UNIVERSIDADE D
COIMBRA

Nelson Rodrigo Carvalho Monteiro

Toward the Explainability of Drug-Target Interactions: End-to-End Deep Learning Architectures for Binding Affinity Prediction

PhD Thesis in Informatics Engineering, Intelligent Systems, supervised by Professor Joel P. Arrais and Professor José L. Oliveira and presented to the Department of Informatics Engineering of the Faculty of Sciences and Technology of the University of Coimbra.

November 2023

1 2 9 0

FACULDADE DE CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE D COIMBRA

Toward the Explainability of Drug–Target Interactions: End-to-End Deep Learning Architectures for Binding Affinity Prediction

Nelson Rodrigo Carvalho Monteiro

Supervisor:
Prof. Joel P. Arrais, Ph.D.

Co-supervisor:
Prof. José L. Oliveira, Ph.D.

Dissertation presented to obtain a Ph.D. degree in Informatics Engineering, Intelligent Systems, at the Faculty of Sciences and Technology of the University of Coimbra

Dissertação de Doutoramento apresentada à Faculdade de Ciências e Tecnologia da Universidade de Coimbra, para prestação de provas de Doutoramento em Engenharia Informática, Sistemas Inteligentes

November, 2023

The studies presented in this thesis were carried out at the Centre for Informatics and Systems of the University of Coimbra (CISUC), Department of Informatics Engineering (DEI), Faculty of Science and Technology of the University of Coimbra (FCTUC), Portugal.

This work was conducted with financial support from the following institutions/programs:

- Ph.D. grant 2020.04741.BD from the FCT - Foundation for Science and Technology, I.P..
- Project CISUC (UIDB/00326/2020) financed by national funds (PIDDAC) through FCT - Foundation for Science and Technology, I.P./MCTES.
- Project D4 - Deep Drug Discovery and Deployment (CENTRO-01-0145-FEDER-029266) financed by national funds (PIDDAC) through FCT - Foundation for Science and Technology, I.P./MCTES.

"You can, you should, and if you're brave enough to start, you will."

Stephen King

Acknowledgments

Quero começar por agradecer ao meu orientador, Joel P. Arrais, por ter-me dado a oportunidade deste longo percurso que aqui termina, por ter sempre acreditado e confiado naquilo que são as minhas capacidades e as minhas ideias, e por todo o apoio dado. Se me perguntassem no dia 10 de Dezembro de 2018, quando enviei um email sobre o meu interesse na área de bioinformática, motivado pela cadeira de Introdução à Bioinformática, que estaria hoje a escrever uma tese de doutoramento, provavelmente diria que isso não iria acontecer. Obrigado por esta jornada não só científica, mas também de enriquecimento pessoal.

Um agradecimento ao meu co-orientador, José L. Oliveira, pela disponibilidade e ajuda demonstrada, por todos os comentários construtivos, e pelas revisões detalhadas e essenciais para todo o progresso desta investigação. Este obrigado também se estende para as várias pessoas do IEETA que tive a oportunidade de conhecer e de partilhar conhecimento.

Agradeço ao Carlos Simões da BSIM Therapeutics por toda a partilha de conhecimento, pela disponibilidade prestada, e por ter sempre apoiado o meu trabalho e o meu progresso.

Agradeço ao laboratório LARN (ou Transformer), em especial ao Tiago, Luís, Daniel, Anuschka, e Maryam, por me terem acompanhado neste percurso e por ter feito também parte do vosso, por todo o espírito de ajuda e de trabalho que sempre existiu, pelas discussões de ideias, pelos convívios, e por todas as histórias que ficam. Fico à espera do próximo almoço/jantar!

Quero também agradecer a amigos e família que Coimbra me deu, Alberto, Rita, Luís Silva, Rodrigo, Mafalda, Vitor, Pipa, Tiago, Luís, Ana, Inês, Chico, Albuquerque, Oshley e Mariana. Um grande e profundo obrigado por serem vocês, por estarem lá de uma forma ou de outra, e por acompanharem este meu longo percurso. Guardo eternamente todos os momentos, todas as histórias, todas as memórias, todas as conversas e todas as partilhas.

Um obrigado aos meus amigos, Fábio, Diogo, e Júlio, que acompanharam tanto engenharia biomédica como agora o doutoramento em engenharia informática ao mesmo tempo que eu. A partilha de todos os memes e das frustrações foram importantes para chegar hoje ao fim!

Ao Hélio, Raquel, João e Violeta. Um grande obrigado a vocês que, apesar de acompanharem este percurso há menos tempo, são imprescindíveis e tiveram uma enorme importância para aquele que foi o meu último ano de doutoramento. Agradeço imenso o vosso apoio e a confiança que sempre me transmitem. São muitas as coisas que já poderia aqui escrever, mas ficam para as nossas conversas no S. José!

À minha família. Acho que é complicado escrever sobres aqueles que desde sempre se sacrificaram, que abdicaram de tanta coisa, e que tudo fizeram para que eu pudesse hoje chegar aqui. Apenas posso agradecer por acreditarem sempre em mim, por todo o apoio, por toda a força, por nunca me terem deixado desistir, pelo orgulho que sempre vi nos vossos olhos e nas vossas palavras, e por terem sempre estado ao meu lado neste percurso como se fosse o vosso. Sem vocês não teria sido possível nem teria feito sentido.

Abstract

The identification of compounds that exhibit selective binding to druggable proteins continues to pose challenges within the realm of pharmaceutical exploration and drug discovery. On that account, the proper assessment of target-specific compound selectivity and the accurate prediction of an unbiased Drug–Target Affinity (DTA) metric are pivotal to promoting the identification of novel Drug–Target Interactions (DTIs), the discovery of potential leads, and the understanding of the binding process. Although significant efforts have been made to increase the effectiveness of traditional *in vitro* and *in vivo* experimental approaches, these methods remain impractical for the vast array of compounds and proteins currently known. Hence, the establishment of effective computational strategies capable of using all available proteomics, chemical, and pharmacological data becomes decisive in the pursuit of new findings in the field of drug discovery.

Despite the plethora of *in silico* solutions to overcome the challenges of traditional *in vitro* and *in vivo* experiments, most of these studies still focus on binary classification, overlooking the importance of characterizing DTIs with unbiased binding strength values to properly distinguish primary interactions from those with off-targets. Moreover, several of these methods usually simplify the entire interaction mechanism, neglecting the multi-domain inter-dependency associated with the proteomics, chemical, and pharmacological spaces, and have yet to consider including explainability into the inner structure of the architectures or providing potential explanations to the predictions, thus, limiting the validity and understanding of the results. Furthermore, the majority of DTA or DTI prediction computational studies have not yet given any special characterization or attention to binding positions or actively integrated information regarding binding pockets during the learning process, leading to the estimation of potential DTIs based on redundant sites or substructures and compromising the reliability of the predictions.

This research endeavors to tackle the challenge of DTA prediction by proposing and

investigating novel Deep Learning (DL) architectures that leverage raw sequential and structural data and focus on modeling the multi-domain representation space of DTIs. Furthermore, it aims to offer potential insights regarding DTIs and enhance prediction understanding by exploring the explainability field of black-box models. This thesis comprises three main contributions.

The first contribution consisted of exploring the reliability of Convolutional Neural Networks (CNNs) in the identification of relevant sequential and structural regions for binding, specifically binding sites and evolutionarily conserved motifs, and the robustness of the deep representations extracted by providing explanations to the model's decisions based on the identification of the input regions that contributed the most to the prediction. This study makes use of an end-to-end DL architecture to predict binding affinity, where CNNs are exploited in their capacity to automatically identify and extract discriminating deep representations from 1D sequential and structural data. The results demonstrated the effectiveness of the deep representations extracted from CNNs in the prediction of binding affinity. CNNs were found to identify and extract features from regions relevant to the interaction without any *a priori* information, where the weight associated with these spots was in the range of those with the highest positive influence given by the CNNs in the prediction. The end-to-end DL model achieved the highest performance both in the prediction of the binding affinity and in the ability to correctly distinguish the interaction strength rank order when compared to baseline approaches. This research study validated the potential applicability of an end-to-end DL architecture in the context of drug discovery beyond the confined space of proteins and ligands with determined 3D structures. Furthermore, it showed the reliability of the deep representations extracted from the CNNs by providing explainability to the decision-making process.

The second contribution is a novel end-to-end Transformer-based architecture, Drug–Target Interaction TRansformer (DTITR), for predicting DTA using 1D raw sequential and structural data to represent the proteins and compounds. This architecture exploits self-attention layers to capture the short and long-term proteomics and chemical context dependencies between the sequential and structural units of the proteins and compounds, respectively, and cross-attention layers to exchange information and learn the pharmacological context associated with the interaction space. The results showed that DTITR is effective in predicting DTA, achieving superior performance in both correctly predicting the value of interaction strength and being able to correctly discriminate the rank order of binding strength compared to state-of-the-art baselines. The combination of multiple Transformer-Encoders was found to result in robust and discriminative aggregate representations of the pro-

teins and compounds for binding affinity prediction, in which the addition of a Cross-Attention Transformer-Encoder was identified to be important for improving the discriminative power of these representations. This research study validated the applicability of an end-to-end Transformer-based architecture in the context of drug discovery, capable of self-providing different levels of potential DTI and prediction understanding due to the nature of the attention blocks.

The last contribution is a novel end-to-end binding-region-guided Transformer-based architecture, TAG-DTA, that simultaneously predicts the 1D binding pocket and the binding affinity of DTI pairs, where the prediction of the 1D binding pocket guides and conditions the prediction of DTA. This architecture uses 1D raw sequential and structural data to represent the proteins and compounds, respectively, and combines multiple Transformer-Encoder blocks to capture and learn the proteomics, chemical, and pharmacological contexts. The predicted 1D binding pocket conditions the attention mechanism of the Transformer-Encoder used to learn the pharmacological space in order to model the inter-dependency amongst binding-related positions. The obtained results outline the predictive performance of TAG-DTA compared to state-of-the-art benchmarks, including in unknown subsets of the proteomics and chemical representation spaces. Moreover, it was found that the 1D binding pocket prediction increases the discriminative power and robustness of the aggregate representation of the pharmacological space and improves the DTA prediction performance. This study demonstrated that combining computationally different yet contextually related tasks is critical to new findings in the DTI domain. Additionally, it showed that TAG-DTA is capable of providing increased DTI and prediction understanding due to the nature of the attention blocks and prediction of the 1D binding pocket.

Keywords: Drug–Target Interaction, Drug–Target Affinity, Binding Pocket, Explainability, Deep Learning

Resumo

A identificação de compostos que exibem ligação seletiva para proteínas farmacologicamente viáveis continua a colocar desafios no âmbito da exploração farmacêutica e da descoberta de fármacos. Nesse sentido, a avaliação adequada da seletividade de compostos específicos de um alvo biológico e a previsão precisa de uma métrica de afinidade fármaco-alvo (DTA) imparcial são cruciais para promover a identificação de novas interações fármaco-alvo (DTI), a descoberta de potenciais fármacos ativos, e a compreensão do processo de ligação. Apesar dos esforços significativos para aumentar a eficácia das abordagens experimentais *in vitro* e *in vivo* tradicionais, estes métodos continuam inexequíveis para a vasta gama de compostos e proteínas atualmente conhecidos. Portanto, estabelecer estratégias computacionais eficazes, que sejam capazes de usar todos os dados proteômicos, químicos e farmacológicos disponíveis, torna-se decisivo na procura de novas descobertas no ramo da descoberta de fármacos.

Apesar das inúmeras soluções *in silico* para superar os desafios das experiências *in vitro* e *in vivo* tradicionais, a maioria destes estudos ainda se foca na classificação binária, subvalorizando a importância de caracterizar DTIs com valores de força de ligação imparciais para distinguir adequadamente as interações primárias das interações com alvos não específicos. Para além disso, muitos destes métodos geralmente simplificam o mecanismo de interação, negligenciado a interdependência multi-domínio associada aos espaços proteômico, químico e farmacológico, e ainda não consideraram incluir explicabilidade na estrutura interna das arquiteturas ou o fornecimento de potenciais explicações para as previsões, limitando, assim, a validade e a compreensão dos resultados. Além disso, a maioria dos estudos computacionais de previsão de DTA ou DTI ainda não deu nenhuma caracterização especial ou atenção às posições de ligação, ou integrou ativamente informação sobre zonas de ligação durante o processo de aprendizagem, levando à previsão de potenciais DTIs com base em locais ou subestruturas redundantes, e comprometendo a confiabilidade das previsões.

Esta investigação visa abordar o desafio da previsão de DTA ao propor e investigar novas arquiteturas de aprendizagem profunda (DL) que aproveitem dados sequenciais e estruturais brutos, e que se foquem na modelação do espaço de representação multi-domínio das DTIs. Para além disso, ela procura oferecer potenciais perceções sobre DTIs e melhorar a compreensão da previsão ao explorar o campo da explicabilidade de modelos caixa preta. Esta tese compreende três contribuições principais.

A primeira contribuição consistiu em explorar a confiabilidade das redes neuronais convolucionais (CNNs) na identificação de regiões sequenciais e estruturais relevantes para a ligação, especificamente sítios de ligação e padrões evolucionariamente conservados, e na robustez das representações profundas extraídas através do fornecimento de explicações para as decisões do modelo com base na identificação das regiões de entrada que mais contribuíram para a previsão. Este estudo faz uso de uma arquitetura de DL de ponta a ponta para prever a afinidade de ligação, onde as CNNs são exploradas na sua capacidade de identificar e extrair automaticamente representações profundas discriminantes de dados sequenciais e estruturais 1D. Os resultados demonstraram a eficácia das representações profundas extraídas pelas CNNs na previsão da afinidade de ligação. As CNNs foram capazes de identificar e extrair características de regiões relevantes para a interação sem qualquer informação *a priori*, onde o peso associado a esses locais estava na gama daqueles com a maior influência positiva dada pelas CNNs na previsão. O modelo de DL de ponta a ponta alcançou o melhor desempenho tanto na previsão da afinidade de ligação como na capacidade de distinguir corretamente a ordem de grandeza da força de interação em comparação com as abordagens de referência. Este estudo validou a potencial aplicabilidade de uma arquitetura de DL de ponta a ponta no contexto da descoberta de fármacos para além do espaço confinado de proteínas e ligantes com estruturas 3D determinadas. Além disso, ele mostrou a confiabilidade das representações profundas extraídas das CNNs ao fornecer explicabilidade ao processo de tomada de decisão.

A segunda contribuição é uma nova arquitetura baseada em *Transformers* de ponta a ponta, *Drug-Target Interaction TRansformer (DTITR)*, para prever DTA usando dados sequenciais e estruturais brutos de 1D para representar as proteínas e os compostos. Esta arquitetura explora camadas de *self-attention* para capturar as dependências do contexto proteómico e químico de curto e longa distância entre as unidades sequenciais e estruturais das proteínas e dos compostos, respetivamente, e camadas de *cross-attention* para trocar informação e aprender o contexto farmacológico associado ao espaço de interação. Os resultados mostraram que a DTITR é eficaz na previsão de DTA, alcançando um desempenho superior tanto na previsão

correta do valor da força de interação como na capacidade de distinguir corretamente a ordem de grandeza da força de ligação em comparação com as abordagens de estado de arte. A combinação de vários *Transformer-Encoders* demonstrou resultar em representações agregadas robustas e discriminativas das proteínas e dos compostos para previsão de afinidade de ligação, em que a adição de um *Cross-Attention Transformer-Encoder* foi identificada como sendo importante para melhorar o poder discriminativo destas representações. Este estudo validou a aplicabilidade de uma arquitetura de ponta a ponta baseada em *Transformers* no contexto da descoberta de fármacos, capaz de fornecer diferentes níveis de potencial compreensão de DTI e previsão devido à natureza das camadas de atenção.

A última contribuição é uma nova arquitetura baseada em *Transformers* de ponta a ponta e guiada pela região de ligação, TAG-DTA, que prevê simultaneamente a cavidade de ligação 1D e a afinidade de ligação de pares DTI, em que a previsão da cavidade de ligação 1D guia e condiciona a previsão de DTA. Esta arquitetura utiliza dados sequenciais e estruturais brutos de 1D para representar as proteínas e os compostos, respetivamente, e combina vários blocos *Transformer-Encoder* para capturar e aprender os contextos proteómico, químico e farmacológico. A previsão da cavidade de ligação 1D condiciona o mecanismo de atenção do *Transformer-Encoder* utilizado para aprender o espaço farmacológico, com o objetivo de modelar a interdependência entre as posições envolvidas na ligação. Os resultados obtidos destacam o desempenho preditivo da TAG-DTA em comparação com o estado de arte, incluindo em subconjuntos desconhecidos dos espaços de representação proteómica e química. Além disso, verificou-se que a previsão da cavidade de ligação 1D aumenta o poder discriminativo e a robustez da representação agregada do espaço farmacológico e melhora o desempenho na previsão de DTA. Este estudo demonstrou que a combinação de tarefas computacionalmente diferentes, contudo, contextualmente relacionadas, é fundamental para novas descobertas no domínio de DTI. Adicionalmente, demonstrou que a TAG-DTA é capaz de fornecer uma maior compreensão de DTI e previsão devido à natureza das camadas de atenção e à previsão da cavidade de ligação 1D.

Keywords: Interação Fármaco-Alvo, Afinidade Fármaco-Alvo, Cavidade de Ligação, Explicabilidade, Aprendizagem Profunda

Contents

List of Acronyms

r^2	Coefficient of Determination. 80, 97, 98, 123–126, 157, 162
Adam	Adaptive Moment Estimation. 232, 233, 259, 265
ADME	Absorption, Distribution, Metabolism, and Excretion. 22, 24, 38
ADMET	Absorption, Distribution, Metabolism, Excretion, and Toxicity. xxxiv, 22, 23
AI	Artificial Intelligence. 59, 60
ANN	Artificial Neural Network. 70
CF	Collaborative Filtering. 41
CI	Concordance Index. 80, 97, 98, 123–126, 129, 151–157, 162, 236
CNN	Convolutional Neural Network. xii, xvi, xxxiv, xxxv, 6, 8, 45–48, 51, 52, 54–56, 71, 73, 85, 86, 88, 90, 93, 94, 97–103, 105–110, 167
CTD	Composition, Transition and Distribution. 43
DL	Deep Learning. xii, xvi, 3–6, 8, 38, 39, 42–46, 48, 51, 54, 55, 59–61, 67, 85, 86, 88, 89, 95, 97, 98, 107, 109, 110, 114, 130, 132, 167, 168, 170, 258

DNA Deoxyribonucleic Acid. xxxiii, 14, 15

DTA Drug–Target Affinity. xi–xiii, xv–xvii, 2, 3, 5–8, 33, 48, 52, 54, 56, 85, 86, 98, 107, 114, 124, 132, 135, 136, 140, 151, 152, 162–164, 167, 168

DTI Drug–Target Interaction. xi–xiii, xv–xvii, xxxvi, xxxviii, xliii, 1–8, 13, 22, 25, 31, 33, 35, 39–48, 53–56, 60, 85, 86, 88, 89, 91–93, 97, 98, 100–102, 107, 109, 110, 113–119, 121, 123, 126–132, 135–138, 140–142, 144, 146, 147, 150–152, 156–160, 162, 163, 165, 167, 168, 268, 269

ECFP Extended-Connectivity Fingerprint. 39, 55, 231

EMEA European Medicines Evaluation Agency. xxxiv, 22, 23

FCNN Fully-Connected Feed-Forward Neural Network. xxxiv, xxxv, xxxviii, 45–48, 51, 52, 54–56, 70, 71, 76–78, 88–90, 114, 117, 121, 124–126, 129, 130, 136, 141, 142, 146, 150, 259–261, 267, 269

FDA Food and Drug Administration. xxxiv, 22, 23, 40

FFN Feed-Forward Neural Network. 38, 44, 48–50

GAT Graph Attention Neural Network. 55

GAT-GCN Graph Attention - Graph Convolutional Neural Network. 55

GBM Gradient Boosting Machine. 39, 69

GBR Gradient Boosting Regression. xliv, 53, 69, 233, 234

GCN Graph Convolutional Neural Network. 47, 53, 55

GELU Gaussian Error Linear Unit. 259, 260, 265–267, 269

GIN Graph Isomorphism Neural Network. 55
GNN Graph Neural Network. 45, 55
GPCR G-Protein-Coupled Receptor. 40, 42, 43, 53

HCS High Content Screening. 39
HTS High-Throughput Screening. 21

IC_{50} half maximal inhibition concentration. 3, 26, 27

K_d dissociation constant. xl, 3, 5, 26, 48–50, 52, 53, 86, 87, 93, 110, 114, 115, 132, 136–138, 230
K_i inhibition constant. 3, 26, 27, 48, 50–53
KNN K-Nearest Neighbor. 38
KRR Kernel Ridge Regression. xliv, 70, 97, 98, 233, 234

LLM Large Language Model. 131, 132, 164
LN Layer Normalization. 74, 76, 119–121
LSTM Long Short-Term Memory. 45, 55

MACCS Molecular Access System. 39, 42, 43
MCC Matthew's Correlation Coefficient. 82, 154, 156, 268
MF Matrix Factorization. 41
MHCA Multi-Head Cross-Attention. 119, 120
MHSA Multi-Head Self-Attention. xxxv, 73–76, 118–121, 143
ML Machine Learning. 3, 8, 31, 38, 39, 42, 43, 48, 49, 51–53, 59–61, 67, 86, 97, 98, 114, 168, 170
MLM Masked Language Modeling. xxv, xxxix, xliv, xlvi, 140, 141, 143, 144, 147–149, 154, 156, 163, 164, 265, 266, 270
mRNA messenger RNA. xxxiii, xxxix, 14–17, 223, 224

MSE Mean Squared Error. 80, 97, 98, 123–126, 129, 151–157, 162, 232, 233, 236, 259, 267–269

NLP Natural Language Processing. 47

NMR Nuclear Magnetic Resonance. 22, 23

PCM Proteochemometric. 33, 39, 42, 45, 48, 52, 60

pIC_{50} Numerical Log-Based Transformation of Half Maximal Inhibitory Concentration. 49

pK_d Numerical Log-Based Transformation of Dissociation Constant. xxxv–xxxviii, xl, 49, 51, 87–93, 107, 116, 118, 121–123, 129, 138, 141, 142, 146, 162, 230

pK_i Numerical Log-Based Transformation of Inhibition Constant. 49, 51

PSSM Position Specific Scoring Matrix. xxxvi, xxxvii, 45, 94, 101, 102, 106–108

PWFFN Position-Wise Feed-Forward Neural Network. xxxv, 73, 74, 78, 118–121, 141, 143, 259, 265–267, 269

QSAR Quantitative Structure-Activity Relationship. 21, 37, 38, 60

RAdam Rectified Adaptive Moment Estimation. 259, 260, 265–267, 269

RBF Radial Basis Function. 68

RBM Restricted Boltzmann Machine. 44

ReLU Rectified Linear Unit. 95, 96, 231, 259, 265

RF Random Forest. 38, 39, 42, 43, 48, 50, 53, 67

RFR Random Forest Regression. xliv, 68, 233, 234

RMSE Root Mean Squared Error. 80, 97, 98, 123–126, 129, 151–157, 162, 232

RNA Ribonucleic Acid. xxxiii, 14, 16

RNN Recurrent Neural Network. 45, 143, 258

rRNA ribosomal RNA. 15

SAR Structure-Activity Relationship. 21

SMILES Simplified Molecular Input Line Entry System. xxv, xxxv–xxxix, xli, xliv, xlvi, 45–48, 54–56, 87–93, 95, 107, 109, 114–124, 126, 127, 129–131, 136–144, 146–149, 154, 156, 158, 163–165, 229–231, 236, 240, 257–261, 265–267, 269, 270

Spearman Spearman Rank Correlation. 81, 97, 98, 123–126, 151, 152, 154, 156, 157, 162

SVM Support Vector Machine. xxxiv, 38–40, 42–44, 48, 53, 68–70

SVR Support Vector Regression. xliv, 49, 69, 98, 233, 234

tRNA transfer RNA. xxxiii, xxxix, 15, 17, 224

VS Virtual Screening. 39, 48, 49

XAI Explainable Artificial Intelligence. 59, 61, 62, 64

List of Figures

List of Tables

Chapter 1

Introduction

The contributions in the bioinformatics and cheminformatics domains have been pivotal in the pursuit and development of Drug–Target Interaction (DTI) prediction algorithms, promoting and accelerating the drug discovery process. Even though many efforts have been devoted to these prediction methodologies, many challenges remain to address, including providing strategies to explain the models' decisions, which is crucial in critical contexts such as drug discovery, and recognizing the inherent complexity and multi-domain inter-independency of the binding process associated with DTIs. This chapter introduces such limitations and traces this thesis's main research goals.

1.1 Motivation

The discovery and development of new drugs remain one of the greatest challenges in the biomedical and pharmaceutical areas. In that regard, the identification of potential DTIs is decisive in drug discovery and drug repositioning processes, especially when considering that the therapeutic effects of active compounds are determined through the observation of DTIs [1]. Despite the efforts on the traditional *in vivo* and *in vitro* drug discovery experiments, conducting low or high-throughput bioassays for the screening of potential leads is time-consuming, labor-intensive, and unfeasible for the vast compound and protein spaces, compromising the effectiveness of these approaches [2]. Moreover, the wide range of unexpected clinical side effects and the lack of knowledge regarding potential off-targets highly influence the success rate of these more conventional experimental methods [3]. Thus, *in silico* (computational) strategies have attracted increasing attention due to their ability to exploit comprehensive chemical and proteomic libraries, improve the efficacy of the early stages of drug discovery, and promote the overall understanding of the biological, chemical, and pharmacological processes involved in DTIs [4]. Furthermore, most of these computational methods learn from already approved drugs, which allows

them to bypass several steps of the traditional *de novo* drug discovery pipeline, given that a considerable amount of the drug candidates have already been through the validation phases [5].

In spite of the major advances and interesting findings obtained by *in silico* approaches, the amount of approved drugs remains a low percentage of all known bioactive compounds [6, 7]. Moreover, the number of new drugs discovered every year is declining, conversely to the number of new variants of already existing infections and diseases [8]. The emergence of multi-resistant pathological conditions caused by new mutations of certain viruses or bacteria is a rising health concern, especially when considering the ineffectiveness of the currently available medicine against the symptoms triggered by these new stripes, leading to potentially life-threatening situations [9, 10, 11]. Additionally, there is an increasingly irrational and injudicious misuse of the currently available medicine, causing a resistance effect to these kinds of agents or, on the other hand, accelerating directly or indirectly the evolution and potential mutations of those bacteria and viruses [12, 13]. On that account, the proposal of novel and efficient methodologies capable of accurately identifying DTIs, and providing potential explanations for the inferring process remains an ongoing challenge in the drug discovery field.

1.1.1 Limitations of Binary DTI Classification

The majority of the computational studies proposed to solve the DTI prediction challenge rely on shallow binary associations to characterize the interaction and conduct the experiments, indicating only if a certain active compound interacts or not with the corresponding biological target [14]. On that account, the importance of Drug–Target Affinity (DTA), which considers all the comprehensive processes involved in the interaction, i.e., reflects the magnitude and rank order of the pair association, is usually overlooked, especially given that predicting DTA is substantially more challenging. Hence, the quality of the predictions is compromised or at least limited, particularly in the identification of primary interactions, leading to a lack of target selectivity, which is crucial in the drug discovery context due to the polypharmacological nature of most existing drug molecules [15]. Furthermore, negative interactions (lack of interaction) are mostly based on the absence of information or possible hypotheses, resulting in the prediction of potential unknown false negatives [16, 17].

Nevertheless, the increase in interactions with known bioactivity measurements and the expansion of binding-related databases, such as ChEMBL [6], BindingDB [18],

or PDBBind [19], have been instrumental in the pursuit of more realistic and informative studies, shifting computational drug discovery toward methodologies focused on the prediction of real-valued interaction strengths [20]. However, several of these studies center their experiments on sources of unreliable binding affinity metrics, e.g., inhibition constant (K_i) [21, 22] or half maximal inhibition concentration (IC_{50}) [23], due to the number of available data points compared to other sources of bioactivity. Even though computational frameworks may perform significantly better with larger datasets, the use of biased bioactivity metrics limits the validity of the results.

Thus, the use of direct and independent binding affinity or bioactivity measurements, such as the dissociation constant (K_d) [24, 25], is essential to accurately and realistically predict DTIs and properly distinguish primary interactions from those with off-targets (secondary interactions).

1.1.2 DTI Understanding and Model Explainability

On account of the progressive advances in computing and the growth of available data to train complex models, Deep Learning (DL) algorithms have been successfully employed in several fields of interest, including critical contexts such as bioinformatics, cheminformatics, health informatics, and pharmaceutical informatics [26]. Hence, most recent studies dealing with DTI or DTA prediction have explored DL strategies, achieving better results than traditional Machine Learning (ML) solutions [27]. The higher modular capability of these architectures to estimate non-linear mapping between data input and output, discover appropriate representations from structured or unstructured raw data, and learn sequential and/or structural motifs, has led to interesting findings in the DTI domain. However, these methods progressively transform the input to increase the representations' selectivity and invariance, resulting in abstract learned features, which are essentially not interpretable by humans. Furthermore, these representations do not provide a tractable path to the input domain, leading to inadequate explanations about the context that is responsible for a specific decision [28]. Thus, DL architectures are often considered highly complex black-box models, especially regarding the explainability of their results and understanding of the underlying aspects around the inner decisions [29].

Considering the context of the problem, where the results presented may have a great impact on the drug discovery process chain, it is vital to understand and provide possible explanations for the reasoning behind the decisions of these complex architectures [30, 31]. Moreover, focusing on explaining these models' decisions may present an important opportunity to validate the results and lead to novel findings

regarding key regions within each binding component and/or related to binding-specific substructures, i.e., provide potential DTI understanding. Additionally, exploring the explainability of DL architectures can provide ways of understanding how to improve methodologies and select adequate input representations of the involving interacting components [32].

1.1.3 DTI Representation and Domain Inter-dependency

The interaction between small molecules and proteins results from the recognition and complementarity of certain active groups (binding regions) and it is supported by the joint action of other individual substructures scattered across the protein and compound [33, 34]. Hence, the protein amino acid sequence, including the binding pocket within the protein sequence, and the compound's chemical structure are determinants for the interaction. However, several computational approaches characterize proteins and compounds using different combinations of conventional and global descriptors, e.g., physicochemical descriptors, which are mostly not robust or discriminating for predicting a real interaction [35]. Even though these descriptors may capture some intricate information regarding the proteomics and chemical domains, they are significantly limiting to DTI understanding and model explainability.

Additionally, most DTI prediction models simplify the interaction mechanism and do not take simultaneously into consideration the magnitude of certain local regions of each binding component and the interacting substructures involved [36]. Thus, the inter-dependency of the sequential and structural units of each binding component (and their intra-associations) or the inter-associations that revolve around the binding substructures (context of the interaction) are usually neglected, leading to predictions based on local and independent (without context) scattered motifs [37]. Consequently, when striving for valid and accurate DTI predictions, it is paramount to consider the proteomics, chemical, and pharmacological (interaction) spaces during the learning process of these complex architectures.

1.1.4 Binding Pockets in the Learning Process

The identification of protein-ligand binding pockets is crucial for understanding the biological functions of proteins and the mechanisms involved in DTIs, especially considering that these active regions are responsible to form bonds [38, 39, 40, 41]. Additionally, having *a priori* knowledge regarding potential binding pockets is essential for the rational design of new therapeutic compounds to modulate protein

functions [42, 43, 44]. However, the majority of DTA or DTI prediction computational studies have not yet given any special characterization or attention to these binding positions or actively integrated information regarding binding pockets during the learning/training process. Moreover, considering the range of different regions across the whole structure of the proteins and compounds, the relevance given to certain spots might introduce bias in the predictions, leading to the estimation of potential DTIs based on redundant sites (or substructures) and limiting the reliability of the results [45].

Furthermore, in order to realistically model DTIs and understand the interaction process, it is critical to properly integrate information related to binding sites during the learning process of these complex architectures, especially given the importance to learn the inter-associations that revolve around the binding substructures, i.e., the context of the interaction. Moreover, optimizing the DTI/DTA prediction frameworks by actively incorporating context related to the binding sites can increase the binding accuracy to the desired target and elucidate the design of ligands with increased selectivity and affinity for the corresponding target.

1.2 Goals and Contributions

This research aims to present innovative computational solutions to bolster drug discovery efforts, effectively addressing some of the primary challenges that persist in this domain. Consequently, it seeks to achieve informative and explainable modeling and prediction of DTIs. To this end, this thesis focused on the establishment of DL architectures specifically tailored for predicting an unbiased DTA metric, particularly K_d. These architectures rely on 1D raw sequential and structural representations of the proteins and compounds for the inferring process.

Moreover, this work delves into the realm of explainability of black-box models. It strives to provide explanations for the decision-making processes employed by the models and endeavors to incorporate explainability into the inner structure of the architectures during the model construction phase. Thus, the computational solutions presented seek to enhance the comprehensibility of the models' predictions, shed light on the reasoning behind the complex mechanisms involved in the inference step, and expand the current understanding of the intricate processes underlying DTIs.

Furthermore, this investigation methodically addressed the multi-domain representation space of DTIs, duly considering the inter-dependencies revolving around the

proteomics, chemical, and pharmacological domains. Alongside the multi-domain inter-dependency, the analysis and integration of information concerning binding pockets into the learning process were of paramount importance to proficiently modeling the pharmacological space and providing explicit evidence of potential key binding-related regions within proteins.

In sum, the reported findings in this thesis might prove helpful in designing future prospective DTI or DTA prediction applications capable of offering potential evidence to support the rationale behind the predictions and the intricate mechanisms and processes involved in the interaction between active compounds and biologically relevant targets.

This work can be subdivided into three main contributions, described in the following subsections.

1.2.1 Explainable Deep Drug–Target Representations

The first part of this thesis revolves around an investigation into the reliability of employing CNNs within the context of DTA prediction. Specifically, this examination focuses on their ability to identify and attribute significance to relevant sequential and structural regions associated with the binding process, all without relying on any *a priori* information. This study employs an end-to-end DL architecture to predict binding affinity, wherein CNNs are utilized to automatically recognize and extract discriminating deep representations from 1D sequential and structural data. Additionally, a post-hoc explainability algorithm is explored and proposed to provide potential explanations for the decision-making process of CNNs, as well as to identify the input regions that contributed positively to the inferential process. The input regions that were identified to positively influence the predictions underwent comprehensive analysis to evaluate their association with relevant regions in the DTI domain, specifically binding pockets and evolutionarily conserved motifs, and assess their respective significance (weight) to the overall prediction.

1.2.2 Intrinsic Explainability and Drug–Target Multi-Domain Inter-Dependency

The second part of this thesis presents a novel end-to-end Transformer-based architecture, designed to model the multi-domain inter-dependency associated with the proteomics, chemical, and pharmacological spaces. The architecture introduces self-attention layers to capture the short and long-term proteomics and chemical con-

text dependencies between the sequential and structural units of the proteins and compounds, respectively, and cross-attention layers to exchange information and learn the pharmacological context associated with the interaction space. Moreover, this study places emphasis on integrating explainability into the model construction process. This integration allows for the provision of various levels of potential understanding regarding DTIs as well as the reasoning behind the predictions. Furthermore, the resulting attention matrices were subjected to visualization to specifically evaluate the attention (weight) that the compound representation assigns to binding-related regions within the proteins.

1.2.3 Binding Region-Guided Strategy to Predict Drug–Target Affinity

The third part of this thesis is dedicated to actively incorporating binding-related information during the learning process of the architectures. To achieve this, an innovative binding-region-guided strategy is presented, aimed at modeling the pharmacological space of the interaction and learning the inter-dependency amongst binding-related positions for the prediction of binding affinity. This study explores a novel end-to-end binding-region-guided Transformer-based architecture that simultaneously predicts the 1D binding pocket and the binding affinity of DTI pairs, where the prediction of the 1D binding pocket guides and conditions the prediction of DTA. The architecture effectively combines multiple Transformer-Encoder blocks to capture and learn the proteomics, chemical, and pharmacological contexts. Moreover, the predicted 1D binding pocket conditions the attention mechanism of the Transformer-Encoder used to learn the pharmacological space in order to model the inter-dependency amongst binding-related positions. The dual focus of this study on two contextually related yet computationally distinct tasks, specifically binding pocket classification and binding affinity regression, leads to an enhanced understanding of DTIs and predictions, attributable to the nature of the attention blocks and the involvement of binding pocket prediction.

1.3 Thesis Outline/Structure

The remainder of this thesis proposal is structured as follows:

- Chapter 2 provides background information related to the proteomics, chemical, and pharmacological contexts associated with DTIs.
- Chapter 3 provides a showdown of the principal computational approaches used

for the prediction of DTIs and DTA. Several research works across different branches within the *in silico* drug discovery domain are detailed in this chapter.

- Chapter 4 details the importance of explainability in the ML and DL fields, along with some of the major strategies designed to provide explainability for the inferring process and/or predictions of the models.
- Chapter 5 introduces the major and recurring methodology employed throughout the research works associated with this thesis.
- Chapter 6 refers to the post-hoc explainability strategy employed to provide potential explanations for the decision-making process of CNNs in the context of DTA prediction. The reliability of the CNNs was evaluated by comparing input regions that had a positive influence on the prediction and relevant sequential regions in the DTI domain.
- Chapter 7 presents the development of an intrinsically explainable end-to-end DL architecture that models the multi-domain inter-dependency associated with the proteomics, chemical, and pharmacological spaces for the prediction of DTA.
- Chapter 8 describes the binding-region-guided strategy employed to model the pharmacological space of the interaction and learn the inter-dependency amongst binding-related positions for the prediction of binding affinity. It also presents the contributions of using two computational yet contextually related tasks for DTI domain representation and DTA prediction performance.
- Chapter 9 concludes this thesis by providing an overview of the main findings resulting from this research and the overall contribution of this work to the field of drug discovery. It also presents future research directions.

1.4 Scientific Contributions

Apart from contributing to the drug discovery process, one of the main goals of this research is to add value to science through the publishing and divulging of new results and methodologies in scientific journals and international and national conferences. During this thesis, several contributions to the field of DTA prediction and drug discovery were made. These include publishing main authored and co-authored articles in international peer-reviewed journals, participating in national conferences, and co-supervising master's degree theses.

Educational and scientific contributions in other research fields were also made. These comprise science communication activities to the general public.

All scientific contributions are enumerated in the following subsections.

1.4.1 Peer-Reviewed Scientific Articles

*J*1 **Monteiro, N.R.C.**, Ribeiro, B., Arrais, J.P.. "Deep Neural Network Architecture for Drug–Target Interaction Prediction", *Artificial Neural Networks and Machine Learning - ICANN 2019: Workshop and Special Sessions Springer International Publishing*, 804-809, doi: 10.1007/978-3-030-30493-5_76 (2019).

*J*2 **Monteiro, N.R.C.**, Ribeiro, B., Arrais, J.P.. "Drug–Target Interaction Prediction: End-to-End Deep Learning Approach", *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 18, 2364-2374, doi: 10.1109/TCBB.2020.2977335 (2020). IF: 3.702

*J*3 Torres, L., **Monteiro, N.R.C.**, Oliveira, J. L., Arrais, J.P., Ribeiro, B.. "Exploring a Siamese Neural Network Architecture for One-Shot Drug Discovery", *2020 IEEE 20th International Conference on Bioinformatics and Bioengineering (BIBE)*, 168-175, doi: 10.1109/BIBE50027.2020.00035 (2020).

*J*4 **Monteiro, N.R.C.**, Simões, C.J.V., Àvila, H.V., Abbasi, M., Oliveira, J. L., Arrais, J.P.. "Explainable deep drug–target representations for binding affinity prediction", *BMC Bioinformatics*, 23, 237, doi: 10.1186/s12859-022-04767-y (2022). IF: 3.327

*J*5 Abbasi, M., Santos, B. P., Pereira, T. C., Sofia, R., **Monteiro, N.R.C.**, Oliveira, Simões, C.J.V., Britos, R. M. M., Ribeiro, B., Oliveira, J. L., Arrais, J.P.. "Designing optimized drug candidates with Generative Adversarial Network", *Journal of cheminformatics*, 14, 40, doi: 10.1186/s13321-022-00623-6 (2022). IF: 8.489

*J*6 **Monteiro, N.R.C.**, Oliveira, J. L., Arrais, J.P.. "DTITR: End-to-end drug–target binding affinity prediction with transformers", *Computers in Biology and Medicine*, 147, 105772, doi: 10.1016/j.compbiomed.2022.105772 (2022). IF: 6.698

*J*7 **Monteiro, N.R.C.**, Pereira, T. O., Machado, A. C. D., Oliveira, J. L., Abbasi, M., Arrais, J.P.. "FSM-DDTR: End-to-End Feedback Strategy for Multi-Objective *De Novo* Drug Design using Transformers", *Computers in Biology*

and Medicine, 164, 107285, doi: 10.1016/j.compbiomed.2023.107285 (2023). IF: 6.698

*J*8 **Monteiro, N.R.C.**, Oliveira, J. L., Arrais, J.P.. "TAG-DTA: Binding-Region-Guided Strategy to Predict Drug–Target Affinity Using Transformers", *Expert Systems with Applications*, 238, 122334, doi: 10.1016/j.eswa.2023.122334 (2023). IF: 8.665

1.4.2 Scientific Articles under Preparation

*J*9 Almeida, B. C., **Monteiro, N.R.C.**, Motresku, U., Oliveira, J. L., Arrais, J.P.., Carvalho, A. "Zinc Ion Binding Site Prediction in Regulatory Proteins with Transformers". *Manuscript under preparation to be submitted to a scientific journal.* (2023).

1.4.3 Participation in Conferences

*C*1 *Poster presentation in national conference:* Torres, L., **Monteiro, N.R.C.**, Oliveira, J. L., Arrais, J.P., Ribeiro, B.. 'Siamese Neural Networks for One-Shot Drug Discovery", *Bioinformatics Open Days, 2020 (BOD 2020).*

*C*2 *Oral presentation in national conference:* **Monteiro, N.R.C.**, Machado, A. C. D., Abbasi, M., Arrais, J.P. "Multi-Optimized Drug Design Using Transformers", *Portuguese Conference on Pattern Recognition, 2022 (RECPAD 2022).*

*C*3 *Poster presentation in national conference:* **Monteiro, N.R.C.**, Oliveira, J. L., Arrais, J.P.. "Intrinsic Explainability and Multi-Domain Inter-Dependency: End-to-End Drug–Target Binding Affinity Prediction with Transformers", *Encontro com a Ciência e a Tecnologia em Portugal, 2023 (Ciência 2023).*

1.4.4 Master's Degree Theses Co-Supervision

*M*1 Machado, A. C. D. "End-to-End Transformer-based Approach for Optimized *De Novo* Drug Design", *Master Thesis dissertation, Faculty of Science and Technology of the University of Coimbra, Biomedical Engineering - Bioinformatics and Clinical Informatics* (2022).

*M*2 Coelho, G. S. "Deep Generative Models for Protein Repositioning", *Master Thesis dissertation, Faculty of Science and Technology of the University of*

Coimbra, Data Science and Engineering (2023).

1.4.5 Science Communication to the General Public

*G*1 *Invited Speaker:* "Artificial Intelligence in Portugal", *Mind Your Data, 2020 (MYD 2020).*

*G*2 *Invited Speaker:* "Alumni: Career Path & Academic Journey", *Alumni: Physics Department of the University of Coimbra, 2023 (Alumni-DF-UC 2023).*

Chapter 2

Proteomics, Chemical, and Pharmacological Contexts

This chapter introduces the main background concepts related to the proteomics, chemical, and pharmacological contexts associated with DTIs. Section 2.1 presents a brief overview of the role of proteins within every organism, the steps involved in protein synthesis, and the different levels of protein structure/complexity. Section 2.2 details the drug development process and the overall importance of compounds in the prevention and treatment of clinical conditions. Section 2.3 describes the most relevant processes involved in the interaction between active compounds and biological targets, including the various stages of the drug's action. Section 2.4 expounds upon the notion of binding affinity and introduces some of the most frequently employed affinity metrics. Section 2.5 presents certain principles pertaining to protein-ligand binding models, as well as the role of binding pockets in dictating interaction specificity.

2.1 Proteins Overview

The biological, chemical, and physiological balance within every organism is regulated and maintained by key working molecules known as proteins. These vital components reside in every organism cell and are responsible for several unique functions, including signaling, substance transport, biochemical reactions, immune responses, cell adhesion, cell cycle, and cell shape [46]. On that account, their biological activity is determined by their unique amino acid sequence, which is organized in one or more polypeptide chains, and by the interactions that occur within the chains, which determine the folding and, consequently, the structure [47]. Most proteins carry out their roles by interacting with other proteins or molecules, modifying their activity depending on the type of binding that occurs. Thus, the function rate of these complex biomolecules can easily be altered based on the interaction with

potential invasive ligands that dominate over the natural ligands at the binding regions, leading to the rise or decline of their natural function. Moreover, proteins are affected by chemical, biological, and environmental factors, which can lead to the loss of shape or functionality or even abnormal oscillations in their function rate [48, 49, 50].

2.1.1 Protein Synthesis

Protein synthesis is carried out inside the cell and it is divided into two main steps, namely transcription and translation. In the transcription phase, the protein-coding genetic information present in the DNA is transferred to the messenger RNA (mRNA), in which a strand of mRNA is created to complement a strand of DNA [46]. The transcription phase begins when the enzyme RNA polymerase and the necessary transcription factors bind to the promoter sequence, which defines the direction of transcription and indicates which DNA strand will be used for transcription (DNA template strand). Nevertheless, for the transcription step to take place, it is necessary to partially unwind the DNA double helix at the promoter region, i.e., break hydrogen linkages between annealed nucleotide bases. The RNA polymerase moves along the DNA template strand and adds RNA nucleotides to the mRNA strand based on complementary base pairing (hydrogen bonds), where the ribonucleotides (RNA nucleotides) are bonded together via phosphodiester linkages. Figure 2.1 depicts the transcription step of protein synthesis, in which a strand of RNA, specifically pre-mRNA, is synthesized to complement the template strand of duplex DNA.

The resulting mRNA (pre-mRNA) usually undergoes post-transcriptional modifications, which include 5' capping (methylated cap), RNA splicing, RNA editing, and polyadenylation (3' tail of adenine bases) [51, 52]. Moreover, the mature mRNA contains two untranslated regions at the 5' and 3' sides, which are associated with several functions, including translation efficiency, subcellular localization, and stability [53]. Figure 2.2 illustrates the post-transcriptional modifications applied to the pre-mRNA.

In the translation step, the genetic information of the mature mRNA is converted into a protein, i.e., the mRNA is used as a template to assemble a chain of amino acids in a specific order according to the genetic code. Each group of three nucleotides (triplets) in the mature mRNA constitutes a codon, and each codon specifies a particular amino acid (see Figure A.1 in Appendix A for more details). The translation phase occurs at the ribosomes (located in the cytoplasm), which are re-

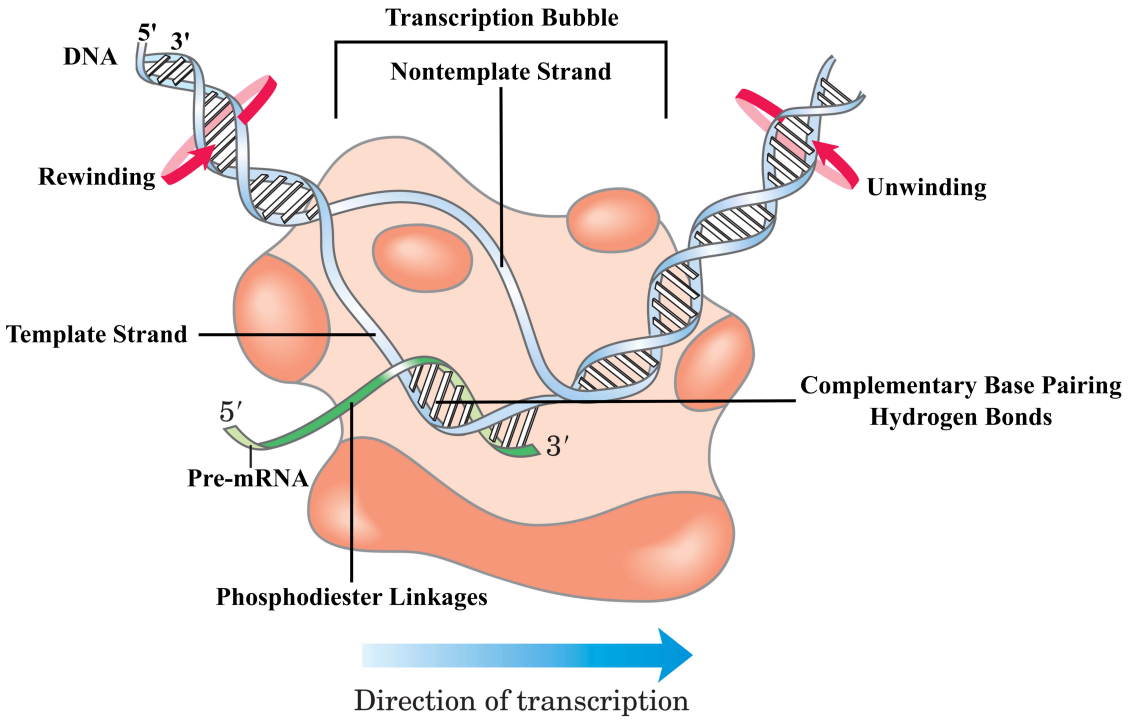


Figure 2.1: Protein synthesis: transcription step. The protein-coding genetic information present in the DNA is transferred to the mRNA. Figure adapted from *Lehninger Principles of Biochemistry, 5th Edition* [46].

sponsible to read the sequence of codons in the mature mRNA [46]. The ribosomal complex is constituted of an ribosomal RNA (rRNA) and proteins, and it is divided into two subunits. The small subunit is responsible to initiate the translation process, where the ribosome binding site of the mRNA binds to the small subunit, and holds the mRNA in place during translation. On the other hand, the larger subunit manages the elongation phase of the translation stage, which corresponds to the assembly of the chain of amino acids (linked by peptide bonds). The amino acids are brought by molecules of transfer RNA (tRNA) to the ribosomal complex according to complementary base pairing between the codons on the mRNA and the anticodons on the tRNA (see Figure A.2 in Appendix A for more details). The initiation and termination phases are based on specific initiation and termination codons of the mature mRNA, respectively, and different initiation, elongation, and termination factors are involved during the translation stage. Figure 2.3 illustrates the different stages of the translation phase of protein synthesis, specifically initiation, elongation, and termination, in which a polypeptide chain is synthesized according to the genetic information (codons) present in the mature mRNA.

Following protein synthesis, the resulting polypeptide chain may undergo additional post-translation events, e.g., proteolysis or protein folding, based on the biological

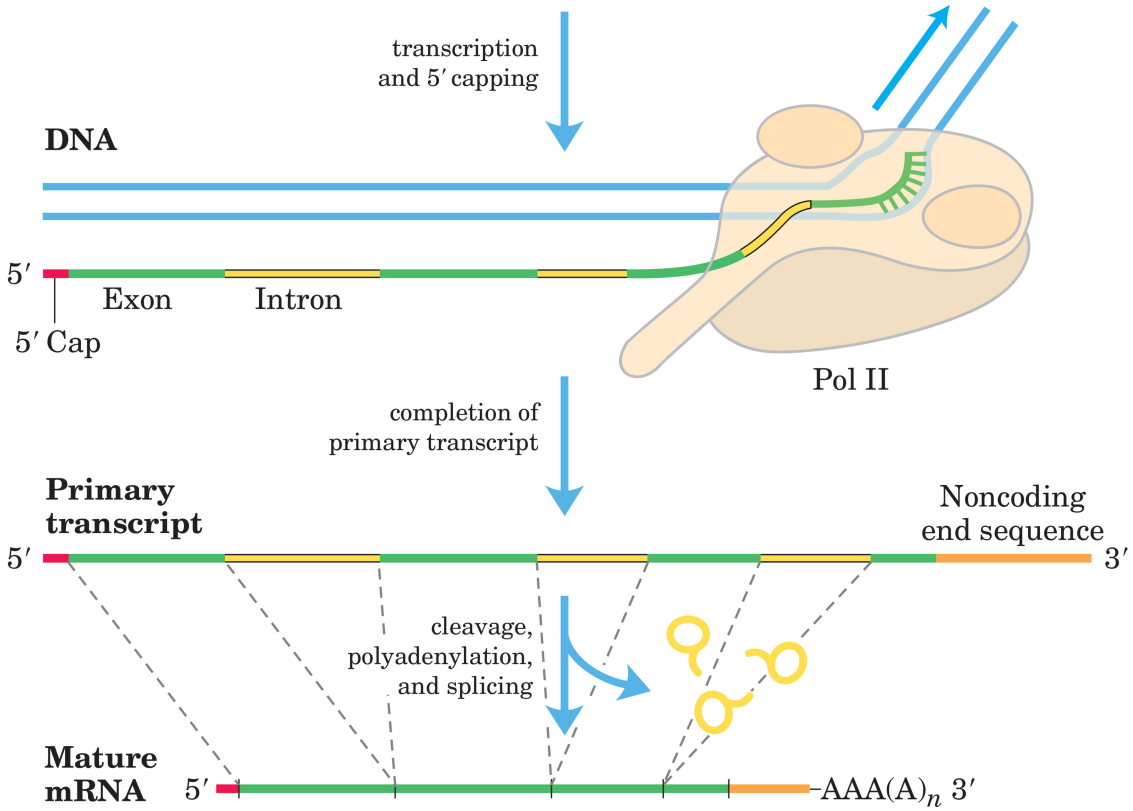


Figure 2.2: Pos-transcriptional modifications of the pre-mRNA: 5' capping, RNA splicing, RNA editing, and polyadenylation. Exon - coding region, Intro - non-coding region, Pol II - RNA polymerase II. Figure adapted from *Lehninger Principles of Biochemistry, 5th Edition* [46].

function to be carried out [54]. On that account, proteins can be free metabolites or constituents that have had common amino acid residues modified after protein synthesis.

2.1.2 Protein Structure

Proteins are composed of amino acids, where each amino acid contains an α-carboxyl group ($-COOH$), an α-amino group ($-NH_2$), a specific R group (side-chain), and a hydrogen atom attached to a central carbon atom (α-carbon). The amino acids are linked by peptide bonds, which are amide covalent chemical bonds between the α-amino group of one amino acid and the α-carboxyl group of another amino acid, resulting in the release of a molecule of water (H_2O) [46]. There are usually 20 amino acids commonly found as residues in proteins, however, other less common amino acids might also occur, including the use of placeholders when it is not possible to conclusively identify the residue (See Tables A.1 and A.2 in Appendix A for more

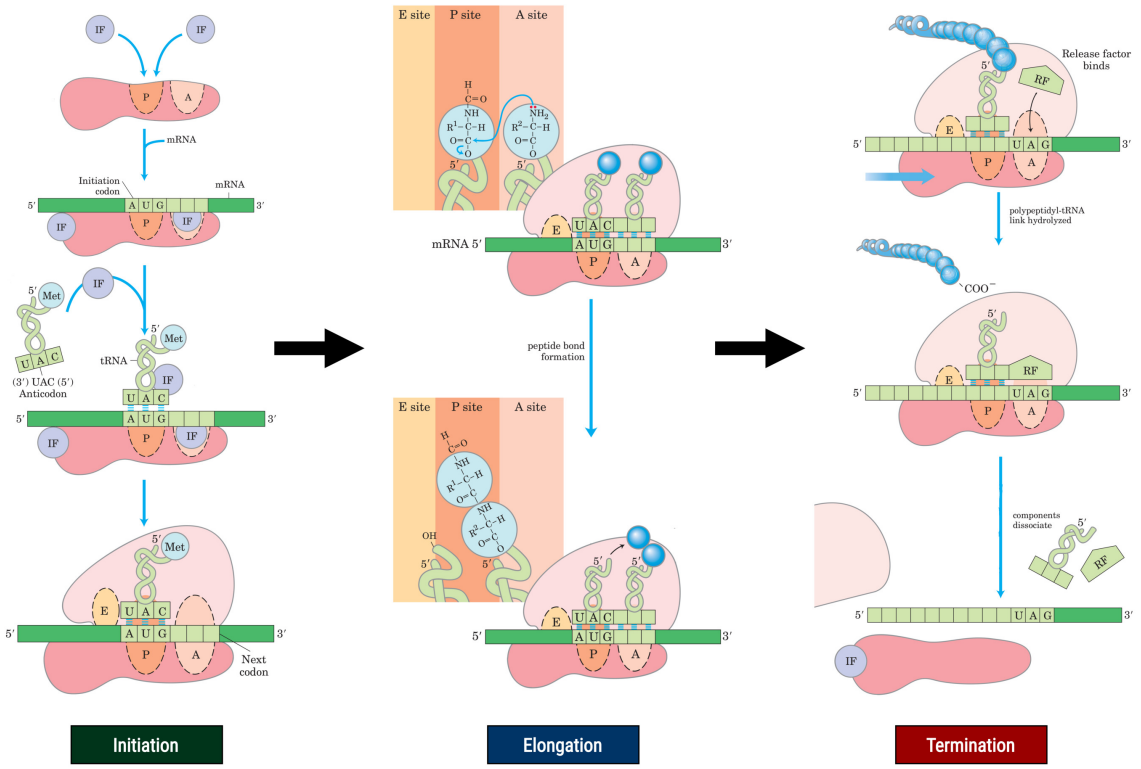
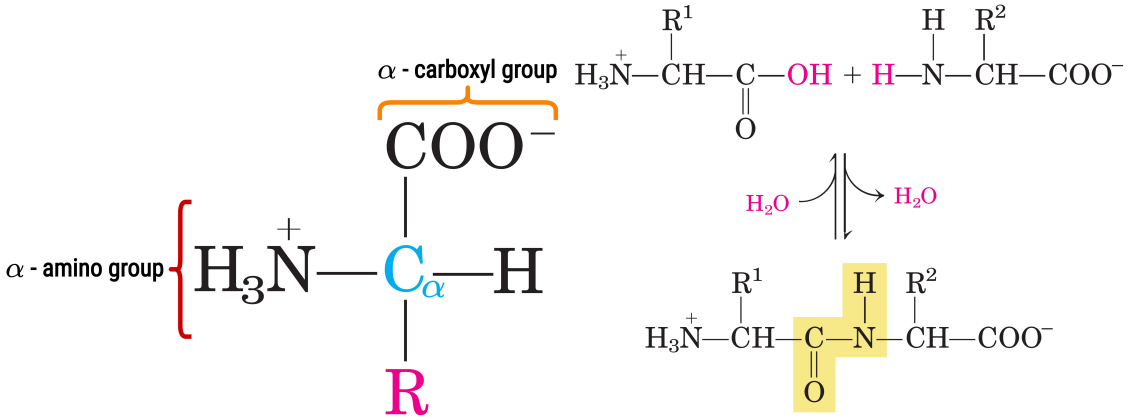


Figure 2.3: Protein synthesis: translation phase. In the initiation step, the small subunit binds to the mature mRNA, the anticodon of the tRNA binds to the initiation codon of the mRNA, and the larger subunit of the ribosome combines with the small subunit. In the elongation step, different amino acids are brought by tRNA molecules according to complementary base pairing between the codons on the mRNA and the anticodons on the tRNA, and peptide bonds are formed between the amino acids. The termination step occurs in response to a termination codon present in the mRNA, resulting in the release of the peptide chain and dissociation of the two subunits of the ribosome. Figure adapted from *Lehninger Principles of Biochemistry, 5th Edition* [46].

details). Moreover, amino acids can be grouped according to the characteristics of the side chains, e.g., polarity and charge at pH 7 [46], dipoles and volume of the side chains [55, 56], or other physicochemical/structural properties [57] (See Tables A.3, A.4, and A.5, respectively, in Appendix A for more details). Figure 2.4 depicts the structure of amino acids and the peptide bond that occurs between the α-amino group of one amino acid and the α-carboxyl group of another amino acid.

The protein structure is usually classified into four different levels of complexity, specifically primary, secondary, tertiary, and quaternary, in which the resulting protein shape or conformation is directly associated with its biological function/activity [46].

The primary structure corresponds to the linear sequence of amino acids in a



(a) **(b)**

Figure 2.4: Amino acids. a) Amino acid structure: α-carboxyl group (orange), α-amino group (red), R group (pink), and a hydrogen atom attached to a central α-carbon (blue). b) Peptide bond (yellow) between the α-amino group of one amino acid and the α-carboxyl group of another amino acid, resulting in the release of a molecule of water (H_2O). Figures adapted from *Lehninger Principles of Biochemistry, 5th Edition* [46].

polypeptide chain linked by peptide bonds (1D information), where each chain has its own set of amino acids and is assembled in a particular order. On that account, the primary structure is critical for the overall conformation of a protein, considering that the order of the side-chain structures and resulting interactions play a critical role in the folding of the protein into more complex structures, especially due to the different chemical properties associated with the R group of each amino acid. Moreover, the order in which the amino acids are connected defines a set of interactions between amino acids, which is crucial for the biological activity and properties of the protein.

The secondary structure is essentially determined by backbone interactions and hydrogen bonds, where the linear sequence of amino acids folds upon itself. On that account, local folded structures occur within the polypeptide chain due to hydrogen bonds between the partially negative oxygen atom and the partially positive nitrogen atom associated with backbone amino acid atoms, thus, amino acid side chains are not involved in these hydrogen bonds. Additionally, the hydrogen bonds can coil or fold the polypeptide chain, resulting in different patterns that contribute to the protein shape. The most common types of secondary structures are the α helix and β sheet, however, other patterns have been identified and categorized, such as β loops or β turns [58, 59, 60].

The tertiary structure is the overall 3D shape of a polypeptide and it is determined

primarily by the interactions that occur between the R groups of the amino acids. Hence, the properties of the R groups highly influence the protein's tertiary structure and global shape, e.g., polar hydrophilic and non-polar hydrophobic amino acids lead to hydrophobic interactions, in which non-polar hydrophobic R groups cluster together at the core of the protein, avoiding contact with the surrounding water. Moreover, several non-covalent interactions are involved in the tertiary structure, including hydrogen bonding, ionic bonding, dipole-dipole interactions, Van der Waals interactions, and London dispersion forces. On that account, the resulting 3D conformation depends on the global energy minimum and stability across all possible interactions between the amino acid residues.

The quaternary structure results from the combination of multiple polypeptide chains (subunits) into a single functional protein, in which mostly weak interactions, e.g., hydrogen bonding, are involved to maintain the structure. Even though this structure is relatively uncommon compared to single polypeptide chain proteins, it can lead to proteins capable of more complex functions, e.g., transporting oxygen throughout the blood (hemoglobin) or cell signaling (G-proteins), and increased stability [61, 62].

Figure 2.5 illustrates the four levels of protein structure, specifically primary, secondary, tertiary, and quaternary.

2.2 Drug Discovery

Drugs have been playing an important role in the overall health and survival of the human race, in which their use has been crucial for the treatment, prevention, and control of a broad range of diseases, illnesses, and other clinical conditions [65]. These substances were initially discovered by evaluating the mechanisms and effects of natural products, e.g., plants and mineral sources, or by mere serendipity [66]. Despite the medical properties of most of these natural compounds, the levels of toxicity were usually considerably high, resulting in harmful effects. In that regard, recognizing that the beneficial and toxic properties of a drug were important, especially that effective drugs should have a higher selectivity for the target microorganism instead of its host [67], was crucial to shifting drug discovery into finding active components within natural substances that account for their pharmacological properties, i.e., isolated products of higher purity [68, 69]. Additionally, microbial natural products or products derived from microbial compounds, i.e., antibiotics, have been vital for the human race's survival by rendering life-threatening bacte-

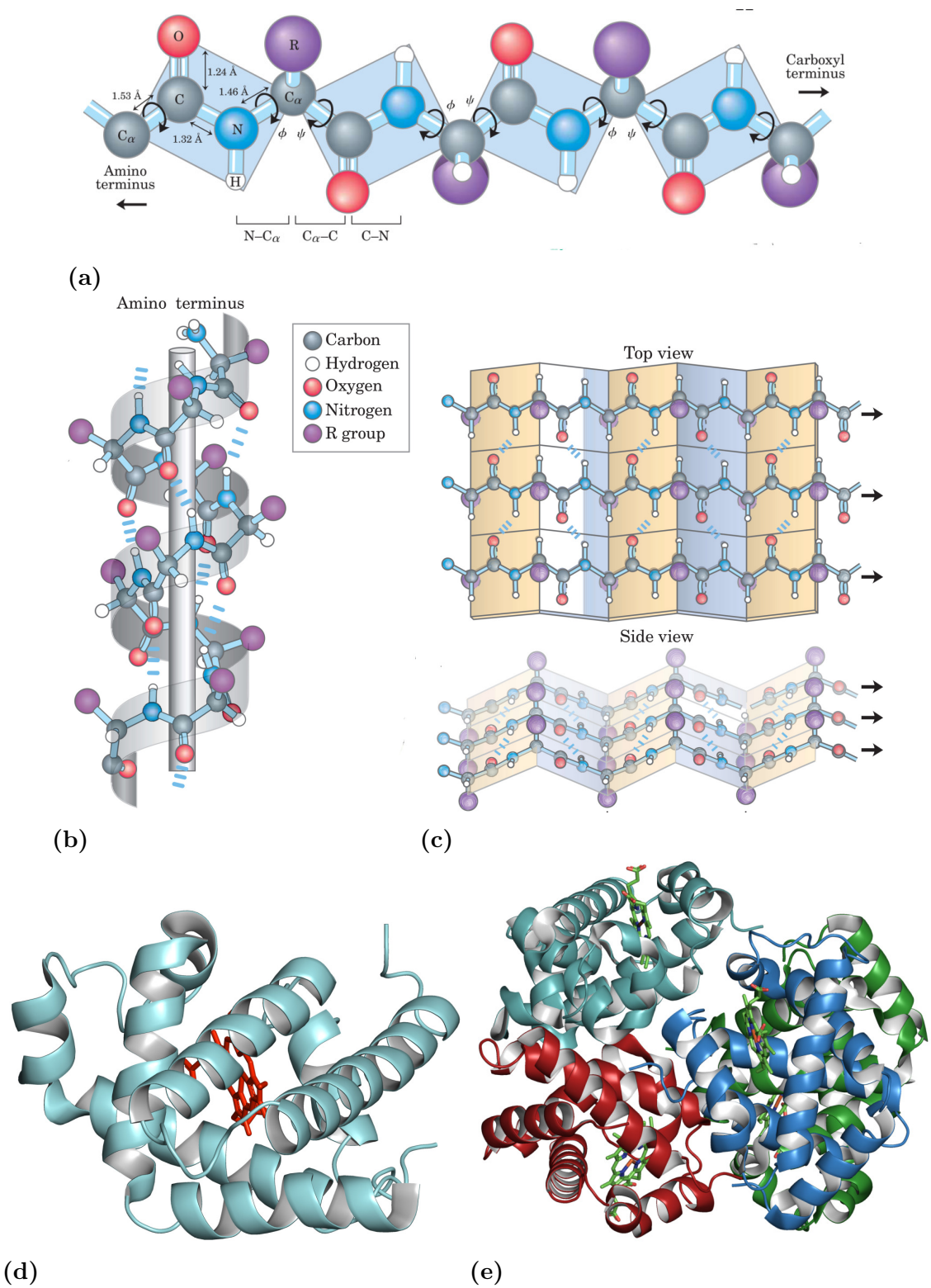


Figure 2.5: Four levels of protein structure. a) Primary structure. b) Secondary structure: α helix. c) Secondary structure: β sheet. c) Tertiary Structure: Myoglobin [63]. d) Quaternary structure: Hemoglobin [64]. Figures a), b), and c) adapted from *Lehninger Principles of Biochemistry, 5th Edition* [46].

rial infections efficiently curable. These compounds, starting with Penicillin, which was discovered by Sir Alexander Fleming in 1928 [70] and used to treat bacterial infections during World War II, were the pinnacle of modern medicine development, leading to major advances in surgery, transplants, and chemotherapy [10, 13].

Most drugs are now derived from combinatory chemistry, chemical synthesis, molecular modification of known drugs, rational design, or related to substances produced by certain microorganisms. Nevertheless, natural compounds isolated from natural sources still represent a major class of molecular drugs that are involved in the treatment of several categorized human diseases and are usually considered a starting point in the discovery of certain agents, e.g., immunosuppressive agents [69].

The drug discovery challenge includes the identification and development of molecules that elicit a certain desired effect in a living organism, the study of proteins involved in key biological pathways or related to certain diseases, and the evaluation and optimization of the organic functional groups and the pharmacophore of potential leads [71]. On that account, the development of High-Throughput Screening (HTS) methods in combination with combinatorial chemistry has led to major findings in the drug discovery field [72, 73, 74]. In the HTS step, compounds from large chemical libraries or collections of synthetic compounds are screened (assayed) against a variety of well-characterized targets in order to identify specific mechanisms of action, including inhibition, activation, modulation, or interference [75]. The compounds that show activity and selectivity in pharmacological and biochemically relevant screening are modified using parallel chemistry approaches (combinatorial chemistry) in an iterative process of synthesis, characterization, and screening to optimize their efficacy and drug-like properties (lead optimization), considering that hit compounds rarely cover the needs in affinity, selectivity, efficacy, and safety [76]. Thus, at each stage of the lead optimization process, Structure-Activity Relationship (SAR) or Quantitative Structure-Activity Relationship (QSAR) studies are usually conducted in order to ascertain physicochemical properties, molecular properties, and the interactions of the drug [77, 78, 79, 80].

The current drug development process is usually divided into six different steps: target discovery, lead discovery, lead optimization, pre-clinical, clinical trials, and regulatory approval [71]. Target discovery comprises studies to identify key molecules in a specific metabolic or cell signaling pathway related to a particular clinical state. Considering that drugs predominantly target proteins, including enzymes, receptors, or transporters, it is essential to accurately identify and evaluate the therapeutic effect and regulation of the discovered targets for the pharmacological action of a

certain drug. Lead discovery includes the identification of chemical compounds or molecules that interact with a certain target with high affinity, efficacy, and selectivity. These molecules should have properties that are likely to be therapeutically useful and are usually the starting point for drug design and development. Lead optimization aims to improve efficacy, selectivity, and pharmacokinetic features, i.e., Absorption, Distribution, Metabolism, and Excretion (ADME) properties of lead compounds. This stage comprises systematic modifications and refinements of the structure of lead compounds in order to evaluate the physicochemical and molecular properties and the relative contribution of each organic functional group. On that account, several analogs of the lead compounds are typically generated through various forms of alterations, wherein the impact of these modifications on the biological activity is assessed. These refinements are usually linked to the size and shape of the carbon skeleton, the spatial arrangement of the lead compound, and the type and extent of substitution. In the pre-clinical stage, *in vitro* and *in vivo* experiments are conducted prior to human consumption in order to determine and evaluate Absorption, Distribution, Metabolism, Excretion, and Toxicity (ADMET) properties, dosage, and efficacy. Clinical trials are conducted on humans and are usually divided into three phases. In phase I, the drug's safety, dosage, and toleration are evaluated in a group of healthy humans. In phase II, the drug is administered to patients that have the condition in order to assess its efficacy. In phase III, a larger sample of patients is selected to conduct a more reliable statistical analysis of the drug's effectiveness and potential side effects. Moreover, reproductive effects, teratogenicity, and immunologic and behavioral toxicities are also evaluated at this stage. Regulatory approval entails the registration and approval by a drug administration department, such as Food and Drug Administration (FDA) in the United States or European Medicines Evaluation Agency (EMEA) in Europe. Figure 2.6 illustrates the drug discovery pipeline.

The contributions in genomics, proteomics, and bioinformatics have been crucial to further development and findings in drug discovery given the rapid and precise discovery of genes/proteins involved in the etiology of certain diseases [81]. On that account, the use of certain proteomics technologies, such as mass spectroscopy, in combination with affinity chromatography or micro-array methods has been important for probing DTIs [82, 83]. Furthermore, X-ray crystallography and Nuclear Magnetic Resonance (NMR) methods have been essential for determining the 3D structure of proteins and/or drug–target complexes, revealing a high level of detail about potential active sites within the protein and providing important insights regarding the interaction between active small molecules and biologically relevant

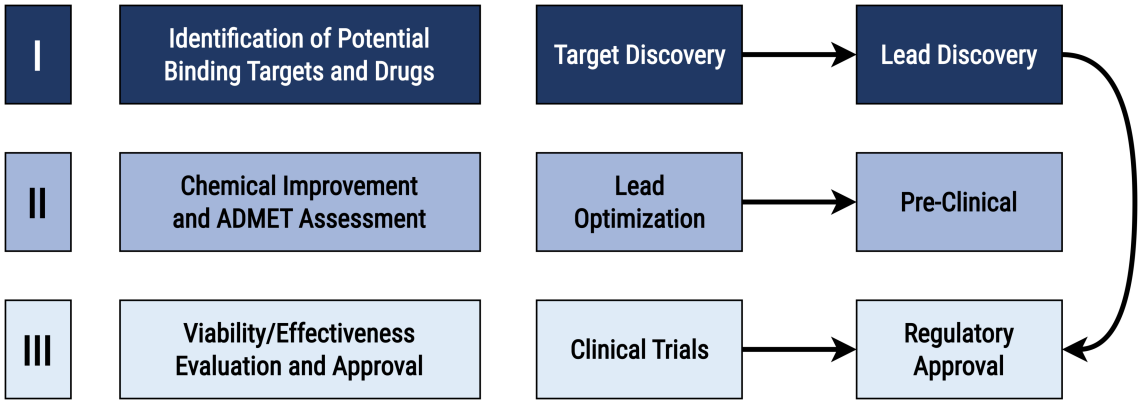


Figure 2.6: Overview of the drug development process. (I) Target and lead discovery focus on identifying which targets interact with a certain drug and which drugs bind to a certain target, respectively. (II) Lead optimization is associated with the improvement of the discovered active compound's chemical properties, including potency, selectivity, and pharmacokinetic attributes. The pre-clinical stage, known as ADMET assessment, enforces that several conditions for consumption are met. (III) The clinical trials comprise several stages (human trials) to meticulously evaluate the effectiveness and viability. The regulatory approval entails the registration and approval by a drug administration department, e.g., FDA or EMEA.

protein targets [84, 85]. NMR spectroscopy has also been effective for the generation of multivariate metabolites in order to improve lead compound selection in drug toxicity screening.

In spite of the interesting findings of screening methods, these traditional drug discovery approaches have a low rate of success. Hence, modern sources focus on computer-aided drug design given the advances in computational methods and increasing knowledge of disease etiology and biological systems associated with it [86]. Computer-aided drug design includes a variety of computational approaches, e.g., structure-based, ligand-based, or *de novo* design, and improves the identification of new leads and design of compounds with increased selectivity, efficacy, and safety [87]. Furthermore, these *in silico* methodologies are usually combined with drug repositioning strategies, making use of well-established chemical substances with known bioactivity as lead or prototype [88, 89, 90]. On that account, it greatly reduces the cost of lead screening and the time required for a drug to reach the market, considering that the search space is around structural congeners, homologs, or analogs of known and approved drugs.

2.3 Pharmacological Activity and Drug–Target Interactions

Drugs are identified as any substance that is used to explore, modify, or control physiological systems or pathological states for the benefit of its host. On that account, these chemical compounds can be applied to a broad range of different conditions, including the provision of certain elements, prevention of a disease or infection, treatment against a bacterial or virus infection, temporary blocking of a normal function, correction of a derange function, or even as diagnostic auxiliary agents [91].

The structure of a drug usually comprises a carbon skeleton and different organic functional groups, where the former is responsible for the size and shape of the molecule and the latter for the overall molecular reactivity [92]. On that account, the structure is usually divided into two main components: pharmacophore and secondary substructures. The pharmacophore is involved in the binding process and is thus responsible for the biological/pharmacological response of the drug. On the other hand, the secondary substructures are associated with transport, storage, bioavailability, chemical and metabolic stability, excretion, and interactions with secondary receptors, which are essential to regulate the effects of the drug molecule in the body [93, 94].

The pharmacological action of a drug is a complex pattern of processes in which several factors are involved, thus, it is not only related to the intrinsic properties of the compounds but also to the interaction with the complementary chemical groups of a specific cellular component (receptor), which initiates various biological and physiological modifications, altering the function rate of that receptor. The drug's action is usually divided into three main phases: pharmaceutical, pharmacokinetic, and pharmacodynamic [76]. In the pharmaceutical phase, the drug's administered form is disintegrated and the compound is pharmaceutical available, i.e., available for absorption. The pharmacokinetic stage is associated with the ADME steps/properties of the compound, in which the drug is available for action (biological availability). Moreover, compounds are usually either stored within the host (drug's metabolism), wherein they can remain intact or undergo further chemical modifications, or eliminated after a certain period of time (drug's excretion). The pharmacodynamic step is associated with the interaction of the active small molecule with its receptor, leading to a series of chemical and biochemical phenomena in order to produce a specific biological/pharmacological effect.

The interaction between a protein and a drug is the consequence of several bond types, including ionic, hydrogen, hydrophobic, Van der Waals, and/or covalent, in which the complementarity of certain active and functional groups in the 3D space is essential to form bonds, i.e., the effectiveness of the drug depends on the complementarity between the molecular shape and stereoelectronic structure (electronic distribution) of the chemical compound and the stereoelectronic structure of the receptor [33]. On that account, the 3D orientation and arrangement of the organic functional groups are determinants for the interaction, e.g., enantiomers and diastereomers of a certain chemical compound usually result in significant differences in the pharmacokinetic and pharmacodynamic behavior and physical and chemical properties, respectively [95, 96]. Nevertheless, several other factors affect the interaction, including physical, chemical, and physiological. The interaction process is usually divided into two types of binding, specifically primary and secondary, where the former is responsible for a firm binding, e.g., ionic bonds, and the latter for supplementary bonds to hold the compound within the interaction complex. Additionally, DTIs can be reversible or irreversible depending on the dynamic equilibrium of the complex, the type of binding bonds that occurs, and the binding region within the protein [97, 98].

Proteins contain pockets, cavities, surface depressions, and other geometric regions where small molecule compounds can easily bind. Thus, not all regions within the protein are responsible to form bonds, only specific spots, denominated of active or binding sites, interact with the drug [99]. However, certain chemical compounds can bind to other areas within the protein surface (allosteric sites), leading to changes in the conformation of the protein binding sites [100]. Moreover, most proteins undergo conformational changes when binding to a substrate, e.g., adjust their shape, in order to promote the interaction [101, 102].

Overall, drugs are potential modulators of the functions performed by several proteins, in which their ability to bind (affinity) and capacity to execute their pharmacological activity (intrinsic activity) determine the role enforced by these chemical molecules. Thus, drugs can be classified as agonists, antagonists, or partial antagonists [103]. Agonists are associated with molecules that induce a biological response upon binding to a receptor, whereas antagonists inhibit or decrease the physiological reactions of the receptor. Nevertheless, some drugs are capable of producing their effects without interacting with a specific receptor (lack of affinity and intrinsic activity).

2.4 Binding Affinity

The interaction between active small compounds and biologically relevant targets is determined by comprehensive processes and factors, including intermolecular interactions and energies, concentrations, and conformations, that are heavily reflected in the binding affinity or bioactivity of the ligand. However, these aspects are not adequately captured in binary relationships, which indicate only the presence or absence of interaction. Hence, it is crucial to consider the binding affinity in order to quantify the strength of the association between compounds and targets and assess the magnitude and rank order of the interaction. On that account, the ligand's binding affinity is vital to differentiate primary interactions from those with secondary targets (known as off-targets). This distinction helps in understanding the specific target of interest and avoids potential confusion by interactions with unintended targets.

The measure of the binding affinity in *in silico* studies is usually calculated using three different metrics: K_d, K_i, and IC_{50} [104]. K_d is a direct measurement of the equilibrium between the receptor-ligand complex and the dissociation components, where lower values indicate higher binding affinity [24, 25]. Considering a protein P, a ligand L, and a protein-ligand complex PL, K_d can be expressed as:

$$\begin{gathered} P + L \underset{k_{on}}{\overset{k_{off}}{\rightleftharpoons}} PL \\ K_d = \frac{[P][L]}{[PL]} = \frac{k_{off}}{k_{on}} = \frac{1}{K_a} \end{gathered} \tag{2.1}$$

, where $[P]$ is the equilibrium concentration of a protein molecule P, $[L]$ is the equilibrium concentration of a ligand molecule L, $[PL]$ is the equilibrium concentration of a protein-ligand molecule PL, k_{off} is the dissociation rate constant, k_{on} is the association rate constant, and K_a is the binding/affinity constant. Moreover, K_d is usually given in terms of molar concentration (M).

K_i is also seen as a dissociation constant but measured in inhibition studies and, therefore, depends on the kinetic mechanisms of inhibition, such as competitive inhibition, uncompetitive inhibition, non-competitive inhibition, or mixed inhibition [21, 22]. In this regard, only when the kinetic mechanism is accurately identified, can K_i values provide an accurate representation of the binding constant. This metric is usually associated with ligands that reduce the catalytic activity of enzymes and, thus, it is calculated by determining rates of enzyme-catalyzed reactions while

independently varying the concentration of substrate and inhibitor. Low values of K_i indicate a strong binding association. On the other hand, IC_{50} is the concentration of inhibitor required to reduce the biological activity to half of the uninhibited value, and it is affected by the measurement conditions, mechanisms of inhibition, and concentrations [23]. Contrarily to K_i, IC_{50} is determined at only one concentration of substrate over a range of inhibitor concentrations. Low values of IC_{50} are associated with higher binding affinity.

IC_{50} can be converted to K_i (and vice-versa) using the Cheng–Prusoff equation [105] when the mechanism of inhibition and the concentration of the substrate are known [21]:

- Competitive inhibition: inhibitor binds only to free enzyme.

$$IC_{50} = K_i(1 + \frac{[S]}{K_M}) \tag{2.2}$$

  , where $[S]$ is the concentration of the substrate and K_M is the Michaelis constant.

- Uncompetitive inhibition: inhibitor binds only to the enzyme-substrate complex.

$$IC_{50} = K_i(1 + \frac{K_M}{[S]}) \tag{2.3}$$

  , where $[S]$ is the concentration of the substrate and K_M is the Michaelis constant.

- Mixed inhibition: inhibitor binds to both free enzyme and enzyme-substrate complex with different inhibition constants.

$$IC_{50} = \frac{[S] + K_M}{(\frac{[S]}{K_{ies}} + \frac{K_M}{K_{ie}})} \tag{2.4}$$

  where $[S]$ is the concentration of the substrate, K_M is the Michaelis constant, K_{ies} is the inhibition constant to the enzyme-substrate complex, and K_{ie} is the inhibition constant to the free enzyme.

- Non-competitive inhibition: a special case of mixed inhibition where substrate binding does not affect inhibitor binding.

$$IC_{50} = K_i \tag{2.5}$$

The equations above are obtained by using derivations of Michaelis–Menten kinetics [106] in order to relate the rate of reaction to the concentration of inhibitor. Consid-

ering an enzyme E, a substrate S, an enzyme-substrate complex ES, and a product P, the standard Michaelis–Menten equation (without any competitive inhibitor) can be expressed as:

$$\begin{gathered} E + S \underset{k_{on}}{\overset{k_{off}}{\rightleftharpoons}} ES \xrightarrow{k_{cat}} E + P \\ V_0 = \frac{V_{max}[S]}{K_M + [S]} \end{gathered} \tag{2.6}$$

, where V_0 is the initial reaction rate, $[S]$ is the concentration of the substrate, K_M is the Michaelis constant, V_{max} is the maximum reaction rate, k_{off} is the dissociation rate constant, k_{on} is the association rate constant, and k_{cat} is the catalytic rate constant. The Michaelis constant can be expressed as:

$$K_M = \frac{k_{off} + k_{cat}}{k_{of}} \tag{2.7}$$

, where k_{off} is the dissociation rate constant, k_{on} is the association rate constant, and k_{cat} is the catalytic rate constant. Given the similarities between K_d and K_M, lower K_M values are usually associated with a higher affinity of the enzymes for the substrates.

2.5 Protein-Ligand Binding Models and Binding Pockets

Apart from the complementary of certain active and functional groups in the 3D space, there are several factors involved in the interaction between proteins and compounds, including physicochemical mechanisms, binding kinetics, thermodynamic profiles, binding driving forces, enthalpy-entropy compensations, and interactions with the adjacent molecules of the surrounding environment, that affect the overall stability of the encounter protein-ligand complex [107, 108]. In that regard, the driving forces leading the association between proteins and ligands are intricately derived from a culmination of diverse interactions and energy exchanges involving the protein, ligand, and adjacent molecules (surrounding environment) [24]. The stability of the resulting protein-ligand complex is usually measured by the magnitude of the negative change in Gibbs free energy (ΔG), i.e., lower values of free energy upon binding are associated with a higher stability of the complex [109, 110]. This measure, corresponding to the difference of energy between bound and unbound

states, can be expressed using enthalpic and entropic contributions:

$$\Delta G = \Delta H - T\Delta S \tag{2.8}$$

, where ΔH is the binding enthalpy, ΔS is the binding entropy, and T in the temperature in Kelvin.

The binding enthalpy is closely associated with the energy changes that arise from the establishment of noncovalent interactions, namely Van der Waals interactions, hydrogen bonds, ionic bonds, and polar or apolar interactions, occurring between the protein and ligand at the binding interface [111]. The disruption of individual noncovalent bonds between the protein and/or ligand with neighboring molecules can also influence the binding enthalpy favorably or unfavorably [112, 113]. Conversely, the binding entropy encompasses the overall increase or decrease in the degrees of freedom exhibited by the proteins, ligands, and adjacent molecules within the surrounding environment [114]. Changes in binding entropy are commonly characterized by various entropic terms, including alterations in the entropy of the surrounding environment, changes in conformational entropy associated with the modification of conformational freedom in both the protein and ligand upon binding, and translational and rotational entropy changes reflecting the loss of translational and rotational degrees of freedom of the protein and ligand upon formation of the complex [115, 116, 117]. Similar to the binding enthalpy, the binding entropy change can have either favorable or unfavorable contributions to the binding free energy. Nonetheless, the overall stability of the resulting binding complex is also influenced by factors such as the structural and thermodynamic properties of the surrounding environment and neighboring molecules, the flexibility of the binding pocket and adjacent regions, the molecular structure and conformation of the ligand, and fluctuations in intermolecular forces throughout the binding process [118, 119, 120]

Given the profound impact of binding thermodynamics and kinetics, and protein dynamics on the binding pockets, various models and mechanics for protein-ligand binding have been proposed and extensively explored, namely the *lock and key* [121, 122], *induced fit* [123, 122], and *conformational selection* [124, 125, 126] models. In the *lock and key* model, both the protein and ligand are regarded as rigid entities, and the binding surfaces exhibit perfect complementarity. Consequently, only ligands that possess precise size and shape can bind to the protein at the binding pocket. This model predominantly relies on entropy-driven processes, as the interaction is primarily dictated by factors such as size, shape, and surface characteristics [127, 128, 129]. Moreover, the most significant contribution to the negative

change in binding free energy of the resulting complex stems from alterations in the entropy of the surrounding environment [130]. In the *induced fit* approach, the receptor's flexibility is taken into consideration, in which the binding pocket within the protein is capable of undergoing conformational changes upon ligand binding to promote the interaction [131]. Contrarily to the *lock and key* model, *induced fit* is predominantly driven by enthalpy factors due to the formation of strong bonds during the restructuring process of the receptor to create a compatible binding site, resulting in a negative change in enthalpy [132, 133, 134]. The *conformational selection* model acknowledges that most proteins inherently exhibit dynamic behavior and significant conformational flexibility [135, 136]. Accordingly, proteins continuously transition between different conformational states or substates that possess similar energies (multiple free energy minima) [137, 138]. To facilitate the interaction, the protein undergoes conformational changes, and the ligand selectively binds to the most suitable conformational state or substate [139]. This protein-ligand binding model encompasses contributions from both entropy and enthalpy factors, which collectively enhance the stability of the protein-ligand complex. Nevertheless, these binding mechanisms may coexist concurrently or occur sequentially, depending on the molecular interaction involved or the context of the interaction [140, 141, 142, 143].

Binding pockets are identified as specific regions (cavities) within the surface or on the interior of a protein that exhibit favorable characteristics for accommodating a ligand. The properties of a binding pocket, including its physicochemical attributes, shape, and positioning within the protein, collectively dictate its functional role [144]. The composition of amino acids within the binding pocket plays a crucial role in determining its properties and, consequently, its ability to effectively bind a ligand [145, 146]. Nevertheless, residues in the neighborhood of the binding pockets can also have long-range effects on the properties of the binding sites, influencing their ligandability [147]. Even though the overall complementarity, geometry, and properties of the binding pockets are important for ligand binding, the intrinsic protein flexibility and conformational adjustments greatly affect receptor-drug binding thermodynamics and kinetics [101, 102, 148, 149].

The mobility of proteins can induce the opening, closing, and adaptation of the binding pocket to regulate the binding process and specific protein functions. Furthermore, the inherent flexibility of proteins can result in subtle modifications to preexisting pockets or even the creation of entirely new pockets [150]. On that account, various categories of binding pockets associated with different protein dynamics properties have been identified: subpocket, adjacent pocket, breathing motion,

channel/tunnel, and allosteric site. Subpockets and adjacents pockets are associated with the appearance/disappearance of pockets in an already existing pocket or the neighborhood of an already existing pocket, respectively. Breathing motion is related to protein motions that lead to the enlargement or contraction of the original binding site. Channels or tunnels correspond to the opening or closing of certain structural gates within the protein structure, i.e., connecting the binding pocket inside the protein with the surrounding environment, that allows or blocks the entrance of the ligand. The allosteric pocket is usually located at another site of the protein, leading to conformational changes of the original pocket or competing with the original pocket at different rates.

Considering the dynamics of several proteins and the distribution of possible conformations with similar energies, some of the aforementioned classes of binding pockets related to small ligands are usually associated with transient states due to the stabilization of energy contributions [43, 151]. On that account, it is challenging to identify and properly characterize these binding pockets, and discover and/or design potential leads that selectively bind to these regions with high affinity [152, 153]. Thus, the identification of protein-ligand binding pockets is crucial for elucidating the biological functions of proteins and the mechanisms involved in DTIs [154, 155, 156, 157]. In that regard, several *in silico* solutions for predicting binding pockets have been proposed and are usually characterized by the strategy of the algorithm, e.g., geometric, template, or learning-based (ML), or by the level of structure data, i.e., sequence or 3D structure-based [158, 159, 160, 161, 162, 163, 164].

Chapter 3

In Silico Drug Discovery

This chapter provides an overview of the state-of-the-art in DTI and DTA prediction, presenting several research works across different branches within the computational drug discovery domain. Section 3.1 briefly describes the standard workflow associated with *in silico* DTI or DTA prediction. Section 3.2 presents various studies centered on the use of 3D structures and complexes for the inferring process. Section 3.3 introduces a range of research works stemming from property-activity similarity concepts. Section 3.4 presents multiple studies that leverage the vast amount of properties available/known to characterize proteins and compounds. Section 3.5 features a selection of studies exclusively related to DTA prediction.

3.1 Computational Drug Discovery Workflow

In silico methods have greatly influenced the drug development pipeline, accelerating the identification of potential DTIs and the discovery of new leads. Considering that several internal and external factors are involved during the binding process of active small molecules and biologically relevant targets, different perspectives and approaches have been proposed over the past years to solve the challenge of identifying new DTIs. Computational methods model this interaction using structural, biological, topological, and/or physicochemical properties as well as an experimentally validated characterization of that interaction, which can either be a binary association or a bioactivity metric. Depending on the type of information used to define and characterize the compounds and/or proteins, these methods can be broadly classified into three dominant categories: Structure-based, Ligand-based, and Chemogenomic/Proteochemometric (PCM) approaches. Figure 3.1 depicts the computational drug discovery workflow.

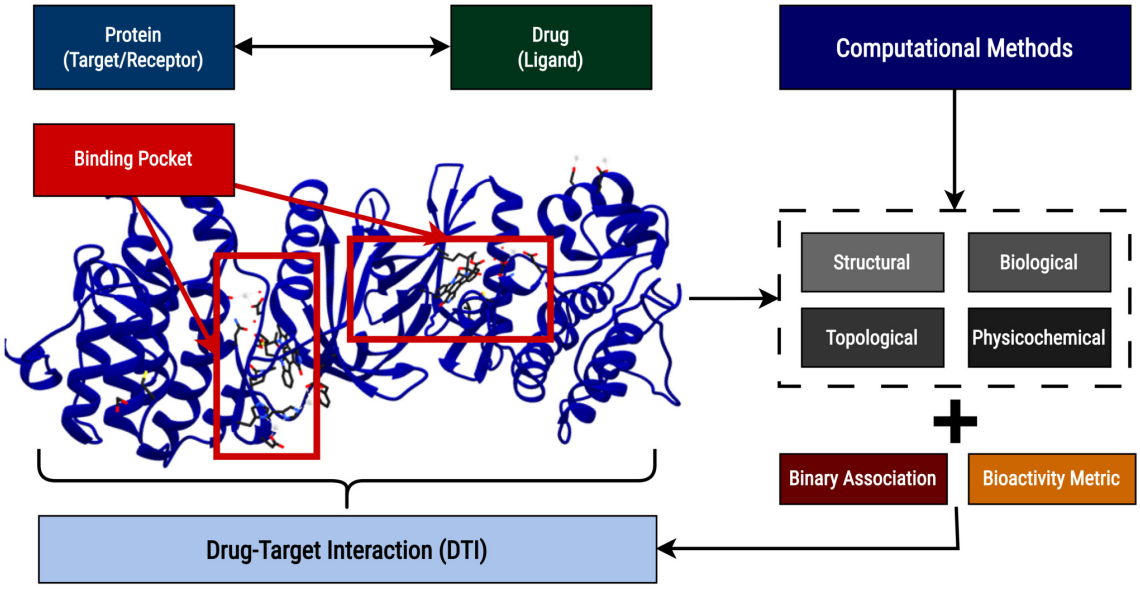


Figure 3.1: Computational Drug Discovery Workflow.

3.2 Structure-based

Structure-based approaches, commonly known as docking simulation, are primarily used to predict the 3D structure of receptor-ligand complexes. These methods simulate the interaction between the receptor and the ligand and score it based on the intermolecular energy and individual contributions from each binding component [176, 177]. The coordinates of the receptor and ligand are used to predict the resulting complex's coordinates based on known potential binding sites, e.g., derived from X-ray crystallography cognate structures, in which the docking search box is centered on these regions (guided docking), or, on the other hand, blindly docking onto the receptor structure, when there is no knowledge available regarding potential binding locations or 3D structures of a complex of the receptor (blind docking) [178]. Docking methods differ in terms of the molecular flexibility considered, the direction of docking (forward or reverse), and the scoring function used.

When predicting the resulting receptor-ligand complex, different degrees of molecular flexibility can be considered: receptor and ligand both rigid, receptor and ligand both flexible, or receptor rigid and ligand flexible [179]. Proteins are in constant motion between different conformation states with similar energies and change their conformation to promote the interaction, thus, it is important to account for receptor flexibility. However, this presents a challenge in most docking simulation approaches due to the number of degrees of freedom associated with all the different possible conformations.

Furthermore, the lack of knowledge regarding the 3D structure of proteins and the number and complexity of possible conformations pose challenges to this methodology. Nevertheless, some approaches can overcome the lack of information associated with the receptor's 3D structure, e.g., homology modeling [180] or AlphaFold [181]. Homology modeling predicts the 3D structure of the receptor based on proteins with high sequence similarity and known 3D structure, however, most of the resulting structures are unreliable given the protein folding complexity [182]. On the other hand, AlphaFold directly predicts the 3D coordinates of all heavy atoms for a given protein using the primary amino acid sequence as input, in which a confidence score is assigned for each residue. Despite the high average confidence score obtained in several structures predicted by AlphaFold, various proteins still present low confidence scores across most 3D predicted substructures.

Moreover, the score functions used in docking simulations often apply various assumptions and simplifications, in which certain energetic or geometric terms are usually not considered due to the computational cost of employing a highly accurate scoring function. Hence, this compromise in the scoring function's accuracy can impact the reliability and validity of the predictions [183].

In spite of the aforementioned limitations, structure-based approaches remain a valuable tool for modeling and predicting DTIs, offering a realistic approach to tackle DTIs. Additionally, they are still widely employed in structure-based drug design due to their ability to provide detailed information about potential active sites and the overall binding process.

Li et al. (2006) [184] developed a valuable tool called Target Fishing Dock (TarFisDock) for target identification. This approach combines a database of potential drug targets with a reverse ligand-protein docking approach to seek and identify potential protein targets for a given small molecule. TarFisDock generates a list of protein targets and then performs docking of the small molecule into the potential binding sites of these proteins. It calculates the interaction energy of the resulting complex using Van der Waals and electrostatic interaction terms in order to assign a score. The database used in TarFisDock consists of proteins that have known 3D structures and have been previously identified as targets in several therapeutic areas. This method takes into account only the flexibility of the ligand and not the flexibility of the receptor. Moreover, TarFisDock has successfully identified targets for vitamin E, and 4H-tamoxifen, which is commonly used in the treatment of breast cancer.

Wang et al. (2012) [185] introduced a web-based tool called idTarget, which uses

reverse docking to screen a vast number of protein structures available in the PDB [186]. This approach employs a divide-and-conquer docking strategy, where the entire receptor surface is explored using overlapping grid maps to identify multiple potential binding sites. For each binding site, an affinity profile is generated to assess the prediction confidence of the target screening results. The scoring function, based on robust versions of the AutoDock scoring function [187], incorporates atomic charges obtained from quantum chemical calculations and employs robust regression analysis to minimize the impact of outliers. To validate the performance of idTarget, the tool was evaluated with an HIV-1 protease inhibitor (DRV), an inhibitor of various protein kinases (6BIO), and an inhibitor of histone deacetylase 2 (LLX).

Gowthaman et al. (2016) [188] discovered inhibitors, validated through biochemical assays, for the human antiapoptotic protein Mcl-1. This protein plays a crucial role in treating various cancers, either as a single agent or in combination with other inhibitors. The proposed method, Docking Approach using Ray-Casting (DARC), involves matching the topography of a surface pocket in the protein with that of a potential ligand using a set of rays originating from a specific point within the protein (binding pocket topography mapping). This method evaluates the intersection of the ray set (from the same origin point within the protein) with both the pocket and the ligand, in which the shape complementary is evaluated based on distance parameters. The origin point is determined by considering the pocket's position within the protein and it is placed directly below the pocket on the protein side. Additionally, the optimal position and orientation of the ligand are determined through derivative-free minimization using the particle-swarm optimizer from Rosetta [189], along with a set of pre-built ligand conformers, i.e., the internal degrees of freedom of the ligand are fixed.

Wang et al. (2019) [190] proposed a consensus inverse docking strategy known as ACID for drug repurposing. This approach combines multiple docking methods, namely AutoDock Vina [191], LEDOCK [192], PLANTS [193], and PSOVina [194], to generate and explore several potential binding poses. Considering that each docking method uses distinct conformation search algorithms and scoring functions, the integration of their results can lead to overall improved performance. The docking results from each method are clustered and merged using an iterative conformational cluster-vote strategy. This strategy identifies the conformational cluster with the highest number of votes, which is then used for binding affinity (binding free energy) calculations based on Molecular Mechanics/Poisson–Boltzmann Surface Area (MM/PBSA) and X-SCORE methods [195, 196].

Zhang et al. (2020) [197] developed a blind protein-ligand docking approach called EDock, which relies on Replica-Exchange Monte Carlo (REMC) simulations [198]. EDock leverages the 3D structure of a given protein or a protein model (generated/predicted) to predict potential binding sites and initial ligand poses using sequence profiling and structure-based comparison searches, in which an initial conformation is generated through graph matching. The final pose is determined by conducting REMC simulations over the initial conformation, involving extensive docking conformation searching and structure refinement. To enhance the reliability of low-resolution docking with predicted protein structures, Van der Waals weightings and binding site distance constraints are incorporated.

3.3 Ligand-based

Ligand-based approaches extrapolate potential interactions by comparing a new ligand with known protein ligands. These methods are based on the premise that similar compounds possess similar properties and, thus, should exhibit similar bioactivity and bind to the same group of proteins [199]. However, they are heavily dependent on the amount of known and available ligands, performing poorly when this number is scarce or there is a lack of knowledge regarding known interactions.

These methods rely on a similarity measure and usually follow QSAR principles, which state that variations in the biological activity associated with a group of ligands are related to variations in their structural, physical, and/or chemical properties. QSAR approaches focus on finding a model, e.g., a statistical-based model, capable of determining the correlation between chemical structures and biological activity [200]. The quality of these models greatly depends on the selection of relevant and discriminating descriptors, which can be related to either molecular (2D QSAR) or 3D geometric (3D QSAR) properties. Furthermore, the choice of a suitable linear or non-linear mapping model is crucial to address the specific problem at hand [79].

Afantitis et al. (2006) [201] developed a linear QSAR model to predict the induction of apoptosis (programmed cell death) by 4-aryl-4H-chromenes, which are identified as promising apoptosis inducers. The study used a set of 43 4-aryl-4H-chromenes with known biological activity, and each compound was characterized using physicochemical, structural, and topological descriptors. These features were reduced to the seven most significant descriptors according to the elimination selection-stepwise regression variable selection method. Multiple linear regression

was employed to derive the QSAR model, resulting in a linear equation capable of predicting apoptosis-inducing activity. The model underwent rigorous validation, including cross-validation, external set validation, and Y-randomization. The results highlighted the utility of QSAR models as alternatives to traditional labor-intensive and expensive experimental procedures.

Keiser et al. (2007) [202] introduced a similarity ensemble approach, SEA, which quantitatively related receptors (proteins) based on the chemical similarity amongst their ligands. The similarity was calculated using ligand topology and expressed as a Tanimoto coefficient, which is considered a distance measure commonly applied to fingerprint representations. A statistical model derived from BLAST [203] was employed to rank the significance of the similarity scores. This approach facilitated the discovery of novel and unexpected associations, as well as the identification of potentially related proteins.

Luo et al. (2014) [204] proposed several binary classification non-linear QSAR models, including K-Nearest Neighbor (KNN), RF, and SVM, to identify novel hit compounds targeting the 5-hydroxytryptamine 1A (5-HT1A) serotonin receptor, which is an important target for potential mood and anxiety disorder treatments. These QSAR models were built using bioactivity data of receptor ligands, where each compound was characterized using 2D molecular descriptors (topological) obtained from the DRAGON software [205]. A threshold of 10 μM was applied to define active and inactive ligands. Furthermore, these models were employed in a consensus fashion for three major types of chemical screening libraries, in which fifteen compounds were selected for experimental testing. On that account, one of the nine confirmed active compounds, Lysergol, was found to have a remarkably high binding activity, which further validated the use of QSAR models.

Ma et al. (2015) [206] explored the use of single-task and multi-task Feed-Forward Neural Networks (FFNs) to predict on-target and/or ADME activities based on QSAR relationships. In this study, molecules were characterized using descriptors associated with the atom type, such as the element and the number of non-hydrogen neighbors, as well as the donor-acceptor pair, e.g., neutral donor or polar. The findings demonstrated that employing DL architectures, specifically FFNs, resulted in superior performance compared to traditional ML methods like RF. Additionally, the multi-task FFN outperformed the single-task FFNs, highlighting the ability of DL architectures to simultaneously model multiple QSAR tasks and leverage larger datasets (increased chemical and molecular information) during the learning process.

Neves et al. (2016) [207] proposed a QSAR-based consensus binary classification

framework to identify potential inhibitors of schistosomiasis, an acute and chronic tropical disease. The approach involved combining multiple ML models, including RF, SVM, and GBM, with various molecular fingerprint representations and descriptors. The fingerprint representations and descriptors used in this study included Morgan or Extended-Connectivity Fingerprints (ECFPs) [208, 209], Atom-Pair fingerprints [210], Molecular Access System (MACCS) structural keys [211], CDK descriptors [212], and DRAGON molecular features [205]. The top predicted inhibitors were further evaluated using Virtual Screening (VS) and High Content Screening (HCS), leading to the discovery of two promising compounds with potential antischistosomal activity.

3.4 Chemogenomic/Proteochemometric

The abundance of useful biological and chemical data and the growth of available computational power have sparked the development of new predictive solutions in the field of DTIs, leading to the PCM approaches [213]. These methods leverage various properties and representations to characterize proteins and compounds, which are usually combined and used as input to ML models and DL architectures due to their improved performance and ability to effectively learn from the available data [27]. Contrarily to structure-based or ligand-based approaches that primarily rely on genomic/proteomics or chemical data, respectively, PCM strategies aim to integrate information from proteomics, chemical, and/or pharmacological spaces during the inference process [214]. Numerous studies in the PCM domain have focused on predicting DTIs, and they are predominantly categorized into two main types: similarity-based and feature-based.

3.4.1 Similarity-Based

Similarity-based methods in the field of DTIs propose that compounds with similar biological, topological, and chemical properties tend to exhibit similar functions and bioactivities and should therefore interact and bind to similar targets [215]. Based on this principle, these methods leverage the shared associations of similar compounds and targets to make new assumptions about their interactions. The compounds and targets are represented using similarity matrices, which are usually derived from similarities between chemical structures or substructures and sequential similarity, respectively. The binding association is represented by an interaction matrix, containing information related to the presence or absence of association for each compound-target pair.

Yamanishi et al. (2008) [216] introduced three supervised statistical methods to infer unknown interactions for four classes of DTI networks in humans: enzymes, ion channels, G-Protein-Coupled Receptors (GPCRs), and nuclear receptors. Their approach involved representing the chemical space using a similarity matrix based on similarity scores between chemical structures and representing the proteomic space using a similarity matrix derived from the Smith-Waterman local alignment algorithm normalized scores. The first method proposed, Nearest Profile, employed the concept of nearest neighbors, in which predictions of a new drug or target were based on the most similar drugs or targets, respectively, and high-scoring drug–target pairs were predicted to interact with each other. In contrast, the second method, Weighted Profile, used the similarity to all other drugs and targets instead of relying solely on the most similar drug or target. The third method, Bipartite Graph Learning, achieved the highest performance by integrating the genomic and chemical spaces into a unified space referred to as the pharmacological space. This space was represented as a bipartite graph projected into a Euclidean space, capturing the interactions between proteins and drugs. This method employed a kernel regression approach to learn the similarity between the chemical/genomic space and the interaction space, enabling the inference of new interactions. The four DTI datasets used in this study have served as benchmarks for numerous studies. Bleakley and Yamanishi (2009) [217] further explored the original methodology proposed by Yamanishi et al. (2008) [216], replacing the kernel regression method with two SVMs: one for the target proteins of a given drug and the other for the target drugs of a given protein. The results obtained from the two SVMs were combined to generate a final prediction for each compound-target pair association.

Cheng et al. (2012) [218] introduced three inference approaches for predicting new DTIs: Drug-based Similarity Inference, Target-based Similarity Inference, and Network-based Inference. The first two methods operate under the assumption that if a drug interacts with a particular target, other drugs with similar properties are likely to interact with that target, and vice versa. The network-based inference method, which exhibited the highest performance among the three approaches, disregards the drug and target similarities and focuses solely on the known bipartite network topology of DTIs. It calculates predictive scores for each drug and unlinked target based on the similarity in network topology. This similarity is represented as a weighted matrix that influences information propagation within the drug–target network. Moreover, this approach exclusively used FDA-approved drug–target binary links to infer new predictions, limiting the predictive capability for new drugs lacking any target information. Experimental validation of some predictions was

conducted through *in vitro* assays.

Zheng et al. (2013) [219] employed Multiple Similarities Collaborative Matrix Factorization (MSCMF) for predicting new DTIs. This approach uses Matrix Factorization (MF) to decompose the connectivity matrix associated with the DTI network (interaction matrix) into two matrices of latent variables representing each drug and target in order to determine the missing interactions that are likely to exist. Additionally, this study extends the MF concept by incorporating Collaborative Filtering (CF) for prediction and multiple similarities for both drugs and targets. CF introduces regularization terms that ensure that unknown drug–target pairs do not contribute to the estimation of the latent variable matrices and that these matrices are factorized representations of the drug and target similarities. Furthermore, instead of relying solely on chemical structure and protein sequence similarities, the method selects the most consistent similarities for the given DTI.

Peng et al. (2017) [220] proposed a semi-supervised framework called NormMulInf, which is based on CF theory. This framework incorporates similarities among the samples and local correlations among the labels into a Robust Principal Component Analysis (RPCA) model. NormMulInf consists of two models: NormDrug and NormTarget. NormDrug treats drugs as samples and targets as labels, and represents the chemical space by combining the similarity matrix based on the chemical structure of drugs and the local associations between drug labels in the DTI network (indicating the likelihood of interaction with targets). NormTarget, on the other hand, treats targets as samples and drugs as labels, and represents the proteomics space by combining the sequence similarity of target proteins and the local correlations of labels among samples in the DTI network. Moreover, the proposed approach masks a portion of interactions for each sample and aims to recover the low-rank DTI matrix using the RPCA model, which is solved using augmented Lagrange multipliers [221].

Ezzat et al. (2017) [222] proposed a Graph−Regularized Matrix Factorization (GRMF) method for DTI prediction. In this approach, the similarity matrices associated with the drugs and targets are sparsified to retain only the similarity values to the nearest neighbors for each drug and target, respectively. This study also employs a Weighted K-Nearest Known Neighbors (WKNKN) pre-processing method to transform binary values in the interaction matrix into interaction likelihood values, promoting the identification of new drugs and targets. Furthermore, a variation of GRMF called Weighted GRMF (WGRMF) is explored to prevent unknown instances (interactions without information) from contributing to the determination

of the latent feature matrices.

3.4.2 Feature-Based

Given that similarity-based approaches have shown lower performance for certain protein classes and that protein sequence similarity may not always be a reliable indicator due to conformational complexity, feature-based methods have garnered significant interest. Feature-based DTI studies involve characterizing each ligand and target as numerical feature vectors, which combine various attributes (features) such as physicochemical, structural, or topological properties. These feature vectors are then used as input in prediction models to uncover unknown DTIs. Considering the amount and variety of information available, numerous features related to receptors and ligands can be extracted. However, not all of these features are relevant or discriminatory to the prediction task. Thus, employing pre-processing methods on the dataset and exploring feature engineering methodologies to assess and extract significant features are usually required in order to improve the performance of the models. Furthermore, when the number of features exceeds the number of samples, the model's performance tends to suffer in new data due to the curse of dimensionality [223]. Nevertheless, the majority of ML and DL studies for DTI prediction primarily focus on feature-based PCM approaches, leveraging these feature vectors to characterize ligand-receptor pairs and learn relationships and patterns within the data.

Yu et al. (2012) [224] proposed a ML framework for inferring new interactions, using RF and SVM as the prediction models. Chemical descriptors were generated using the DRAGON software [205], which encompasses constitutional, topological, and various other molecular properties. Conversely, protein descriptors were generated using the PROFEAT webserver [225], which primarily includes structural and physicochemical properties. These chemical and protein descriptors were concatenated to create feature vectors that characterize drug–target pairs. The proposed framework was validated using four distinct datasets associated with biologically relevant targets, specifically human enzymes, ion channels, GPCRs, and nuclear receptors [216].

Cao et al. (2014) [226] combined chemical, biological, and network properties into feature vectors for predicting DTIs, in which RF was selected as the prediction model. The chemical space was represented using MACCS fingerprints and/or substructure fingerprints, while the proteomic space was represented by protein descriptors such as amino acid composition and distribution, and other physicochemical

properties. Additionally, network properties were described using binary profiles indicating the presence or absence of interaction. The model was validated using the four independent datasets corresponding to human enzymes, ion channels, GPCRs, and nuclear receptors [216].

Coelho et al. (2016) [227] introduced a computational pipeline designed for screening potential DTIs for drug repositioning, applicable to any microbial proteome. This framework combines network metrics calculated for the interactome of the target bacterial organism with predictions from an RF classification model to identify potential DTIs. The compounds and proteins were represented using various descriptors from the PyDPI package [228]. Protein descriptors encompassed amino acid composition, Moran autocorrelation, and Composition, Transition and Distribution (CTD) features. Drug descriptors included molecular constitution, molecular connectivity, molecular property, kappa shape and charge descriptors, MACCS keys, and E-state fingerprints. Moreover, molecular docking experiments were conducted on the highest-scoring DTI pairs, demonstrating the pipeline's effectiveness in identifying new leads for drug repositioning.

Peng et al. (2017) [229] presented a novel DTI prediction framework called PUDTI for identifying potential DTIs and new drug repositioning candidates. This study incorporates feature selection methods, positive-unlabeled learning models, specifically Spy and Rocchio techniques [230, 231], and the K-means clustering algorithm to identify strong positive and negative DTIs. Drug molecules were represented using various descriptors from the PaDEL-Descriptor Software [232], while proteins were characterized using protein domain properties, pseudo amino acid composition features, and position-specific scores. Furthermore, an SVM-based optimization model called SVM with Similarity Weights (SVM-SW) was used to identify DTI candidates, in which ambiguous samples were regulated based on similarity weights to improve classification accuracy.

Considering the progressive advances in computing and the availability of large-scale datasets to train complex models, DL algorithms have emerged as state-of-the-art techniques in several research fields [233]. These architectures, usually consisting of multiple and different hidden layers [234], are capable of uncovering intricate and hidden patterns within the data without relying on feature engineering tools. Moreover, they provide robust and discriminating feature representations from raw input data and are capable of exploiting unknown structures within the data. Additionally, DL models perform significantly better when applied to large datasets, surpassing the traditional ML approaches.

Tian et al. (2016) [235] proposed Deep Learning for Compound-Protein Interactions (DL-CPI), a deep neural network approach for predicting compound-protein interactions. The model combines chemical fingerprints and protein domains, which are binary vectors indicating the presence or absence of specific features, as input for a deep FFN architecture. This architecture consists of stacked dense layers, in which all neurons are interconnected and the information flows in one direction. Additionally, regularization techniques such as weight penalty coefficient, sparsity coefficient, and dropout layers were employed to prevent overfitting of the model.

Peng Wei et al. (2016) [236] proposed a unified framework called MFDR (Multi-Scale Features Deep Representations) that combines an autoencoder with an SVM classifier to predict DTIs. The autoencoder is used to extract deep and low-dimensional representations from molecular fingerprints and protein sequence descriptors. Autoencoders are unsupervised neural networks designed to learn an approximation of the identity function, compressing the input space into a smaller and representative latent-space representation while aiming to reconstruct the input accurately. The autoencoder architecture is divided into two subnetworks: the encoder, responsible to compress the input data into a latent-space representation, and the decoder, which reconstructs the input from the latent-space representation. The resulting latent-space representation is used as input for the SVM classifier to predict the binary association of compound-target pairs.

Wen et al. (2017) [237] introduced a DL framework known as DeepDTIs, based on Deep Belief Neural Networks, for predicting DTIs. DeepDTIs uses features extracted from chemical substructures and protein sequence order information as input. Deep Belief Neural Networks stack Restricted Boltzmann Machines (RBMs), which are two-layered stochastic neural networks (one visible and one hidden) connected by a fully bipartite graph. This architecture learns a probability distribution from the input data and is capable of extracting deep hierarchical features by modeling the joint distribution between the training sample vector x and the hidden layers l:

$$P(x,h^1,...,h^l) = (\prod_{k=0}^{l-2} P(h^k|h^{k+1})P(h^{l-1},h^l) \tag{3.1}$$

, where x=h^0 (sample vector), $P(h^{k-1}|h^k)$ is a conditional distribution for the visible units conditioned on the hidden units of the RBM at level k, and $P(h^{l-1},h^l)$ is the visible-hidden joint distribution in the top level RBM.

Wang et al. (2020) [238] developed a DL framework called DeepLSTM for identifying unknown DTIs. In this architecture, drug molecules are encoded as fingerprint

features, and protein sequences are encoded using position-specific scores obtained through Legendre Moments (LMs) applied to the Position Specific Scoring Matrix (PSSM), which contains evolutionary information. These features are combined to characterize the DTI pair and then subjected to dimensionality reduction using the Sparse Principal Component Analysis (SPCA) method, which reduces the dimension of the feature space and decreases information redundancy. The resulting representation is used as input for DeepLSTM, which consists of stacked Long Short-Term Memory (LSTM) layers, to predict the binary association of DTI pairs. LSTM is a special type of Recurrent Neural Network (RNN) that includes memory blocks with self-connection memory cells for storing temporal states, as well as input, output, and forget gates to control the flow of information.

To address the limitations of using global descriptors, which are mostly not robust or discriminating for predicting real interactions, recent studies in the field of PCM have explored the use of 1D structures, such as amino acid sequences and Simplified Molecular Input Line Entry System (SMILES) strings, as well as graph representations. These representations, in combination with CNNs, Graph Neural Networks (GNNs), and FCNNs, have been explored for their potential in improving DTI prediction accuracy [239, 240, 241].

Tsubaki et al. (2019) [239] proposed an end-to-end DL framework that combines CNNs and GNNs for the prediction of DTIs. In their approach, protein sequences are transformed into sequential representations using overlapping n-gram amino acids and a learnable dictionary lookup matrix, which assigns a learnable continuous vector (embedding) to each protein subword. These sequential representations are used as input for a CNN, which captures deep patterns within the data. On the other hand, SMILES strings are converted into graph representations, where atoms and bonds correspond to nodes and edges, respectively, and propagated through a GNN in order to obtain a molecular vector representation. To enhance representation learning and address the limited number of parameters associated with atom types and chemical bonds, the framework incorporates r-radius subgraphs based on neighboring vertices and edges within a specific radius into the molecular graph representation. Furthermore, a neural attention mechanism, inspired by the work of Bahdanau et al. (2016) [242], is employed to assign different weights to protein subword representations based on their relevance to the compound. The resulting protein and compound representations are concatenated and provided as input to a dense layer, which predicts the binary association of the DTI pairs.

Lee et al. (2019) [240] introduced a DL framework called DeepConv-DTI, which

combines CNNs and FCNNs for the prediction of DTIs. Proteins are represented using raw 1D sequential data, specifically amino acid sequences, and each residue is converted into a learnable continuous vector (embedding) via a learnable dictionary lookup matrix (embedding layer). Following the embedding layer, a CNN is employed to capture local residue patterns within the protein sequences. The final protein representation is obtained by applying a global max pooling layer to the resulting feature maps from the CNN. Conversely, compounds are represented using Morgan fingerprints [208] and used as input for an FCNN in order to obtain a latent drug representation. The resulting protein and compound representations are concatenated and fed into another FCNN, which predicts the binary association of the DTI pairs. Additionally, regularization techniques such as dropout and batch normalization were employed to prevent overfitting of the model and improve the learning process.

Monteiro et al. (2021) [241] presented a DL architecture that leverages the particular ability of CNNs to extract 1D representations from protein amino acid sequences and compounds SMILES strings. In their framework, the protein sequences and SMILES strings are first encoded using a one-hot encoding layer, where each character in the sequence or string is converted into a binary vector representation. Two parallel series of 1D convolutional layers are then employed to capture deep patterns from the protein sequences and SMILES strings, respectively. To reduce the spatial size of each feature map to its maximum representative feature, a global max pooling layer is applied to the resulting feature maps, leading to deep representations of the protein sequences and SMILES strings. The resulting deep representations from both protein sequences and SMILES strings are concatenated into a single feature vector and used as input for an FCNN, which acts as a binary classifier, predicting the presence or absence of interaction. Moreover, the results of this study demonstrated that using 1D raw sequential data instead of global descriptors leads to overall improved performance.

Despite the increasing modeling ability of DL architectures to learn sequential and/or structural motifs and extract robust representations, the resulting predictions often lack interpretability or potential DTI explainability. Furthermore, these predictions typically rely on scattered and local motifs, disregarding the interdependency among the sequential and structural components of each binding entity (independent motifs), or the inter-associations revolving around the binding substructures and, thus, assigned equal weight to all extracted motifs. Nevertheless, some approaches have been proposed to address some of these limitations, particularly attention-based models [239, 243, 244, 245]. These methods focus on learning

short and long-term context dependencies among the units of proteins and/or compounds in order to condition the weight given to input elements based on their relevance.

Chen et al. (2020) [243] presented a Transformer-based architecture named TransformerCPI, which employs a classification encoder-decoder scheme to predict DTIs. Transformers [172] have shown remarkable success in several computational domains, particularly in Natural Language Processing (NLP), due to their ability to capture features between two input sequences and effective modeling of the relationships and dependencies within the sequences (attention mechanisms). In their approach, protein sequences are initially split into overlapping 3-gram amino acid sequences and then transformed into real-valued embedding using a pre-trained Word2Vec model [246, 247]. The resulting sequential representations are used as input to the encoder, which is based on gated 1D CNNs. On the other hand, SMILES strings are converted into graph representations and propagated through a Graph Convolutional Neural Network (GCN) to obtain atomic features. These molecular features are then fed to the Transformer-Decoder, which captures the relationships and dependencies between the atom sequence embedding and the protein sequence embedding (encoder output) and learns the interaction sequence. The resulting interaction sequence is used as input for an FCNN, which outputs the binary association of the DTI pairs. Moreover, this study conducted label reversal experiments to effectively assess the learning capacity of the architecture.

Huang et al. (2022) [244] introduced an architecture called Molecular Interaction Transformer (MolTrans), which leverages the effectiveness of Transformers-Encoders to extract an augmented contextual representation of the input [173]. This framework employs two Transformer-Encoders in parallel to capture the semantic relations and learn the intra-associations amongst 1D substructures in proteins and compounds, respectively. The resulting augmented contextual representations of the protein sequences and SMILES strings are transformed into a 2D interaction map using the dot-product operation. To model high-order interactions and capture and aggregate information from relevant sub-structure pairs, a CNN block is applied on top of the interaction map [248]. The resulting output is flattened and passed through a dense layer, which outputs a probability indicating the likelihood of interaction between the compound and protein.

Zhao et al. (2022) [245] proposed a bio-inspired end-to-end approach named HyperAttentionDTI, which combines 1D CNNs, a special attention block, and FCNNs to predict DTIs. This framework employs two sets of 1D CNNs operating in parallel

to identify local patterns and extract features from protein amino sequences and SMILES strings, respectively. The resulting feature maps from the 1D CNNs are then used as input in a sigmoid-based attention block, which is designed to model the semantic inter-dependencies between drug subsequences and protein subsequences across both spatial and channel dimensions. The latent feature matrices obtained from the 1D CNNs are combined with the latent feature matrices derived from the attention block, followed by global max pooling in order to obtain two feature vectors. These feature vectors are concatenated and used as input for an FCNN, which acts as a binary classifier and predicts the association of the DTI pairs.

3.5 Binding Affinity Prediction

Apart from docking simulation and certain ligand-based approaches, the majority of ML and DL studies conducted in the field of PCM methods primarily focus on binary classification tasks, predicting whether a compound interacts positively or negatively (lacks interactions) with a target. In spite of the interesting results and findings obtained in the field of DTI classification, the use of binary associations to conduct the experiments limits the quality of the results, leading to an increasing lack of target selectivity.

The availability and expansion of specific databases, such as ChEMBL [6], containing detailed information about interactions with known binding affinity and bioactivity metrics, have played a crucial role in shifting the field of computational drug discovery toward DTA prediction. Considering the limitations of certain original score metrics employed in structure-based VS, DTA prediction methods have initially focused on improving and incorporating more information, such as energetic terms, into these functions. Machine learning methods, including RF and SVM, and DL architectures like FFN, have been proposed as alternatives to traditional scoring functions. These approaches aim to predict the putative strengths of protein-ligand complexes based on various features mostly associated with their 3D structures [249, 250, 251, 252, 253, 254, 255, 256].

Ballester et al (2010) [249] introduced a scoring function called RF-score for predicting protein-ligand binding affinities, in which RF was selected as the regression model. The dataset used in this study was extracted from PDB [186] and filtered to only contain complexes with known K_d and K_i. Each 3D protein-ligand complex was characterized using a combination of intermolecular interaction features, such as the occurrence count of specific protein-ligand atom type pairs within a

defined distance range. The binding affinities were transformed into the logarithmic space and merged into single binding constants. The proposed approach was evaluated and compared to several original scoring functions used in docking simulation, demonstrating superior performance. Ballester et al (2014) [250] further improved the initial proposed RF-score [249] by incorporating chemical information relevant to the binding process, including structural interaction fingerprints, atom type, and interaction definitions. These additional chemical terms improved the accuracy of the scoring function in predicting protein-ligand binding affinities.

Durrant et al. (2010) [251] proposed NNScore as one of the initial neural network approaches for replacing scoring functions used in docking simulation. The dataset was collected from PDB [186], where complexes were categorized as either good or poor binders based on a K_d threshold of 25 μM. Each complex was characterized using enthalpic and entropic factors, considered to have the greatest influence on ligand binding affinity. Enthalpic factors include atom-atom interactions such as electrostatic and Van der Waals force, whereas entropic factors are related to the number of ligand rotatable bonds. The neural network architecture employed was an FFN. Durrant et al. (2011) [252] further improved the original NNScore, creating NNScore 2.0 by incorporating additional binding features and shifting from binary classification to estimating the binding affinity measured in pK_d. Each receptor-ligand complex was described using features from Autodock Vina [191] and BINANA [257]. Autodock Vina [191] includes three steric terms, a hydrophobic term, and a hydrogen-bond term. On the other hand, BINANA [257] provides several binding properties, including the number of hydrogen bonds and active-site flexibility. The neural network architecture employed was also an FFN, however, the output layer was replaced to return a continuous value (pK_d) rather than a binary value.

Li et al. (2013) [253] proposed a ML framework known as ID-Score to predict protein-ligand binding affinity, employing SVR as the prediction model. The dataset was collected from PDBbind [19] and filtered to contain complexes with binding activity measured in terms of pK_i, pK_d, and pIC_{50} (log-transformed). Each structure was characterized using a comprehensive set of features related to the binding process, including Van der Waals interaction, hydrogen bonding, electrostatic interaction, π-system interaction, metal-ligand bonding, desolvation effect, entropic loss effect, shape matching, and surface property matching. The results demonstrated the effectiveness of the ID-Score method compared to state-of-the-art scoring functions used in VS, and its ability to correctly differentiate structurally similar ligands.

Li et al. (2014) [254] explored two regression models, specifically multivariate linear

regression and RF, for predicting binding affinity. The multivariate linear regression model employed was Cyscore [258], an empirical scoring function in an additive functional form of four energetic terms: hydrophobic free energy, Van der Waals interaction energy, hydrogen bond interaction energy, and ligand conformational entropy. RF was evaluated using three sets of features: Cyscore energetic terms, AutoDock Vina features (Gauss1, Gauss2, Repulsion, Hydrophobic, Hydrogen Bonding, and the number of rotatable bonds) [191], and the features used in the research work of Ballester et al. (2010) [249]. The dataset was extracted from PDBbind [19], and each complex was characterized by either K_d or K_i. RF achieved superior performance compared to the Cyscore model, particularly when incorporating structural features.

Kumar et al. (2021) [255] introduced Substructural Molecular and Protein-Ligand Interaction Pattern Score (SMPLIP-Score) for predicting binding affinity based on the use of a straightforward and interpretable featurization process. The protein-ligand complexes were collected from PDBbind [19], and processed using the KNIME Analytic Platform [259], which iteratively refined the structures by adding hydrogen atoms, correcting the bond order, and removing water molecules. Each complex was characterized using features related to interaction patterns (fingerprints), interaction distances, and molecular substructural fragments. The authors explored two regression models, namely RF and FFN, where the former exhibited the highest prediction performance. Additionally, the features used to characterize the complexes were evaluated based on their scoring power, ranking power, and robustness. The evaluation demonstrated that the features possessed sufficient discriminatory and predictive capabilities.

Meli et al. (2021) [256] explored the use of atomic environment vectors (AEVs) and FFNs for the prediction of protein-ligand binding affinity. The proposed framework called AEScore describes every atom in the protein-ligand binding site (within a certain distance) using atom-centered symmetry functions (ACSFs), specifically radial and angular symmetry functions, to capture the local chemical environment. The resulting AEVs are propagated across atom-specific FNNs, where all atomic contributions are summed together to predict the binding affinity value. Moreover, the authors explored combining their approach with AutoDock Vina [191] in order to learn potential corrections to the classical scoring function. These results demonstrated sufficient predictive capabilities of binding affinity while retaining the docking and screening power of AutoDock Vina [191].

Given that the interaction between an active compound and a protein occurs in the

3D space and that the conformational space accessed by the ligand plays a critical role in ensuring optimal interactions with the protein and achieving high binding affinity, recent research endeavors have focused on employing 3D CNNs in conjunction with 3D single-instance learning due to the exceptional capability of 3D CNNs to effectively capture spatial context [260, 261, 262, 263]. However, these approaches are limited by the availability and complexity of the 3D structures, resulting in complex models with reduced reproducibility. Furthermore, 3D single-instance learning does not consider the range of possible ligand and protein conformations. Many 3D conformations would have to be taken into consideration and multiple 3D instances would be necessary to represent a single object, in which ML and DL approaches have been considering a single 3D instance due to the complexity/limitation of multi-instance learning.

Gomes et al. (2017) [260] presented a novel 3D spatial convolutional approach in order to learn atomic-level chemical interactions based on atomic coordinates. The proposed framework predicts the energy gap, specifically the binding free energy, between a protein-ligand complex in a bound state and its unbound state. The 3D crystal structures and corresponding K_i values used in this study were collected from PDBbind [19]. In this approach, two primitive convolutional operations were introduced, specifically atom-type convolutional and radial pooling. The atom-type convolutional uses a neighbor-listed distance matrix to extract features from the Cartesian atomic coordinates. On the other hand, the radial pooling filters, characterized by learnable mean and variance, are employed to extract information regarding the atom's environment, resulting in a representation that is invariant to atom ordering and the orientation of the complex. The output of the radial pooling layer is used as input to an atomistic FCNN that estimates the energy of each atom, in which the sum of all of these atomic energies gives the total energy of the molecule.

Stepniewska-Dziubinska et al. (2018) [261] proposed a deep neural network approach called Pafnucy based on 3D CNNs for predicting binding affinity. Drug–target complexes were extracted from PDBbind [19] with their corresponding binding affinities expressed in pK_d or pK_i. Each complex was represented by a 4D tensor, in which the first three dimensions correspond to the Cartesian coordinates (obtained by a 3D grid) and the last dimension to a vector of atom features. These features were computed using Open Babel [264], representing several atom properties, including atom type, hybridization, and partial charge. Even though each complex was represented by a 4D tensor, the molecular complex is seen as a 3D image with multiple color channels, wherein each position is characterized by the vector of atom features. Moreover, a regularization technique was employed to ensure that both proteins and

ligands have the same atom types.

Jiménez et al. (2018) [262] introduced K_{DEEP} for predicting binding affinity based on the use of 3D CNNs. The dataset used in this study was extracted from PDBbind [19] and filtered to include only complexes with known K_d or K_i. Additionally, the 3D complexes were pre-processed based on structural resolution and the experimental precision of the binding measurement. Both proteins and ligands were characterized by a 3D voxelized representation of the binding site by assigning a Van der Waals radius to each atom type. On that account, pharmacophoric-like properties, including hydrophobic, hydrogen-bond donor or acceptor, aromatic, positive or negative ionizable, metallic, and total excluded volume, were computed by centering a fixed subgrid on the geometric center of the ligand. The resulting 3D voxelized representation comprised 16 different channels to account for both protein and ligand. This representation was used as input for a 3D CNN followed by an FCNN, predicting the binding affinity value.

Jones et al. (2021) [263] explored a deep fusion inference framework that combines the outputs obtained from a 3D CNN and a Spatial Graph Convolutional Neural Network (SG-CNN) for the prediction of binding affinity. The 3D complexes were represented using various atomic features, including element type, atom hybridization, number of heavy atom bonds, number of bonds with other heteroatoms, structural properties, partial charge, molecule type, and Van der Waals radius. The 3D CNN was applied to capture 3D atomic features and implicit atomic interactions, while the SG-CNN was employed to capture noncovalent interactions, which play a crucial role in modeling complex biological structures. Contrary to traditional molecular graph representations, the SG-CNN applies explicit distance thresholds to determine which pairs of atoms should be considered for pairwise interactions, e.g., covalent or noncovalent. Both covalent and noncovalent bonds are represented through the use of an adjacency matrix. Overall, the results demonstrated that the integration of heterogeneous feature representations obtained from the two models of the fusion framework leads to improved prediction performance of binding affinity.

To circumvent the limitations of 3D single instance learning and the confined space of proteins and ligands with known/determined 3D structures, recent research studies have been exploring PCM approaches based on chemogenomic and lower structure information, e.g., 1D and 2D structures, to conduct their experiments, leading to more realistic and reproducible methodologies in the DTA prediction domain. In addition to traditional ML methods, several studies have centered their studies on CNN-based frameworks, such as 1D CNNs, 2D CNNs,

or GCNs, to extract knowledge and meaningful information from different protein and compound representations, including 1D structures, 2D similarity matrices, feature vectors, or molecular graphs, for the prediction of binding affinity [265, 266, 267, 268, 269, 270, 271, 272, 273, 274, 275]. Moreover, three benchmark datasets associated with the studies by Davis et al. (2011) [276], Metz et al. (2011) [277], and Tang et al. (2014) [278], measured in K_d, K_i, and KIBA scores, respectively, have been the focus of several of these studies to establish and evaluate the regression models.

Pahikkala et al (2014) [265] presented a Kronecker-Regularized Least Squares (Kronecker RLS) algorithm to predict binding affinity. The datasets used in this study were collected from the research works of Davis et al. (2011) [276] and Metz et al. (2011) [277]. Proteins and compounds were represented by their pairwise similarity score matrices, which were obtained from the PubChem structure clustering server [279] and Smith-Waterman local alignment algorithm, respectively. Kronecker-RLS focus on minimizing the following objective function:

$$J(f) = \sum_{i=1}^{m}(y_i - f(x_i))^2 + \lambda||f||_k^2 \tag{3.2}$$

, where f is the prediction function, x_i the input, y_i the interaction affinity, λ>0 a user-provided regularization parameter, and k the kernel function, which is associated with the protein and compounds similarity matrices.

Shar et al (2016) [266] explored two regression ML models, specifically RF and SVM, for the prediction of binding affinity measured in terms of K_i. The dataset was collected from the Psychoactive Drug Screening Program K_i database [280], which comprises bioactivity data of several molecules targeting GPCRs, transporters, and ion channels. The proteins and compounds were characterized by features generated from the DRAGON software [205] and PROFEAT web server [225], respectively. The results of this study demonstrate that using 2D autocorrelation, topological charge indices, and 3D-MoRSE descriptors to characterize compounds, and amphiphilic pseudo amino acid composition, autocorrelation features, and quasi-sequence order descriptors to represent the proteins leads to improved K_i prediction. Moreover, both regression ML models achieved similar prediction performance.

He et al. (2017) [267] introduced SimBoost for predicting the binding affinity of DTI pairs, employing GBR as the regression model. The proposed method constructs features for each drug, target, and DTI pair by using similarity and network information. Three types of features were extracted using feature engineering: occurrence

and pairwise similarities information, drug-drug and target-target similarity network features, and drug–target interaction network features. The compound and protein similarity matrices were obtained from the PubChem structure clustering server [279] and the Smith-Waterman algorithm, respectively. Furthermore, a variation of SimBoost called SimBoostQuant was explored to compute a confidence score for the prediction interval of a given DTI pair based on quantile regression. The datasets used in this study were collected from the research studies of Davis et al. (2011) [276], Metz et al. (2011) [277], and Tang et al. (2014) [278].

Öztürk et al. (2018) [268] proposed a DL pipeline called DeepDTA, which is based on 1D CNNs, for the prediction of binding affinity. The authors validated DeepDTA using datasets obtained from the studies conducted by Davis et al. (2011) [276] and Tang et al. (2014) [278]. In their framework, proteins and compounds were represented using 1D sequential information, specifically amino acid sequences and SMILES strings, respectively. Two parallel CNNs were employed to uncover underlying patterns and extract deep representations from the input data. The resulting deep representations were concatenated and used as input for an FCNN, which predicted the binding affinity value. *Öztürk* et al. (2019) [269] extended their previous approach using a text-based method for predicting DTA by shifting from a character-based sequence representation to a word-based sequence representation. Only chemical and biological textual sequence information was considered in this new framework, and four types of information were collected: protein amino acid sequences, protein motifs and domains, compound SMILES strings, and ligand maximum common substructures. Protein sequences were represented using sets of 3-residue words, which consisted of groups of three consecutive amino acids. Protein domains and motifs were extracted from the PROSITE database [281] and represented using 3-residue subsequences. The SMILES strings were initially converted to Deep SMILES [282] and then transformed into consecutive overlapping 8-character words. Ligand maximum common substructures [283] were employed to extract chemical words from the SMILES strings, representing certain patterns capable of distinguishing sets of molecules. The proposed framework called WideDTA employed four parallel CNNs to extract a deep representation from each specific type of information. The resulting deep representations were concatenated and used as input for an FCNN.

Feng et al. (2018) [270] introduced a DL framework known as Protein And Drug Molecule Interaction Prediction (PADME) to predict real-valued interaction strength. Compounds were represented using molecular graphs, where nodes represented atoms and edges represented bonds. Proteins were described using protein

sequence composition descriptors and a binary entry to characterize the phosphorylation status. The proposed architecture leverages the ability of GCNs to capture deep patterns from molecular graph representations, where the resulting deep representations were combined with the protein features to characterize the DTI pairs. The final DTI representations were used as input for an FCNN, which predicted the binding affinity values. Moreover, a variation of PADME was explored, wherein the graph representation of the compounds was replaced with an ECFP-based representation. The authors validated PADME using datasets obtained from the studies conducted by Davis et al. (2011) [276], Metz et al. (2011), and Tang et al. (2014) [278], and a dataset containing toxicology data measured in AC_{50} (activity concentration at 50% of maximal activity). Furthermore, the proposed approach was capable of handling cold-target/cold-drug problems, which are associated with targets and drugs that have never appeared in the training data, respectively.

Nguyen et al. (2020) [271] presented GraphDTA for predicting binding affinity based on the use of GNNs and 1D CNNs. This study explored multiples types of GNNs, including Graph Isomorphism Neural Network (GIN) [284], Graph Attention Neural Network (GAT) [285], GCN, and Graph Attention - Graph Convolutional Neural Network (GAT-GCN), to extract deep features from molecular graph representations of the compounds. Each node in the molecular graphs was characterized by various atom features such as atom symbol, atom degree, number of bonded neighbors and hydrogens, total number of hydrogens, implicit atom value, and aromaticity. The edges of the molecular graphs were represented by the presence or absence of interaction. On the other hand, proteins were encoded using their 1D amino acid sequences and fed into 1D CNNs to extract deep representations. The resulting representations from both GNNs and 1D CNNs were concatenated and used as input for an FCNN to predict the binding affinity value. The proposed approach was validated using datasets from the studies conducted by Davis et al. (2011) [276] and Tang et al. (2014) [278].

Abbasi et al. (2020) [272] introduced DeepCDA, a DL framework for predicting binding affinity that combines CNNs and LSTMs. This framework employs two parallel blocks of CNNs followed by LSTMS to extract discriminating representations from protein amino acid sequences and compound SMILES strings, respectively. The resulting representations are used as input for a two-sided attention mechanism to encode the interaction strength between protein and compound substructures, leading to a binding map containing the weights (strength) of each interaction. This binding map is used as input for an FCNN, which outputs the binding affinity value. Furthermore, the authors explored an adversarial domain adaptation technique to

improve the generalization of the model and solve the problem of training and testing data being sampled from different domains with different contributions. The authors validated their approach using datasets collected from the research works of Davis et al. (2011) [276] and Tang et al. (2014) [278], and from the BindingDB database [18].

Shim et al. (2021) [273] explored a similarity-based model called SimCNN-DTA to predict binding affinity based on 2D CNNs. Proteins and compounds were represented by their pairwise similarity score matrices, which were computed using the Smith-Waterman local alignment algorithm ans Tanimoto distance metric, respectively. The outer products between the column vectors of the two similarity matrices were used as input for 2D CNNs in order to extract deep patterns from the DTI representation space. The output of the 2D CNNs was used as input for an FCNN, which predicted the binding affinity value. The authors validated their approach using datasets collected from the studies conducted by Davis et al. (2011) [276] and Tang et al. (2014) [278].

Wang et al. (2021) [274] presented DeepDTAF for predicting binding affinity based on 1D CNNs and 1D dilated CNNs. This framework consists of three parallel modules to extract local and global contextual features from proteins, compounds, and binding pockets. Proteins were represented by their 1D amino acid sequence and secondary structure and physicochemical characteristics, compounds by their SMILES strings, and binding pockets by sequential and structural properties. 1D dilated CNNs were employed to capture multiscale long-range interactions from protein features and ligand SMILES, while traditional 1D CNNs were applied to uncover and extract deep patterns within the binding pocket representation. The resulting deep features from the three modules were concatenated and used as input for an FCNN, which predicted the binding affinity value. The authors validated their approach using data collected from the PDBbind database [19].

Rifaioglu et al. (2021) [275] proposed a hybrid pairwise input deep neural network called MDeePred (Multi-channel Deep Proteochemometric Predictor for Binding Affinity) to estimate DTA. In their approach, multiple types of protein features such as sequential, structural, evolutionary, and physicochemical properties were incorporated within multiple 2D vectors and used as input for a CNN-based model. On the other hand, compounds were represented by molecular fingerprints-based vectors and fed to an FCNN. The resulting representations obtained from the two modules were concatenated and used as input for another FCNN, which predicted the binding affinity value. The authors validated their approach using different

datasets collected from ChEMBL [6], PDBbind [19], and the research study by Davis et al. (2011) [276].

Chapter 4

Explainable Artificial Intelligence

This chapter presents an overview of explainability and interpretability concepts in the context of Artificial Intelligence (AI) methods. Section 4.1 details the need for providing compelling explanations concerning ML and DL approaches in critical domains such as drug discovery. Section 4.2 introduces relevant and recurring terminology employed in the context of Explainable Artificial Intelligence (XAI) methods, and describes the concepts of explainability and interpretability. Section 4.3 summarizes various strategies within the realm of XAI capable of providing explanations to the inferring process and/or predictions of the models.

4.1 Explaining Models' Decisions

Despite the continuous advances in the field of Artificial Intelligence (AI) to produce autonomous systems capable of perceiving, learning, and reasoning on their own, the underlying mathematical models often remain elusive to interpretation by the human mind, limiting the effectiveness of these systems [286]. Moreover, the majority of AI approaches are based on models/heuristics of difficult interpretation, which may result in an inadequate explanation of the input context that leads to a specific choice of a particular decision. Given the current pace of ML and DL in real-world applications and critical contexts, such as health care and drug discovery, there is an increasing demand to assure specific criteria, develop high-performing models under rigorous conditions, and provide explanations in a human-intelligible format [287, 288].

Several AI models have been successfully adopted for computer-assisted drug discovery [289, 290], in which DL architectures have been surpassing most traditional ML methods [291, 292] due to their ability to model and capture complex and intricate nonlinear relationships between input data and the associate output by stacking

multiple processing layers. Furthermore, DL has been pivotal to broaden *in silico* drug discovery across different subdomains, namely molecular design [293, 294], chemical synthesis [295, 296], protein structure prediction [297, 181], and DTI identification [268, 241]. However, DL architectures are considered highly black-box models, devoided of transparency and explainability in their inner operations and decision-making process. Moreover, it is difficult to understand the error surface, and obtain detailed explanations of their behavior [298, 299, 300].

Some prior attempts have been explored to extract, represent, and explain features from medical and chemical knowledge to better fit human intuition [301, 302]. In the particular case of drug discovery, efforts have been made to explain QSAR models in terms of algorithmic insights and molecular analysis in order to understand and correlate biological effects with physicochemical properties [303, 304, 305, 306]. Furthermore, recent PCM approaches, whose focus is based on the use of target and ligand information for the inferring process, have been providing interpretable prediction results by analyzing the importance of certain descriptors for the prediction of DTIs, resulting in the identification of important ligand features and/or target features [40, 307, 308, 309, 310]. However, the rise of DL architectures reduces the willingness to sacrifice prediction performance in favor of explainability, especially when considering the ability of these architectures to model nonlinear associations, perform pattern recognition and feature extraction from low-level data representations, and achieve state-of-the-art prediction performance.

Drug discovery is unequivocally complex and not straightforward, posing many domain-specific challenges [311]. Several comprehensive factors are involved in DTIs, in which identifying important substructures remains an ongoing task. Even though ML and DL models may identify complex hidden patterns within sequential and structural data, these are usually not perceptible by humans or require domain knowledge to be explained. Moreover, the representation of the proteomics, chemical, and pharmacological spaces plays a crucial role in AI-assisted drug discovery, considering that there are no raw and complete representations of the proteins, compounds, and binding pockets. Thus, the choice of the representation models conditions the explainability and performance of the resulting prediction architecture given that they determine the context, type, and interpretability of the information retained [31]. Overall, there is a need for further understanding due to the lack of knowledge regarding all the biological, chemical, and pharmacological processes involved, and the inability to formulate infallible mathematical models and corresponding explanations [32, 312].

The lack of interpretability of the underlying models and the need to augment human reasoning and decision-making have been motivating the emergence of the Explainable Artificial Intelligence (XAI) field [313, 314]. Providing potential and informative explanations for the models' decisions helps to guarantee decisive criteria concerning ML and DL in critical contexts: i) ensure that the decision-making is impartial and understandable; ii) avoid correct predictions for the wrong decisions; iii) advert possible disturbances or unfair biases that may alter the prediction; iv) ensure a real underlying causality in the reasoning of the model by identifying the most significant variables that infer the output; and v) bridge the gap between ML/DL and domain knowledge [32, 315, 316, 312]. On that account, XAI bears the promise to undertake informed actions while concurrently taking into account domain knowledge, model logic, and an understanding of the limitations inherent to the model [317].

4.2 Explainability and XAI Terminology

DL is increasingly being employed in several areas of interest, including critical contexts, e,g., medical and pharmaceutical areas, where the decision provided may have a great impact on the overall well-being. The efficient learning algorithms associated with these architectures have proven pivotal in attaining exceptional levels of performance, while simultaneously solving progressively intricate computational tasks. Contrarily to simple linear or rule-based models, where it is possible to search for a direct understanding of the mechanisms involved in the model (transparency), DL architectures are mostly opaque and complex black-box models due to the vast parametric space, which compromises the interpretability of these architectures [29]. In that regard, it is crucial to provide explanations that support the output of a model given the risk of employing decisions that are not justifiable, reliable, or capable of providing detailed explanations [318]. Moreover, given the rise of more complex data across various domains, it is vital to provide fundamental support in the interpretation and analysis of the results [319, 320]. Hence, XAI focuses on producing explainable and high-performing models that enable humans to understand, comprehend, and trust these systems.

In the context of ML/DL, it is important to distinguish interpretability from explainability, considering that these two terms are mostly not interchangeable. Interpretability is associated with the intrinsic characteristics of the model to provide the user with the ability to understand the processes involved, i.e., to be, on its own, understandable for a human (transparency). On that account, interpretable

models make it possible for humans to know the influence of the input variables on the overall performance, understand the marginal relationship between input variables and the target, and predict a future output by analyzing the input. Moreover, these models can feature different levels of understandability/intelligibility [321, 315]. On the other hand, explainability concerns the ability of the model to provide details with the intent of clarifying or detailing its internal functions and reasoning in human terms [315]. In that regard, explainable methods focus on understanding sets of decisions instead of the internal structure of the model or the intrinsic algorithmic processes involved [312]. Hence, explainability emphasizes converting non-interpretable models into explainable ones. Nevertheless, the degree of explainability relies on the capability of the users to understand the knowledge contained in the model (comprehensibility) [322].

4.3 XAI Methods

Methods designed for explainability can be categorized according to multiple criteria, including interpretability grade, model dependency, explainability range, type of data, and results of the interpretation method [31, 312, 315, 320]. However, XAI methods broadly fall under three primary domains: i) intrinsic interpretability or post-hoc explainability, ii) model-specific or model-agnostic, and iii) local or global explainability.

The achievement of intrinsic interpretability is accomplished by altering the structure of the model in a manner that enables the user to discern the features that influenced the overall inferring process. Furthermore, the design should be oriented in a way that facilitates the visualization of the marginal contribution of each input variable to the model, and that the error surface can be understood and substantiated [323]. These approaches are primarily designed for transparency, i.e., need to convey some degree of interpretability by themselves, and are built under the premise that all parts of the model can be understandable to a human without the need for additional methods or tools [324]. Intrinsic interpretability is commonly associated with simpler models, such as logistic regression, low-depth decision trees, decision rules, sparse linear regression, generalized linear models, and generalized additive models. In spite of the accurate explanations provided by these models, the compromise between complexity and explainability often leads to a reduction in prediction performance. Nevertheless, recent approaches focus on incorporating explainability directly into the structural units of the architectures (*intrinsic explainability*). In that regard, attention mechanisms have been explored and incorporated

into the architectures to condition the learning process, and provide explainability through the visualization of the input elements that were given more attention (weight) [325, 326].

Post-hoc techniques aim to provide explanations for an already existing model, imparting insights into the parameters, learned representations, individual predictions, or the behavior of the model. Typically, this is accomplished by constructing supplementary models or employing external methods on models that are not inherently interpretable, with the intention of augmenting their interpretability. Nonetheless, these approaches possess limitations in their inherent approximation, whereby if the model demonstrates bias, the explanations will similarly be biased [327]. Post-hoc explainability can be attained through diverse perspectives such as:

- Surrogate models/functions: explain the predictions of a complex model by locally approximating it with a simple interpretable surrogate function, e.g., decision trees, decision rules, or linear models. Ribeiro et al. (2016) [328] introduced Local Interpretable Model-Agnostic Explanations (LIME), a technique that involves generating local surrogate models to approximate the behavior of the primary model around a given prediction. The LIME approach samples data points within the neighborhood of the specific instance of interest, subsequently evaluating the model at these points, and then tries to fit the surrogate function in a manner that closely approximates the behavior of the primary model for that particular instance. Lundberg et al. (2017) [329] presented Shapley Additive Explanations (SHAP), a method designed to explain the prediction of a certain instance by computing the contribution of each feature to the prediction. This approach relies on the concept of Shapley values, which enables the provision of explanations via an additive feature attribution technique. The method leverages a linear combination of binary variables to attribute the contributions of individual features to the overall prediction.

- Local Perturbations: attempt to explain the model by modifying or removing parts of the input and measuring the respective changes in the output. Zeiler et al. (2014) [330] explored an occlusion sensitivity method, which systematically occludes different portions of the input in order to measure the importance of the input dimension. Zintgraf et al. (2017) [331] proposed a method based on prediction difference analysis, which uses conditional sampling within the pixel neighborhood of an analyzed feature to measure how the prediction changes if the feature is unknown.

- Propagation-based: leverages the model's internal structure and propagates a

score in order to measure how much a change around a local neighborhood of the input corresponds to a change in the output. Methods such as Deconvolution [330, 332], Layer Relevance Propagation [333] and Gradient-Weighted Class Activation Mapping [334] highlight the critical regions in the input for the prediction of the concept, in which the feature activity, the relevance score, and the gradients of the model's outcome are backpropagated to the input domain, respectively.

- Instance-based: identifies data points, such as relevant feature subsets, that are essential for retaining or altering the predictions of a given model. This category includes anchor algorithms, counterfactual instance search, and contrastive explanation methods [31]. Anchor algorithms are designed to derive a subset of if-then rules based on one or more features that must be satisfied to maintain the predicted outcome [335]. Counterfactual explanations, on the other hand, aim to discover data points most similar to a specific instance that can lead to a different prediction outcome [336, 337]. Contrastive explanation methods combine the key concepts of anchor algorithms and counterfactual instance search approaches to provide explanations. These methods identify the smallest set of features required for the model to predict the correct outcome and the smallest set of features needed to be absent to ensure sufficient distinction from other potential outcomes [338].

XAI-generated explanations can also be categorized into model-specific or model-agnostic. The model-specific category includes methods designed for specific models, such as propagation-based approaches and saliency maps used for neural networks [339, 334, 332, 330]. Intrinsically interpretable models are also included in this category considering that intrinsic interpretability is inherently tied to the design of the models. On the other hand, model-agnostic techniques are not restricted to any particular model and focus on establishing relationships between input and output pairs by extracting information from the prediction procedure of the models. These methods are typically applied post-hoc, as they lack access to the model's internal logic and intrinsic operations [315, 312].

Additionally, the resulting explainability can be classified according to its range, namely local or global explainability [340]. Local explainability addresses explainability by segmenting the solution space and providing explanations for individual predictions or a small subset of neighboring samples. Hence, these methods only explain part of the model's functioning. In contrast, global explainability seeks to provide an overall understanding of the model's behavior across the entire dataset or a specific domain. Thus, global explainability focuses on analyzing the model's

patterns and general behaviors and summarizing the relevance of input features that consistently influence the decisions of the model.

Chapter 5

Methods, Models, and Architectures

This chapter details some of the major and recurring methodology employed throughout the research work associated with this thesis. Section 5.1 details ML models used to compare the prediction performance. Section 5.2 explains the main concepts regarding DL architectures involved in the design of the frameworks explored in this thesis. Section 5.3 describes the similarity methods applied to the proteins and compounds. Section 5.4 presents the evaluation metrics explored to assess the prediction performance.

5.1 Machine Learning Models

5.1.1 Random Forest

Random Forest (RF) is an ensemble learning method that generates a chosen number of uncorrelated decision trees and returns the class or value that is the mode of the classes (majority voting) or the average of the values, respectively, across the output of each decision tree [165]. Decision trees are the building blocks of the forest and they can be defined as a series of if-then-else rules (nodes) that divide the dataset into smaller subsets until the predicted class or value is achieved or when the impurity can no longer be reduced. The nodes are based on a single feature and a specific threshold according to the combination that generates less impurity, e.g., entropy, for the tree. Each decision tree is created using bootstrapping, in which only a randomly selected portion of the dataset is used to build the tree. The non-sampled data (out-of-the-bag) is used to evaluate the performance (generalization capacity). Moreover, each tree at each node only considers a subset of features that are randomly chosen. The randomness of the whole process increases the diversity amongst the trees, making them grow dissimilar and uncorrelated. This method is highly adaptive to different

data types given its capacity to describe the relationship between independent and dependent variables. Figure 5.1 illustrates the RF architecture for classification and regression tasks (RFR).

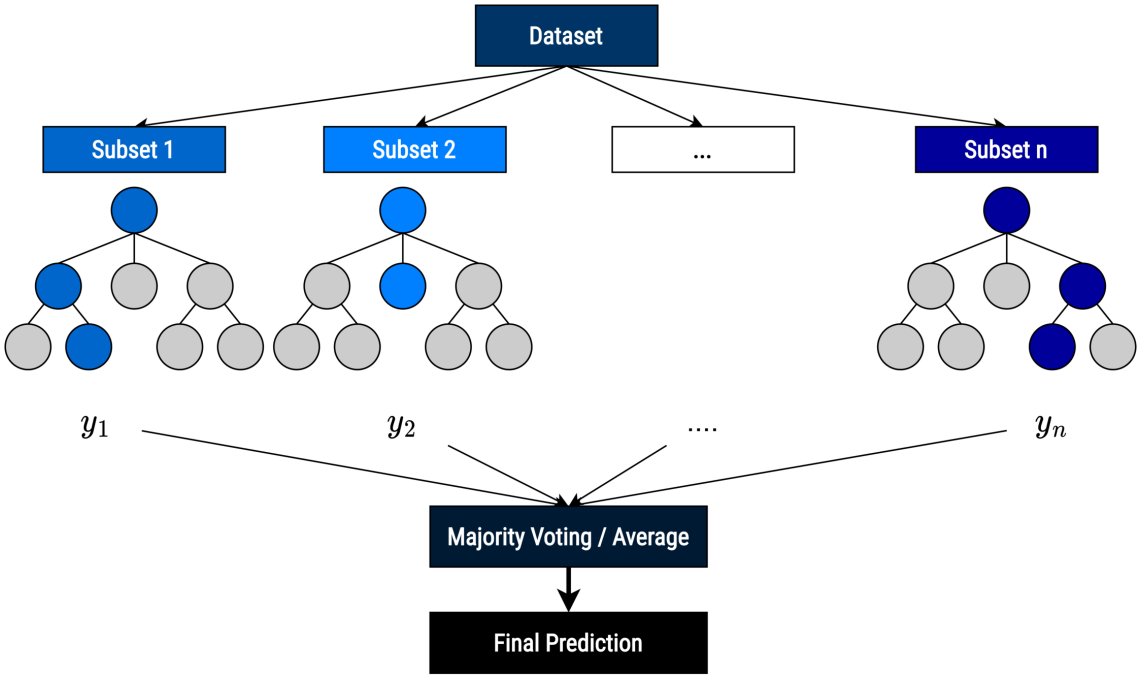


Figure 5.1: Random Forest, where the majority voting approach is applied for classification problems and the average for regression tasks.

5.1.2 Support Vector Machine

Support Vector Machine (SVM) identifies an optimal hyperplane that maximizes the separation margin between different classes [166]. For problems that are not linearly separable, SVM applies two different approaches, namely soft-margins and kernel tricks. Soft-margin tolerates violations of the margins and controls the trade-off between margin-width maximization and misclassified sample minimization. The tolerance is represented by the penalty term C, which is responsible for the number of violations allowed. Kernel tricks, which are identified as functions capable of transforming the data, are used to map data to higher dimensional spaces where it is possible to classify with linear decision surfaces. Gaussian Radial Basis Function (RBF) is one of the most used kernels for handling nonlinear problems and it replaces each point in the feature space by the Gaussian of the squared Euclidean distance from support vectors. Considering x_1 and x_2 two different feature vectors in the original input, RBF can be expressed as:

$$K(x_1,x_2) = exp(-\frac{||x_1 - x_2||^2}{2\sigma^2}) \tag{5.1}$$

, where σ is the scale parameter that is related to the Gaussian width. Support Vector Regression (SVR) is an extension of the SVM for regression tasks, in which instead of defining a hyperplane that maximizes the separation margin between different classes, it finds the best-fit line corresponding to the hyperplane that has the maximum number of points [167]. The regression line margin in the SVR is controlled by a parameter ϵ. Figure 5.2 illustrates the SVM architecture applied to a classification problem.

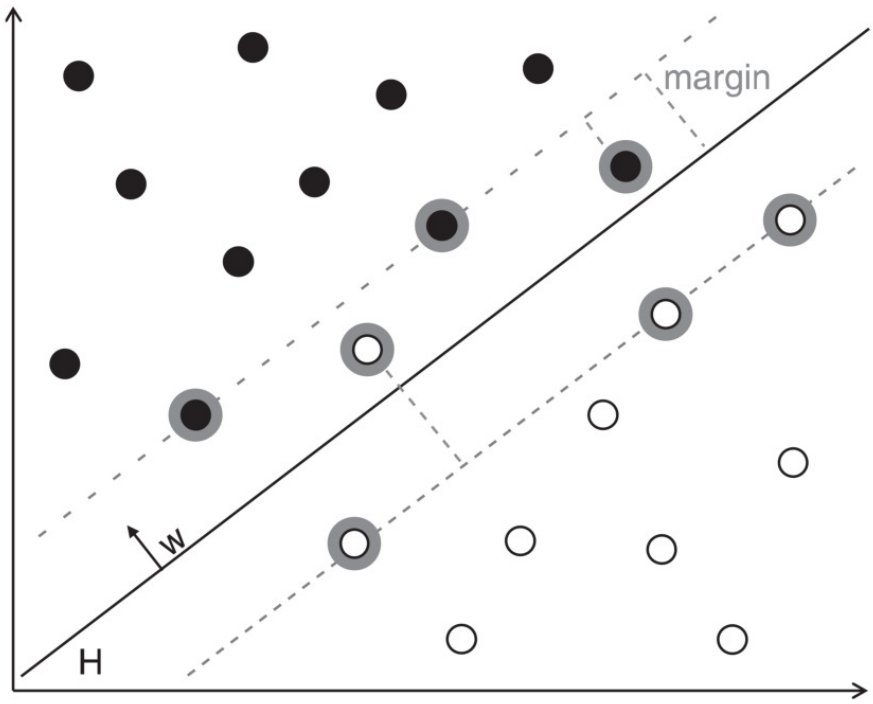


Figure 5.2: Support Vector Machine applied to a binary classification problem. Figure adapted from "Support vector machines for drug discovery" [168].

5.1.3 Gradient Boosting Regression

Gradient Boosting Regression (GBR) is an ensemble learning method derived from the Gradient Boosting Machine (GBM) model [169, 170, 171], which iteratively adds decision trees in order to improve the objective function. GBR makes predictions using the following equation:

$$\widehat{y_i} = \Phi(x_i) = \sum_{k=1}^{K} f_k(x_i), f_k \in F \tag{5.2}$$

, where $\widehat{y_i}$ is the predicted value, K is the number of regression trees, x_i is the input, and F the space of all possible trees. In order to learn the set of trees f_k, a

regularized objective function is employed:

$$L(\Phi) = \sum_{i} l(\hat{y}_i,y_i) + \sum_{k} \Omega(f_k) \tag{5.3}$$

, where l is the loss function and Ω is a tuning parameter that measures the complexity of the model to avoid overfitting.

5.1.4 Kernel Ridge Regression

Kernel Ridge Regression (KRR) combines Ridge regression with kernels, in which the learning process is similar to the SVM. This method estimates a regression function f by solving an optimization problem over the reproducing kernel Hilbert space of functions $\mathcal{H}$:

$$\hat{f} = \underset{f \in \mathcal{H}}{argmin} \sum_{i=1}^{m} (y_i - f(x_i))^2 + \lambda ||f||_k^2 \tag{5.4}$$

, where $||f||_k^2$ is norm of the regression function, x_i is the input, y_i is the target, λ>0 is a user provided regularization parameter, and k is the kernel function. The regularized loss function (minimizer) for the above objective can be expressed as :

$$f(x) = \sum_{i=1}^{m} \alpha_i k(x,x_i) \tag{5.5}$$

, where α is the data dependent weights and $k(x,x_i)$ is the kernel function centered in the input x.

5.2 Deep Learning Architectures

5.2.1 Fully-Connected Feed-Forward Neural Network

Fully-Connected Feed-Forward Neural Networks (FCNNs) are similar to traditional Artificial Neural Networks (ANNs), comprising an input layer, multiple hidden layers, and an output layer. The input layer is associated with independent values (features) that are fed to the working units, denominated artificial neurons, which constitute the hidden layers. Each one of the hidden layers is composed of multiple neurons, which are interlinked across the layers. The output is the result of the weighted sum of all the outputs given by the previous layers and to which is applied an activation function. Each artificial neuron is organized into five building ele-

ments: input, weight, bias, activation function, and output. The output associated with the ith neuron can be expressed as:

$$f(\alpha_i) = f(\sum_{j=1}^{n} W_{ij} X_j + b_i) \tag{5.6}$$

, where W is the weight, X is the input value, f is the activation function, b is the bias, and n is the number of neurons from the previous layer connected to the ith neuron. In this type of architecture, the information flows in one direction, from the input layer, going through the hidden layers (middle layers), to the output layer. Figure 5.3 shows the architecture of an FCNN.

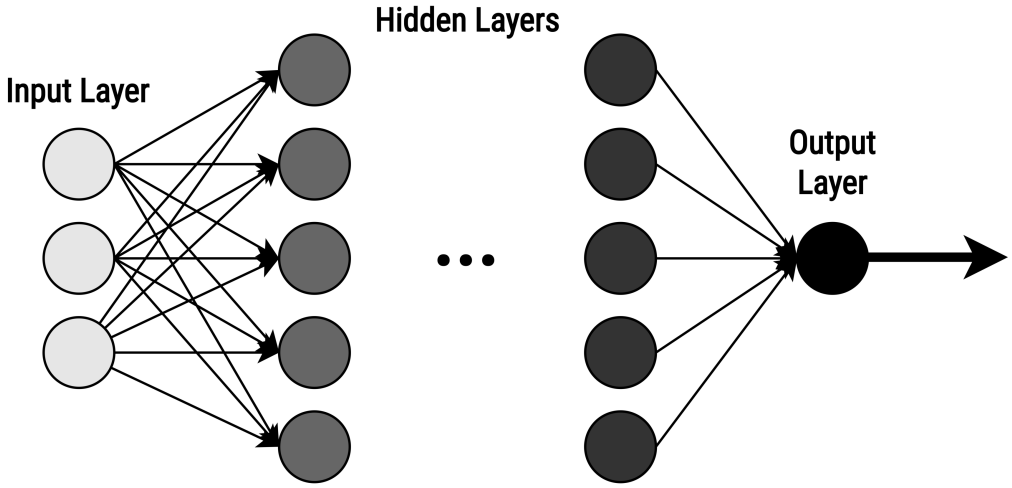


Figure 5.3: Fully-Connected Feed-Forward Neural Network architecture, wherein the information flows in one direction and all neurons are interlinked across the multiple hidden layers.

5.2.2 Convolutional Neural Network

Convolutional Neural Networks (CNNs) are inspired in the visual cortex, specifically in the receptive fields, where some neurons are only activated in the presence of stimulus in certain orientations, i.e., restricted regions of the visual space. The architecture of a typical CNN is organized as a series of layers, comprising convolutional layers and pooling layers.

Convolutional layers are composed of filters, which are arrays of weights, that slide over the entire input and convolute at each particular location, originating activation (feature) maps. Convolution is a specialized type of linear operation, described as an element-by-element multiplication between a particular location of the input (local patch) and the filter, followed by the sum of the results and to which is applied

an activation function. On that account, this type of neural network performs scatter interactions, limiting the number of connections for each input, conversely to traditional neural networks, where all the neurons are interlinked across the hidden layers. Figure 5.4 illustrates the convolution operation.

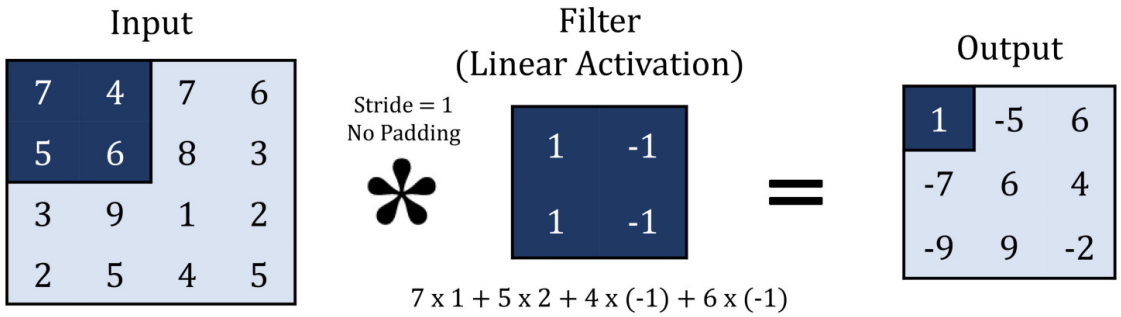


Figure 5.4: Convolution operation: element-by-element multiplication between local patches of the input and the filter, followed by the sum of the results and to which is applied an activation function.

The filters work as feature identifiers and can only extract a single type of features due to the parameter sharing, i.e., the weights associated with the filter are used in every position of the input where it slides over (local patches) and, therefore, in order to learn more types of features, it is necessary to use additional filters in parallel. On the other hand, the activation maps are identified as learnable feature maps and used as the input of the next layer. In that regard, each convolutional layer detects local conjunctions of features from the previous layer. Additionally, the output volume depth (number of feature maps) is equal to the number of filters and the depth of the filter has to be the same as the depth of the input. The output of each filter can be given by:

$$Output = \frac{Input\_Size - Filter\_Size + 2 * Padding}{Stride} + 1 \tag{5.7}$$

, where *stride* corresponds to the number of steps that a filter moves along the input (sliding size) and *padding* to the output size control.

Pooling layers reduce the spatial size of each feature map by replacing local patches of units with a single unit based on a specific function, e.g., max pooling extracts the maximum value and average pooling extracts the average of all values within the local patch. These layers are usually applied for dimensionality reduction and to preserve only the features associated with a certain motif rather than its exact location, considering that the relative positions of the characteristics forming a motif

may change. Moreover, this layer promotes the invariance of the input to translations and reduces the number of parameters to be learned in the following layers. The output of a pooling layer can be given by:

$$Output\_Pooling = \frac{Input\_Size - Pool\_Size}{Stride} + 1 \tag{5.8}$$

Overall, CNNs are identified as motif detectors and feature extractors, capable of retrieving deep patterns from the data by moving from low-level features to abstract concepts using learnable feature maps. Furthermore, they are capable of describing complex interactions between many variables using fewer interactions, resulting in better generalization capacity and reduced training costs. There are mainly three types of CNNs, specifically 1D, 2D, and 3D, according to the depth of the input. Figure 5.5 depicts the architecture of an CNN.

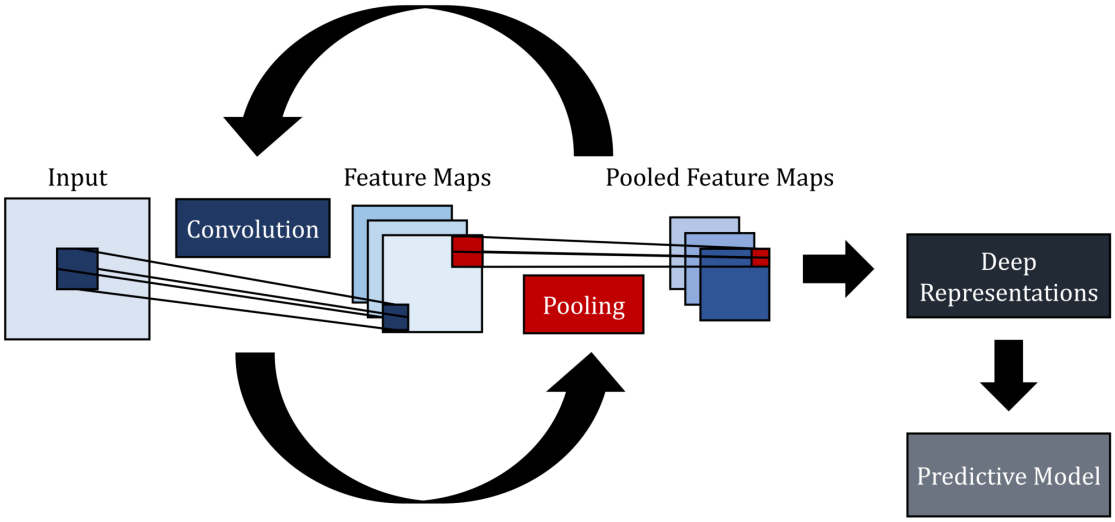


Figure 5.5: Convolutional Neural Network architecture, which comprises convolutional and pooling layers to extract deep representations from the input data.

5.2.3 Transformer-Encoder

Transformer-Encoders transform each token of the input sequence into a robust and contextual representation, which reflects the short and long-term dependencies, through the use of self-attention mechanisms [172]. Moreover, it is possible to extract an aggregated representation of the input using this architecture, reflecting the overall inter-dependencies amongst the tokens of the input sequence [173].

The Transformer-Encoder architecture stacks N identical blocks, where each block comprises a Multi-Head Self-Attention (MHSA) layer and a Position-Wise Feed-

Forward Neural Network (PWFFN). Residual connections are applied after each subunit followed by Layer Normalization (LN) to mitigate vanishing gradients. Additionally, dropout is added after each MHSA layer and after each dense layer of the PWFFN to prevent overfitting. Considering x^1 and x^2 the outputs of the MHSA layer and the PWFFN block, respectively, the output of the kth Transformer-Encoder block can be expressed as:

$$\begin{aligned} x_k^1 &= \mathbf{LN}(x_{k-1}^2 + \mathbf{dropout}(\mathbf{MHSA}(x_{k-1}^2))) \\ x_k^2 &= \mathbf{LN}(x_k^1 + \mathbf{PWFFN}(x_k^1)) \end{aligned} \tag{5.9}$$

, where x_k^1, $x_k^2 \in R^{N\ x\ d_{model}}$, N is the number of tokens in the input sequence, and d_{model} is the embedding dimension.

Figure 5.6 illustrates the architecture of the Transformer-Encoder.

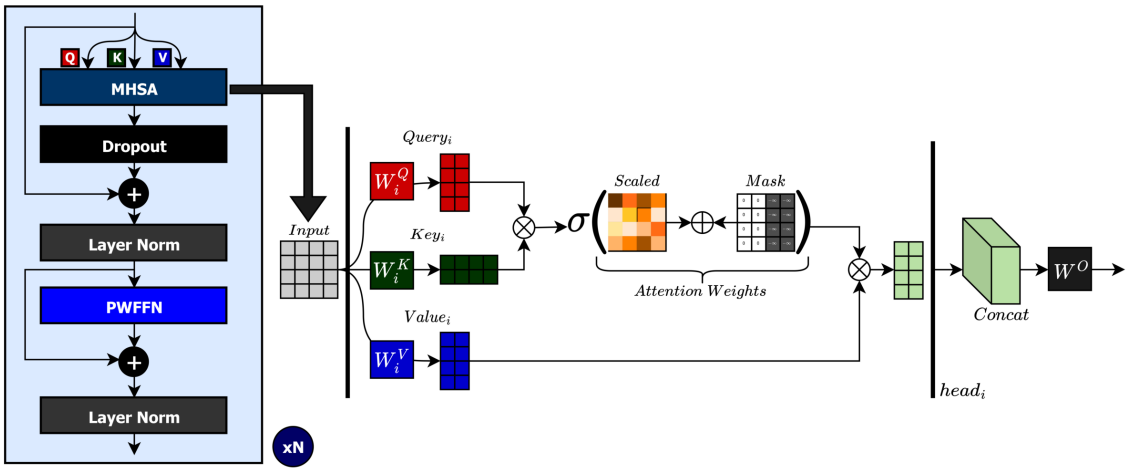


Figure 5.6: Transformer-Encoder architecture, where each block is composed of an MHSA layer and a PWFFN. The MHSA layer computes self-attention across h heads of attention, where each $head_i$ computes a weighted sum of V_{proj}^i. The attention weights are determined by applying a softmax (σ) to the scaled dot-product between Q_{proj}^i and K_{proj}^i, in which PAD tokens are masked. The PWFFN is applied to the last dimension of the MHSA outputs in order to give them an individually more robust representation.

5.2.3.1 Multi-Head Self-Attention

The MHSA layer determines the short and long-term inter-dependencies between the input elements by applying self-attention multiple times in parallel, resulting in a robust and contextual representation for each token of the sequence. This layer takes the input in the form of three parameters, specifically Query (Q), Key (K), and Value (V), which are generated from the same input sequence, and computes

attention across h heads of attention. On that account, Q, K, and V are linearly projected and divided into h sub-dimensions, where each $head_{1,..,h}$ computes a weighted sum of $V^{proj}_{1,...,h}$, i.e., each head maps a query and a set of key-value pairs to an output. The attention weights assigned to each element of $V^{proj}_{1,..,h}$ are determined by applying a softmax to the scaled dot-product between $Q^{proj}_{1,...,h}$ and $K^{proj}_{1,...,h}$. Furthermore, a masking matrix is added before the softmax to prevent the model from attention to certain tokens, e.g., PAD tokens. This masking matrix is obtained by assigning an extremely negative value (close to minus infinity) to the positions of the non-attending tokens, considering that in the softmax function values close to minus infinity lead to a probability of zero. The outputs of each head of attention are concatenated and linearly projected, resulting in the same dimensions as the input Q.

$$\begin{gathered}\mathbf{attn}(Q,K,V) = \mathbf{softmax}(\frac{QK^T}{\sqrt{d_K}} + Mask)V \\ \mathbf{MHSA}(Q,K,V) = [\mathbf{attn}(Q^{proj}_1,K^{proj}_1,V^{proj}_1); ...; \mathbf{attn}(Q^{proj}_h,K^{proj}_h,V^{proj}_h)]\mathbf{W}^O \\ Q^{proj}_{1,..,h} = QW^Q_{1,...,h},\ K^{proj}_{1,..,h} = KW^K_{1,...,h},\ V^{proj}_{1,..,h} = VW^V_{1,...,h}\end{gathered} \tag{5.10}$$

, where $Q \in R^{N\ x\ d_{model}}$, $K \in R^{N\ x\ d_{model}}$, $V \in R^{N\ x\ d_{model}}$, $W^Q_{1,...,h} \in R^{d_{model}\ x\ d_Q}$ are the Q projection matrices, $W^K_{1,...,h} \in R^{d_{model}\ x\ d_K}$ are the K projection matrices, $W^V_{1,...,h} \in R^{d_{model}\ x\ d_V}$ are the V projection matrices, $W^O \in R^{h\ x\ d_V\ x\ d_{model}}$ is the output projection matrix, $[;]$ denotes concatenation, $d_Q = d_K = d_V = \frac{d_{model}}{h}$, $Mask \in R^{N\ x\ N}$, N is the number of tokens in the input sequence, and d_{model} is the embedding dimension.

Figure 5.7 illustrates the architecture of an MHSA layer with h heads of attention in parallel.

5.2.3.2 Dropout Layer

Deep neural network architectures have many non-linear hidden layers and, thus, many complex relationships can be learned between inputs and outputs. On that account, neurons can overly adapt to each other during training (overfitting), resulting in increased noise. Dropout is a regularization strategy that helps reduce learning inter-dependency and improves the generalization of the model [174]. This approach deactivates a percentage p of randomly selected neurons and their connections during the training stage of the architecture, reducing the possibility of neurons developing co-dependency amongst each other. At the testing step (inferring process), the weights associated with the units that remain activated are multiplied by

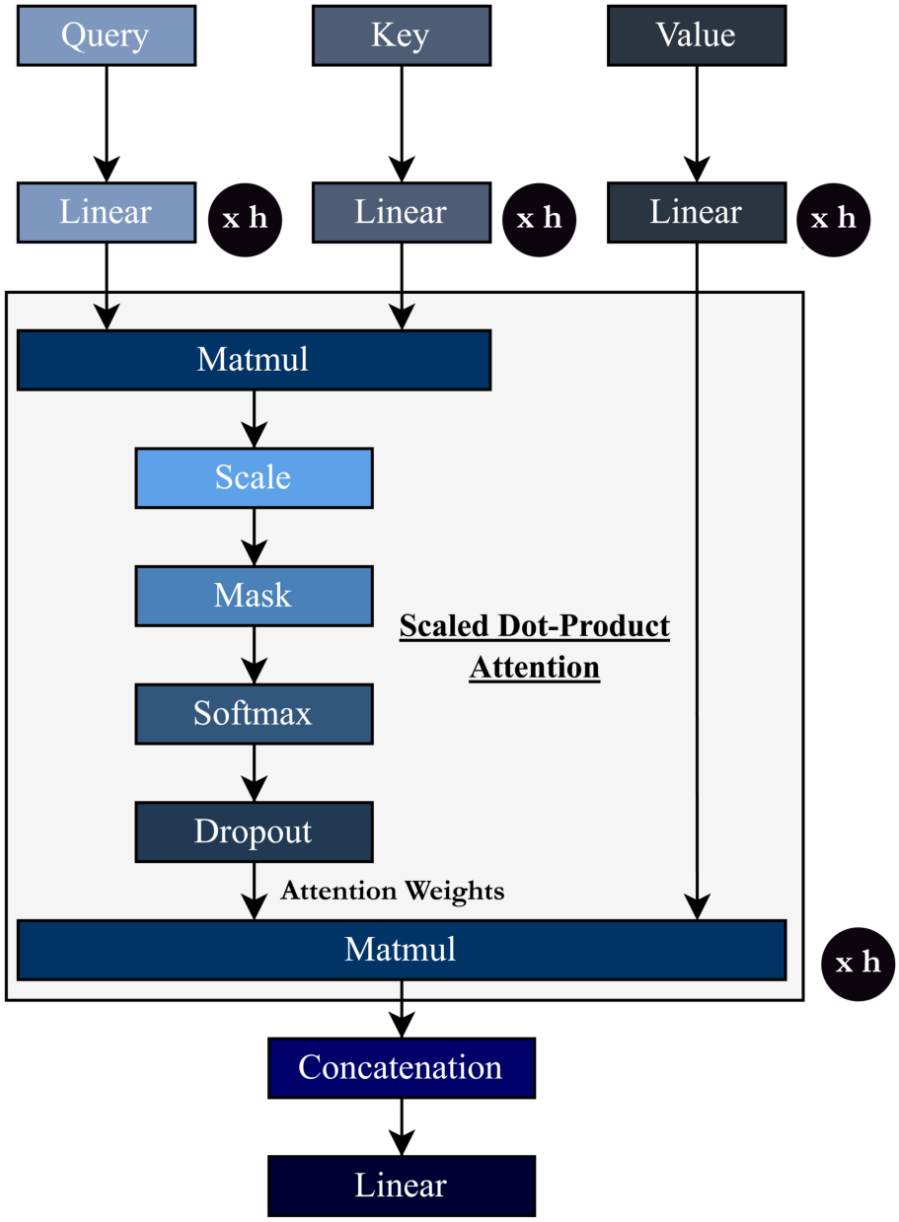


Figure 5.7: MHSA architecture, where each head of attention maps a query and set of key-value pairs to an output, which is computed as a weighted sum of the values. h is the number of heads of attention and *mask* corresponds to the masking of the PAD tokens.

the training forgetting rate p. Figure 5.8 illustrates the dropout method applied to a standard FCNN.

5.2.3.3 Layer Normalization

LN reduces the training time and enhances the generalization capacity of deep neural networks by redistributing the input values of each layer to a mean of approximately

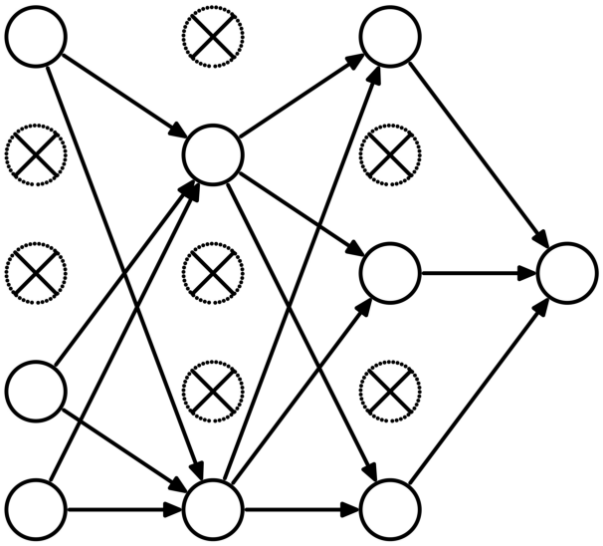


Figure 5.8: Dropout applied to a standard FCNN with 2 hidden layers. Figure adapted from "Dropout: A Simple Way to Prevent Neural Networks from Overfitting" [174].

zero and a standard deviation of one [175]. Moreover, this strategy is robust toward the scale and shift of the weight matrix and variations in the input scale. Consequently, it mitigates the occurrence of significant discrepancies in neuron values across different layers' inputs and reduces the probability of vanishing gradients. The mean (μ_l) and standard deviation (σ_l) associated with the neurons of a certain layer l are computed as follows:

$$\mu_l = \frac{1}{H_l}\sum_{j=1}^{H_l} x_{lj} \ , \ \sigma_l = \sqrt{\frac{1}{H_l}(\sum_{j=1}^{H_l} x_{lj} - \mu_l)^2} \tag{5.11}$$

, where H is the number of hidden units in the layer l and x are the inputs connected to the neurons associated with layer l. These two statistical measures are used to normalize x:

$$x'_{lj} = \frac{x_{lj} - \mu_l}{\sqrt{\sigma_l^2 + \epsilon}} \tag{5.12}$$

, where ϵ is a stability factor.

Additionally, this transformation is independent of the batch size and performs identically during the training and testing stages.

5.2.3.4 Position-Wise Feed-Forward Network

The PWFFN block improves the robustness of the representation of each token and increases the learning capacity of the architecture by projecting the last dimension (position-wise) of the attention outputs. The architecture of the PWFFN is similar to the FCNN, where all neurons are interlinked and the information flows in one direction. This block is composed of two dense layers, where the first dense layer projects the attention outputs to a higher dimension and the second dense layer projects it back to the initial last dimension. Thus, this block is usually compared to two 1x1 convolution layers. Moreover, dropout layers are applied after each dense layer of the PWFFN.

$$\mathbf{PWFFN}(x) = \mathbf{dropout}(\mathbf{F}_2(\mathbf{dropout}(\mathbf{F}_1(x\mathbf{W}_1 + \mathbf{b}_1)\mathbf{W}_2 + \mathbf{b}_2))) \tag{5.13}$$

, where $W_1 \in R^{d_{model}\ x\ \pi}$, $W_2 \in R^{\pi\ x\ d_{model}}$, $b_1 \in R^{\pi}$, $b_2 \in R^{d_{model}}$, F is the activation function, π is the expansion ratio, and d_{model} is the embedding dimension.

5.3 Similarity Methods

5.3.1 Smith-Waterman Algorithm

The Smith-Waterman algorithm is usually applied for local sequence alignment and to determined similar regions between protein sequences. It is a dynamic programming algorithm used to find the optimal local alignment with respect to the scoring system that is selected. This method initializes a matrix F, indexed by (i,j), where i and j correspond to the sequence length of the two protein sequences, respectively, and systematically fills the matrix based on a scoring function:

$$F(i,j) = max \begin{cases} 0 \\ F(i-1,j) - d \\ F(i,j-1) - d \\ F(i-1,j-1) + s(x_i,y_j) \end{cases}$$

, where d is the penalty for opening or extending gaps, and $s(x_i, y_j)$ is the score for matches or mismatches, where a substitution matrix, e.g., BLOSUM62, can be used instead. Furthermore, this method initializes the top left $(F(i,0),F(0,j))$ with zero, and finds the best local alignment using traceback from the highest score until it finds the first zero. The alignment score (similarity) for the two protein

sequences corresponds to the sum of the scores for each position associated with the optimal local alignment. Nevertheless, this method can lead to various local alignments. Figure 5.9 illustrates the Smith-Waterman algorithm applied to two protein fragments.

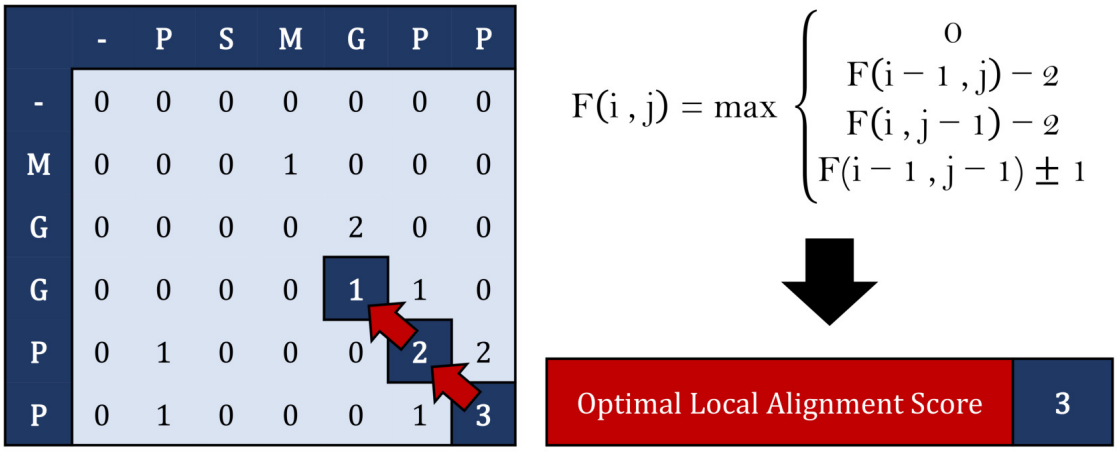


Figure 5.9: Smith-Waterman algorithm: optimal local alignment between "MG-GPP" and "PSMGPP", using d=-2 and $s(x_i,y_j)$=±1 (+1 for match and -1 for mismatch).

5.3.2 Tanimoto Coefficient

The Tanimoto coefficient is a distance metric used to determine the similarity between two finite sample sets and takes into account the ratio between the intersection and union of the two sample sets. Considering A and B two different finite sample sets, the Tanimoto coefficient (T) is given by:

$$T(A,B) = \frac{|A \cap B|}{|A| + |B| - |A \cap B|} \tag{5.14}$$

This method is usually applied to determine the similarity between chemical compounds based on their hashed binary chemical fingerprint representations, which are bitmap strings that contain information the presence or absence of particular substructures. Considering i and j the vector (fingerprint) representations of two different compounds, respectively, the Tanimoto coefficient (T) can be expressed as:

$$T(i,j) = \frac{i \cdot j}{|i|^2 + |j|^2 - i \cdot j} \tag{5.15}$$

5.4 Evaluation Metrics

There are many metrics used to evaluate the performance and the capacity of the models as predictors, where the choice of which ones to use highly depends on the context of the problem and on the distribution of the target vector (labels). In spite of the fundamental purpose of comparing predicted labels with true labels, each evaluation metric assesses specifics aspects and is influenced differently by the distribution of the labels and outcomes.

5.4.1 Binding Affinity Prediction (Regression)

- **Mean Squared Error (MSE)**: measures the average squared difference between the predicted values and the real values.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{5.16}$$

  , where n is the number of samples, y_i is the real value, and $\hat{y}_i$ is the predicted value.

- **Root Mean Squared Error (RMSE)**: measures the square root of the average squared difference between the predicted values and the real values.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{5.17}$$

  , where n is the number of samples, y_i is the real value, and $\hat{y}_i$ is the predicted value.

- **Concordance Index (CI)**: measures the probability of non-equal pairs being correctly predicted in terms of order.

$$CI = \frac{1}{Z}\sum_{y_i > y_j} h(p_i - p_j),\ h(p) = \begin{cases} 1, p > 0 \\ 0.5, p = 0 \\ 0, p < 0 \end{cases} \tag{5.18}$$

  , where Z corresponds to the number of non-equal pairs, p_i to the predicted value for the larger affinity y_i, and p_j to the predicted value for the smaller affinity y_j.

- **Coefficient of Determination (r^2)**: measures the ratio between the total vari-

ance explained by the model and the total variance.

$$r^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y_i})^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} \tag{5.19}$$

, where $\hat{y_i}$ is the predicted value, y_i is the real value, and $\overline{y}$ is the mean of the real values.

- **Spearman Rank Correlation (Spearman)**: measures the strength and direction of association between two ranked variables (non-parametric).

$$Spearman = \frac{\frac{1}{n}\sum_{i=1}^{n}(R(y_i) - \overline{R(y)}) \cdot (R(\hat{y_i}) - \overline{R(\hat{y})})}{\sqrt{(\frac{1}{n}\sum_{i=1}^{n}(R(y_i) - \overline{R(y)})^2) \cdot (\frac{1}{n}\sum_{i=1}^{n}(R(\hat{y_i}) - \overline{R(\hat{y})})^2)}} \tag{5.20}$$

, where $R(\hat{y_i})$ is the predicted value rank, $R(y_i)$ is the real value rank, $\overline{R(\hat{y})}$ is the mean of the predicted values ranks, and $\overline{R(y)}$ is the mean of the real values ranks.

5.4.2 Binding Pocket Prediction (Binary Classification)

- **Balanced Accuracy**: arithmetic mean of sensitivity and specificity.

$$\begin{aligned} Balanced\ Accuracy &= \frac{Sensitivity + Specificity}{2} \\ Sensitivity &= \frac{TP}{TP + FN} \\ Specificity &= \frac{TN}{TN + FP} \end{aligned} \tag{5.21}$$

, where TP are the True positives, TN are the True negatives, FP are the False positives, and FN are the False negatives.

- **Precision**: proportion of true positives to all positive predictions.

$$Precision = \frac{TP}{TP + FP} \tag{5.22}$$

, where TP are the True positives and FP are the False positives.

- **Recall**: rate of positives correctly classified.

$$Recall = \frac{TP}{TP + FN} \tag{5.23}$$

, where TP are the True positives and FN are the False negatives.

- **F1-Score**: harmonic mean of precision and recall.

$$\begin{aligned} F1 - Score &= 2 * \frac{precision * recall}{precision + recall} \\ Precision &= \frac{TP}{TP + FP} \\ Recall &= \frac{TP}{TP + FN} \end{aligned} \tag{5.24}$$

  , where TP are the True positives, FP are the False positives, and FN are the False negatives.

- **Matthew's Correlation Coefficient (MCC)**: correlation between two binary variables.

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{5.25}$$

  , where TP are the True positives, TN are the True negatives, FP are the False positives, and FN are the False negatives.

Chapter 6

Explainable Deep Drug–Target Representations

This chapter concerns the use of a post-hoc explainability algorithm to explore and provide potential explanations for the decision-making process of complex models such as CNNs, which are employed to identify and extract deep patterns from input data, in the context of DTA prediction. This study also probes the correlation between input regions that had a positive influence on the prediction and relevant regions in the DTI domain.

The content of this chapter is based on a journal article published in *BMC Bioinformatics* [341]. Section 6.1 presents the study context. Section 6.2 details the materials and methods used in this study. Section 6.3 reports the results and discusses the obtained findings. Section 6.4 provides some final reflections and the limitations of this study.

6.1 Study Context

The accurate identification of novel DTIs and the understanding of the binding process are determinants in the discovery of potential hit-to-lead compounds. Despite the *in silico* research advances, the majority of these studies rely on binary associations to conduct their experiments, neglecting the importance of the binding affinity [14]. Thus, the quality of the predictions is usually compromised or least limited, particularly when considering secondary interactions with off-targets [16, 17, 342]. Moreover, most studies seldom characterize proteins and compounds using sequential and structural data, making use of global descriptors and certain topological or physicochemical properties, which are mostly not robust and remarkably limiting to DTI understanding and model explainability [35].

DL architectures have gained acceptance in recent years due to their ability to exploit

and learn from comprehensive chemical and proteomic libraries, retrieve unprecedented knowledge in DTIs, and identify complex and discriminating patterns within the input data [27]. Furthermore, these architectures typically do not require feature engineering tools given their capacity to extract features with increased selectivity from structured or unstructured raw data and, thus, outperform most traditional ML algorithms. However, these complex models are still considered opaque and devoided of transparency in their inner decisions and results, despite the high performance achieved [28, 29]. Moreover, the binding process between an active compound and a protein is unequivocally complex, especially given the range of different regions across the whole structure of proteins and compounds that directly or indirectly participate in the interaction [34, 39]. Hence, the lack of explainability in the predictions compromises the comprehension and identification of the underlying aspects of the interaction, considering that it is not possible to directly associate the output with the input domain in DL models [300]. Additionally, when taking into account the context of the problem, wherein the decision presented may have a great impact on the drug discovery process chain, it is vital to understand and provide possible explanations for the reasoning behind the decisions of these complex architectures [31, 311].

This study explores the use of an end-to-end DL approach to predict DTA measured in terms of the dissociation constant (K_d), where 1D sequential and structural data, protein sequences and SMILES strings, are used to represent the targets and compounds, respectively. Furthermore, it aims to provide explainability and validate the decision-making process of CNNs when extracting deep features from protein sequences and SMILES strings in the context of DTIs. On that account, three critical points were investigated in this work: a) efficiency of the deep representations in the prediction of a real-valued interaction strength; b) reliability of CNNs in the identification of important sequential and structural regions for the binding process; and c) robustness of the features extracted from relevant sequential regions.

6.2 Materials and Methods

6.2.1 Binding Affinity Prediction

6.2.1.1 Drug–Target Interaction Pairs

In order to establish the binding affinity prediction model, it was necessary to collect DTI pairs characterized with binding affinity measured in terms of K_d. However,

standard experimentally validated, also known as *gold standard*, datasets are extremely scarce in the context of the problem. On that account, in order to conduct this study and the experiments, the data from the Davis et al. (2011) [276] research study was explored, considering that it is the only benchmark dataset that contains interactions characterized with their respective dissociation constant values. The Davis et al. (2011) [276] dataset comprises selectivity assays related to the human catalytic protein kinome, resulting in a total of 31 824 interactions between 72 kinase inhibitors (compounds) and 442 kinases (proteins). Furthermore, all compounds and proteins in this dataset are interlinked, which is critical to preserve their overall representability, given that the number of observations for each protein and compound has a great impact on their relative importance and influence during the learning stage.

The protein sequences of the Davis dataset were collected from UniProt [57] based on the corresponding accession numbers (identifiers). Proteins are characterized by a unique amino acid sequence, resulting in varying sequence lengths. Thus, to standardize the number of features and avoid the loss of relevant sequential information or increased noise due to excessive padding, the protein sequence length was fixed between 264 and 1400 residues based on a 95% information density threshold. Protein sequences shorter than the maximum length were padded.

The SMILES strings of the Davis dataset were extracted from PubChem [343] based on their compound identifiers (CIDs). To ensure a consistent notation to represent the chemical structure of all compounds, the RDKit [344] canonical transformation was applied to every SMILES string. Even though the canonical notation does not include stereochemical information, it is a unique representation, where the atoms are consistently numbered. Similar to the protein sequences, the sequence length of the SMILES strings was fixed between 38 and 72 chemical characters based on a 95% information density threshold. SMILES strings shorter than the maximum length were padded.

The distribution of the Davis K_d values is significantly skewed toward K_d equal to 10 000 nM (22 400 interaction pairs out of 31 824), which is associated with extremely weak or almost non-existing interactions. Furthermore, the variance of this distribution is considerably high, since it ranges from values close to zero (strong interaction) to high values (weak binding). Hence, in order to reduce the effects of the high variance of this distribution on the learning loss, the K_d values were transformed into the logarithmic space (pK_d) using Equation 6.1. The distribution

of the pK_d values spans from 5 (10 000 nM) to approximately 11.

$$pK_d = -log_{10}(\frac{K_d}{10^9}) \tag{6.1}$$

Table 6.1 summarizes the statistics of the original and pre-processed Davis dataset.

Table 6.1: Original and pre-processed Davis dataset [276]: unique proteins, compounds, and DTIs.

Davis Kinase Dataset					
	Proteins	**Compounds**	**DTI**	**pKd = 5**	**pKd > 5**
Original	442	72	31 824	22 400	9424
Pre-Processed	423	69	29 187	20 479	8708

See Figures B.1 and B.2 in Section B.1 of Appendix B for more details regarding the distribution of the Davis dataset.

6.2.1.2 Data Representation and Encoding

Protein sequences and SMILES strings are constituted by different sequential and structural characters, respectively, which are used as input for the binding affinity prediction model. A dictionary-based approach was considered to encode each one of the characters into an integer according to the number of unique tokens, resulting in a 20-character dictionary for the protein sequences and a 26-character dictionary for the SMILES strings. In order to normalize the importance of each one of these integer values and preserve only the structural information, one-hot encoding was applied, assigning a binary variable for each unique integer value and converting every integer into a binary vector. Figure 6.1 illustrates the dictionary-based approach and the one-hot encoding applied to the AKK1 kinase.

6.2.1.3 Binding Affinity Prediction Model

In order to predict real-valued DTI strength measured in pK_d, an end-to-end DL model based on CNNs and FCNNs was explored, where 1D sequential and structural information, specifically protein amino acid sequences and SMILES strings, respectively, were used as input.

The protein sequences and SMILES strings were initially processed based on their length and then encoded according to the dictionary-based approach mentioned in Section 6.2.1.2. Considering that these integer values are recognized as categorical variables, a one-hot encoder layer was assigned to both protein sequences and

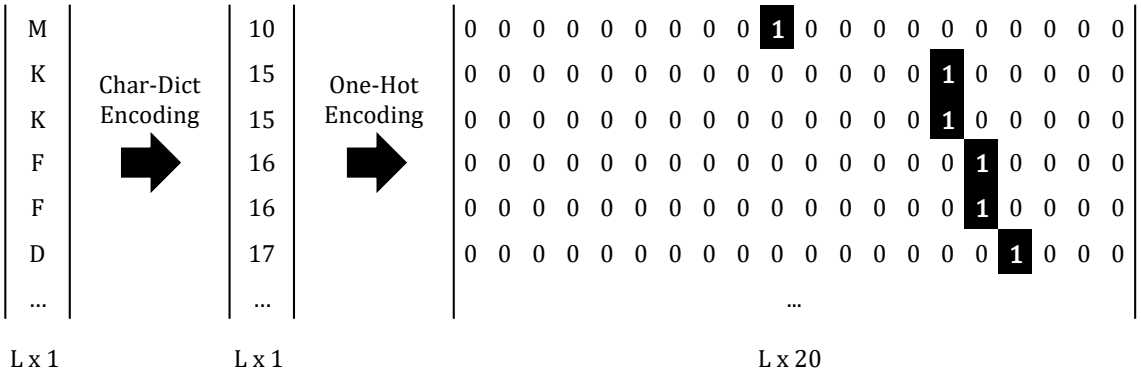


Figure 6.1: Dictionary-based encoding followed by one-hot encoding applied to the kinase AKK1, where L is the length of the protein sequence.

SMILES strings, respectively. Following the one-hot encoding, two parallel series of 1D convolutional layers were considered, one for the protein sequences and the other for the SMILES strings. These sets of convolutional layers were employed to uncover deep patterns within the input data and automatically surmise and identify important sequential and structural regions for the interaction. Following each series of convolutional layers, global max pooling was applied to reduce the spatial size of each feature map to its maximum representative feature, i.e., to capture the most relevant motifs. The resulting deep representations were concatenated into a single vector, comprising the most significant sequential and structural features, and used as input for an FCNN, in which dropout regularization was applied between each fully-connected layer. Following the FCNN, an output layer composed of one neuron was employed to return the real-valued interaction strength measured in pK_d. The proposed end-to-end DL architecture is illustrated in Figure 6.2.

6.2.1.4 Chemogenomic Representative K-Fold

In order to find the best model architecture and set of parameters, a variation of the stratified K-fold cross-validation approach was explored. In the context of the problem, a randomly split of the dataset leads to non-representative sets of DTIs, especially when considering the imbalanced distribution of pK_d values and that the majority of the interactions are characterized with a pK_d value equal to 5 (extremely weak interactions). Moreover, considering that 1D raw data is used to characterize the proteins and compounds, specifically the amino acid sequence and the SMILES string, respectively, the overall representability of the data plays an important role in the learning process of the architecture. Thus, in order to split the dataset into representative sets, it is crucial to consider the protein sequences similarity, the SMILES strings similarity, and the pK_d values distribution during the splitting

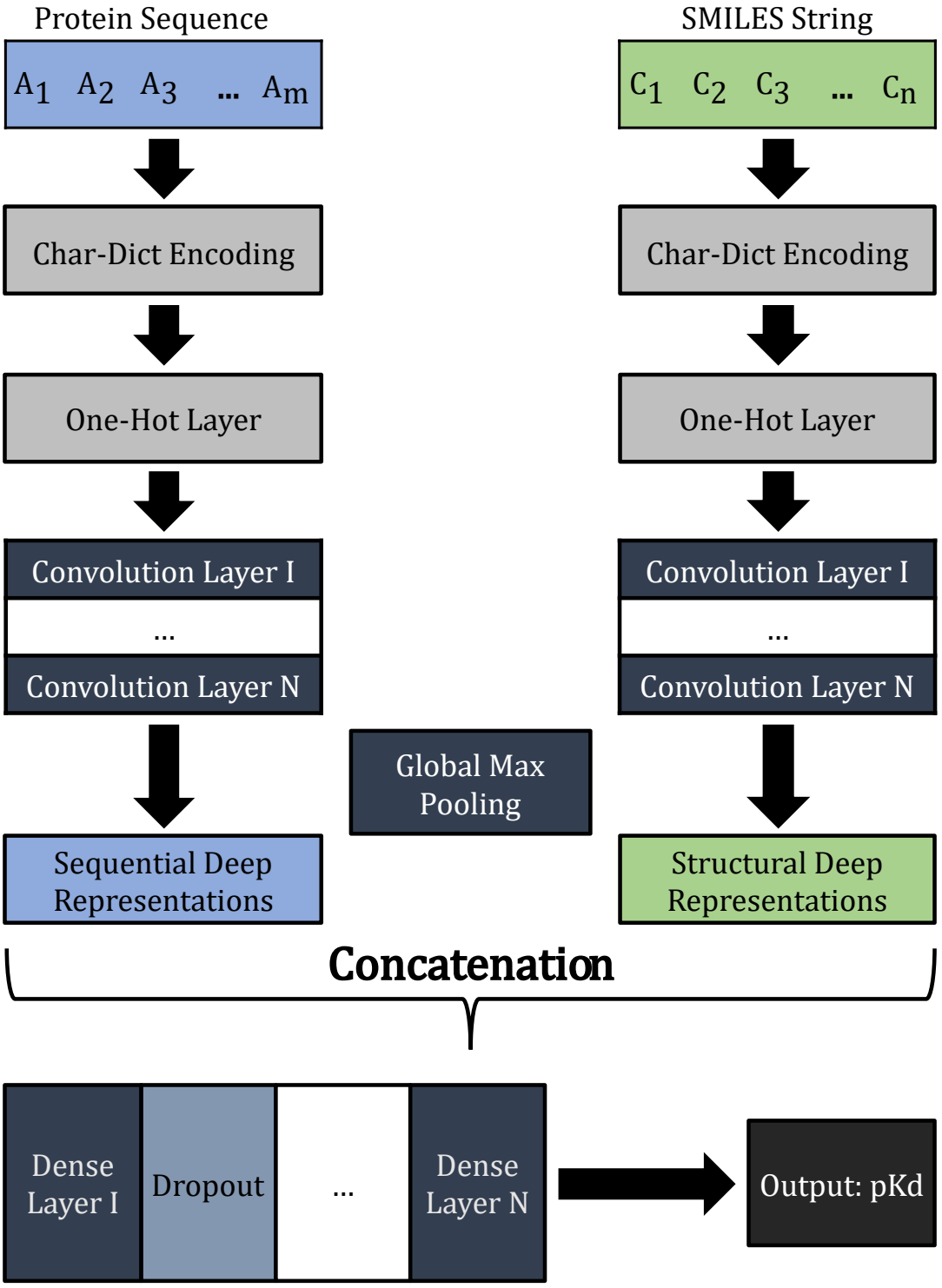


Figure 6.2: CNN-FCNN binding affinity prediction model. Two parallel series of 1D CNNs uncover deep patterns and extract deep representations from protein sequences and SMILES strings, respectively. The resulting deep representations, comprising the most relevant and significant sequential and structural motifs, are concatenated and used as input for an FCNN, which predicts the binding affinity measured in terms of pK_d.

process.

The proposed method, Chemogenomic Representative K-Fold, initially splits the data into two different groups according to the pK_d value, specifically greater than 5 or equal to 5, respectively. Following the sampling process, the samples with a pK_d value greater than 5 are initially distributed across the different K folds based on the lowest similarity score (dissimilarity score). The first K samples of this group are assigned to each K set in order to initialize each fold, and the remaining $N_I - K$ samples (N_I is the number of DTI pairs in the dataset with a pK_d value greater than 5) are distributed based on their dissimilarity score. The dissimilarity score corresponds to the lowest similarity score between the sample and each K set, in which the sample is assigned to the set with the lowest similarity score. The similarity score is computed as the weighted mean between the median value across all the protein sequences' similarity scores and the median value across all the SMILES strings' similarity scores, which are calculated (e.g., obtained from similarity matrices) between the sample and each entry in the corresponding set, i.e., between the protein sequence of the sample and all the protein sequences in the corresponding set, and between the SMILES string of the sample and all the SMILES strings in the corresponding set. In order to guarantee that each set is equally sized, only sets that had not previously been assigned a sample are considered at each step (until it is reset), thus, the dissimilarity score corresponds to the lowest similarity convex combination across all $K - m$ sets, where $m = 1,..,K-1$ is associated with the number of sets that had previously been assigned a sample. Following the pairs with a pK_d value greater than 5, this process is repeated for the remaining N_{II} samples, which correspond to the DTI pairs with a pK_d value equal to 5 (weak interactions).

This approach leads to equally sized representative sets, prioritizing the relevant interactions. Furthermore, considering that this method splits the data according to the lowest similarity score (improved representability), it is possible to extract an independent testing set to evaluate the model's generalization capacity. The Chemogenomic Representative K-Fold is illustrated in Figure 6.3.

6.2.2 Explainable Binding Affinity Prediction

6.2.2.1 Binding Sites

The interaction between compounds and proteins results from the recognition and complementarity of certain active and functional groups (binding sites). On that

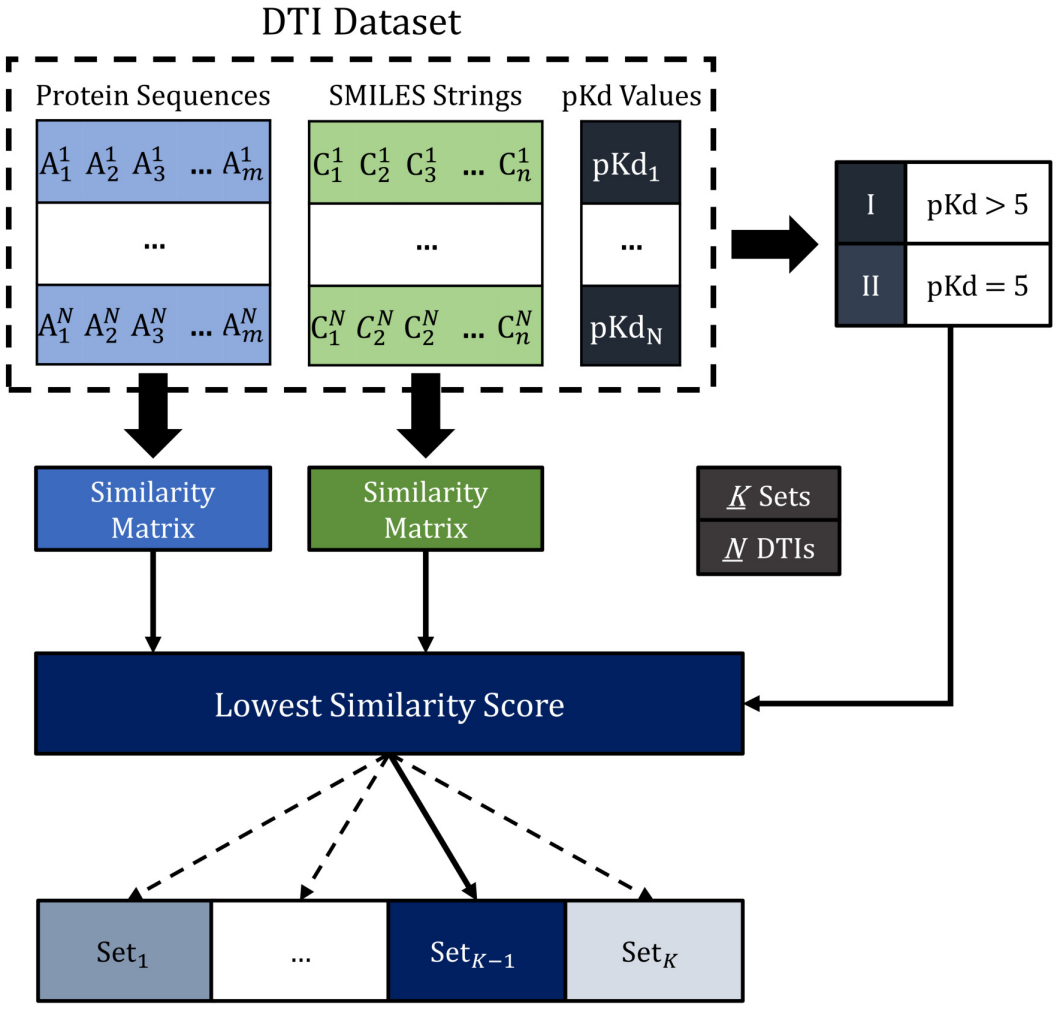


Figure 6.3: Chemogenomic Representative K-Fold, where DTI pairs are distributed based on the pK_d value, protein sequence similarity, and SMILES string similarity. The DTI pairs with a $pK_d > 5$ are initially assigned to the K set with the lowest similarity score followed by the DTI pairs with a $pK_d = 5$. The similarity score corresponds to the weighted mean between the median value across all the protein sequences' similarity scores and the median value across all the SMILES strings' similarity scores, which are computed between the sample and each entry in the corresponding set.

account, the protein amino acid sequence, specifically the binding regions within the protein, and the compound's chemical structure are determinants for the binding. Considering the range of different regions across the whole structure of proteins and compounds, respectively, the relevance given to certain spots might introduce bias in the predictions, compromising the validity of the inferring process of the model. Thus, apart from providing visual explanations to the predictions inferred by the proposed model, it is determinant to evaluate the relevance and significance given to the regions identified as important for the prediction, i.e., the model's reliability

in identifying binding spots as regions of interest.

In order to conduct this evaluation, it was necessary to identify the binding regions of the interaction pairs in the dataset, although the number of DTIs with the exact binding regions known or available represents only a small subset of the whole DTI universe. Nevertheless, the sc-PDB database [345], which is an annotated database of druggable binding sites, contains some DTI pairs with interactions sites known and, thus, it was explored to collect DTIs with binding sites annotated. The DTI pairs from this database were initially pre-processed, where only entries belonging to the taxonomic identifier 9606 (*Homo sapiens*) were selected, considering that the proposed model was trained using DTI pairs associated with the human catalytic protein kinome. The remaining samples were then processed according to the protein and SMILES string length thresholds, where any entry with a sequential size outside the thresholds defined in Section 6.2.1.1 was removed, respectively. Moreover, the real-valued interaction strength measured in K_d is not known for the majority of the DTI pairs in this database, thus, their binding affinity had to be initially predicted by the proposed model, where only pairs with a predicted pK_d greater than 5 were considered. The final DTI pairs were then divided into two groups, specifically those that are also present in the Davis dataset (Davis ∩ sc-PDB pairs) and the ones that are exclusively from the sc-PDB database. Table 6.2 summarizes the number of DTIs, the average number of binding sites for each DTI, and the number of unique proteins and compounds for the two datasets.

Table 6.2: Statistics of collected binding sites datasets from the sc-PDB database [345].

	DTI	**$\overline{\text{Binding Sites}}$**	**Proteins**	**Compounds**
Davis ∩ sc-PDB	32	16	27	8
sc-PDB	266	12	64	249

$\overline{\text{Binding Sites}}$ corresponds to the average number of binding sites annotated for each DTI.

6.2.2.2 Protein Evolutionary Conserved Motifs

Many proteins are functionally and evolutionarily related, where certain regions (motifs/profiles), usually associated with important protein functions/activities, e.g., binding, folding, or secondary interactions, are conserved. Thus, apart from understanding if the CNNs are identifying and assigning importance to the binding sites, it is also relevant to explore if there is any association between the input regions

selected by the model that are not in the vicinity of the binding regions and the motifs that are usually conserved.

These profiles can be obtained using PSI-BLAST [346], which iteratively searches for regions of similarity between the protein query sequence and a target protein database. This method scores the matches using PSSMs instead of pre-defined scoring matrices, where final high-scoring sequences (filtered using an expectation threshold) found at each iteration step are multiple aligned to produce a new PSSM to be used in the next iteration. The resulting PSSM provides evolutionary information for the protein query sequence, where each position (a, l), $a = \{1,2,..,20\}$ and $l = \{1,2,..., L\}$ corresponding to the number of possible amino acids and the protein sequence length, respectively, is the probability of the pattern l in the protein sequence diverge to another amino acid a. On that account, the amino acids that have a high score (probability) of not diverging are considered to be evolutionarily conserved and thus associated with important activities of the protein.

In order to obtain the PSSMs for the Davis ∩ sc-PDB and sc-PDB pairs, a standalone version of PSI-BLAST [346] from blast+ 2.11.0 [347] was explored. The database selected was the non-redundant (nr), the number of iterations was fixed at 3, the E-value chosen was 0.001, and the search was restricted to the taxonomic group 9606. Considering that the PSSM scores range from negative values up to a maximum of 10 (highest probability), different thresholds were considered to select the conserved motifs, specifically from 5 to 10. Table 6.3 summarizes the average number of conserved motifs for the Davis ∩ sc-PDB and sc-PDB pairs across different thresholds.

Table 6.3: Average number of conserved motifs across different thresholds for the Davis ∩ sc-PDB and sc-PDB pairs.

	PSSM Motifs Threshold					
	≥ 5	**≥ 6**	**≥ 7**	**≥ 8**	**≥ 9**	**≥ 10**
Davis ∩ sc-PDB	276	162	88	45	19	11
sc-PDB	191	121	69	36	15	8

6.2.2.3 Gradient-Weighted Regression Activation Mapping

Gradient-Weighted Class Activation Mapping (Grad-CAM) [334] is a gradient-based method that provides visual explanations for the decisions associated with CNN-based architectures, producing coarse localization maps that highlight the important

regions for prediction. This method is a generalization of the Class Activation Mapping (CAM) [348] and it uses the gradient information flowing into the last convolutional layer to assign importance to each neuron for a particular decision of interest. The class discriminative localization maps are obtained by performing a linear (weighted) combination of the forward feature maps of the convolutional layer with the neuron importance weights, which is followed by a Rectified Linear Unit (ReLU) in order to obtain the features that have a positive influence on the class of interest.

$$L^c_{Grad-CAM} \in R^{uxv} = ReLU(\sum_k \alpha^c_k A^k) \tag{6.2}$$

, where $L^c_{Grad-CAM} \in R^{uxv}$ is the class discriminative localization map of width u and height v for the class of interest c, k is the number of feature maps, A^k is the kth feature map activations, and α^c_k is the neuron importance weights connecting the kth feature map activations with the cth class.

In order to obtain the neuron importance weights α^c_k, which capture the importance of the feature map k for the target class c, the gradients of the score for the class of interest (y_c) concerning the feature map activations A^k of the convolutional layer are computed through backpropagation and global average pooled over the width and height dimensions of the feature map.

$$\alpha^c_k = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A^k_{ij}} \tag{6.3}$$

, where $\frac{1}{Z} \sum_i \sum_j$ corresponds to the global average pooling (Z is the number of points/pixels in the feature map), $\frac{\partial y^c}{\partial A^k_{ij}}$ to the gradient of the score of class c with the respect to the feature map activations A^k, and i and j to the width and height dimensions, respectively, of the feature map.

The DL framework employed in this study focuses on a regression task instead of a classification problem. Thus, in the context of the problem, it was important to identify the discriminative regions toward the regression outcome, specifically the sequential and structural regions in the protein sequences and SMILES strings, respectively, that were considered to be important for the prediction of binding affinity. On that account, an adaptation of the Grad-CAM approach called Gradient Weighted Regression Activation Mapping (Grad-RAM) was proposed, which computes the gradients of the regression outcome with respect to the feature map activations. Similar to the Grad-CAM method, these gradients are global average pooled, leading to neuron importance weights that capture the importance of the

feature map activations for the interaction strength. The resulting regression discriminative localization maps are capable of explaining the output layer decisions by identifying the relevant sequential and structural regions for prediction.

$$L_{Grad-RAM} = ReLU(\sum_{k}(\frac{1}{Z}\sum_{i}\sum_{j}\frac{\partial\hat{y}}{\partial A_{ij}^{k}})A^{k}) \tag{6.4}$$

, where $L_{Grad-RAM}$ is the regression discriminative localization map for the predicted value $\hat{y}$, $\frac{1}{Z}\sum_i\sum_j$ corresponds to the global average pooling, $\frac{\partial\hat{y}}{\partial A_{ij}^{k}}$ to the gradient of the regression outcome $\hat{y}$ with respect to the feature map activations A^k of the convolutional layer, and i and j to the width and height dimensions, respectively, of the feature map.

Global Max Pooling. In image or object localization/detection tasks, global average pooling encourages the network to identify the complete extent of the object, considering that the average of a feature map takes into account both discriminative and low-activation regions. However, in the context of the problem, the interaction is determined by structural and sequential regions scattered in a 1D dimension. Hence, global max pooling was of special interest since the goal was to identify single discriminative spots.

$$L_{Grad-RAM} = ReLU(\sum_{k} max(\frac{\partial\hat{y}}{\partial A^{k}})A^{k}) \tag{6.5}$$

, where $L_{Grad-RAM}$ is the regression discriminative localization map for the predicted value $\hat{y}$, max corresponds to the global max pooling, $\frac{\partial\hat{y}}{\partial A^{k}}$ to the gradient of the regression outcome $\hat{y}$ with respect to the feature map activations A^k of the convolutional layer.

Guided (Positive) Gradients. In the work of Selvaraju et al. (2020) [334], the authors proposed an adaptation of their Grad-CAM method by combining their visualizations with the Guided backpropagation approach proposed by Springenberg et al. (2015) [349], in which negative gradients are suppressed when backpropagating through ReLU layers. Considering that visualizing the sequential and structural regions that have the highest positive influence on the prediction of binding affinity was of special interest, a variation of Grad-RAM was also explored by masking all the gradient positions associated with negatives values or where the activations of the feature maps were not greater than zero.

$$\frac{\partial \hat{y}}{\partial A^k} = (A^k > 0) \cdot (\frac{\partial \hat{y}}{\partial A^k} > 0) \cdot \frac{\partial \hat{y}}{\partial A^k} \tag{6.6}$$

, where $\frac{\partial \hat{y}}{\partial A^k}$ is the gradient of the regression outcome $\hat{y}$ with respect to the feature map activations A^k of the convolutional layer.

6.3 Results and Discussion

6.3.1 Prediction efficiency of the deep representations

The accurate and reliable prediction of a real-valued interaction strength is critical in the path of new findings regarding DTIs. In this study, an end-to-end DL architecture was proposed, in which CNNs were leveraged due to their capacity to automatically identify and extract deep representations from relevant and important sequential and structural regions associated with DTIs. In order to validate the prediction efficiency of the proposed architecture (CNN-FCNN), the performance was evaluated and compared with different state-of-the-art baselines. Additionally, the efficiency of the features extracted from the CNNs was further validated by evaluating and comparing the performance of using those deep representations as input for traditional ML models (see Section B.2.1 in Appendix B for more details regarding the experimental setup conducted in this study). Table 6.4 reports the binding affinity prediction results over the Davis independent testing set in terms of five different metrics: MSE, RMSE, CI, r^2, and Spearman (See Table B.4 in Section B.3.1 of Appendix B for the binding affinity predictions results using the original split methodology of the state-of-the-art research works).

The results demonstrate that the CNN-FCNN model achieved the highest performance in terms of MSE (0.177), RMSE (0.421), CI (0.915), Spearman (0.725), and r^2(0.789) compared to state-of-the-art baselines. Hence, it exceeds the other models in its capacity to correctly predict the binding affinity value (lower MSE and RMSE) and distinguish the binding strength rank order across DTI pairs (higher CI).

Regarding the efficiency of the deep representations, the results validate the effectiveness of CNNs in their capacity to extract relevant deep representations from sequential and structural data, especially when considering the performance achieved in terms of CI, which is significantly high across all models and superior to the state-of-the-art baselines (except the KRR model). Albeit the accurate prediction of the interaction strength value, assessed in terms of MSE and RMSE, is important in the context of the problem, the ability to correctly distinguish the binding

Table 6.4: Binding affinity prediction results over the Davis testing set.

Method	Protein Rep.	Compound Rep.	↓ MSE	↓ RMSE	↑ CI	↑ r^2	↑ Spearman
Baseline Methods							
KronRLS [265]	Smith-Waterman	PubChem-Sim	0.443	0.665	0.847	0.473	0.624
GraphDTA-GCN [271]	1D	Graph	0.315	0.561	0.879	0.625	0.676
GraphDTA-GATNet [271]	1D	Graph	0.307	0.554	0.875	0.634	0.670
SimBoost [267]	Smith-Waterman	PubChem-Sim	0.277	0.526	0.891	0.670	0.694
Sim-CNN-DTA [273]	Smith-Waterman	PubChem-Sim	0.266	0.516	0.884	0.683	0.674
GraphDTA-GIN [271]	1D	Graph	0.255	0.505	0.889	0.696	0.690
GraphDTA-GAT-GCN [271]	1D	Graph	0.254	0.504	0.885	0.697	0.683
DeepDTA [268]	1D	1D	0.222	0.472	0.888	0.735	0.678
DeepCDA [272]	1D	1D	0.202	0.449	0.882	0.760	0.668
Proposed Method							
CNN-FCNN	1D	1D	**0.177**	**0.421**	**0.915**	**0.789**	**0.725**
Deep Representations Eval.							
SVR	CNN Deep Representations		0.203	0.450	0.907	0.759	0.714
GBR	CNN Deep Representations		0.271	0.520	0.894	0.677	0.699
RFR	CNN Deep Representations		0.283	0.532	0.895	0.663	0.703
KRR	CNN Deep Representations		0.453	0.673	0.848	0.461	0.630

Bold indicates the best performance value associated with each evaluating metric.
RFR-Random Forest Regressor, SVR-Support Vector Regressor, GBR-Gradient Boosting Regressor, KRR-Kernel Ridge Regression

strength rank order between two different DTI pairs is of special interest, since it allows to differentiate primary from secondary or not so relevant interactions. On that account, the deep representations extracted from the CNNs are efficient and discriminating in their capacity to describe DTIs and distinguish interactions based on their binding affinity values.

Additionally, the performance of the SVR model in terms of MSE (0.203), RMSE (0.450), CI (0.907), Spearman (0.714) and r^2 (0.759) is considerably high and overall superior to all state-of-the-art baselines, despite it being a traditional ML approach. These findings demonstrate that the input data's quality and discriminatory power greatly influence the performance, validating once more the efficiency of the deep representations extracted from the CNNs in the prediction process.

Overall, the use of an end-to-end DL architecture to predict binding affinity demonstrates not only the ability of DL to automatically identify and extract discriminating features from drug and protein data collection, but also the capacity to learn complex and hidden patterns related to DTIs for the prediction of binding affinity.

Figure 6.4 illustrates the predictions from the proposed model against the actual (true) binding affinity values for the Davis independent testing set, where it is possible to observe a significant density around the *predicted = true value* reference line (perfect model).

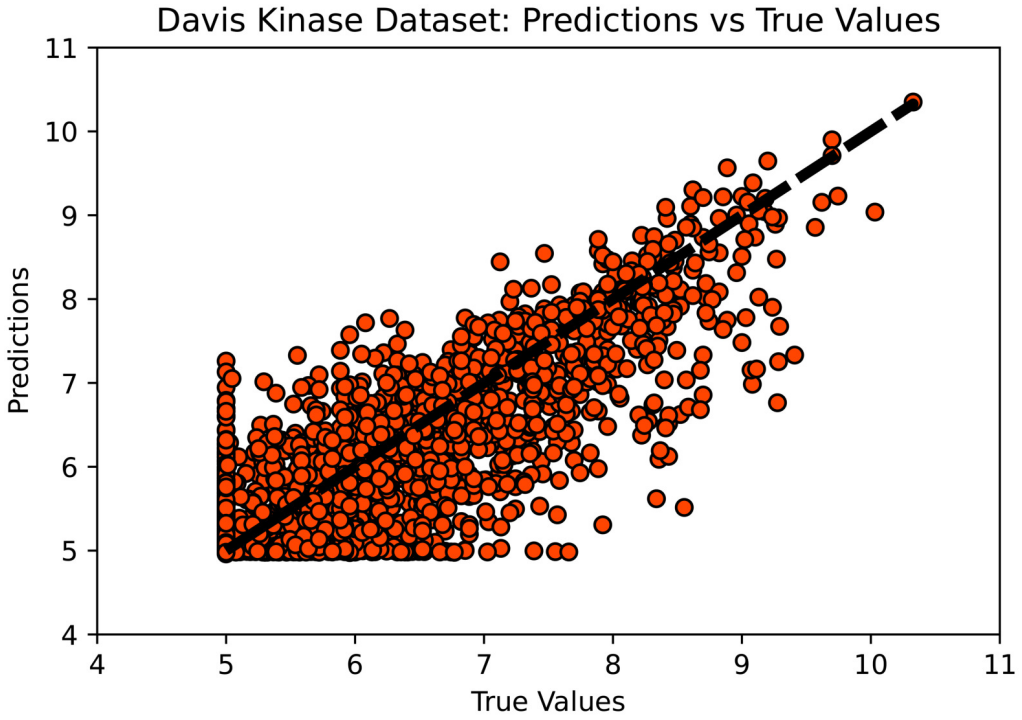


Figure 6.4: CNN-FCNN model predictions against the true values for the Davis kinase binding affinity testing set, where the diagonal line is the reference line (*predicted = true value*).

6.3.2 Reliability of the CNNs in the identification of important regions for binding

Despite the prediction efficiency achieved, it is not possible to directly extract explanations for the decision-making process solely based on the deep representations, considering that they are not interpretable by humans. In this study, Grad-RAM was proposed to obtain regression discriminative localization maps, which provide information related to the regions of the input that had a positive influence on the prediction. In order to evaluate the reliability of the CNNs in the identification of important regions for binding, the correlation between the input regions that had a positive influence on the prediction and the window-based pockets related to binding sites and motifs was explored. Table 6.5 and 6.6 report the $L_{Grad-RAM}$ matching (see Section B.2.2.1 in Appendix B for more details) results for the binding sites of the Davis ∩ sc-PDB and sc-PDB pairs, respectively, across different window lengths and for the different formulations of the $L_{Grad-RAM}$.

Regarding the differences in the formulation of $L_{Grad-RAM}$, specifically between employing a global max pooling (GMP) instead of a global average pooling (GAP), and between using guided gradients (G) instead of non-guided gradients (NG), the

Table 6.5: Davis ∩ sc-PDB Binding Sites - $L_{Grad-RAM}$ matching (Equation B.4) results across different window lengths and for the different formulations of the $L_{Grad-RAM}$. Lower and higher percentage values are associated with lower and higher numbers of window-based binding pockets, in which information is extracted from at least one position, across all the DTI pairs.

Window Length	GMP-G	GMP-NG	GAP-G	GAP-NG
0	20.74	20.74	20.74	19.57
1	46.32	46.32	46.32	42.83
2	53.29	53.29	53.29	50.39
3	56.98	56.98	56.98	54.07
4	60.66	60.66	60.66	57.95
5	61.24	61.24	61.24	58.72

GMP: Global Max Pooling, GAP - Global AVG Pooling, G - Guided Gradients, NG - Non Guided Gradients

Table 6.6: sc-PDB Binding Sites - $L_{Grad-RAM}$ matching (Equation B.4) results across different window lengths and for the different formulations of the $L_{Grad-RAM}$. Lower and higher percentage values are associated with lower and higher numbers of window-based binding pockets, in which information is extracted from at least one position, across all the DTI pairs.

Window Length	GMP-G	GMP-NG	GAP-G	GAP-NG
0	16.51	16.51	16.51	15.08
1	39.14	39.14	39.14	36.93
2	49.37	49.37	49.37	46.89
3	56.92	56.92	56.92	53.99
4	63.33	63.33	63.33	60.53
5	66.85	66.85	66.85	64.44

GMP: Global Max Pooling, GAP - Global AVG Pooling, G - Guided Gradients, NG - Non Guided Gradients

results demonstrate that there was no significant difference, except for GAP-NG, which generated worse localization maps. Considering that regions with the highest positive influence are of special interest in the context of the problem, GMP-G was determined to be the most consistent combination and, thus, used for all evaluations and comparisons.

The Binding sites - $L_{Grad-RAM}$ matching results demonstrate that the CNNs are identifying and extracting features from the window-based binding pockets without any *a priori* information, considering that there is relevant information being detected at every window length. Furthermore, the highest $L_{Grad-RAM}$ matching increase occurs between a window length 0 and 1, and between a window length 1

and 2 (20.74 - 46.32 - 53.29% and 16.51 - 39.14 - 49.37% for the Davis ∩ sc-PDB and sc-PDB pairs, respectively), showing that the CNNs are extracting information essentially within the closer regions to the exact binding site location, in which with a window length of 2, the DTI pairs have in average around 50% or more of their window-based binding sites identified. Nevertheless, the $L_{Grad-RAM}$ matching values in the Davis ∩ sc-PDB pairs are essentially higher for the lower window lengths when compared to the sc-PDB pairs, which is in agreement with the fact that sc-PDB pairs are not associated only with kinases (representability).

Regarding the motifs, the $L_{Grad-RAM}$ matching was evaluated across different PSSM thresholds, window lengths, and data collections, where subsets of these datasets, specifically related to the filtering process of the motifs inside the entire binding region, were also considered. Figure 6.5 illustrates the $L_{Grad-RAM}$ matching in terms of a heatmap for the PSSM motifs across different thresholds and window lengths for the Davis ∩ sc-PDB, Davis ∩ sc-PDB with the motifs inside the binding region filtered out, sc-PDB, and sc-PDB with the motifs inside the binding region filtered out pairs, respectively (see Tables B.5, B.6, B.7, and B.8 in Section B.3.2.1 of Appendix B for more details regarding the results).

The motifs - $L_{Grad-RAM}$ matching results demonstrate that the CNNs are identifying and extracting features from window-based motifs across different thresholds and window lengths. Similar to the binding sites, the highest $L_{Grad-RAM}$ matching increase occurs between a window length 0 and 1, and between a window length 1 and 2 (e.g., 11.28 - 20.26 - 26.52% for the PSSM threshold ≥ 5, and 13.3 - 28.25 - 47.65% for the PSSM threshold ≥ 10 for the Davis ∩ sc-PDB pairs). The sc-PDB pairs (Figures 6.5c and 6.5d) present higher $L_{Grad-RAM}$ matching values, demonstrating that the CNNs are especially focusing on the conserved motifs positions, which reflects the absence of the protein domain similarity. Furthermore, higher PSSM thresholds (≥ 8) are associated with higher $L_{Grad-RAM}$ matching values across the different window lengths, suggesting that the CNNs are focusing on the highly conserved motifs, which are usually associated with important protein functions. Nevertheless, the filtering process of the motifs inside the entire binding region (Figures 6.5b and 6.5d) resulted in overall lower $L_{Grad-RAM}$ matching values, showing that the CNNs are identifying and extracting features simultaneously from binding sites and motifs.

Figure 6.6 illustrates the $L_{Grad-RAM}$ maps for some of the protein sequences associated with the Davis ∩ sc-PDB and sc-PDB DTI pairs, in which the binding sites are annotated, i.e., known and available.

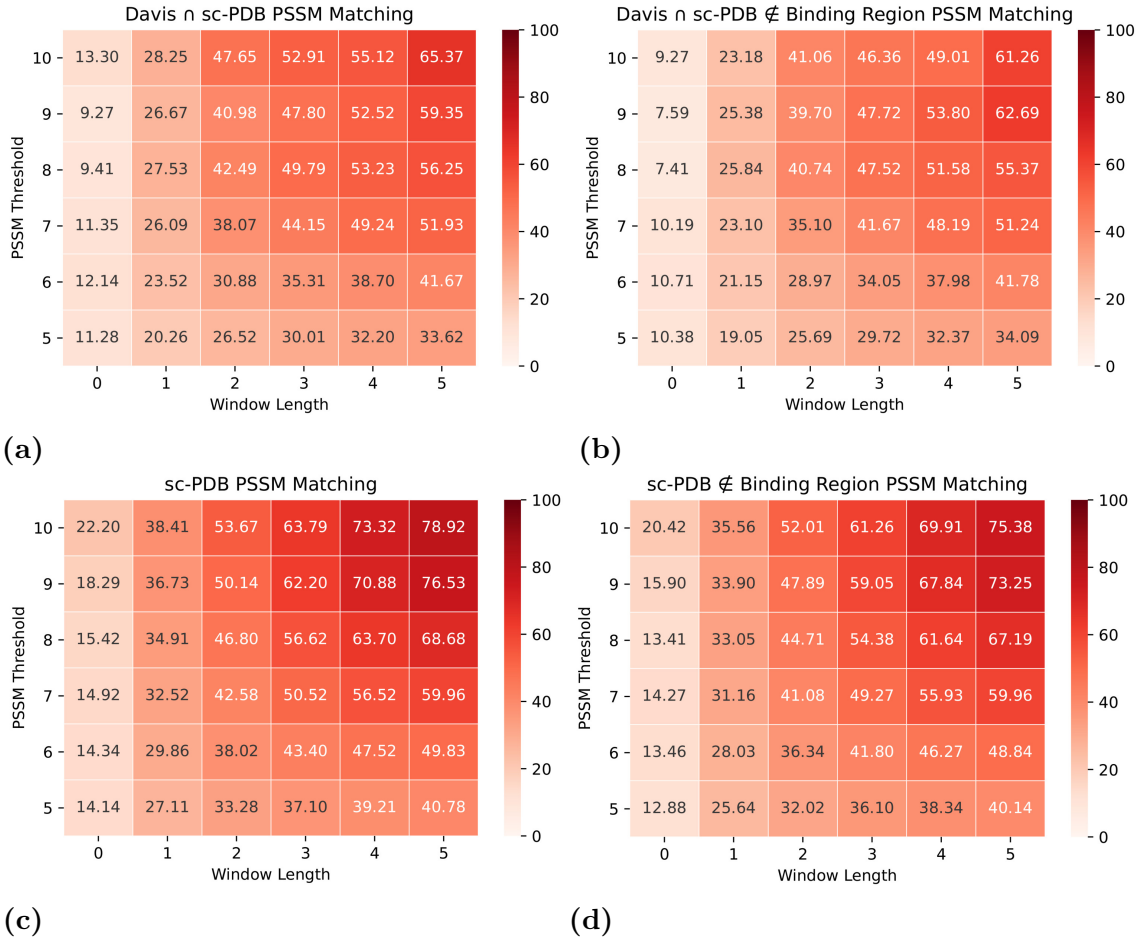


Figure 6.5: PSSM Motifs - $L_{Grad-RAM}$ matching results (Equation B.4) across different window lengths and PSSM thresholds, where weaker and deeper red colors are associated with lower and higher $L_{Grad-RAM}$ matching values, respectively. a) Davis ∩ sc-PDB pairs; b) Davis ∩ sc-PDB pairs (filtered*); c) sc-PDB pairs; d) sc-PDB pairs (filtered*). *Motifs inside the binding region filtered out.

6.3.2.1 3D Interaction Space Analysis (Docking)

In order to further validate the reliability of the CNNs in the identification of important regions for binding, and the previous Binding sites - $L_{Grad-RAM}$ matching results, it was critical to explore the 3D interaction for DTI pairs without any binding information available, i.e., where the interacting protein residues are not annotated or available (contrarily to the pairs represented in Figure 6.6 and the ones used for the Binding sites - $L_{Grad-RAM}$ matching results). On that account, two DTI pairs from the Davis kinase binding affinity testing set, specifically ABL1(E255K)-phosphorylated - SKI-606 and DDR1 - Foretinib, were selected and their 3D interaction space was explored using docking approaches, wherein the resulting 3D complexes were thoroughly assessed in order to make a fair comparison with the $L_{Grad-RAM}$ hits. Figures 6.7 and 6.8 depict the 3D receptor-ligand com-

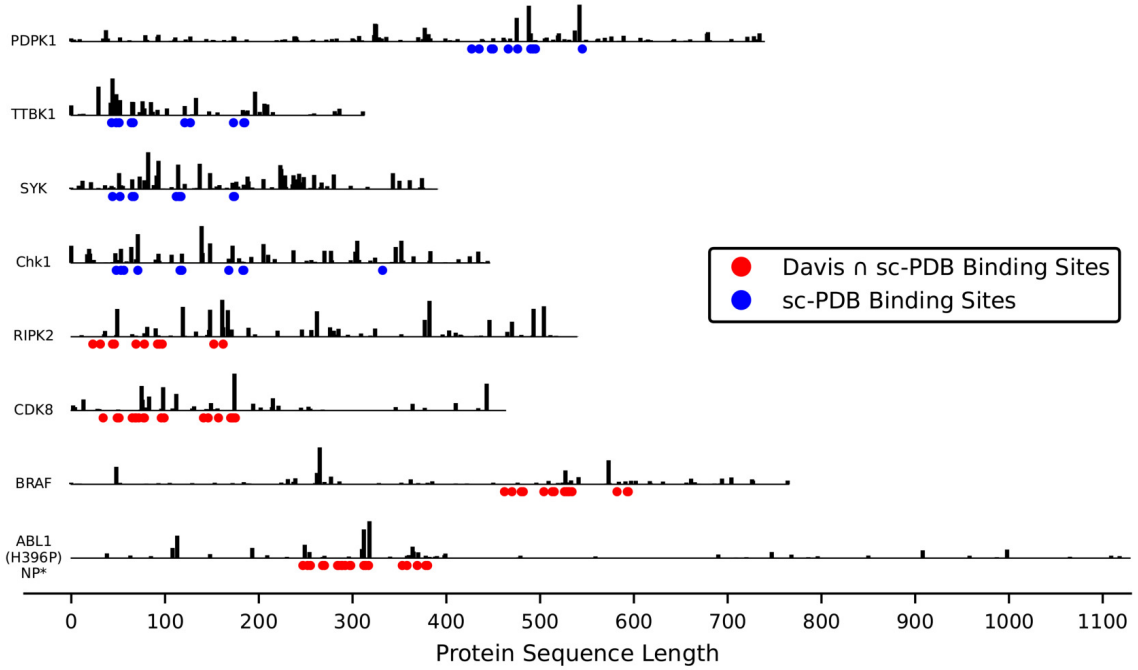
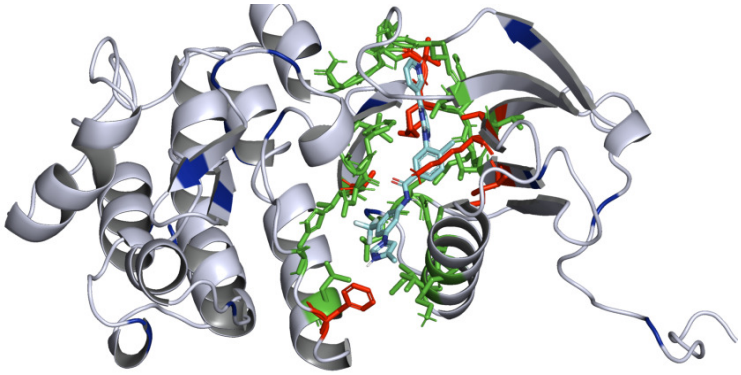


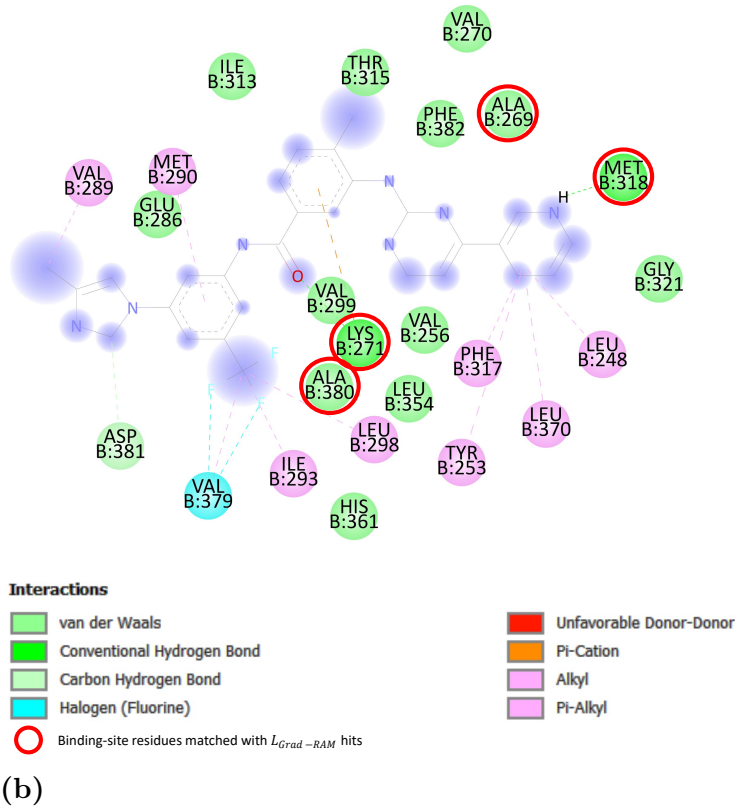
Figure 6.6: $L_{Grad-RAM}$ maps for some of the protein sequences of the Davis ∩ sc-PDB pairs and sc-PDB pairs, where the binding sites are represented by the red and blue circles, respectively. The height of the vertical lines corresponds to the importance (weight) of the feature extracted from the corresponding position (amino acid).*NP: non-phosphorylated

plex, in which the potential binding sites (≤ 5 Å) and the information retrieved from the $L_{Grad-RAM}$ are annotated, and the 2D interaction diagram, where the matched binding - $L_{Grad-RAM}$ positions are annotated, for the ABL1(E255K)-phosphorylated receptor and DDR1 receptor, respectively.

Consistent with the previous findings related to the $L_{Grad-RAM}$ matching results, Figures 6.7a and 6.8a show that the CNNs are not aimlessly identifying regions to extract features from when predicting binding affinity, especially considering that there are $L_{Grad-RAM}$ hits matched with the potential binding sites (also represented in Figures 6.7b and 6.8b) and other hits near the neighborhood of these interaction spots. Regarding the $L_{Grad-RAM}$ hits close to the main binding pocket and also those not in the vicinity of the binding pocket, their spacial positions suggest they bear relation to conserved regions or other potential interaction pockets/subpockets, e.g., some of these hits are near α-helices, which are usually important for the structure and function of the protein, and for certain interactions given their polarity. In particular, for the case of the DDR1 kinase, some of these $L_{Grad-RAM}$ hits were found to be matched or nearly matched with certain experimental validated critical interacting residues.



(a)



(b)

Figure 6.7: SKI-606 in complex with ABL1(E255K)-phosphorylated. a) Annotated 3D complex obtained from docking, where the potential binding sites ($\leq$ 5 Å), the $L_{Grad-RAM}$ hits, and the matched binding - $L_{Grad-RAM}$ positions are represented by the green, blue and red colors, respectively. b) 2D Interaction Diagram, in which the matched binding - $L_{Grad-RAM}$ hits are shown delimited by red circles.

See Section B.3.2.2 in Appendix B for more details regarding the docking process and analysis of the resulting 3D complexes.

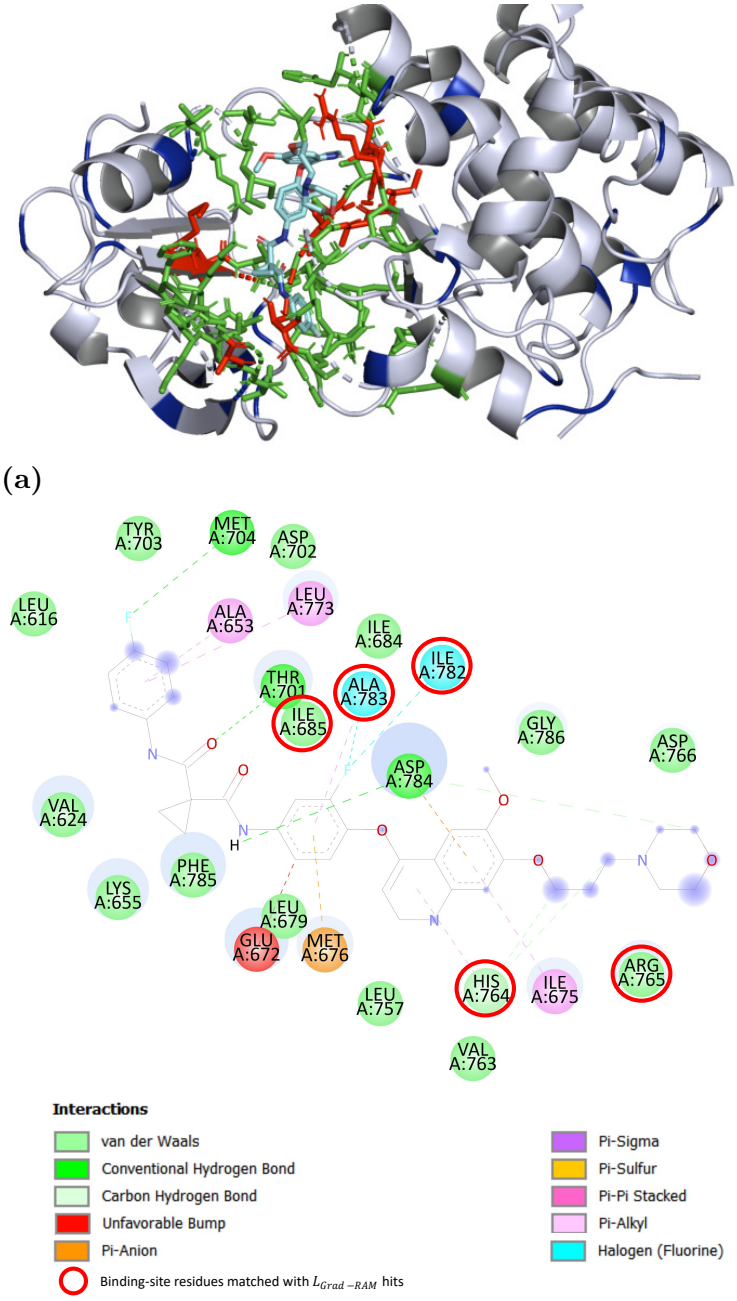
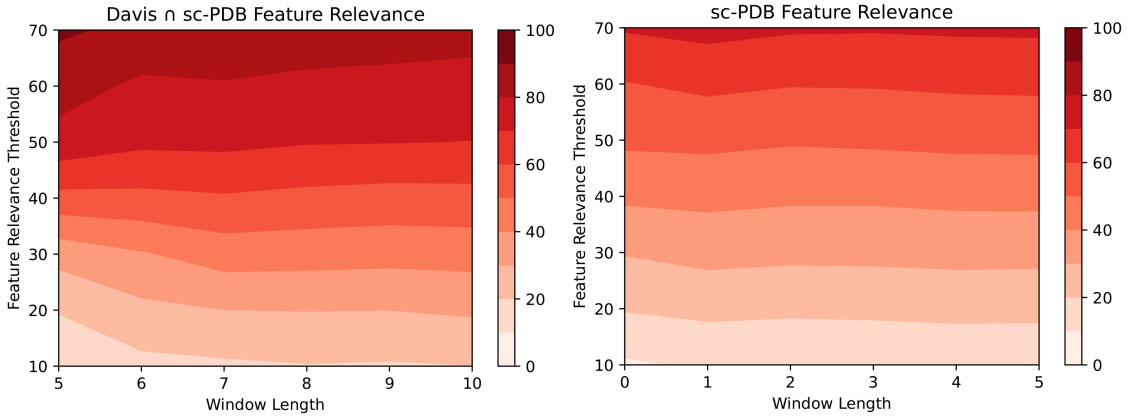


Figure 6.8: Foretinib in complex with DDR1. a) Annotated 3D complex obtained from docking, where the potential binding sites ($\leq$ 5 Å), the $L_{Grad-RAM}$ hits, and the matched binding - $L_{Grad-RAM}$ positions are represented by the green, blue and red colors, respectively. b) 2D Interaction Diagram, in which the matched binding - $L_{Grad-RAM}$ hits are shown delimited by red circles.

6.3.3 Robustness of the deep representations

Apart from validating the reliability of the CNNs in the identification of important regions for binding, it is critical to understand the robustness (significance) of the

deep representations. On that account, the feature relevance correlation between the positive-valued features in the input domain and the ones extracted from the window-based binding sites and motifs was also evaluated and explored. Figure 6.9 illustrates the $L_{Grad-RAM}$ feature relevance (see Section B.2.2.2 in Appendix B for more details) in terms of a density map for the binding sites across the different feature relevance thresholds and window length values for the Davis ∩ sc-PDB and sc-PDB pairs (see Tables B.10 and B.11 in Section B.3.3.1 of Appendix B for more details regarding the results).



(a) **(b)**

Figure 6.9: Binding sites - $L_{Grad-RAM}$ feature relevance (Equation B.5) results across different feature relevance thresholds and window lengths, where weaker and deeper red colors are associated with lower and higher $L_{Grad-RAM}$ feature relevance values, respectively. a) Davis ∩ sc-PDB pairs; b) sc-PDB pairs

The results demonstrate that at every feature importance threshold and window length value, the Binding sites - $L_{Grad-RAM}$ feature relevance values are superior to the corresponding threshold, i.e., the positive-valued features extracted from the window-based binding pockets are in the range of those with the highest influence. In particular, Figure 6.9a shows that at every feature significance threshold, the $L_{Grad-RAM}$ feature relevance value is roughly 10% higher than the corresponding threshold. Regarding the window length, there is no significant difference across the different thresholds, corroborating the Binding sites - $L_{Grad-RAM}$ matching results, where CNNs were shown to extract information within the closer regions to the binding sites. Overall, CNNs are not aimlessly identifying and extracting features from each window-based binding pocket, but essentially assigning significance to these regions when predicting binding affinity.

Regarding the motifs, the $L_{Grad-RAM}$ feature relevance was evaluated across different PSSM thresholds, feature significance thresholds, window lengths, and data

collections, including the subsets related to the filtering process of the motifs inside the entire binding region. However, since the window length did not represent any significant difference in the results, the mean value across the different window lengths was considered for visualization. Figure 6.10 illustrates the $L_{Grad-RAM}$ feature relevance in terms of a heatmap for the motifs across different PSSM thresholds and feature significance thresholds for the Davis ∩ sc-PDB, Davis ∩ sc-PDB with the motifs inside the binding region filtered out, sc-PDB, and sc-PDB with the motifs inside the binding region filtered out pairs, respectively (see Tables B.12, B.13, B.14, and B.15 in Section B.3.3.2 of Appendix B for more details regarding the results).

The motifs - $L_{Grad-RAM}$ feature relevance results demonstrate that the CNNs are also assigning significance to the conserved motifs, although inferior to the one given to the window-based binding pockets, considering that the $L_{Grad-RAM}$ feature relevance is essentially lower in filtered pairs and even below the corresponding feature significance threshold values in some cases (illustrated when comparing Figures 6.10a and 6.10b, and Figures 6.10c and 6.10d). These results are in agreement with the fact that the binding sites and the overall binding region have a great impact on the interaction process, which is expressed in terms of the binding affinity (regression outcome). Additionally, across the different feature importance thresholds, higher $L_{Grad-RAM}$ feature relevance values are essentially associated with higher PSSM thresholds (≥ 8), except for Davis ∩ sc-PDB unfiltered (Figure 6.10a) and Davis ∩ sc-PDB filtered (Figure 6.10b), in which at the feature relevance threshold 60 % and 70 % it was not verified. Nonetheless, the results are consistent with the previous findings, where the conserved motifs associated with the higher PSSM thresholds were found to be associated with the highest $L_{Grad-RAM}$ matching values.

6.4 Conclusions

6.4.1 Final Remarks

This study explored an end-to-end DL architecture to predict DTA measured in pK_d, where CNNs were exploited to automatically identify and extract discriminating deep representations from protein sequences and SMILES strings. The deep representations were found to be efficient and discriminating in their capacity to describe DTIs and distinguish interactions based on their binding affinity values (interaction strength rank order). Furthermore, the CNN-FCNN model yielded better results compared to state-of-the-art baselines, demonstrating its viability for practical use.

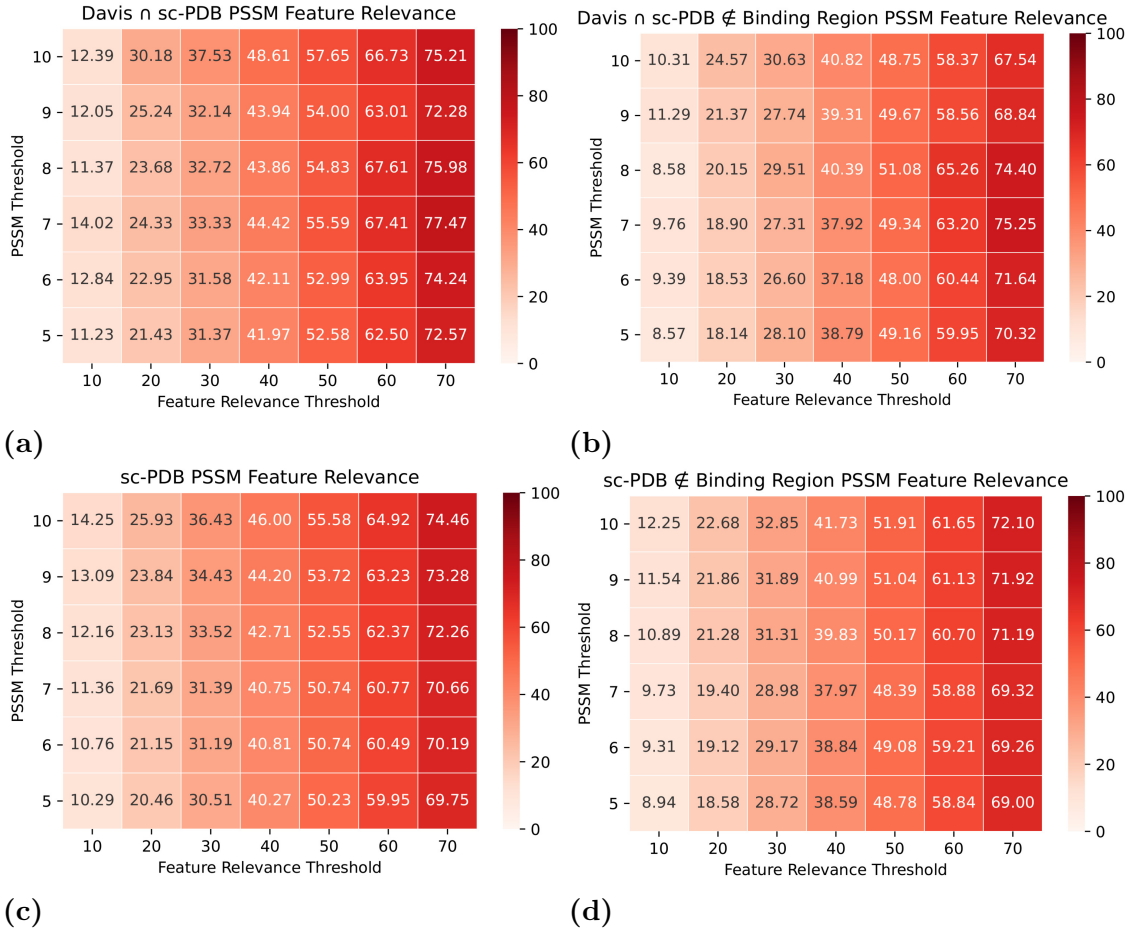


Figure 6.10: PSSM Motifs - $L_{Grad-RAM}$ feature relevance (Equation B.2.2.2) results* across different PSSM thresholds and feature significance thresholds, where weaker and deeper red colors are associated with lower and higher $L_{Grad-RAM}$ matching values, respectively. a) Davis ∩ sc-PDB pairs; b) Davis ∩ sc-PDB pairs (filtered**); c) sc-PDB pairs; d) sc-PDB pairs (filtered**). * Each value corresponds to the mean value across the different window lengths. **Motifs inside the binding region filtered out.

Additionally, this work provided explainability to the predictions by connecting the deep representations extracted from the CNNs to the input domain, exploring the reliability of CNNs in the identification of important sequential regions, specifically binding sites and evolutionarily conserved motifs, when predicting binding affinity, and evaluating the significance of the deep representations extracted from relevant sequential spots. On that account, the results demonstrated that the CNNs are identifying and extracting features simultaneously from window-based binding sites and motifs without any *a priori* information. Moreover, CNNs were found to extract information essentially within the closer regions to the exact binding or motif location, respectively, validating the effectiveness of these architectures in drug discovery. Additionally, the features extracted from window-based relevant regions for

binding were shown to be in the range of those with the highest positive influence, particularly in the case of the interaction sites.

Overall, the major contribution of this study relies on an efficient end-to-end DL architecture to predict binding affinity beyond the confined space of proteins and ligands with determined 3D structures, in which explanations for the predictions are presented and explored.

6.4.2 Study Limitations and Future Work

In spite of the discriminating power and robustness of the deep representations extracted from the protein sequences and SMILES strings for the prediction of binding affinity, the CNNs do not model the intra-associations amongst the units of proteins or compounds, i.e., the inter-dependency of the proteomics and chemical spaces. Thus, the predictions are based solely on local and independent scattered motifs, in which the global context is not taken into consideration during the feature-extracting process. Furthermore, the proposed CNN-FCNN model only considers the magnitude of certain local regions of each binding component for the prediction of binding affinity, neglecting the inter-associations that revolve around the binding substructures (context of interaction), and the contributions of the interacting substructures involved. On that account, the multi-domain inter-dependency associated with the proteomics, chemical, and pharmacological spaces is not captured in the learning process of the proposed model, compromising the robustness of the predictions and the validity of the DTI representation space for the estimation of binding affinity.

Post-hoc explainability approaches have demonstrated their efficiency and potency as tools for generating potential explanations concerning the predictions and inference processes of black box models, specifically complex DL architectures. In the context of this study, the implemented and proposed Grad-RAM method played a crucial role in identifying and visualizing the input regions that positively influenced the prediction of binding affinity. In particular, interacting residues were found to be associated with the features extracted from the CNNs. However, conducting external evaluations may result in explanations that stem from artifacts learned by the model rather than actual knowledge derived from the data. Consequently, the effectiveness and usefulness of these explanations are limited, particularly given the inherent bias associated with the traceback process of the Grad-RAM approach. On that account, it is critical to consider interpretability and explainability during the model construction in order to improve the performance and reliability of the predictions.

Even though it was shown that CNNs are identifying and extracting features from window-based binding sites, especially within closer regions to the exact binding locations, without any *a priori* information, the reliability of the predictions is limited considering that information about binding sites/pockets is not actively integrated into the learning process. Therefore, given the range of different regions across the whole structure of the proteins and compounds, the relevance given to certain spots introduces bias in the predictions, resulting in the estimation of potential DTIs based on certain redundant sites. Moreover, in order to realistically model DTIs and understand the interaction process, it is crucial to learn the inter-associations that revolve around the binding substructures based on information related to binding sites.

DL-based architectures perform significantly better when the dataset becomes larger. Thus, collecting and building a larger and valid DTI dataset measured in terms of the K_d constant, which is one of the few unbiased binding affinity/activity metrics, may lead to superior prediction performance. However, the choice of using the Davis dataset (exclusively) to establish the grounds/basis of the proposed approach relied upon the fact that it is the benchmark dataset that contains only DTIs measured in terms of the dissociation constant and, thus, representing a source of more direct/reliable binding affinity metric, offering less noise/error, and better serving the purpose of centering the focus of the work on the proposed methodology. Furthermore, the number of DTIs with binding affinity measured in K_d is reduced compared to other bioactivities, hence, exploring bioactivity-related domain adaptation methods might also lead to interesting findings.

Additionally, considering the polypharmacological nature associated with most active small compounds, in which these drugs interfere with different disease pathways, it is relevant to extend this work to validate the identification of important components/substructures within the compound space, especially to reduce potential off-target effects and toxicity. Moreover, identifying relevant substructures in hit-to-lead compounds greatly reduces the search space around structural congeners, homologs, or analogs during the lead optimization step of the drug discovery pipelines.

Chapter 7

Intrinsic Explainability and Drug–Target Multi-Domain Inter-Dependency

This chapter concerns the use of multiple attention mechanisms to model the multi-domain inter-dependency associated with the proteomics, chemical, and pharmacological spaces for the prediction of binding affinity. This study also focuses on incorporating explainability during model construction in order to provide different levels of potential DTI and prediction understanding.

The content of this chapter is based on a journal article published in *Computers in Biology and Medicine* [350]. Section 7.1 presents the study context. Section 7.2 details the materials and methods used in this study. Section 7.3 reports the results and discusses the obtained findings. Section 7.4 provides some final reflections and the limitations of this study.

7.1 Study Context

The therapeutic effects of active compounds are determined through the observation of DTIs, where the role enforced by the drug (pharmacological activity) regulates the target's biological process. Therefore, identifying new molecules with relevant binding activity against targets with biological interest is crucial in the early drug discovery stages, considering that the ability of a drug to bind plays an important role in the execution of its intrinsic activity [1, 103].

In recent years, *in silico* DTI prediction has attracted increasing attention and holds broad interest to address several challenges, including target fishing, drug repositioning, and polypharmacology studies. These computational methods through the scanning of large amounts of pharmacogenomic data in shorter periods of time and

leveraging of the knowledge available to characterize the proteins and/or compounds have been determinant in the discovery of new drugs, new findings for existing drugs, and improving the overall understanding of the biological, chemical and pharmacological processes involved in the DTIs [4]. In spite of the encouraging results and performances obtained by numerous computational studies proposed to solve the DTI prediction challenge, most of these methodologies rely either on shallow binary associations or biased bioactivities to characterize the interaction and conduct the experiments [14].

The interaction between compounds and proteins results from the recognition and complementarity of certain groups (binding regions) and it is supported by the joint action of other individual substructures scattered across the protein and compound [24, 108]. However, most DTI prediction models simplify the interaction mechanism and do not take simultaneously into consideration the magnitude of certain local regions of each binding component and the interacting substructures involved [36]. Thus, the multi-domain inter-dependency is usually not captured or learned in these approaches, limiting the inferring process and the validity of the results.

Given the advances in computational power, most recent studies dealing with DTI or DTA prediction explore DL strategies, achieving better results than traditional ML solutions [27]. Despite the increased modular ability of these architectures to learn sequential and/or structural motifs and extract robust representations, the final predictions are mostly not interpretable by humans, which affects the understanding of the underlying aspects around the inner decisions. Hence, it is crucial to consider explainability during model construction [28, 31].

This study explores a novel end-to-end Transformer-based architecture for predicting DTA measured in terms of the dissociation constant (K_d), where 1D sequential and structural data, specifically protein sequences and SMILES strings, are used to represent the targets and compounds, respectively. The proposed architecture employs three Transformer-Encoder blocks, particularly a protein encoder, a compound encoder, and a protein-compound encoder, and concatenates the resulting aggregate representations to feed into an FCNN. Moreover, this architecture, Drug–Target Interaction TRansformer (DTITR), leverages the use of self-attention layers to learn the short and long-term proteomics and chemical context dependencies between the sequential and structural units of the proteins and compounds, respectively, and cross-attention layers to exchange information and make the interaction between the proteomic and chemical domains (pharmacological space). Additionally, the proposed model's emphasis is not only on the predictive performance, but

also on the self-capability of the architecture to provide different levels of potential DTI and prediction understanding due to the nature of the attention blocks, which give information about the overall importance of the input components and their associations to the model.

7.2 Material and Methods

7.2.1 Binding Affinity Dataset

The proposed model was evaluated using the Davis et al. (2011) [276] research study dataset, which contains a total of 31 824 interactions between 72 kinase inhibitors (compounds) and 442 kinases (proteins). This dataset covers a large percentage of the human catalytic protein kinome, and the binding strength of the DTI pairs is measured in terms of a quantitative dissociation constant (K_d), which expresses a direct measurement (unbiased) of the equilibrium between the receptor-ligand complex and dissociation components, in which lower values are associated with strong interactions.

The protein sequences of the Davis dataset were extracted from the UniProt [57] database based on the corresponding accession numbers (identifiers). In order to avoid increased noise due to excessive padding or loss of relevant sequential information potentially related to binding regions, only proteins with a length between 264 and 1400 residues were selected, corresponding to 95.7% of the information present in the dataset.

Davis compound SMILES strings were collected in their canonical notation from the PubChem [343] database based on their compound identifiers (CIDs). Even though the canonical notation is unique, where the atoms are consistently numbered, there are some differences in the representation across different data sources. On that account, the canonical transformation from the RDKit [344] package was applied in order to guarantee a consistent notation to represent the chemical structure of the compounds and increase the overall reproducibility. Similar to the protein sequences, only SMILES strings with a length between 38 and 72 chemical characters were selected, corresponding to 95.8% of the information present in the dataset.

The Davis binding strength distribution ranges from low values (strong interactions) to high values (weak interactions), in which the majority of the DTI pairs are characterized by a binding affinity equal to 10 000 nM. Hence, in order to reduce the effects of the high variance of this distribution on the learning loss, a normalization

strategy (Equation 7.1) was employed to the Kd values, transforming them into the logarithmic space (pK_d). The distribution of the pK_d values ranges from 5 (10 000 nM) to approximately 11.

$$pK_d = -log_{10}(\frac{K_d}{10^9}) \tag{7.1}$$

Table 7.1 summarizes the statistics of the original and pre-processed Davis Dataset.

Table 7.1: Original and pre-processed Davis dataset: unique proteins, compounds, and DTIs.

Davis Kinase Dataset					
	Proteins	**Compounds**	**DTI**	**pKd = 5**	**pKd > 5**
Original	442	72	31 824	22 400	9424
Pre-Processed	423	69	29 187	20 479	8708

7.2.2 Input Representation

In order to represent the structural characters of the SMILES strings, an integer-based encoding was applied, in which the different SMILES in the Davis dataset were scanned and 26 categories (unique characters) were extracted. This 26-character dictionary was used to encode each character into the corresponding integer. SMILES strings shorter than the maximum length threshold of 72 characters were padded. Figure 7.1 illustrates the integer-based encoding applied to the SMILES string associated with the Dasatinib compound.

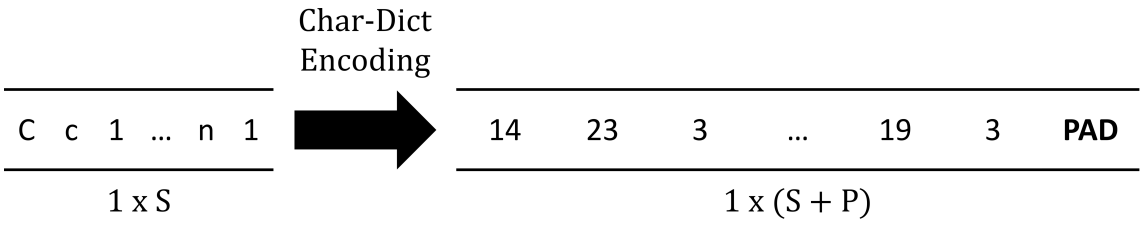


Figure 7.1: Integer-based encoding applied to the Dasatinib SMILES string, where each character is encoded into the corresponding integer. S is the length of the SMILES string and P is the number of padding tokens (zeros).

In the case of protein sequences, it was not reasonable to apply the same encoding method of the SMILES strings given the computational complexity of the self-attention layers of $O(n^2)$ with respect to the sequence length. On that account, the approach proposed in the research study by Huang et al. (2021) [244] was employed, which combines a Frequent Consecutive Subsequence (FCS) mining method with the Byte Pair Encoding (BPE) algorithm. The FCS method examines large amounts

of unlabeled data to discover frequent substructures and create a set of recurring subsequences (subwords). On the other hand, BPE decomposes the sequence into an order of discovered frequent subsequences, where each subsequence must be exclusive and must not overlap, and the aggregation of all subsequences must recover the original sequence. The hierarchy set of frequent subsequences contains a total of 16 693 different subwords, which results in a maximum length of 556 subwords for the protein sequences present in the Davis dataset. Similar to the SMILES strings, protein sequences shorter than this maximum length were padded. Figure 7.2 depicts the FCS and BPE encoding approach applied to the AAK1 kinase.

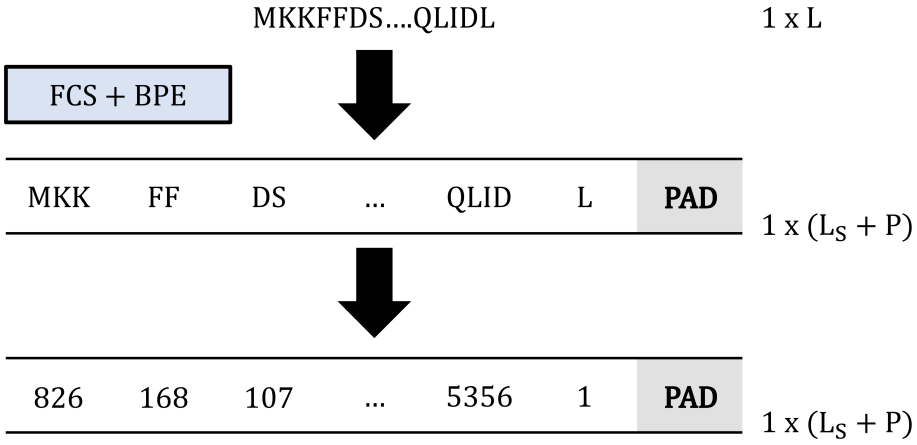


Figure 7.2: FCS and BPE encoding applied to the AAK1 kinase amino acid sequence, where the sequence is decomposed into an order of discovered frequent subsequences followed by integer encoding. L is the length of the amino acid sequence, L_S is the length of the sequence decomposed into subsequences, and P is the number of padding tokens (zeros).

See Figure C.1 in Section C.1 of Appendix C for more details regarding the distribution of the Davis dataset based on the FCE and BPE encoding approach.

7.2.3 DTITR Framework

The DTITR framework learns to predict the binding strength of DTIs, where 1D sequential and structural information, protein sequences and SMILES strings, respectively, are used as input. This architecture makes use of two parallel Transformer-Encoders to compute a contextual embedding of the protein sequences and SMILES strings. The outputs are then fed into a Cross-Attention Transformer-Encoder block, which comprises cross-attention and self-attention layers, to exchange information and model the interaction space. The resulting aggregate representations, which correspond to the final hidden states of the start tokens added to the protein sequences and SMILES strings, are concatenated and used as input for an FCNN.

The final layer, which is composed of a single neuron, outputs the binding affinity measured in terms of pK_d.

7.2.3.1 Embedding Block

The protein sequences and SMILES strings are initially processed based on their length (Section 7.2.1) and then encoded according to the approaches mentioned in Section 7.2.2. Similar to the BERT architecture [173], special tokens of regression R_P and R_S have been added to the beginning of every protein sequence and SMILES string, respectively. An embedding layer was assigned to the protein sequences and SMILES strings, generating a learned embedding to every token with a fixed size of d^P_{model} and d^S_{model}, respectively, via a learnable dictionary matrix. Following the embedding layers, the embedding values were multiplied by $\sqrt{d^P_{model}}$ and $\sqrt{d^S_{model}}$ to initially rescale their value.

Considering that the Transformer-Encoder is permutation invariant, it is necessary to include additional information about the relative or absolute position of the tokens in the sequence. On that account, the same approach used in the study by Vaswani et al. (2017) [172] was applied in order to assign a positional encoding for each token of the input sequences. This method is based on sine and cosine functions of different frequencies and outputs a unique encoding for each position (see Section C.2.1 in Appendix C for more details). The final embeddings for the ith and jth input tokens of the protein sequence ($E^{P_k}_i$) and SMILES string ($E^{S_k}_j$), respectively, associated with the kth DTI pair are given by the sum of the token embedding and the positional embedding:

$$\begin{aligned} E^{P_k}_i &= E^{P_k}_{token_i} + E^{P_k}_{pos_i} \\ E^{S_k}_j &= E^{S_k}_{token_j} + E^{S_k}_{pos_j} \end{aligned} \tag{7.2}$$

, where $E^{P_k}_{token_i} \in R^{d^P_{model}}$ and $E^{S_k}_{token_j} \in R^{d^S_{model}}$, and $E^{P_k}_{pos_i} \in R^{d^P_{model}}$ and $E^{S_k}_{pos_j} \in R^{d^S_{model}}$ are the token embeddings and the positional embeddings for the ith and jth inputs tokens of the protein sequence P_k and SMILES string S_k, respectively.

Following the sum of the two types of embedding, a dropout layer was added.

7.2.3.2 Transformer-Encoder

In order to capture the proteomics and chemical context information present in the protein sequences and SMILES strings, respectively, two Transformer-Encoders in parallel were explored. The Transformer-Encoder architecture is composed of a stack of identical blocks, where each block contains a MHSA with an PWFFN. Residual

connections are applied after every block followed by LN, and dropout is applied after each MHSA layer and after each Dense layer of the PWFFN. Considering B^1 the output of the first subunit and B^2 the output of the second subunit, the output of the kth block can be expressed as:

$$\begin{aligned} B_k^1 &= \mathbf{LN}(B_{k-1}^2 + \mathbf{dropout}(\mathbf{MHSA}(B_{k-1}^2))) \\ B_k^2 &= \mathbf{LN}(B_k^1 + \mathbf{PWFFN}(B_k^1)) \end{aligned} \tag{7.3}$$

, where $B_k^1, B_k^2 \in R^{N_P \; x \; d_{model}^P}$ in the case of the protein sequences (N_P is the number of protein subwords), and $B_k^1, B_k^2 \in R^{N_S \; x \; d_{model}^S}$ in the case of the SMILES strings (N_S is the number of SMILES characters).

Overall, these two stacked Transformer-Encoders in parallel compute a contextual embedding for the protein sequences and SMILES strings, in which the self-attention mechanisms condition the weight given to input elements by learning the short and long-term context dependencies between the individual units.

7.2.3.3 Cross-Attention Transformer-Encoder

Apart from attending individually and learning the context dependencies between the individual units of each element of the DTI pair (Section 7.2.3.2), it is crucial for the compounds and proteins to attend mutually to each other, i.e., to exchange information, especially when considering that DTIs are primarily substructural, where the complementarity of certain regions is key for the binding process. Hence, a Cross-Attention Transformer-Encoder block was proposed to learn the pharmacological context information associated with the interaction space. The Cross-Attention Transformer-Encoder architecture is composed of a stack of two parallel identical blocks, where each block contains a Multi-Head Cross-Attention (MHCA), an MHSA, and an PWFFN. Similar to the Transformer-Encoder, residual connections are applied after every block followed by LN, and dropout is applied after each MHCA and MHSA layers and each Dense layer of the PWFFN.

The two MHCA layers are responsible for the exchange of information between the proteins and compounds, and to model the substructural space of the interaction. Instead of employing a full attention approach, i.e., the whole protein and compound attending to the whole compound and protein, respectively, which is computationally expensive and complex, and also redundant since the two attention matrices would have to satisfy the condition $W_{P-S} = W_{S-P}^T$, the R_P and R_S tokens are used for the exchange of context information [351]. These tokens previously learn

(Section 7.2.3.2) the overall proteomics and chemical context information amongst the individual units of the protein sequence and SMILES string, respectively, and therefore are considered an aggregate representation. On that account, these can be efficiently used as the attending agents (Query) in a Multi-Head Attention Layer, where each one of these tokens attends to the information present in the corresponding interaction component, i.e., the R_P token attends to the tokens of the SMILES string and the R_S token attends to the tokens of the protein sequence. Hence, these tokens interact and learn the context information present in the corresponding binding component, which further enriches their representation. The MHCA layers work similarly to the MHSA layer (Equation 5.10), but instead of the input attending to itself, i.e., the Query, Key, and Value being generated from the same input sequence, the Query will correspond to R_P or R_S token, and the Key and Value to the concatenation of the R_P or R_S token with the corresponding interaction component tokens. Considering X_P and X_S the representation of the protein sequence and SMILES string, respectively, the outputs for the two MHCA subunits associated with the kth Cross-Attention Transformer-Encoder block (X_P^k and X_S^k) can be expressed as:

$$\begin{aligned}
X_P^{k-1} = [R_P^{k-1} \,\|\, T_P^{k-1}]\,,\, X_S^{k-1} = [R_S^{k-1} \,\|\, T_S^{k-1}] \\
\mathbf{Q_P} = R_P^{k-1}\,,\, \mathbf{Q_S} = R_S^{k-1} \\
\mathbf{K_P/V_P} = [R_P^{k-1} \,\|\, T_S^{k-1}]\,,\, \mathbf{K_S/V_S} = [R_S^{k-1} \,\|\, T_P^{k-1}] \\
R_P^k = \mathbf{LN}(R_P^{k-1} + \mathbf{dropout}(\mathbf{MHCA}(Q_P,K_P,V_P))) \\
R_S^k = \mathbf{LN}(R_S^{k-1} + \mathbf{dropout}(\mathbf{MHCA}(Q_S,K_S,V_S))) \\
X_P^K = [R_P^k \,\|\, T_P^{k-1}]\,,\, X_S^K = [R_S^K \,\|\, T_S^{k-1}]
\end{aligned} \tag{7.4}$$

, where X_P^{k-1}, $X_P^k \in R^{N_P \; x \; d_{model}^P}$; R_P^{k-1}, $R_P^k \in R^{d_{model}^P}$; T_P^{k-1}, $T_P^k \in R^{(N_P-1) \; x \; d_{model}^P}$; X_S^{k-1}, $X_S^k \in R^{N_S \; x \; d_{model}^S}$; R_S^{k-1}, $R_S^k \in R^{d_{model}^S}$; and T_S^{k-1}, $T_S^k \in R^{(N_S-1) \; x \; d_{model}^S}$.

Following each one of these MHCA layers, an MHSA layer is applied in order to improve the internal connections between the individual units and enhance the representation of each token based on the learnt cross-attention context information. Similar to the Transformer-Encoder, an PWFFN is added and applied to the output of each MHSA layer. Considering B^1 the output of the first subunit, B^2 the output of the second subunit, and B^3 the output of the third subunit, the outputs of the kth Cross-Attention Transformer-Encoder block can be expressed as:

$$
\begin{aligned}
B^1_{k_P-1} &\xrightarrow{\textbf{Eq.7.4}} B^1_{k_P}, B^1_{k_S-1} \xrightarrow{\textbf{Eq.7.4}} B^1_{k_S} \\
B^2_{k_P} &= \mathbf{LN}(B^1_{k_P} + \mathbf{dropout}(\mathbf{MHSA}(B^1_{k_P}))) \\
B^2_{k_S} &= \mathbf{LN}(B^1_{k_S} + \mathbf{dropout}(\mathbf{MHSA}(B^1_{k_S}))) \\
B^3_{k_P} &= \mathbf{LN}(B^2_{k_P} + \mathbf{PWFFN}(B^2_{k_P})) \\
B^3_{k_S} &= \mathbf{LN}(B^2_{k_S} + \mathbf{PWFFN}(B^2_{k_S}))
\end{aligned}
\tag{7.5}
$$

, where $B^1_{k_P-1}$, $B^1_{k_P}$, $B^2_{k_P}$, $B^3_{k_P} \in R^{N_P \; x \; d^P_{model}}$, and $B^1_{k_S-1}$, $B^1_{k_S}$, $B^2_{k_S}$, $B^3_{k_S} \in R^{N_S \; x \; d^S_{model}}$.

7.2.3.4 Fully-Connected Feed-Forward

The final hidden states of the aggregate representations, which correspond to the start R_P and R_S tokens added to the protein sequences and SMILES strings, respectively, are concatenated and used as input for an FCNN, which is essentially a Multilayer Perceptron (MLP). After each Dense layer of this block, a dropout layer was employed. Following the FCNN, a Dense layer with a single neuron is applied to predict the binding affinity of the DTI pair measured in terms of the logarithmic-transformed dissociation constant (pK_d).

Figure 7.3 illustrates the proposed DTITR architecture.

7.2.4 Hyperparameter Optimization Approach

The most common approach to determine the model's best architecture and set of parameters is grid search with cross-validation, in which the dataset is split across different folds under different conditions depending on the methodology used, e.g., stratified K-fold splits the dataset into different folds by taking into consideration the distribution of the classes. However, in the context of the problem, traditional cross-validation approaches are usually not satisfactory or representative, especially when considering that the Davis dataset is extremely imbalanced toward the pK_d values distribution and that 1D raw sequential and structural data is used to characterize the proteins and compounds. On that account, the DTI representability of each fold is determinant in the learning process of the architecture.

The Chemogenomic Representative K-Fold method [341] was applied to split the dataset into representative folds and determine the hyperparameters. This method takes into consideration the pK_d values distribution, the protein sequences similarity, and the SMILES strings similarity during the splitting process. It initially distributes the DTI pairs with a pK_d value greater than 5 (relevant interactions) across the different K folds based on the lowest similarity score. This metric corresponds to

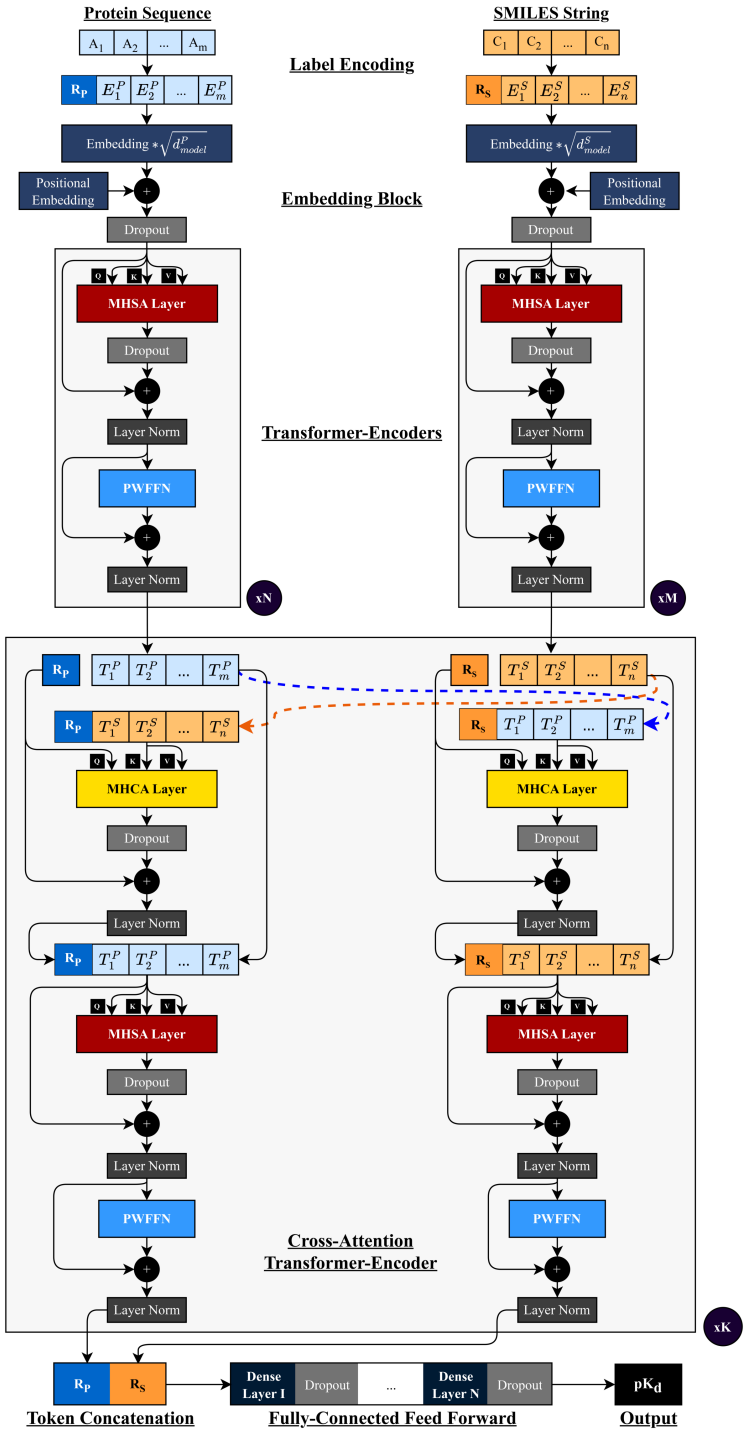


Figure 7.3: DTITR: End-to-End Transformer-based architecture. Two parallel Transformer-Encoders compute a contextual embedding of the protein sequences and SMILES strings, and a Cross-Attention Transformer-Encoder models the interaction space and learns the pharmacological context of the interaction. The resulting aggregate representations of the proteins (R_P) and compounds (R_S) are concatenated and used as input for a FCNN. The final dense layer outputs the binding affinity measured in terms of pK_d

the weighted mean between the median value across all the protein sequences' similarity scores and the median value across all the SMILES strings' similarity scores, which are calculated between the sample and each entry in the corresponding set. Additionally, this method also guarantees that every set is equally sized, thus, only sets that had not previously been assigned a sample are considered at each step (until it is reset). Following the pairs with a pK_d value greater than 5, this process is repeated for the DTIs with a pK_d value equal to 5 (weak interactions).

Considering the improved representability of each fold obtained by this splitting methodology, it is also possible to extract an independent testing set in order to estimate the model's performance in the context and chemogenomic domain of the problem and evaluate the generalization capacity.

7.3 Results and Discussion

7.3.1 Predictive Performance Evaluation

In the context of drug discovery and drug repositioning, it is crucial to accurately predict the binding strength of DTI pairs to properly identify and distinguish main interactions from those with secondary targets (off-targets). In order to validate the performance of the proposed DTITR architecture, the prediction efficiency was evaluated and compared with different state-of-the-art binding affinity regression models (see Section C.3 in Appendix C for more details regarding the experimental setup conducted in this study). Table 7.2 reports the binding affinity prediction results over the Davis independent testing set in terms of five different metrics: MSE, RMSE, CI, r^2, and Spearman.

The proposed DTITR architecture achieved superior performance across almost all metrics, specifically MSE (0.192), RMSE (0.438), CI (0.907), and r^2 (0.771) when compared to the state-of-the-art baselines. The lower MSE and RMSE scores demonstrate the capacity of the model to correctly predict the binding strength values, and the higher CI score indicates the ability of the architecture to correctly distinguish the binding strength rank order across DTI pairs, which is not only crucial in the drug discovery context to differentiate primary from secondary or weak interactions, but also of special interest given the imbalance nature of the pK_d values distribution of the Davis dataset.

Contrarily to the majority of the baseline methods, where either only individual representations of the proteins and compounds are being learnt by the model or

Table 7.2: Binding affinity prediction results over the Davis independent testing set.

Method	Protein Rep.	Compound Rep.	↓ MSE	↓ RMSE	↑ CI	↑ r^2	↑ Spearman
Baseline Methods							
KronRLS [265]	Smith-Waterman	PubChem-Sim	0.443	0.665	0.847	0.473	0.624
GraphDTA-GCNNet [271]	1D-Subseq	Graph	0.311	0.558	0.883	0.630	0.681
GraphDTA-GATNet [271]	1D-Subseq	Graph	0.286	0.535	0.881	0.660	0.688
SimBoost [267]	Smith-Waterman	PubChem-Sim	0.277	0.526	0.891	0.670	0.694
GraphDTA-GAT-GCN [271]	1D-Subseq	Graph	0.269	0.518	0.874	0.680	0.670
Sim-CNN-DTA [273]	Smith-Waterman	PubChem-Sim	0.266	0.516	0.884	0.683	0.674
GraphDTA-GINConvNet [271]	1D-Subseq	Graph	0.238	0.488	0.899	0.717	**0.741**
DeepDTA [268]	1D-Subseq	1D	0.215	0.464	0.891	0.743	0.691
DeepCDA [272]	1D-Subseq	1D	0.208	0.457	0.895	0.752	0.689
Proposed Method							
DTITR	1D-Subseq	1D	**0.192**	**0.438**	**0.907**	**0.771**	0.712

Bold indicates the best performance value associated with each evaluating metric.

only the mutual interaction space is being considered during the inferring process, the DTITR architecture takes simultaneously into consideration the magnitude of certain local regions of each binding component (and their intra-associations) and the involving interaction substructures, resulting in robust representations of the protein sequences and SMILES strings. On that account, the results demonstrate that the DTITR model is properly learning the proteomics, chemical, and pharmacological context information of the proteins, compounds, and protein-compounds interactions, respectively, considering that the final aggregate representations are robust and discriminative for the prediction of binding affinity.

Figure 7.4 illustrates the predictions from the DTITR model against the actual (true) binding affinity values for the Davis testing set, where it is possible to observe a significant density around the *predicted = true value* reference line (perfect model).

7.3.2 Ablation study

In order to further validate the DTITR architecture, three different alternatives for the DTITR model were explored, specifically (i) DTITR architecture without the Cross-Attention Transformer-Encoder block, (ii) DTITR architecture without the FCNN block, and (iii) FCS and BPE encoding applied to the SMILES strings instead of the integer-based character-dictionary method. Table 7.3 reports the binding affinity prediction results over the Davis independent testing set in terms of the five different metrics for the different alternatives of the DTITR model.

To properly assess the efficacy of the Cross-Attention Transformer-Encoder block,

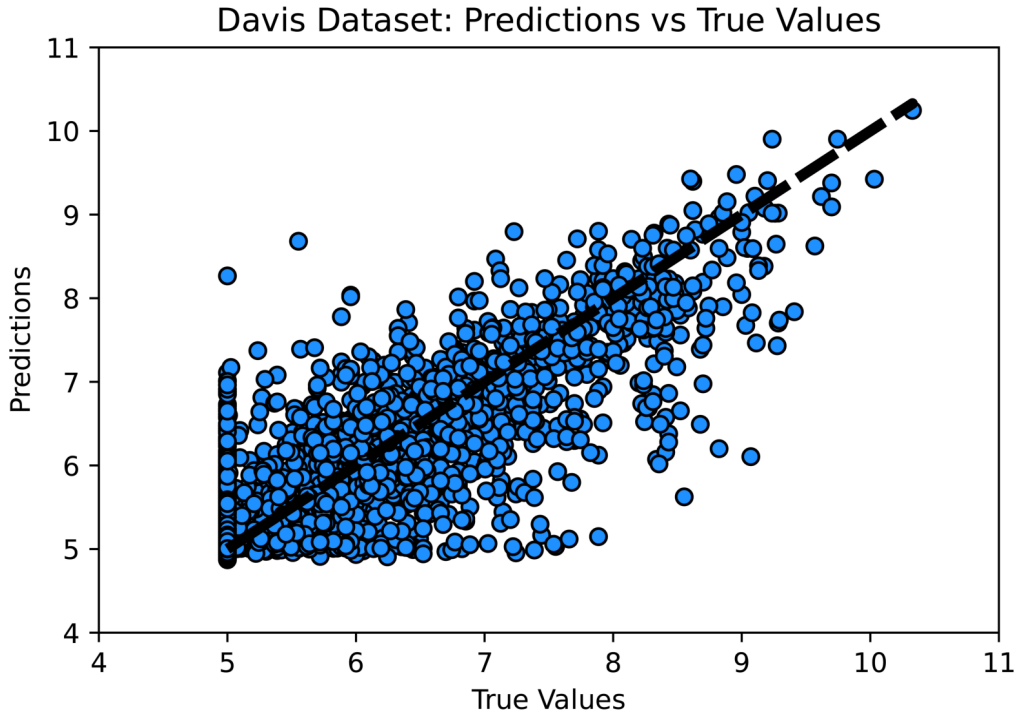


Figure 7.4: DTITR predictions against the true values for the Davis testing set, where the diagonal line is the reference line (*predicted = true value*).

Table 7.3: Binding affinity prediction results over the Davis independent testing set for the different alternatives of the DTITR model.

Method	Protein Rep.	Compound Rep.	↓ MSE	↓ RMSE	↑ CI	↑ r^2	↑ Spearman
DTITR - I	1D-Subseq	1D	0.232	0.481	0.906	0.724	0.712
DTITR - II	1D-Subseq	1D-Subseq	0.205	0.453	0.905	0.756	0.712
DTITR - III	1D-Subseq	1D	0.196	0.443	0.899	0.766	0.703
DTITR	1D-Subseq	1D	**0.192**	**0.438**	**0.907**	**0.771**	**0.712**

I - Without FCNN Block, II - Both Subseq (FCS and BPE Encoding), III - Without Cross-Block. Bold indicates the best performance value associated with each evaluating metric.

which is responsible for the exchange of context information between proteins and compounds (pharmacological space), the model prediction efficiency with and without this module was evaluated. The DTITR architecture with the Cross-Attention Transformer-Encoder block resulted in overall better performance in terms of the MSE (0.192), RMSE (0.438), CI (0.907), r^2 (0.771) and Spearman (0.712) scores when compared to the DTITR architecture without the Cross-Block (MSE - 0.196, RMSE - 0.443, CI - 0.899, r^2 - 0.766 and Spearman - 0.703). These results demonstrate that using the Cross-Attention Transformer-Encoder block to learn the pharmacological context information associated with the interaction space improves the discriminative power of the final aggregate representation hidden states for the pre-

diction of binding affinity. Moreover, it indicates that the use of only the individual proteomics and chemical contextual information of the protein sequences and SMILES strings, respectively, leads to worse performance when compared to combining the proteomics, chemical, and pharmacological contexts, which is in agreement with the fact that DTIs result from the recognition and complementarity of certain substructures (pharmacological space) but are supported by the joint action of other individual substructures scattered across the proteins (proteomics space) and compounds (chemical space).

Regarding the prediction efficiency of the DTITR model without the FCNN block, the performance obtained over the independent testing set is worse in terms of the MSE (0.231), RMSE (0.481), and r^2 (0.724) scores when compared to the DTITR architecture with this block (MSE - 0.192, RMSE - 0.438, and r^2 - 0.771). These results demonstrate that the use of the FCNN increases the learning capacity of the architecture and aids in the generalization from the concatenated aggregate representations space, which describes the DTI, to the output space.

Additionally, the differences in the prediction performance of the model by applying the same encoding approach of the protein sequences to the SMILEs strings instead of using the character-dictionary encoding method mentioned in Section 7.2.2 were also explored. The performance achieved is substantially worse (MSE - 0.205, RMSE - 0.453 and r^2 - 0.756), except for the CI (0.905) and Spearman (0.712) scores, when compared to using the proposed integer-based encoding method. These results suggest that employing the FCS and BPE algorithms to represent the SMILES strings reduces the learning capacity of the DTITR model, which might be a consequence of the restrictive representation of the SMILES strings since this encoding method results in a maximum length of 15 for the SMILES strings in the Davis dataset.

Overall, the use of an end-to-end Transformer-based architecture for predicting binding affinity demonstrates the ability to use Transformer-Encoders to learn robust and discriminative aggregate representations of the protein sequences and SMILES strings. Moreover, it shows the capacity of the self-attention layers to learn the context dependencies between the sequential and structural units of the proteins and compounds, respectively, and the cross-attention layers to exchange information and model the interaction space.

7.3.3 Attention Maps

DTIs are primarily substructural, where the recognition and complementarity of certain substructures are crucial for the interaction, but the support of the joint

action of other individual substructures scattered across the protein and compound also plays a key role in the overall binding process. On that account, visualizing the overall importance of the input components and their associations to the model may potentially lead not only to understanding the model prediction but also to significant findings in the DTI domain. The DTITR architecture contains three different levels of attention: (i) self-attention over the individual units of the protein sequences and SMILES strings; (ii) cross-attention between the protein sequences and SMILES strings; and (iii) self-attention over the individual units of the protein sequences and SMILES strings after the cross-attention (interaction). The first level of attention provides information about the overall importance of the individual units (substructures) and intra-associations of protein sequences and SMILES strings prior to the interaction, i.e., the individual importance of the proteomics space and chemical space. On the other hand, the second level provides clues about which protein and compound substructures lead to the interaction, in particular, which compound substructures the protein attends to and vice versa. The third level of attention provides information about how the individual proteomics and chemical importance shifts after the interaction, i.e., how the pharmacological information affects the overall importance of the individual units and intra-associations of protein sequences and SMILES strings.

In order to visualize the attention levels, heat maps for the second level of attention were generated, specifically for the attention of the R_S token over the protein substructures (subwords). Four different DTI pairs were selected, particularly ABL1(E255K)-phosphorylated - SKI-606, DDR1 - Foretinib, ERBB4 - Lapatinib, and BRAF - PLX-4720, where only subwords associated with interaction residues were considered for visualization. In the case of the ERBB4 - Lapatinib and BRAF - PLX-4720 DTI pairs, the binding positions were collected from the sc-PDB [345] database, which is a specialized structure database focused on ligand binding site in ligandable proteins, i.e., contains some experimental 3D interaction complexes with the binding regions known/available. On the other hand, the BL1(E255K)-phosphorylated - SKI-606 and DDR1 - Foretinib DTI pairs do not have experimental 3D interaction complexes available/known, thus, the 3D interaction space was explored using docking approaches [341]. On that account, potential binding positions were selected based on a distance threshold of ≤ 5 Å from the resulting 3D receptor-ligand complexes, which were obtained by using guided docking (AutoDock Vina [352]) based on the highest scoring binding pocket from the DoGSiteScorer [161] platform. Figure 7.5 illustrates the attention heat maps for ABL1(E255K)-phosphorylated - SKI-606, DDR1 - Foretinib, ERBB4 - Lapatinib, and BRAF -

PLX-4720, where the attention weights were normalized across all the positions for each head of attention.

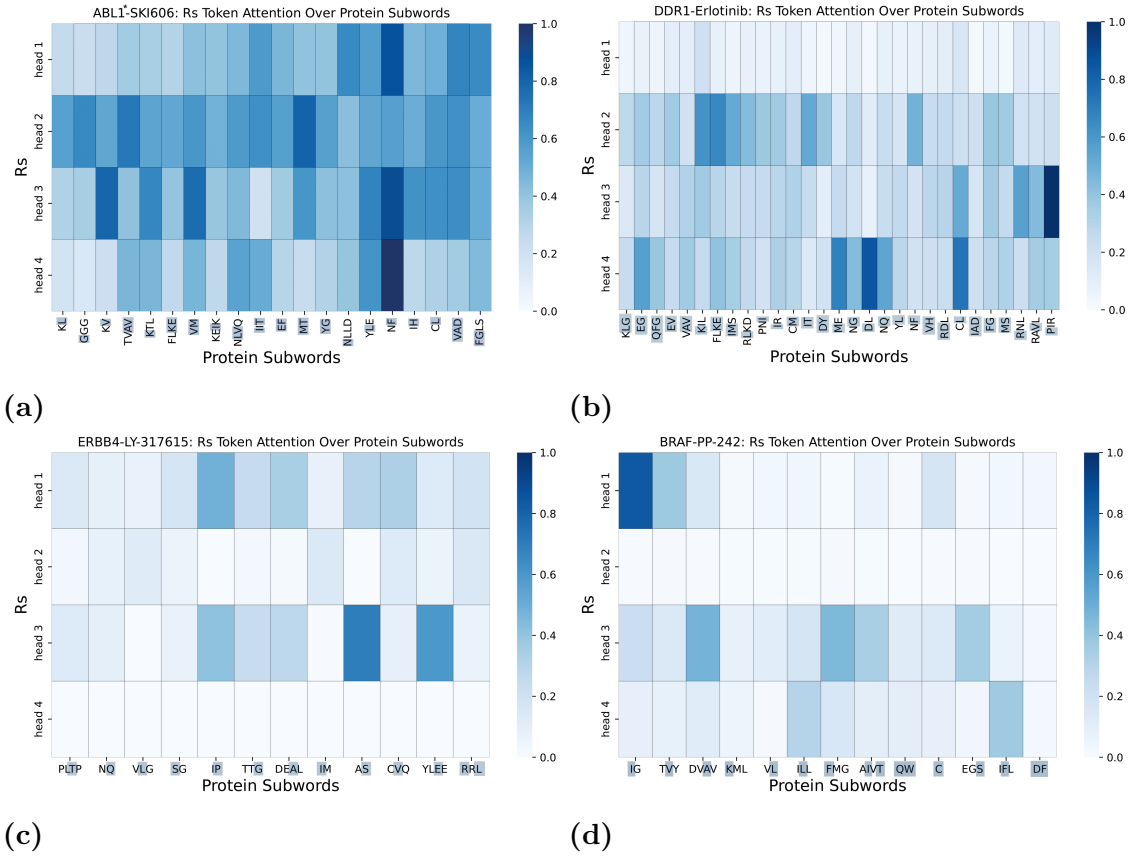


Figure 7.5: Attention maps for the attention of the R_S token over the protein substructures, where the interacting residues within the protein subwords are highlighted in gray. a) ABL1(E255K)-phosphorylated - SKI-606; b) DDR1 - Foretinib; c) ERBB4 - Lapatinib; d) BRAF - PLX-4720.

These visual results show that the R_S token, which is an aggregate representation of the compound, is attending, i.e., giving weight, to substructures of the protein sequences associated with binding residues. For each one of these DTI pairs, there are binding-related substructures with a high percentage of significance (weight) in almost every head of attention, e.g., head 4 - motif NF, head 3 - motif PIR, head 3 - motif AS, and head 1 - motif IG for the ABL1(E255K)-phosphorylated - SKI-606, DDR1 - Foretinib, ERBB4 - Lapatinib and BRAF - PLX-4720 interaction pairs, respectively. Moreover, in the particular case of the ABL1(E255K)-phosphorylated - SKI-606 interaction pair, all heads of attention highly attend to almost every substructure. Overall, these findings demonstrate that the Cross-Attention Transformer-Encoder block is learning the pharmacological context of the DTIs, indicating that the DTITR architecture is capable of providing reasonable

evidence for understanding the model prediction and potentially leading to new knowledge about DTIs.

7.4 Conclusions

7.4.1 Final Remarks

In this research study, an end-to-end Transformer-based architecture (DTITR) is proposed for predicting the logarithmic-transformed quantitative dissociation constant (pK_d) of DTI pairs, where self-attention layers are exploited to learn the short and long-term proteomics and chemical context dependencies between the sequential and structural units of the protein sequences and compound SMILES strings, respectively, and cross-attention layers to exchange information and learn the pharmacological context associated with the interaction space. The architecture makes use of two parallel Transformer-Encoders to compute a contextual embedding of the protein sequences and SMILES strings, and a Cross-Attention Transformer-Encoder block to model the interaction, where the resulting aggregate representations are concatenated and used as input for an FCNN. The experiments were performed on the Davis kinase binding affinity dataset, and the performance of the proposed model was compared with different state-of-the-art binding affinity regression baselines.

The proposed model yielded better results than state-of-the-art baselines. It obtained lower MSE and RMSE values and a higher CI score, demonstrating the model's ability to correctly predict the value of the binding strength and correctly distinguish the rank order of binding strength between the DTI pairs, respectively. In addition, the DTITR architecture is shown to efficiently learn the proteomics, chemical, and pharmacological context of the proteins, compounds, and protein-compound interactions, respectively, given the robustness and discriminative power of the resulting aggregate representations of the protein sequences and SMILES strings.

Additionally, various formulations of the DTITR architecture were examined. It was found that the Cross-Attention Transformer-Encoder, which is responsible for the exchange of information between the protein sequences and SMILES strings and learning the pharmacological context, leads to better performance than when only the two initial parallel Transformer-Encoders are used. These results show that combining the proteomics, chemical, and pharmacological contexts improves the robustness and discriminative power of the aggregate representations compared to using only the individual proteomics and chemical context information of the

protein sequences and SMILES strings. In addition, the FCNN block was found to improve the learning capacity of the architecture as it can improve the generalization from the concatenated aggregate representations space to the output space.

Considering the nature of the attention layers, which give information about the overall importance of the input components and their associations to the model, the DTITR architecture provides three different levels of potential DTI and prediction understanding. The attention maps for the second level of attention (Cross-Attention Transformer-Encoder block), specifically for the attention of the aggregate representation of the compounds over the protein sequences substructures, were visualized. The results show that the compounds are attending to subwords of the protein sequences associated with binding residues, confirming the ability of this block to properly learn the pharmacological context of the DTIs. It also demonstrates that the DTITR architecture is capable of providing reasonable model understanding and potentially leading to new insights in the DTI field.

The major contribution of this study is an efficient and novel end-to-end Transformer-based DL architecture for predicting binding affinity that simultaneously considers the magnitude of certain local regions of each binding component (proteomics and chemical context) and the interacting substructures involved (pharmacological context). Moreover, this architecture provides three different levels of potential DTI and prediction understanding, which is critical in the context of drug discovery.

7.4.2 Study Limitations and Future Work

Despite the improved capacity of the proposed architecture to model the inter-dependency of the sequential and structural units of each binding component (and their intra-associations) and the inter-associations that revolve around the binding substructures (context of the interaction), i.e., the multi-domain inter-dependency of the proteomics, chemical, and pharmacological spaces, DTITR is unequivocally computationally complex, especially regarding the self-attention layers, which have a computational complexity of $O(n^2)$ with respect to the sequence length. In that regard and given that the main efficiency component in the Transformer-Encoder is its attention mechanism, DTITR is limited for long-sequence proteins. Even though the FCS/BPE encoding method employed in this study reduces the sequential representation space of the protein sequences, it drastically increases the number of parameters in the embedding block due to the cardinality of the protein subwords dictionary. Moreover, the FCS/BPE encoding method deepens the compromise of

the input data's representability in the architecture's learning process.

Large Language Models (LLMs) such as Transformer-Encoders have been proven to lead to improved performance and learning capacity by initially pre-training with a large corpus associated with the input domain in an unsupervised fashion, i.e., leveraging a vast amount of data points to learn the context, semantics, and inter-dependencies within the domain space. On that account, considering that the aggregate representations of the protein sequences and SMILES strings are not only used as attending agents in the cross-attention block to capture the pharmacological context and inter-dependencies amongst the binding structures but also for the prediction of binding affinity, pre-training the Transformer-Encoders blocks associated with the protein sequences and SMILES strings can improve the overall learning capacity and increase the discriminative power and robustness of the aggregate representations. Furthermore, pre-training these Transformer-Encoder Blocks, which learn the short and long-term proteomics and chemical context dependencies between the sequential and structural units of the protein sequences and compound SMILES strings, i.e., the proteomics and chemical contexts associated with DTIs, can reduce the effects of the proteomics and chemical domains representability in the learning stage of DTITR, and greatly improve the training speed of the architecture. Nevertheless, this requires considerable resources, especially in the case of protein sequences.

DTITR provides multiple levels of potential DTI and prediction understanding due to the nature of the attention blocks, which give information about the overall importance of the input components and their intra-associations to the model. Even though it was shown that the aggregate representation of the compound attends to substructures of the protein sequences associated with binding residues, DTITR does not take into account information about binding sites/interaction regions during the training process or actively integrates binding-related information during the learning stage of the architecture. Thus, it does not model the inter-dependency amongst binding-related tokens, i.e., the inter-dependency within the binding region, or exclusively the interaction between the compound and the binding region. On that account, the aggregate representation of the compound also attends to redundant positions during the learning stage of the pharmacological space, introducing noise in the resulting aggregate representation. Furthermore, actively integrating information about binding sites in the training process is crucial to properly learning the pharmacological space considering that the protein's binding pocket is paramount in the interaction mechanism involved in the DTIs. Furthermore, to realistically model DTIs and improve the reliability of the predictions, it is vital to consider binding

pockets in the learning process.

DL models, especially LLM-related architectures, perform significantly better when the dataset becomes larger due to their capacity to learn the context of the input domain and the inter-dependency amongst the units of the input data. Hence, focusing on building a larger and more valid DTI dataset measured in terms of the K_d constant is essential to increase the generalization capacity of the architecture beyond the limited space of the Kinase domain. Nevertheless, the Davis Kinase dataset remains an important benchmark to validate the prediction efficiency of the architectures in the context of DTA prediction.

Chapter 8

Binding-Region-Guided Strategy to Predict Drug–Target Affinity

This chapter concerns the use of a binding-region-guided strategy to model the pharmacological space of the interaction and learn the inter-dependency amongst binding-related positions for the prediction of binding affinity. This study also focuses on two contextually related yet computationally different tasks, specifically binding pocket classification and binding affinity regression. This chapter resulted in a novel approach capable of providing increased DTI and prediction understanding due to the nature of the attention blocks and prediction of the binding pocket.

The content of this chapter is based on a journal article published in *Expert Systems with Applications* [353]. Section 8.1 presents the study context. Section 8.2 details the materials and methods used in this study. Section 8.3 reports the results and discusses the obtained findings. Section 8.4 provides some final reflections and the limitations of this study.

8.1 Study Context

The discovery of compounds that selectively bind to relevant and ligandable proteins remains one of the greatest challenges in drug discovery. In spite of the existing comprehensive chemical and proteomic libraries and numerous computational approaches in the DTI and DTA prediction fields, the proper modeling of the multi-domain inter-dependency of DTIs is still limited, compromising the validity and reliability of the results and inferring process [4].

Recent research endeavors dealing with DTI or DTA prediction have been exploring attention-based frameworks to learn short and long-term term context intra-dependencies within proteins and/or compounds and inter-dependencies amongst in-

teracting substructures, promoting the prediction performance and ability to provide reasonable model understanding in the DTI domain [245, 243, 244]. However, these approaches have yet to actively include information about binding sites/pockets into the learning process, introducing noise when modeling the pharmacological space and the inter-associations around the binding substructures. Moreover, the identification of protein-ligand binding pockets is paramount to understanding the biological functions of proteins and the mechanisms involved in DTIs [156, 157]. Therefore, the explainability in the DTI domain is particularly limited and partially compromised, given the lack of explicit evidence to support the pharmacological space representation and the redundancy introduced by attending to potentially inaccurate binding regions when learning the interaction domain.

This study explores an end-to-end Transformer-based framework to simultaneously predict the 1D binding pocket and the DTA measured in terms of the dissociation constant (K_d), where the prediction of the binary binding vector, which states the binding nature of each protein residue, guides (conditions) the prediction of the binding affinity. The targets and compounds are represented using 1D sequential and structural information, specifically protein sequences and SMILES strings, respectively. This architecture, TAG-DTA, consists of two Transformer-Encoder-based prediction models, specifically a 1D binding pocket classifier and a binding affinity regressor, and shares three cores layers, including lower Transformer-Encoders and a condition-based concatenation block. The framework leverages the use of lower self-attention layers to learn the short and long-term proteomics and chemical context dependencies between the sequential and structural units of the proteins and compounds, respectively, and a condition-based concatenation layer to represent the pharmacological (interaction) space. The binding pocket Transformer-Encoder uses the pharmacological space representation for binary token labeling, where the predicted binary 1D binding pocket is used to condition the attention mechanism of the binding affinity Transformer-Encoder, resulting in the exchange of information between the proteomics and chemical domains over binding-related residues. The resulting aggregate representations of the proteomics, chemical, and binding-region-based pharmacological spaces are concatenated and fed into a fully-connected feed-forward network (FCNN), which predicts the binding strength of DTI pairs. Furthermore, the proposed framework leads to increasing DTI and model understanding not only due to the nature of the attention blocks, which give information about the overall importance of the input components and their intra-associations, but also due to the prediction of the 1D binding pocket, which determines the attention mechanism over the interaction space and shows explicit evidence of potential

key residues within the protein sequences for the binding process.

8.2 Materials and Methods

8.2.1 Binding Affinity Dataset

To establish the binding affinity prediction model, drug–target pairs were collected from the Davis et al. [276] research study, which comprises selectivity assays associated with the human catalytic protein kinome measured in terms of a quantitative dissociation constant (K_d). This study covers the interaction between 442 kinases and 72 kinase inhibitors, resulting in 31 824 DTIs. K_d expresses a direct and unbiased measurement of the equilibrium between the receptor-ligand complex and the dissociation components, where lower values are associated with strong interactions.

The protein sequences of the Davis dataset were collected from UniProt [57] using the corresponding accession numbers. Proteins are characterized by a unique amino acid sequence, resulting in varying sequence lengths. To standardize the number of features and avoid the loss of relevant sequential information or increasing noise, the sequence length was fixed between 264 and 1400 residues based on a 95% information density threshold. Considering the computational complexity of the self-attention layers of $O(n^2)$ concerning the sequence length, the approach proposed in the Huang et al. (2021) [244] research study, which combines a frequent consecutive subsequence (FCS) mining method with the byte pair encoding (BPE) algorithm, was applied to represent and encode the protein sequences. This method decomposes each protein sequence into an ordered set of non-overlapping frequent subsequences, where the aggregation of all subsequences must recover the original sequence. The hierarchy dictionary of frequent subsequences (V_P) comprises 16 693 different subwords.

The SMILES strings of the Davis dataset were extracted from PubChem [343] based on their PubChem CIDs. To ensure a consistent notation to represent the chemical structure of all compounds, the RDKit [344] canonical transformation was applied to every SMILES string. Similarly to the protein sequences, the sequential length of the SMILES strings was fixed between 38 and 72 chemical characters. To represent and encode the SMILES strings into numerical values, an integer-based encoding based on a character-integer dictionary was applied to convert each chemical token into the corresponding numerical value. The character-integer dictionary (V_S) contains a total of 72 unique letters (labels) resulting from the scanning of approximately 1.3 M SMILES strings from the ChEMBL [6] database.

The distribution of the Davis K_d values is significantly skewed toward K_d equal to 10 000 nM, which is associated with extremely weak or almost non-existing interactions. Moreover, the variance of the distribution is considerably high, thus, to avoid high learning losses, the K_d values were transformed into the log space (pK_d) using Equation 8.1. The distribution of the pK_d values spans from 5 (10 000 nM) to approximately 11.

$$pK_d = -log_{10}(\frac{K_d}{10^9}) \tag{8.1}$$

The Davis dataset was split into six different folds using the chemogenomic representative k-fold [341] method, where one of the folds was selected as an independent test set to estimate the performance and generalization capacity of the architecture and the remaining folds to determine the hyperparameters of the binding affinity prediction model. The chemogenomic representative k-fold approach takes into consideration the pK_d values distribution, the protein sequences similarity, and the SMILES strings similarity during the splitting process, leading to representative folds in the context of the problem.

8.2.2 1D Binding Pocket Dataset

The interaction between compounds and proteins results from the recognition and complementarity of particular functional groups (binding sites) in the 3D space. In order to construct the 1D binding pocket dataset, 3D complexes with binding information available, i.e., complexes with the interacting residues annotated, and associated with drug-like molecules were collected from scPDB [345], PDBBind [19], and BioLiP [354]. In order to filter and select biologically relevant ligands, the HET group lists from BioLiP [354] and P2Rank [163] were employed, and complexes with less than five binding residues were excluded. Single-chain 3D complexes were regarded as single DTI pairs and multiple-chain 3D complexes were split into single-chain interaction pairs. Most of these 3D complex-based DTI pairs, however, correspond to specific fractions of the protein sequence in the 1D space. Thus, to identify the positions of the interacting residues in the whole protein sequence, it was necessary to map these fragments onto the corresponding UniProt [57] sequence. On that account, the pairwise sequential local alignment function from Biopython [355] was applied to determine the best alignment and identify the corresponding binding positions, where entries with a mismatch greater than 50% between the residues of the original and alignment binding pocket were removed from the dataset. Furthermore, the binding information of single-chain pairs belonging to the same PDB

complex and with identical UniProt sequence was unified into single 1D binding pockets.

The binding sites are usually determined based on protein residues whose distance toward ligands is below a certain threshold. However, the definition of a binding pocket is inconsistent across multiple studies or databases, leading to some noise in identifying the correct interacting residues, especially in a 1D representation. Furthermore, residues in the neighborhood of a particular binding residue likely influence its ligandability, which is consistent with the distribution of the binding sites in a 1D representation. These positions are non-consecutive in the 1D sequence, however, they are prone to be concentrated across scattered local binding regions. Hence, in order to reasonably define the 1D binding pocket, the neighborhood of every single interacting position was also taken into account, i.e., for each binding position p, the residues within the interval $]p-k, p+k[$, where k was fixed at 3 [341], were also considered as binding-related positions. The resulting 1D binding pockets were converted into binary binding vectors of the same length as the corresponding protein sequences, where ones and zeros represent binding and non-binding residues, respectively. Figure 8.1 depicts the process to generate the 1D binding pocket associated with the Penicillin G acylase - Homogentisic acid complex (PDB: 1AJP chain B).

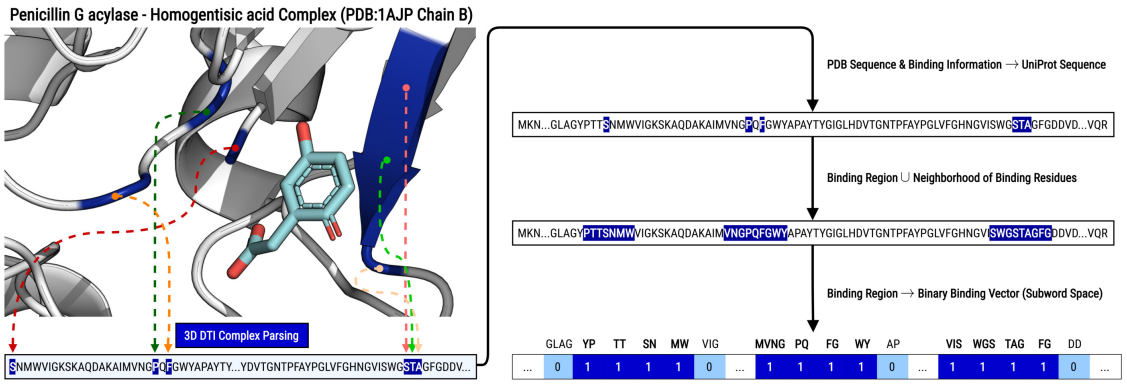


Figure 8.1: Generation of the 1D binding pocket for the Penicillin G acylase - Homogentisic acid complex (PDB: 1AJP chain B). The 3D complex is collected from one of the binding-related databases (scPDB [345], PDBBind [19], or BioLiP [354]) and parsed to the 1D space, in which the protein sequence fragment and the binding positions are retrieved. The 1D binding information is mapped onto the corresponding UniProt [57] sequence using the Biopython [355] package, where the neighborhood of each binding position is also taken into consideration. The resulting 1D binding pocket is converted into a binary binding vector, where ones and zeros represent binding and non-binding residues (subwords), respectively.

The protein sequences and SMILES strings associated with the resulting 1D binding

pocket dataset were processed and encoded based on similar approaches to those applied to the binding affinity dataset. On that account, only proteins with a length between 30 and 575 subwords and SMILES strings with a sequential length between 10 and 100 chemical tokens were selected.

In order to select the hyperparameters of the 1D binding pocket prediction model, the resulting 1D binding pocket dataset was split into a 90/10 % training/validation dataset ratio. On the other hand, to estimate the binding site prediction model's performance, the COACH [162] test dataset, which is widely used in several studies related to binding site prediction, was selected. The COACH test dataset was processed identically to the 1D binding pocket dataset and duplicated PDB complexes were removed from the 1D binding pocket dataset.

8.2.3 SMILES Pre-Train MLM Dataset

In order to pre-train the SMILES Transformer-Encoder block using the Masked Language Modeling (MLM) approach, SMILES strings associated with small compounds that follow the Lipinski's rule of five (zero violations) were collected from ChEMBL [6]. Lipinski's rule defines boundaries for certain physicochemical properties, including molecular weight, lipophilicity, polar surface area, number of hydrogen bond acceptors, number of hydrogen bond donors, and number of rotatable bonds, to determine the drug-likeness (orally active) of the molecules.

The sequential length of the SMILES strings was fixed between 10 and 100 chemical tokens and each character was encoded into the corresponding integer using the 72-character-integer dictionary. Furthermore, the resulting SMILES pre-train MLM dataset was split into a 90/10 % training/validation ratio to select the hyperparameters of the SMILES Transformer-Encoder block.

Table 8.1 summarizes the statistics of the binding affinity, 1D binding pocket, and SMILES pre-train MLM datasets.

See Figure D.1 in Section D.1 of Appendix D for more details regarding the distributions associated with the binding pocket datasets.

8.2.4 TAG-DTA Framework

The TAG-DTA framework simultaneously learns to predict the 1D binding pocket and binding strength of DTIs, where the prediction of the binding sites vector guides and conditions the prediction of DTA. This framework comprises two models, specifically a 1D binding pocket classifier and a binding affinity regressor, and shares three

Table 8.1: Statistics of collected binding affinity, 1D binding pocket, and SMILES pre-train MLM datasets.

Dataset	Proteins	Compounds	DTI	pKd = 5	pKd > 5	Bind Res.[a]	Non-Bind Res.[a]
Binding Affinity							
Davis Original	442	72	31824	22400	9424	-	-
Davis Pre-Processed	423	69	29187	20479	8708	-	-
1D Binding Pocket							
SPB[b]	14256	24151	81696	-	-	1816960 (4.4%)	39340137 (95.6%)
SPB[b,c]	14256	24151	81696	-	-	2464345 (15.9%)	13011113 (84.1%)
COACH Test	402	332	490	-	-	6708 (3.4%)	190626 (96.6%)
COACH Test[c]	402	332	490	-	-	11852 (15.8%)	63194 (84.2%)
SMILES Pre-Train							
ChEMBL	-	1321328	-	-	-	-	-

[a] Residues
[b] scPDB ∪ PDBBind ∪ BioLiP
[c] FCS/BPE Encoding + Neighborhood

core layers, including lower Transformer-Encoders and a condition-based concatenation block. The architecture uses two parallel Transformer-Encoders to compute contextual embeddings and capture the proteomics and chemical context present in the protein sequences and SMILES strings, respectively, where the SMILES Transformer-Encoder is pre-trained using an MLM approach. The aggregate representation of the SMILES string, which corresponds to the final hidden state of the start token added to the SMILES strings, is concatenated with the resulting protein tokens, followed by conditional and positional encoding. The binding site classifier block, which comprises a Transformer-Encoder with a PWFFN, uses the resulting condition-based concatenated tokens as input for binary token labeling learning, predicting the 1D binding pocket. The predicted 1D binding pocket is used to condition the attention mechanism of the Transformer-Encoder of the binding affinity regressor, which also uses the condition-based concatenated tokens as input, by masking non-binding residues. On that account, it learns the pharmacological space and the inter-dependencies amongst the binding-related subwords. The resulting aggregate representations of the binding affinity Transformer-Encoder, protein Transformer-Encoder, and SMILES Transformer-Encoder are concatenated and used as input for an FCNN, which outputs the binding affinity measured in terms of pK_d. Figure 8.2 illustrates the proposed TAG-DTA architecture.

8.2.4.1 Embedding Block

Protein sequences and SMILES strings are tokenized according to the FCS/BPE and character-integer encoding methods, respectively, where each token is converted to a numeric value and each sequence/string is padded up to a maximum value of

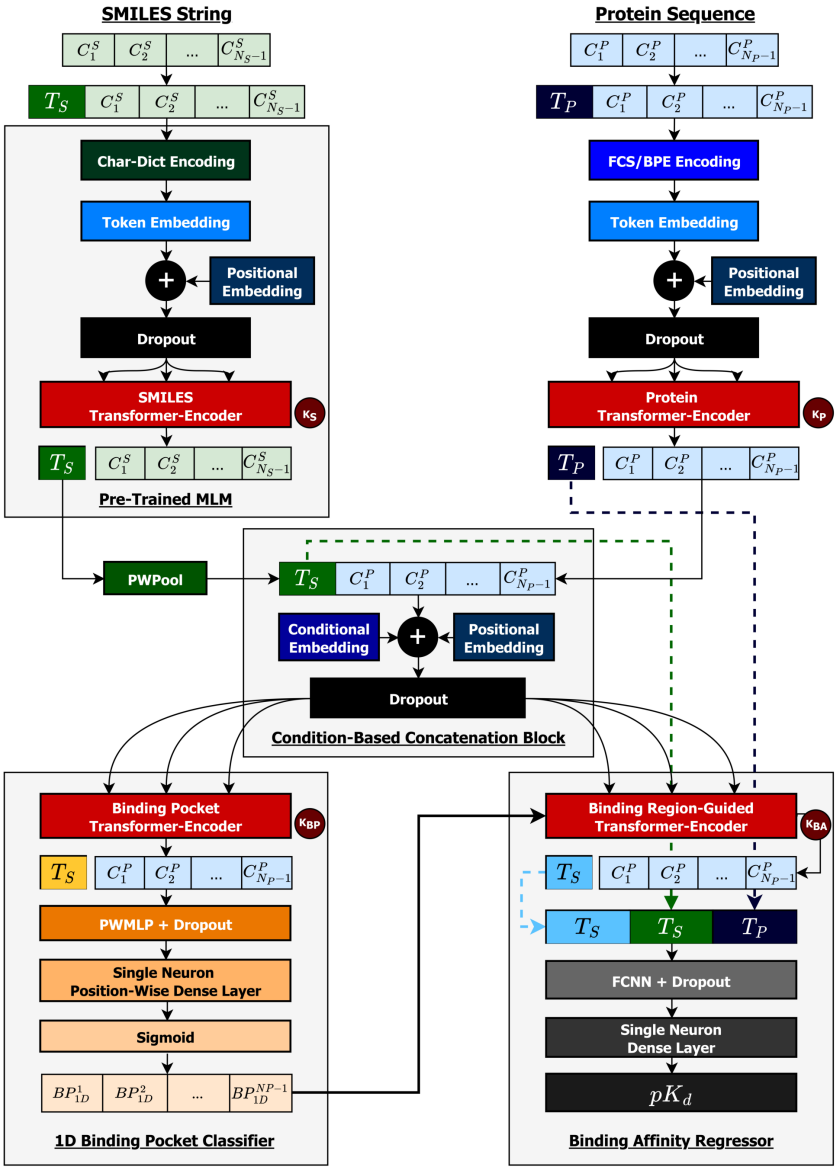


Figure 8.2: TAG-DTA: Binding-Region-Guided Transformer-based architecture. Two parallel Transformer-Encoders capture the proteomics and chemical context present in the protein sequences and SMILES strings, respectively. A condition-based concatenation block concatenates the projected T_S (green) token from the SMILES Transformer-Encoder with the resulting protein tokens from the protein Transformer-Encoder to represent the pharmacological space. The resulting concatenated DTI representation is used as input to the 1D binding pocket classifier for binary token labeling, determining the binding nature of each protein subword. The predicted binary binding vector conditions the attention mechanism of the binding-region-guided Transformer-Encoder, resulting in the learning of the interaction context based on binding-related positions. The resulting aggregate representations of the binding affinity Transformer-Encoder (blue T_S), protein Transformer-Encoder (T_P), and SMILES Transformer-Encoder (green T_S) are concatenated and used as input for an FCNN, which outputs the binding affinity measured in $\mathrm{pK_d}$.

N_P/N_S. Additionally, special start tokens have been added to the beginning of every protein sequence (T_P) and SMILES string (T_S). In order to map semantic meaning into a geometric space, an embedding layer was assigned to the protein sequences and SMILES strings, transforming each token into a learned continuous vector (embedding) with a fixed size of d^P_{model} and d^S_{model} via a learnable dictionary lookup matrix $W^P_{token} \in R^{d^P_{model}\ x\ |V^P|}$ and $W^S_{token} \in R^{d^S_{model}\ x\ |V^S|}$, respectively.

Contrarily to RNNs, Transformer-Encoders do not have a built-in mechanism to deal with the order of sequences, i.e., they are entirely invariant to sequence order. To provide absolute or relative positional information of the tokens in the sequence to the model, a positional embedding was included via a learnable dictionary lookup matrix $W^P_{pos} \in R^{d^P_{model}\ x\ N_P}$ and $W^S_{pos} \in R^{d^S_{model}\ x\ N_S}$. The final embeddings E^P_i and E^S_j associated with the ith and jth input tokens of the protein sequence and SMILES string, respectively, are given by the sum of the token embedding ($E^P_{token_i}$ and $E^S_{token_j}$) and the positional embedding ($E^P_{pos_i}$ and $E^S_{pos_j}$), followed by a dropout layer:

$$\begin{aligned} \mathbf{E}^P_i &= \mathbf{dropout}(\mathbf{E}^P_{token_i} + \mathbf{E}^P_{pos_i}) \\ \mathbf{E}^S_j &= \mathbf{dropout}(\mathbf{E}^S_{token_j} + \mathbf{E}^S_{pos_j}) \end{aligned} \tag{8.2}$$

, where E^P_i, $E^P_{token_i}$, $E^P_{pos_i} \in R^{d^P_{model}}$ and E^S_j, $E^S_{token_j}$, $E^S_{pos_j} \in R^{d^S_{model}}$.

The parameters of the token and positional embedding layers associated with the SMILES strings are initialized with the ones learned during the MLM pre-training stage.

8.2.4.2 Transformer-Encoder

Two Transformer-Encoders in parallel are used to capture the proteomics and chemical context and model the inter-dependency amongst the substructures and molecular components of the protein sequences and SMILES strings, respectively. Transformer-Encoders stack N identical blocks, where each block comprises a MHSA layer and a PWFFN. Residual connections are applied after each subunit followed by LN to mitigate vanishing gradients. Additionally, dropout is added after each MHSA layer and after each dense layer of the PWFFN to prevent overfitting. Considering x^1 and x^2 the outputs of the MHSA layer and the PWFFN block, respectively, the output of the kth Transformer-Encoder block can be expressed as:

$$\begin{aligned} x^1_k &= \mathbf{LN}(x^2_{k-1} + \mathbf{dropout}(\mathbf{MHSA}(x^2_{k-1}))) \\ x^2_k &= \mathbf{LN}(x^1_k + \mathbf{PWFFN}(x^1_k)) \end{aligned} \tag{8.3}$$

, where x_k^1, $x_k^2 \in R^{N_{P/S}\ x\ d_{model}^{P/S}}$.

The weights of the SMILES Transformer-Encoder block are initialized with the ones learned during the MLM pre-training step.

8.2.4.3 Condition-Based Concatenation Block

In order to model the interaction and exchange of information between proteins and compounds, it is crucial to represent the pharmacological space of the interaction. On that account, a condition-based concatenation block is proposed to concatenate the T_S token from the SMILES Transformer-Encoder with the resulting protein tokens from the protein Transformer-Encoder. The T_S token previously learns the overall chemical context and inter-dependency amongst the individual units of the SMILES strings and is thus considered an aggregate representation. On the other hand, each resulting token of the protein sequences from the protein Transformer-Encoder is characterized by a robust and contextual representation based on the learned short and long-term dependencies. Considering X_P and X_S the outputs of the protein and SMILES Transformer-Encoder, respectively, the output of the concatenation layer (X_{DT}) can be expressed as:

$$X_P = [T_P \,\|\, C_P]\ ,\ X_S = [T_S \,\|\, C_S] \longrightarrow X_{DT} = [T_S^{Pool} \,\|\, C_P] \tag{8.4}$$

, where $X_P \in R^{N_P\ x\ d_{model}^P}$, $T_P \in R^{d_{model}^P}$, $C_P \in R^{(N_P-1)\ x\ d_{model}^P}$, $X_S \in R^{N_S\ x\ d_{model}^S}$, $T_S \in R^{d_{model}^S}$, $C_S \in R^{(N_S-1)\ x\ d_{model}^S}$, $T_S^{Pool} \in R^{d_{model}^P}$, and $X_{DT} \in R^{N_P\ x\ d_{model}^P}$.

Considering that the T_S token is used to condition and interact with the proteins tokens, i.e., it is not exclusively the attending agent, a conditional embedding via a learnable dictionary lookup matrix $W_{cond}^{DT} \in R^{d_{model}^P\ x\ N_P}$ was included to distinguish the T_S token from the resulting protein tokens. In order to update the positional information of the tokens in the concatenated DTI representation, a positional embedding via a learnable dictionary lookup matrix $W_{pos}^{DT} \in R^{d_{model}^P\ x\ N_P}$ was added. Following the sum of the conditional embedding and positional embedding, a dropout layer was applied. The final representation E_i^{DT} associated with the ith token of the resulting concatenated tokens can be expressed as:

$$\mathbf{E}_i^{DT} = \mathbf{dropout}(\mathbf{E}_i^{DT} + \mathbf{E}_{cond_i}^{DT} + \mathbf{E}_{pos_i}^{DT}) \tag{8.5}$$

, where $E_i^{DT} \in R^{d_{model}^P}$, $E_{cond_i}^{DT} \in R^{d_{model}^P}$ is the conditional embedding, and $E_{pos_i}^{DT} \in R^{d_{model}^P}$ is the positional embedding.

In order to map the T_S token to the last dimension of the protein tokens (C_P), a position-wise pooling (PWPool) block, which comprises a dense layer applied to the last dimension of T_S and an LN layer, was employed.

$$\mathbf{PWPool}(T_S) = \mathbf{LN}(\mathbf{F}_{T_S}(\mathbf{W}_{T_S}T_S \ + \ \mathbf{b}_{T_S})) \tag{8.6}$$

, where F is the activation function, $W_{T_S} \in R^{d^S_{model} \ x \ d^P_{model}}$, and $b_{T_S} \in R^{d^P_{model}}$

8.2.4.4 1D Binding Pocket Classifier

The 1D binding pocket classifier learns to identify the binding and non-binding positions within the protein sequence, i.e., it predicts each protein subword as a binding or non-binding spot. This block is composed of a Transformer-Encoder followed by a Position-Wise Multi-Layer Perceptron (PWMLP). The binding pocket Transformer-Encoder uses the output of the condition-based concatenation block as input and determines the interaction (and inter-dependencies) between the aggregate representation of the compound and the protein tokens. On that account, the T_S token and the protein tokens attend mutually to each other, where T_S conditions the representation of each token of the protein sequences based on their potential selectivity toward the compound. The resulting protein tokens are fed to the PWMLP, which stacks dense layers applied to the last dimension of the protein tokens (embedding/representation space), in order to increase the learning capacity of this block. Additionally, dropout layers are added after each dense layer of the PWMLP. Following the PWMLP, a single neuron position-wise dense layer is added to classify the binding nature of each protein token, where a sigmoid activation function is applied for binary token labeling. Considering X_{DT} the output of condition-based concatenation block, X^1_{DT} the output of the binding pocket Transformer-Encoder, and X^2_{DT} the output of the PWMLP, the output of the 1D binding pocket classifier (BP_{1D}) can be expressed as:

$$\begin{aligned} X^1_{DT} &= \mathbf{TransformerEncoder}_{BP}(X_{DT}) \\ X^2_{DT} &= \mathbf{PWMLP}(X^1_{DT}) \\ BP_{1D} &= \sigma(X^2_{DT}\mathbf{W}_{BP_{1D}} \ + \ \mathbf{b}_{BP_{1D}}) \end{aligned} \tag{8.7}$$

, where $X_{DT} \in R^{N_P \ x \ d^P_{model}}$, $X^1_{DT} \in R^{N_P \ x \ d^P_{model}}$, $X^2_{DT} \in R^{(N_P-1) \ x \ \pi^2_{DT}}$, $W_{BP_{1D}} \in R^{\pi^2_{DT} \ x \ 1}$, $b_{BP_{1D}} \in R^1$, $BP_{1D} \in \mathbb{Z}_2^{(N_P-1)}$, σ is the sigmoid activation function, and π^2_{DT} is expansion/contraction ratio of the final dense layer of the PWMLP.

8.2.4.5 Binding Affinity Regressor

The interaction between active compounds and proteins results from the recognition and complementarity of certain groups (binding regions) and it is supported by the joint action of other individual substructures scattered across the protein and compound. Hence, to effectively predict binding affinity, it is important to consider the proteomics, chemical, and pharmacological spaces. In order to learn the pharmacological context information associated with the interaction space for the prediction of binding affinity, a binding-region-guided Transformer-Encoder is proposed, where the output of the 1D binding pocket classifier guides the attention mechanism and the output of the condition-based concatenation block is used as input. The binding-region-guided Transformer-Encoder architecture is similar to the standard Transformer-Encoder, however, instead of applying global self-attention, it takes into consideration tokens that are potential binding residues using the predicted 1D binding pocket. On that account, the padding masking matrix is combined with the predicted 1D binding pocket, masking non-binding residues and PAD tokens. Thus, this layer learns the inter-dependencies amongst binding-related residues and the interaction between T_S (compound representation) and the binding tokens of the protein sequences. The resulting T_S token is considered an aggregate representation of the pharmacological space and expresses the short and long-term dependencies between the compound and the binding region.

The final hidden states of the aggregate representation of the protein Transformer Encoder, SMILES Transformer-Encoder, and binding-region-guided Transformer-Encoder, respectively, are concatenated, followed by an LN layer, and used as input for an FCNN. Similarly to the PWMLP, dropout is added after each dense layer of the FCNN. Following the FCNN, a dense layer with a single neuron is applied to predict the binding affinity of the DTI pair measured in terms of the logarithmic-transformed dissociation constant ($\mathrm{pK_d}$). Considering X^1_{Aff} the output of the binding-region-guided Transformer-Encoder and X^2_{Aff} the output of the FCNN, the output of the binding affinity regressor (BF_{pK_D}) can be expressed as:

$$\begin{aligned} X^1_{Aff} &= \mathbf{TransformerEncoder}_{Aff}(X_{DT}) \\ X^2_{Aff} &= \mathbf{FCNN}(X^1_{Aff}) \; , \; X^2_{Aff} = [T_{Aff} \,\|\, C_{Aff}] \\ T_{DTI} &= \mathbf{LN}([T_P \; ; \; T^{Pool}_S \; ; \; T_{Aff}]) \\ BF_{pK_D} &= T_{DTI}\mathbf{W}_{BF_{pK_D}} \; + \; \mathbf{b}_{BF_{pK_D}} \end{aligned} \tag{8.8}$$

, where $X_{DT} \in R^{N_P \; x \; d^P_{model}}$, $X^1_{Aff} \in R^{N_P \; x \; d^P_{model}}$, $T_P \in R^{d^P_{model}}$, $T^{Pool}_S \in R^{d^P_{model}}$,

$T_{Aff} \in R^{d^P_{model}}$, $C_{Aff} \in R^{(N_P-1)\ x\ d^P_{model}}$, $T_{DTI} \in R^{d^{DTI}_{model}}$ is concatenated representation of the final hidden states of the aggregate representations, $W_{BF_{pK_D}} \in R^{d^{DTI}_{model}\ x\ 1}$, $b_{BF_{pK_D}} \in R^1$, $BF_{pK_D} \in R^1$, $d^{DTI}_{model} = d^P_{model} + d^P_{model} + d^P_{model}$, and $[;]$ denotes concatenation.

In order to avoid potential negative transfer of information from the 1D binding pocket classifier, i.e., incapable of identifying any binding position for a certain DTI pair, some flexibility is added to the attention mechanism in the binding-region-guided Transformer-Encoder. On that account, when the predicted binary 1D binding pocket does not contain any value equal to 1 (binding spot), which would lead to the T_S token attending exclusively to itself, only the PAD tokens are masked before the softmax function. Thus, the attention mechanism corresponds to the standard global self-attention, learning the short and long-term inter-dependencies amongst all tokens of the input sequence, except PAD tokens.

$$\textbf{Attention} = \begin{cases} \textbf{Conditioned} & \textbf{if} \quad \exists i = 1,...,N_P - N_{PAD} - 1 : BP_{1D}(i) = 1 \\ \textbf{Global} & \textbf{if} \quad \forall i \in \{1,..,N_P - N_{PAD} - 1\}, BP_{1D}(i) = 0 \end{cases} \tag{8.9}$$

, where *Conditioned* corresponds to the binding-region-guided attention, *Global* to the standard global self-attention, and N_{PAD} to the number of padding tokens.

Figure 8.3 illustrates the masking matrix applied in the binding-region-guided Transformer-Encoder when the predicted binary 1D binding pocket identifies binding spots (binding-region-guided attention).

8.2.4.6 SMILES Pre-Train Masked Language Modeling

MLM leverages a large number of data points in an unsupervised fashion to capture the context and semantics of the input domain. This approach masks certain tokens of the input sequence and designs the model to predict the original tokens based on the unaltered sentence units. On that account, the model needs to learn the statistical and distribution properties as well as the short and long-term dependencies amongst the tokens of the sequence, considering that tokens can have different meanings in different positions. Hence, the model learns deep and multiple representations of the tokens, improving the performance levels in downstream tasks.

In the context of this work, the MLM approach is used to pre-train the SMILES Transformer-Encoder in order to capture the molecular context within the SMILES

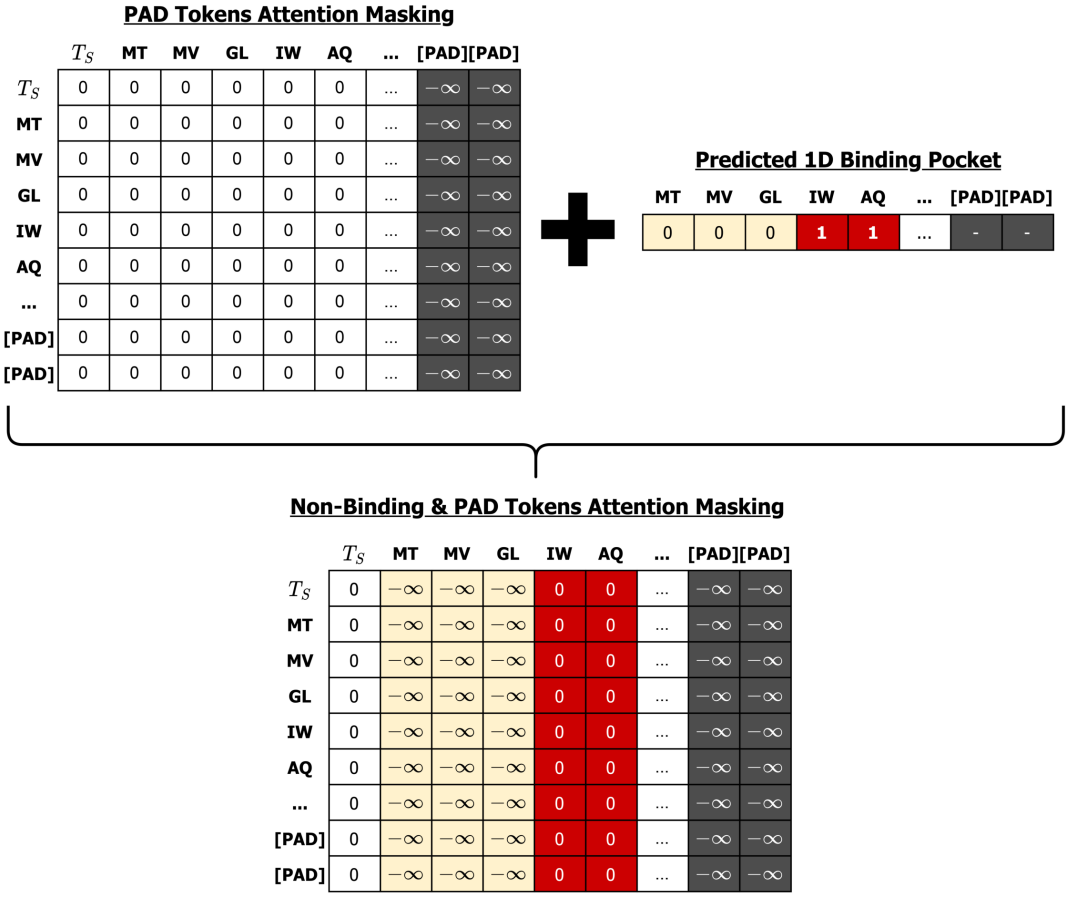


Figure 8.3: Binding region-guided attention masking matrix, where the PAD tokens masking matrix is combined with the predicted 1D binding pocket.

strings. The traditional MLM masking setup used in BERT [173] is followed, which randomly masks 15% of the tokens of the input sequence. The selected tokens are replaced with a special $[MASK]$ token, replaced with a random token from the SMILES dictionary, or remain unaltered based on an 80%, 10%, and 10% probability rate, respectively.

The architecture of the SMILES pre-train MLM includes the SMILES token embedding layer, the SMILES positional embedding layer, the SMILES Transformer-Encoder, a dense layer, and a softmax activation function. The dense layer projects the output of the Transformer-Encoder into the dimension of the SMILES vocabulary V_S. The softmax function normalizes the output of the dense layer to a probability distribution over the cardinality of the vocabulary, indicating the likelihood of each token in the vocabulary for each position of the input sequence. Considering E^S the output of the SMILES embedding block and X_S the output of the SMILES Transformer-Encoder, the output of the SMILES pre-train MLM (X_S^{MLM}) can be

expressed as:

$$X_S = \textbf{TransformerEncoder}_{SMILES}(E^S)$$
$$X_S^{MLM} = \textbf{Softmax}(X_S \mathbf{W}_S^{MLM} + \mathbf{b}_S^{MLM}) \tag{8.10}$$

, where $E^S \in R^{N_S \, x \, d_{model}^S}$, $X^S \in R^{N_S \, x \, d_{model}^S}$, $W_S^{MLM} \in R^{d_{model}^S \, x \, |V^S|}$, $b_S^{MLM} \in R^{|V^S|}$, and $X_S^{MLM} \in R^{N_S \, x \, |V^S|}$.

Figure 8.4 shows the SMILES pre-train MLM architecture.

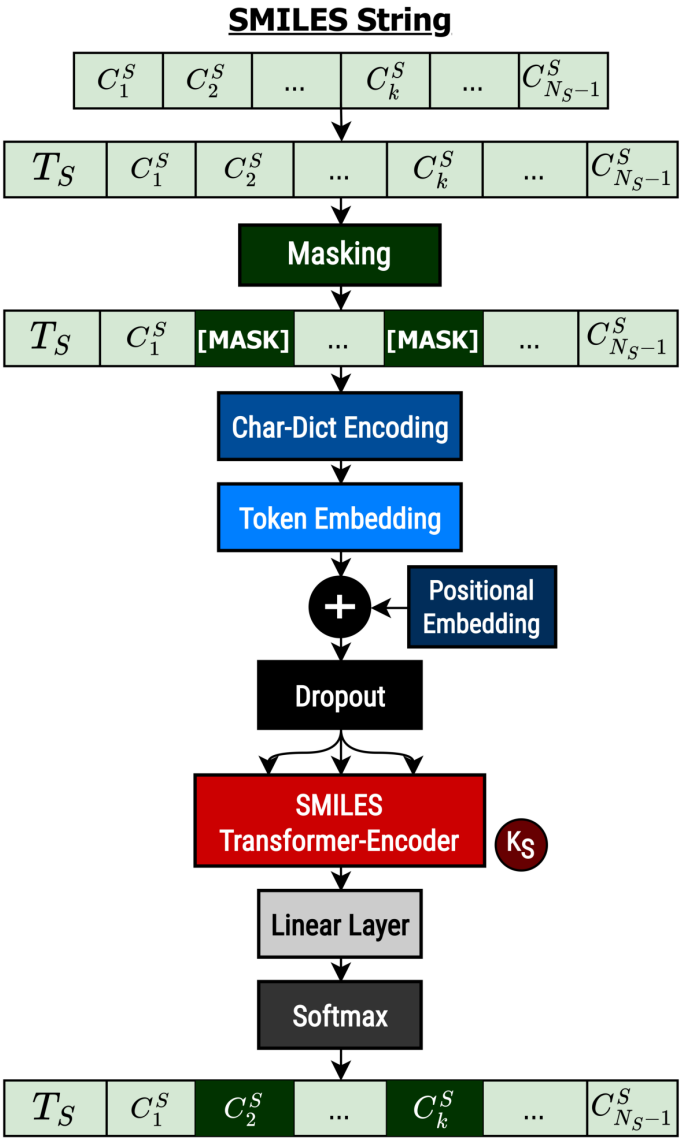


Figure 8.4: Pre-training of the SMILES Transformer-Encoder using an MLM approach, where the model learns to predict the $[MASK]$ tokens based on the unaltered input units.

8.2.5 TAG-DTA Training Strategy

In order to train the TAG-DTA framework, the training strategy is divided into three stages: (I) pre-training of the 1D binding pocket classifier; (II) training of the binding affinity regressor; and (III) training of the 1D binding pocket classifier. The 1D binding pocket subunit, which consists of the three shared blocks, the binding pocket Transformer-Encoder, and the PWMLP, is initially pre-trained in order to partially converge and optimize TAG-DTA toward the prediction of the 1D binding pocket, considering that binding sites vector is used to guide and condition the prediction of binding affinity. Following the first stage, the binding affinity subunit, which is composed of the three shared blocks, the binding region-Guided Transformer-Encoder, and the FCNN, and the 1D binding pocket subunit are alternatively trained to optimize TAG-DTA to simultaneously predict the 1D binding pocket and the binding affinity of DTIs. On that account, two different training modes are considered, where only binding pocket or binding-affinity-related blocks are trainable during each corresponding prediction task, respectively, i.e., the binding pocket Transformer-Encoder and PWMLP weights are frozen during the training of the binding affinity regressor, and the binding-region-guided Transformer-Encoder and FCNN weights are frozen during the training of the 1D binding pocket classifier. Hence, this training strategy aims to reduce the discrepancy between the 1D binding pocket classification and the binding affinity regression and avoid the negative transfer of information between computationally different tasks yet contextually related. Moreover, instead of considering the convergence of the TAG-DTA architecture at the end of each training epoch, the framework is optimized at the end of each training cycle, which corresponds to the end of the two training modes, i.e., the training of the binding affinity regressor and the training of the 1D binding pocket classifier.

8.3 Results and Discussion

Binding Affinity Prediction Performance Evaluation

In the context of drug discovery and drug repositioning, it is essential to properly assess the target selectivity of potential leads. Thus, establishing models capable of accurately predicting unbiased bioactivities associated with the interaction of biologically relevant targets and active small molecules is pivotal on the road to new insights in the DTI field. In order to validate the performance of the proposed TAG-DTA architecture in the prediction of binding affinity, the prediction efficiency was evaluated and compared with different state-of-the-art binding affinity regression

and binary classification models (see Section D.2 of Appendix D for more details regarding the experimental setup conducted in this study) . Table 8.2 reports the binding affinity prediction results over the Davis independent test set in terms of MSE, RMSE, CI, r^2, and Spearman.

Table 8.2: Binding affinity prediction results over the Davis independent test set.

Method	Protein Rep.	Compound Rep.	↓ MSE	↓ RMSE	↑ CI	↑ r^2	↑ Spearman
Baseline Methods							
KronRLS [265]	Smith-Waterman	PubChem-Sim	0.443	0.665	0.847	0.473	0.624
GraphDTA-GCNNet [271]	1D-Subseq	Graph	0.311	0.558	0.883	0.630	0.681
TransformerCPI [243]	1D-Subseq	1D	0.291	0.539	0.852	0.511	0.507
GraphDTA-GATNet [271]	1D-Subseq	Graph	0.286	0.535	0.881	0.660	0.688
SimBoost [267]	Smith-Waterman	PubChem-Sim	0.277	0.526	0.891	0.670	0.694
GraphDTA-GAT-GCN [271]	1D-Subseq	Graph	0.269	0.518	0.874	0.680	0.670
Sim-CNN-DTA [273]	Smith-Waterman	PubChem-Sim	0.266	0.516	0.884	0.683	0.674
GraphDTA-GINConvNet [271]	1D-Subseq	Graph	0.238	0.488	0.899	0.717	**0.741**
HyperAttentionDTI [245]	1D-Subseq	1D	0.227	0.477	0.890	0.729	0.690
DeepDTA [268]	1D-Subseq	1D	0.215	0.464	0.891	0.743	0.691
DeepCDA [272]	1D-Subseq	1D	0.208	0.457	0.895	0.752	0.689
DTITR [350]	1D-Subseq	1D	0.192	0.438	0.907	0.771	0.712
Proposed Method							
TAG-DTA	1D-Subseq	1D	**0.185**	**0.430**	**0.917**	**0.780**	0.729

Bold indicates the best performance value associated with each evaluating metric.

The results demonstrate that the proposed TAG-DTA framework achieved the highest performance in terms of MSE (0.185), RMSE (0.430), CI (0.917), and r^2 (0.780) compared to the state-of-the-art baselines. Thus, it exceeds the other models in its ability to correctly predict the binding affinity values (lower MSE and RMSE scores) and distinguish the binding strength rank order across DTI pairs (higher CI score). Furthermore, the significant increase in the CI metric shows the superior capacity of the architecture to correctly assess the target selectivity, which is crucial in identifying potential leads and differentiating primary from secondary interactions. Moreover, these findings are consistent with the results of Table 8.3, which reports the binding affinity prediction results over the Davis dataset using the original split methodology of the state-of-the-art research works. The TAG-DTA showed a significant increase in performance across all metrics, specifically MSE (0.199 ± 0.003), RMSE (0.446 ± 0.003), CI (0.898 ± 0.001), r^2 (0.752 ± 0.004), and Spearman rank correlation (0.710 ± 0.002), and lower standard deviation values compared to the state-of-the-art baselines, which furthers validates the efficiency of the proposed architecture in the prediction of binding affinity and shows superior learning stability.

In order to properly predict DTA it is essential to learn the inter-dependency of the sequential and structural units of each binding component and the inter-associations

Table 8.3: Binding affinity prediction results over the Davis dataset using the original split methodology, where the standard deviations are given in parentheses.

Method	Protein Rep.	Compound Rep.	↓ MSE	↓ RMSE	↑ CI	↑ r^2	↑ Spearman
Baseline Methods							
KronRLS [265]*	Smith-Waterman	PubChem-Sim	0.379	0.616	0.871 (0.001)	-	-
GraphDTA-GCNNet [271]	1D-Subseq	Graph	0.322 (0.047)	0.566 (0.039)	0.850 (0.020)	0.598 (0.059)	0.636 (0.032)
TransformerCPI [243]	1D-Subseq	1D	0.285 (0.024)	0.533 (0.022)	0.839 (0.017)	0.493 (0.043)	0.472 (0.023)
GraphDTA-GAT-GCN [271]	1D-Subseq	Graph	0.285 (0.004)	0.534 (0.004)	0.862 (0.006)	0.644 (0.005)	0.656 (0.009)
GraphDTA-GATNet [271]	1D-Subseq	Graph	0.284 (0.012)	0.533 (0.012)	0.872 (0.002)	0.646 (0.015)	0.681 (0.010)
SimBoost [267]*	Smith-Waterman	PubChem-Sim	0.282	0.531	0.872 (0.002)	-	-
Sim-CNN-DTA [273]	Smith-Waterman	PubChem-Sim	0.276 (0.008)	0.525 (0.008)	0.869 (0.006)	0.656 (0.010)	0.666 (0.011)
GraphDTA-GINConvNet [271]	1D-Subseq	Graph	0.256 (0.004)	0.506 (0.004)	0.875 (0.005)	0.680 (0.005)	0.708 (0.011)
HyperAttentionDTI [245]	1D-Subseq	1D	0.241 (0.005)	0.491 (0.005)	0.879 (0.002)	0.699 (0.006)	0.680 (0.004)
DeepDTA [268]	1D-Subseq	1D	0.235 (0.006)	0.485 (0.006)	0.871 (0.006)	0.707 (0.007)	0.670 (0.014)
DeepCDA [272]	1D-Subseq	1D	0.232 (0.004)	0.482 (0.004)	0.879 (0.003)	0.710 (0.005)	0.680 (0.005)
DTITR [350]	1D-Subseq	1D	0.216 (0.006)	0.465 (0.006)	0.880 (0.005)	0.730 (0.007)	0.681 (0.008)
Proposed Method							
TAG-DTA	1D-Subseq	1D	**0.199 (0.003)**	**0.446 (0.003)**	**0.898 (0.001)**	**0.752 (0.004)**	**0.710 (0.002)**

*Baselines results from Öztürk et al. [268].
The standard deviations are given in parentheses.
Bold indicates the best performance value associated with each evaluating metric.

that revolve around the binding-related substructures (pharmacological space). The majority of the baseline methods, however, either focus on learning individual representations of the proteins and compounds, which are usually combined for the inferring process, or only take into consideration the mutual interaction space for the prediction of binding affinity. Moreover, the mere use of Transformers, such as TransformerCPI [243], or attention mechanisms, e.g., HyperAttentionDTI [245], do not necessarily guarantee a performance increase, considering that the short and long-distance interactions within the proteins and compounds are crucial for the DTA prediction performance. On that account, DTITR [350] showed superior performance than all previous state-of-the-art baselines since it models the intra-associations within each individual binding component and the inter-dependency between the involving interacting components. However, DTITR [350] does not model the inter-dependency amongst binding-related tokens or the interaction between the compound and the binding region. TAG-DTA overcomes this limitation by employing a binding-region-guided Transformer-Encoder to learn the pharmacological space based on the short and long-term dependencies between the compound and the binding region and the inter-dependency within the binding region. Additionally, it takes into consideration the magnitude of the local regions of each binding component and their intra-associations. Overall, the results demonstrate that the TAG-DTA model is properly learning the proteomics, chemical, and pharmacological context of the proteins, compounds, and DTIs, respectively, and that guiding the learning of the pharmacological space based on potential binding positions leads to improved DTA prediction performance. Figure 8.5 depicts the predictions from

the TAG-DTA model against the actual binding affinity values for the Davis independent test set, where is it possible to observe a significant density around the *predicted = true value* reference line (perfect model).

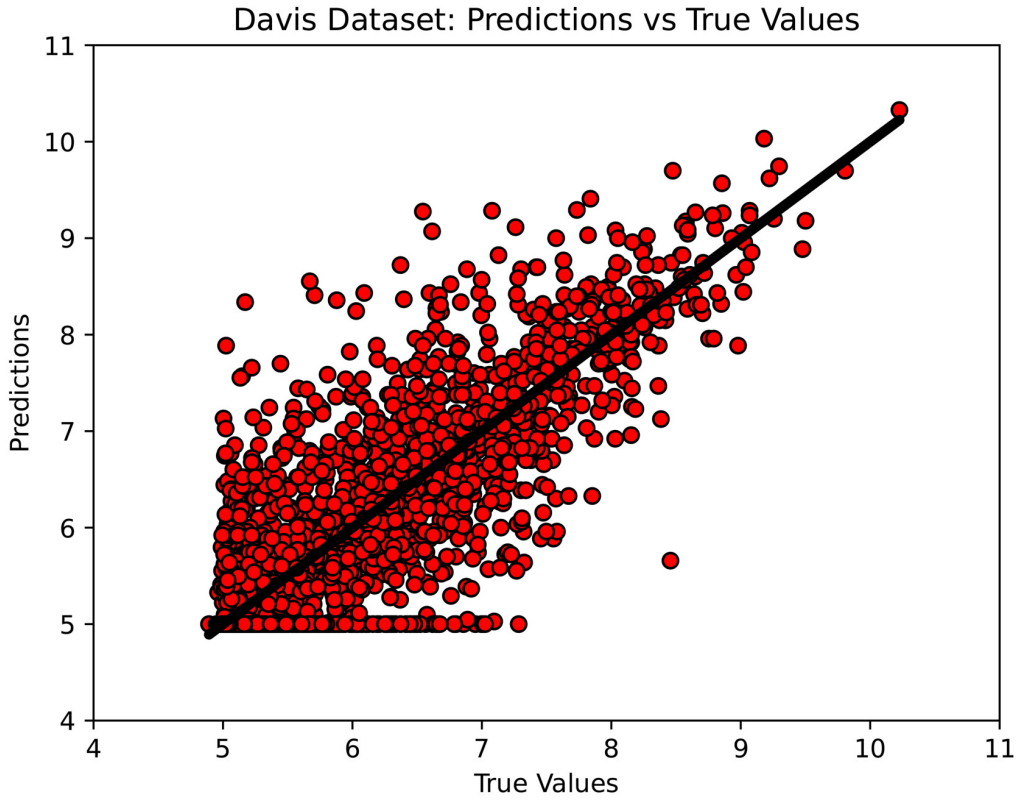


Figure 8.5: TAG-DTA binding affinity predictions against the true values for the Davis affinity testing set, where the black diagonal line corresponds to the reference line (*predicted = true value*).

Considering the importance of the models to generalize toward unknown subsets of the proteomics and/or chemical representation spaces, the performance of the proposed framework in the prediction of binding affinity was explored and evaluated for three different experimental settings, specifically novel compounds, novel proteins, and novel protein-compounds pairs. Table 8.4 reports the binding affinity prediction results over a 5-fold random split of the Davis affinity dataset for these three different experimental settings in terms of MSE, RMSE, and CI. Regarding the novel compound and novel protein-compound pairs settings, TAG-DTA achieved lower values of MSE (0.600 ± 0.064 and 0.609 ± 0.057) and RMSE (0.775 ± 0.042 and 0.780 ± 0.036), and higher CI scores (0.715 ± 0.031 and 0.670 ± 0.058) compared to the state-of-the-art baselines. However, there is no significant difference between the proposed architecture and attention-based baselines, specifically DTITR [350], HyperAttentionDTI [245], and TransformerCPI [243], in these two experimental set-

tings. In the case of the novel protein experimental setting, the proposed TAG-DTA framework achieved an MSE of 0.372 ± 0.044, an RMSE of 0.549 ± 0.035, and a CI score of 0.833 ± 0.013, demonstrating superior and competitive performance to generalize in unknown subsets of the proteomics space. Additionally, all models showed improved performance in the novel protein setting compared to the novel compound and novel protein-compound pair settings, which is consistent with the Kinase representability properties of the Davis affinity dataset. Furthermore, the proposed TAG-DTA model showed overall lower standard deviations amongst the validation sets across all three experimental settings, which demonstrates improved learning stability.

8.3.1 TAG-DTA Ablation Study

In order to further validate the TAG-DTA architecture, different alternatives for the TAG-DTA model were explored, specifically (i) TAG-DTA architecture without pre-training the SMILES Transformer-Encoder, (ii) TAG-DTA architecture without the binding affinity regression block, and (iii) TAG-DTA architecture without the 1D binding pocket classification block. Table 8.5 reports the binding affinity and 1D binding pocket prediction results over the Davis and COACH test set, respectively, for the different alternatives of the TAG-DTA model.

To properly assess the efficacy of pre-training the SMILES Transformer-Encoder using an MLM approach, which leverages a vast amount of data points in an unsupervised fashion to learn the context and semantics of the chemical domain, the model prediction efficiency was evaluated with and without initializing the SMILES Transformer-Encoder related layers weights with the ones learned during the MLM pre-training step (see Table D.3 in Section D.3.1 of Appendix D for the MLM pre-training results). The TAG-DTA architecture with the pre-trained SMILES Transformer-Encoder resulted in overall better performance in terms of the MSE (0.185), RMSE (0.430), CI (0.917), r^2 (0.780), and Spearman (0.729) when compared to the TAG-DTA architecture without pre-training the SMILES Transformer-Encoder (MSE - 0.190, RMSE - 0.436, CI - 0.912, r^2 - 0.773, and Spearman - 0.721) for the prediction of binding affinity. Moreover, in the case of the 1D binding pocket prediction, the TAG-DTA with the pre-trained SMILES Transformer-Encoder demonstrated a significant increase in performance in terms of balanced accuracy (87.10 %), recall (81.32 %), precision (68.17 %), F1-score (74.16 %), and MCC (0.692) when compared to the TAG-DTA architecture without pre-training the SMILES Transformer-Encoder (balanced accuracy - 85.99 %, recall - 79.25 %,

Table 8.4: Binding affinity prediction results over a 5-fold random split of the Davis affinity dataset for three different experimental settings: novel compounds, novel proteins, and novel protein-compound pairs.

Method	Protein Rep.	Compound Rep.	↓ MSE	↓ RMSE	↑ CI
Novel Compound					
Baseline Methods					
GraphDTA-GCNNet [271]	1D-Subseq	Graph	0.808 (0.134)	0.896 (0.074)	0.628 (0.056)
GraphDTA-GATNet [271]	1D-Subseq	Graph	0.793 (0.101)	0.889 (0.056)	0.594 (0.068)
GraphDTA-GAT-GCN [271]	1D-Subseq	Graph	0.763 (0.116)	0.871 (0.066)	0.628 (0.093)
GraphDTA-GINConvNet [271]	1D-Subseq	Graph	0.636 (0.052)	0.797 (0.032)	0.685 (0.056)
DeepDTA [268]	1D-Subseq	1D	0.641 (0.054)	0.800 (0.034)	0.696 (0.058)
DeepCDA [272]	1D-Subseq	1D	0.620 (0.057)	0.786 (0.036)	0.708 (0.048)
HyperAttentionDTI [245]	1D-Subseq	1D	0.612 (0.075)	0.782 (0.049)	0.706 (0.053)
TransformerCPI [243]	1D-Subseq	1D	0.606 (0.072)	0.778 (0.046)	0.687 (0.092)
DTITR [350]	1D-Subseq	1D	0.604 (0.052)	0.777 (0.032)	0.710 (0.027)
Proposed Method					
TAG-DTA	1D-Subseq	1D	**0.600 (0.064)**	**0.775 (0.042)**	**0.715 (0.031)**
Novel Protein					
Baseline Methods					
GraphDTA-GCNNet [271]	1D-Subseq	Graph	0.588 (0.048)	0.766 (0.031)	0.744 (0.010)
GraphDTA-GATNet [271]	1D-Subseq	Graph	0.652 (0.021)	0.807 (0.013)	0.733 (0.014)
GraphDTA-GAT-GCN [271]	1D-Subseq	Graph	0.579 (0.064)	0.759 (0.042)	0.739 (0.018)
GraphDTA-GINConvNet [271]	1D-Subseq	Graph	0.495 (0.031)	0.703 (0.022)	0.770 (0.011)
DeepDTA [268]	1D-Subseq	1D	0.399 (0.062)	0.630 (0.049)	0.807 (0.026)
DeepCDA [272]	1D-Subseq	1D	0.396 (0.045)	0.628 (0.037)	0.815 (0.015)
HyperAttentionDTI [245]	1D-Subseq	1D	0.396 (0.057)	0.628 (0.045)	0.814 (0.020)
TransformerCPI [243]	1D-Subseq	1D	0.385 (0.046)	0.620 (0.037)	0.785 (0.033)
DTITR [350]	1D-Subseq	1D	0.380 (0.052)	0.615 (0.043)	0.827 (0.017)
Proposed Method					
TAG-DTA	1D-Subseq	1D	**0.372 (0.044)**	**0.549 (0.035)**	**0.833 (0.013)**
Novel Protein-Compound Pair					
Baseline Methods					
GraphDTA-GCNNet [271]	1D-Subseq	Graph	0.815 (0.129)	0.900 (0.073)	0.575 (0.066)
GraphDTA-GATNet [271]	1D-Subseq	Graph	0.788 (0.101)	0.886 (0.057)	0.573 (0.104)
GraphDTA-GAT-GCN [271]	1D-Subseq	Graph	0.791 (0.107)	0.887 (0.062)	0.598 (0.072)
GraphDTA-GINConvNet [271]	1D-Subseq	Graph	0.698 (0.087)	0.834 (0.053)	0.674 (0.079)
DeepDTA [268]	1D-Subseq	1D	0.630 (0.058)	0.793 (0.036)	0.664 (0.053)
DeepCDA [272]	1D-Subseq	1D	0.627 (0.053)	0.791 (0.033)	0.665 (0.053)
HyperAttentionDTI [245]	1D-Subseq	1D	0.620 (0.070)	0.787 (0.046)	0.667 (0.055)
TransformerCPI [243]	1D-Subseq	1D	0.615 (0.073)	0.784 (0.050)	0.652 (0.082)
DTITR [350]	1D-Subseq	1D	0.612 (0.085)	0.782 (0.051)	0.665 (0.062)
Proposed Method					
TAG-DTA	1D-Subseq	1D	**0.609 (0.057)**	**0.780 (0.036)**	**0.670 (0.058)**

The standard deviations are given in parentheses.
Bold indicates the best performance value associated with each evaluating metric.

Table 8.5: Binding affinity and 1D binding pocket prediction results over the Davis and COACH test set, respectively, for the different alternatives of the TAG-DTA model: (I) TAG-DTA without pre-training the SMILES Transformer-Encoder related block; (II) TAG-DTA without the binding affinity regression block; (III) TAG-DTA without the 1D binding pocket classification block.

1D Binding Pocket Prediction							
Method	**Protein Rep.**	**Compound Rep.**	**↑ Balanced Accuracy**	**↑ Recall**	**↑ Precision**	**↑ F1-Score**	**↑ MCC**
TAG-DTA - I	1D-Subseq	1D	85.99	79.25	67.80	72.93	0.676
TAG-DTA - II	1D-Subseq	1D	**87.99**	**81.45**	**73.70**	**77.38**	**0.730**
TAG-DTA	1D-Subseq	1D	87.18	81.26	68.82	74.52	0.696
Binding Affinity Prediction							
Method	**Protein Rep.**	**Compound Rep.**	**↓ MSE**	**↓ RMSE**	**↑ CI**	**↑ r^2**	**↑ Spearman**
TAG-DTA - I	1D-Subseq	1D	0.190	0.436	0.912	0.773	0.721
TAG-DTA - III	1D-Subseq	1D	0.199	0.446	0.906	0.763	0.713
TAG-DTA	1D-Subseq	1D	**0.185**	**0.430**	**0.917**	**0.780**	**0.729**

Bold indicates the best performance value associated with each evaluating metric.

precision - 67.80 %, F1-score - 72.93 %, and MCC - 0.676). These results show that pre-training the SMILES Transformer-Encoder using an MLM approach to learn the context and semantics of the chemical domains improves the overall learning capacity of the TAG-DTA architecture and increases the discriminating power and robustness of the aggregate representation of the SMILES strings, which is not only used to interact with the protein tokens (and condition their representation) but also for the prediction of binding affinity.

Regarding the prediction efficiency of the TAG-DTA model without the binding affinity regression block, i.e., to exclusively predict the 1D binding pocket, the performance obtained over the COACH test dataset is slightly superior in terms of balanced accuracy (87.99 %), recall (81.45 %), precision (73.70 %), F1-score (77.38 %), and MCC (0.730) when compared to the TAG-DTA model with the dual nature in the inferring process (balanced accuracy - 87.18 %, recall - 81.26 %, precision - 68.82 %, F1-score - 74.52 %, and MCC - 0.696). These results suggest that using TAG-DTA to simultaneously predict the 1D binding pocket and binding affinity of DTIs reduces the learning capacity of the TAG-DTA to predict the 1D binding pocket, which is expected considering that predicting the 1D binding pocket is computationally complex due to the imbalanced nature of binding and non-binding positions and that these two prediction tasks are computationally different. Nevertheless, the small performance gap demonstrates that the proposed TAG-DTA architecture is capable of converging toward two different prediction tasks, and the overall good performance in the prediction of the 1D binding pocket indicates the

viability of the TAG-DTA to learn and identify binding-related positions.

Additionally, the contribution of the prediction of the 1D binding pocket for the prediction of binding affinity was evaluated by training the TAG-DTA without the 1D binding pocket classification block. The prediction efficiency of the TAG-DTA architecture without the 1D binding pocket classification block resulted in significantly worse performance in terms of the MSE (0.199), RMSE (0.446), CI (0.906), r^2 (0.763), and Spearman (0.713) when compared to the TAG-DTA architecture with the 1D binding pocket classification block (MSE - 0.185, RMSE - 0.430, CI - 0.917, r^2 - 0.780, and Spearman - 0.729). These results demonstrate that using the predicted 1D binding pocket to guide and condition the attention mechanism of the Transformer-Encoder of the binding affinity regression block improves the discriminating power of the final aggregate representation hidden states associated with the interaction space and the capacity of this block to learn the pharmacological context information associated with the interaction space. Moreover, it indicates that learning the pharmacological space based on the short and long-term dependencies between the compound and the binding region (and the intra-associations within the binding region) leads to improved performance, which is in agreement with the fact that protein binding pockets are crucial in the interaction mechanism involved in DTIs. Furthermore, the superior binding affinity prediction performance of the proposed TAG-DTA architecture demonstrates that the 1D binding pocket and binding affinity prediction models are contextually related and that the model is capable of converging toward two different prediction tasks.

Overall, the use of an end-to-end binding-region-guided Transformer-based architecture, which simultaneously predicts the 1D binding pocket and binding strength of DTIs, demonstrates that actively integrating information about binding sites in the training process is crucial to properly learning the pharmacological space of the interaction and leads to improved binding affinity prediction performance. Additionally, it shows the capacity of the attention mechanisms of the Transformer-Encoders to learn the inter-dependencies between the compound and the protein tokens for the prediction of the 1D binding pocket, and the ability of the Transformer-Encoders to learn robust and discriminating aggregate representations of the proteins, compounds, and pharmacological space for the prediction of binding affinity.

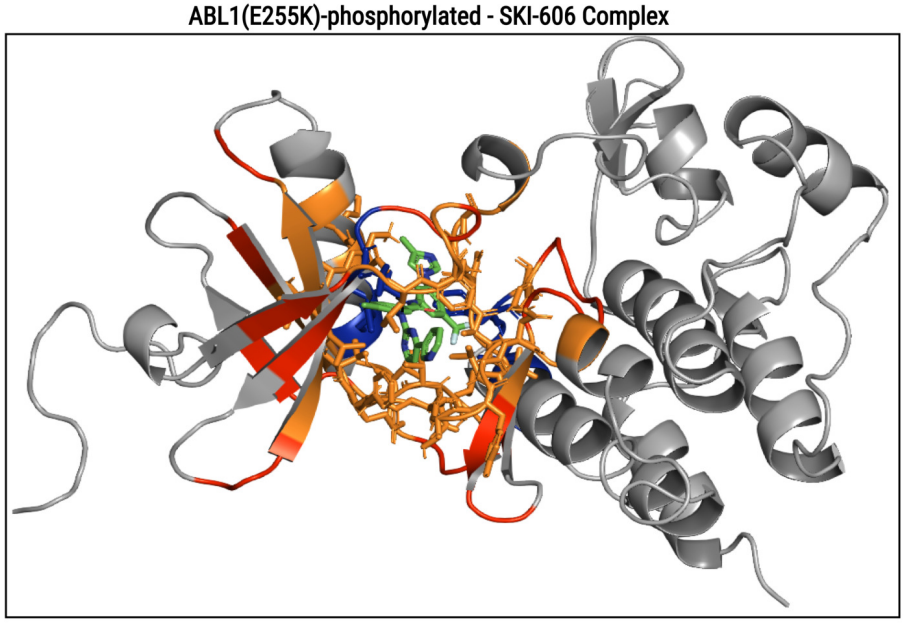
8.3.2 DTI and Model Understanding

In spite of the increasing performance of the models to correctly predict the binding strength or binary association of DTIs, it is crucial to understand the model

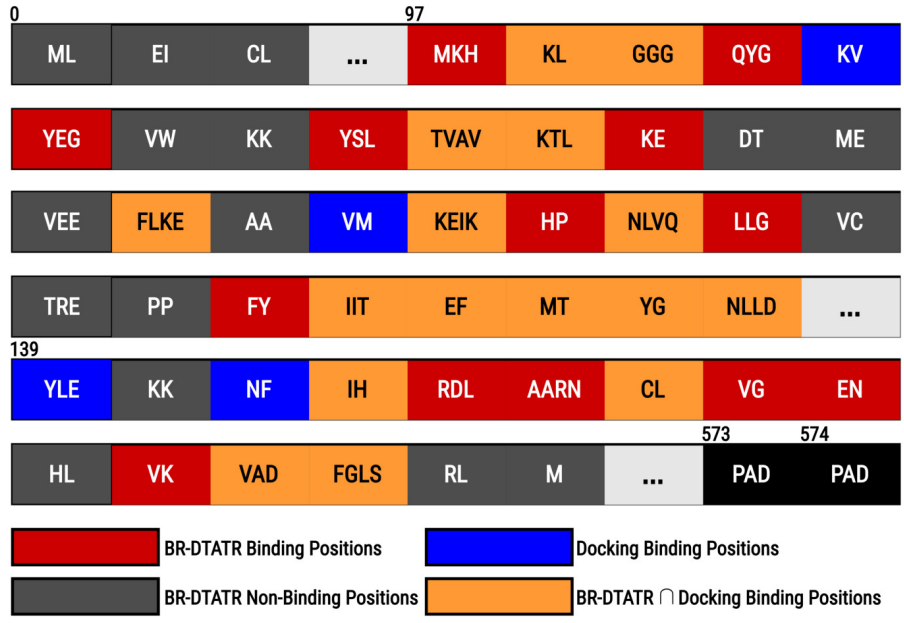
prediction and the overall importance of the input intra-associations and inter-dependencies to the model in the context of drug discovery. Moreover, given that DTIs revolve around specific substructures between the binding components and are supported by individual substructures within each interacting element, providing potential model understanding may lead to significant findings in the DTI domain. Considering the nature of the attention blocks, which give information about the overall importance of the input components (and their associations) to the model, the TAG-DTA architecture provides four different levels of attention: (I) protein sequences self-attention; (II) SMILES strings self-attention; (III) 1D binding pocket condition-based self-attention; and (IV) binding-region-guided conditioned-based self-attention. The first and second levels of attention provide information about the overall importance of the individual units (substructures) and intra-associations of the protein sequences (proteomics context) and SMILES strings (chemical context), respectively. The third level of attention gives information about the inter-associations between the compound representation and protein tokens (and their intra-associations) for the prediction of the 1D binding pocket, i.e., how the chemical information affects the overall importance of the individual units and intra-associations of protein sequences for the prediction of the binding pocket. The fourth level of attention provides information about the inter-associations between the compound representation and binding-related protein tokens, and the inter-associations between the binding tokens, i.e., it reflects the interaction and selectivity to the ligand based on the binding pocket.

The visualization and analysis of the different levels of attention provide a reasonable model and DTI understanding, however, they do not show explicit evidence of potential key residues within the protein sequences for the binding process, which is essential to promote the identification and selection of potential leads and understanding of the biological functions of proteins and mechanisms involved in DTIs. On that account, the proposed TAG-DTA may provide increasing DTI and model understanding considering that it predicts the 1D binding pocket of the DTIs, which also conditions the prediction of the binding affinity. In order to further validate the reliability of the TAG-DTA in the prediction of the 1D binding pocket and that it is capable of providing reasonable evidence for understanding the model prediction, the ABL1(E255K)-phosphorylated - SKI-606 DTI pair, which does not have the 3D interaction space available or annotated, was explored. Potential binding positions ($\leq$ 5 Å) were selected based on the research study by Monteiro et al. (2022) [341], where the 3D interaction space was explored and thoroughly assessed using guided docking, and compared with the predicted 1D binding pocket from the TAG-DTA

architecture. Figure 8.6 depicts the 3D receptor-ligand complex and the TAG-DTA 1D binding pocket for the ABL1(E255K)-phosphorylated - SKI-606 DTI pair.



(a)



(b)

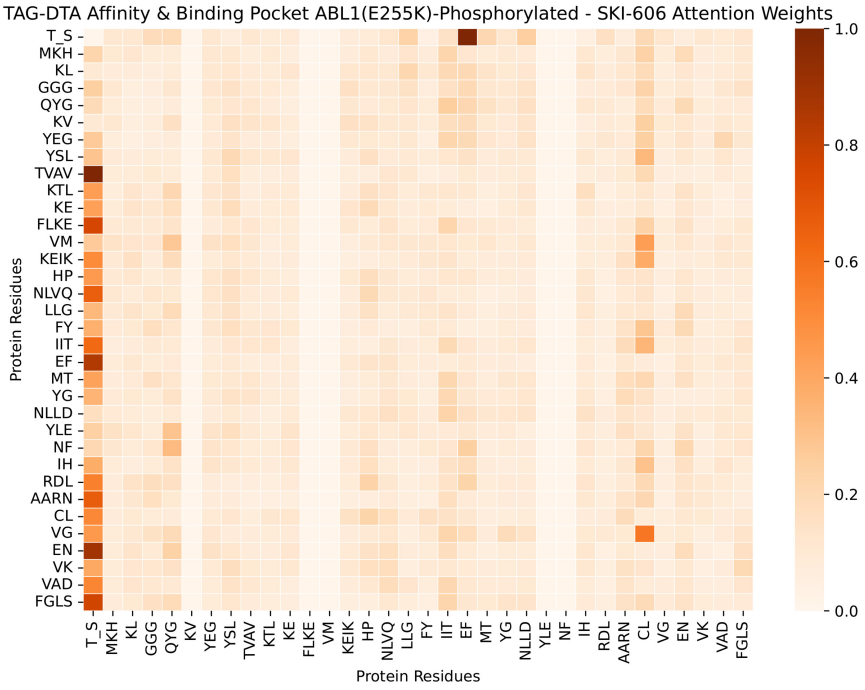
Figure 8.6: SKI-606 in complex with ABL1(E255K)-phosphorylated. **(a)** Annotated 3D complex obtained from docking [341]. **(b)** TAG-DTA 1D binding pocket. The docking binding sites ($\leq$ 5), TAG-DTA binding positions, TAG-DTA non-binding positions, and matched binding positions are represented by the blue, red, gray, and orange colors, respectively.

These visual results show that the TAG-DTA 1D binding pocket identifies almost

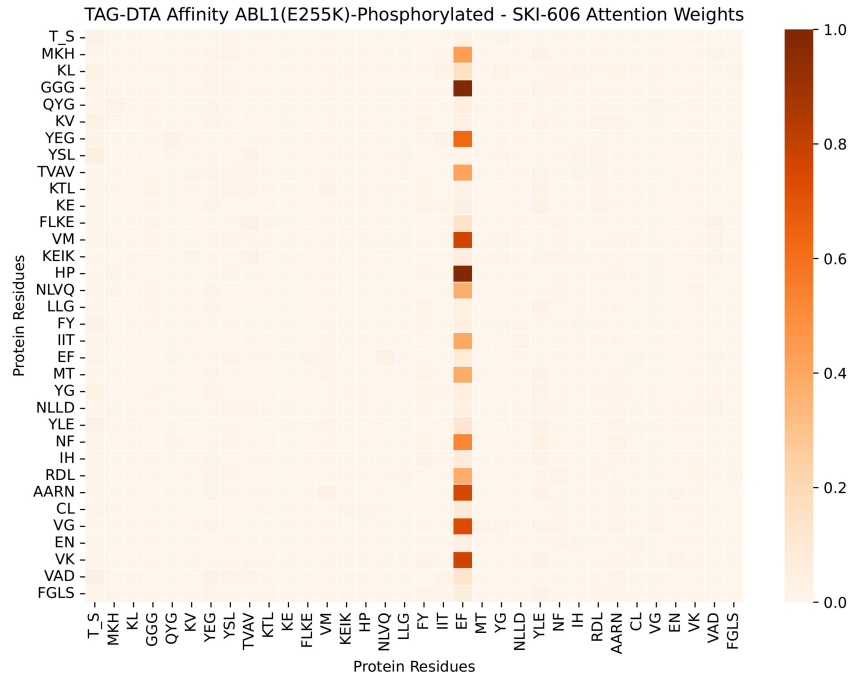
all the docking binding hits ($\leq$ 5 Å), which further validates the capacity of the TAG-DTA architecture to learn and identify binding-related positions. Moreover, the other TAG-DTA binding hits are the in the neighborhood of the docking binding pocket, where the spacial position of some of these hits (Figure 8.6a) suggests that they bear relation to conserved regions of the proteins or other potential interaction pockets/subpockets, e.g., some of these hits are near β-strands, which are usually important for the structure and function of the protein. Overall, these findings demonstrate the TAG-DTA is capable of providing increasing DTI and model understanding, and improves the validity of the learning stage of the pharmacological space based on binding-related positions.

In order to further validate the contribution of the 1D binding pocket block in the learning process of TAG-DTA, heat maps for the fourth level of attention were generated, specifically for the attention related to the inter-associations between the compound representation and binding-related protein tokens (and inter-associations between binding subwords). The ABL1(E255K)-phosphorylated - SKI-606 DTI pair was selected for examination, and visual analysis of the attention disparities across binding hits was conducted between TAG-DTA without the 1D binding pocket classification block and TAG-DTA with the dual nature incorporated into the inference process. Figure 8.7 illustrates the attention heat maps for ABL1(E255K)-phosphorylated - SKI-606 DTI pair across the two TAG-DTA configurations, where only subwords predicted as binding hits or related to docking binding hits were considered for visualization.

These visual observations indicate that TAG-DTA without the 1D binding pocket classification block (Figure 8.7b) is not attending, i.e., assigning significance, to the majority of protein subwords associated with binding. Furthermore, it fails to learn the inter-associations between the compound representation (T_S) and binding-related protein tokens. Notably, the T_S token does not direct its attention to any binding-related tokens, except for the EF motif. Consequently, this architecture configuration introduces bias into the predictions, resulting in the estimation of potential DTIs based on redundant sites. Conversely, TAG-DTA (Figure 8.7a), which incorporates a dual nature in the inference process, assigns significance to all binding hits (except for certain docking-related tokens) and comprehensively learns the pharmacological space of the interaction based on the inter-associations amongst binding-related substructures. Furthermore, TAG-DTA demonstrates a lower absolute prediction error for DTA prediction in comparison to TAG-DTA without the 1D binding pocket classification block. In the particular case of ABL1(E255K)-phosphorylated - SKI-606, the pK_d values are 10.33 for the experiment, 10.23 for



(a)



(b)

Figure 8.7: Attention maps associated with the binding-region-guided Transformer-Encoder for the SKI-606 in complex with ABL1(E255K)-phosphorylated. a) TAG-DTA. b) TAG-DTA without the 1D binding pocket classification block. The attention weights were normalized across all the positions for each head of attention and the maximum value was selected for visualization.

TAG-DTA, and 10.04 for TAG-DTA without the 1D binding pocket classification block. These findings underscore the critical importance of actively integrating information about binding sites throughout the training process to effectively learn the pharmacological space of the interaction and to enhance the predictive performance of DTA. Moreover, they highlight the capability of TAG-DTA to provide reasonable evidence for understanding model predictions.

8.4 Conclusion

8.4.1 Final Remarks

In this research study, an end-to-end binding-region-guided Transformed-based architecture (TAG-DTA) is proposed to simultaneously predict the 1D binding pocket and the binding affinity in terms of the logarithmic-transformed dissociation constant (pK_d) of DTI pairs, where the prediction of the 1D binding vector conditions the prediction of DTA. The architecture comprises two models, specifically a 1D binding pocket classifier and a binding affinity regressor, and shares three core layers, including lower Transformer-Encoders and a condition-based concatenation block. The prediction of the 1D binding pocket conditions the attention mechanism of the binding affinity Transformer-Encoder, resulting in the exchange of information between the proteomics and chemical domains over binding-related residues. To perform the experiments, DTIs with binding information annotated were collected from various binding-related databases, and DTIs with binding affinity available were extracted from the Davis kinase binding affinity dataset. The performance of the proposed TAG-DTA model was compared with different state-of-the-art baselines and in different experimental setups based on unknown subsets of the proteomics and chemical representation spaces.

The proposed model yielded better results than state-of-the-art baselines for the prediction of binding affinity, resulting in lower MSE and RMSE values and higher CI and r^2 scores in the Davis independent test dataset, and lower MSE and RMSE values and higher CI, r^2, and Spearman scores in the original Davis split folds. These results demonstrate the model's ability to accurately predict the value of binding strength and to properly assess the target selectivity (distinguish the rank order of binding strength between the DTI pairs). In the novel compounds, novel proteins, and novel protein-compounds pairs experimental settings, TAG-DTA achieved lower values of MSE and RMSE, and higher CI scores when compared to the baselines, demonstrating a superior capacity to generalize toward unknown subsets of

the proteomics and chemical representation spaces. Furthermore, the TAG-DTA architecture is shown to efficiently learn the proteomics and chemical context of the proteins and compounds, respectively, and that learning the pharmacological space based on the short and long-term dependencies between the compound and the binding region (and the intra-associations within the binding region) leads to improved DTA prediction performance.

The influence of different blocks on the prediction efficiency of the TAG-DTA architecture was also explored. It was found that pre-training the SMILES Transformer-Encoder using an MLM approach resulted in overall better performance in the prediction of the binding affinity and 1D binding pocket as it increases the discriminating power and robustness of the aggregate representation of the SMILES string, i.e., it improves the learning capacity of this block to capture chemical context. In addition, it was further demonstrated that conditioning the attention mechanism of the binding affinity Transformer-Encoder based on the predicted 1D binding pocket leads to significantly improved DTA prediction performance. Moreover, the results validated that combining computationally different yet contextually related tasks (1D binding pocket classification and binding affinity regression) is crucial for the DTI domain representation and DTA prediction performance.

TAG-DTA provides different levels of potential DTI and prediction understanding due to the nature of the attention layers, which give information about the overall importance of the input components (and their associations) to the model. Moreover, it showed increasing model understanding due to the dual nature of the prediction process, specifically due to the prediction of the 1D binding pocket, which conditions the prediction of the binding affinity and presents explicit evidence of potential key regions within the protein sequences, including in DTI pairs without the 3D complex annotated.

The major contribution of this study is an efficient and novel end-to-end binding-region-guided Transformer-based architecture capable of simultaneously predicting the 1D binding pocket and binding affinity of DTI pairs, where binding information is actively integrated into the training process. Moreover, it models the interdependency of the proteomics, chemical, and binding-region-related pharmacological spaces, and provides increased DTI and model understanding due to the nature of the attention blocks and prediction of the 1D binding pocket.

8.4.2 Study Limitations and Future Work

The pre-training step of the SMILES Transformer-Encoder based on the MLM approach improved the prediction performance and increased the robustness and discriminating power of the aggregate representation of the SMILES strings. These results corroborate that pre-training LLMs with a large corpus associated with the input domain leads to improved performance and learning capacity. On that account, extending the pre-training stage to the protein sequences can result in superior prediction performance, improve the robustness of the aggregate representation of the protein sequences, and reduce the effects of the proteomics domain representability in the learning process of TAG-DTA, which is of special interest considering the dual nature associated with the training and inferring stages of this architecture. Furthermore, pre-training LLMs, specifically Transformer-Encoders, greatly decreases the training time required to converge in the downstream tasks. Nevertheless, pre-training the Transformer-Encoder associated with the protein sequences is computationally complex and requires considerable resources due to the computational complexity of $O(n^2)$ concerning the sequence length in the attention layers.

The unequivocal complexity surrounding the binding process between biologically relevant targets and active small compounds, and the importance of the complementarity of certain functional groups in the 3D space have led to a plethora of studies in the realm of binding pocket prediction. In that regard, the state-of-the-art in binding pocket prediction focuses on using 3D information to determine various potential pocket surfaces within the protein structures. Moreover, the 3D conformation and flexibility of the protein sequences and the 3D orientation and arrangement of the organic functional groups of the compounds are crucial for the binding process. Thus, exploring different levels of structural information and integrating multiple representations of the proteins and compounds can lead to improved performance and increased reliability in the learning process of the proteomics, chemical, and pharmacological spaces. Furthermore, the stereoelectronic structure of the proteins and compounds plays an important role in the binding selectivity and overall reactivity, hence, incorporating additional intricate terms associated with the proteomics and chemical domains into the learning process might lead to interesting findings and superior validity in the inferring process.

TAG-DTA demonstrated that learning the pharmacological space based on the short and long-term dependencies between the compound and the binding region (and the intra-associations within the binding region) leads to superior DTA prediction per-

formance. Moreover, TAG-DTA provides increasing DTI and model understanding due to the prediction of the 1D binding pocket. However, stereochemical information is not included in the representation of the compounds (canonical SMILES), and the pharmacophore and relevant secondary functional groups associated with the compounds are not actively integrated into the learning process of the architecture. Hence, TAG-DTA is limited in its capacity to model the pharmacokinetic, pharmacodynamic, and physicochemical differences across constitutional isomers or stereoisomers (enantiomers or diastereomers) of certain chemical compounds, and the overall polypharmacological nature associated with most active small molecules. Additionally, the identification of relevant components and substructures within the compound space is important to increase the reliability of the predictions and to promote the following steps of the drug discovery pipeline, i.e., the lead optimization stage. In that regard, extending the TAG-DTA to actively model the interdependencies between the pharmacophore of the compounds and the binding region within the proteins can increase the robustness and validity of the results and lead to new insights in the DTI domain.

Chapter 9

Conclusions

This chapter highlights the primary contributions and offers a comprehensive overview of the conducted research. Furthermore, it discusses and delves into future research directions.

9.1 Overview of the Main Contributions

This thesis presents a comprehensive study on novel and explainable DL-based solutions for predicting Drug–Target Affinity (DTA) using 1D raw sequential and structural representations of the proteins and compounds. These solutions are capable of providing potential evidence to support the rationale behind the predictions and shed light on the mechanisms involved in the interaction between active compounds and biologically relevant targets. Furthermore, this research progressively addresses some of the primary challenges that persist in the drug discovery domain, including the multi-domain representation space of DTIs and the importance to proficiently modeling the pharmacological space based on information concerning binding pockets.

In Chapter 6, a post-hoc explainability algorithm was proposed and explored to provide potential explanations for the decision-making process of CNNs in the context of DTA prediction. The deep representations extracted by the CNNs were shown to be efficient and discriminating for the prediction of binding affinity. Moreover, CNNs were found to identify and extract features from regions relevant for the interaction, specifically binding sites and evolutionarily conserved motifs, without any *a priori* information during the learning stage. The weight associated with these spots was also in the range of those with the highest positive influence.

In Chapter 7, the multi-domain inter-dependency associated with DTIs was addressed by combining self and cross-attention mechanisms to learn the proteomics, chemical, and pharmacological contexts. Multiple Transformer-Encoders were

stacked into a novel end-to-end Transformer-based architecture (DTITR) for predicting DTA. The results demonstrated that combining the proteomics, chemical, and pharmacological contexts improves the prediction efficiency compared to using only the individual proteomics and chemical context information of the proteins and compounds. Furthermore, the architecture is capable of providing different levels of potential DTI and prediction understanding due to the nature of the attention mechanisms. In that regard, the compounds were shown to be attending to binding-related residues.

In Chapter 8, a binding-region-guided strategy was proposed to model the pharmacological space of the interaction and learn the inter-dependency amongst binding-related positions. Additionally, two computationally different yet contextually related tasks, specifically 1D binding pocket classification and binding affinity regression, were combined into an end-to-end Transformer-based framework (TAG-DTA). The results demonstrated that actively integrating information concerning binding pockets during the learning stage of TAG-DTA leads to significantly improved DTA prediction performance. Moreover, this framework presents explicit evidence of potential key regions within the protein sequence for the prediction of binding affinity due to the dual nature of the inferring process.

Overall, endeavors were undertaken to design computational solutions with the aim of enhancing the drug discovery process chain and focusing on tackling pressing challenges related to the development of prospective applications for predicting DTI or DTA. Furthermore, the reported findings bridge the gap between ML/DL and domain knowledge, and highlight possible paths to incorporating these solutions into existing and stacked drug discovery and development pipelines.

9.2 Future Research Directions

In spite of the considerable and consistently increasing investments in drug design and development, numerous opportunities for expansion endure. The investigation undertaken in this study centers on certain cornerstones of drug design and development. Nevertheless, the unequivocal complexity inherent in the comprehensive binding process between biologically relevant targets and active compounds underscores the multi-domain and multi-objective nature of drug discovery. Consequently, future research endeavors encompass a range of pertinent aspects.

In light of the polypharmacological nature commonly attributed to most existing compounds and the potential for synergistic effects, it is imperative to incorporate

additional intricate properties associated with the chemical domain into the learning process for accurate compound characterization [15, 356]. Furthermore, the 3D orientation and arrangement of the organic functional groups significantly influence the pharmacokinetic, pharmacodynamic, and physicochemical properties of active small molecules [95, 96]. Hence, exploring various levels of structural information and integrating multiple representations of compounds to accurately depict their stereoelectronic structure, which plays a pivotal role in binding selectivity and overall reactivity, may yield valuable insights and enhance the predictive validity of the computational solutions reported in this study.

The identification of relevant components and substructures within the compound space is indispensable for facilitating various phases of the drug discovery pipeline, such as lead discovery and lead optimization. Modern drug discovery sources prioritize the application of *de novo* drug design based on *in silico* techniques to generate novel and synthesizable small compounds endowed with desired pharmacological properties and heightened selectivity for biologically relevant targets [357]. On that account, extending the computational frameworks put forth and investigated in this thesis to actively model the inter-dependencies between the organic functional groups of compounds and the binding regions within the proteins can effectively narrow the search space and reduce the requirement for numerous systematic modifications and refinements of the structure of lead compounds. Furthermore, it can augment the robustness and validity of the findings and further amplify the explainability and informativeness of the ensuing computational frameworks.

The majority of proteins inherently manifest dynamic behavior and substantial conformational flexibility, undergoing transitions between various conformational states or substates that maintain comparable energy levels [136, 138]. Moreover, numerous proteins experience conformational changes to facilitate their interaction with specific ligands. In this context, the characteristics and positioning of the binding pockets within the proteins also exert influence on the binding dynamics [150, 144]. Therefore, it is imperative to delve into 3D multi-instance learning to accurately characterize proteins (including their binding pockets) with respect to the range of possible conformations. Additionally, certain binding pockets are linked to transient states due to the stabilization of energy contributions [151], hence, the incorporation of 3D multi-instance learning may serve to facilitate the discovery and/or design of potential leads that selectively bind to these regions with high affinity.

Research within the realm of pharmacogenomics has revealed that the pharmacokinetics of various compounds are subject to the influence of genomic variations

[358, 359]. These genetic alterations result in distinct transcriptomic profiles and, consequently, modifications in proteomic profiles. Thus, it is relevant to incorporate considerations of individual variability, particularly genomic and proteomics profiles, to effectively identify potential lead compounds. This is particularly decisive for specific clinical conditions marked by a notable degree of heterogeneity [360, 361]. Moreover, in the era of precision medicine it is paramount to account for all individual variability in order to accurately design pharmacological strategies..

Regardless of the inherent challenges entailed in the pursuit of scientific research and the indispensable requirement for sufficient resources to facilitate these endeavors, it is of paramount importance to enhance the communication between ML/DL researchers and domain experts. This is essential to further validate the applicability of complex computational frameworks within crucial domains like drug discovery. Furthermore, active engagement with multidisciplinary experts assumes a vital role in achieving a deeper understanding of the challenges associated with this research area and designing explainable and informative computational solutions [31, 312].

Bibliography

[1] Hughes, J. P. and Rees, S. and Kalindjian, S. B. and Philpott, K. L., Principles of early drug discovery, British Journal of Pharmacology 162 (6) (2011) 1239–1249. doi:10.1111/j.1476-5381.2010.01127.x.

[2] Paul, Steven M. and Mytelka, Daniel S. and Dunwiddie, Christopher T. and Persinger, Charles C. and Munos, Bernard H. and Lindborg, Stacy R. and Schacht, Aaron L., How to improve R&D productivity: the pharmaceutical industry's grand challenge, Nature Reviews Drug Discovery 9 (2010) 203 EP –. doi:10.1038/nrd3078.

[3] Zhou, Hongyi and Gao, Mu and Skolnick, Jeffrey, Comprehensive prediction of drug-protein interactions and side effects for the human proteome, Scientific Reports 5 (2015) 11090 EP –. doi:10.1038/srep11090.

[4] Schneider, Petra and Walters, W. Patrick and Plowright, Alleyn T. and Sieroka, Norman and Listgarten, Jennifer and Goodnow, Robert A. and Fisher, Jasmin and Jansen, Johanna M. and Duca, José S. and Rush, Thomas S. and Zentgraf, Matthias and Hill, John Edward and Krutoholow, Elizabeth and Kohler, Matthias and Blaney, Jeff and Funatsu, Kimito and Luebkemann, Chris and Schneider, Gisbert, Rethinking drug design in the artificial intelligence era, Nature Reviews Drug Discovery 19 (5) (2020) 353–364. doi:10.1038/s41573-019-0050-3.

[5] Shameer, Khader and Dudley, Ben Readhead and Joel T., Computational and Experimental Advances in Drug Repositioning for Accelerated Therapeutic Stratification, Current Topics in Medicinal Chemistry 15 (1) (2015) 5–20. doi:10.2174/1568026615666150112103510.

[6] Gaulton, Anna and Hersey, Anne and Nowotka, Michał and Bento, A. Patrícia and Chambers, Jon and Mendez, David and

Mutowo, Prudence and Atkinson, Francis and Bellis, Louisa J. and Cibrián-Uhalte, Elena and Davies, Mark and Dedman, Nathan and Karlsson, Anneli and Magariños, María Paula and Overington, John P. and Papadatos, George and Smit, Ines and Leach, Andrew R., The ChEMBL database in 2017, Nucleic Acids Research 45 (D1) (2017) D945–D954. doi:10.1093/nar/gkw1074.

[7] Wang, Wenjing and Sun, Qiu, Novel targeted drugs approved by the NMPA and FDA in 2019, Signal Transduction and Targeted Therapy 5 (1) (2020) 65. doi:10.1038/s41392-020-0164-4.

[8] Boni, Maciej F. and Lemey, Philippe and Jiang, Xiaowei and Lam, Tommy Tsan-Yuk and Perry, Blair W. and Castoe, Todd A. and Rambaut, Andrew and Robertson, David L., Evolutionary origins of the SARS-CoV-2 sarbecovirus lineage responsible for the COVID-19 pandemic, Nature Microbiologydoi:10.1038/s41564-020-0771-4.

[9] Gould, Ian M. and Bal, Abhijit M., New antibiotic agents in the pipeline and how they can help overcome microbial resistance, Virulence 4 (2) (2013) 185–191. doi:10.4161/viru.22507.

[10] Sengupta, Saswati and Chattopadhyay, Madhab and Grossart, Hans-Peter, The multifaceted roles of antibiotics and antibiotic resistance in nature, Frontiers in Microbiology 4 (2013) 47. doi:10.3389/fmicb.2013.00047.

[11] Ventola, C. Lee, The antibiotic resistance crisis: part 1: causes and threats, P & T : a peer-reviewed journal for formulary management 40 (4) (2015) 277–283.

[12] Aslam, Bilal and Wang, Wei and Arshad, Muhammad Imran and Khurshid, Mohsin and Muzammil, Saima and Rasool, Muhammad Hidayat and Nisar, Muhammad Atif and Alvi, Ruman Farooq and Aslam, Muhammad Aamir and Qamar, Muhammad Usman and Salamat, Muhammad Khalid Farooq and Baloch, Zulqarnain, Antibiotic resistance: a rundown of a global crisis, Infect Drug Resist 11 (2018) 1645–1658. doi:10.2147/IDR.S173867.

[13] Hegemann, Julian D. and Birkelbach, Joy and Walesch, Sebastian and Müller, Rolf, Current developments in antibiotic discovery, EMBO reports 24 (1) (2023) e56184. doi:10.15252/embr.202256184.

[14] D'Souza, Sofia and Prema, K. V. and Balaji, Seetharaman, Machine learning models for drug–target interactions: current knowledge and future directions, Drug Discovery Today 25 (4) (2020) 748–756. doi:10.1016/j.drudis.2020.03.003.

[15] Kabir, Abbas and Muth, Aaron, Polypharmacology: The science of multi-targeting molecules, Pharmacological Research 176 (2022) 106055. doi:10.1016/j.phrs.2021.106055.

[16] Ezzat, Ali and Wu, Min and Li, Xiao-Li and Kwoh, Chee-Keong, Drug-target interaction prediction via class imbalance-aware ensemble learning, BMC Bioinformatics 17 (19) (2016) 509. doi:10.1186/s12859-016-1377-y.

[17] Tayebi, Aida and Yousefi, Niloofar and Yazdani-Jahromi, Mehdi and Kolanthai, Elayaraja and Neal, Craig J. and Seal, Sudipta and Garibay, Ozlem O., UnbiasedDTI: Mitigating Real-World Bias of Drug-Target Interaction Prediction by Using Deep Ensemble-Balanced Learning (2022). doi:10.3390/molecules27092980.

[18] Gilson, Michael K. and Liu, Tiqing and Baitaluk, Michael and Nicola, George and Hwang, Linda and Chong, Jenny, BindingDB in 2015: A public database for medicinal chemistry, computational chemistry and systems pharmacology, Nucleic Acids Research 44 (D1) (2016) D1045–D1053. doi:10.1093/nar/gkv1072.

[19] Wang, Renxiao and Fang, Xueliang and Lu, Yipin and Wang, Shaomeng, The PDBbind Database: Collection of Binding Affinities for Protein-Ligand Complexes with Known Three-Dimensional Structures, Journal of Medicinal Chemistry 47 (12) (2004) 2977–2980. doi:10.1021/jm0305801.

[20] Thafar, Maha and Raies, Arwa Bin and Albaradei, Somayah and Essack, Magbubah and Bajic, Vladimir B., Comparison Study of Computational Prediction Tools for Drug-Target Binding Affinities, Frontiers in Chemistry 7. doi:10.3389/fchem.2019.00782.

[21] Burlingham, Benjamin T. and Widlanski, Theodore S., An Intuitive Look at the Relationship of Ki and IC50: A More General Use for the Dixon Plot, Journal of Chemical Education 80 (2) (2003) 214. doi:10.1021/ed080p214.

[22] Bachmann, Kenneth A. and Lewis, Jeffrey D., Predicting Inhibitory Drug—Drug Interactions and Evaluating Drug Interaction Reports Using Inhibition Constants, Annals of Pharmacotherapy 39 (6) (2005) 1064–1072. doi:10.1345/aph.1E508.

[23] Hulme, Edward C. and Trevethick, Mike A., Ligand binding assays at equilibrium: validation and interpretation, British Journal of Pharmacology 161 (6) (2010) 1219–1237. doi:10.1111/j.1476-5381.2009.00604.x.

[24] Du, Xing and Li, Yi and Xia, Yuan-Ling and Ai, Shi-Meng and Liang, Jing and Sang, Peng and Ji, Xing-Lai and Liu, Shu-Qun, Insights into Protein–Ligand Interactions: Mechanisms, Models, and Methods (2016). doi:10.3390/ijms17020144.

[25] Ma, Weina and Yang, Liu and He, Langchong, Overview of the detection methods for equilibrium dissociation constant KD of drug-receptor interaction, Journal of Pharmaceutical Analysis 8 (3) (2018) 147–152. doi:10.1016/j.jpha.2018.05.001.

[26] Ravì, D. and Wong, C. and Deligianni, F. and Berthelot, M. and Andreu-Perez, J. and Lo, B. and Yang, G.-Z., Deep Learning for Health Informatics, IEEE Journal of Biomedical and Health Informatics 21 (1) (2017) 4–21. doi:10.1109/JBHI.2016.2636665.

[27] Rifaioglu, Ahmet Sureyya and Atas, Heval and Martin, Maria Jesus and Cetin-Atalay, Rengul and Atalay, Volkan and Doğan, Tunca, Recent applications of deep learning and machine intelligence on in silico drug discovery: methods, tools and databases, Briefings in Bioinformatics 20 (5) (2018) 1878–1912. doi:10.1093/bib/bby061.

[28] London, Alex John, Artificial Intelligence and Black-Box Medical Decisions: Accuracy versus Explainability, Hastings Center Report 49 (1) (2019) 15–21. doi:10.1002/hast.973.

[29] Castelvecchi, Davide, Can we open the black box of AI?, Nature 538 (2016) 20–23. doi:10.1038/538020a.

[30] Preuer, Kristina and Klambauer, Günter and Rippmann, Friedrich and Hochreiter, Sepp and Unterthiner, Thomas, Interpretable Deep Learning in Drug Discovery, Springer International Publishing, Cham, 2019, pp. 331–345. doi:10.1007/978-3-030-28954-6_18.

[31] Jiménez-Luna, José and Grisoni, Francesca and Schneider, Gisbert, Drug discovery with explainable artificial intelligence, Nature Machine Intelligence 2 (10) (2020) 573–584. doi:10.1038/s42256-020-00236-4.

[32] Finale Doshi-Velez and Been Kim, Towards A Rigorous Science of Interpretable Machine Learning (2017). arXiv:1702.08608v2.

[33] Krasner, Joseph, Drug-Protein Interaction, Pediatric Clinics of North America 19 (1) (1972) 51–63. doi:10.1016/S0031-3955(16)32666-9.

[34] Guo, Fei and Wang, Lusheng, Computing the protein binding sites, BMC bioinformatics 13 Suppl 10 (Suppl 10) (2012) S2–S2. doi:10.1186/1471-2105-13-S10-S2.

[35] Cheng, Tiejun and Hao, Ming and Takeda, Takako and Bryant, Stephen H. and Wang, Yanli, Large-Scale Prediction of Drug-Target Interaction: a Data-Centric Review, The AAPS Journal 19 (5) (2017) 1264–1275. doi:10.1208/s12248-017-0092-6.

[36] Agamah, Francis E. and Mazandu, Gaston K. and Hassan, Radia and Bope, Christian D. and Thomford, Nicholas E. and Ghansah, Anita and Chimusa, Emile R., Computational/in silico methods in drug target and lead prediction, Briefings in Bioinformatics 21 (5) (2020) 1663–1675. doi:10.1093/bib/bbz103.

[37] Hanson, Jack and Paliwal, Kuldip K. and Litfin, Thomas and Yang, Yuedong and Zhou, Yaoqi, Getting to Know Your Neighbor: Protein Structure Prediction Comes of Age with Contextual Machine Learning, Journal of Computational Biology 27 (5) (2020) 796–814. doi:10.1089/cmb.2019.0193.

[38] Konc, Janez and Janežič, Dušanka, Binding site comparison for function prediction and pharmaceutical discovery, Current Opinion in Structural Biology 25 (2014) 34–39. doi:10.1016/j.sbi.2013.11.012.

[39] Paricharak, Shardul and Cortés-Ciriano, Isidro and IJzerman, Adriaan P. and Malliavin, Thérèse E. and Bender, Andreas, Proteochemometric modelling coupled to in silico target prediction: an integrated approach for the simultaneous prediction of polypharmacology and binding affinity/potency of small molecules, Journal of Cheminformatics 7 (1) (2015) 15. doi:10.1186/s13321-015-0063-9.

[40] Rasti, Behnam and Karimi-Jafari, Mohammad H. and Ghasemi, Jahan B., Quantitative Characterization of the Interaction Space of the Mammalian Carbonic Anhydrase Isoforms I, II, VII, IX, XII, and XIV and their Inhibitors, Using the Proteochemometric Approach, Chemical Biology & Drug Design 88 (3) (2016) 341–353. doi:10.1111/cbdd.12759.

[41] Christmann-Franck, Serge and van Westen, Gerard J. P. and Papadatos, George and Beltran Escudie, Fanny and Roberts, Alexander and Overington, John P. and Domine, Daniel, Unprecedently Large-Scale Kinase Inhibitor Set Enabling the Accurate Prediction of Compound–Kinase Activities: A Way toward Selective Promiscuity by Design?, Journal of Chemical Information and Modeling 56 (9) (2016) 1654–1675. doi:10.1021/acs.jcim.6b00122.

[42] Pérot, Stéphanie and Sperandio, Olivier and Miteva, Maria A. and Camproux, Anne-Claude and Villoutreix, Bruno O., Druggable pockets and binding site centric chemical space: a paradigm shift in drug discovery, Drug Discovery Today 15 (15) (2010) 656–667. doi:10.1016/j.drudis.2010.05.015.

[43] Zheng, Xiliang and Gan, LinFeng and Wang, Erkang and Wang, Jin, Pocket-Based Drug Design: Exploring Pocket Space, The AAPS Journal 15 (1) (2013) 228–241. doi:10.1208/s12248-012-9426-6.

[44] Tibaut, T. and Borišek, J. and Novič, M. and Turk, D., Comparison of in silico tools for binding site prediction applied for structure-based design of autolysin inhibitors, SAR and QSAR in Environmental Research 27 (7) (2016) 573–587. doi:10.1080/1062936X.2016.1217271.

[45] Capra, John A. and Singh, Mona, Predicting functionally important residues from sequence conservation, Bioinformatics 23 (15) (2007) 1875–1882. doi:10.1093/bioinformatics/btm270.

[46] Nelson, David L. and Cox, Michael, Lehninger Principles of Biochemistry. 5th ed, W. H. Freeman and Company, W. H. Freeman and Company, 2008.

[47] Jez, Joseph M., Revisiting protein structure, function, and evolution in the genomic era, Journal of Invertebrate Pathology 142 (2017) 11–15. doi:10.1016/j.jip.2016.07.013.

[48] Talley, Kemper and Alexov, Emil, On the pH-optimum of activity and stability of proteins, Proteins: Structure, Function, and Bioinformatics 78 (12) (2010) 2699–2706. doi:10.1002/prot.22786.

[49] Studer, Romain and Dessailly, Benoit and Orengo, Christine, Residue mutations and their impact on protein structure and function: Detecting beneficial and pathogenic changes, The Biochemical journal 449 (2013) 581–94. doi:10.1042/BJ20121221.

[50] Bhattacharya, Roshni and Rose, Peter W. and Burley, Stephen K. and Prlic, Andreas, Impact of genetic variation on three dimensional structure and function of proteins, PLOS ONE 12 (3) (2017) e0171355. doi:10.1371/journal.pone.0171355.

[51] Laurencikiene, Jurga and Källman, Annika M. and Fong, Nova and Bentley, David L. and Öhman, Marie, RNA editing and alternative splicing: the importance of co-transcriptional coordination, EMBO reports 7 (3) (2006) 303–307. doi:10.1038/sj.embor.7400621.

[52] Jurado, Ashley R. and Tan, Dazhi and Jiao, Xinfu and Kiledjian, Megerditch and Tong, Liang, Structure and Function of Pre-mRNA 5'-End Capping Quality Control and 3'-End Processing, Biochemistry 53 (12) (2014) 1882–1898. doi:10.1021/bi401715v.

[53] Mignone, Flavio and Gissi, Carmela and Liuni, Sabino and Pesole, Graziano, Untranslated regions of mRNAs, Genome Biology 3 (3) (2002) reviews0004.1. doi:10.1186/gb-2002-3-3-reviews0004.

[54] Ramazi, Shahin and Zahiri, Javad, Post-translational modifications in proteins: resources, tools and prediction methods, Database 2021 (2021) baab012. doi:10.1093/database/baab012.

[55] Shen, Juwen and Zhang, Jian and Luo, Xiaomin and Zhu, Weiliang and Yu, Kunqian and Chen, Kaixian and Li, Yixue and Jiang, Hualiang, Predicting protein–protein interactions based only on sequences information, Proceedings of the National Academy of Sciences 104 (11) (2007) 4337–4341. doi:10.1073/pnas.0607879104.

[56] Yu, Chi-Yuan and Chou, Lih-Ching and Chang, Darby Tien-Hao, Predicting protein-protein interactions in unbalanced data using the primary structure of proteins, BMC Bioinformatics 11 (1) (2010) 167. doi:10.1186/1471-2105-11-167.

[57] Consortium, The UniProt, UniProt: the Universal Protein Knowledgebase in 2023, Nucleic Acids Research 51 (D1) (2023) D523–D531. doi:10.1093/nar/gkac1052.

[58] Kabsch, Wolfgang and Sander, Christian, Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features, Biopolymers 22 (12) (1983) 2577–2637. doi:10.1002/bip.360221211.

[59] Andersen, Claus A. F. and Palmer, Arthur G. and Brunak, Søren and Rost, Burkhard, Continuum secondary structure captures protein flexibility, Structure 10 (2) (2002) 175–184. doi:10.1016/s0969-2126(02)00700-1.

[60] Cooley, Richard B. and Arp, Daniel J. and Karplus, P. Andrew, Evolutionary Origin of a Secondary Structure: π-Helices as Cryptic but Widespread Insertional Variations of α-Helices That Enhance Protein Functionality, Journal of Molecular Biology 404 (2) (2010) 232–246. doi:10.1016/j.jmb.2010.09.034.

[61] Lukin, Jonathan A. and Kontaxis, Georg and Simplaceanu, Virgil and Yuan, Yue and Bax, Ad and Ho, Chien, Quaternary structure of hemoglobin in solution, Proceedings of the National Academy of Sciences 100 (2) (2003) 517–520. doi:10.1073/pnas.232715799.

[62] Navarro, Gemma and Cordomí, Arnau and Zelman-Femiak, Monika and Brugarolas, Marc and Moreno, Estefania and Aguinaga, David and Perez-Benito, Laura and Cortés, Antoni and Casadó, Vicent and Mallol, Josefa and Canela, Enric I. and Lluís, Carme and Pardo, Leonardo and García-Sáez, Ana J. and McCormick, Peter J. and Franco, Rafael, Quaternary structure of a G-protein-coupled receptor heterotetramer in complex with Gi and Gs, BMC Biology 14 (1) (2016) 26. doi:10.1186/s12915-016-0247-4.

[63] Watson, H. C. and Kendrew, J.C., The stereochemistry of the protein myoglobin (2017). doi:10.2210/pdb1MBN/pdb.

[64] Paoli, Massimo and Liddington, Robert and Tame, Jeremy and Wilkinson, Anthony and Dodson, Guy, Oxy T State Haemoglobin - Oxygen bound at all four haems (1996). doi:10.2210/pdb1GZX/pdb.

[65] Drews, Jürgen, Drug Discovery: A Historical Perspective, Science 287 (5460) (2000) 1960. doi:10.1126/science.287.5460.1960.

[66] Chin, Young-Won and Balunas, Marcy J. and Chai, Hee Byung and Kinghorn, A. Douglas, Drug discovery from natural sources, The AAPS Journal 8 (2) (2006) 28. doi:10.1007/BF02854894.

[67] Bosch, Fèlix and Rosich, Laia, The Contributions of Paul Ehrlich to Pharmacology: A Tribute on the Occasion of the Centenary of His Nobel Prize, Pharmacology 82 (3) (2008) 171–179. doi:10.1159/000149583.

[68] Lombardino, Joseph G. and Lowe, John A., The role of the medicinal chemist in drug discovery — then and now, Nature Reviews Drug Discovery 3 (10) (2004) 853–862. doi:10.1038/nrd1523.

[69] Pina, Ana Sofia and Hussain, Abid and Roque, Ana Cecília A., An Historical Overview of Drug Discovery, Humana Press, Totowa, NJ, 2010, pp. 3–12. doi:10.1007/978-1-60761-244-5_1.

[70] Gaynes, Robert, The Discovery of Penicillin–New Insights After More Than 75 Years of Clinical Use, Emerging Infectious Diseases 23 (5) (2017) 849–853, pMC5403050[pmcid]. doi:10.3201/eid2305.161556.

[71] Mohs, Richard C. and Greig, Nigel H., Drug discovery and development: Role of basic biological research, Alzheimer's & Dementia: Translational Research & Clinical Interventions 3 (4) (2017) 651–657. doi:10.1016/j.trci.2017.10.005.

[72] Broach, James R. and Thorner, Jeremy, High-throughput screening for drug discovery, Nature 384 (6604 SUPPL.) (1996) 14–16, review article. doi:10.1038/384014a0.

[73] Appell, Ken and Baldwin, John J. and Egan, William J., 2 - Combinatorial Chemistry and High-Throughput Screening in Drug Discovery and Development, Vol. 3 of Handbook of Modern Pharmaceutical Analysis, Academic Press, 2001, pp. 23–56. doi:10.1016/S0149-6395(01)80004-0.

[74] Entzeroth, Michael and Flotow, Horst and Condron, Peter, Overview of High-Throughput Screening, Current Protocols in Pharmacology 44 (1) (2009) 9.4.1–9.4.27. doi:10.1002/0471141755.ph0904s44.

[75] Williams, Michael and Raddatz, Rita and Mehlin, Christopher and Triggle, David J., Receptor Targets in Drug Discovery, Reviews in

Cell Biology and Molecular Medicine, 2006, major Reference Works. doi:10.1002/3527600906.mcb.200500063.

[76] Lemke, Thomas L. and Williams, David A. and Roche, Victoria F. and Zito, S. William, Foye's Principles of Medicinal Chemistry, Seventh Edition, Wolters Kluwer Health Adis (ESP), 2013, book.

[77] Kerns, Edward H. and Di, Li, Pharmaceutical profiling in drug discovery, Drug Discovery Today 8 (7) (2003) 316–323. doi:10.1016/S1359-6446(03)02649-7.

[78] Di, Li and Kerns, Edward H., Profiling drug-like properties in discovery research, Current Opinion in Chemical Biology 7 (3) (2003) 402–408. doi:10.1016/S1367-5931(03)00055-3.

[79] Perkins, Roger and Fang, Hong and Tong, Weida and Welsh, William J., Quantitative structure-activity relationship methods: Perspectives on drug discovery and toxicology, Environmental Toxicology and Chemistry 22 (8) (2003) 1666–1679. doi:10.1897/01-171.

[80] Yengi, Lilian G. and Leung, Louis and Kao, John, The Evolving Role of Drug Metabolism in Drug Discovery and Development, Pharmaceutical Research 24 (5) (2007) 842–858. doi:10.1007/s11095-006-9217-9.

[81] Sleno, Lekha and Emili, Andrew, Proteomic methods for drug target discovery, Current Opinion in Chemical Biology 12 (1) (2008) 46–54. doi:10.1016/j.cbpa.2008.01.022.

[82] Edwards, Aled M. and Arrowsmith, Cheryl H. and Christendat, Dinesh and Dharamsi, Akil and Friesen, James D. and Greenblatt, Jack F. and Vedadi, Masoud, Protein production: feeding the crystallographers and NMR spectroscopists, Nature Structural Biology 7 (11) (2000) 970–972. doi:10.1038/80751.

[83] Katayama, Hiroyuki and Oda, Yoshiya, Chemical proteomics for drug discovery based on compound-immobilized affinity chromatography, Journal of Chromatography B 855 (1) (2007) 21–27. doi:10.1016/j.jchromb.2006.12.047.

[84] Pellecchia, Maurizio and Sem, Daniel S. and Wüthrich, Kurt, Nmr in drug discovery, Nature Reviews Drug Discovery 1 (3) (2002) 211–219. doi:10.1038/nrd748.

[85] Maveyraud, Laurent and Mourey, Lionel, Protein X-ray Crystallography and Drug Discovery (2020). doi:10.3390/molecules25051030.

[86] Zhou, Shu-Feng and Zhong, Wei-Zhu, Drug Design and Discovery: Principles and Applications (2017). doi:10.3390/molecules22020279.

[87] Mouchlis, Varnavas D. and Afantitis, Antreas and Serra, Angela and Fratello, Michele and Papadiamantis, Anastasios G. and Aidinis, Vassilis and Lynch, Iseult and Greco, Dario and Melagraki, Georgia, Advances in De Novo Drug Design: From Conventional to Machine Learning Methods (2021). doi:10.3390/ijms22041676.

[88] Ashburn, Ted T. and Thor, Karl B., Drug repositioning: identifying and developing new uses for existing drugs, Nature Reviews Drug Discovery 3 (8) (2004) 673–683. doi:10.1038/nrd1468.

[89] Dudley, Joel T. and Deshpande, Tarangini and Butte, Atul J., Exploiting drug–disease relationships for computational drug repositioning, Briefings in Bioinformatics 12 (4) (2011) 303–311. doi:10.1093/bib/bbr013.

[90] Ekins, Sean and Williams, Antony J. and Krasowski, Matthew D. and Freundlich, Joel S., In silico repositioning of approved drugs for rare and neglected diseases, Drug Discovery Today 16 (7) (2011) 298–310. doi:10.1016/j.drudis.2011.02.016.

[91] FDA, The Drug Development Process.
URL https://www.fda.gov/patients/learn-about-drug-and-device-approvals/drug-development-process

[92] Mao, Fei and Ni, Wei and Xu, Xiang and Wang, Hui and Wang, Jing and Ji, Min and Li, Jian, Chemical Structure-Related Drug-Like Criteria of Global Approved Drugs (2016). doi:10.3390/molecules21010075.

[93] Golan, David E and Tashjian, Armen H and Armstrong, Ehrin J, Principles of pharmacology: the pathophysiologic basis of drug therapy, Lippincott Williams & Wilkins, 2011.

[94] Hage, David S. and Jackson, Abby and Sobansky, Matthew R. and Schiel, John E. and Yoo, Michelle J. and Joseph, K. S., Characterization of drug–protein interactions in blood using high-performance affinity chromatography, Journal of Separation Science 32 (5-6) (2009) 835–853. doi:10.1002/jssc.200800640.

[95] McConathy, Jonathan and Owens, Michael J., Stereochemistry in drug action, Prim. Care Companion J. Clin. Psychiatry 5 (2) (2003) 70–73, pMC353039. doi:10.4088/pcc.v05n0202.

[96] Brooks, W. H. and Guida, W. C. and Daniel, K. G., The significance of chirality in drug design and development, Curr. Top. Med. Chem. 11 (7) (2011) 760–770, pMC5765859. doi:10.2174/156802611795165098.

[97] Zhang, Tinghu and Hatcher, John M. and Teng, Mingxing and Gray, Nathanael S. and Kostic, Milka, Recent advances in selective and irreversible covalent ligand development and validation, Cell Chem. Biol. 26 (11) (2019) 1486–1500, pMC6886688. doi:10.1016/j.chembiol.2019.09.012.

[98] Kiely-Collins, Hannah and Winter, Georg E. and Bernardes, Gonçalo J.L., The role of reversible and irreversible covalent chemistry in targeted protein degradation, Cell Chemical Biology 28 (7) (2021) 952–968. doi:10.1016/j.chembiol.2021.03.005.

[99] Oravcova´, Jana and Bo¨hs, Barbara and Lindner, Wolfgang, Drug-protein binding studies new trends in analytical and experimental methodology, Journal of Chromatography B: Biomedical Sciences and Applications 677 (1) (1996) 1–28. doi:10.1016/0378-4347(95)00425-4.

[100] Christopoulos, Arthur, Allosteric binding sites on cell-surface receptors: novel targets for drug discovery, Nature Reviews Drug Discovery 1 (3) (2002) 198–210. doi:10.1038/nrd746.

[101] Cozzini, Pietro and Kellogg, Glen E. and Spyrakis, Francesca and Abraham, Donald J. and Costantino, Gabriele and Emerson, Andrew and Fanelli, Francesca and Gohlke, Holger and Kuhn, Leslie A. and Morris, Garrett M. and Orozco, Modesto and Pertinhez, Thelma A. and Rizzi, Menico and Sotriffer, Christoph A., Target Flexibility: An Emerging Consideration in Drug Discovery and Design, Journal of Medicinal Chemistry 51 (20) (2008) 6237–6255. doi:10.1021/jm800562d.

[102] Amaral, M. and Kokh, D. B. and Bomke, J. and Wegener, A. and Buchstaller, H. P. and Eggenweiler, H. M. and Matias, P. and Sirrenberg, C. and Wade, R. C. and Frech, M., Protein conforma-

tional flexibility modulates kinetics and thermodynamics of drug binding, Nature Communications 8 (1) (2017) 2276. doi:10.1038/s41467-017-02258-w.

[103] Lambert, D. G., Drugs and receptors, Continuing Education in Anaesthesia Critical Care & Pain 4 (6) (2004) 181–184. doi:10.1093/bjaceaccp/mkh049.

[104] Newton, Philip and Harrison, Paula and Clulow, Stephen, A Novel Method for Determination of the Affinity of Protein: Protein Interactions in Homogeneous Assays, Journal of Biomolecular Screening 13 (7) (2008) 674–682. doi:10.1177/1087057108321086.

[105] Yung-Chi, Cheng and Prusoff, William H., Relationship between the inhibition constant (KI) and the concentration of inhibitor which causes 50 per cent inhibition (I50) of an enzymatic reaction, Biochemical Pharmacology 22 (23) (1973) 3099–3108. doi:10.1016/0006-2952(73)90196-2.

[106] Johnson, Kenneth A. and Goody, Roger S., The Original Michaelis Constant: Translation of the 1913 Michaelis–Menten Paper, Biochemistry 50 (39) (2011) 8264–8269. doi:10.1021/bi201284u.

[107] Chaires, Jonathan B., Calorimetry and Thermodynamics in Drug Design, Annual Review of Biophysics 37 (1) (2008) 135–151. doi:10.1146/annurev.biophys.36.040306.132812.

[108] Steinbrecher, Thomas and Labahn, Andreas, Towards Accurate Free Energy Calculations in Ligand Protein-Binding Studies, Current Medicinal Chemistry 17 (8) (2010) 767–785. doi:doi:10.2174/092986710790514453.

[109] Gibbs, Josiah Willard, A method of geometrical representation of the thermodynamic properties by means of surfaces, The Collected Works of J. Willard Gibbs, Ph. D., LL. D (1957) 33–54.

[110] Gilson, Michael K. and Zhou, Huan-Xiang, Calculation of Protein-Ligand Binding Affinities, Annual Review of Biophysics and Biomolecular Structure 36 (1) (2007) 21–42. doi:10.1146/annurev.biophys.36.040306.132550.

[111] Li, HuiMin and Xie, YueHui and Liu, CiQuan and Liu, ShuQun, Physicochemical bases for protein folding, dynamics, and protein-ligand binding, Science China Life Sciences 57 (3) (2014) 287–302. doi:10.1007/s11427-014-4617-2.

[112] Perozzo, Remo and Folkers, Gerd and Scapozza, Leonardo, Thermodynamics of Protein–Ligand Interactions: History, Presence, and Future Aspects, Journal of Receptors and Signal Transduction 24 (1-2) (2004) 1–52. doi:10.1081/RRS-120037896.

[113] Liu, Shu-Qun and Ji, Xing-Lai and Tao, Yan and Tan, De-Yong and Zhang, Ke-Qin and Fu, Yun-Xin, Protein Folding, Binding and Energy Landscape: A Synthesis, IntechOpen, Rijeka, 2012, p. Ch. 10. doi:10.5772/30440.

[114] Amzel, L.Mario, Calculation of entropy changes in biological processes: Folding, binding, and oligomerization, Vol. 323 of Energetics of Biological Macromolecules, Part C, Academic Press, 2000, pp. 167–177. doi:10.1016/S0076-6879(00)23366-1.

[115] Amzel, L. Mario, Loss of translational entropy in binding, folding, and catalysis, Proteins: Structure, Function, and Bioinformatics 28 (2) (1997) 144–149. doi:10.1002/(SICI)1097-0134(199706)28:2<144::AID-PROT2>3.0.CO;2-F.

[116] MacRaild, Christopher A. and Daranas, Antonio Hernández and Bronowska, Agnieszka and Homans, Steve W., Global Changes in Local Protein Dynamics Reduce the Entropic Cost of Carbohydrate Binding in the Arabinose-binding Protein, Journal of Molecular Biology 368 (3) (2007) 822–832. doi:10.1016/j.jmb.2007.02.055.

[117] Bronowska, Agnieszka K., Thermodynamics of Ligand-Protein Interactions: Implications for Molecular Design, IntechOpen, Rijeka, 2011, p. Ch. 1.

[118] Dunitz, J. D., Win some, lose some: enthalpy-entropy compensation in weak intermolecular interactions, Chemistry & biology 2 (11) (1995) 709–712. doi:10.1016/1074-5521(95)90097-7.

[119] Chodera, John D. and Mobley, David L., Entropy-Enthalpy Compensation: Role and Ramifications in Biomolecular Ligand Recognition and Design, Annual Review of Biophysics 42 (1) (2013) 121–142. doi:10.1146/annurev-biophys-083012-130318.

[120] Breiten, Benjamin and Lockett, Matthew R. and Sherman, Woody and Fujita, Shuji and Al-Sayah, Mohammad and Lange, Heiko and Bowers, Carleen M. and Heroux, Annie and Krilov,

Goran and Whitesides, George M., Water Networks Contribute to Enthalpy/Entropy Compensation in Protein–Ligand Binding, Journal of the American Chemical Society 135 (41) (2013) 15579–15584. doi:10.1021/ja4075776.

[121] Cramer, Friedrich, Emil Fischer's Lock-and-Key Hypothesis after 100 years—Towards a Supracellular Chemistry, Perspectives in Supramolecular Chemistry, Perspectives in Supramolecular Chemistry, 1994, pp. 1–23. doi:10.1002/9780470511411.ch1.

[122] Koshland Jr., Daniel E., The Key–Lock Theory and the Induced Fit Theory, Angewandte Chemie International Edition in English 33 (23-24) (1995) 2375–2378. doi:10.1002/anie.199423751.

[123] Koshland, D. E., Application of a Theory of Enzyme Specificity to Protein Synthesis*, Proceedings of the National Academy of Sciences 44 (2) (1958) 98–104. doi:10.1073/pnas.44.2.98.

[124] Tobi, Dror and Bahar, Ivet, Structural changes involved in protein binding correlate with intrinsic motions of proteins in the unbound state, Proceedings of the National Academy of Sciences 102 (52) (2005) 18908–18913. doi:10.1073/pnas.0507603102.

[125] Csermely, Peter and Palotai, Robin and Nussinov, Ruth, Induced fit, conformational selection and independent dynamic segments: an extended view of binding events, Nature Precedingsdoi:10.1038/npre.2010.4422.1.

[126] Holyoak, Todd, Molecular Recognition: Lock-and-Key, Induced Fit, and Conformational Selection, Springer Berlin Heidelberg, Berlin, Heidelberg, 2013, pp. 1584–1588. doi:10.1007/978-3-642-16712-6_468.

[127] Odriozola, G. and Jiménez-Ángeles, F. and Lozada-Cassou, M., Entropy driven key-lock assembly, The Journal of Chemical Physics 129 (11) (2008) 111101. doi:10.1063/1.2981795.

[128] Sacanna, S. and Irvine, W. T. M. and Chaikin, P. M. and Pine, D. J., Lock and key colloids, Nature 464 (7288) (2010) 575–578. doi:10.1038/nature08906.

[129] Kraft, Daniela J. and Ni, Ran and Smallenburg, Frank and Hermes, Michiel and Yoon, Kisun and Weitz, David A. and van Blaaderen, Alfons and Groenewold, Jan and Dijkstra, Marjolein and Kegel, Willem K., Surface roughness directed self-assembly of

patchy particles into colloidal micelles, Proceedings of the National Academy of Sciences 109 (27) (2012) 10787–10792. doi:10.1073/pnas.1116820109.

[130] Schneider, Hans-Jörg, Limitations and Extensions of the Lock-and-Key Principle: Differences between Gas State, Solution and Solid State Structures (2015). doi:10.3390/ijms16046694.

[131] Bosshard, Hans Rudolf, Molecular Recognition by Induced Fit: How Fit is the Concept?, Physiology 16 (4) (2001) 171–173. doi:10.1152/physiologyonline.2001.16.4.171.

[132] Chang, Chia-En and Gilson, Michael K., Free Energy, Entropy, and Induced Fit in Host-Guest Recognition: Calculations with the Second-Generation Mining Minima Algorithm, Journal of the American Chemical Society 126 (40) (2004) 13156–13164. doi:10.1021/ja047115d.

[133] Corbett, Peter T. and Tong, Lok H. and Sanders, Jeremy K. M. and Otto, Sijbren, Diastereoselective Amplification of an Induced-Fit Receptor from a Dynamic Combinatorial Library, Journal of the American Chemical Society 127 (25) (2005) 8902–8903. doi:10.1021/ja050790i.

[134] Hariharan, Parameswaran and Guan, Lan, Insights into the Inhibitory Mechanisms of the Regulatory Protein IIA¡sup¿Glc¡/sup¿ on Melibiose Permease Activity *, Journal of Biological Chemistry 289 (47) (2014) 33012–33019. doi:10.1074/jbc.M114.609255.

[135] Henzler-Wildman, Katherine and Kern, Dorothee, Dynamic personalities of proteins, Nature 450 (7172) (2007) 964–972. doi:10.1038/nature06522.

[136] Boehr, David D. and Nussinov, Ruth and Wright, Peter E., The role of dynamic conformational ensembles in biomolecular recognition, Nature Chemical Biology 5 (11) (2009) 789–796. doi:10.1038/nchembio.232.

[137] Foote, J. and Milstein, C., Conformational isomerism and the diversity of antibodies., Proceedings of the National Academy of Sciences 91 (22) (1994) 10370–10374. doi:10.1073/pnas.91.22.10370.

[138] KUMAR, SANDEEP and MA, BUYONG and TSAI, CHUNG-JUNG and SINHA, NEETI and NUSSINOV, RUTH, Folding and binding cascades: Dynamic landscapes and population shifts, Protein Science 9 (1) (2000) 10–19. doi:10.1110/ps.9.1.10.

[139] Ma, Buyong and Wolfson, Haim J. and Nussinov, Ruth, Protein functional epitopes: hot spots, dynamics and combinatorial libraries, Current Opinion in Structural Biology 11 (3) (2001) 364–369. doi:10.1016/S0959-440X(00)00216-5.

[140] Vogt, Austin D. and Di Cera, Enrico, Conformational Selection or Induced Fit? A Critical Appraisal of the Kinetic Mechanism, Biochemistry 51 (30) (2012) 5894–5902. doi:10.1021/bi3006913.

[141] Kastritis, Panagiotis L. and Bonvin, Alexandre M. J. J., On the binding affinity of macromolecular interactions: daring to ask why proteins interact, Journal of The Royal Society Interface 10 (79) (2013) 20120835. doi:10.1098/rsif.2012.0835.

[142] Nussinov, Ruth and Ma, Buyong and Tsai, Chung-Jung, Multiple conformational selection and induced fit events take place in allosteric propagation, Biophysical Chemistry 186 (2014) 22–30. doi:10.1016/j.bpc.2013.10.002.

[143] Greives, Nicholas and Zhou, Huan-Xiang, Both protein dynamics and ligand concentration can shift the binding mechanism between conformational selection and induced fit, Proceedings of the National Academy of Sciences 111 (28) (2014) 10197–10202. doi:10.1073/pnas.1407545111.

[144] Gao, Mu and Skolnick, Jeffrey, A Comprehensive Survey of Small-Molecule Binding Pockets in Proteins, PLOS Computational Biology 9 (10) (2013) e1003302. doi:10.1371/journal.pcbi.1003302.

[145] Hopkins, Andrew L. and Groom, Colin R., The druggable genome, Nature Reviews Drug Discovery 1 (9) (2002) 727–730. doi:10.1038/nrd892.

[146] Henrich, Stefan and Salo-Ahen, Outi M. H. and Huang, Bingding and Rippmann, Friedrich F. and Cruciani, Gabriele and Wade, Rebecca C., Computational approaches to identifying and characterizing protein binding sites for ligand design, Journal of Molecular Recognition 23 (2) (2010) 209–219. doi:10.1002/jmr.984.

[147] del Sol, Antonio and Fujihashi, Hirotomo and Amoros, Dolors and Nussinov, Ruth, Residue centrality, functionally important residues, and active site shape: Analysis of enzyme and non-enzyme families, Protein Science 15 (9) (2006) 2120–2128. doi:10.1110/ps.062249106.

[148] Carlson, Heather A., Protein flexibility and drug design: how to hit a moving target, Current Opinion in Chemical Biology 6 (4) (2002) 447–452. doi:10.1016/S1367-5931(02)00341-1.

[149] Surade, Sachin and Blundell, Tom L., Structural Biology and Drug Discovery of Difficult Targets: The Limits of Ligandability, Chemistry & Biology 19 (1) (2012) 42–50. doi:10.1016/j.chembiol.2011.12.013.

[150] Stank, Antonia and Kokh, Daria B. and Fuller, Jonathan C. and Wade, Rebecca C., Protein Binding Pocket Dynamics, Accounts of Chemical Research 49 (5) (2016) 809–815. doi:10.1021/acs.accounts.5b00516.

[151] Voll, Andreas M. and Meyners, Christian and Taubert, Martha C. and Bajaj, Thomas and Heymann, Tim and Merz, Stephanie and Charalampidou, Anna and Kolos, Jürgen and Purder, Patrick L. and Geiger, Thomas M. and Wessig, Pablo and Gassen, Nils C. and Bracher, Andreas and Hausch, Felix, Macrocyclic FKBP51 Ligands Define a Transient Binding Mode with Enhanced Selectivity, Angewandte Chemie International Edition 60 (24) (2021) 13257–13263. doi:10.1002/anie.202017352.

[152] Umezawa, Koji and Kii, Isao, Druggable Transient Pockets in Protein Kinases (2021). doi:10.3390/molecules26030651.

[153] Lerma Romero, Jorge A. and Meyners, Christian and Christmann, Andreas and Reinbold, Lisa M. and Charalampidou, Anna and Hausch, Felix and Kolmar, Harald, Binding pocket stabilization by high-throughput screening of yeast display libraries, Frontiers in Molecular Biosciences 9, original Research. doi:10.3389/fmolb.2022.1023131.

[154] Leis, Simon and Schneider, Sebastian and Zacharias, Martin, In Silico Prediction of Binding Sites on Proteins, Current Medicinal Chemistry 17 (15) (2010) 1550–1562. doi:doi:10.2174/092986710790979944.

[155] Chen, Ke and Mizianty, Marcin J. and Gao, Jianzhao and Kurgan, Lukasz, A Critical Comparative Assessment of Predictions of Protein-Binding Sites for Biologically Relevant Organic Compounds, Structure 19 (5) (2011) 613–621. doi:10.1016/j.str.2011.02.015.

[156] Roche, Daniel B. and Brackenridge, Danielle A. and McGuffin, Liam J., Proteins and Their Interacting Partners: An Introduction

to Protein–Ligand Binding Site Prediction Methods (2015). doi:10.3390/ijms161226202.

[157] Broomhead, Neal K. and Soliman, Mahmoud E., Can We Rely on Computational Predictions To Correctly Identify Ligand Binding Sites on Novel Protein Drug Targets? Assessment of Binding Site Prediction Methods and a Protocol for Validation of Predicted Binding Sites, Cell Biochemistry and Biophysics 75 (1) (2017) 15–23. doi:10.1007/s12013-016-0769-y.

[158] Brylinski, Michal and Skolnick, Jeffrey, A threading-based method (FINDSITE) for ligand-binding site prediction and functional annotation, Proc Natl Acad Sci U S A 105 (1) (2007) 129–134, pMC2224172. doi:10.1073/pnas.0707684105.

[159] Capra, John A. and Laskowski, Roman A. and Thornton, Janet M. and Singh, Mona and Funkhouser, Thomas A., Predicting Protein Ligand Binding Sites by Combining Evolutionary Sequence Conservation and 3D Structure, PLOS Computational Biology 5 (12) (2009) e1000585. doi:10.1371/journal.pcbi.1000585.

[160] Roy, Ambrish and Yang, Jianyi and Zhang, Yang, COFACTOR: an accurate comparative algorithm for structure-based protein function annotation, Nucleic Acids Research 40 (W1) (2012) W471–W477. doi:10.1093/nar/gks372.

[161] Volkamer, Andrea and Kuhn, Daniel and Grombacher, Thomas and Rippmann, Friedrich and Rarey, Matthias, Combining Global and Local Measures for Structure-Based Druggability Predictions, Journal of Chemical Information and Modeling 52 (2) (2012) 360–372. doi:10.1021/ci200454v.

[162] Yang, Jianyi and Roy, Ambrish and Zhang, Yang, Protein–ligand binding site recognition using complementary binding-specific substructure comparison and sequence profile alignment, Bioinformatics 29 (20) (2013) 2588–2595. doi:10.1093/bioinformatics/btt447.

[163] Kriv´ák, Radoslav and Hoksza, David, P2Rank: machine learning based tool for rapid and accurate prediction of ligand binding sites from protein structure, Journal of Cheminformatics 10 (1) (2018) 39. doi:10.1186/s13321-018-0285-8.

[164] Cui, Yifeng and Dong, Qiwen and Hong, Daocheng and Wang, Xikun, Predicting protein-ligand binding residues with deep convolutional neural networks, BMC Bioinformatics 20 (1) (2019) 93. doi:10.1186/s12859-019-2672-1.

[165] Breiman, Leo, Random Forests, Machine Learning 45 (1) (2001) 5–32. doi:10.1023/A:1010933404324.

[166] Cortes, Corinna and Vapnik, Vladimir, Support-vector networks, Machine Learning 20 (3) (1995) 273–297. doi:10.1007/BF00994018.

[167] Drucker, Harris and Burges, Christopher J. C. and Kaufman, Linda and Smola, Alex and Vapnik, Vladimir, Support Vector Regression Machines, in: Advances in Neural Information Processing Systems, Vol. 9, MIT Press, 1996.

[168] Heikamp, Kathrin and Bajorath, Jürgen, Support vector machines for drug discovery, Expert Opinion on Drug Discovery 9 (1) (2014) 93–104. doi:10.1517/17460441.2014.866943.

[169] Friedman, Jerome H., Greedy Function Approximation: A Gradient Boosting Machine, The Annals of Statistics 29 (5) (2001) 1189–1232, full publication date: Oct., 2001. doi:10.1214/aos/1013203451.

[170] Tianqi Chen and Tong He, Higgs Boson Discovery with Boosted Trees, in: Proceedings of the NIPS 2014 Workshop on High-energy Physics and Machine Learning, Vol. 42 of Proceedings of Machine Learning Research, PMLR, Montreal, Canada, 2015, pp. 69–80.

[171] Chen, Tianqi and Guestrin, Carlos, XGBoost: A Scalable Tree Boosting System, KDD '16, Association for Computing Machinery, New York, NY, USA, 2016, p. 785–794. doi:10.1145/2939672.2939785.

[172] Ashish Vaswani and Noam Shazeer and Niki Parmar and Jakob Uszkoreit and Llion Jones and Aidan N. Gomez and Lukasz Kaiser and Illia Polosukhin, Attention Is All You Need (2017). arXiv:1706.03762v5.

[173] Jacob Devlin and Ming-Wei Chang and Kenton Lee and Kristina Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding (2019). arXiv:1810.04805.

[174] Nitish Srivastava and Geoffrey Hinton and Alex Krizhevsky and Ilya Sutskever and Ruslan Salakhutdinov, Dropout: A Simple Way to Prevent Neural Networks from Overfitting, Journal of Machine Learning Research 15 (56) (2014) 1929–1958.

[175] Jimmy Lei Ba and Jamie Ryan Kiros and Geoffrey E. Hinton, Layer Normalization (2016). arXiv:1607.06450.

[176] Sousa, Sérgio Filipe and Fernandes, Pedro Alexandrino and Ramos, Maria João, Protein-ligand docking: Current status and future challenges, Proteins: Structure, Function, and Bioinformatics 65 (1) (2006) 15–26. doi:10.1002/prot.21082.

[177] Giovanni, A. Lavecchia and C. Di, Virtual Screening Strategies in Drug Discovery: A Critical Review, Current Medicinal Chemistry 20 (23) (2013) 2839–2860. doi:10.2174/09298673113209990001.

[178] Pujadas, Gerard and Vaqué, Montserrat and Ardèvol, Anna and Bladé, Cinta and Salvadó, Maria-Josepa and Blay, Mayte and Fernandez-Larrea, Juan-Bautista and Arola, Lluis, Protein-ligand Docking: A Review of Recent Advances and Future Perspectives, Vol. 4 of Current Pharmaceutical Analysis, Current Pharmaceutical Analysis, 2008. doi:10.2174/157341208783497597.

[179] Pagadala, Nataraj S. and Syed, Khajamohiddin and Tuszynski, Jack, Software for molecular docking: a review, Biophysical reviews 9 (2) (2017) 91–102. doi:10.1007/s12551-016-0247-1.

[180] Waterhouse, Andrew and Bertoni, Martino and Bienert, Stefan and Studer, Gabriel and Tauriello, Gerardo and Gumienny, Rafal and Heer, Florian T. and de Beer, Tjaart A. P. and Rempfer, Christine and Bordoli, Lorenza and Lepore, Rosalba and Schwede, Torsten, SWISS-MODEL: homology modelling of protein structures and complexes, Nucleic acids research 46 (W1). doi:10.1093/nar/gky427.

[181] Jumper, John and Evans, Richard and Pritzel, Alexander and Green, Tim and Figurnov, Michael and Ronneberger, Olaf and Tunyasuvunakool, Kathryn and Bates, Russ and Žídek, Augustin and Potapenko, Anna and Bridgland, Alex and Meyer, Clemens and Kohl, Simon A. A. and Ballard, Andrew J. and Cowie, An-

drew and Romera-Paredes, Bernardino and Nikolov, Stanislav and Jain, Rishub and Adler, Jonas and Back, Trevor and Petersen, Stig and Reiman, David and Clancy, Ellen and Zielinski, Michal and Steinegger, Martin and Pacholska, Michalina and Berghammer, Tamas and Bodenstein, Sebastian and Silver, David and Vinyals, Oriol and Senior, Andrew W. and Kavukcuoglu, Koray and Kohli, Pushmeet and Hassabis, Demis, Highly accurate protein structure prediction with AlphaFold, Nature 596 (7873) (2021) 583–589. doi:10.1038/s41586-021-03819-2.

[182] Thilagavathi, N. and Amudha, T., An Analytical Study of NP-Hard Protein Folding Problems, in: 2014 International Conference on Intelligent Computing Applications, 2014, pp. 184–188. doi:10.1109/ICICA.2014.47.

[183] L. Teodoro, Miguel and Phillips, George and Kavraki, Lydia, Molecular Docking: A Problem With Thousands Of Degrees Of Freedom, Vol. 1 of Proceedings - IEEE International Conference on Robotics and Automation, Proceedings - IEEE International Conference on Robotics and Automation, 2002. doi:10.1109/ROBOT.2001.932674.

[184] Zhang, Hailei and Li, Honglin and Jiang, Hualiang and Shen, Jianhua and Chen, Kaixian and Yang, Kun and Yu, Kunqian and Kang, Ling and Zhu, Weiliang and Luo, Xiaomin and Wang, Xicheng and Gao, Zhenting, TarFisDock: a web server for identifying drug targets with docking approach, Nucleic Acids Research 34 (suppl_2) (2006) W219–W224. doi:10.1093/nar/gkl114.

[185] Wang, Jui-Chih and Chu, Pei-Ying and Chen, Chung-Ming and Lin, Jung-Hsin, idTarget: a web server for identifying protein targets of small chemical molecules with robust scoring functions and a divide-and-conquer docking approach, Nucleic Acids Research 40 (W1) (2012) W393–W399. doi:10.1093/nar/gks496.

[186] Berman, H. M. and Westbrook, J. and Feng, Z. and Gilliland, G. and Bhat, T. N. and Weissig, H. and Shindyalov, I. N. and Bourne, P. E., The Protein Data Bank, Nucleic Acids Res 28 (1) (2000) 235–242. doi:10.1093/nar/28.1.235.

[187] Huey, Ruth and Morris, Garrett M. and Olson, Arthur J. and Goodsell, David S., A semiempirical free energy force field with charge-

based desolvation, Journal of Computational Chemistry 28 (6) (2007) 1145–1152. doi:10.1002/jcc.20634.

[188] Gowthaman, Ragul and Miller, Sven A. and Rogers, Steven and Khowsathit, Jittasak and Lan, Lan and Bai, Nan and Johnson, David K. and Liu, Chunjing and Xu, Liang and Anbanandam, Asokan and Aubé, Jeffrey and Roy, Anuradha and Karanicolas, John, DARC: Mapping Surface Topography by Ray-Casting for Effective Virtual Screening at Protein Interaction Sites, Journal of Medicinal Chemistry 59 (9) (2016) 4152–4170. doi:10.1021/acs.jmedchem.5b00150.

[189] Leaver-Fay, Andrew and Tyka, Michael and Lewis, Steven M. and Lange, Oliver F. and Thompson, James and Jacak, Ron and Kaufman, Kristian W. and Renfrew, P. Douglas and Smith, Colin A. and Sheffler, Will and Davis, Ian W. and Cooper, Seth and Treuille, Adrien and Mandell, Daniel J. and Richter, Florian and Ban, Yih-En Andrew and Fleishman, Sarel J. and Corn, Jacob E. and Kim, David E. and Lyskov, Sergey and Berrondo, Monica and Mentzer, Stuart and Popović, Zoran and Havranek, James J. and Karanicolas, John and Das, Rhiju and Meiler, Jens and Kortemme, Tanja and Gray, Jeffrey J. and Kuhlman, Brian and Baker, David and Bradley, Philip, Chapter nineteen - Rosetta3: An Object-Oriented Software Suite for the Simulation and Design of Macromolecules, Vol. 487 of Computer Methods, Part C, Academic Press, 2011, pp. 545–574. doi:10.1016/B978-0-12-381270-4.00019-6.

[190] Wang, Fan and Wu, Feng-Xu and Li, Cheng-Zhang and Jia, Chen-Yang and Su, Sun-Wen and Hao, Ge-Fei and Yang, Guang-Fu, ACID: a free tool for drug repurposing using consensus inverse docking strategy, Journal of Cheminformatics 11 (1) (2019) 73. doi:10.1186/s13321-019-0394-z.

[191] Trott, Oleg and Olson, Arthur J., AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading, Journal of Computational Chemistry 31 (2) (2010) 455–461. doi:10.1002/jcc.21334.

[192] Lephar, LEDOCK.
URL http://www.lephar.com/

[193] Korb, Oliver and Stützle, Thomas and Exner, Thomas E., Empirical Scoring Functions for Advanced Protein-Ligand Docking with PLANTS, Journal of Chemical Information and Modeling 49 (1) (2009) 84–96. doi:10.1021/ci800298z.

[194] Ng, Marcus C. K. and Fong, Simon and Siu, Shirley W. I., PSOVina: The hybrid particle swarm optimization algorithm for protein–ligand docking, Journal of Bioinformatics and Computational Biology 13 (03) (2015) 1541007. doi:10.1142/S0219720015410073.

[195] Wang, Renxiao and Lai, Luhua and Wang, Shaomeng, Further development and validation of empirical scoring functions for structure-based binding affinity prediction, Journal of Computer-Aided Molecular Design 16 (1) (2002) 11–26. doi:10.1023/A:1016357811882.

[196] Obiol-Pardo, Cristian and Rubio-Martinez, Jaime, Comparative Evaluation of MMPBSA and XSCORE To Compute Binding Free Energy in XIAP-Peptide Complexes, Journal of Chemical Information and Modeling 47 (1) (2007) 134–142. doi:10.1021/ci600412z.

[197] Zhang, Wenyi and Bell, Eric W. and Yin, Minghao and Zhang, Yang, EDock: blind protein–ligand docking by replica-exchange monte carlo simulation, Journal of Cheminformatics 12 (1) (2020) 37. doi:10.1186/s13321-020-00440-9.

[198] Swendsen, Robert H. and Wang, Jian-Sheng, Replica Monte Carlo Simulation of Spin-Glasses, Phys. Rev. Lett. 57 (1986) 2607–2609. doi:10.1103/PhysRevLett.57.2607.

[199] Geppert, Hanna and Vogt, Martin and Bajorath, Jürgen, Current Trends in Ligand-Based Virtual Screening: Molecular Representations, Data Mining Methods, New Application Areas, and Performance Evaluation, Journal of Chemical Information and Modeling 50 (2) (2010) 205–216. doi:10.1021/ci900419k.

[200] Dudek, Arkadiusz Z. and Galvez, Tomasz Arodz and Jorge, Computational Methods in Developing Quantitative Structure-Activity Relationships (QSAR): A Review, Combinatorial Chemistry & High Throughput Screening 9 (3) (2006) 213–228. doi:10.2174/138620706776055539.

[201] Afantitis, Antreas and Melagraki, Georgia and Sarimveis, Haralambos and Koutentis, Panayiotis A. and Markopoulos, John

and Igglessi-Markopoulou, Olga, A novel QSAR model for predicting induction of apoptosis by 4-aryl-4H-chromenes, Bioorganic & Medicinal Chemistry 14 (19) (2006) 6686–6694. doi:10.1016/j.bmc.2006.05.061.

[202] Keiser, Michael J. and Roth, Bryan L. and Armbruster, Blaine N. and Ernsberger, Paul and Irwin, John J. and Shoichet, Brian K., Relating protein pharmacology by ligand chemistry, Nature Biotechnology 25 (2007) 197 EP –. doi:10.1038/nbt1284.

[203] Pearson, William R., Empirical statistical estimates for sequence similarity searches, Journal of Molecular Biology 276 (1) (1998) 71–84. doi:10.1006/jmbi.1997.1525.

[204] Luo, Man and Wang, Xiang Simon and Roth, Bryan L. and Golbraikh, Alexander and Tropsha, Alexander, Application of Quantitative Structure-Activity Relationship Models of 5-HT1A Receptor Binding to Virtual Screening Identifies Novel and Potent 5-HT1A Ligands, Journal of Chemical Information and Modeling 54 (2) (2014) 634–647. doi:10.1021/ci400460q.

[205] Talete: Milano, Italy, DRAGON for Windows and Linux.
URL http://www.talete.mi.it/help/dragon_help

[206] Ma, Junshui and Sheridan, Robert P. and Liaw, Andy and Dahl, George E. and Svetnik, Vladimir, Deep Neural Nets as a Method for Quantitative Structure–Activity Relationships, Journal of Chemical Information and Modeling 55 (2) (2015) 263–274. doi:10.1021/ci500747n.

[207] Neves, Bruno J. and Dantas, Rafael F. and Senger, Mario R. and Melo-Filho, Cleber C. and Valente, Walter C. G. and de Almeida, Ana C. M. and Rezende-Neto, João M. and Lima, Elid F. C. and Paveley, Ross and Furnham, Nicholas and Muratov, Eugene and Kamentsky, Lee and Carpenter, Anne E. and Braga, Rodolpho C. and Silva-Junior, Floriano P. and Andrade, Carolina Horta, Discovery of New Anti-Schistosomal Hits by Integration of QSAR-Based Virtual Screening and High Content Screening, Journal of Medicinal Chemistry 59 (15) (2016) 7075–7088. doi:10.1021/acs.jmedchem.5b02038.

[208] Morgan, H. L., The Generation of a Unique Machine Description for Chemical Structures-A Technique Developed at Chemical Abstracts Service., Journal

of Chemical Documentation 5 (2) (1965) 107–113. doi:10.1021/c160017a018.

[209] Rogers, David and Hahn, Mathew, Extended-Connectivity Fingerprints, Journal of Chemical Information and Modeling 50 (5) (2010) 742–754. doi:10.1021/ci100050t.

[210] Carhart, Raymond E. and Smith, Dennis H. and Venkataraghavan, R., Atom pairs as molecular features in structure-activity studies: definition and applications, Journal of Chemical Information and Computer Sciences 25 (2) (1985) 64–73. doi:10.1021/ci00046a002.

[211] Durant, Joseph L. and Leland, Burton A. and Henry, Douglas R. and Nourse, James G., Reoptimization of MDL Keys for Use in Drug Discovery, Journal of Chemical Information and Computer Sciences 42 (6) (2002) 1273–1280. doi:10.1021/ci010132r.

[212] Willighagen, Egon L. and Mayfield, John W. and Alvarsson, Jonathan and Berg, Arvid and Carlsson, Lars and Jeliazkova, Nina and Kuhn, Stefan and Pluskal, Tomáš and Rojas-Chertó, Miquel and Spjuth, Ola and Torrance, Gilleain and Evelo, Chris T. and Guha, Rajarshi and Steinbeck, Christoph, The Chemistry Development Kit (CDK) v2.0: atom typing, depiction, molecular formulas, and substructure searching, Journal of Cheminformatics 9 (1) (2017) 33. doi:10.1186/s13321-017-0220-4.

[213] Brandon J. Bongers and Adriaan. P. IJzerman and Gerard J.P. Van Westen, Proteochemometrics - recent developments in bioactivity and selectivity modeling, Drug Discovery Today: Technologies 32-33 (2019) 89–98, artificial Intelligence. doi:https://doi.org/10.1016/j.ddtec.2020.08.003.

[214] Ezzat, Ali and Wu, Min and Li, Xiao-Li and Kwoh, Chee-Keong, Computational prediction of drug-target interactions using chemogenomic approaches: an empirical survey, Briefings in Bioinformaticsdoi:10.1093/bib/bby002.

[215] Wang, Chen and Kurgan, Lukasz, Review and comparative assessment of similarity-based methods for prediction of drug-protein interactions in the druggable human proteome, Briefings in Bioinformaticsdoi:10.1093/bib/bby069.

[216] Gutteridge, Alex and Araki, Michihiro and Kanehisa, Minoru and Honda, Wataru and Yamanishi, Yoshihiro, Prediction

of drug-target interaction networks from the integration of chemical and genomic spaces, Bioinformatics 24 (13) (2008) i232–i240. doi:10.1093/bioinformatics/btn162.

[217] Bleakley, Kevin and Yamanishi, Yoshihiro, Supervised prediction of drug-target interactions using bipartite local models, Bioinformatics 25 (18) (2009) 2397–2403. doi:10.1093/bioinformatics/btp433.

[218] Cheng, Feixiong and Liu, Chuang and Jiang, Jing and Lu, Weiqiang and Li, Weihua and Liu, Guixia and Zhou, Weixing and Huang, Jin and Tang, Yun, Prediction of Drug-Target Interactions and Drug Repositioning via Network-Based Inference, PLOS Computational Biology 8 (5) (2012) e1002503. doi:10.1371/journal.pcbi.1002503.

[219] Zheng, Xiaodong and Ding, Hao and Mamitsuka, Hiroshi and Zhu, Shanfeng, Collaborative matrix factorization with multiple similarities for predicting drug-target interactions, Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2013.

[220] Peng, L. and Liao, B. and Zhu, W. and Li, Z. and Li, K., Predicting Drug–Target Interactions With Multi-Information Fusion, IEEE Journal of Biomedical and Health Informatics 21 (2) (2017) 561–572. doi:10.1109/JBHI.2015.2513200.

[221] Zhouchen Lin and Minming Chen and Yi Ma, The Augmented Lagrange Multiplier Method for Exact Recovery of Corrupted Low-Rank Matrices (2013). arXiv:1009.5055v3.

[222] Ezzat, A. and Zhao, P. and Wu, M. and Li, X.-L. and Kwoh, C.-K., Drug-Target Interaction Prediction with Graph Regularized Matrix Factorization, IEEE/ACM Transactions on Computational Biology and Bioinformatics 14 (3) (2017) 646–656. doi:10.1109/TCBB.2016.2530062.

[223] Debie, Essam and Shafi, Kamran, Implications of the Curse of Dimensionality for Supervised Learning Classifier Systems: Theoretical and Empirical Analyses, Pattern Anal. Appl. 22 (2) (2019) 519–536. doi:10.1007/s10044-017-0649-0.

[224] Yu, Hua and Chen, Jianxin and Xu, Xue and Li, Yan and Zhao, Huihui and Fang, Yupeng and Li, Xiuxiu and Zhou, Wei and Wang,

Wei and Wang, Yonghua, A Systematic Prediction of Multiple Drug-Target Interactions from Chemical, Genomic, and Pharmacological Data, PLOS ONE 7 (5) (2012) e37608. doi:10.1371/journal.pone.0037608.

[225] Li, Z. R. and Lin, H. H. and Han, L. Y. and Jiang, L. and Chen, X. and Chen, Y. Z., PROFEAT: a web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence, Nucleic Acids Research 34 (suppl_2) (2006) W32–W37. doi:10.1093/nar/gkl305.

[226] Cao, Dong-Sheng and Zhang, Liu-Xia and Tan, Gui-Shan and Xiang, Zheng and Zeng, Wenbin and Xu, Qingsong and Chen, Alex, Computational Prediction of DrugTarget Interactions Using Chemical, Biological, and Network Features, Vol. 33 of Molecular Informatics, Molecular Informatics, 2014. doi:10.1002/minf.201400009.

[227] Coelho, Edgar D. and Arrais, Joel P. and Oliveira, José Luís, Computational Discovery of Putative Leads for Drug Repositioning through Drug-Target Interaction Prediction, PLOS Computational Biology 12 (11) (2016) e1005219. doi:10.1371/journal.pcbi.1005219.

[228] Cao, Dong-Sheng and Liang, Yi-Zeng and Yan, Jun and Tan, Gui-Shan and Xu, Qing-Song and Liu, Shao, PyDPI: Freely Available Python Package for Chemoinformatics, Bioinformatics, and Chemogenomics Studies, Journal of Chemical Information and Modeling 53 (11) (2013) 3086–3096. doi:10.1021/ci400127q.

[229] Peng, Lihong and Zhu, Wen and Liao, Bo and Duan, Yu and Chen, Min and Chen, Yi and Yang, Jialiang, Screening drug-target interactions with positive-unlabeled learning, Scientific Reports 7 (1) (2017) 8087. doi:10.1038/s41598-017-08079-7.

[230] Liu, Bing and Lee, Wee Sun and Yu, Philip S. and Li, Xiaoli, Partially Supervised Classification of Text Documents, in: Proceedings of the Nineteenth International Conference on Machine Learning, ICML '02, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2002, p. 387–394.

[231] Li, Xiaoli and Liu, Bing, Learning to Classify Texts Using Positive and Unlabeled Data., Proceedings of Eighteenth International Joint Conference on Artificial Intelligence (IJCAI-03): 2003; Acapulco, Mexico, Proceedings of

Eighteenth International Joint Conference on Artificial Intelligence (IJCAI-03): 2003, 2003.

[232] Yap, Chun Wei, PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints, Journal of Computational Chemistry 32 (7) (2011) 1466–1474. doi:10.1002/jcc.21707.

[233] Mamoshina, Polina and Vieira, Armando and Putin, Evgeny and Zhavoronkov, Alex, Applications of Deep Learning in Biomedicine, Molecular Pharmaceutics 13 (5) (2016) 1445–1454. doi:10.1021/acs.molpharmaceut.5b00982.

[234] LeCun, Yann and Bengio, Yoshua and Hinton, Geoffrey, Deep learning, Nature 521 (2015) 436 EP –. doi:10.1038/nature14539.

[235] Tian, Kai and Shao, Mingyu and Wang, Yang and Guan, Jihong and Zhou, Shuigeng, Boosting compound-protein interaction prediction by deep learning, Methods 110 (2016) 64–72. doi:10.1016/j.ymeth.2016.06.024.

[236] Peng-Wei and Chan, Keith and You, Zhu-Hong, Large-scale prediction of drug-target interactions from deep representations, 2016 International Joint Conference on Neural Networks (IJCNN) (2016) 1236–1243doi:10.1109/IJCNN.2016.7727339.

[237] Wen, Ming and Zhang, Zhimin and Niu, Shaoyu and Sha, Haozhi and Yang, Ruihan and Yun, Yonghuan and Lu, Hongmei, Deep-Learning-Based Drug-Target Interaction Prediction, Journal of Proteome Research 16 (4) (2017) 1401–1409. doi:10.1021/acs.jproteome.6b00618.

[238] Wang, Yan-Bin and You, Zhu-Hong and Yang, Shan and Yi, Hai-Cheng and Chen, Zhan-Heng and Zheng, Kai, A deep learning-based method for drug-target interaction prediction based on long short-term memory neural network, BMC Medical Informatics and Decision Making 20 (2) (2020) 49. doi:10.1186/s12911-020-1052-0.

[239] Tsubaki, Masashi and Tomii, Kentaro and Sese, Jun, Compound–protein interaction prediction with end-to-end learning of neural networks for graphs and sequences, Bioinformatics 35 (2) (2019) 309–318. doi:10.1093/bioinformatics/bty535.

[240] Lee, Ingoo and Keum, Jongsoo and Nam, Hojung, DeepConv-DTI: Prediction of drug-target interactions via deep learning with convolution on

protein sequences, PLOS Computational Biology 15 (6) (2019) e1007129. doi:10.1371/journal.pcbi.1007129.

[241] Monteiro, N. R. C. and Ribeiro, B. and Arrais, J. P., Drug-Target Interaction Prediction: End-to-End Deep Learning Approach, IEEE/ACM Transactions on Computational Biology and Bioinformatics 18 (6) (2021) 2364–2374. doi:10.1109/TCBB.2020.2977335.

[242] Dzmitry Bahdanau and Kyunghyun Cho and Yoshua Bengio, Neural Machine Translation by Jointly Learning to Align and Translate (2016). arXiv:1409.0473v7.

[243] Chen, Lifan and Tan, Xiaoqin and Wang, Dingyan and Zhong, Feisheng and Liu, Xiaohong and Yang, Tianbiao and Luo, Xiaomin and Chen, Kaixian and Jiang, Hualiang and Zheng, Mingyue, TransformerCPI: improving compound–protein interaction prediction by sequence-based deep learning with self-attention mechanism and label reversal experiments, Bioinformatics 36 (16) (2020) 4406–4414. doi:10.1093/bioinformatics/btaa524.

[244] Huang, Kexin and Xiao, Cao and Glass, Lucas M. and Sun, Jimeng, MolTrans: Molecular Interaction Transformer for drug–target interaction prediction, Bioinformatics 37 (6) (2021) 830–836. doi:10.1093/bioinformatics/btaa880.

[245] Zhao, Qichang and Zhao, Haochen and Zheng, Kai and Wang, Jianxin, HyperAttentionDTI: improving drug–protein interaction prediction by sequence-based deep learning with attention mechanism, Bioinformatics 38 (3) (2022) 655–662. doi:10.1093/bioinformatics/btab715.

[246] Tomas Mikolov and Kai Chen and Greg Corrado and Jeffrey Dean, Efficient Estimation of Word Representations in Vector Space (2013). arXiv:1301.3781v3.

[247] Tomas Mikolov and Ilya Sutskever and Kai Chen and Greg Corrado and Jeffrey Dean, Distributed Representations of Words and Phrases and their Compositionality (2013). arXiv:1310.4546.

[248] Yichen Gong and Heng Luo and Jian Zhang, Natural Language Inference over Interaction Space (2018). arXiv:1709.04348v2.

[249] Ballester, Pedro J. and Mitchell, John B. O., A machine learning approach to predicting protein-ligand binding affinity with applications

to molecular docking, Bioinformatics 26 (9) (2010) 1169–1175. doi:10.1093/bioinformatics/btq112.

[250] Ballester, Pedro J. and Schreyer, Adrian and Blundell, Tom L., Does a More Precise Chemical Description of Protein-Ligand Complexes Lead to More Accurate Prediction of Binding Affinity?, Journal of Chemical Information and Modeling 54 (3) (2014) 944–955. doi:10.1021/ci500091r.

[251] Durrant, Jacob D. and McCammon, J. Andrew, NNScore: A Neural-Network-Based Scoring Function for the Characterization of Protein-Ligand Complexes, Journal of Chemical Information and Modeling 50 (10) (2010) 1865–1871. doi:10.1021/ci100244v.

[252] Durrant, Jacob D. and McCammon, J. Andrew, NNScore 2.0: A Neural-Network Receptor-Ligand Scoring Function, Journal of Chemical Information and Modeling 51 (11) (2011) 2897–2903. doi:10.1021/ci2003889.

[253] Li, Guo-Bo and Yang, Ling-Ling and Wang, Wen-Jing and Li, Lin-Li and Yang, Sheng-Yong, ID-Score: A New Empirical Scoring Function Based on a Comprehensive Set of Descriptors Related to Protein-Ligand Interactions, Journal of Chemical Information and Modeling 53 (3) (2013) 592–600. doi:10.1021/ci300493w.

[254] Li, Hongjian and Leung, Kwong-Sak and Wong, Man-Hon and Ballester, Pedro J., Substituting random forest for multiple linear regression improves binding affinity prediction of scoring functions: Cyscore as a case study, BMC Bioinformatics 15 (1) (2014) 291. doi:10.1186/1471-2105-15-291.

[255] Kumar, Surendra and Kim, Mi-hyun, SMPLIP-Score: predicting ligand binding affinity from simple and interpretable on-the-fly interaction fingerprint pattern descriptors, Journal of Cheminformatics 13 (1) (2021) 28. doi:10.1186/s13321-021-00507-1.

[256] Meli, Rocco and Anighoro, Andrew and Bodkin, Mike J. and Morris, Garrett M. and Biggin, Philip C., Learning protein-ligand binding affinity with atomic environment vectors, Journal of Cheminformatics 13 (1) (2021) 59. doi:10.1186/s13321-021-00536-w.

[257] Durrant, Jacob D. and McCammon, J. A., BINANA: A novel algorithm for ligand-binding characterization, Journal of Molecular Graphics and Modelling 29 (6) (2011) 888–893. doi:10.1016/j.jmgm.2011.01.004.

[258] Cao, Yang and Li, Lei, Improved protein-ligand binding affinity prediction by using a curvature-dependent surface-area model, Bioinformatics 30 (12) (2014) 1674–1680. doi:10.1093/bioinformatics/btu104.

[259] Berthold, Michael R. and Cebron, Nicolas and Dill, Fabian and Gabriel, Thomas R. and Kötter, Tobias and Meinl, Thorsten and Ohl, Peter and Thiel, Kilian and Wiswedel, Bernd, KNIME - the Konstanz Information Miner: Version 2.0 and Beyond, SIGKDD Explor. Newsl. 11 (1) (2009) 26–31. doi:10.1145/1656274.1656280.

[260] Joseph Gomes and Bharath Ramsundar and Evan N. Feinberg and Vijay S. Pande, Atomic Convolutional Networks for Predicting Protein-Ligand Binding Affinity (2017). arXiv:1703.10603.

[261] Stepniewska-Dziubinska, Marta M. and Zielenkiewicz, Piotr and Siedlecki, Pawel, Development and evaluation of a deep learning model for protein-ligand binding affinity prediction, Bioinformatics 34 (21) (2018) 3666–3674. doi:10.1093/bioinformatics/bty374.

[262] Jiménez, José and Škalič, Miha and Martínez-Rosell, Gerard and De Fabritiis, Gianni, KDEEP: Protein–Ligand Absolute Binding Affinity Prediction via 3D-Convolutional Neural Networks, Journal of Chemical Information and Modeling 58 (2) (2018) 287–296. doi:10.1021/acs.jcim.7b00650.

[263] Jones, Derek and Kim, Hyojin and Zhang, Xiaohua and Zemla, Adam and Stevenson, Garrett and Bennett, W. F. Drew and Kirshner, Daniel and Wong, Sergio E. and Lightstone, Felice C. and Allen, Jonathan E., Improved Protein–Ligand Binding Affinity Prediction with Structure-Based Deep Fusion Inference, Journal of Chemical Information and Modeling 61 (4) (2021) 1583–1592. doi:10.1021/acs.jcim.0c01306.

[264] O'Boyle, Noel M. and Banck, Michael and James, Craig A. and Morley, Chris and Vandermeersch, Tim and Hutchison, Geoffrey R., Open Babel: An open chemical toolbox, Journal of Cheminformatics 3 (1) (2011) 33. doi:10.1186/1758-2946-3-33.

[265] Pahikkala, Tapio and Airola, Antti and Pietilä, Sami and Shakyawar, Sushil and Szwajda, Agnieszka and Tang, Jing and Aittokallio, Tero, Toward more realistic drug-target interaction predictions, Briefings in Bioinformatics 16 (2) (2014) 325–337. doi:10.1093/bib/bbu010.

[266] Shar, Piar Ali and Tao, Weiyang and Gao, Shuo and Huang, Chao and Li, Bohui and Zhang, Wenjuan and Shahen, Mohamed and Zheng, Chunli and Bai, Yaofei and Wang, Yonghua, Pred-binding: large-scale protein-ligand binding affinity prediction, Journal of Enzyme Inhibition and Medicinal Chemistry 31 (6) (2016) 1443–1450. doi:10.3109/14756366.2016.1144594.

[267] He, Tong and Heidemeyer, Marten and Ban, Fuqiang and Cherkasov, Artem and Ester, Martin, SimBoost: a read-across approach for predicting drug-target binding affinities using gradient boosting machines, Journal of Cheminformatics 9. doi:10.1186/s13321-017-0209-z.

[268] Öztürk, Hakime and Özgür, Arzucan and Ozkirimli, Elif, DeepDTA: deep drug-target binding affinity prediction, Bioinformatics 34 (17) (2018) i821–i829. doi:10.1093/bioinformatics/bty593.

[269] Hakime Öztürk and Elif Ozkirimli and Arzucan Özgür, WideDTA: prediction of drug-target binding affinity (2019). arXiv:1902.04166.

[270] Qingyuan Feng and Evgenia Dueva and Artem Cherkasov and Martin Ester, PADME: A Deep Learning-based Framework for Drug-Target Interaction Prediction (2018). arXiv:1807.09741.

[271] Nguyen, Thin and Le, Hang and Quinn, Thomas P and Nguyen, Tri and Le, Thuc Duy and Venkatesh, Svetha, GraphDTA: Predicting drug-target binding affinity with graph neural networks, BioinformaticsBtaa921. doi:10.1093/bioinformatics/btaa921.

[272] Abbasi, Karim and Razzaghi, Parvin and Poso, Antti and Amanlou, Massoud and Ghasemi, Jahan B. and Masoudi-Nejad, Ali, DeepCDA: deep cross-domain compound–protein affinity prediction through LSTM and convolutional neural networks, Bioinformatics 36 (17) (2020) 4633–4642. doi:10.1093/bioinformatics/btaa544.

[273] Shim, Jooyong and Hong, Zhen-Yu and Sohn, Insuk and Hwang, Changha, Prediction of drug–target binding affinity using similarity-based convolutional neural network, Scientific Reports 11 (1) (2021) 4416. doi:10.1038/s41598-021-83679-y.

[274] Wang, Kaili and Zhou, Renyi and Li, Yaohang and Li, Min, DeepDTAF: a deep learning method to predict protein–ligand binding affinity, Briefings in Bioinformatics 22 (5) (2021) bbab072. doi:10.1093/bib/bbab072.

[275] Rifaioglu, A. S. and Cetin Atalay, R. and Cansen Kahraman, D. and Doğan, T. and Martin, M. and Atalay, V., MDeePred: novel multi-channel protein featurization for deep learning-based binding affinity prediction in drug discovery, Bioinformatics 37 (5) (2021) 693–704. doi:10.1093/bioinformatics/btaa858.

[276] Davis, Mindy I. and Hunt, Jeremy P. and Herrgard, Sanna and Ciceri, Pietro and Wodicka, Lisa M. and Pallares, Gabriel and Hocker, Michael and Treiber, Daniel K. and Zarrinkar, Patrick P., Comprehensive analysis of kinase inhibitor selectivity, Nature Biotechnology 29 (11) (2011) 1046–1051. doi:10.1038/nbt.1990.

[277] Metz, James T. and Johnson, Eric F. and Soni, Niru B. and Merta, Philip J. and Kifle, Lemma and Hajduk, Philip J., Navigating the kinome, Nature Chemical Biology 7 (4) (2011) 200–202. doi:10.1038/nchembio.530.

[278] Tang, Jing and Szwajda, Agnieszka and Shakyawar, Sushil and Xu, Tao and Hintsanen, Petteri and Wennerberg, Krister and Aittokallio, Tero, Making Sense of Large-Scale Kinase Inhibitor Bioactivity Data Sets: A Comparative and Integrative Analysis, Journal of Chemical Information and Modeling 54 (3) (2014) 735–743. doi:10.1021/ci400709d.

[279] Kim, Sunghwan and Han, Lianyi and Yu, Bo and Hähnke, Volker D. and Bolton, Evan E. and Bryant, Stephen H., PubChem structure-activity relationship (SAR) clusters, Journal of cheminformatics 7 (2015) 33–33. doi:10.1186/s13321-015-0070-x.

[280] Roth, Bryan L. and Lopez, Estelle and Patel, Shamil and Kroeze, Wesley K., The Multiplicity of Serotonin Receptors: Uselessly Diverse Molecules or an Embarrassment of Riches?, The Neuroscientist 6 (4) (2000) 252–262. doi:10.1177/107385840000600408.

[281] Sigrist, Christian J. A. and Cerutti, Lorenzo and de Castro, Edouard and Langendijk-Genevaux, Petra S. and Bulliard, Virginie and Bairoch, Amos and Hulo, Nicolas, PROSITE, a protein domain database for functional characterization and annotation, Nucleic Acids Research 38 (suppl_1) (2009) D161–D166. doi:10.1093/nar/gkp885.

[282] Noel O'Boyle and Andrew Dalke, DeepSMILES: An Adaptation of SMILES for Use in Machine-Learning of Chemical Structures (2018). doi:

10.26434/chemrxiv.7097960.v1.

[283] Wozniak, Michal and Wolos, Agnieszka and Modrzyk, Urszula and Górski, Rafal L. and Winkowski, Jan and Bajczyk, Michal and Szymkuc, Sara and Grzybowski, Bartosz A. and Eder, Maciej, Linguistic measures of chemical diversity and the "keywords" of molecular collections, Scientific Reports 8 (1) (2018) 7598. doi:10.1038/s41598-018-25440-6.

[284] Keyulu Xu and Weihua Hu and Jure Leskovec and Stefanie Jegelka, How Powerful are Graph Neural Networks? (2019). arXiv:1810.00826v3.

[285] Petar Veličković and Guillem Cucurull and Arantxa Casanova and Adriana Romero and Pietro Liò and Yoshua Bengio, Graph Attention Networks (2018). arXiv:1710.10903v3.

[286] Rudin, Cynthia, Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, Nature Machine Intelligence 1 (5) (2019) 206–215. doi:10.1038/s42256-019-0048-x.

[287] Goebel, R. and Chander, A. and Holzinger, K. and Lecue, F. and Akata, Z. and Stumpf, S. and Kieseberg, P. and Holzinger, A., Explainable AI: The new 42?, in: CD-MAKE 2018, Vol. 11015, 2018. doi:10.1007/978-3-319-99740-7_21.

[288] Zachary C. Lipton, The Mythos of Model Interpretability (2017). arXiv:1606.03490v3.

[289] Gawehn, Erik and Hiss, Jan A. and Schneider, Gisbert, Deep Learning in Drug Discovery, Molecular Informatics 35 (1) (2016) 3–14. doi:10.1002/minf.201501008.

[290] Zhang, Lu and Tan, Jianjun and Han, Dan and Zhu, Hao, From machine learning to deep learning: progress in machine intelligence for rational drug discovery, Drug Discovery Today 22 (11) (2017) 1680–1685. doi:10.1016/j.drudis.2017.08.010.

[291] Lenselink, Eelke B. and ten Dijke, Niels and Bongers, Brandon and Papadatos, George and van Vlijmen, Herman W. T. and Kowalczyk, Wojtek and IJzerman, Adriaan P. and van Westen, Gerard J. P., Beyond the hype: deep neural networks outperform estab-

lished methods using a ChEMBL bioactivity benchmark set, Journal of Cheminformatics 9 (1) (2017) 45. doi:10.1186/s13321-017-0232-0.

[292] Carpenter, Kristy A. and Cohen, David S. and Jarrell, Juliet T. and Huang, Xudong, Deep learning and virtual drug screening, Future Medicinal Chemistry 10 (21) (2018) 2557–2567. doi:10.4155/fmc-2018-0314.

[293] Merk, Daniel and Friedrich, Lukas and Grisoni, Francesca and Schneider, Gisbert, De Novo Design of Bioactive Small Molecules by Artificial Intelligence, Molecular Informatics 37 (1-2) (2018) 1700153. doi:10.1002/minf.201700153.

[294] Zhavoronkov, Alex and Ivanenkov, Yan A. and Aliper, Alex and Veselov, Mark S. and Aladinskiy, Vladimir A. and Aladinskaya, Anastasiya V. and Terentiev, Victor A. and Polykovskiy, Daniil A. and Kuznetsov, Maksim D. and Asadulaev, Arip and Volkov, Yury and Zholus, Artem and Shayakhmetov, Rim R. and Zhebrak, Alexander and Minaeva, Lidiya I. and Zagribelnyy, Bogdan A. and Lee, Lennart H. and Soll, Richard and Madge, David and Xing, Li and Guo, Tao and Aspuru-Guzik, Alán, Deep learning enables rapid identification of potent DDR1 kinase inhibitors, Nature Biotechnology 37 (9) (2019) 1038–1040. doi:10.1038/s41587-019-0224-x.

[295] Schwaller, Philippe and Gaudin, Théophile and Lányi, Dávid and Bekas, Costas and Laino, Teodoro, "Found in Translation": predicting outcomes of complex organic chemistry reactions using neural sequence-to-sequence models, Chemical Science 9 (28) (2018) 6091–6098. doi:10.1039/C8SC02339E.

[296] Coley, Connor W. and Green, William H. and Jensen, Klavs F., Machine Learning in Computer-Aided Synthesis Planning, Accounts of Chemical Research 51 (5) (2018) 1281–1289. doi:10.1021/acs.accounts.8b00087.

[297] Senior, Andrew W. and Evans, Richard and Jumper, John and Kirkpatrick, James and Sifre, Laurent and Green, Tim and Qin, Chongli and Žídek, Augustin and Nelson, Alexander W. R. and Bridgland, Alex and Penedones, Hugo and Petersen, Stig and Simonyan, Karen and Crossan, Steve and Kohli, Pushmeet and Jones, David T. and Silver, David and Kavukcuoglu, Koray and Hassabis, Demis, Improved protein structure prediction using potentials

from deep learning, Nature 577 (7792) (2020) 706–710. doi:10.1038/s41586-019-1923-7.

[298] Kenji Kawaguchi, Deep Learning without Poor Local Minima (2016). arXiv:1605.07110v3.

[299] Datta, Anupam and Sen, Shayak and Zick, Yair, Algorithmic Transparency via Quantitative Input Influence: Theory and Experiments with Learning Systems, in: 2016 IEEE Symposium on Security and Privacy (SP), 2016, pp. 598–617. doi:10.1109/SP.2016.42.

[300] Guidotti, Riccardo and Monreale, Anna and Ruggieri, Salvatore and Turini, Franco and Giannotti, Fosca and Pedreschi, Dino, A Survey of Methods for Explaining Black Box Models, ACM Comput. Surv. 51 (5). doi:10.1145/3236009.

[301] Hirst, Jonathan D. and King, Ross D. and Sternberg, Michael J. E., Quantitative structure-activity relationships by neural networks and inductive logic programming. I. The inhibition of dihydrofolate reductase by pyrimidines, Journal of Computer-Aided Molecular Design 8 (4) (1994) 405–420. doi:10.1007/BF00125375.

[302] Fiore, M. and Sicurello, F. and Indorato, G., An integrated system to represent and manage medical knowledge, Medinfo 8 Pt 2 (1995) 931–933.

[303] Marcou, G. and Horvath, D. and Solov'ev, V. and Arrault, A. and Vayer, P. and Varnek, A., Interpretability of SAR/QSAR Models of any Complexity by Atomic Contributions, Molecular Informatics 31 (9) (2012) 639–642. doi:10.1002/minf.201100136.

[304] Riniker, Sereina and Landrum, Gregory A., Similarity maps - a visualization strategy for molecular fingerprints and machine-learning methods, Journal of Cheminformatics 5 (1) (2013) 43. doi:10.1186/1758-2946-5-43.

[305] Marchese Robinson, Richard L. and Palczewska, Anna and Palczewski, Jan and Kidley, Nathan, Comparison of the Predictive Performance and Interpretability of Random Forest and Linear Models on Benchmark Data Sets, Journal of Chemical Information and Modeling 57 (8) (2017) 1773–1792. doi:10.1021/acs.jcim.6b00753.

[306] Chen, Ya and Stork, Conrad and Hirte, Steffen and Kirchmair, Johannes, NP-Scout: Machine Learning Approach for the Quantification

and Visualization of the Natural Product-Likeness of Small Molecules (2019). doi:10.3390/biom9020043.

[307] Christmann-Franck, Serge and van Westen, Gerard J. P. and Papadatos, George and Beltran Escudie, Fanny and Roberts, Alexander and Overington, John P. and Domine, Daniel, Unprecedently Large-Scale Kinase Inhibitor Set Enabling the Accurate Prediction of Compound–Kinase Activities: A Way toward Selective Promiscuity by Design?, Journal of Chemical Information and Modeling 56 (9) (2016) 1654–1675. doi:10.1021/acs.jcim.6b00122.

[308] Subramanian, Vigneshwari and Ain, Qurrat Ul and Henno, Helena and Pietilä, Lars-Olof and Fuchs, Julian E. and Prusis, Peteris and Bender, Andreas and Wohlfahrt, Gerd, 3D proteochemometrics: using three-dimensional information of proteins and ligands to address aspects of the selectivity of serine proteases, MedChemComm 8 (5) (2017) 1037–1045. doi:10.1039/C6MD00701E.

[309] Giblin, Kathryn A. and Hughes, Samantha J. and Boyd, Helen and Hansson, Pia and Bender, Andreas, Prospectively Validated Proteochemometric Models for the Prediction of Small-Molecule Binding to Bromodomain Proteins, Journal of Chemical Information and Modeling 58 (9) (2018) 1870–1888. doi:10.1021/acs.jcim.8b00400.

[310] Hariri, Safoura and Ghasemi, Jahan B. and Shirini, Farhad and Rasti, Behnam, Probing the origin of dihydrofolate reductase inhibition via proteochemometric modeling, Journal of Chemometrics 33 (2) (2019) e3090. doi:10.1002/cem.3090.

[311] Schneider, Petra and Schneider, Gisbert, De Novo Design at the Edge of Chaos, Journal of Medicinal Chemistry 59 (9) (2016) 4077–4086. doi:10.1021/acs.jmedchem.5b01849.

[312] Christoph Molnar, Interpretable Machine Learning, 2nd Edition, 2022. URL https://christophm.github.io/interpretable-ml-book

[313] Guidotti, Riccardo and Monreale, Anna and Ruggieri, Salvatore and Turini, Franco and Giannotti, Fosca and Pedreschi, Dino, A Survey of Methods for Explaining Black Box Models, ACM Comput. Surv. 51 (5). doi:10.1145/3236009.

[314] Murdoch, W. James and Singh, Chandan and Kumbier, Karl and Abbasi-Asl, Reza and Yu, Bin, Definitions, methods, and applications in interpretable machine learning, Proceedings of the National Academy of Sciences 116 (44) (2019) 22071–22080. doi:10.1073/pnas.1900654116.

[315] Alejandro Barredo Arrieta and Natalia Díaz-Rodríguez and Javier Del Ser and Adrien Bennetot and Siham Tabik and Alberto Barbado and Salvador García and Sergio Gil-López and Daniel Molina and Richard Benjamins and Raja Chatila and Francisco Herrera, Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI (2019). arXiv:1910.10045v2.

[316] Lapuschkin, Sebastian and Wäldchen, Stephan and Binder, Alexander and Montavon, Grégoire and Samek, Wojciech and Müller, Klaus-Robert, Unmasking Clever Hans predictors and assessing what machines really learn, Nature Communications 10 (1) (2019) 1096. doi:10.1038/s41467-019-08987-4.

[317] Q. Vera Liao and Daniel Gruen and Sarah Miller, Questioning the AI: Informing Design Practices for Explainable AI User Experiences, in: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, ACM, 2020. doi:10.1145/3313831.3376590.

[318] Gunning, David and Aha, David, DARPA's Explainable Artificial Intelligence (XAI) Program, AI Magazine 40 (2) (2019) 44–58. doi:10.1609/aimag.v40i2.2850.

[319] Ciallella, Heather L. and Zhu, Hao, Advancing Computational Toxicology in the Big Data Era by Artificial Intelligence: Data-Driven and Mechanism-Driven Modeling for Chemical Toxicity, Chemical Research in Toxicology 32 (4) (2019) 536–547. doi:10.1021/acs.chemrestox.8b00393.

[320] Erico Tjoa and Cuntai Guan, A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI, IEEE Transactions on Neural Networks and Learning Systems 32 (11) (2021) 4793–4813. doi:10.1109/tnnls.2020.3027314.

[321] Leilani H. Gilpin and David Bau and Ben Z. Yuan and Ayesha Bajwa and Michael Specter and Lalana Kagal, Explaining Ex-

planations: An Overview of Interpretability of Machine Learning (2019). arXiv:1806.00069v3.

[322] Lydia P. Gleaves and Reva Schwartz and David A. Broniatowski, The Role of Individual User Differences in Interpretable and Explainable Machine Learning Systems (2020). arXiv:2009.06675.

[323] James, Gareth and Witten, Daniela and Hastie, Trevor and Tibshirani, Robert, An Introduction to Statistical Learning: with Applications in R, Springer New York, New York, NY, 2013, pp. 59–126.

[324] Lou, Yin and Caruana, Rich and Gehrke, Johannes, Intelligible Models for Classification and Regression, in: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '12, Association for Computing Machinery, New York, NY, USA, 2012, p. 150–158. doi:10.1145/2339530.2339556.

[325] Kelvin Xu and Jimmy Ba and Ryan Kiros and Kyunghyun Cho and Aaron Courville and Ruslan Salakhutdinov and Richard Zemel and Yoshua Bengio, Show, Attend and Tell: Neural Image Caption Generation with Visual Attention (2015). arXiv:arXiv:1502.03044.

[326] Kyle Yingkai Gao and Achille Fokoue and Heng Luo and Arun Iyengar and Sanjoy Dey and Ping Zhang, Interpretable Drug Target Prediction Using Deep Neural Representation, in: Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18, 2018, pp. 3371–3377. doi:10.24963/ijcai.2018/468.

[327] Thibault Laugel and Marie-Jeanne Lesot and Christophe Marsala and Xavier Renard and Marcin Detyniecki, The Dangers of Post-hoc Interpretability: Unjustified Counterfactual Explanations (2019). arXiv:1907.09294.

[328] Marco Tulio Ribeiro and Sameer Singh and Carlos Guestrin, "Why Should I Trust You?": Explaining the Predictions of Any Classifier (2016). arXiv:1602.04938v3.

[329] Lundberg, Scott M. and Lee, Su-In, A Unified Approach to Interpreting Model Predictions, in: Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17, Curran Associates Inc., Red Hook, NY, USA, 2017, p. 4768–4777.

[330] Zeiler, Matthew D. and Fergus, Rob, Visualizing and Understanding Convolutional Networks, in: Computer Vision – ECCV 2014, Springer International Publishing, Cham, 2014, pp. 818–833. doi:10.1007/978-3-319-10590-1_53.

[331] Luisa M Zintgraf and Taco S Cohen and Tameem Adel and Max Welling, Visualizing Deep Neural Network Decisions: Prediction Difference Analysis (2017). arXiv:arXiv:1702.04595.

[332] M. D. Zeiler and D. Krishnan and G. W. Taylor and R. Fergus, Deconvolutional networks, in: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2010, pp. 2528–2535. doi:10.1109/CVPR.2010.5539957.

[333] Bach, Sebastian AND Binder, Alexander AND Montavon, Grégoire AND Klauschen, Frederick AND Müller, Klaus-Robert AND Samek, Wojciech, On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation, PLOS ONE 10 (7) (2015) 1–46. doi:10.1371/journal.pone.0130140.

[334] Selvaraju, Ramprasaath R. and Cogswell, Michael and Das, Abhishek and Vedantam, Ramakrishna and Parikh, Devi and Batra, Dhruv, Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization, International Journal of Computer Vision 128 (2) (2020) 336–359. doi:10.1007/s11263-019-01228-7.

[335] Ribeiro, Marco Tulio and Singh, Sameer and Guestrin, Carlos, Anchors: High-Precision Model-Agnostic Explanations, Proceedings of the AAAI Conference on Artificial Intelligence 32 (1). doi:10.1609/aaai.v32i1.11491.

[336] Sandra Wachter and Brent Mittelstadt and Chris Russell, Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR (2018). arXiv:1711.00399v3.

[337] Arnaud Van Looveren and Janis Klaise, Interpretable Counterfactual Explanations Guided by Prototypes (2020). arXiv:1907.02584v2.

[338] Amit Dhurandhar and Pin-Yu Chen and Ronny Luss and Chun-Chen Tu and Paishun Ting and Karthikeyan Shanmugam and Payel Das, Explanations based on the Missing: Towards Contrastive Explanations with Pertinent Negatives (2018). arXiv:1802.07623v2.

[339] Avanti Shrikumar and Peyton Greenside and Anshul Kundaje, Learning Important Features Through Propagating Activation Differences, in: Proceedings of the 34th International Conference on Machine Learning, Vol. 70 of Proceedings of Machine Learning Research, PMLR, 2017, pp. 3145–3153.

[340] Lundberg, Scott M. and Erion, Gabriel and Chen, Hugh and DeGrave, Alex and Prutkin, Jordan M. and Nair, Bala and Katz, Ronit and Himmelfarb, Jonathan and Bansal, Nisha and Lee, Su-In, From local explanations to global understanding with explainable AI for trees, Nature Machine Intelligence 2 (1) (2020) 56–67. doi:10.1038/s42256-019-0138-9.

[341] Monteiro, Nelson R. C. and Simões, Carlos J. V. and Ávila, Henrique V. and Abbasi, Maryam and Oliveira, José L. and Arrais, Joel P., Explainable deep drug–target representations for binding affinity prediction, BMC Bioinformatics 23 (1) (2022) 237. doi:10.1186/s12859-022-04767-y.

[342] Hopkins, Andrew L., Predicting promiscuity, Nature 462 (7270) (2009) 167–168. doi:10.1038/462167a.

[343] Kim, Sunghwan and Chen, Jie and Cheng, Tiejun and Gindulyte, Asta and He, Jia and He, Siqian and Li, Qingliang and Shoemaker, Benjamin A and Thiessen, Paul A and Yu, Bo and Zaslavsky, Leonid and Zhang, Jian and Bolton, Evan E, PubChem in 2021: new data content and improved web interfaces, Nucleic Acids Research 49 (D1) (2020) D1388–D1395. doi:10.1093/nar/gkaa971.

[344] Greg Landrum, RDKit: Open-source cheminformatics (2021).
URL http://www.rdkit.org

[345] Desaphy, Jérémy and Bret, Guillaume and Rognan, Didier and Kellenberger, Esther, sc-PDB: a 3D-database of ligandable binding sites—10 years on, Nucleic Acids Research 43 (D1) (2015) D399–D404. doi:10.1093/nar/gku928.

[346] Altschul, Stephen F. and Madden, Thomas L. and Schäffer, Alejandro A. and Zhang, Jinghui and Zhang, Zheng and Miller, Webb and Lipman, David J., Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, Nucleic Acids Research 25 (17) (1997) 3389–3402. doi:10.1093/nar/25.17.3389.

[347] Camacho, Christiam and Coulouris, George and Avagyan, Vahram and Ma, Ning and Papadopoulos, Jason and Bealer, Kevin and Madden, Thomas L., BLAST+: architecture and applications, BMC Bioinformatics 10 (1) (2009) 421. doi:10.1186/1471-2105-10-421.

[348] Zhou, Bolei and Khosla, Aditya and Lapedriza, Agata and Oliva, Aude and Torralba, Antonio, Learning Deep Features for Discriminative Localization, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2921–2929. doi:10.1109/CVPR.2016.319.

[349] Jost Tobias Springenberg and Alexey Dosovitskiy and Thomas Brox and Martin Riedmiller, Striving for Simplicity: The All Convolutional Net (2015). arXiv:1412.6806v3.

[350] Monteiro, Nelson R.C. and Oliveira, José L. and Arrais, Joel P., DTITR: End-to-end drug–target binding affinity prediction with transformers, Computers in Biology and Medicine 147 (2022) 105772. doi:10.1016/j.compbiomed.2022.105772.

[351] Chun-Fu Chen and Quanfu Fan and Rameswar Panda, CrossViT: Cross-Attention Multi-Scale Vision Transformer for Image Classification (2021). arXiv:2103.14899v2.

[352] Eberhardt, Jerome and Santos-Martins, Diogo and Tillack, Andreas F. and Forli, Stefano, AutoDock Vina 1.2.0: New Docking Methods, Expanded Force Field, and Python Bindings, Journal of Chemical Information and Modeling 61 (8) (2021) 3891–3898. doi:10.1021/acs.jcim.1c00203.

[353] Monteiro, Nelson R.C. and Oliveira, José L. and Arrais, Joel P., TAG-DTA: Binding-region-guided strategy to predict drug-target affinity using transformers, Expert Systems with Applications 238 (2024) 122334. URL https://www.sciencedirect.com/science/article/pii/S0957417423028361

[354] Yang, Jianyi and Roy, Ambrish and Zhang, Yang, BioLiP: a semi-manually curated database for biologically relevant ligand–protein interactions, Nucleic Acids Research 41 (D1) (2013) D1096–D1103. doi:10.1093/nar/gks966.

[355] Cock, Peter J. A. and Antao, Tiago and Chang, Jeffrey T. and Chapman, Brad A. and Cox, Cymon J. and Dalke, Andrew and

Friedberg, Iddo and Hamelryck, Thomas and Kauff, Frank and Wilczynski, Bartek and de Hoon, Michiel J. L., Biopython: freely available Python tools for computational molecular biology and bioinformatics, Bioinformatics 25 (11) (2009) 1422–1423. doi:10.1093/bioinformatics/btp163.

[356] Pemovska, Tea and Bigenzahn, Johannes W. and Superti-Furga, Giulio, Recent advances in combinatorial drug screening and synergy scoring, Current Opinion in Pharmacology 42 (2018) 102–110. doi:10.1016/j.coph.2018.07.008.

[357] Mak, Kit-Kay and Pichika, Mallikarjuna Rao, Artificial intelligence in drug development: present status and future prospects, Drug Discovery Today 24 (3) (2019) 773–780. doi:10.1016/j.drudis.2018.11.014.

[358] Evans, William E. and Relling, Mary V., Moving towards individualized medicine with pharmacogenomics, Nature 429 (6990) (2004) 464–468. doi:10.1038/nature02626.

[359] Spreafico, Roberto and Soriaga, Leah B. and Grosse, Johannes and Virgin, Herbert W. and Telenti, Amalio, Advances in Genomics for Drug Development (2020). doi:10.3390/genes11080942.

[360] Iorio, Francesco and Knijnenburg, Theo A. and Vis, Daniel J. and Bignell, Graham R. and Menden, Michael P. and Schubert, Michael and Aben, Nanne and Gonçalves, Emanuel and Barthorpe, Syd and Lightfoot, Howard and Cokelaer, Thomas and Greninger, Patricia and van Dyk, Ewald and Chang, Han and de Silva, Heshani and Heyn, Holger and Deng, Xianming and Egan, Regina K. and Liu, Qingsong and Mironenko, Tatiana and Mitropoulos, Xeni and Richardson, Laura and Wang, Jinhua and Zhang, Tinghu and Moran, Sebastian and Sayols, Sergi and Soleimani, Maryam and Tamborero, David and Lopez-Bigas, Nuria and Ross-Macdonald, Petra and Esteller, Manel and Gray, Nathanael S. and Haber, Daniel A. and Stratton, Michael R. and Benes, Cyril H. and Wessels, Lodewyk F.A. and Saez-Rodriguez, Julio and McDermott, Ultan and Garnett, Mathew J., A Landscape of Pharmacogenomic Interactions in Cancer, Cell 166 (3) (2016) 740–754. doi:10.1016/j.cell.2016.06.017.

[361] Chiu, Yu-Chiao and Chen, Hung-I Harry and Gorthi, Aparna and Mostavi, Milad and Zheng, Siyuan and Huang, Yufei and Chen, Yidong, Deep learning of pharmacogenomics resources: moving towards precision oncology, Briefings in Bioinformatics 21 (6) (2020) 2066–2083. doi:10.1093/bib/bbz144.

[362] H. Pagés and P. Aboyoun and R. Gentleman and S. DebRoy, Biostrings: Efficient manipulation of biological strings, r package version 2.50.2 (2019).

[363] Pedregosa, Fabian and Varoquaux, Gaël and Gramfort, Alexandre and Michel, Vincent and Thirion, Bertrand and Grisel, Olivier and Blondel, Mathieu and Prettenhofer, Peter and Weiss, Ron and Dubourg, Vincent, Scikit-learn: Machine learning in Python, the Journal of machine Learning research 12 (2011) 2825–2830.

[364] Chan, Wayne W. and Wise, Scott C. and Kaufman, Michael D. and Ahn, Yu Mi and Ensinger, Carol L. and Haack, Torsten and Hood, Molly M. and Jones, Jennifer and Lord, John W. and Lu, Wei Ping and Miller, David and Patt, William C. and Smith, Bryan D. and Petillo, Peter A. and Rutkoski, Thomas J. and Telikepalli, Hanumaiah and Vogeti, Lakshminarayana and Yao, Tony and Chun, Lawrence and Clark, Robin and Evangelista, Peter and Gavrilescu, L. Cristina and Lazarides, Katherine and Zaleskas, Virginia M. and Stewart, Lance J. and Van Etten, Richard A. and Flynn, Daniel L., Crystal structure of human ABL1 kinase domain in complex with DCC-2036 (2011). doi:10.2210/pdb3QRI/pdb.

[365] Canning, Peter and Tan, Li and Chu, Kiki and Lee, Sam W. and Gray, Nathanael S. and Bullock, Alex N., Crystal structure of the human DDR1 kinase domain in complex with imatinib (2013). doi:10.2210/pdb4BKJ/pdb.

[366] Systèmes, Dassault, BIOVIA Discovery Studio (2021).
URL http://accelrys.com

[367] Molecular Graphics Laboratory, AutoDockTools 1.5.6.
URL https://ccsb.scripps.edu/mgltools/

[368] Schrödinger, LLC, The PyMOL Molecular Graphics System, Version 2.5 (May 2021).

[369] Stephen F. Altschul and Warren Gish and Webb Miller and Eugene W. Myers and David J. Lipman, Basic local alignment search tool, Journal of Molecular Biology 215 (3) (1990) 403–410. doi:https://doi.org/10.1016/S0022-2836(05)80360-2.

[370] Lemeer, Simone and Bluwstein, Andrej and Wu, Zhixiang and Leberfinger, Julia and Müller, Konrad and Kramer, Karl and Kuster, Bernhard, Phosphotyrosine mediated protein interactions of the discoidin domain receptor 1, Journal of Proteomics 75 (12) (2012) 3465–3477. doi:10.1016/j.jprot.2011.10.007.

[371] Dan Hendrycks and Kevin Gimpel, Gaussian Error Linear Units (GELUs) (2020). arXiv:1606.08415v4.

[372] Liyuan Liu and Haoming Jiang and Pengcheng He and Weizhu Chen and Xiaodong Liu and Jianfeng Gao and Jiawei Han, On the Variance of the Adaptive Learning Rate and Beyond (2021). arXiv:1908.03265v4.

Appendices

Chapter A

Appendix Background

A.1 Proteins

First letter of codon (5′ end)

Second letter of codon

	U		C		A		G	
U	UUU	Phe	UCU	Ser	UAU	Tyr	UGU	Cys
	UUC	Phe	UCC	Ser	UAC	Tyr	UGC	Cys
	UUA	Leu	UCA	Ser	UAA	Stop	UGA	Stop
	UUG	Leu	UCG	Ser	UAG	Stop	UGG	Trp
C	CUU	Leu	CCU	Pro	CAU	His	CGU	Arg
	CUC	Leu	CCC	Pro	CAC	His	CGC	Arg
	CUA	Leu	CCA	Pro	CAA	Gln	CGA	Arg
	CUG	Leu	CCG	Pro	CAG	Gln	CGG	Arg
A	AUU	Ile	ACU	Thr	AAU	Asn	AGU	Ser
	AUC	Ile	ACC	Thr	AAC	Asn	AGC	Ser
	AUA	Ile	ACA	Thr	AAA	Lys	AGA	Arg
	AUG	Met	ACG	Thr	AAG	Lys	AGG	Arg
G	GUU	Val	GCU	Ala	GAU	Asp	GGU	Gly
	GUC	Val	GCC	Ala	GAC	Asp	GGC	Gly
	GUA	Val	GCA	Ala	GAA	Glu	GGA	Gly
	GUG	Val	GCG	Ala	GAG	Glu	GGG	Gly

Figure A.1: Genetic code: mapping of triplets of nucleotides (codons) in the mRNA to specific amino acids. The initiation and termination codons are highlighted in green and pink, respectively. Figure adapted from *Lehninger Principles of Biochemistry, 5th Edition* [46].

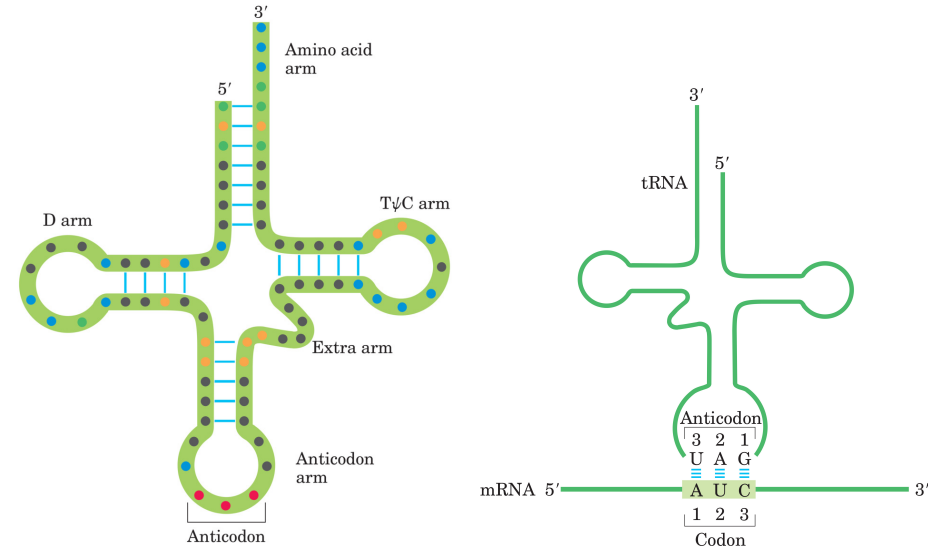


Figure A.2: Transfer RNA. a) General cloverleaf secondary structure of tRNA: D arm, anticodon arm, extra arm, TψC arm, and amino acid arm. b) Pairing relationship of codon and anticodon: complementary base pairing between the codon on the mRNA and the anticodon on the tRNA. Figures adapted from *Lehninger Principles of Biochemistry, 5th Edition* [46].

Table A.1: Standard amino acids.

Amino Acid	3-letter	1-letter	Occurrence in proteins(%)
Alanine	Ala	A	9.06
Arginine	Arg	R	5.84
Asparagine	Asn	N	3.79
Aspartate	Asp	D	5.47
Cysteine	Cys	C	1.28
Glutamine	Gln	Q	3.79
Glutamate	Glu	E	6.24
Glycine	Gly	G	7.29
Histidine	His	H	2.21
Isoleucine	Ile	I	5.55
Leucine	Leu	L	9.87
Lysine	Lys	K	4.92
Methionine	Met	M	2.33
Phenylalanine	Phe	F	3.89
Proline	Pro	P	4.97
Serine	Ser	S	6.78
Threonine	Thr	T	5.54
Tryptophan	Trp	W	1.30
Tyrosine	Tyr	Y	2.88
Valine	Val	V	6.88

*Occurrence percentage values extracted from the UniProt database [57].

Table A.2: Uncommon amino acids and placeholders.

Amino Acid	3-letter	1-letter	Amino acids included
Selenocysteine	Sec	U	-
Pyrrolysine	Pyl	O	-
Any/Unknown	Xaa	X	All
Asparagine/Aspartate	Asx	B	D,N
Glutamine/Glutamate	Glx	Z	E,Q
Leucine/Isoleucine	Xle	J	I,L

Table A.3: Amino acids categories according to polarity and charge of the side chains at pH 7 [46].

Amino Acid	3-letter	1-letter	Occurrence in proteins(%)
Nonpolar, aliphatic R groups			
Glycine	Gly	G	7.29
Alanine	Ala	A	9.06
Proline	Pro	P	4.97
Valine	Val	V	6.88
Leucine	Leu	L	9.87
Isoleucine	Ile	I	5.55
Methionine	Met	M	2.33
Aromatic R groups			
Phenylalanine	Phe	F	3.89
Tyrosine	Tyr	Y	2.88
Tryptophan	Trp	W	1.30
Polar, uncharged R groups			
Serine	Ser	S	6.78
Threonine	Thr	T	5.54
Cysteine	Cys	C	1.28
Asparagine	Asn	N	3.79
Glutamine	Gln	Q	3.79
Positively charged R groups			
Lysine	Lys	K	4.92
Histidine	His	H	2.21
Arginine	Arg	R	5.84
Negatively charged R groups			
Aspartate	Asp	D	5.47
Glutamate	Glu	E	6.24

*Occurrence percentage values extracted from the UniProt database [57].

Table A.4: Amino acids categories according to dipoles and volume of the side chains [55, 56].

Amino Acid	3-letter	1-letter	Occurrence in proteins(%)
		Group 1	
Glycine	Gly	G	7.29
Alanine	Ala	A	9.06
Valine	Val	V	6.88
		Group 2	
Isoleucine	Ile	I	5.55
Leucine	Leu	L	9.87
Phenylalanine	Phe	F	3.89
Proline	Pro	P	4.97
		Group 3	
Tyrosine	Tyr	Y	2.88
Methionine	Met	M	2.33
Threonine	Thr	T	5.54
Serine	Ser	S	6.78
		Group 4	
Histidine	His	H	2.21
Asparagine	Asn	N	3.79
Glutamine	Gln	Q	3.79
Tryptophan	Trp	W	1.30
		Group 5	
Arginine	Arg	R	5.84
Lysine	Lys	K	4.92
		Group 6	
Aspartate	Asp	D	5.47
Glutamate	Glu	E	6.24
		Group 1	
Cysteine	Cys	C	1.28

*Occurrence percentage values extracted from the UniProt database [57].

Table A.5: Amino acids categories according to physicochemical/structural properties of the side chains [57].

Amino Acid	3-letter	1-letter	Occurrence in proteins(%)
Aliphatic R Groups			
Glycine	Gly	G	7.29
Alanine	Ala	A	9.06
Proline	Pro	P	4.97
Valine	Val	V	6.88
Leucine	Leu	L	9.87
Isoleucine	Ile	I	5.55
Aromatic R Groups			
Phenylalanine	Phe	F	3.89
Tyrosine	Tyr	Y	2.88
Tryptophan	Trp	W	1.30
Acidic R Groups			
Aspartate	Asp	D	5.47
Glutamate	Glu	E	6.24
Basic R groups			
Arginine	Arg	R	5.84
Histidine	His	H	2.21
Lysine	Lys	K	4.92
Hydroxylic R groups			
Serine	Ser	S	6.78
Threonine	Thr	T	5.54
Sulphur-containing R groups			
Methionine	Met	M	2.33
Cysteine	Cys	C	1.28
Amidic R groups			
Asparagine	Asn	N	3.79
Glutamine	Gln	Q	3.79

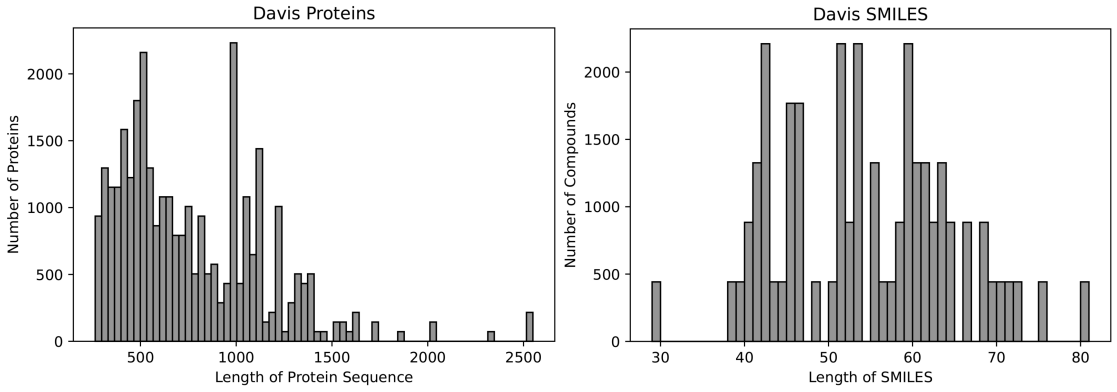
*Occurrence percentage values extracted from the UniProt database [57].

Chapter B

Appendix Explainable Deep Drug–Target Representations

B.1 Supplementary Materials

B.1.1 Davis Kinase Binding Affinity Dataset Distributions



(a) **(b)**

Figure B.1: Davis kinase binding affinity dataset distributions associated with the input vectors. a) Protein sequences length distribution; b) SMILES string length distribution.

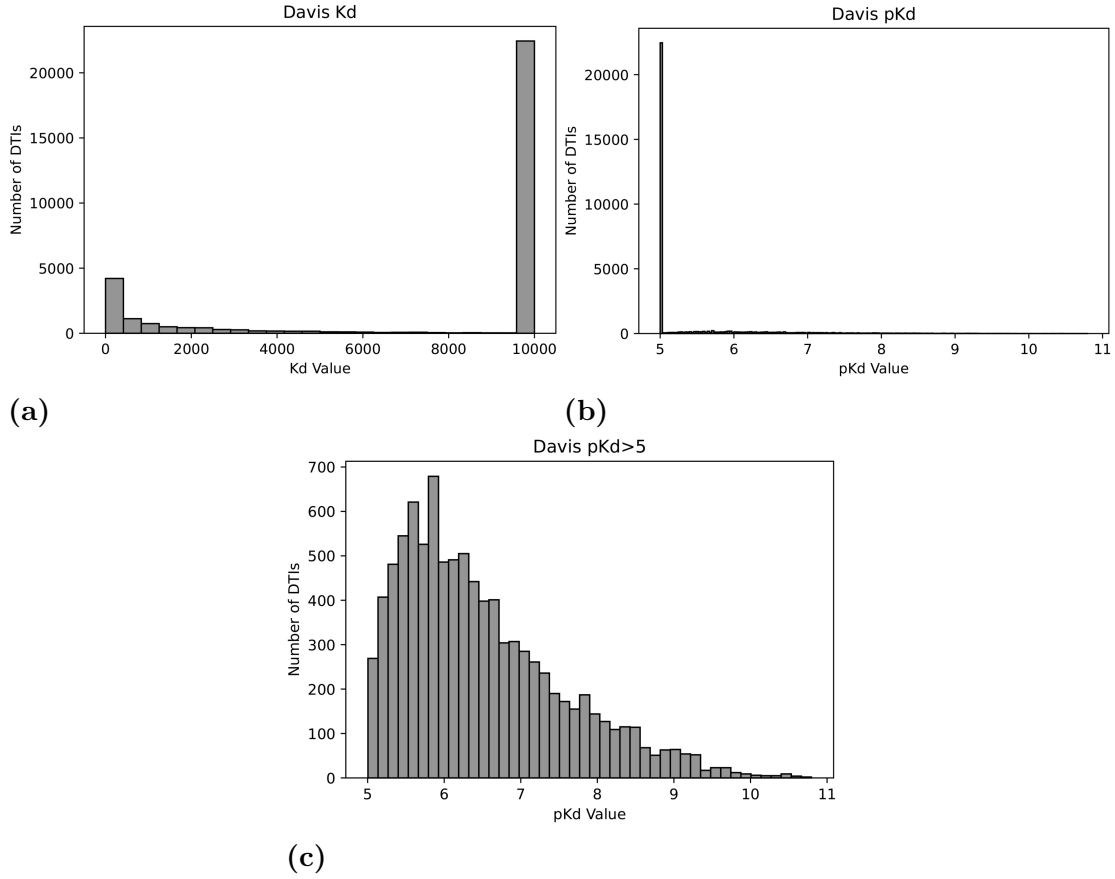


Figure B.2: Davis kinase binding affinity dataset distributions associated with the output (target) vector. a) K_d values distribution; b) pK_d values distribution; c) $pK_d > 5$ values distribution.

B.2 Supplementary Experimental Setup

B.2.1 Binding Affinity Prediction

The optimized architecture and set of parameters for the proposed model were determined by the Chemogenomic K-Fold Cross-Validation methodology, which requires a similarity matrix for all the pairs of protein sequences and SMILES strings. The similarity for the protein pairs was obtained using the Smith-Waterman algorithm, which is usually applied for local sequence alignment and to determine similar regions between two protein sequences. This method was implemented using the Biostrings R Package [362], where the substitution matrix selected was the BLOSUM62, and the gap penalty for opening and extension was fixed at 10 and 0.5, respectively. Furthermore, the final alignment scores were normalized to a [0,1] range using the approach mentioned in the work of Yamanishi et al. (2008) [216]:

$$SW_{Normalized}(p_1,p_2) = \frac{SW(p_1,p_2)}{\sqrt{SW(p_1,p_1)} * \sqrt{SW(p_2,p_2)}} \tag{B.1}$$

, where SW stands for Smith-Watermann, and p_1 and p_2 for two proteins associated with a certain pair (p_1, p_2). On the other hand, the similarity for the SMILES pairs was determined by the Tanimoto Coefficient, which is a distance metric usually applied to calculate the similarity between two molecules based on their bitmap representation (fingerprints). In order to calculate this coefficient, the SMILES strings were initially converted to the Morgan circular fingerprints with a radius of 3, representing the presence or absence of particular substructures across the bitmap. Morgan fingerprints are similar to ECFPs, and the features generated using this representation are based on the neighborhood (fragments) of each non-hydrogen atom of the molecule up to a certain radius and mapped into integer codes using a hashing procedure. The Tanimoto distance coefficient and the SMILES strings fingerprint transformation were implemented using the RDKit Python package [344]. Consequently, the dataset was split into six different folds, in which one fold was used to evaluate the generalization capacity of the model (independent test set) and the remaining folds for hyperoptimization. Table B.1 summarizes the statistics of the different folds obtained from the Chemogenomic K-Fold Cross-Validation approach.

Table B.1: Number of DTIs for the different Davis train/validation folds and independent test fold.

	DTI	**pKd = 5**	**pKd > 5**
Train/Validation Fold 0	4864	3413	1451
Train/Validation Fold 1	4864	3412	1452
Train/Validation Fold 2	4864	3413	1451
Train/Validation Fold 3	4864	3413	1451
Train/Validation Fold 4	4864	3413	1451
Independent Test Fold	4867	3415	1452

Several parameters were hyperoptimized, including the number of convolutional layers, the number of dense layers, the number of filters for each convolutional layer, the filter length, the number of neurons for each dense layer, the dropout rate, and the optimizer learning rate. A considerable range of possible values was assigned for each hyperparameter and the search was narrowed down to the best parameter values.

The Rectified Linear Unit (ReLU) was selected as the activation function for every layer, except for the final output dense layer which uses a linear activation. This function preserves the main properties of the linear and non-linear activation func-

tions, returning zero if it receives any negative input (non-linear) or the value itself in the case of a positive input (linear). Furthermore, it is simple to compute and easier to optimize with gradient-based methods.

$$f(x) = max(0,x) \tag{B.2}$$

Considering that the proposed model focuses on a regression task, the loss function selected was the MSE, which measures the average squared differences between the predicted values and the real values. Regarding the optimizer function, Adaptive Moment Estimation (Adam) was used to update the network weights in each iteration of the training process. This function, identified as a combination of the RMSprop (Root Mean Square Propagation) and SGD (Stochastic Gradient Descent) with momentum, is an adaptive learning rate optimization algorithm that computes individual learning rates for each parameter.

$$W_t = W_{t\_1} - \eta \frac{\hat{m}_t}{\sqrt{\hat{v}_t + \epsilon}} \tag{B.3}$$

, where W is the weight, η the learning rate, m and v the moving averages, and ϵ a small value to avoid division by zero.

Furthermore, early stopping with a patience of 30 and model checkpoint were also considered in order to avoid potential overfitting, where the RMSE was evaluated at each epoch by these two callbacks. Early stopping allows the interruption of the training process if there is no improvement of the evaluation metric after a chosen number of epochs (patience). On the other hand, model checkpoint saves the best model, including the parameters, for the training run, independently of the finishing epoch. Overall, the hyperparameter combination that provided the best average RMSE score over the validation sets was selected as the best set of parameters to establish the optimized model and evaluate the generalization capacity on the test set.

Table B.2 summarizes the parameter settings for the CNN-FCNN model.

In order to validate the prediction efficiency of the end-to-end deep learning architecture (CNN-FCNN), the performance was evaluated and compared with different state-of-the-art baselines, specifically KronRLS [265], SimBoost [267], Sim-CNN-DTA [273], DeepDTA [268], DeepCDA [272], and all the different formulations of the GraphDTA [271]. The original hyperparameter settings described in each one of these works were applied, except for DeepCDA [272], in which it was necessary to

Table B.2: CNN-FCNN parameter settings.

Parameter	Value
Number of Convolution Layers	3
Number of Dense Layers	3
Number of Filters	[64,64,128]
Filter Length (Proteins)	[4,4,5]
Filter Length (Compounds)	[4,4,5]
Filter Padding	'same'
Hidden Neurons	[1024,512,1024]
Dropout Rate	[0.5,0.1]
Optimizer	Adam
Learning Rate	1e-04
Activation Function (CNN)	ReLU
Activation Function (FCNN)	ReLU
Activation Function (Output)	Linear
Loss Function	MSE
Epochs*	500

*Initial number of epochs to allow convergence of the model, where early stopping and model checkpoint were applied to avoid overfitting.

conduct a hyperparameter search considering that the authors did not provide any reference values.

To further evaluate the efficiency of the CNN deep representations, the performance was compared with RFR, SVR, GBR, and KRR. Scikit-learn [363] was used to implement these models and the parameters were obtained using the Chemogenomic K-Fold Cross-Validation approach. Table B.3 summarizes the parameter settings for the deep representations evaluation baseline models.

The model was developed using Python 3.7.9 and Tensorflow 2.4.1, and the experiments were run on 2.20GHz Intel i7-8750H and GeForce GTX 1060 6GB.

B.2.2 Explainable Binding Affinity Prediction

In order to provide explainability to the predictions, Grad-RAM was applied to the implemented trained model, specifically to the last convolutional layers. Even though Grad-RAM connects the features extracted from the CNNs to the input domain, the sole visualization of the input regions that had a positive influence on the prediction does not provide enough explainability without any domain knowledge. Hence, the matching and feature relevance correlation between input regions that

Table B.3: Parameters settings for the deep representations evaluation baseline models. a) RFR; b) KRR; c) SVR; d) GBR.

(a)

Parameters	Value
n_estimators	300
criterion	mse
max_features	auto

(b)

Parameters	Value
alpha	0.01
kernel	poly
degree	5

(c)

Parameters	Value
C	5
kernel	rbf
gamma	scale

(d)

Parameters	Value
n_estimators	900
criterion	friedman_mse
learning_rate	0.1
max_features	None

had a positive influence on the prediction and the spots associated with binding sites or motifs were also assessed and explored.

Additionally, binding sites (and motifs) are mostly non-consecutive and scattered in a 1D representation. Thus, to reasonably evaluate the reliability of the CNNs in the identification of these regions as relevant for prediction, the neighborhood of every single position was also taken into consideration. On that account, for each position p associated with a binding (or motif) region, the resulting pocket is given by an interval $]p - s_w, p + s_w[$, where s_w is the size of the window. Nevertheless, the interval is always left or right-bounded in the presence of another binding site (or motif) in order to avoid overlapping.

B.2.2.1 $L_{Grad-RAM}$ Matching

The regression discriminative localization map provides information regarding the regions of the input that positive-influenced the prediction, and their relative importance (weight). On that account, the first evaluation step consisted in verifying if the CNNs are identifying the binding sites as relevant for the prediction of the binding affinity. Different window lengths were considered, specifically ranging from 0 (exact matching) to 5, in order to determine if in these window-based binding pockets, the CNNs are extracting information from at least one position, considering that the binding spots are non-consecutive single positions. Moreover, $L_{Grad-RAM}$ only contains values equal to zero or greater than zero (positive influence). Thus, to evaluate the $L_{Grad-RAM}$ matching, it is necessary to verify if there is at least one value greater than zero in the window-based binding pocket. Overall, this informa-

tion is presented as a matching percentage corresponding to the weighted average of the average number of binding sites, wherein information is being extracted from at least one position, across all the DTI pairs.

$$\sum_{p=1}^{P} \frac{B_p}{\sum_{p=1}^{P} B_p} * \frac{1}{B_p} \sum_{b=1}^{B_p} 1,\ \exists i = 1,...,W : window_b(i) > 0 \tag{B.4}$$

, where P is the number of DTI pairs, B is the number of binding sites associated with a certain DTI pair p, and W is the total length of the window-based pocket.

In the case of the conserved motifs, the $L_{Grad-RAM}$ matching for the positions outside the entire binding region, i.e., from the first to the last binding position, was also evaluated.

B.2.2.2 $L_{Grad-RAM}$ Feature Relevance

In addition to the $L_{Grad-RAM}$ matching, it is critical to understand the significance of the features extracted from the window-based pockets, specifically if these features are in the range of those with the highest positive influence. On that matter, different thresholds of significance, ranging from the 10% to the 70% highest positive-valued features, were defined in order to perceive what percentage of the features extracted from the window-based pocket regions actually fall into these $L_{Grad-RAM}$ feature threshold distributions. Overall, the $L_{Grad-RAM}$ feature relevance is presented as the weighted average of the average number of positive features extracted from the window-based pocket regions that belong to the feature threshold distribution across all the DTI pairs.

$$\begin{aligned} &\sum_{p=1}^{P} \frac{F_p}{\sum_{p=1}^{P} F_p} * \frac{1}{F_P} \sum_{f=1}^{F_P} 1 \iff F_P(f) \in \\ &\{x \in L_p^{GR>0} || \{y \in L_p^{GR>0} | x \leq y\}| \leq \lambda |L_p^{GR>0}|\} \end{aligned} \tag{B.5}$$

, where P is the number of DTI pairs, F is the number of positive features extracted from all the window-based pockets, L_p^{GR} is the regression discriminative localization map, and λ is the significance threshold.

B.3 Supplementary Results

B.3.1 Binding Affinity Prediction

In order to further validate the proposed architecture and increase the fairness in the comparisons with the state-of-the-art models, the binding affinity performance of the CNN-FCNN model was also evaluated and compared using the experimental settings, i.e., the split methodology of the Davis dataset, proposed in these baselines. The protein sequences and SMILES strings were truncated to the maximum lengths defined in Section 6.2.1.1 of Chapter 6, and the hyperparameter combination of Table B.2 was selected to establish the optimized model. Table B.4 reports the average MSE and CI scores over the independent test set using the five different training sets for the Davis dataset.

Table B.4: Binding affinity prediction results over the Davis dataset using the original split methodology.

Method	Protein Rep.	Compound Rep.	↓ MSE	↑ CI
Baseline Methods				
KronRLS [265]	Smith-Waterman	PubChem-Sim	0.379	0.871
Sim-CNN-DTA [273]	Smith-Waterman	PubChem-Sim	0.306	0.855
SimBoost [267]	Smith-Waterman	PubChem-Sim	0.282	0.872
DeepDTA [268]	1D	1D	0.261	0.878
GraphDTA-GCN [271]	1D	Graph	0.254	0.880
DeepCDA [272]	1D	1D	0.248	0.891
GraphDTA-GAT-GCN [271]	1D	Graph	0.245	0.881
GraphDTA-GATNet [271]	1D	Graph	0.232	0.892
GraphDTA-GIN [271]	1D	Graph	0.229	0.893
Proposed Method				
CNN-FCNN	1D	1D	**0.198 (0.003)**	**0.902 (0.002)**

The standard deviations for the proposed architecture are given in parentheses.

B.3.2 $L_{Grad-RAM}$ Matching

B.3.2.1 PSSM Motifs

Table B.5: PSSM Motifs - $L_{Grad-RAM}$ Matching (Equation B.4) for the Davis ∩ sc-PDB pairs across different PSSM thresholds. a) PSSM Threshold ≥ 5; b) PSSM Threshold ≥ 6; c) PSSM Threshold ≥ 7; d) PSSM Threshold ≥ 8; e) PSSM Threshold ≥ 9; f) PSSM Threshold ≥ 10.

(a)

Window Length	GMP-G	GMP-NG	GAP-G	GAP-NG
0	11.28	11.28	11.28	10.37
1	20.26	20.26	20.26	19.07
2	26.52	26.52	26.52	25.11
3	30.01	30.01	30.01	28.59
4	32.20	32.20	32.20	30.73
5	33.62	33.62	33.62	32.12

(b)

Window Length	GMP-G	GMP-NG	GAP-G	GAP-NG
0	12.14	12.14	12.14	11.23
1	23.52	23.52	23.52	22.40
2	30.88	30.88	30.88	29.19
3	35.31	35.31	35.31	33.69
4	38.70	38.70	38.70	37.24
5	41.67	41.67	41.67	40.24

(c)

Window Length	GMP-G	GMP-NG	GAP-G	GAP-NG
0	11.35	11.35	11.35	10.39
1	26.09	26.09	26.09	25.20
2	38.07	38.07	38.07	36.16
3	44.15	44.15	44.15	42.42
4	49.24	49.24	49.24	47.61
5	51.93	51.93	51.93	50.44

(d)

Window Length	GMP-G	GMP-NG	GAP-G	GAP-NG
0	9.41	9.41	9.41	7.87
1	27.53	27.53	27.53	26.69
2	42.49	42.49	42.49	40.17
3	49.79	49.79	49.79	48.10
4	53.23	53.23	53.23	52.25
5	56.25	56.25	56.25	55.41

(e)

Window Length	GMP-G	GMP-NG	GAP-G	GAP-NG
0	9.27	9.27	9.27	6.83
1	26.67	26.67	26.67	25.37
2	40.98	40.98	40.98	37.89
3	47.80	47.80	47.80	45.04
4	52.52	52.52	52.52	51.06
5	59.35	59.35	59.35	58.21

(f)

Window Length	GMP-G	GMP-NG	GAP-G	GAP-NG
0	13.30	13.30	13.30	9.42
1	28.25	28.25	28.25	26.59
2	47.65	47.65	47.65	42.94
3	52.91	52.91	52.91	48.48
4	55.12	55.12	55.12	53.19
5	65.37	65.37	65.37	64.27

Table B.6: PSSM Motifs - $L_{Grad-RAM}$ Matching (Equation B.4) for the Davis ∩ sc-PDB pairs with the motifs inside the entire binding region filtered out across different PSSM thresholds. a) PSSM Threshold ≥ 5; b) PSSM Threshold ≥ 6; c) PSSM Threshold ≥ 7; d) PSSM Threshold ≥ 8; e) PSSM Threshold ≥ 9; f) PSSM Threshold ≥ 10.

(a)

Window Length	GMP-G	GMP-NG	GAP-G	GAP-NG
0	10.38	10.38	10.38	9.50
1	19.05	19.05	19.05	17.95
2	25.69	25.69	25.69	24.54
3	29.72	29.72	29.72	28.51
4	32.37	32.37	32.37	31.11
5	34.09	34.09	34.09	32.78

(b)

Window Length	GMP-G	GMP-NG	GAP-G	GAP-NG
0	10.71	10.71	10.71	9.83
1	21.15	21.15	21.15	20.15
2	28.97	28.97	28.97	27.69
3	34.05	34.05	34.05	32.82
4	37.98	37.98	37.98	36.90
5	41.78	41.78	41.78	40.74

(c)

Window Length	GMP-G	GMP-NG	GAP-G	GAP-NG
0	10.19	10.19	10.19	9.24
1	23.10	23.10	23.10	22.10
2	35.10	35.10	35.10	33.43
3	41.67	41.67	41.67	40.19
4	48.19	48.19	48.19	46.86
5	51.24	51.24	51.24	50.05

(d)

Window Length	GMP-G	GMP-NG	GAP-G	GAP-NG
0	7.41	7.41	7.41	5.78
1	25.84	25.84	25.84	25.02
2	40.74	40.74	40.74	39.02
3	47.52	47.52	47.52	45.80
4	51.58	51.58	51.58	50.77
5	55.37	55.37	55.37	54.65

(e)

Window Length	GMP-G	GMP-NG	GAP-G	GAP-NG
0	7.59	7.59	7.59	4.56
1	25.38	25.38	25.38	23.86
2	39.70	39.70	39.70	36.23
3	47.72	47.72	47.72	44.69
4	53.80	53.80	53.80	52.71
5	62.69	62.69	62.69	61.61

(f)

Window Length	GMP-G	GMP-NG	GAP-G	GAP-NG
0	9.27	9.27	9.27	4.97
1	23.18	23.18	23.18	21.52
2	41.06	41.06	41.06	36.09
3	46.36	46.36	46.36	41.72
4	49.01	49.01	49.01	47.35
5	61.26	61.26	61.26	59.93

Table B.7: PSSM Motifs - $L_{Grad-RAM}$ Matching (Equation B.4) for the sc-PDB pairs across different PSSM thresholds. a) PSSM Threshold ≥ 5; b) PSSM Threshold ≥ 6; c) PSSM Threshold ≥ 7; d) PSSM Threshold ≥ 8; e) PSSM Threshold ≥ 9; f) PSSM Threshold ≥ 10.

(a)

Window Length	GMP-G	GMP-NG	GAP-G	GAP-NG
0	14.14	14.14	14.14	12.84
1	27.11	27.11	27.11	24.96
2	33.28	33.28	33.28	30.82
3	37.10	37.10	37.10	34.54
4	39.21	39.21	39.21	36.61
5	40.78	40.78	40.78	38.15

(b)

Window Length	GMP-G	GMP-NG	GAP-G	GAP-NG
0	14.34	14.34	14.34	13.07
1	29.86	29.86	29.86	27.52
2	38.02	38.02	38.02	35.32
3	43.40	43.40	43.40	40.53
4	47.52	47.52	47.52	44.53
5	49.83	49.83	49.83	46.82

(c)

Window Length	GMP-G	GMP-NG	GAP-G	GAP-NG
0	14.92	14.92	14.92	13.61
1	32.52	32.52	32.52	30.12
2	42.58	42.58	42.58	39.74
3	50.52	50.52	50.52	47.39
4	56.52	56.52	56.52	53.29
5	59.96	59.96	59.96	56.72

(d)

Window Length	GMP-G	GMP-NG	GAP-G	GAP-NG
0	15.42	15.42	15.42	13.97
1	34.91	34.91	34.91	32.80
2	46.80	46.80	46.80	43.90
3	56.62	56.62	56.62	53.22
4	63.70	63.70	63.70	60.15
5	68.68	68.68	68.68	65.28

(e)

Window Length	GMP-G	GMP-NG	GAP-G	GAP-NG
0	18.29	18.29	18.29	16.45
1	36.73	36.73	36.73	34.25
2	50.14	50.14	50.14	47.36
3	62.20	62.20	62.20	58.72
4	70.88	70.88	70.88	67.46
5	76.53	76.53	76.53	73.36

(f)

Window Length	GMP-G	GMP-NG	GAP-G	GAP-NG
0	22.20	22.20	22.20	19.87
1	38.41	38.41	38.41	35.23
2	53.67	53.67	53.67	50.09
3	63.79	63.79	63.79	60.12
4	73.32	73.32	73.32	69.87
5	78.92	78.92	78.92	75.74

Table B.8: PSSM Motifs - $L_{Grad-RAM}$ Matching (Equation B.4) for the sc-PDB pairs with the motifs inside the entire binding region filtered out across different PSSM thresholds. a) PSSM Threshold ≥ 5; b) PSSM Threshold ≥ 6; c) PSSM Threshold ≥ 7; d) PSSM Threshold ≥ 8; e) PSSM Threshold ≥ 9; f) PSSM Threshold ≥ 10.

(a)

Window Length	GMP-G	GMP-NG	GAP-G	GAP-NG
0	12.88	12.88	12.88	11.55
1	25.64	25.64	25.64	23.57
2	32.02	32.02	32.02	29.59
3	36.10	36.10	36.10	33.60
4	38.34	38.34	38.34	35.80
5	40.14	40.14	40.14	37.56

(b)

Window Length	GMP-G	GMP-NG	GAP-G	GAP-NG
0	13.46	13.46	13.46	12.12
1	28.03	28.03	28.03	25.79
2	36.34	36.34	36.34	33.73
3	41.80	41.80	41.80	39.06
4	46.27	46.27	46.27	43.32
5	48.84	48.84	48.84	45.89

(c)

Window Length	GMP-G	GMP-NG	GAP-G	GAP-NG
0	14.27	14.27	14.27	12.81
1	31.16	31.16	31.16	28.76
2	41.08	41.08	41.08	38.30
3	49.27	49.27	49.27	46.19
4	55.93	55.93	55.93	52.75
5	59.96	59.96	59.96	56.72

(d)

Window Length	GMP-G	GMP-NG	GAP-G	GAP-NG
0	13.41	13.41	13.41	11.90
1	33.05	33.05	33.05	30.94
2	44.71	44.71	44.71	41.79
3	54.38	54.38	54.38	51.26
4	61.64	61.64	61.64	58.29
5	67.19	67.19	67.19	63.92

(e)

Window Length	GMP-G	GMP-NG	GAP-G	GAP-NG
0	15.90	15.90	15.90	13.99
1	33.90	33.90	33.90	31.38
2	47.89	47.89	47.89	45.06
3	59.05	59.05	59.05	55.79
4	67.84	67.84	67.84	64.70
5	73.25	73.25	73.25	70.25

(f)

Window Length	GMP-G	GMP-NG	GAP-G	GAP-NG
0	20.42	20.42	20.42	17.78
1	35.56	35.56	35.56	32.37
2	52.01	52.01	52.01	48.41
3	61.26	61.26	61.26	57.84
4	69.91	69.91	69.91	66.79
5	75.38	75.38	75.38	72.31

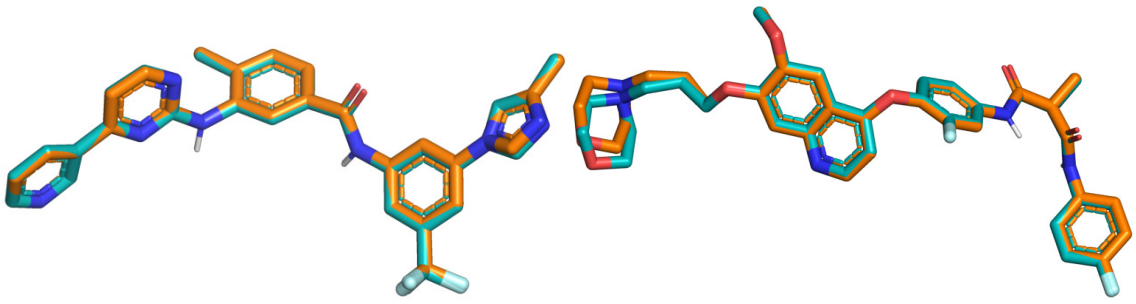
B.3.2.2 3D Interaction Space Analysis (Docking)

Apart from visualizing and exploring the $L_{Grad-RAM}$ matching results in the 1D space and for DTI pairs with binding information known and available, it is essential to validate the reliability of the CNNs in the identification of important regions for binding and the binding sites - $L_{Grad-RAM}$ matching results for DTI pairs without any binding information available, especially in the 3D interaction space. On that account, two DTI pairs with an extremely low absolute prediction error from the Davis testing set, specifically ABL1(E255K)-phosphorylated - SKI-606 and DDR1 - Foretinib, were selected to conduct 3D interaction space analysis.

The 3D structures of the proteins associated with these DTI pairs were collected from the PDB database [186], in which the 3QRI [364] and 4BKJ [365] structures were selected for the ABL1(E255K)-phosphorylated and DDR1 kinases, respectively. These structures were processed using the Discovery Studio Visualizer 4.5 [366] and converted into the PDBQT format using the AutoDockTools 1.5.6 [367]. On the other hand, OpenBabel 3.1.1 [264] was used to generate the 3D coordinates and convert the SMILES strings into the PDB format, and AutoDockTools 1.5.6 [367]

to convert the ligands in the PDB format to the PDBQT format.

In order to generate and predict the 3D receptor-ligand complexes for the two aforementioned DTI pairs, docking experiments were conducted using AutoDock Vina 1.2.0 [352]. The docking process was divided into different stages: blind docking and guided docking. In the blind docking step, search boxes larger than the receptors were employed (increased searching space), specifically 60x60x45 Å for ABL1(E255K)-phosphorylated and 75x55x50 Å for DDR1, and the exhaustiveness was set to 2000. In the guided docking stage, the DoGSiteScorer [161] platform was explored to perform an unbiased assessment of the most likely binding regions (high drug scores) for the ABL1(E255K)-phosphorylated and DDR1 kinases. On that account, the search boxes, with a size of 30x30x30 Å, were centered around the highest-scoring binding pockets for each receptor, respectively, and the exhaustiveness was set to 200. The two docking approaches presented very similar results, in which the RMSD (Root Mean Squared Deviation) between the resulting blind and guided docking best poses was under 0.1 Å for the SKI-606 (ABL1(E255K)-phosphorylated ligand) and was equal to 0.2 Å for the Foretinib (DDR1 ligand). Figure B.3 illustrates the superimposed blind and guided docking best poses for the ligands associated with each receptor, where it is possible to observe that the best binding poses obtained from each docking approach almost completely overlap.



(a) **(b)**

Figure B.3: Overlapped blind and guided docking best poses. a) SKI-606 (ABL1(E255K) - phosphorylated ligand); b) Foretinib (DDR1 ligand). Blind Docking - Blue, Guided Docking - Orange.

Table B.9 reports the blind and guided docking scores measured in terms of kcal/mol (binding affinity) for the best three poses of the ligands associated with each receptor, specifically SKI-606 (ABL1(E255K)-phosphorylated ligand) and Foretinib (DDR1 ligand).

Consistent with visual findings observed in Figure B.3, the docking scores for the best pose of the blind and guided docking methods associated with the ABL1(E255K)-

Table B.9: Blind and guided docking scores, measured in terms of kcal/mol, for the best three poses of the ligands associated with each receptor, specifically SKI-606 (ABL1(E255K)-phosphorylated ligand) and Foretinib (DDR1 ligand).

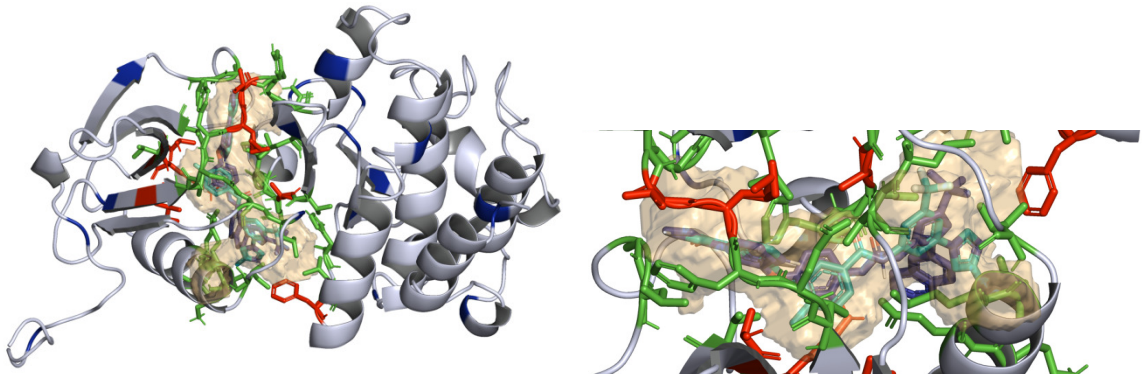
		Binding Affinity (kcal/mol)	
Drug–Target Interaction Pair	**Ligand Pose**	**Blind Docking**	**Guided Docking**
ABL1(E255K)-phosphorylated - SKI-606	**1**	-12.4	-12.4
	2	-11.6	-11.7
	3	-11.1	-11.7
DDR1 - Foretinib	**1**	-10.8	-10.8
	2	-10.7	-10.7
	3	-10.6	-10.7

phosphorylated - SKI-606 and DDR1 - Foretinib interaction pairs, respectively, are in the same strength order.

To further validate the resulting 3D complexes for each one of the selected DTI pairs, the information present in the 3D structures of the receptors in the PDB database [186], i.e., the X-ray crystallography structures of ligands in complex with these receptors (cognates), was used to conduct pocket surface visual evaluations. On that account, it was evaluated if the pocket surface associated with the X-ray crystallography structure of the ligand in complex with each one of the receptors, which is considered a region of high binding probability, contains the docked ligand, i.e., if it falls inside this binding surface. The DoGSiteScorer [161] platform was used to extract the pocket surface, and PyMol [368] for the representation and annotation of these structures.

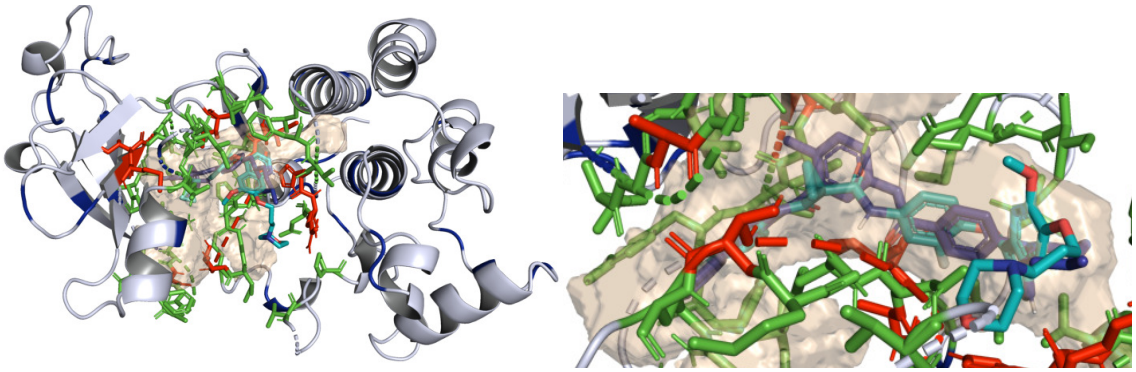
Considering that binding spots are of special interest in the context of this study, PyMol [368] was used to select and identify potential interacting residues within the protein sequences based on a distance threshold of ≤ 5 Å from the docked ligand molecule. However, protein crystal structures from the PDB [186] repository are usually associated with certain fragments of the whole protein 1D amino acid sequence (e.g., protein sequence from the UniProt database [57]). Thus, BLASTP [369] was employed to align the PDB fragments with the protein sequences used to characterize ABL1(E255K)-phosphorylated and DDR1 in the Davis dataset, respectively. The resulting 3D interaction complexes were annotated based on the potential binding sites (≤ 5 Å), $L_{Grad-RAM}$ hits, matched binding-$L_{Grad-RAM}$ positions, and pocket surface.

Figures B.4 and B.5 illustrate the 3D receptor-ligand complexes (both cognate and docked ligand) associated with the ABL1(E255K)-phosphorylated - SKI-606 and DDR1 - Foretinib DTI pairs.



(a) **(b)**

Figure B.4: Annotated 3D structure for the ABL1(E255K)-phosphorylated receptor in complex with the cognate ligand and docked ligand (SKI-606), where the potential binding sites (≤ 5 Å), the $L_{Grad-RAM}$ hits, the matched binding - $L_{Grad-RAM}$ positions, and the pocket surface are represented by the green, blue, red and orange colors, respectively. a) Full representation of the 3D complex; b) Pocket surface in detail. Cognate Ligand - Dark Blue, Docked Ligand - Cyan.



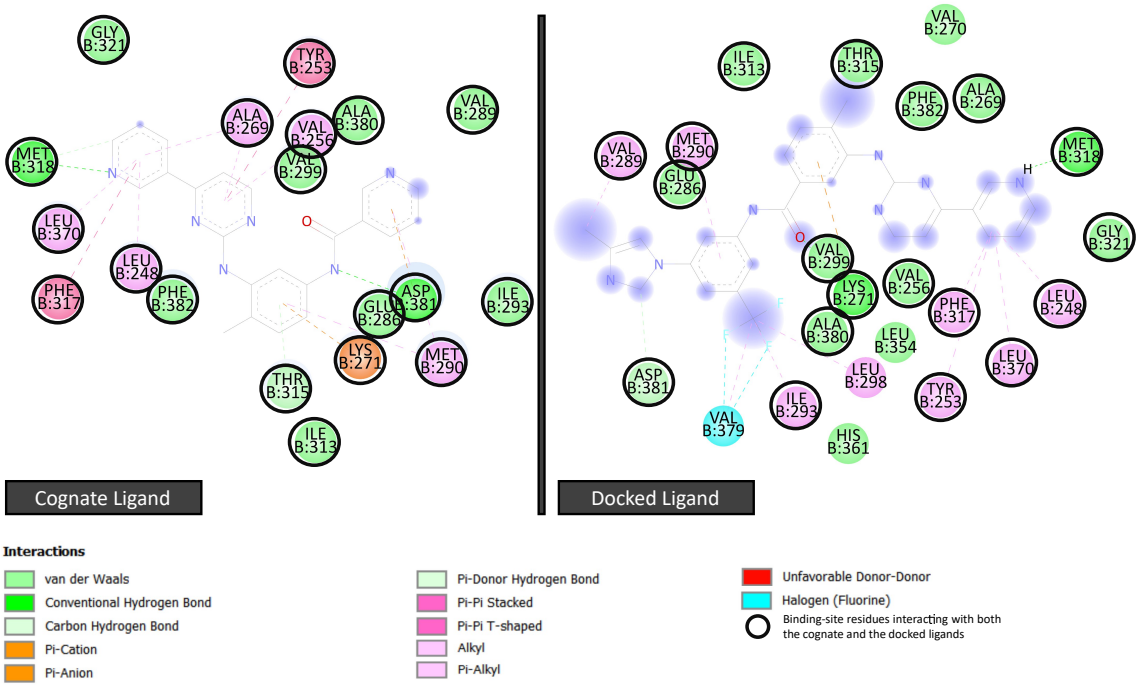
(a) **(b)**

Figure B.5: Annotated 3D structure for the DDR1 receptor in complex with the cognate ligand and docked ligand (Foretinib), where the potential binding sites (≤ 5 Å), the $L_{Grad-RAM}$ hits, the matched binding - $L_{Grad-RAM}$ positions, and the pocket surface are represented by the green, blue, red and orange colors, respectively. a) Full representation of the 3D complex; b) Pocket surface in detail. Cognate Ligand - Dark Blue, Docked Ligand - Cyan.

The visual findings demonstrate that the docked ligand falls inside the pocket surface associated with the cognate ligand for each one of the receptors considered, thus, improving the significance of the binding pose of the docked ligands and the overall docking approach. Moreover, it is possible to observe in Figures B.4b and B.5b that the binding poses of the docked ligand and the cognate ligand in each one of the receptors seem to be correlated. On that account, the correlation of the binding residues associated with each one of these ligands was further evaluated.

Discovery Studio Visualizer 4.5 [366] was used to generate 2D Interaction Diagrams, representing the type of directed bonds between protein and ligand and the interacting protein residues. Figures B.6 and B.7 depict the 2D Interaction Diagrams for the ABL1(E255K)-phosphorylated receptor in complex with the cognate ligand and the docked ligand, and the DDR1 receptor in complex with the cognate ligand and the docked ligand, respectively.

Figure B.6: ABL1(E255K)-phosphorylated 2D Interaction Diagram, in which the binding residues interacting with both the cognate and docked ligands are shown delimited by black circles. a) Cognate Ligand; b) Docked Ligand (SKI-606).

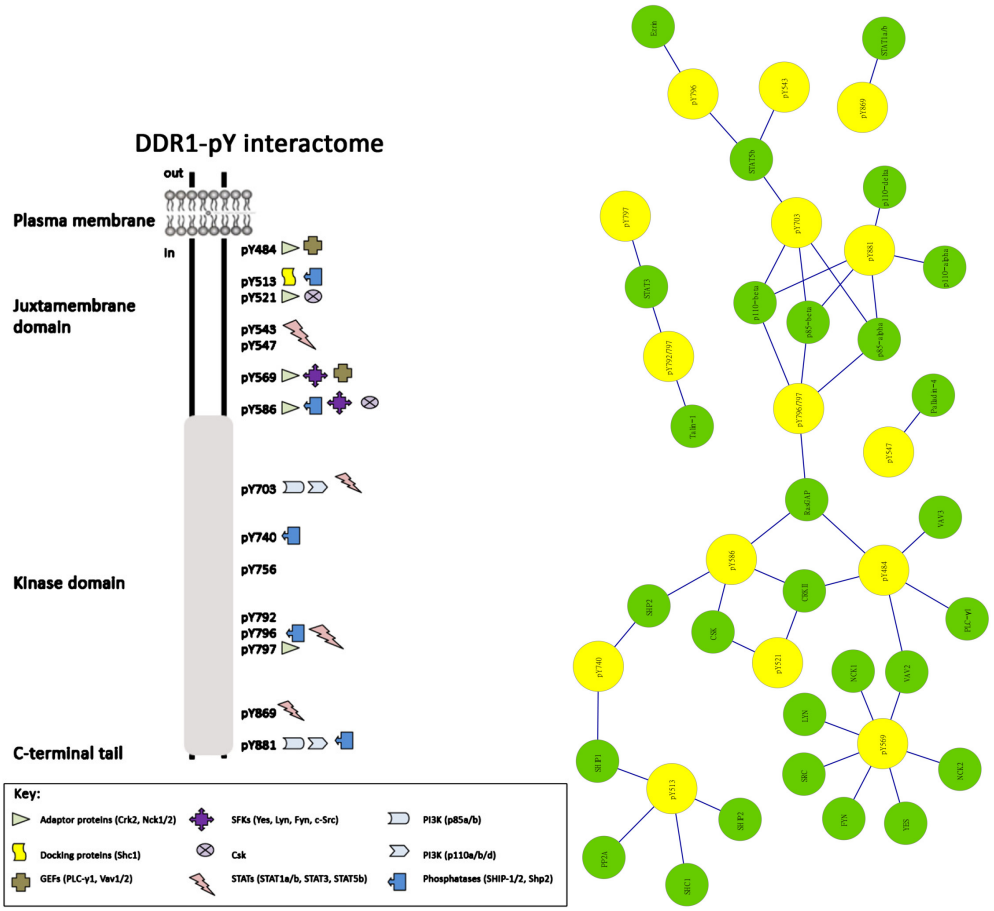


The 2D interaction diagrams for both receptors corroborate the previous visual findings (Figures B.4 and B.5), where it is possible to observe that the majority of the binding residues are interacting with both the cognate and docked ligands. Overall, these results increase the significance of 3D complexes obtained from docking, and the comparisons with the $L_{Grad-RAM}$ hits.

In addition to exploring the resulting 3D complexes for the two selected DTI pairs, and assessing the visual correlation between the binding sites and $L_{Grad-RAM}$ hits, it was relevant to check any potential meaning for the $L_{Grad-RAM}$ hits close to the binding pocket and those not in the vicinity of the binding pocket. In particular, for the DDR1 kinase, which is considered an important therapeutic target due to its implication in pressing contexts, e.g., cancer, it was found that some of these hits are correlated with certain experimental validated critical interacting residues

Figure B.7: DDR1 2D Interaction Diagram, in which the binding residues interacting with both the cognate and docked ligands are shown delimited by black circles. a) Cognate Ligand; b) Docked Ligand (Foretinib).

[370], specifically pY703, pY740, pY756, pY792 and pY869. On that account, the $L_{Grad-RAM}$ hits are matched with pY703 (near main binding pocket), pY740 (far away from main binding pocket), and pY869 (far away from main binding pocket), and nearly matched (1 position away) with pY756 and pY792 (far away from binding pocket). Figure B.8 illustrates the DDR1 kinase domain interactome related to experimental validated critical interacting residues identified in the research work of Lemeer et al. (2012) [370].



(a) (b)

Figure B.8: DDR1 kinase domain interactome. a) Interaction map of DDR1.DDR1-pY interactome based on phosphotyrosine peptide pulldowns performed in human placenta tissue [370]; b) Network map of DDR1 interactions, where the interacting residues and the interactors are represented by the yellow and green colors, respectively.

B.3.3 $L_{Grad-RAM}$ Feature Relevance

B.3.3.1 Binding Sites

Table B.10: Binding Sites - $L_{Grad-RAM}$ Feature Relevance (Equation B.5) for the Davis - sc-PDB pairs across different feature significance thresholds. a) Feature Relevance 10%; b) Feature Relevance 20%; c) Feature Relevance 30%; d) Feature Relevance 40%; e) Feature Relevance 50%; f) Feature Relevance 60%; g) Feature Relevance 70%.

(a)

Window Length	GMP-G	GMP-NG	GAP-G	GAP-NG
0	13.08	13.08	13.08	12.87
1	17.24	17.24	17.24	20.08
2	18.42	18.42	18.42	20.62
3	19.58	19.58	19.58	21.45
4	19.09	19.09	19.09	20.27
5	19.84	19.84	19.84	20.83

(b)

Window Length	GMP-G	GMP-NG	GAP-G	GAP-NG
0	20.56	20.56	20.56	17.82
1	27.59	27.59	27.59	28.03
2	30.00	30.00	30.00	30.79
3	30.30	30.30	30.30	30.92
4	30.08	30.08	30.08	30.96
5	31.52	31.52	31.52	32.50

(c)

Window Length	GMP-G	GMP-NG	GAP-G	GAP-NG
0	33.64	33.64	33.64	30.69
1	39.08	39.08	39.08	39.33
2	44.74	44.74	44.74	45.48
3	44.06	44.06	44.06	44.89
4	43.36	43.36	43.36	44.32
5	43.97	43.97	43.97	45.00

(d)

Window Length	GMP-G	GMP-NG	GAP-G	GAP-NG
0	57.01	57.01	57.01	58.42
1	57.47	57.47	57.47	61.09
2	58.95	58.95	58.95	61.86
3	57.34	57.34	57.34	60.10
4	56.22	56.22	56.22	58.80
5	56.61	56.61	56.61	58.96

(e)

Window Length	GMP-G	GMP-NG	GAP-G	GAP-NG
0	76.64	76.64	76.64	73.27
1	72.03	72.03	72.03	72.80
2	72.37	72.37	72.37	72.88
3	70.63	70.63	70.63	71.32
4	70.33	70.33	70.33	71.05
5	69.84	69.84	69.84	70.62

(f)

Window Length	GMP-G	GMP-NG	GAP-G	GAP-NG
0	84.11	84.11	84.11	85.15
1	78.16	78.16	78.16	79.92
2	79.21	79.21	79.21	81.07
3	77.62	77.62	77.62	79.55
4	76.76	76.76	76.76	78.62
5	75.88	75.88	75.88	77.71

(g)

Window Length	GMP-G	GMP-NG	GAP-G	GAP-NG
0	91.59	91.59	91.59	93.07
1	87.36	87.36	87.36	89.12
2	86.84	86.84	86.84	87.57
3	85.55	85.55	85.55	86.28
4	85.06	85.06	85.06	85.97
5	83.85	83.85	83.85	84.58

Table B.11: Binding Sites - $L_{Grad-RAM}$ Feature Relevance (Equation B.5) for the sc-PDB pairs across different feature significance thresholds: a) Feature Relevance 10%; b) Feature Relevance 20%; c) Feature Relevance 30%; d) Feature Relevance 40%; e) Feature Relevance 50%; f) Feature Relevance 60%; g) Feature Relevance 70%.

(a)

Window Length	GMP-G	GMP-NG	GAP-G	GAP-NG
0	8.48	8.48	8.48	8.42
1	11.83	11.83	11.83	12.42
2	11.85	11.85	11.85	12.65
3	11.97	11.97	11.97	12.86
4	12.11	12.11	12.11	12.90
5	11.82	11.82	11.82	12.54

(b)

Window Length	GMP-G	GMP-NG	GAP-G	GAP-NG
0	20.71	20.71	20.71	20.73
1	22.55	22.55	22.55	23.26
2	21.69	21.69	21.69	22.47
3	22.08	22.08	22.08	22.56
4	22.91	22.91	22.91	23.39
5	22.81	22.81	22.81	23.11

(c)

Window Length	GMP-G	GMP-NG	GAP-G	GAP-NG
0	30.57	30.57	30.57	30.67
1	33.40	33.40	33.40	33.33
2	32.45	32.45	32.45	32.59
3	32.58	32.58	32.58	32.74
4	33.15	33.15	33.15	33.62
5	33.03	33.03	33.03	33.35

(d)

Window Length	GMP-G	GMP-NG	GAP-G	GAP-NG
0	41.81	41.81	41.81	43.63
1	42.65	42.65	42.65	43.48
2	41.53	41.53	41.53	42.46
3	41.52	41.52	41.52	42.36
4	42.42	42.42	42.42	43.30
5	42.57	42.57	42.57	43.16

(e)

Window Length	GMP-G	GMP-NG	GAP-G	GAP-NG
0	51.87	51.87	51.87	52.48
1	52.45	52.45	52.45	52.58
2	51.05	51.05	51.05	51.60
3	51.67	51.67	51.67	52.42
4	52.44	52.44	52.44	53.15
5	52.62	52.62	52.62	53.20

(f)

Window Length	GMP-G	GMP-NG	GAP-G	GAP-NG
0	59.57	59.57	59.57	59.61
1	62.18	62.18	62.18	61.82
2	60.56	60.56	60.56	60.74
3	60.76	60.76	60.76	61.18
4	61.66	61.66	61.66	61.88
5	62.00	62.00	62.00	61.98

(g)

Window Length	GMP-G	GMP-NG	GAP-G	GAP-NG
0	71.01	71.01	71.01	71.27
1	73.18	73.18	73.18	73.26
2	71.27	71.27	71.27	71.90
3	70.98	70.98	70.98	71.65
4	71.55	71.55	71.55	72.14
5	71.79	71.79	71.79	72.14

B.3.3.2 PSSM Motifs

Table B.12: PSSM Motifs - $L_{Grad-RAM}$ Feature Relevance (Equation B.5) for the Davis ∩ sc-PDB pairs across different feature significance thresholds, window lengths, and PSSM Thresholds. a) PSSM Threshold ≥ 5; b) PSSM Threshold ≥ 6; c) PSSM Threshold ≥ 7; d) PSSM Threshold ≥ 8; e) PSSM Threshold ≥ 9; f) PSSM Threshold ≥ 10.

(a)

		GMP-G							GMP-NG							GAP-G							GAP-NG						
		Feature Relevance Threshold																											
		10	20	30	40	50	60	70	10	20	30	40	50	60	70	10	20	30	40	50	60	70	10	20	30	40	50	60	70
Window Length	**0**	11.82	22.55	32.06	44.39	56.21	66.53	75.35	11.82	22.55	32.06	44.39	56.21	66.53	75.35	11.82	22.55	32.06	44.39	56.21	66.53	75.35	12.10	22.36	32.50	44.71	56.82	66.85	76.34
	1	12.17	22.52	32.24	43.10	53.92	63.69	73.30	12.17	22.52	32.24	43.10	53.92	63.69	73.30	12.17	22.52	32.24	43.10	53.92	63.69	73.30	12.47	23.09	33.03	43.65	54.61	64.44	74.04
	2	11.69	21.94	32.05	43.02	53.50	63.53	74.08	11.69	21.94	32.05	43.02	53.50	63.53	74.08	11.69	21.94	32.05	43.02	53.50	63.53	74.08	11.96	22.47	32.48	43.11	54.22	63.83	74.09
	3	10.98	21.13	31.40	41.49	51.70	61.75	72.47	10.98	21.13	31.40	41.49	51.70	61.75	72.47	10.98	21.13	31.40	41.49	51.70	61.75	72.47	11.17	21.72	31.72	41.58	52.34	62.20	72.61
	4	10.53	20.33	30.30	39.95	50.13	59.84	70.17	10.53	20.33	30.30	39.95	50.13	59.84	70.17	10.53	20.33	30.30	39.95	50.13	59.84	70.17	10.40	20.87	30.52	39.99	50.42	60.11	70.20
	5	10.20	20.10	30.19	39.84	50.01	59.66	70.03	10.20	20.10	30.19	39.84	50.01	59.66	70.03	10.20	20.10	30.19	39.84	50.01	59.66	70.03	10.16	20.64	30.35	39.92	50.28	60.05	70.03

(b)

		GMP-G							GMP-NG							GAP-G							GAP-NG						
		Feature Relevance Threshold																											
		10	20	30	40	50	60	70	10	20	30	40	50	60	70	10	20	30	40	50	60	70	10	20	30	40	50	60	70
Window Length	**0**	12.38	23.65	31.90	44.60	56.51	68.57	78.73	12.38	23.65	31.90	44.60	56.51	68.57	78.73	12.38	23.65	31.90	44.60	56.51	68.57	78.73	12.52	22.98	32.08	44.25	57.80	69.30	79.25
	1	13.94	24.21	32.83	43.18	54.20	64.62	73.31	13.94	24.21	32.83	43.18	54.20	64.62	73.31	13.94	24.21	32.83	43.18	54.20	64.62	73.31	14.46	24.04	33.15	43.53	54.87	65.10	73.80
	2	13.33	23.13	31.16	41.38	52.19	63.49	74.46	13.33	23.13	31.16	41.38	52.19	63.49	74.46	13.33	23.13	31.16	41.38	52.19	63.49	74.46	13.93	23.41	31.79	42.02	53.76	63.87	73.93
	3	12.59	21.98	31.02	41.11	51.86	62.92	73.80	12.59	21.98	31.02	41.11	51.86	62.92	73.80	12.59	21.98	31.02	41.11	51.86	62.92	73.80	13.06	22.33	31.60	41.44	52.84	63.15	73.08
	4	12.66	22.55	31.46	41.11	51.66	62.25	73.00	12.66	22.55	31.46	41.11	51.66	62.25	73.00	12.66	22.55	31.46	41.11	51.66	62.25	73.00	12.88	22.78	31.92	41.32	52.39	62.37	72.44
	5	12.10	22.19	31.12	41.25	51.52	61.82	72.12	12.10	22.19	31.12	41.25	51.52	61.82	72.12	12.10	22.19	31.12	41.25	51.52	61.82	72.12	12.30	22.52	31.50	41.19	52.13	61.75	71.56

(c)

		GMP-G							GMP-NG							GAP-G							GAP-NG						
		Feature Relevance Threshold																											
		10	20	30	40	50	60	70	10	20	30	40	50	60	70	10	20	30	40	50	60	70	10	20	30	40	50	60	70
Window Length	**0**	6.23	14.95	23.68	36.76	48.29	64.49	76.01	6.23	14.95	23.68	36.76	48.29	64.49	76.01	6.23	14.95	23.68	36.76	48.29	64.49	76.01	8.16	14.29	24.49	38.10	49.66	65.65	78.91
	1	14.20	25.19	34.57	45.19	56.54	68.52	77.04	14.20	25.19	34.57	45.19	56.54	68.52	77.04	14.20	25.19	34.57	45.19	56.54	68.52	77.04	15.18	25.52	35.21	46.47	57.33	69.11	78.14
	2	15.38	25.87	35.22	46.46	57.54	68.55	78.58	15.38	25.87	35.22	46.46	57.54	68.55	78.58	15.38	25.87	35.22	46.46	57.54	68.55	78.58	15.95	26.15	35.95	47.53	58.95	69.31	79.11
	3	16.44	26.46	35.50	46.36	57.16	67.78	77.67	16.44	26.46	35.50	46.36	57.16	67.78	77.67	16.44	26.46	35.50	46.36	57.16	67.78	77.67	16.70	26.60	36.05	46.73	57.73	68.28	77.48
	4	16.30	27.20	36.11	46.17	57.34	68.03	78.41	16.30	27.20	36.11	46.17	57.34	68.03	78.41	16.30	27.20	36.11	46.17	57.34	68.03	78.41	16.33	27.29	36.63	46.70	58.00	68.34	78.19
	5	15.58	26.33	34.89	45.56	56.69	67.07	77.13	15.58	26.33	34.89	45.56	56.69	67.07	77.13	15.58	26.33	34.89	45.56	56.69	67.07	77.13	15.66	26.60	35.36	45.60	57.24	67.28	76.78

		GMP-G							GMP-NG							GAP-G							GAP-NG						
		Feature Relevance Threshold																											
		10	20	30	40	50	60	70	10	20	30	40	50	60	70	10	20	30	40	50	60	70	10	20	30	40	50	60	70
Window Length	0	5.97	19.40	26.87	38.06	51.49	66.42	73.88	5.97	19.40	26.87	38.06	51.49	66.42	73.88	5.97	19.40	26.87	38.06	51.49	66.42	73.88	7.14	20.54	31.25	42.86	54.46	70.54	81.25
	1	11.45	23.57	33.04	42.73	52.86	67.40	74.23	11.45	23.57	33.04	42.73	52.86	67.40	74.23	11.45	23.57	33.04	42.73	52.86	67.40	74.23	12.65	24.36	33.49	44.03	53.40	68.38	76.58
	2	11.17	23.51	32.21	43.77	54.81	67.92	76.23	11.17	23.51	32.21	43.77	54.81	67.92	76.23	11.17	23.51	32.21	43.77	54.81	67.92	76.23	11.50	23.28	32.26	44.18	56.24	69.14	78.12
	3	12.80	24.97	34.13	45.68	55.78	67.95	76.90	12.80	24.97	34.13	45.68	55.78	67.95	76.90	12.80	24.97	34.13	45.68	55.78	67.95	76.90	13.01	24.69	34.15	45.49	56.17	68.52	78.42
	4	13.63	25.61	35.50	46.70	57.55	68.92	78.30	13.63	25.61	35.50	46.70	57.55	68.92	78.30	13.63	25.61	35.50	46.70	57.55	68.92	78.30	13.46	25.35	35.67	46.64	57.97	69.31	79.45
	5	13.19	25.00	34.57	46.22	56.48	67.05	76.31	13.19	25.00	34.57	46.22	56.48	67.05	76.31	13.19	25.00	34.57	46.22	56.48	67.05	76.31	13.24	24.84	34.48	45.67	56.78	67.32	77.21

(d)

		GMP-G							GMP-NG							GAP-G							GAP-NG						
		Feature Relevance Threshold																											
		10	20	30	40	50	60	70	10	20	30	40	50	60	70	10	20	30	40	50	60	70	10	20	30	40	50	60	70
Window Length	0	10.53	31.58	36.84	52.63	61.40	66.67	71.93	10.53	31.58	36.84	52.63	61.40	66.67	71.93	10.53	31.58	36.84	52.63	61.40	66.67	71.93	14.29	38.10	47.62	59.52	73.81	83.33	90.48
	1	12.12	27.27	35.35	45.96	55.05	66.16	71.72	12.12	27.27	35.35	45.96	55.05	66.16	71.72	12.12	27.27	35.35	45.96	55.05	66.16	71.72	14.61	29.78	37.64	46.07	57.87	69.66	76.97
	2	10.33	20.67	27.96	40.12	51.06	61.40	71.43	10.33	20.67	27.96	40.12	51.06	61.40	71.43	10.33	20.67	27.96	40.12	51.06	61.40	71.43	11.99	22.26	30.48	40.75	55.14	64.73	76.37
	3	11.22	20.70	27.93	38.15	48.88	58.35	70.07	11.22	20.70	27.93	38.15	48.88	58.35	70.07	11.22	20.70	27.93	38.15	48.88	58.35	70.07	12.78	21.67	29.44	38.33	51.39	60.83	73.33
	4	14.15	25.00	31.98	42.83	53.68	63.18	74.61	14.15	25.00	31.98	42.83	53.68	63.18	74.61	14.15	25.00	31.98	42.83	53.68	63.18	74.61	15.50	25.69	33.76	43.31	55.84	64.76	77.71
	5	13.93	26.23	32.79	43.93	53.93	62.30	73.93	13.93	26.23	32.79	43.93	53.93	62.30	73.93	13.93	26.23	32.79	43.93	53.93	62.30	73.93	14.56	27.00	34.28	44.58	55.77	63.59	76.55

(e)

		GMP-G							GMP-NG							GAP-G							GAP-NG						
		Feature Relevance Threshold																											
		10	20	30	40	50	60	70	10	20	30	40	50	60	70	10	20	30	40	50	60	70	10	20	30	40	50	60	70
Window Length	0	12.50	37.50	43.75	56.25	60.42	66.67	70.83	12.50	37.50	43.75	56.25	60.42	66.67	70.83	12.50	37.50	43.75	56.25	60.42	66.67	70.83	17.65	47.06	58.82	67.65	73.53	82.35	91.18
	1	16.94	41.13	50.00	58.87	65.32	76.61	79.84	16.94	41.13	50.00	58.87	65.32	76.61	79.84	16.94	41.13	50.00	58.87	65.32	76.61	79.84	21.30	46.30	55.56	62.04	70.37	82.41	87.96
	2	10.00	24.55	32.27	43.64	54.55	65.00	75.45	10.00	24.55	32.27	43.64	54.55	65.00	75.45	10.00	24.55	32.27	43.64	54.55	65.00	75.45	12.50	27.08	35.94	46.35	60.42	69.79	83.33
	3	9.96	23.11	30.68	40.64	52.59	62.15	72.51	9.96	23.11	30.68	40.64	52.59	62.15	72.51	9.96	23.11	30.68	40.64	52.59	62.15	72.51	12.22	24.89	33.03	42.53	56.56	66.06	79.64
	4	12.38	26.35	33.33	45.08	55.87	65.08	76.19	12.38	26.35	33.33	45.08	55.87	65.08	76.19	12.38	26.35	33.33	45.08	55.87	65.08	76.19	14.18	28.37	35.82	46.45	59.22	67.73	82.62
	5	12.56	28.46	35.13	47.18	57.18	64.87	76.41	12.56	28.46	35.13	47.18	57.18	64.87	76.41	12.56	28.46	35.13	47.18	57.18	64.87	76.41	13.20	30.34	37.08	48.60	59.55	67.13	81.46

(f)

Table B.13: PSSM Motifs - $L_{Grad-RAM}$ Feature Relevance (Equation B.5) for the Davis ∩ sc-PDB pairs with the motifs inside the entire binding region filtered out across different feature significance thresholds, window lengths, and PSSM Thresholds. a) PSSM Threshold ≥ 5; b) PSSM Threshold ≥ 6; c) PSSM Threshold ≥ 7; d) PSSM Threshold ≥ 8; e) PSSM Threshold ≥ 9; f) PSSM Threshold ≥ 10.

(a)

		GMP-G							GMP-NG							GAP-G							GAP-NG						
		Feature Relevance Threshold																											
		10	20	30	40	50	60	70	10	20	30	40	50	60	70	10	20	30	40	50	60	70	10	20	30	40	50	60	70
Window Length	0	9.97	19.13	28.14	39.89	51.78	62.57	71.31	9.97	19.13	28.14	39.89	51.78	62.57	71.31	9.97	19.13	28.14	39.89	51.78	62.57	71.31	9.85	18.51	28.51	39.85	52.39	62.99	72.39
	1	8.32	18.10	28.30	39.20	49.69	60.87	70.86	8.32	18.10	28.30	39.20	49.69	60.87	70.86	8.32	18.10	28.30	39.20	49.69	60.87	70.86	8.36	18.39	28.79	39.34	50.57	61.42	71.51
	2	8.64	18.45	28.79	40.07	50.37	61.35	72.33	8.64	18.45	28.79	40.07	50.37	61.35	72.33	8.64	18.45	28.79	40.07	50.37	61.35	72.33	8.74	18.68	29.03	39.54	51.04	61.55	72.22
	3	8.40	18.13	28.31	38.72	48.61	59.55	70.77	8.40	18.13	28.31	38.72	48.61	59.55	70.77	8.40	18.13	28.31	38.72	48.61	59.55	70.77	8.43	18.50	28.40	38.34	49.27	60.03	70.87
	4	8.15	17.53	27.44	37.36	47.21	57.63	68.30	8.15	17.53	27.44	37.36	47.21	57.63	68.30	8.15	17.53	27.44	37.36	47.21	57.63	68.30	7.80	17.88	27.39	36.89	47.47	57.78	68.21
	5	7.93	17.48	27.60	37.48	47.33	57.75	68.37	7.93	17.48	27.60	37.48	47.33	57.75	68.37	7.93	17.48	27.60	37.48	47.33	57.75	68.37	7.70	17.84	27.43	37.10	47.60	57.96	68.17

(b)

		GMP-G							GMP-NG							GAP-G							GAP-NG						
		Feature Relevance Threshold																											
		10	20	30	40	50	60	70	10	20	30	40	50	60	70	10	20	30	40	50	60	70	10	20	30	40	50	60	70
Window Length	0	10.80	19.25	25.35	37.32	50.00	63.15	74.88	10.80	19.25	25.35	37.32	50.00	63.15	74.88	10.80	19.25	25.35	37.32	50.00	63.15	74.88	10.23	17.39	25.06	35.55	51.41	63.68	74.68
	1	8.95	18.67	27.29	37.57	48.29	60.99	70.50	8.95	18.67	27.29	37.57	48.29	60.99	70.50	8.95	18.67	27.29	37.57	48.29	60.99	70.50	9.07	17.79	26.97	36.98	49.35	61.01	70.67
	2	9.27	18.31	26.01	36.40	47.09	60.09	72.42	9.27	18.31	26.01	36.40	47.09	60.09	72.42	9.27	18.31	26.01	36.40	47.09	60.09	72.42	9.62	18.03	26.28	36.06	48.80	60.18	71.39
	3	8.74	17.25	25.81	36.27	46.96	59.48	71.35	8.74	17.25	25.81	36.27	46.96	59.48	71.35	8.74	17.25	25.81	36.27	46.96	59.48	71.35	8.92	17.27	26.12	35.93	48.01	59.52	70.27
	4	9.47	18.84	27.48	37.42	47.62	59.42	70.60	9.47	18.84	27.48	37.42	47.62	59.42	70.60	9.47	18.84	27.48	37.42	47.62	59.42	70.60	9.38	18.82	27.65	37.09	48.45	59.33	69.70
	5	9.12	18.87	27.68	38.11	48.05	59.53	70.10	9.12	18.87	27.68	38.11	48.05	59.53	70.10	9.12	18.87	27.68	38.11	48.05	59.53	70.10	9.13	19.03	27.73	37.58	48.79	59.13	69.18

(c)

		GMP-G							GMP-NG							GAP-G							GAP-NG						
		Feature Relevance Threshold																											
		10	20	30	40	50	60	70	10	20	30	40	50	60	70	10	20	30	40	50	60	70	10	20	30	40	50	60	70
Window Length	0	3.74	8.41	15.89	27.57	41.12	59.35	74.30	3.74	8.41	15.89	27.57	41.12	59.35	74.30	3.74	8.41	15.89	27.57	41.12	59.35	74.30	4.12	6.70	16.49	26.80	42.27	60.31	77.32
	1	8.65	18.61	28.01	37.22	49.06	64.10	74.62	8.65	18.61	28.01	37.22	49.06	64.10	74.62	8.65	18.61	28.01	37.22	49.06	64.10	74.62	9.24	18.47	28.31	37.95	50.40	64.86	75.90
	2	10.71	20.54	29.02	39.62	50.22	63.28	76.00	10.71	20.54	29.02	39.62	50.22	63.28	76.00	10.71	20.54	29.02	39.62	50.22	63.28	76.00	11.04	20.05	29.29	39.74	51.86	64.23	76.71
	3	11.51	20.56	28.91	39.98	50.53	63.01	74.96	11.51	20.56	28.91	39.98	50.53	63.01	74.96	11.51	20.56	28.91	39.98	50.53	63.01	74.96	11.60	20.39	29.28	39.66	51.36	63.80	74.93
	4	12.36	23.09	31.61	41.75	52.63	65.06	76.54	12.36	23.09	31.61	41.75	52.63	65.06	76.54	12.36	23.09	31.61	41.75	52.63	65.06	76.54	12.15	22.87	31.86	41.80	53.71	65.46	76.42
	5	11.59	22.20	30.40	41.41	52.47	64.39	75.07	11.59	22.20	30.40	41.41	52.47	64.39	75.07	11.59	22.20	30.40	41.41	52.47	64.39	75.07	11.46	22.22	30.57	40.92	53.35	64.60	74.67

		GMP-G							GMP-NG							GAP-G							GAP-NG						
		Feature Relevance Threshold																											
		10	20	30	40	50	60	70	10	20	30	40	50	60	70	10	20	30	40	50	60	70	10	20	30	40	50	60	70
Window Length	0	4.88	12.20	19.51	28.05	41.46	58.54	68.29	4.88	12.20	19.51	28.05	41.46	58.54	68.29	4.88	12.20	19.51	28.05	41.46	58.54	68.29	6.25	12.50	23.44	29.69	43.75	62.50	76.56
	1	8.58	21.60	31.95	41.72	52.37	68.34	74.85	8.58	21.60	31.95	41.72	52.37	68.34	74.85	8.58	21.60	31.95	41.72	52.37	68.34	74.85	9.75	22.01	32.08	42.14	52.83	69.81	77.36
	2	9.01	22.18	31.37	42.46	52.86	66.72	75.91	9.01	22.18	31.37	42.46	52.86	66.72	75.91	9.01	22.18	31.37	42.46	52.86	66.72	75.91	9.61	21.07	31.05	41.77	53.97	68.39	78.00
	3	9.40	21.23	30.77	42.45	51.99	65.38	75.36	9.40	21.23	30.77	42.45	51.99	65.38	75.36	9.40	21.23	30.77	42.45	51.99	65.38	75.36	9.83	20.27	30.26	41.15	51.89	66.26	77.00
	4	10.35	22.74	32.73	44.40	54.75	67.63	77.38	10.35	22.74	32.73	44.40	54.75	67.63	77.38	10.35	22.74	32.73	44.40	54.75	67.63	77.38	10.55	21.86	32.40	43.58	54.89	68.23	78.91
	5	9.26	20.95	30.74	43.26	53.05	64.95	74.63	9.26	20.95	30.74	43.26	53.05	64.95	74.63	9.26	20.95	30.74	43.26	53.05	64.95	74.63	9.63	20.38	30.23	41.97	53.16	65.45	75.86

(d)

		GMP-G							GMP-NG							GAP-G							GAP-NG						
		Feature Relevance Threshold																											
		10	20	30	40	50	60	70	10	20	30	40	50	60	70	10	20	30	40	50	60	70	10	20	30	40	50	60	70
Window Length	0	5.71	11.43	14.29	31.43	45.71	54.29	60.00	5.71	11.43	14.29	31.43	45.71	54.29	60.00	5.71	11.43	14.29	31.43	45.71	54.29	60.00	9.52	9.52	19.05	38.10	57.14	76.19	90.48
	1	12.41	25.52	33.79	45.52	57.24	66.90	72.41	12.41	25.52	33.79	45.52	57.24	66.90	72.41	12.41	25.52	33.79	45.52	57.24	66.90	72.41	15.63	28.13	36.72	46.88	60.94	72.66	80.47
	2	10.97	20.68	27.43	38.40	47.68	57.38	69.20	10.97	20.68	27.43	38.40	47.68	57.38	69.20	10.97	20.68	27.43	38.40	47.68	57.38	69.20	13.59	22.33	30.58	39.81	51.94	62.14	76.21
	3	12.12	20.88	27.61	36.36	45.12	53.54	67.68	12.12	20.88	27.61	36.36	45.12	53.54	67.68	12.12	20.88	27.61	36.36	45.12	53.54	67.68	14.50	22.14	29.39	37.02	47.71	56.87	72.14
	4	14.06	25.00	32.03	41.93	51.30	60.42	72.40	14.06	25.00	32.03	41.93	51.30	60.42	72.40	14.06	25.00	32.03	41.93	51.30	60.42	72.40	16.09	25.57	33.91	43.10	53.45	62.64	76.44
	5	12.47	24.73	31.29	42.23	50.98	58.86	71.33	12.47	24.73	31.29	42.23	50.98	58.86	71.33	12.47	24.73	31.29	42.23	50.98	58.86	71.33	14.08	25.30	32.70	43.44	52.74	60.86	74.70

(e)

		GMP-G							GMP-NG							GAP-G							GAP-NG						
		Feature Relevance Threshold																											
		10	20	30	40	50	60	70	10	20	30	40	50	60	70	10	20	30	40	50	60	70	10	20	30	40	50	60	70
Window Length	0	7.14	14.29	17.86	32.14	39.29	50.00	53.57	7.14	14.29	17.86	32.14	39.29	50.00	53.57	7.14	14.29	17.86	32.14	39.29	50.00	53.57	13.33	13.33	26.67	40.00	46.67	66.67	86.67
	1	17.24	39.08	47.13	56.32	64.37	73.56	75.86	17.24	39.08	47.13	56.32	64.37	73.56	75.86	17.24	39.08	47.13	56.32	64.37	73.56	75.86	23.29	45.21	54.79	61.64	71.23	82.19	87.67
	2	9.87	23.68	29.61	38.16	46.05	56.58	69.08	9.87	23.68	29.61	38.16	46.05	56.58	69.08	9.87	23.68	29.61	38.16	46.05	56.58	69.08	13.28	26.56	34.38	42.19	52.34	62.50	79.69
	3	9.55	21.91	28.09	35.39	44.38	53.93	66.29	9.55	21.91	28.09	35.39	44.38	53.93	66.29	9.55	21.91	28.09	35.39	44.38	53.93	66.29	12.50	23.68	30.92	38.16	48.68	58.55	75.66
	4	9.72	24.07	30.56	40.74	48.61	58.33	69.91	9.72	24.07	30.56	40.74	48.61	58.33	69.91	9.72	24.07	30.56	40.74	48.61	58.33	69.91	12.17	25.93	33.33	43.39	52.38	61.90	78.31
	5	8.36	24.36	30.55	42.18	49.82	57.82	70.55	8.36	24.36	30.55	42.18	49.82	57.82	70.55	8.36	24.36	30.55	42.18	49.82	57.82	70.55	10.12	25.91	32.39	44.53	52.23	60.73	76.92

(f)

Table B.14: PSSM Motifs - $L_{Grad-RAM}$ Feature Relevance (Equation B.5) for the sc-PDB pairs across different feature significance thresholds, window lengths, and PSSM Thresholds. a) PSSM Threshold ≥ 5; b) PSSM Threshold ≥ 6; c) PSSM Threshold ≥ 7; d) PSSM Threshold ≥ 8; e) PSSM Threshold ≥ 9; f) PSSM Threshold ≥ 10.

(a)

		GMP-G							GMP-NG							GAP-G							GAP-NG						
		Feature Relevance Threshold																											
		10	20	30	40	50	60	70	10	20	30	40	50	60	70	10	20	30	40	50	60	70	10	20	30	40	50	60	70
Window Length	0	10.72	21.02	31.02	40.75	50.19	59.68	69.61	10.72	21.02	31.02	40.75	50.19	59.68	69.61	10.72	21.02	31.02	40.75	50.19	59.68	69.61	10.76	20.98	30.92	40.79	50.36	59.41	69.36
	1	10.76	20.98	31.23	40.91	50.78	60.52	70.22	10.76	20.98	31.23	40.91	50.78	60.52	70.22	10.76	20.98	31.23	40.91	50.78	60.52	70.22	10.79	21.08	31.28	41.08	51.20	60.66	70.43
	2	10.34	20.40	30.47	40.17	50.08	59.84	69.55	10.34	20.40	30.47	40.17	50.08	59.84	69.55	10.34	20.40	30.47	40.17	50.08	59.84	69.55	10.37	20.44	30.57	40.27	50.27	59.95	69.69
	3	10.13	20.30	30.33	40.18	50.30	60.01	69.78	10.13	20.30	30.33	40.18	50.30	60.01	69.78	10.13	20.30	30.33	40.18	50.30	60.01	69.78	10.18	20.33	30.39	40.21	50.43	60.07	69.86
	4	9.97	20.12	30.10	39.92	50.11	59.88	69.66	9.97	20.12	30.10	39.92	50.11	59.88	69.66	9.97	20.12	30.10	39.92	50.11	59.88	69.66	9.98	20.12	30.12	39.87	50.15	59.89	69.71
	5	9.82	19.94	29.89	39.70	49.92	59.77	69.68	9.82	19.94	29.89	39.70	49.92	59.77	69.68	9.82	19.94	29.89	39.70	49.92	59.77	69.68	9.84	19.99	29.93	39.73	49.98	59.81	69.76

(b)

		GMP-G							GMP-NG							GAP-G							GAP-NG						
		Feature Relevance Threshold																											
		10	20	30	40	50	60	70	10	20	30	40	50	60	70	10	20	30	40	50	60	70	10	20	30	40	50	60	70
Window Length	0	10.46	20.96	30.77	40.63	50.57	60.08	69.98	10.46	20.96	30.77	40.63	50.57	60.08	69.98	10.46	20.96	30.77	40.63	50.57	60.08	69.98	10.60	20.98	30.75	40.97	51.00	60.11	69.84
	1	11.56	22.01	32.24	41.75	51.78	61.38	70.85	11.56	22.01	32.24	41.75	51.78	61.38	70.85	11.56	22.01	32.24	41.75	51.78	61.38	70.85	11.54	22.13	32.33	41.96	52.19	61.53	71.01
	2	11.08	21.40	31.52	41.08	50.91	60.82	70.35	11.08	21.40	31.52	41.08	50.91	60.82	70.35	11.08	21.40	31.52	41.08	50.91	60.82	70.35	11.14	21.60	31.74	41.21	51.22	60.96	70.54
	3	10.73	21.16	31.14	40.75	50.64	60.49	70.24	10.73	21.16	31.14	40.75	50.64	60.49	70.24	10.73	21.16	31.14	40.75	50.64	60.49	70.24	10.80	21.28	31.17	40.74	50.70	60.49	70.15
	4	10.47	20.79	30.80	40.38	50.31	60.10	69.81	10.47	20.79	30.80	40.38	50.31	60.10	69.81	10.47	20.79	30.80	40.38	50.31	60.10	69.81	10.55	20.93	30.80	40.30	50.33	60.06	69.79
	5	10.26	20.61	30.68	40.30	50.24	60.09	69.89	10.26	20.61	30.68	40.30	50.24	60.09	69.89	10.26	20.61	30.68	40.30	50.24	60.09	69.89	10.34	20.67	30.64	40.18	50.24	60.01	69.83

(c)

		GMP-G							GMP-NG							GAP-G							GAP-NG						
		Feature Relevance Threshold																											
		10	20	30	40	50	60	70	10	20	30	40	50	60	70	10	20	30	40	50	60	70	10	20	30	40	50	60	70
Window Length	0	10.95	20.53	29.56	39.07	49.33	59.45	69.97	10.95	20.53	29.56	39.07	49.33	59.45	69.97	10.95	20.53	29.56	39.07	49.33	59.45	69.97	11.33	20.65	30.08	39.87	49.82	59.34	69.60
	1	12.06	22.15	31.62	40.84	50.65	60.89	70.67	12.06	22.15	31.62	40.84	50.65	60.89	70.67	12.06	22.15	31.62	40.84	50.65	60.89	70.67	12.11	22.17	31.62	40.92	50.75	60.58	70.64
	2	11.74	22.47	32.49	41.70	51.48	61.42	71.16	11.74	22.47	32.49	41.70	51.48	61.42	71.16	11.74	22.47	32.49	41.70	51.48	61.42	71.16	11.91	22.53	32.59	41.89	51.54	61.21	71.17
	3	11.58	22.17	32.01	41.37	51.18	61.13	70.90	11.58	22.17	32.01	41.37	51.18	61.13	70.90	11.58	22.17	32.01	41.37	51.18	61.13	70.90	11.71	22.31	32.01	41.55	51.30	61.01	70.75
	4	11.06	21.62	31.53	40.96	51.05	60.96	70.76	11.06	21.62	31.53	40.96	51.05	60.96	70.76	11.06	21.62	31.53	40.96	51.05	60.96	70.76	11.26	21.85	31.55	41.09	51.06	60.77	70.67
	5	10.78	21.20	31.12	40.58	50.73	60.74	70.49	10.78	21.20	31.12	40.58	50.73	60.74	70.49	10.78	21.20	31.12	40.58	50.73	60.74	70.49	10.95	21.33	31.09	40.64	50.81	60.65	70.45

		GMP-G							GMP-NG							GAP-G							GAP-NG						
		Feature Relevance Threshold																											
		10	20	30	40	50	60	70	10	20	30	40	50	60	70	10	20	30	40	50	60	70	10	20	30	40	50	60	70
Window Length	0	13.30	25.05	36.60	45.67	55.07	64.27	73.94	13.30	25.05	36.60	45.67	55.07	64.27	73.94	13.30	25.05	36.60	45.67	55.07	64.27	73.94	13.49	25.13	37.29	46.63	55.00	64.20	73.09
	1	13.06	24.16	34.40	43.43	53.62	63.50	73.17	13.06	24.16	34.40	43.43	53.62	63.50	73.17	13.06	24.16	34.40	43.43	53.62	63.50	73.17	12.95	23.97	34.19	43.39	53.10	62.76	72.28
	2	12.31	23.07	33.18	42.27	52.23	62.20	72.30	12.31	23.07	33.18	42.27	52.23	62.20	72.30	12.31	23.07	33.18	42.27	52.23	62.20	72.30	12.17	22.99	33.04	42.39	51.95	61.76	71.59
	3	11.83	22.51	32.65	41.88	51.58	61.44	71.47	11.83	22.51	32.65	41.88	51.58	61.44	71.47	11.83	22.51	32.65	41.88	51.58	61.44	71.47	11.74	22.56	32.41	41.96	51.50	61.26	70.87
	4	11.34	22.16	32.32	41.67	51.54	61.33	71.28	11.34	22.16	32.32	41.67	51.54	61.33	71.28	11.34	22.16	32.32	41.67	51.54	61.33	71.28	11.44	22.31	32.17	41.71	51.45	61.12	70.89
	5	11.16	21.85	31.97	41.36	51.28	61.46	71.42	11.16	21.85	31.97	41.36	51.28	61.46	71.42	11.16	21.85	31.97	41.36	51.28	61.46	71.42	11.29	21.96	31.87	41.30	51.22	61.20	70.98

(d)

		GMP-G							GMP-NG							GAP-G							GAP-NG						
		Feature Relevance Threshold																											
		10	20	30	40	50	60	70	10	20	30	40	50	60	70	10	20	30	40	50	60	70	10	20	30	40	50	60	70
Window Length	0	15.36	26.82	38.41	48.46	57.40	65.50	75.14	15.36	26.82	38.41	48.46	57.40	65.50	75.14	15.36	26.82	38.41	48.46	57.40	65.50	75.14	15.84	28.42	40.53	50.31	58.23	65.84	74.84
	1	13.41	23.88	33.96	43.75	53.54	63.21	73.80	13.41	23.88	33.96	43.75	53.54	63.21	73.80	13.41	23.88	33.96	43.75	53.54	63.21	73.80	13.08	23.97	34.18	43.77	52.49	62.52	72.04
	2	13.18	23.23	33.38	43.21	52.99	63.04	73.52	13.18	23.23	33.38	43.21	52.99	63.04	73.52	13.18	23.23	33.38	43.21	52.99	63.04	73.52	12.95	23.57	33.67	43.61	52.51	62.69	72.22
	3	12.64	23.14	33.85	43.51	52.92	62.51	72.41	12.64	23.14	33.85	43.51	52.92	62.51	72.41	12.64	23.14	33.85	43.51	52.92	62.51	72.41	12.61	23.60	33.83	43.89	52.81	62.70	71.67
	4	12.11	23.00	33.51	43.05	52.68	62.25	72.06	12.11	23.00	33.51	43.05	52.68	62.25	72.06	12.11	23.00	33.51	43.05	52.68	62.25	72.06	12.31	23.55	33.70	43.51	52.81	62.53	71.63
	5	11.86	22.98	33.45	43.25	52.79	62.89	72.74	11.86	22.98	33.45	43.25	52.79	62.89	72.74	11.86	22.98	33.45	43.25	52.79	62.89	72.74	11.90	23.29	33.51	43.53	52.83	62.81	72.13

(e)

		GMP-G							GMP-NG							GAP-G							GAP-NG						
		Feature Relevance Threshold																											
		10	20	30	40	50	60	70	10	20	30	40	50	60	70	10	20	30	40	50	60	70	10	20	30	40	50	60	70
Window Length	0	17.14	31.05	42.34	52.62	62.10	69.56	77.42	17.14	31.05	42.34	52.62	62.10	69.56	77.42	17.14	31.05	42.34	52.62	62.10	69.56	77.42	17.57	33.11	44.59	55.18	63.06	69.82	77.70
	1	14.10	26.05	35.95	45.28	54.81	64.33	74.04	14.10	26.05	35.95	45.28	54.81	64.33	74.04	14.10	26.05	35.95	45.28	54.81	64.33	74.04	13.98	26.40	36.34	46.27	54.45	64.08	72.77
	2	13.97	24.35	34.38	43.87	53.54	63.68	73.82	13.97	24.35	34.38	43.87	53.54	63.68	73.82	13.97	24.35	34.38	43.87	53.54	63.68	73.82	13.59	24.72	34.63	45.18	53.40	63.43	72.43
	3	13.89	24.95	36.00	45.48	54.91	64.44	74.09	13.89	24.95	36.00	45.48	54.91	64.44	74.09	13.89	24.95	36.00	45.48	54.91	64.44	74.09	13.98	25.74	36.22	46.59	54.80	64.65	73.34
	4	13.13	24.29	34.75	43.97	53.78	63.27	73.22	13.13	24.29	34.75	43.97	53.78	63.27	73.22	13.13	24.29	34.75	43.97	53.78	63.27	73.22	13.41	25.11	35.21	45.05	53.85	63.72	72.52
	5	13.27	24.90	35.19	44.78	54.34	64.23	74.17	13.27	24.90	35.19	44.78	54.34	64.23	74.17	13.27	24.90	35.19	44.78	54.34	64.23	74.17	13.31	25.34	35.38	45.52	54.35	64.40	73.29

(f)

Table B.15: PSSM Motifs - $L_{Grad-RAM}$ Feature Relevance (Equation B.5) for the sc-PDB pairs with the motifs inside the entire binding region filtered out across different feature significance thresholds, window lengths, and PSSM Thresholds. a) PSSM Threshold ≥ 5; b) PSSM Threshold ≥ 6; c) PSSM Threshold ≥ 7; d) PSSM Threshold ≥ 8; e) PSSM Threshold ≥ 9; f) PSSM Threshold ≥ 10.

(a)

		GMP-G							GMP-NG							GAP-G							GAP-NG						
		Feature Relevance Threshold																											
		10	20	30	40	50	60	70	10	20	30	40	50	60	70	10	20	30	40	50	60	70	10	20	30	40	50	60	70
Window Length	**0**	8.43	18.00	28.45	38.16	47.81	58.08	68.41	8.43	18.00	28.45	38.16	47.81	58.08	68.41	8.43	18.00	28.45	38.16	47.81	58.08	68.41	8.27	17.97	28.11	38.02	48.02	57.81	67.95
	1	9.37	18.93	29.36	39.14	49.35	59.34	69.40	9.37	18.93	29.36	39.14	49.35	59.34	69.40	9.37	18.93	29.36	39.14	49.35	59.34	69.40	9.21	18.86	29.14	38.92	49.53	59.14	69.22
	2	9.06	18.60	28.75	38.61	48.74	58.73	68.85	9.06	18.60	28.75	38.61	48.74	58.73	68.85	9.06	18.60	28.75	38.61	48.74	58.73	68.85	8.94	18.52	28.75	38.40	48.75	58.65	68.77
	3	9.02	18.74	28.75	38.75	49.07	59.02	69.17	9.02	18.74	28.75	38.75	49.07	59.02	69.17	9.02	18.74	28.75	38.75	49.07	59.02	69.17	8.96	18.67	28.73	38.51	48.99	58.90	69.05
	4	8.93	18.68	28.63	38.56	49.00	59.01	69.10	8.93	18.68	28.63	38.56	49.00	59.01	69.10	8.93	18.68	28.63	38.56	49.00	59.01	69.10	8.84	18.63	28.57	38.28	48.85	58.87	68.96
	5	8.82	18.52	28.39	38.30	48.74	58.86	69.08	8.82	18.52	28.39	38.30	48.74	58.86	69.08	8.82	18.52	28.39	38.30	48.74	58.86	69.08	8.72	18.53	28.34	38.11	48.63	58.74	68.98

(b)

		GMP-G							GMP-NG							GAP-G							GAP-NG						
		Feature Relevance Threshold																											
		10	20	30	40	50	60	70	10	20	30	40	50	60	70	10	20	30	40	50	60	70	10	20	30	40	50	60	70
Window Length	**0**	8.35	17.93	27.98	37.52	47.79	58.00	68.37	8.35	17.93	27.98	37.52	47.79	58.00	68.37	8.35	17.93	27.98	37.52	47.79	58.00	68.37	8.29	18.05	27.89	37.79	48.32	58.29	67.98
	1	9.95	19.65	29.93	39.47	50.00	59.93	69.77	9.95	19.65	29.93	39.47	50.00	59.93	69.77	9.95	19.65	29.93	39.47	50.00	59.93	69.77	9.71	19.66	29.89	39.35	50.22	59.95	69.69
	2	9.63	19.38	29.48	39.20	49.31	59.57	69.47	9.63	19.38	29.48	39.20	49.31	59.57	69.47	9.63	19.38	29.48	39.20	49.31	59.57	69.47	9.47	19.46	29.68	39.07	49.50	59.58	69.57
	3	9.50	19.46	29.30	39.06	49.23	59.40	69.55	9.50	19.46	29.30	39.06	49.23	59.40	69.55	9.50	19.46	29.30	39.06	49.23	59.40	69.55	9.37	19.42	29.22	38.82	49.09	59.24	69.33
	4	9.31	19.21	29.21	38.94	49.10	59.20	69.17	9.31	19.21	29.21	38.94	49.10	59.20	69.17	9.31	19.21	29.21	38.94	49.10	59.20	69.17	9.21	19.21	29.08	38.59	48.87	58.94	68.96
	5	9.12	19.09	29.13	38.88	49.02	59.15	69.22	9.12	19.09	29.13	38.88	49.02	59.15	69.22	9.12	19.09	29.13	38.88	49.02	59.15	69.22	9.06	19.06	28.94	38.46	48.75	58.81	68.93

(c)

		GMP-G							GMP-NG							GAP-G							GAP-NG						
		Feature Relevance Threshold																											
		10	20	30	40	50	60	70	10	20	30	40	50	60	70	10	20	30	40	50	60	70	10	20	30	40	50	60	70
Window Length	**0**	8.91	17.71	26.29	34.82	45.68	56.38	67.63	8.91	17.71	26.29	34.82	45.68	56.38	67.63	8.91	17.71	26.29	34.82	45.68	56.38	67.63	8.95	18.09	26.62	35.51	46.22	56.68	67.21
	1	10.38	19.89	29.31	37.88	48.15	58.74	69.06	10.38	19.89	29.31	37.88	48.15	58.74	69.06	10.38	19.89	29.31	37.88	48.15	58.74	69.06	10.04	19.89	29.16	37.65	47.98	58.30	68.68
	2	10.18	20.38	30.44	39.38	49.52	59.87	70.16	10.18	20.38	30.44	39.38	49.52	59.87	70.16	10.18	20.38	30.44	39.38	49.52	59.87	70.16	10.00	20.45	30.48	39.33	49.48	59.64	69.99
	3	10.02	19.97	29.64	38.87	49.13	59.53	69.93	10.02	19.97	29.64	38.87	49.13	59.53	69.93	10.02	19.97	29.64	38.87	49.13	59.53	69.93	9.83	20.09	29.55	38.83	49.05	59.30	69.54
	4	9.62	19.57	29.50	38.83	49.35	59.69	69.92	9.62	19.57	29.50	38.83	49.35	59.69	69.92	9.62	19.57	29.50	38.83	49.35	59.69	69.92	9.50	19.73	29.34	38.70	49.09	59.29	69.55
	5	9.26	18.91	28.73	38.05	48.53	59.07	69.26	9.26	18.91	28.73	38.05	48.53	59.07	69.26	9.26	18.91	28.73	38.05	48.53	59.07	69.26	9.19	19.02	28.52	37.82	48.33	58.72	68.90

		GMP-G							GMP-NG							GAP-G							GAP-NG						
		Feature Relevance Threshold																											
		10	20	30	40	50	60	70	10	20	30	40	50	60	70	10	20	30	40	50	60	70	10	20	30	40	50	60	70
Window Length	0	11.61	22.69	33.37	41.14	51.09	61.04	71.61	11.61	22.69	33.37	41.14	51.09	61.04	71.61	11.61	22.69	33.37	41.14	51.09	61.04	71.61	11.57	23.25	34.00	42.29	51.05	61.57	70.79
	1	12.08	22.42	32.32	40.24	50.86	61.44	71.81	12.08	22.42	32.32	40.24	50.86	61.44	71.81	12.08	22.42	32.32	40.24	50.86	61.44	71.81	11.80	22.43	32.04	40.09	50.06	60.57	70.69
	2	11.04	21.34	31.16	39.65	49.94	60.68	71.35	11.04	21.34	31.16	39.65	49.94	60.68	71.35	11.04	21.34	31.16	39.65	49.94	60.68	71.35	10.66	21.33	30.89	39.44	49.36	60.02	70.43
	3	10.75	20.80	30.58	39.38	49.64	60.27	70.84	10.75	20.80	30.58	39.38	49.64	60.27	70.84	10.75	20.80	30.58	39.38	49.64	60.27	70.84	10.53	20.97	30.18	39.17	49.20	59.75	69.92
	4	10.07	20.51	30.55	39.55	49.96	60.38	70.76	10.07	20.51	30.55	39.55	49.96	60.38	70.76	10.07	20.51	30.55	39.55	49.96	60.38	70.76	10.02	20.75	30.20	39.37	49.52	59.89	70.14
	5	9.78	19.92	29.87	39.06	49.57	60.40	70.75	9.78	19.92	29.87	39.06	49.57	60.40	70.75	9.78	19.92	29.87	39.06	49.57	60.40	70.75	9.81	20.05	29.54	38.77	49.23	59.87	70.11

(d)

		GMP-G							GMP-NG							GAP-G							GAP-NG						
		Feature Relevance Threshold																											
		10	20	30	40	50	60	70	10	20	30	40	50	60	70	10	20	30	40	50	60	70	10	20	30	40	50	60	70
Window Length	0	12.42	23.55	34.48	43.90	53.96	61.88	72.59	12.42	23.55	34.48	43.90	53.96	61.88	72.59	12.42	23.55	34.48	43.90	53.96	61.88	72.59	12.65	25.55	36.74	45.99	55.23	63.50	72.51
	1	11.95	21.72	31.48	40.07	50.25	60.27	71.72	11.95	21.72	31.48	40.07	50.25	60.27	71.72	11.95	21.72	31.48	40.07	50.25	60.27	71.72	12.00	22.77	32.15	40.58	49.30	59.79	69.82
	2	12.66	22.53	31.93	40.86	50.68	61.50	72.58	12.66	22.53	31.93	40.86	50.68	61.50	72.58	12.66	22.53	31.93	40.86	50.68	61.50	72.58	12.57	23.41	32.45	41.37	50.06	61.01	71.21
	3	11.39	21.27	31.30	40.46	50.41	60.90	71.64	11.39	21.27	31.30	40.46	50.41	60.90	71.64	11.39	21.27	31.30	40.46	50.41	60.90	71.64	11.58	22.20	31.45	41.08	50.21	60.83	70.87
	4	10.62	21.17	31.28	40.41	50.64	61.04	71.45	10.62	21.17	31.28	40.41	50.64	61.04	71.45	10.62	21.17	31.28	40.41	50.64	61.04	71.45	10.97	22.07	31.54	41.09	50.66	61.07	70.97
	5	10.20	20.89	30.88	40.22	50.31	61.21	71.56	10.20	20.89	30.88	40.22	50.31	61.21	71.56	10.20	20.89	30.88	40.22	50.31	61.21	71.56	10.41	21.51	31.04	40.64	50.28	60.93	70.86

(e)

		GMP-G							GMP-NG							GAP-G							GAP-NG						
		Feature Relevance Threshold																											
		10	20	30	40	50	60	70	10	20	30	40	50	60	70	10	20	30	40	50	60	70	10	20	30	40	50	60	70
Window Length	0	13.82	27.06	38.24	47.94	59.12	65.59	74.41	13.82	27.06	38.24	47.94	59.12	65.59	74.41	13.82	27.06	38.24	47.94	59.12	65.59	74.41	14.19	29.39	40.88	51.01	60.47	67.23	74.32
	1	12.36	22.33	32.02	40.03	49.72	59.69	70.22	12.36	22.33	32.02	40.03	49.72	59.69	70.22	12.36	22.33	32.02	40.03	49.72	59.69	70.22	12.93	24.13	33.44	42.11	50.00	60.41	68.93
	2	12.52	21.49	31.13	39.85	49.75	60.74	71.74	12.52	21.49	31.13	39.85	49.75	60.74	71.74	12.52	21.49	31.13	39.85	49.75	60.74	71.74	12.26	22.56	31.94	41.32	49.30	60.54	70.10
	3	12.25	22.08	32.74	41.74	51.75	62.03	72.81	12.25	22.08	32.74	41.74	51.75	62.03	72.81	12.25	22.08	32.74	41.74	51.75	62.03	72.81	12.46	23.30	33.10	43.11	51.29	62.14	71.87
	4	11.41	21.44	31.47	40.23	50.46	60.70	71.60	11.41	21.44	31.47	40.23	50.46	60.70	71.60	11.41	21.44	31.47	40.23	50.46	60.70	71.60	11.72	22.48	31.89	41.42	50.17	60.87	70.63
	5	11.15	21.68	31.47	40.59	50.64	61.17	71.79	11.15	21.68	31.47	40.59	50.64	61.17	71.79	11.15	21.68	31.47	40.59	50.64	61.17	71.79	11.31	22.33	31.57	41.39	50.29	61.02	70.60

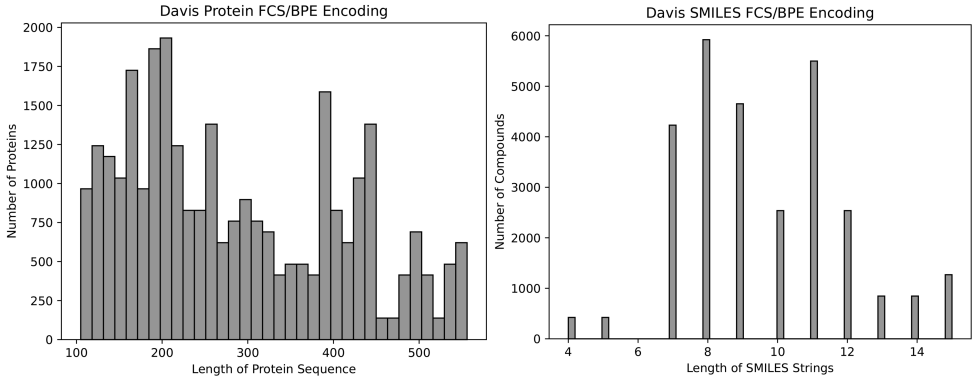
(f)

Chapter C

Appendix Intrinsic Explainability and Drug–Target Multi-Domain Inter-Dependency

C.1 Supplementary Materials

C.1.1 Davis Kinase Binding Affinity Dataset Distributions



(a) **(b)**

Figure C.1: Davis kinase binding affinity dataset distributions. a) Protein sequences length distribution based on the FCS/BPE encoding; SMILES string length distribution based on the FCS/BPE Encoding.

C.2 Supplementary Methods

C.2.1 Sinusoidal Positional Encoding

The position of each token in the input sequence is determinant for the context and meaning, in which a single modification of the order can result in a different inter-

pretation. On that account, DL models must have information about the relative or absolute position of each element of the input when dealing with raw sequential data.

Contrarily to RNN architectures, Transformer-based architectures do not contain hidden states that keep information about the relative or absolute position of each token in the input sequence, hence, it is necessary to include this information. The sinusoidal positional encoding proposed by Vaswani et al. [172] is based on sine and cosine functions of different frequencies, resulting in a unique encoding for each position of the sequence. The positional encoding for the kth token of the sequence and ith position of the embedding vector can be given by:

$$\begin{aligned} PE(k,2i) &= sin(\frac{k}{n^{\frac{2i}{d_{model}}}}) \\ PE(k,2i+1) &= cos(\frac{k}{n^{\frac{2i}{d_{model}}}}) \end{aligned} \tag{C.1}$$

, where k is the position of the token in the sequence, i is the position in the embedding vector, d_{model} is the embedding dimension, n is a user-defined scalar usually fixed at 10 000, and $PE \in R^{N\ x\ d_{model}}$ (N is the number of tokens in the input sequence).

Moreover, this encoding approach is not influenced by the input sequence length and it is based on the relative positions of the tokens.

C.3 Supplementary Experimental Setup

The hyperparameters for the DTITR architecture were determined by the chemogenomic K-fold cross-validation method (Section 7.2.4). The protein sequences similarity matrix was obtained using the Smith-Waterman local alignment algorithm, which was implemented using the Biostrings R Package [362]. The substitution matrix selected was the BLOSUM62, and the gap penalty for opening and extension was fixed at 10 and 0.5, respectively. The final alignment scores were normalized to a [0,1] range [216]. On the other hand, the SMILES similarity matrix was obtained by computing the Tanimoto Coefficient, where the SMILES strings were initially converted to the Morgan circular fingerprints with a radius of 3 using the RDKit Python package [344].

The dataset was split into six different folds, where one of the folds was selected to evaluate the generalization capacity of the model (independent test set) and the

remaining folds to determine the hyperparameters of the architecture. Several parameters were hyperoptimized: number of protein Transformer-encoders, number of SMILES Transformer-encoders, number of cross-attention Transformer-encoder blocks, number of heads for the self-attention and cross-attention layers, embedding dimension for the protein sequences and the SMILES strings, PWFFN hidden neurons, FCNN number of layers, FCNN hidden neurons, dropout rate, optimizer learning rate, and optimizer weight decay. A wide range of values was initially considered for each hyperparameter and then the search range was narrowed around the best-performing parameter values.

The Gaussian Error Linear Unit (GELU) [371] was selected as the activation function for every layer, with the exception of the final output dense layer which uses a linear activation. The GELU function weighs its input by its value rather than gating the input depending upon its sign, thus, it can be seen as a smoother ReLU. Moreover, this activation function avoids the *dead neurons* problem and is able to more easily approximate complicated functions due to the increased curvature and non-monotonicity.

$$\begin{aligned} GELU(x) &= xP(X \leq x) = x\Phi(x) \\ &\approx 0.5x(1 + tanh[\sqrt{2/\pi}(x + 0.044715x^3)]) \end{aligned} \tag{C.2}$$

, where $\Phi(x)$ is the cumulative distribution function for the standard normal distribution (Gaussian) and $P(X) \sim N(0,1)$.

Considering that the context of the problem focuses on a regression task, the loss function selected was the MSE, which measures the average squared difference between the predicted values and the real values.

Regarding the optimizer function, Rectified Adaptive Moment Estimation (RAdam) [372] was used to update the network weights in each iteration of the training process. This function is an improved version of the Adam optimizer and it dynamically adjusts the adaptive learning rate based on the underlying divergence of the variance. Thus, it avoids the need to use a warmup heuristic, which is usually required for adaptive learning rate optimizers due to the excessive variance of the initial training steps.

In order to avoid potential overfitting, two callbacks were considered during the training process, specifically early stopping with a patience of 30 and model checkpoint. The hyperparameter combination that provided the best average MSE score over the validation sets was selected to establish the optimized model and evaluate

the generalization capacity on the independent test set. Table C.1 summarizes the parameter settings for the DTITR architecture.

Table C.1: DTITR architecture parameter settings.

Parameter	Value
Protein Transformer-Encoders	3
SMILES Transformer-Encoders	3
Cross-Attention Transformer-Encoders	1
Protein Self-Attention Heads	4
SMILES Self-Attention Heads	4
Cross-Attention Heads	4
Protein Embedding Dim	128
SMILES Embedding Dim	128
Protein PWFNN Hidden Neurons	512
SMILES PWFNN Hidden Neurons	512
Activation Function	GELU
Activation Function (Output)	Linear
Dropout Rate	0.1
FCNN Dense Layers	3
FCNN Hidden Neurons	[512,512,512]
Loss Function	Mean Squared Error
Optimizer Function	RAdam
Optimizer Learning Rate	1e-04
Optimizer Beta 1	0.9
Optimizer Beta 2	0.999
Optimizer Epsilon	1e-08
Optimizer Weight Decay	1e-05
Batch Size	32
Epochs*	500

*Initial number of epochs to allow convergence of the model, where early stopping and model checkpoint were applied to avoid overfitting.

In order to validate and assess the prediction efficiency of the proposed DTITR architecture, the performance was evaluated and compared with different state-of-the-art binding affinity regression baselines: KronRLS [265], SimBoost [267], SimCNN-DTA [273], DeepDTA [268], DeepCDA [272], and all the different formulations of the GraphDTA [271]. The same folds obtained from the chemogenomic K-fold cross-validation methodology were considered to train these models and the testing fold to evaluate their performance. Additionally, the same encoding approach was applied to the protein sequences (Section 7.2.2) in the research works where the proteins are represented by their 1D amino acid sequence in order to ensure fairness in the comparisons.

Apart from evaluating the prediction efficiency of the proposed architecture, different

alternatives for the DTITR model were also explored and evaluated, specifically the efficacy of the Cross-Attention Transformer-Encoder block (Section 7.2.3.3) by applying and training the model with and without this module, the differences in the prediction efficiency of the architecture by employing the FCS and BPE encoding approach (Section 7.2.2) to the SMILES strings instead of the character-dictionary integer-based method, and the increasing learning capacity of the model due to the FCNN block (Section 7.2.3.4) by applying and training the model with and without this module.

The model was developed using Python 3.9.6 and Tensorflow 2.6.0, and the experiments were run on AMD Ryzen 9 3900X and GeForce RTX 3070 8GB.

Chapter D

Binding-Region-Guided Strategy to Predict Drug–Target Affinity

D.1 Supplementary Materials

D.1.1 Binding Pocket Datasets Distributions

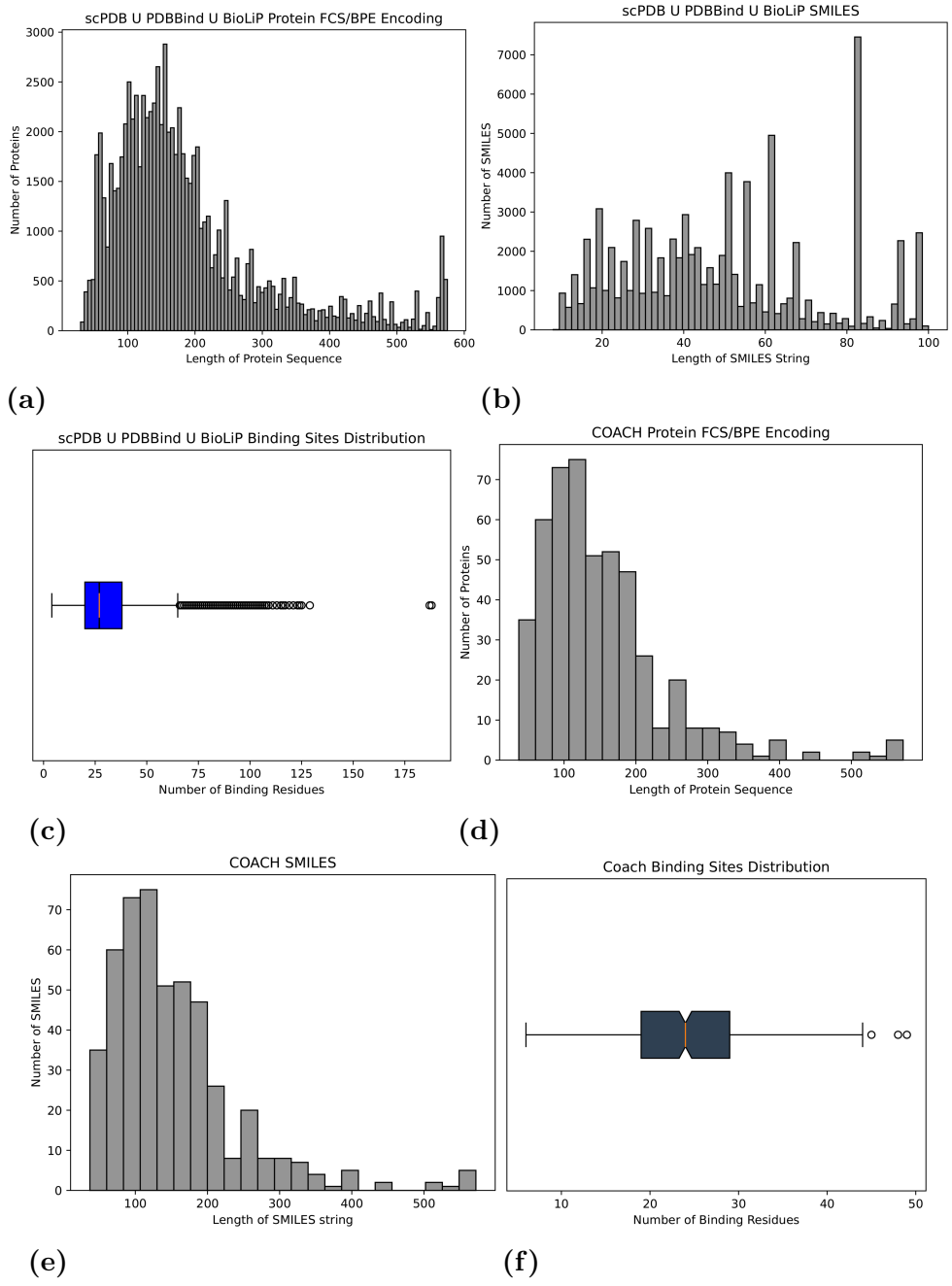


Figure D.1: Binding pocket datasets distributions. a) scPDB ∪ PDBBind ∪ BioLiP protein sequences length distribution based on the FCS/BPE encoding; scPDB ∪ PDBBind ∪ BioLiP SMILES strings length distribution. c) scPDB ∪ PDBBind ∪ BioLiP binding residues distribution based on the FCS/BPE encoding and neighborhood. d) COACH protein sequences length distribution based on the FCS/BPE encoding. e) COACH SMILES strings length distribution. f) COACH binding residues distribution based on the FCS/BPE encoding and neighborhood.

D.2 Supplementary Experimental Setup

D.2.1 SMILES Pre-Train MLM Optimization

The hyperparameters for the SMILES pre-train MLM architecture were determined using the 10 % hold-out validation approach, in which the SMILES ChEMBL dataset was randomly split into a 90/10 & training/validation ratio. Several parameters were hyperoptimized, including the embedding dimension of the SMILES strings, the number of SMILES Transformer-Encoder layers, the number of attention heads, the PWFFN expansion ratio, the dropout rate, the optimizer learning rate, the optimizer weight drop, the optimizer warm-up ratio, and the batch size. A considerable range of possible values was assigned for each hyperparameter and the search was narrowed down to the best parameter values.

The GELU [371] was selected as the activation function for every layer, except for the final output dense layer which uses a softmax activation. The GELU function is a smoother version of the ReLU activation and weighs its input by its value rather than gating the input depending upon its sign. The optimizer function considered to update the network weights in each iteration of the training process was the RAdam [372]. This algorithm is an enhanced version of the traditional Adam optimizer and it dynamically adjusts the adaptive learning rate based on the underlying divergence of the variance using a variance rectification term. Even though this method usually avoids the need to use a warm-up heuristic, a warm-up learning rate scheduler was used to promote the training process.

Considering that the SMILES pre-train MLM focuses on the prediction of masked input tokens, the loss function chosen was the categorical cross-entropy (CCE), which measures the divergence between probability distributions for multi-class classification problems. Additionally, all non-masked input chemical tokens were excluded from the calculation of the loss value.

$$CCE = -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m} y_{ij} log(p_{ij}) \tag{D.1}$$

, where n is the number of samples, m is the number of classes, y is the true label, and $p(y)$ is the predicted probability.

Moreover, to avoid potential overfitting of the model, two callbacks were considered during the training process, specifically early stopping with a patience of 30 and model checkpoint. The hyperparameter combination that provided the best accu-

racy of correctly predicting the masked input tokens over the validation set was selected to establish the optimized model. The parameters of the SMILES token embedding layer, SMILES positional embedding layer, and SMILES Transformer-Encoder of the resulting model were used to initialize the parameters of the corresponding blocks in the TAG-DTA framework. Table D.1 summarizes the parameter settings for the SMILES pre-train MLM architecture.

Table D.1: SMILES Pre-Train MLM parameter settings.

Parameter	Value
SMILES Strings Length (N_S)	101
SMILES Transformer-Encoders	3
SMILES Self-Attention Heads	8
SMILES Embedding Dim	512
SMILES PWFFN Hidden Neurons	2048
Activation Function	GELU
Activation Function (Output)	Softmax
Dropout Rate	0.1
Loss Function	CCE
Optimizer Function	RAdam
Optimizer Learning Rate	1e-03
Optimizer Minimum Learning Rate	1e-05
Optimizer Beta 1	0.9
Optimizer Beta 2	0.999
Optimizer Epsilon	1e-08
Optimizer Weight Decay	1e-04
Optimizer Warm Up Proportion	0.01
Optimizer Total steps	512500
Batch Size	232
Epochs*	500

*Initial number of epochs to allow convergence of the model, where early stopping and model checkpoint were applied to avoid overfitting.

D.2.2 TAG-DTA Optimization

The hyperoptimization approach applied to the TAG-DTA architecture is based on 10 % hold-out validation and chemogenomic representative k-fold cross-validation [341], where the former is applied to the binding sites dataset and the latter to the binding affinity dataset. These two methods were combined with grid-search, where early stopping and model checkpoint were considered over the training cycles

in order to avoid overfitting. Several parameters were hyperoptimized: embedding dimension of the protein sequences, number of protein Transformer-Encoder layers, number of protein Transformer-Encoder attention heads, protein Transformer-Encoder PWFFN expansion ratio, number of binding-pocket Transformer-Encoder layers, number of binding-pocket Transformer-Encoder attention heads, binding-pocket Transformer-Encoder PWFFN expansion ratio, PWMLP number of layers, PWMLP hidden neurons, number of binding region-guided Transformer-Encoder layers, number of binding region-guided Transformer-Encoder attention heads, binding region-guided Transformer-Encoder PWFFN expansion ratio, FCNN number of layers, FCNN hidden neurons, 1D binding pocket optimizer learning rate, 1D binding pocket optimizer weight decay, binding affinity optimizer learning rate, binding Affinity optimizer weight decay, SMILES Transformer-Encoder optimizer learning rate, SMILES Transformer-Encoder optimizer weight decay, dropout rate, number of epochs for the pre-training of the binding sites classifier, number of epochs for the training of the binding affinity regressor, and number of epochs for the training of the binding sites classifier. A wide range of values was initially considered for each hyperparameter and then the search range was narrowed around the best parameter values.

The GELU [371] activation function was selected for every layer, except for the final output dense layer of the 1D binding pocket classifier and binding affinity regressor, which uses a sigmoid function and a linear activation, respectively. The optimizer selected to update the weights of the pre-trained layers, 1D binding pocket classifier, and binding affinity regressor was the RAdam [372]. Additionally, a step decay learning rate scheduler was applied in the case of the optimizers used in the 1D binding pocket and binding affinity training modes. The step decay scheduler drops the learning rate by a factor of 0.5 every 25 training cycles.

$$Step\_Decay(TC) = LR_{init} * factor^{floor(TC/TC_{drop})} \tag{D.2}$$

, where LR_{init} is the initial learning rate, $factor$ is the learning rate dropping factor, TC is the current training cycle, and TC_{drop} is the number of training cycles that the learning rate keeps constant.

Considering that the context of the problem focuses on a classification and a regression task, the loss function selected for the prediction of the 1D binding pocket and the binding affinity of DTIs was the binary cross-entropy (BCE) and the MSE, respectively. BCE measures the divergence between two probability distributions and is a special case of the CCE. Additionally, considering the imbalanced nature of

the distribution of binding and non-binding positions, different class weights were introduced.

$$BCE = -\frac{1}{n}\sum_{i=1}^{n}\left[y_i \log(p_i) + (1 - y_i)\log(1 - p_i)\right] \tag{D.3}$$

, where n is the number of samples, y is the true label, and $p_i(y)$ is the predicted probability. On the other hand, MSE measures the average squared difference between the predicted values and the real values.

The hyperparameter combination that provided the best average MSE score over the validation sets in the case of binding affinity and the best MCC in the case of the 1D binding pocket was selected to establish the optimized model. Table D.2 summarizes the parameter settings for the TAG-DTA framework.

To validate and assess the prediction efficiency of binding affinity of the proposed TAG-DTA architecture, the performance was evaluated and compared with different state-of-the-art binding affinity regression and binary DTI classification baselines: KronRLS [265], SimBoost [267], Sim-CNN-DTA [273], DeepDTA [268], DeepCDA [272], TransformerCPI [243], HyperAttentionDTI [245], DTITR [350], and all the different formulations of the GraphDTA [271]. In the case of the binary DTI classification baselines, such as TransformerCPI [243] and HyperAttentionDTI [245], they were transformed into regression models by modifying their final layers, i.e., the output of the last layer was replaced with a single neuron dense layer, and by altering their loss function to MSE. Furthermore, the FCS/BPE encoding approach was applied to the protein sequences in the research works where the proteins are represented by their 1D amino acid sequence in order to ensure fairness in the comparisons. To further validate the binding affinity performance of the proposed framework and increase the fairness in the comparisons with the state-of-the-art baselines, the results were evaluated and compared using the standard experimental settings of these baselines, i.e., the same split method of the Davis binding affinity dataset.

Considering the existing unexplored space in the drug discovery domain and the importance of the models to generalize toward unknown subsets of the proteomics and/or chemical representation spaces, the performance of the proposed TAG-DTA framework was evaluated and compared in three different experimental settings, specifically novel target proteins, novel compounds, and novel compound-target pairs. On that account, the Davis binding affinity dataset was randomly split into three groups of five folds each based on these three different restrictions, i.e., proteins, compounds, and DTIs absent from the corresponding training set, respectively.

Table D.2: TAG-DTA parameter settings.

Parameter	Value
SMILES Strings Length (N_S)	101
SMILES Transformer-Encoders	3
SMILES Self-Attention Heads	8
SMILES Embedding Dim	512
SMILES PWFFN Hidden Neurons	2048
Protein Sequences Length (N_P)	576
Protein Transformer-Encoders	3
Protein Self-Attention Heads	4
Protein Embedding Dim	256
Protein PWFFN Hidden Neurons	1024
Binding Pocket Transformer-Encoders	1
Binding Pocket Self-Attention Heads	4
Binding Pocket PWFFN Hidden Neurons	1024
PWMLP Dense Layers	3
PWMLP Hidden Neurons	[128,64,32]
Binding Region-Guided Transformer-Encoders	1
Binding Region-Guided Self-Attention Heads	4
Binding Region-Guided PWFFN Hidden Neurons	1024
FCNN Dense Layers	3
FCNN Hidden Neurons	[1536,1536,1536]
Activation Function	GELU
Activation Function (1D Binding Pocket Output)	Sigmoid
Activation Function (Binding Affinity Output)	Linear
Dropout Rate	0.1
1D Binding Pocket Loss Function	BCE
1D Binding Pocket Loss Class Weights	0: 0.4, 1: 0.6
Binding Affinity Loss Function	MSE
SMILES Pre-Trained Layers Optimizer	RAdam, LR:1e-05, Beta 1: 0.9, Beta 2: 0.999, Epsilon: 1e-08, Weight Decay: 1e-05
1D Binding Pocket Classifier Optimizer	RAdam, LR:1e-04, Beta 1: 0.9, Beta 2: 0.999, Epsilon: 1e-08, Weight Decay: 1e-05
Binding Affinity Regressor Optimizer	RAdam, LR:1e-04, Beta 1: 0.9, Beta 2: 0.999, Epsilon: 1e-08, Weight Decay: 1e-05
Batch Size	32
1D Binding Pocket Pre-Train Epochs	20
Binding Affinity Train Epochs	3
1D Binding Pocket Train Epochs	1
Epochs*	500

*Initial number of epochs to allow convergence of the model, where early stopping and model checkpoint were applied to avoid overfitting.

Apart from evaluating the performance and the generalization capacity of the proposed TAG-DTA framework in the prediction of the binding affinity of DTI pairs, it is critical to explore the efficacy and contribution of certain blocks in the learning

capacity of the TAG-DTA architecture, especially considering the duality nature of the prediction process associated with this model. On that account, different alternatives of the TAG-DTA model were explored, specifically the contribution to the learning capacity of the architecture by pre-training the SMILES Transformer-Encoder blocks using the MLM approach, the performance of the TAG-DTA architecture to exclusively predict the 1D binding pocket, and the contribution of the prediction of the 1D binding pocket for the prediction of binding affinity, i.e., the performance of the TAG-DTA architecture to exclusively predict binding affinity without limiting the attention mechanism of the Transformer-Encoder associated with the binding affinity regression block.

All models were developed using Python 3.9.6 and Tensorflow 2.8.0, and the experiments were run on AMD Ryzen 9 3900X and GeForce RTX 3080 10 GB.

D.3 Supplementary Results

D.3.1 SMILES Pre-Train MLM

Table D.3: SMILES Pre-Train MLM: masked token prediction results over a randomly chosen 10% hold-out validation set.

	(Sparse) CCE	**Masked Token Accuracy (%)**
Training	0.0669	97.55
Validation	0.1056	96.80