Syndrome
MERS-CoV	Middle East Respiratory Syndrome Coronavirus
ML	Machine Learning
Mpro	Main Protease
MWT	Molecular weight
NSP	Non-Structural Protein
ODDT	Open Drug Discovery Toolkit

ORF	Open Reading Frame
PC	Principal Component
PCA	Principal Component Analysis
PDB	Protein Data Bank
PLEC	Protein-Ligand extended connectivity fingerprints
PLpro	Papain-like protease
PR-AUC	Precision-Recall Area Under Curve
RMSD	Root Mean Square Deviation
RNA	Ribonucleic Acid
ROC-AUC	Receiver Operating Characteristics Area Under Curve
RTC	Replication-Transcription Complexes
SARS	Severe Acute Respiratory Syndrome
SARS-CoV	Severe Acute Respiratory Syndrome Coronavirus
SARS-CoV-2	Severe Acute Respiratory Syndrome Coronavirus 2
SF	Scoring Function
SVM	Support Vector Machine
WHO	World Health Organization

Chapter 1: Introduction

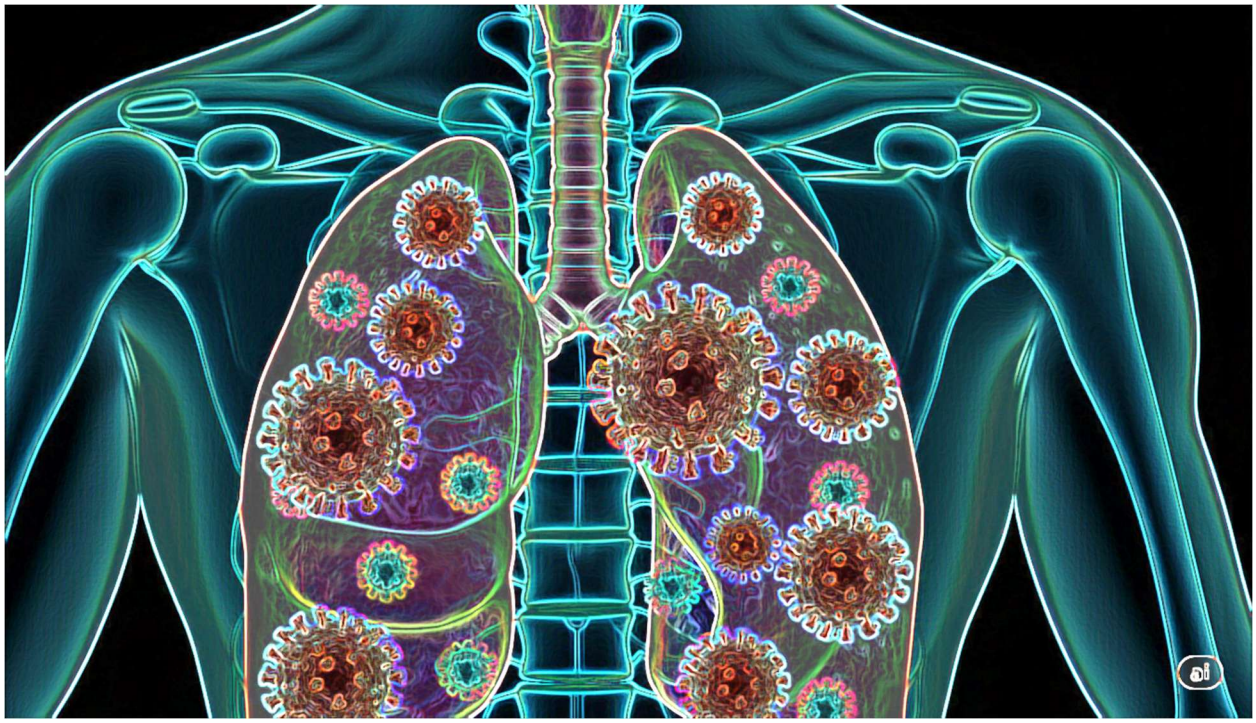


Image adapted from Gemini 2.5. This is a mock image to describe the content of this chapter, and it was not used for any scientific purposes.

1.1 COVID-19 Pandemic Brings the Light back on Coronaviruses

Coronaviruses are single-stranded RNA viruses that have repeatedly crossed species barriers to infect humans, causing significant pathological effects in multiple instances.[1] The first recorded human coronavirus infections date back to the 1960s[2], yet these viruses received limited scientific attention for decades. This changed with episodic spillover events, such as the 2003 outbreak of Severe Acute Respiratory Syndrome (SARS) caused by Severe Acute Respiratory Syndrome Coronavirus (SARS-CoV) which was transmitted from civet cats to humans.[3] Another notable instance was the Middle East Respiratory Syndrome (MERS) outbreak in 2012 caused by Middle East Respiratory Syndrome Coronavirus (MERS-CoV), which emerged from camels[4]. Despite these occurrences, sustained global focus on coronaviruses remained insufficient until the emergence of a novel coronavirus in late 2019, which precipitated a global pandemic claiming the lives of millions of people (Fig. 1.1A) while fundamentally disrupting the global economy.[5,6] Early genomic analyses of this pathogen revealed a 96 percent sequence similarity with SARS-CoV, hence giving its strain name, SARS-CoV-2.[7]

Building on this emergence, SARS-CoV-2, the causative agent of COVID-19, rapidly demonstrated its capacity for global transmission and severe public health impact. Within months of the initial outbreak, the virus spread rapidly across populations worldwide, driving a sharp escalation in infections and fatalities. By March 2020, COVID-19 spread to over 100 countries and claimed the lives of thousands of persons, prompting the World Health Organization (WHO) to declare it a pandemic, necessitating a global lockdown to curb the spread of the disease.[8] Different variants have emerged throughout the pandemic, exhibiting various mutations and degrees of virulence. To track emerging variants, the WHO established a classification system based on their impact and behavior. A variant is first labeled a Variant Under Monitoring (VUM) when it shows signs of growing faster than existing variants. If this growth advantage is confirmed, it is reclassified as a Variant of Interest (VOI). If the VOI is later found to cause more severe disease or have a greater public health impact, it is then designated as a Variant of Concern (VOC).[9] Notable examples include: (i) VUM:

BA.2.86 (Pirola), monitored for its high number of mutations;[10] (ii) VOI: JN.1, a descendant of Pirola (BA.2.86), recognized for its rapid global spread around 2024;[11] (iii) VOC: Alpha (B.1.1.7), Delta (B.1.617.2), and Omicron (B.1.1.529), all of which caused waves of increased transmission and severe virulence.[12,13]

By the end of 2020, two COVID-19 vaccines namely the Pfizer-BioNTech vaccine and the Moderna vaccine were approved by the Food and Drug Administration (FDA) for emergency use, initiating a global immunization campaign.[14] A sustained decline in the mortality rate was observed after February 2022, coinciding with the administration of over 10 billion vaccine dosages worldwide (Fig. 1.1A). By 2023, most lockdown protocols were lifted, and the WHO officially declared an end to the global public health emergency. Nevertheless, SARS-CoV-2 remains in global circulation, continuously evolving through the emergence of new variants. The evolutionary trajectory depicted in Fig. 1.1B shows the successive replacement of earlier dominant strains, such as Delta and the initial Omicron subvariants, by increasingly transmissible descendants. In late 2023 and into 2024, the JN.1 subvariant, a descendant of BA.2.86 (Pirola), emerged from this diverse Omicron background to become the globally dominant strain, characterized by its rapid spread and enhanced immune escape potential.

SARS-CoV-2 is also exhibiting patterns of seasonal transmission, like other respiratory viruses, prompting the need for continuous surveillance.[15] Global tracking systems, including wastewater monitoring, have been established to detect community-level viral trends and the emergence of novel variants.[16] However, the virus's evolutionary potential suggests that future variants could exhibit enhanced virulence or further evade population immunity, posing renewed public health challenges.[11,17] In response, WHO has emphasized the importance of pandemic preparedness through global strategies that include not only surveillance but also the development of effective diagnostic tools and antiviral therapeutics targeting coronaviruses.[18] Establishing a robust molecular toolkit will be essential to rapidly diagnose infections and mitigate the impact of potential future outbreaks.

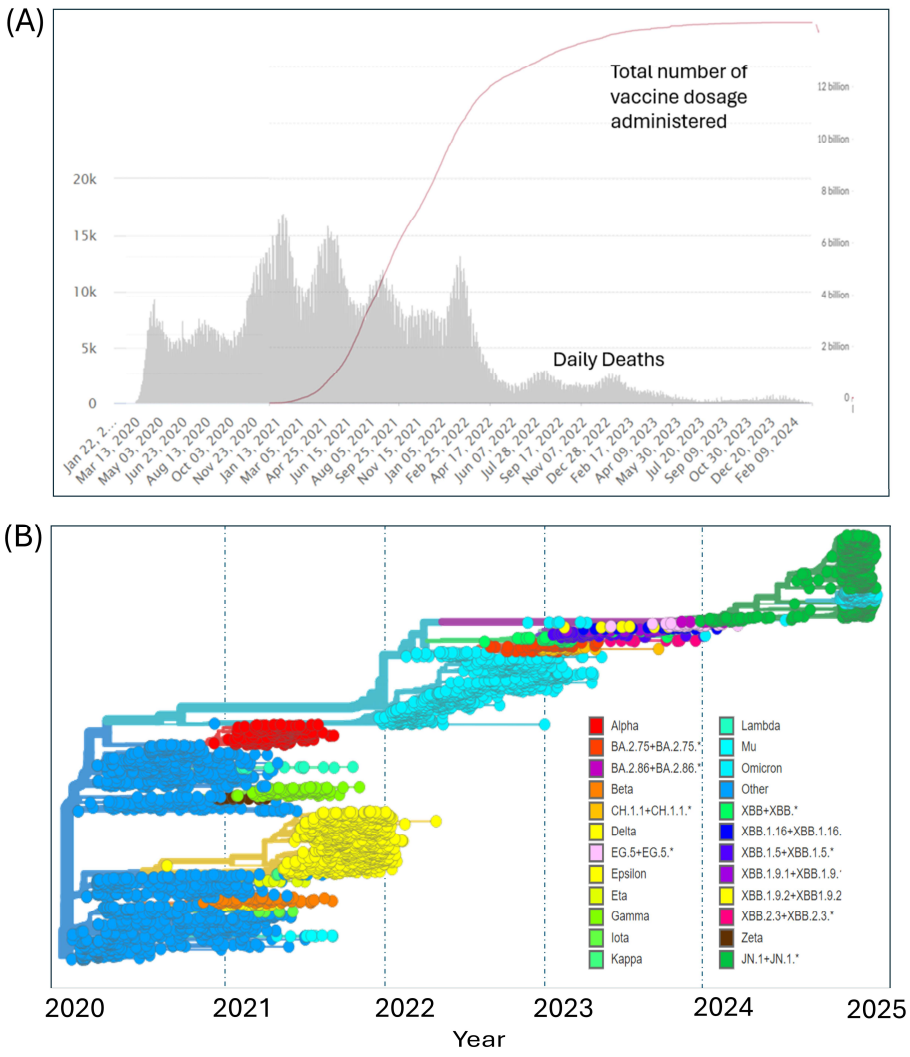


Fig. 1.1: Relationship between Vaccination, Death Rate, and Strain Emergence for COVID-19 Pandemic. (A) Overlaid plots for daily deaths attributed to SARS-CoV-2 and for the total number of SARS-CoV-2 vaccine administered. The daily deaths are shown as a histogram from January 2020 to February 2024. The total number of vaccines administered is shown in red from January 2021 to February 2024. The plot for daily death was obtained from the Woldometer website[19] while the plot for the total number of vaccine dosages administered was obtained from ourworldindata website[20]. **(B)** SARS-CoV-2 Variants Phylodynamic Tree**.** The plot was adapted from the GISAID Database[21], and it highlights the evolution of the genomic sequence of SARS-CoV-2, labeling each sequence according to the

SARS-CoV-2 variant it belongs to. 24 strains are shown in total with the respective color assignment shown as a legend in the plot.

1.2 SARS-CoV-2: The Etiological Agent of COVID-19 Pandemic

SARS-CoV-2 is a single-stranded positive-sense ribonucleic acid (RNA) virus with a 30-kb long genomic RNA.[7] The latter consists of 2 open reading frames (ORFs), ORF1a and ORF1b, encoding 16 non-structural proteins (NSPs) in total and additional segments encoding 4 structural proteins, amongst other proteins (Fig. 1.2A).[22] The structural proteins (spike glycoprotein, membrane protein, envelope glycoprotein, and nucleocapsid protein) form the building blocks of the virus and collectively determine host tropism.[22] Following host cell entry, the NSPs take over, facilitating viral replication while simultaneously impeding the host immune response.[22–25] These NSPs are initially expressed as polyproteins, which are subsequently cleaved to release individual NSPs. SARS-CoV-2 synthesizes two kinds of polyproteins: pp1a, comprising 11 NSPs translated from ORF1a, and pp1ab, comprising 15 NSPs translated from both ORF1a and ORF1b.[13] These polyproteins remain functionally inactive until proteolytic cleavage releases the individual NSPs.

The intricate lifecycle of SARS-CoV-2 begins with the spike protein, the structural glycoprotein exposed on the viral surface[26] (Fig. 1.2B), which mediates initial host cell recognition. The spike's receptor binding domain specifically recognizes and binds to the angiotensin-converting enzyme 2 (ACE2) receptor, which is abundantly expressed in pulmonary and extrapulmonary tissues.[27,28] The widespread distribution and high conservation of ACE2 in humans explain SARS-CoV-2's broad tissue tropism.[29] Once the virus enters the cell, the NSPs previously described come into play. The viral RNA genome hijacks the host's ribosomes for translation, a process facilitated by two eukaryotic-like motifs that ensure recognition by the cellular machinery:[30] (i) 5' methyl cap that prevents RNA degradation and serves as a ribosomal recognition site; (ii) 3' poly(A) tail that enhances RNA stability.

This molecular mimicry enables the virus to integrate into the host's translation system, resulting in the production of the polyproteins pp1a and pp1ab. Papain-like protease (PLpro)

and Mpro cleave these polyproteins to release individual NSPs, which then assemble into Replication-Transcription Complexes (RTCs)[30]. These RTCs coordinate both viral genome replication and protein synthesis, ultimately producing the components needed for virion assembly. Newly formed viral particles are subsequently released from the host cell, continuing the infection cycle. This overall lifecycle is highly conserved across coronaviruses, including SARS-CoV and MERS-CoV, reflecting shared mechanisms of host entry, replication, and assembly.[26]

Given the spike protein's crucial role in mediating host cell infection, it has become a primary research focus, leading to the resolution of numerous crystal and cryo-EM structures deposited in the Protein Data Bank (PDB).[31,32] The spike protein became the primary target for COVID-19 vaccine development, whereby robust neutralizing host antibody responses would be triggered by the spike-based antigen design.[33] However, the spike protein has high mutational tendencies, which pose significant challenges. For example, 6 major mutations (N501Y, E484K, L452R, D614G, K417N/T, and P681R) have been identified that either enhance viral infectivity or facilitate immune escape.[34] This has prompted researchers to explore alternative viral targets, such as NSPs, to broaden therapeutic strategies against SARS-CoV-2.

Importantly, the viral proteases PLpro and Mpro have emerged as attractive targets due to their essential roles in viral replication. PLpro, encoded by nsp3, employs a catalytic triad (cysteine, histidine, and aspartic acid) to cleave polyproteins at LXGG↓XX motifs (where X represents any acid), specifically between NSP1-NSP2, NSP2-NSP3, and NSP3-NSP4.[27] In contrast, Mpro (encoded by nsp5) utilizes a catalytic dyad (cysteine and histidine) to cleave at Q↓X sites (where X is S, A, or N).[27] While PLpro releases NSPs 1-3, Mpro processes the remaining NSPs 4-16, demonstrating broader proteolytic scope. Inhibiting these proteases effectively blocks the proteolytic processing of viral polyproteins, thereby halting viral maturation.[27] Consequently, small-molecule inhibitors of Mpro and PLpro represent promising therapeutic strategies for antiviral development.

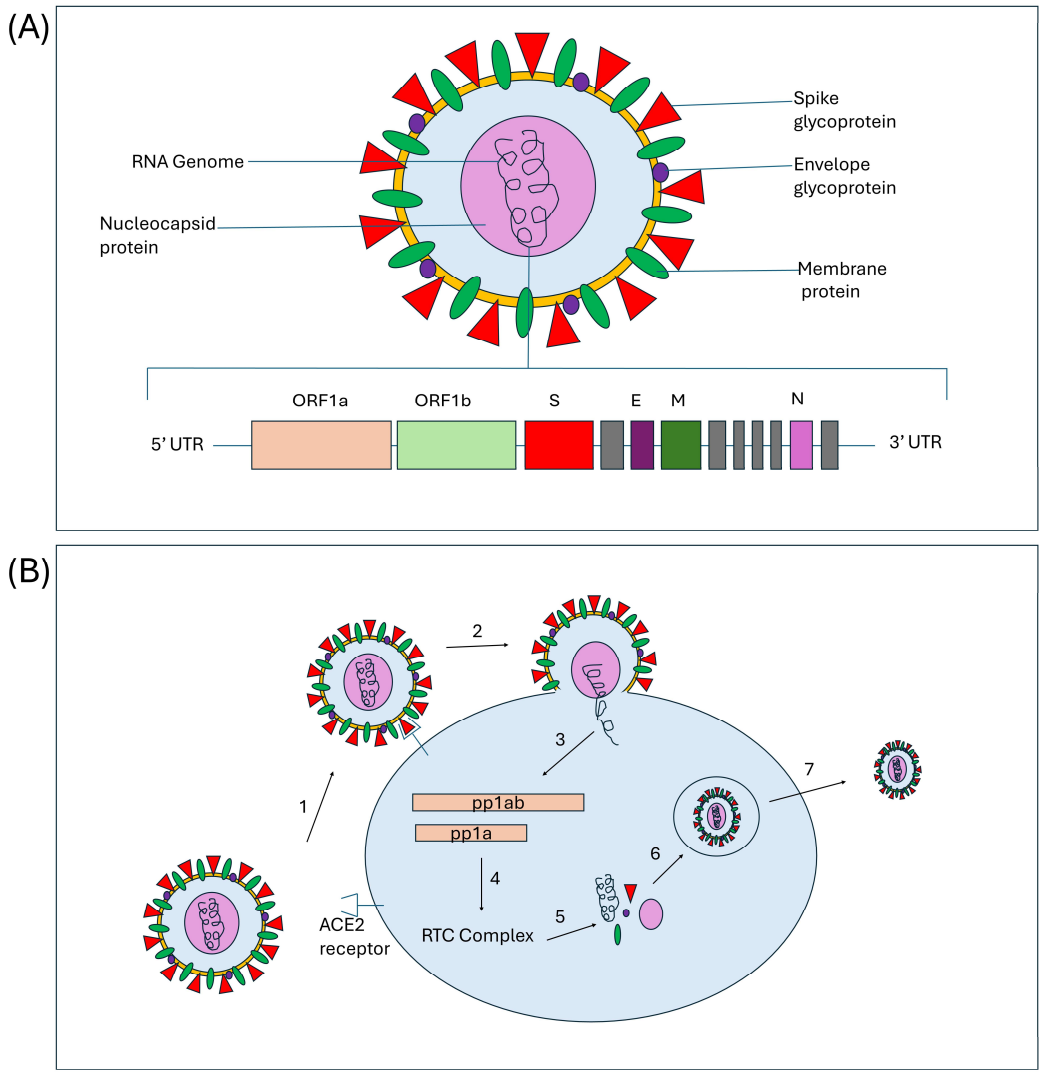


Fig. 1.2: Schematic of the SARS-CoV-2 Virion and its Replicative Cycle.[22] **(A)** Architecture of SARS-CoV-2 virion. The key structural components of the virion are represented schematically: (i) spike glycoprotein (red), (ii) membrane protein (green), (iii) envelope glycoprotein (dark purple), and (iv) nucleocapsid protein (light purple). Key sequences encoding the NPS (ORF1a, ORF1b), the spike protein (S), the envelope glycoprotein (E), the membrane protein (M), and the nucleocapsid (N) are highlighted in the open reading frame of the RNA genome. **(B)** Simplified lifecycle of SARS-CoV-2. Step 1 represents the attachment of the spike protein of the virus to the ACE2 receptor of the host cell. Step 2 is the release of the viral genome in the cell's cytoplasm. Step 3 is the translation of the viral genome, yielding polyproteins pp1a and pp1ab. Step 4 is the formation of the RTC after proteolytic processing

of the polyproteins by PLpro and Mpro. Step 5 is the production of viral components by the RTC. Step 6 is the assembly of the newly made viral components into a new viral particle. Step 7 is the release of the viral particle from the host cell.

1.3 Main Protease: A Promising Target for COVID-19 Therapeutics

Given the importance of Mpro in the replication cycle of coronaviruses, understanding Mpro's structure-function relationship becomes crucial for drug design. This has driven significant research efforts into elucidating the structure of SARS-CoV-2 Mpro and identifying potent inhibitors.[35–40] This enzyme functions as a homodimer, with each protomer consisting of 306 amino acids organized into 3 distinct domains (Fig. 1.3A): domain I (residues 8-101) and II (residues 102-184) form the catalytic core, while domain III (residues 201-303) mediates dimerization.[41] The active site, located in a cleft between domains I and II, houses the catalytic dyad His41 and Cys145 that executes proteolytic cleavage.[41]

Substrate specificity is governed by the 4 well-defined subsites within this catalytic cleft: S1, S1', S2, and S4 (Fig. 1.3A).[42] Each subsite accommodates a corresponding substrate position (P1, P1', P2, and P4, respectively), creating a precise molecular recognition system (Fig. 1.3B). The S1 subsite, also referred to as the lateral pocket, exhibits the highest selectivity, exclusively accepting glutamine or glutamine mimetics at P1.[42] Hydrogen bonding interaction between P1 and His163 (residue from S1) has been shown to enhance substrate binding affinity.[43] Dimerization is essential for the activity of Mpro as the N-terminal serine from one protomer (N-finger) interacts with the side-chain of Glu166 and the backbone of Phe140 of the other protomer to stabilize the S1 subsite for substrate binding.[44] In contrast, the S1' subsite's spatial constraints accommodate only small amino acids at P1'.[42] The S1 and S1' subsites are linked by an oxyanion loop (residues 138 to 145) which is involved in stabilizing the intermediate state formed during proteolysis.[45,46] Specifically, the backbone amides of Gly143, Ser144, and Cys145 form hydrogen bond interactions with the tetrahedral acyl transition state for product formation.[45] The S2 pocket preferentially binds hydrophobic residues at P2 to match its hydrophobic nature, while the S4 pocket requires small residues

at P4 due to spatial constraints.[42] Fig. 1.3C provides a comprehensive mapping of residues defining each subsite.

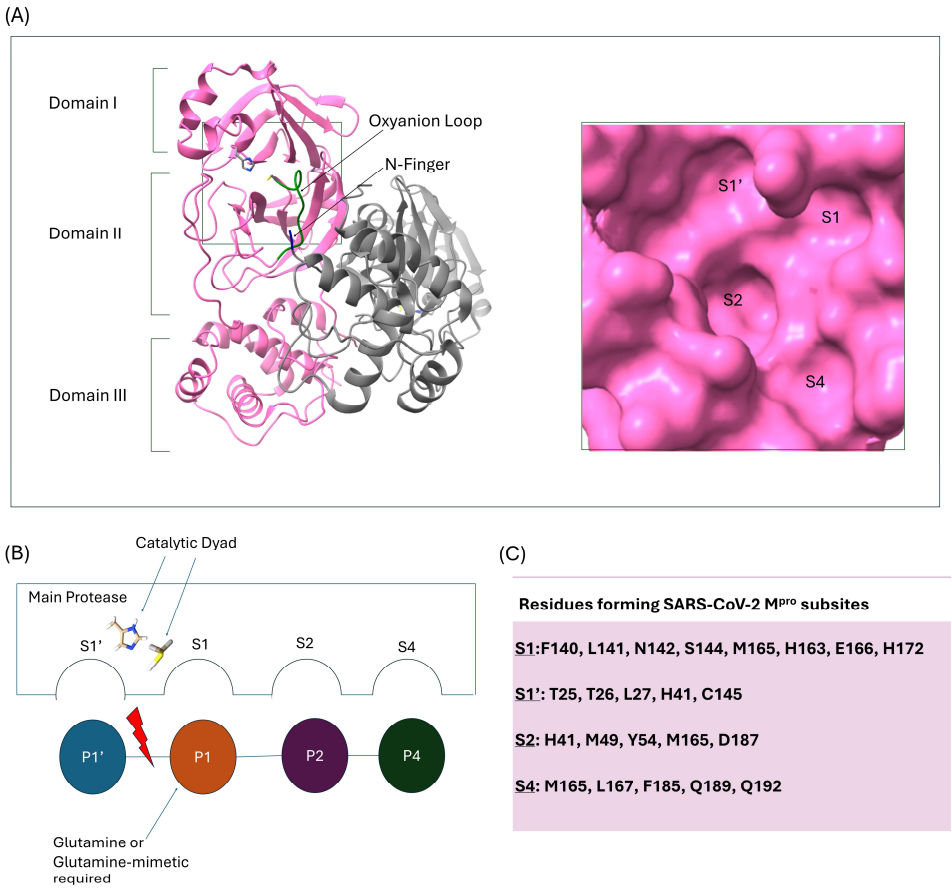


Fig. 1.3: Subsite Architecture of SARS-CoV-2 Mpro. **(A)** The dimeric state of Mpro is shown in cartoon representation using PDB code 6Y2G. One protomer is shown in pink and the other in grey. Domains 1 to 3 are highlighted for one protomer. The oxyanion loop is colored in green on one protomer while the N-finger is colored in blue in the other protomer. The catalytic dyad, His41 and Cys145, is shown in stick representation. The active site is boxed in one protomer and a zoomed-in view in surface representation is given highlighting the location of the 4 principal subsites: S1, S1', S2, and S4. The image was created using UCSF ChimeraX.[47] **(B)** Schematic diagram illustrating substrate binding to Mpro. The 4 major subsites are shown along with their complementary substrate moieties. At P1 position, either a glutamine or a glutamine-mimetic is required and the bond between P1 and P1' is hydrolyzed during polyprotein processing. **(C)** Identities of residues lining SARS-CoV-2 Mpro

subsites.[42] The identities of residues comprising each of the 4 major subsites (S1, S1', S2, S4) of Mpro are listed.

Extensive computational screening[48–54] and experimental characterization[35–40] efforts against Mpro have identified several promising inhibitors, for example GC-376, X77, and boceprevir to name a few (Fig. 1.4A-C). GC-376, a covalent inhibitor, exhibits the highest potency among the three with an IC50 of 30 nM[39], efficiently occupying all four subsites (S1, S1', S2, and S4) and forming hydrogen bonds with Cys145, Glu166, His163, His164, Gly143, and GLn189 (Fig. 1.4A). In contrast, boceprevir, also a covalent inhibitor, is less potent, with an IC50 of 4.1 μM[39], though it similarly spans all subsites and interacts with His41, Gly143, His164, and Glu166 via hydrogen bonding and Met49, and Met165 via hydrophobic contact (Fig. 1.4C). X77, a non-covalent inhibitor, has a reported IC50 value of 4.1 μM[55], binding all subsites except S4[56]. It forms hydrogen bonds with Glu166, His163, Gly143, and engages in a distinctive stacking interaction with His141, enhancing its stability within the active site (Fig. 1.4B).

Such rigorous efforts culminated in the development of Paxlovid, the first FDA-approved oral antiviral for COVID-19.[57] Paxlovid combines Nirmatrelvir (the Mpro inhibitor) with ritonavir (a pharmacokinetic booster). Nirmatrelvir evolved from PF-0083521, originally developed for SARS-CoV, leveraging the 96% sequence similarity between SARS-CoV and SARS-CoV-2 Mpro.[7] Key optimizations included: (i) a nitrile warhead at P1' for covalent binding to Cys145, (ii) a cyclic leucine mimetic at P2 to complement the hydrophobic S2 pocket, and (iii) trifluoroacetamide at P4 to fit the constrained S4 pocket(Fig. 1.4D).[7,27] Similarly, Ensitrelvir (oral Mpro inhibitor first approved in Japan, now marketed as Xocova) emerged from systematic screening focused on compounds that could hydrogen bond with His163 and Glu166 in the S1 pocket, while optimizing hydrophobic interactions with the S2 pocket (Fig. 1.4D).[37,57] The successful development of these clinically approved antivirals supports the structure-based design approach for Mpro inhibition. The systematic optimization strategies employed, from initial scaffold identification through subsite-specific modifications, provide a robust framework that can be applied to develop next-generation inhibitors with enhanced properties.

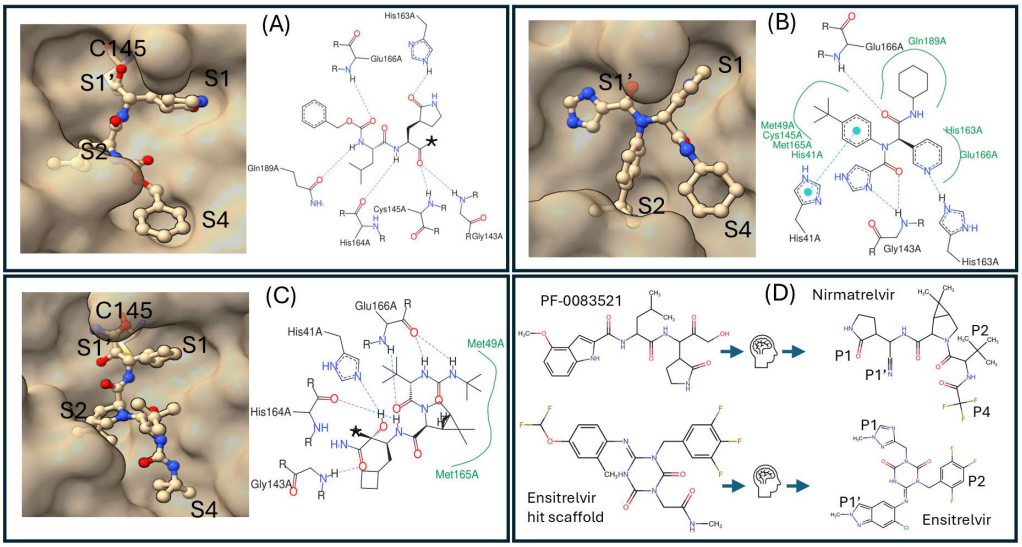


Fig. 1.4: Binding Modes and Chemical Scaffolds of SARS-CoV-2 Mpro Inhibitors. The binding modes of **(A)** GC-376 (PDB Code: 7D1M), **(B)** X77 (PDB Code: 6W63), and **(C)** Boceprevir (PDB Code: 7K40) are given. The subpocket fit for each ligand was illustrated using UCSF ChimeraX[47] with the protein backbone in surface representation and the ligand in ball-and-stick representation. Each Mpro subsite (S1, S1', S2, S4) is annotated on the diagram and CYS145 is annotated for covalent ligands (A) GC-376 and (C) Boceprevir. The 2D interaction was generated using PoseEdit[58] from the ProteinsPlus server.[59] Non-bonded interactions represented in the 2D interaction diagram are: (i) hydrogen bonds in blue dotted lines linking the hydrogen bond donors and acceptors, (ii) hydrophobic contacts in green contour lines surrounding the ligand and the residues color coded in green, and (iii) aromatic stacking interaction in turquoise dotted lines linking the centroids (also rendered in turquoise) of the aromatic rings involved. For covalent ligands, an asterisk (*) is used to denote the covalent bond linking Cys145, as in the case for (A) GC-376 and (C) Boceprevir. **(D)** Hit-to-Lead optimization for oral Mpro inhibitors Nirmatrelvir and Ensitrelvir. The moieties targeted by the optimization step are labeled on the lead compounds. The 2D diagram was created using the Chemical Sketch Tool powered by ChemAxon on PDB.[60]

1.4 Coronavirus Mpro: Targets for Pan-Coronavirus Therapeutics

The coronavirus family (Coronaviridae) comprises four genera: *Alphacoronavirus* (17 species), *Betacoronavirus* (16 species), *Gammacoronavirus* (2 species), and *Deltacoronavirus* (7 species) (Fig. 1.5A).[61,62] Phylogenetic analysis reveals distinct host tropism: *Alpha-* and *Betacoronaviruses* primarily infect mammals while *Gamma-* and *Deltacoronaviruses* predominantly infect birds, though some exhibit cross-species transmission to mammalian hosts.[63] This phylogenetic distribution underscores the zoonotic potential of coronaviruses, particularly betacoronaviruses, which display greater genetic diversity and adaptability to human hosts.[64] Indeed, betacoronaviruses have emerged as persistent threats to global health, causing 3 major outbreaks within 2 decades: SARS (2003), MERS (2012), and COVID-19 (2020-present) (Fig. 1.5A). Betacoronaviruses rely on conserved proteases, including Mpro, for viral maturation.[26]

Remarkably, the Mpro of betacoronaviruses exhibits a high degree of sequence conservation, particularly within the catalytic cleft. As shown in Fig. 1.5B-E, active residues remain highly conserved across betacoronaviruses, with pairwise sequence identities to SARS-CoV-2 Mpro ranging from approximately 50% to 96% (Appendix: Table A2). Specifically, SARS-CoV-2 Mpro shares ~96% sequence identity with SARS-CoV Mpro, while identity decreases to ~50% with human coronavirus HKU1 (HKU1) Mpro. To date, numerous crystal structures of Mpro have been determined for key betacoronaviruses, including SARS-CoV (68 structures), MERS (34 structures), HKU1 (1 structure), Tylonycteris bat coronavirus HKU4 (HKU4) (6 structures), and SARS-CoV-2 (1257 structures).[60] This abundance of SARS-CoV-2 Mpro reflects the unprecedented research focus during the COVID-19 pandemic, as this protease became an attractive therapeutic target.

Despite the inherently high mutation rate of RNA viruses, Mpro maintains remarkable structural integrity and sequence conservation within its active site across evolving strains. This conservation likely reflects evolutionary constraints imposed by essential catalytic requirements and oxidative defense mechanisms. Moreover, specific residues prove critical for maintaining enzymatic function: mutation of His163 to alanine in SARS-CoV-2 Mpro

triggers formation of a disulfide bond between catalytic Cys145 and Cys117, resulting in enzyme inactivation.[65] Similar inactivation occurs in SARS-CoV Mpro upon Asn28-alanine mutation.[66] Importantly, mutation surveillance during the COVID-19 pandemic demonstrated that frequently observed SARS-CoV-2 Mpro mutations, G15S, T21I, L89F, K90R, P132H, L205V, and A260V (Fig. 1.5F), map exclusively outside the active site and did not compromise inhibitor binding.[67] This suggests that resistance mutations within the active site may compromise enzymatic function.

Given this structurally constrained nature, Mpro has become an attractive target for pan-coronavirus inhibitor development.[68–71] Moreover, human proteases that are closely related to Mpro do not require the presence of a glutamine/glutamine-mimetic at their cut sites, so that the risk of off-targeting is low.[68] Current efforts have yielded promising early results. For instance, plant-derived compounds such as Hypericin, Isorhamnetin, and Xanthohumol have demonstrated inhibitory activity against multiple coronaviral Mpro, though further optimization is required to enhance their potency.[68,69] Additionally, Nirmatrelvir, the active component of Paxlovid, has shown broad-spectrum activity against both *Alpha*- and *Betacoronaviruses* in vitro, lending support to the pan-inhibitor concept[72]. However, clinical translation faces limitations, including the need for pharmacokinetic boosting and the lack of comprehensive animal model validation.[72,73]

These efforts align with the WHO's pandemic preparedness agenda, which advocates for proactive therapeutic development to counter future coronavirus outbreaks.[8] Therefore, this thesis aims to contribute to this global effort by developing a molecular toolkit to probe ligand-Mpro interactions in betacoronaviruses with known 3D structures, specifically HKU1, HKU4, MERS-CoV, SARS-CoV, and SARS-CoV-2.

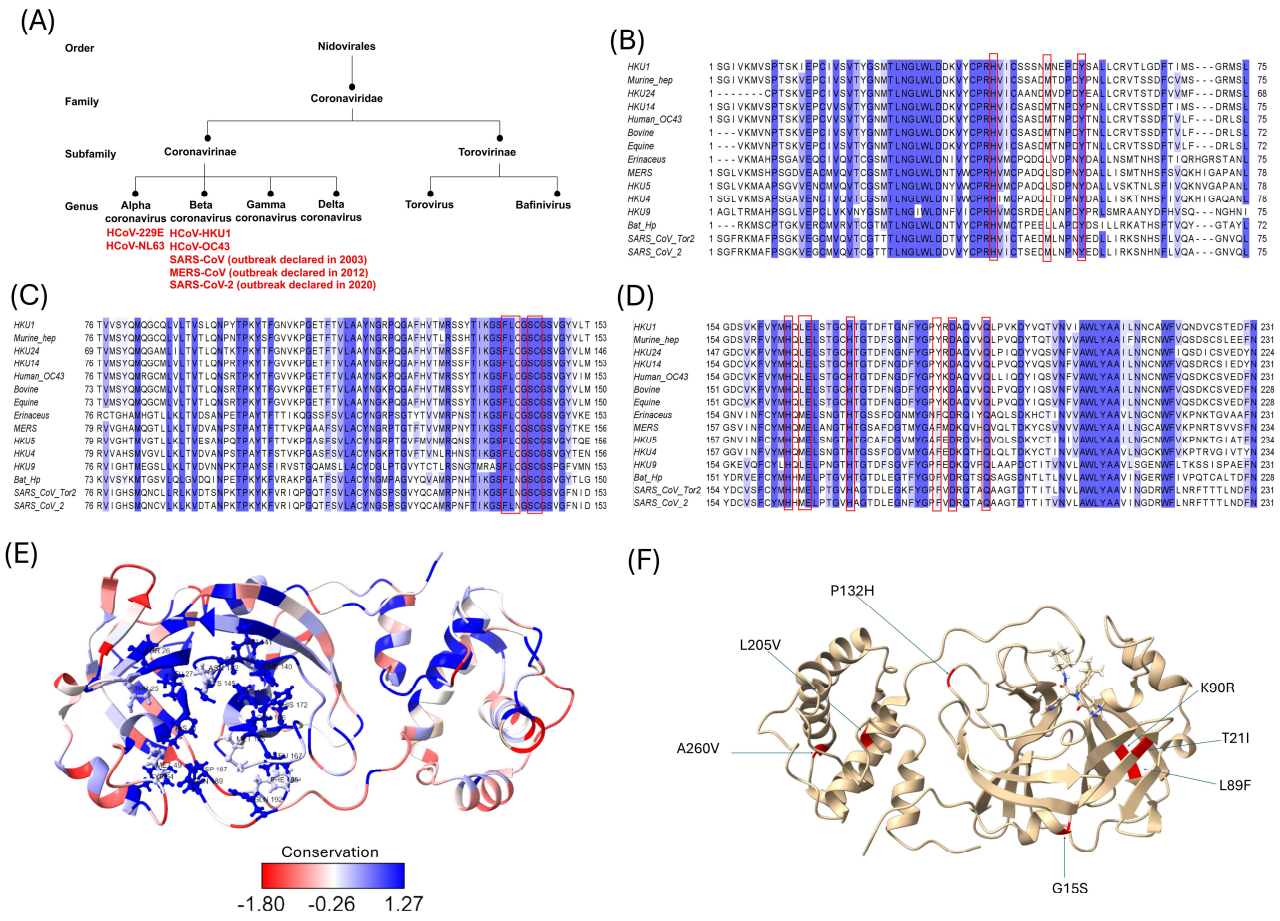


Fig. 1.5: Coronavirus Mpro as a Conserved Target for Broad-Spectrum Antiviral Drug Design. (A) Phylogenetic tree of coronavirus[61,74]. Coronaviruses belong to the order of *Nidovirales* which can be broken down into 2 subfamilies. *Coronaviridae* is the subfamily of interest as it contains 2 genera that can infect humans, *Alphacoronavirus* and *Betacoronavirus*. Species from those genera that pose a threat to human health are highlighted in red, along with the date of first recorded outbreak where applicable. **(B)-(D)** Sequence alignment of *Betacoronavirus* Mpro. Clustal Omega,[75] accessible via the European Molecular Biology Laboratory's European Bioinformatics Institute,[76] was used to perform the sequence alignment using the Mpro NCBI sequence reference sequence for 15 betacoronaviral strains.[62] Their NCBI accession codes are given in Table A1 (Appendix). Only portions of the generated alignment plot corresponding to the subsite residues are shown and these are boxed in red.[77] **(E)** Mpro structure color-coded according to sequence

conservation across betacoronaviruses. The residues from the protomer of SARS-CoV-2 Mpro (PDB: 6W63) are color coded according their conservation score calculated using Clustal Omega.[75] The color palette ranges from red (least conserved) to blue (highly conserved). Residues that were not assigned a conservation score are colored in tan. The protein backbone is styled in cartoon representation while subsite residues are labeled and shown in ball-and-stick representations. The image was created using UCSF ChimeraX.[47] **(F)** Mutations of Mpro. The commonly observed mutation hotspots in Mpro are colored in red and labeled according to the residue substitution, while the protein backbone is colored in tan and rendered in cartoon representation. The PDB code 6W63 was used for this representation and the ligand X77 is shown in stick representation to highlight the location of the active site. UCSF ChimeraX was to generate the image.[47]

1.5 Hypothesis and Objectives

Hypothesis

Despite the significant accumulation of structural data for coronavirus Mpro, particularly for SARS-CoV-2 Mpro with over 1000 deposited structures, there remains a critical gap in leveraging this information cohesively for pan-coronavirus drug discovery. The high sequence conservation of Mpro active sites across betacoronaviruses (50-96% identity) and the structural constraints that render resistance mutations within the catalytic region unfavorable suggest that systematic analysis of ligand-Mpro interactions could reveal design principles for pan-inhibitors.

This thesis hypothesizes that by developing an integrated computational platform that combines structural database mining and machine learning-guided pose selection, insights into molecular scaffolds and binding modes may be obtained to assist with drug design for pan-inhibition.

Objective 1: Building a New Database to Facilitate Comparative SARS-CoV-2 Mpro-Ligand Exploration

Despite the accumulation of crystal structures of SARS-CoV-2 Mpro complexed with diverse ligands, current bioinformatics tools and databases restrict analyses to individual protein-ligand complexes. This fragmentation prevents researchers from leveraging collective structural insights for rational drug design.

A comprehensive platform will be developed using modern web technologies and specialized cheminformatics toolkits, to create an integrated database dedicated to Mpro-ligand interactions. This platform will feature automated extraction of ligand-bound Mpro complexes from the PDB, real-time calculation of interaction fingerprints, comparative analysis of subsite-occupancies, assessment of Lipinski properties[78], identification of binding site water molecules and both 2D and 3D visualization capabilities of ligand involvement with Mpro. The system will employ specialized plotting libraries for generating

interactive structural diagrams and implement plugins for multi-structure comparative analyses to facilitate scaffold hopping and medicinal chemistry optimization.

Expected Outcome: A fully functional web-based platform enabling researchers to perform rapid comparative analyses across deposited and user-supplied ligand-bound Mpro complexes will be created. The database will feature user-friendly toolkits to guide the ligand exploration of SARS-CoV-2 Mpro.

Objective 2: Benchmarking a Machine Learning Model for Ligand Pose Selection in Betacoronaviruses Mpro Complexes

Despite molecular docking's ability to generate near-native ligand poses within protein binding sites, integrated scoring functions often fail to correctly prioritize optimal conformations. This pose selection problem represents a critical bottleneck in virtual screening campaigns, where incorrect pose identification can propagate errors throughout the drug discovery pipeline and amplify financial burdens.

A GPU docking pipeline will be created for large-scale redocking of ligand-bound *Betacoronavirus* Mpro deposited in the PDB. A published machine learning (ML) classification model (ClassyPose[79]), will be benchmarked to distinguish between near-native and non near-native binding poses of Mpro ligands. Unlike traditional scoring functions that rely on explicitly defined interaction energy terms, this ML model will utilize protein-ligand extended connectivity fingerprints[80] (PLEC) that implicitly define the interaction profile of the ligand for the classification task.

Expected Outcome: A GPU docking pipeline tailored for large-scale docking of *Betacoronavirus* Mpro will be created. The efficiency of the ML model to support accurate pose selection from large number of poses obtained through virtual screening during the initial stages of drug discovery efforts against *Betacoronavirus* Mpro.

Objective 3: Integrating a Fine-tuned, Mpro Specific Pose Classification Model to Support Antiviral Search

The efficiency of ML model is highly dependent on the quality of data on which it is trained. Currently, there are limited pose selection models which have been specifically developed and optimized for *Betacoronavirus* Mpro. Specifically, the ML model (ClassyPose in Objective 2) chosen for this thesis was initially developed as a general-purpose pose classifier. Therefore, its performance might improve when retrained with Mpro specific data (Chapter 4).

For this purpose, this objective will focus on finetuning the ML model (ClassyPose) using all of the known crystal structures of *Betacoronavirus* M^{pro}, reported in the PDB. Thousands of docking poses generated through GPU-docking (in Objective 2) will be used in the retraining process of ClassyPose. Initially, detailed Mpro-ligand fingerprints ($PLEC^{80}$) of all the docking poses are calculated and used for retraining the ML model. Automated pipelines are developed in-house for all preprocessing, such as data cleaning, molecular fitting, and classification.

Expected Outcome: A finetuned ML model will be implemented for reliable Mpro ligand pose classification, which is an inherent challenge with traditional virtual screening. The optimized model will be integrated into the in-house developed Mpro-specific database (CoviProDigy in Chapter 3) to support rapid ML classification. This will offer a robust framework for supplementing virtual screening campaigns for antiviral design and discovery.

Chapter 2: Methodology

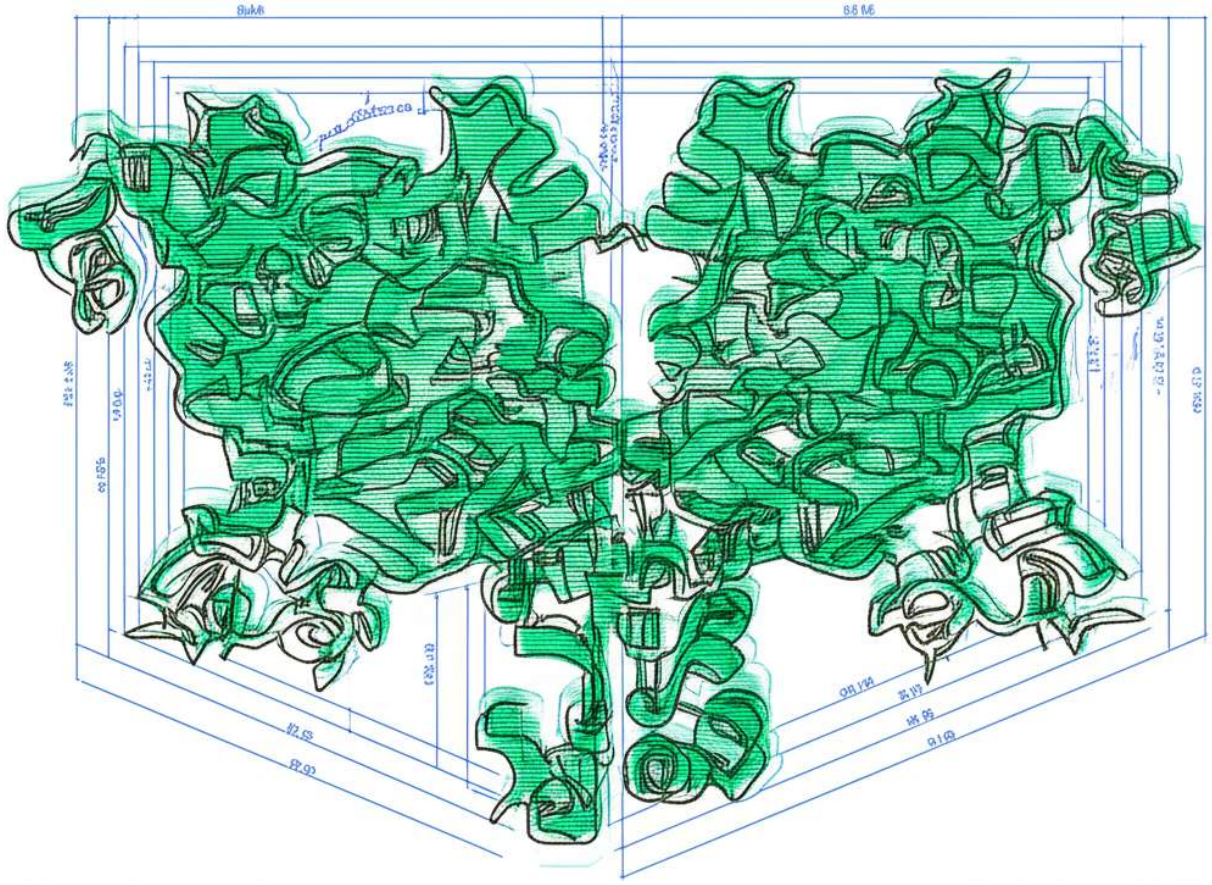


Image adapted from Gemini 2.5. This is a mock image to describe the content of this chapter, and it was not used for any scientific purposes.

2.1 Frontend Web Development

2.1.1 Overview and Rationale

Frontend web development refers to the creation of the visual and interactive elements of web applications, which are responsible for accepting user input and interactions.[81] It is essential for delivering a user-friendly interface that simplifies access to drug discovery platforms. Typically, command-line interfaces are used to run drug discovery plugins,[82–84] which may impose a barrier to those without extensive programming knowledge. Frontend web development bridges this gap by adding intuitive interfaces that transform communication with drug discovery plugins into a matter of clicks.[81] Modern frontend technologies facilitate the development of interactive platforms that can handle large-scale molecular data while maintaining performance and usability.

2.1.2 Core Programming Languages

Frontend web development stems from 3 core programming languages:[81] HyperText Markup Language (HTML), Cascading Style Sheets (CSS), and JavaScript (JS). HTML provides the blueprint of web applications by defining the elements, such as click buttons, input fields, drop-down menus, to name a few, to be incorporated and how to organize them.[85] CSS controls the aesthetic aspect, enabling consistent styling and responsive design across different devices and screen sizes.[81] JS serves as the programming language that renders the web application interactive while also handling communication with the drug discovery plugins running in the background.[81,86] However, JS is a dynamically typed language, and scripting errors are only identified at runtime, hindering the development process.[86] To mitigate this, TypeScript is employed, which is a statically typed superset of JS.[86] With TypeScript, errors are caught at compile time rather than run time, leading to more robust and maintainable code bases.[86]

2.1.3 React as a Framework for Frontend Web Development

Don't Repeat Yourself (DRY) is a common software development practice aimed at minimizing code duplication for creating maintainable, scalable, and robust codebases.[87] React is a specialized JS library designed by Facebook (now Meta) for frontend web development that is anchored on the DRY principle.[88] It allows the splitting of the entire codebase into a component-based architecture where each encapsulated component is now reusable.[88] React also offers performance optimization by introducing a virtual Document Object Model (DOM).[88] This minimizes the need to interact with the browser's DOM, which may be inefficient, and it allows for selective component re-rendering so that unchanged components are not unnecessarily reloaded.[88]

2.1.4 Graphics and Visualization

Scalable Vector Graphics (SVG) is an XML-based markup language for generating 2D graphics.[89] Unlike popular image formats such as JPG and PNG, SVGs are resolution-independent, making them scalable without loss of quality.[89] SVG components are compatible with both CSS for styling and JS for image manipulation.[89] Thus, the main benefit of SVG components is their direct manipulation with the core programming languages for frontend web development. D3.js is a powerful JS library that leverages this key characteristic of SVG components for creating visual elements in web applications.[90] D3.js offers fine-grained control over every step involved in this process, making it an attractive library for generating complex visual elements.

2.2 Backend Web Development

2.2.1 Overview and Rationale

Backend web development refers to setting up the server-side infrastructure for handling computational requests initiated by the frontend, which include data retrieval from databases and data processing involving specialized plugins.[91] Typically, those specialized plugins are accessible via a command-line interface, without an explicit framework for handling job submission, job queuing, and result manipulation. Backend web development

bridges this gap by first exposing these plugins to a user-friendly interface and by setting up a pipeline for efficient job execution and result manipulation.[91]

2.2.2 Flask as a Framework for Backend Web Development

Backend Web Development requires the creation of specialized endpoints, also known as Application Programming Interface (API), through which the frontend can communicate with the backend.[92] Flask is a Python package often used to establish these APIs, owing to its minimalistic nature compared to other options such as Django.[93] Another advantage of using Flask is the implementation of Python for backend development, particularly in the context of drug discovery. Specialized cheminformatics toolkits such as RDKit,[82] OpenBabel,[83] and Open Drug Discovery Toolkit[84] (ODDT) all have Python bindings. Thus, Flask offers a framework for the direct integration of these popular libraries, avoiding environment-related conflicts.

2.2.3 Cheminformatics Toolkits

RDKit is a comprehensive open-source cheminformatics toolkit available for drug discovery applications.[82] The toolkit provides extensive functionality for molecular manipulation, including reading and writing various chemical file formats, generating molecular fingerprints, calculating physicochemical properties, and performing substructure searches.[82] RDKit's integration with Python enables seamless incorporation into web applications, allowing server-side processing of chemical structures.

OpenBabel is an open-source cheminformatics toolbox designed to handle the diverse range of file formats encountered in computational chemistry.[83] The toolkit excels at format interconversion, supporting over 110 chemical file formats.[83] Beyond simple conversion, OpenBabel provides functionality for adding hydrogens, generating 3D coordinates, and calculating molecular properties.[83] Its Python bindings make it easily integrable into automated workflows.

Lastly, ODDT is another open-source cheminformatics toolkit built on existing state-of-the-art open source drug discovery tools and algorithms.[84] This library supports a variety of file

formats as it was built around existing toolkits such as RDKit and OpenBabel.[82–84] ODDT provides methods for characterizing protein-ligand interactions and for filtering small molecules based on their physicochemical properties.[84] Moreover, ODDT also offers a user-friendly docking stack that can be accelerated for large-scale docking, which includes machine learning-based scoring functions for assessing docked poses.[84]

2.3 Web Deployment

2.3.1 Deployment Overview

Web deployment refers to transforming a local web application into an online platform that can be accessed via a URL. This requires the hosting of backend and frontend components on a cloud platform.[94] One critical step during the deployment process is establishing the native runtime environments for both frontend and backend components on the cloud platform. Improper configurations will lead to inconsistent behavior of the web application across development and production environments.

2.3.2 Heroku for Hosting Web Applications

Heroku offers a comprehensive solution for web deployment as it handles the infrastructure and deployment complexity involved.[95] It is viewed as a beginner-friendly service offering simple deployment workflows for the deployment process. One example is the "buildpack" system utilized by Heroku to automate the configuration of runtime environments for web applications.[95] A major challenge with this approach is that it is not guaranteed to work, especially when scaling up applications. However, this issue is easily overcome by using Heroku Container Registry, where Docker containers encapsulating the native runtime environment of the web application can be deployed.[96]

2.3.3 Docker for Containerization

Docker is a software used to package an application along with its dependencies in a standalone executable package known as a container.[96] This container allows the app to be compiled in systems other than its native one, without any conflict, as the application is now

supplied along with its independent compiler.[97] The requirements for setting up such compiler are listed in a Dockerfile. The latter defines a series of steps for creating a container which include: (i) Base image selection (e.g., Python) (ii) Source code for the application (iii) Dependencies to be installed (iv) Exposure to network ports for web access.

2.4 Interaction Fingerprinting

2.4.1 Overview

Non-covalent interactions such as hydrogen bonding, hydrophobic contacts, and stacking interactions play a key role in biological processes.[44,65,98,99] An example is the stabilization of the S1 subsite of Mpro by the N-finger of its protomer via hydrogen bonding.[44] The presence and strength of non-covalent interactions are influenced by the interaction-type specific distance and angle requirements that must be met by the interacting partners. Thus, an accurate computer representation of these interactions can assist with computer-aided drug discovery applications such as predicting binding affinities, and studying ligand unbinding pathways from molecular dynamics simulation.[100,101] This class of representation is known as interaction fingerprints (IFPs), which encode the non-bonded interactions between a protein and a ligand in a vector format for machine representation.[102]

2.4.2 Predicting Non-Bonded Interactions

Accurate prediction of non-bonded interactions is a requisite to ensure the reliability of the interaction fingerprints. Hydrogen bonding interactions occur when a hydrogen atom, covalently bonded to an electronegative atom (donor) such as oxygen or nitrogen, interacts with another electronegative atom (acceptor). The difference in electronegativities between the donor atom and the hydrogen atom creates an asymmetric distribution of the electron cloud of the covalent bond, favoring the donor atom while imparting a positive charge on the hydrogen atom. The electropositive hydrogen can now form electrostatic interactions with other electronegative atoms. The geometric criteria for hydrogen bond detection involve a donor-acceptor distance of less than 3.5 Å and a donor-H-acceptor angle greater than 120°.[103] Hydrophobic interactions occur when non-polar moieties interact with each other,

typically excluding water molecules, thereby creating a favorable free energy change. The geometric criterion used to identify hydrophobic interactions involves a distance cutoff of less than 4 Å between hydrophobic carbon atoms.[103] Lastly, stacking interaction occurs due to the overlapping of p-orbitals between electron-rich and electron-deficient aromatic rings. The geometric criterion used to identify stacking interactions is a distance cutoff of less than 5 Å between the centroids of the aromatic rings.[103] An implementation of this interaction prediction guideline can be found in ODDT.[84]

2.4.3 Algorithms for Generating IFPs

Different algorithms have been developed for converting the interacting space between a protein and a ligand into a vector format. For instance, the Structural Interaction Fingerprint (SIFt) creates a 7-bit representation encoding the interacting amino acid by its interaction type such as hydrophobic, polar, and hydrogen bonding interactions.[102] The involvement of the residue's side chain or backbone is also noted in the SIFt fingerprint.[102] Moreover, this fingerprinting technique was optimized to include aromatic and ionic interactions.[102] A more advanced representation is the PLEC, which is suited for machine learning.[80] The PLEC is an extension of the Extended Connectivity Fingerprint (ECFP), a fingerprinting technique primarily based on atomic features, which considers only interacting atoms between the protein and ligand.[80,84] With the PLEC, protein-ligand interactions are implicitly represented via the pairing of their ECFP environment. Recently, a machine-learning model for addressing the ligand pose classification problem was trained using the PLEC from docked complexes.[79]

2.5 Tanimoto Coefficient

2.5.1 Overview

The Tanimoto coefficient, also known as the Jaccard Index, is a mathematical framework for computing molecular similarity based on shared structural motifs. It requires a vector representation, computed using fingerprinting methods, of the molecules to be compared. While other frameworks are available for assessing molecular similarities, several

benchmark studies have consistently supported the appropriateness of the Tanimoto coefficient for drug discovery-related applications.[104–106] The coefficient ranges from 0 (no similarity) to 1 (identical molecules), providing an intuitive scale for assessing structural relatedness. The Tanimoto coefficient is dependent on the fingerprinting method used, and its value should be specified along with the fingerprinting method used for a meaningful overview of the similarity assessment.[107]

2.5.2 Morgan Fingerprints

The Tanimoto coefficient necessitates separate bit representations of the molecules to be assessed for similarity. A popular method for generating computer representations of compounds is the Morgan fingerprint, also known as the ECFP.[108,109] ECFP has been successful in drug discovery applications such as ligand-based virtual screening and target prediction involving machine learning models.[110,111] The principle behind generating ECFP is to capture the local chemical environment of each atom within a compound by assessing the neighboring atom types.[108] An implementation of ECFP can be found in the RDKit Python package, and it accepts user-defined parameters for determining the size of the search space for the local chemical environment and for determining the length of the output vector.[82]

2.5.3 Tanimoto Coefficient Formula

The formula for computing the Tanimoto coefficient between two molecular fingerprints is given by:[112] Tanimoto(A, B) = c / (a + b - c), where (i) a represents the number of features present in A; (ii) b represents the number of features present in B; and (iii) c represents the features common in both A and B. An implementation of the Tanimoto coefficient can be found in the RDKit Python package.[82]

2.6 Lipinski Properties

2.6.1 Overview

Assessing drug-likeness is a critical step in drug discovery because it helps identify compounds likely to become bioavailable drugs before costly clinical trials. This helps with streamlining screening campaigns by narrowing the search space to compounds with promising bioavailability. Moreover, drug-likeness can guide hit-to-lead optimization to address the poor bioavailability profiles of promising hits while maintaining their therapeutic activities. The mode of administration is an important consideration when predicting the drug-likeness of a drug. For oral administration, the Lipinski properties, also known as Lipinski's Rule of 5, are suitable for predicting oral bioavailability.[113] This guideline was created after assessing 4 physicochemical parameters of orally active drugs that reached the phase II clinical trial stage.[113] These parameters are: (i) Molecular weight (MWT); (ii) octanol-water partition coefficient (cLogP); (iii) Hydrogen Bond Donors Count (HBD); (iv) Hydrogen Bond Acceptors Count (HBA).

2.6.2 Rule of Five

Favorable oral bioavailability may be observed if (i) MWT is less than 500 Daltons; (ii) cLogP is less than 5; (iii) HBD is less than 5; (iv) HBA is less than 10.[78] Higher MWT compounds are less likely to be orally active because they display poor intestinal and blood-brain barrier permeability.[114,115] The cLogP is a measure of the amphipathic nature of a small molecule, and the latter should be lipophilic enough to cross lipid membranes while still maintaining solubility in the blood.[116–118] A high HBA or HBD interferes with the molecular membrane permeability as it is more likely to be retained by the aqueous environment due to the collective strengths of multiple hydrogen bonding interactions. [78,119,120]

2.7 Molecular Docking

2.7.1 Overview

Molecular docking is a computational approach used to study the binding of a ligand to a target. The binding of the ligand to its target is predicted by altering the ligand's conformation and orientation within the binding site.[121] The resulting conformation and orientation of the ligand are collectively referred to as its pose.[121] The binding site of the ligand in the target must be defined unless blind docking is to be performed. A search algorithm is then employed to sample multiple poses of the ligand within the identified binding site. Commonly used search algorithms are (i) systematic search; and (ii) stochastic search.[122] Afterwards, a scoring function is used to filter out prominent docked poses.

2.7.2 Systematic Search

Under this approach, the ligand is allowed to explore different conformations under three different algorithms.[122] If the exhaustive algorithm is used, all possible conformations of the ligands are explored under a set of constraints, and the newly generated conformations are used to generate the binding poses.[122] If the fragmentation algorithm is used, the newly generated conformations are fragmented, and the resulting fragments are then used to generate the ligand poses.[122] Lastly, if the ensemble algorithm is used, the newly generated conformations of the ligands are docked against the target without altering the conformation of the ligand in the process.[122]

2.7.3 Stochastic Search

Under this approach, the ligand poses are generated by making random changes to the conformation of the ligand.[122] A probability function is then employed to keep or discard the changes made.[122] 4 different probabilistic criteria may be used:[122] Monte Carlo methods (MC), evolutionary algorithms (EA), Tabu search methods (TSH), and swarm optimization methods (SO). Firstly, MC employs a Boltzmann probability function relying on the energies of the original and proposed ligand poses and the absolute temperature of the system.[122] Secondly, EA employs a genetic algorithm to filter out poses based on an evolutionary

process. The algorithm assigns each pose a score representative of its fitness (dependent on its binding mode with the receptor) to prioritize poses with high fitness for subsequent generation of docking conformations.[123] Next, TSH employs an algorithm for comparing the proposed changes with the pool of previously accepted ligand poses and these changes are discarded if the suggested pose has been observed already among the existing poses, as determined by a root mean square deviation (RMSD) cutoff.[122] Lastly, SO employs an algorithm for learning the trend observed in the current pool of accepted poses for generating novel poses with similar characteristics as the accepted poses.[122]

2.7.4 Scoring Function

The scoring function (SF) is used to quantify the binding affinity of the ligand towards its target. Commonly used scoring functions are (i) Physics-based; (ii) Empirical-based; and (iii) knowledge-based.[122,124] Physics-based SFs include force-field based SFs, solvation models and quantum mechanics methods.[124] The force field-based scoring function utilizes mathematical expressions (force field) to estimate the contribution of non-bonded interactions to the binding of the ligand to its target while neglecting the effects of solvation and entropy.[124] Accuracy is improved by the incorporation of solvation models, which treat water molecules explicitly (computationally expensive) or implicitly.[122,124] Conversely, the empirical-based scoring function employs weighted energy factors to estimate binding affinity.[122] Factors considered include non-bonded interactions, steric effects, and "desolvation" effects, among others.[122,124] The weights assigned to each factor are optimized over a training set of data.[124] Lastly, the knowledge-based scoring relies on a statistical model built on the distribution of interacting atom pairs from protein-ligand complexes to assess docked poses.[124] The statistical model is employed to compute a pairwise interaction potential term for each pair of ligand and receptor atoms, which is then used to calculate the overall energy terms.[124]

2.8 Machine Learning in Structure-Based Drug Design

2.8.1 Overview

Structure-based drug design is a fundamental approach that incorporates methods relying on the 3D structure of the receptor target to guide drug design.[125] It is implemented in drug discovery pipelines to guide the identification of candidate hits showing promising therapeutic potential to their lead optimization for subsequent clinical trials.[125] Understanding protein-ligand interactions is crucial for structure-based drug design as their interpretations are used to assess the binding mode of the ligand and to drive medicinal chemistry optimizations for enhanced binding.[125,126] Traditional computational methods may struggle with the complexity of these interactions, and expert knowledge may be required to interpret results.[126] However, machine learning (ML) models are continuously being optimized to address these gaps. With the growing availability of structural data, ML approaches are trained to identify key patterns from protein-ligand complexes for accurate predictions.[79,127–129] Notably, ML models have been implemented to address the ligand pose classification problem.[130–133]

2.8.2 Ligand Pose Classification with Machine Learning

Ligand pose classification remains a critical challenge in virtual screening campaigns.[134] While docking algorithms can successfully generate near-native poses among their outputs, their inherent scoring functions often fail to prioritize these optimal conformations.[79] Incorrect pose selections can lead to a flawed drug discovery campaign, further amplifying its financial burden. ML models have been developed to address the ligand pose classification problem.[130–133] Such methods have been found to employ supervised learning algorithms, trained on known native and computer-generated non-native poses of crystal ligands.[130–133] Unlike docking algorithms relying on scoring functions built on energy terms, ML classification models are trained on machine representations of the interaction space between the protein and the ligand without any assigned bias.[130–133] As a result, distinct patterns are learnt by the ML models to distinguish between native and non-native poses. Recently, Tran-Nguyen et al. have demonstrated the superiority of ML models such as

support vector machine (SVM), random forest, and neural network models over classical scoring functions in solving the ligand pose classification problem.[79]

2.8.3 Training and Benchmarking ML Models

A dataset must be curated for the task that an ML model will be implemented. It should be split into a training dataset and a testing dataset.[135] The training dataset is used by the ML model for identifying patterns that may be generalized over the entire dataset for the task it will be assigned for.[135] In contrast, the testing dataset is used to assess how well the model performs after training.[135] Commonly used train-test split ratios are 70-30 and 80-20.[135] While a large training dataset is conducive to the learning of the model, a larger testing dataset ensures a fair evaluation of the model's performance.[135]

During testing, specialized metrics are computed to assess the model's performance and these include:[136,137] (i) Recall, (ii) Precision, (iii) Specificity, (iv) Receiver Operating Characteristics Area Under Curve (ROC-AUC), and (v) Precision-Recall Area Under Curve (PR-AUC). Recall measures the ability of the model to classify all true positives (true good poses) from the testing dataset correctly.[137] Precision, however, indicates how often a model returns a false positives.[137] The balance between recall and precision is highlighted by the PR-AUC which measures the bias between the model's false positive rate and its ability to detect true good poses.[137] The lower the PR-AUC score, the more biased the model is towards outputting a high false positive rate in an attempt to label all true good poses correctly. Specificity measures the ability of the model to classify all true negatives from the testing dataset correctly.[137] Lastly, the ROC-AUC score measures how well the model performs, as a binary classifier.[138] A score greater than 0.5 means that model will outperform another classifier operating at random.[138]

2.9 Multivariate Data Analysis

2.9.1 Overview

Multivariate Data Analysis refers to statistical techniques for examining complex datasets containing multiple variables.[139] These methods enable the extraction of meaningful

patterns from high-dimensional data, commonly encountered in drug discovery pipelines.[98,99,140,140–142] Principal Component Analysis[143] (PCA) and clustering[144] are 2 approaches used for multivariate analysis in computational drug discovery.

2.9.2 PCA & Clustering

PCA is a dimensionality reduction technique for transforming high-dimensional data into a lower-dimensional representation while retaining maximum variance to minimize information loss.[143] In drug discovery applications, PCA operates by identifying orthogonal linear combinations of original variables, called principal components, that capture the most significant variations in the dataset.[143] Each principal component (PC) has an associated eigenvalue, which indicates the proportion of the total variance, arising from the dataset, that the PC is accounts for.[143]

Clustering algorithms are employed to group data points based on similarity metrics and the classification process can be unsupervised, eliminating the need for specifying the clustering categories.[144] Hierarchical clustering is an example of an unsupervised clustering algorithm that utilizes a dendrogram for grouping the data points in a series of partitions at different level of similarities.[144,145] The optimal number of clusters is determined by identifying the largest gap between merging clusters along the dendrogram.[145]

Workflow

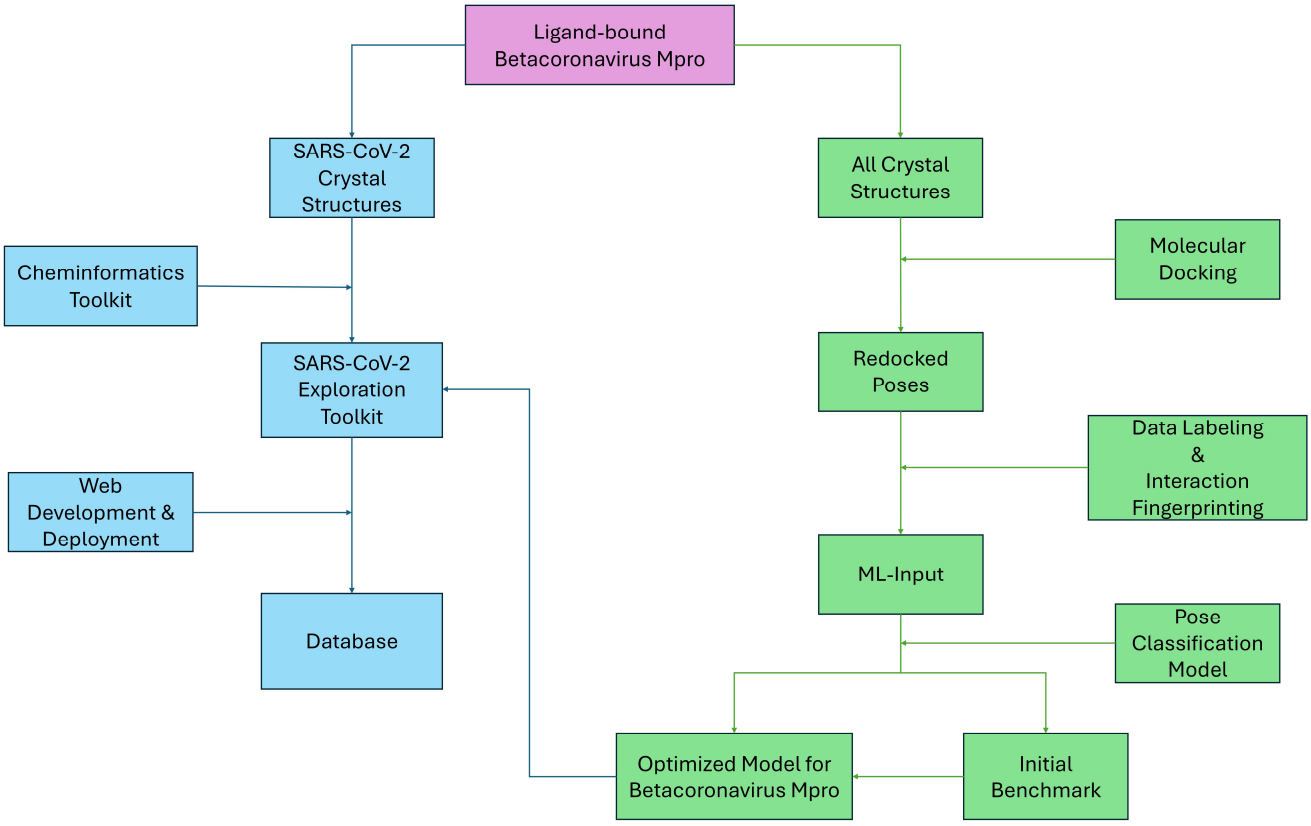


Fig. 2.1: Methodology Workflow Diagram. This diagram illustrates how the techniques introduced in this chapter are utilized to achieve the 3 objectives as listed in Chapter 1. Objective 1, which pertains to the database development, is highlighted in blue. Objectives 2 & 3, which pertain to benchmarking a pose classification model and integrating an optimized model for *Betacoronavirus* Mpro into the database, are highlighted in green.

Chapter 3: A New Database to Facilitate Comparative SARS-CoV-2 Mpro-Ligand Exploration

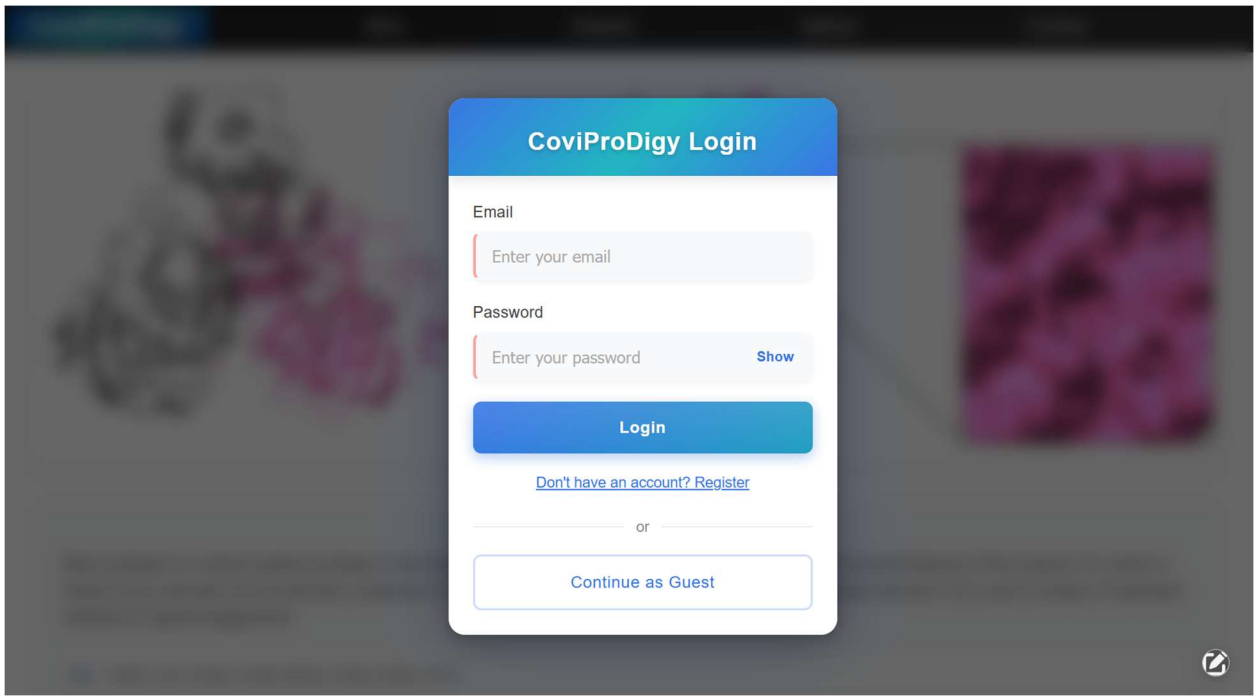


Image obtained from the login page of CoviProDigy

3.1 Introduction

The global health crisis initiated by SARS-CoV-2, a positive-sense RNA virus, highlighted the need for effective antiviral treatments. A central process in the viral life cycle is the proteolytic cleavage of 2 large polyproteins, pp1a and pp1ab, which are translated from the viral RNA upon entry into the host cell.[22] This cleavage, mediated by viral proteases, releases functional NSPs that are essential for viral replication.[22]

Mpro is a cysteine protease that plays a predominant role in this process, executing cleavage at 11 distinct sites along the polyproteins.[42] Consequently, a thorough understanding of how Mpro recognizes and binds its substrates became a critical focus of research during the pandemic to guide the design of potent antiviral drugs.[43,49,56,146,147] Key insights into this mechanism were provided by the crystal structures of catalytically inactive Mpro mutants in complex with their native peptide substrates.[42] These structures revealed a 3-domain architecture with the substrate-binding pocket situated at a cleft between domains I and II.[42] This pocket is further divided into several subsites, primarily S1, S1', S2, and S4 which accommodate the corresponding P1, P2, P4, and P1' residues of the substrate, respectively (Fig. 3.1).[42]

Analysis of the 11 Mpro cleavage sites revealed a strong conservation of specific amino acids at key positions.[42] A glutamine residue is highly conserved at the P1 position.[42] The structural basis for this high specificity is evident in the crystal structure of Mpro with its natural substrates. The S1 subsite forms a precise hydrogen bond network with the P1 glutamine, involving the side chain of His163, the backbone of Phe140, and a conserved water molecule.[42] This intricate network establishes a stringent requirement for glutamine at P1. The P2 position shows a strong preference for leucine, which is present in 9 of the 11 cleavage sites.[42] This is due to the deep, hydrophobic nature of the S2 subsite, which readily accommodates large, nonpolar side chains such as leucine, valine, and phenylalanine.[42] In contrast, the S4 subsite is a smaller, more constrained pocket that favors residues with smaller side chains, such as alanine, valine, proline or threonine, through mainly hydrophobic interactions.[42] Finally, the shallow S1' subsite can only accommodate small

residues, with serine being the most common at the P1' position.[42] While other subpockets exist, the S1, S1', S2 and S4 sites are the primary determinants of substrate specificity and binding.[7,27,35,148–152] This detailed characterization of Mpro's subpockets has served as a blueprint for designing inhibitors since the pandemic began. Extensive screening campaigns, guided by these subsite requirements, have led to the discovery of numerous drug candidates.[43,49,56,72,146,147] These inhibitors can be broadly classified as either peptidomimetic or non-peptidomimetic.[153]

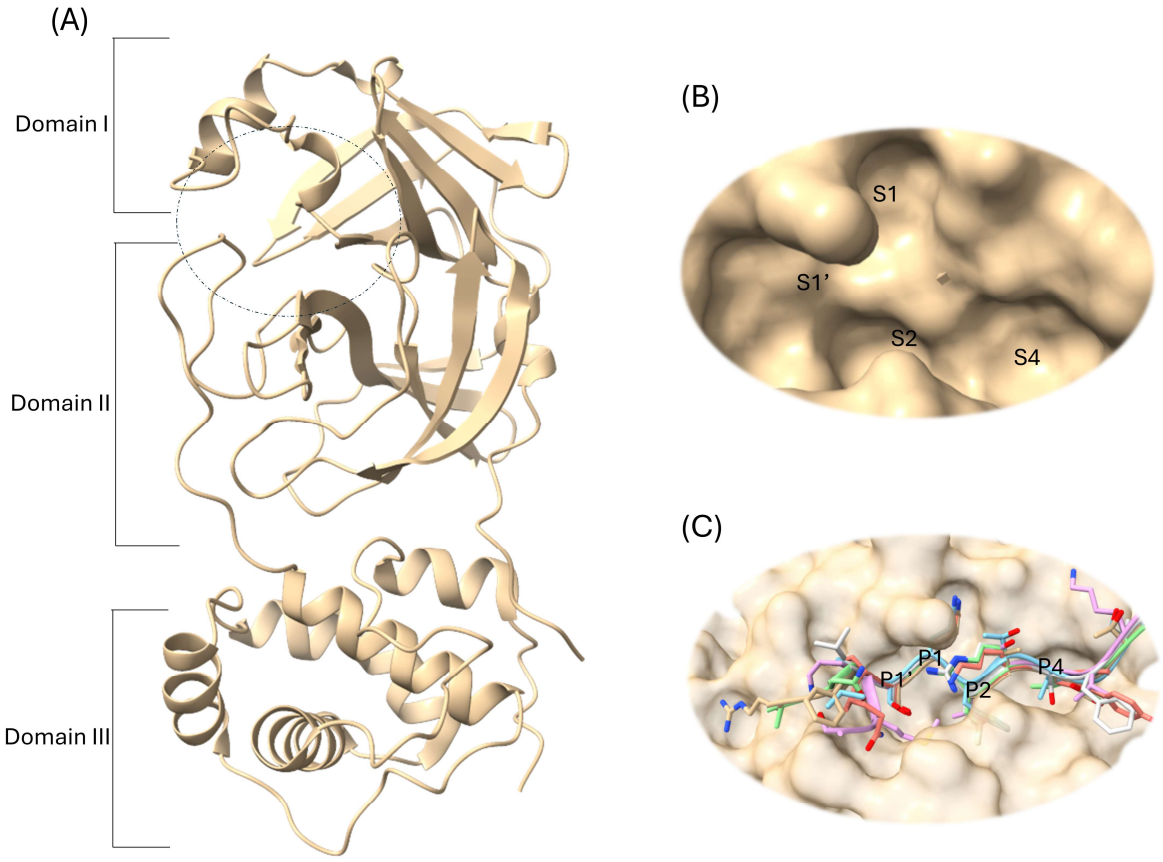


Fig. 3.1: Active Site Architecture of SARS-CoV-2 Mpro. (A) Domain architecture of Mpro protomer. Domains I-III are labeled for the cartoon representation (colored in tan) of SARS-CoV-2 Mpro (PDB: 7DVP). The active site, circled in black dotted lines, is located at a cleft between domains I-II. (**B)** Expanded view of SARS-CoV-2 Mpro active site. The active site is shown in surface representation with the location of the 4 major subsites indicated: S1, S1', S2, S4. **(C)** SARS-CoV-2 Mpro in complex with cleavage substrates. 6 cleavage substrates

spanning 6 different Mpro cut-sites along the polyproteins are shown in complex with SARS-CoV-2 Mpro. Their moieties are labeled to the subsites (S1, S1', S2, S4) they are interacting with. The protein backbone is rendered in cartoon representation and colored in tan. The cleavage substrates are shown in licorice representation with their side chain atoms in stick representation. The cut-sites spanned by the substrates are[42]: nsp4|5 (PDB: 7DVP, tan), nsp5|6 (PDB: 7DVW blue), nsp6|7 (PDB: 7DVX, purple), nsp9|10 (PDB: 7DVY, green), nsp14|15 (PDB: 7DW6, red), nsp15|16 (PDB: 7DW0, white). All images were created using UCSF ChimeraX.[47]

Peptidomimetic inhibitors are based on the scaffold of natural Mpro substrates with careful optimizations at the P1, P2, P4, and P1' positions for optimal binding with the corresponding subpockets.[153] Nirmatrelvir and Ibuzatrelvir are notable covalent peptidomimetic inhibitors. Their electrophilic warhead (nitrile group) at P1' allows for reversible covalent interaction with the catalytic Cys145.[154] Nirmatrelvir is an FDA approved inhibitor of SARS-CoV-2 Mpro and key optimizations leading to its discovery included: (i) the replacement of its original P1' ketone moiety by a nitrile warhead at P1' for increased oral bioavailability and covalent binding with Cys145, (ii) a cyclic leucine mimetic at P2 to complement the hydrophobic S2 pocket, and (iii) trifluoroacetamide at P4 to fit the constrained S4 pocket(Fig. 9).[7,27] This resulted in an improved interaction profile of Nirmatrelvir with Mpro involving hydrogen bonding interactions with key residues from each subpocket (Fig. 3.2A). Notably, Glu166 is involved in a precise hydrogen bonding network with different polar moieties of the ligand. Despite the high potency of Nirmatrelvir (K_i = 3.11nM[7]), it displays a poor pharmacokinetic profile, requiring the co-administration of a cytochrome inhibitor, ritonavir.[154] Hence, Pfizer developed Ibuzatrelvir, a Nirmatrelvir-derived second-generation Mpro inhibitor.[154] The need for a pharmacokinetic booster was overcome with Ibuzatrelvir by 2 key substitutions: (i) P2 leucine analogue with trifluroromethyl proline residue, and (ii) P4 trifluoroacetamide to methyl carbamate.[154] Ibuzatrelvir shares the same binding mode (Fig. 3.2B) and inhibitory potency as its predecessor (K_i = 2.48nM), without any compromise on its stability.[155]

Non-peptidomimetic inhibitors, on the other hand, lack the peptide-like backbone of natural substrates.[153] These compounds are often identified through high-throughput screening of

existing molecule libraries and are subsequently optimized to fit the Mpro subsites.[153] They can be further categorized as non-covalent, such as X77, or covalent, such as 16a (potent X77-mimetics).[153] X77 has an IC50 value of 4.1 μM[55], binding all subsites except S4.[38,56] It forms hydrogen bonds with Glu166, His163, Gly143, and engages in a distinctive stacking interaction with His141, enhancing its stability within the active site (Fig. 3.2C). To enhance its potency, the imidazole ring was substituted with a chloromethyl group giving the compound 16a an IC50 value of 0.42 μM.[55] This 10-fold increase in potency was due to the chlorine being a suitable leaving group for covalent bond formation of 16a with Cys145. Despite the imidazole replacement, 16a still maintains the same interaction profile as X77 (Fig. 3.2C, D)

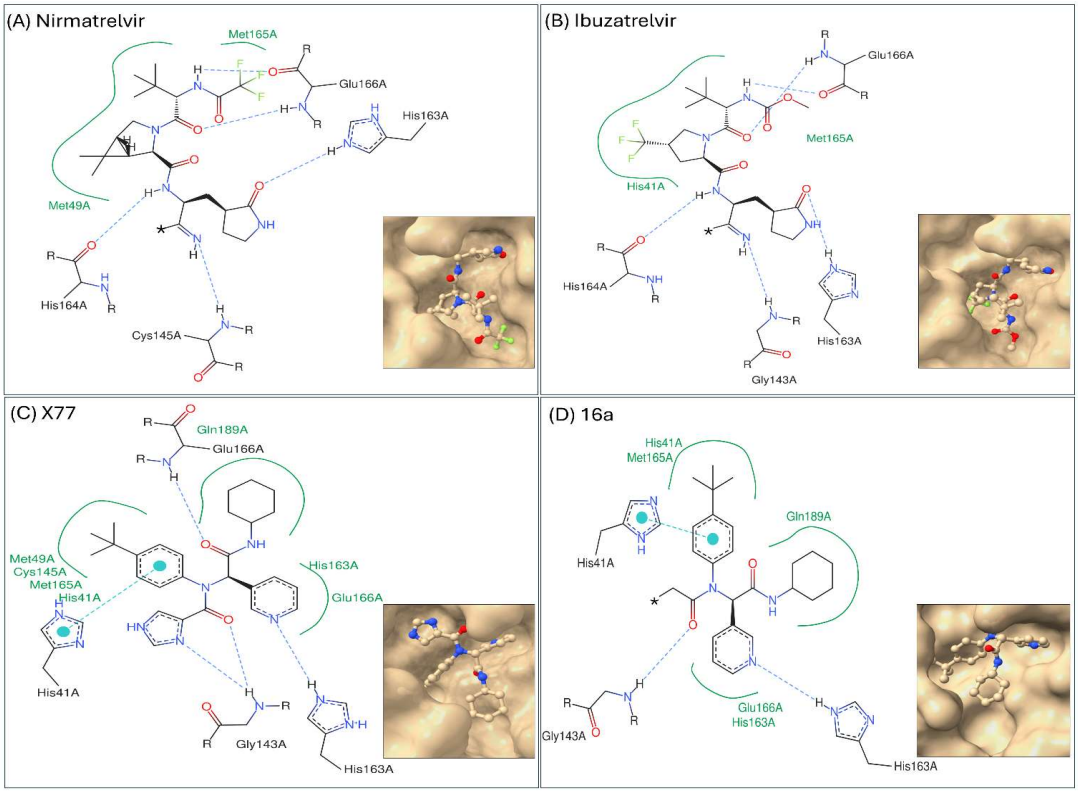


Fig. 3.2: Binding Modes of Selected Peptidomimetic and Non-Peptidomimetic SARS-CoV-2 Mpro Inhibitors. The 2D interaction diagrams and 3D subpocket fit are shown for (A) Nirmatrelvir (PDB: 7RFS), (B) Ibuzatrelvir (PDB: 9BPF), (C) X77 (PDB: 6W63), and (D) 16a (PDB: 7MLF). Non-bonded interactions are represented in the 2D interaction diagram: (i) hydrogen bonds in blue dotted lines linking the hydrogen bond donors and acceptors, (ii)

hydrophobic contacts in green contour lines surrounding the ligand and the residues color coded in green, and (iii) aromatic stacking interaction in turquoise dotted lines linking the centroids (also rendered in turquoise) of the aromatic rings involved. For covalent ligands, an asterisk (*) is used to denote the covalent bond linking Cys145, as in the case for (B) Ibuzatrelvir and (D) 16a. In 3D representation, each ligand is styled as ball and stick with the atoms colored according to their atom types while the active site of SARS-CoV-2 Mpro is rendered in surface representation and colored in beige. PoseEdit[58] was used for generating the 2D interaction diagrams via the ProteinsPlus server[59], while UCSF ChimeraX[47] was used for generating the 3D depictions.

Moreover, the high degree of conservation of Mpro across the *Betacoronavirus* family makes it an ideal target for the development of pan-coronavirus inhibitors (Fig. 1.5). A series of studies have already screened thousands of compounds, establishing extensive structure-activity relationships that provide a solid foundation for rational inhibitor design.[72,147,151,153,156] However, a major bottleneck is the difficulty in accessing and comparatively analyzing these rich structural insights using current database tools.

The PDB is the primary archive for 3D structure of biomolecules but only allows for the retrieval and visualization of entries on an individual basis, lacking tools for comparative analysis (Appendix: Fig. A1).[60] PDBsum provides useful overviews of PDB entries, including secondary structure assessment and ligand interaction profiling, but also operates on a case-by-case basis (Appendix: Fig. A2).[157] While Fragalysis offers a step forward by allowing the visualization of multiple ligands within a target's binding site, it does not provide quantitative or textual analysis of the interactions, leaving the comparative assessment to the user (Appendix: Fig.A3).[158] None of these tools offer a toolkit for assisting with the medicinal chemistry optimization of promising SARS-CoV-2 inhibitors which may include an assessment of the subpocket occupancies to guide hit-lead optimizations, the availability of a dedicated ligand dataset which may be filtered using Lipinski's metrics or clustered according to a similarity index. To address these limitations, this thesis introduces CoviProDigy (Covid-19 Mpro Digging), a novel toolkit specifically designed for the comprehensive analysis of ligands targeting SARS-CoV-2 Mpro. This platform facilitates

comparative analysis of multiple Mpro complexes, enabling the identification of common and unique interaction patterns. This functionality is intended to support medicinal chemistry efforts, such as scaffold hopping and lead optimization, by providing deeper molecular-level insights. This toolkit will be a valuable resource in the ongoing effort to develop novel antivirals against SARS-CoV-2 Mpro.

3.2 Methodology

3.2.1 CoviProDigy Design Strategy

CoviProDigy is built on 2-layer architecture, comprising a frontend and a backend. The frontend was developed using HTML, CSS, and TypeScript, with the React framework employed to create a component-based and maintainable structure. All data visualizations, including charts and 2D interaction diagrams, are rendered as SVG. The D3.js library[90] is utilized to generate the SVG code for the charts, while the 2D interaction diagrams are produced by an in-house script. The interactive 3D molecular viewer is implemented using the 3DMolJS library, with custom patches applied to the source code to introduce new, Mpro-specific features.

The backend is powered by Python, using the Flask framework for web development. A suite of cheminformatics packages including RDKit, Open Babel, and the ODDT, are used for processing PDB files.[82–84] PDB files are either uploaded by the user or fetched directly from the PDB[60] using the Python requests library. Both the frontend and backend components, along with their dependencies, were containerized using Docker and deployed via Heroku's container registry.

3.2.2 Data Input and Validation

CoviProDigy is designed to process only ligand-bound structures of SARS-CoV-2 Mpro. Upon input, the sequence of the provided structure is aligned against the wild-type SARS-CoV-2 Mpro sequence using the Needleman-Wunsch algorithm[159], implemented with the NumPy[160] library. A structure is accepted for further processing if the sequence similarity score exceeds 90%. Following this, the tool verifies that a ligand is present in the active site by

identifying any ligands within a 5 Å radius of the active site residues, a task performed using Open Babel.

3.2.3 Data Pre-processing

Once a structure is validated, CoviProDigy performs an initial pre-processing step. This involves identifying the ligand's structural formula, detecting any alternative conformations (which are treated as separate entities), calculating its Lipinski properties[78], and determining if it forms a covalent bond with the catalytic Cys145. Open Babel[83] is used to generate the SVG for the ligand's 2D structure and to detect covalent linkages, while RDKit[82] is used to predict the Lipinski properties.

3.2.4 Data Mining and Interaction Profiling

CoviProDigy calculates the ligand subpocket occupancy by measuring the distance between the geometric center of each subpocket (S1, S1', S2, S4) and the closest ligand atom. The geometric center of a subpocket is defined as the average 3D coordinates of all backbone atoms of its constituting residues. Interaction profiling is performed using ODDT[84] where non-bonded interactions, such as hydrogen bonds, hydrophobic contacts, and aromatic stacking interactions, between the ligand and Mpro are identified.[84]

3.2.5 Data Clustering

CoviProDigy employs 3 distinct methods to cluster the ligands in its database. The first approach uses a series of histograms to group ligands based on each Lipinski metric.[78] Bins within these histograms are further stratified to show the proportion of ligands occupying 1, 2, 3, or 4 subpockets. The second method uses PCA[143] to cluster ligands based on their collective Lipinski metrics. The resulting scatter plot visualizes the ligands in a reduced dimensional space, color-coded by their compliance with Lipinski's Rule of 5. This analysis is implemented using the Scikit-learn library.[161] The final approach clusters ligands based on their structural similarity. Morgan fingerprints[108] are generated for each ligand using RDKit[82] to create a machine-readable representation. A Tanimoto[112] similarity matrix is then

calculated for all ligand pairs , and this matrix is used to perform hierarchical clustering[162] with the SciPy library[163], resulting in a dendrogram that groups structurally similar ligands.

3.2.6 Structural Alignment

To overlay multiple Mpro structures, a two-step structural alignment protocol is used. First, the amino sequences of the Mpro protomer for all structures are globally aligned against a selected Mpro protomer using the BioPython[164] pairwise2 module. Based on this sequence alignment, the corresponding C-alpha atoms' coordinates are extracted from the paired residues. These are used as input for applying the Kabsch algorithm[165] to superpose the structures, yielding an optimal rotation matrix that minimizes the root-mean-square deviation (RMSD) between the paired C-alpha atoms. All non-reference structures are aligned onto the reference structure by first translating them onto the reference structure followed by a rotation operation to overlay the structures.

3.2.7 Data Curation and Management

The CoviProDigy database contains pre-calculated data for all orthosteric ligands from SARS-CoV-2 Mpro crystal structures available in the PDB. This includes Lipinski properties, subpocket occupancies, and SMILES codes, which are stored in a JSON file for rapid access. User authentication is handled by Google's Firebase platform, which is also used to provide cloud storage for saving user sessions and notepad content for later retrieval.

3.3 Results

The extensive focus on SARS-CoV-2 Mpro since the start of the COVID-19 pandemic has led to an unprecedented number of PDB depositions for a single target. After a steady deposition of approximately 200 annually until 2022, a prominent surge occurred in 2023 with nearly 600 new crystal structures, largely driven by the COVID Moonshot consortium's release of 466 structures (Fig. 3.3).[147] Extracting subpocket-level molecular insights from this vast dataset would typically require manual analysis with specialized scripts, as no dedicated database exists for this purpose. CoviProDigy was developed to fill this gap.

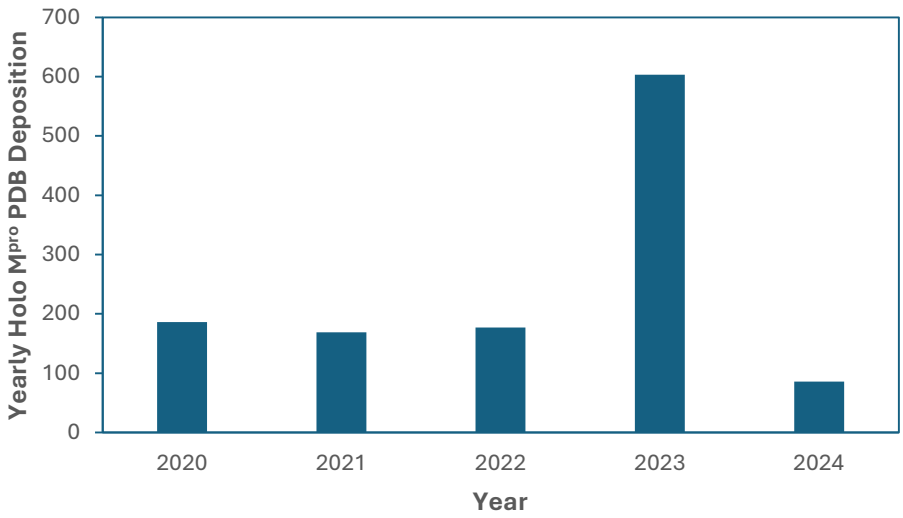


Fig. 3.3: Annual Deposition of Ligand-Bound SARS-CoV-2 Mpro in the PDB. The number of PDB structures of ligands in complex with the orthosteric site of SARS-CoV-2 Mpro is given on a yearly basis from the year 2020 to 2024. The holo SARS-CoV-2 Mpro crystal structures were retrieved and dated using the API access of PDB.[60] UCSF ChimeraX[47] was used to confirm the presence of ligands in the orthosteric site of Mpro. The number of depositions was stable from the years 2020 to 2022 at a yearly deposition of approximately 200 holo crystal structures. A spike in the number of deposition was observed in 2023 due to the COVID Moonshot consortium's release of 466 structures.[147] In 2024, the number of yearly depositions dropped to around 100.

3.3.1 Optimized Algorithm for Characterizing Subpocket Occupancy

To provide a quantitative and automated assessment of subpocket engagement, CoviProDigy employs a custom algorithm to characterize ligand occupancy. This determination is based on a defined distance cut-off between a subsite reference point and the nearest ligand atom. To ensure accuracy, both the reference point and the distance cut-off for each subpocket were empirically optimized. This was achieved by graphically correlating the number of protein-ligand interactions within a given subsite against the distance to the closest ligand atom (excluding hydrogen atoms), a process performed across the entire dataset of SARS-CoV-2 Mpro crystal structures (Appendix: Fig. A8). The final distance cut-off was selected at the point where the interaction count became minimal or

was not primarily confined to residues at the subsite's entrance. Through this process, the optimal reference point was confirmed to be the geometric center of the backbone atoms from a curated set of lining residues for each subpocket. The finalized subpocket definitions and their corresponding distance cut-offs are detailed in Table 3.1.

Table 3.1: Optimized Subpocket Definitions and Distance Cutoff for Assessing Subpocket Occupancies. The reference subpocket residues and distance cutoff used to determine whether a ligand occupies a given subpocket are given for the subpockets S1, S1', S2, and S4.

Subpocket	Subpocket Residues	Distance cutoff (Å)
S1	F140, L141, N142, H163, M165, H172	3
S1'	T25, T26, L27, H41, V42, N142, S144, C145	3.5
S2	M49, Y54, M165, E166, D187	2.3
S4	L167, P168, F185, R188, Q189, T190, A191, Q192	3

Application of this algorithm to the entire ligand dataset yielded key insights into subpocket preferences. The distribution plots for the S1 and S2 subsites are negatively skewed, with over 80% and 75% of all crystal ligands falling within their respective occupancy cut-offs of 3.0 Å and 2.3 Å (Fig. 3.4A). This indicates a strong preference for ligands to engage with these 2 subsites. In contrast, the S4 subsite is occupied by less than 30% of the ligand population (Fig. 3.4A), reflecting its constrained nature which accommodates only smaller functional groups from ligands having a scaffold of sufficient length to span into the subsite.[42]

The S1' subsite was also occupied by less than 30% of the ligands, based on its 3.5 Å distance cut-off (Fig. 3.4A). Further analysis revealed that the average distances from the S1' geometric center to the catalytic residues His41 (Nε atom) and Cys145 (Sγ atom) were approximately 5.2 ±0.7 Å and 5.0 ± 0.4 Å (Appendix: Table A3), respectively. This spatial

separation suggests that while a ligand may interact with the catalytic dyad, it must span additional surface area to occupy the S1' subpocket. For each subsite, two ligand examples, one classified as an occupant and the other classified as a non-occupant, are illustrated in Fig. 3.4B, highlighting the need for a sufficiently long ligand scaffold for subsite fitting.

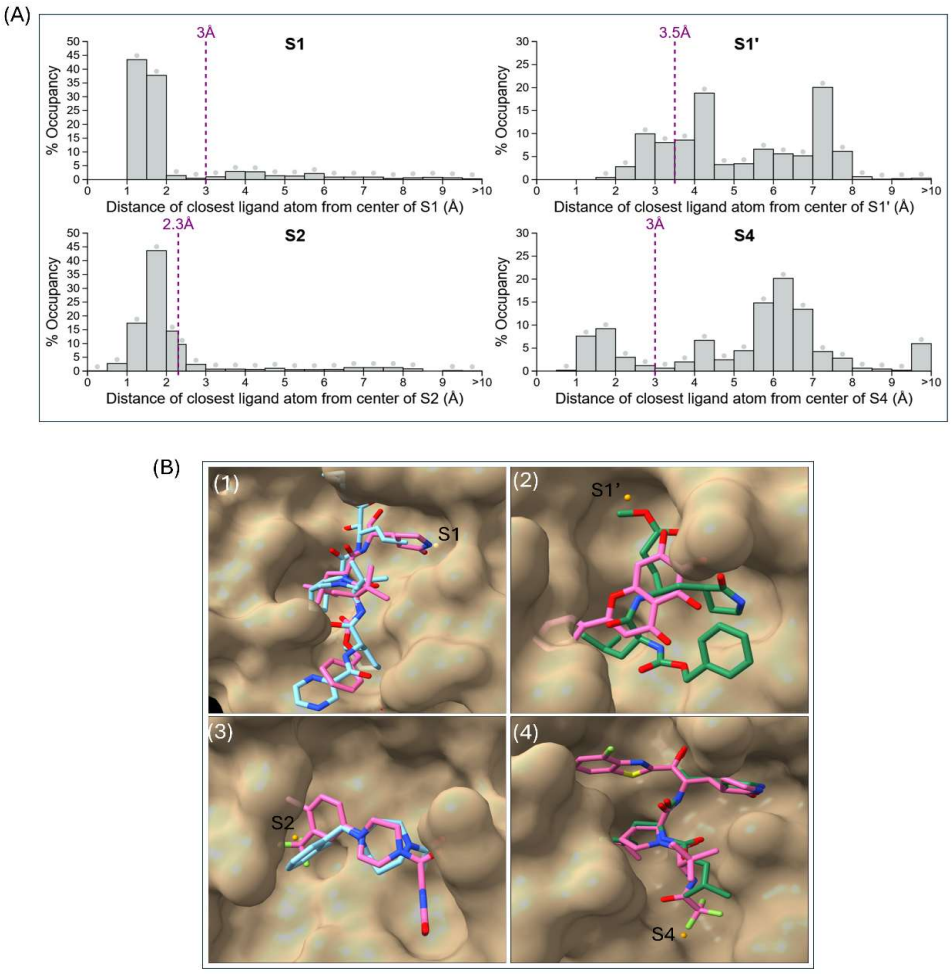


Fig. 3.4: Characterizing Subpocket Occupancies of SARS-CoV-2 Mpro Ligands. (A) Histograms showing the distribution of the closest ligand atom distance to the geometric center for subpockets S1, S1', S2, and S4. The distance threshold for occupancy is marked in purple along with a purple dotted line: 3 Å for S1, 3.5 Å for S1', 2.3 Å for S2, and 3 Å for S4. The subpocket occupancies for all SARS-CoV-2 crystal ligands were determined using the optimized subpocket definition in Table 3.1. UCSF ChimeraX[47] was used for the distance measurements and the histograms were generated using D3.js.[90] **(B)** Example

classifications of subpocket occupants and non-occupants. For each quadrant, one ligand classified as an occupant and one as a non-occupant are shown for a specific subpocket: **(B1)** S1: Occupant is 7XT (PDB: 7RVU, pink); non-occupant is SV6 (PDB: 7K6D, blue); **(B2)** S2**:** Occupant is 5Z3 (PDB: 7RME, pink); non-occupant is T4Y (PDB: 5REY, blue); **(B3)** S1'**:** Occupant is VQR (PDB: 8OKB, green); non-occupant is 3WL (PDB: 6M2N, blue); and **(B4)** S4: Occupant is T2L (PDB: 9ARS, green); non-occupant is WEL (PDB: 8E64, blue). In all panels of (B), ligands are shown in stick representation, subpocket centers are depicted as orange spheres, and the protein backbone is colored tan and styled in surface representation. Structures were aligned using the matchmaker command in UCSF ChimeraX[47], with the occupant's PDB structure serving as the reference for each alignment. Image generation was done using the same software.

3.3.2. Crystal Ligand Dataset

The Dataset tab in CoviProDigy provides a comprehensive overview of all orthosteric ligands co-crystallized with SARS-CoV-2 Mpro. It consists of summary plots that give a quick overview of all the ligands based on key features. Firstly, physicochemical properties are used in conjunction with subpocket occupancy to illustrate the relationship between them. Interactive histograms are implemented to showcase the distribution of key Lipinski metrics[78] (MWT, cLogP, HBD, and HBA) stratified by the number of subpockets each ligand occupies. It can immediately be observed that over 80% of the ligand population has a molecular weight between 200 and 600 g/mol (Fig. 3.5A). Moreover, a clear trend is observed where ligands with higher molecular weight tend to occupy more subpockets, as larger molecules can physically span a larger surface area to fit into the subpockets of the binding site. For instance, in the 200-400 g/mol range, ligands primarily occupy 1 or 2 subpockets, while in the 400-600 g/mol range, there is a significant increase in the proportion of ligands occupying 3 or 4 subpockets (Fig. 3.5A).

Furthermore, around 80% of the crystal ligands possess a HBD between 0 and 4 (Fig. 3.5B). While ligands falling into this category occupy primarily 1 or 2 subpockets, ligands having a higher HBD count occupy 2 or 3 subpockets mostly. Interestingly, no distinction in ligand

population is observed for complete subpocket engagement across all HBD counts. The HBA count, however, shows greater variability in subpocket occupancy. Ligands having an HBA count in the range of 0 to 4 occupied at most 2 subpockets while ligands with higher HBA count shows greater subpocket engagement. Moreover, the fraction of ligands engaging with all 4 subpockets is higher for the HBA count in the range of 4-8, which is the intermediate range for this property. This suggests that a higher HBA count does not correlate with a higher subpocket engagement, and this is plausible as some Mpro subsites are hydrophobic,[42] rejecting polar hydrogen bond acceptors.

Lastly, around 60% of all ligands have a cLogP value between -0.8 and 2.8 (Fig. 3.5D), with most ligands occupying 2 subpockets. The same trend is observed for cLogP values between 2.8 to 6.4 but with a significant drop in the number of ligands occupying 1 subpocket. The increase in lipophilicity may have possibly been coupled with bigger ligand scaffolds, allowing most of the ligands to span into more subsites. However, this observation does not carry over for the fraction of ligands occupying more than 2 subpockets, suggesting that increasing lipophilicity alone is not sufficient to drive subpocket engagement.

The Dataset tab also features a PCA Compliance Clustering plot, which visualizes all ligands in a PC space defined by their Lipinski metrics (Fig. 3.5E-H) whereby the first 2 PCs captured 89.3% of the total variance. In this plot, ligands are styled based on their compliance with Lipinski's Rule of 5, revealing 2 distinct clusters of compliant and non-compliant compounds. Generally, ligands that occupy 1 or 2 subpockets tend to be compliant with Lipinski's rules (Fig. 3.5E, F), whereas those occupying 3 or 4 subpockets were more often non-compliant (Fig. 3.5G, H). This suggests a potential trade-off between increasing molecular size to achieve higher subpocket occupancy and maintaining drug-like properties.

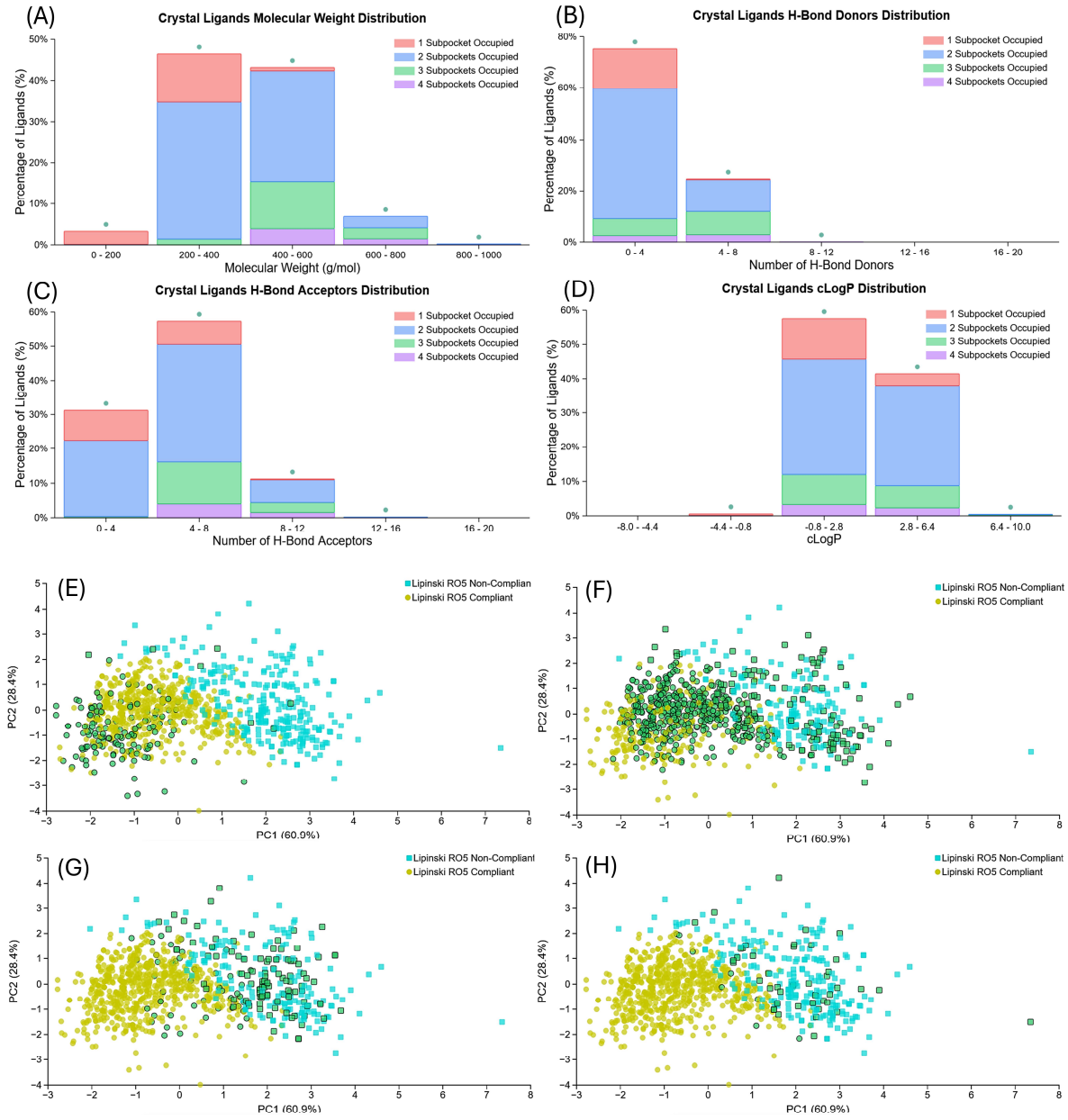


Fig. 3.5: Physicochemical Properties Characterization of SARS-CoV-2 Mpro Crystal Ligands. Distributions of Lipinski's metrics for all crystallographically determined SARS-CoV-2 Mpro ligands, including **(A)** Molecular Weight, **(B)** Hydrogen Bond Donor Count, **(C)** Hydrogen Bond Acceptor Count, and **(D)** Octanol-Water Partition Coefficient. Each histogram is represented as stacked bars, where the colors in each bin correspond to the number of subpockets occupied by the ligands (Red: 1 subpocket occupancy; blue: 2

subpocket occupancy; green: 3 subpocket occupancy; purple: 4 subpocket occupancy). The subpocket occupancies for all SARS-CoV-2 crystal ligands were determined using the optimized subpocket definition in Table 3.1. **(E)-(H):** PCA Decomposition of Lipinski's Rule of 5 Metrics for SARS-CoV-2 Mpro Ligands Stratified by Subpocket Occupancy. Crystallographically determined SARS-CoV-2 Mpro ligands are decomposed via PCA on their Lipinski's metrics along PC1 (capturing 60.9% of variance) and PC2 (capturing 28.4%). The ligands are depicted according to their compliance with Lipinski's Rule of 5: a golden sphere if compliant and a turquoise-colored square if not. The number of subpockets occupied by the ligands is further highlighted in green in **(A)** 1 Subpocket Occupancy, **(B)** 2 Subpocket Occupancy **(C)** 3 subpocket Occupancy, and **(D)** 4 Subpocket Occupancy, whereby the subpocket occupancies were determined using the optimized algorithm discussed previously. The Lipinski's properties were calculated using RDKit[82] if they could not be retrieved from Pubchem[166], and PCA[143] analysis was implemented using the Scikit-learn library.[161]

With a large ligand dataset at hand, it may be difficult to find structurally related ligands from the available databases, such as PDB[60] or Fragalysis.[158] For example, the ligand O6K has 6 published crystal structures in the PDB, from 3 different research groups.[35,167,168] Manual verification of each PDB entry revealed additional related structures that were crystallized within the same studies. However, no indication is given about how closely related these structures are and whether there exist other related structures originating from different studies not involving the ligand O6K. To address this gap, CoviProDigy provides a Ligand Tanimoto Similarity Tree that enables the easy identification of structurally related ligands deposited in the PDB at various similarity thresholds (Fig. 3.6). This tree is a dendrogram generated by hierarchical clustering[162] based on Morgan fingerprints based Tanimoto scores[108,112] and allows for the exploration of related ligand families. For example, the tool can highlight the path from the main cluster to a specific ligand, revealing its parent, scaffold, and child clusters. With the example of the ligand O6K, a child cluster consisting of 5 entries (including O6K) with a local similarity average of 74.4% was picked from the similarity tree (Fig. 3.6). Closer inspection revealed that this cluster consisted of a geometric

isomer of O6K (V9R), 2 structural isomers (A1ICP & A1H9D) lacking the pyridine and aromatic rings from O6K, and another small molecule (MXU) having virtually the same scaffold as O6K minus the phenyl ring (Fig. 3.6).

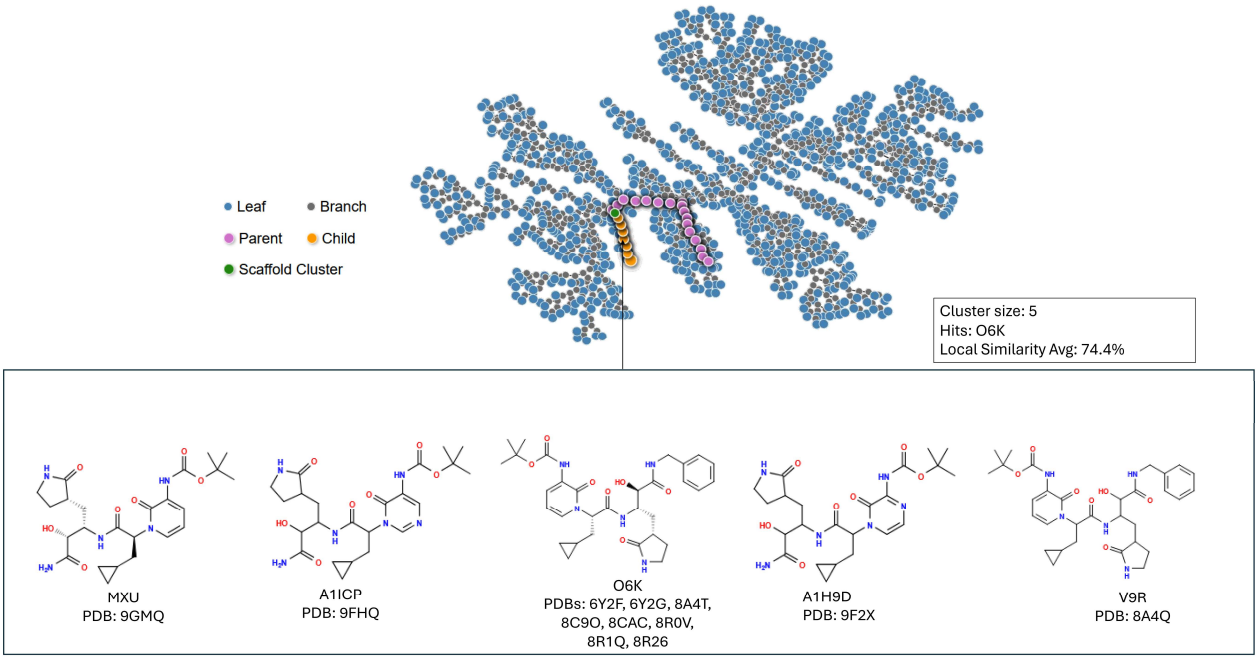


Fig. 3.6: Identifying Structurally Related Ligands Across SARS-CoV-2 Crystal Ligands Dataset via Morgan Fingerprint-based Tanimoto Similarity Clustering. A dendrogram (Top) created by hierarchical clustering of the Morgan fingerprint-based Tanimoto scores[108,112] of all possible SARS-CoV-2 Mpro ligands is depicted in a series of nodes: (i) Leaf node in blue, (ii) Branch node in grey. The ancestral path leading to the leaf containing the ligand O6K is highlighted: (i) Parent node in pink, (ii) Scaffold node in green, and (iii) child node in orange. The 2D structures of the ligands identified in the 4th cluster counting from leaf node of O6K is given below, along with the PDB codes of their crystal structures. This cluster contains 5 ligands in total with a local similarity average of 74.4%. RDKit[82] was used to compute the Morgan Fingerprints[108] and the Tanimoto scores[112] while SciPy[163] was used for the hierarchical clustering[162] of the resulting similarity matrix.

All the plots incorporated in the Dataset page are interactive and “queryable” to allow for guided ligand exploration. For instance, the leaf nodes of the similarity reveal their respective

ligand compositions when hovered on. Moreover, when they are clicked, the ancestral paths revealing the partitioning events leading their formation are highlighted (Fig. 3.7A). Along these paths, ligands at defined similarity thresholds with respect to the leaf nodes are revealed, which can be probed further. Additionally, A ligand locator widget is implemented to allow users to search for a specific ligand name/PDB and highlight its corresponding data points across all 3 summary plots, providing a unified view of its properties and relationships (Fig. 3.7A-C).

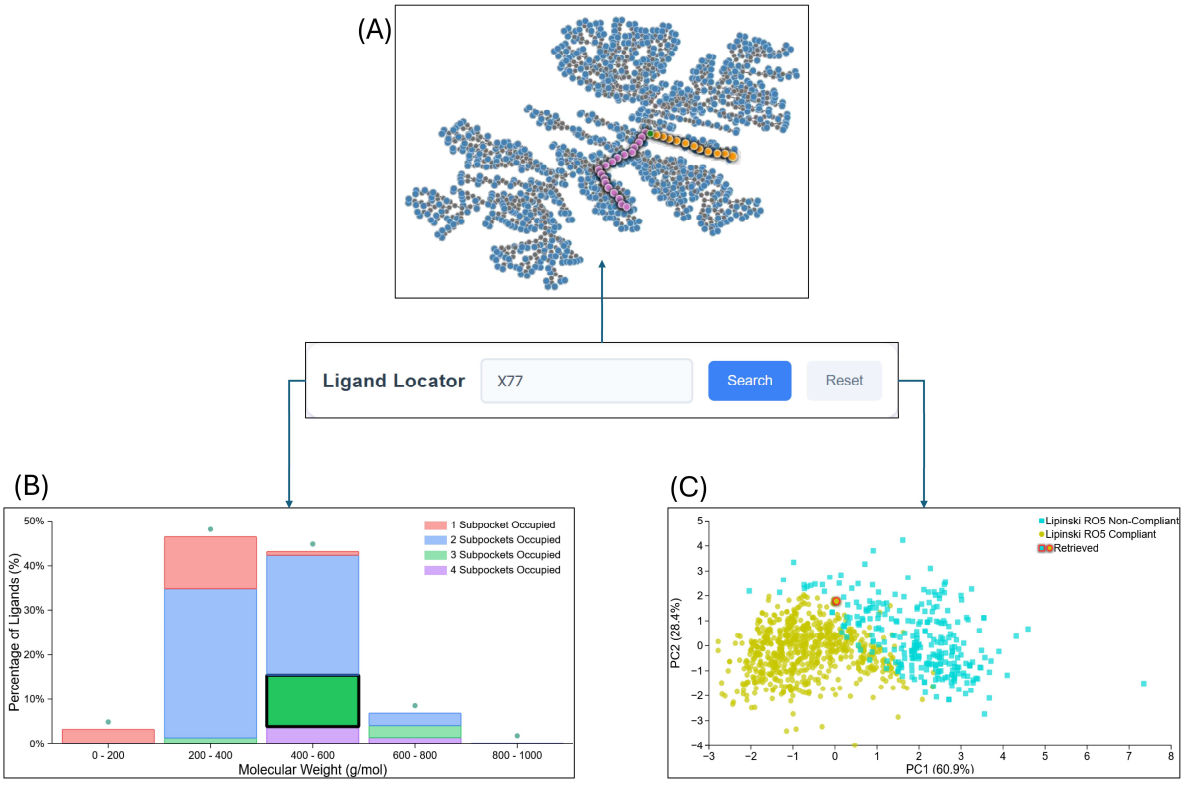


Fig. 3.7: Identifying Ligands of Interest across Summary Plots. The ligand locator was used to highlight the ligand X77 across all the summary plots in the Dataset tab. **(A)** Tanimoto Similarity tree with the ancestral path leading to X77 is highlighted: (i) Parent node in pink, (ii) Scaffold node in green, and (iii) child node in orange. Other leaf nodes are shown in blue while other branches are shown in grey. RDKit[82] was used to compute the Morgan Fingerprints[108] and the Tanimoto scores[112] while SciPy[163] was used for the hierarchical clustering[162] of the resulting similarity matrix. **(B)** Stratified histogram for the molecular weight distribution of all crystal ligand and X77 is highlighted in bold in the molecular weight

range of 400-600 (g/mol) at the green segment representing ligands occupying 3 subpockets. Other subpocket occupancies are indicated by red (1), blue (2), and purple (4). The molecular weight was calculated using RDKit[82] if it could not be retrieved from Pubchem.[166]**(C)** PCA decomposition plot based on the Lipinski's metrics of the crystal ligands with the datapoint corresponding to X77 highlighted with a red border. The subpocket occupancies for all SARS-CoV-2 crystal ligands were determined using the optimized subpocket definition in Table 3.1. The Lipinski's properties were calculated using RDKit[82] if they could not be retrieved from Pubchem[166], and PCA[143] analysis was implemented using the Scikit-learn library.[161] PC1 accounts for 60.9% of the total variance observed in the dataset while PC2 accounts for 28.4% only , collectively amounting to 89.3%. Plots **(A)-(C)** were generated using D3.js.[90]

3.3.3 Data Exploration

The Xplorer tab provides tools for filtering and sorting the 980 curated SARS-CoV-2 Mpro ligands based on their Lipinski metrics and subpocket occupancy (Fig. 3.8). Out of those ligands, the majority (~63%) occupy 2 subpockets, most commonly the S1 and S2 sites. Approximately 15% occupy a single subpocket, with the S4 subpocket being the least frequently occupied. Another 15% occupy 3 subpockets, while a complete subpocket engagement is the rarest (~5%). Notably, only 2 ligands (UGS from PDB 5RGY and A1BNP from PDB 7HUE) occupy the S4 subpocket only while most of the ligands occupying this subpocket tend to occupy at least 2 other subpockets.

For each ligand curated in the Xplorer, a quick overview of its PDB source(s), Lipinski data and a comparison of how similar it is against the remainder of the curated dataset are readily available (Fig. 3.8H, Appendix: Fig. A4(A)). Moreover, an interactive option for filtering the curated ligand dataset is also incorporated, whereby data points can be manually selected from any of the plots in the Dataset tab for exploratory analysis in the Xplorer tab (Appendix: Fig. A5(B)). The Data Import widget is then used for navigating between the collected datasets so that multiple mining sessions can be run at the same time (Fig. 3.8A). Naturally, 3D exploration of the selected ligands is the next step, and this is performed in the Mine tab,

where an Mpro dedicated drug discovery toolkit is readily available for extracting insights from the protein-ligand complex.

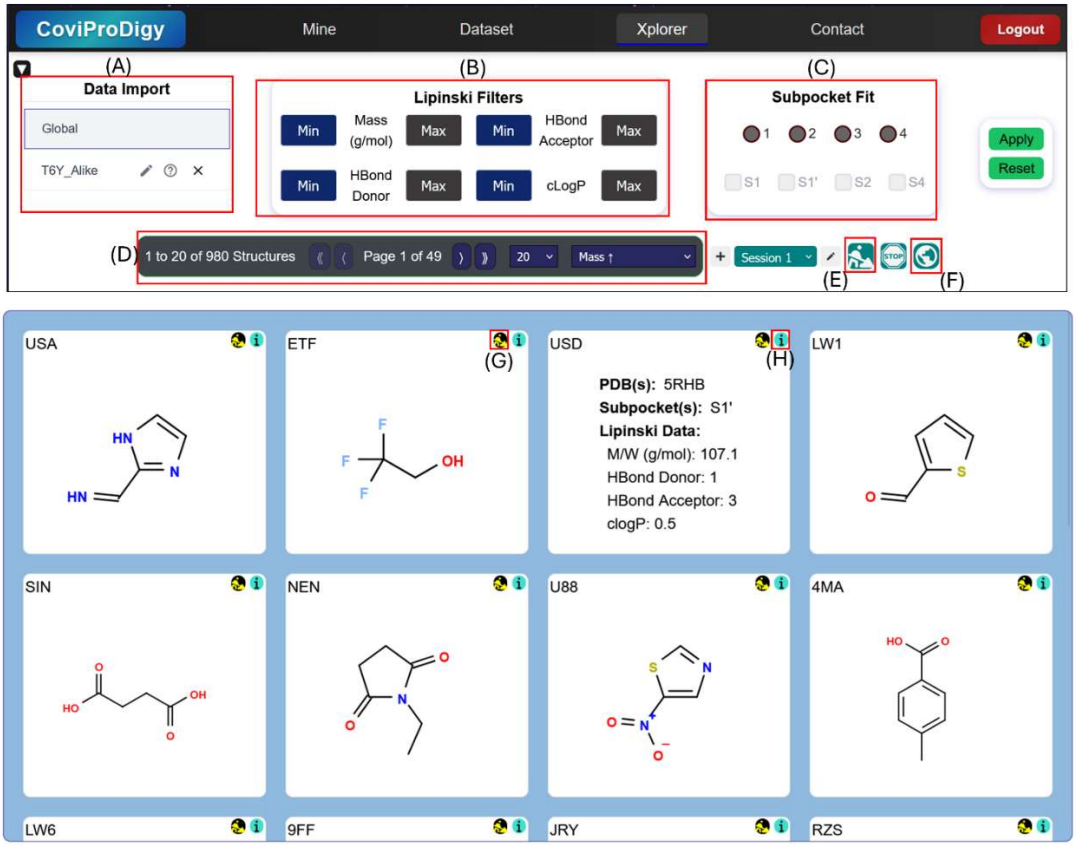


Fig. 3.8: Xplorer Tab Interface. The Xplorer tab in CoviProDigy provides an interactive interface for analyzing and filtering the ligand database. The top panel contains widgets for data import **(A)** and for filtering ligands by physicochemical properties **(B)** such as MWT in g/mol, HBD (indicated as HBond Donor) Donor, HBA (indicated as HBond Acceptor), cLogP (also indicated as cLogP). Additional filtering may be done according to subpocket occupancy **(C)**. The subpocket occupancies for all SARS-CoV-2 crystal ligands were determined using the optimized subpocket definition in Table 3.1. Below these, a central panel allows for navigating and sorting the retrieved hits **(D)**. Users can then view detailed information for a specific ligand **(H)**, analyze Tanimoto similarity **(F)**, or send selected ligand, either individually **(G)** or as a group **(E)**, to the 'Mine' tab for 3D structural exploration

3.3.4 3D Structure Mining and Interaction Profiling

The Mine tab is the core of CoviProDigy's 3D analysis capabilities. Users can input Mpro-ligand structures via PDB code, file upload, or by selecting them from the Xplorer tab (Fig. 3.9A). The platform includes a customized 3D viewer with several Mpro-specific features (Fig. 3.9B). These include toggles to display all Cys residues (highlighting potential disulfide bridges) and to show water molecules within the binding site, which can be crucial for mediating ligand interactions. A key feature is the ability to automatically zoom into the binding site and render the residues of each subpocket, with spheres marking their geometric centers, thereby facilitating a detailed assessment of subpocket engagement.

To complement the 3D view, the Mine tab also generates a 2D interaction diagram and an IFP heatmap for each ligand. These visuals detail the non-bonded interactions (hydrogen bonds, hydrophobic contacts, and aromatic stacking) between the ligand and specific Mpro residues, which are categorized by their respective subpockets. Additionally, covalent interaction between the ligand and the catalytic Cys145 is also depicted in the 2D interaction diagram. For example, analysis of the ligand X77 (PDB: 6W63) shows that it engages the S1, S1', and S2 subsites through multiple interactions types, including an aromatic stacking interaction with the catalytic His41, and a hydrogen bond with His163, a key residue for stabilizing ligand binding (Fig. 3.9C, D).[43] Its limited engagement with the S4 subpocket, as shown in the IFP heatmap (Fig. 3.9D), is consistent with its poor fit in that site.[56]

The platform also benchmarks the subpocket fit of a given ligand against the entire dataset. The fit is quantified as the distance from the ligand to the geometric center of a subpocket. The analysis of X77, for instance, confirms it is within the distance cut-off for the S1, S1' and S2 subsites but not for S4, corroborating the IFP data (Appendix: Fig. A5(A)). Finally, the ligand's Lipinski properties are displayed on the PCA plot and a separate radar chart, providing a quick assessment of its drug-likeness (Fig. A5(B)).

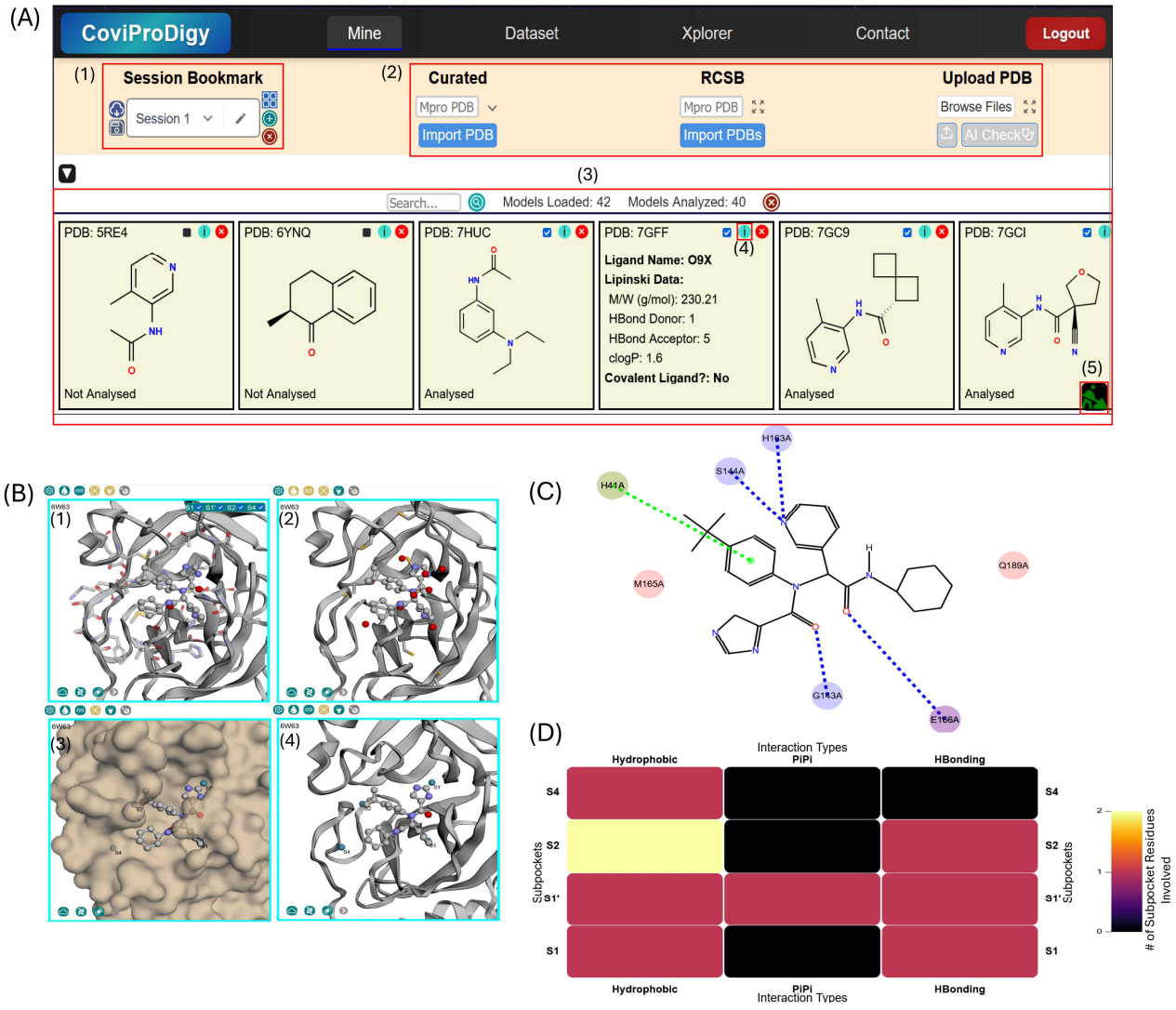


Fig. 3.9: Architecture of Mine Tab and SARS-CoV-2 Mpro Dedicated Toolkits. (A) The user interface of the 'Mine' tab, showing features for session management (A1), PDB data input (A2), and viewing loaded structures(A3). Within the model panel (A3), additional information about the loaded ligand such as its physicochemical properties, its interaction type with Cys145 (covalent or not) is available by clicking the info button (A4). Subpocket-guided insights extraction is initiated by activating the mine button (A5). **(B)** Examples of tailored visualization options for the Mpro active site, using the inhibitor X77 (PDB: 6W63) to highlight subpocket residues in stick representation (B1), subsite water molecules as red spheres (B2), all Cys residues in stick representation (B2), and subsite-specific geometric centers as dummy atoms (B3, B4). **(C)** A 2D interaction diagram for X77, showing hydrogen bonds (blue),

residues in hydrophobic contact (red), and aromatic stacking (green). **(D)** The corresponding interaction fingerprint map, which groups interactions according to subpocket involvement and the interaction type. Interactions for **(C)** and **(D)** were generated using ODDT.[84] RDKit[82] was used to flatten the 3D ligands coordinates for SVG depiction. **(D)** was created using D3.js.[90]

Beyond the standalone analysis of individual ligands on the Mine page, CoviProDigy features a powerful overlay mode for the comparative analysis of multiple ligands. To illustrate this, a subset of ligands sharing a similar scaffold to X77, identified using the Ligand Locator and Tanimoto Similarity Tree on the Dataset page, was loaded for direct comparison. In this overlay mode, the 3D visualizer renders all selected ligands simultaneously within the same Mpro binding site, leveraging the backend structural alignment protocol to superpose the protein structures (Fig. 3.10A). The 2D interaction diagram transforms into a summary plot that reveals the involvement of specific residues in non-bonded interactions across the entire subset (Fig. 3.10A). For the X77-like ligands, this analysis highlights that residues Glu166, His41, Gln189, and Gly143 are the most common points of interaction (Fig. 3.10A). This comparative view also extends to subpocket occupancies and Lipinski properties, enabling the rapid identification of key structural/physicochemical differences (Fig. 3.10B). For instance, while X77 does not occupy the S4 subpocket (Fig. 3.10C, D), the analysis quickly reveals that a structurally similar ligand, QCO (PDB: 7GI5), successfully engages this site (Fig. 3.10C, E). To our knowledge, there are no tools capable of comparative 2D interaction analyses for comprehending conserved and non-conserved interactions.

To address the common challenge of session discontinuity in online drug discovery platforms, CoviProDigy incorporates robust session management features. Each analysis session initiated in the Mine tab can be saved to the cloud through Firebase integration, allowing users to retrieve their work for later exploration (Fig. 3.9A). This functionality ensures that complex analyses are preserved over time. Furthermore, the platform includes an integrated notepad widget where users can record session-relevant observations. The content of this notepad is also automatically backed up to the cloud, ensuring that user annotations are securely saved and synchronized with their corresponding analysis session.

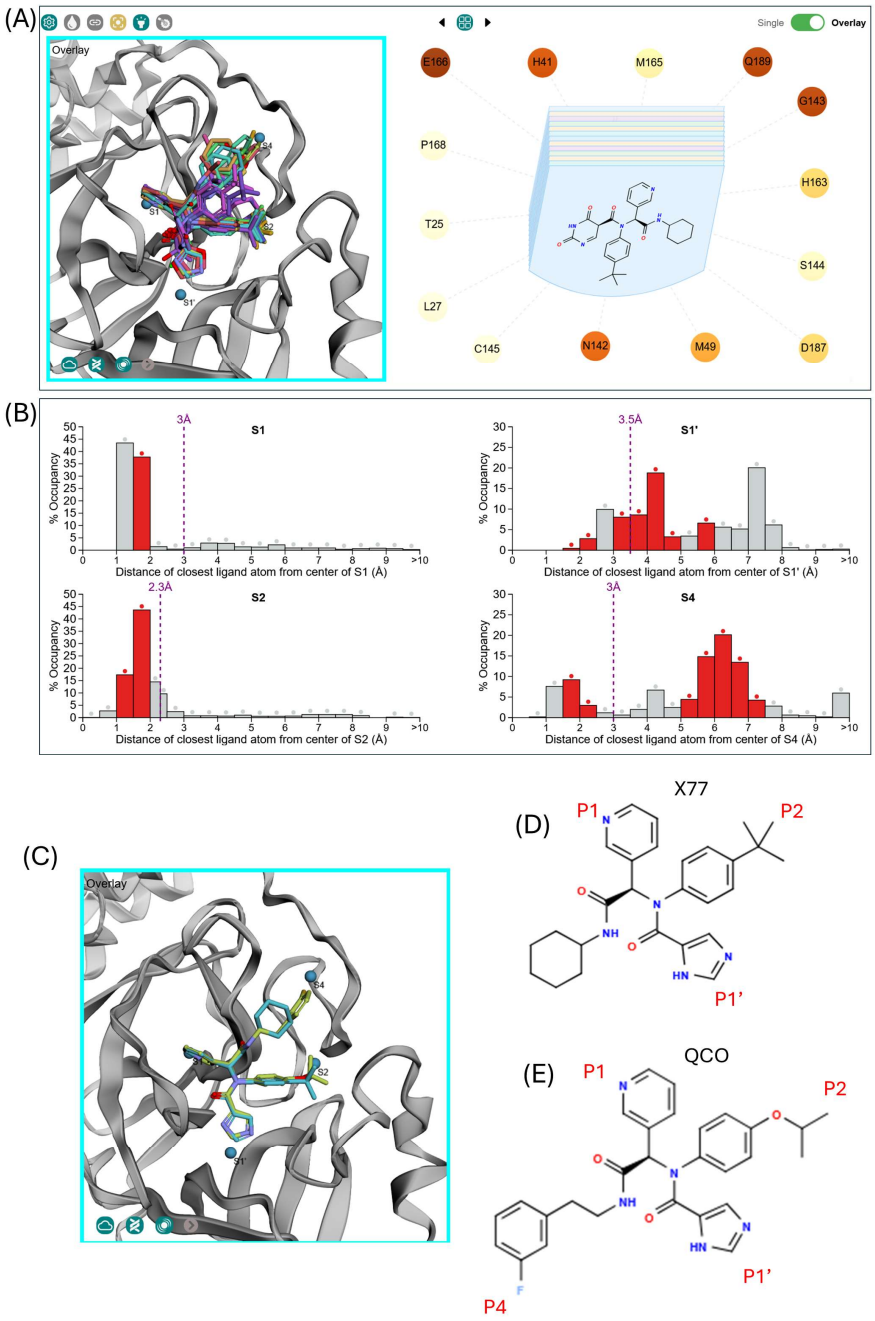


Fig. 3.10: Comparison of Binding Modes for the X77 and Related Analogues. (A) Left: Structural imposition of the ligands into the orthosteric site of Mpro. Ligands are shown in stick representation while subsite-specific geometric centers are rendered as dummy atoms. Right: Overlaid 2D interaction profiles for all ligands, with all interacting residues color-coded according to the local conservation of their respective interactions across the

local dataset. The higher the observation, the darker the color of the residue node. ODDT[84] was used to compute ligand interactions and the overlaid interaction was drawn using D3.js. **(B)** Distribution of subpocket occupancies for all 4 subpockets across all crystallographically determined ligands with all ligands' locations highlighted in red. **(C)** Structural superimposition of both ligands (PDB 6W63 for X77 and PDB 7GI5 for QCO) into the orthosteric site of Mpro. Both ligands are shown in stick representation with X77 colored in turquoise and QCO colored in dirty green. The geometric center for each subpocket is rendered as dummy atoms. **(D) – (E):** 2D structures of X77 (D) and QCO (E) with P1, P1', P2, and P4 moieties annotated according to their respective subpocket occupancies.

3.4 Discussion

The unprecedented research effort targeting the SARS-CoV-2 Mpro has resulted in a wealth of structural data, creating both a valuable resource and a significant analytical challenge. While essential archival platforms like PDB[60], and PDBsum[157] exist, they lack the integrated tools necessary for the comparative analysis required in modern drug discovery. Other platforms such as Fragalysis[158] offer improved visualization but still fall short of providing a comprehensive toolkit for medicinal chemistry optimization, particularly regarding quantitative subpocket analysis and systematic dataset exploration. CoviProDigy was developed to address this critical gap, providing a novel, web-based toolkit that transforms a static collection of PDB files into a dynamic and "queryable" knowledge base specifically tailored for Mpro inhibitor design.

The core strength of CoviProDigy lies in its integrated components tailored for supporting the decision-making process in medicinal chemistry. A researcher can seamlessly transition from a high-level overview of the entire ligand landscape to a granular, atomic-level comparison of specific inhibitors. For example, using the Dataset tab, one can identify trends, such as the trade-off between increased subpocket occupancy and Lipinski rule compliance[78], or discover novel scaffolds through Tanimoto similarity clustering. A subset of promising candidates can then be isolated in the Xplorer tab and subjected to in-depth analysis in the Mine tab's overlay mode. The utility of this approach is exemplified by the

comparative analysis of X77 and its analogues. The platform makes it immediately apparent that while these ligands share a common scaffold and engage many of the same key residues, a subtle modification allows the ligand QCO to successfully engage the S4 subpocket where X77 fails. This rapid generation of detailed structure-subpocket engagement relationships is crucial for guiding rational design efforts like scaffold hopping and lead optimization.

Beyond its analytical capabilities, CoviProDigy is designed to facilitate modern, collaborative research. By enabling users to save and retrieve analysis sessions from the cloud, the platform addresses the challenge of session discontinuity and supports long-term projects that can be accessed from multiple devices. This, combined with the integrated notepad for synchronized annotations, creates a persistent workspace for individuals and teams.

This toolkit aims to serve as a powerful platform for the development of potent SARS-CoV-2 Mpro inhibitors. The future development of CoviProDigy is focused on expanding its capabilities. Key enhancements will include: (i) expanding the input strain from SARS-CoV-2 to betacoronaviruses; (ii) incorporating a fragment library for combinatorial ligand design; (iii) implementing additional filtering and sorting options based on Tanimoto similarity scoring; (iv) providing comprehensive data download capabilities; and (v) curating the reported activity data for ligands against SARS-CoV-2 Mpro. A key aspect of this evolution is the recent integration of a fine-tuned machine learning model for classifying ligand poses in *Betacoronavirus* Mpro targets (discussed in Chapter 4). This feature will assist in validating ligand poses from large-scale screening campaigns, thereby ensuring higher data quality and reliability for future drug discovery efforts.

3.5 Conclusion

The search for potent Mpro inhibitors, initiated during the COVID-19 pandemic, has resulted in an unprecedented wealth of structural data, with no dedicated infrastructure to capitalize on it. CoviProDigy provides a unique and powerful solution to this challenge by offering a dedicated platform for the ligand exploration of SARS-CoV-2. By integrating dataset-wide

clustering, interactive filtering, and detailed comparative 3D analysis into a single, intuitive platform, it empowers researchers to efficiently extract actionable molecular insights. This toolkit is a valuable resource intended to accelerate the knowledge-driven design of potent and specific inhibitors for SARS-CoV-2 Mpro and to aid in the critical mission of preparing for future coronavirus outbreaks.

Chapter 4: Machine Learning for Ligand Pose Selection in Betacoronaviruses Mpro Complexes

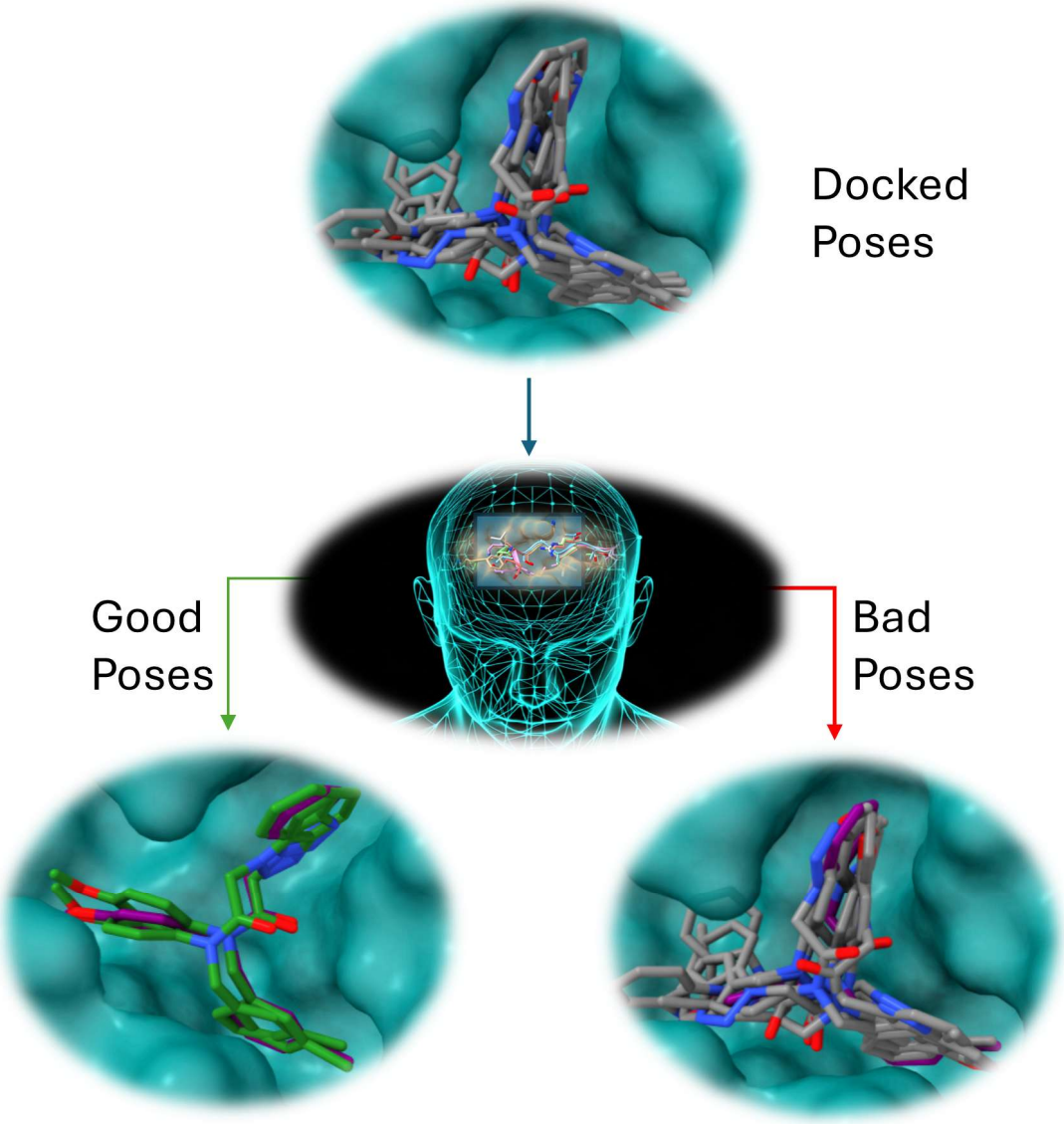


Image adapted from Gemini 2.5. Docked complex images were created using UCSF ChimeraX[47]. This is a mock image to describe the content of this chapter, and it was not used for any scientific purposes.

4.1 Introduction

The COVID-19 pandemic has rekindled research interest in betacoronaviruses, specifically the novel strain SARS-CoV-2.[27,40,61,62,169] While COVID-19 vaccines have been developed successfully[33], alternative therapeutic approaches were also explored, leading to an unprecedented surge of scientific data focused on a specific target: the SARS-CoV-2 Mpro.[43,61,65,134,147,170] Mpro is a cysteine protease that plays a critical role in the viral life cycle. It processes 2 large polyproteins, pp1a and pp1ab, releasing functional NSPs essential for viral transcription and replication.[61] Therefore, inhibiting the activity of this enzyme is a highly plausible therapeutic strategy for controlling the spread of the virus.

A key feature of Mpro is the high degree of conservation in its active site across the *Betacoronavirus* genus, which includes other notable human pathogens such as SARS-CoV and MERS-CoV (Fig. 4.1).[171,172] Specifically, the catalytic dyad (formed by His41 and Cys145) and most key residues lining each subsite show high conservation across the genus (Fig. 4.1). For subsite residues that are not conserved, their chemical properties are often conserved. For instance, Met49 from the S2 subsite (Fig. 1.3C) of SARS-CoV-2 Mpro is substituted for Leu in other strains, maintaining hydrophobicity at this position (Fig. 4.1). Another example is Phe185 from the S4 subsite (Fig. 1.3C) of SARS-CoV-2 Mpro, which is substituted for tyrosine in other strains, maintaining aromatic residues at this position.

This structural conservation presents a unique opportunity for the development of pan-coronavirus inhibitors, broad-spectrum therapeutics that could be effective against current and arising coronavirus threats. Significant screening campaigns were initiated in the search of COVID-19 therapeutics while at the same time curating a useful database for driving the development of broad-spectrum therapeutics. These ranged from large-scale, open-science consortia such as the COVID Moonshot project[147], which focused on developing a globally accessible SARS-CoV-2 Mpro inhibitor, to numerous academic laboratories aiming to identify promising inhibitors.[40,146,169,170,173,174]

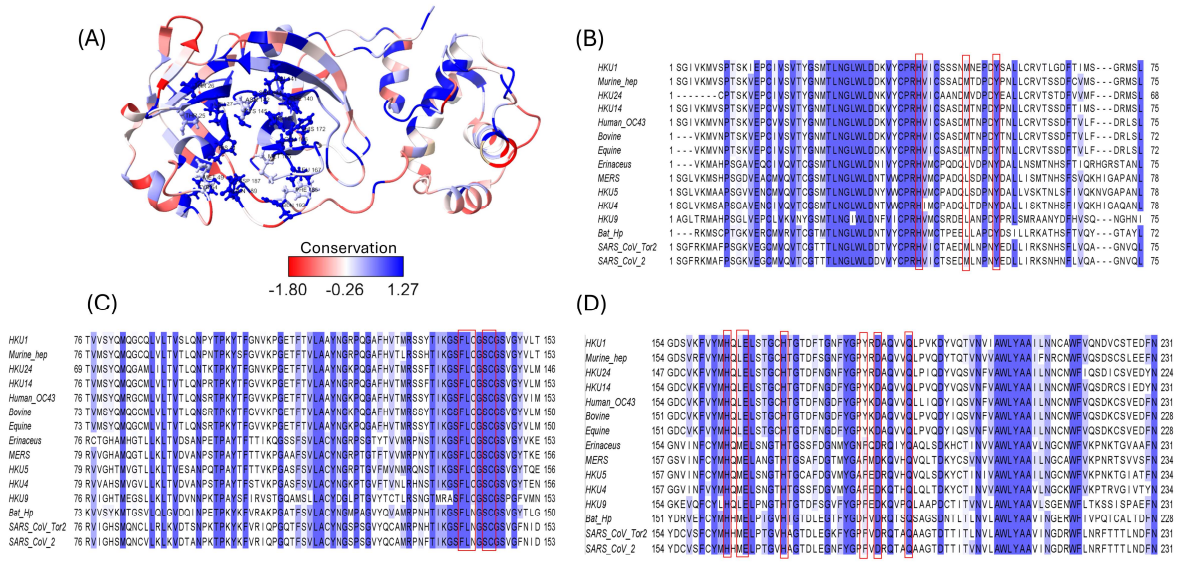


Fig. 4.1: Conservation of *Betacoronavirus* Active Site Residues. (A) Mpro structure color-coded according to sequence conservation across betacoronaviruses. The residues from the protomer of SARS-CoV-2 Mpro (PDB: 6W63) are color coded according their conservation score calculated using Clustal Omega.[75] The color palette ranges from red (least conserved) to blue (highly conserved). Residues that were not assigned a conservation score are colored tan. The protein backbone is styled in cartoon representation while subsite residues are labeled and shown in ball-and-stick representations. The image was created using UCSF ChimeraX.[47] **(B)-(C)** Sequence alignment of *Betacoronavirus* Mpro. Clustal Omega,[75] accessible via the European Molecular Biology Laboratory's European Bioinformatics Institute,[76] was used to perform the sequence alignment using the Mpro NCBI sequence reference sequence for 15 betacoronaviral strains.[62] Their NCBI accession codes are given in the Appendix (Table A1). Only portions of the generated alignment plot corresponding to the subsite residues are shown and these are boxed in red.[77]

To process the wealth of structural data from these screening campaigns and efficiently navigate the diverse chemical space, computational drug discovery methods were used intensively. Structure-based virtual screening, which involves docking small molecules into the protein's crystal structure, has been a primary approach.[134,175,176] The resulting hits are often further refined using more computationally intensive methods such as molecular dynamics (MD) simulations coupled with binding free energy calculations.[177] More recently,

Artificial Intelligence models have been employed to design de novo small molecules for targeting SARS-CoV-2 Mpro and to predict their potencies.[178–180]

Despite advances made in the landscape of computational tools, a fundamental challenge persists: accurately predicting the correct binding pose of a ligand within the target's active site. This is evidenced by the screening campaigns aimed at identifying Mpro inhibitors which failed to prioritize ligands with correct binding modes even though the docking algorithms were successful in generating plausible poses.[134,176] The root of this limitation lies in the classical scoring functions inherent to docking algorithms. These functions must strike a difficult balance between predictive accuracy and computational cost.[133] During large-scale virtual screening, this trade-off necessitates a compromise: accuracy is often sacrificed for speed to accommodate time and resource constraints. Consequently, these scoring inaccuracies lead to the wasteful pursuit of false positives while causing true inhibitors to be prematurely rejected.

To address this bottleneck, standalone scoring functions based on ML models have been developed. These ML models have been successful in solving the pose classification problem in drug discovery campaigns.[124,181–187] A major determinant of the performance of ML models is their learning algorithms, built on patterns extracted from the training set. In the context of pose classification, these patterns would be dependent on a set of descriptors representing the protein-ligand interaction environment.[79] These descriptors translate the complex protein-ligand interaction architecture into a "machine-friendly" fingerprint. During training, the ML model processes these fingerprints to fine-tune its algorithm, learning to recognize the specific patterns that define a stable, native-like interaction. This data-driven approach ultimately allows the model to make more nuanced predictions than those based on the simplified energy terms of classical functions.

Recently, Tran-Nguyen et al. implemented this approach to develop ClassyPose, an SVM model that uses PLEC of docked poses for quality assessment.[79] ClassyPose offers a rapid and lightweight alternative to resource-intensive methods like MD simulations and has shown promising performance in general pose classification.[79] However, its training dataset

was curated prior to the COVID-19 pandemic[188–191] and therefore excludes the sheer volume of SARS-CoV-2 Mpro structural data that has since become available. This suggests that ClassyPose may not be optimally suited for filtering Mpro docking results. The goal of this work, therefore, is to optimize the pose classification model using *Betacoronavirus* Mpro data, for creating a more accurate and sensitive tool for assessing modern screening campaigns against this important class of viral targets.

4.2 Methodology

4.2.1 *Betacoronavirus* Mpro Dataset Curation

The PDB codes of all *Betacoronavirus* Mpro was extracted from the PDB via its API access. Filters were applied using the taxon names for *Betacoronavirus* strains, published by the International Committee on Taxonomy of Viruses[192], and the PDB macromolecule names of Mpro which are 3C-like proteinase and ORF1a protein. A time-based filtering was applied to differentiate between SARS-CoV and SARS-CoV-2 entries. UCSF ChimeraX[47] was used to filter out apo crystal structures, keeping only Mpro with ligands bound in the orthosteric site.

4.2.2 *Betacoronavirus* Mpro Redocking

All *Betacoronavirus* crystal structures were first aligned to a reference SARS-CoV-2 Mpro crystal (PDB: 6LZE) using the command matchmaker from UCSF ChimeraX[47] before extracting the aligned receptors and orthosteric ligands separately. The prepare_receptor4.py and prepare_ligand4.py scripts from AutoDockTools[193] were used to add missing hydrogens and Gasteiger charges to the ligands and receptors, respectively, creating PDBQT files. These served as input for redocking all orthosteric ligands back to their original receptors using Vina-GPU 2.0.[194] A docking grid box of size 20 along the x, y, and z axes was defined to guide the redocking along the orthosteric site of Mpro. A total of 9 poses was captured per ligand.

4.2.3 Adapting Docking Results for Supervised Learning

An in-house script was used to calculate the RMSD between the docked poses and the corresponding crystal ligand poses. Open Babel[83] was used to create a one-to-one mapping between the heavy atoms while maintaining the same atom connectivity. This relationship was used to calculate the RMSD value. A docked pose was classified as “Good” if its RMSD value was less than 2 Å, and “Bad” otherwise. Next, the interacting environment between the docked ligand and its receptor was captured as a bits representation of length 4092 using the PLEC[80] method from ODDT.[84] At this stage, each docked pose had a label (Good/Bad) and a PLEC consisting of 4092 bits capturing the interaction environment between the protein and the ligand.

4.2.4 Training Set-Test Set Partition

The data partitioning process began with crystal ligands having at least one "Good" redocked pose. A similarity matrix was constructed from the PLEC of all possible crystal ligand pairs using the pdist method (Euclidean metric) from SciPy.[163] Hierarchical clustering[162] was applied to this matrix using SciPy's linkage function (Ward’s method[162]), grouping ligands with similar crystal binding modes into a dendrogram. The optimal number of clusters was determined by identifying the largest gap between merging clusters along the dendrogram.[145] PCA,[143] implemented using SciPy,[163] was used to visualize the cluster assignment of the PLEC in the 2D space. From each cluster, 80% of the ligands were allocated to the training set and the remaining 20% to the test set. Ligands that did not produce any "Good" poses were assigned to a separate "hard" test set.

4.2.5 Training and Benchmarking the ML model

The supervised learning algorithm, SVM[195], was used to train the ML model using the PLEC and their associated labels from the training dataset. This was implemented using the support vector classification method from Scikit-learn[161] using the default hyperparameters.[79] The model was configured with a radial basis function (RBF) kernel, a regularization parameter (C) of 1.0, and a scaled kernel coefficient (gamma='scale'). No further hyperparameter tuning was performed. The pan coronavirus optimized model and

the published ClassyPose[79] model were benchmarked using the test and hard datasets based on 5 metrics implemented from Scikit-learn[136,137,161]: (i) Specificity, (ii) Recall, (iii) Precision, (iv) ROC-AUC, and (v) PR-AUC.

4.3 Results

The crystal structures of ligand-bound *Betacoronavirus* Mpro were retrieved from the PDB[60] for a total of 4 strains: (i) HKU4, (ii) MERS-CoV, (iii) SARS-CoV, and (iv) SARS-CoV-2 (Table 4.1). These strains have been responsible for respiratory disease outbreaks in the past.[61] The SARS-CoV-2 strain had 91 ligands duplicated across multiple PDB entries while the SARS-CoV strain had 1 ligand duplicated across 3 PDB entries. A data cleaning step was performed to remove duplicate entries by comparing the RMSD values within each common ligand set, and removing ligand candidates until the RMSD value(s) for all possible remaining ligand pair(s) was/were above 2 Å. A total of 265 SARS-CoV-2 and 1 SARS-CoV ligand duplicates were removed in this step. Interestingly, after duplicate conformation pruning, 6 ligands were still crystallized in alternate conformations (RMSD > 2 Å) across different PDB entries (Fig. 4.2): (i) K36 (SARS-CoV-2 PDBs: 6WTJ, 6WTT), (ii) AG7 (SARS-CoV-2 PDBs: 7L8I, 7P35), (iii) UED (SARS-CoV PDBs: 7LCP, 8IG5; SARS-CoV-2 PDBs: 6WTK, 8IG7*)*, (iv) FHR (SARS-CoV-2 PDBs: 6LZE, 8YWZ), (v) B1S (SARS-CoV-2 PDBs: 6WTT, 7CB7), and (vi) ALD (SARS-CoV-2 PDBs: 7BE7, 7WQK).

Table 4.1: Retrieval Pan Coronavirus Crystal Structures. The crystal structures of 4 *Betacoronavirus* strains were retrieved from the PDB[60]: (i) HKU4, (ii) MERS-CoV, (iii) SARS-CoV, (iv) SARS-CoV-2. HKU4 had the lowest number of depositions (4) while SARS-CoV-2 had the highest number (1148). MERS-CoV and SARS-CoV had similar numbers of depositions: 28 and 21. SARS-CoV-2 had 199 duplicate crystal structures, which were pruned during the data cleaning step.

Betacoronavirus Strain	Number of crystal ligands retrieved	Number of crystal ligands retrieved after duplicate removal
HKU4	4	4
MERS-Cov	55	55
SARS-CoV	46	38
SARS-CoV-2	1434	1421

4.3.1 Redocking of PanCoronavirus Mpro Ligands

Given the high conservation of the Mpro active site among betacoronaviruses, all pruned crystal structures were structurally aligned to a single reference structure (SARS-CoV-2 PDB: 6LZE). This approach ensured that a consistent docking configuration was maintained throughout the process. Since a large-scale redocking was to be performed, a GPU-based pipeline powered by Vina-GPU 2.0[194] was established. Receptor and ligand prep for docking were automated via the Python shell of AutoDock Tools[193] for executing the docking prep python scripts prepare_receptor4.py and prepare_ligand4.py. The prepared biological molecules were processed with the docking pipeline along with a universal docking configuration file in order to be redocked using Vina-GPU 2.0.[194] Redocking failed for 18 ligands due to errors with the docking software invoked along the redocking pipeline.

The docking results highlight the limitation in Vina-GPU 2.0[194] scoring function's ability to identify correct ligand poses. For the HKU4 strains, the scoring function performed counter-intuitively: the most inaccurate docked poses (RMSD approximately 8 Å) were ranked as the

most stable, while near-native poses (RMSD <= 2 Å) were scored as the least stable (Fig. 4.2A). A similar trend was observed for MERS-CoV, where the scoring function prioritized only a small subset of near-native poses. The remaining near-native poses received scores that were indistinguishable from those of inaccurate poses, preventing their reliable identification (Fig. 4.2B). For SARS-CoV and SARS-CoV-2, there was also an absence of a monotonic relationship between docking energies and RMSD values (Fig. 4.2B). Ultimately, for all strains tested, the scoring function of Vina-GPU 2.0[194] failed to reliably prioritize near-native redocked poses.

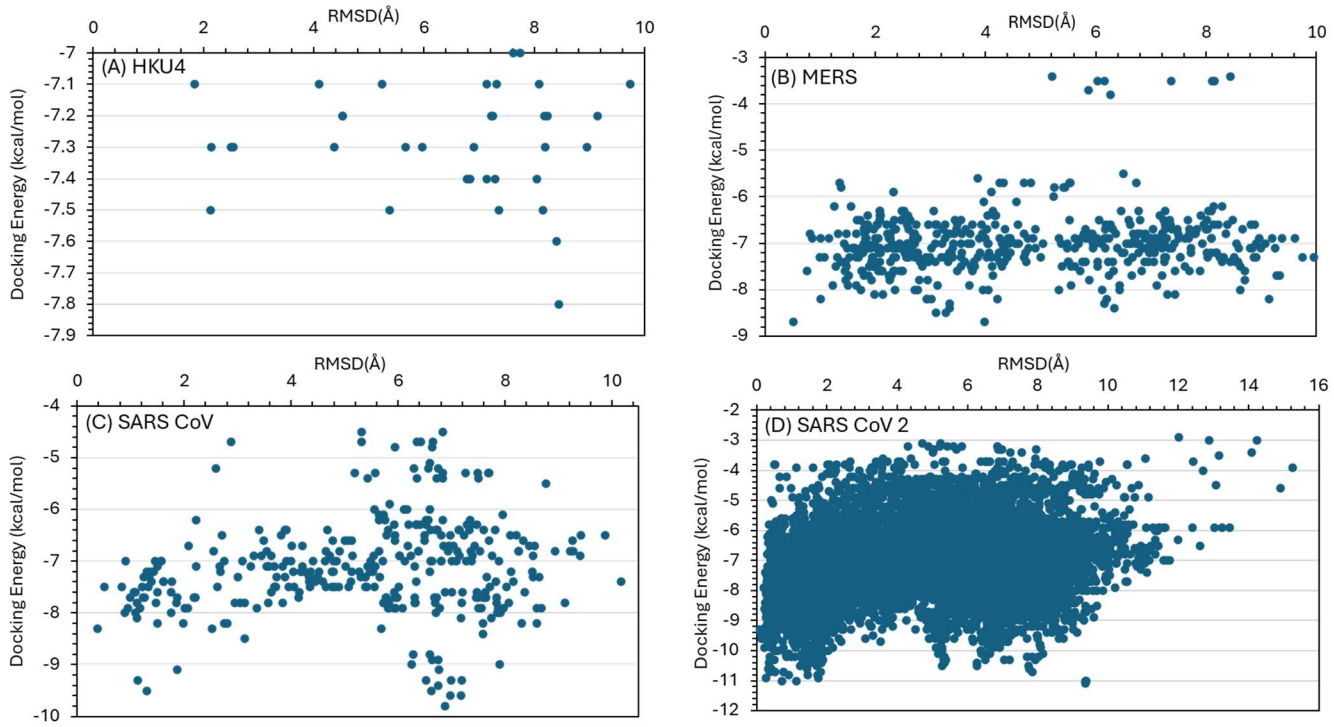


Fig. 4.2: Benchmarking the Scoring Function of Vina-GPU 2.0[194] via Redocking of Pan Coronavirus Mpro Ligands. Vina-GPU 2.0[194] was used to redock all pan coronavirus Mpro ligands back to their originating receptor crystal structure. The RMSD was computed for all redocked ligands against their corresponding crystal poses using an in-house script implemented using Open Babel.[83] The docking energies were plotted against the RMSD values for 4 *Betacoronavirus* strains separately: **(A)** HKU4, **(B)** MERS, **(C)** SARS-CoV, and **(D)** SARS-CoV-2. For all strains tested, the scoring function of Vina-GPU 2.0[194] failed to reliably

prioritize near-native redocked poses. A non-monotonic relationship existed between the RMSD values and the docking energies.

4.3.2 Fingerprinting Redocked Pan Coronavirus Ligand Poses for ML Classification

To address the limitations of Vina-GPU 2.0[194] scoring function, the ML classifier ClassyPose[79] was implemented. The initial step of this approach is to translate the 3D protein-ligand poses into a format suitable for the model. This translation was achieved by converting each docked pose into a PLEC, a bit string representation of the interaction environment between the ligand and the receptor.[80]

The PLEC method captures all protein atoms interacting with each ligand atom within a defined distance. For this analysis, a distance cutoff of 4.5 Å was used. This concept is illustrated in Fig. 4.3A, whereby all Mpro active site residues that fall within cutoff are highlighted for the ligand X77 (PDB: 6W63). The generation of the fingerprint is a multi-step process. An illustration is provided in Fig. 4.3B, using the nitrogen atom on the pyridine ring of X77 as example. First, the algorithm identifies all protein residues interacting with this atom.[80,84] 9 residues are within this distance cutoff (Phe140, Leu141, Asn142, Ser144, Cys145, His163, Met165, Glu166, and His172).[80,84] A preliminary receptor-independent fingerprint is generated for the ligand's nitrogen atom by considering its local atomic environment (neighboring atoms up to 2 bonds away), as shown in Fig. 4.3B.[80,84]

Similarly, a ligand-independent fingerprint is generated for each of the identified protein atoms. For example, the side-chain hydroxyl group of Ser144 interacts with the pyridine nitrogen. A fingerprint for this hydroxyl group is created by capturing its local environment (neighboring atoms up to 4 bonds away) (Fig. 4.3B).[80,84] The ligand-side fingerprint (from the pyridine nitrogen) and the protein-side fingerprint (from the Ser144 hydroxyl) are then combined into a bit string representing this specific interaction.[80,84] This process is repeated for the interactions between the nitrogen atom and the other 8 residues, creating a complete interaction profile for that ligand atom.[80,84] Finally, this entire procedure is carried out for every atom in the ligand. The information from all protein-ligand atomic interactions is

encoded into a final PLEC of 4092 bits. This serves as a machine representation of the binding pose's interacting environment, ready for assessment by ClassyPose.[79]

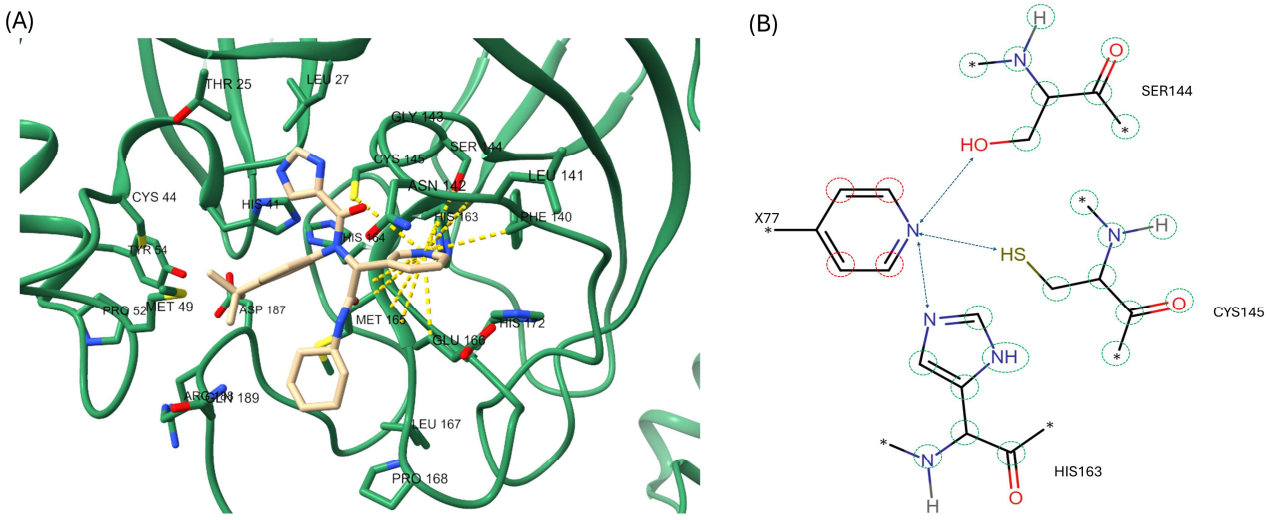


Fig. 4.3: PLEC Generation of SARS-CoV-2 Ligand X77. (A) Interacting environment of X77 (PDB: 6W63) is depicted using UCSF ChimeraX.[47] Backbone protein atoms are colored green and shown in licorice representation. Sidechain atoms for all protein residues within a distance of 4.5 Å from any X77 ligand atoms are represented in stick representation and labeled according to their corresponding residue names and numbers. Sidechain carbon atoms are colored green while sidechain oxygen, nitrogen, and sulfur atoms are colored red, blue, and yellow, respectively. The ligand atoms are shown in stick representation with all carbon atoms colored beige and remaining nitrogen, and oxygen atoms colored blue and red, respectively. Dotted yellow lines are used to link residue atoms with the pyridine nitrogen atom of X77 if the distance between them is less than 4.5 Å, involving 9 protein residues in total (Phe140, Leu141, Asn142, Ser144, Cys145, His163, Met165, Glu166, and His172). **(B)** Breakdown of the algorithm used to generate PLEC. The pyridine nitrogen atom from X77 is used as an example, and all ligand atoms circled in red are within 2 covalent bonds away from it. These are considered from generating a receptor-independent fingerprint representation of the nitrogen atom. 3 residues from 9 interacting residues with the pyridine nitrogen are used as examples to illustrate the generation of ligand-independent fingerprint representations: Ser144, Cys145, and His163. The protonation state of His163

was selected based on its crystallized protonation state. The residue atoms circled in green are within 4 covalent bonds away from the main residue atom interacting with the pyridine nitrogen, as indicated by the dual dotted arrows. The independent fingerprints from each pyridine nitrogen and residue interacting pair are combined, generating a final PLEC of 4092 bits.

4.3.3 Benchmarking ClassyPose using Redocked PanCoronavirus Ligands

To benchmark the performance of ClassyPose, each redocked pose was described by two key inputs: its RMSD value relative to the crystal pose and its corresponding PLEC. The model's ability to distinguish between near-native (correct) and non-near-native (incorrect) poses was then assessed. ClassyPose classifies poses using a probability function, where a probability greater than 0.5 designates a "good" pose.[79,161] The performance was evaluated over five training-test runs using five standard metrics[136,137]: PR-AUC, ROC-AUC, Specificity, Recall, and Precision.

The analysis revealed that the published ClassyPose model demonstrated high recall, achieving an average value of approximately 0.8 (Fig. 4.4A). This indicates that the model correctly identified around 80% of the total available good poses. Furthermore, the average ROC-AUC score was greater than 0.6, suggesting ClassyPose performs better than a random classifier.[137] However, the model's high recall was contrasted by poor specificity. With an average specificity of approximately 0.45, the model correctly identified only about 45% of the incorrect poses.[137] This highlighted the model's difficulty in labeling all available bad poses correctly.

This issue was further compounded by a low precision score, which was less than 0.4 on average. This score implies that when ClassyPose labeled a pose as "good," it was correct less than 40% of the time.[137] The practical consequence is that while the model finds most of the true good poses (high recall), they are buried within a larger pool of false positives.

Finally, the trade-off between high recall and low precision is highlighted by the PR-AUC score of less than 0.4. This low value confirms that the model is biased towards a high false-

positive rate in its effort to correctly label all good poses, limiting its utility for reliably filtering docking results.[136]

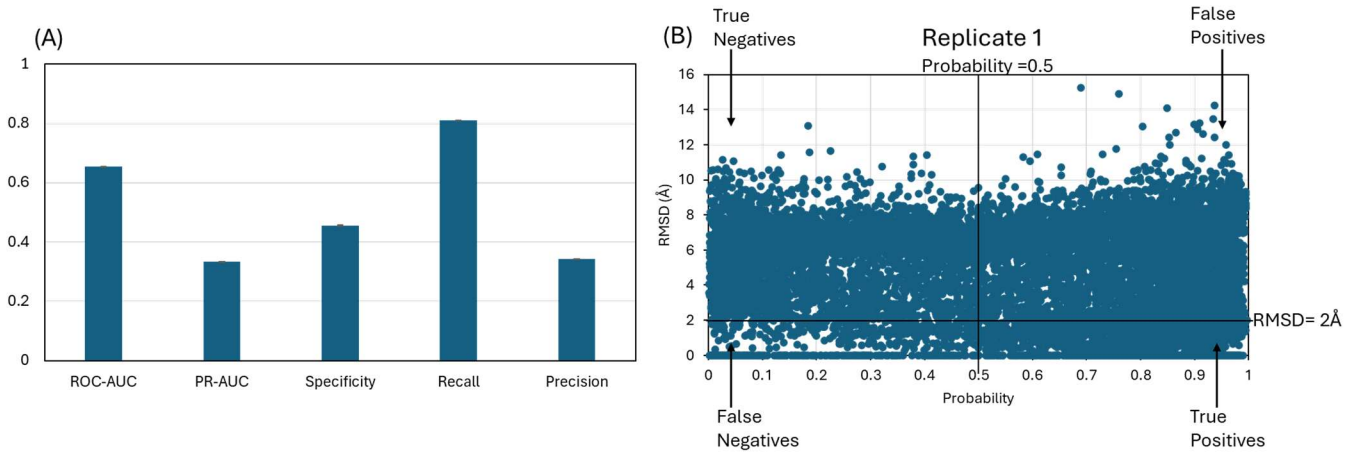


Fig. 4.4: Outcome of Pan Coronavirus Redocked Poses Classification by ClassyPose. (A) Benchmark metrics used to assess the performance of ClassyPose. 5 training-test runs were performed, and the averaged benchmark metrics are reported along with their corresponding standard deviations. 5 benchmark metrics were used: (i) PR-AUC (Precision-Recall Area Under Curve), (ii) ROC-AUC (Receiver Operating Characteristic Area Under Curve), (iii) Specificity, (iv) Recall, (v) Precision. With a recall score of approximately 0.8, ClassyPose was able to label approximately 80% of the good pose population correctly. With an ROC-AUC greater than 0.5, ClassyPose would outperform another binary classifier operating solely at random. However, with a specificity of approximately 0.45, Classypose struggled to classify true negative poses correctly, resulting in a true negative rate of approximately 45%. Lastly, while Classypose correctly identified true good poses, it also displayed a high false positive rate, with both a precision score and PR-AUC score of less than approximately 0.4. **(B)** Graphical representation of ClassyPose's classification for the first replicate. ClassyPose assigned probability scores were plotted against the corresponding RMSD values of the poses. The plot was divided into 4 quadrants by splitting along the probability axis at 0.5 and along the RMDS axis at 2: (i) true negatives quadrant (probability <= 0.5 & RMSD > 2 Å), (ii) false positives quadrant (probability > 0.5 & RMSD > 2

Å), (iii) true positives quadrant (probability > 0.5 & RMSD <= 2), and (iv) false negatives quadrant (probability <= 0.5 & RMSD <= 2)

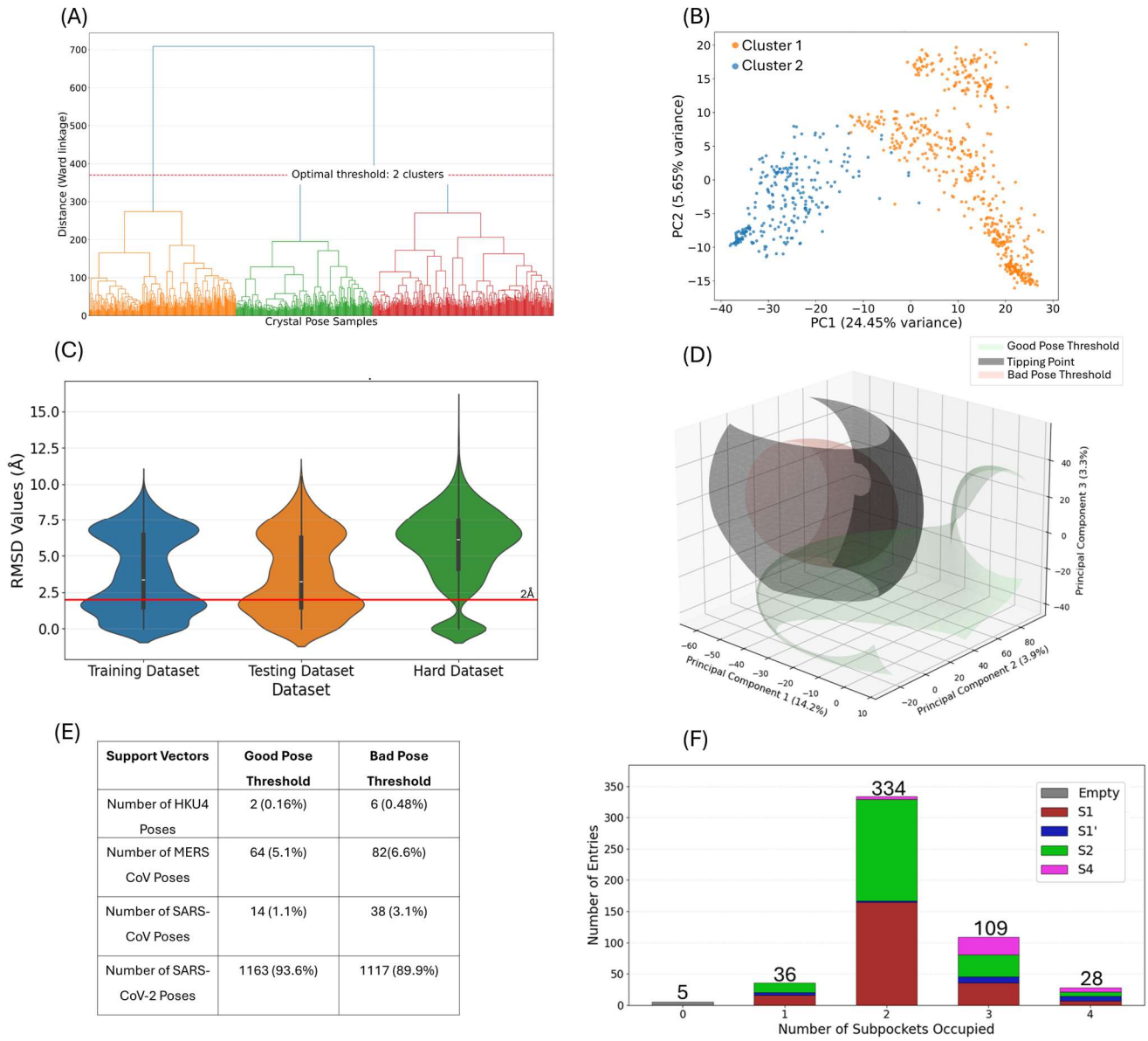
4.3.4 Optimizing ClassyPose for Pan Coronavirus Mpro Ligand Poses Classification

The published ClassyPose model was trained on a dataset curated before the COVID-19 pandemic.[188–191] This lack of exposure to Mpro-specific fingerprints likely explains its suboptimal performance in the previous benchmark. Therefore, to optimize its performance, ClassyPose was retrained on a new dataset composed entirely of pan-coronavirus Mpro data. The first step was to construct new training, test, and "hard" datasets. A preliminary analysis of the 1518-ligand population revealed that Vina GPU 2.0[194] produced at least one good redocked pose for 739 ligands. The latter were used to construct the training and test sets. The remaining ligands, for which no good pose was generated, were used to create the hard dataset, an approach adapted from Tran-Nguyen et al. for assessing model performance on more challenging, "unseen" data.[79]

To ensure domain compatibility between the training and test sets,[196] the selected ligands were first clustered hierarchically based on the PLEC of their crystal poses. An optimal threshold cutoff revealed 2 distinct clusters (Fig. 4.5A). PCA was then performed on the PLEC to visualize this clustering. The resulting PCA plot confirmed a clear segregation of the two clusters in the 2D space, although it also suggested the possible presence of additional sub-clusters (Fig. 4.5B). However, because the first two principal components encoded only 30.1% of the total variance, a reliable conclusion about sub-clusters could not be drawn from the PCA plot alone. The datasets were then populated by splitting each of the two clusters: 80% of the ligands from each cluster were assigned to the training set, and the remaining 20% were assigned to the test set. Analysis of the RMSD distributions showed that for the training and test sets, the two most frequent RMSD values were near 2 Å and 7 Å. In contrast, the hard dataset was composed primarily of poses with RMSD values around 7 Å (Fig. 4.6C). The final training dataset contained 6010 poses, 36.1% of which were good poses (RMSD <= 2 Å). The test set had a similar distribution but a smaller size (1560 poses), while

the hard set contained 4820 poses, of which only 10% were good poses (crystal poses only). A detailed strain-specific composition is given in Table 4.2 for all 3 datasets.

The SVM model was then trained using the PLEC and their corresponding classifications ("Good" or "Bad") from the new training dataset. During this supervised learning process, the SVM maps the fingerprints to a multi-dimensional space and defines a hyperplane that optimally segregates the data points.[195] This hyperplane represents the decision boundary the model uses to classify new poses (depicted as the "tipping point" in Fig. 4.5D).[195] A margin is defined on both sides of this boundary, marking the regions where classifications are made with high confidence (labeled "Good Pose Threshold" and "Bad Pose Threshold" in Fig. 4.5D).[195] The definition of this margin is determined by a critical subset of training data points known as support vectors.[195] Analysis of the support vectors for the retrained model revealed that they were diverse, including 1243 "good" poses and 1243 "bad" poses, spanning across all 4 *Betacoronavirus* strains (Fig. 4.5E). The support vectors, representative of good poses, captured the interaction profiles across all four Mpro subpockets (Fig. 4.5F). Importantly, the model captured a key binding characteristic identified in Chapter 3 (CoviProDigy), showing a preference for the S1 and S2 subsites, which were the most occupied subpockets. A minor engagement with the S1' subsite was observed across these support vectors. However, from chapter 3, it was shown that a ligand could still interact with the catalytic dyad from this subsite despite being classified as a non-occupant. Thus, a minor involvement of S1' across the good support may not necessarily equate to a disregard of interactions involving the catalytic dyad when classifying a pose.



Support Vectors	Good Pose Threshold	Bad Pose Threshold
Number of HKU4 Poses	2 (0.16%)	6 (0.48%)
Number of MERS CoV Poses	64 (5.1%)	82(6.6%)
Number of SARS-CoV Poses	14 (1.1%)	38 (3.1%)
Number of SARS-CoV-2 Poses	1163 (93.6%)	1117 (89.9%)

Fig. 4.5: Supervised Learning Protocol of SVM ML Model for Pan Coronavirus Mpro Optimization. (A) Hierarchical Clustering of Crystal Pan Coronavirus Mpro Ligand PLEC**.** The crystal PLEC of all pan coronavirus Mpro ligands having at least 1 good redocked pose (RMSD <= 2 Å) were clustered hierarchically using the linkage (Ward's method[162]) function from the Python package SciPy[163]. The outcome of this clustering is shown as a dendrogram with the crystal pose input on the x-axis and the distance (obtained using Ward's method) on the y-axis. The optimal number of clusters was determined by the largest gap in vertical heights between the recombinant clusters, which occurred with a cluster size of 2. A

horizontal red dotted line is drawn at the threshold for the optimal number of clusters. **(B)** Projection of Crystal PLEC in 2D Space for Visualizing Clustering Outcome**.** The crystal PLEC of all pan Coronavirus Mpro ligands having at least 1 good redocked pose (RMSD <= 2 Å) were projected into the 2D space using the PCA method from the Python package SciKit-learn[161]. The first 2 major PCs collectively accounted for 28.44% of the total variance. The datapoints were colored according to their cluster assignment, determined by the hierarchical clustering performed previously. Datapoints belonging to cluster 1 were colored gold while datapoints belonging to cluster 2 were colored blue. The PLEC have high dimensionality (4092 bits), and a reduction into 2D was not sufficient for capturing most of the variability in the untransformed dataset. This illustration was done for a visualization of the outcome of clustering hierarchically the PLEC. **(C)** Pose RMSD Distribution Across Training, Testing and Hard Dataset. Violin plots[197] are used to visualize the distribution of RMSD values across the training dataset (blue), testing dataset (orange), and hard dataset (green). The seaborn library[198] was implemented in Python to render the plots, and the threshold for pose classification was marked as a red horizontal line where RMSD = 2 Å. The plots broaden horizontally at RMSD values in the neighborhoods of 2 and 7 Å for the training and test datasets, indicating the most frequently observed RMSD values in those datasets. For the hard dataset, the most frequently observed RMSD values was approximately 7 Å, and the heavy tailing in the violin plot was due to some poses having high RMSD values (outliers). **(D)** Visualizing the Optimized Decision Boundary after Supervised Learning**.** The SVM model, implemented using SciKit-learn[161], was optimized using the training dataset, and its decision boundary was projected onto the 3D space defined by the first 3 PCs of the PCA decomposition of the training dataset. The tipping point surface (shown in black) represents the hyperplane that optimally separates the 2 classes of data (Good/Bad) from the training set. The good pose threshold (green) and bad pose threshold (black) are constructed using the datapoints from the training site that lie optimally at the interface between good and bad pose classification (also known as the support vectors)[195]. **(E)** Characterizing Support Vectors for Good and Bad Pose Thresholds. The poses selected as support vectors for the good and bad pose thresholds were characterized according to the number of strain-specific

poses present (HKU4, MERS-CoV, SARS-CoV, and SARS-CoV-2), and the number of unique ligands present. Support vectors belonging to HKU4 strain were not observed for either threshold, as HKU4 poses were not assigned to the training set. **(F)** Characterizing Subpocket Occupancies of Support Vectors for Good Pose Threshold**.** The algorithm for assessing subpocket occupancy in CoviProDigy (Chapter 3) was adapted to assess the subpocket occupancy of all poses chosen as the support vectors for the Good Pose Threshold. 6 ligands were omitted from this analysis as they did not occupy any Mpro subpocket. In the histogram, each bar, representing the number of subpocket occupied, was stratified according to the involvement of each subpocket (S1 (crimson), S1' (blue), S2 (green), S4 (prink)) at the given level. Zero subpocket involvement was indicated in grey color. S1 and S2 subsites showed the highest ligand involvement among the support vectors.

Table 4.2: Characteristics of the Training, Testing, and Hard Datasets. A breakdown of the poses assigned to each dataset (Training, Testing, Hard) is given according to betacoronaviral strains, and pose quality (Good pose: Redocked pose RMSD < 2 Å or Crystal pose)

Dataset	**Training**	**Testing**	**Hard**
Number of assigned Poses	6010	1560	4820
Number of HKU4 Poses	10	0	30
Number of MERS-CoV Poses	290	30	230
Number of SARS-CoV Poses	110	60	200
Number of SARS-CoV-2 Poses	5600	1470	4360
Total Number of Good Poses	2172	564	482
Percentage of Good Poses (%)	36.1	36.2	10

4.3.5 Benchmarking ClassyPose and the Optimized SVM model on Testing Dataset

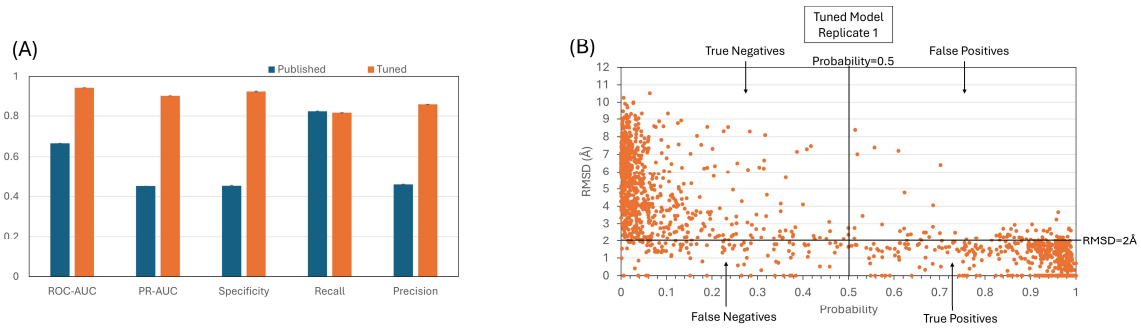
To evaluate the benefits of retraining, the performance of the published ClassyPose model was benchmarked against the newly optimized SVM model. Both models were assessed in five replicates on the same testing dataset, which was composed of 36.2% good poses (Table 4.2). The initial comparison revealed one area of shared strength: recall. Both the original and the optimized models demonstrated a high capacity to identify the correct poses within the dataset, achieving average recall scores around 0.8 (Fig. 4.6A). This indicates that both models successfully retrieved approximately 80% of the available good poses.[137] This is visually confirmed in Fig. 4.6B & 4.6C, where the data point distribution is far more prominent in the true positives quadrant than in the false negatives quadrant for both models.

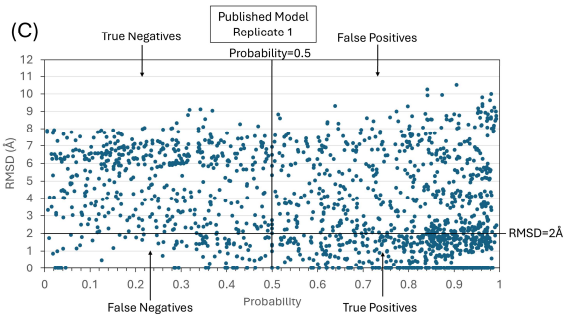
However, this was the only point of similarity. The optimized model showed considerable improvements across all other metrics. ClassyPose struggled significantly with identifying incorrect poses, achieving a low average specificity score of approximately 0.45.[137] In contrast, the optimized SVM model excelled, correctly classifying bad poses nearly 90% of the time (average specificity ≈ 0.9).[137] This represents a two-fold improvement in the model's ability to reject false positives (Fig. 4.6B, C). While both models found most of the good poses (high recall), the reliability of their predictions differed significantly. ClassyPose's predictions were correct only about 45% of the time (average precision approximately 0.45). The optimized model, however, demonstrated a much higher true positive rate, with its predictions being correct approximately 85% of the time (average precision approximately 0.85). This superior precision is evident in Fig. 4.6B, where data points with a probability > 0.5 are clustered in the true positives quadrant, a pattern not seen for the original model (Fig. 4.6C). Consequently, the optimized model achieved a suitable balance between recall and precision, reflected in a high PR-AUC score of approximately 0.9. ClassyPose, hindered by its high false-positive rate, scored only around 0.45. As a binary classifier, the optimized

model also significantly outperformed ClassyPose, with an average ROC-AUC score exceeding 0.9, indicating a near-perfect classification performance (Fig. 4.6A).[136]

A decomposition of the results by virus strain revealed the source of ClassyPose's poor performance. The original model was unable to correctly classify any incorrect poses from three of the four *Betacoronavirus* strains (HKU4, MERS, and SARS-CoV), demonstrating a critical flaw (Fig. 4.6D). Only for the SARS-CoV-2 strain did it achieve a modest specificity of approximately 50%. The optimization step successfully addressed these deficiencies. The retrained model was able to correctly classify most poses from the MERS, SARS-CoV, and SARS-CoV-2 strains.

Even with its high performance, the optimized model was not perfect. The testing dataset contained 564 good poses, which the model classified as either true positives or false negatives. A detailed analysis revealed that 15 specific ligands, all from SARS-CoV-2 Mpro structures, were consistently misclassified as false negatives (Fig. 4.6F): 4N0 (PDB: 7P2G), 5IL (PDB: 7V1T), A1AA0 (PDB: 8VDJ), A1ALF (PDB: 9BBQ), A1ALW (PDB: 9BBP), A1BVZ (PDB: 7I13), A1JGD (PDB: 9RI1), ASC (PDB: 7MPB), KMF (PDB: 7GB1), M9U (PDB: 7GDB), NZK (PDB: 7GEZ), RV8 (PDB: 7AQJ), RWO (PDB: 7GNE), RZS (PDB: 5R82), and U2I (PDB: 8DII). A subpocket engagement analysis revealed the following (Fig. 4.6G): (i) 1 ligand occupied no subpocket; (ii) 6 ligands occupied 1 subpocket; (iii) 7 ligands occupied 2 subpockets; and (iv) 1 ligand occupied 3 subpockets. The commonly occupied subpockets were S1 and S2, which mirrored the characteristics of the support vectors of the good pose support vectors (Fig. 4.5F). The misclassification may be due to the nature of the ligands, causing their interaction environments to deviate from that of the good pose support vectors.





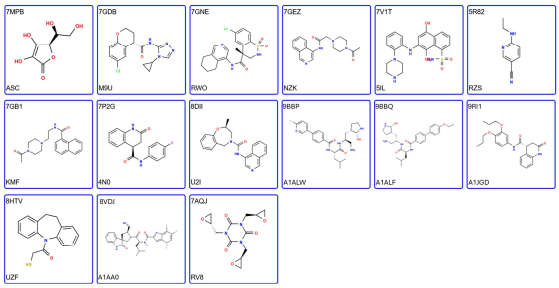
(D)

Published Model (Replicate 1)		HKU4	MERS	SARS_CoV	SARS_CoV_2
Good Pose	True Positives	0	9	27	428
	False Positives	0	16	28	503
Bad Pose	True Negatives	0	5	5	439
	False Negatives	0	0	0	100

(E)

Trained Model (Replicate 1)		HKU4	MERS	SARS_CoV	SARS_CoV_2
Good Pose	True Positives	0	6	26	428
	False Positives	0	0	3	71
Bad Pose	True Negatives	0	21	30	871
	False Negatives	0	3	1	100

(F)



(G)

		Subpocket Engagement			
Ligand	PDB	S1	S1'	S2	S4
4N0	7P2G	Yes	No	Yes	No
5IL	7V1T	No	Yes	Yes	No
A1AA0	8VDJ	Yes	Yes	Yes	No
A1ALF	9BBQ	Yes	No	Yes	No
A1ALW	9BBP	Yes	No	Yes	No
A1BVZ	7I13	No	No	No	No
A1JGD	9RI1	Yes	No	Yes	No
ASC	7MPB	Yes	No	No	No
KMF	7GB1	No	Yes	Yes	No
M9U	7GDB	Yes	No	No	No
NZK	7GEZ	Yes	No	No	No
RV8	7AQJ	Yes	No	No	No
RWO	7GNE	Yes	No	Yes	No
RZS	5R82	No	No	Yes	No
U2I	8DII	Yes	No	No	No

Fig. 4.6: Outcome of Benchmarking ClassyPose and the Optimized SVM Model on the Testing Dataset. (A) Benchmark metrics used to compare the output from ClassyPose (Blue) and the optimized SVM model (Orange) on the testing dataset. 5 training-test runs

were performed for the optimized model, while 5 test runs were used for ClassyPose, and the averaged benchmark metrics are reported along with their corresponding standard deviations. 5 benchmark metrics were used: (i) PR-AUC (Precision-Recall Area Under Curve), (ii) ROC-AUC (Receiver Operating Characteristic Area Under Curve), (iii) Specificity, (iv) Recall, (v) Precision. **(B)** & **(C)** Graphical representation of the optimized model's classification and ClassyPose's classification, respectively. Only the first replicate is shown The assigned probability scores were plotted against the corresponding RMSD values of the poses. The plot was divided into 4 quadrants by splitting along the probability axis at 0.5 and along the RMDS axis at 2: (i) true negatives quadrant (probability <= 0.5 & RMSD > 2 Å), (ii) false positives quadrant (probability > 0.5 & RMSD > 2 Å), (iii) true positives quadrant (probability > 0.5 & RMSD <= 2), and (iv) false negatives quadrant (probability <= 0.5 & RMSD <= 2). **(D) & (E)** Strain's specific pose classification outcome from ClassyPose and the optimized model, respectively. ClassyPose could not classify any bad poses from HKU4, MERS, and SARS-CoV correctly. While ClassyPose could identify bad poses from SARS-CoV-2 correctly, it also output a greater amount of false positives. The optimization step with the training dataset addressed these issues with the optimized model. **(F)** 2D structures of ligands classified as false negatives only by the optimized model. The image was obtained from CoviProDigy (Chapter 3) and the ligands depicted are: 4N0 (7P2G), 5IL (7V1T), A1AA0 (8VDJ), A1ALF (9BBQ), A1ALW (9BBP), A1BVZ (7I13), A1JGD (9RI1), ASC (7MPB), KMF (7GB1), M9U (7GDB), NZK (7GEZ), RV8 (7AQJ), RWO (7GNE), RZS (5R82), and U2I (8DII). **(G)** SARS-CoV-2 Mpro subpocket occupancy characterization of the ligands classified as false negatives exclusively**.** The characterization was done using CoviProDigy (Chapter 3).

4.3.6 Benchmarking ClassyPose and Optimized SVM model on Hard Dataset

The final benchmark was performed on the "hard dataset", which contained ligands for which no good redocked poses were generated by the docking algorithm. This test was designed to evaluate model performance on the most challenging, unseen data and to compare it against the performance on the standard testing dataset.

When tasked with identifying the few "good" poses (crystal poses) within this challenging dataset, ClassyPose marginally outperformed the optimized model in its ability to find them, as shown by its slightly higher recall score (Fig. 4.7A). However, this came at a significant cost to precision. The predictions made by ClassyPose were accurate only about 10% of the time (average precision ≈ 0.1). In contrast, the optimized SVM model demonstrated a 6-fold higher accuracy, with an average precision score of approximately 0.6. This low precision severely impacted ClassyPose's overall balance, resulting in a significantly lower PR-AUC score, as it could not maintain prediction accuracy while being sensitive enough to retrieve all good poses (Fig. 4.7C).

Although the optimized model clearly outperformed ClassyPose, its own performance declined on the hard dataset compared to the testing dataset (Fig. 4.6A, 4.7A). To investigate this, a subpocket occupancy analysis was performed on the optimized model's true positive and false negative predictions. The analysis revealed a distinct difference in binding modes: (i) True Positives**:** Ligands correctly identified as "good" predominantly occupied two subpockets, primarily S1 and S2 (Fig. 4.8D); (ii) False Negatives: Ligands incorrectly rejected as "bad" primarily occupied one subpocket, primarily the S1' subpocket (Fig. 4.7E). This subpocket occupancy profile for the false negatives deviates significantly from the interaction patterns of the model's "good pose" support vectors (Fig. 4.5F). This suggests that the model failed to classify these ligands correctly because their interaction fingerprints represented a chemical space the model had not learned to associate with correct poses.

The models' abilities to correctly identify incorrect poses were also compared. ClassyPose continued to struggle, as highlighted by the symmetrical distribution of data points between the true negatives and false positives quadrants in Fig. 4.7C. The optimized SVM model, however, maintained its exceptional ability to reject bad poses. The data distribution for the optimized model was strongly skewed away from the false positives quadrant, demonstrating its robust performance. It achieved a high specificity score greater than 0.9, consistent with its performance on the standard testing dataset (Figs. 4.6A, 4.7A). In summary, the optimized SVM model once again outperformed ClassyPose as a binary classifier on this challenging dataset, achieving a strong ROC-AUC score of approximately

0.85. While this was slightly lower than its near-perfect score on the testing dataset, it confirms that the optimized model reliably distinguishes between correct and incorrect poses when faced with a challenging dataset.

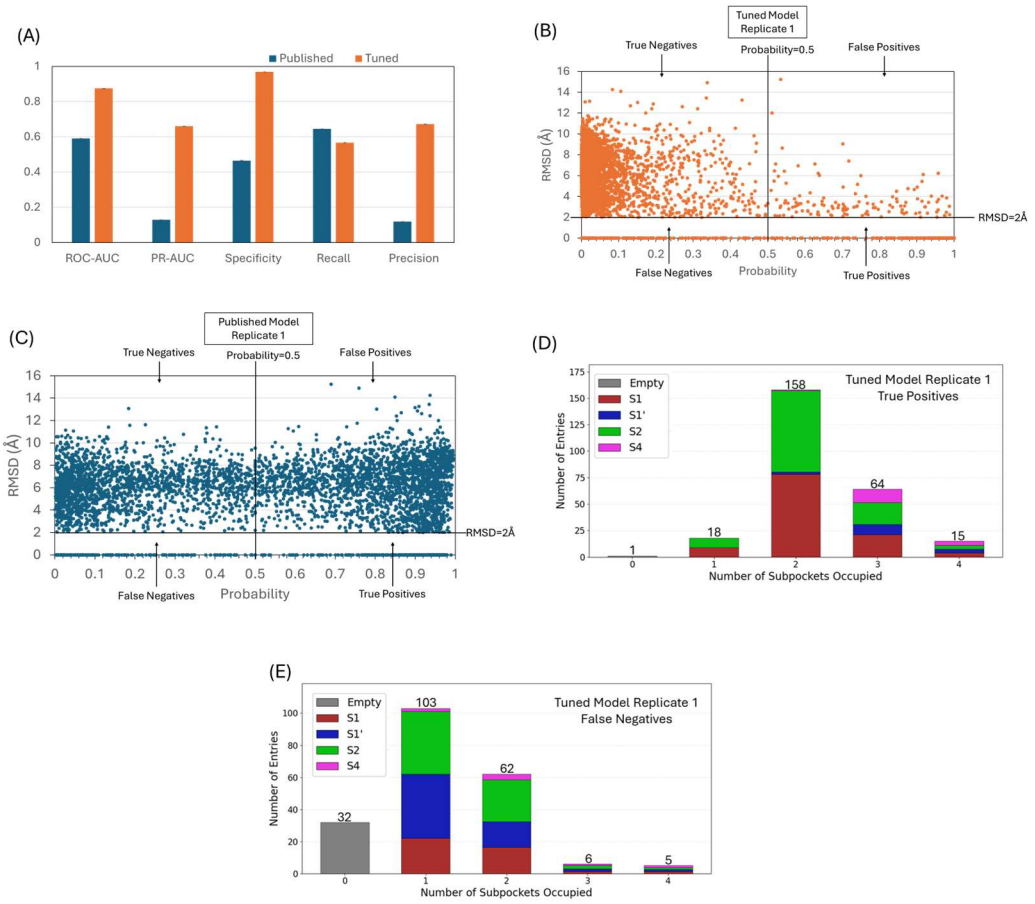


Fig. 4.7: Outcome of Benchmarking ClassyPose and The Optimized SVM Model on the Hard Dataset. (A) Benchmark metrics used to compare the output from ClassyPose (Blue) and the optimized SVM model (Orange) on the testing dataset. 5 training-test runs were performed for the optimized model while 5 test runs were used for ClassyPose, and the averaged benchmark metrics are reported along with their corresponding standard deviations. 5 benchmark metrics were used: (i) PR-AUC (Precision-Recall Area Under Curve), (ii) ROC-AUC (Receiver Operating Characteristic Area Under Curve), (iii) Specificity, (iv) Recall, (v) Precision. **(B)** & **(C)** Graphical representation of the optimized model's classification and ClassyPose's classification respectively. Only the first replicate is shown The assigned probability scores were plotted against the corresponding RMSD values of the

poses. The plot was divided into 4 quadrants by splitting along the probability axis at 0.5 and along the RMDS axis at 2: (i) true negatives quadrant (probability <= 0.5 & RMSD > 2 Å), (ii) false positives quadrant (probability > 0.5 & RMSD > 2 Å), (iii) true positives quadrant (probability > 0.5 & RMSD <= 2), and (iv) false negatives quadrant (probability <= 0.5 & RMSD <= 2). **(D) & (E)** Subpocket occupancy assessment for ligands labeled as true positives and false positives by the optimized model respectively. Each histogram bar is stratified to represent the involvement of the specific Mpro subpockets (grey for zero, red for S1, blue for S1', green for S2, and pink for S4).

4.3.7 Integration of the Optimized SVM Model in CoviProDigy

The optimized model was implemented successfully in CoviProDigy (Chapter 3) for quick pose quality assessment of user-uploaded structures (Fig. 4.8). Unlike ClassyPose, which requires the user to manually generate the fingerprint input for poses to be assessed and accessed via a command-line interface, CoviProDigy has automated this process and provides a user-intuitive interface for initiating the analysis and visualizing the output. Currently, only SARS-CoV-2 docked complexes can be assessed within CoviProDigy. Future work will include the expansion of CoviProDigy to additional betacoronaviral strains.

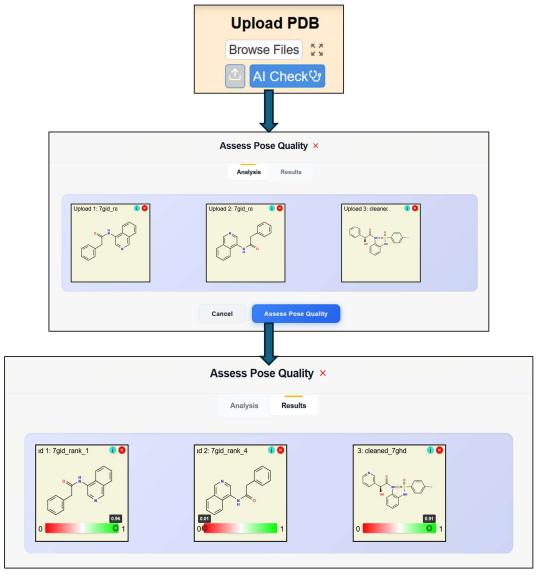


Fig. 4.8: Pose Quality Assessment in CoviProDigy. In CoviProDigy, the optimized SVM model is available under the "AI Check" feature in the "Upload PDB" widget. The interface

for using this feature is depicted. It has a 2 tab architecture, one for viewing the input complexes to be analyzed and another tab for the output of the computation. The probability assigned by the model to each complex is shown as a heatmap ranging from 0 (Bad Pose with high confidence) to 1 (Good pose with high confidence), along with a color scale conveying the same notion: Red (Bad), White (Uncertain Prediction), Green (Good pose).

4.4 Discussion

The accurate prediction of ligand binding poses is a persistent bottleneck in computational drug discovery, often limiting the efficiency of large-scale virtual screening campaigns. This was evident during the search for COVID-19 therapeutics, where the limitations of classical docking algorithms in prioritizing native-like poses for SARS-CoV-2 Mpro hindered progress.[134] While machine learning-based rescoring methods such as ClassyPose were developed to address this challenge, this work demonstrated that its pre-pandemic training data rendered it suboptimal for assessing poses for the *Betacoronavirus* Mpro family.

To address this, a new SVM model was trained exclusively on a curated pan-coronavirus Mpro dataset. The core of this approach relied on PLEC to encode the complex protein-ligand interaction environment into a machine-readable format. This strategy is well-suited for this target class due to the highly conserved architecture of the Mpro active site across the *Betacoronavirus* genus (Fig. 4.1). By learning the specific interaction patterns common to this family, the model could build a more relevant understanding of what constitutes a "good" pose. Indeed, the analysis of the model's support vectors confirmed that it had learned key binding features, such as the preferential occupancy of the S1 and S2 subsites, while still capturing interaction diversity from multiple *Betacoronavirus* strains. The resulting optimized model demonstrated a significant performance improvement. On both the standard testing and the challenging "hard" datasets, it maintained a high ROC-AUC score, indicating its ability to generalize reliably to unseen data. Crucially, it overcame the primary weakness of the original ClassyPose model: a low specificity and precision driven by a high false-positive rate. The optimized model, by contrast, excelled at correctly identifying both good and bad poses.

Despite its success, the model is not without limitations. The analysis of false negatives revealed a weakness in classifying ligands that occupy only a single subpocket, particularly the S1' subsite. This is likely because the training process did not explicitly differentiate ligands based on the number of subpockets they occupy, causing the model to be biased towards the more common, multi-subpocket binders. While this may not be a major drawback, as ligand potency often correlates with broader subpocket engagement, it highlights an area for future improvement. A subpocket-guided training protocol could enhance the model's applicability, creating an even more robust and versatile tool for the discovery of pan-coronavirus inhibitors.

4.5 Conclusion

The challenge of accurately prioritizing correct ligand binding poses remains a significant obstacle in virtual screening, a limitation observed during the extensive search for SARS-CoV-2 Mpro inhibitors. In this work, a dedicated computational pipeline was established to address this pose classification problem, specifically for the *Betacoronavirus* Mpro family. A GPU-accelerated docking protocol was first developed to enable large-scale screening, which was used to generate comprehensive datasets for training and optimizing a machine learning model. The resulting model demonstrated an improved ability to classify ligand poses for *Betacoronavirus* Mpro correctly. This complete pipeline represents a valuable resource to support the rational design of potent, pan-coronavirus inhibitors, aiding in the critical mission of preparing for future pandemic threats.

Chapter 5: Summary and Conclusion

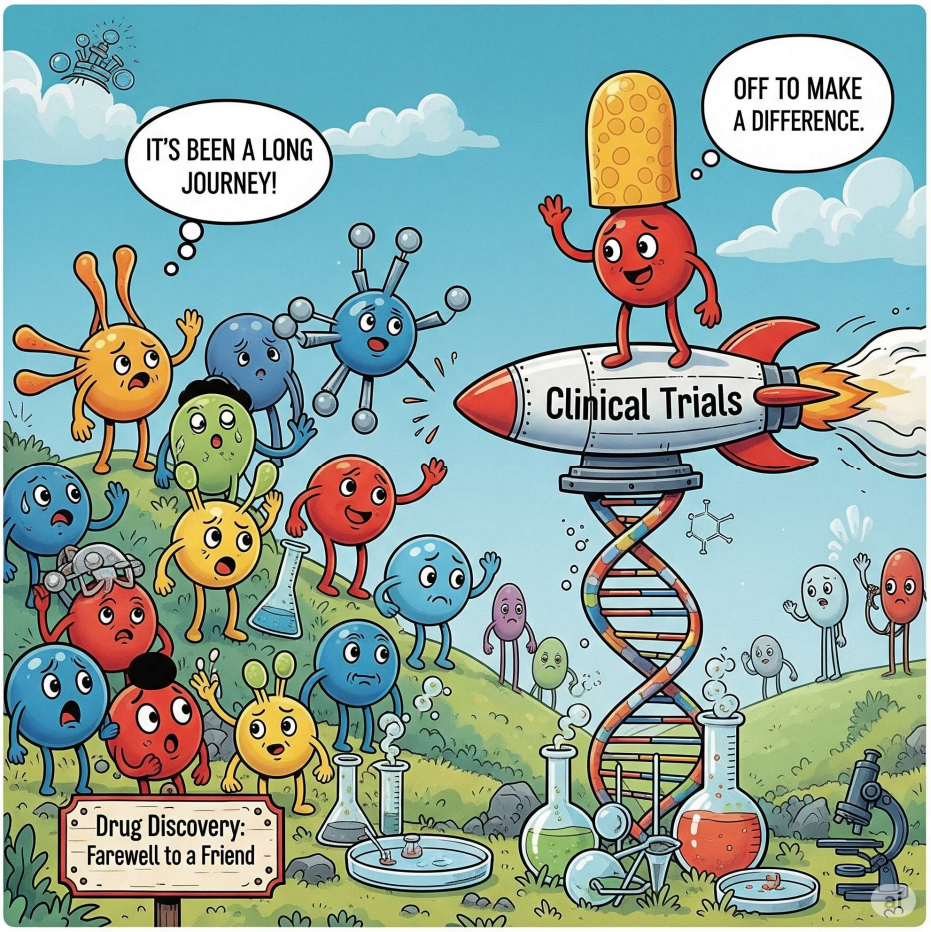


Image adapted from Gemini 2.5. This is a mock image to describe the content of this chapter, and it was not used for any scientific purposes.

5.1 Overview

The COVID-19 pandemic, caused by the *Betacoronavirus* SARS-CoV-2, exposed a vulnerability in global public health: a lack of readily available, broad-spectrum antiviral therapeutics. While vaccine development progressed at an unprecedented pace,[199] the need for effective treatments remained urgent. The viral Mpro, a cysteine protease essential for processing viral polyproteins and enabling replication, quickly emerged as a promising drug target.[27] Its function is critical for the maturation of the virus, and importantly, its active site architecture is highly conserved across the *Betacoronavirus* genus, including pathogens responsible for SARS and MERS. This conservation renders Mpro an ideal target for developing therapeutics that could be effective against both current and future coronavirus threats, specifically pan-coronavirus inhibitors.

5.2 The Gap in Current Drug Discovery Pipeline

The global scientific response produced an unparalleled volume of structural data, with over 1,000 crystal structures of SARS-CoV-2 Mpro in complex with various ligands deposited into the PDB. This wealth of information created a new challenge: a data-to-knowledge gap. Existing bioinformatics platforms are designed as archives for individual entries and lack the integrated tools for systematic, comparative analysis. Researchers could not easily extract insights across hundreds of complexes, especially at the subpocket level. This fragmentation represented a significant bottleneck, hindering the rational, knowledge-driven design of next-generation inhibitors.

Therefore, the central objective of this thesis was to address this critical gap by developing an integrated computational platform. The goal was to build a molecular toolkit that could not only manage and analyze the vast amount of Mpro structural data but also provide specialized tools to enhance the accuracy and efficiency of the drug discovery pipeline, thereby accelerating the search for potent pan-coronavirus therapeutics.

5.3 Key Achievements of this Thesis

This thesis has delivered a novel, integrated molecular toolkit designed to accelerate the discovery of Mpro inhibitors. The work is centered on 2 primary contributions: first, the development of CoviProDigy, a web-based platform for the comprehensive and comparative analysis of Mpro-ligand interactions; and second, the creation of a fine-tuned machine learning model for accurately classifying ligand binding poses, a critical step in virtual screening. The first major achievement, detailed in Chapter 3, was the design and implementation of CoviProDigy. This platform was built to overcome the limitations of existing databases by providing a suite of specialized tools for Mpro inhibitor analysis. CoviProDigy automates the mining of structural data, allowing for real-time calculation of interaction fingerprints, comparative analysis of subpocket occupancies, assessment of drug-like properties (Lipinski's Rule of 5), and interactive 2D and 3D visualization.

A key innovation within CoviProDigy is its algorithm for quantifying ligand subpocket occupancy, which provides a metric for comparing how different chemical scaffolds engage the S1, S1', S2, and S4 subpockets. Through its integrated analysis tools, the platform revealed critical trends across the dataset, such as the trade-off between increasing molecular weight to achieve multi-subpocket occupancy and maintaining drug-like properties. Furthermore, its powerful overlay mode enables the direct comparison of multiple ligands, facilitating the rapid generation of insights into conserved/unique interaction patterns. By transforming a static collection of PDB files into a dynamic and “queryable” knowledge base, CoviProDigy provides a powerful resource for guiding medicinal chemistry efforts like scaffold hopping and lead optimization.

The second key achievement, presented in Chapter 4, addressed a fundamental bottleneck in structure-based virtual screening: the failure of classical docking algorithms to reliably distinguish between correct (near-native) and incorrect ligand binding poses.[134] While universal machine learning models like ClassyPose[79] exist, we demonstrated that its performance was suboptimal for the *Betacoronavirus* Mpro family, suffering from a high false-positive rate. To solve this, we developed and trained a new SVM model[195] exclusively

on a curated dataset of pan-coronavirus Mpro structures (SARS-CoV-2, SARS-CoV, MERS-CoV, and HKU4). By training on PLEC which encodes the specific interaction patterns of this protein family, the model learned a target-specific definition of a "good" pose. The resulting optimized model demonstrated a significant performance improvement, effectively eliminating the high false-positive rate of the generic model. Moreover, the model showed an ability to generalize its predictions on a hard dataset, which was not included in its training set, indicating its applicability to “unseen” data. This validated ML model provides a rapid and reliable tool for filtering virtual screening results, ensuring that computational resources are focused on the most promising candidates.

5.4 Overall Significance

Collectively, these contributions provide a computational framework that addresses key challenges in modern drug discovery. The thesis delivers a comprehensive toolkit where CoviProDigy serves as the engine for knowledge extraction and hypothesis generation from existing structural data, while the fine-tuned ML model provides the validation step for new data generated from virtual screening. This work provides a robust platform to accelerate the knowledge-driven design of potent inhibitors for SARS-CoV-2 and, more broadly, contributes directly to the vital mission of pandemic preparedness by enabling more efficient search for pan-coronavirus therapeutics.

5.5 Case Study: Investigating the Binding Mode of a Novel Molecule Docked against SARS-CoV-2 Mpro

A novel molecule was created by fusing the moieties from X77-analogues, identified using CoviProDigy. This molecule was docked into the active site of SARS-CoV-2 Mpro (PDB: 7RFS) using MOE[200] and 5 docked poses were generated. In this case study, the docked poses were assessed using CoviProDigy and the optimized model collectively to identify the correct docked pose as well as extracting additional insights from the dedicated Mpro toolkit available in CoviProDigy. This case study highlights the complementarity of the toolkits offered by CoviProDigy and the optimized model in assessing the binding modes of ligands.

These structures were loaded into CoviProDigy via the file upload feature, and the pose quality was assessed using the AI check feature (Chapter 3) which is an implementation of the ML model optimized for *Betacoronavirus* Mpro. The third ranked pose (pose 3) was scored the highest with a probability score of 0.72 (Fig. 5.1A), suggesting it might be a good docking conformation. Conversely, the top ranked pose (pose 1) was assigned a probability score of 0.27, suggesting that it might be a bad docking conformation. The 2D interaction diagrams of both ligands indicated that the top ranked pose showed greater engagement with the Mpro subsite (Fig. 5.1B). It was involved in 4 hydrogen bonds, 1 stacking interaction, and hydrophobic contacts with Mpro while the third ranked pose was involved in 2 hydrogen bonding interactions and hydrophobic contacts with the active site (Fig. 5.1B-C). This might explain why the scoring function of MOE ranked the poses as such.

To get a better understanding of the binding mode of this novel molecule, a similarity search was performed, retrieving X77 and X77-analogues (Appendix: Fig. A6, A7). Their overlaid 2D interactions revealed a highly conserved non-bonded interaction with Glu166 from the S1 subsite (Fig. 1.3C, Fig. 5.1D). Moreover, their overlaid structures revealed a consistent orientation of their pyridine ring in the S1 subsite (Fig. 5.1E). Both poses of the novel molecule did maintain this interaction but with different ligand moiety involvement. In pose 1, Glu166 was in hydrophobic contact with a cyclic alkane group (Fig. 5.1F) whereas in pose 3, Glu166 was in contact with the pyridine ring (Fig. 5.1G).

Since the pyridine rings of X77 and its analogues were mostly oriented in the S1 subsite, this explained why pose 3 was assigned a higher probability than pose 1 by the optimized model. In Fig. 5.1F, the cyclo-alkane group of pose 1 is oriented towards the S1 subsite, whereas the pyridine ring of pose 3 is oriented towards the S1 subsite (Fig. 5.1G). Moreover, pose 1 did not meet the threshold to be classified as an occupant of the S1 subsite while pose 3, however, was classified as an occupant (Fig. 5.1I). Pose 1 did not maintain the conserved binding mode observed across similar crystal poses (Fig. 5.1E, F, I), so that it was classified as bad.

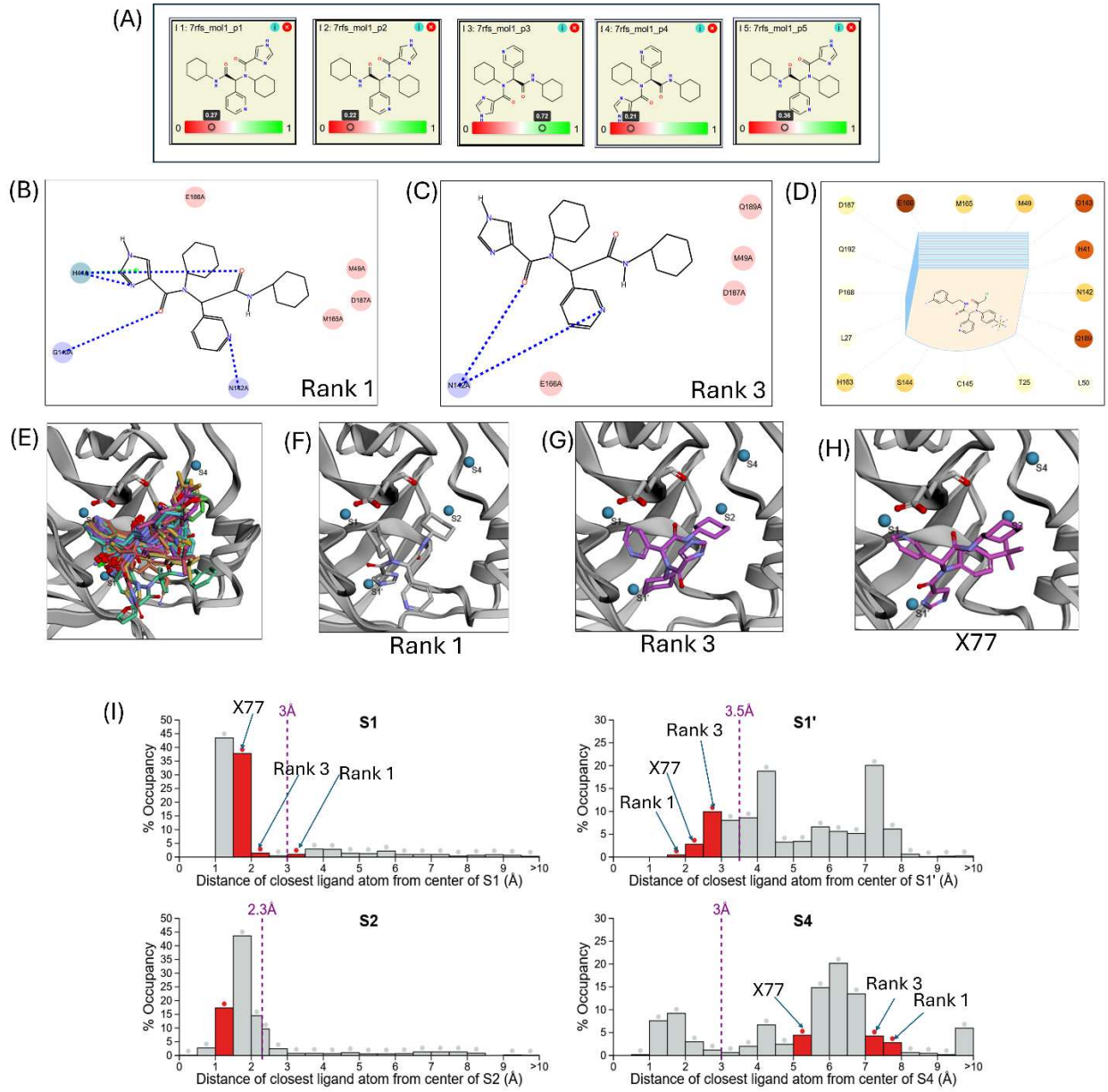


Fig. 5.1: Characterizing the Binding Mode of the Docked Poses of the Novel Fused Molecule and Novel Fused Molecule Analogues. (A) Classification of the docked poses by the optimized ML model. The assigned probability scores were (in ascending order of Docked Pose ranking): 0.27, 0.22, 0.72, 0.21, and 0.38. **(B)–(C)** 2D interaction diagram of pose 1 (ranked 1st) and pose 3 (ranked 3rd), respectively, showing hydrogen bonds (blue), residues in hydrophobic contact (red), and aromatic stacking (green). **(D)** Overlaid 2D interaction of all Novel Fused Molecule Analogues, with all interacting residues color-coded according to the

local conservation of their respective interactions across the local dataset. The higher the conservation, the darker the color of the residue node. **(E)** Structural overlap of all the analogues across a single Mpro binding site. The ligands are styled in stick representations and the geometric centers for each subpocket are shown as dummy atoms. Glu166 is shown in stick representation. **(F)-(H)** Subpocket fit of pose 1, pose 3 and X77 respectively. The ligands are styled in stick representations and the geometric centers for each subpocket are shown as dummy atoms. Glu166 is shown in stick representation. **(I)** Comparing the subpocket occupancies of pose 1, pose 3, and X77. All computations were done using CoviProDigy and all images were adapted from its widgets.

5.6 Current Limitations

While this thesis delivers novel computational toolkits for Mpro inhibitor design, it is important to acknowledge the boundaries of the current work and outline the promising avenues for future research that these limitations reveal. The primary contributions of this thesis have inherent limitations that provide clear directions for enhancement. First, the CoviProDigy platform, while powerful, is currently focused exclusively on SARS-CoV-2 Mpro. It does not yet incorporate structural data from other critical betacoronaviruses such as SARS-CoV and MERS-CoV, limiting its direct application for pan-inhibitor analysis.

Furthermore, the database currently lacks integrated experimental activity data (e.g., IC50, Ki), preventing a direct correlation between structural interaction patterns and inhibitor potency. Moreover, CoviProDigy does not offer any toolkit for designing novel fusion molecules using fragments from the curated ligand dataset, which could assist with combinatorial ligand design. Lastly, CoviProDigy is optimized for exploring orthosteric ligands. However, other druggable sites are included in Mpro, notably allosteric sites around the dimer interface, which are targeted to inhibit Mpro dimerization, inactivating the enzyme Mpro.[201] Drug candidates have been proposed to prevent Mpro dimerization,[202] but they cannot be explored in CoviProDigy, to leverage its tailored Mpro toolkit, in its current architecture.

Lastly, the fine-tuned ML model, despite its high accuracy, exhibited a specific weakness. The analysis of its false negatives revealed difficulty in correctly classifying ligands that adopt less common binding modes, particularly those that occupy only a single subpocket (S1'). The model's training may have been biased towards the more prevalent multi-subpocket binders, making it less sensitive to the minority, which still represents potentially valuable interaction profiles. Moreover, the active oligomeric state of Mpro was not sustained across the entire training dataset, and the model might not consider the N-finger[44] involvement in dimeric complexes during classification.

5.7 Future Directions

The immediate next step is to expand the CoviProDigy database beyond its current focus on SARS-CoV-2 to support other betacoronaviruses. This would transform it into a true pan-coronavirus analysis tool, allowing for cross-species comparative analysis to identify conserved and unique binding features essential for designing broad-spectrum inhibitors. Furthermore, the knowledge base of CoviProDigy will integrate the experimental activity data of Mpro ligands reported in the literature. This will empower users to correlate ligand binding mode and subpocket engagement with inhibitor potency. This can give useful insights into the medicinal chemistry optimization of ligands.

Another key future optimization is to broaden the druggable landscape of CoviProDigy so that other reported druggable sites of Mpro are considered by its drug discovery toolkit, eliminating any bias towards any given Mpro site. This will give users the flexibility to collect insights from any site for guiding rational drug design. AI approaches for predicting potential binding sites[127–129] will also be implemented, allowing users to explore Mpro from the perspectives of its binding sites rather than solely from the context of different ligands. Moreover, a fragment library will be implemented to support combinatorial drug design, allowing users to explore a customized chemical space in their ligand exploration analyses of Mpro.

Furthermore, the training protocol of the ML model will be adapted to avoid any subpocket-induced bias. An approach to address this will include the implementation of a subpocket-

guided supervised learning protocol whereby each subpocket is treated independently. In this training scheme, the model will be trained to recognize the distinct interaction fingerprints associated with each subpocket (S1, S1', S2, and S4) independently. This will allow it to learn the key features of stable binding within each individual subpocket, preventing it from being biased by the more prevalent multi-subpocket binders. Moreover, the pan coronavirus dataset will be updated using dimeric ligand bound Mpro structures only. This will be used in conjunction with the subpocket-guided training protocol so that the resulting ML model will offer a precise understanding of ligand engagement, making it a reliable tool in virtual screening campaigns involving betacoronaviuses Mpro.

5.8 Concluding Remarks

This thesis began by identifying a major bottleneck in the post-COVID-19 pandemic era: the inability to translate efficiently the massive volume of structural data into actionable knowledge for drug discovery. CoviProDigy was built to address this problem, empowering researchers to extract subpocket-based Mpro insights to guide medicinal chemistry optimizations for designing the next generation of SARS-CoV-2 inhibitors. CoviProDigy will be expanded to support betacoronaviruses Mpro as input so that it can be leveraged for the pan-inhibitor development against *Betacoronavirus* Mpro.

Another key challenge addressed by this work was the accurate classification of ligand docked poses. To this end, a GPU-based docking pipeline was created to generate data for the optimization of an existing pose classification model. This model was trained specifically for the classification of *Betacoronavirus* Mpro ligand poses, with high reliability. This ML model was implemented in CoviProDigy and collectively, they offer a strong foundation for developing a platform capable of supporting screening campaigns for the search of broad-spectrum *Betacoronavirus* Mpro inhibitors.

References

(1) Weiss, S. R.; Leibowitz, J. L. Chapter 4 - Coronavirus Pathogenesis. In *Advances in Virus Research*; Maramorosch, K., Shatkin, A. J., Murphy, F. A., Eds.; Academic Press, 2011; Vol. 81, pp 85–164. https://doi.org/10.1016/B978-0-12-385885-6.00009-2.

(2) Tyrrell, D. a. J.; Bynoe, M. L. Cultivation of a Novel Type of Common-Cold Virus in Organ Cultures. *Br Med J* **1965**, *1* (5448), 1467–1470. https://doi.org/10.1136/bmj.1.5448.1467.

(3) Guan, Y.; Zheng, B. J.; He, Y. Q.; Liu, X. L.; Zhuang, Z. X.; Cheung, C. L.; Luo, S. W.; Li, P. H.; Zhang, L. J.; Guan, Y. J.; Butt, K. M.; Wong, K. L.; Chan, K. W.; Lim, W.; Shortridge, K. F.; Yuen, K. Y.; Peiris, J. S. M.; Poon, L. L. M. Isolation and Characterization of Viruses Related to the SARS Coronavirus from Animals in Southern China. *Science* **2003**, *302* (5643), 276–278. https://doi.org/10.1126/science.1087139.

(4) Azhar, E. I.; El-Kafrawy, S. A.; Farraj, S. A.; Hassan, A. M.; Al-Saeed, M. S.; Hashem, A. M.; Madani, T. A. Evidence for Camel-to-Human Transmission of MERS Coronavirus. *N. Engl. J. Med.* **2014**, *370* (26), 2499–2505. https://doi.org/10.1056/NEJMoa1401505.

(5) 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.; Shi, Z.-L. A Pneumonia Outbreak Associated with a New Coronavirus of Probable Bat Origin. *Nature* **2020**, *579* (7798), 270–273. https://doi.org/10.1038/s41586-020-2012-7.

(6) Nicola, M.; Alsafi, Z.; Sohrabi, C.; Kerwan, A.; Al-Jabir, A.; Iosifidis, C.; Agha, M.; Agha, R. The Socio-Economic Implications of the Coronavirus Pandemic (COVID-19): A Review. *Int. J. Surg.* **2020**, *78*, 185–193. https://doi.org/10.1016/j.ijsu.2020.04.018.

(7) 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.; Zhu, Y. An Oral SARS-CoV-2 Mpro Inhibitor Clinical Candidate for the Treatment of COVID-19. *Science* **2021**, *374* (6575), 1586–1593. https://doi.org/10.1126/science.abl4784.

(8) CDC. *CDC Museum COVID-19 Timeline*. Centers for Disease Control and Prevention. https://www.cdc.gov/museum/timeline/covid19.html (accessed 2023-11-30).

(9) *Updated working definitions and primary actions for SARSCoV2 variants*. https://www.who.int/publications/m/item/updated-working-definitions-and-primary-actions-for--sars-cov-2-variants (accessed 2023-12-01).

(10) Focosi, D.; Spezia, P. G.; Maggi, F. SARS-CoV-2 BA.2.86 (“Pirola”): Is It Pi or Just Another Omicron Sublineage? *Vaccines* **2023**, *11* (11), 1634. https://doi.org/10.3390/vaccines11111634.

(11) Lu, Y.; Ao, D.; He, X.; Wei, X. The Rising SARS-CoV-2 JN.1 Variant: Evolution, Infectivity, Immune Escape, and Response Strategies. *MedComm* **2024**, *5* (8), e675. https://doi.org/10.1002/mco2.675.

(12) Tao, K.; Tzou, P. L.; Nouhin, J.; Gupta, R. K.; de Oliveira, T.; Kosakovsky Pond, S. L.; Fera, D.; Shafer, R. W. The Biological and Clinical Significance of Emerging SARS-CoV-2 Variants. *Nat. Rev. Genet.* **2021**, *22* (12), 757–773. https://doi.org/10.1038/s41576-021-00408-x.

(13) Wu, C.; Yin, W.; Jiang, Y.; Xu, H. E. Structure Genomics of SARS-CoV-2 and Its Omicron Variant: Drug Design Templates for COVID-19. *Acta Pharmacol. Sin.* **2022**, *43* (12), 3021–3033. https://doi.org/10.1038/s41401-021-00851-w.
(14) Chaudhary, N.; Weissman, D.; Whitehead, K. A. mRNA Vaccines for Infectious Diseases: Principles, Delivery and Clinical Translation. *Nat. Rev. Drug Discov.* **2021**, *20* (11), 817–838. https://doi.org/10.1038/s41573-021-00283-5.
(15) Townsend, J. P.; Hassler, H. B.; Lamb, A. D.; Sah, P.; Alvarez Nishio, A.; Nguyen, C.; Tew, A. D.; Galvani, A. P.; Dornburg, A. Seasonality of Endemic COVID-19. *mBio 14* (6), e01426-23. https://doi.org/10.1128/mbio.01426-23.
(16) Naughton, C. C.; Roman, F. A.; Alvarado, A. G. F.; Tariqi, A. Q.; Deeming, M. A.; Kadonsky, K. F.; Bibby, K.; Bivins, A.; Medema, G.; Ahmed, W.; Katsivelis, P.; Allan, V.; Sinclair, R.; Rose, J. B. Show Us the Data: Global COVID-19 Wastewater Monitoring Efforts, Equity, and Gaps. *FEMS Microbes* **2023**, *4*, xtad003. https://doi.org/10.1093/femsmc/xtad003.
(17) Carabelli, A. M.; Peacock, T. P.; Thorne, L. G.; Harvey, W. T.; Hughes, J.; de Silva, T. I.; Peacock, S. J.; Barclay, W. S.; de Silva, T. I.; Towers, G. J.; Robertson, D. L. SARS-CoV-2 Variant Biology: Immune Escape, Transmission and Fitness. *Nat. Rev. Microbiol.* **2023**, *21* (3), 162–177. https://doi.org/10.1038/s41579-022-00841-7.
(18) *WHO launches new initiative to improve pandemic preparedness*. https://www.who.int/news/item/26-04-2023-who-launches-new-initiative-to-improve-pandemic-preparedness (accessed 2025-07-17).
(19) *COVID - Coronavirus Statistics - Worldometer*. https://www.worldometers.info/coronavirus/ (accessed 2024-03-12).
(20) Mathieu, E.; Ritchie, H.; Rodés-Guirao, L.; Appel, C.; Giattino, C.; Hasell, J.; Macdonald, B.; Dattani, S.; Beltekian, D.; Ortiz-Ospina, E.; Roser, M. Coronavirus Pandemic (COVID-19). *Our World Data* **2020**.
(21) Re3data.Org. GISAID. re3data.org - Registry of Research Data Repositories 2012. https://doi.org/10.17616/R3Q59F.
(22) Brant, A. C.; Tian, W.; Majerciak, V.; Yang, W.; Zheng, Z.-M. SARS-CoV-2: From Its Discovery to Genome Structure, Transcription, and Replication. *Cell Biosci.* **2021**, *11*, 136. https://doi.org/10.1186/s13578-021-00643-z.
(23) Wang, W.-H.; Lin, C.-Y.; Chang, M. R.; Urbina, A. N.; Assavalapsakul, W.; Thitithanyanont, A.; Chen, Y.-H.; Liu, F.-T.; Wang, S.-F. The Role of Galectins in Virus Infection - A Systemic Literature Review. *J. Microbiol. Immunol. Infect.* **2020**, *53* (6), 925–935. https://doi.org/10.1016/j.jmii.2019.09.005.
(24) Pablos, I.; Machado, Y.; de Jesus, H. C. R.; Mohamud, Y.; Kappelhoff, R.; Lindskog, C.; Vlok, M.; Bell, P. A.; Butler, G. S.; Grin, P. M.; Cao, Q. T.; Nguyen, J. P.; Solis, N.; Abbina, S.; Rut, W.; Vederas, J. C.; Szekely, L.; Szakos, A.; Drag, M.; Kizhakkedathu, J. N.; Mossman, K.; Hirota, J. A.; Jan, E.; Luo, H.; Banerjee, A.; Overall, C. M. Mechanistic Insights into COVID-19 by Global Analysis of the SARS-CoV-2 3CLpro Substrate Degradome. *Cell Rep.* **2021**, *37* (4), 109892. https://doi.org/10.1016/j.celrep.2021.109892.
(25) Cheng, N.; Liu, M.; Li, W.; Sun, B.; Liu, D.; Wang, G.; Shi, J.; Li, L. Protein Post-Translational Modification in SARS-CoV-2 and Host Interaction. *Front. Immunol.* **2023**, *13*, 1068449. https://doi.org/10.3389/fimmu.2022.1068449.
(26) V'kovski, P.; Kratzel, A.; Steiner, S.; Stalder, H.; Thiel, V. Coronavirus Biology and Replication: Implications for SARS-CoV-2. *Nat. Rev. Microbiol.* **2021**, *19* (3), 155–170. https://doi.org/10.1038/s41579-020-00468-6.

(27) Li, G.; Hilgenfeld, R.; Whitley, R.; De Clercq, E. Therapeutic Strategies for COVID-19: Progress and Lessons Learned. *Nat. Rev. Drug Discov.* **2023**, *22* (6), 449–475. https://doi.org/10.1038/s41573-023-00672-y.
(28) Zhang, H.; Penninger, J. M.; Li, Y.; Zhong, N.; Slutsky, A. S. Angiotensin-Converting Enzyme 2 (ACE2) as a SARS-CoV-2 Receptor: Molecular Mechanisms and Potential Therapeutic Target. *Intensive Care Med.* **2020**, *46* (4), 586–590. https://doi.org/10.1007/s00134-020-05985-9.
(29) Fam, B. S. O.; Vargas-Pinilla, P.; Amorim, C. E. G.; Sortica, V. A.; Bortolini, M. C. ACE2 Diversity in Placental Mammals Reveals the Evolutionary Strategy of SARS-CoV-2. *Genet. Mol. Biol. 43* (2), e20200104. https://doi.org/10.1590/1678-4685-GMB-2020-0104.
(30) Malone, B.; Urakova, N.; Snijder, E. J.; Campbell, E. A. Structures and Functions of Coronavirus Replication–Transcription Complexes and Their Relevance for SARS-CoV-2 Drug Design. *Nat. Rev. Mol. Cell Biol.* **2022**, *23* (1), 21. https://doi.org/10.1038/s41580-021-00432-z.
(31) Wrapp, D.; Wang, N.; Corbett, K. S.; Goldsmith, J. A.; Hsieh, C.-L.; Abiona, O.; Graham, B. S.; McLellan, J. S. Cryo-EM Structure of the 2019-nCoV Spike in the Prefusion Conformation. *Science* **2020**, *367* (6483), 1260–1263. https://doi.org/10.1126/science.abb2507.
(32) Walls, A. C.; Park, Y.-J.; Tortorici, M. A.; Wall, A.; McGuire, A. T.; Veesler, D. Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein. *Cell* **2020**, *181* (2), 281-292.e6. https://doi.org/10.1016/j.cell.2020.02.058.
(33) Li, M.; Wang, H.; Tian, L.; Pang, Z.; Yang, Q.; Huang, T.; Fan, J.; Song, L.; Tong, Y.; Fan, H. COVID-19 Vaccine Development: Milestones, Lessons and Prospects. *Signal Transduct. Target. Ther.* **2022**, *7* (1), 1–32. https://doi.org/10.1038/s41392-022-00996-y.
(34) Chakraborty, C.; Bhattacharya, M.; Sharma, A. R. Present Variants of Concern and Variants of Interest of Severe Acute Respiratory Syndrome Coronavirus 2: Their Significant Mutations in S-Glycoprotein, Infectivity, Re-Infectivity, Immune Escape and Vaccines Activity. *Rev. Med. Virol.* **2022**, *32* (2), e2270. https://doi.org/10.1002/rmv.2270.
(35) Zhang, L.; Lin, D.; Sun, X.; Curth, U.; Drosten, C.; Sauerhering, L.; Becker, S.; Rox, K.; Hilgenfeld, R. Crystal Structure of SARS-CoV-2 Main Protease Provides a Basis for Design of Improved α-Ketoamide Inhibitors. *Science* **2020**, *368* (6489), 409–412. https://doi.org/10.1126/science.abb3405.
(36) Zhang, L.; Lin, D.; Kusov, Y.; Nian, Y.; Ma, Q.; Wang, J.; von Brunn, A.; Leyssen, P.; Lanko, K.; Neyts, J.; de Wilde, A.; Snijder, E. J.; Liu, H.; Hilgenfeld, R. α-Ketoamides as Broad-Spectrum Inhibitors of Coronavirus and Enterovirus Replication: Structure-Based Design, Synthesis, and Activity Assessment. *J. Med. Chem.* **2020**, *63* (9), 4562–4578. https://doi.org/10.1021/acs.jmedchem.9b01828.
(37) Unoh, Y.; Uehara, S.; Nakahara, K.; Nobori, H.; Yamatsu, Y.; Yamamoto, S.; Maruyama, Y.; Taoda, Y.; Kasamatsu, K.; Suto, T.; Kouki, K.; Nakahashi, A.; Kawashima, S.; Sanaki, T.; Toba, S.; Uemura, K.; Mizutare, T.; Ando, S.; Sasaki, M.; Orba, Y.; Sawa, H.; Sato, A.; Sato, T.; Kato, T.; Tachibana, Y. Discovery of S-217622, a Noncovalent Oral SARS-CoV-2 3CL Protease Inhibitor Clinical Candidate for Treating COVID-19. *J. Med. Chem.* **2022**, *65* (9), 6499–6512. https://doi.org/10.1021/acs.jmedchem.2c00117.
(38) Dai, W.; Jochmans, D.; Xie, H.; Yang, H.; Li, J.; Su, H.; Chang, D.; Wang, J.; Peng, J.; Zhu, L.; Nian, Y.; Hilgenfeld, R.; Jiang, H.; Chen, K.; Zhang, L.; Xu, Y.; Neyts, J.; Liu, H. Design, Synthesis, and Biological Evaluation of Peptidomimetic Aldehydes as Broad-Spectrum Inhibitors against Enterovirus and SARS-CoV-2. *J. Med. Chem.* **2022**, *65* (4), 2794–2808. https://doi.org/10.1021/acs.jmedchem.0c02258.
(39) Ma, C.; Sacco, M. D.; Hurst, B.; Townsend, J. A.; Hu, Y.; Szeto, T.; Zhang, X.; Tarbet, B.; Marty, M. T.; Chen, Y.; Wang, J. Boceprevir, GC-376, and Calpain Inhibitors II, XII Inhibit SARS-CoV-2 Viral

Replication by Targeting the Viral Main Protease. *Cell Res.* **2020**, *30* (8), 678–692. https://doi.org/10.1038/s41422-020-0356-z.
(40) 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.; Yang, H. Structure of Mpro from SARS-CoV-2 and Discovery of Its Inhibitors. *Nature* **2020**, *582* (7811), 289–293. https://doi.org/10.1038/s41586-020-2223-y.
(41) Ullah, A.; Ullah, K. Inhibition of SARS-CoV-2 3CL Mpro by Natural and Synthetic Inhibitors: Potential Implication for Vaccine Production Against COVID-19. *Front. Mol. Biosci.* **2021**, *8*.
(42) Zhao, Y.; Zhu, Y.; Liu, X.; Jin, Z.; Duan, Y.; Zhang, Q.; Wu, C.; Feng, L.; Du, X.; Zhao, J.; Shao, M.; Zhang, B.; Yang, X.; Wu, L.; Ji, X.; Guddat, L. W.; Yang, K.; Rao, Z.; Yang, H. Structural Basis for Replicase Polyprotein Cleavage and Substrate Specificity of Main Protease from SARS-CoV-2. *Proc. Natl. Acad. Sci.* **2022**, *119* (16), e2117142119. https://doi.org/10.1073/pnas.2117142119.
(43) Weng, Y. L.; Naik, S. R.; Dingelstad, N.; Lugo, M. R.; Kalyaanamoorthy, S.; Ganesan, A. Molecular Dynamics and in Silico Mutagenesis on the Reversible Inhibitor-Bound SARS-CoV-2 Main Protease Complexes Reveal the Role of Lateral Pocket in Enhancing the Ligand Affinity. *Sci. Rep.* **2021**, *11* (1), 7429. https://doi.org/10.1038/s41598-021-86471-0.
(44) Arutyunova, E.; Khan, M. B.; Fischer, C.; Lu, J.; Lamer, T.; Vuong, W.; van Belkum, M. J.; McKay, R. T.; Tyrrell, D. L.; Vederas, J. C.; Young, H. S.; Lemieux, M. J. N-Terminal Finger Stabilizes the S1 Pocket for the Reversible Feline Drug GC376 in the SARS-CoV-2 Mpro Dimer. *J. Mol. Biol.* **2021**, *433* (13), 167003. https://doi.org/10.1016/j.jmb.2021.167003.
(45) Fornasier, E.; Macchia, M. L.; Giachin, G.; Sosic, A.; Pavan, M.; Sturlese, M.; Salata, C.; Moro, S.; Gatto, B.; Bellanda, M.; Battistutta, R. A New Inactive Conformation of SARS-CoV-2 Main Protease. *Acta Crystallogr. Sect. Struct. Biol.* **2022**, *78* (Pt 3), 363–378. https://doi.org/10.1107/S2059798322000948.
(46) Ramos-Guzmán, C. A.; Ruiz-Pernía, J. J.; Tuñón, I. Unraveling the SARS-CoV-2 Main Protease Mechanism Using Multiscale Methods. *ACS Catal.* **2020**, *10* (21), 12544–12554. https://doi.org/10.1021/acscatal.0c03420.
(47) Pettersen, E. F.; Goddard, T. D.; Huang, C. C.; Meng, E. C.; Couch, G. S.; Croll, T. I.; Morris, J. H.; Ferrin, T. E. UCSF ChimeraX: Structure Visualization for Researchers, Educators, and Developers. *Protein Sci. Publ. Protein Soc.* **2021**, *30* (1), 70–82. https://doi.org/10.1002/pro.3943.
(48) Odhar, H. A.; Ahjel, S. W.; Albeer, A. A. M. A.; Hashim, A. F.; Rayshan, A. M.; Humadi, S. S. Molecular Docking and Dynamics Simulation of FDA Approved Drugs with the Main Protease from 2019 Novel Coronavirus. *Bioinformation* **2020**, *16* (3), 236–244. https://doi.org/10.6026/97320630016236.
(49) Gentile, D.; Patamia, V.; Scala, A.; Sciortino, M. T.; Piperno, A.; Rescifina, A. Putative Inhibitors of SARS-CoV-2 Main Protease from A Library of Marine Natural Products: A Virtual Screening and Molecular Modeling Study. *Mar. Drugs* **2020**, *18* (4), 225. https://doi.org/10.3390/md18040225.
(50) Ullrich, S.; Nitsche, C. The SARS-CoV-2 Main Protease as Drug Target. *Bioorg. Med. Chem. Lett.* **2020**, *30* (17), 127377. https://doi.org/10.1016/j.bmcl.2020.127377.
(51) Padhi, A. K.; Shukla, R.; Saudagar, P.; Tripathi, T. High-Throughput Rational Design of the Remdesivir Binding Site in the RdRp of SARS-CoV-2: Implications for Potential Resistance. *iScience* **2021**, *24* (1), 101992. https://doi.org/10.1016/j.isci.2020.101992.

(52) Nukoolkarn, V.; Lee, V. S.; Malaisree, M.; Aruksakulwong, O.; Hannongbua, S. Molecular Dynamic Simulations Analysis of Ritronavir and Lopinavir as SARS-CoV 3CLpro Inhibitors. *J. Theor. Biol.* **2008**, *254* (4), 861–867. https://doi.org/10.1016/j.jtbi.2008.07.030.
(53) Huynh, T.; Wang, H.; Luan, B. In Silico Exploration of the Molecular Mechanism of Clinically Oriented Drugs for Possibly Inhibiting SARS-CoV-2's Main Protease. *J. Phys. Chem. Lett.* **2020**, *11* (11), 4413–4420. https://doi.org/10.1021/acs.jpclett.0c00994.
(54) Wang, Z.; Yang, L. Turning the Tide: Natural Products and Natural-Product-Inspired Chemicals as Potential Counters to SARS-CoV-2 Infection. *Front. Pharmacol.* **2020**, *11*. https://doi.org/10.3389/fphar.2020.01013.
(55) Stille, J. K.; Tjutrins, J.; Wang, G.; Venegas, F. A.; Hennecker, C.; Rueda, A. M.; Sharon, I.; Blaine, N.; Miron, C. E.; Pinus, S.; Labarre, A.; Plescia, J.; Burai Patrascu, M.; Zhang, X.; Wahba, A. S.; Vlaho, D.; Huot, M. J.; Schmeing, T. M.; Mittermaier, A. K.; Moitessier, N. Design, Synthesis and in Vitro Evaluation of Novel SARS-CoV-2 3CLpro Covalent Inhibitors. *Eur. J. Med. Chem.* **2022**, *229*, 114046. https://doi.org/10.1016/j.ejmech.2021.114046.
(56) Luttens, A.; Gullberg, H.; Abdurakhmanov, E.; Vo, D. D.; Akaberi, D.; Talibov, V. O.; Nekhotiaeva, N.; Vangeel, L.; De Jonghe, S.; Jochmans, D.; Krambrich, J.; Tas, A.; Lundgren, B.; Gravenfors, Y.; Craig, A. J.; Atilaw, Y.; Sandström, A.; Moodie, L. W. K.; Lundkvist, Å.; van Hemert, M. J.; Neyts, J.; Lennerstrand, J.; Kihlberg, J.; Sandberg, K.; Danielson, U. H.; Carlsson, J. Ultralarge Virtual Screening Identifies SARS-CoV-2 Main Protease Inhibitors with Broad-Spectrum Activity against Coronaviruses. *J. Am. Chem. Soc.* **2022**, *144* (7), 2905–2920. https://doi.org/10.1021/jacs.1c08402.
(57) Lenharo, M. New Pill Helps COVID Smell and Taste Loss Fade Quickly. *Nature* **2023**. https://doi.org/10.1038/d41586-023-03244-7.
(58) Diedrich, K.; Krause, B.; Berg, O.; Rarey, M. PoseEdit: Enhanced Ligand Binding Mode Communication by Interactive 2D Diagrams. *J. Comput. Aided Mol. Des.* **2023**, *37* (10), 491–503. https://doi.org/10.1007/s10822-023-00522-4.
(59) Schöning-Stierand, K.; Diedrich, K.; Ehrt, C.; Flachsenberg, F.; Graef, J.; Sieg, J.; Penner, P.; Poppinga, M.; Ungethüm, A.; Rarey, M. ProteinsPlus: A Comprehensive Collection of Web-Based Molecular Modeling Tools. *Nucleic Acids Res.* **2022**, *50* (W1), W611–W615. https://doi.org/10.1093/nar/gkac305.
(60) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28* (1), 235–242.
(61) Pang, X.; Xu, W.; Liu, Y.; Li, H.; Chen, L. The Research Progress of SARS-CoV-2 Main Protease Inhibitors from 2020 to 2022. *Eur. J. Med. Chem.* **2023**, *257*, 115491. https://doi.org/10.1016/j.ejmech.2023.115491.
(62) Helmy, Y. A.; Fawzy, M.; Elaswad, A.; Sobieh, A.; Kenney, S. P.; Shehata, A. A. The COVID-19 Pandemic: A Comprehensive Review of Taxonomy, Genetics, Epidemiology, Diagnosis, Treatment, and Control. *J. Clin. Med.* **2020**, *9* (4), 1225. https://doi.org/10.3390/jcm9041225.
(63) Woo, P. C. Y.; Lau, S. K. P.; Lam, C. S. F.; Lau, C. C. Y.; Tsang, A. K. L.; Lau, J. H. N.; Bai, R.; Teng, J. L. L.; Tsang, C. C. C.; Wang, M.; Zheng, B.-J.; Chan, K.-H.; Yuen, K.-Y. Discovery of Seven Novel Mammalian and Avian Coronaviruses in the Genus *Deltacoronavirus* Supports Bat Coronaviruses as the Gene Source of *Alphacoronavirus* and *Betacoronavirus* and Avian Coronaviruses as the Gene Source of *Gammacoronavirus* and *Deltacoronavirus*. *J. Virol.* **2012**, *86* (7), 3995–4008. https://doi.org/10.1128/JVI.06540-11.
(64) Cui, J.; Li, F.; Shi, Z.-L. Origin and Evolution of Pathogenic Coronaviruses. *Nat. Rev. Microbiol.* **2019**, *17* (3), 181–192. https://doi.org/10.1038/s41579-018-0118-9.

(65) Tran, N.; Dasari, S.; Barwell, S. A. E.; McLeod, M. J.; Kalyaanamoorthy, S.; Holyoak, T.; Ganesan, A. The H163A Mutation Unravels an Oxidized Conformation of the SARS-CoV-2 Main Protease. *Nat. Commun.* **2023**, *14* (1), 5625. https://doi.org/10.1038/s41467-023-40023-4.
(66) Barrila, J.; Gabelli, S. B.; Bacha, U.; Amzel, L. M.; Freire, E. Mutation of Asn28 Disrupts the Dimerization and Enzymatic Activity of SARS 3CLpro. *Biochemistry* **2010**, *49* (20), 4308–4317. https://doi.org/10.1021/bi1002585.
(67) Ullrich, S.; Ekanayake, K. B.; Otting, G.; Nitsche, C. Main Protease Mutants of SARS-CoV-2 Variants Remain Susceptible to Nirmatrelvir. *Bioorg. Med. Chem. Lett.* **2022**, *62*, 128629. https://doi.org/10.1016/j.bmcl.2022.128629.
(68) Shahhamzehei, N.; Abdelfatah, S.; Efferth, T. In Silico and In Vitro Identification of Pan-Coronaviral Main Protease Inhibitors from a Large Natural Product Library. *Pharmaceuticals* **2022**, *15* (3), 308. https://doi.org/10.3390/ph15030308.
(69) Lin, Y.; Zang, R.; Ma, Y.; Wang, Z.; Li, L.; Ding, S.; Zhang, R.; Wei, Z.; Yang, J.; Wang, X. Xanthohumol Is a Potent Pan-Inhibitor of Coronaviruses Targeting Main Protease. *Int. J. Mol. Sci.* **2021**, *22* (22), 12134. https://doi.org/10.3390/ijms222212134.
(70) Xu, J.; Zhu, Q.; Li, W.; Yin, X.; Li, J. Structural Basis for the Inhibition of the HCoV-NL63 Main Protease Mpro by X77. *Biochem. Biophys. Res. Commun.* **2024**, *724*, 150231. https://doi.org/10.1016/j.bbrc.2024.150231.
(71) Jiang, H.; Li, W.; Zhou, X.; Zhang, J.; Li, J. Crystal Structures of Coronaviral Main Proteases in Complex with the Non-Covalent Inhibitor X77. *Int. J. Biol. Macromol.* **2024**, *276*, 133706. https://doi.org/10.1016/j.ijbiomac.2024.133706.
(72) Joyce, R. P.; Hu, V. W.; Wang, J. The History, Mechanism, and Perspectives of Nirmatrelvir (PF-07321332): An Orally Bioavailable Main Protease Inhibitor Used in Combination with Ritonavir to Reduce COVID-19-Related Hospitalizations. *Med. Chem. Res.* **2022**, *31* (10), 1637–1646. https://doi.org/10.1007/s00044-022-02951-6.
(73) Saravolatz, L. D.; Depcinski, S.; Sharma, M. Molnupiravir and Nirmatrelvir-Ritonavir: Oral Coronavirus Disease 2019 Antiviral Drugs. *Clin. Infect. Dis.* **2023**, *76* (1), 165–171. https://doi.org/10.1093/cid/ciac180.
(74) *SARS | Basics Factsheet | CDC*. https://www.cdc.gov/sars/about/fs-sars.html (accessed 2023-12-04).
(75) Sievers, F.; Wilm, A.; Dineen, D.; Gibson, T. J.; Karplus, K.; Li, W.; Lopez, R.; McWilliam, H.; Remmert, M.; Söding, J.; Thompson, J. D.; Higgins, D. G. Fast, Scalable Generation of High-Quality Protein Multiple Sequence Alignments Using Clustal Omega. *Mol. Syst. Biol.* **2011**, *7*, 539. https://doi.org/10.1038/msb.2011.75.
(76) Thakur, M.; Buniello, A.; Brooksbank, C.; Gurwitz, K. T.; Hall, M.; Hartley, M.; Hulcoop, D. G.; Leach, A. R.; Marques, D.; Martin, M.; Mithani, A.; McDonagh, E. M.; Mutasa-Gottgens, E.; Ochoa, D.; Perez-Riverol, Y.; Stephenson, J.; Varadi, M.; Velankar, S.; Vizcaino, J. A.; Witham, R.; McEntyre, J. EMBL's European Bioinformatics Institute (EMBL-EBI) in 2023. *Nucleic Acids Res.* **2024**, *52* (D1), D10–D17. https://doi.org/10.1093/nar/gkad1088.
(77) Zhao, Y.; Zhu, Y.; Liu, X.; Jin, Z.; Duan, Y.; Zhang, Q.; Wu, C.; Feng, L.; Du, X.; Zhao, J.; Shao, M.; Zhang, B.; Yang, X.; Wu, L.; Ji, X.; Guddat, L. W.; Yang, K.; Rao, Z.; Yang, H. Structural Basis for Replicase Polyprotein Cleavage and Substrate Specificity of Main Protease from SARS-CoV-2. *Proc. Natl. Acad. Sci. U. S. A.* **2022**, *119* (16), e2117142119. https://doi.org/10.1073/pnas.2117142119.
(78) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and Computational Approaches to Estimate Solubility and Permeability in Drug Discovery and Development Settings. *Adv. Drug Deliv. Rev.* **1997**, *23* (1), 3–25. https://doi.org/10.1016/S0169-409X(96)00423-1.

(79) Tran-Nguyen, V.-K.; Camproux, A.-C.; Taboureau, O. ClassyPose: A Machine-Learning Classification Model for Ligand Pose Selection Applied to Virtual Screening in Drug Discovery. *Adv. Intell. Syst.* **2024**, *6* (12), 2400238. https://doi.org/10.1002/aisy.202400238.
(80) Wójcikowski, M.; Kukiełka, M.; Stepniewska-Dziubinska, M. M.; Siedlecki, P. Development of a Protein–Ligand Extended Connectivity (PLEC) Fingerprint and Its Application for Binding Affinity Predictions. *Bioinformatics* **2019**, *35* (8), 1334–1341. https://doi.org/10.1093/bioinformatics/bty757.
(81) Kasenda, R. Y.; Tenda, J. O.; Iman, E. W. R.; Manantung, J. A. M.; Moekari, Z. J. S.; Pantas, M. C. The Role and Evolution of Frontend Developers in the Software Development Industry. *J. Syntax Admiration* **2024**, *5* (11), 5191–5196. https://doi.org/10.46799/jsa.v5i11.1852.
(82) *The RDKit Documentation — The RDKit 2023.09.6 documentation*. https://www.rdkit.org/docs/index.html (accessed 2024-03-13).
(83) *Open Babel: An open chemical toolbox | Journal of Cheminformatics | Full Text*. https://jcheminf.biomedcentral.com/articles/10.1186/1758-2946-3-33 (accessed 2025-08-02).
(84) Wójcikowski, M.; Zielenkiewicz, P.; Siedlecki, P. Open Drug Discovery Toolkit (ODDT): A New Open-Source Player in the Drug Discovery Field. *J. Cheminformatics* **2015**, *7* (1), 26. https://doi.org/10.1186/s13321-015-0078-2.
(85) *HTML Standard*. https://html.spec.whatwg.org/multipage/introduction.html#background (accessed 2025-08-14).
(86) Emmanni, P. S. *The Role of TypeScript in Enhancing Development with Modern JavaScript Frameworks*. International Journal of Science and Research (IJSR). https://www.ijsr.net/getabstract.php?paperid=SR24401234212 (accessed 2025-08-14).
(87) Balaban, G.; Grytten, I.; Rand, K. D.; Scheffer, L.; Sandve, G. K. Ten Simple Rules for Quick and Dirty Scientific Programming. *PLoS Comput. Biol.* **2021**, *17* (3), e1008549. https://doi.org/10.1371/journal.pcbi.1008549.
(88) Chen, S.; Thaduri, U. R.; Ballamudi, V. K. R. Front-End Development in React: An Overview. *Eng. Int.* **2019**, *7* (2), 117–126. https://doi.org/10.18034/ei.v7i2.662.
(89) Jackson, D. Scalable Vector Graphics (SVG): The World Wide Web Consortium's Recommendation for High Quality Web Graphics. In *ACM SIGGRAPH 2002 conference abstracts and applications*; SIGGRAPH '02; Association for Computing Machinery: New York, NY, USA, 2002; p 319. https://doi.org/10.1145/1242073.1242327.
(90) Bostock, M.; Ogievetsky, V.; Heer, J. D^3 Data-Driven Documents. *IEEE Trans. Vis. Comput. Graph.* **2011**, *17* (12), 2301–2309. https://doi.org/10.1109/TVCG.2011.185.
(91) Dubey, A.; Chauhan, G. S.; Dubey, A.; Singh, J.; Girdhar, P. Customized Framework for Backend Using Node JS. In *2023 International Conference on Sustainable Emerging Innovations in Engineering and Technology (ICSEIET)*; 2023; pp 01–05. https://doi.org/10.1109/ICSEIET58677.2023.10303462.
(92) Shahi, A.; Barge, A.; Butala, A.; Patil, S. Rest API Based Web Interface for Blogging Application. In *2023 International Conference on Sustainable Computing and Smart Systems (ICSCSS)*; 2023; pp 673–677. https://doi.org/10.1109/ICSCSS57650.2023.10169662.
(93) Idris, N.; Foozy, C. F. M.; Shamala, P. A Generic Review of Web Technology: DJango and Flask. *Int. J. Adv. Sci. Comput. Eng.* **2020**, *2* (1), 34–40. https://doi.org/10.62527/ijasce.2.1.29.
(94) Dalal, N.; Shenoy, G. S. Deployment of Full Stack Applications with Backend REST API Using Spring Boot and Frontend Using ReactJS on A WS Cloud. In *2024 International Conference on IoT Based Control Networks and Intelligent Systems (ICICNIS)*; 2024; pp 93–98. https://doi.org/10.1109/ICICNIS64247.2024.10823252.

(95) Danielsson, P.; Postema, T.; Munir, H. Heroku-Based Innovative Platform for Web-Based Deployment in Product Development at Axis. *IEEE Access* **2021**, *9*, 10805–10819. https://doi.org/10.1109/ACCESS.2021.3050255.
(96) *Container Registry & Runtime (Docker Deploys) | Heroku Dev Center*. https://devcenter.heroku.com/articles/container-registry-and-runtime (accessed 2025-08-14).
(97) Singh, S.; Singh, N. Containers & Docker: Emerging Roles & Future of Cloud Technology. In *2016 2nd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT)*; 2016; pp 804–807. https://doi.org/10.1109/ICATCCT.2016.7912109.
(98) Ly, N.; Ramgoolam, S. D.; Ganesan, A. Computational Mapping of Conformational Dynamics and Interaction Hotspots of Human VISTA with pH-Selective Antibodies. *Biochemistry* **2025**. https://doi.org/10.1021/acs.biochem.4c00863.
(99) Petgrave, M.; Ramgoolam, S. D.; Ganesan, A. Deciphering the Molecular Association of Human CRIP1a with an Agonist-Bound Cannabinoid Receptor 1. *J. Chem. Inf. Model.* **2024**, *64* (2), 499–517. https://doi.org/10.1021/acs.jcim.3c01579.
(100) Kokh, D. B.; Doser, B.; Richter, S.; Ormersbach, F.; Cheng, X.; Wade, R. C. A Workflow for Exploring Ligand Dissociation from a Macromolecule: Efficient Random Acceleration Molecular Dynamics Simulation and Interaction Fingerprint Analysis of Ligand Trajectories. *J. Chem. Phys.* **2020**, *153* (12), 125102. https://doi.org/10.1063/5.0019088.
(101) Harren, T.; Gutermuth, T.; Grebner, C.; Hessler, G.; Rarey, M. Modern Machine-Learning for Binding Affinity Estimation of Protein–Ligand Complexes: Progress, Opportunities, and Challenges. *WIREs Comput. Mol. Sci.* **2024**, *14* (3), e1716. https://doi.org/10.1002/wcms.1716.
(102) Fellinger, C.; Seidel, T.; Merget, B.; Schleifer, K.-J.; Langer, T. GRADE and X-GRADE: Unveiling Novel Protein–Ligand Interaction Fingerprints Based on GRAIL Scores. *J. Chem. Inf. Model.* **2025**, *65* (5), 2456–2475. https://doi.org/10.1021/acs.jcim.4c01902.
(103) Zhang, J.; Chen, H. De Novo Molecule Design Using Molecular Generative Models Constrained by Ligand–Protein Interactions. *J. Chem. Inf. Model.* **2022**, *62* (14), 3291–3306. https://doi.org/10.1021/acs.jcim.2c00177.
(104) Bajusz, D.; Rácz, A.; Héberger, K. Why Is Tanimoto Index an Appropriate Choice for Fingerprint-Based Similarity Calculations? *J. Cheminformatics* **2015**, *7*, 20. https://doi.org/10.1186/s13321-015-0069-3.
(105) Whittle, M.; Gillet, V. J.; Willett, P.; Alex, A.; Loesel, J. Enhancing the Effectiveness of Virtual Screening by Fusing Nearest Neighbor Lists: A Comparison of Similarity Coefficients. *J. Chem. Inf. Comput. Sci.* **2004**, *44* (5), 1840–1848. https://doi.org/10.1021/ci049867x.
(106) Todeschini, R.; Consonni, V.; Xiang, H.; Holliday, J.; Buscema, M.; Willett, P. Similarity Coefficients for Binary Chemoinformatics Data: Overview and Extended Comparison Using Simulated and Real Data Sets. *J. Chem. Inf. Model.* **2012**, *52* (11), 2884–2901. https://doi.org/10.1021/ci300261r.
(107) Cereto-Massagué, A.; Ojeda, M. J.; Valls, C.; Mulero, M.; Garcia-Vallvé, S.; Pujadas, G. Molecular Fingerprint Similarity Search in Virtual Screening. *Methods* **2015**, *71*, 58–63. https://doi.org/10.1016/j.ymeth.2014.08.005.
(108) Morgan, H. L. The Generation of a Unique Machine Description for Chemical Structures-A Technique Developed at Chemical Abstracts Service. *J. Chem. Doc.* **1965**, *5* (2), 107–113. https://doi.org/10.1021/c160017a018.
(109) Capecchi, A.; Probst, D.; Reymond, J.-L. One Molecular Fingerprint to Rule Them All: Drugs, Biomolecules, and the Metabolome. *J. Cheminformatics* **2020**, *12* (1), 43. https://doi.org/10.1186/s13321-020-00445-4.

(110) Riniker, S.; Landrum, G. A. Open-Source Platform to Benchmark Fingerprints for Ligand-Based Virtual Screening. *J. Cheminformatics* **2013**, *5*, 26. https://doi.org/10.1186/1758-2946-5-26.
(111) Awale, M.; Reymond, J.-L. Polypharmacology Browser PPB2: Target Prediction Combining Nearest Neighbors with Machine Learning. *J. Chem. Inf. Model.* **2019**, *59* (1), 10–17. https://doi.org/10.1021/acs.jcim.8b00524.
(112) *T. T. Tanimoto, (1958) "An elementary mathematical theory of classification and prediction," IBM Internal Report. - References - Scientific Research Publishing*. https://www.scirp.org/reference/referencespapers?referenceid=12779 (accessed 2025-08-14).
(113) Lipinski, C. A. Lead- and Drug-like Compounds: The Rule-of-Five Revolution. *Drug Discov. Today Technol.* **2004**, *1* (4), 337–341. https://doi.org/10.1016/j.ddtec.2004.11.007.
(114) Navia, M. A.; Chaturvedi, P. R. Design Principles for Orally Bioavailable Drugs. *Drug Discov. Today* **1996**, *1* (5), 179–189. https://doi.org/10.1016/1359-6446(96)10020-9.
(115) Pardridge, W. M. Transport of Small Molecules through the Blood-Brain Barrier: Biology and Methodology. *Adv. Drug Deliv. Rev.* **1995**, *15* (1), 5–36. https://doi.org/10.1016/0169-409X(95)00003-P.
(116) Artursson, P.; Palm, K.; Luthman, K. Caco-2 Monolayers in Experimental and Theoretical Predictions of Drug Transport1. *Adv. Drug Deliv. Rev.* **2001**, *46* (1), 27–43. https://doi.org/10.1016/S0169-409X(00)00128-9.
(117) Seddon, A. M.; Casey, D.; Law, R. V.; Gee, A.; Templer, R. H.; Ces, O. Drug Interactions with Lipid Membranes. *Chem. Soc. Rev.* **2009**, *38* (9), 2509–2519. https://doi.org/10.1039/B813853M.
(118) Milanetti, E.; Raimondo, D.; Tramontano, A. Prediction of the Permeability of Neutral Drugs Inferred from Their Solvation Properties. *Bioinformatics* **2016**, *32* (8), 1163–1169. https://doi.org/10.1093/bioinformatics/btv725.
(119) Abraham, M. H.; Chadha, H. S.; Whiting, G. S.; Mitchell, R. C. Hydrogen Bonding. 32. An Analysis of Water-Octanol and Water-Alkane Partitioning and the ΔLog *P* Parameter of Seiler. *J. Pharm. Sci.* **1994**, *83* (8), 1085–1100. https://doi.org/10.1002/jps.2600830806.
(120) Paterson, D. A.; Conradi, R. A.; Hilgers, A. R.; Vidmar, T. J.; Burton, P. S. A Non-Aqueous Partitioning System for Predicting the Oral Absorption Potential of Peptides. *Quant. Struct.-Act. Relatsh.* **1994**, *13* (1), 4–10. https://doi.org/10.1002/qsar.19940130103.
(121) Torres, P. H. M.; Sodero, A. C. R.; Jofily, P.; Silva-Jr, F. P. Key Topics in Molecular Docking for Drug Design. *Int. J. Mol. Sci.* **2019**, *20* (18), 4574. https://doi.org/10.3390/ijms20184574.
(122) Huang, S.-Y.; Zou, X. Advances and Challenges in Protein-Ligand Docking. *Int. J. Mol. Sci.* **2010**, *11* (8), 3016–3034. https://doi.org/10.3390/ijms11083016.
(123) Jones, G.; Willett, P.; Glen, R. C. Molecular Recognition of Receptor Sites Using a Genetic Algorithm with a Description of Desolvation. *J. Mol. Biol.* **1995**, *245* (1), 43–53. https://doi.org/10.1016/S0022-2836(95)80037-9.
(124) Li, J.; Fu, A.; Zhang, L. An Overview of Scoring Functions Used for Protein–Ligand Interactions in Molecular Docking. *Interdiscip. Sci. Comput. Life Sci.* **2019**, *11* (2), 320–328. https://doi.org/10.1007/s12539-019-00327-w.
(125) Isert, C.; Atz, K.; Schneider, G. Structure-Based Drug Design with Geometric Deep Learning. *Curr. Opin. Struct. Biol.* **2023**, *79*, 102548. https://doi.org/10.1016/j.sbi.2023.102548.
(126) Zou, Y.; Guo, T.; Fu, Z.; Guo, Z.; Bo, W.; Yan, D.; Wang, Q.; Zeng, J.; Xu, D.; Wang, T.; Chen, L. A Structure-Based Framework for Selective Inhibitor Design and Optimization. *Commun. Biol.* **2025**, *8* (1), 422. https://doi.org/10.1038/s42003-025-07840-3.

(127) Lu, W.; Wu, Q.; Zhang, J.; Rao, J.; Li, C.; Zheng, S. TANKBind: Trigonometry-Aware Neural NetworKs for Drug-Protein Binding Structure Prediction. bioRxiv June 6, 2022, p 2022.06.06.495043. https://doi.org/10.1101/2022.06.06.495043.
(128) Stärk, H.; Ganea, O.-E.; Pattanaik, L.; Barzilay, R.; Jaakkola, T. EquiBind: Geometric Deep Learning for Drug Binding Structure Prediction. arXiv June 4, 2022. https://doi.org/10.48550/arXiv.2202.05146.
(129) Corso, G.; Stärk, H.; Jing, B.; Barzilay, R.; Jaakkola, T. DiffDock: Diffusion Steps, Twists, and Turns for Molecular Docking. arXiv February 11, 2023. https://doi.org/10.48550/arXiv.2210.01776.
(130) Ballester, P. J.; Mitchell, J. B. O. A Machine Learning Approach to Predicting Protein-Ligand Binding Affinity with Applications to Molecular Docking. *Bioinforma. Oxf. Engl.* **2010**, *26* (9), 1169–1175. https://doi.org/10.1093/bioinformatics/btq112.
(131) Durrant, J. D.; McCammon, J. A. NNScore 2.0: A Neural-Network Receptor–Ligand Scoring Function. *J. Chem. Inf. Model.* **2011**, *51* (11), 2897–2903. https://doi.org/10.1021/ci2003889.
(132) Sun, H.; Pan, P.; Tian, S.; Xu, L.; Kong, X.; Li, Y.; Dan Li, null; Hou, T. Constructing and Validating High-Performance MIEC-SVM Models in Virtual Screening for Kinases: A Better Way for Actives Discovery. *Sci. Rep.* **2016**, *6*, 24817. https://doi.org/10.1038/srep24817.
(133) Afifi, K.; Al-Sadek, A. F. Improving Classical Scoring Functions Using Random Forest: The Non-Additivity of Free Energy Terms' Contributions in Binding. *Chem. Biol. Drug Des.* **2018**, *92* (2), 1429–1434. https://doi.org/10.1111/cbdd.13206.
(134) Macip, G.; Garcia-Segura, P.; Mestres-Truyol, J.; Saldivar-Espinoza, B.; Ojeda-Montes, M. J.; Gimeno, A.; Cereto-Massagué, A.; Garcia-Vallvé, S.; Pujadas, G. Haste Makes Waste: A Critical Review of Docking-based Virtual Screening in Drug Repurposing for SARS-CoV-2 Main Protease (M-pro) Inhibition. *Med. Res. Rev.* **2021**, *42* (2), 744. https://doi.org/10.1002/med.21862.
(135) Sivakumar, M.; Parthasarathy, S.; Padmapriya, T. Trade-off between Training and Testing Ratio in Machine Learning for Medical Image Processing. *PeerJ Comput. Sci.* **2024**, *10*, e2245. https://doi.org/10.7717/peerj-cs.2245.
(136) Richardson, E.; Trevizani, R.; Greenbaum, J. A.; Carter, H.; Nielsen, M.; Peters, B. The Receiver Operating Characteristic Curve Accurately Assesses Imbalanced Datasets. *Patterns* **2024**, *5* (6), 100994. https://doi.org/10.1016/j.patter.2024.100994.
(137) Erickson, B. J.; Kitamura, F. Magician's Corner: 9. Performance Metrics for Machine Learning Models. *Radiol. Artif. Intell.* **2021**, *3* (3), e200126. https://doi.org/10.1148/ryai.2021200126.
(138) Tafvizi, A.; Avci, B.; Sundararajan, M. Attributing AUC-ROC to Analyze Binary Classifier Performance. arXiv May 24, 2022. https://doi.org/10.48550/arXiv.2205.11781.
(139) Hazra, A.; Gogtay, N. Biostatistics Series Module 10: Brief Overview of Multivariate Methods. *Indian J. Dermatol.* **2017**, *62* (4), 358–366. https://doi.org/10.4103/ijd.IJD_296_17.
(140) Giuliani, A. The Application of Principal Component Analysis to Drug Discovery and Biomedical Data. *Drug Discov. Today* **2017**, *22* (7), 1069–1076. https://doi.org/10.1016/j.drudis.2017.01.005.
(141) Wenderski, T. A.; Stratton, C. F.; Bauer, R. A.; Kopp, F.; Tan, D. S. Principal Component Analysis as a Tool for Library Design: A Case Study Investigating Natural Products, Brand-Name Drugs, Natural Product-Like Libraries, and Drug-Like Libraries. *Methods Mol. Biol. Clifton NJ* **2015**, *1263*, 225–242. https://doi.org/10.1007/978-1-4939-2269-7_18.
(142) Talevi, A.; Bellera, C. L. Clustering of Small Molecules: New Perspectives and Their Impact on Natural Product Lead Discovery. *Front. Nat. Prod.* **2024**, *3*. https://doi.org/10.3389/fntpr.2024.1367537.

(143) Jolliffe, I. T.; Cadima, J. Principal Component Analysis: A Review and Recent Developments. *Philos. Transact. A Math. Phys. Eng. Sci.* **2016**, *374* (2065), 20150202. https://doi.org/10.1098/rsta.2015.0202.
(144) Yin, H.; Aryani, A.; Petrie, S.; Nambissan, A.; Astudillo, A.; Cao, S. A Rapid Review of Clustering Algorithms. arXiv January 14, 2024. https://doi.org/10.48550/arXiv.2401.07389.
(145) Gere, A. Recommendations for Validating Hierarchical Clustering in Consumer Sensory Projects. *Curr. Res. Food Sci.* **2023**, *6*, 100522. https://doi.org/10.1016/j.crfs.2023.100522.
(146) Andrianov, A. M.; Kornoushenko, Y. V.; Karpenko, A. D.; Bosko, I. P.; Tuzikov, A. V. Computational Discovery of Small Drug-like Compounds as Potential Inhibitors of SARS-CoV-2 Main Protease. *J. Biomol. Struct. Dyn.* 1–13. https://doi.org/10.1080/07391102.2020.1792989.
(147) Boby, M. L.; Fearon, D.; Ferla, M.; Filep, M.; Koekemoer, L.; Robinson, M. C.; The COVID Moonshot Consortium; Chodera, J. D.; Lee, A. A.; London, N.; von Delft, A.; von Delft, F. Open Science Discovery of Potent Noncovalent SARS-CoV-2 Main Protease Inhibitors. *Science* **2023**, *382* (6671), eabo7201. https://doi.org/10.1126/science.abo7201.
(148) Bai, Y.; Ye, F.; Feng, Y.; Liao, H.; Song, H.; Qi, J.; Gao, G. F.; Tan, W.; Fu, L.; Shi, Y. Structural Basis for the Inhibition of the SARS-CoV-2 Main Protease by the Anti-HCV Drug Narlaprevir. *Signal Transduct. Target. Ther.* **2021**, *6* (1), 51. https://doi.org/10.1038/s41392-021-00468-9.
(149) Lockbaum, G. J.; Reyes, A. C.; Lee, J. M.; Tilvawala, R.; Nalivaika, E. A.; Ali, A.; Kurt Yilmaz, N.; Thompson, P. R.; Schiffer, C. A. Crystal Structure of SARS-CoV-2 Main Protease in Complex with the Non-Covalent Inhibitor ML188. *Viruses* **2021**, *13* (2), 174. https://doi.org/10.3390/v13020174.
(150) Kneller, D. W.; Phillips, G.; O'Neill, H. M.; Jedrzejczak, R.; Stols, L.; Langan, P.; Joachimiak, A.; Coates, L.; Kovalevsky, A. Structural Plasticity of SARS-CoV-2 3CL Mpro Active Site Cavity Revealed by Room Temperature X-Ray Crystallography. *Nat. Commun.* **2020**, *11* (1), 3202. https://doi.org/10.1038/s41467-020-16954-7.
(151) Xia, Z.; Sacco, M.; Hu, Y.; Ma, C.; Meng, X.; Zhang, F.; Szeto, T.; Xiang, Y.; Chen, Y.; Wang, J. Rational Design of Hybrid SARS-CoV-2 Main Protease Inhibitors Guided by the Superimposed Cocrystal Structures with the Peptidomimetic Inhibitors GC-376, Telaprevir, and Boceprevir. *ACS Pharmacol. Transl. Sci.* **2021**, *4* (4), 1408–1421. https://doi.org/10.1021/acsptsci.1c00099.
(152) Melano, I.; Lo, Y.-C.; Su, W.-C. Characterization of Host Substrates of SARS-CoV-2 Main Protease. *Front. Microbiol.* **2023**, *14*, 1251705. https://doi.org/10.3389/fmicb.2023.1251705.
(153) Yang, Y.; Luo, Y.-D.; Zhang, C.-B.; Xiang, Y.; Bai, X.-Y.; Zhang, D.; Fu, Z.-Y.; Hao, R.-B.; Liu, X.-L. Progress in Research on Inhibitors Targeting SARS-CoV-2 Main Protease (Mpro). *ACS Omega* **2024**, *9* (32), 34196–34219. https://doi.org/10.1021/acsomega.4c03023.
(154) Chen, P.; Van Oers, T. J.; Arutyunova, E.; Fischer, C.; Wang, C.; Lamer, T.; van Belkum, M. J.; Young, H. S.; Vederas, J. C.; Lemieux, M. J. A Structural Comparison of Oral SARS-CoV-2 Drug Candidate Ibuzatrelvir Complexed with the Main Protease (Mpro) of SARS-CoV-2 and MERS-CoV. *JACS Au* **2024**, *4* (8), 3217–3227. https://doi.org/10.1021/jacsau.4c00508.
(155) Brewitz, L.; Schofield, C. J. Fixing the Achilles Heel of Pfizer's Paxlovid for COVID-19 Treatment. *J. Med. Chem.* **2024**, *67* (14), 11656. https://doi.org/10.1021/acs.jmedchem.4c01342.
(156) Rungruangmaitree, R.; Phoochaijaroen, S.; Chimprasit, A.; Saparpakorn, P.; Pootanakit, K.; Tanramluk, D. Structural Analysis of the Coronavirus Main Protease for the Design of Pan-Variant Inhibitors. *Sci. Rep.* **2023**, *13* (1), 7055. https://doi.org/10.1038/s41598-023-34305-6.

(157) Laskowski, R. A.; Hutchinson, E. G.; Michie, A. D.; Wallace, A. C.; Jones, M. L.; Thornton, J. M. PDBsum: A Web-Based Database of Summaries and Analyses of All PDB Structures. *Trends Biochem. Sci.* **1997**, *22* (12), 488–490. https://doi.org/10.1016/S0968-0004(97)01140-7.
(158) Skyner,R.; von Delft, F. Fragalysis. https://fragalysis.diamond.ac.uk/.
(159) Needleman, S. B.; Wunsch, C. D. A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins. *J. Mol. Biol.* **1970**, *48* (3), 443–453. https://doi.org/10.1016/0022-2836(70)90057-4.
(160) *Array programming with NumPy | Nature*. https://www.nature.com/articles/s41586-020-2649-2 (accessed 2025-08-14).
(161) 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.; Duchesnay, É. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12* (85), 2825–2830.
(162) Murtagh, F.; Legendre, P. Ward's Hierarchical Clustering Method: Clustering Criterion and Agglomerative Algorithm. *J. Classif.* **2014**, *31* (3), 274–295. https://doi.org/10.1007/s00357-014-9161-z.
(163) Virtanen, P.; Gommers, R.; Oliphant, T. E.; Haberland, M.; Reddy, T.; Cournapeau, D.; Burovski, E.; Peterson, P.; Weckesser, W.; Bright, J.; van der Walt, S. J.; Brett, M.; Wilson, J.; Millman, K. J.; Mayorov, N.; Nelson, A. R. J.; Jones, E.; Kern, R.; Larson, E.; Carey, C. J.; Polat, İ.; Feng, Y.; Moore, E. W.; VanderPlas, J.; Laxalde, D.; Perktold, J.; Cimrman, R.; Henriksen, I.; Quintero, E. A.; Harris, C. R.; Archibald, A. M.; Ribeiro, A. H.; Pedregosa, F.; van Mulbregt, P. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nat. Methods* **2020**, *17* (3), 261–272. https://doi.org/10.1038/s41592-019-0686-2.
(164) Cock, P. J. A.; Antao, T.; Chang, J. T.; Chapman, B. A.; Cox, C. J.; Dalke, A.; Friedberg, I.; Hamelryck, T.; Kauff, F.; Wilczynski, B.; de Hoon, M. J. L. Biopython: Freely Available Python Tools for Computational Molecular Biology and Bioinformatics. *Bioinformatics* **2009**, *25* (11), 1422–1423. https://doi.org/10.1093/bioinformatics/btp163.
(165) Kabsch, W. A Solution for the Best Rotation to Relate Two Sets of Vectors. *Acta Crystallogr. A* **1976**, *32* (5), 922–923. https://doi.org/10.1107/S0567739476001873.
(166) Kim, S.; Chen, J.; Cheng, T.; Gindulyte, A.; He, J.; He, S.; Li, Q.; Shoemaker, B. A.; Thiessen, P. A.; Yu, B.; Zaslavsky, L.; Zhang, J.; Bolton, E. E. PubChem in 2021: New Data Content and Improved Web Interfaces. *Nucleic Acids Res.* **2021**, *49* (D1), D1388–D1395. https://doi.org/10.1093/nar/gkaa971.
(167) Ibrahim, M.; Sun, X.; Martins de Oliveira, V.; Liu, R.; Clayton, J.; El Kilani, H.; Shen, J.; Hilgenfeld, R. Why Is the Omicron Main Protease of SARS-CoV-2 Less Stable than Its Wild-Type Counterpart? A Crystallographic, Biophysical, and Theoretical Study. *hLife* **2024**, *2* (8), 419–433. https://doi.org/10.1016/j.hlife.2024.06.003.
(168) Cooper, M. S.; Zhang, L.; Ibrahim, M.; Zhang, K.; Sun, X.; Röske, J.; Göhl, M.; Brönstrup, M.; Cowell, J. K.; Sauerhering, L.; Becker, S.; Vangeel, L.; Jochmans, D.; Neyts, J.; Rox, K.; Marsh, G. P.; Maple, H. J.; Hilgenfeld, R. Diastereomeric Resolution Yields Highly Potent Inhibitor of SARS-CoV-2 Main Protease. *J. Med. Chem.* **2022**, *65* (19), 13328–13342. https://doi.org/10.1021/acs.jmedchem.2c01131.
(169) Poli, A. N. R.; Tietjen, I.; Nandwana, N. K.; Cassel, J.; Messick, T. E.; Register, E. T.; Keeney, F.; Rajaiah, R.; Verma, A. K.; Pandey, K.; Acharya, A.; Byrareddy, S. N.; Montaner, L. J.; Salvino, J. M. Design of Novel and Highly Selective SARS-CoV-2 Main Protease Inhibitors. *Antimicrob. Agents Chemother.* **2024**, *68* (10), e00562-24. https://doi.org/10.1128/aac.00562-24.
(170) Westberg, M.; Su, Y.; Zou, X.; Huang, P.; Rustagi, A.; Garhyan, J.; Patel, P. B.; Fernandez, D.; Wu, Y.; Hao, C.; Lo, C.-W.; Karim, M.; Ning, L.; Beck, A.; Saenkham-Huntsinger, P.; Tat, V.;

Drelich, A.; Peng, B.-H.; Einav, S.; Tseng, C.-T. K.; Blish, C.; Lin, M. Z. An Orally Bioavailable SARS-CoV-2 Main Protease Inhibitor Exhibits Improved Affinity and Reduced Sensitivity to Mutations. *Sci. Transl. Med.* **2024**, *16* (738), eadi0979. https://doi.org/10.1126/scitranslmed.adi0979.

(171) Dhama, K.; Patel, S. K.; Sharun, K.; Pathak, M.; Tiwari, R.; Yatoo, M. I.; Malik, Y. S.; Sah, R.; Rabaan, A. A.; Panwar, P. K.; Singh, K. P.; Michalak, I.; Chaicumpa, W.; Martinez-Pulgarin, D. F.; Bonilla-Aldana, D. K.; Rodriguez-Morales, A. J. SARS-CoV-2 Jumping the Species Barrier: Zoonotic Lessons from SARS, MERS and Recent Advances to Combat This Pandemic Virus. *Travel Med. Infect. Dis.* **2020**, *37*, 101830. https://doi.org/10.1016/j.tmaid.2020.101830.

(172) Hilgenfeld, R.; Peiris, M. From SARS to MERS: 10 Years of Research on Highly Pathogenic Human Coronaviruses. *Antiviral Res.* **2013**, *100* (1), 286–295. https://doi.org/10.1016/j.antiviral.2013.08.015.

(173) Ruatta, S. M.; Prada Gori, D. N.; Fló Díaz, M.; Lorenzelli, F.; Perelmuter, K.; Alberca, L. N.; Bellera, C. L.; Medeiros, A.; López, G. V.; Ingold, M.; Porcal, W.; Dibello, E.; Ihnatenko, I.; Kunick, C.; Incerti, M.; Luzardo, M.; Colobbio, M.; Ramos, J. C.; Manta, E.; Minini, L.; Lavaggi, M. L.; Hernández, P.; Šarlauskas, J.; Huerta García, C. S.; Castillo, R.; Hernández-Campos, A.; Ribaudo, G.; Zagotto, G.; Carlucci, R.; Medrán, N. S.; Labadie, G. R.; Martinez-Amezaga, M.; Delpiccolo, C. M. L.; Mata, E. G.; Scarone, L.; Posada, L.; Serra, G.; Calogeropoulou, T.; Prousis, K.; Detsi, A.; Cabrera, M.; Alvarez, G.; Aicardo, A.; Araújo, V.; Chavarría, C.; Mašič, L. P.; Gantner, M. E.; Llanos, M. A.; Rodríguez, S.; Gavernet, L.; Park, S.; Heo, J.; Lee, H.; Paul Park, K.-H.; Bollati-Fogolín, M.; Pritsch, O.; Shum, D.; Talevi, A.; Comini, M. A. Garbage in, Garbage out: How Reliable Training Data Improved a Virtual Screening Approach against SARS-CoV-2 MPro. *Front. Pharmacol.* **2023**, *14*. https://doi.org/10.3389/fphar.2023.1193282.

(174) Krüger, N.; Kronenberger, T.; Xie, H.; Rocha, C.; Pöhlmann, S.; Su, H.; Xu, Y.; Laufer, S. A.; Pillaiyar, T. Discovery of Polyphenolic Natural Products as SARS-CoV-2 Mpro Inhibitors for COVID-19. *Pharm. Basel Switz.* **2023**, *16* (2), 190. https://doi.org/10.3390/ph16020190.

(175) Lavecchia, A.; Giovanni, C. D. Virtual Screening Strategies in Drug Discovery: A Critical Review. *Curr. Med. Chem. 20* (23), 2839–2860.

(176) Llanos, M. A.; Gantner, M. E.; Rodriguez, S.; Alberca, L. N.; Bellera, C. L.; Talevi, A.; Gavernet, L. Strengths and Weaknesses of Docking Simulations in the SARS-CoV-2 Era: The Main Protease (Mpro) Case Study. *J. Chem. Inf. Model.* **2021**, *61* (8), 3758–3770. https://doi.org/10.1021/acs.jcim.1c00404.

(177) Santos, L. H. S.; Ferreira, R. S.; Caffarena, E. R. Integrating Molecular Docking and Molecular Dynamics Simulations. In *Docking Screens for Drug Discovery*; de Azevedo Jr., W. F., Ed.; Springer: New York, NY, 2019; pp 13–34. https://doi.org/10.1007/978-1-4939-9752-7_2.

(178) Saramago, L. C.; Santana, M. V.; Gomes, B. F.; Dantas, R. F.; Senger, M. R.; Oliveira Borges, P. H.; Ferreira, V. N. dos S.; dos Santos Rosa, A.; Tucci, A. R.; Dias Miranda, M.; Lukacik, P.; Strain-Damerell, C.; Owen, C. D.; Walsh, M. A.; Ferreira, S. B.; Silva-Junior, F. P. AI-Driven Discovery of SARS-CoV-2 Main Protease Fragment-like Inhibitors with Antiviral Activity In Vitro. *J. Chem. Inf. Model.* **2023**, *63* (9), 2866–2880. https://doi.org/10.1021/acs.jcim.3c00409.

(179) Chakraborty, A.; Krishnan, V.; Thamotharan, S. Generative Adversarial Network (GAN) Model-Based Design of Potent SARS-CoV-2 Mpro Inhibitors Using the Electron Density of Ligands and 3D Binding Pockets: Insights from Molecular Docking, Dynamics Simulation, and MM-GBSA Analysis. *Mol. Divers.* **2025**, *29* (4), 3059–3075. https://doi.org/10.1007/s11030-024-11047-9.

(180) Hu, F.; Wang, L.; Hu, Y.; Wang, D.; Wang, W.; Jiang, J.; Li, N.; Yin, P. A Novel Framework Integrating AI Model and Enzymological Experiments Promotes Identification of SARS-CoV-2 3CL Protease Inhibitors and Activity-Based Probe. *Brief. Bioinform.* **2021**, *22* (6), bbab301. https://doi.org/10.1093/bib/bbab301.

(181) Tran-Nguyen, V.-K.; Simeon, S.; Junaid, M.; Ballester, P. J. Structure-Based Virtual Screening for PDL1 Dimerizers: Evaluating Generic Scoring Functions. *Curr. Res. Struct. Biol.* **2022**, *4*, 206–210. https://doi.org/10.1016/j.crstbi.2022.06.002.
(182) Wang, C.; Chen, Y.; Zhang, Y.; Li, K.; Lin, M.; Pan, F.; Wu, W.; Zhang, J. A Reinforcement Learning Approach for Protein-Ligand Binding Pose Prediction. *BMC Bioinformatics* **2022**, *23* (1), 368. https://doi.org/10.1186/s12859-022-04912-7.
(183) Ghislat, G.; Rahman, T.; Ballester, P. J. Recent Progress on the Prospective Application of Machine Learning to Structure-Based Virtual Screening. *Curr. Opin. Chem. Biol.* **2021**, *65*, 28–34. https://doi.org/10.1016/j.cbpa.2021.04.009.
(184) Ballester, P. J. Selecting Machine-Learning Scoring Functions for Structure-Based Virtual Screening. *Drug Discov. Today Technol.* **2019**, *32–33*, 81–87. https://doi.org/10.1016/j.ddtec.2020.09.001.
(185) Wójcikowski, M.; Ballester, P. J.; Siedlecki, P. Performance of Machine-Learning Scoring Functions in Structure-Based Virtual Screening. *Sci. Rep.* **2017**, *7*, 46710. https://doi.org/10.1038/srep46710.
(186) Ain, Q. U.; Aleksandrova, A.; Roessler, F. D.; Ballester, P. J. Machine-Learning Scoring Functions to Improve Structure-Based Binding Affinity Prediction and Virtual Screening. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2015**, *5* (6), 405–424. https://doi.org/10.1002/wcms.1225.
(187) Shen, C.; Hu, X.; Gao, J.; Zhang, X.; Zhong, H.; Wang, Z.; Xu, L.; Kang, Y.; Cao, D.; Hou, T. The Impact of Cross-Docked Poses on Performance of Machine Learning Classifier for Protein-Ligand Binding Pose Prediction. *J. Cheminformatics* **2021**, *13* (1), 81. https://doi.org/10.1186/s13321-021-00560-w.
(188) Tran-Nguyen, V.-K.; Jacquemard, C.; Rognan, D. LIT-PCBA: An Unbiased Data Set for Machine Learning and Virtual Screening. *J. Chem. Inf. Model.* **2020**, *60* (9), 4263–4273. https://doi.org/10.1021/acs.jcim.0c00155.
(189) Wang, R.; Fang, X.; Lu, Y.; Wang, S. The PDBbind Database: Collection of Binding Affinities for Protein-Ligand Complexes with Known Three-Dimensional Structures. *J. Med. Chem.* **2004**, *47* (12), 2977–2980. https://doi.org/10.1021/jm030580l.
(190) Mysinger, M. M.; Carchia, M.; Irwin, J. J.; Shoichet, B. K. Directory of Useful Decoys, Enhanced (DUD-E): Better Ligands and Decoys for Better Benchmarking. *J. Med. Chem.* **2012**, *55* (14), 6582–6594. https://doi.org/10.1021/jm300687e.
(191) Hartshorn, M. J.; Verdonk, M. L.; Chessari, G.; Brewerton, S. C.; Mooij, W. T. M.; Mortenson, P. N.; Murray, C. W. Diverse, High-Quality Test Set for the Validation of Protein-Ligand Docking Performance. *J. Med. Chem.* **2007**, *50* (4), 726–741. https://doi.org/10.1021/jm061277y.
(192) Siddell, S. G.; Smith, D. B.; Adriaenssens, E.; Alfenas-Zerbini, P.; Dutilh, B. E.; Garcia, M. L.; Junglen, S.; Krupovic, M.; Kuhn, J. H.; Lambert, A. J.; Lefkowitz, E. J.; Łobocka, M.; Mushegian, A. R.; Oksanen, H. M.; Robertson, D. L.; Rubino, L.; Sabanadzovic, S.; Simmonds, P.; Suzuki, N.; Van Doorslaer, K.; Vandamme, A.-M.; Varsani, A.; Zerbini, F. M. Virus Taxonomy and the Role of the International Committee on Taxonomy of Viruses (ICTV). *J. Gen. Virol.* **2023**, *104* (5), 001840. https://doi.org/10.1099/jgv.0.001840.
(193) Morris, G. M.; Huey, R.; Lindstrom, W.; Sanner, M. F.; Belew, R. K.; Goodsell, D. S.; Olson, A. J. AutoDock4 and AutoDockTools4: Automated Docking with Selective Receptor Flexibility. *J. Comput. Chem.* **2009**, *30* (16), 2785–2791. https://doi.org/10.1002/jcc.21256.
(194) Ding, J.; Tang, S.; Mei, Z.; Wang, L.; Huang, Q.; Hu, H.; Ling, M.; Wu, J. Vina-GPU 2.0: Further Accelerating AutoDock Vina and Its Derivatives with Graphics Processing Units. *J. Chem. Inf. Model.* **2023**, *63* (7), 1982–1998. https://doi.org/10.1021/acs.jcim.2c01504.
(195) Cortes, C.; Vapnik, V. Support-Vector Networks. *Mach. Learn.* **1995**, *20* (3), 273–297. https://doi.org/10.1007/BF00994018.

(196) Valiant, L. G. A Theory of the Learnable. *Commun ACM* **1984**, *27* (11), 1134–1142. https://doi.org/10.1145/1968.1972.
(197) Hintze, J. L.; Nelson, R. D. Violin Plots: A Box Plot-Density Trace Synergism. *Am. Stat.* **1998**, *52* (2), 181–184. https://doi.org/10.1080/00031305.1998.10480559.
(198) Waskom, M. L. Seaborn: Statistical Data Visualization. *J. Open Source Softw.* **2021**, *6* (60), 3021. https://doi.org/10.21105/joss.03021.
(199) Chakraborty, C.; Bhattacharya, M.; Dhama, K. SARS-CoV-2 Vaccines, Vaccine Development Technologies, and Significant Efforts in Vaccine Development during the Pandemic: The Lessons Learned Might Help to Fight against the Next Pandemic. *Vaccines* **2023**, *11* (3), 682. https://doi.org/10.3390/vaccines11030682.
(200) Molecular Operating Environment (MOE).
(201) Singh, A.; Jangid, K.; Nehul, S.; Dhaka, P.; Rani, R.; Pareek, A.; Sharma, G.; Kumar, P.; Tomar, S. Structural and Mechanistic Insights into Main Protease (Mpro) Dimer Interface Destabilization Inhibitor: Unveiling New Therapeutic Avenues against SARS-CoV-2. ChemRxiv October 7, 2024. https://doi.org/10.26434/chemrxiv-2023-qf6rb-v3.
(202) Yang, Y.; Zhao, Z.; Li, X.; Chen, Y.; Liu, L.; Zhang, S.-L.; Yang, A. Design and Evaluation of Peptide Inhibitors Targeting the Dimerization of SARS-CoV-2 Main Protease. *ChemBioChem* **2025**, *26* (1), e202400688. https://doi.org/10.1002/cbic.202400688.

Appendix

Table A1: Species of *Betacoronavirus* and their corresponding Mpro reference sequence from the National Center for Bioinformatics Information (NCBI) used for the sequence alignment.[62]

***Betacoronavirus* Specie**	**NCBI Reference Sequence**
Murine hepatitis virus	YP_009915692.1
Human coronavirus HKU1	YP_009944273.1
***Betacoronavirus* HKU24**	YP_009113022.1
Rabbit coronavirus HKU14	YP_009944258.1
Human coronavirus OC43	YP_009924323.1
Bovine coronavirus	NP_150073.3
Equine coronavirus	UOI88117.1
***Betacoronavirus* Erinaceus/VMC/DEU/2012**	YP_009513008.1
MERS coronavirus	YP_009047233.1
Pipistrellus bat coronavirus HKU5	YP_009944358.1
Tylonycteris bat coronavirus HKU4	YP_009944329.1
Rousettus bat coronavirus HKU9	YP_009924387.1
Bat Hp-*Betacoronavirus*/Zhejiang2013	YP_009072438.1
SARS coronavirus (Tor2 isolate)	YP_009944370.1
SARS coronavirus 2	YP_009742612.1

Table A2: Pairwise Sequence Similarity for Species of *Betacoronavirus*. Sequence similarity was calculated using Clustal Omega with strains' reference sequences given in Table A1.

Strain	HKU1	Murine hep	HKU24	HKU14	Human OC43	Bovine	Equine	Erinaceus	MERS	HKU5	HKU4	HKU9	Bat Hp	SARS CoV Tor2	SARS CoV 2
HKU1	100	85.48	82.07	85.81	82.51	84.01	84.01	52.38	55.12	52.81	52.15	48.18	50.68	48.51	49.17
Murine hep	85.48	100	82.76	86.8	82.51	84.01	84.01	51.7	53.8	53.8	53.8	51.16	51.02	50.5	50.83
HKU24	82.07	82.76	100	86.9	86.21	87.24	87.24	51.72	54.14	52.76	53.1	47.59	49.66	49.31	49.66
HKU14	85.81	86.8	86.9	100	91.75	92.86	92.86	51.02	53.8	54.13	52.81	48.51	48.3	48.51	48.51
Human OC43	82.51	82.51	86.21	91.75	100	98.3	98.3	51.7	53.47	52.48	52.15	48.84	48.98	48.51	48.18
Bovine	84.01	84.01	87.24	92.86	98.3	100	100	52.04	54.08	52.72	52.72	48.64	49.32	48.64	48.3
Equine	84.01	84.01	87.24	92.86	98.3	100	100	52.04	54.08	52.72	52.72	48.64	49.32	48.64	48.3
Erinaceus	52.38	51.7	51.72	51.02	51.7	52.04	52.04	100	78.45	76.43	72.73	51.02	51.36	51.02	52.04
MERS	55.12	53.8	54.14	53.8	53.47	54.08	54.08	78.45	100	82.68	81.05	51.49	51.36	51.82	50.83
HKU5	52.81	53.8	52.76	54.13	52.48	52.72	52.72	76.43	82.68	100	83.66	51.82	50.68	50.83	50.17
HKU4	52.15	53.8	53.1	52.81	52.15	52.72	52.72	72.73	81.05	83.66	100	52.15	52.04	51.16	50.17
HKU9	48.18	51.16	47.59	48.51	48.84	48.64	48.64	51.02	51.49	51.82	52.15	100	50.17	51.96	51.96
Bat Hp	50.68	51.02	49.66	48.3	48.98	49.32	49.32	51.36	51.36	50.68	52.04	50.17	100	69.02	68.35
SARS CoV Tor2	48.51	50.5	49.31	48.51	48.51	48.64	48.64	51.02	51.82	50.83	51.16	51.96	69.02	100	96.08
SARS CoV 2	49.17	50.83	49.66	48.51	48.18	48.3	48.3	52.04	50.83	50.17	50.17	51.96	68.35	96.08	100

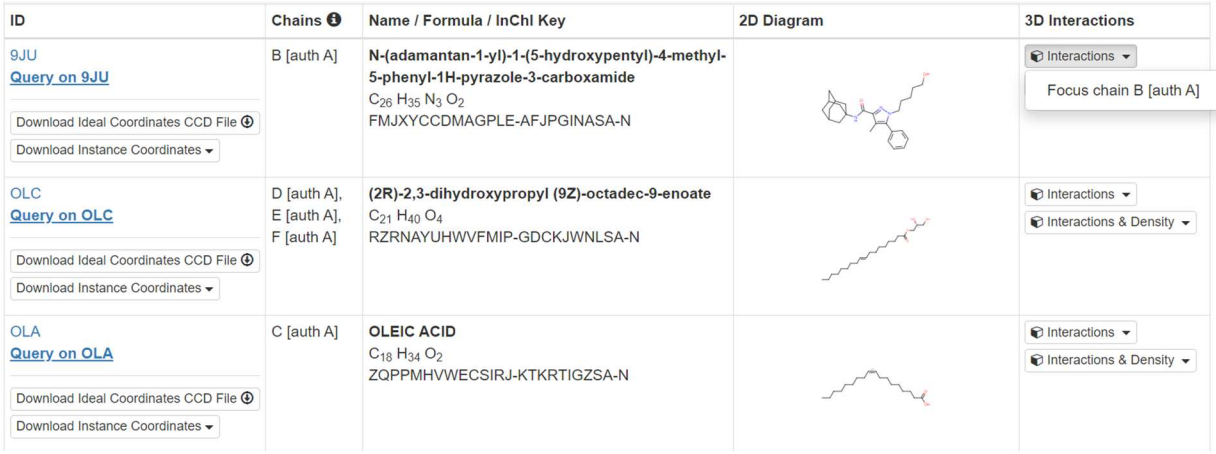


Fig. A1: Segregated ligand interaction analysis in PDB website[60]. These ligands are from the PDB entry 5zty. In the ligand information section for this PDB entry, the ligand engagement can be assessed with its target. However, it is not possible to select multiple ligands for this assessment, as indicated by each ligand entry having its separate 3D interaction option.

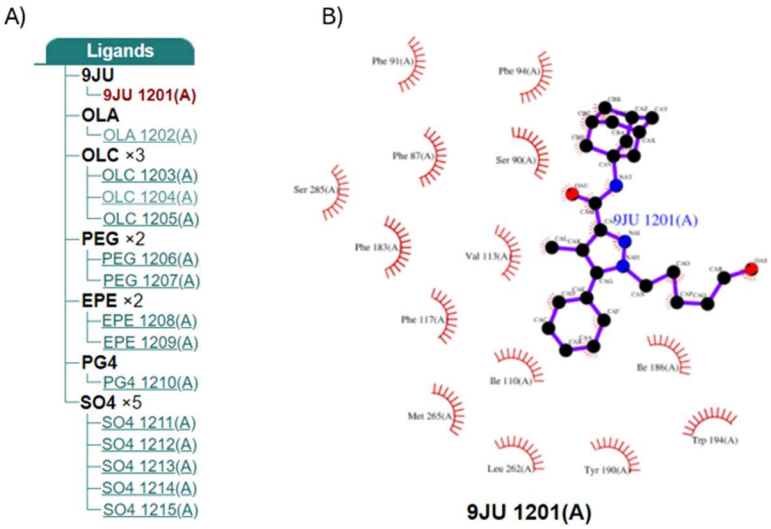


Fig. A2: 2D Ligand interaction generator from PDBsum.[157] The ligand information shown pertains to the PDB Code: 5zty. A) shows the ligand selector, which does not allow for multiple selections. B) shows the 2D-interaction diagram of the chosen ligand, 9JU 1201(A).

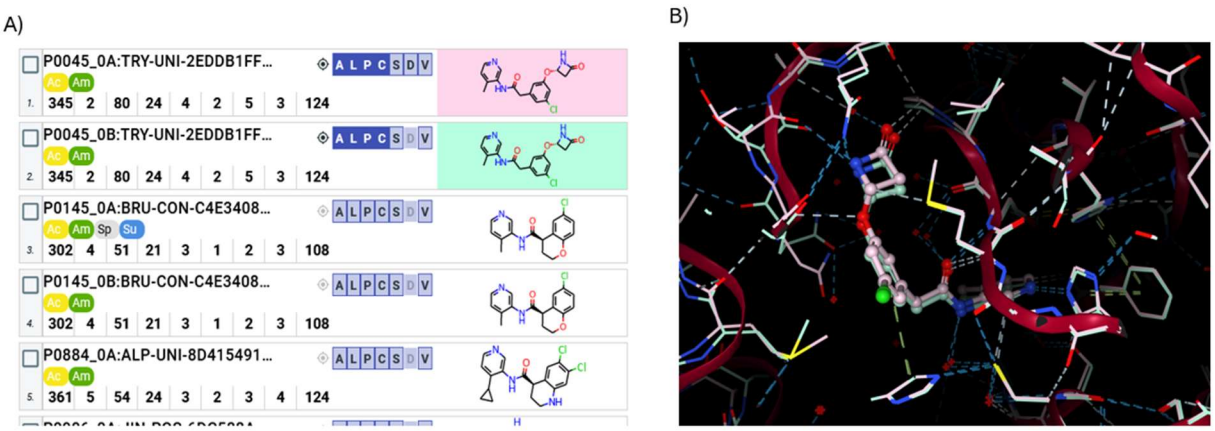


Fig. A3: Architecture of Fragalysis.[158] A) represents the ligand selector table. Multiple ligands may be picked by checking their corresponding checkboxes. B) represents the molecular visualizer, which is currently showing the 3D interactions of the first two ligands in A) with Mpro. The interactions involved are not quoted in text and any comparative analysis has to be made visually. This may become a tedious task if multiple ligands are chosen.

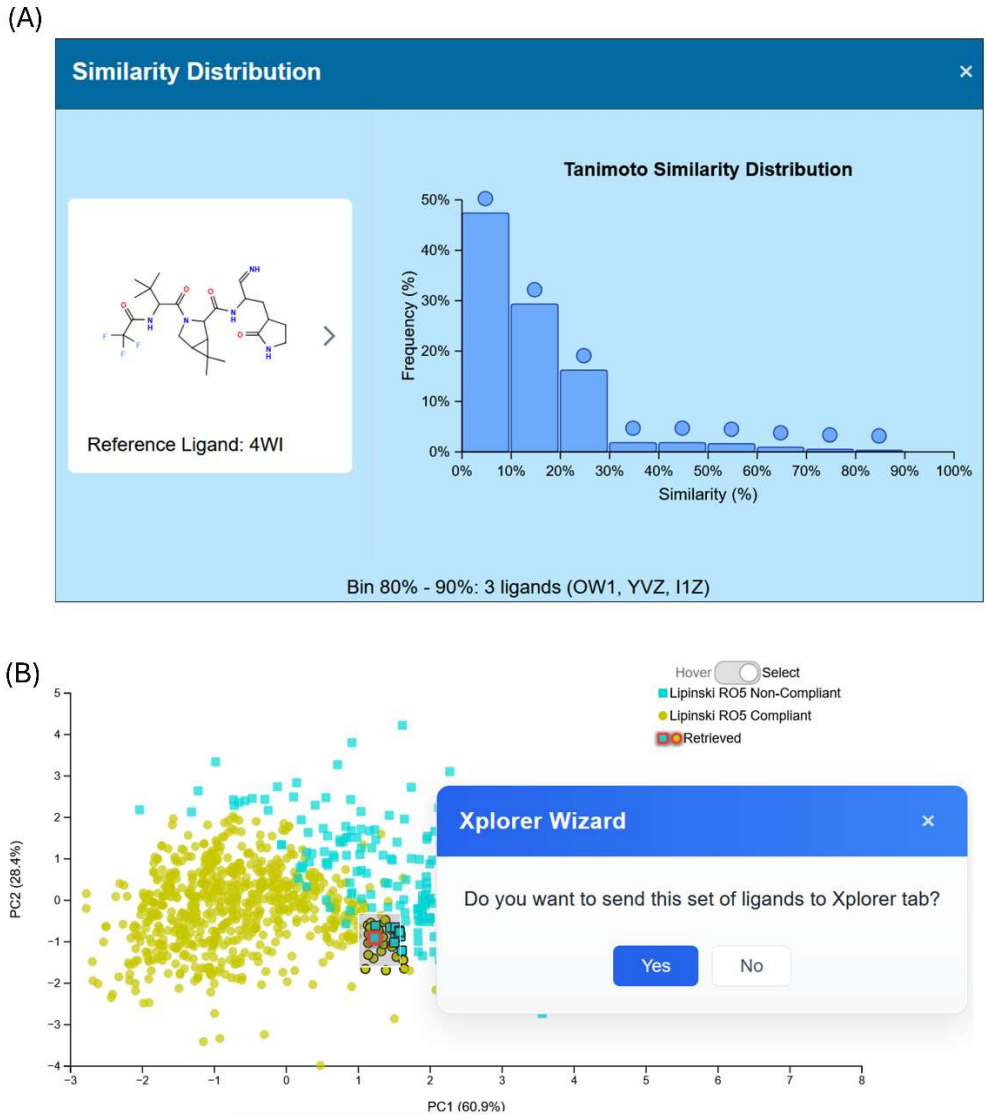


Fig. A4: Interactive Features Implemented in the Xplorer Tab. (A) Tanimoto-based similarity widget for assessing ligand similarity across a loaded dataset in the Xplorer Tab given a ligand of interest. As an example, the similarity was calculated for all SARS-CoV-2 crystal ligands with respect to the ligand 4WI (Nirmatrelvir). **(B)** PCA Decomposition plot from the Dataset tab whereby a group of ligand datapoints were selected directly from the plot and about to be loaded in the Xplorer tab as an additional dataset.

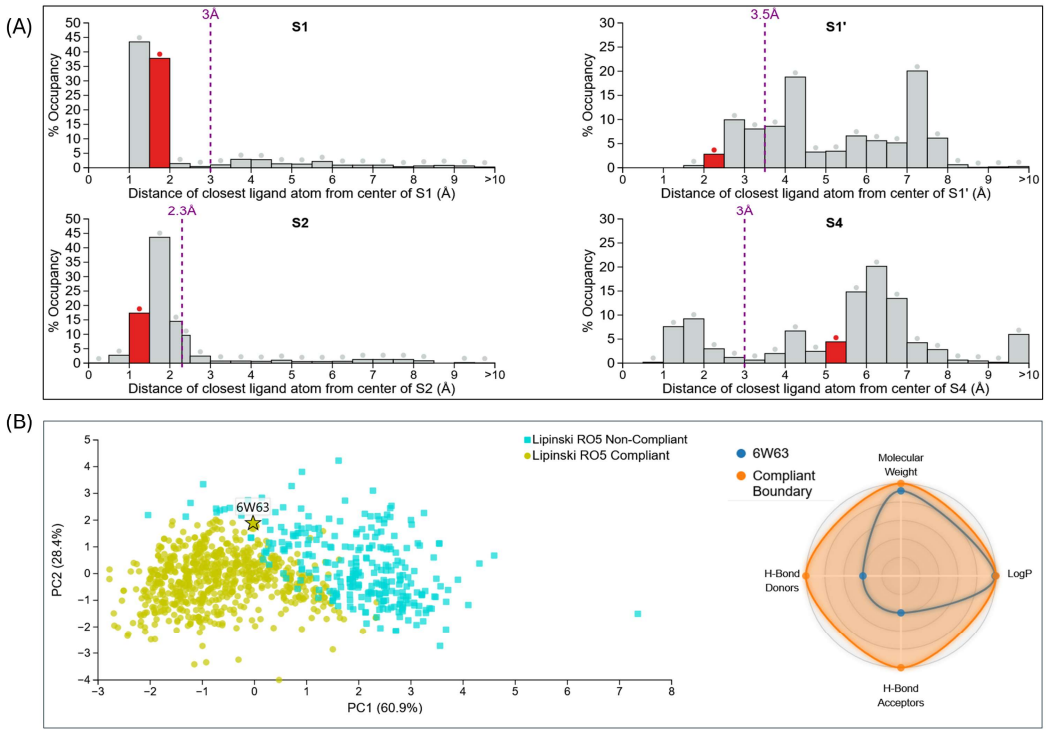


Fig. A5: Characterizing Subpocket Occupancy and Lipinski's Properties of X77 (PDB: 6W63). (A) Distribution of subpocket occupancies for all 4 subpockets across all crystallographically determined ligands with X77's location highlighted in red. **(B)** Left: PCA decomposition plot based on the Lipinski's metrics of the crystal ligands, with the datapoint corresponding to X77 starred. Right: Radar chart assessing compliance of X77 with Lipinski's Rule of 5.

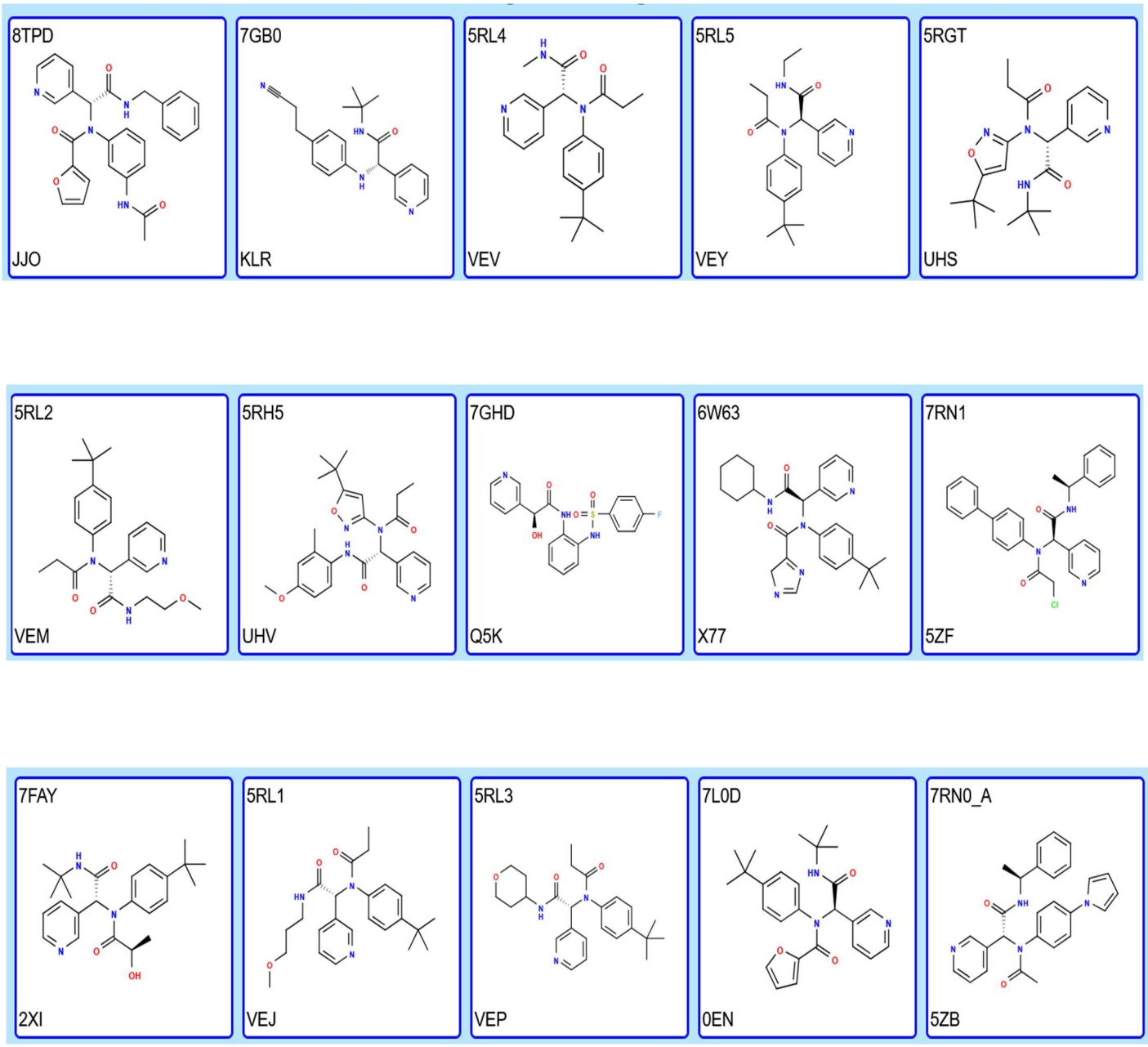


Fig. A6: Part 1 of X77 Analogues used to assess the novel fused compound. A ligand similarity search was conducted using CoviProDigy using the novel fused compound as reference. Part 1 of 2 of the total retrieved compounds is shown here.

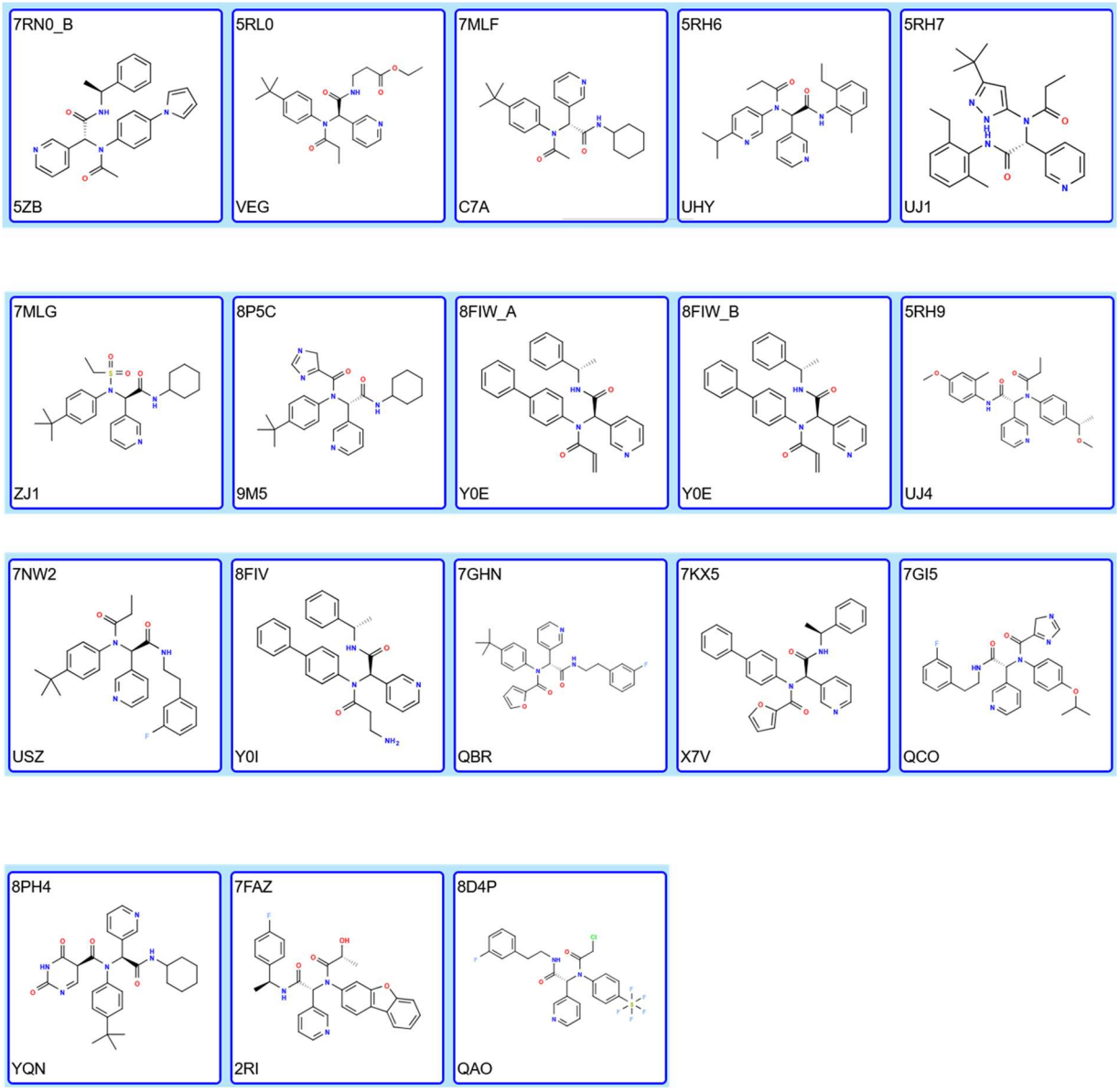


Fig. A7: Part 2 of X77 Analogues used to assess the novel fused compound. A ligand similarity search was conducted using CoviProDigy using the novel fused compound as reference. Part 2 of 2 of the total retrieved compounds is shown here.

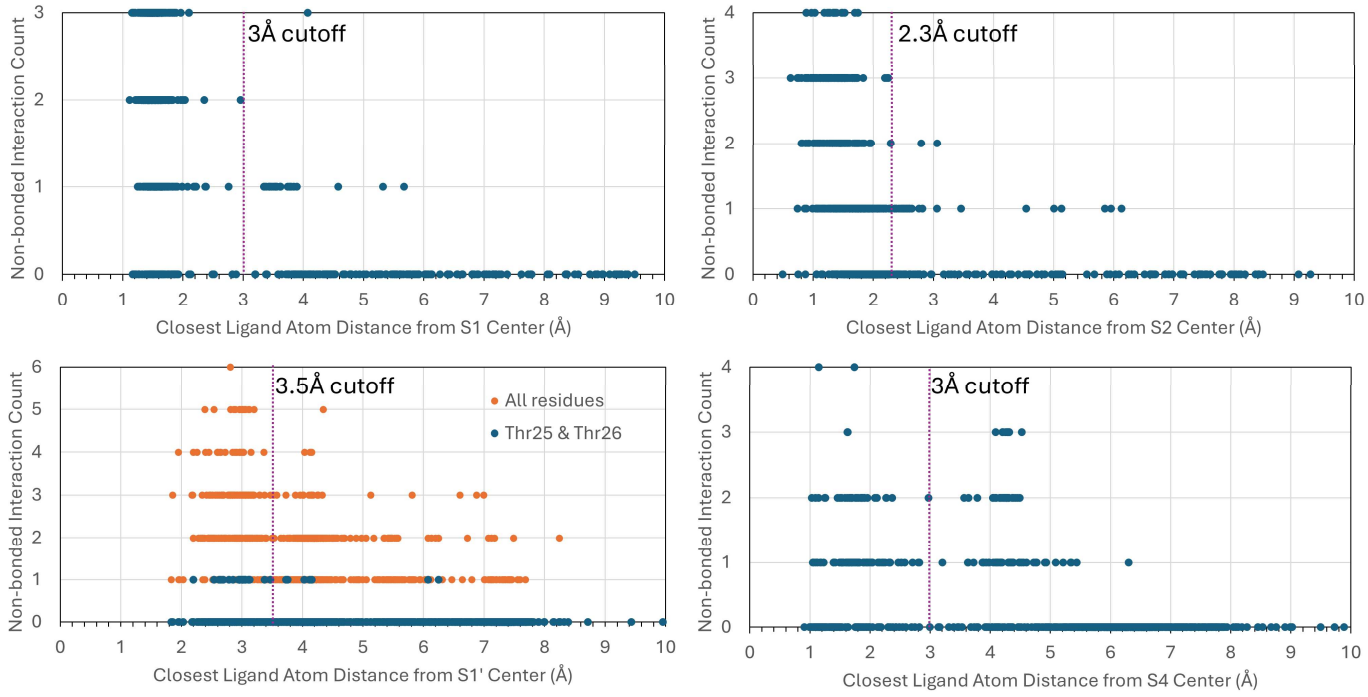


Fig. A8: Optimizing CoviProDigy's Algorithm for Assessing Ligand Subpocket Occupancies. The non-bonded interaction count (hydrogen bond, hydrophobic interaction, and stacking interaction) involving each subpocket is plotted against the distance between that subpocket's geometric center and the nearest ligand atom. For the S1' subpocket, interactions involving either Thr25 or Thr26 are highlighted in orange, while interactions involving any S1' residue are highlighted in blue. Thr25 and Thr26 do not line the entrance of the S1' subpocket, and thus, offer a better reference point for assessing subpocket occupancy.

Table A3: Distribution of the S1' geometric center distances from the catalytic residues His41 (Nε atom) and Cys145 (Sγ atom).

PDB_Codes	His_41_NE2_Dist (Angstroms)	Cys_145_SG_Dist (Angstroms)
5R7Y	5.918908571	5.168711147
5R7Z	4.515304897	5.091234047
5R80	4.603677734	5.13951576
5R81	4.79997909	4.96619374
5R82	4.895132148	5.088788903
5R83	4.861616515	5.151417119
5R84	4.67957623	5.142345076
5RE4	4.856349158	4.980477571
5RE9	4.914435286	5.062848605
5REB	4.809099245	5.165790682
5REH	4.879851596	4.911695052
5REJ	4.736027855	5.33661555
5REK	4.737428858	5.247532587
5REL	4.788136315	5.662489834
5REM	4.818394738	5.455172286
5REN	5.204468068	5.292845784
5REO	4.739204533	5.488864435
5REP	4.74028759	5.261503053
5RER	4.701150863	5.546971178
5RES	4.755283291	5.302052951
5RET	4.775290292	5.640316669
5REU	4.793484725	5.466904197
5REV	4.801222922	5.177821253
5REW	5.907720249	5.173357582
5REX	4.749996418	6.133525187
5REY	4.600440649	5.620622584
5REZ	4.824435859	5.162525259
5RF1	4.855201374	5.150459616
5RF3	4.845477928	5.086670163
5RF6	6.023441508	5.124415154
5RF7	4.740617949	4.891877813
5RF8	4.885485875	5.207162992
5RFE	4.819570074	5.199403453
5RFF	4.709219295	5.256961241
5RFG	4.955491425	5.218777529
5RFH	4.994645612	5.76668401

5RFI	4.8657788	5.579405027
5RFJ	4.959530426	5.397723854
5RFK	4.934068895	5.487124872
5RFL	4.876676689	5.450676245
5RFM	4.789539648	5.364654198
5RFN	4.76573946	5.170754331
5RFO	4.986383613	5.277481286
5RFP	4.69430816	4.202032065
5RFQ	4.85066406	5.453847192
5RFR	4.899259211	5.675552666
5RFS	4.758699818	5.752803715
5RFT	4.771192363	5.553708845
5RFU	4.75030237	5.247293747
5RFV	4.817873353	5.645325142
5RFW	4.827693525	5.765340408
5RFX	4.911958737	5.395249914
5RFY	5.036675636	5.244677918
5RFZ	4.7801868	5.786448881
5RG0	4.707064456	5.499648549
5RG1	5.908150786	4.978763792
5RG2	4.731224101	4.165134264
5RG3	4.803588872	4.121401482
5RGH	4.815968242	5.160062771
5RGI	5.913067614	4.538788486
5RGK	4.885043945	5.069531361
5RGL	4.853429329	5.919407229
5RGM	4.914936176	5.653144578
5RGN	4.744538333	5.229248702
5RGO	4.857775121	5.694690323
5RGP	4.746670375	5.388025696
5RGT	4.639984481	4.903144774
5RGU	4.943047203	4.995003419
5RGV	4.793633346	4.871397287
5RGW	4.825002791	5.026196182
5RGX	4.817674792	4.935759861
5RGY	4.745408448	5.160946101
5RGZ	4.796445272	4.952736523
5RH0	5.011591183	5.054055122
5RH1	4.948175644	5.132596402
5RH2	4.555680766	5.005913846
5RH3	4.568635599	4.875087287
5RH5	4.934430079	4.793763229
5RH6	4.813977777	4.577052249

5RH7	4.866425449	4.779238194
5RH8	4.726422361	4.866698056
5RH9	4.677665514	4.604299523
5RHA	4.876426635	5.728967155
5RHB	4.897796802	4.932404265
5RHC	4.752442496	4.90987909
5RHD	4.904650142	5.088673576
5RHE	4.81651849	5.160486447
5RHF	4.970421696	5.293767558
5RL0	4.891649487	4.885581115
5RL1	4.82075464	4.753427341
5RL2	4.795173197	4.783289563
5RL3	4.765036099	4.83516239
5RL4	4.793317651	4.742144068
5RL5	4.685636251	4.667203012
5SMM	4.729008309	4.714603651
5SMN	4.8178811	5.170547558
6LZE	5.940195204	5.016973796
6M0K	5.999669362	5.167407978
6M2N	5.954905667	5.300408074
6W63	5.501749334	4.331285534
6WNP	5.062471318	4.830752609
6WTJ	4.859119154	5.404011138
6WTK	4.776497417	5.307077338
6WTT	4.780716958	5.219747085
6XB2	4.600962396	5.23327436
6XHM	5.53361312	5.23499173
6XMK	4.633799571	5.069516862
6XQS	5.507850652	5.035246089
6XQT	5.553227504	5.10854705
6XQU	5.274806417	5.150103693
6XR3	7.500127381	5.151793725
6Y2F	4.834742491	5.100941091
6Y2G	5.011655792	5.247326077
6YNQ	5.692625142	4.016759473
6Z2E	4.502821299	5.458296938
7A1U	5.68718304	4.944990837
7ADW	5.832894939	4.178152728
7AEH	4.992123811	2.872483511
7AHA	5.843621682	5.189619123
7AK4	4.903403052	4.879784258
7ANS	5.792186718	5.105500744
7AP6	5.673821812	4.898311571

7AQE	5.74867416	5.102860544
7AQJ	5.717985589	4.764824116
7AU4	4.8452675	4.547849699
7AVD	5.636438859	4.823202511
7AWS	5.109499846	3.875913404
7AWU	5.436352569	5.029303418
7AY7	4.81671673	5.232070382
7B2J	4.698332682	4.871658879
7B2U	4.717162267	4.94981933
7B5Z	4.718524699	4.95698319
7B77	4.760038465	4.949288983
7B83	4.651019174	5.095088689
7BE7	5.524492258	5.039049434
7BGP	4.745365595	5.025912575
7BIJ	4.760794955	5.030731393
7BRP	5.285523042	5.244852448
7BUY	5.701795118	4.877808608
7C6S	5.248275131	5.0348744
7C6U	4.661104974	5.402614265
7C7P	5.27871976	4.961955669
7C8R	4.894028901	5.761959336
7C8T	5.700887873	5.107160394
7C8U	4.779149023	5.151934286
7CA8	6.535687095	4.363550389
7CB7	4.574631649	5.276986087
7CBT	5.642929157	5.186517089
7COM	5.238590681	4.991016919
7CX9	8.316441153	4.381132796
7D1M	4.694389673	5.275594199
7D1O	5.400701732	5.237890619
7D3I	4.777546836	5.702936022
7DDC	5.351350264	4.564995382
7DGB	5.614231529	5.36973092
7DGF	5.816066861	4.972723359
7DGG	5.893424453	5.216256821
7DGH	4.833498562	5.487054894
7DGI	4.709204232	5.231237957
7DHJ	5.735542479	5.312567962
7DPP	8.00132782	4.400188548
7DPU	5.34302801	4.32592938
7DPV	5.946424241	4.926784494
7E19	7.465149294	5.211887585
7E6K	5.850596929	5.016415277

7EN8	5.91095277	5.016049155
7EN9	5.820887685	4.788225443
7FAY	6.173953311	4.863315888
7FAZ	5.874152425	4.732546259
7GAV	4.97223387	4.948081544
7GAW	4.926924906	4.9437786
7GAX	4.802396613	5.067781059
7GAY	4.734981508	5.121139186
7GAZ	4.802088982	5.006780979
7GB0	4.843113074	5.017758021
7GB1	4.898240463	5.351870118
7GB2	4.874892345	5.433226729
7GB3	4.704188523	3.887921228
7GB4	4.823222595	5.090775685
7GB5	4.494972877	4.855128993
7GB6	4.43353758	5.030467061
7GB7	4.628438781	4.984394201
7GB8	4.635381153	4.935815598
7GB9	4.606993879	5.055453507
7GBA	4.842145049	5.444217889
7GBB	4.745340398	5.570689101
7GBC	4.731945291	5.019694935
7GBD	4.792144045	5.174201489
7GBE	4.559281526	4.847643774
7GBF	4.722688458	5.079730993
7GBG	4.634579872	5.024258058
7GBH	4.854055383	4.613044254
7GBI	4.649248541	4.917927624
7GBJ	4.679587773	5.071069979
7GBK	4.539101957	5.0843694
7GBL	4.792683581	5.063106615
7GBM	4.659577262	5.030190325
7GBN	4.875957234	5.200911622
7GBO	4.678539552	5.077466324
7GBP	4.74725648	5.113386276
7GBQ	4.478419729	5.03762791
7GBR	4.566771884	5.010174904
7GBS	4.560524321	4.989140641
7GBT	4.801696644	5.077074832
7GBU	4.735967557	5.253181989
7GBV	4.726786521	4.978286414
7GBW	4.627860371	4.946197964
7GBX	4.586536119	4.788807825

7GBY	4.676315447	5.01543602
7GBZ	4.652757185	5.146041189
7GC0	4.638269014	5.135700166
7GC1	4.690318296	5.143735289
7GC2	4.606528813	5.005696357
7GC3	4.69583094	4.936351066
7GC4	4.661860745	4.480495031
7GC5	4.621712043	5.578025134
7GC6	4.724772984	5.526938755
7GC7	4.757902018	5.05752076
7GC8	4.623357322	4.802038479
7GC9	4.506938952	5.14515441
7GCA	4.640629893	5.00432052
7GCB	4.704794471	4.972174456
7GCC	4.683694052	5.059036912
7GCD	4.875774784	4.99964118
7GCE	4.782657134	5.04789616
7GCF	4.659258381	4.971657895
7GCG	5.786211837	4.975434088
7GCI	4.765869907	5.087869695
7GCJ	4.777016338	5.26899785
7GCK	4.682879005	4.909711749
7GCL	4.911147311	4.947926122
7GCM	4.723805437	5.477245857
7GCN	4.749857799	4.877655737
7GCO	4.651017117	4.988832601
7GCP	4.575870733	5.395553641
7GCQ	4.483897114	4.889274219
7GCR	4.798075347	5.164480599
7GCS	4.718292336	4.965512468
7GCT	4.962654857	4.958506451
7GCU	4.6332565	4.936535126
7GCV	4.840907657	4.759209606
7GCW	4.710144502	4.997863041
7GCX	4.673811565	4.967767542
7GCY	4.852997425	4.729122323
7GCZ	5.801351156	4.980456523
7GD0	4.720153477	4.893800214
7GD1	4.774896704	5.054104035
7GD2	4.933106199	4.999497215
7GD3	4.798532342	5.026723984
7GD4	4.825053721	5.079277831
7GD5	4.695814542	4.962360951

7GD6	4.690499769	4.983325103
7GD7	4.51870926	5.02960137
7GD8	4.666501153	5.014511736
7GD9	4.819829824	5.256715636
7GDA	4.719871971	4.897780166
7GDB	4.780289207	4.949752755
7GDC	4.575858907	4.951202959
7GDD	4.589162402	4.751362001
7GDE	4.544121255	4.730791555
7GDF	4.764736332	4.971347969
7GDG	4.787969503	4.92095374
7GDH	4.715096426	5.228232732
7GDI	4.628009812	5.074515612
7GDJ	4.59549097	4.986555074
7GDK	4.497534428	4.809747368
7GDL	4.840860594	4.900840953
7GDM	4.661050288	5.008287186
7GDN	4.755746325	5.131153981
7GDO	4.736474184	5.11813392
7GDP	4.605710702	5.003844392
7GDQ	4.751356176	5.047852818
7GDR	4.86415546	5.143013042
7GDS	4.996641512	5.788157341
7GDU	4.840347229	5.049319458
7GDV	4.846511173	5.109448801
7GDW	4.614114086	5.190758427
7GDX	4.726938541	5.093921902
7GDY	5.016824652	5.318661894
7GDZ	4.552822006	4.797502003
7GE0	4.644505175	5.010323762
7GE1	4.717107723	4.963918458
7GE2	4.675867208	5.002279102
7GE3	4.542878974	4.784286088
7GE4	4.674623366	4.927227611
7GE5	4.678202603	5.252423164
7GE6	4.588183239	5.034295314
7GE7	4.616307409	5.060149698
7GE8	4.89999548	5.040082001
7GE9	4.582099637	4.985866358
7GEA	4.681774208	4.944432252
7GEB	4.710006022	5.021840223
7GEC	4.672339944	5.02146063
7GED	4.647342641	4.916443646

7GEE	4.735121122	5.043174742
7GEF	4.783455799	5.275547307
7GEG	4.720407382	4.877556801
7GEH	4.836059096	5.005359985
7GEI	4.862976665	4.804556436
7GEJ	4.64419178	4.853425585
7GEK	4.630277934	4.75314058
7GEL	4.807554673	5.073608159
7GEM	4.713208159	5.061915531
7GEN	4.748972406	4.991599748
7GEO	4.626551226	4.990526074
7GEQ	4.64362965	4.859175485
7GER	4.505041107	4.875264377
7GES	4.706617566	5.160255363
7GET	4.718213571	4.888846016
7GEU	4.613967022	4.878438601
7GEV	4.807498194	5.076547789
7GEW	4.577053728	4.880237041
7GEX	4.585571658	4.917731419
7GEY	4.685623169	5.176109553
7GEZ	4.933439974	5.052753306
7GF0	4.577401689	4.886154179
7GF1	4.599853019	4.819550775
7GF2	4.638310032	4.900657476
7GF3	4.484887707	4.897709892
7GF4	4.806875201	5.162174089
7GF5	4.666962782	4.988809407
7GF6	4.821530354	4.941814857
7GF7	5.009772421	5.686057374
7GF8	4.895998877	5.434563144
7GF9	4.851141602	5.374422628
7GFA	4.407039113	5.036099128
7GFB	4.582466532	4.842968926
7GFC	4.752279173	4.864477666
7GFD	4.810880832	5.031667617
7GFE	4.76360756	4.874643972
7GFF	4.866029648	5.091298364
7GFG	4.669859375	4.819492007
7GFH	4.656838184	4.88729763
7GFI	4.803352327	4.875906866
7GFJ	4.767044272	5.006022539
7GFK	4.756416218	5.0068135
7GFL	4.739584705	4.950071481

7GFM	4.976115447	4.917971895
7GFN	4.852333706	4.930630869
7GFO	4.658359278	4.958253748
7GFP	4.695125088	4.873696085
7GFQ	4.72741413	4.977832898
7GFR	4.680121525	4.846186947
7GFS	4.815587376	4.938655348
7GFT	4.693263358	4.796002425
7GFU	4.695560552	4.838570207
7GFV	4.816476511	5.641135246
7GFW	4.764767319	5.102816377
7GFX	4.594035208	4.936582914
7GFY	4.866631859	5.635584328
7GFZ	4.848689526	5.131970844
7GG0	4.685502136	4.9766731
7GG1	4.911770869	5.180674257
7GG2	4.529486759	5.132192726
7GG3	4.835876754	4.96485804
7GG4	4.607829009	4.949258952
7GG5	4.885879123	4.924247023
7GG6	4.826675325	4.90871228
7GG7	4.802005405	4.947715373
7GG8	4.867475201	4.91731471
7GG9	4.728870707	4.841236461
7GGA	4.810983746	4.888496322
7GGB	4.789076556	4.907416901
7GGC	4.937904496	4.779025548
7GGD	4.714297087	5.061759306
7GGE	4.738766447	5.047046481
7GGF	4.578848745	5.077400795
7GGG	4.753996061	4.790874678
7GGH	4.712823324	5.023172459
7GGI	4.807513314	4.851865293
7GGJ	4.833754817	4.88393837
7GGK	4.688463117	4.855005692
7GGL	4.723154144	4.672104271
7GGM	4.809422597	4.717589728
7GGN	4.841875789	4.871427893
7GGO	4.926586275	5.02212388
7GGP	4.957893186	4.80424629
7GGQ	4.70059114	4.919882824
7GGR	4.584519691	4.796664679
7GGS	4.773130648	5.09091882

7GGT	4.621607448	5.024261298
7GGU	4.624297533	4.859559817
7GGV	4.585486933	4.731286067
7GGW	4.90201603	5.18766988
7GGX	4.753602457	5.008800281
7GGY	4.713672906	4.849670545
7GGZ	4.879967976	5.075334608
7GH0	4.708112679	5.090407001
7GH1	4.83679721	5.237352999
7GH2	4.766775927	4.901826061
7GH3	4.60984299	4.995421043
7GH4	4.555580527	4.563450347
7GH5	4.773056685	5.12436421
7GH6	4.791939076	5.03857386
7GH7	4.581737242	5.009998462
7GH8	4.66346594	5.013075585
7GH9	4.706722146	5.061453008
7GHA	4.962616318	5.358319726
7GHB	4.674151625	5.085109897
7GHC	5.914694336	5.114590058
7GHD	4.906935017	5.094968806
7GHE	4.628604243	5.107815603
7GHF	4.592159104	5.048700234
7GHG	4.685757173	3.944119844
7GHH	4.866569435	5.731871199
7GHI	4.887142877	2.812947009
7GHJ	4.840548941	5.660783111
7GHL	4.831259358	5.097564908
7GHM	5.755935016	4.836590429
7GHN	5.654673491	4.495880174
7GHO	5.71990354	4.929588821
7GHP	5.615688417	5.011287941
7GHQ	5.67889483	4.78623587
7GHR	5.817818032	4.912888372
7GHS	5.677672189	4.867330394
7GHT	5.791658105	4.987525399
7GHU	5.667830795	4.916555392
7GHV	5.590040512	4.919984793
7GHW	5.941870981	4.922591417
7GHX	5.713914875	5.020968763
7GHY	5.753182894	4.927259821
7GHZ	5.667257311	4.842393429
7GI0	5.656386903	4.948853262

7GI1	5.681917327	4.996402594
7GI2	5.686082299	4.954271671
7GI3	5.755861045	4.709131906
7GI4	5.727755542	4.884779017
7GI5	5.84321991	4.473766043
7GI6	5.628274045	4.868908166
7GI7	5.926416525	4.739633498
7GI8	5.630632444	4.867418884
7GI9	5.722157031	4.87608205
7GIA	5.810232098	4.676249923
7GIB	5.842843096	4.884918155
7GIC	5.829877948	4.682593667
7GID	5.783941095	4.869269616
7GIE	5.786390715	4.799345789
7GIF	5.772506441	4.854961546
7GIG	5.765592628	4.763455367
7GIH	5.660460767	4.860816775
7GII	5.740540021	4.728234797
7GIJ	5.72627682	4.851785834
7GIK	5.681808993	4.607755473
7GIL	5.620635495	4.651666495
7GIM	5.75349051	4.60067283
7GIN	5.644533178	4.689161051
7GIO	5.610625382	4.672128984
7GIP	5.63252112	4.78637559
7GIQ	5.696891019	4.792768273
7GIR	5.530103887	4.802066788
7GIS	5.68432908	4.814032058
7GIT	5.617205835	4.748593089
7GIU	5.55847512	4.820421879
7GIV	5.704810876	4.71454996
7GIW	5.72047971	4.918846364
7GIX	5.22886392	4.730511365
7GIY	5.814375428	4.863421448
7GIZ	5.691965496	4.816009521
7GJ0	5.75862703	4.834648477
7GJ1	5.82030614	4.749553091
7GJ2	5.770133416	4.818964925
7GJ3	5.726475232	4.848510585
7GJ4	5.696926731	4.776473025
7GJ5	5.723990545	4.6166136
7GJ6	5.651256007	4.609756388
7GJ7	5.739541143	4.787083563

7GJ8	5.759187126	4.748637474
7GJ9	5.633180785	4.720607364
7GJA	5.58482704	4.520334508
7GJB	5.739015422	4.731298317
7GJC	4.837395184	4.777966452
7GJD	5.579375425	4.786528368
7GJE	5.637380241	4.494111861
7GJF	5.820910292	4.887372857
7GJG	5.79652916	4.826422316
7GJH	5.62193453	4.730734713
7GJI	5.908517011	4.736961792
7GJJ	5.777121731	4.878444942
7GJK	5.736221679	4.763794472
7GJL	4.938331669	4.486102476
7GJM	5.718942287	4.752794869
7GJN	5.778387542	4.743052817
7GJO	5.812377195	4.611932448
7GJP	5.714384263	4.773315469
7GJQ	5.653251762	4.683092313
7GJR	5.990489515	4.770347596
7GJS	5.816936102	4.750237276
7GJT	5.875367388	4.827683244
7GJU	5.645421984	4.673629465
7GJV	5.758213075	4.723560264
7GJW	5.866294938	4.953702446
7GJX	5.710052655	4.783694918
7GJY	5.899942928	4.842168651
7GJZ	5.695519519	4.795218761
7GK0	5.753437752	4.879901238
7GK1	5.463372699	4.637646797
7GK2	6.003194918	4.84037774
7GK3	5.770701712	4.685389264
7GK4	5.803951734	4.709860745
7GK5	5.653471941	4.780179231
7GK6	5.646435354	4.742710435
7GK7	4.992667219	4.888094947
7GK8	5.707552682	4.778862161
7GK9	5.685342979	4.805462065
7GKA	5.720682167	4.809245076
7GKB	5.578978079	4.729961697
7GKC	5.747433331	4.821915603
7GKD	5.748297774	4.603498057
7GKE	5.578452882	4.783130799

7GKF	5.607292368	4.67201685
7GKG	5.748806664	4.762681406
7GKH	5.674588111	4.718127625
7GKI	5.696839149	4.749254577
7GKJ	5.994309554	4.846286551
7GKK	5.636531359	4.838982487
7GKL	5.730160857	4.872494062
7GKM	5.744349488	4.767798179
7GKN	5.839432566	4.840997554
7GKO	5.660072522	4.775339112
7GKP	5.718284223	4.786883298
7GKQ	5.667688777	4.837457959
7GKR	5.751056017	4.885665218
7GKS	5.633630359	4.870223586
7GKT	5.797587783	4.845115205
7GKU	5.683614936	4.866182955
7GKV	5.741962027	4.818561195
7GKW	5.712476633	4.79144908
7GKX	5.714906907	4.917804232
7GKY	5.71801862	4.807092586
7GKZ	6.054519891	4.878242703
7GL0	5.614637647	4.770847949
7GL1	5.746437659	4.894217477
7GL2	5.685696825	4.867706738
7GL3	4.993871675	4.910867865
7GL4	5.679292391	4.828719324
7GL5	5.729310118	4.862597292
7GL6	5.781024283	4.953077391
7GL7	5.774607888	4.741535835
7GL8	5.716849124	4.93120238
7GL9	5.259400562	4.680749542
7GLA	5.635712904	4.842427903
7GLB	5.654833789	4.932395431
7GLC	5.925071552	4.946348567
7GLD	5.661397852	4.802136544
7GLE	5.632798669	4.846526808
7GLF	5.604945311	4.833862495
7GLG	5.566306227	4.86789725
7GLH	5.763834287	4.91833816
7GLI	5.667409393	4.823527557
7GLJ	5.67603859	4.925873731
7GLK	5.603065307	4.901470094
7GLL	5.655875591	4.856556478

7GLM	5.4955073	4.825671499
7GLN	5.594747422	4.843732551
7GLO	5.64972615	4.810362552
7GLP	5.694996868	4.907454108
7GLQ	5.655486604	4.801676762
7GLR	5.891790603	4.956856244
7GLS	5.654226772	5.061260356
7GLT	5.558964653	4.923219083
7GLU	5.671686314	4.861449503
7GLV	5.626044266	4.974539748
7GLW	5.847304547	4.909238793
7GLX	5.572421038	4.92562438
7GLY	5.598755037	4.840456051
7GLZ	5.559196583	4.924731308
7GM0	5.594800488	4.949303227
7GM1	5.718083672	4.881951199
7GM2	5.526471135	4.84411313
7GM3	5.539922746	4.873462363
7GM4	5.661603814	4.871732828
7GM5	5.662076853	4.857322389
7GM6	5.695112102	4.809591307
7GM7	5.633523598	4.861740687
7GM8	5.632047401	4.812141174
7GM9	5.538132997	4.768251798
7GMA	4.845836516	4.796422942
7GMB	5.79668686	4.844264572
7GMC	5.574909674	4.829989188
7GMD	5.631116272	4.818316398
7GME	5.686893602	4.843432766
7GMF	5.644069314	4.94483194
7GMG	5.650646295	4.93654885
7GMH	5.727819109	4.854202483
7GMI	5.780025206	4.797961599
7GMJ	5.609792487	4.884152204
7GMK	5.721698924	4.860342305
7GML	5.640917177	4.821667188
7GMM	5.649393886	4.884866531
7GMN	5.476470261	4.380733902
7GMO	5.731370804	4.897948274
7GMP	5.687474826	4.74627832
7GMQ	5.664040029	4.679012684
7GMR	5.642426684	4.868909889
7GMS	5.879007318	4.880258937

7GMT	5.701400272	4.830274417
7GMU	5.717156611	4.818737604
7GMV	5.65054339	4.951664023
7GMW	5.742977708	4.943662226
7GMX	5.691261784	4.816502213
7GMY	5.69591135	4.880138047
7GMZ	5.925001314	4.824560123
7GN0	5.688986014	4.825899216
7GN1	5.630541071	4.844821669
7GN2	4.759700409	4.934322514
7GN3	5.871846929	4.832801683
7GN4	5.693003715	4.974459926
7GN5	5.660145654	4.966822382
7GN6	5.630404171	4.858904726
7GN7	5.861035076	4.676992741
7GN8	5.598082625	4.789290215
7GN9	5.585576436	4.892979875
7GNA	5.803447629	4.806777891
7GNB	5.613001952	4.967829969
7GNC	4.776273266	5.216433827
7GND	5.764498018	4.98284338
7GNE	5.607491952	4.876532309
7GNF	5.939549436	4.969864975
7GNG	5.778686634	4.945876366
7GNH	5.710044395	4.924720429
7GNI	4.803573843	4.966960096
7GNJ	4.783208807	5.216689161
7GNK	4.794090274	5.066489896
7GNL	4.770741603	4.833114628
7GNM	4.823857049	4.860173474
7GNN	4.871404435	4.904371708
7GNO	4.873599752	4.928607389
7GNP	4.811589649	4.848589963
7GNQ	4.912424092	4.853162181
7GNR	4.720483428	5.283502033
7GNS	5.834428973	5.02318346
7GNT	4.915930319	5.026597553
7GNU	4.818623827	5.109269788
7GRE	5.895977915	4.882921571
7GRF	5.740907868	4.838689139
7GRG	5.934651226	4.882693613
7GRH	6.004003389	4.745747967
7GRI	6.020761196	4.769125791

7GRJ	6.019978609	4.904641037
7GRK	5.910097164	4.760393423
7GRL	5.926694471	4.889104824
7GRM	5.879456091	4.914621138
7GRN	5.760182545	4.753189772
7GRO	6.081518348	4.904644104
7GRP	5.99920998	4.810836051
7GRQ	5.833832542	4.797549284
7GRR	5.932815149	4.88048208
7GRS	5.801035341	4.90064834
7GRT	5.945667111	4.793578278
7GRU	5.880059532	4.849263878
7GRV	5.999419711	4.847859836
7GRW	5.720070504	4.78079849
7GRX	5.900762936	4.927745856
7GRY	5.859583696	4.892436992
7GS2	5.766403597	4.856692812
7GS3	5.757983733	4.677812634
7GS4	5.856026154	4.785778902
7GS5	4.758861142	5.061722989
7GS6	6.067601377	4.81646657
7HUC	6.042730422	5.029586142
7HUD	4.839826709	5.096973012
7HUE	4.82378372	5.134653159
7JKV	7.701497349	5.405730481
7JP0	4.936956933	5.303869137
7JPZ	6.073476987	5.276804225
7JQ0	4.93359056	5.395312874
7JQ1	4.930182101	5.101697485
7JQ2	4.929032972	5.196670332
7JQ3	5.556209826	5.112678469
7JQ4	4.914787719	5.150953353
7JQ5	4.919217214	5.00966517
7JSU	4.793179854	5.302119657
7JT0	5.221457594	4.655785289
7JT7	4.652268028	5.377407001
7JU7	4.840558073	5.084425375
7JW8	4.800414112	5.403672879
7JYC	5.462237688	5.120073504
7K0E	4.766928808	5.130209537
7K0F	4.96501264	5.24852355
7K40	5.423027851	5.227859835
7K6D	5.460790325	4.989128146

7K6E	5.415597499	4.938953776
7KVL	4.629619131	4.676985627
7KX5	5.653002524	4.683820886
7KYU	7.390722612	4.168068581
7L0D	5.514429083	4.617780722
7L10	5.746757566	4.966215582
7L11	5.657658803	5.076419562
7L12	4.907043177	5.080305335
7L13	5.679124164	4.76292375
7L14	5.794766824	5.117049124
7L5D	5.750536153	4.985989952
7L8I	9.052536071	5.266872437
7L8J	8.599121622	5.255426584
7LB7	5.352295656	5.176314668
7LCO	5.556619323	5.258172253
7LCR	5.417601393	5.053382062
7LCS	4.668920754	5.289482004
7LCT	4.357070949	5.375785948
7LDL	4.38595154	5.364187014
7LFE	4.920035413	4.380341506
7LKE	5.804136508	5.031721299
7LKR	4.660029386	5.103636419
7LKS	4.782174588	5.162564035
7LKT	4.709637617	5.254355525
7LKU	4.671184502	5.081085559
7LKV	4.668721329	5.198933602
7LKW	4.72408708	5.247910827
7LKX	4.697558318	5.242174355
7LMD	5.976524744	5.042348773
7LME	5.745440503	4.962031672
7LMF	4.4739048	4.988377215
7LTJ	6.064097666	5.005176821
7LTN	4.899133076	4.392782639
7LYH	4.698969871	5.630036703
7LYI	5.791462609	5.144111375
7LZT	4.726754418	5.478411176
7LZU	4.719857766	5.244706083
7LZV	4.636248944	5.211942214
7LZW	4.748358224	5.058569741
7LZX	4.63785607	5.101774004
7LZY	4.824676652	5.577432007
7LZZ	4.899598398	5.565979139
7M00	4.820178759	5.482058385

7M01	4.707428324	5.283663666
7M02	4.794333264	5.205794885
7M03	4.715826711	5.107053365
7M04	4.811365897	5.262799319
7M8M	5.692078973	4.904343181
7M8N	5.647182505	4.738724881
7M8O	5.334241768	4.950736907
7M8P	5.520937574	4.71231137
7M8X	4.706093032	5.008988353
7M8Y	5.880602435	5.019424213
7M8Z	5.57250548	4.855774045
7M90	4.637197802	4.91014774
7M91	5.553731765	4.749222682
7MAT	5.2647714	5.62279139
7MAU	4.865317847	4.971232151
7MAV	5.061779494	5.218159017
7MAW	4.781995352	5.597093167
7MAX	4.664623969	5.160866963
7MAZ	5.809143095	5.26499528
7MB0	4.88151346	5.311295973
7MB1	4.662313131	5.151699148
7MB2	6.150295941	5.366926343
7MB3	5.731550361	5.176271024
7MBI	5.446956728	4.840743401
7MLF	4.787402821	3.214699493
7MLG	4.958875377	3.495083789
7MNG	4.847178001	5.240506799
7MPB	5.477075441	4.958273592
7N44	4.753481998	5.12108175
7N8C	6.282404391	4.93888194
7NBR	5.482307649	5.116739181
7NBS	5.426628885	5.03672167
7NBT	4.765377318	4.811454339
7NEO	4.778751141	4.936214145
7NF5	4.540846823	5.153539928
7NG3	4.57553488	5.024374391
7NG6	4.532544371	5.057038336
7NT1	4.943087314	4.806316594
7NT2	5.659698516	4.995432405
7NT3	4.78866898	5.101687279
7NTQ	5.834647146	4.986911609
7NTV	4.809727725	4.853041653
7NUK	4.652679045	5.117586915

7NW2	5.738390043	4.682631282
7O46	4.879261848	4.52807742
7P2G	5.092592314	4.238604719
7P35	6.018760104	5.31955978
7P51	5.767656767	4.877270044
7Q5E	4.808177138	5.222850185
7Q5F	5.768940251	5.157535708
7QBB	4.759486974	4.549656773
7QKA	4.79045899	5.474139332
7QL8	5.207567024	4.894339281
7QT5	5.737611946	5.004836221
7QT6	4.832638186	4.572413103
7QT7	4.477346008	4.659844139
7QT8	4.872409146	2.904943827
7R7H	4.578741748	5.134245736
7RBZ	5.194355385	4.317163279
7RC0	5.503397194	4.354303566
7RFR	7.64607274	5.281861968
7RFS	4.957639601	5.398019577
7RFU	7.852088707	5.288092023
7RFW	4.941854302	5.451966412
7RLS	5.96834396	5.025313765
7RM2	6.046900362	4.952120251
7RMB	6.012861777	4.927454598
7RME	6.048521628	4.906741568
7RMT	6.011877187	4.987355606
7RMZ	6.089151969	4.956372598
7RN0	5.736111299	4.92275597
7RN1	5.824001185	4.862818825
7RN4	6.033542499	4.945414956
7RNH	6.023348866	4.850363067
7RNK	5.676730476	4.92593563
7RVM	4.714595812	5.333938535
7RVN	4.731919449	5.221783225
7RVO	5.520033643	5.214830683
7RVP	5.797893078	4.909168708
7RVQ	4.907761719	5.218904629
7RVR	4.762315218	5.087346016
7RVS	4.690251074	5.16986262
7RVT	4.703130074	5.261850821
7RVU	4.831205765	5.178875622
7RVV	7.09763958	5.255332986
7RVW	5.703282438	4.945334348

7RVX	5.530118595	5.189784478
7RVY	4.964643914	5.004355747
7RVZ	4.716897705	5.271203422
7RW0	4.910840461	4.988339738
7RW1	5.51047444	5.318084899
7S3K	4.659501342	4.837399289
7S3S	4.864344169	4.743875388
7S4B	4.858805755	4.774083072
7S6W	5.892921681	5.176686604
7S6X	4.864343398	4.983515765
7S6Y	4.910433316	5.068725651
7S6Z	4.85782205	4.990914703
7S70	4.784651381	4.954836667
7S71	4.768823647	4.900622024
7S72	4.974360826	5.049844132
7S73	4.770732555	5.017198371
7S74	4.705747771	5.217553859
7S75	4.770536583	5.074294863
7SD9	4.527485742	5.2733757
7SDA	4.516750099	5.398686127
7SDC	4.779973963	5.131118685
7SET	5.217195533	5.202407842
7SF1	5.126422502	5.156774668
7SF3	5.055911277	5.238849942
7SFB	5.102171978	5.17533097
7SFH	5.084957202	5.297251675
7SFI	5.345214265	5.281402528
7SGH	4.939838526	5.297181483
7SH7	5.924534403	3.107356566
7SH8	4.580778951	3.056049226
7SH9	5.275116135	2.794263064
7SHB	5.313199526	2.99090143
7SI9	4.855335445	5.147653055
7T42	4.597884146	5.186963243
7T43	4.678516071	5.214136248
7T44	4.61630397	5.227396511
7T45	4.660596599	5.143032468
7T46	4.694688295	5.026983222
7T48	4.802020973	5.5198261
7T49	4.758324799	5.133292154
7T4A	4.708013986	5.09444529
7T4B	4.675139558	5.18451991
7TDU	6.546036817	5.133440552

7TE0	4.74739624	5.224049618
7TEH	5.935599402	5.273239216
7TEK	5.825811533	5.066871452
7TEL	5.968283343	5.033539057
7TFR	5.982182215	5.262410954
7TGR	4.723633159	5.526533647
7TIA	4.909352335	5.279172092
7TIU	4.768349592	5.240832731
7TIV	5.894713707	4.955861977
7TIW	4.774335622	5.360523028
7TIX	4.753892276	5.1331159
7TIY	4.683391244	5.431839408
7TIZ	4.802160914	5.168526568
7TJ0	5.735954638	4.9321561
7TLL	4.887962805	5.389077705
7TOB	4.730705075	5.579997648
7TQ2	4.872165678	5.496476634
7TQ3	4.739014583	5.049267989
7TQ4	4.81738499	5.094544253
7TQ5	4.680951546	5.116888874
7TQ6	4.729109321	5.211962601
7TUU	6.32484745	4.68868265
7TVX	5.727742431	4.993725244
7U28	4.81309679	5.458430449
7U29	4.567049214	5.558039815
7U92	4.968707887	5.249641685
7UJG	4.675301355	5.16383633
7UJU	4.71802299	5.27993296
7UKK	4.800927113	5.50457226
7UR9	6.005530273	4.710229353
7URB	6.019208576	4.837397736
7US4	6.114998635	4.875708147
7UUA	4.770179046	4.943893274
7UUB	4.655955788	5.228861604
7UUC	4.602847618	5.191352845
7UUD	4.812871053	4.969544441
7UUE	7.14038359	3.0202626
7UUG	5.192966079	5.166988679
7UUP	4.882354731	5.355761287
7V1T	4.746907179	5.554439071
7VH8	5.048426829	5.472953026
7VIC	4.903694745	5.414846607
7VLQ	5.767691031	5.299359164

7VTH	7.575271913	4.45801528
7VU6	7.634275159	4.432505945
7VVP	5.795832573	5.020983447
7VVT	4.901385464	4.670271168
7WO1	4.743990394	5.371811785
7WO2	4.786327155	5.41075584
7WO3	5.895268554	5.526495523
7WOF	5.831778345	5.129058044
7WOH	5.753678012	4.678652488
7WQB	4.857312669	5.174087305
7WQK	5.004998218	4.349809081
7WYM	5.650821178	3.823832772
7WYP	5.823769522	4.499095341
7X6J	6.77424067	4.401476745
7X6K	5.158244032	4.425366509
7XAR	4.754626571	4.543124189
7Z0P	5.319783443	5.09820919
7Z59	4.941923165	5.002434639
7ZQV	4.796251961	5.543530548
8A4Q	4.820197402	5.184032328
8A4T	5.251512779	5.475495985
8ACD	5.899702904	4.948135784
8ACL	5.736512412	4.729005858
8AEB	4.60187468	5.025224919
8AJ1	5.064300375	5.171273976
8B2T	5.8906649	5.389821115
8BGA	4.745210195	5.453668202
8BGD	6.114656945	5.413092895
8C9O	5.202744188	5.199798947
8CAC	5.015238186	5.060233739
8CYU	4.809377818	4.830803782
8CYZ	5.854884325	4.828812511
8CZ4	4.824915984	4.790047495
8CZ7	4.869920031	4.818123167
8CZW	4.689013816	5.238792407
8CZX	4.708454492	5.222072094
8D4K	5.35214473	5.321927936
8D4M	5.605646154	5.533929819
8D4P	4.511296124	5.018138521
8DCZ	5.650736072	5.595616697
8DD1	5.829480097	5.420663711
8DD9	4.890824753	5.454106259
8DDM	5.342478092	5.318719316

8DGB	4.542523397	5.23558496
8DIB	5.688564398	5.142381425
8DIC	5.998841663	4.918908528
8DID	5.650906717	5.163281791
8DIE	5.730806397	5.389174423
8DIF	4.745337614	5.240142449
8DIG	5.031028831	4.739763519
8DIH	5.580719369	4.768564497
8DII	5.983256339	4.61295024
8DL9	4.948664541	5.810065312
8DLB	6.050915774	5.661765191
8DMD	4.870785779	5.637402564
8DOX	7.688250776	5.250487536
8DOY	7.653515291	5.156853661
8DPR	7.850276034	5.310729152
8DS2	5.265129123	4.825221646
8DSU	4.569723226	5.138110431
8DT9	4.53235516	5.092180493
8DZ0	7.752149965	4.490739279
8DZ2	4.670650109	5.027200325
8DZ6	5.768162175	4.861950063
8DZ9	5.679606383	5.315871264
8DZA	5.635429748	5.066766349
8DZB	5.164735742	5.343985671
8DZC	5.104931284	5.071297877
8E1Y	5.651445905	5.032696206
8E26	5.652183588	5.07693472
8E4W	4.972725912	5.033454467
8E5C	4.592323164	5.317805837
8E5X	4.649626736	5.11917617
8E5Z	4.833444075	5.264650148
8E61	4.972027365	5.536326512
8E63	4.688807967	5.320763981
8E64	4.733568655	5.18675378
8E65	4.690103014	5.370146364
8E68	4.695575441	5.227277217
8E69	4.906977989	5.408233306
8E6A	4.863413137	5.404713212
8EOY	4.554143465	2.787134751
8EZV	5.160535424	5.273175987
8EZZ	5.120907529	5.298228057
8F02	5.122125279	5.315291079
8F2C	5.132508718	5.209833748

8F2D	5.0973886	5.212249392
8F44	4.693220117	5.291974366
8F45	4.99872298	5.277644775
8F46	4.717763276	5.323029901
8FIG	4.775504412	5.465444402
8FIV	5.602608183	4.646578144
8FIW	5.854687069	4.786022773
8FTC	5.640263634	5.468064275
8FTL	5.311594793	5.102462904
8FY6	5.40683371	5.16955647
8FY7	5.330762185	5.052269657
8GQT	5.694059038	4.767997163
8GTV	6.115500332	5.01346644
8GTW	5.988368798	4.906167271
8GXG	4.98732302	4.914904873
8GXH	4.997376416	5.097745701
8GXI	5.107347435	5.158208902
8GZB	5.103460702	4.180377569
8H3G	7.643684523	4.448070077
8H3K	7.654922748	4.455056512
8H3L	7.707657328	4.52036227
8H4Y	5.508087224	5.43910729
8H51	5.408783027	5.560033258
8H57	4.731437781	5.541910893
8H5F	4.667483978	5.560690627
8H5P	4.681008112	5.497036156
8H6I	6.009320192	5.575095356
8H6N	4.783981703	5.638507365
8H7W	4.721022011	5.605098835
8H82	4.704823273	5.481082389
8HBK	7.518542004	4.351822125
8HEF	7.661772625	4.439940463
8HHT	4.858629662	2.951741119
8HHU	5.012473858	2.95410349
8HI9	5.528046984	5.443505393
8HOM	7.495814049	4.403032152
8HOZ	4.361651701	5.257113116
8HQG	7.151029359	5.385653354
8HQH	7.198905485	5.396486122
8HQI	7.482496082	5.511054509
8HQJ	7.771628156	5.31952643
8HTV	6.071281875	5.255482481
8HUR	7.528883541	4.438030861

8HUV	7.684998285	4.431838225
8HUW	7.519983034	4.402562996
8HUX	7.769552295	4.43490875
8HVK	4.745919239	5.367666065
8HVL	4.923732339	5.405447118
8HVM	4.734592865	5.28432833
8HVN	5.856152644	5.351208037
8HVO	4.7279092	5.299066955
8HVU	4.691192741	5.275202788
8HVV	4.727010536	5.258941067
8HVW	4.921512524	5.23687731
8HVX	4.822844374	5.221717228
8HVY	4.72128058	5.2421902
8HVZ	4.774203245	5.310698472
8HZR	5.710019008	4.923381975
8I30	4.844493041	2.832468305
8IFP	4.536177573	5.552089334
8IFQ	4.78849234	5.662103602
8IFR	4.627976392	5.463648231
8IFS	4.736679826	4.915333682
8IFT	5.948047904	5.65575605
8IG7	4.729855936	5.452748612
8IG8	4.679247693	5.358104257
8IG9	4.67950574	5.382238402
8IGA	4.69378062	5.208273995
8IGB	4.557932648	5.293762747
8IGN	4.81049128	5.353987204
8IGX	5.857326996	5.638244632
8IGY	4.728939601	5.601626728
8INU	4.73116981	5.623641899
8INW	4.698121423	5.327098545
8INX	7.587155978	4.459575767
8INY	7.442013301	4.436493481
8J32	4.736091237	5.205925394
8J35	4.765907664	5.291168773
8J36	4.75983869	5.187011614
8J37	4.687085157	5.27744864
8J38	4.612311461	5.324960459
8J39	4.649199363	5.230902817
8J3A	4.8170349	5.410112104
8J3B	4.616237445	5.203975394
8JCJ	7.261740953	3.252003534
8JCK	5.731138943	3.200165879

8JCL	5.384329505	3.089476661
8JCM	5.835321606	3.146110793
8JCN	5.69990288	3.162314791
8JCO	4.822152305	5.198938436
8JOP	6.111735404	5.050409441
8K6D	6.055019827	5.028922689
8OKB	4.87780253	5.188177633
8OKC	4.927492159	5.314297529
8OKK	5.618477676	4.846422106
8OKL	5.666415643	4.866528986
8OKM	4.668958236	5.051153194
8OKN	5.803968294	4.90151415
8P54	5.610610975	4.984807387
8P55	5.650352357	4.995442128
8P56	5.778646085	4.448503657
8P57	5.76998362	4.474832433
8P58	5.668860426	4.27016351
8P5A	5.695090618	4.303892596
8P5C	5.732509387	4.746415595
8P86	4.539710803	5.118115943
8P87	5.702754986	4.40101798
8PH4	5.707801693	4.969977188
8R0V	5.17287347	5.192241784
8R11	5.894246466	4.733989639
8R12	5.989161155	4.958392478
8R14	5.900334693	4.84959589
8R16	5.988212669	4.789341001
8R1Q	5.147394576	5.306936991
8R26	5.24440916	5.242337801
8RI4	4.996659359	4.121005318
8RJZ	5.965500048	5.032110457
8SK4	6.148643922	5.258834081
8SKH	4.90871047	5.44624694
8STY	4.556873109	5.074903279
8STZ	5.595378685	5.083630934
8SXR	4.662157544	4.669666687
8TBE	5.577869872	5.429708849
8TPD	5.846516992	5.388432866
8TQH	4.634646147	5.27929274
8TQJ	4.642294708	5.139101563
8TQL	4.790526247	5.141832152
8TQT	5.571778892	4.964839376
8TY3	5.268423413	5.149917563

8TY4	5.32930696	5.247617568
8TY5	4.671047841	5.09374694
8U40	5.389384035	5.231757302
8U9H	4.81112323	5.178069894
8U9K	5.468971694	5.107133397
8U9M	4.704582678	5.138380735
8U9N	4.755935305	5.153808788
8U9T	5.63646447	5.112390831
8U9U	4.846828742	5.089643155
8U9V	4.621042971	5.234620978
8U9W	4.666586197	5.372201501
8UAB	4.687945379	5.477423527
8UDJ	4.692290964	5.135098519
8UDM	4.901538484	5.934135148
8UDO	4.836874156	4.712674477
8UDP	4.747242814	4.699525184
8UDQ	4.82675448	4.267762287
8UDW	4.973773085	5.242666021
8UDX	5.817375401	4.225043882
8UDY	4.861494901	5.437837098
8UE0	4.993853009	4.171852451
8UEA	4.7754361	5.824760602
8UEB	4.835877595	5.783203993
8UEF	4.790580751	4.789224591
8UEG	4.826781426	4.580360323
8UEH	4.889746191	5.027196796
8UEI	4.771892186	5.826576501
8UH5	7.771362678	5.322536308
8UH9	6.395297655	5.187016475
8UHO	4.765955002	5.003062673
8UIA	5.813722639	5.241828563
8UIF	4.764132713	5.189222296
8UPS	5.165043108	5.165160651
8UPV	4.787019878	5.318840738
8UPW	4.930090688	5.370170375
8UR9	4.787113921	4.971176117
8UTE	4.797220134	5.093197156
8V4U	5.010046129	5.357977215
8V8E	7.617579067	4.336323
8V8G	7.645019906	4.308808701
8VDJ	5.615668567	5.205314924
8VQX	5.75525117	5.27957485
8VSG	4.870335034	4.941030752

8W1T	6.191650796	4.90755472
8W1U	6.020624534	4.854137008
8WS3	4.842523897	3.329782155
8WTI	4.782765121	2.835750047
8WUR	6.079626645	4.449376925
8WZP	5.708479362	4.755205509
8WZQ	4.777085112	4.910585757
8Y42	4.774052054	5.164377094
8Y44	4.726393948	5.141335576
8Y4D	5.925067236	4.953351693
8Y4G	5.989527	4.959135213
8Y4H	5.780915912	5.108637144
8Y7T	5.949466277	4.797676467
8Y7U	5.817588701	4.880987648
8YKM	5.780675507	4.478894521
8YKO	5.754947757	4.414929625
8YKP	5.893509093	4.422004579
8YKQ	5.830059263	4.325164923
8YLS	5.556526486	5.193063176
8YRH	5.713364619	4.885630469
8Z1H	4.508348037	5.057439257
8Z46	6.004294729	5.485832521
9ARQ	6.089719102	5.139662476
9ARS	6.69657287	5.320290795
9ART	6.576491353	5.293715727
9ASV	4.68200073	5.095115562
9ASW	4.693585259	5.077576004
9ASY	4.815591787	5.241827735
9ASZ	4.79684464	5.027570045
9AT0	4.852660904	5.117715229
9AT1	5.652949258	5.058889225
9AT3	4.754037864	5.06497249
9AT4	4.784317918	5.131081252
9AT5	4.769583852	5.074648959
9AT6	4.790818327	5.133977843
9AT7	4.720582222	5.170310141
9AUJ	4.79931652	5.541330216
9AUK	4.711817483	5.510728723
9AUL	4.812507017	5.474619631
9AUM	4.7665671	5.472129697
9AVQ	4.747681608	5.525954779
9BO0	4.918981207	5.021004635
9BO1	5.671141366	4.704263174

9BO2	4.777026248	4.941466243
9BO3	5.380850136	4.429321817
9BO4	4.710409838	4.612218671
9BO5	4.745504875	5.269650055
9BO6	4.881154864	5.216318172
9BO7	4.706978576	5.53376269
9BO8	4.584099331	5.0937837
9BO9	4.569584333	5.263752867
9BOA	4.574589768	5.409151442
9BOB	4.516812508	5.4144784
9BOC	4.569767897	5.453159535
9BOD	4.563406568	5.428485838
9BOE	4.801777443	5.23479973
9BPF	5.860214757	5.335187976
9C80	5.863524989	5.295147075
9C8Q	5.723158114	3.269882908
9CDK	4.802324753	5.123187695
9CDL	4.760633468	5.224932143
9CDM	4.7154138	5.164497016
9CEC	4.536960654	5.192757574
9CED	4.4256683	5.214194344
9CEK	4.459595906	4.841395957
9CF9	4.331932195	4.866599617
9CFB	5.776276044	4.739032111
9CMS	7.701170512	4.420200385
9CMU	7.793394246	4.396858716
9EL4	4.605875864	5.163053047
9ELV	4.605875864	5.163053047
9F2X	5.06961461	5.263678233
9F39	4.936443949	5.102138111
9F3A	4.915750639	5.116006991
9FHQ	5.153304984	5.244584207
9GIJ	5.830650794	4.883809118
9GIL	6.026261721	4.918410994
9GLV	5.655570088	5.625472999
9GMQ	5.072212769	5.269514126
9GV2	4.899291653	4.695078935
9HAJ	5.827253068	4.847891353
9HAK	5.884061005	4.811117195
9HBQ	4.658783138	5.00671308
9HC1	5.756600752	4.805652046
9HD8	4.804604623	4.163336567
9HDC	5.854011608	5.124028077

9HDJ	4.549700376	4.749994896
9HDN	5.85773903	4.415459938
9HJH	6.04936406	4.787117377
9IK2	9.370151938	5.254303833
9MEI	4.707118972	5.189878692