UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE INFORMÁTICA

In silico exploration of protein structural units for the discovery of new therapeutic targets

Pedro Frederico Matos Regalado de Oliveira

Mestrado em Engenharia Informática

Trabalho de Projeto orientado por:
Prof. Doutor André Osório E Cruz De Azerêdo Falcão
Profª. Doutora Rita Alexandra do Nascimento Cardoso Guedes

2023

Acknowledgments

I want to thank Dr André Falcão for believing in me and recommending this project. I also want to thank the guidance and availability which made my life way easier. I would also like to thank FCT (Fundação para a Ciência e a Tecnologia) for financing the scholarship and this project in the context of the project ”ChemSlotProtein: Expanding the druggable universe to unlock chemists’ creativity”(EXPL/QUI-OUT/1288/2021).

Resumo

O tempo médio de produção de um fármaco situa-se entre os 10 e os 15 anos e este processo tem custos muito elevados que têm vindo a aumentar. Para tentar reduzir a duração deste processo, reduzir os custos e cumprir com regulações e questões éticas, é frequente a aplicação de métodos in silico. Estes métodos incluem triagem virtual, uma prática que envolve a análise de conjuntos de compostos com a finalidade de identificar aqueles que possuam potencial terapêutico, aproveitando as diversas bases de dados disponíveis. Métodos de triagem virtual baseados nos ligandos, em que são utilizados os atributos químicos e biológicos de ligandos para os quais se sabe a atividade no alvo em estudo, são capazes de prever com eficácia as iterações com alvos extensivamente estudados. Uma técnica bastante utilizada que pertence a esta categoria é a dos modelos Relação Estrutura-Atividade Quantitativa (QSAR), que tipicamente se baseia no uso de modelos de machine learning para encontrar a relação entre as características estruturais dos compostos e as suas atividades. Contudo, não é possível fazer uso destas metodologias se não existir informação sobre o alvo que se pretende estudar. De forma a ultrapassar esta contrariedade, este trabalho introduz uma metodologia nova que utiliza informações estruturais sobre as proteínas alvo em modelos de machine learning. Este conjunto de dados é utilizado para treinar os modelos de forma a que estes consigam capturar as relações entre os atributos estruturais dos compostos e os atributos estruturais das proteínas e distinguir nuances ocorrentes nas interações de vários alvos. O objetivo é melhorar a sua capacidade de prever as iterações proteína-composto, tornar possível fazer previsões quando não há informação disponível sobre atividades de compostos na proteína-alvo nas várias bases de dados e fazer previsões para compostos ainda não testados.

Para testar a eficácia e a versatilidade desta metodologia, vários conjuntos de dados incluindo toda a informação disponível para alvos da família Receptores Acoplados a Proteínas G humana disponível na base de dados do CHEMBL englobando dois tipos de bioatividades: Ki, contante de inibição, e IC50, concentração inibitória média. Além de conjuntos de dados com a informação disponível para cada um destes tipos de bioatividades, foram utilizados conjuntos de dados que consistiam em toda a informação disponível. O AlphaFold 2, uma ferramenta de inteligência artificial que faz previsões das estruturas de proteínas a partir das suas sequências de aminoácidos, foi utilizado para extrair a informação estrutural dos alvos incluídos nos conjuntos de dados. O processo de extração de informação estrutural, tirando inspiração nos métodos de extração de "fingerprints" em compostos, é inovador e nunca tinha sido utilizado neste contexto. O processo consiste em, a partir dos ficheiros gerados pelo AlphaFold 2, definir e aplicar um raio com centro

num átomo e obter o conjunto de aminoácidos dentro desse raio, sendo que um aminoácido é considerado pertencente a essa esfera se algum dos seus átomos estiver a uma distância menor que o raio. Após obter os conjuntos de aminoácidos para cada átomo, estes foram convertidos num único vetor binário que, subsequentemente, sofreu uma redução dimensional. Para esse efeito foi utilizada a Análise de Componentes Principais (PCA), que é uma técnica de análise estatística para reduzir a dimensionalidade de conjuntos de dados. As fingerprints dos ligandos foram obtidas com recurso à ferramenta RDKit, que entre outros tipos de fingerprints, permite a geração de Morgan Fingerprints a partir dos identificadores SMILES de compostos. Estes identificadores foram extraídos da base de dados do CHEMBL. No total, foram criados 12 conjuntos de dados, 4 para cada tipo de bioatividade, contando o conjunto de dados em que foi utilizada informação sobre Ki e IC50 bioatividades em conjunto como um tipo de bioatividade. Para comparar os resultados da abordagem seguida, foram desenvolvidos modelos para cada um dos alvos que utilizavam apenas as informações estruturais dos compostos para prever as suas bioatividades. Este procedimento foi seguido para todos os alvos que cumprissem os requisitos mínimos. Assim, é possível ter um termo de comparação que permite comparar os resultados obtidos com aqueles que seriam os resultados se se tivesse seguido a metodologia QSAR para cada um dos alvos. Para cada alvo foram utilizados 3 algoritmos de machine learning para criar os modelos: Support Vector-Machine, Random Forests e Gradient-Boosting. Estes modelos foram gerados com auxílio da biblioteca SkLearn e apenas o melhor resultado dos três foi utilizado nas comparações. Os parâmetros de cada modelo foram definidos através de um processo de 5-fold cross-validation, usando depois um conjunto de validação com 20% do conjunto de dados (atividades pertencentes ao alvo) para obter os resultados finais dos modelos.

Três tipos de modelos foram utilizados para testar a abordagem seguida neste trabalho. No primeiro tipo a partição do conjunto de dados em conjunto de treino e de teste era feita aleatoriamente. Este tipo de modelo serviu para testar numa primeira instância se um modelo de machine learning conseguia capturar as relações entre os atributos estruturais das proteínas e os atributos dos ligandos e correlacioná-los com as atividades. Os resultados mostraram que o modelo tinha esta capacidade para conjuntos de dados de todos os tipos de bioatividades e com as fingerprints das proteínas geradas com raio de 5 Å e 7 Å. Para os conjuntos de dados com as fingerprints não reduzidas, os resultados obtidos ficaram sempre entre 0.167 e 0.17 para RMSE (raiz do erro quadrático médio) e 0.681 e 0.633 para RVE (rácio da variância explicada). Por outro lado, nos conjuntos de dados com fingerprints reduzidas, os resultados obtidos ficaram entre 0.199 e 0.21 para RMSE e 0.502 e 0.525 para RVE. Para comparação, os modelos QSAR desenvolvidos obtiveram 0.167 para RMSE e 0.644 e 0.683 para RVE. Estes resultados indicam que é possível obter bons resultados usando informação estrutural das proteínas e enriquecer conjuntos de dados com dados de vários alvos.

O segundo tipo de modelo tinha como objetivo simular a previsão de bioatividades para um alvo sobre o qual não havia informações disponíveis. Neste tipo de modelo, os alvos a serem testados eram escolhidos aleatoriamente e constituíam o conjunto de teste, com o conjunto de treino

a ser constituído pelos restantes alvos. Contudo, o foco depressa foi desviado deste modelo após os resultados iniciais não serem satisfatórios. Como o objetivo final deste trabalho era desenvolver um procedimento que permita desenvolver um modelo capaz de fazer previsões para alvos sem atividades disponíveis, foi depois testado um terceiro tipo de modelo.

No terceiro tipo de modelo, 15 alvos com foram escolhidos para fazer parte do conjunto de teste. Os alvos foram escolhidos por fazerem parte dos pares de alvos mais semelhantes, com cada elemento do par a ter pelo menos 100 atividades. Todos os dados relativos a estes alvos de teste constituíram o conjunto de teste, com o conjunto de treino a ser formado pelos dados relativos aos restantes alvos do conjunto de dados. Neste modelo já houve variabilidade entre conjuntos de dados e os resultados em geral não foram tão positivos, embora isso fosse expectável devido à forma como a partição foi feita. No entanto, em todos os conjuntos de dados existem alvos para os quais o modelo é capaz de captar um sinal distinto. Os dois conjuntos de dados com bioatividades Ki e com fingerprints das proteínas geradas com raio 5 Å são os que melhor demonstram o potencial desta abordagem. No conjunto em que as fingerprints dos ligantes não foram reduzidas, foi obtido RVE acima de 0.2 para 6 dos alvos e acima de 0.5 para 2 alvos e, no conjunto em que as fingerprints dos ligantes foram reduzidas, foi obtido RVE acima de 0.2 para 9 dos alvos e acima de 0.5 para 3 alvos.

Palavras-chave: Descoberta de drogas; modelação Relaçao Estrutura-Atividade; Triagem virtual baseada na estrutura; Machine Learning; Estrutura de proteínas.

Abstract

In silico drug development is increasingly being seen as a valuable approach, offering a faster pace for drug discovery while concurrently driving down costs. Ligand-based methods, anchored on the chemical and biological attributes of recognized ligands, are effective in predicting interactions with extensively studied target proteins. However, a hurdle presents itself when these methodologies are tasked with making predictions for targets that haven't been previously explored. Addressing this shortcoming, a novel methodology is introduced that melds structural data of target proteins—sourced from AlphaFold 2's predictive structures—into machine learning models, thereby elevating their ability to predict protein-molecule interactions. This innovative strategy refines and extracts 3D structural protein fingerprints and then amalgamates them with the structural data of ligands This enriched dataset trains our machine learning model to discern the nuanced dynamics between ligand attributes and distinct structural nuances of myriad target proteins, facilitating predictions for previously uncharted molecules and protein targets.

To assess the efficacy of the introduced model, comprehensive datasets, encompassing two distinct bioactivity types and detailing the entirety of available information for Human G-Protein Coupled Receptors, were employed. The derived insights highlighted that this innovative strategy parallels the prowess of contemporary traditional ligand-based methodologies. It was found that, in some cases, the model exhibited a unique ability to accurately predict interactions for target proteins that were not included in its training phase. This ability hinges on the presence of some level of similarity between these external proteins and those within the training set. Such capabilities underscore the approach's expansive potential, offering a promising avenue to advance drug research into previously uncharted target proteins.

In summary, this work introduces a novel in silico drug discovery technique, adeptly merging ligand-centric methodologies with structural data integration to predict protein-molecule interactions. By integrating both protein and ligand structural information within a machine learning framework, the model paves the way for robust, automated predictions, even for targets that were absent during the training process. This innovation signifies a breakthrough in drug discovery, presenting far-reaching implications for the future landscape of the pharmaceutical industry.

Keywords: Drug Discovery; Quantitative Structure-Activity Relationship modeling; Structure-Based Virtual Screening; Machine Learning; Protein structure.

Contents

Lista de Figuras **xv**

Lista de Tabelas **xvii**

1 Introduction **1**

2 Background and related work **5**

2.1 Proteins and their function 5

2.2 Protein structure similarity 6

2.3 Protein Structures 7

2.4 Molecules and their function 8

2.5 Drug Discovery 9

2.6 Molecular similarity 10

2.7 Quantitative Structure-Activity Relationship 12

2.8 Multi-target QSAR 14

2.9 Performance Metrics 15

2.10 Related work 15

3 Data and methods **19**

3.1 Protein fingerprints 19

3.2 Adapting machine learning methods for protein structure 21

3.3 Data retrieving and curation 21

3.4 Baseline models 24

3.5 Global and individual models 25

3.6 Dataset 26

3.7 Pairs 26

4 Results **35**

4.1 Ki datasets 35

4.2 Comparative analysis of Ki and IC50 datasets 39

5 Conclusion **47**

Abreviaturas **51**

Bibliografia **52**

A Results data **63**

List of Figures

1.1 Visual representation of the traditional QSAR approach on Figure 1.1a, where the descriptors of the molecules with known activities are used to create a model capable of predicting activities for untested molecules. On Figure 1.1b, there is the visual representation of the proposed methodology where descriptors of multiple targets and molecules are used to create the model 2

3.1 Extration of the protein fingerprints 20
3.2 Overview of the workflow to the fingerprints 21
3.3 Workflow overview for obtaining baseline results. This figure provides an overview of the workflow followed to obtain the baseline results. The process was applied to each target and with Ki bioactivities. The dataset used in the workflow includes molecule fingerprints along with their corresponding bioactivity values. 24
3.4 The representation of the different modeling techniques used in this work 25
3.5 Tanimoto coefficient of the pairs for each bioactivity type 28

4.1 RMSE and RVE validation results for the Ki dataset with radius=5Å without PCA 36
4.2 RMSE and RVE validation results for the Ki dataset with radius=5Å with PCA . 37
4.3 RMSE and RVE validation results for the Ki dataset with radius=7Å without PCA 40
4.4 RMSE and RVE validation results for the Ki dataset with radius=7Å with PCA . 41
4.5 RMSE and RVE validation results for the IC50 dataset with radius=5Å without PCA 42
4.6 RMSE and RVE validation results for the IC50 dataset with radius=5Å with PCA 43
4.7 RMSE and RVE validation results for the IC50+Ki dataset with radius=5Å without PCA 44
4.8 RMSE and RVE validation results for the IC50+Ki dataset with radius=5Å with PCA 45

List of Tables

3.1 List of targets for the datasets with Ki activities and with protein fingerprints generate with radius=5 . . . 29
3.2 List of targets for the datasets with Ki activities and with protein fingerprints generate with radius=7 . . . 30
3.3 List of targets for the datasets with IC50 activities and with protein fingerprints generate with radius=5 . . . 31
3.4 List of targets for the datasets with IC50 activities and with protein fingerprints generate with radius=7 . . . 32
3.5 List of targets for the datasets with IC50+Ki activities and with protein fingerprints generate with radius=5 . . . 33
3.6 List of targets for the datasets with IC50+Ki activities and with protein fingerprints generate with radius=7 . . . 34

4.1 Results and training and testing set sizes for all models - Ki with radius=5Å and without PCA . . . 35
4.2 Results and training and testing set sizes for all models - Ki with radius=5Å and with PCA . . . 35
4.3 Comparison of results for all 15 validation targets of the semi-blind model for the Ki dataset with radius=5Å and without PCA . . . 38
4.4 Comparison of results for all 15 validation targets of the semi-blind model for the Ki dataset with radius=5Å and with PCA . . . 39

Chapter 1

Introduction

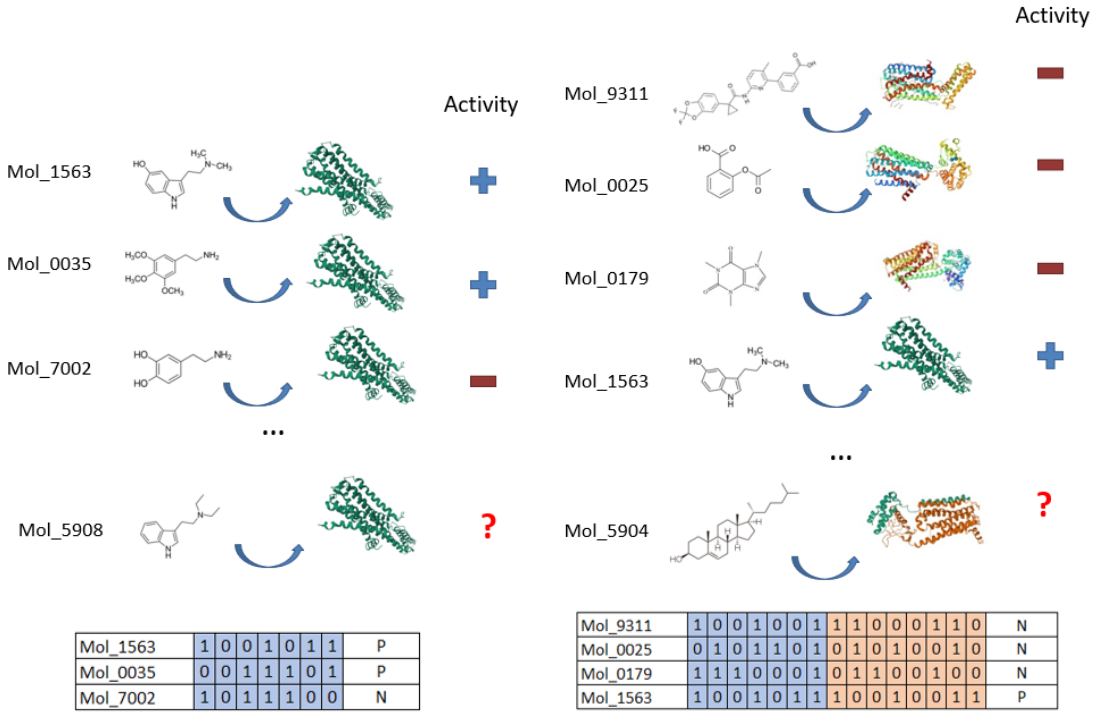
In silico approaches have been attracting considerable interest because of their potential to accelerate drug discovery in terms of time, labor, and costs. Many new drug compounds have been successfully developed using computational methods [81, 19]. There are several in silico methods used in drug discovery. Some of these methods include computational identification of potential drug targets, virtual screening of large chemical libraries for effective drug candidates, further optimization of candidate compounds, and in silico assessment of their potential toxicity. There are two main methods used for in silico drug discovery: ligand-based drug discovery (LBDD) and structure-based drug discovery (SBDD) [11, 78, 81]. SBDD techniques make use of the 3D structure of the target proteins to try to predict how molecules might bind and try to find molecules with high affinity. SBDD includes some popular methods like molecular docking and structure-based virtual screening (SBVD) [58, 81, 6, 54]. On the other hand, LBDD focuses on the chemical and structural characteristics of the ligands themselves and is used to identify and design new drugs based on the properties of known ligands or molecules that interact with the target of interest. The use of quantitative structure-activity relationship (QSAR) models is one of the most used LBDD approaches [49, 60].

QSAR modeling is an approach that utilizes the chemical and structural characteristics of known ligands to design and predict the activity of new compounds. QSAR models establish a relation between the molecular descriptors of ligands and their biological activities, enabling the prediction of activity for untested compounds. QSAR models are a classic and common method in drug discovery.

QSAR models, as most Machine Learning models, perform better when the dependent variable is the same for all observations, therefore in the large majority of QSAR studies, modelers select one specific target for which there is already some data with actual observations and try to identify the characteristics of the molecules that potentiate similar behavior to the active ones in the data set (Figure 1.1a). Depending on the quality and quantity of available data, it is nonetheless possible to develop good predictive models for individual targets for which there exists reliable molecular activity data [105, 95, 90, 14, 60].

The hypothesis guiding this study is that it is possible to use Machine Learning models with structural information of the targets and molecules to predict the activity of untested molecules.

Hence the goal of the current work is to surpass this limitation of QSAR models by including structural information of the proteins in the modeling procedure, so that the model may understand the characteristics of the proteins that are relevant and therefore be able to make larger models that may be able to reproduce the results of small individual models and also capable of making predictions targets that were never in the training set and that may even never have been recorded or assessed. To accomplish the above goal, the structural information for proteins will come from AlphaFold [40] predictions, which is an artificial intelligence (AI) system that predicts the 3D structure of proteins from their amino acid sequences, transformed into a vector format that will be added to the structural information of small molecules for developing predictive models (Figure 1.1b).



(a) Traditional QSAR (b) Proposed methodology

Figure 1.1: Visual representation of the traditional QSAR approach on Figure 1.1a, where the descriptors of the molecules with known activities are used to create a model capable of predicting activities for untested molecules. On Figure 1.1b, there is the visual representation of the proposed methodology where descriptors of multiple targets and molecules are used to create the model

Furthermore, this work demonstrates the feasibility of developing a methodology utilizing the information offered by AlphaFold. The abundance and quality of data provided by AlphaFold may explain the absence of a similar methodology in the past. In 2018, the Protein Data Bank (PDB) contained approximately 140,000 unique protein structures [8, 101]. While this number has increased to about 180,000 structures presently available, it still pales in comparison to the over 200 million structures accessible in the AlphaFold Database [34]. Although the predictions generated by AlphaFold may not be entirely flawless, it is important to note that the structural

data available in the PDB also has its limitations, including missing atoms, side-chains, water molecules, and solvent information [21, 88]. Numerous software programs have been developed to address these deficiencies [61, 36, 70], but the utilization of PDB-formatted files provided by AlphaFold offers a simpler approach.

Chapter 2

Background and related work

In this chapter, we will explore key aspects related to the context in which the techniques employed in this thesis are used, known as the state-of-the-art. Additionally, we will discuss the techniques and tools that will be utilized in the development of this work, as well as important conceptual foundations that underpin the research.

2.1 Proteins and their function

The function of proteins is diverse and essential for life. Proteins can act as enzymes, which catalyze biochemical reactions in cells, facilitating various metabolic processes. Enzymes are involved in processes such as digestion, energy production, DNA replication, and cell signaling. Proteins also play important roles in cell signaling and communication. They can act as receptors, receiving signals from the environment or other cells, and transmitting these signals to the interior of the cell, leading to specific cellular responses. Proteins are also involved in transporting molecules and ions across cell membranes, regulating gene expression, and defending against pathogens as part of the immune system.

Proteins are fundamental biomolecules that play crucial roles in various biological processes. They are composed of long chains of amino acids and are essential for the structure, function, and regulation of cells and organisms.

Proteins have four levels of structure: primary, secondary, tertiary, and quaternary. The primary structure of a protein is determined by the sequence of amino acids that make up the protein chain. There are 20 different amino acids that can be combined in different sequences to form a vast number of proteins with diverse structures and functions. The sequence of amino acids in a protein is encoded in the genetic material (DNA) of an organism. The Secondary structure refers to local folding patterns of the protein chain, such as alpha helices and beta sheets. Tertiary structure refers to the overall three-dimensional folding of the protein molecule, which is crucial for its function. Quaternary structure refers to the arrangement of multiple protein subunits, if present, in a larger protein complex.

In summary, proteins are essential biomolecules with diverse structures and functions. They are involved in almost every aspect of cellular and organismal processes, including metabolism,

signaling, structure, and defense. Understanding the structure and function of proteins is crucial for advancing our knowledge of biology and developing new treatments for diseases.

2.2 Protein structure similarity

There are multiple reasons for comparing protein structures. Analyzing conformational changes in ligand binding, by comparing the ligand-bound and unbound structures, can help to assist rational drug design. The binding of a ligand to an active site in a protein often induces a structural change and the degree of structural change can be determined by this comparison. Comparing protein structures can also be useful in the detection of distant evolutionary relationships because the protein structure is more highly conserved than the protein sequence. Analyses of structural families have shown that homologous proteins (related by evolutionary divergence from a common ancestor) frequently share fewer than 15% identical residues but, usually, the structure remains similar. In some protein families, the sequence can vary significantly without translating into any change in the structure. The structural variability varies significantly between protein families and while in some families that are highly variable structural changes have no impact on functions, in other families the opposite happens. Protein comparisons can help identify parts of the structure that are structurally conserved across a family, such as active sites. Another possible use is to look for common structural motifs that occur due to constraints in packing the secondary structure. The analysis of these protein similarities can help understand the physicochemical requirements of protein folding and secondary structure packing [66].

There are several algorithms for comparing protein structures. These algorithms are comprised of two important components: techniques for scoring similarities in the structural features of proteins and the use of an optimization strategy that can identify an alignment that maximizes the structural similarities measured. The majority of methods compare the geometric properties and/or the secondary structure elements or residues along the carbon backbone but sometimes other non-geometric properties like physicochemical properties such as hydrophobicity are also included [66].

Usually, the approaches for comparing protein structures fall into two categories: intermolecular and intramolecular. Intermolecular approaches compare geometric properties, like residue positions in 3-D coordinate space. The most common method is to superimpose the protein structures and search for the superimposition that minimizes the intermolecular distances. In distant relatives, information about intramolecular relationships between residues or secondary structures is often included to obtain the correct alignment. The intramolecular approaches align the protein structures on the basis of information about internal relationships within each protein, like structural relationships between residues within the protein [66].

Challenges abound for protein structure comparison techniques, regardless of their reliance on intermolecular or intramolecular methods. Distant homologs may exhibit substantial insertions that lead to additional secondary structures. Additionally, mutations can trigger shifts in the orientations of equivalent secondary structures when these mutations result in residues of different sizes

occupying the same positions. To address these challenges, methodologies have been developed that employ varied strategies. These include excluding the variable loop regions where insertions and substitutions predominantly occur, segmenting proteins into fragments, and adopting more resilient optimization techniques [66].

In summary, the realm of protein structure comparison is intricate, necessitating sophisticated algorithms and methodologies. While these comparisons offer valuable insights into evolutionary relationships and drug design, the challenges they pose, particularly in handling insertions and substitutions, underscore the need for continuous advancement in this field.

2.3 Protein Structures

AlphaFold is a highly accurate protein structure prediction algorithm. It uses a combination of bioinformatics and physical approaches to predict the 3D structure of a protein from its amino acid sequence [40]. The problem of predicting the protein structure from the amino acids sequence has been a research problem for more than 50 years and a major problem was the number of ways that a protein can fold [91]. It is this structure that helps to determine its function and it is also important in the case of drug discovery because, for example, if we want to target some protein to block what it is doing, we need to understand how/where the compound is going to bind on the protein. Before AlphaFold, these structures could only be obtained experimentally and one way to do that is x-ray crystallography [85]. The problem is that these processes are time very consuming and there were only around 100,000 unique protein structures available which are a small portion of the sequences known [102], and Alphafold performs at a similar level or even better than experimental approaches in predicting protein structures even though this experimental methods are still important for validation. Now AlphaFold DataBase contains more than 200 million structures [34] and it is also publicly available to predict the 3D structures corresponding to user input amino acid strings.

AlphaFold is an advanced algorithm for predicting protein structures with remarkable accuracy. Combining bioinformatics and physical approaches, AlphaFold can determine the three-dimensional (3D) structure of a protein based on its amino acid sequence [40]. The prediction of protein structures from amino acid sequences has been a challenging research problem for over half a century, mainly due to the vast number of possible ways a protein can fold [91].

The 3D structure of a protein is critical for understanding its function and is particularly relevant in drug discovery. When targeting specific proteins for therapeutic intervention, it is essential to comprehend how and where a compound will bind to the protein [85]. Previously, experimental methods such as X-ray crystallography were employed to obtain protein structures, but these techniques are time-consuming. AlphaFold has revolutionized this landscape by predicting protein structures with comparable or even superior accuracy to experimental methods [107, 40], although experimental validation remains important. This breakthrough in protein structure prediction has significantly expanded the availability of structural information, providing valuable insights for various scientific fields, including drug discovery.

While no published research has yet employed an approach identical to the one presented here, there have been successful applications of AlphaFold's predictions in various studies. For instance, Kim et al. [44] harnessed AlphaFold to predict protein structures and assess the binding affinities of antipsychotics such as clozapine, olanzapine, and quetiapine with neuropharmacological, immunological, and metabolic receptors. This was achieved using GNINA [56], and the results indicated a high binding affinity of these drugs to diverse receptors, suggesting that cyclosporin A and everolimus could inform the development of novel antipsychotics.

In a separate study, a high-throughput analysis was designed and executed to examine Iron-Sulfur and Zinc binding sites on over 360,000 predicted structures within the AlphaFold database [98]. This large-scale analysis uncovered numerous potential binding sites previously unidentified in the corresponding UniProt accession. Additionally, it was revealed that the majority of known Fe-S cluster and Zn binding sites documented in UniProt were recovered by the AlphaFold2 structures.

However, it is crucial to note that AlphaFold's predictions are not flawless [96, 4, 79], and should be utilized with care.

2.4 Molecules and their function

Molecules are the smallest fundamental units of a chemical compound that retain the chemical properties of that compound. They consist of one or more atoms bonded together and can range in terms of complexity, from simple molecules like oxygen (O2) to more complex ones like DNA. Drug discovery is a multifaceted endeavor, aimed at identifying novel compounds that have therapeutic potential against specific diseases, so molecules play an essential role. Understanding the role and function of molecules within the drug discovery process is critical for the successful identification and development of new therapeutic agents.

The fundamental principle of drug action is the ability of a molecule (potential drug) to interact with specific targets within the body, such as proteins, nucleic acids, or other macromolecules. The interaction between a drug molecule and its target can either activate (agonists) or inhibit (antagonists) the function of the target, leading to a desired therapeutic effect. Beyond the primary interaction with their target, drug molecules are also designed with consideration for their pharmacokinetic properties. This encompasses the drug's absorption, distribution, metabolism, and excretion (often abbreviated as ADME). A molecule's physicochemical properties, such as solubility and stability, play crucial roles in determining its suitability as a drug. Furthermore, not all interactions of a drug molecule are therapeutic. Some interactions can lead to undesirable side effects. Therefore, a comprehensive understanding of a molecule's potential interactions in the biological system is crucial to predict and mitigate potential adverse effects.

Molecular design is a cornerstone of drug discovery. The process involves tailoring molecules to enhance their therapeutic potential and minimize adverse effects. This is achieved by making iterative changes to the molecular structure, based on feedback from biological testing. Techniques such as structure-based drug design, molecular modeling, and quantitative structure-activity rela-

tionships (QSAR) modeling are often employed to predict and understand the effects of molecular changes on biological activity.

2.5 Drug Discovery

The process of drug discovery is a complex and costly endeavor, often taking up to 15 years and exceeding $1 billion in expenses. This process can be divided into four stages, with the early drug discovery stage encompassing several key steps:

- **Target identification and validation.** One of the initial and crucial steps in drug discovery is identifying a suitable target that is "druggable," meaning it can be targeted by a drug molecule to elicit a desired biological response. Target validation involves understanding the target's role in the disease phenotype through a combination of in vitro and in vivo techniques, employing a multi-validation approach to increase confidence [35, 22].

- **Hit identification and validation.** During this stage of the drug discovery process, various screening assays are employed to identify "hit" compounds that exhibit the desired activity in a compound screen. High throughput screening (HTS) is a widely used method in which an entire compound library is screened against a specific target. However, in recent years, virtual screening (VS) techniques have gained popularity as a means to reduce time and costs in drug discovery [78]. VS techniques can be categorized into two main types: SBVS and LBVS. SBVS is employed when structural information about the target is available. It involves a docking approach where the ligand is docked to the biological target, and a score based on their affinity is calculated. LBVS, on the other hand, is utilized when information about the target structure is lacking. It relies on known structure-activity data of active molecules and employs approaches such as pharmacophore-based search, QSAR modeling, and similarity search. These techniques can also be combined to enhance hit identification [78].

  Following hit identification, the number of compounds is further reduced, and the remaining hit compounds are clustered based on their structure-activity relationship (SAR). Representative members of each cluster are then evaluated for absorption, distribution, metabolism, and excretion (ADME) properties, as well as pharmacokinetic (PK) properties. This analysis aids in the selection of a series that minimizes in vivo side effects [35].

- **Hit to lead and lead optimization**. During this stage of the drug discovery process, the initial hit compounds are refined to generate more potent and selective compounds with optimized PK properties. This process involves conducting SAR investigations to enhance the selectivity and potency of the compounds, thereby reducing the likelihood of undesirable interactions with other targets. Attention is also given to the ADME properties of the compounds. Furthermore, assessments of solubility and permeability are performed to determine the optimal route of administration[35, 47, 48].

> The objective of lead optimization is to maintain favorable properties in the lead compounds while addressing any deficiencies in their structures. It is crucial to continue searching for backup molecules during this stage, as a selected compound may fail in subsequent preclinical or clinical research [35].

Preclinical research represents the second stage of the drug discovery process, involving rigorous testing of drug candidates through both *in vitro* (laboratory-based) and *in vivo* (animal-based) experiments. The primary objective of this stage is to assess whether the drug candidates exhibit the desired pharmacological activity and to evaluate their safety profiles. These studies help researchers understand how the compounds interact with biological systems and provide crucial information for subsequent clinical trials [25, 48, 80].

Following preclinical research, the drug candidate progresses to the clinical research stage, which involves conducting trials on human subjects. Clinical trials consist of four distinct phases, each with specific objectives. The overarching goal is to ensure the safety of the drug and determine the appropriate dosage. Phase I trials focus on assessing the drug's safety profile, dosage range, and potential side effects in a small group of healthy volunteers or patients. Phase II trials involve a larger cohort to evaluate the drug's efficacy and further assess its safety. Phase III trials expand the participant pool to gather additional evidence of effectiveness and monitor adverse reactions in a larger patient population. Finally, phase IV trials, which occur after regulatory approval, involve post-market surveillance to monitor the drug's safety and effectiveness in real-world settings [25, 48, 80].

The focus of this project is on virtual screening (VS) techniques, which find application within the drug discovery stage. These techniques can be utilized for hit discovery and lead optimization, providing valuable insights into the potential bioactivities of candidate compounds.

2.6 Molecular similarity

Molecular similarity is a fundamental concept extensively studied and applied in ligand-based virtual screening, chemical informatics, and medicinal chemistry [53]. The underlying motivation for exploring molecular similarity is the "similar property principle," which asserts that molecules with similar structures tend to exhibit similar properties [39]. However, quantifying molecular similarity lacks an absolute measure, and results can be evaluated and interpreted in various ways [53, 23].

Among the different methods available for comparing molecules, molecular fingerprints combined with a similarity/dissimilarity metric are commonly used. The concept of molecular fingerprints is based on the idea that the structural and chemical characteristics of a molecule can be encoded into a binary string. Each bit in the fingerprint represents the presence or absence of a specific substructure or feature of a molecule [12, 46]. By capturing the structural information of molecules and allowing the comparison of compounds based on their structural features using similarity measures, such as the Tanimoto coefficient [67, 5]. This approach is preferred

over computationally demanding and inefficient methods like direct graph comparisons and graph similarity calculations [53]. Several types of molecular fingerprints exist:

- Substructure keys-based fingerprints set bits in a bit string based on the presence of specific substructures or features from a predetermined list of structural keys [13]. Examples of substructure keys-based fingerprints include MACCS and PubChem fingerprint. However, one limitation of this fingerprint type is that features not represented by the bits in the array are not captured, similar to the limitation of substructure searches that require precise matching of molecular activities and atom types [13, 53, 7, 100].

- Topological fingerprints analyze the fragments of a molecule by following paths up to a certain number of bonds and hashing each path to create the fingerprint [13]. In fingerprints with a reduced number of bits, one bit can be set by multiple fragments, making it impossible to determine the specific fragment associated with the bit [13]. An example of topological fingerprints is topological atom pair descriptors, which count the occurrences of atom pairs with specific features at a predefined topological distance [55].

- Circular fingerprints, such as the ECFP fingerprints based on the Morgan algorithm, examine the environment of each atom up to a predetermined radius instead of paths. They are commonly used for full structure similarity searching [13, 53, 59]. The Morgan algorithm assigns initial identifiers to each atom in the molecule, and in each iteration, the identifiers of neighboring atoms are combined and hashed to generate new identifiers. The final fingerprint is constructed by hashing the identifiers of all atoms in the molecule [12].

- Hybrid fingerprints combine multiple approaches to set the bits in the bit string using different methods [13].

- Pharmacophore fingerprints encode information about the presence of features from a predefined list in a molecule while considering the distances between these features, often using a list of distance ranges [13].

- Other types of fingerprints adopt different approaches. For instance, there are text-based fingerprints calculated based on the canonical SMILES representation of molecules, as well as protein-ligand interaction fingerprints that encode information about protein-ligand interactions [13]. Additionally, ongoing efforts are being made to develop molecular representations that incorporate three-dimensional spatial relationships, conformational dynamics, and kinetic pathways, although these methods are still in the developmental stage [16].

While various similarity metrics exist, the Tanimoto coefficient is the most commonly used [53]. However, due to the use of different fingerprints, establishing a universal threshold for similarity is challenging, and the results obtained can also vary depending on the class of compounds being compared.

Given the diverse ways to represent molecules and evaluate their similarity, an emerging trend is the use of data fusion [13, 99]. Data fusion involves performing multiple search calculations using different reference molecules, similarity measures, or molecular representations, thus capturing diverse chemical information [89, 13, 99]. Several studies have demonstrated that data fusion techniques yield results that are at least as effective as individual searches using conventional approaches [99].

2.7 Quantitative Structure-Activity Relationship

The SAR refers to the correlation between the chemical structure of a molecule and its biological activity. Understanding this relationship is crucial as it allows researchers to identify the structural features that are important for a molecule's activity in a specific target. In the field of drug discovery, QSAR models are commonly employed techniques for screening libraries of molecules, using information about both active and inactive compounds for a given target. A QSAR model can be described as a function that takes descriptors derived from chemical features as input parameters and predicts the value of the desired biological activity.

While QSAR modeling has been successfully applied for many years, the field continues to evolve, and there is a growing recognition of the need for robust validation. Several papers have emphasized the importance of validation in QSAR modeling [15, 1, 95, 93, 30, 29]. It is generally considered essential that models undergo external testing, in addition to internal testing, to ensure their reliability for making predictions on external datasets and for regulatory purposes [30, 29]. In 2004, the Organisation for Economic Co-operation and Development (OECD) adopted principles for the validation of QSAR models [37] and modified them into five principles, which include defining the endpoint, using an unambiguous algorithm, defining the domain of applicability, employing appropriate measures of goodness-of-fit, robustness, and predictivity, and providing a mechanistic interpretation if possible [62, 14]. However, despite these guidelines, Dearden et al. reported 21 errors that violated the OECD principles, demonstrating the need for continuous improvement in QSAR modeling practices [20].

The importance of good practices in QSAR has been emphasized in highly cited papers [95, 14]. One critical aspect is chemical data curation, as it is often assumed that experimentally obtained input data is correct. However, a study cited in both papers [64] demonstrated that, on average, there are two chemical annotation errors per medicinal chemistry publication, with an overall error rate in the popular commercial WOMBAT database reaching as high as 8% (Tropsha et al., 2010)[65, 95]. This finding is particularly significant considering other studies have shown that the type of chemical descriptors used has a greater impact on the prediction performance of QSAR models than the nature of the model optimization [106, 92, 95].

In 2010, Fourches et al. developed a strategy for chemical data curation, offering best practices and recommendations for implementation [27]. They also suggested several accessible software tools. Their proposed workflow enables the identification and correction of structural errors, which involves removing inorganics, organometallics, counterions, and mixtures that most QSAR

descriptor generation programs are ill-equipped to handle or that lead to confounding duplicates when simplified (e.g., desalted). Additional curation elements include structural cleaning (e.g., detection of valence violations), ring aromatization, normalization of specific chemotypes, and standardization of tautomeric forms. Postprocessing steps involve deleting duplicates resulting from curation, standardization, and normalization, as well as performing manual checks on complex cases [14].

Tropsha also outlines two additional steps prior to model validation: dataset balancing and detection/removal of outliers [95]. Dataset balancing refers to ensuring that the number of compounds in the dataset falls within a certain range. While the specific upper limit may vary depending on the situation, Tropsha suggests a minimum dataset size of 40 compounds, with 20 for training, and 10 each for testing and validation sets. For classification tasks, it is recommended to have at least 20 compounds from each class in the training set and a minimum of 5 compounds from each class in the test and validation sets. Activity values also need to meet certain requirements. In the case of continuous variables, the total range of activities should be at least 5 times the experimental error, and consecutive activity values should not have gaps exceeding 10% to 15% of the entire range. In classification tasks, it is desirable to have a balanced number of compounds per class, but this is often not feasible. To address this, Tropsha suggests calculating a distance matrix between compounds from different classes and excluding those that exceed a similarity threshold. This approach accounts for the fact that smaller classes typically have fewer compounds, and these compounds are more likely to exhibit structural similarity.

In addition to dataset size and balancing, the presence of outliers must be considered. Outliers can adversely affect the reliability of the model and should be removed. This paper classifies outliers into two categories: structural outliers and activity outliers. Structural outliers refer to compounds that do not fit into any cluster when the dataset is subjected to clustering using available techniques. Activity outliers, on the other hand, are defined as ”activity cliffs” and represent regions in the dataset where small changes in compound structure lead to significant changes in molecular properties. These activity cliffs are particularly important because QSAR models assume that similar compounds have similar properties [95].

In 2010, Tropsha emphasized the importance of using an external validation set, separate from the model-building process, to accurately assess the predictive power of a QSAR model [95]. This external validation set is crucial because typically multiple models are constructed, and the set is used to select the best-performing model. Additionally, the selection of the validation set itself is a critical aspect of the validation process. Ideally, an entirely independent dataset with new experimental data would be used for validation. However, this is not always feasible, and it is acknowledged that different data splits can yield different results [82, 93].

To simulate validation with a new dataset, one approach is to utilize a time-split validation set, where the compounds tested more recently are chosen [82]. It has been demonstrated that this time-split approach yields results most similar to prospective validation, with random selection tending to overestimate the results and leave-class-out underestimating them. Furthermore, some

users of CHEMBL, a widely used chemical database, employ the date of publication as a threshold to split the data [93]. This helps ensure that the validation set closely simulates the conditions of real-world applications.

When developed with sufficient care, QSAR models can serve as powerful tools in drug discovery, particularly in virtual screening (VS) approaches for lead optimization and the identification of new drugs [69, 95]. For instance, Peterson et al. conducted a study where QSAR models were employed to screen a large collection of commercially available chemicals, leading to the identification of hits with novel scaffolds. This study also demonstrated the limitations of traditional chemical similarity methods in identifying these hits [69].

By giving careful attention to the development of QSAR models, researchers can harness their potential to drive drug discovery efforts and uncover promising chemical candidates that might otherwise go unnoticed.

2.8 Multi-target QSAR

Traditional QSAR models have two significant limitations: they typically predict the biological activities of drugs against a single target and are often limited to homogeneous compounds [93, 86, 60]. This restricts the exploration of structural patterns that may be relevant to the desired activity. Moreover, many pharmaceutical agents interact with multiple targets, which can be advantageous but may also lead to undesired interactions and adverse effects, often impeding further development [60, 76]. Consequently, there is a growing interest among researchers in utilizing multi-target QSAR to predict biological activities for structurally diverse compounds against multiple targets, particularly in the context of drug discovery [93].

One notable difference between these models and conventional QSAR models developed using popular methods lies in the descriptors employed. Various studies have utilized different types of descriptors within the same model [86, 87], including descriptors that are uncommon or rarely used in typical QSAR studies [71]. Although the models may employ different methodologies, with some studies highlighting the advantages of artificial neural networks compared to linear discriminant models [86, 87, 71], the overall results obtained from these models are generally promising. Importantly, these models offer the advantage of predicting activities against multiple targets simultaneously.

Multi-target QSAR can be helpful for identifying potential drug candidates and for designing new drugs since these models can predict the activity of compounds against a wider range of targets [45, 75, 9]. However, one challenge of multi-target QSAR is that it can be more difficult to develop models that are accurate for multiple targets. This is because the relationships between molecular structure and biological activity can be more complex when multiple targets are involved [74, 75]. As research in multi-target QSAR continues, it is possible that the accuracy of these models will improve. This could lead to the development of new drugs that are more effective and have fewer side effects [33].

2.9 Performance Metrics

A diverse range of metrics is available for the evaluation of machine learning models. In assessing the performance of the developed models, two key metrics were employed: Root Mean Squared Error (RMSE) and the Ratio of Explained Variance (RVE).

RMSE is a widely used metric for regression models, providing a measure of the average deviation between predicted and actual values. Mathematically, it is defined as:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{2.1}$$

Here, y_i represents the actual values, $\hat{y}_i$ signifies the predicted values, and n is the number of observations. RMSE penalizes larger errors more heavily and is expressed in the same units as the target variable, offering an interpretable measure of prediction accuracy.

The RVE quantifies the proportion of variance in the dependent variable that the model is capable of explaining. Although there are different ways to calculate this metric, in this work it was used the R^2 formula which is defined as:

$$R^2 = 1 - \frac{\text{Var}(y - \hat{y})}{\text{Var}(y)} \tag{2.2}$$

Here, $\text{Var}(y - \hat{y})$ denotes the variance of the residuals, representing the differences between actual (y) and predicted ($\hat{y}$) values, and $\text{Var}(y)$ is the variance of the true values. The R^2 score ranges between 0 and 1, with higher values indicating a better fit and 1 indicating a perfect fit.

These metrics were chosen to provide a comprehensive assessment of model performance, considering both the accuracy of predictions (RMSE) and the ability of the model to explain the variability in the target variable (RVE).

2.10 Related work

In this section, some works with similar objectives, namely the prediction of compounds' bioactivities in one or more targets, will be mentioned. As explained above, there are multiple approaches to this problem, so the focus will be on works that utilize SAR.

There are numerous ways to approach SAR studies. Following an approach that focuses on molecular space visualization may help to explore the diversity of heterogeneous chemical data and increase the understanding of SAR. This was the approach followed by Kausar et al. in a work where the goal was to compute a probabilistic surface of molecular activity (PSMA) [43]. The molecular distances were determined based on their similarity and then four dimensionality reduction methods were used in order to determine the one that in the end led to better results and to represent the molecules in a 2D space. Then it was used a non-parametric 2D kernel density estimation function to map the activity surfaces. The results of the test set showed that it can also be a good classification model.

In a paper regarding a well-known virus, Alves et al. present a computational approach to identify potential inhibitors of the SARS-CoV-2 main protease, a key target for antiviral drugs [3]. They used a combination of molecular docking, quantitative structure-activity relationship (QSAR) modeling, and experimental assays to screen a library of over 6,000 compounds, including FDA-approved drugs and natural products. The authors identified several compounds that showed promising activity in vitro, including some drugs that are already approved for other indications. The results suggest that these compounds could be further developed as potential treatments for COVID-19. This study provides valuable insights into the repurposing of existing drugs for the treatment of COVID-19 and highlights the potential of computational methods to accelerate drug discovery efforts.

Multi-target QSAR is also a very useful methodology that is related to the work being developed. As an example will be mentioned the work of Alejandro Speck-Planche et al. where was introduced a novel chemoinformatic multi-target approach for virtual screening, prediction, and in silico design of compounds anti-Breast Cancer against several Breast Cancer cell lines [86]. This work used 3 different types of structural descriptors and a Linear Discriminant Analysis model as the classification model. The classification results obtained were very good but another advantage of this methodology is the possibility to calculate the quantitative contribution of any fragment to anti-Breast Cancer activity which is important for in silico design of new compounds.

Chapter 3

Data and methods

3.1 Protein fingerprints

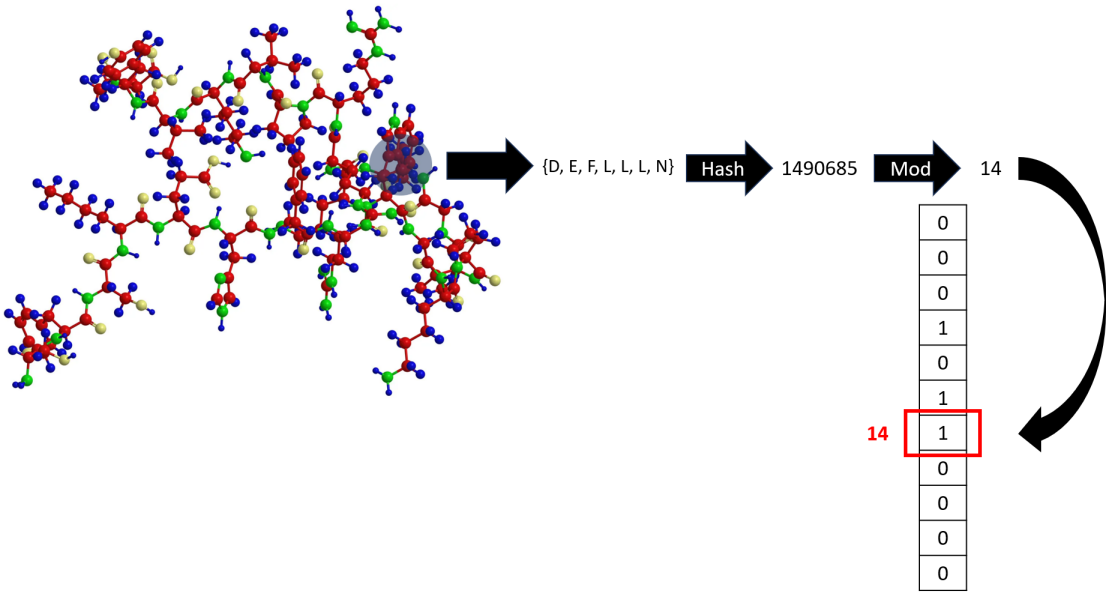
The primary aim of this work is to investigate the feasibility of incorporating proteins' structural information into QSAR modeling to enable predictions for multiple targets. To achieve this, a method was required to represent this information as a binary vector, where each bit corresponds to the presence of a specific structural feature.

The developed method involves generating sets of amino acids within a specified radius and encoding them into a binary vector. Since the optimal radius for obtaining effective fingerprints was unknown, two distances were considered: 5 Å and 7 Å. The distances were calculated at the atom level using the coordinates extracted from the protein's PDB files.

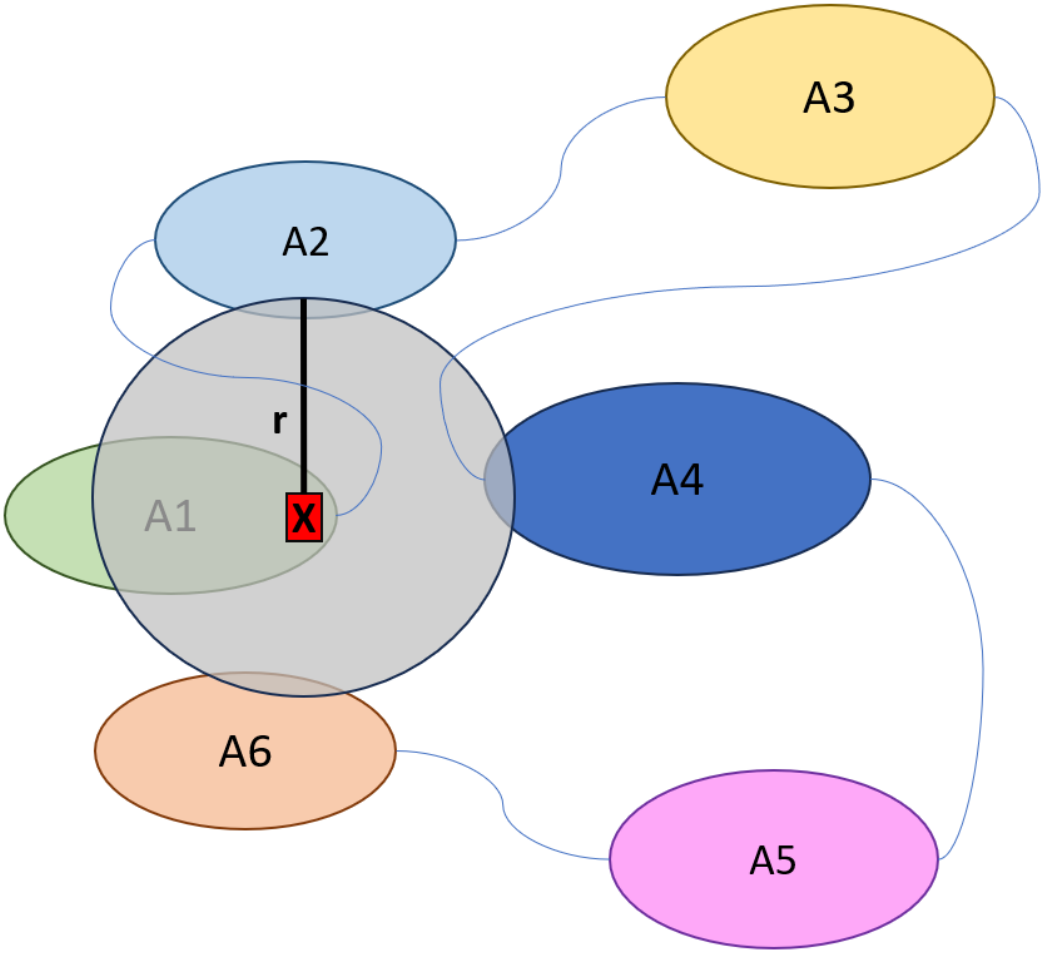
An amino acid is considered part of the "central" amino acid's set if any of its atoms fall within the specified radius of the central amino acid's atoms. Multiple sets can be obtained by varying the atom used to define the radius within the same amino acid, resulting in a comprehensive search across all atoms of the protein. Sets consisting of a single amino acid (where no other amino acid fell within the radius) were excluded. The result of this search was a collection of sets representing all amino acid sets found within the specified distance.

Next, this comprehensive set of amino acid sets was encoded into a binary vector using a hash function. This ensured that the presence of a specific amino acid set was consistently represented by setting the corresponding bit to 1. The chosen length for the binary vector was 16381. This length was chosen in order to retain the maximum amount of information by avoiding collisions in the hash function. Collisions happen when the hash function uses the same bit to represent two different sets of information, in this case, two different sets of amino acids.

In summary, the method involved defining a radius, identifying amino acids within that radius, encoding them into sets, aggregating these sets, and finally encoding the comprehensive set into a binary vector using a hash function. The resulting binary vector represented the presence or absence of specific amino acid sets within the defined distance (Figure 3.1).



(a) The figure illustrates the process of extracting protein fingerprints, although it should be noted that the illustration is not an exact representation.



(b) Here is a more detailed representation of how the amino acid sets are obtained. The Atom X belongs to the amino acid A1 and it is the center of the search. We can see that part of the amino acids A2, A4, and A6 fall within the circle with radius r and center in the Atom X. So it would be obtained a set containing the set {A1,A2,A4,A6}, and then the process presented in (a) would be followed

Figure 3.1: Extration of the protein fingerprints

3.2 Adapting machine learning methods for protein structure

The objective of this work is to develop a predictive model that integrates the structural information of proteins with molecular data to enable bioactivity predictions for multiple targets. Unlike traditional Quantitative Structure-Activity Relationship (QSAR) models, which focus on a single target, this study aims to leverage the protein structural information alongside the molecular information to enhance predictive capabilities.

Addressing the issue of the considerable size of the protein fingerprints, which are represented by an array of 16,381 bits, and the substantial size of the dataset, direct integration of protein and molecule fingerprints proved computationally challenging and hard to overcome given the resources and time available available. Consequently, Principal Component Analysis (PCA) was employed as a method for dimensionality reduction to mitigate this issue. This technique was the first choice to reduce the dimensionality of the protein fingerprints because not only it is widely used and known, but also is available in the Scikit-learn library [68]. An autoencoder was also explored for the purpose of dimensionality reduction, but the initial outcomes were not superior to those obtained using PCA. Given the complexity of the autoencoder technique, the decision was made to focus on enhancing the quality of fingerprints obtained through PCA, rather than further pursue the autoencoder approach. In one dataset variant, PCA was separately applied to both the protein and molecule fingerprints to reduce their dimensions before their subsequent merger. Conversely, in the other dataset variant, PCA was exclusively applied to the protein fingerprints. This strategy effectively amalgamates the structural information while managing the computational complexities associated with handling large-scale datasets.

3.3 Data retrieving and curation

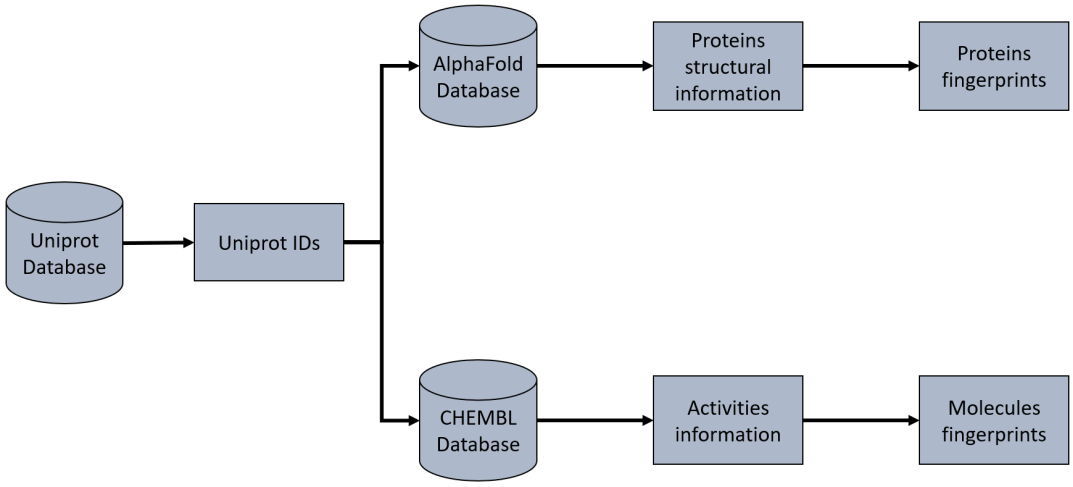


Figure 3.2: Overview of the workflow to the fingerprints

The data obtained for this study can be divided into two main categories: compound bioactiv-

ity and protein structures (Figure 3.2). To retrieve the compound bioactivity data, two websites, UniProt and CHEMBL, were used [57, 17]. A specific family of proteins was chosen, namely G-protein-coupled receptors (GPCRs). GPCRs are important because they are involved in a wide range of physiological processes in the body, including sensory perception, hormone regulation, neurotransmission, and immune response. GPCRs are the largest family of membrane proteins in the human genome, and they are also the most common targets for drugs [104]. In fact, approximately 30% of all FDA-approved drugs target GPCRs, making them a crucial area of focus for drug discovery and development [103]. Understanding the structure and function of GPCRs is therefore essential for advancing our knowledge of human biology and developing new treatments for a variety of diseases.

The UniProt website was employed to search for the desired proteins and obtain their corresponding identifiers (IDs). The search was filtered to include only human proteins that had undergone review (Swiss-Prot). These protein IDs were then utilized to search the CHEMBL database and retrieve the related target IDs. Subsequently, all the bioactivities of interest, specifically those of type Ki and IC50, associated with each target were collected.

Given the lack of standardization across results published in CHEMBL, certain rules were established to ensure the creation of a consistent table for each target, containing the following columns for each molecule: molecule ID on CHEMBL, pCHEMBL-Value (or value if the bioactivity type is inhibition), molecule SMILES, and observations. The rules followed were as follows:

1. Bioactivities with incorrect units for the given bioactivity type were excluded. All other bioactivity values were converted to nanomolar (nM), which was the predominant unit and the unit used in the pCHEMBL-Value calculation formula.

2. Bioactivities with a bioactivity relation of "=" and no value, or no bioactivity relation and no bioactivity comment, or an bioactivity comment indicating that the molecule was active but without a value, were excluded. Only active compounds have pCHEMBL-Values, so if a pCHEMBL-Value was available, it was used. If there were two pCHEMBL-Values available, their difference was checked, and if it exceeded one order of magnitude, the most recent pCHEMBL-Value was chosen; otherwise, the average of the pCHEMBL-Values was calculated. If more than two pCHEMBL-Values existed, the retrieved pCHEMBL-Value would be the average of all values.

3. If there were one or more bioactivities with pCHEMBL-Values but at least one bioactivity with "<" or "<=" in the bioactivity relation (indicating bioactivity below a certain threshold, but the exact value is unknown), this information was saved in the observations. In the case of multiple bioactivities with "<" or "<=" relations, the bioactivity with the lowest value was chosen.

4. For bioactivities marked as active and having values but no pCHEMBL-Value, the pCHEMBL-Value was calculated, and these bioactivities were treated similarly to bioactivities with ex-

plicit values. The pCHEMBL-Value formula is:

$$9 - \log_{10}(x) \tag{3.1}$$

5. If there were no bioactivities with pCHEMBL-Values but one or more bioactivities with "<" or "<=" in the bioactivity relation, the pCHEMBL-Value would be calculated for the bioactivity with the lowest value. The observations would note that this bioactivity relation was "<" or "<=" the recorded value.

6. No distinction was made for bioactivities with a bioactivity relation of ">" or ">=" (as these do not imply bioactivity, but rather inactivity) and bioactivities marked as not active. If these were the only types of bioactivities available, the most recent one was chosen, and either the bioactivity's comment or ">" or ">=" followed by the value was recorded in the observations. The corresponding pCHEMBL-Value was always set to 0.

After making the aforementioned adjustments, the pCHEMBL-Values were further transformed to ensure that they fell within the range of 0 to 1. This transformation was necessary due to the specific interpretations associated with different ranges of pCHEMBL-Values. Values below a certain threshold indicate that a compound is not active, as they require a high concentration for bioactivity, whereas values above a certain threshold correspond to very low concentrations of the compounds. This transformation is crucial for improving the predictive performance of the models. The transformation was performed using the following formula:

$$x = \begin{cases} 0 & \text{if pCHEMBL-Value} \leq 5 \\ 1 & \text{if pCHEMBL-Value} \geq 9 \\ \frac{\text{pCHEMBL-Value} - 5}{4} & \text{otherwise} \end{cases} \tag{3.2}$$

It is important to highlight that a pCHEMBL-Value of 5 corresponds to a compound concentration of 10,000 nM, indicating a relatively high concentration required for bioactivity. On the other hand, a pCHEMBL-Value of 9 corresponds to a concentration of 1 nM, indicating a very low concentration of the compounds. The transformation of pCHEMBL-Values, as described earlier, was employed to rescale the values within the range of 0 to 1. This normalization facilitates improved interpretation and utilization of the pCHEMBL-Values in subsequent modeling processes.

In this work, Morgan Fingerprints were selected as the type of fingerprints to be utilized. These fingerprints were generated from the molecular structures' SMILES representation. To perform this task, the RDKit toolkit, an open-source cheminformatics toolkit [31], was employed. RDKit provided the required functionality for generating Morgan Fingerprints from the molecular structures.

The second category of data retrieved in this study pertains to the structural information of proteins. The PDB files were obtained from the AlphaFold Database, and the process described previously was followed to derive the corresponding protein fingerprints from the protein structures.

3.4 Baseline models

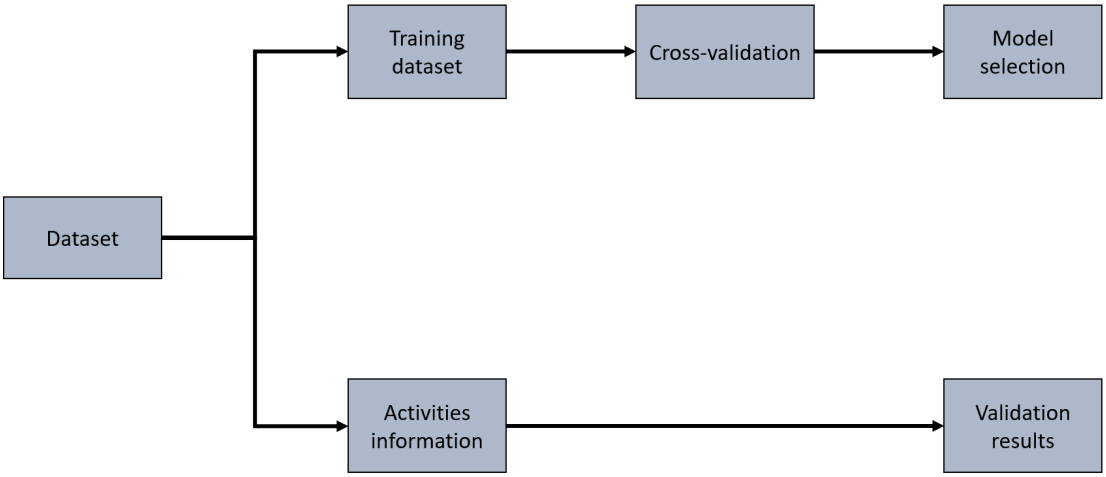


Figure 3.3: Workflow overview for obtaining baseline results. This figure provides an overview of the workflow followed to obtain the baseline results. The process was applied to each target and with Ki bioactivities. The dataset used in the workflow includes molecule fingerprints along with their corresponding bioactivity values.

To compare the results of the approach being investigated in this study, a classic QSAR approach was employed for each target that met the specified criteria. The workflow for this approach was implemented using the Scikit-learn library [68], and the following tasks were performed (Figure 3.3):

- Dataset Split: The dataset was divided into training and validation sets, with the training set comprising 80% of the data. Targets with less than 40 bioactivities were excluded from the analysis due to insufficient information.

- Model Evaluation: The models were evaluated using 5-fold cross-validation to identify the optimal parameters that yielded the best results on the training set. The evaluation metrics used were Root Mean Squared Error (RMSE) and Explained Variance.

- Prediction and Comparison: Once the best parameters were determined, the models were used to make predictions on the validation set, and the results were recorded. The models used were Random Forests (RF) [10], Support Vector Machines (SVM) for Regression [18] and Gradient Boosting Methods (GBM) [28]. The performance of the best model was considered for subsequent comparisons.

By following this workflow, the classic QSAR approach was applied to each target, allowing for a comparison with the results obtained from the tested approach in this study.

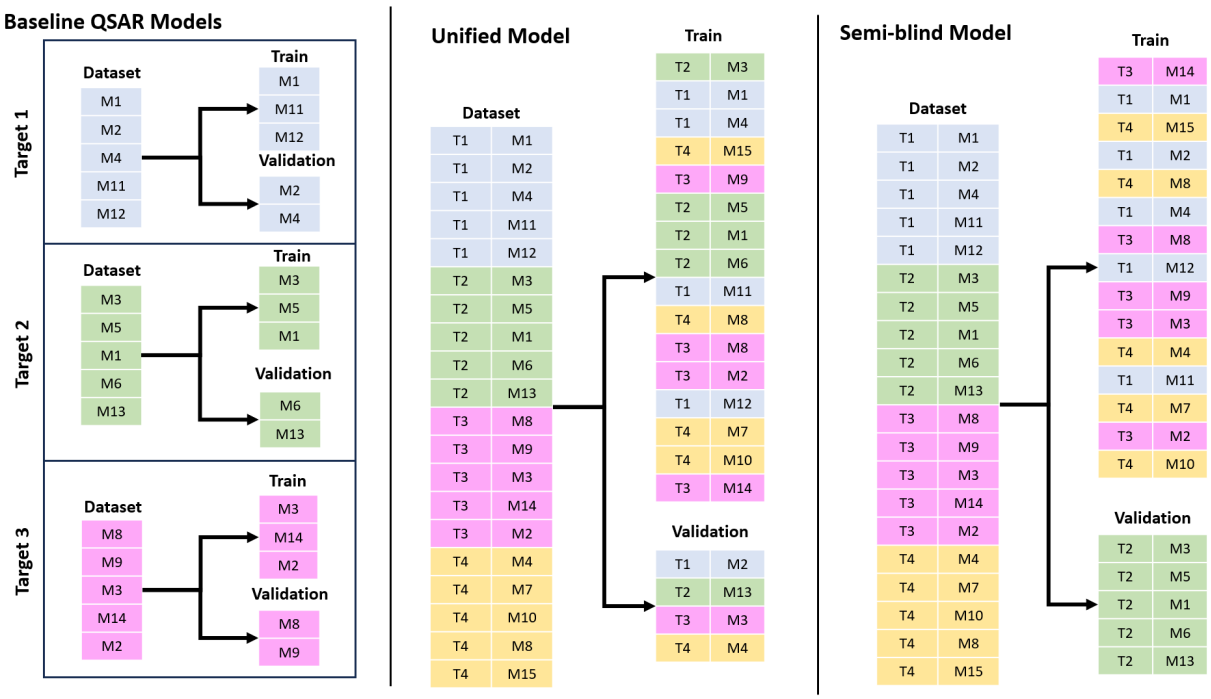


Figure 3.4: The representation of the different modeling techniques used in this work

3.5 Global and individual models

Besides the baseline QSAR technique described used for comparison, the new approach was tested using three different techniques:

1. Unified model: The dataset was randomly divided into training and test sets, with the training set comprising 80% of the data and the test set 20%, without any specific criteria regarding the presence or absence of target-related fingerprints in either set.

2. Blind model: one or more targets would be randomly selected from the dataset to be the testing set and the remaining targets constitute the training set. This ensures that the model makes predictions only for targets not present in the training set, simulating the procedure that would take place with a target never tested before.

3. Semi-Blind model: Sixteen pairs of targets were selected based on their similarity. For each pair, the target with the most bioactivities was included in the training set, while the other target was assigned to the test set. The selection ensured that a target present in the training set did not appear in the testing set, although it could be repeated in the training set. The same principle applied to targets in the testing set. The Tanimoto coefficient, computed using the unreduced fingerprints, was employed to measure the similarity between targets.

Initially, the Unified model was used to assess the feasibility of predicting bioactivities for multiple targets solely based on the structural information of molecules and proteins. As presented in the subsequent sections, the obtained results were promising, indicating that the models were

capable of correlating this information with bioactivities. Encouraged by these findings, it was decided to try the Blind model approach, however, this did not produce good results and was abandoned. From this attempt, the Semi-Blind model was developed to try to correct a possible weakness of the Blind model, which was the fact that, by choosing the target to be tested randomly, this target might not have a sufficiently similar target in the training set.

Due to the dataset's size, the cross-validation approach utilized for the baseline models was not adopted. Instead, all models employed were Random Forests (RF) with 200 components. To ensure reliable and robust results, each model was executed ten times for each distinct dataset.

3.6 Dataset

This study involved the utilization of twelve datasets to investigate the GPCR family of targets. The datasets were constructed based on the available information on two types of bioactivities: Ki and IC50, as well as the radius applied in the extraction of protein fingerprints: 5 Å and 7 Å. Specifically, four datasets were exclusively composed of Ki bioactivities, while another four datasets exclusively contained IC50 bioactivities. For each type of bioactivity, two datasets were created—one with both protein and molecule fingerprints dimensionally reduced using a radius of 5 Å, and another with a radius of 7 Å. Additionally, two datasets were generated with only protein fingerprints reduced using the respective radii.

Furthermore, four additional datasets were formed by merging all available information. In cases where both Ki and IC50 bioactivity values were present for the same target, the decision was made to exclude the IC50 bioactivity and retain only the Ki bioactivity for further analysis and modeling.

A total of 197 targets with available Ki bioactivities and 246 targets with IC50 bioactivities were included in this study. When considering both types of bioactivities together, the total number of targets was 253.

Within the Ki group, a total of 69,879 different molecules and 141,225 different bioactivities (including the combinations of proteins with their respective molecules) were recorded. In the IC50 group, there were a total of 77,518 different molecules and 101,865 different bioactivities.

When merging the Ki and IC50 datasets, a total of 232,666 bioactivities were included, involving 143,450 unique molecules and the aforementioned 253 targets.

3.7 Pairs

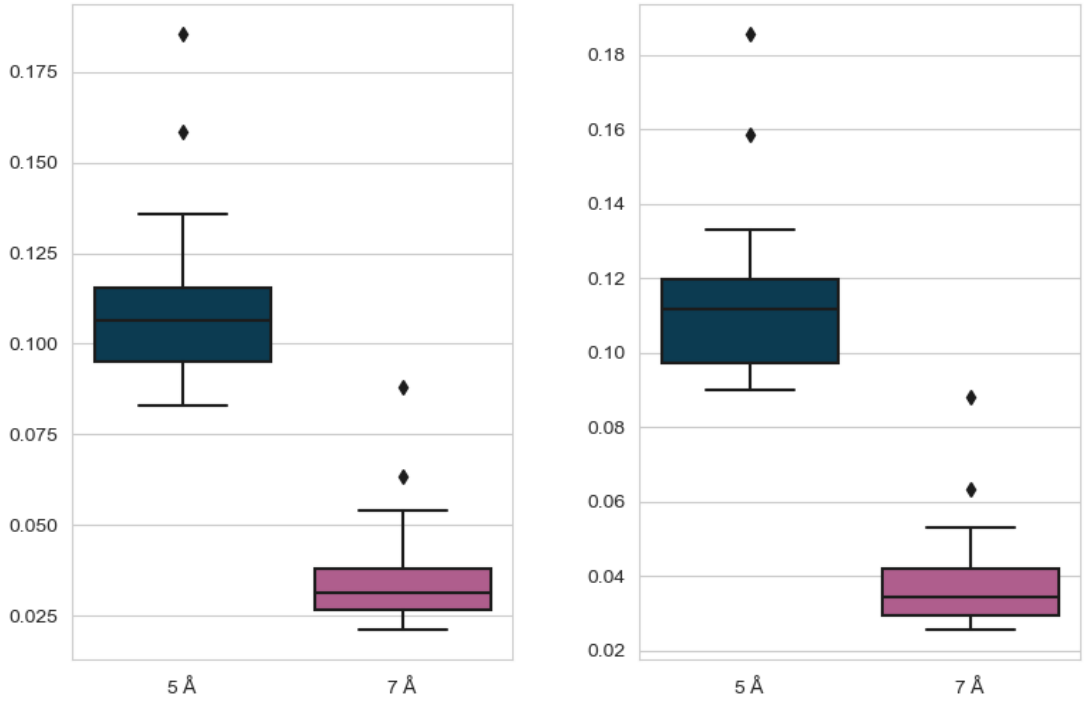
Besides the aforementioned criteria, an additional threshold was applied to exclude targets with a low number of bioactivities, ensuring more reliable results. The specific threshold varied depending on the type of bioactivities present in each dataset:

- For the Ki dataset, each target was required to have a minimum of 100 Ki bioactivities.
- For the IC50 dataset, each target was required to have a minimum of 100 IC50 bioactivities.

- For the IC50+Ki dataset, each target was required to have a minimum of either 100 IC50 bioactivities or 100 Ki bioactivities.

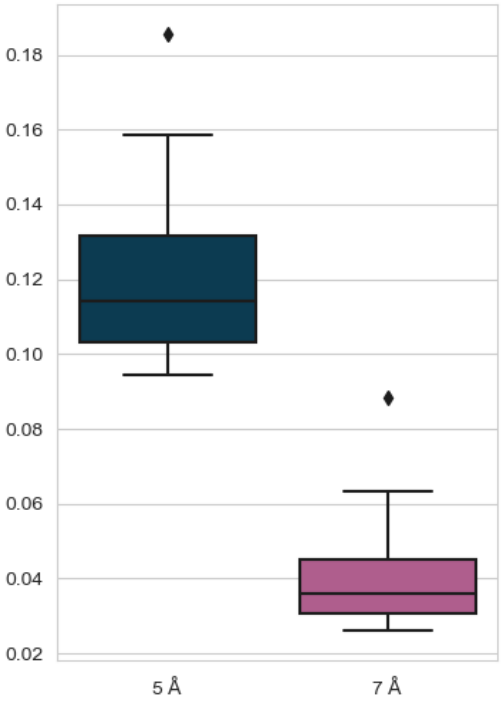
In all the datasets, the number of unique training and testing targets was 15. Although some target pairs were selected for multiple datasets, the specific pairings were different for each dataset. As shown in Figure 3.5, there is little variation in the set of target pairs across different bioactivity types. The main factor influencing the dissimilarity between pair elements is the distance used in fingerprint generation. This makes sense since bigger a radius is able to capture sets with more amino acids increasing the number of different amino acids. Figure 3.5 presents the similarity values between the elements of each pair, categorized by bioactivity type and fingerprint distance.

The list of unique targets used for testing in each combination of bioactivity type and distance used for generating the protein fingerprints is presented in Table 3.1 and Table 3.2 (for Ki), Table 3.3 and Table 3.4 (for IC50), and Table 3.5 and Table 3.6 (for IC50+Ki). As mentioned, the target used for testing always had fewer bioactivities available in the dataset than the other element of the pair used in the training set. For datasets in which the bioactivity type is IC50+Ki, the decision on which target goes into training or testing was based on the sum of the IC50 and Ki bioactivities. For example, a target could have fewer Ki bioactivities than the other target but still be used for training.



(a) Ki

(b) IC50



(c) IC50+Ki

Figure 3.5: Tanimoto coefficient of the pairs for each bioactivity type

Testing pair	Testing pair name	N activities	Training pair name	Similarity
O43613	Orexin/Hypocretin receptor type 1	1574	Orexin receptor type 2	0.110
P08173	Muscarinic acetylcholine receptor M4	853	Muscarinic acetylcholine receptor M2	0.159
P08912	Muscarinic acetylcholine receptor M5	780	Muscarinic acetylcholine receptor M1	0.115
P13945	Beta-3 adrenergic receptor	335	Beta-1 adrenergic receptor	0.087
P18089	Alpha-2B adrenergic receptor	558	Alpha-2A adrenergic receptor	0.105
P18825	Alpha-2C adrenergic receptor	683	Alpha-2A adrenergic receptor	0.116
P21918	D(1B) dopamine receptor	526	D(1A) dopamine receptor	0.136
P25024	C-X-C chemokine receptor type 1	102	C-X-C chemokine receptor type 2	0.186
P28222	5-hydroxytryptamine receptor 1B	1040	5-hydroxytryptamine receptor 1D	0.096
P28335	5-hydroxytryptamine receptor 2C	2633	5-hydroxytryptamine receptor 2A	0.094
P30872	Somatostatin receptor type 1	185	Somatostatin receptor type 4	0.083
P32246	C-C chemokine receptor type 1	160	C-C chemokine receptor type 3	0.096
P35462	D(3) dopamine receptor	5106	D(2) dopamine receptor	0.098
P41145	Kappa-type opioid receptor	3907	Delta-type opioid receptor	0.090
P51681	C-C chemokine receptor type 5	160	C-C chemokine receptor type 2	0.116

Table 3.1: List of targets for the datasets with Ki activities and with protein fingerprints generate with radius=5

Testing pair	Testing pair name	N activities	Training pair name	Similarity
P25024	C-X-C chemokine receptor type 1	102	C-X-C chemokine receptor type 2	0.088
P08173	Muscarinic acetylcholine receptor M4	853	Muscarinic acetylcholine receptor M2	0.063
P21918	D(1B) dopamine receptor	526	D(1A) dopamine receptor	0.054
P51681	C-C chemokine receptor type 5	160	C-C chemokine receptor type 2	0.042
P08912	Muscarinic acetylcholine receptor M5	780	Muscarinic acetylcholine receptor M3	0.037
P41145	Kappa-type opioid receptor	3907	Delta-type opioid receptor	0.035
P32246	C-C chemokine receptor type 1	160	C-C chemokine receptor type 3	0.034
O43613	Orexin/Hypocretin receptor type 1	1574	Orexin receptor type 2	0.032
P28335	5-hydroxytryptamine receptor 2C	2633	5-hydroxytryptamine receptor 2A	0.031
P18825	Alpha-2C adrenergic receptor	683	Alpha-2A adrenergic receptor	0.028
P35462	D(3) dopamine receptor	5106	D(2) dopamine receptor	0.027
P18089	Alpha-2B adrenergic receptor	558	Alpha-2A adrenergic receptor	0.026
P28222	5-hydroxytryptamine receptor 1B	1040	5-hydroxytryptamine receptor 1D	0.026
P30872	Somatostatin receptor type 1	185	Somatostatin receptor type 4	0.022
P35346	Somatostatin receptor type 5	322	Somatostatin receptor type 3	0.021

Table 3.2: List of targets for the datasets with Ki activities and with protein fingerprints generate with radius=7

Testing pair	Testing pair name	N activities	Training pair name	Similarity
P25024	C-X-C chemokine receptor type 1	286	C-X-C chemokine receptor type 2	0.186
P08173	Muscarinic acetylcholine receptor M4	319	Muscarinic acetylcholine receptor M2	0.159
P34947	G protein-coupled receptor kinase 5	209	G protein-coupled receptor kinase 6	0.133
Q13255	Metabotropic glutamate receptor 1	873	Metabotropic glutamate receptor 5	0.131
P41597	C-C chemokine receptor type 2	1715	C-C chemokine receptor type 5	0.116
P18825	Alpha-2C adrenergic receptor	137	Alpha-2A adrenergic receptor	0.116
P08912	Muscarinic acetylcholine receptor M5	329	Muscarinic acetylcholine receptor M1	0.115
Q14832	Metabotropic glutamate receptor 3	156	Metabotropic glutamate receptor 2	0.113
O43614	Orexin receptor type 2	1821	Orexin/Hypocretin receptor type 1	0.110
P18089	Alpha-2B adrenergic receptor	142	Alpha-2A adrenergic receptor	0.105
P35462	D(3) dopamine receptor	372	D(2) dopamine receptor	0.098
P28222	5-hydroxytryptamine receptor 1B	297	5-hydroxytryptamine receptor 1D	0.096
P51677	C-C chemokine receptor type 3	769	C-C chemokine receptor type 1	0.096
P28335	5-hydroxytryptamine receptor 2C	1069	5-hydroxytryptamine receptor 2A	0.094
P41145	Kappa-type opioid receptor	1196	Delta-type opioid receptor	0.090

Table 3.3: List of targets for the datasets with IC50 activities and with protein fingerprints generate with radius=5

Testing pair	Testing pair name	N activities	Training pair name	Similarity
P25024	C-X-C chemokine receptor type 1	286	C-X-C chemokine receptor type 2	0.088
P08173	Muscarinic acetylcholine receptor M4	319	Muscarinic acetylcholine receptor M2	0.063
P34947	G protein-coupled receptor kinase 5	209	G protein-coupled receptor kinase 6	0.053
Q14832	Metabotropic glutamate receptor 3 (mGluR3)	156	Metabotropic glutamate receptor 2	0.043
P41597	C-C chemokine receptor type 2	1715	C-C chemokine receptor type 5	0.042
Q13255	Metabotropic glutamate receptor 1 (mGluR1)	873	Metabotropic glutamate receptor 5	0.042
P08912	Muscarinic acetylcholine receptor M5	329	Muscarinic acetylcholine receptor M3	0.037
P41145	Kappa-type opioid receptor	1196	Delta-type opioid receptor	0.035
P51677	C-C chemokine receptor type 3	769	C-C chemokine receptor type 1	0.034
O43614	Orexin receptor type 2	1821	Orexin/Hypocretin receptor type 1	0.032
P28335	5-hydroxytryptamine receptor 2C	1069	5-hydroxytryptamine receptor 2A	0.031
P18825	Alpha-2C adrenergic receptor	137	Alpha-2A adrenergic receptor	0.028
P35462	D(3) dopamine receptor	372	D(2) dopamine receptor	0.027
P18089	Alpha-2B adrenergic receptor	142	Alpha-2A adrenergic receptor	0.026
P28222	5-hydroxytryptamine receptor 1B	297	5-hydroxytryptamine receptor 1D	0.026

Table 3.4: List of targets for the datasets with IC50 activities and with protein fingerprints generate with radius=7

Testing pair	Testing pair name	N activities	Training pair name	Similarity
O43613	Orexin/Hypocretin receptor type 1	3422	Orexin receptor type 2	0.110
P08173	Muscarinic acetylcholine receptor M4	1172	Muscarinic acetylcholine receptor M2	0.159
P08912	Muscarinic acetylcholine receptor M5	1109	Muscarinic acetylcholine receptor M1	0.115
P18089	Alpha-2B adrenergic receptor	700	Alpha-2A adrenergic receptor	0.105
P18825	Alpha-2C adrenergic receptor	820	Alpha-2A adrenergic receptor	0.116
P21918	D(1B) dopamine receptor	543	D(1A) dopamine receptor	0.136
P25024	C-X-C chemokine receptor type 1	388	C-X-C chemokine receptor type 2	0.186
P28222	5-hydroxytryptamine receptor 1B	1337	5-hydroxytryptamine receptor 1D	0.096
P28335	5-hydroxytryptamine receptor 2C	3702	5-hydroxytryptamine receptor 2A	0.094
P32246	C-C chemokine receptor type 1	974	C-C chemokine receptor type 3	0.096
P35462	D(3) dopamine receptor	5478	D(2) dopamine receptor	0.098
P41597	C-C chemokine receptor type 2	2005	C-C chemokine receptor type 5	0.116
P43250	G protein-coupled receptor kinase 6	228	G protein-coupled receptor kinase 5	0.133
Q13255	Metabotropic glutamate receptor 1 (mGluR1)	968	Metabotropic glutamate receptor 5	0.131
Q14832	Metabotropic glutamate receptor 3 (mGluR3)	241	Metabotropic glutamate receptor 2	0.113

Table 3.5: List of targets for the datasets with IC50+Ki activities and with protein fingerprints generate with radius=5

Testing pair	Testing pair name	N activities	Training pair name	Similarity
O43613	Orexin/Hypocretin receptor type 1	3422	Orexin receptor type 2	0.032
P08173	Muscarinic acetylcholine receptor M4	1172	Muscarinic acetylcholine receptor M2	0.063
P08912	Muscarinic acetylcholine receptor M5	1109	Muscarinic acetylcholine receptor M1	0.037
P18089	Alpha-2B adrenergic receptor	700	Alpha-2A adrenergic receptor	0.026
P18825	Alpha-2C adrenergic receptor	820	Alpha-2A adrenergic receptor	0.028
P21918	D(1B) dopamine receptor	543	D(1A) dopamine receptor	0.054
P25024	C-X-C chemokine receptor type 1	388	C-X-C chemokine receptor type 2	0.0.08
P28335	5-hydroxytryptamine receptor 2C	3702	5-hydroxytryptamine receptor 2A	0.031
P32246	C-C chemokine receptor type 1	974	C-C chemokine receptor type 3	0.034
P35462	D(3) dopamine receptor	5478	D(2) dopamine receptor	0.027
P41145	Kappa-type opioid receptor (K-OR-1) (KOR-1)	5103	Delta-type opioid receptor (D-OR-1) (DOR-1)	0.035
P41597	C-C chemokine receptor type 2	2005	C-C chemokine receptor type 5	0.042
P43250	G protein-coupled receptor kinase 6	228	G protein-coupled receptor kinase 5	0.053
Q13255	Metabotropic glutamate receptor 1 (mGluR1)	968	Metabotropic glutamate receptor 5	0.042
Q14832	Metabotropic glutamate receptor 3 (mGluR3)	241	Metabotropic glutamate receptor 2	0.043

Table 3.6: List of targets for the datasets with IC50+Ki activities and with protein fingerprints generate with radius=7

Chapter 4

Results

To facilitate a comprehensive comparison of various models' outcomes, assessments were confined to the targets that were shared between the baseline models and other models. Such confinement stemmed from the aforementioned criterion that only targets with a minimum of 40 bioactivities were considered. A select group of targets, exhibiting a Ratio of Explained Variance (RVE) below -1, were eliminated due to their distinct outlier nature. Consequently, for datasets containing Ki bioactivities, outcomes were garnered for 144 targets using radii of both 5 Å and 7 Å. In the IC50 bioactivity datasets, results were obtained for 189 and 190 targets using the 5 Å and 7 Å radii respectively. Meanwhile, for the IC50+Ki bioactivity datasets, results were amassed for 201 targets across both radii.

4.1 Ki datasets

Model	Training set size	Testing set size	RMSE	RVE
Baseline QSAR	112217	28122	0.167	0.683
Unified model	112248	28091	0.167	0.681
Semi-blind model	121737	18602	0.301	0.208

Table 4.1: Results and training and testing set sizes for all models - Ki with radius=5Å and without PCA

Model	Training set size	Testing set size	RMSE	RVE
Baseline QSAR	112217	28122	0.167	0.683
Unified model	112197	28043	0.208	0.518
Semi-blind model	121638	18602	0.296	0.209

Table 4.2: Results and training and testing set sizes for all models - Ki with radius=5Å and with PCA

Due to the multiplicity of datasets available, emphasis is predominantly directed towards results garnered from the Ki datasets, specifically with a radius of 5Å. Typically, this bioactivity type and radius combination has consistently yielded superior outcomes. The forthcoming sections will

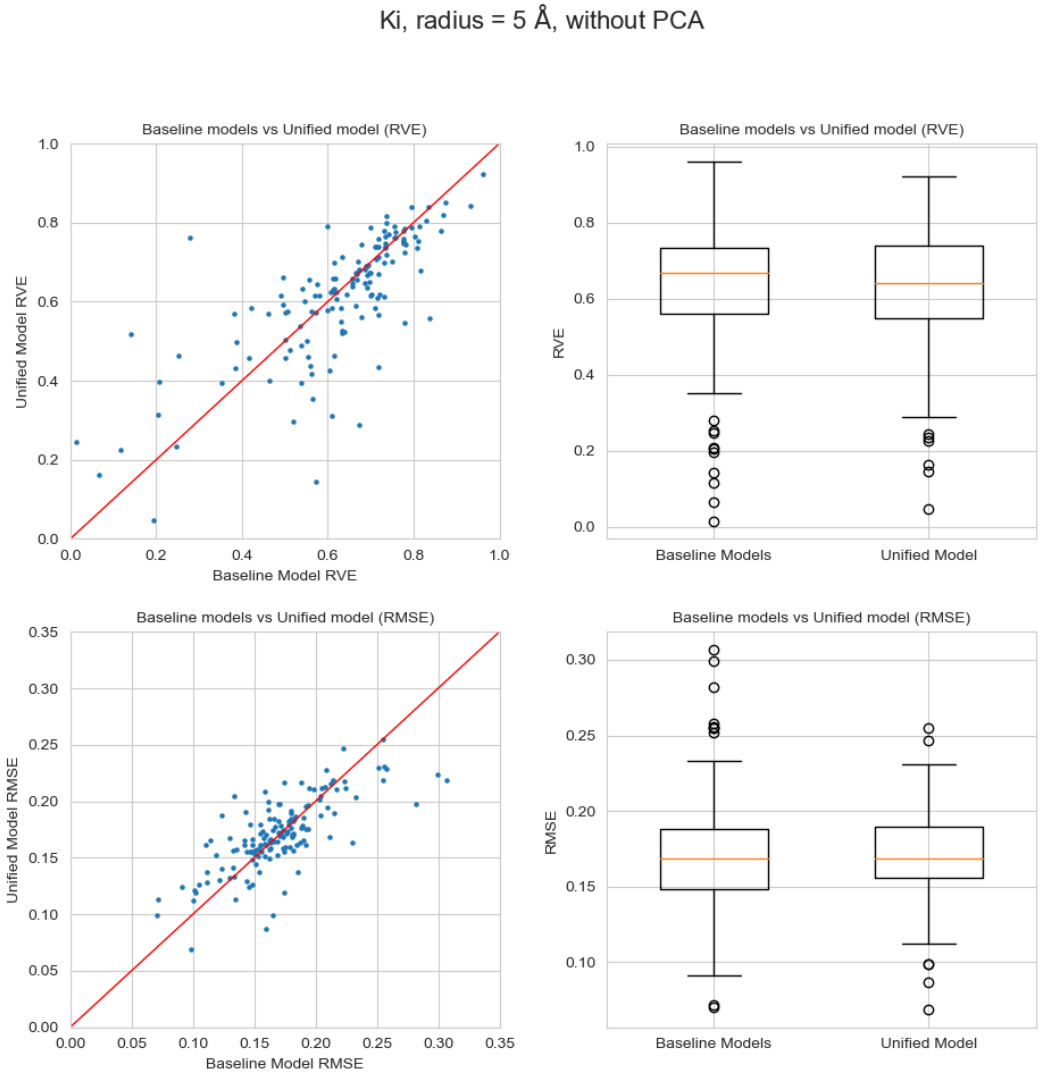


Figure 4.1: RMSE and RVE validation results for the Ki dataset with radius=5Å without PCA

present and dissect the results, both with and without the application of Principal Component Analysis (PCA) for molecular fingerprint dimensionality reduction. This distinction is pivotal as the disparities observed between these two datasets bear resemblance to patterns discerned in other datasets.

The cumulative results of the Baseline model were deduced as the weighted average of results from individual models. The Unified model, however, necessitated multiple runs owing to the potentially fluctuating proportions of ligands for various targets across training and testing datasets. Summatively, the Unified model underwent 10 iterations for each dataset, with the final outcome being an average of results across all targets. The Semi-Bind model employed a similar strategy, but with an added layer of complexity: all bioactivities of test targets were allocated to the validation set, and hence, the conclusive results reflect the weighted average of individual target results.

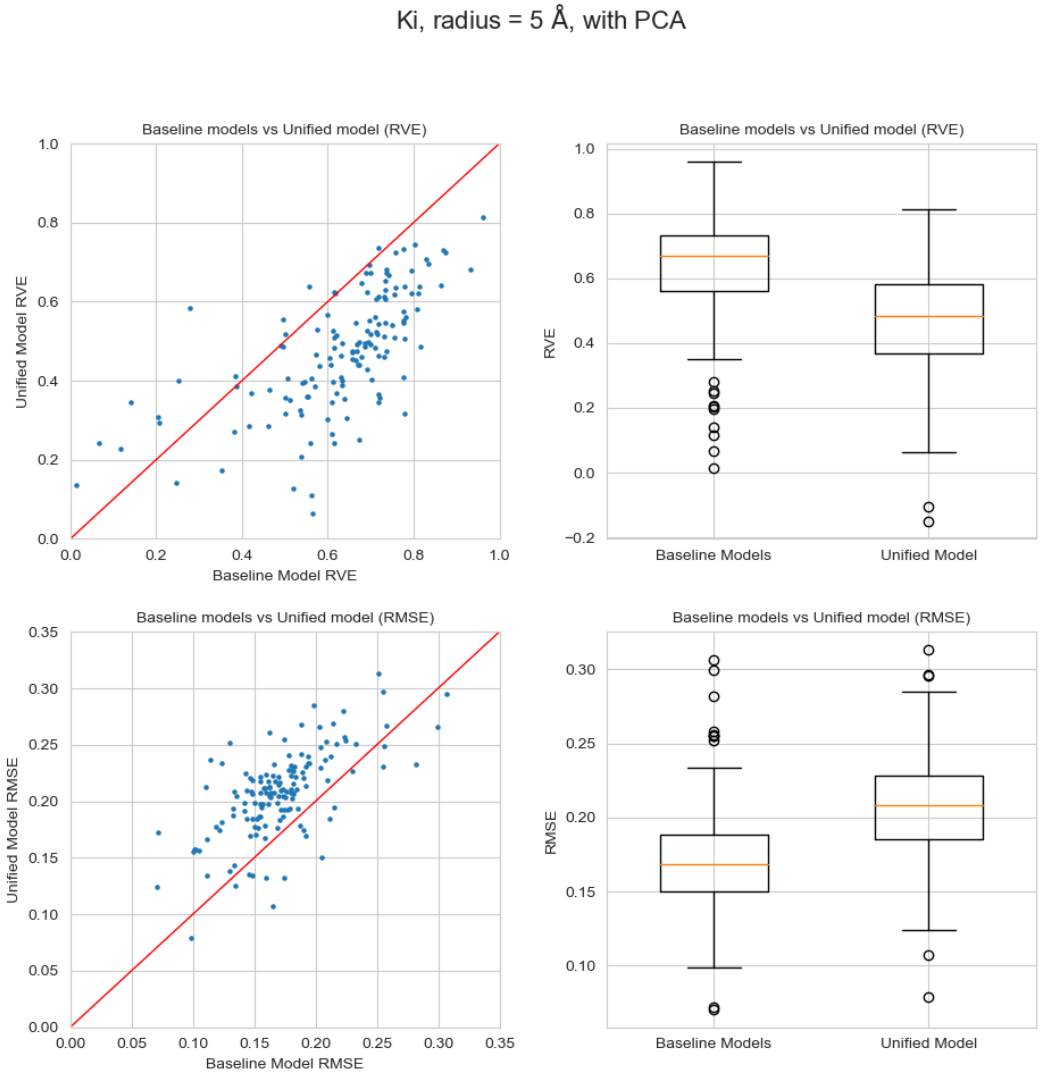


Figure 4.2: RMSE and RVE validation results for the Ki dataset with radius=5Å with PCA

Observations from the aggregated results for the Ki dataset (radius=5Å without molecular fingerprint reduction, as referenced in Table 4.1) indicate marginal differences between the Baseline and Unified models. A visual representation in Figure 4.1 shows that most results closely align with the 45º line for both RMSE and RVE metrics. A few isolated target results noticeably favor one model over the other, yet without offering a consistent advantage to either model. An imperative takeaway is that while certain targets seemingly benefit from the Baseline model, an almost equal number of targets prefer the Unified model. Boxplot observations further underscore this by highlighting the presence of a greater number of low-performing outliers within baseline models, suggesting that expanding the information used to train the models by adding information about other targets might have a positive impact on the capacity of the models to make predictions for targets difficult to model for QSAR models. Remarkably, the outcomes are commendable, especially considering the all-inclusive data utilization without explicit target distinctions. Such

	Baseline Models		Unified Model		Semi-blind Model	
Target ID	RMSE	RVE	RMSE	RVE	RMSE	RVE
O43613	0.163	0.668	0.163	0.671	0.280	0.034
P08173	0.147	0.776	0.155	0.779	0.222	0.555
P08912	0.159	0.618	0.162	0.658	0.238	0.284
P13945	0.142	0.709	0.165	0.585	0.249	0.098
P18089	0.169	0.500	0.165	0.570	0.216	0.242
P18825	0.187	0.616	0.164	0.698	0.285	0.166
P21918	0.154	0.756	0.137	0.791	0.280	0.191
P25024	0.136	0.608	0.159	0.581	0.263	-0.140
P28222	0.182	0.741	0.172	0.771	0.238	0.565
P28335	0.173	0.658	0.174	0.646	0.299	0.026
P30872	0.190	0.696	0.163	0.663	0.243	0.445
P32246	0.192	0.252	0.161	0.461	0.261	0.129
P35462	0.171	0.695	0.170	0.694	0.359	0.164
P41145	0.174	0.734	0.168	0.748	0.304	0.297
P51681	0.255	0.116	0.223	0.204	0.427	-0.037

Table 4.3: Comparison of results for all 15 validation targets of the semi-blind model for the Ki dataset with radius=5Å and without PCA

findings bolster the notion that this methodology can indeed architect models proficient in delineating structural ligand-target relationships, producing results on par with extant QSAR models.

In contrast, when evaluating the aggregated results for the Ki dataset at radius=5Å with molecular fingerprint reduction (refer to Table 4.2), the outcomes, although still commendable, depict distinct deviations. These observations insinuate inherent challenges associated with capturing intricate ligand-target relationships when ligand fingerprints undergo reduction.

Turning to the Semi-Blind model results for the Ki dataset at radius=5Å without fingerprint reduction, the outcomes weren't as optimal, yet predominantly remained positive. The aggregate RVE, approximately 0.2, reflects the model's adeptness at discerning relationships between ligand and target structural data in relation to their bioactivities. A more granular perspective, presented in Table 4.3, sheds light on results acquired through varied approaches (Baseline models, Unified model, and Semi-Blind model) for each of the 15 test targets. Notably, targets P08173 and P28222 merit special mention due to their RVE exceeding 0.5.

Conversely, when examining the Ki dataset at radius=5Å with fingerprint reduction, the Semi-Blind model showcased a slight improvement, particularly in the aggregate RMSE. An in-depth examination of individual results, as illustrated in Table 4.3, reveals the model's proficiency in predicting bioactivities for three targets (P08173, P28222, P30872) with an RVE surpassing 0.5—a slight uptick from the two targets (P08173 and P28222) in the preceding dataset.

For the Ki datasets employing a radius of 7Å, the Unified model presented consistent outcomes irrespective of PCA application for molecular fingerprint dimensionality reduction. Yet, the Semi-Blind model highlighted a decrement in predictive accuracy—a comprehensive breakdown of which is available in the appendix.

	Baseline Models		Unified Model		Semi-blind Model	
Target ID	RMSE	RVE	RMSE	RVE	RMSE	RVE
O43613	0.163	0.668	0.204	0.493	0.288	0.059
P08173	0.147	0.776	0.169	0.735	0.221	0.576
P08912	0.159	0.618	0.168	0.623	0.270	0.357
P13945	0.142	0.709	0.191	0.486	0.244	0.222
P18089	0.169	0.500	0.177	0.519	0.204	0.358
P18825	0.187	0.616	0.178	0.624	0.267	0.214
P21918	0.154	0.756	0.186	0.621	0.280	0.270
P25024	0.136	0.608	0.205	0.348	0.279	-0.248
P28222	0.182	0.741	0.206	0.667	0.243	0.556
P28335	0.173	0.658	0.211	0.476	0.283	0.103
P30872	0.190	0.696	0.175	0.695	0.249	0.536
P32246	0.192	0.252	0.169	0.402	0.235	0.045
P35462	0.171	0.695	0.217	0.497	0.338	0.118
P41145	0.174	0.734	0.204	0.630	0.315	0.242
P51681	0.255	0.116	0.233	0.223	0.394	-0.040

Table 4.4: Comparison of results for all 15 validation targets of the semi-blind model for the Ki dataset with radius=5Å and with PCA

4.2 Comparative analysis of Ki and IC50 datasets

The datasets dedicated to Ki bioactivities consistently exhibited superior performance. At first glance, one might assume that datasets singularly based on IC50 bioactivities would yield results akin to those of the Ki datasets due to the homogeneity in their bioactivity profiles. Contrarily, blended datasets, which combined both Ki and IC50 bioactivities, consistently delivered enhanced outcomes, compared to the datasets containing only IC50 bioactivities.

Several plausible explanations could elucidate these observed discrepancies. One could surmise that the relational dynamics between targets and ligands in these merged datasets might be intrinsically more intricate to model. Alternatively, the methodologies employed in data processing might be inherently more congruent with Ki bioactivities than with IC50.

This analytical effort focused primarily on the RVE metrics, as higher RVE values suggest that a model is proficient in capturing the fundamental associations between the structural intricacies of proteins and ligands. It is important to note that the RMSE outcomes, observed across a range of datasets with diverse bioactivity types, did not exhibit significant variations substantial enough to necessitate special attention.

During this research, fingerprints were generated using two distinct radii. It became apparent that the composition of amino acid sets varied significantly based on the chosen radius, particularly when using a radius of 7Å, which encompassed an expanded amino acid subset. Given this variability, the selected radius in fingerprint generation was expected to have a substantial impact on a model's ability to discern relationships between fingerprints and bioactivity, thus influencing its predictive accuracy. Nevertheless, with a few exceptions, notably in the Ki datasets for

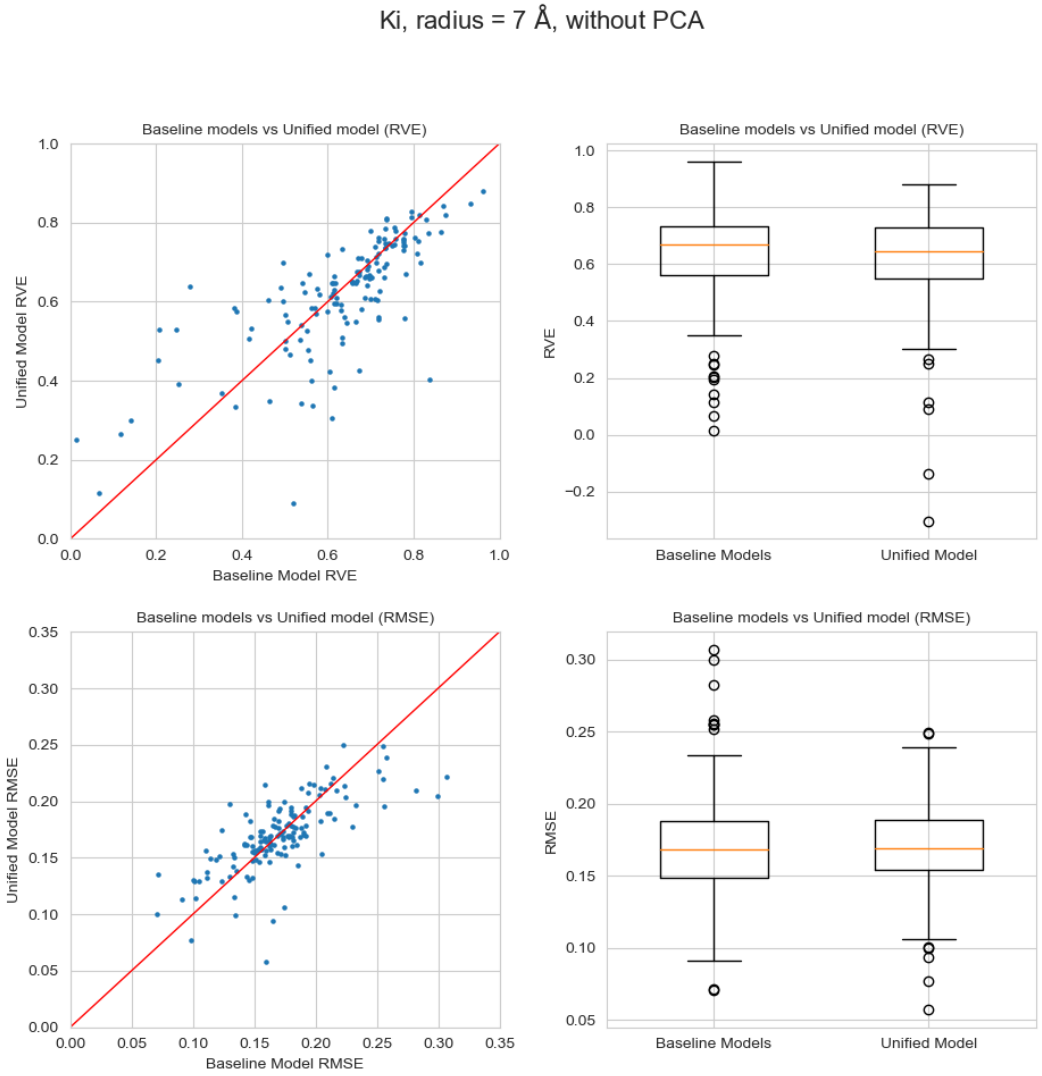


Figure 4.3: RMSE and RVE validation results for the Ki dataset with radius=7Å without PCA

the Semi-Blind model, the anticipated divergence in performance based on radius choice was not clearly evident. Particularly in the Semi-Blind models, datasets lacking testing set target information in the training phase exhibited this variability, suggesting a potential role of proximal amino acids in mediating interactions between proteins and ligands.

An additional observation of interest pertains to the performance of the Unified model within datasets where ligand fingerprints underwent dimensionality reduction via PCA. In such instances, the RVE values consistently plummeted, and the RMSE values surged noticeably. Conversely, Unified model outcomes for other datasets mirrored the results exhibited by the Baseline models. Interestingly, the Semi-Blind model's performance hinted towards an inverse trend, albeit the variations were more nuanced.

This research has resulted in the development and comprehensive testing of an innovative

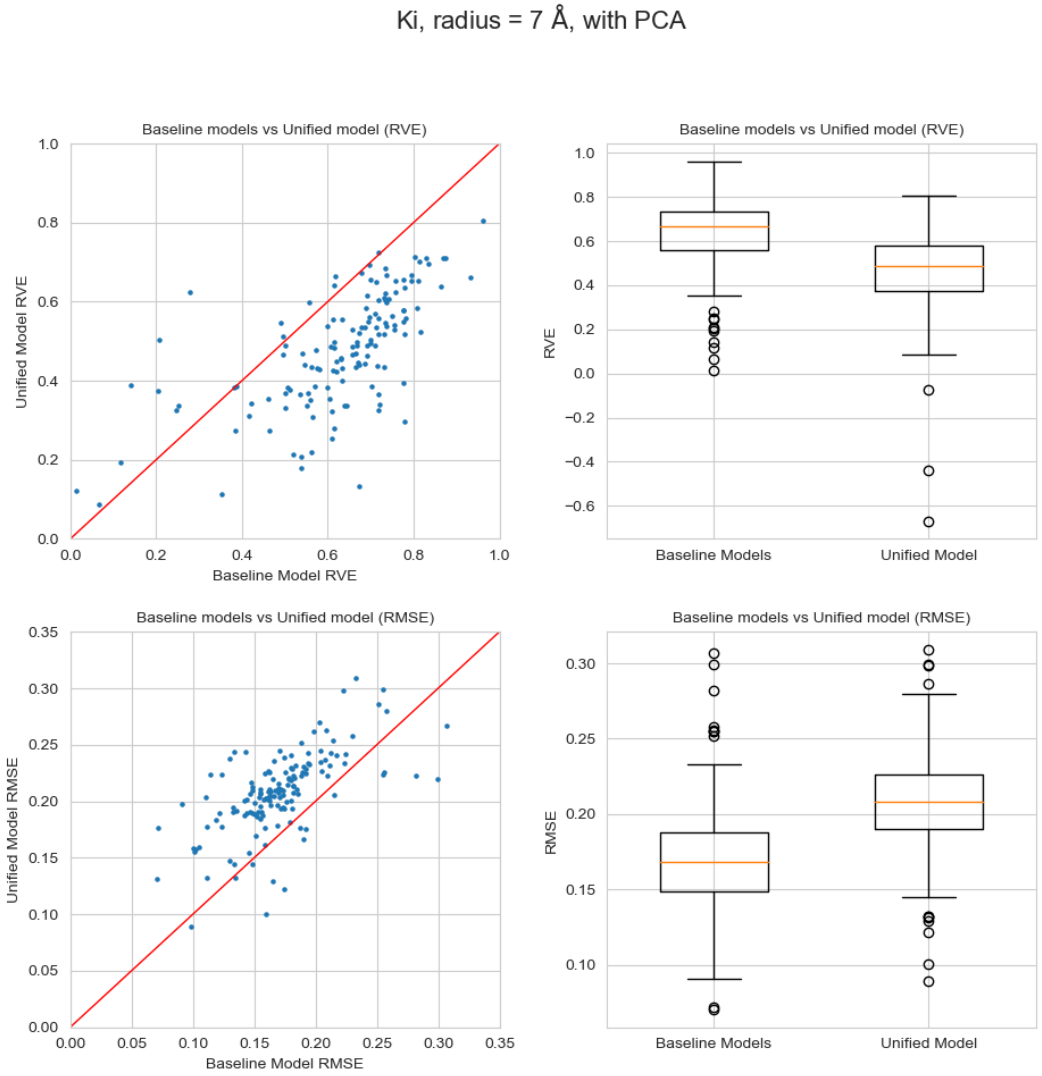


Figure 4.4: RMSE and RVE validation results for the Ki dataset with radius=7Å with PCA

methodology. The primary goal was to incorporate protein structural data into machine learning models, addressing a crucial limitation in QSAR-based approaches—constrained predictions for singular or a few correlated targets, while maintaining the methodology's inherent simplicity. This specific constraint poses a significant hurdle to QSAR approaches, demanding substantial initial data from laboratory assays and an ample number of assays for model training. The central idea was to utilize structural data of targets within the model, potentially enabling predictions of ligand bioactivities for previously untested targets. This shift in approach aims to transform the previous ligand-centric framework into a structure-driven one. The essential structural data, extracted from AlphaFold, facilitates a straightforward workflow, directing 3D information for seamless integration into machine learning models.

The outcomes derived from this endeavor show promise, indicating a methodology with poten-

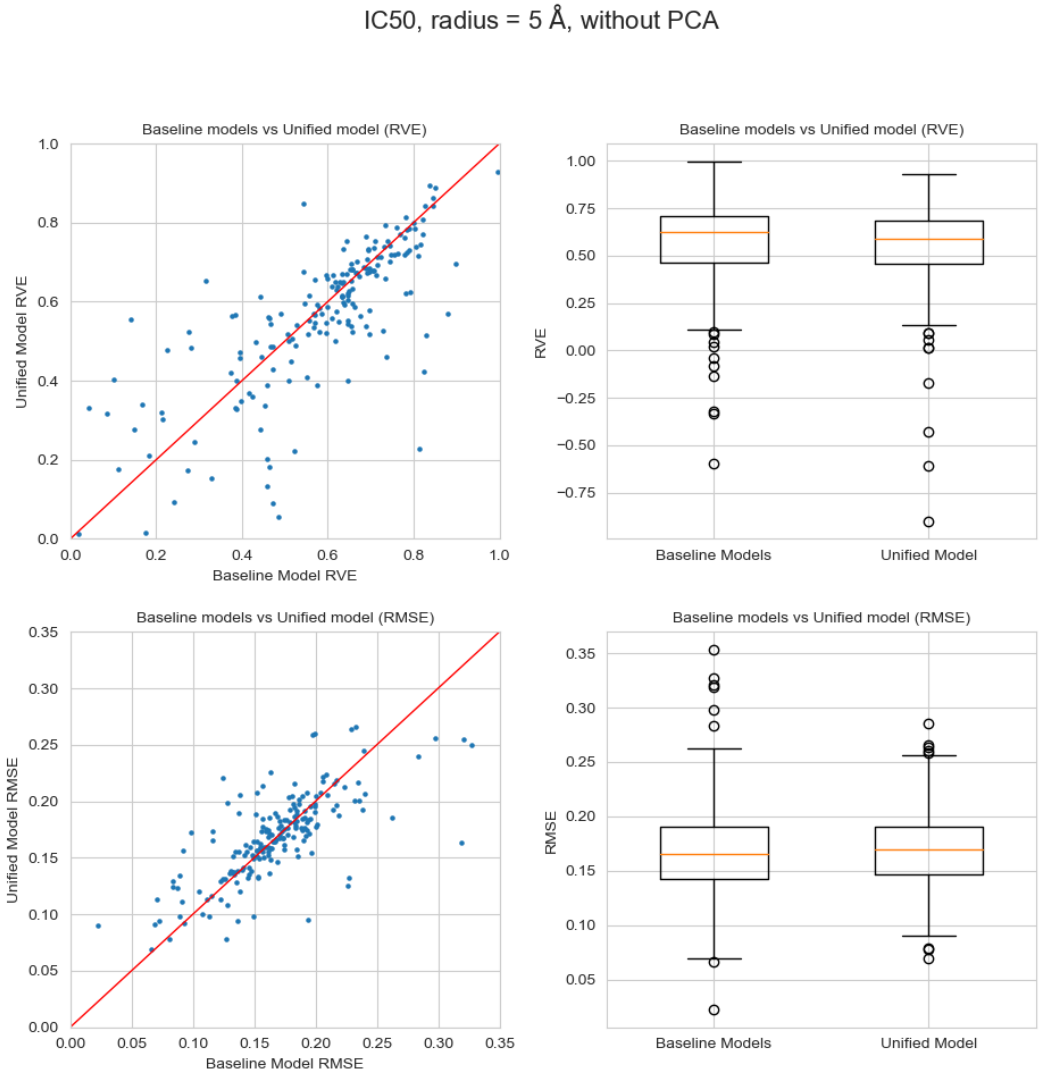


Figure 4.5: RMSE and RVE validation results for the IC50 dataset with radius=5Å without PCA

tial for further development and significant capabilities to expedite the drug discovery process. The performance of the Unified model is noteworthy, demonstrating predictive abilities comparable to contemporary state-of-the-art QSAR models, each meticulously tailored for a single target. This revelation prompts further research exploration, particularly to evaluate whether this model can address the prediction gap for targets with limited data, which is insufficient for reliable QSAR model training. The Semi-blind model also displays commendable predictive potential for untrained targets, although it falls slightly behind the prowess of the Unified model. While not all targets yield predictions, the outcomes of the Ki dataset with radius=5Å highlight the model's capacity, achieving RVE values hovering around 0.2 or higher for numerous targets. This underscores the utility of the approach in predicting outcomes for unexplored targets. Despite these noteworthy advancements, the path forward is laden with challenges. Further research and refinements are necessary to develop models capable of providing reliable predictions across the entire spectrum of targets. Varied outcomes resulting from different radius values underscore the ongoing

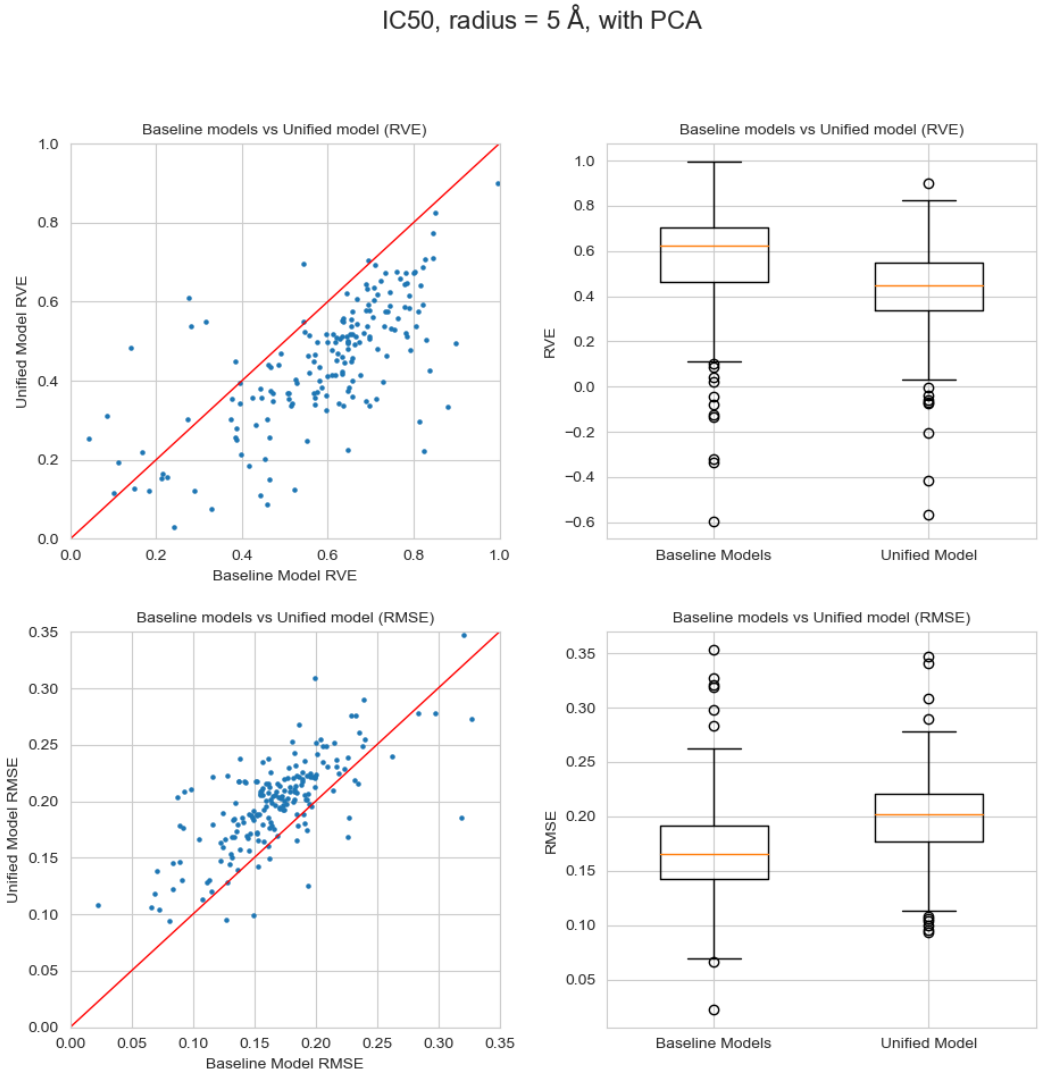


Figure 4.6: RMSE and RVE validation results for the IC50 dataset with radius=5Å with PCA

need to refine the structural information extraction process, either to identify the optimal radius or to effectively amalgamate diverse radii. Equally crucial is the exploration of distinct dimensionality reduction methodologies or innovative fingerprint representations, given the current challenges associated with dimensions. This need is further supported by PCA-derived ligand fingerprint reductions, which the Unified model predicted with less precision. Additionally, the formulation of an acceptable target applicability domain, similar to QSAR practices, could offer valuable insights. Another avenue ripe for exploration is the adoption of deep learning models. While initial investigations into this domain yielded limited benefits, these models could potentially leverage AlphaFold's widespread structural data availability for bioactivity-informed targets.

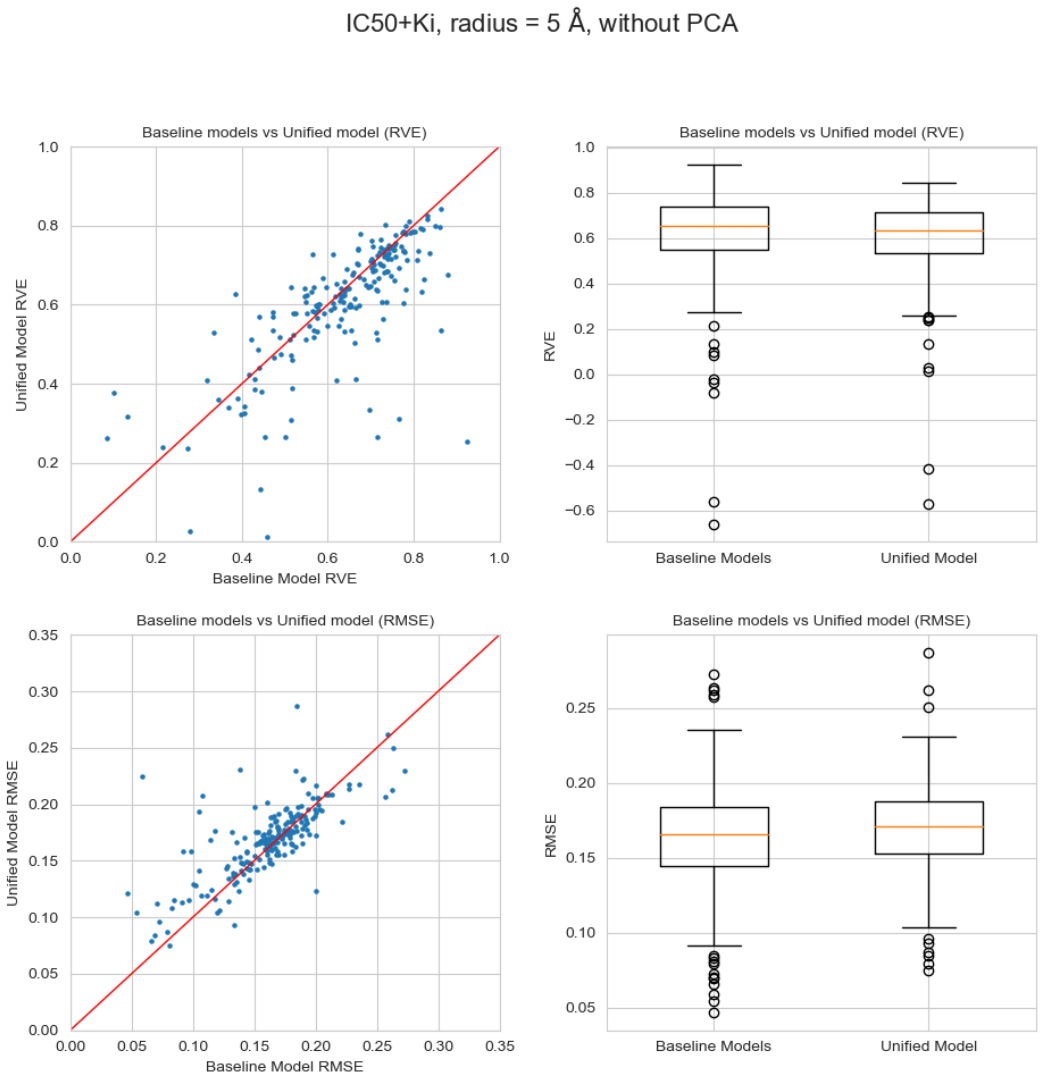


Figure 4.7: RMSE and RVE validation results for the IC50+Ki dataset with radius=5Å without PCA

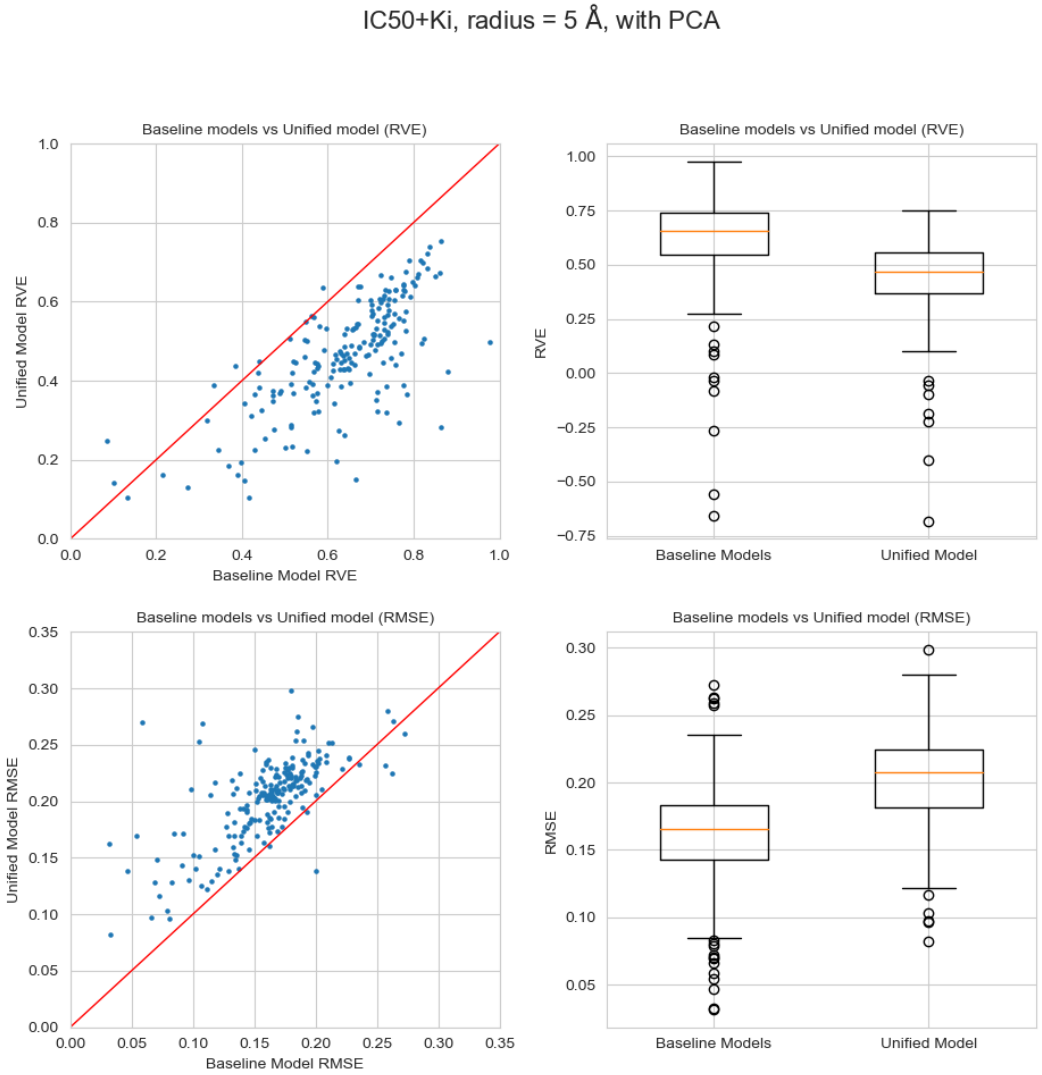


Figure 4.8: RMSE and RVE validation results for the IC50+Ki dataset with radius=5Å with PCA

Chapter 5

Conclusion

This research has culminated in the development and thorough testing of a methodology. The primary objective was to integrate protein structural data into machine learning models, addressing a critical shortcoming in QSAR-based approaches—restricted predictions for singular or a handful of correlated targets, all while upholding the methodology's inherent simplicity. This particular limitation poses a significant challenge to QSAR approaches, necessitating substantial initial data from lab assays and an ample number of assays for model training. The core premise was to leverage structural data of targets within the model, potentially enabling predictions of ligand bioactivities for hitherto untested targets. This paradigm shift essentially seeks to transform the erstwhile ligand-centric framework into a structure-driven approach. The pivotal structural data, extracted from AlphaFold, engenders a straightforward workflow, channeling 3D information for seamless integration into machine learning models.

The findings obtained from this study show promise, suggesting a methodology that warrants further development and possesses significant potential to streamline the drug discovery process. The performance of the Unified model is noteworthy, demonstrating predictive capabilities comparable to contemporary state-of-the-art QSAR models, each specifically designed for a single target. This discovery prompts additional research exploration, especially to evaluate the model's ability to fill the prediction gap for targets with limited data, which is inadequate for reliable QSAR model training. The Semi-blind model also displays commendable predictive potential for untrained targets, albeit with a slightly lower proficiency compared to the Unified model. Although not all targets yield predictions, the Ki dataset's radius=5Å outcomes underline the model's capacity, achieving RVE values hovering around 0.2 or higher for numerous targets. This underscores the approach's utility in predicting outcomes for unexplored targets.

Despite these notable advancements, the path forward is laden with challenges. Further research and refinements are essential to cultivate models capable of providing reliable predictions across the entire spectrum of targets. Varied outcomes arising from different radius values highlight the ongoing necessity to enhance the structural information extraction process, either by identifying the optimal radius or by effectively integrating diverse radii. Equally crucial is the exploration of alternative dimensionality reduction methodologies or innovative fingerprint representations, given the existing challenges associated with dimensions. This need is underscored by

PCA-derived ligand fingerprint reductions, which the Unified model predicted with less precision. Additionally, the establishment of a robust target applicability domain, similar to QSAR practices, could offer valuable insights. Another avenue for exploration is the adoption of deep learning models. While initial investigations into this domain yielded limited benefits, these models have the potential to leverage the extensive structural data provided by AlphaFold for bioactivity-informed targets.

In conclusion, the convergence of protein structural data with machine learning marks a novel frontier in drug discovery—one that is holistic, predictive, and potentially transformative. This study emphasizes the essential need to advance traditional methodologies, push the boundaries of existing knowledge, and embrace the intricacies of the molecular world. Through harnessing the strengths of both QSAR and structural data, this research lays the groundwork for more comprehensive, accurate, and adaptable models. Such advancements are particularly vital in an era where swift drug discovery and adaptability are paramount.

In closing, while there remain challenges to address and refine, the foundation laid by this research offers a beacon for future endeavors in the realm of drug discovery. It is hoped that this work serves as a catalyst, inspiring further innovations and ultimately contributing to the improvement of human health.

Abreviaturas

ADME Absorption, distribution, metabolism, and excretion.

AI Artificial intelligence.

BLAST Basic Local Alignment Search Tool.

ECFP Extended-Connectivity Fingerprints.

GB Gradient Boosting.

HTS High throughput screening.

LBDD Ligand-based drug discovery.

OECD Organisation for Economic Co-operation and Development.

PDB Protein Data Bank.

PK Pharmacokinetic.

QSAR Quantitative structure-activity relationship.

RF Random Forests.

RMSE Root Mean Square Error.

RVE Ratio of the Explained Variance.

SAR Structure-activity relationship.

SBDD Structure-based drug discovery.

SVM Support Vector Machine.

VS Virtual screening.

Bibliography

[1] David LJ Alexander, Alexander Tropsha, and David A Winkler. Beware of r 2: simple, unambiguous assessment of the prediction accuracy of qsar and qspr models. *Journal of chemical information and modeling*, 55(7):1316–1322, 2015.

[2] Stephen F Altschul, Warren Gish, Webb Miller, Eugene W Myers, and David J Lipman. Basic local alignment search tool. *Journal of molecular biology*, 215(3):403–410, 1990.

[3] Vinicius M. Alves, Tesia Bobrowski, Cleber C. Melo-Filho, Daniel Korn, Scott Auerbach, Charles Schmitt, Eugene N. Muratov, and Alexander Tropsha. Qsar modeling of sars-cov mpro inhibitors identifies sufugolix, cenicriviroc, proglumetacin, and other drugs as candidates for repurposing against sars-cov-2. *Molecular Informatics*, 40(1):2000113, 2021.

[4] Fodil Azzaz, Nouara Yahi, Henri Chahinian, and Jacques Fantini. The epigenetic dimension of protein structure is an intrinsic weakness of the alphafold program. *Biomolecules*, 12(10):1527, 2022.

[5] Dávid Bajusz, Anita Rácz, and Károly Héberger. Why is tanimoto index an appropriate choice for fingerprint-based similarity calculations? *Journal of cheminformatics*, 7(1):1–13, 2015.

[6] Maria Batool, Bilal Ahmad, and Sangdun Choi. A structure-based drug discovery paradigm. *International journal of molecular sciences*, 20(11):2783, 2019.

[7] Andreas Bender and Robert C Glen. Molecular similarity: a key technique in molecular informatics. *Organic & biomolecular chemistry*, 2(22):3204–3218, 2004.

[8] Helen M Berman, John Westbrook, Zukang Feng, Gary Gilliland, Talapady N Bhat, Helge Weissig, Ilya N Shindyalov, and Philip E Bourne. The protein data bank. *Nucleic acids research*, 28(1):235–242, 2000.

[9] Nicolas Bosc, Francis Atkinson, Eloy Felix, Anna Gaulton, Anne Hersey, and Andrew R Leach. Large scale comparison of qsar and conformal prediction methods and their applications in drug discovery. *Journal of cheminformatics*, 11:1–16, 2019.

[10] Leo Breiman. Random forests. *Machine learning*, 45:5–32, 2001.

[11] Simone Brogi, Teodorico Castro Ramalho, Kamil Kuca, José L Medina-Franco, and Marian Valko. In silico methods for drug design and discovery, 2020.

[12] Alice Capecchi, Daniel Probst, and Jean-Louis Reymond. One molecular fingerprint to rule them all: drugs, biomolecules, and the metabolome. *Journal of cheminformatics*, 12(1):1–15, 2020.

[13] Adrià Cereto-Massagué, María José Ojeda, Cristina Valls, Miquel Mulero, Santiago Garcia-Vallvé, and Gerard Pujadas. Molecular fingerprint similarity search in virtual screening. *Methods*, 71:58–63, 2015. Virtual Screening.

[14] Artem Cherkasov, Eugene N Muratov, Denis Fourches, Alexandre Varnek, Igor I Baskin, Mark Cronin, John Dearden, Paola Gramatica, Yvonne C Martin, Roberto Todeschini, et al. Qsar modeling: where have you been? where are you going to? *Journal of medicinal chemistry*, 57(12):4977–5010, 2014.

[15] Nicola Chirico and Paola Gramatica. Real external predictivity of qsar models: how to evaluate it? comparison of different validation criteria and proposal of using the concordance correlation coefficient. *Journal of chemical information and modeling*, 51(9):2320–2335, 2011.

[16] Kangway V Chuang, Laura M Gunsalus, and Michael J Keiser. Learning molecular representations for medicinal chemistry: Miniperspective. *Journal of Medicinal Chemistry*, 63(16):8705–8722, 2020.

[17] The UniProt Consortium. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Research*, 49(D1):D480–D489, 11 2020.

[18] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20:273–297, 1995.

[19] Wenqiang Cui, Adnane Aouidate, Shouguo Wang, Qiuliyang Yu, Yanhua Li, and Shuguang Yuan. Discovering anti-cancer drugs via computational methods. *Frontiers in pharmacology*, 11:733, 2020.

[20] JC Dearden, Mark TD Cronin, and Klaus LE Kaiser. How not to develop a quantitative structure–activity or structure–property relationship (qsar/qspr). *SAR and QSAR in Environmental Research*, 20(3-4):241–266, 2009.

[21] Kristina Djinovic-Carugo and Oliviero Carugo. Missing strings of residues in protein crystal structures. *Intrinsically disordered proteins*, 3(1):e1095697, 2015.

[22] Jurgen Drews. Drug discovery: a historical perspective. *science*, 287(5460):1960–1964, 2000.

[23] Hanna Eckert and Jürgen Bajorath. Molecular similarity analysis in virtual screening: foundations, limitations and novel approaches. *Drug discovery today*, 12(5-6):225–233, 2007.

[24] Michael Eisenstein. Artificial intelligence powers protein-folding predictions. *Nature*, 599:706–708, 2021.

[25] FDA. The drug development process, 04 2018.

[26] Robert D Finn, Jody Clements, and Sean R Eddy. Hmmer web server: interactive sequence similarity searching. *Nucleic acids research*, 39(suppl_2):W29–W37, 2011.

[27] Denis Fourches, Eugene Muratov, and Alexander Tropsha. Trust, but verify: on the importance of chemical structure curation in cheminformatics and qsar modeling research. *Journal of chemical information and modeling*, 50(7):1189, 2010.

[28] Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189 – 1232, 2001.

[29] Alexander Golbraikh and Alexander Tropsha. Beware of q2! *Journal of molecular graphics and modelling*, 20(4):269–276, 2002.

[30] Paola Gramatica. Principles of qsar models validation: internal and external. *QSAR & combinatorial science*, 26(5):694–701, 2007.

[31] Brian Kelley Ric sriniker gedeck Riccardo Vianello NadineSchneider David Cosgrove Eisuke Kawashima Andrew Dalke Dan N Gareth Jones Brian Cole Matt Swain Samo Turk AlexanderSavelyev Alain Vaucher Maciej Wójcikowski Ichiru Take Daniel Probst Kazuya Ujihara Vincent F. Scalfani guillaume godin Axel Pahl Francois Berenger JLVarjo strets123 JP DoliathGavid Greg Landrum, Paolo Tosco. rdkit/rdkit: 2022_09_1b1 (q3 2022) release, October 2022.

[32] Brian Kelley Ric sriniker gedeck Riccardo Vianello NadineSchneider David Cosgrove Eisuke Kawashima Andrew Dalke Dan N Gareth Jones Brian Cole Matt Swain Samo Turk AlexanderSavelyev Alain Vaucher Maciej Wójcikowski Ichiru Take Daniel Probst Kazuya Ujihara Vincent F. Scalfani guillaume godin Axel Pahl Francois Berenger JLVarjo strets123 JP DoliathGavid Greg Landrum, Paolo Tosco. rdkit/rdkit: 2022_09_1b1 (q3 2022) release, October 2022.

[33] Amit Kumar Halder and M Natalia Dias Soeiro Cordeiro. Qsar-co-x: an open source toolkit for multitarget qsar modelling. *Journal of Cheminformatics*, 13:1–18, 2021.

[34] Demis Hassabis. Alphafold reveals the structure of the protein universe, 07 2022.

[35] James P Hughes, Stephen Rees, S Barrett Kalindjian, and Karen L Philpott. Principles of early drug discovery. *British journal of pharmacology*, 162(6):1239–1249, 2011.

[36] William Humphrey, Andrew Dalke, and Klaus Schulten. Vmd: Visual molecular dynamics. *Journal of Molecular Graphics*, 14(1):33–38, 1996.

[37] Joanna S Jaworska, M Comber, C Auer, and CJ Van Leeuwen. Summary of a workshop on regulatory acceptance of (q) sars for human health and environmental endpoints. *Environmental health perspectives*, 111(10):1358–1360, 2003.

[38] VA Jisna and PB Jayaraj. Protein structure prediction: conventional and deep learning perspectives. *The Protein Journal*, 40(4):522–544, 2021.

[39] Mark A Johnson and Gerald M Maggiora. *Concepts and applications of molecular similarity*. Wiley, 1990.

[40] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.

[41] Wolfgang Kabsch and Christian Sander. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers: Original Research on Biomolecules*, 22(12):2577–2637, 1983.

[42] Samina Kausar and Andre O Falcao. An automated framework for qsar model building. *Journal of cheminformatics*, 10(1):1–23, 2018.

[43] Samina Kausar and Andre O Falcao. A visual approach for analysis and inference of molecular activity spaces. *Journal of cheminformatics*, 11(1):1–17, 2019.

[44] Kim Seyong Kim Yangsik. Target discovery using deep learning-based molecular docking and predicted protein structures with alphafold for novel antipsychotics. *Psychiatry Investig*, 20(6):504–514, 2023.

[45] Valeria V Kleandrova, Luciana Scotti, Francisco Jaime Bezerra Mendonca Junior, Eugene Muratov, Marcus T Scotti, and Alejandro Speck-Planche. Qsar modeling for multi-target drug discovery: Designing simultaneous inhibitors of proteins in diverse pathogenic parasites. *Frontiers in Chemistry*, 9:634663, 2021.

[46] Hiroyuki Kuwahara and Xin Gao. Analysis of the effects of related fingerprints on molecular similarity using an eigenvalue entropy approach. *Journal of Cheminformatics*, 13:1–12, 2021.

[47] Laura Elizabeth Lansdowne. A journey through the drug discovery pipeline.

[48] Laura Elizabeth Lansdowne. Exploring the drug development process, 08 2022.

[49] Sumudu P. Leelananda and Steffen Lindert. Computational methods in drug discovery. *Beilstein Journal of Organic Chemistry*, 12:2694–2718, 2016.

[50] Felipe Llinares-López, Quentin Berthet, Mathieu Blondel, Olivier Teboul, and Jean-Philippe Vert. Deep embedding and alignment of protein sequences. *Nature Methods*, 20(1):104–111, 2023.

[51] Ingrid Lobo. Basic local alignment search tool (blast), 2008.

[52] Tom Madden. The blast sequence analysis tool. *The NCBI handbook*, 2003.

[53] Gerald Maggiora, Martin Vogt, Dagmar Stumpfe, and Jurgen Bajorath. Molecular similarity in medicinal chemistry: miniperspective. *Journal of medicinal chemistry*, 57(8):3186–3204, 2014.

[54] Eduardo Habib Bechelane Maia, Letícia Cristina Assis, Tiago Alves De Oliveira, Alisson Marques Da Silva, and Alex Gutterres Taranto. Structure-based virtual screening: from classical to artificial intelligence. *Frontiers in chemistry*, 8:343, 2020.

[55] Andrea Mauri, Viviana Consonni, Roberto Todeschini, et al. Molecular descriptors. In *Handbook of computational chemistry*, pages 2065–2093. Springer International Publishing, 2017.

[56] Andrew T McNutt, Paul Francoeur, Rishal Aggarwal, Tomohide Masuda, Rocco Meli, Matthew Ragoza, Jocelyn Sunseri, and David Ryan Koes. Gnina 1.0: molecular docking with deep learning. *Journal of cheminformatics*, 13(1):1–20, 2021.

[57] David Mendez, Anna Gaulton, A Patrícia Bento, Jon Chambers, Marleen De Veij, Eloy Félix, María Paula Magariños, Juan F Mosquera, Prudence Mutowo, Michał Nowotka, et al. Chembl: towards direct deposition of bioassay data. *Nucleic acids research*, 47(D1):D930–D940, 2019.

[58] Xuan-Yu Meng, Hong-Xing Zhang, Mihaly Mezei, and Meng Cui. Molecular docking: a powerful approach for structure-based drug discovery. *Current computer-aided drug design*, 7(2):146–157, 2011.

[59] Harry L Morgan. The generation of a unique machine description for chemical structures-a technique developed at chemical abstracts service. *Journal of chemical documentation*, 5(2):107–113, 1965.

[60] Eugene N. Muratov, Jürgen Bajorath, Robert P. Sheridan, Igor V. Tetko, Dmitry Filimonov, Vladimir Poroikov, Tudor I. Oprea, Igor I. Baskin, Alexandre Varnek, Adrian Roitberg, Olexandr Isayev, Stefano Curtalolo, Denis Fourches, Yoram Cohen, Alan Aspuru-Guzik, David A. Winkler, Dimitris Agrafiotis, Artem Cherkasov, and Alexander Tropsha. Qsar without borders. *Chemical Society Reviews*, 49(11):3525–3564, 2020.

[61] Osita Sunday Nnyigide, Tochukwu Olunna Nnyigide, Sun-Gu Lee, and Kyu Hyun. Protein repair and analysis server: a web server to repair pdb structures, add missing heavy atoms and hydrogen atoms, and assign secondary structures by amide interactions. *Journal of Chemical Information and Modeling*, 62(17):4232–4246, 2022.

[62] OECD. Oecd principles for the validation, for regulatory purposes, of (quantitative) structure-activity relationship models, 2004.

[63] Daiki Okada, Fumihiko Ino, and Kenichi Hagihara. Accelerating the smith-waterman algorithm with interpair pruning and band optimization for the all-pairs comparison of base sequences. *BMC bioinformatics*, 16:1–15, 2015.

[64] Marius Olah, Maria Mracec, Liliana Ostopovici, Ramona Rad, Alina Bora, Nicoleta Hadaruga, Ionela Olah, Magdalena Banda, Zeno Simon, Mircea Mracec, and Tudor I. Oprea. *WOMBAT: World of Molecular Bioactivity*, chapter 9, pages 221–239. John Wiley Sons, Ltd, 2005.

[65] Marius Olah, Maria Mracec, Liliana Ostopovici, Ramona Rad, Alina Bora, Nicoleta Hadaruga, Ionela Olah, Magdalena Banda, Zeno Simon, Mircea Mracec, and Tudor I. Oprea. *WOMBAT: World of Molecular Bioactivity*, chapter 9, pages 760–786. John Wiley Sons, Ltd, 2005.

[66] Christine Orengo, David Jones, and Janet Thornton. *Bioinformatics: genes, proteins and computers*, chapter 6. BIOS Scientific Publishers Limited, 2003.

[67] Noel M O'Boyle and Roger A Sayle. Comparing structural fingerprints using a literature-based similarity benchmark. *Journal of cheminformatics*, 8(1):1–14, 2016.

[68] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[69] Yuri K Peterson, Xiang S Wang, Patrick J Casey, and Alexander Tropsha. Discovery of geranylgeranyltransferase-i inhibitors with novel scaffolds by the means of quantitative structure- activity relationship modeling, virtual screening, and experimental validation. *Journal of medicinal chemistry*, 52(14):4210–4220, 2009.

[70] Eric F Pettersen, Thomas D Goddard, Conrad C Huang, Gregory S Couch, Daniel M Greenblatt, Elaine C Meng, and Thomas E Ferrin. Ucsf chimera—a visualization system for exploratory research and analysis. *Journal of computational chemistry*, 25(13):1605–1612, 2004.

[71] Francisco J Prado-Prado, Xerardo García-Mera, and Humberto González-Díaz. Multi-target spectral moment qsar versus ann for antiparasitic drugs against different parasite species. *Bioorganic & medicinal chemistry*, 18(6):2225–2231, 2010.

[72] Anita Rácz, Dávid Bajusz, and Károly Héberger. Life beyond the tanimoto coefficient: similarity measures for interaction fingerprints. *Journal of cheminformatics*, 10(1):1–12, 2018.

[73] Peter Rice, Ian Longden, and Alan Bleasby. Emboss: the european molecular biology open software suite. *Trends in genetics*, 16(6):276–277, 2000.

[74] Raquel Rodríguez-Pérez and Jürgen Bajorath. Evaluation of multi-target deep neural network models for compound potency prediction under increasingly challenging test conditions. *Journal of Computer-Aided Molecular Design*, 35:285–295, 2021.

[75] Lars Rosenbaum, Alexander Dörr, Matthias R Bauer, Frank M Boeckler, and Andreas Zell. Inferring multi-target qsar models with taxonomy-based multi-task learning. *Journal of cheminformatics*, 5(1):1–20, 2013.

[76] Bryan L Roth, Douglas J Sheffler, and Wesley K Kroeze. Magic shotguns versus magic bullets: selectively non-selective drugs for mood disorders and schizophrenia. *Nature reviews Drug discovery*, 3(4):353–359, 2004.

[77] Kunal Roy, Supratik Kar, and Rudra Narayan Das. *Understanding the basics of QSAR for applications in pharmaceutical sciences and risk assessment*. Academic press, 2015.

[78] Victor T Sabe, Thandokuhle Ntombela, Lindiwe A Jhamba, Glenn EM Maguire, Thavendran Govender, Tricia Naicker, and Hendrik G Kruger. Current trends in computer aided drug design and a highlight of drugs discovered via computational techniques: A review. *European Journal of Medicinal Chemistry*, 224:113705, 2021.

[79] Valeria Scardino, Juan I Di Filippo, and Claudio N Cavasotto. How good are alphafold models for docking-based virtual screening? *Iscience*, 26(1), 2023.

[80] Courtney Schaal. The drug discovery pipeline, 01 2018.

[81] Bilal Shaker, Sajjad Ahmad, Jingyu Lee, Chanjin Jung, and Dokyun Na. In silico methods and tools for drug discovery. *Computers in biology and medicine*, 137:104851, 2021.

[82] Robert P Sheridan. Time-split cross-validation as a method for estimating the goodness of prospective prediction. *Journal of chemical information and modeling*, 53(4):783–790, 2013.

[83] Eugene G Shpaer, Max Robinson, David Yee, James D Candlin, Robert Mines, and Tim Hunkapiller. Sensitivity and selectivity in protein similarity searches: a comparison of smith–waterman in hardware to blast and fasta. *Genomics*, 38(2):179–191, 1996.

[84] Temple F Smith and Michael S Waterman. Identification of common molecular subsequences. *Journal of molecular biology*, 147(1):195–197, 1981.

[85] MS Smyth and JHJ Martin. x ray crystallography. *Molecular Pathology*, 53(1):8, 2000.

[86] Alejandro Speck-Planche, Valeria V Kleandrova, Feng Luan, and M Natália DS Cordeiro. Chemoinformatics in anti-cancer chemotherapy: multi-target qsar model for the in silico discovery of anti-breast cancer agents. *European journal of pharmaceutical sciences*, 47(1):273–279, 2012.

[87] Alejandro Speck-Planche, Valeria V Kleandrova, Feng Luan, and M Natália DS Cordeiro. Rational drug design for anti-cancer chemotherapy: multi-target qsar models for the in silico discovery of anti-colorectal cancer agents. *Bioorganic & medicinal chemistry*, 20(15):4848–4855, 2012.

[88] Ashutosh Srivastava, Tetsuro Nagai, Arpita Srivastava, Osamu Miyashita, and Florence Tama. Role of computational methods in going beyond x-ray crystallography to explore protein structure and dynamics. *International Journal of Molecular Sciences*, 19(11), 2018.

[89] Dagmar Stumpfe and Jürgen Bajorath. Similarity searching. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 1(2):260–282, 2011.

[90] Hao Tang, Xiang S Wang, Xi-Ping Huang, Bryan L Roth, Kyle V Butler, Alan P Kozikowski, Mira Jung, and Alexander Tropsha. Novel inhibitors of human histone deacetylase (hdac) identified by qsar modeling of known inhibitors, virtual screening, and experimental validation. *Journal of chemical information and modeling*, 49(2):461–476, 2009.

[91] The AlphaFold team. Alphafold: a solution to a 50-year-old grand challenge in biology, 11 2020.

[92] Igor V Tetko, Iurii Sushko, Anil Kumar Pandey, Hao Zhu, Alexander Tropsha, Ester Papa, Tomas Oberg, Roberto Todeschini, Denis Fourches, and Alexandre Varnek. Critical assessment of qsar models of environmental toxicity against tetrahymena pyriformis: focusing on applicability domain and overfitting by variable selection. *Journal of chemical information and modeling*, 48(9):1733–1746, 2008.

[93] Andrey A Toropov and Alla P Toropova. Qspr/qsar: State-of-art, weirdness, the future. *Molecules*, 25(6):1292, 2020.

[94] Wouter G Touw, Coos Baakman, Jon Black, Tim AH Te Beek, Elmar Krieger, Robbie P Joosten, and Gert Vriend. A series of pdb-related databanks for everyday needs. *Nucleic acids research*, 43(D1):D364–D368, 2015.

[95] Alexander Tropsha. Best practices for qsar model development, validation, and exploitation. *Molecular informatics*, 29(6-7):476–488, 2010.

[96] Mihaly Varadi and Sameer Velankar. The impact of alphafold protein structure database on the fields of life sciences. *PROTEOMICS*, page 2200128, 2022.

[97] Panagiotis D Vouzis and Nikolaos V Sahinidis. Gpu-blast: using graphics processors to accelerate protein sequence alignment. *Bioinformatics*, 27(2):182–188, 2011.

[98] Zachary J. Wehrspan, Robert T. McDonnell, and Adrian H. Elcock. Identification of iron-sulfur (fe-s) cluster and zinc (zn) binding sites within proteomes predicted by deepmind's alphafold2 program dramatically expands the metalloproteome. *Journal of Molecular Biology*, 434(2):167377, 2022.

[99] Peter Willett. Fusing similarity rankings in ligand-based virtual screening. *Computational and structural biotechnology journal*, 5(6):e201302002, 2013.

[100] Peter Willett, John M Barnard, and Geoffrey M Downs. Chemical similarity searching. *Journal of chemical information and computer sciences*, 38(6):983–996, 1998.

[101] wwPDB consortium. Protein Data Bank: the single global archive for 3D macromolecular structure data. *Nucleic Acids Research*, 47(D1):D520–D528, 10 2018.

[102] wwPDB consortium. Protein Data Bank: the single global archive for 3D macromolecular structure data. *Nucleic Acids Research*, 47(D1):D520–D528, 10 2018.

[103] Dehua Yang, Qingtong Zhou, Viktorija Labroska, Shanshan Qin, Sanaz Darbalaei, Yiran Wu, Elita Yuliantie, Linshan Xie, Houchao Tao, Jianjun Cheng, Qing Liu, Suwen Zhao, Wenqing Shui, Yi Jiang, and Ming-Wei Wang. G protein-coupled receptors: structure- and function-based drug discovery. *Signal Transduction and Targeted Therapy*, 6(1):7, 2021.

[104] Ru Zhang and Xin Xie. Tools for gpcr drug discovery. *Acta Pharmacologica Sinica*, 33(3):372–384, 2012.

[105] Shuxing Zhang, Linyi Wei, Ken Bastow, Weifan Zheng, Arnold Brossi, Kuo-Hsiung Lee, and Alexander Tropsha. Antitumor agents 252. application of validated qsar models to database mining: discovery of novel tylophorine derivatives as potential anticancer agents. *Journal of computer-aided molecular design*, 21:97–112, 2007.

[106] Hao Zhu, Alexander Tropsha, Denis Fourches, Alexandre Varnek, Ester Papa, Paola Gramatica, Tomas Oberg, Phuong Dao, Artem Cherkasov, and Igor V Tetko. Combinatorial qsar modeling of chemical toxicants tested against tetrahymena pyriformis. *Journal of chemical information and modeling*, 48(4):766–784, 2008.

[107] Wensi Zhu, Aditi Shenoy, Petras Kundrotas, and Arne Elofsson. Evaluation of AlphaFold-Multimer prediction on multi-chain protein complexes. *Bioinformatics*, 39(7):btad424, 07 2023.

Appendix A

Results data

Model	Training set size	Testing set size	RMSE	RVE
Baseline QSAR	112217	28122	0.167	0.683
Unified model	112264	28075	0.168	0.678
Semi-blind model	121750	18589	0.339	0.073

Results and training and testing set sizes for all models - Ki with radius=7Å and without PCA

Model	Training set size	Testing set size	RMSE	RVE
Baseline QSAR	112217	28122	0.167	0.683
Unified model	112266	28073	0.206	0.525
Semi-blind model	121750	18589	0.314	0.138

Results and training and testing set sizes for all models - Ki with radius=7Å and with PCA

Model	Training set size	Testing set size	RMSE	RVE
Baseline QSAR	80646	20260	0.167	0.644
Unified model	80719	20187	0.168	0.633
Semi-blind model	91216	9690	0.292	0.046

Results and training and testing set sizes for all models - IC50 with radius=5Å and without PCA

Model	Training set size	Testing set size	RMSE	RVE
Baseline QSAR	80646	20260	0.167	0.644
Unified model	80737	20213	0.199	0.503
Semi-blind model	91260	9690	0.295	0.009

Results and training and testing set sizes for all models - IC50 with radius=5Å and with PCA

Model	Training set size	Testing set size	RMSE	RVE
Baseline QSAR	80646	20260	0.167	0.644
Unified model	80671	20143	0.168	0.636
Semi-blind model	91124	9690	0.316	0.007

Results and training and testing set sizes for all models - IC50 with radius=7Å and without PCA

Model	Training set size	Testing set size	RMSE	RVE
Baseline QSAR	80646	20260	0.167	0.644
Unified model	80779	20171	0.201	0.502
Semi-blind model	91260	9690	0.296	0.028

Results and training and testing set sizes for all models - IC50 with radius=7Å and with PCA

Model	Training set size	Testing set size	RMSE	RVE
Baseline QSAR	185264	46419	0.167	0.682
Unified model	185364	46319	0.170	0.674
Semi-blind model	209662	22021	0.322	0.111

Results and training and testing set sizes for all models - IC50+Ki with radius=5Å and without PCA

Model	Training set size	Testing set size	RMSE	RVE
Baseline QSAR	185264	46419	0.167	0.682
Unified model	185548	46396	0.209	0.514
Semi-blind model	209662	22021	0.298	0.151

Results and training and testing set sizes for all models - IC50+Ki with radius=5Å and with PCA

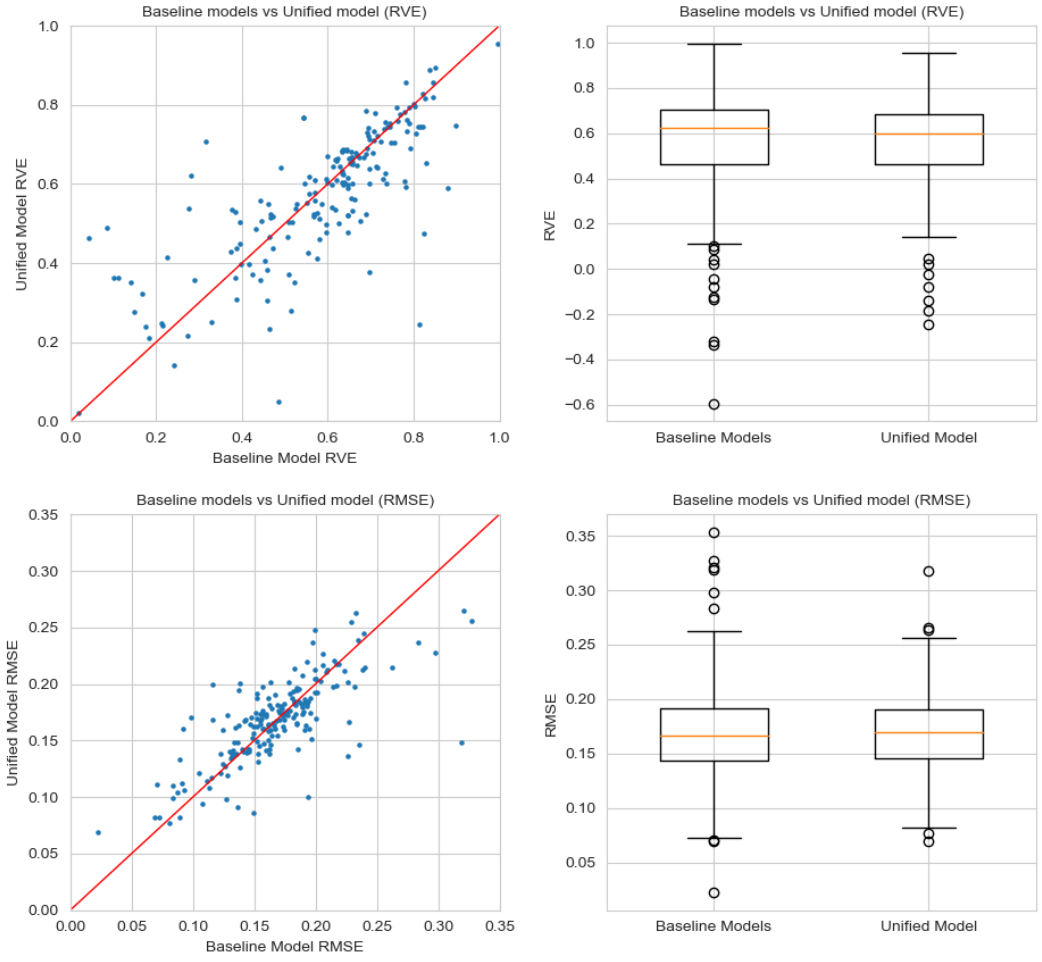
Model	Training set size	Testing set size	RMSE	RVE
Baseline QSAR	185264	46419	0.167	0.682
Unified model	185548	46396	0.170	0.670
Semi-blind model	206286	25658	0.329	0.108

Results and training and testing set sizes for all models - IC50+Ki with radius=7Å and without PCA

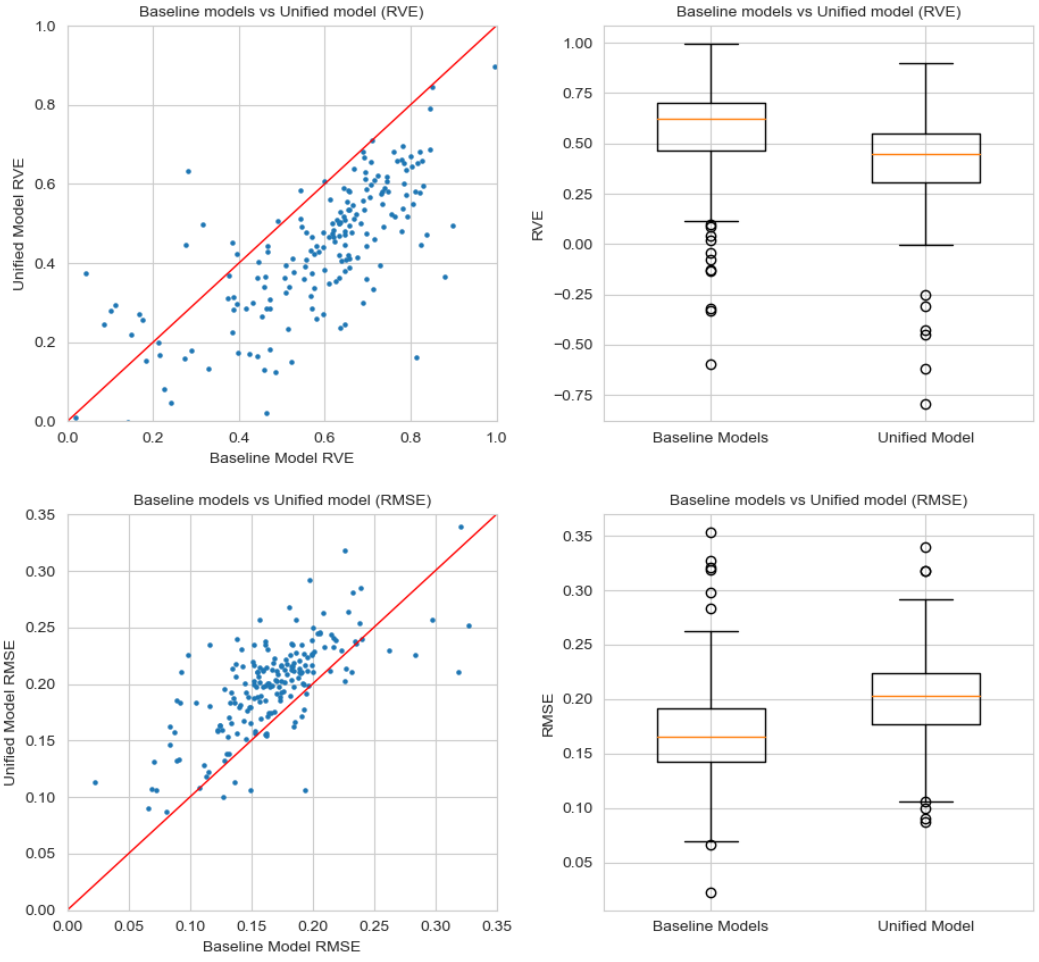
Model	Training set size	Testing set size	RMSE	RVE
Baseline QSAR	185264	46419	0.167	0.682
Unified model	185442	46366	0.210	0.512
Semi-blind model	206150	25658	0.300	0.145

Results and training and testing set sizes for all models - IC50+Ki with radius=7Å and with PCA

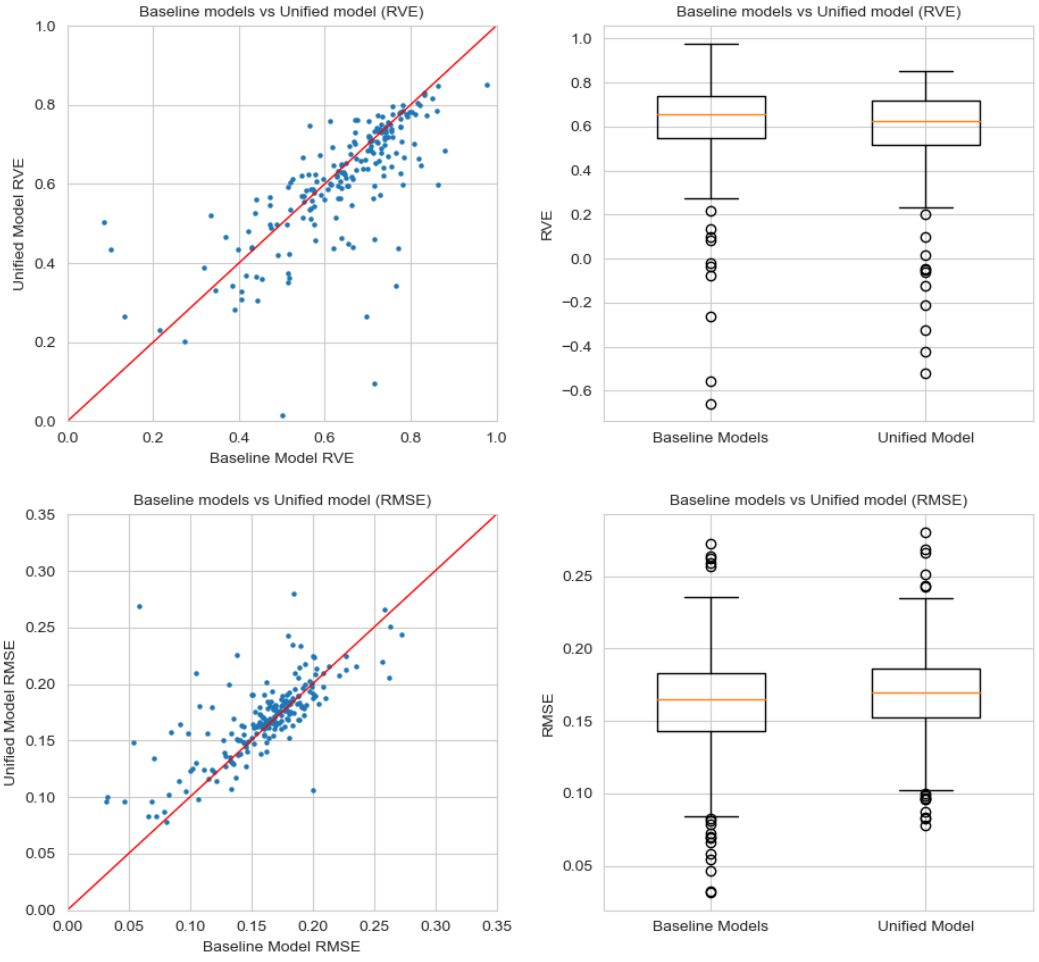
IC50, radius = 7 Å, without PCA



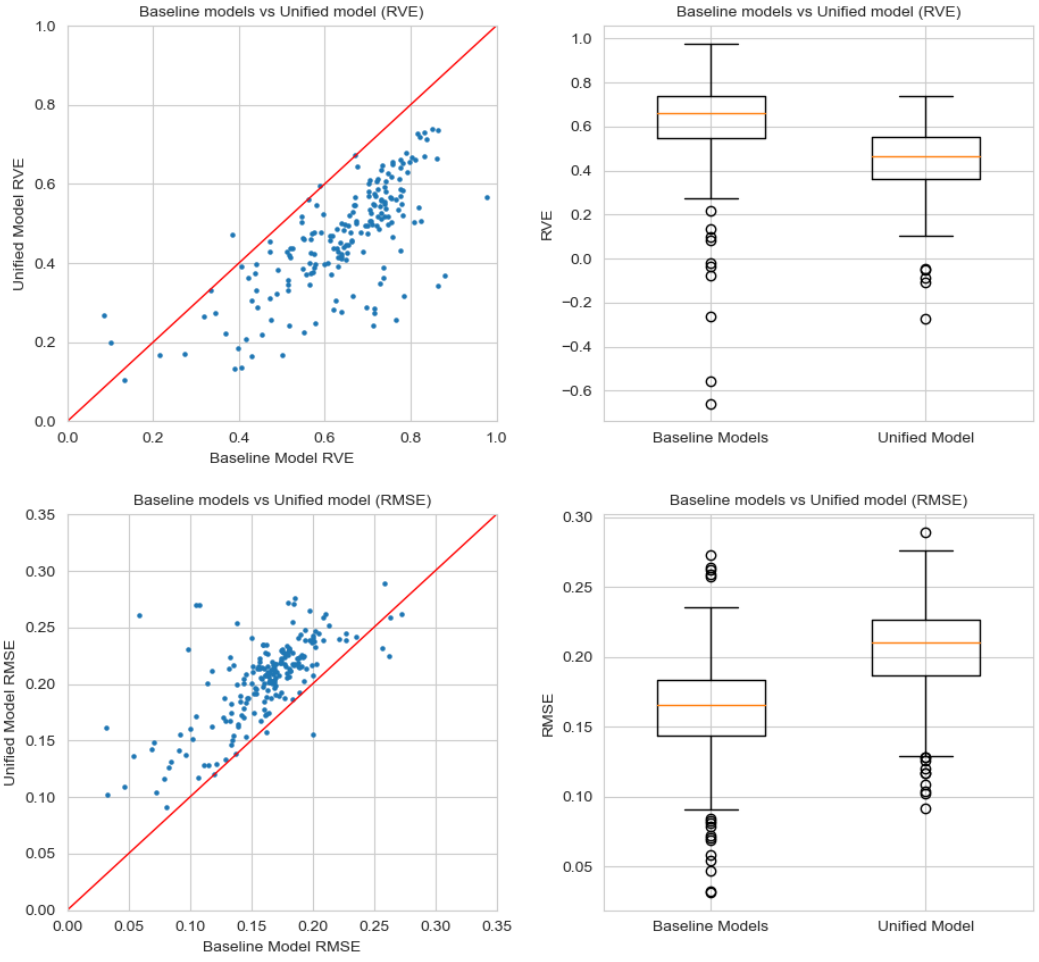
IC50, radius = 7 Å, with PCA



IC50+Ki, radius = 7 Å, with PCA



IC50+Ki, radius = 7 Å, with PCA



	Baseline Models		Unified Model		Semi-blind Model	
Target ID	RMSE	RVE	RMSE	RVE	RMSE	RVE
O43613	0.163	0.668	0.163	0.675	0.293	-0.012
P08173	0.147	0.776	0.168	0.750	0.314	0.117
P08912	0.159	0.618	0.168	0.648	0.294	0.058
P18089	0.169	0.500	0.169	0.568	0.249	-0.014
P18825	0.187	0.616	0.168	0.665	0.299	0.163
P21918	0.154	0.756	0.146	0.743	0.274	0.240
P25024	0.136	0.608	0.138	0.646	0.259	-0.131
P28222	0.182	0.741	0.177	0.748	0.334	0.115
P28335	0.173	0.658	0.173	0.650	0.344	0.018
P30872	0.190	0.696	0.172	0.660	0.300	0.125
P32246	0.192	0.252	0.169	0.393	0.251	0.036
P35346	0.133	0.863	0.153	0.778	0.315	0.054
P35462	0.171	0.695	0.172	0.691	0.389	0.058
P41145	0.174	0.734	0.168	0.747	0.341	0.127
P51681	0.255	0.116	0.219	0.266	0.398	0.024

Comparison of results for all 15 validation targets of the semi-blind model for the Ki dataset with radius=7Å and without PCA

	Baseline Models		Unified Model		Semi-blind Model	
Target ID	RMSE	RVE	RMSE	RVE	RMSE	RVE
O43613	0.163	0.668	0.205	0.491	0.316	0.036
P08173	0.147	0.776	0.190	0.655	0.320	0.104
P08912	0.159	0.618	0.161	0.663	0.332	0.018
P18089	0.169	0.500	0.178	0.490	0.246	0.074
P18825	0.187	0.616	0.176	0.642	0.273	0.148
P21918	0.154	0.756	0.203	0.530	0.245	0.426
P25024	0.136	0.608	0.192	0.322	0.292	-0.204
P28222	0.182	0.741	0.222	0.606	0.331	0.137
P28335	0.173	0.658	0.210	0.485	0.320	0.073
P30872	0.190	0.696	0.167	0.694	0.293	0.243
P32246	0.192	0.252	0.175	0.337	0.234	-0.033
P35346	0.133	0.863	0.191	0.638	0.334	-0.033
P35462	0.171	0.695	0.204	0.551	0.358	0.064
P41145	0.174	0.734	0.206	0.622	0.271	0.347
P51681	0.255	0.116	0.224	0.195	0.284	0.081

Comparison of results for all 15 validation targets of the semi-blind model for the Ki dataset with radius=7Å and with PCA

	Baseline Models		Unified Model		Semi-blind Model	
Target ID	RMSE	RVE	RMSE	RVE	RMSE	RVE
O43614	0.149	0.673	0.152	0.665	0.294	-0.035
P08173	0.163	0.646	0.186	0.543	0.274	-0.019
P08912	0.153	0.580	0.156	0.522	0.272	-0.205
P18089	0.185	0.644	0.157	0.752	0.282	0.144
P18825	0.227	0.543	0.132	0.847	0.270	0.381
P25024	0.116	0.510	0.173	0.401	0.321	-0.060
P28222	0.186	0.490	0.152	0.568	0.185	0.372
P28335	0.170	0.518	0.176	0.507	0.245	0.038
P34947	0.124	0.609	0.132	0.558	0.338	-0.196
P35462	0.162	0.696	0.175	0.676	0.312	0.082
P41145	0.157	0.706	0.177	0.679	0.279	0.178
P41597	0.171	0.647	0.178	0.622	0.284	0.068
P51677	0.152	0.803	0.151	0.784	0.397	0.124
Q13255	0.177	0.665	0.177	0.672	0.327	-0.048
Q14832	0.144	0.633	0.155	0.646	0.239	0.045

Comparison of results for all 15 validation targets of the semi-blind model for the IC50 dataset with radius=5Å and without PCA

	Baseline Models		Unified Model		Semi-blind Model	
Target ID	RMSE	RVE	RMSE	RVE	RMSE	RVE
O43614	0.149	0.673	0.186	0.499	0.286	-0.040
P08173	0.163	0.646	0.213	0.376	0.269	0.057
P08912	0.153	0.580	0.171	0.401	0.288	-0.194
P18089	0.185	0.644	0.188	0.620	0.287	0.167
P18825	0.227	0.543	0.186	0.696	0.336	0.063
P25024	0.116	0.510	0.180	0.355	0.312	-0.177
P28222	0.186	0.490	0.179	0.470	0.200	0.267
P28335	0.170	0.518	0.203	0.342	0.259	0.029
P34947	0.124	0.609	0.160	0.415	0.363	-0.605
P35462	0.162	0.696	0.215	0.512	0.296	0.104
P41145	0.157	0.706	0.205	0.560	0.285	0.184
P41597	0.171	0.647	0.203	0.498	0.290	0.045
P51677	0.152	0.803	0.187	0.675	0.382	0.052
Q13255	0.177	0.665	0.221	0.513	0.338	-0.170
Q14832	0.144	0.633	0.170	0.551	0.237	0.069

Comparison of results for all 15 validation targets of the semi-blind model for the IC50 dataset with radius=5Å and with PCA

	Baseline Models		Unified Model		Semi-blind Model	
Target ID	RMSE	RVE	RMSE	RVE	RMSE	RVE
O43614	0.149	0.673	0.153	0.668	0.348	0.033
P08173	0.163	0.646	0.182	0.522	0.261	0.066
P08912	0.153	0.580	0.170	0.512	0.292	-0.217
P18089	0.185	0.644	0.164	0.687	0.311	-0.029
P18825	0.227	0.543	0.166	0.768	0.388	0.056
P25024	0.116	0.510	0.168	0.372	0.253	-0.073
P28222	0.186	0.490	0.143	0.640	0.298	-0.173
P28335	0.170	0.518	0.176	0.504	0.248	0.023
P34947	0.124	0.609	0.129	0.540	0.271	-0.182
P35462	0.162	0.696	0.188	0.639	0.320	-0.006
P41145	0.157	0.706	0.174	0.680	0.279	0.204
P41597	0.171	0.647	0.181	0.615	0.342	0.029
P51677	0.152	0.803	0.149	0.796	0.358	0.103
Q13255	0.177	0.665	0.176	0.680	0.365	-0.268
Q14832	0.144	0.633	0.139	0.684	0.239	0.036

Comparison of results for all 15 validation targets of the semi-blind model for the IC50 dataset with radius=7Å and without PCA

	Baseline Models		Unified Model		Semi-blind Model	
Target ID	RMSE	RVE	RMSE	RVE	RMSE	RVE
O43614	0.149	0.673	0.180	0.523	0.293	0.021
P08173	0.163	0.646	0.214	0.381	0.283	0.065
P08912	0.153	0.580	0.202	0.259	0.303	-0.147
P18089	0.185	0.644	0.197	0.589	0.294	0.083
P18825	0.227	0.543	0.214	0.584	0.320	0.176
P25024	0.116	0.510	0.180	0.325	0.290	-0.089
P28222	0.186	0.490	0.167	0.507	0.249	-0.003
P28335	0.170	0.518	0.203	0.339	0.248	0.023
P34947	0.124	0.609	0.164	0.348	0.283	-0.081
P35462	0.162	0.696	0.234	0.442	0.325	0.070
P41145	0.157	0.706	0.201	0.568	0.272	0.210
P41597	0.171	0.647	0.209	0.481	0.307	0.072
P51677	0.152	0.803	0.199	0.645	0.347	0.058
Q13255	0.177	0.665	0.213	0.547	0.355	-0.271
Q14832	0.144	0.633	0.167	0.502	0.239	0.089

Comparison of results for all 15 validation targets of the semi-blind model for the IC50 dataset with radius=7Å and with PCA

	Baseline Models		Unified Model		Semi-blind Model	
Target ID	RMSE	RVE	RMSE	RVE	RMSE	RVE
O43613	0.138	0.748	0.153	0.716	0.307	0.052
P08173	0.170	0.729	0.173	0.698	0.273	0.336
P08912	0.161	0.579	0.168	0.602	0.230	0.204
P18089	0.169	0.567	0.158	0.643	0.221	0.297
P18825	0.175	0.699	0.165	0.711	0.256	0.405
P21918	0.150	0.741	0.154	0.731	0.311	-0.016
P25024	0.184	0.546	0.163	0.621	0.282	-0.081
P28222	0.157	0.779	0.169	0.735	0.377	0.108
P28335	0.174	0.639	0.174	0.632	0.309	0.051
P32246	0.168	0.695	0.185	0.644	0.335	0.079
P35462	0.179	0.669	0.171	0.700	0.375	0.150
P41597	0.169	0.663	0.180	0.614	0.331	0.042
P43250	0.083	0.820	0.108	0.633	0.229	0.387
Q13255	0.163	0.712	0.188	0.638	0.343	-0.090
Q14832	0.118	0.824	0.177	0.664	0.343	-0.009

Comparison of results for all 15 validation targets of the semi-blind model for the IC50+Ki dataset with radius=5Å and without PCA

	Baseline Models		Unified Model		Semi-blind Model	
Target ID	RMSE	RVE	RMSE	RVE	RMSE	RVE
O43613	0.138	0.748	0.194	0.539	0.270	0.109
P08173	0.170	0.729	0.200	0.606	0.245	0.422
P08912	0.161	0.579	0.176	0.539	0.248	0.189
P18089	0.169	0.567	0.173	0.561	0.223	0.282
P18825	0.175	0.699	0.196	0.605	0.259	0.390
P21918	0.150	0.741	0.183	0.626	0.319	0.002
P25024	0.184	0.546	0.205	0.461	0.312	0.106
P28222	0.157	0.779	0.206	0.627	0.252	0.433
P28335	0.174	0.639	0.212	0.451	0.296	0.049
P32246	0.168	0.695	0.228	0.466	0.298	0.108
P35462	0.179	0.669	0.211	0.544	0.360	0.177
P41597	0.169	0.663	0.209	0.468	0.294	0.018
P43250	0.083	0.820	0.128	0.494	0.242	-0.233
Q13255	0.163	0.712	0.215	0.532	0.304	0.074
Q14832	0.118	0.824	0.216	0.506	0.330	0.018

Comparison of results for all 15 validation targets of the semi-blind model for the IC50+Ki dataset with radius=5Å and with PCA

	Baseline Models		Unified Model		Semi-blind Model	
Target ID	RMSE	RVE	RMSE	RVE	RMSE	RVE
O43613	0.138	0.748	0.152	0.719	0.311	0.050
P08173	0.170	0.729	0.163	0.737	0.302	0.195
P08912	0.161	0.579	0.161	0.607	0.238	0.149
P18089	0.169	0.567	0.171	0.586	0.286	0.066
P18825	0.175	0.699	0.173	0.683	0.344	0.237
P21918	0.150	0.741	0.152	0.731	0.295	0.051
P25024	0.184	0.546	0.164	0.620	0.273	-0.071
P28335	0.174	0.639	0.176	0.629	0.294	0.028
P32246	0.168	0.695	0.179	0.662	0.346	0.069
P35462	0.179	0.669	0.170	0.700	0.425	0.046
P41145	0.181	0.721	0.173	0.742	0.314	0.309
P41597	0.169	0.663	0.180	0.618	0.310	0.020
P43250	0.083	0.820	0.102	0.665	0.210	0.087
Q13255	0.163	0.712	0.179	0.680	0.326	0.019
Q14832	0.118	0.824	0.179	0.647	0.340	0.029

Comparison of results for all 15 validation targets of the semi-blind model for the IC50+Ki dataset with radius=7Å and without PCA

	Baseline Models		Unified Model		Semi-blind Model	
Target ID	RMSE	RVE	RMSE	RVE	RMSE	RVE
O43613	0.138	0.748	0.199	0.519	0.275	0.083
P08173	0.170	0.729	0.218	0.544	0.288	0.205
P08912	0.161	0.579	0.172	0.548	0.218	0.345
P18089	0.169	0.567	0.197	0.462	0.248	0.122
P18825	0.175	0.699	0.222	0.476	0.314	0.157
P21918	0.150	0.741	0.201	0.537	0.262	0.271
P25024	0.184	0.546	0.187	0.504	0.293	0.127
P28335	0.174	0.639	0.219	0.418	0.289	0.048
P32246	0.168	0.695	0.217	0.494	0.323	0.053
P35462	0.179	0.669	0.210	0.545	0.340	0.146
P41145	0.181	0.721	0.214	0.606	0.301	0.298
P41597	0.169	0.663	0.212	0.458	0.302	0.041
P43250	0.083	0.820	0.126	0.541	0.204	0.198
Q13255	0.163	0.712	0.208	0.551	0.342	-0.146
Q14832	0.118	0.824	0.212	0.508	0.293	0.111

Comparison of results for all 15 validation targets of the semi-blind model for the IC50+Ki dataset with radius=7Å and with PCA