Machine Learning Model for Repurposing Drugs to Target Viral Diseases

by

Justine Williams

A thesis

presented to the University of Waterloo

in fulfillment of the

thesis requirement for the degree of

Master of Science

in

Chemistry

Waterloo, Ontario, Canada, 2023

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

With recent events, such as the Covid-19 pandemic, it is increasingly important to develop strategies to combat viral diseases. Due to technological advancements, computer-aided drug design and machine learning (ML)-based hit identification strategies have gained popularity. Applying these techniques to identify novel scaffolds and/or repurpose existing therapeutics for viral diseases is a promising approach. As an avenue to improve existing classification models for antiviral applications, this thesis aimed to make improvements to non-binding data selection within these models. We created a classification model using molecular fingerprints to assess the performance of machine learning predictions when the model is trained using randomly selected and rationally selected non-binding datasets. Our analyses revealed that machine learning predictions can be improved using a rational selection approach. We further used this approach and trained three machine learning models based on XGBoost, Random Forest, and Support Vector Machine to predict potential inhibitors for the SARS-CoV2 main protease (Mpro) enzyme. Probability-ranked hits from the combined model were further analyzed using classical structure-based methods. The binding modes and affinities of the hits were identified using AutoDock Vina, and molecular dynamics simulations-enabled MM-GBSA calculations. The top hits identified from this multi-step screening approach revealed potential candidates that show improved affinity and stability than existing non-covalent Mpro inhibitors. Thus, our approach and the model could be useful for screening large ligand libraries.

Acknowledgements

I would like to thank my supervisor, Dr. Subha Kalyaanamoorthy, for her guidance and support throughout my studies, as well as the other members of the SK lab.

Table of Contents

List of Figures

List of Tables

Chapter 1
Introduction

Just over 100 years ago, when the Spanish Flu spread across the world resulting in the deaths of 20-50 million people, no effective treatments were available. The only approaches were mandatory masks, physical distancing, and washing of hands [1]. Over the next century, we learned from this pandemic and others. In current days, we have immunizations and drug treatments available, but as new diseases arise, we need to develop ways to combat them more quickly. Towards this goal, we now have updated public health and research strategies at our disposal, including high-throughput screening of therapeutic candidates for testing and employing innovative computational methods, such as machine learning and structure-based drug discovery.

Investigation and development of drugs to combat these diseases is a crucial strategy. This process has a high cost, of approximately $1.3 billion per drug [2], and is highly selective, resulting in a less than 10 % chance for a drug to reach market approval [3]. The approved drugs have undergone rigorous testing with a pipeline of preclinical, phase I and phase II trials, taking a significant investment of time and cost, as shown in **Figure 1**.

De Novo Drug discovery; 10-17 years.

Target Discovery	Discovery & Screening	Lead Optimization	ADMET	Development	Registration
2-3 years	0.5-1 years	1-3 years	1-2 years	5-6 years	1-2 years

Drug repurposing; 7-12 years.

Compound identification	Compound acquisition	Development	Registration
1-2 years	0-2 years	5-6 years	1-2 years

Figure 1. Drug discovery pipeline.Reprinted (adapted) with permission from [4]. Copyright 2004 Nature Reviews Drug Discovery.

In the past, drugs that could be repurposed were identified through serendipitous discoveries and experimental measures, and then more recently through computational efforts, known as "Computer-Aided Drug Design" (CADD). These CADD strategies have been attributed to the discovery of protein inhibitors Saquinavir and Indinavir for HIV-1 protease through structure-based computational approaches in 1995 and 1996, respectively [5]. Of these strategies, structure-based drug design (SBDD), where 3D information on the protein target is used to find ligands that bind and provide a 'hit,' remains a promising strategy in CADD. When 3D structure is not available for use, ligand-based virtual screening (LBVS) can be employed to identify hit candidates provided there is some knowledge regarding the known binder. In LBVS, the binders are used as a template to compare and identify other ligands that share similar structure and/or properties as similar ligands are assumed to have similar activity [6].

Recently, there have been some impressive antiviral screening efforts using machine learning (ML). One example is the 'deep docking' approach, a structure-based deep learning model that uses a sampling procedure to efficiently compute scores for a large number of structures. This approach was used for screening a library of 1.3 billion drug-like compounds, leading to the identification of 11 hits as potential inhibitors for SARS-CoV-2, and 585 unique scaffolds [7]. Another study reported the supervised machine learning screening of two subsets of the ZINC15 database, one comprising 39,442 compounds and another containing 1577 approved drugs. A third screening set contained 115 natural products, extracted from literature [8]. This combination approach used virtual screening, through AutoDock Vina/DOCK6, and a machine learning classification model to find hits for three SARS-CoV-2 target proteins: spike, nucleocapsid, and 2'-o-ribose methyltransferase. The top 100 hits per dataset and target underwent further processing, and the final results indicated that some anti-hepatitis C drugs could be promising candidates for treatment against SARS-CoV-2. Similarly, other targets from different viruses, e.g., protein targets from West Nile and Dengue viruses have also undergone computational screening. One example is a notable pharmacophore-based

approach, which screened over a billion compounds from PubChem. This effort resulted in a shortlist of ranked lead compounds [9].

Despite the development of multiple strategies, identifying promising hits and transforming them to marketable drugs is highly challenging due to safety and efficacy problems. However, by screening the set of drugs meeting human safety protocols, the pipeline can be shortened, resulting in both time and cost savings. Additionally, re-use of previously approved drugs for a new indication (drug repurposing) also reduces the risk of drug-induced toxic side effects. In the case of viral infections, a drug can be used on the same target protein against a new virus, such as an approved antiviral drug targeting proteases on one virus that can be repurposed for targeting proteases in another virus [10]. Currently, nirmatrelvir and ritonavir are antiviral drugs that are repurposed drugs used in combination to treat SARS-CoV-2 [11]. On the other hand, a drug may also be reused for a new target protein with a new indication, as with an anti-cancer drug being reused to treat a respiratory infection. Since these drugs have already undergone clinical trials and have been approved for an alternate use with no adverse effects, drug repurposing has a higher likelihood of approval since phases of the drug trials may be bypassed (preclinical/phase I). Through use of computational methods, either by way of large-scale screening efforts to find novel scaffolds, or drug-repurposing computer-aided drug design, these methods are used with the goal of minimizing time and costs within the drug discovery pipeline.

The aim of this thesis is three-fold. First, to study the impact of molecular features and proportions of binding/non-binding data on the prediction accuracy using a basic classification model. Second, to apply the optimal classification model and a probability-based ranking to screen a small molecule database to identify potential drug-like candidates with antiviral properties. Finally, to understand the mode of interactions, stability, and affinities of the top hits from the machine learning model through structure-based methods such as molecular docking and dynamics simulations.

Chapter 2 outlines the tools and methodology. For model building, there is an overview of datasets available that are used for model training, and along with this, there is a description

of machine learning model algorithms and procedures. Following this, methods for “hit” validation, including molecular docking, molecular dynamics, and binding free energy calculations, are described.

Chapter 3 outlines the training and testing of the classification machine learning models and reports the impact of the quantity and type of non-binding data on model performance using the DUD-E dataset. In this chapter, we also apply this classification model to the Mpro antiviral case, where two Mpro external datasets are tested with the optimal classification model.

Finally, Chapter 4 investigates the screened hits from the combined machine learning model for the Mpro target protein, through use of the Chem-space virtual screening dataset. Unique poses are then validated structurally for interactions, binding mode, stability, and affinity with molecular dynamics simulations, to determine the most favourable binders for future experimental validations.

Chapter 2
Methods

This chapter outlines methods required to complete the steps of an iterative virtual screening procedure, to determine protein-ligand binding for an Mpro protein that is screened against a set of ligands, or 'drug' molecules. Initially, databases containing binding information for proteins with small molecules and drug-like compound libraries are introduced. Next, a description of machine learning algorithms and methods applied to the drug screening pipeline and the common evaluation metrics used for assessing the qualities of the ML models is provided. Finally, other structure-based analysis methods, such as protein-ligand docking, molecular dynamics, and binding free energy calculations used for validating a subset of ML-screened 'hits' are described.

2.1 Datasets

A machine learning-based predictive model was trained using the features from the three-dimensional (3D) structures of the ligand-bound complexes of the biological targets. Further, we used different small molecule libraries for training and testing the performance of the ML models. The data sources used for this are listed within this section.

2.1.1 Binding Information

The Protein Data Bank (PDB) [12], is a comprehensive repository for experimentally resolved structures of biomolecules. It is comprised of the three-dimensional (3D) structures of protein (187,536), DNA (9055), RNA (6141), NA-hybrid (230), and other molecules. Each four digit 'PDB ID' value points to and describes a 3D structure. The binding data for ligands were obtained from the PDBbind dataset [13], which is a curated set from the full PDB collection and includes entries with K_d and K_i experimental binding data. The PDBbind set contains multiple subsets: the 'refined' data are checked and validated entries of the highest quality, and the 'general' set is the remainder of the PDBbind entries that have a lower crystal structure resolution. These PDBbind datasets are typically released annually, with updated crystal

structures extracted from the PDB. From these sets, benchmarking protocols have been created, such as with the 2007 collection, named by the year it was created, containing 1300 proteins. This benchmark set is typically used to compare regression models and instructs users to follow a specific protocol. These datasets are widely used for benchmarking, training, and testing of ML models. Since this curated data is considered as a gold-standard dataset, we used the 2007 dataset from the PDBbind for initial training and testing procedures to determine a fingerprint threshold value.

In contrast, The Directory of Useful Decoys – Enhanced (DUD-E) [14], contains a much smaller number of target proteins (102) along with corresponding ‘bound’ or ‘active’ ligands, and ‘non-binding’ or ‘decoy’ ligands. The set of ‘decoys’ contains 50 times more data points than the positive set and were identified based on physicochemical characteristics and similarity measures. **Table 1** provides the summary of all databases used for ML model training and testing.

Table 1. Protein databases used for ML model training and testing.

Database Name	Quantity	Description
Protein Databank (PDB) [12]	191,565 biological macromolecular structures	Collection of experimentally determined crystal structures
PDBbind 2020 [13]	23,496 biomolecular structures: 19,443 protein-ligand, 2,852 protein-protein, 1,052 protein-nucleic acid, and 149 nucleic acid-ligand complexes 2007 benchmark: 1300 protein-ligand complexes	Curated to contain only binding pairs that have corresponding experimental information Updated bi-annually
Directory of Useful Decoys – Enhanced (DUD-E) [14]	102 proteins Actives ~224 per target Decoys ~11,200 per target	Collection of 102 proteins. Each protein has a set of 'active' and 'decoy' small molecule binders

2.1.2 Small Molecule Screening Libraries

Multiple libraries of small molecules are available for *in silico* screening of hit candidates. For drug repurposing applications, databases such as DrugBank [15] provide a collection of

approved, investigational, and experimental molecules that may be screened. However, this database is small, with only 11,682 molecules. To expand the diversity of the hits and identify novel scaffolds, we performed virtual screening of the 'Chem-space virtual screening' dataset, which includes 3,878,821 compounds [16]. Given the higher efficiency of current computational methods, especially ML-based techniques, larger databases may be screened to find novel potential inhibitors for a given target.

2.1.3 Molecular Representations and Similarity Indices

Molecular fingerprints are a method of describing the chemical features of a molecule; two fingerprints may be compared to one another to evaluate similarity between two molecules [17]. Thus, this method is employed in ligand-based virtual screening, a procedure where only the structure of a known active ligand molecule is used in a screening procedure to search for similar molecules that could be a potential binder for a given target. Specifically, two-dimensional (2D) molecular fingerprints are used as descriptors by way of a bit-encoded representation of a particular molecule.

To compare these bit-based representations, similarity between molecules can be calculated using mathematical coefficients that include Tanimoto, dice, cosine, etc. [18]. These mathematical methods/similarity metrics are used to compare the fingerprints to each other and yield a score between 0 and 1 depending on their similarity, where the score range is from 0 (completely dissimilar) to 1 (identical). Fingerprint similarity comparison is outlined in **Figure 2**. The simplistic representations of fingerprints allow for faster comparison between molecules and thus, can be used for virtual screening efforts for a large database.

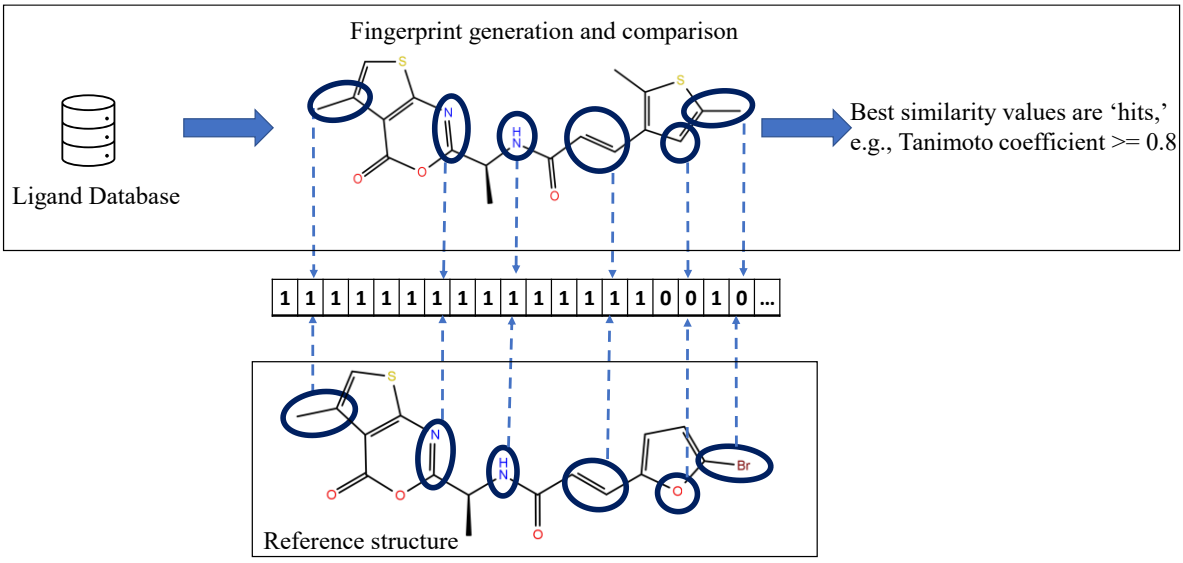


Figure 2. Example of structural comparison between two molecules. Reprinted (adapted) with permission from [19]. Copyright 2022 Elsevier Inc. Substructures that are equal result in a '1' within the array. Differing substructures result in a '0.' Total similarity score is calculated using a mathematical coefficient, such as the Tanimoto coefficient, to determine whether the fingerprint is a 'hit' (≥ 0.8) or not (< 0.8), when compared to the reference structure using a particular threshold value.

Fingerprints have different strengths/weaknesses. For example, topological, or path-based fingerprints give more weightage towards structures with a similar 'skeleton.' In contrast, fingerprints, such as MACCS (Molecular ACCess System), are key-based, which can aid with 'scaffold hopping' and finding hits with different core molecular structures [20]. To use these differences as an advantage, combined scoring strategies can be used to minimize individual weaknesses, where a hit is deemed suitable based on the average of the fingerprint group. To incorporate these into computational workflows, fingerprints are included in common molecular toolkits, such as RDKit, as it contains base functionality of molecule and protein data structures.

Types of fingerprints that will be utilized in this study are substructure, path-based, and circular fingerprints. Specifically, fingerprints were chosen from RDKit and include Daylight-based substructure (RDKit implemented), topological (substructure/layered fingerprint), key-based (MACCS), and circular (extended-connectivity subtype, Morgan). The 'RDKit' implementation of the Daylight fingerprint is a method that uses hashing, and linear pathways to encode a molecule [21]. Bits are not assigned to one specific characteristic in the Daylight fingerprint. Next, the layered fingerprint utilized in this study is an adaptation of the topological fingerprint, where a more generic approach is used, with pre-defined substructures, but the same method of searching through the molecule. Alternatively, 'MACCS' is a fragment/key-based approach that uses 166 structural keys, representing important chemical features, to determine the bit-representation. Finally, 'Morgan' fingerprints are circular fingerprints, where a circular radius is used for determination of the fingerprint encoding. These fingerprints are also known as ECFP (extended-connectivity fingerprints); a radius of 6 was used for the Morgan fingerprint within this study.

2.2 Machine Learning Methods

Machine learning methods learn from a large amount of data and make a 'target' prediction by determining a mathematical relationship between a set of input 'features' and the 'target' value [22]. In cheminformatic applications, the feature data could consist of molecular weight, charge, solvent accessible surface area, number of hydrogen bond donors, and more. In the ideal case, a machine learning model trained with large amounts of high-quality data will learn the overall trends within the data and accurately predict a property of interest [23]. Machine learning is typically performed through 'supervised,' 'unsupervised,' 'semi-supervised' or 'reinforcement learning,' the four main types of machine learning.

Of these, the two most commonly used learning methods are 'supervised' and 'unsupervised.' First, the 'supervised' model requires labelled input and output data that is to be predicted by the model [24]. Supervised learning can be further categorized into classification-based and regression-based models. In the classification scheme, the aim is to use classifier labels (e.g., binary) to train a model to predict that label. For example, a

classification model in hit identification can predict a molecule to be a ‘binder’ (1) or a ‘nonbinder’ (0). Pre-classifier probabilities can also be extracted and utilized to provide additional information for categorical predictions. In the regression scheme, a model is trained on quantitative data and learns to predict the specific values or rank them. For example, a simple regression model can predict the ‘binding affinity’ values for a hit molecule. In contrast, unsupervised learning aims to predict trends within data, but with no set target label [24]. This approach is applicable for categorization; for example, unsupervised learning can be used for clustering protein families based on the protein sequences or binding sites.

2.2.1 Algorithms

2.2.1.1 Random Forest

The random forest (RF) algorithm uses the ‘bagging’ method, where a quantity of decision trees each receive a random set of features to produce an output from [25]. There is an equal probability for any feature to be used. A consensus score is determined from the decision tree predictions. Either the trees make a classification prediction (0 or 1), with the majority taken as the prediction, or have a regression prediction with the average number as the outcome; this process is visualized in **Figure 3** [25, 26].

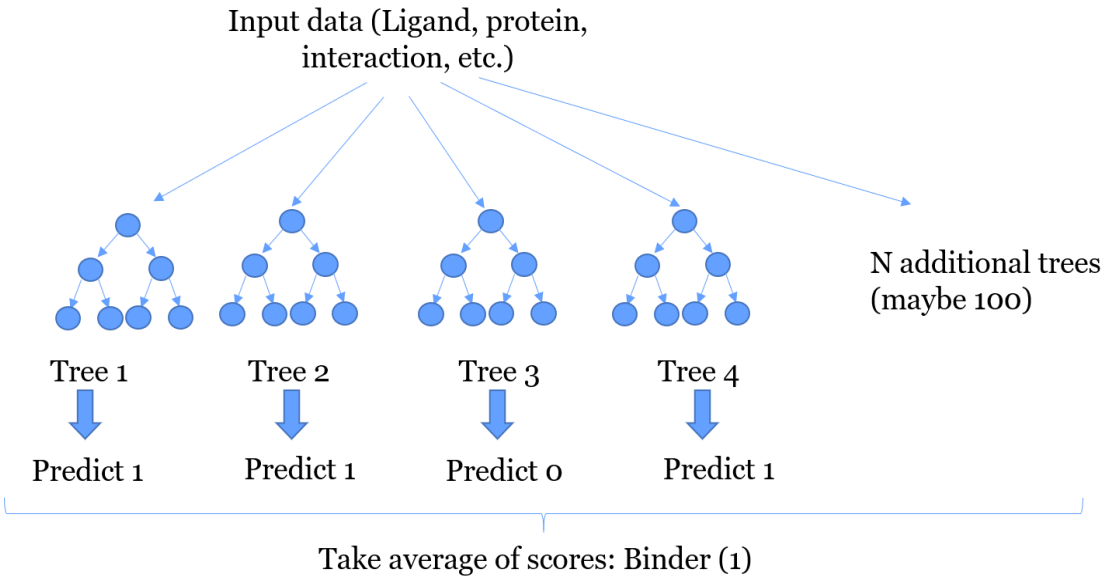


Figure 3. Visualization of random forest algorithm. Visualization of use of 'bagging' method. The result is from the averaged trees, and any feature has an equal probability to contribute to this, since each decision tree that makes up the result is comprised of a random selection of features.

2.2.1.2 XGBoost

In contrast to 'bagging,' 'boosting' is another method that uses decision trees to make a prediction. In this case, each decision tree is built upon to improve the prediction, where the feature weighting for the next decision tree is based on the last one, as depicted by **Figure 4**. Extreme gradient boosting refines gradient boost and provides advantages such as faster execution speed and improved handling of noise and sparse data [27].

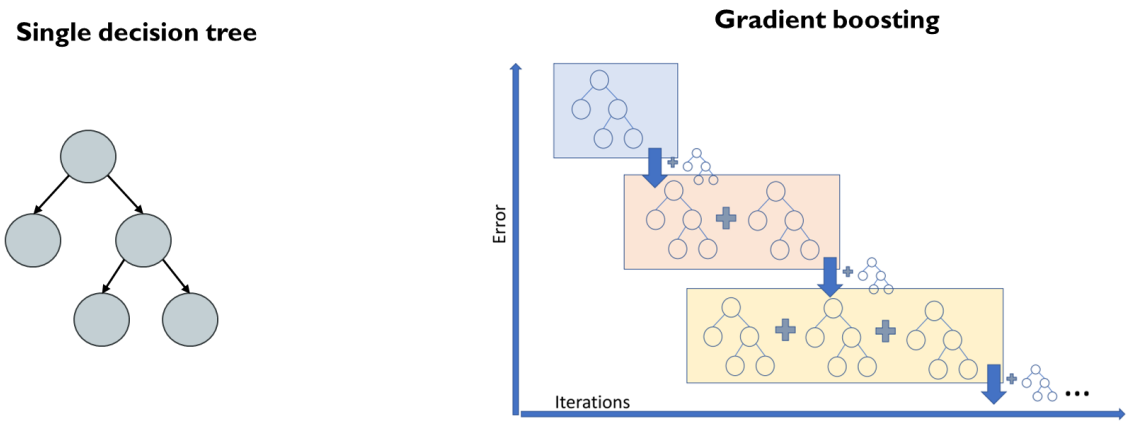


Figure 4. Gradient boosting.Rows represent the ability for this method to learn from previous decision trees, to make an improved prediction. Duplicated from [28].

2.2.1.3 Support Vector Machine

Support Vector Machine (SVM) is an algorithm that was created to use feature data to group inputs into two categories [29]. In this sense, it is an excellent choice for a model requiring 'binding (1)' and 'non-binding (0)' categorization. The categorization of SVM occurs through creating cut-off values for feature data and clustering the binding and non-binding features of each. Through use of mathematical operations, all feature data can be plotted onto a graph that represents N-dimensional space, in order to find the 'hyperplane' that separates the data categories, illustrated in **Figure 5** [24]. Understandably, points closer to the hyperplane are the most crucial, as they dictate where the optimal hyperplane will lie. This further emphasizes the importance of validated data inputs, as a few crucial data points close to the hyperplane could make a significant difference in overall model accuracy.

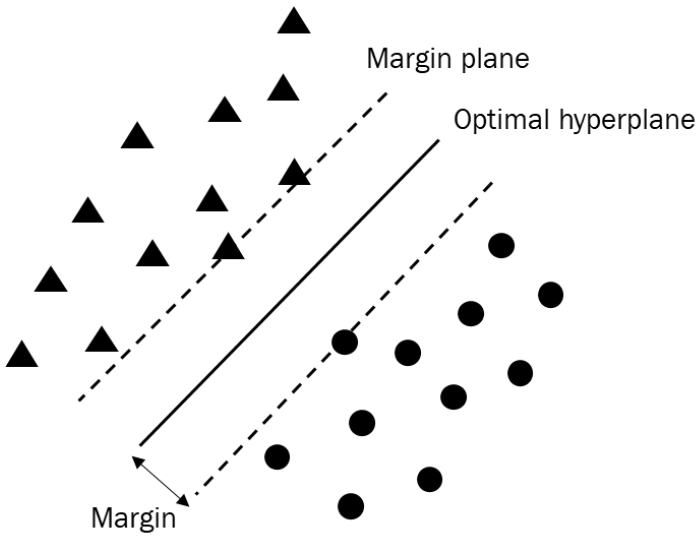


Figure 5. Support vector machine graphical depiction. Categorization of the SVM algorithm shows the division of data on either side of a defined hyperplane with the optimal separation between datapoints of two different categories. Reprinted (adapted) with permission from [29]. Copyright 2005 American Chemical Society.

2.2.2 Evaluation

Model evaluation can occur through the use of standardized datasets and performance metrics. This gives an objective gauge of model effectiveness between computational methods [20]. In addition to the common evaluation metrics of accuracy, precision, and recall used in ML model evaluation, AUC (Area Under the Curve) and Cohen Kappa score are also used in structure-based virtual screening applications. A summary of important metrics is outlined in **Table 2**.

Table 2. Evaluation metrics and their mathematical significance. Within the table: TP = true positive, FP = false positive, FN = false negative and TN = true negative, and N = TP + FP + TN + FN [30].

Metric	**Description**
Area Under the Curve	Measure of ability to distinguish between classes, compared to random chance.
Precision	$\frac{TP}{TP + FP}$
Recall/Sensitivity/True positive rate	$\frac{TP}{TP + FN}$
Specificity/Selectivity/True negative rate	$\frac{TN}{TN + FP}$
F1 Score	$2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$
Accuracy	$\frac{TP + TN}{(TP + FN + TN + FP)}$
FM Index	$\frac{TP}{\sqrt{(TP + FP) \cdot (FP + FN)}}$
Matthew's Correlation Coefficient	Balanced measure, useful when data classes are of different sizes; ranges from -1 to 1
Jaccard Index	$\frac{TP}{TP + FP + FN}$
Cohen Kappa	Level of agreement between two class labels, ranges from -1 to 1
False Positive Rate	$\frac{FP}{FP + TN}$
False Negative Rate	$\frac{FN}{FN + TP}$
Informedness	$\frac{TP}{TP + FN} + \frac{TN}{TN + FP} - 1$
Markedness	$\frac{TN}{TN + FN} + \frac{TN}{TN + FN} - 1$

Specifically, the AUC metric is determined graphically by counting backwards from highest to lowest ranked compounds and plotting that against the number of actives [16]. The area under the curve is calculated, with a higher value indicating more accurate predictions. Matthew's correlation coefficient, Jaccard Index, Cohen Kappa (ranging from -1 to 1), and informedness/markedness are useful metrics when evaluating dataset performance where true positives and true negatives are equally important [30]. The informedness and markedness metrics are variations on recall and precision, respectively [31]. These metrics take into consideration all aspects of the confusion matrix (true positive, true negative, false positive and false negative), and range from a -1 to 1 score value. Informedness builds upon recall and indicates probability of predicting true positives and true negatives. Similarly, markedness builds upon precision and is an indication of the trustworthiness of these predictions.

2.2.3 Over and under-sampling

For imbalanced datasets, SMOTE (Synthetic Minority Oversampling Technique) oversampling, and under-sampling steps can be added to minimize biases that are introduced due to class imbalance. Over-sampling procedures create additional data points by estimating according to the feature space [32]. These procedures perform well within drug-target interaction prediction strategies [33]. Additionally, their combination with under-sampling techniques has been shown to provide good performance. Under-sampling removes samples to equalize the quantities for class data [34]. These steps can be helpful to avoid biases within machine learning models, while still taking advantage of an imbalanced dataset as input.

2.3 Additional Computational Analysis Methods

Either used as standalone methods, or in combination, molecular docking, molecular dynamics, and binding free energy (MM-PBSA/MM-GBSA) methods are utilized to evaluate binding affinity predictions.

2.3.1 Molecular Docking

Molecular docking is a high throughput computational approach used to predict the mode of binding of a molecule and its binding affinity to a target [35]. The model of binding is obtained through an exhaustive sampling of ligand conformations in the protein's binding site, and the affinity of the molecule is predicted using a scoring function. We used molecular docking to obtain the binding modes and affinities of the hits identified through the ML model. All docking calculations were performed using AutoDock Vina v1.1.2, an efficient docking program, which uses a mixture of empirical and knowledge-based methods to make a prediction [36]. This program was selected for the virtual screening processes in this study due to AutoDock Vina's fast performance; this program can utilize multiple CPUs in parallel to further increase performance.

2.3.2 Molecular Dynamics

Molecular dynamics (MD) simulations use mathematical approximations to describe the motions and time-dependent conformational changes of molecules. In drug discovery, it is commonly used for studying the protein-ligand dynamics, interactions, and their affinities. Forces that act on both bonded and non-bonded atoms are estimated using the functions and parameters described by a force field. Bonds, atom to atom angles, dihedral angles, and non-bonded atoms are approximated with springs, sinusoidal function, and Coulomb's Law, respectively [37]. The MD simulations in this thesis were carried out using the 'AMBERff14SB' force field of the Amber18 package. The ligands were parameterized using General Amber Forcefield (GAFF2) as shown in **Equation 1,** and atom/bond types and partial charges assigned through antechamber's AM1-BCC package [38].

Equation 1. Amber forcefield functional form [39].

$$E_{total} = \sum_{bonds} K_r\left(r - r_{eq}\right)^2 + \sum_{angles} K_\theta\left(\theta - \theta_{eq}\right)^2 + \sum_{dihedrals} \frac{V_n}{2}[1 + cos(n\phi - \gamma)] + \sum_{i<j}\left[\frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^{6}} + \frac{q_i q_j}{\epsilon R_{ij}}\right]$$

Next, parmchk2 checked for missing forcefield parameters. The tleap program was utilized to prepare the system for the molecular dynamics simulations, and with this program, a 12 Å box was created with TIP3P water molecules, and NaCl counter ions in a concentration of 150 mM. Simulations consisted of five steps: 1) energy minimization, 2) heating, 3) equilibration, 4) pre-production, and 5) production.

First, five stages of energy minimization were performed; initial stages were carried out with 10,000 steps and strong harmonic constraints of 100, 50, 25, 10, and 5 kcal/mol/Å applied to solute atoms. A final minimization with 20,000 steps and no constraints was then performed. Three steps of equilibration took place with restraint masks of 5, 0.1, 0.01 (backbone only), each 100 ps in length at a constant 310 K. A fourth equilibration step was performed with no constraints and for 400 ps, followed by a 2 ns pre-production phase held at constant temperature and pressure. Temperature throughout the simulations was controlled with the Langevin thermostat [40], and pressure, with the Berendsen barostat [41]. Finally, three production runs were performed, at 10 ns each, with sampling at 2 fs, resulting in 15,000 production frames. These simulations were run on the University of Waterloo's 'tqtmaster' cluster.

To analyze these trajectory files, Root Mean Squared Deviation (RMSD) plots are a method to visualize how atom placement and protein structure changes over time [42]. Different aspects can be visualized, such as the conformational changes of a ligand over the course of a molecular dynamics simulation, or protein backbone fluctuations that are taking place. Another analysis method is Root Mean Squared Fluctuation (RMSF), which operates on the same principal, but in this case, specific residue changes are recorded and visualized. Through RMSF analysis, regions of high and low flexibility may be identified within a protein. Additionally, hydrogen bonding was examined; strong hydrogen bonding aids in ligand stability within a protein binding site.

2.3.3 Binding Free Energy

‘Molecular mechanics Poisson−Boltzmann surface area’ (MM-PBSA) and ‘molecular mechanics generalized Born surface area’ (MM-GBSA) are two major endpoint methods used widely in drug discovery for calculating the binding free energy of the hit molecules towards their target proteins. These methods use implicit solvent models to estimate the free energy of binding of ligands from the dynamic trajectories and have been shown to have better accuracy in relative ranking of the hit molecules [43].

The main difference between these two methods is that the polar solvation term is calculated differently; in MM-PBSA, this term is calculated through the Poisson Boltzmann equation, while MM-GBSA solves this term using the Generalized Born equation. To solve the polar solvation term of MM-PBSA, the AMBER program contains finite difference solution methods (a manner of approximating differential equations): four linear and six nonlinear, for obtaining this term [43]. Alternatively, polar solvation is calculated in the Generalized Born method through the approximation of molecules as charged spheres with the goal of determining the electric field strength based on molecule proximity and amount of ‘de-screening’ if two molecules are close to each other [43]. Of the two, the Poisson Boltzmann method is more resource intensive and time consuming, but also more accurate. However, studies have shown that the MM-GBSA method can provide results that are close to that of MM-PBSA, depending on the system, and specifically that this method also can provide accurate relative binding affinities when comparing bound ligands for a system.

Chapter 3
Classification Model

3.1 Introduction

This study was comprised of three main components. First, an investigation of individual and combination fingerprints through use of the known binders and non-binders of the DUD-E dataset was carried out, where the selected fingerprints would be compared against the random selection method. Next, there was further exploration using these fingerprints within a smaller dataset to select a similarity threshold value. Finally, these non-binding dataset parameters were utilized for creation of a model that was subsequently tested with two Mpro external test sets. The final model was utilized to assist in a virtual screening procedure of the 'Chem-space virtual screening' dataset, containing 3,878,821 compounds.

Databases such as the RCSB Protein Data Bank [44], and PDBbind database [45] provide high quality structural data and curated experimental binding/inhibition/dissociation constants. However, these datasets, by definition, contain only binding information for these complexes, and do not contain known non-binding crystal structure complexes and information. Despite the lack of non-binding data, it is important to provide well-balanced datasets to be used in machine learning models. Existing strategies supplement this data through two main approaches.

The first approach is through random selection, where a potential non-binder is randomly chosen from the binder pool of ligands. In this case, it is assumed that a randomly chosen ligand is likely not to bind. Another strategy is the use of specifically generated 'decoy' molecules, such as with the DUD-E dataset [14]. This is a more rigorous approach, where potential non-binders are created based on a protein and its binding dataset. However, using specific property information can introduce biases. This strategy is very restricted in terms of the target space covered and computationally expensive to generate the non-binder dataset prior to training. To implement an efficient process for selecting non-binders for training a model, we propose to use a fingerprint-based selection approach. Fingerprints are widely used

for finding molecules with similar or dissimilar structures [17]. Utilizing the fingerprints to pick the non-binders could offer a more rational way for training models and may improve the overall learning when compared to the traditional random non-binder selection method. Nevertheless, it is important to verify the general application of this approach for various targets, as the performance could change with different targets.

Therefore, we studied the impact of the fingerprint-based negative data selection method on the model performance. We used a simple classification model and compared the performance of the models using gold-standard datasets. We explored different factors, including the type of fingerprint, the proportion of the non-binding data, and the performance with different classification algorithms. Types of fingerprints that will be utilized in this study were chosen from RDKit and include Daylight-based substructure ('RDKit' implemented), 'Topological' (substructure/layered fingerprint), key-based ('MACCS'), and circular (extended-connectivity subtype, 'Morgan').

3.2 DUD-E Fingerprint Screening

To evaluate the DUD-E dataset of actives and decoys, various fingerprints and their combinations were examined and assessed through analysis of top scores across DUD-E targets for binding/non-binding predictability and contribution breakdown within multi-fingerprints.

3.2.1 Methodology

For each DUD-E target (102 total), molecular fingerprints were calculated for both the 'binding/active' (1) and 'non-binding/decoy' (0) ligand sets using Python with the RDKit package [46] and then stored as bit strings, along with the ligand's 'actual value,' meaning its label of 'active' (1) or 'decoy' (0). Each of the DUD-E target sets contains subsets of known binders, as well as computationally generated 'decoys.' For each molecule in the combined set of 'actives' and 'decoys,' a randomly selected binder was chosen to compute the fingerprint similarity via the Tanimoto coefficient per each fingerprint type. This procedure was replicated 50 times with unique active molecules chosen, and then an average value was computed per

the selected individual fingerprint types (RDKit, Morgan, MACCS, Topological). Data values were stored in 'Pandas' data frames per target, to be used for further calculations. This procedure is outlined in **Figure 6**.

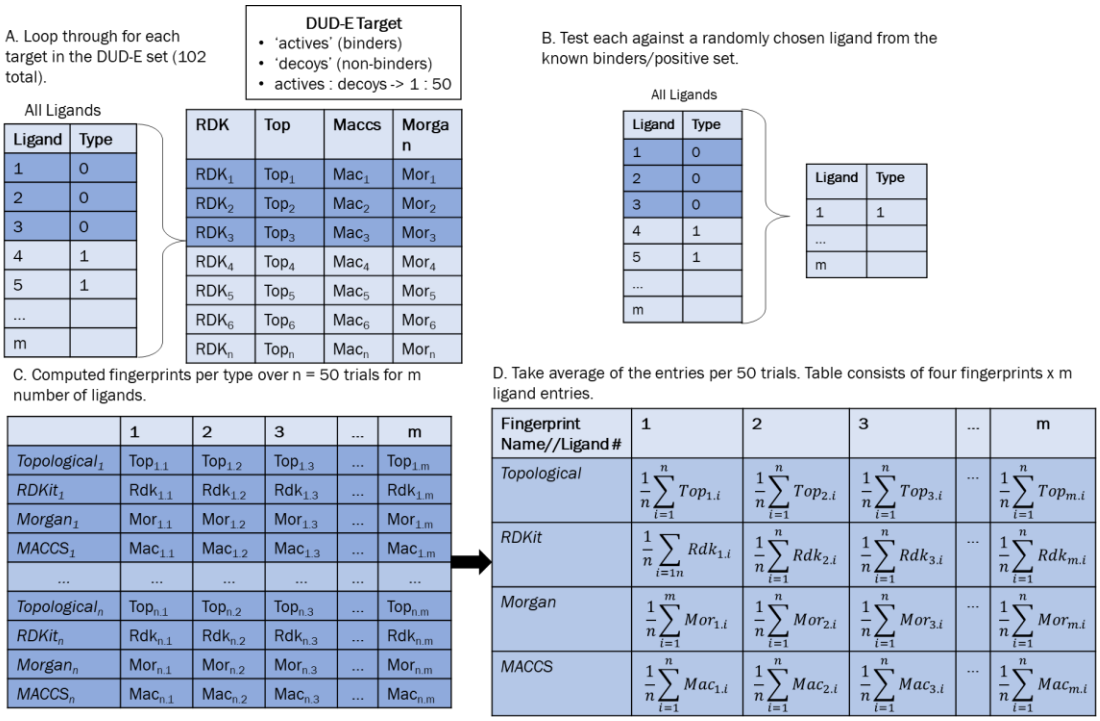


Figure 6. Workflow for DUD-E fingerprint generation. A. Each target in the DUD-E dataset was processed (102 total). B. A randomly chosen ligand was tested against the known binders/positive set. C. Fingerprints were computed per type over n = 50 trials for m number of ligands; an average of 11,424 ligands per target. D: The entries were averaged across the 50 trials per each target and each ligand.

Next, combination fingerprint scores were determined through using each of the ‘minimum’, ‘maximum’, and ‘average’ methods. Given an array of single fingerprint values for a particular ligand of a target protein, the ‘minimum’ function took the smallest value from the set as the score for that ligand, the ‘maximum’ took the largest value, and the ‘average’ took the average value. Total combination fingerprints tested were:

1. Pairwise fingerprints (6): minimum, maximum, average (18 total)
2. Triplet fingerprints (4): minimum, maximum, average (12 total)
3. All combined fingerprints (1): minimum, maximum, average (3 total)

3.2.1.1 DUD-E Prediction Ability of Binders/Non-Binders

The first method evaluated was a 1:1 ratio (actives vs. decoys) dataset, where the total number of ‘actives’ along with their computed similarities were extracted, and then a random and equal subset of the total decoys was extracted from the total dataset. Once this data was compiled, the ideal fingerprints were determined through examination of the prediction ability of the 1:1 subset to evaluate the ‘binding’ and ‘non-binding’ abilities, and which of the minimum/maximum/average methods to pursue in the next steps. This subset was sorted, smallest to largest for the ‘non-binders’ set, and largest to smallest for the ‘binders’ set, and the top 1%, 2%, and 10% were examined for best performance across the DUD-E targets. Additionally, the top 1% of predictions were also examined through heatmap visualization between individual targets.

3.2.1.2 Whole Dataset Prediction Ability for Individual Targets

Next, all ligands from the dataset were examined across all targets for the ‘non-binding’ case. In this case, particular threshold percentages of correct ‘non-binding’ predictions across each method per target were selected (starting with 55%, and then proceeding in increments of 5%) and method type and threshold. The targets that fit these criteria were summed, with a final comparison across the thresholds and methods being made.

3.2.1.3 Unique Ligand Prediction Ability

To further examine the individual ligand prediction abilities among actives and decoys from each DUD-E target for the 50 trials, a uniqueness test was performed. In this case, out of a proportion of the data, any correctly predicted non-binder was extracted from a fingerprint type and assigned an ID value. Among each of the 11 fingerprints (average was taken for the combined), any subtype with the same prediction as another was summed into the 'combined' fingerprint, and individual fingerprints were counted separately. In the case of the two fingerprint combined methods, per each target, the proportion of the values predicted by each individual fingerprint and the combination fingerprint were computed, where an overall average per these combined fingerprints yields the overall contribution of these combined fingerprints versus them alone, with a score between 0-1, where a score of 0 would indicate each of the two fingerprints predict only different ligands, and no ligands are predicted the same. A score of 1 would indicate that all predicted ligands are the same, and that these methods complement each other. The scores between fingerprint methods can be compared to each other, and visualization with heatmaps is an efficient method to determine the individual fingerprints that contribute the most per each of the combined fingerprints.

3.2.2 Results and Discussion

Although a 1:50 proportion of fingerprint data exists for each target within the dataset, the 1:1 prediction ratio was used for this aspect due to the largely disproportionate data, and to be able to observe trends more clearly among the binding/non-binding predictions. Although random chance selection is calculated, the 1:1 approach additionally means that both 'binding' and 'non-binding' strategies are uniform and can be compared to each other; the positive and negative predictions may be considered with equal chance of random selection, and equal weight towards the fingerprint selection criteria, while examining the entire dataset through other analysis methods.

3.2.2.1 DUD-E Binding Ability Prediction

Based on this evaluation, the ‘maximum’ and ‘minimum’ prediction methods performed very similarly, as exemplified in **Figure 7**. Notably, fingerprints perform almost two times better compared to the random selection method for the 1%, 2%, and 10% predictions among the minimum and maximum methods. Additionally, for the top 10% of predictions, the best performance was observed among the ‘MACCS’ and ‘Morgan’ individual fingerprints. Furthermore, from the minimum/maximum top 10% predictions, several of the pairwise fingerprints, namely ‘RDKit/Topological’, ‘Morgan/Topological’, and ‘RDKit/Morgan’, and multi-combined fingerprints, specifically ‘MACCS/RDKit/Topological’, ‘Morgan/RDKit/Topological’ and ‘MACCS/RDKit/Topological/Morgan’, showed the best binding predictive capabilities.

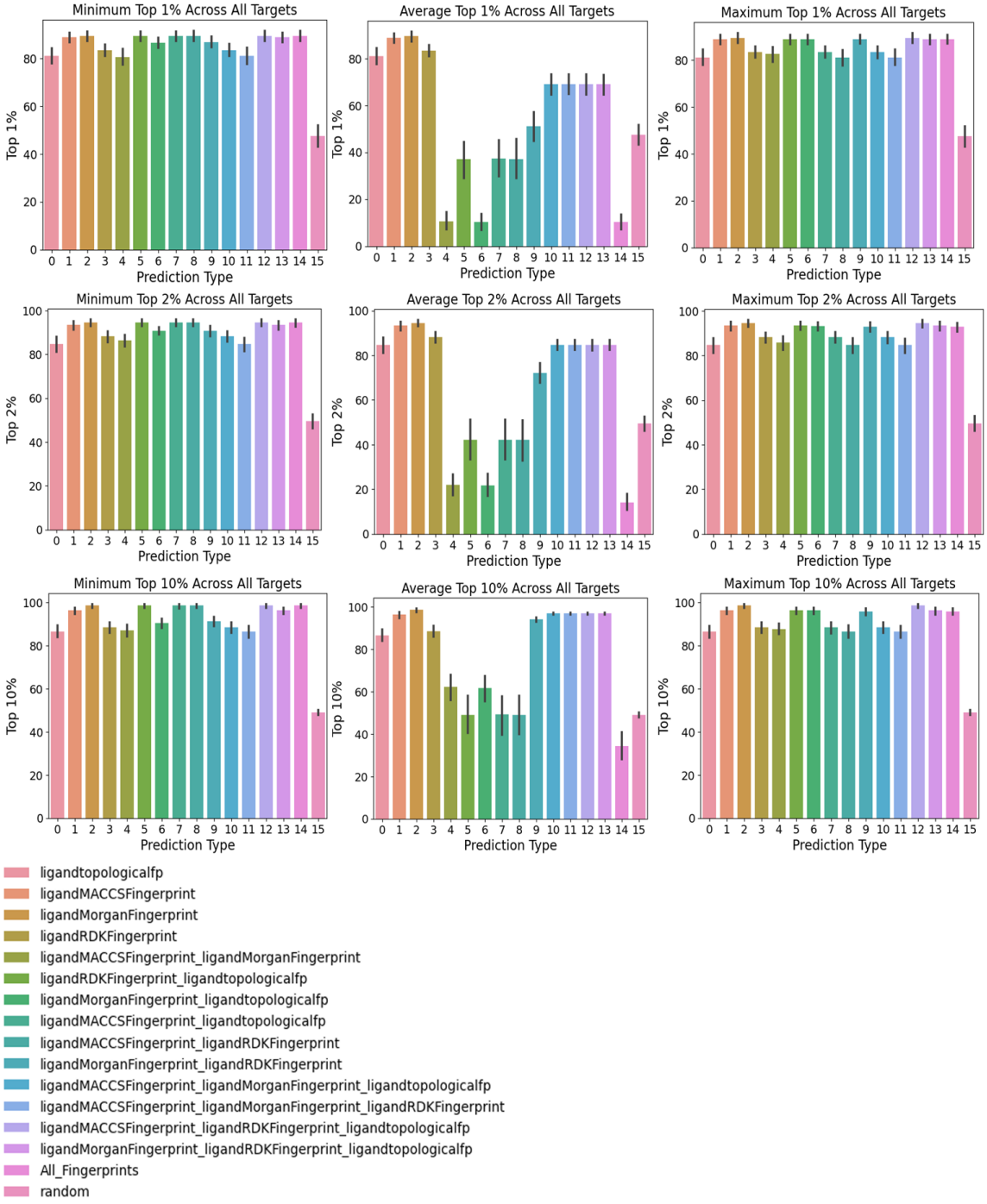


Figure 7. Positive predictions 1:1. Top 1%, 2%, and 10% scores. Scores are calculated per target as a percentage based on the maximum total based on the number of ligands within the top percentage.

3.2.2.2 DUD-E Non-binding Ability Prediction

As shown in **Figure 8**, when the ‘minimum’, ‘average’, and ‘maximum’ fingerprints are compared across the top 1% values, the ‘minimum’ method yields higher predictive values for the individual fingerprints, which remain constant across each of the 1, 2, and 10% metrics, when compared to the combinations. In contrast, the ‘average’ and ‘maximum’ methods perform similarly across methods, with the ‘average’ performing slightly better throughout. Within these methods, and the top 10%, fingerprints providing the highest performance were ‘Morgan’ and ‘MACCS/RDKit/Topological’. However, all fingerprint predictions were only about 5% different from these top predictions. Similar to the positive 1:1 prediction method, fingerprints perform just under two times better compared to the random method for the top 1% of predictions.

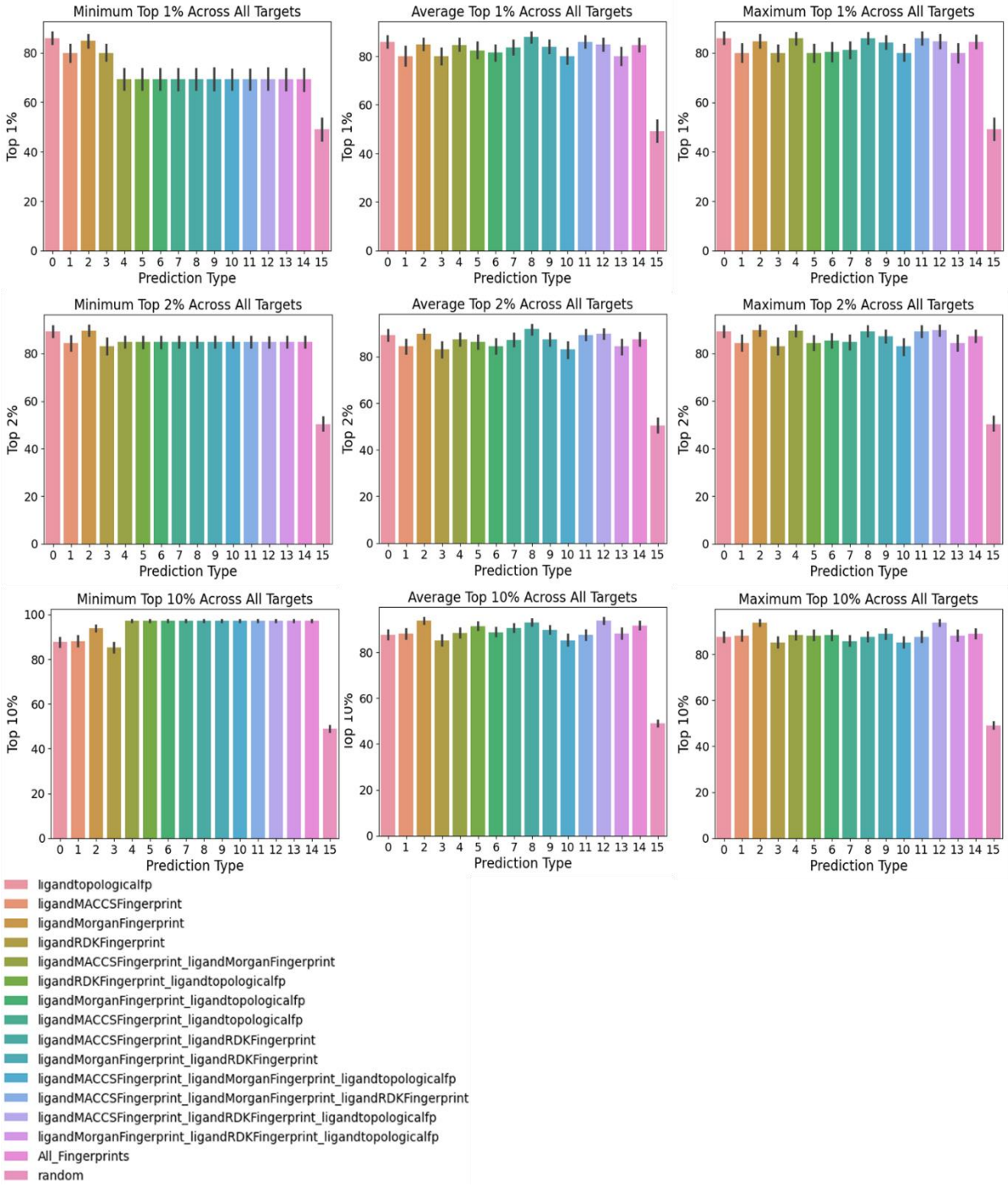


Figure 8. Negative predictions 1:1. Top 1%, 2%, and 10% scores. Scores are calculated per target as a percentage based on the maximum total based on the number of ligands within the top percentage.

To further illustrate the trend, **Figure 9** displays the normalized heatmaps per target for the negative and positive 1% prediction calculations. To account for the differences in top prediction value totals among the targets, the percent of correct predictions was computed per target. Lighter areas of the heatmap show the fingerprint types that predicted non-binders in a better capacity than the darker portions. From this visualization, it is clear that the 'random' method performs significantly worse when compared to any of the fingerprint methods. For example, the random method showed decreased performance when compared to fingerprint methods for targets 'comt,' 'kith', and 'ptn1.' Furthermore, there are targets with poor performance across all methods: 'thb' and 'xiap' target predictions are close to zero. However, this visualization does not give a good indication as to how each fingerprint performs compared to each other; additional trends cannot be well observed in this case. This is why additional analyses were performed, so that fingerprint performance could be visualized in different capacities.

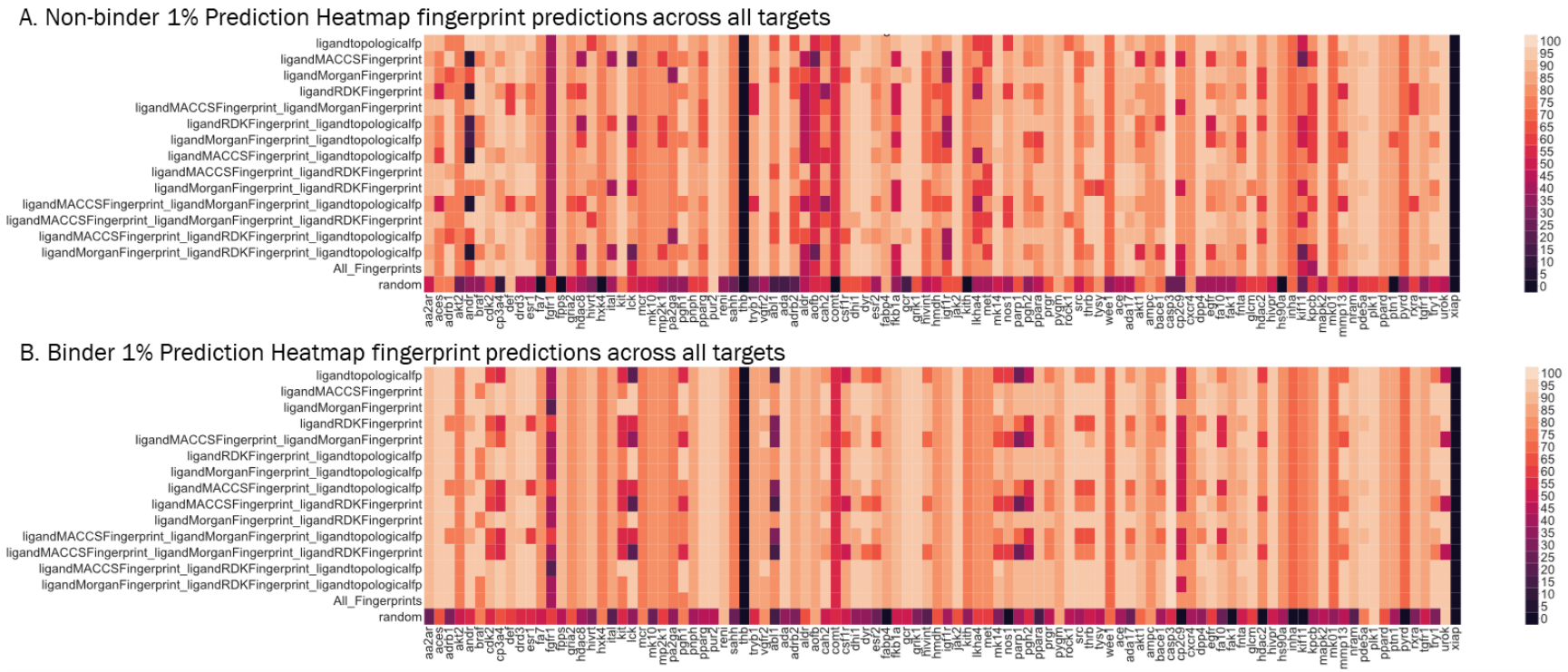


Figure 9. Heatmaps that display top 1% predictions across all targets. A: Top 1% predictions across all targets in predicting non-binders. B: Top 1% predictions across all targets in predicting binders. Lighter regions show more accurate predictions. Clearly, fingerprints predict the top 1% of each of these targets with increased accuracy compared to random selection (last row).

3.2.2.3 Best Fingerprint Based on Top Prediction Ability

To more easily visualize the performance differences among the various method types, comparison per target of total prediction ability was performed, from 55% to 95%. This analysis and graphical representation in **Figure 10** outlines performance across targets. The y-axis represents how many targets are predicted, and the value is determined based on how many targets fit the percentage criteria. In the case of 55%, a target is tallied for a particular selection method if 55% or greater of the ligands are predicted correctly. For this 55% prediction, all prediction methods besides ‘random’ have a high prediction accuracy for the binding ability of 55% of the ligands or greater. The ‘random’ method gives no correct predictions when targets are only tallied where 85% or greater have been correctly predicted. There is increased separation between the prediction ability of the fingerprints until 95%, where ‘Morgan’ and ‘MACCS/RDKit/Topological’ methods predict 95% or greater of the ligands correctly for about 70/102 targets. Clearly, there is an improvement among the combined fingerprints compared to the individual fingerprints.

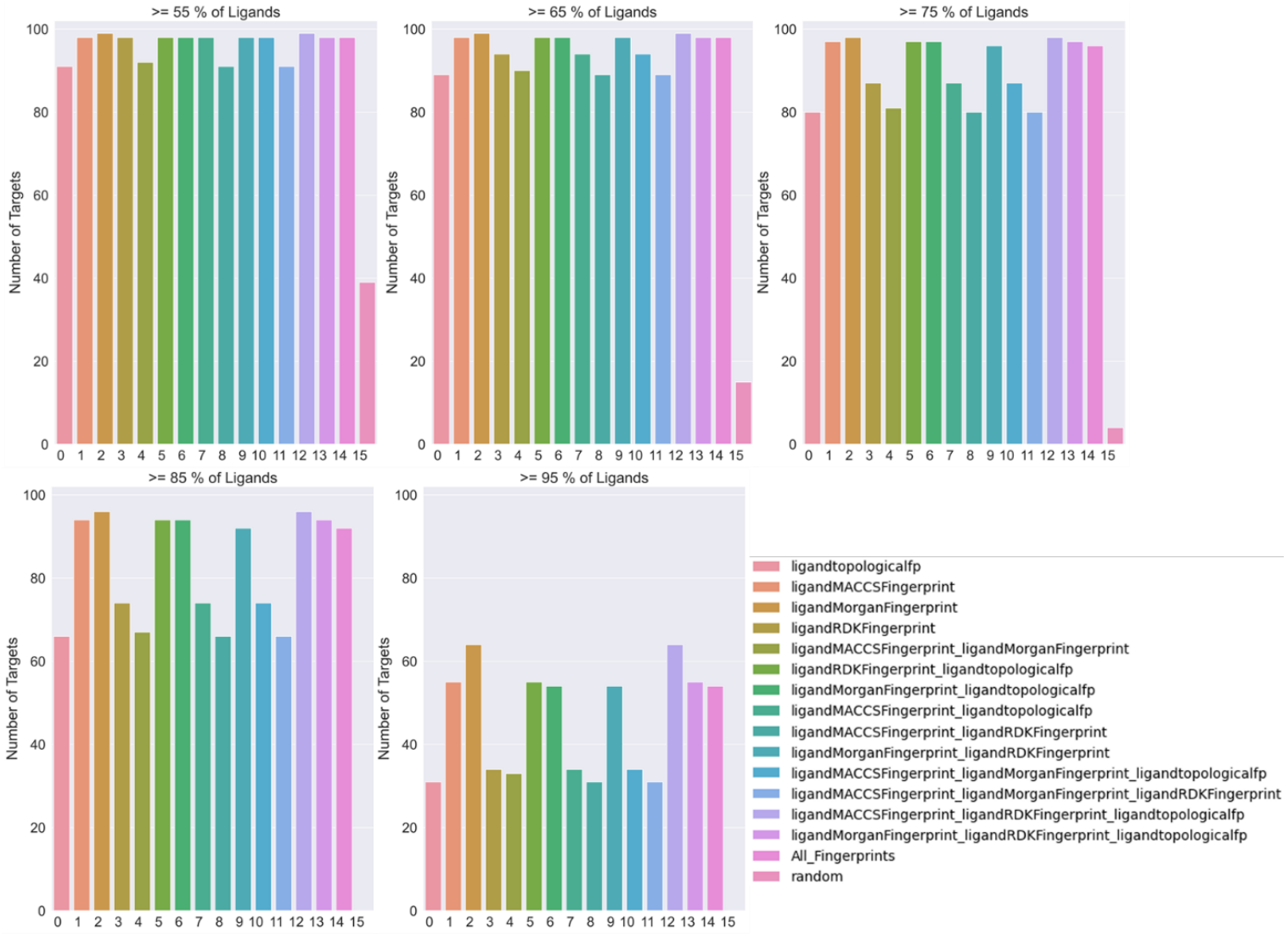


Figure 10. Percentage of correct predictions across all targets using fingerprints and random selection method.

3.2.2.4 Pairwise Fingerprint Comparison

Next, paired results were extracted from the top 50% of predictions, and checked for complementary effects, as seen in **Figure 11**. Since a threshold to define 'binder' vs 'non-binder' has not yet been imposed, using 100% of predictions would not yield informative results. Thus, the top 50% of predictions were selected for this analysis. Per target and fingerprint pair type, a fraction was computed per each of the contributing fingerprints to the pair. This value was then averaged across all targets to give an overall contribution of each individual fingerprint's contribution to the score, and the combined contribution to the score (0-1). These combined values were organized into the heatmap, as shown in **Figure 11**, where the highest combination is between the 'RDKit' and 'Topological' fingerprints, where a fraction of 0.62, greater than half of the predictions, are through a joint contribution. These observations would indicate that, since the performance of this fingerprint combination is greater than the individual fingerprint contributions, these combination fingerprints show a complementary effect. The 'Morgan/MACCS' fingerprint combination has the next best score, with the joint contribution equaling almost half (0.49) of the total contributions.

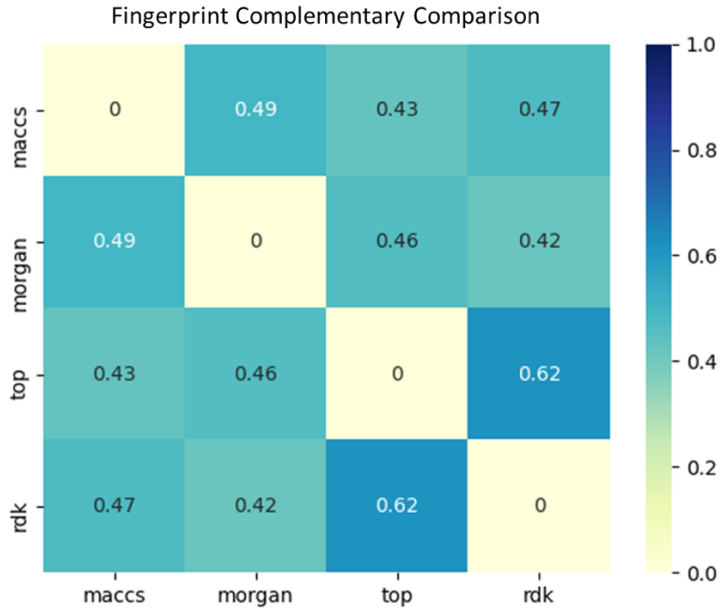


Figure 11. Combination fingerprint pairwise uniqueness value comparison.Comparison among combined fingerprints for the top predicted 50% per fingerprint type.

3.2.2.5 Combined Fingerprint Unique Prediction Ability

Although this method does not consider all sub-combinations within the three and four type fingerprints, these are considered for the triplet and pairwise when they are compared to each other with other methods. The objective is to see how each type of combination fingerprints complement each other. Among the fingerprint types, the fingerprints that were the best at predicting the correct binders were the total combination, or 'all fingerprints.' The top three were: 1) All_fingerprints (0.76), 2) MACCS/Topological/RDKit (0.63), and 3) RDKit/Topological (0.62), represented graphically in **Figure 12**.

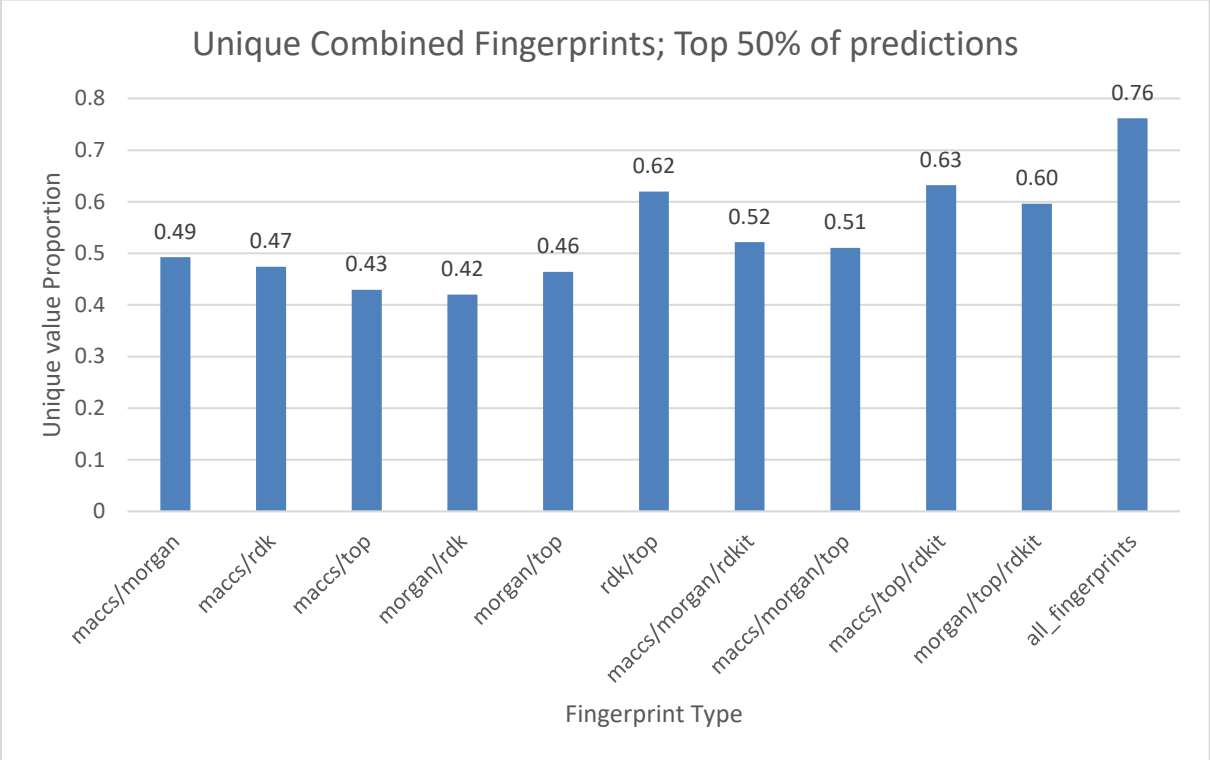


Figure 12. Numbers of unique total combination fingerprints. Values were normalized in the range of 0-1 based on the total number of ligands per target.

3.2.2.6 Conclusions

Based on the 1:1 prediction of non-binders, the ‘average’ method was deemed to be the most suitable for use with other analysis methods, and to make a non-binder prediction. However, the ‘minimum’ and ‘maximum’ methods performed very similarly across fingerprints. Within the non-binder predictions, fingerprints that proved to be the best at making predictions were ‘Morgan’ and ‘MACCS/RDK/Topological’. Within the 1:1 ‘binder’ predictions, the ‘minimum’ and ‘maximum’ methods were the best performing and showed improved predictive capabilities with the paired combined fingerprints ‘RDKit/Topological’, ‘Morgan/Topological’, and ‘RDKit/Morgan’; triplet fingerprints, ‘MACCS/RDKit/Topological’ and ‘Morgan/RDKit/Topological’; and the all-combined ‘MACCS/RDKit/Topological/Morgan’ fingerprint.

Further investigation into the performance of the non-binder ‘average’ method was performed. When all targets were observed for these through a heatmap visualization, it was clearly observed that fingerprint methods outperformed the random selection method. Next, through ‘top ability’ predictions, notably the ‘Morgan’ and ‘MACCS/RDKit/Topological’, fingerprints correctly predicted 95% of ligands for 70/102 targets. When complementary effects are investigated for the pairwise fingerprints, the ‘RDKit’ and ‘topological’ fingerprints were deemed to complement each other the most. Finally, the ‘unique prediction ability’ was investigated within the top 50% of the dataset. Within this analysis, the combination with the most complementary effects was the ‘MACCS/RDKit/Topological/Morgan,’ or ‘all fingerprints’ combination.

Based on the performance of the various fingerprint methods, the ‘Morgan’, ‘MACCS/RDKit/Topological’, and ‘MACCS/RDKit/Topological/Morgan’ methods were selected to be further tested with a machine learning classification model, and further compared with the ‘random’ method.

3.3 Classification Model Refinement

3.3.1.1 Dataset Preparation

Given a set of fingerprints to test, non-binding data was generated to be used with a classification machine learning model. To generate non-binding data per sample, the binding data was processed with a Python script, which split the protein features from the ligand features. Next, each PDB ID/sample was iterated through, and a random ligand entry was selected and checked if it met the threshold. Threshold values ranged from 0.2 to 0.6, with intermediate values with a step value of 0.1 between 0.2 and 0.6. In the case of single fingerprints, the Tanimoto similarity between the bound ligand and a potential non-binder were calculated through use of the ligand bit vectors corresponding to this fingerprint. Multiple fingerprints followed this same procedure, but a combined value was used to check against the threshold similarity. If the calculated score value was greater than the threshold, the ligand was logged, so that it was not re-checked, and the procedure was repeated with another random ligand. To improve efficiency, and for simplicity with memory constraints for any large dataset, the non-binders were computed in subsets, and then merged afterwards.

3.3.1.2 Threshold Value Determination

The non-binder fingerprint generation method was next tested with a machine learning classification model from the Scikit-learn Python library [47] using default model parameters and a fixed seed value. Initially, a trial set between the fingerprints was performed with the benchmark PDBbind subset (1300 proteins) and a limited feature set (280 binding site features and 280 ligand features) to test the fingerprint selection threshold values on a small dataset across 1) data proportion, 2) machine learning model type, and 3) fingerprint method types. Proportions of non-binding data were generated to be 1, 2, 5, 10, 20, and 50 times the binding data with the selected ligand chosen based on: A) random selection, and B) fingerprint selection across threshold values (0.2, 0.3, 0.4, 0.5, 0.6). With the PDBbind benchmarking set, top combinations of fingerprints were iterated through for 25 trials each, averaging the AUC score values and calculating the metrics based on the averaged values for true positive, true

negative, false positive, and false negative. This procedure was repeated with machine learning models of random forest, support vector machine, and XGboost, to test for an ideal threshold value to proceed with.

3.3.1.3 Model Training and Cross-Validation

A more comprehensive feature set, of ligand (280), binding site (280), and sequence (1810) features, were generated for the full PDBbind dataset, using a mixture of the Python libraries ODDT [48], RDKit [46], and PyBioMed [49]. Over and under-sampling steps were performed on this dataset as well. Feature data was generated for the PDBbind full dataset, which contained 18,100 known protein binders, after data generation and cleaning. The chosen threshold value was now used for testing with the full PDBbind dataset, to investigate non-binding dataset selection ability with differences between 1) data proportion size, 2) machine learning model type, 3) fingerprint selection threshold values, and 4) fingerprint method types, where each fingerprint used was one from among the top fingerprints determined through the investigation of the DUD-E methods. Models were evaluated through 10-fold cross validation; to provide a rigorous comparison, each cross-validation fold was computed and saved prior to training, to ensure that models were provided with an identical fold of data.

Then, the full PDBbind dataset was used to determine the proportion as well as the fingerprint types to use within the selected threshold value, and comparison for each fingerprint and proportion in the threshold set, along with the random selection was performed. Among the top three fingerprints, three aspects were compared: 1) the summation of the best score value per category, 2) an average across all indices, and 3) a percent improvement metric, comparing the average cross-validation score to the equivalent model random selection.

3.3.1.4 Model Testing with External Test Set

Finally, two Mpro sets were prepared and used as external test sets for models containing non-binding data determined by the optimal fingerprint, and its varying proportions. The first Mpro test set was generated from Protein Data Bank (PDB) extracted Mpro crystal structures. A similarity search was performed in the PDB, and then the corresponding PDB IDs were saved.

These PDB IDs were run using the same Python code for feature generation as per the PDBbind sets. The structures were filtered to be independent from PDBbind (258 unique entries), and 10 times the proportion of binders were used for generating the non-binder dataset using the fingerprint method.

To provide further validation for this model, a second Mpro dataset was used as an additional screening dataset. The binding data was downloaded from Chem-space and named "Part 3 Fight COVID2 set," containing 5346 inhibitors for PDB id = 5RF7. These structures from Chem-space [16] were compiled from structure-based screening calculations. A proportion equal to 10 non-binder dataset was also created for this dataset with the fingerprint method.

Both datasets were evaluated with a range of metrics, and additionally, the prediction probability threshold values were considered, through choosing an optimal threshold value based on the dataset's Cohen Kappa value.

3.3.2 Results and Discussion

3.3.2.1 Threshold Value Determination

A comparison between threshold values was made to determine the optimal value, based on observations made from models trained using the PDBbind benchmark dataset with the limited feature set. This smaller set was used to generate non-binding data for each threshold value (6), model type (3), proportion (6), and filtering type (fingerprint or random) (4), totaling 342 models. The three fingerprints tested were those determined to be the best performing based on the DUD-E analyses. This threshold value was selected through observation of the numerical trends between important metrics, which are exemplified through heatmaps and line graphs in **Figure 13**, which highlight an overall trend.

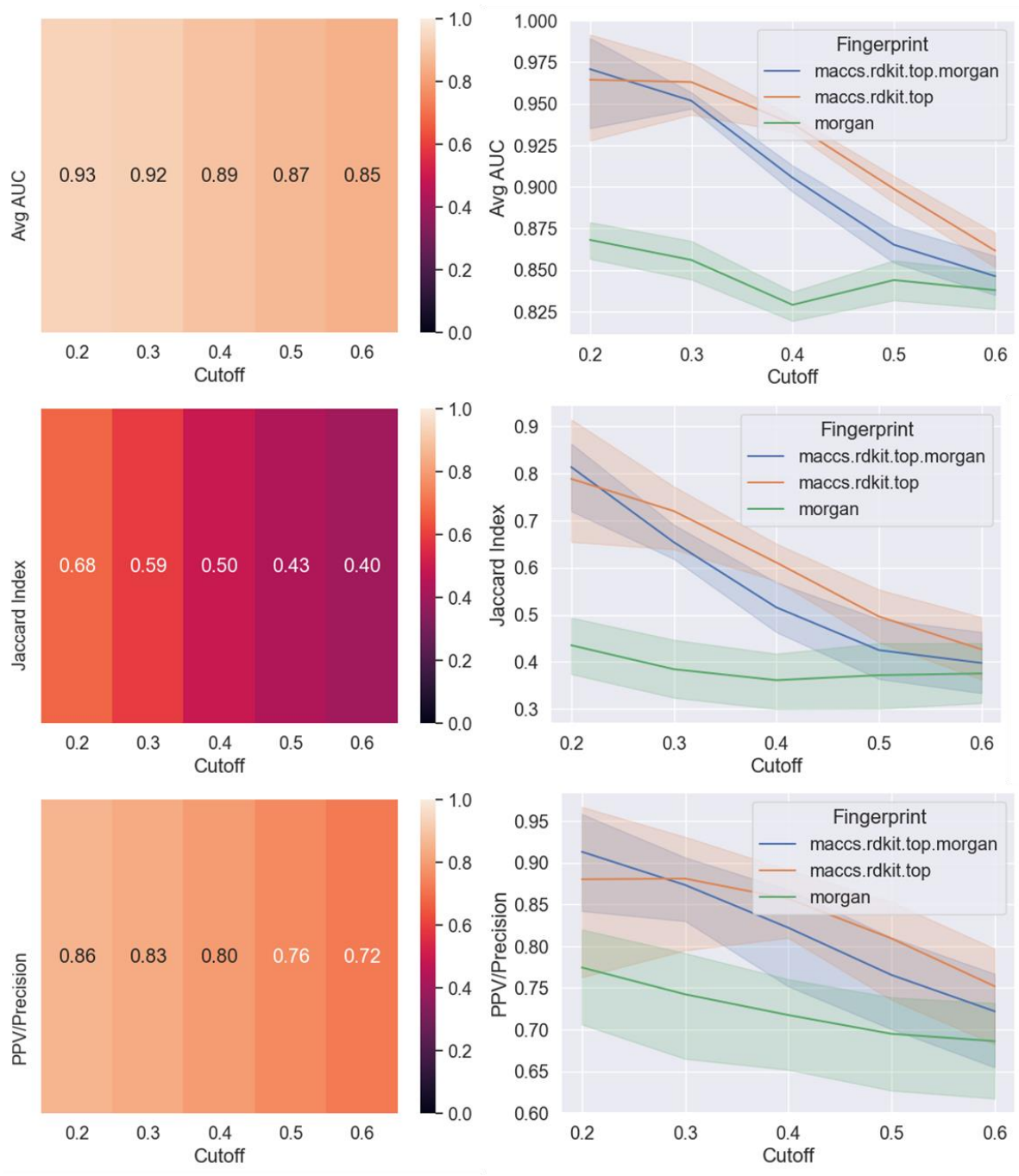


Figure 13. Threshold value selection. Heatmaps and line graphs were generated from average values between the random forest, support vector machine, and XGBoost models.

When the trend across the averaged dataset is observed, clearly, a lower threshold value used for generation of non-binding data resulted in higher observed scores. This can be clearly observed in **Figure 13**, where the averaged results between the three models (XGBoost, support vector machine, and random forest) illustrate the trends between the metrics AUC, Jaccard index, and PPV (positive predictive value/precision).

3.3.2.2 Model Training and Cross-Validation

Next, to evaluate the full PDBbind model, 10-fold cross validation was performed. The averaged values are shown in **Table 3**, where results are summarized per model and type. This summary includes top results for the tallied number of metrics that were 1) the maximum per metric within each of the three model types (support vector machine/XGBoost/random forest), 2) the average value of the computed indices, and 3) an improvement score, which is a percent improvement calculated based on the random score per proportion. Within the 'Max' column, for XGBoost, random forest, and support vector machine the total number is 18, which corresponds to the total number of metrics.

Table 3. Summary of 10-fold cross-validation. Categorized based on proportions (P) and fingerprint types for XGBoost (XGB), random forest (RF), and support vector machine (SVM) models. Within the table, 'P' indicates the proportion of the non-binding data compared to the binding data. The 'Max' column shows, out of the 18 metrics, how many times the fingerprint model combination achieved the maximum metric value, compared to other fingerprint models with the same machine learning algorithm. 'Avg' is the average among the eight indices studied, and 'I' is the importance score, or a percent improvement/decrease compared to the random value. The highest improvement score result among each column is highlighted within this table.

		XGB			RF			SVM		
Fingerprint	**P**	**Max**	**Avg**	**I**	**Max**	**Avg**	**I**	**Max**	**Avg**	**I**
All fingerprints	1	0	0.92	18.51	0	0.89	21.84	0	0.94	0.73
Maccs/rdkit/top	1	5	0.95	22.83	7	0.94	28.13	5	0.96	3.13
Morgan	1	0	0.77	-0.02	0	0.74	1.15	0	0.75	-19.76
Random	1	0	0.77	0.00	0	0.73	0.00	0	0.93	0.00
All fingerprints	2	0	0.93	17.67	0	0.90	17.82	0	0.94	1.39
Maccs/rdkit/top	2	0	0.96	21.46	2	0.94	23.81	2	0.96	3.93
Morgan	2	0	0.79	0.57	0	0.77	0.78	0	0.77	-16.69
Random	2	0	0.79	0.00	0	0.76	0.00	0	0.93	0.00
All fingerprints	5	0	0.94	19.98	0	0.90	17.93	0	0.94	0.52
Maccs/rdkit/top	5	1	0.96	23.40	0	0.94	23.60	0	0.96	2.90
Morgan	5	0	0.79	1.37	0	0.77	1.42	0	0.78	-16.97
Random	5	0	0.78	0.00	0	0.76	0.00	0	0.94	0.00
All fingerprints	10	0	0.94	21.03	0	0.90	20.34	0	0.95	1.55
Maccs/rdkit/top	10	0	0.96	24.42	0	0.93	24.91	1	0.96	3.30
Morgan	10	0	0.78	0.83	0	0.76	1.09	0	0.78	-16.39
Random	10	0	0.78	0.00	0	0.75	0.00	0	0.93	0.00
All fingerprints	20	0	0.94	25.03	0	0.91	25.71	0	0.94	1.86
Maccs/rdkit/top	20	1	0.96	28.53	0	0.93	28.81	0	0.96	3.62
Morgan	20	0	0.77	1.99	0	0.74	2.39	0	0.77	-16.35
Random	20	0	0.75	0.00	0	0.72	0.00	0	0.93	0.00
All fingerprints	50	0	0.94	30.24	3	0.91	32.42	10	0.94	1.33
Maccs/rdkit/top	50	11	0.97	34.36	6	0.93	36.33	0	0.62	-32.71
Morgan	50	0	0.74	2.34	0	0.71	3.15	0	0.74	-20.16
Random	50	0	0.72	0.00	0	0.68	0.00	0	0.93	0.00

3.3.2.2.1 XGBoost Proportion Comparison

When compared to the average 'random' score per proportion, the 'MACCS/RDKit/Topological' combination yields the highest improvements that range from 21 – 34 %, and 'MACCS/RDKit/Topological/Morgan' combination scores range from 18 – 30 %. The 'Morgan' fingerprint scores the lowest of the three, with improvements from 0 – 2 %. Clearly, the 'MACCS/RDKit/Topological' and 'MACCS/RDKit/Topological/Morgan' fingerprints yield the highest performance. Additionally, the 'Morgan' fingerprint performs either similarly, or slightly better than, the 'random' method, depending on the data proportion. The highest performing fingerprint, 'MACCS/RDKit/Topological', performs the best at a proportion = 50 with an average score of 0.97. However, it should be noted that there is only a small difference between this highest prediction and the lowest, where the proportion = 1 and score = 0.95. Additionally, when the best performing selection type ('MACCS/RDKit/Topological'), is compared to 'MACCS/RDKit/Topological/Morgan', the addition of the 'Morgan' fingerprint yields a score only slightly lower, within a range of 2.1 – 3.3 %, across proportions.

3.3.2.2.2 Random Forest Proportion Comparison

Next, the random forest algorithm displayed a similar trend to XGBoost, with the 'MACCS/RDKit/Topological' and 'MACCS/RDKit/Topological/Morgan' methods showing 23.60 – 36.33 % and 17.82 – 36.33 % improvement over the random method, respectively. These two methods also perform very similarly to each other, with differences ranging between 2.2 – 5.3 % across all proportions. In contrast, the 'Morgan' fingerprint has only slight improvements compared to the random method (0.78 – 3.15 %).

3.3.2.2.3 Support Vector Machine Proportion Comparison

Lastly, the support vector machine cross validation results had varied performance among the proportions and fingerprints. The 'RDKit/Topological/MACCS' method again performed the best, but unlike the other methods, the best score was with proportion = 2. Unlike the other

methods, the ‘Morgan’ fingerprint received scores lower than the random method, within the range of -32.71 to -16.39 %.

3.3.2.2.4 Further Discussion

Based on the initial DUD-E fingerprint results, which indicated that the ‘Morgan’ fingerprint performed highly, it was expected that the ‘Morgan’ fingerprint would perform more similarly to the two combined fingerprints. However, other analyses indicated that the combination fingerprints provide complementary effects, so when this is considered, the higher results for the ‘RDKit/Topological/MACCS’ combined fingerprint and the ‘MACCS/RDKit/Topological/Morgan’ are in line with expectations, and, as illustrated in **Table 3,** were expected to perform more similarly. Additionally, a combined fingerprint with greater diversity would be expected to perform better. It is also consistent that if the individual ‘Morgan’ fingerprint is underperforming on this dataset, that the combination fingerprint ‘MACCS/RDKit/Topological/Morgan’ would also have decreased performance.

Additionally, when scores are compared among the proportions for the 10-fold cross validation, results are extremely similar across the machine learning methods, which can be visualized in **Figure 14** XGBoost shows a slight improvement, but besides the score for the support vector machine method with proportion = 50, these results indicate that models using any of the tested proportions perform very similarly compared to each other.

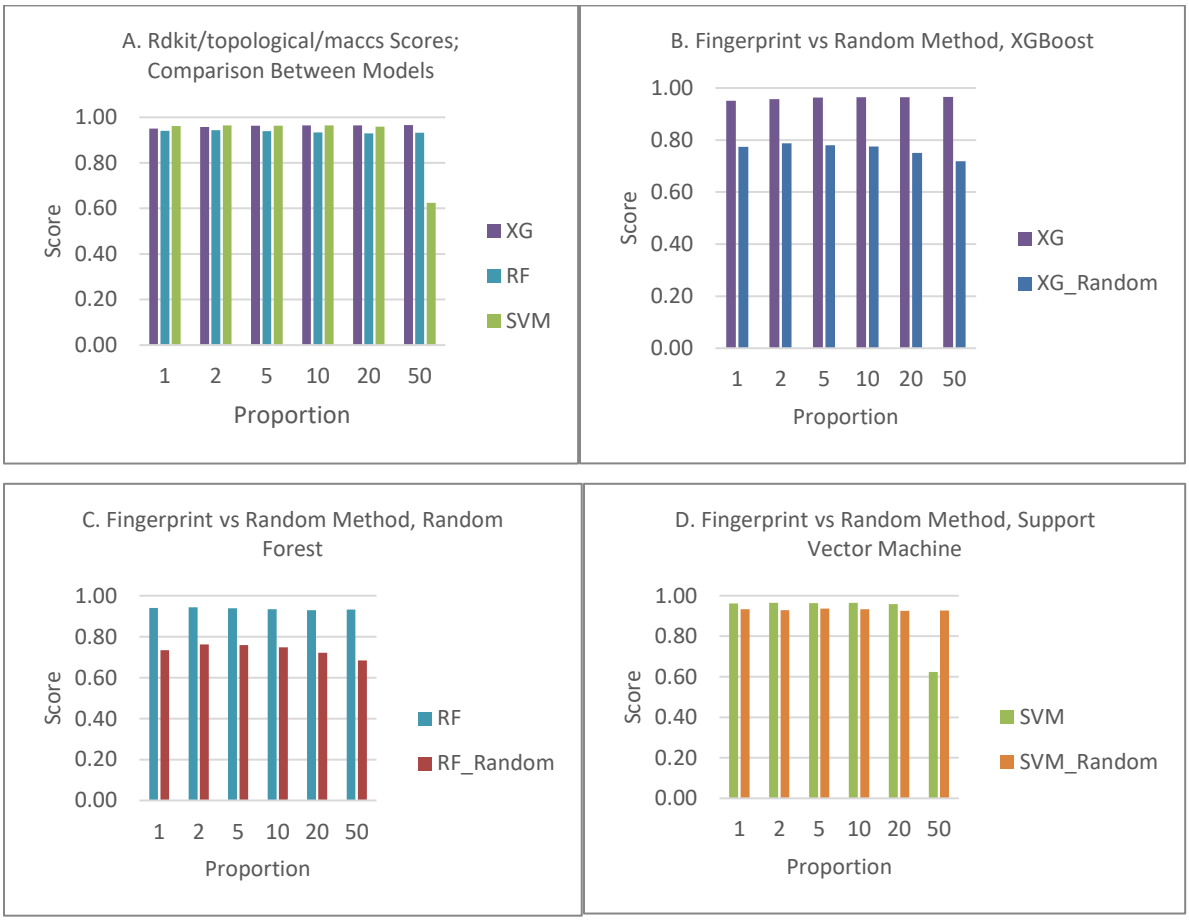


Figure 14. Visualization of machine learning model scores for XGBoost, random forest, and support vector machine non-binding proportions. A. 'RDKit/Topological/MACCS' scores graphed by proportion and compared among the XGBoost, random forest, and support vector machine models. B. Comparison between 'RDKit/Topological/MACCS' fingerprint model and the random model, using XGBoost. C. Comparison between 'RDKit/Topological/MACCS' fingerprint model and the random model, using random forest. D. Comparison between 'RDKit/Topological/MACCS' fingerprint model and the random model, using support vector machine.

Finally, based on the 10-fold cross validation results, the best performing fingerprint combination was the 'RDKit/Topological/MACCS'. Performance among the proportions per model was similar, so these proportions were maintained for further investigation with this 'RDKit/Topological/MACCS' combined fingerprint on the external datasets.

3.3.2.3 Model Testing with External Test Sets

3.3.2.3.1 Mpro Crystal Structure Dataset Observations

Key metrics for the full PDB model tested against an Mpro crystal structure dataset are outlined in **Table 4.** First, for the XGBoost model, very low thresholds yield the best Cohen Kappa scores. Proportion = 10 is best performing, with lower data proportions (5, 2, 1) performing similarly. Proportions 20 and 50 appear to be unfavourable for this model.

Table 4. Summarized results from the Mpro crystal structure dataset, across proportions and models.

Model	P	Threshold	TPR	Precision	Recall	Accuracy	AUC	Fbeta	F1	Jaccard	MCC	Cohen kappa
XG	1	0.90	0.52	0.42	0.52	0.89	0.72	0.49	0.46	0.30	0.40	0.40
	2	0.10	0.54	0.46	0.54	0.90	0.74	0.52	0.49	0.33	0.44	0.44
	5	0.05	0.52	0.44	0.52	0.90	0.73	0.50	0.48	0.31	0.42	0.42
	10	0.05	0.54	0.47	0.54	0.90	0.74	0.52	0.50	0.34	0.45	0.45
	20	0.05	0.29	0.44	0.29	0.90	0.63	0.31	0.35	0.21	0.31	0.30
	50	0.05	0.21	0.38	0.21	0.90	0.59	0.23	0.27	0.16	0.23	0.22
RF	1	0.65	0.41	0.43	0.41	0.90	0.68	0.42	0.42	0.27	0.37	0.37
	2	0.45	0.81	0.45	0.81	0.89	0.86	0.70	0.58	0.41	0.56	0.53
	5	0.45	0.92	0.60	0.92	0.94	0.93	0.83	0.73	0.57	0.72	0.70
	10	0.45	0.94	0.67	0.94	0.95	0.95	0.87	0.78	0.64	0.77	0.75
	20	0.40	0.97	0.66	0.97	0.95	0.96	0.88	0.79	0.65	0.78	0.76
	50	0.40	0.96	0.70	0.96	0.96	0.96	0.90	0.81	0.68	0.80	0.79
SVM	1	0.85	0.40	0.39	0.40	0.89	0.67	0.39	0.39	0.24	0.33	0.33
	2	0.75	0.40	0.38	0.40	0.89	0.67	0.39	0.39	0.24	0.32	0.32
	5	0.85	0.37	0.40	0.37	0.89	0.66	0.38	0.38	0.24	0.32	0.32
	10	0.75	0.40	0.37	0.40	0.88	0.66	0.39	0.38	0.24	0.32	0.32
	20	0.85	0.37	0.40	0.37	0.89	0.66	0.38	0.39	0.24	0.33	0.33
	50	0.80	0.36	0.41	0.36	0.89	0.66	0.37	0.39	0.24	0.33	0.33

For random forest, proportions 10, 20, and 50 are promising due to their Cohen Kappa scores. Threshold values for optimal Cohen Kappa scores were close to 0.5, and the true positive prediction ranges from 0.92 - 0.97 for proportions 5, 10, 20, and 50. The 20 and 50 proportions are the best performing, especially in terms of true positive score.

The support vector machine model exhibits very similar performance across all proportions. Threshold values have been adjusted to ~0.8 on average, for the optimal Cohen Kappa score. Specifically, the true positive rates are slightly higher with the proportions 1, 2, and 10. However, these differences are less than 10% between the highest and lowest true positive score. Overall, the support vector machine predictions for the Mpro crystal structure dataset are not ideal. Cohen Kappa scores range from -1 to 1, so the 0.3 scores are showing some overall prediction ability according to this metric. However, since the rest of the predictions should be > 0.5 to show an informed prediction, this model is not giving adequate performance. Additionally, due to the similar scores, one proportion cannot be chosen over another for the model on this test set.

To note, accuracy scores are high due to the high number of predicted true negative values. An example calculation of this is shown in **Equation 2**, where the accuracy calculation from the first row of **Table 4** is highlighted. Other metrics do not score as well because, unlike the accuracy measure, true negatives are not incorporated into the numerator of these equations.

Equation 2. Accuracy calculation example; first row of **Table 3**.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} = \frac{133 + 2395}{133 + 2395 + 185 + 125} = 0.89$$

For example, precision, recall, and F1 calculations do not incorporate true negative values. Precision, is calculated from true positive, and false positive values, as shown in **Equation 3**. In an ideal model, a low number of false positives gives a high precision score that is close to or equal to 1.

Equation 3. Precision calculation example; first row of **Table 3**.

$$Precision = \frac{TP}{TP + FP} = \frac{133}{133 + 185} = 0.42$$

Recall (also named true positive rate or sensitivity) is calculated in a similar manner but as shown in **Equation 4**, is calculated from the set of known positives; those that are correctly categorized as true positives, and those that are mislabeled (false negative).

Equation 4. Recall calculation example.

$$Recall = \frac{TP}{TP + FN} = \frac{133}{133 + 125} = 0.52$$

Next, F1 score uses true positive, false negative and false positive value to compute the harmonic mean between precision and recall, as shown with an example calculation in **Equation 5** [31].

Equation 5. F1 calculation example; first row of **Table 4**.

$$F1 = \frac{precision \cdot recall}{precision + recall} = 2 \times \frac{0.42 \cdot 0.52}{0.42 + 0.52} = 0.46$$

3.3.2.3.2 Chem-space Inhibitor Dataset Observations

Key metrics for the full PDB model tested against the Chem-space dataset are outlined in **Table 5**. First, at the optimal Cohen kappa values, the XGBoost model has a wide range of threshold values (0.05-0.95). However, within this range, proportions 2 and 5 have mid-range threshold values (0.45), close to the default value of 0.5. In terms of prediction of true positives, proportions 1, 2, and 10 predict binders very well. However, in this case, it appears that proportion = 2 would be the best suited for the model, because this proportion has the best Cohen Kappa value, indicating high predictability for both binders and non-binders, and yields the highest true positive rate.

Next, random forest gives the strongest predictions overall. Notably, the Cohen Kappa and true positive values are the highest. Since the proportion = 5 performs the best across all categories, this would seem the most reasonable to be selected for the random forest model.

Table 5. Summarized results from the Chem-space inhibitor dataset, across proportions and models.

Model	P	Threshold	TPR	Precision	Recall	Accuracy	AUC	Fbeta	F1	Jaccard	MCC	Cohen kappa
XG	1	0.95	0.98	0.86	0.98	0.98	0.98	0.95	0.92	0.85	0.91	0.91
	2	0.45	0.98	0.93	0.98	0.99	0.99	0.97	0.96	0.92	0.95	0.95
	5	0.45	0.92	0.91	0.92	0.98	0.96	0.92	0.92	0.85	0.91	0.91
	10	0.25	0.97	0.91	0.97	0.99	0.98	0.96	0.94	0.89	0.94	0.94
	20	0.05	0.82	0.97	0.82	0.98	0.91	0.85	0.89	0.80	0.88	0.88
	50	0.05	0.58	0.97	0.58	0.96	0.79	0.63	0.73	0.57	0.74	0.71
RF	1	0.70	0.99	0.92	0.99	0.99	0.99	0.98	0.95	0.91	0.95	0.95
	2	0.70	0.99	0.94	0.99	0.99	0.99	0.98	0.96	0.93	0.96	0.96
	5	0.60	0.99	0.90	0.99	0.99	0.99	0.97	0.94	0.89	0.94	0.94
	10	0.55	0.99	0.89	0.99	0.99	0.99	0.97	0.94	0.89	0.93	0.93
	20	0.50	0.98	0.90	0.98	0.99	0.99	0.97	0.94	0.89	0.94	0.93
	50	0.45	0.98	0.96	0.98	0.99	0.99	0.97	0.97	0.93	0.96	0.96
SVM	1	0.85	0.65	0.64	0.65	0.93	0.81	0.65	0.65	0.48	0.61	0.61
	2	0.75	0.73	0.61	0.73	0.93	0.84	0.70	0.66	0.50	0.63	0.63
	5	0.80	0.86	0.53	0.86	0.92	0.89	0.76	0.65	0.48	0.63	0.61
	10	0.75	0.74	0.61	0.74	0.93	0.85	0.71	0.67	0.50	0.64	0.63
	20	0.80	0.83	0.54	0.83	0.92	0.88	0.75	0.66	0.49	0.63	0.61
	50	0.75	0.81	0.55	0.81	0.92	0.87	0.74	0.66	0.49	0.63	0.62

The performance of the three selected models across probabilities, for the combined RDKit/Topological/MACCS fingerprint are visualized in **Figure 15**, where there is shown to be an optimal separation between the binder and non-binder predictions in the XGBoost and random forest models. However, the support vector machine model predicts some non-binders incorrectly as binders within the 0.8-1.0 probability range.

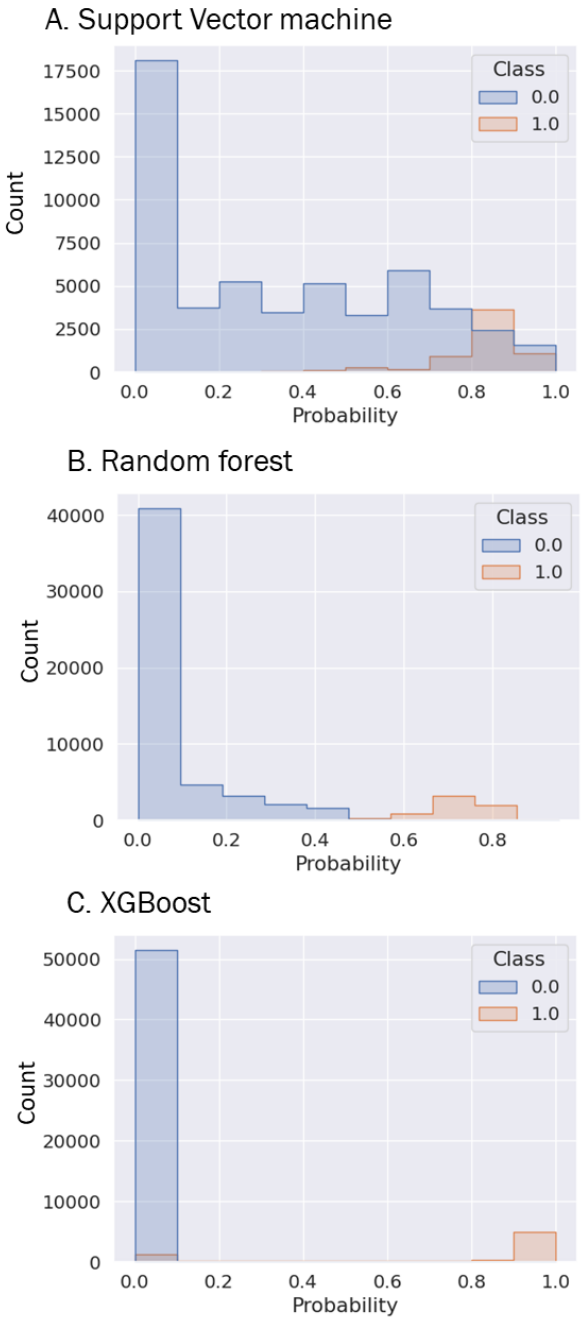


Figure 15. Predictions of binders and non-binders within the three machine learning models.Class 0 represent the non-binders, and class 1 represent binders. A. Support vector machine, proportion = 5. B. RF, proportion = 5. C. XGBoost, proportion = 2.

To address the high scores and concerns of overfitting, specifically for the random forest model, a check of the feature contribution was performed, shown **in Figure 16;** this model contains a high fraction of ligand features, which could be biasing the model to give overfitted scores. Among these selected models, the features can also be visualized for only the random forest and XGBoost predictions. There is no feature breakdown for the version of support vector machine that was utilized in this case, as the format of this model used does not provide feature results that can be extracted. Random forest has a majority of ligand features contributing to the prediction, and only a small number of sequence features. This may explain the very high predictions observed on the Chem-space inhibitors set, as using a model with a high proportion of ligand features could be overfitting/contributing to bias. However, the XGBoost model yields a more equitable distribution of features among the three major categories, with the sequence features contributing to about half of the overall feature weightage, then the ligand features, and then binding site features.

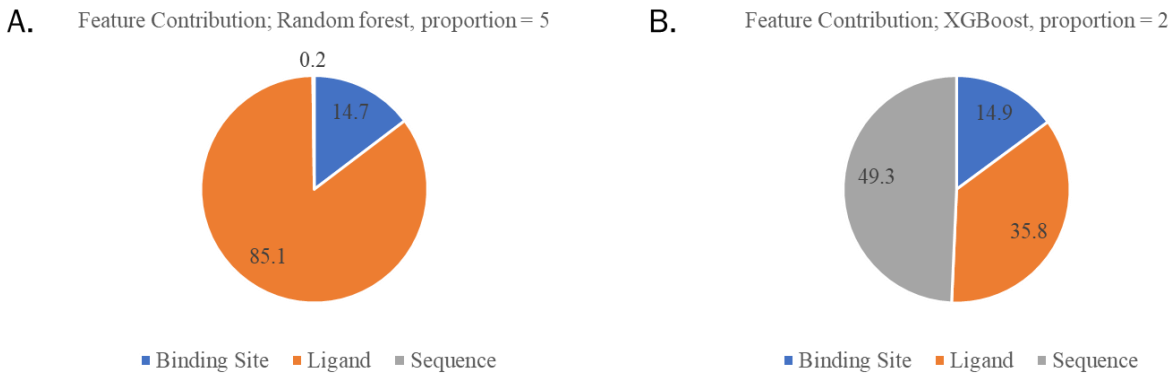


Figure 16. Feature contributions from random forest and XGBoost. In this case, the 'MACCS/RDKit/Topological' fingerprint model was utilized. A. The random forest model with a proportion equal to 5. B. The XGBoost model with a proportion equal to 2.

Finally, a feature importance comparison was made among the top three fingerprint types with a constant proportion value of 2 for the XGBoost model, as illustrated in **Figure 17**. All fingerprint-generated models yielded higher performance compared to the 'random' method. Additionally, within the 'MACCS/RDKit/Topological' model (**Figure 17A**) is a more evenly distributed feature importance, with emphasis on the sequence features (~50 %). This is an indication that the selected model is trained from a variety of types of features, and not biasing results highly towards a particular feature type, especially compared to the feature importance of the other fingerprint selection models and random method (**Figures 17 B, C, D**), where sequence importance is highly weighted, compared to other feature types.

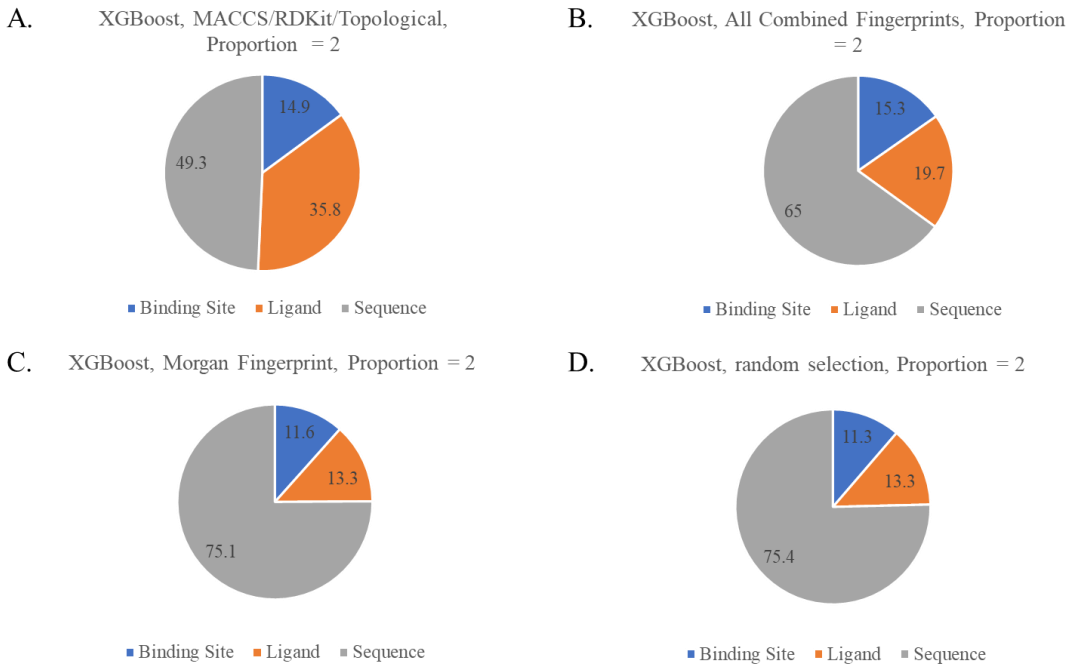


Figure 17. Feature importance comparison among the top three fingerprints. Among the trained XGBoost models with a proportion of 2, graphs depict the percentage contribution of each of the three categories of features for the top three fingerprint types. A. 'MACCS/ RDKit/Topological' fingerprint selection. B. 'MACCS/RDKit/Topological/Morgan' fingerprint selection. C. 'Morgan' fingerprint selection. D. 'Random' fingerprint selection.

Another feature importance comparison was made among the top three fingerprint types with a constant proportion value of 5 and the random forest model, shown in **Figure 18**. The random forest fingerprints consist primarily of ligand and sequence features. Fingerprint selection types were compared, and the selected fingerprint combination, 'MACCS/RDKit/Topological' (**Figure 18A)** utilizes ligand features as a majority. As discussed, random forest received high scores across both external test sets. However, since feature importance is skewed towards ligand ('MACCS/RDKit/Topological' and 'All Combined'), or sequence ('Morgan' and 'random'), these results indicate that overfitting may be occurring in this case. Since both combined fingerprints with a high importance of ligand features received high scores within the 10-fold cross-validation check, this feature type would be the most likely cause of potential overfitting for the random forest model.

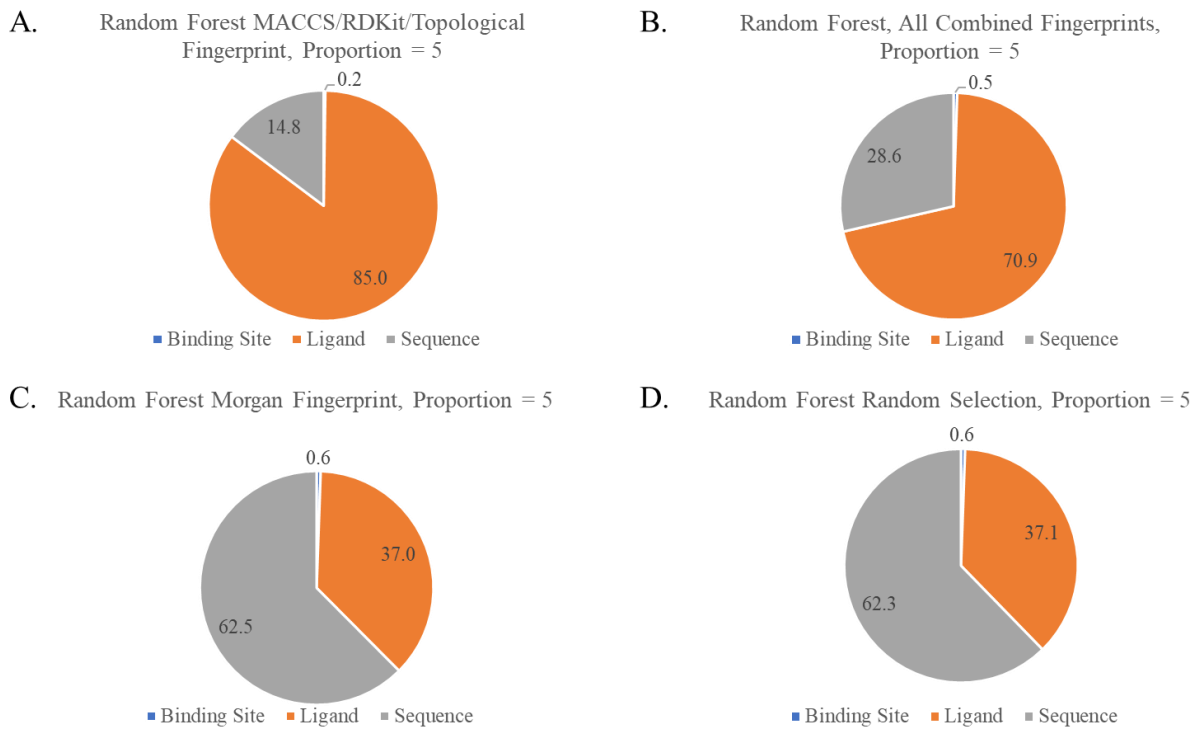


Figure 18. Feature importance comparison among the top three fingerprints. Within the trained random forest models with a proportion of 5, graphs depict the percentage contribution of each of the three categories of features for the top three fingerprint types. A. 'MACCS/RDKit/ Topological' fingerprint selection. B. 'MACCS/RDKit/Topological/ Morgan' fingerprint selection. C. 'Morgan' fingerprint selection. D. 'Random' fingerprint selection.

3.3.2.4 Conclusions

An efficient approach to supplement a machine learning model with informed non-binding data was tested. This approach utilized individual and combination molecular fingerprints, which were determined through an initial effectiveness evaluation with the DUD-E library. The best performing fingerprints, as determined from the DUD-E methods, were the individual 'Morgan' fingerprint, and the combination fingerprint of 'MACCS/RDKit/Topological' and

‘MACCS/RDKit/Topological/Morgan.’ These fingerprints and a corresponding threshold value were evaluated with a benchmark PDBbind dataset across proportions 1, 2, 5, 10, 20, and 50, the machine learning models support vector machine, XGBoost, and random forest, and the selected fingerprints. A random selection method was also tested with each of the proportions, and models. Across methods, the lowest threshold, 0.2, was deemed to yield the highest machine learning scores on the PDBbind benchmark dataset.

Additional evaluations were performed with the selected fingerprints and random selection method, proportions, and models, with 10-fold cross validation on the full PDBbind dataset. Per model and fingerprint, results were very similar across the metrics, with the ‘MACCS/RDKit/Topological’ fingerprint yielding the highest scores. All proportion combinations of this combination fingerprint were then tested against two external datasets: a PDBbind derived Mpro dataset, and a Chem-space inhibitors dataset. Although the dataset proportions provided similar performance, from these external test sets it was determined that the best proportions for the ‘MACCS/RDKit/Topological’ fingerprint per model are proportion = 5 for random forest, proportion = 5 for SVM, and proportion = 2 for XGBoost.

3.3.2.5 Future Directions

Future directions for this work are to expand the type and quantity of fingerprints that were used in the initial DUD-E dataset screening. Additionally, we plan to implement more testing and checks with another dataset to ensure that no forms of bias are contributing to the performance. Ideally, an experimental dataset would be used for testing, and not a computationally generated dataset. Although diversity is typically important for fingerprint screening, we still would want to be careful when considering very well-performing fingerprints, such as 3D pharmacophore, as this may introduce its own biases into the non-binding data for the dataset.

Chapter 4
Structure-Based Analysis

4.1 Introduction

High throughput virtual screening, through use of molecular docking, can be used as a standalone strategy to screen for compounds with favourable binding affinity [50]. Given the advancements in ML-based modeling, it would be efficient to combine this approach with traditional methods to improve the speed and hit rate of the screening process. Therefore, we performed classical structure-based analyses of the machine learning model's pre-filtered hit candidates from the 'Chem-space' dataset to assess the quality of the hits, mode of binding, and their binding affinities. We believe that this would not only help reduce the screening time but might improve the quality of the hits.

The antiviral target chosen for this screening effort was the SARS-CoV-2 Main Protease (Mpro). As a potential target for SARS-CoV-2, its main protease is attractive, since this protein is not homologous to human proteases, and therefore the chance of negative off-target effects is decreased [51]. Additionally, this protein is a dimer that is characterized as having three domains, with the active site between domains I (residues 1-101) and II (residues 102-184), and joined by linker regions, as shown in **Figure 19** [52]. At the binding site, HIS41 and CYS145 form a catalytic dyad. The Mpro target has been studied through various methods, and interactions between residues HIS41, GLY143, SER144, GLU166, and GLN189 have been identified as important for potential drugs [53].

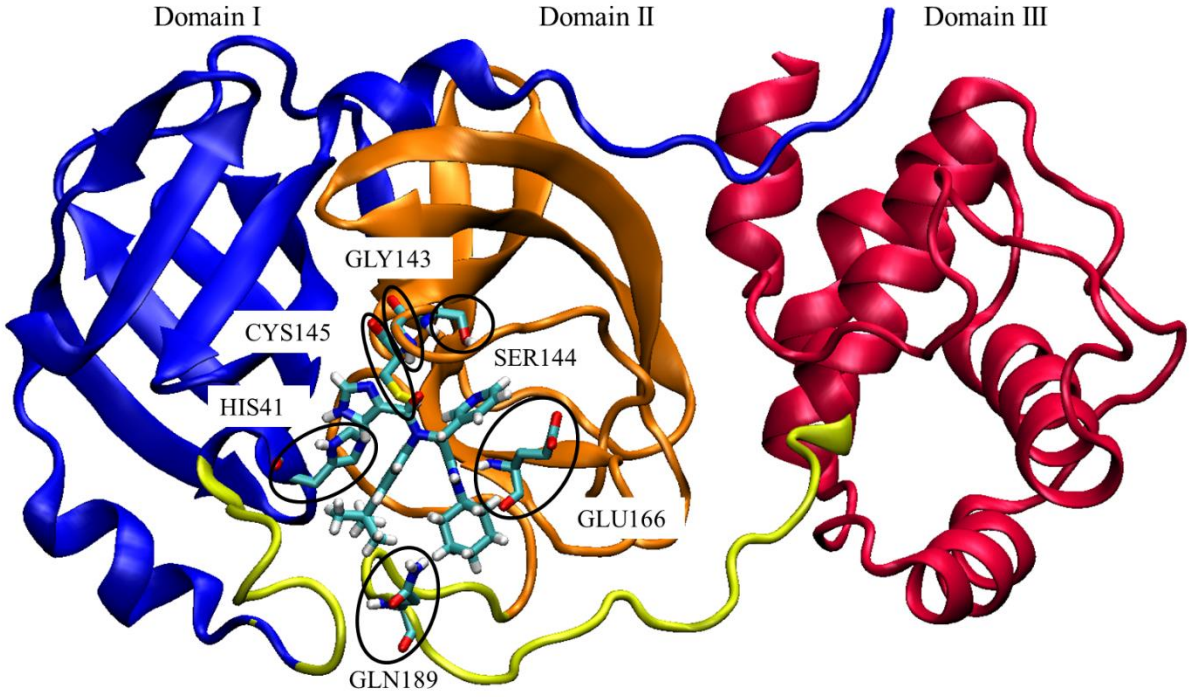


Figure 19. Mpro visualization of domains. Depiction of reference ligand bound in active site of monomeric Mpro, along with key residues, generated in Visual Molecular Dynamics (VMD) [54].

4.2 Methodology

4.2.1 Molecular Docking and Pose Filtering

The determined machine learning model combinations were used to make predictions for 3.9 million compounds within the Chem-space virtual screening database. The dataset was split into 785 sections, due to memory constraints. To ensure that a manageable number could be docked, 0.1% of the total dataset, or the top ~5000 structures were selected for molecular docking. Since a probability threshold was required to be selected, probabilities starting at 0.9, with a 0.1 probability interval, were tested to find the combined threshold value resulting in a

value of total structures closest to 0.1%. Next, this ligand set was further processed through docking with AutoDock Vina.

Bash and Python scripts preprocessed each ligand; these scripts automated the conversion of the Chem-space molecules from the .SMILES format to the .PDB 3D format (using the Python package RDKit functionality). To prepare for docking, scripts from the Scripps Institute converted protein and ligand files to .PDBQT format. To generate configuration files, calculation of the centre of mass of the bound ligand and a grid box size of 2.9 × the radius of gyration of each potential ligand binder were carried out. After the docking procedure, Python scripts collected important information (binding affinity output and the corresponding molecule IDs) into a summary file, where results were ranked according to binding affinity of the top pose from each molecule.

For each of the ten compounds with the highest binding affinity, our in-house PyMOL plugin 'PoseFilter,' was used for filtering the docked poses [55]. A combination of Simple Interaction, Structural Protein Ligand Interaction Fingerprint (SPLIF), and Root Mean Squared (RMS) comparisons were made to identify three top-scoring unique poses. Fingerprint similarity is computed using a similarity coefficient (dice, or a tailored metric for SPLIF) to determine the similarity between two protein-ligand complexes. Results are sorted into folders based on a user specified 'uniqueness' threshold value and outputted into graphs to visualize similarity. The initial parameters of 2.0 Å for RMS threshold and 0.5 for simple interaction and SPLIF (default PoseFilter parameters) sorted uniqueness. Parameters were incrementally changed to achieve a consensus of three unique poses for fingerprints and RMS by 0.1 (minimum of 0.3) and 1.0 Å, respectively. Further adjustments were made until a consensus pose was determined; parameters and poses are shown in **Appendix A**.

4.2.1 Molecular Dynamics

Preparation and determination of protonation states were computed for the Mpro (PDB id = 6w63) protein at a neutral pH with Schrodinger Maestro's Propka [56], and then PyMOL was used for final protein and ligand preparation. Molecular dynamics simulations using Amber 18

were run for these 30 structures (3 poses, each with 10 hits), each with a simulation time of 30 ns. The protein ligand complexes were described using the AMBERff14SB and GAFF2 forcefield as outlined in Chapter 2. The complexes were solvated using a TIP3P water box and electro-neutralized using NaCl- ions. The MD protocol described in Chapter 2 was followed for equilibrating the complexes, and the dynamic trajectories were analyzed for their stability, conformational changes, ligand binding affinity and interactions.

The free energy of binding for the ligands was computed using the MM-GBSA calculations. The binding free energy values for the ligand poses was estimated by calculating the energy difference between the protein-ligand complex, and the receptor and ligand (complex – receptor – ligand), using the last 10 ns of the trajectory, with a total of 1000 frames. The dynamics of the hits were compared to the dynamics of the crystal structure of Mpro bound to a non-covalent X77 inhibitor (PDB ID = 6W63, positive control in this thesis).

4.3 Results and Discussion

4.3.1 Molecular Docking and Pose Filtering

The combined machine learning model screening resulted in 4878 Chem-space molecules. Of these molecules, the best scoring, or lowest docking energy values, were ≤ -10 kcal/mol. From the docking results, the maximum screened value was -6.3 kcal/mol, and the median/mean was -8 kcal/mol. The distribution of these scores is shown in **Figure 20.**

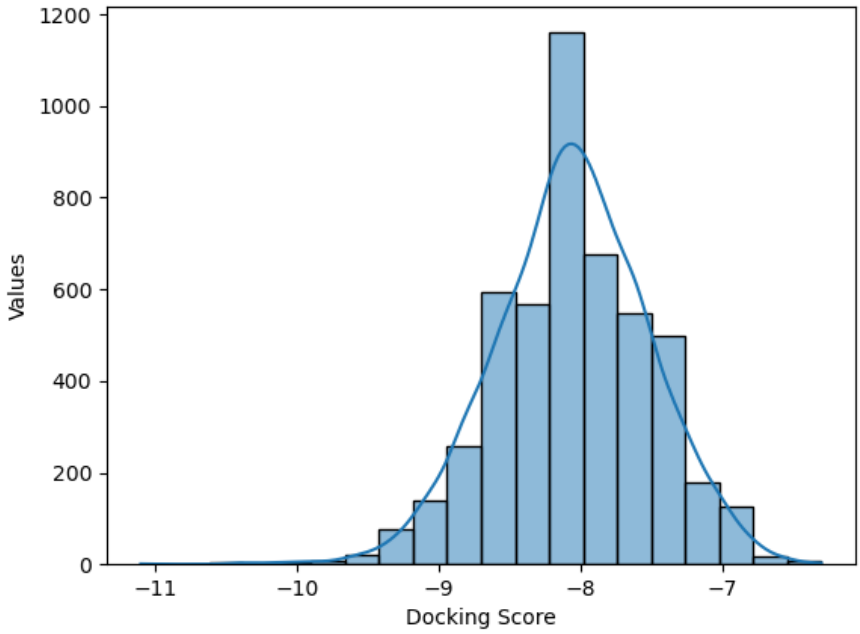


Figure 20. Distribution of top 0.1 % docking scores. Distribution of the docking affinity of 4878 Chem-space molecules.

In addition, the ID values assigned to the top ten poses, along with their binding affinity, are listed in **Table 6**. Further information about these structures, such as their SMILES strings, 2D structures, and Chem-space IDs, are outlined in **Appendix B.** As a note, the naming convention used for structures was based on the batch number and molecule number out of the batch. When docked poses are examined, the final value after the underscore corresponds to the pose number.

Table 6. AutoDock Vina docking scores of the three unique poses of the top ten hits.

Complex	Unique (Pose1)	Binding affinity (Pose 1)	Unique (Pose2)	Binding affinity (Pose 2)	Unique (Pose3)	Binding affinity (Pose 3)
264_4209	1	-11.1	6	-8.5	7	-8.3
267_3318	1	-10.4	2	-9.9	9	-8.8
268_874	1	-10.2	3	-8.5	5	-8.2
268_2452	1	-10.0	3	-9.1	5	-8.8
269_1725	1	-10.4	2	-10.2	7	-10.0
269_1816	1	-10.1	4	-9.4	5	-9.3
269_3556	1	-10.0	2	-9.2	4	-8.9
373_2580	1	-10.2	5	-8.5	6	-8.4
413_1761	1	-10.4	3	-8.8	4	-8.4
574_2168	1	-10.0	4	-8.9	6	-8.6

4.3.1.1 MM-GBSA Calculations

All three unique poses for the top 10 hits were subjected to 30 ns molecular dynamics simulations, and the binding free energy for the ligands was calculated using the MM-GBSA method. **Figure 21** provides a summary of the estimated free energies and the standard error values. Our previous analyses on the Mpro crystal structures showed that < -30 kcal/mol free energy is a good threshold for classifying strong binding Mpro inhibitors [52]. Further, as can be noted, the positive control (X77, 6W63 ligand, green bar on the far right in **Figure 21**) used in this study also had a binding free energy of -30.8 kcal/mol. The free energy plot illustrates that at least one of the three poses for the hits has significant affinity comparable to that of the

control molecule. This emphasizes the importance of the combined screening protocol incorporated in this thesis. While the affinities are significant, it is essential to check whether the bound ligands remain stable in the binding pocket and satisfy the key interactions observed in other Mpro inhibitors.

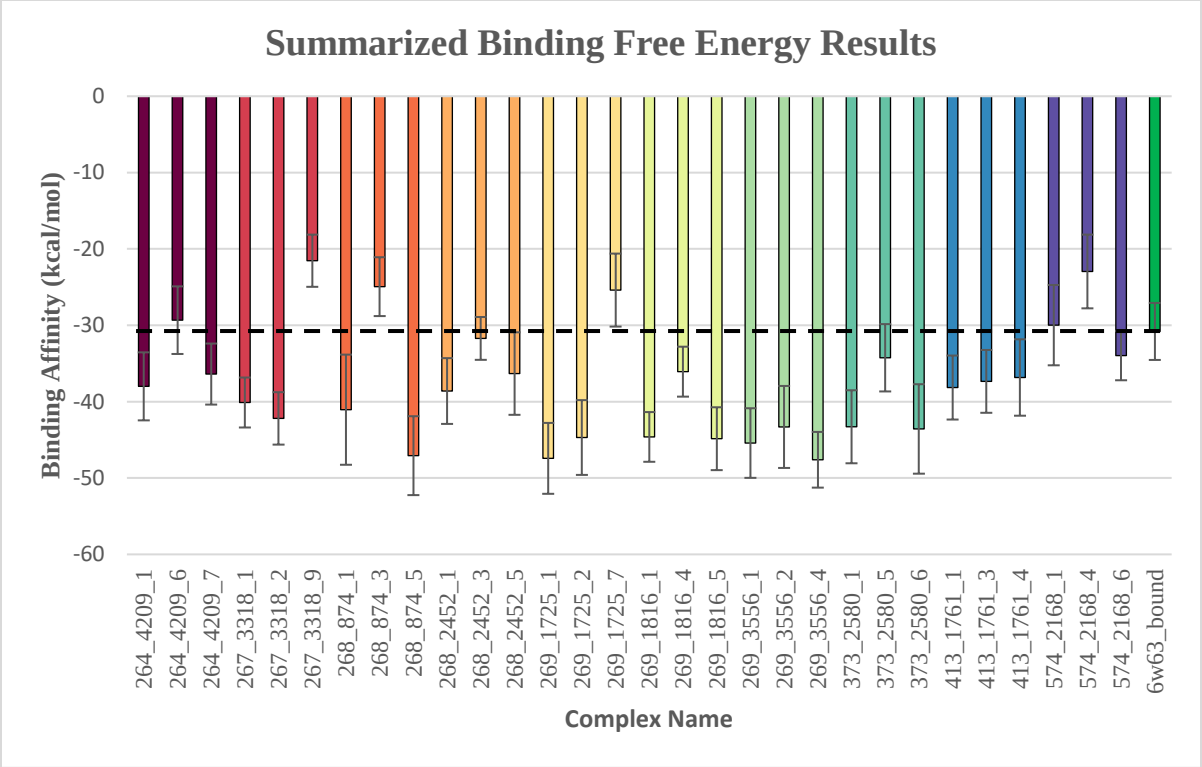


Figure 21. MM-GBSA calculations for the ten potential hits, three unique poses each, along with the native binder from PDB ID = 6W63 (green bar on the far right).

4.3.1.2 RMSD Results

Overall protein RMSD results indicate that of the 31 screened structures, almost all have relatively low fluctuations within the backbone, with all but one structure below 2.5 Å RMSD, as shown in **Figure 22**. The majority of structures (27/30) also lie within 1.5 – 2.0 Å, the same range as the bound crystal structure (2.0 Å).

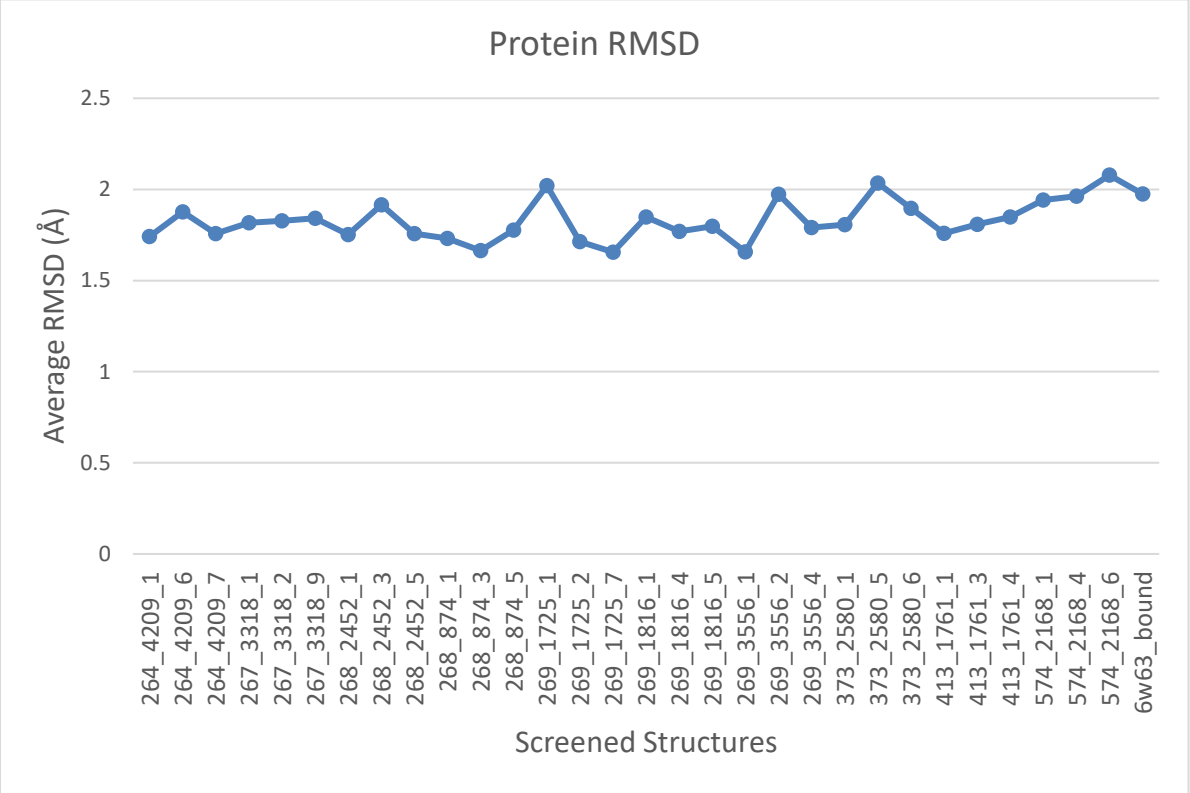


Figure 22. Average RMSD per complex.

Ligand RMSD varies more widely than the protein RMSD. Discounting the natively bound 6w63, which achieved the lowest ligand RMSD (0.6 Å), of the screened compounds, 9 had ligand fluctuations of less than 1.5 Å, and 15 structures achieved an RMSD less than 2 Å. Ranked ligand RMSD results are outlined in **Figure 23**, plotted against their binding free energy. All ligands achieved less than 3.5 Å of fluctuations.

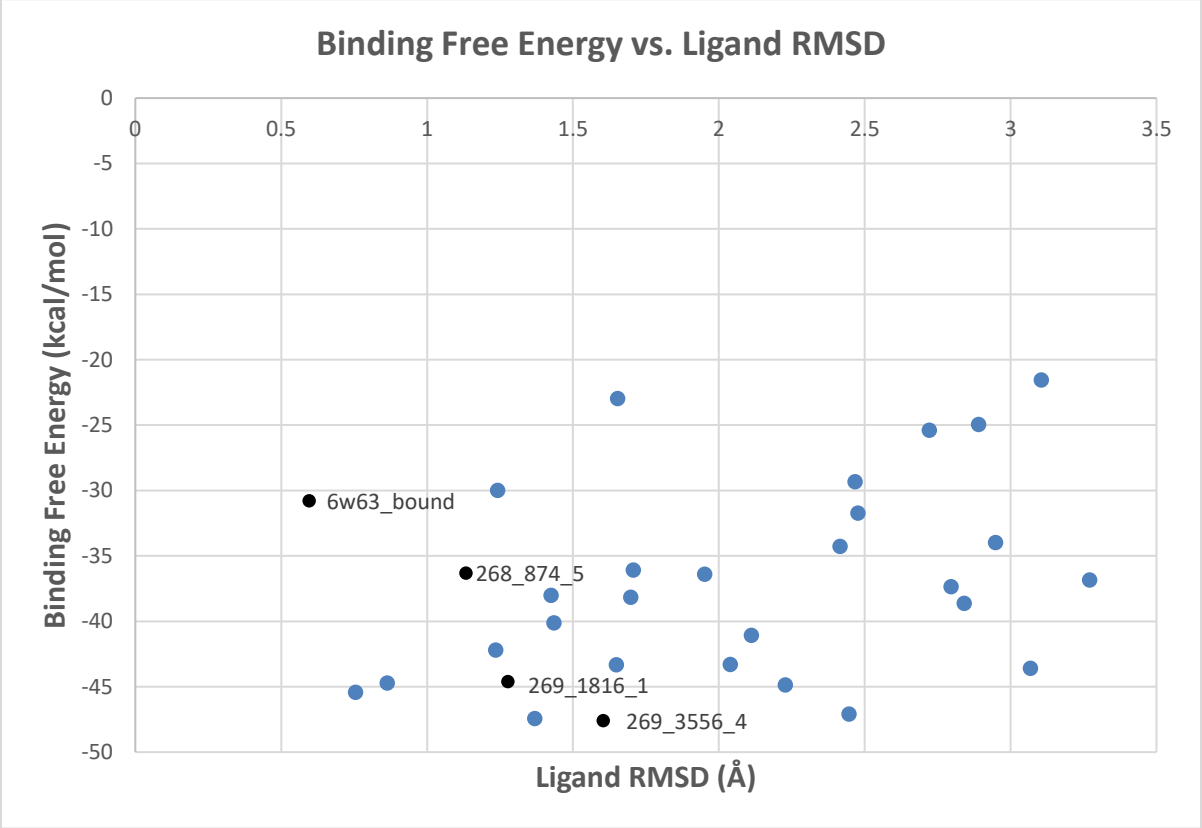


Figure 23. Average ligand RMSD per structure.

4.3.1.3 Energy Decomposition and Hydrogen Bonding

Among the 30 ‘hit-based’ results, MM-GBSA overall energetics were examined, along with their hydrogen bonding. Since the key contributing residues to binding within the Mpro catalytic site include residues GLU166, GLN189, HIS163, SER144, and GLY143 [52], an emphasis on these residues was made when examining these structures for hydrogen bonds and the corresponding decomposition plots. Structures with promising decomposition results were screened for the occurrence of hydrogen bonding, and distance between the donor and acceptor pairs was plotted using VMD.

First, the 6W63 crystal structure complex had stable hydrogen bond interactions with residues GLY143, GLU166, and HIS163 and a weaker hydrogen bonding interaction with

ASN 142. Within the energy decomposition graph, in **Figure 24A**, the ASN142 residue contributes mostly through electrostatic interactions, and MET165 contributions are through electrostatic and van der Waals forces. Notably, the ligand RMSD is very stable for this crystal structure bound ligand.

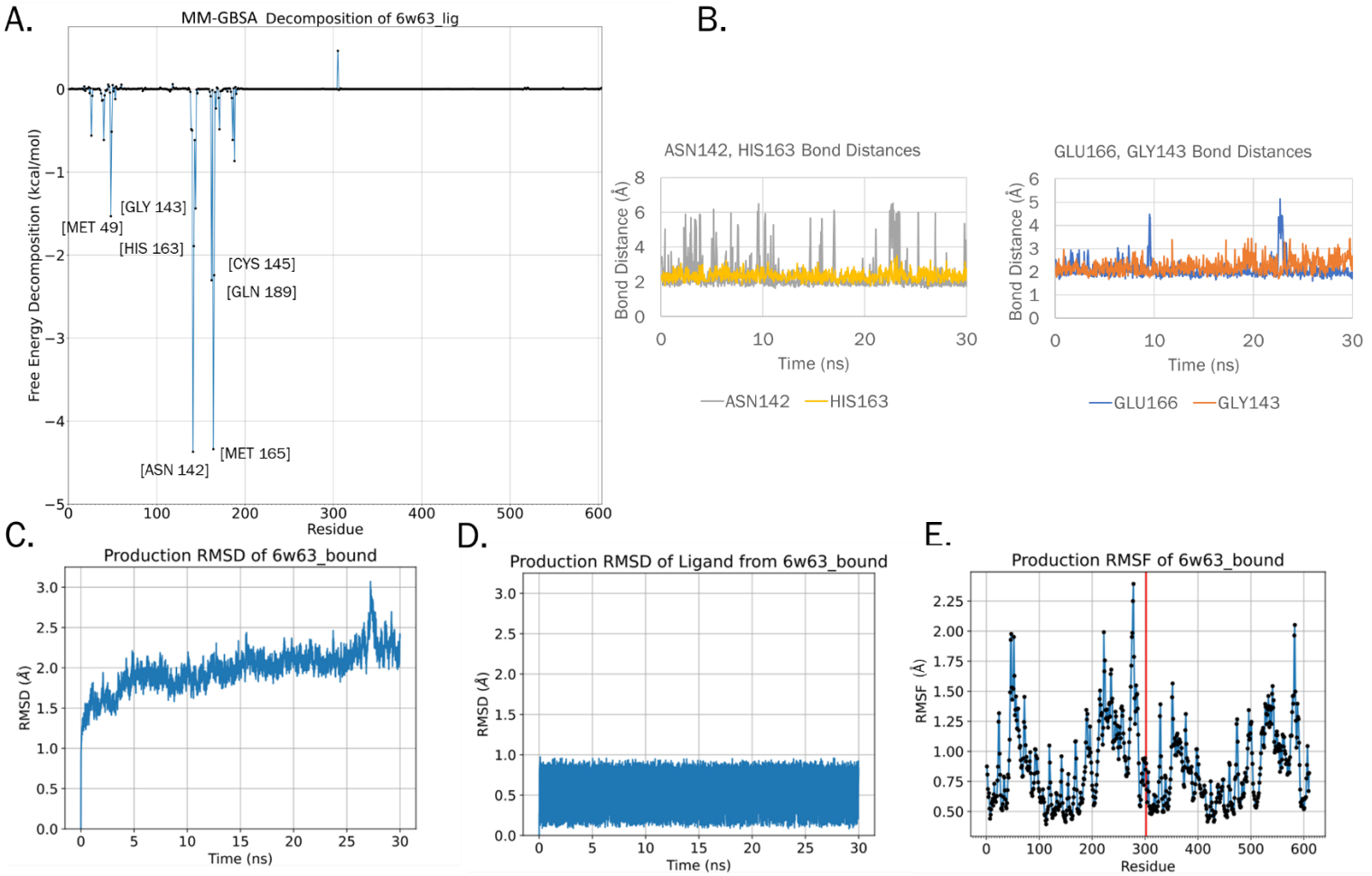


Figure 24. Combined results for 6W63 native ligand. A: Energy decomposition per residue. B: graphical depiction of hydrogen bonding interactions between 6W63 native molecule and ASN142, HIS163, GLU166, and GLY143. C: overall production RMSD. D: production RMSD of protein and ligand. E. RMSF plot for reference molecule.

To further visualize the placement and interactions occurring in the binding site, the structure of the 6W63 positive control and its hydrogen bonding is represented in **Figure 25** with a 2D interaction diagram, along with a corresponding 3D visualization of the residue interactions with this molecule**.** Within this diagram, the interactions between GLU166, HIS163, and GLY143 are highlighted.

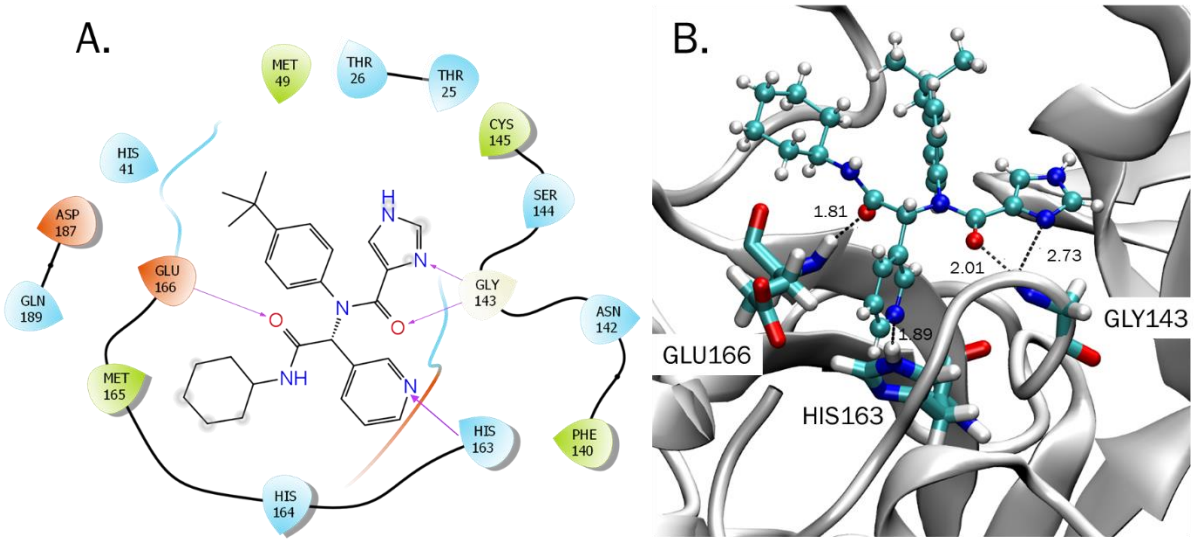


Figure 25. 6W63 crystal structure interactions depicted in 2D and 3D structures. A. 2D representation of hydrogen bonding, generated with Schrödinger Maestro. Strong hydrogen bonds, represented by the purple lines, are formed with GLU166, HIS163, and GLY143. B: 3D visualization of hydrogen bonds and interactions within the 6W63 binding pocket.

Based on MM-GBSA, protein and ligand fluctuations, and key hydrogen bond interactions, three molecules were selected as the top three out of the molecular dynamics simulations that were run. First, molecule 268_874_5 yielded a binding free energy of -36.3 kcal/mol. According to the breakdown of decomposition energies, as shown in **Figure 26A**, residues MET165 and GLN189 contribute through van der Waals forces, and residue THR26 contributes through a mixture of Van der Waals and electrostatic interactions. This molecule

exhibited hydrogen bonding with residues GLN192, THR26, and CYS145. To visualize these interactions, the hydrogen bonding distances were plotted and are shown in **Figure 26B**. Similar to the reference crystal structure, this molecule has a key interaction with GLN192 within the binding site, and MET165/GLN189 are energetic contributors common to the reference structure. However, the THR26 interaction does not occur in the reference crystal structure. In this case, the protein backbone remains relatively stable over the course of the simulation (**Figure 26C**), with an average value of 1.8 Å. Additionally, the ligand remains stable within the binding pocket, yielding an average value of 1.1 Å (**Figure 26D**), however, it is less stable than the reference structure.

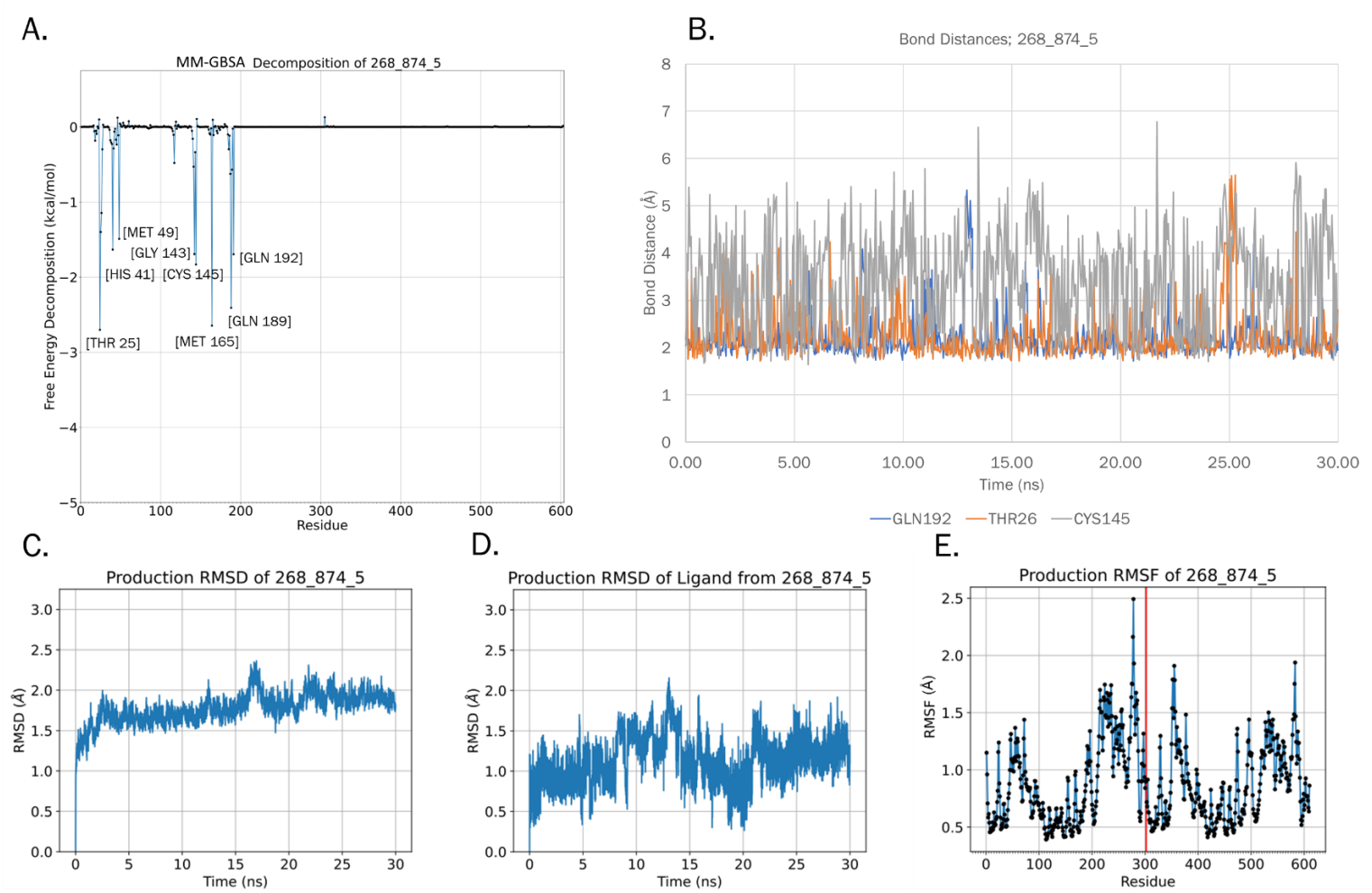


Figure 26. Combined results for molecule 268_874_5. A. Energy decomposition per residue. B. Graphical depiction of hydrogen bonding interactions between molecule 268_874_5 and GLY143 and GLN189. C. Overall production RMSD. D. Production RMSD of protein and ligand, E. RMSF plot for molecule 268_874_5.

The 268_874_5 molecule's hydrogen bonding interactions can be more closely examined in **Figure 27A**, where additional hydrogen bond lines were added to the 2D image to correspond to the frame capture of the 3D structure in **Figure 27B**.

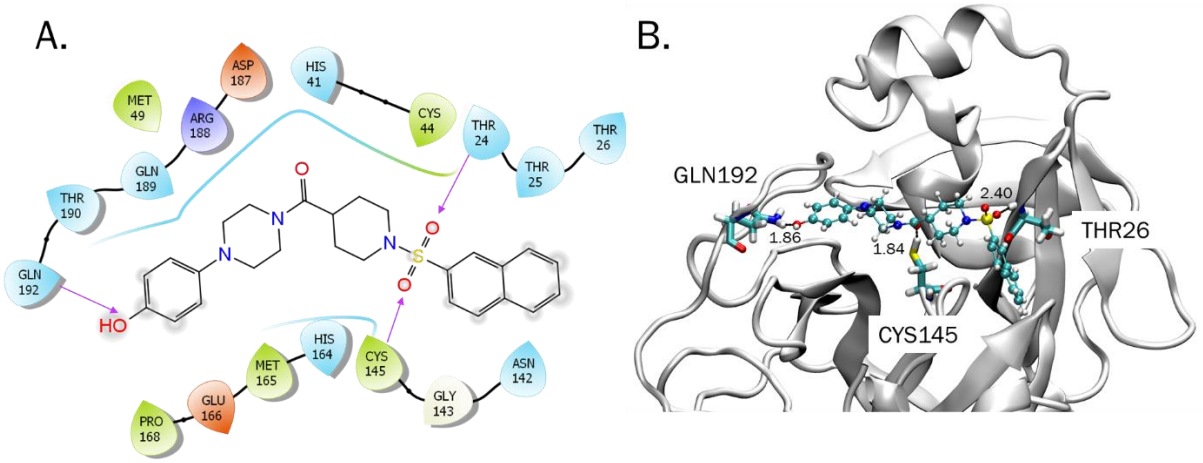


Figure 27. Hydrogen bond interactions for structure 268_874_5. A. 2D diagram generated using Schrödinger showing hydrogen bonds between molecule 268_874_5 and GLN192, CYS145, THR26. B. 3D depiction of hydrogen bonding occurring during the production simulation with VMD.

Next, the 269_1816_1 molecule resulted in a very favourable binding free energy value of -44.6 kcal/mol. The decomposition breakdown of this molecule within **Figure 28A** revealed that the MET165 energy contribution was primarily through van der Waals interactions, and GLN189 was through electrostatic interactions. Additionally, brief hydrogen bonding was observed between this molecule and residues GLN189, GLN192, and GLY143; distance plots (**Figure 28B**) show that these residues are within range for hydrogen bonding for some of the simulation, but there is movement within the ligand, and these hydrogen bonds are not sustained. As with the crystal structure, interactions consist of the potential for hydrogen bonding with residues GLN192 and GLY143. Additionally, MET165 and GLN189 are key

energetic contributors in the reference structure as well. Protein (**Figure 28C**) and ligand (**Figure 28D**) production RMSD graphs show good overall stability over the course of the simulations, with average fluctuations of 1.8 Å and 1.7 Å, respectively. Protein stability is comparable to the reference structure, and the ligand stability is greater than that of the reference.

Although this molecule shows some non-ideal characteristics, the highly favourable binding affinity for it is very promising. To further investigate, first steps would be to run a longer simulation, and then to investigate whether potential structural changes could be made to this molecule hit to further stabilize it, since, although transient, the potential hydrogen bond interactions are key interactions that have the potential to greatly stabilize the ligand within the binding pocket if sustained.

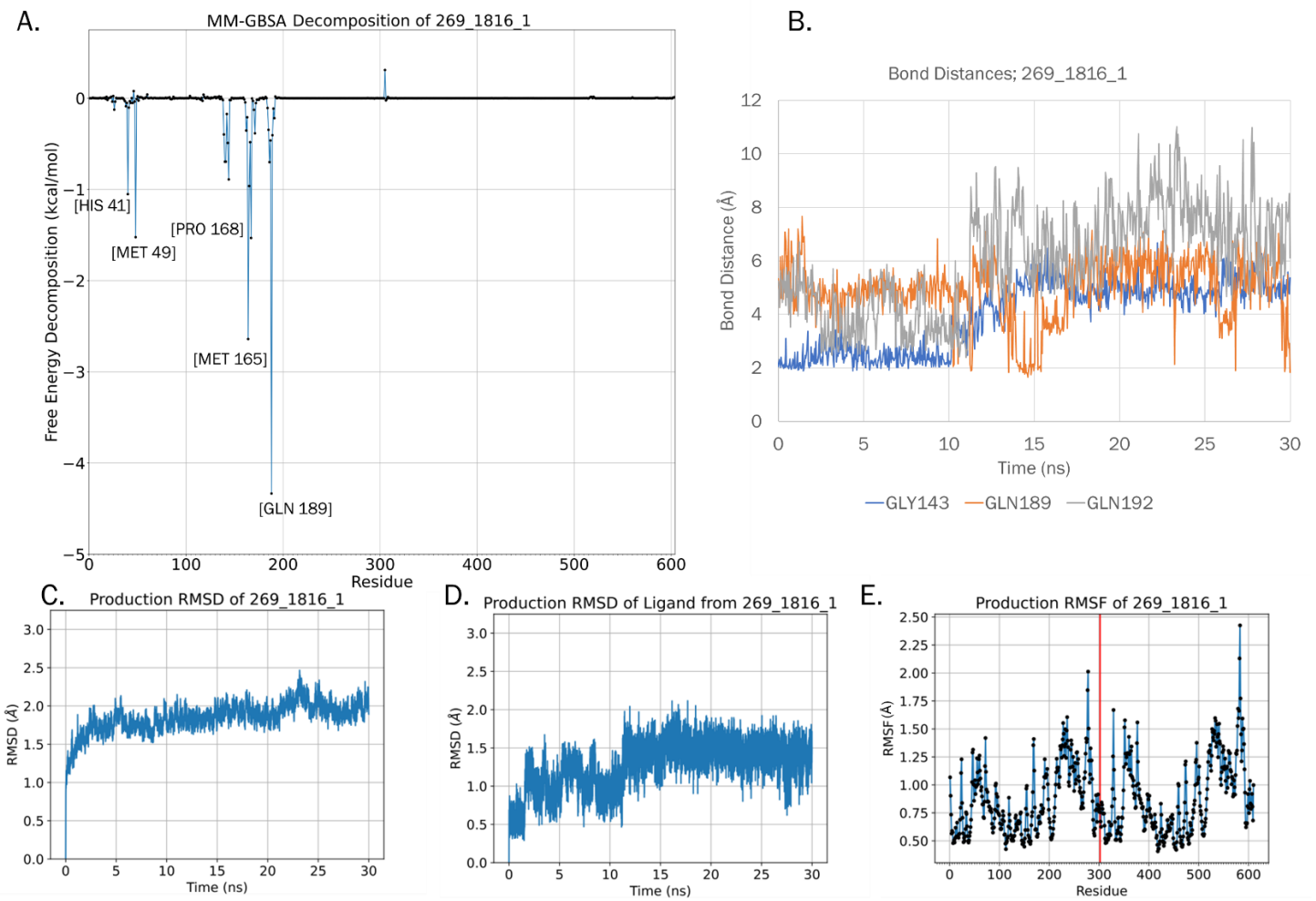


Figure 28. Combined results for molecule 269_1816_1. A. Energy decomposition per residue. B. Graphical depiction of hydrogen bonding interactions between molecule 269_1816_1 and CYS145 and GLN189. C. Overall production RMSD. D. Production RMSD of protein and ligand. E. RMSF plot of molecule 269_1816_1.

A snapshot of molecule 269_1816_1 undergoing transient hydrogen bonding interactions is shown in **Figure 29**. Bonding with residues 189, GLN192, and GLY143 can be seen in 2D and 3D diagrams.

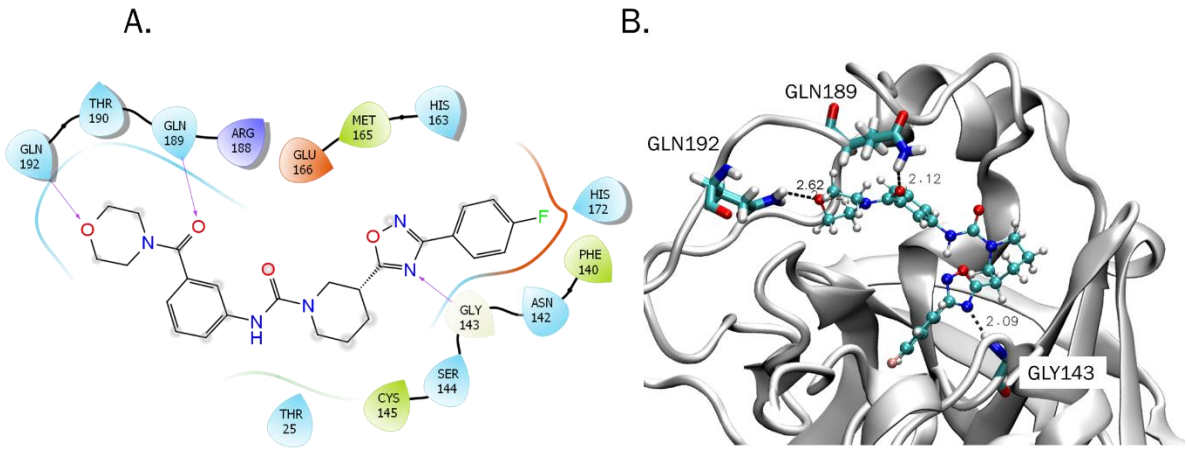


Figure 29. Hydrogen bond interactions for structure 269_1816_1. A. 2D diagram generated using Schrödinger showing hydrogen bonds between molecule 269_1816_1 and GLN192, GLN189, GLY143. B. 3D depiction of hydrogen bonding occurring during the production simulation with VMD.

Finally, molecule 269_3556_4 yielded an overall binding free energy of -47.6 kcal/mol. Of the individual residues that are contributing the most energetically (**Figure 30A**), CS145 contributes through van der Waals and electrostatic interactions, and MET165 and MET49 each contribute through van der Waals interactions. This molecule exhibited hydrogen bonding with residues GLY143 and HIS41 (**Figure 30B**). CYS145 showed some potential to create hydrogen bonds, but the distance and angle of the interaction was not as ideal for this type of bonding. For this molecule, overall protein backbone RMSD was 1.8 Å (**Figure 30C**) and ligand RMSD was 1.6 Å (**Figure 30D**). Common to the reference crystal structure are the CYS145, MET165, and GLY143 interactions.

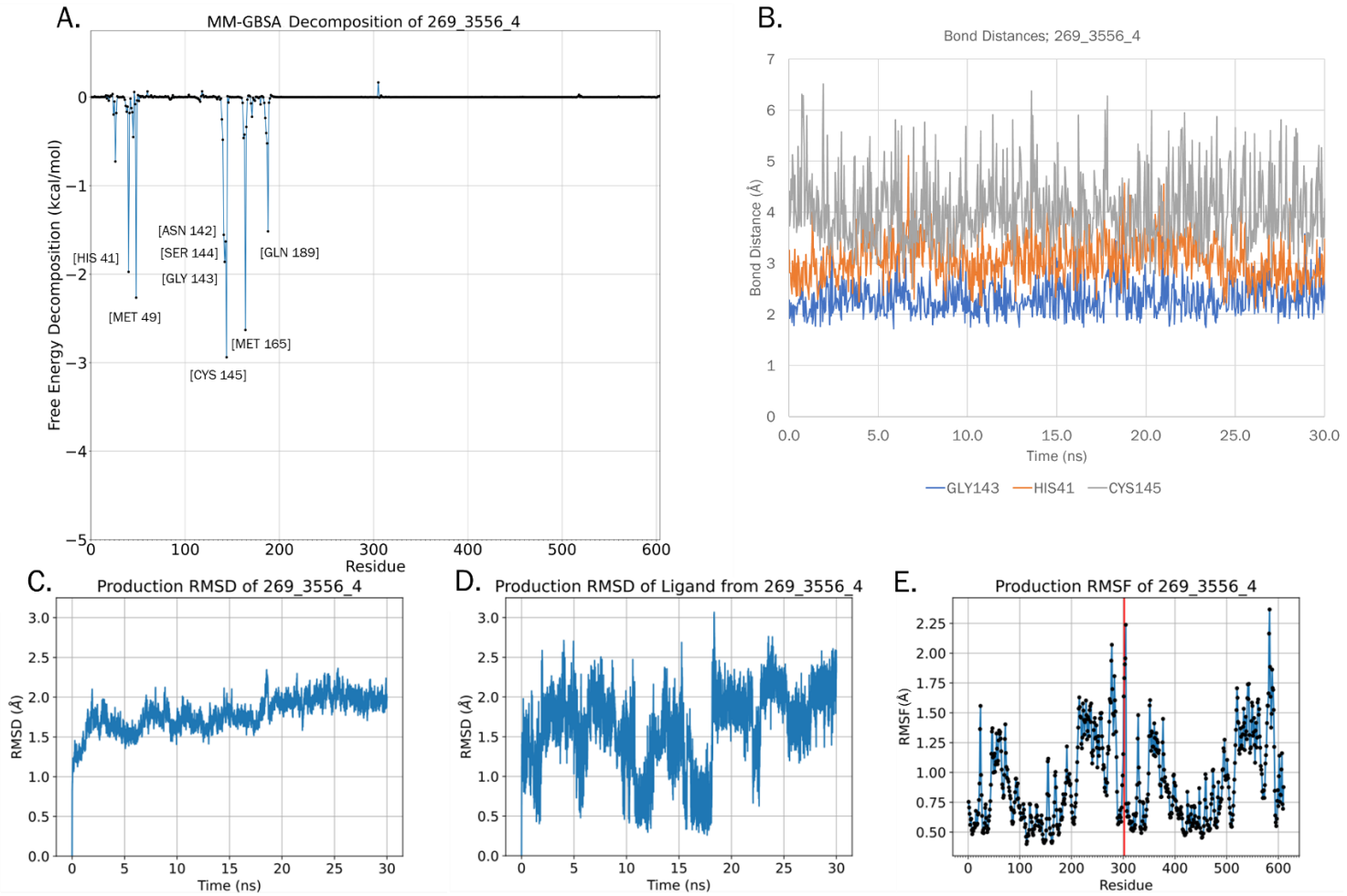


Figure 30. Combined results for molecule 269_3556_4 A. Energy decomposition per residue. B. Graphical depiction of hydrogen bonding interactions between molecule 269_3556_4 and GLY143, HIS41, and CYS145. C. Overall production RMSD. D. Production RMSD of protein and ligand. E. RMSF plot of molecule 269_3556_4.

A 2D interaction image of the hydrogen bonding between residues GLY143, HIS41, and molecule 269_3556_4 is shown in **Figure 31**, along with the corresponding 3D visualization of this hydrogen bonding.

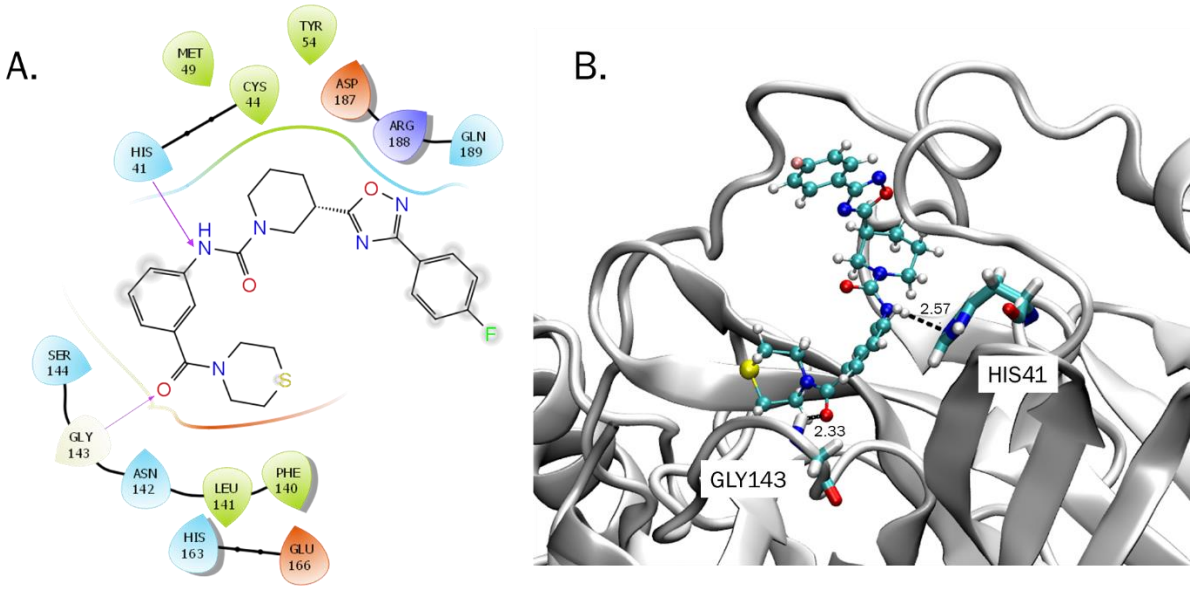


Figure 31. Hydrogen bond interactions for structure 269_3556_4. A. 2D diagram generated using Schrödinger showing hydrogen bonds between molecule 269_3556_4 and GLY143 and HIS41. B. 3D depiction of hydrogen bonding occurring during the production simulation with VMD.

4.3.1.4 Conclusions

Top structures from the machine learning model were selected, and then 0.1% of this dataset, or 4878 Chem-space molecules, were docked with AutoDock Vina. Structures that yielded the best binding affinity were examined further with molecular dynamics simulations. These structures were compared to the 6W63 reference structure. Among the three compounds

investigated, all exhibited some similar interactions with the reference molecule; notably molecule 268_874_5 had common interactions with MET165 and GLN189. This structure also exhibited three hydrogen bond interactions and yielded the lowest ligand RMSD out of the potential hits (1.1 Å).

The 269_1816 molecule was selected as another potential hit. Although this molecule did not exhibit strong hydrogen bonding throughout the simulation, it yielded a very favourable binding affinity (-44.6 kcal/mol), and residues MET165 and GLN189 were again energetic contributors common to the reference molecule. Although there is potential for strong hydrogen bonding interactions with residues GLY143, GLN189, and GLN192 to occur, this may need to be performed with a lead optimization step, where functional groups can be optimized to better stabilize the protein.

Finally, the 269_3556_4 molecule also obtained an excellent binding affinity value, of -47.6 kcal/mol, and this molecule shared key interactions of CYS145, MET165, and GLY143 with the reference molecule. This molecule formed two hydrogen bonds, with GLY143, and with HIS41, with reasonable protein and ligand RMSD: 1.8 and 1.6 Å, respectively. This molecule would also benefit from further investigation and lead optimization procedures, to determine if stability in the binding pocket that is closer to the reference molecule can be achieved.

Chapter 5
Summary and Outlook

Computer-aided drug design is a strategy that can add efficiency to the drug discovery process. These strategies are especially useful for drug development in the case of antivirals, as viral infections can quickly spread, as with the case of the Covid-19 pandemic. This thesis examined the creation and use of a high-throughput machine learning approach, where non-binding data for the machine learning training set was supplemented in a rational manner, through use of 2D molecular fingerprints, and a screening procedure was performed against an Mpro antiviral target.

Initially, 2D fingerprints and their combinations were evaluated using the DUD-E dataset, where individual and combination fingerprint selection approaches were compared against each other, as well as random selection. Fingerprint methods provided more accurate predictions when compared to random selection, and from the fingerprint types tested, the 'MACCS/RDKit/Topological', 'MACCS/RDKit/Topological/Morgan', and 'Morgan' fingerprints provided the most accurate ligand non-binder predictions. Next, this approach was applied to a small PDBbind dataset, where non-binder threshold values were tested (values of 0.2 to 0.6), and where a threshold value of 0.2 provided the highest results across data proportions, fingerprints, and machine learning models. This threshold was further tested with 10-fold cross validation and the full PDBbind dataset, across fingerprint types and proportions. In this case, the 'MACCS/RDKit/Topological' fingerprint combination yielded the highest results.

Models were then validated with two Mpro antiviral datasets, one extracted from PDB, and another from Chem-space with the random forest, support vector machine, and XGBoost machine learning algorithms and with the 'MACCS/RDKit/Topological' fingerprint selection method. Various proportions were compared and provided similar results. However, the highest performance resulted in a proportion = 5 for the random forest model, an XGBoost proportion = 2, and a support vector machine proportion = 5.

With a combination of these models, probability predictions were computed for the large Chem-space dataset, where the top 0.1% of predictions were processed through an AutoDock Vina (rigid protein/flexible receptor) virtual screening procedure. The top ten Vina results, with three unique poses from each, were then further validated through molecular dynamics procedures. Three notable hits were identified, with more favourable binding affinity compared to the Mpro/6W63 reference structure. These hits contained similar key residue interactions to the reference, although the strong hydrogen bonds with key residues were not as abundant in these examined potential hits. Although the molecules from our screening procedure show promise, further investigation will be required, through longer molecular dynamics simulations and lead optimization steps to further determine the stability of these molecules compared to the reference structure, and to further stabilize these molecules through additional interactions.

5.1 Future Work with Regression

Additionally, a regression-based model, adapted from the classification model's feature set, is in progress. The goal of this model is to build upon the classification approach, and to add features that encode protein-ligand binding information to predict binding affinity, and also to determine the most ideal pose. In order to add the protein-ligand features, the training set molecules must be docked. This regression strategy currently utilizes the PDBbind dataset for training purposes, with five docked poses, generated from AutoDock Vina. The model contains identical feature information to the classification model (protein, ligand, and binding site), but additional computed protein-ligand interaction information based on the docked pose. To provide a varying 'binding affinity' parameter, the $-\log K_d/K_i$ value from this dataset was adjusted in two ways. Half of this property was adjusted based on a binned RMS value, when compared to the crystal structure pose within the PDBbind dataset. The other half was adjusted according to an average fingerprint value between the 'SPLIF' and 'Simple Interaction' fingerprints, using the crystal structure pose as the reference.

Limitations with this method are that each structure must be docked, since this is used as a parameter and to compute protein-ligand information. Although docking all complexes from a training set may be achievable, if a large-scale dataset is desired for a screening protocol, this would be much more resource-intensive, and reduce the advantages of time/computational cost savings that are typical with machine learning methods. However, perhaps using ‘deep dock’ to determine poses may be an approach to build upon, so that additional protein-ligand complex features may be added to a model to increase performance. A description of the feature breakdown and a preliminary visualization of the 10-fold cross validation results is shown in **Appendix C**. Future work with this model will involve the addition and optimization of features, as well as testing with external datasets, where accuracy of binding affinity per ligand binder is assessed, as well as model capabilities in determining a pose closest to the ideal for a ligand.

References

[1] Y. J. Park *et al.*, "Fighting the War Against COVID-19 via Cell-Based Regenerative Medicine: Lessons Learned from 1918 Spanish Flu and Other Previous Pandemics," in *Stem Cell Reviews and Reports*, ed, 2020.

[2] O. J. Wouters, M. McKee, and J. Luyten, "Estimated Research and Development Investment Needed to Bring a New Medicine to Market, 2009-2018," *JAMA*, vol. 323, no. 9, pp. 844-853, 2020, doi: 10.1001/jama.2020.1166.

[3] J. M. Reichert, "Trends in development and approval times for new therapeutics in the United States," (in eng), *Nat Rev Drug Discov*, vol. 2, no. 9, pp. 695-702, Sep 2003, doi: 10.1038/nrd1178.

[4] T. T. Ashburn and K. B. Thor, "Drug repositioning: identifying and developing new uses for existing drugs," *Nature Reviews Drug Discovery*, vol. 3, no. 8, pp. 673-683, 2004/08/01 2004, doi: 10.1038/nrd1468.

[5] V. T. Sabe *et al.*, "Current trends in computer aided drug design and a highlight of drugs discovered via computational techniques: A review," *European Journal of Medicinal Chemistry*, vol. 224, p. 113705, 2021/11/15/ 2021, doi: https://doi.org/10.1016/j.ejmech.2021.113705.

[6] D. E. Clark, "What has virtual screening ever done for drug discovery?," (in eng), *Expert Opin Drug Discov*, vol. 3, no. 8, pp. 841-51, Aug 2008, doi: 10.1517/17460441.3.8.841.

[7] A. T. Ton, F. Gentile, M. Hsing, F. Ban, and A. Cherkasov, "Rapid Identification of Potential Inhibitors of SARS-CoV-2 Main Protease by Deep Docking of 1.3 Billion Compounds," in *Molecular Informatics* vol. 39, ed, 2020, pp. 1-7.

[8] O. Kadioglu, M. Saeed, H. J. Greten, and T. Efferth, "Identification of novel compounds against three targets of SARS CoV-2 coronavirus by combined virtual screening and supervised machine learning," (in eng), *Comput Biol Med*, vol. 133, p. 104359, Jun 2021, doi: 10.1016/j.compbiomed.2021.104359.

[9] B. Geoffrey, A. Sanker, R. Madaj, M. S. V. Tresanco, M. Upadhyay, and J. Gracia, "A program to automate the discovery of drugs for West Nile and Dengue virus-programmatic screening of over a billion compounds on PubChem, generation of drug leads and automated in silico modelling," (in eng), *J Biomol Struct Dyn*, vol. 40, no. 10, pp. 4293-4300, Jul 2022, doi: 10.1080/07391102.2020.1856185.

[10] V. Parvathaneni and V. Gupta, "Utilizing drug repurposing against COVID-19 – Efficacy, limitations, and challenges," *Life Sciences*, vol. 259, p. 118275, 2020/10/15/ 2020, doi: https://doi.org/10.1016/j.lfs.2020.118275.

[11] S. Dotolo, A. Marabotti, A. Facchiano, and R. Tagliaferri, "A review on drug repurposing applicable to COVID-19," *Briefings in bioinformatics*, vol. 22, no. 2, pp. 726-741, 2021.

[12] H. M. Berman *et al.*, "The Protein Data Bank," *Nucleic acids research*, vol. 28, no. 1, pp. 235-242, 2000, doi: gkd090 [pii].

[13] M. Su *et al.*, "Comparative Assessment of Scoring Functions: The CASF-2016 Update," in *Journal of Chemical Information and Modeling* vol. 59, ed: American Chemical Society, 2019, pp. 895-913.
[14] M. M. Mysinger, M. Carchia, J. J. Irwin, and B. K. Shoichet, "Directory of Useful Decoys, Enhanced (DUD-E): Better Ligands and Decoys for Better Benchmarking," in *J. Med. Chem* vol. 55, ed, 2012, p. 6594.
[15] D. S. Wishart *et al.*, "DrugBank 5.0: a major update to the DrugBank database for 2018," (in eng), *Nucleic Acids Res,* vol. 46, no. D1, pp. D1074-d1082, Jan 4 2018, doi: 10.1093/nar/gkx1037.
[16] *Chem-space*, Chem-space,
[17] I. Muegge and P. Mukherjee, "An overview of molecular fingerprint similarity search in virtual screening," (in eng), *Expert Opin Drug Discov,* vol. 11, no. 2, pp. 137-48, 2016, doi: 10.1517/17460441.2016.1117070.
[18] J. Duan, S. L. Dixon, J. F. Lowrie, and W. Sherman, "Analysis and comparison of 2D fingerprints: insights into database screening performance using eight fingerprint methods," (in eng), *J Mol Graph Model,* vol. 29, no. 2, pp. 157-70, Sep 2010, doi: 10.1016/j.jmgm.2010.05.008.
[19] J. C. Williams, S. Opare, S. K. Sugadoss, A. Ganesan, and S. Kalyaanamoorthy, "Virtual screening techniques in pharmaceutical research," in *Contemporary Chemical Approaches for Green and Sustainable Drugs*: Elsevier, 2022, pp. 89-128.
[20] M. Vogt, D. Stumpfe, H. Geppert, and J. Bajorath, "Scaffold Hopping Using Two-Dimensional Fingerprints: True Potential, Black Magic, or a Hopeless Endeavor? Guidelines for Virtual Screening," *Journal of Medicinal Chemistry*, vol. 53, no. 15, pp. 5707-5715, 2010/08/12 2010, doi: 10.1021/jm100492z.
[21] A. Cereto-Massagué, M. J. Ojeda, C. Valls, M. Mulero, S. Garcia-Vallvé, and G. Pujadas, "Molecular fingerprint similarity search in virtual screening," *Methods,* vol. 71, pp. 58-63, 2015/01/01/ 2015, doi: https://doi.org/10.1016/j.ymeth.2014.08.005.
[22] T. Micheal, *Machine learning*. 2017, pp. 40-48.
[23] X. Yang, Y. Wang, R. Byrne, G. Schneider, and S. Yang, "Concepts of Artificial Intelligence for Computer-Assisted Drug Discovery," in *Chemical Reviews* vol. 119, ed, 2019, pp. 10520-10594.
[24] J. Hurwitz and D. Kirsch, *Machine Learning for Dummies, IBM Limited Edition*. Noboken, NJ: John Wiley & Sons, Inc., 2018.
[25] L. Breiman, "Random forests," *Machine learning,* vol. 45, no. 1, pp. 5-32, 2001.
[26] Y. Zhang *et al.*, "A combined drug discovery strategy based on machine learning and molecular docking," *Chemical Biology & Drug Design,* vol. 93, no. 5, pp. 685-699, 2019.
[27] S. Ramraj, N. Uzir, R. Sunil, and S. Banerjee, "Experimenting XGBoost algorithm for prediction and classification of different datasets," *International Journal of Control Theory and Applications,* vol. 9, no. 40, 2016.
[28] I. Baturynska and K. Martinsen, "Prediction of geometry deviations in additive manufactured parts: comparison of linear regression with machine learning

algorithms," *Journal of Intelligent Manufacturing*, vol. 32, no. 1, pp. 179-200, 2021/01/01 2021, doi: 10.1007/s10845-020-01567-0.

[29] R. N. Jorissen and M. K. Gilson, "Virtual screening of molecular databases using a support vector machine," *Journal of chemical information and modeling*, vol. 45, no. 3, pp. 549-561, 2005.

[30] J. Brownlee, *Imbalanced Classification with Python: Better Metrics, Balance Skewed Classes, Cost-Sensitive Learning*. Machine Learning Mastery, 2020.

[31] D. M. Powers, "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation," *arXiv preprint arXiv:2010.16061*, 2020.

[32] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321-357, 2002.

[33] S. M. H. Mahmud, W. Chen, H. Jahan, Y. Liu, N. I. Sujan, and S. Ahmed, "iDTi-CSsmoteB: Identification of Drug–Target Interaction Based on Drug Chemical Structure and Protein Sequence Using XGBoost With Over-Sampling Technique SMOTE," *IEEE Access*, vol. 7, pp. 48699-48714, 2019, doi: 10.1109/ACCESS.2019.2910277.

[34] A. Fernández, S. García, M. Galar, R. C. Prati, B. Krawczyk, and F. Herrera, *Learning from imbalanced data sets*. Springer, 2018.

[35] I. A. Guedes, C. S. de Magalhães, and L. E. Dardenne, "Receptor-ligand molecular docking," (in eng), *Biophysical reviews*, vol. 6, no. 1, pp. 75-87, 2014, doi: 10.1007/s12551-013-0130-2.

[36] O. Trott and A. J. Olson, "AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading," (in eng), *Journal of computational chemistry*, vol. 31, no. 2, pp. 455-461, 2010, doi: 10.1002/jcc.21334.

[37] J. D. Durrant and J. A. McCammon, "Molecular dynamics simulations and drug discovery," *BMC biology*, vol. 9, pp. 71-7007-9-71, 2011, doi: 10.1186/1741-7007-9-71 [doi].

[38] J. Wang, W. Wang, P. A. Kollman, and D. A. Case, "Automatic atom type and bond type perception in molecular mechanical calculations," (in eng), *J Mol Graph Model*, vol. 25, no. 2, pp. 247-60, Oct 2006, doi: 10.1016/j.jmgm.2005.12.005.

[39] W. D. Cornell *et al.*, "A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules," *Journal of the American Chemical Society*, vol. 117, no. 19, pp. 5179-5197, 1995, doi: 10.1021/ja00124a002.

[40] R. L. Davidchack, R. Handel, and M. Tretyakov, "Langevin thermostat for rigid body dynamics," *The Journal of chemical physics*, vol. 130, no. 23, p. 234101, 2009.

[41] H. J. Berendsen, J. v. Postma, W. F. Van Gunsteren, A. DiNola, and J. R. Haak, "Molecular dynamics with coupling to an external bath," *The Journal of chemical physics*, vol. 81, no. 8, pp. 3684-3690, 1984.

[42] C. Margreitter and C. Oostenbrink, "MDplot: Visualise Molecular Dynamics," (in eng), *R j*, vol. 9, no. 1, pp. 164-186, May 10 2017.

[43] E. Wang *et al.*, "End-Point Binding Free Energy Calculation with MM/PBSA and MM/GBSA: Strategies and Applications in Drug Design," *Chem Rev*, vol. 119, no. 16, pp. 9478-9508, Aug 28 2019, doi: 10.1021/acs.chemrev.9b00055.
[44] P. W. Rose *et al.*, "The RCSB protein data bank: integrative view of protein, gene and 3D structural information," *Nucleic Acids Research*, vol. 45, no. D1, pp. D271-D281, 2017, doi: 10.1093/nar/gkw1000.
[45] Z. Liu *et al.*, "PDB-wide collection of binding data: current status of the PDBbind database," *Bioinformatics*, vol. 31, no. 3, pp. 405-412, 2015, doi: 10.1093/bioinformatics/btu626.
[46] *RDKit : Open-source cheminformatics; http://www.rdkit.org*.
[47] L. Buitinck *et al.*, "API design for machine learning software: experiences from the scikit-learn project," *arXiv preprint arXiv:1309.0238*, 2013.
[48] M. Wójcikowski, P. J. Ballester, and P. Siedlecki, "Performance of machine-learning scoring functions in structure-based virtual screening," in *Scientific Reports* vol. 7, ed: Nature Publishing Group, 2017, pp. 1-10.
[49] J. Dong *et al.*, "PyBioMed: a python library for various molecular representations of chemicals, proteins and DNAs and their interactions," *Journal of Cheminformatics*, vol. 10, no. 1, p. 16, 2018/03/20 2018, doi: 10.1186/s13321-018-0270-2.
[50] P. Śledź and A. Caflisch, "Protein structure-based drug design: from docking to molecular dynamics," (in eng), *Curr Opin Struct Biol*, vol. 48, pp. 93-102, Feb 2018, doi: 10.1016/j.sbi.2017.10.010.
[51] Q. Hu *et al.*, "The SARS-CoV-2 main protease (Mpro): Structure, function, and emerging therapies for COVID-19," *MedComm*, vol. 3, no. 3, p. e151, 2022.
[52] Y. L. Weng, S. R. Naik, N. Dingelstad, M. R. Lugo, S. Kalyaanamoorthy, and A. Ganesan, "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," *Scientific Reports*, vol. 11, no. 1, p. 7429, 2021/04/01 2021, doi: 10.1038/s41598-021-86471-0.
[53] R. Yoshino, N. Yasuo, and M. Sekijima, "Identification of key interactions between SARS-CoV-2 main protease and inhibitor drug candidates," *Scientific Reports*, vol. 10, no. 1, p. 12493, 2020/07/27 2020, doi: 10.1038/s41598-020-69337-9.
[54] W. Humphrey, A. Dalke, and K. Schulten, "VMD: visual molecular dynamics," *Journal of molecular graphics*, vol. 14, no. 1, pp. 33-38, 1996.
[55] J. C. Williams and S. Kalyaanamoorthy, "PoseFilter: a PyMOL plugin for filtering and analyzing small molecule docking in symmetric binding sites," *Bioinformatics*, 2021, doi: 10.1093/bioinformatics/btab188.
[56] M. H. M. Olsson, C. R. Søndergaard, M. Rostkowski, and J. H. Jensen, "PROPKA3: Consistent Treatment of Internal and Surface Residues in Empirical pKa Predictions," *Journal of Chemical Theory and Computation*, vol. 7, no. 2, pp. 525-537, 2011/02/08 2011, doi: 10.1021/ct100578z.
[57] M. Wójcikowski, P. Zielenkiewicz, and P. Siedlecki, "Open Drug Discovery Toolkit (ODDT): a new open-source player in the drug discovery field," *Journal of cheminformatics*, vol. 7, no. 1, pp. 1-6, 2015.

Appendix A

Molecular Docking Pose Results

Screening with Random Forest proportion = 5, XGBoost proportion = 2, Support Vector Machine proportion = 5

264_4209; unique poses => 1, 6, 7

Type	**Input Parameter**	**Unique**
SI	0.45	1, 3, 6, 9
SPLIF	0.45	1, 6, 7
RMS	3.5 Å	1, 6, 7

267_3318; unique poses => 1, 2, 9

Type	Input Parameter	Unique
SI	0.3	1, 4, 9
SPLIF	0.3	1, 2, 4, 7, 8, 9
RMS	6.5 Å	1, 2, 9

268_874l; unique poses => 1, 3, 5

Type	Input Parameter	Unique
SI	0.3	1, 3, 8
SPLIF	0.3	1, 3, 5, 7, 8
RMS	6.5 Å	1, 5, 7

268_2452; unique poses => 1, 3, 5

Type	Input Parameter	Unique
SI	0.3	1, 3, 5
SPLIF	0.3	1, 3, 5, 8
RMS	6.5 Å	1, 3, 7

269_1725; unique poses => 1, 2, 7

Type	Input Parameter	Unique
SI	0.4	1, 2, 7
SPLIF	0.3	1, 2, 6, 7
RMS	6.5 Å	1, 2, 7

269_1816; unique poses => 1, 4, 5

Type	Input Parameter	Unique
SI	0.3	1, 4, 5, 8
SPLIF	0.3	1, 4, 5
RMS	7.0 Å	1, 4, 5, 8

269_3556; unique poses => 1, 2, 4

Type	Input Parameter	Unique
SI	0.35	1, 3, 4
SPLIF	0.3	1, 2, 4, 5, 7

RMS	6.5 Å	1, 2, 5

373_2580; unique poses => 1, 5, 6

Type	Input Parameter	Unique
SI	0.4	1, 3, 4
SPLIF	0.3	1, 5, 6
RMS	6.0 Å	1, 5, 6

413_1761; unique poses => 1, 3, 4

Type	Input Parameter	Unique
SI	0.3	1, 3
SPLIF	0.3	1, 3, 4, 6, 9
RMS	6.5 Å	1, 4, 6

574_2168; unique poses => 1, 4, 6

Type	Input Parameter	Unique
SI	0.3	1, 2, 4, 6
SPLIF	0.3	1, 4, 6, 7
RMS	7.5 Å	1, 6, 7

Appendix B

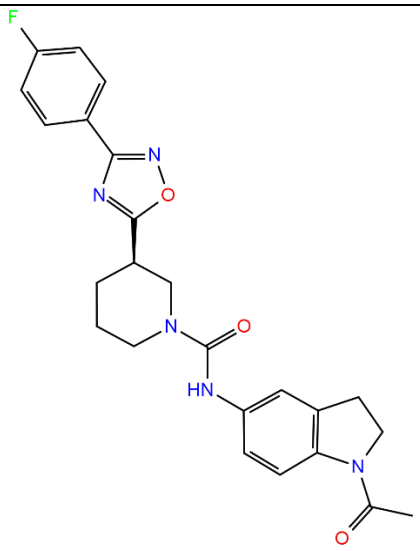
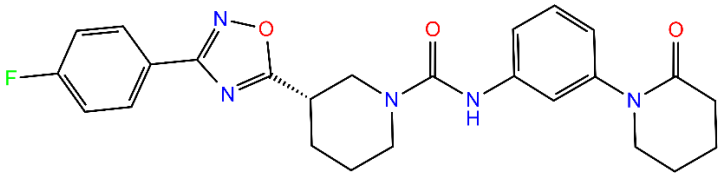
Docking Results

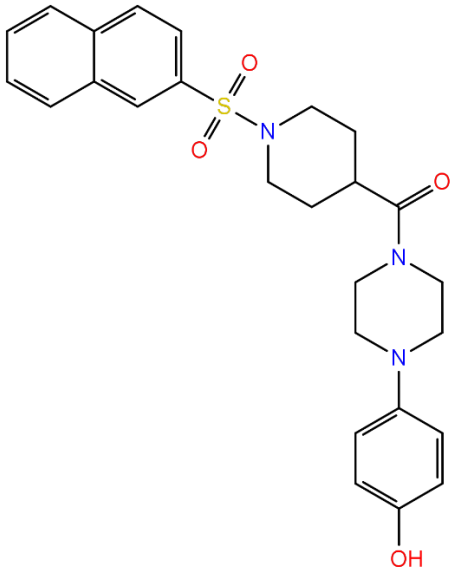
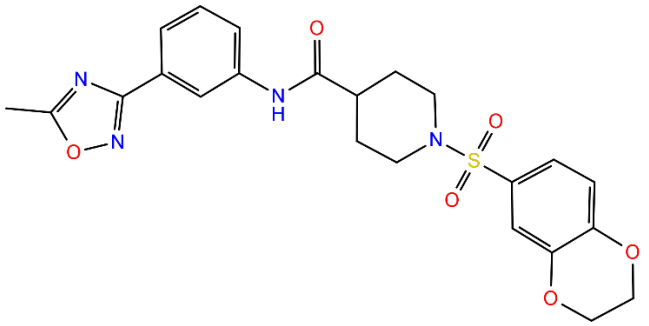
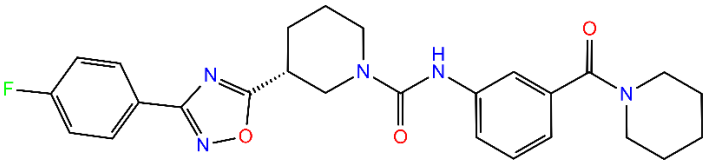
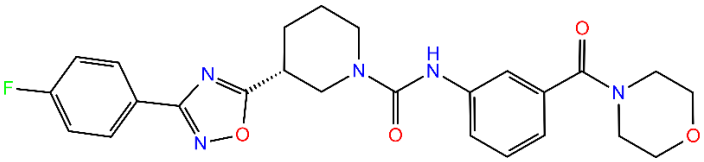
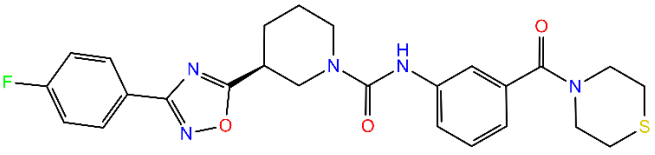
Table 7. Smiles, Chem-space ID, and probability per top hits.

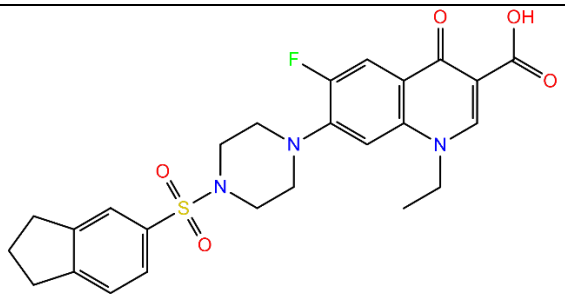
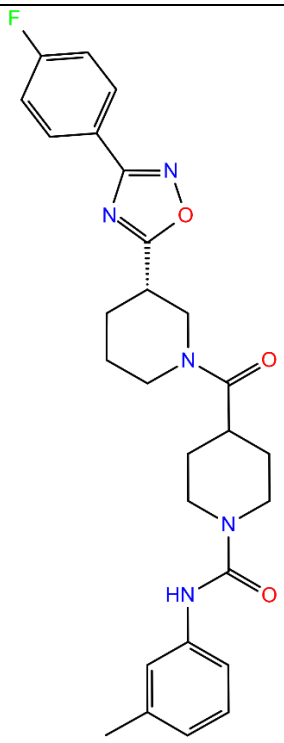
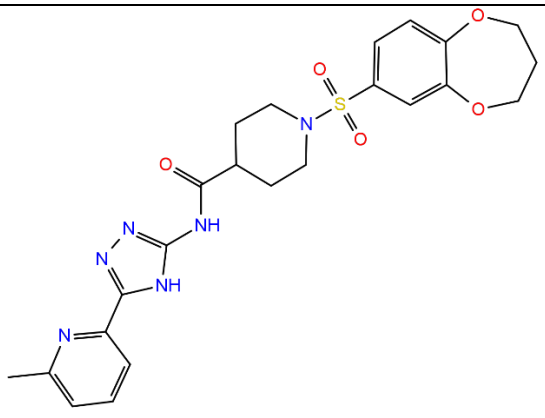
Molecule	SMILES	Chem-space ID	Probability
264_4209	CC(=O)N1CCC2=CC(NC(=O)N3CCCC(C4=NC(C5=CC=C(F)C=C5)=NO4)C3)=CC=C21	CSCS00663872595	0.9602
267_3318	O=C(NC1=CC=CC(N2CCCCC2=O)=C1)N1CCCC(C2=NC(C3=CC=C(F)C=C3)=NO2)C1	CSCS00671901626	0.9675
268_874	O=C(C1CCN(S(=O)(=O)C2=CC=C3C=CC=CC3=C2)CC1)N1CCN(C2=CC=C(O)C=C2)CC1	CSCS00674329345	0.9670
268_2452	CC1=NC(C2=CC=CC(NC(=O)C3CCN(S(=O)(=O)C4=CC=C5OCCOC5=C4)CC3)=C2)=NO1	CSCS00674867711	0.9660
269_1725	O=C(NC1=CC=CC(C(=O)N2CCCCC2)=C1)N1CCCC(C2=NC(C3=CC=C(F)C=C3)=NO2)C1	CSCS00676663707	0.9680
269_1816	O=C(NC1=CC=CC(C(=O)N2CCOCC2)=C1)N1CCCC(C2=NC(C3=CC=C(F)C=C3)=NO2)C1	CSCS00676801443	0.9693
269_3556	O=C(NC1=CC=CC(C(=O)N2CCSCC2)=C1)N1CCCC(C2=NC(C3=CC=C(F)C=C3)=NO2)C1	CSCS00678658197	0.9645
373_2580	CCN1C=C(C(=O)O)C(=O)C2=CC(F)=C(N3CCN(S(=O)(=O)C4=CC=C5CCCC5=C4)CC3)C=C21	CSCS01392107013	0.9628
413_1761	CC1=CC=CC(NC(=O)N2CCC(C(=O)N3CCCC(C4=NC(C5=CC=C(F)C=C5)=NO4)C3)CC2)=C1	CSCS01506988776	0.9733

574_2168	CC1=CC=CC(C2=NN=C(NC(=O)C3CCN(S(=O)(=O)C4=CC=C5OCCCOC5=C4)CC3)[NH]2)=N1	CSCS0228_1696192	0.9733

Table 8. Molecules and their corresponding 2D structures.

Molecule	2D Structure
264_4209	
267_3318	

268_874	
268_2452	
269_1725	
269_1816	
269_3556	

373_2580	
413_1761	
574_2168	

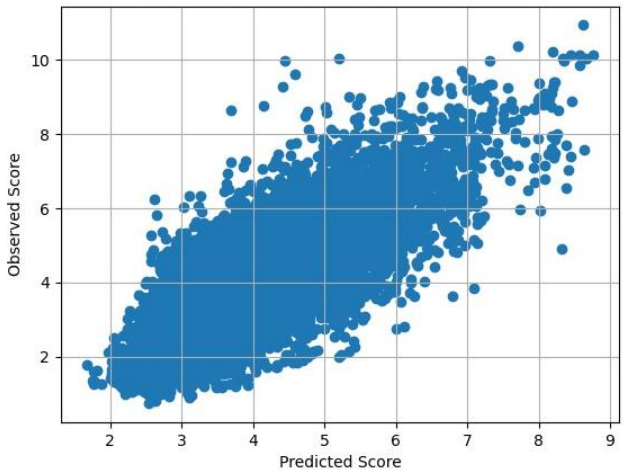
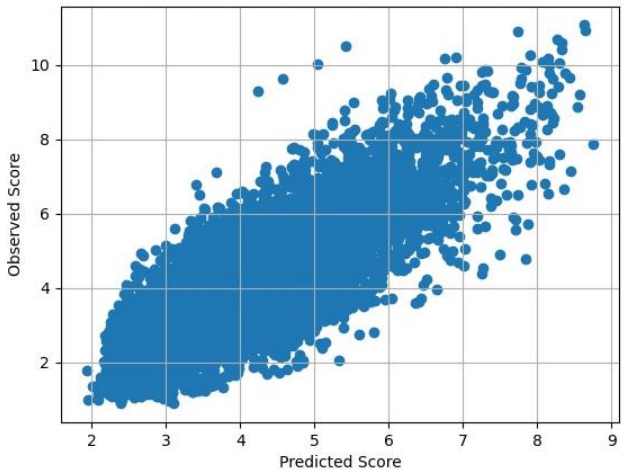
Appendix C

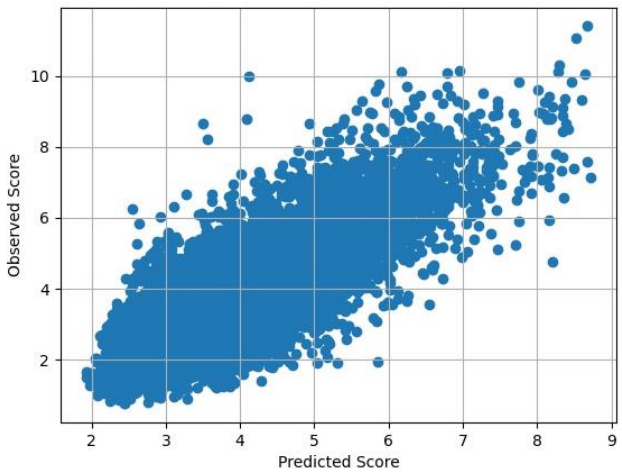
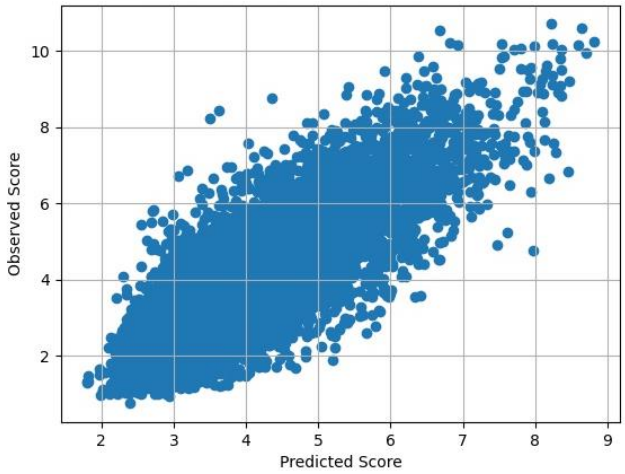
Regression Model

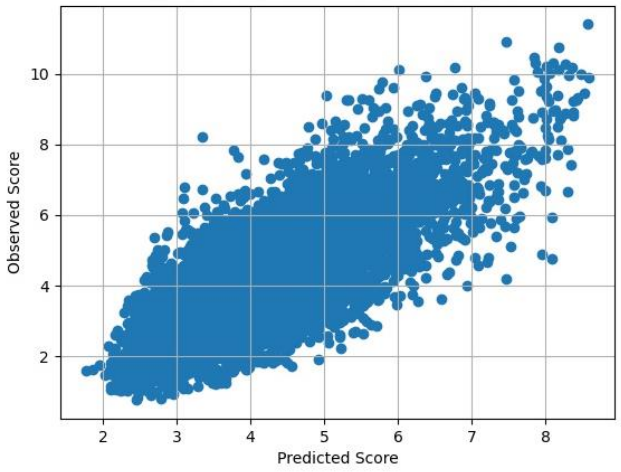
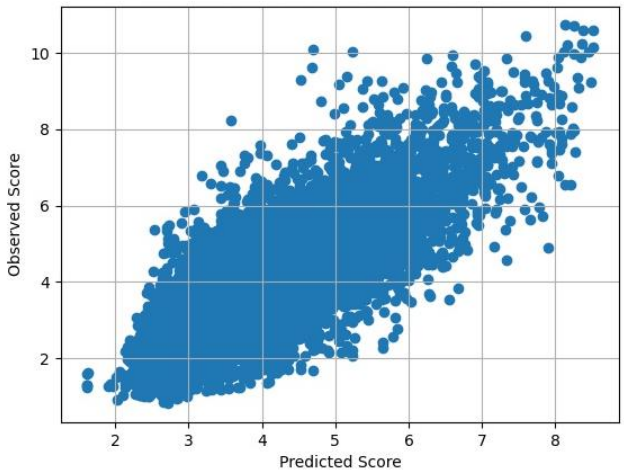
Table 9. Features used in regression model.

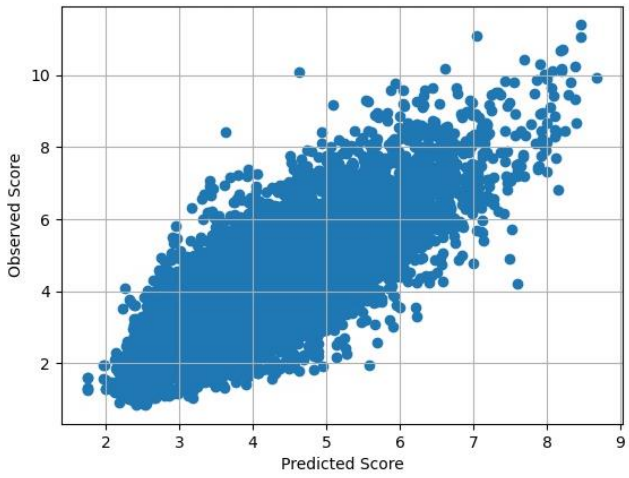
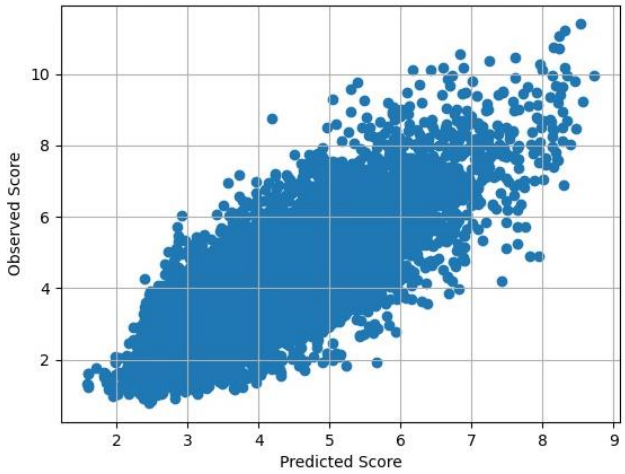
Feature Type	Source
Ligand/binding site	PyBioMed [49] /RDKit [46]
Sequence	PyBioMed [49]
Docking: docking score, mode, RMS	AutoDock Vina [36]
Atom proximity binning: 0-12 (6 bins, containing 2 Å each)	Open Drug Discovery Toolkit (ODDT) [57]
Protein-ligand interaction fingerprint: SInteraction (0-1) SPLIF (0-1)	Open Drug Discovery Toolkit (ODDT) [57]
Modified y-value predictor	$Prediction\ value = \frac{\log(binding\ affinity) \times (1 - entry)}{\max RMSD}$ 0-1 scale, highest for lowest RMSD but still scaled according to the total max RMSD

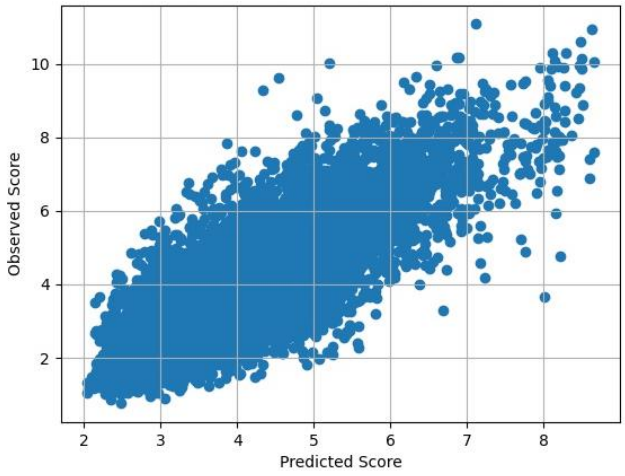
Table 10. Visualized 10-fold Cross Validation Regression Results.

Fold	R^2	10-Fold Cross Validation Results
1	0.731	
2	0.724	

3	0.723	 Observed Score Predicted Score 2 4 6 8 10 2 3 4 5 6 7 8 9
4	0.726	 Observed Score Predicted Score 2 4 6 8 10 2 3 4 5 6 7 8 9

5	0.725	 Observed Score Predicted Score 2 4 6 8 10 2 3 4 5 6 7 8
6	0.731	 Observed Score Predicted Score 2 4 6 8 10 2 3 4 5 6 7 8

7	0.726	 Observed Score Predicted Score 2 4 6 8 10 2 3 4 5 6 7 8 9
8	0.723	 Observed Score Predicted Score 2 4 6 8 10 2 3 4 5 6 7 8 9

9	0.727	
10	0.725