INTEGRATION AND ANALYSIS OF BIOLOGICAL DATA FOR COMPUTATIONAL DRUG DISCOVERY

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF INFORMATICS OF
THE MIDDLE EAST TECHNICAL UNIVERSITY
BY

HEVAL ATAŞ GÜVENİLİR

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE
OF DOCTOR OF PHILOSOPHY
IN
MEDICAL INFORMATICS

JUNE 2023

Approval of the thesis:

INTEGRATION AND ANALYSIS OF BIOLOGICAL DATA FOR COMPUTATIONAL DRUG DISCOVERY

Submitted by Heval Ataş Güvenilir in partial fulfillment of the requirements for the degree of **Doctor of Philosophy in Health Informatics Department, Middle East Technical University** by,

Prof. Dr. Banu Günel Kılıç
Dean, **Graduate School of Informatics** _____

Assoc. Prof. Dr. Yeşim Aydın Son
Head of Department, **Health Informatics** _____

Prof. Dr. Mehmet Volkan Atalay
Supervisor, **Computer Engineering, METU** _____

Assoc. Prof. Dr. Tunca Doğan
Co-Supervisor, **Computer Engineering, Hacettepe University** _____

Examining Committee Members:

Assoc. Prof. Dr. Yeşim Aydın Son
Health Informatics, METU _____

Prof. Dr. Mehmet Volkan Atalay
Computer Engineering, METU _____

Asst. Prof. Dr. Aybar Can Acar
Health Informatics, METU _____

Assoc. Prof. Dr. Özlen Konu Karakayalı
Molecular Biology and Genetics, Bilkent University _____

Assoc. Prof. Dr. Nurcan Tunçbağ
Chemical and Biological Engineering, Koç University _____

Date: *05.06.2023*

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last name : Heval Ataş Güvenilir

Signature : _____

ABSTRACT

INTEGRATION AND ANALYSIS OF BIOLOGICAL DATA FOR COMPUTATIONAL DRUG DISCOVERY

Ataş Güvenilir, Heval

Ph.D., Department of Health Informatics

Supervisor: Prof. Dr. Mehmet Volkan Atalay

Co-Supervisor: Assoc. Prof. Dr. Tunca Doğan

June 2023, 161 pages

Drug discovery and development is a slow and costly process that comprises identifying bioactive compounds against biomolecular targets and evaluating their efficacy and safety. Computational drug/compound–target/protein interaction (DTI/CPI) prediction approaches have emerged as valuable tools to streamline this process and minimize expenses. In recent years, the integration of artificial intelligence (AI) based methods in DTI prediction has gained considerable attention, but challenges persist due to limitations in existing approaches and the complex nature of this biological problem. This thesis study aims to contribute to the effective utilization of AI in drug discovery by addressing current obstacles and developing innovative DTI prediction models. The main goal is to establish a reliable standard for designing robust and industry-applicable computational systems. The study is divided into three parts, each addressing a different aspect of the problem. In the first part, we performed a comprehensive benchmark for machine learning-based DTI prediction to achieve better data representations and more successful learning, and proposed high-quality bioactivity datasets for a fair and reliable comparison. In the second part, we utilized the knowledge graph (KG) data structure to leverage heterogeneous biological data for improved drug discovery, and constructed the KG module of our biological data integration system (CROssBAR) by incorporating essential relationships among multiple types of biomedical entities. In the last part, we proposed HetCPI, a systems-level CPI representation and prediction framework, which utilizes cutting-edge heterogeneous graph representation learning algorithms to extract hidden knowledge from multi-layered biomedical data, i.e., CROssBAR KGs, and demonstrates a considerable performance improvement in challenging scenarios. The outputs of this thesis study are expected to aid experimental and computational work in biomedical sciences, especially in drug discovery and repurposing.

Keywords: machine/deep learning, bioactivity modeling, protein representation, biomedical knowledge graph, heterogeneous graph representation learning

ÖZ

İŞLEMSEL İLAÇ KEŞFİ İÇİN BİYOLOJİK VERİNİN ENTEGRASYONU VE ANALİZİ

Ataş Güvenilir, Heval

Doktora, Sağlık Bilişimi Bölümü

Tez Yöneticisi: Prof. Dr. Mehmet Volkan Atalay

Tez Eş Yöneticisi: Doç. Dr. Tunca Doğan

Haziran 2023, 161 sayfa

İlaç keşfi ve geliştirme süreci, biyomoleküler hedeflere karşı biyoaktif bileşiklerin tanımlanması ve etkinlik ile güvenliklerinin değerlendirilmesini içeren yavaş ve maliyetli bir süreçtir. Hesaplamalı ilaç/bileşik-hedef/protein etkileşimi (İHE/BPE) tahmin yaklaşımları, bu süreci hızlandırmak ve maliyetleri azaltmak için değerli araçlar olarak ortaya çıkmıştır. Son yıllarda, İHE tahmininde yapay zeka (YZ) temelli yöntemlerin entegrasyonu önem kazanmıştır, ancak mevcut yaklaşımlardaki sınırlamalar ve bu biyolojik problemin karmaşıklığı nedeniyle İHE tahminindeki zorluklar devam etmektedir. Bu tez çalışması, mevcut sorunları ele alarak ve yenilikçi İHE tahmin modelleri geliştirerek YZ' nin ilaç keşfi alanında etkili bir şekilde kullanımına katkıda bulunmayı amaçlamaktadır. Temel hedef, sağlam ve endüstriye uygun hesaplamalı sistemlerin tasarımı için güvenilir bir standart oluşturmaktır. Çalışma, problemin farklı yönlerini ele alan üç bölümden oluşmaktadır. İlk bölümde, daha iyi veri temsilleri elde etmek ve daha başarılı öğrenme sağlamak amacıyla makine öğrenmesi temelli İHE tahmini için kapsamlı bir karşılaştırma yapılmış ve adil ve güvenilir bir kıyaslama için yüksek kaliteli biyoaktivite veri setleri oluşturulmuştur. İkinci bölümde, heterojen biyolojik veriyi daha etkili bir şekilde kullanarak ilaç keşfini iyileştirmek için bilgi çizgesi (BÇ) veri yapısından yararlanılmış ve çeşitli biyomedikal olgular arasındaki temel ilişkileri bir araya getiren biyolojik veri entegrasyon sistemimizin (CROssBAR) BÇ modülü oluşturulmuştur. Son bölümde ise, CROssBAR BÇ' leri üzerinden çok katmanlı biyomedikal veride gömülü bilgiyi ortaya çıkarmak için modern çizge tabanlı temsil öğrenme algoritmalarını kullanan HetCPI adlı sistem düzeyinde bir BPE temsil ve tahmin yapısı geliştirilmiş ve zorlu senaryolarda önemli bir performans artışı elde edilmiştir. Bu tezin çıktılarının, biyomedikal bilimlerdeki -özellikle ilaç keşfi ve yeniden konumlandırmada- deneysel ve hesaplamalı çalışmalar için oldukça faydalı olması beklenmektedir.

Anahtar Sözcükler: makine öğrenmesi, derin öğrenme, biyoaktivite modelleme, protein temsili, biyomedikal bilgi grafiği, heterojen çizge tabanlı öğrenme

To all those who played an important role in becoming the person I am today

ACKNOWLEDGEMENTS

First of all, I would like to express my sincere gratitude to my supervisors, Dr. Tunca Doğan and Dr. Volkan Atalay, for their invaluable guidance, academic support, and positive attitude that made this work possible. Dr. Tunca Doğan, in particular, has been an exceptional mentor, providing continuous feedback and encouragement that have played a crucial role in my academic and personal growth.

Besides my supervisors, I would like to thank my thesis committee members Dr. Nurcan Tunçbağ, Dr. Özlen Konu Karakayalı, Dr. Yeşim Aydın Son, and Dr. Aybar Can Acar for their valuable advices and insightful comments that enriched the quality of my research.

I would also like to thank the department staff, especially Hakan Güler, for their assistance in managing the administrative aspects of my research.

I would like to extend my heartfelt thanks to my colleagues and friends, particularly Selin Gerekçi, İpek Karasu, Fulya Çıray, and Esra Nalbant, for their countless support and motivational words during my stressful times.

My deepest appreciation goes to my father, Halil Ataş. His unwavering support, encouragement, and belief in my abilities have been a constant source of strength for me. I am also so grateful to my mother Aysel Ataş and my siblings for their unconditional love and support.

To my husband and best friend, Taylan Güvenilir, I am incredibly thankful for his endless support, understanding, and patience throughout this journey.

Finally, I would like to acknowledge TÜBİTAK for the financial support by BİDEB 2211-E Domestic PhD Scholarship and YÖK for the 100/2000 PhD fellowship during my research.

TABLE OF CONTENTS

ABSTRACT iv
DEDICATION vi
ACKNOWLEDGEMENTS vii
TABLE OF CONTENTS viii
LIST OF TABLES xi
LIST OF FIGURES xiii
LIST OF ABBREVIATIONS xvi
CHAPTERS
1. INTRODUCTION 1
1.1. Problem Statement 5
1.2. Scope and Objectives 6
1.3. Structure of the Thesis 7
2. LITERATURE REVIEW 9
2.1. Classical Machine Learning Applications in DTI Prediction 9
2.2. Cutting-edge Deep Learning Applications in DTI Prediction 12
3. A COMPREHENSIVE BENCHMARK ANALYSIS FOR ML-BASED DRUG-TARGET INTERACTION PREDICTION 17
3.1. Chapter Overview 17
3.2. Introduction 18
3.3. Materials and Methods 22
3.3.1. Dataset Construction and Splitting 22
3.3.2. Types of Featurization for Proteins and Compounds 26
3.3.3. Modelling Approaches 30
3.3.4. t-SNE Projection of Protein Representations on Large-Scale Datasets 32
3.3.5. Performance Evaluation 34
3.4. Results and Discussion 35
3.4.1. Exploration of Data Characteristics 36
3.4.2. Small-Scale Analysis (Target Feature-based Modelling) 44
3.4.3. Medium-Scale Analysis (PCM Modelling) 46
3.4.4. Large-Scale Analysis (PCM Modelling) 49

3.5. Conclusion 62
4. CROssBAR: GENERATION AND ANALYSIS OF BIOMEDICAL KNOWLEDGE GRAPHS 65
4.1. Chapter Overview 65
4.2. Introduction 66
4.3. Materials and Methods 68
4.3.1. Construction of the Prototype Hepatocellular Carcinoma (HCC) Network 68
4.3.2. Automating the Query-Based KG Construction Process of CROssBAR 70
4.3.3. Generation of CROssBAR COVID-19 KGs 72
4.3.4. Node Filtering via Overrepresentation Analysis 75
4.4. Results and Discussion 77
4.4.1. Prototype HCC Network 77
4.4.2. Use-Case Study on CROssBAR Web-Service (Query: TFP + Gastric Cancer) 78
4.4.3. Literature-Based Validation of COVID-19 KGs 80
4.4.4. Analysis of Knowledge Graph Diversity and Stability 83
4.4.5. Graph Construction Runtime Tests 91
4.5. Conclusion 93
5. LARGE-SCALE PREDICTION OF DRUG-TARGET INTERACTIONS VIA GRAPH REPRESENTATION LEARNING 95
5.1. Chapter Overview 95
5.2. Introduction 95
5.3. Materials and Methods 98
5.3.1. Dataset Construction 98
5.3.2. The Representation of Graph Nodes 101
5.3.3. Model Design and Architecture 102
5.4. Results and Discussion 105
5.4.1. Preliminary Results 106
5.4.2. Utilization of Alternative Node Attributes and Integrated CROssBAR KGs for Model Construction 110
5.4.3. Model Performance Evaluation for Different HetCPI Architectures 112
5.4.4. Performance Comparison with State-of-the-Art (SOTA) Models 115
5.4.5. Exploring the Predictive Power of HetCPI for Extreme Values 117

5.4.6. Use-Case Study: Evaluation of Bioactivity Predictions for Druggable and Undruggable Proteins 119
5.5. Conclusion and Future Directions 124
6. CONCLUSION 127
REFERENCES 131
APPENDICES 151
APPENDIX A 151
APPENDIX B 157
CURRICULUM VITAE 160

LIST OF TABLES

Table 3.1. Properties of the selected protein descriptor sets and representations used in our benchmarks. 33
Table 3.2. Protein family-based average Spearman scores of the best models and baseline models in each dataset split. 57
Table 3.3. Prediction error percentages of transformer-avg models with different thresholds on random, dissimilar-compound, and fully-dissimilar splits of transferases family dataset. 61
Table 4.1. Literature based information for new potential COVID-19 based repurposing of CROssBAR COVID-19 knowledge graph drugs. 83
Table 5.1. Node and edge statistics of protein family-specific KG datasets for (a) fully-dissimilar-split, (b) dissimilar-compound-split, (c) random-split strategy 99
Table 5.2. Node-type statistics of protein family-specific KG datasets 100
Table 5.3. (a) Node- and (b) edge-type statistics of the integrated CROssBAR KGs. 101
Table 5.4. Hyperparameters of top ten models on dissimilar-compound-split of proteases dataset. 106
Table 5.5. Performance scores of top ten models on dissimilar-compound-split of proteases dataset. 107
Table 5.6. Performance scores of the models in different design setups 108
Table 5.7. Test performance scores of the models in the ablation study. 109
Table 5.8. Test performance scores of RF regression models on dissimilar-compound-split of human proteases bioactivity dataset. 110
Table 5.9. Test performance scores of models constructed using different node attribute types and alternative KG versions on dissimilar-compound-split bioactivity datasets of human proteases and transferases. 111
Table 5.10. Hyperparameters of the finalized HetCPI models on fully-dissimilar-split (FDS), dissimilar-compound-split (DCS), and random-split (RS) datasets of transferases. 113
Table 5.11. Performance comparison with SOTA models on the filtered Davis dataset. Standard deviations are given in parentheses. 117
Table 5.12. Evaluation of HetCPI and RF model predictions for extreme bioactivity values in the dissimilar-compound-split of the transferases test set. 119
Table 5.13. Representative compound structures of over-, close- and under-estimated clusters along with experimental and predicted bioactivities. 123
Table 3.1. Model performance scores (in terms of MCC) in the small-scale analysis (on the compound-centric datasets) for; (a) random forest, and (b) SVM models. The 3 best performances for each dataset are shown in bold font. 151
Table 3.2. Model performance scores in the medium-scale analysis (on the mDavis dataset). The best performance for each metric is shown in bold font. 153

Table 3.3. Model performance scores (in terms of the median corrected MCC) in the large-scale analysis on the protein family specific datasets of; (a) the random-split, (b) dissimilar-compound-split, and (c) the fully-dissimilar-split. The 3 best performances for each protein family are shown in bold font (ran200_ran-ecfp4: random200_ random-ecfp4, only-ran-ecfp4: only-random-ecfp4). 154
Table 5.1. Comparison of test performance scores for different architecture alternatives of HetCPI models on the transferases bioactivity dataset. The best performance for each split is shown in bold font. 155

LIST OF FIGURES

Figure 1.1. The pipeline of drug discovery and development process (Rifaioglu et al., 2019) 2
Figure 1.2. The typical steps involved in constructing a supervised learning-based AI model for DTI prediction 5
Figure 2.1. The flowchart of the Pred-binding method (Shar et al., 2016) 11
Figure 2.2. The overall view of the DeepCDA framework (Abbasi et al., 2020) 13
Figure 2.3. Overview of the MGraphDTA model (Yang et al., 2022) 15
Figure 3.1. The schematic overview of the study in this chapter. 21
Figure 3.2. t-SNE based visualization of conventional (apaac, k-sep_pssm, pfam, taap) and learned (protvec, seqvec, transformer-avg, unirep1900) protein representations on; (a) enzymes including hydrolases, oxidoreductases, proteases, transferases, and other-enzymes groups, and (b) non-enzyme protein families including epigenetic regulators, ion channels, membrane receptors, transcription factors, and transporters, in different colors. 38
Figure 3.3. Pairwise similarity distributions of (a) proteins and (b) compounds for "train vs. train", "test vs. test", and "train vs. test" samples in random-split, dissimilar-compound-split, and fully-dissimilar-split of the transferases dataset (shown in the logarithmic scale). 39
Figure 3.4. Histogram plots displaying bioactivity distributions of transferase, ion channel, and membrane receptor families based on train (green bars) and test (orange bars) samples of; (a) random-split, (b) dissimilar-compound-split, and (c) fully-dissimilar-split datasets, along with their median values shown as vertical dashed lines. 41
Figure 3.5. KS distance (between train and test samples) score distributions of (a) apaac, and (b) transformer-avg representations among random, dissimilar-compound, and fully-dissimilar splits in the transferases family proteins. 42
Figure 3.6. t-SNE projections of train-test samples (i.e., compound-protein pairs) of transferase and ion channel families for k-sep_pssm and unirep1900 representations on; (a) the random-split, (b) dissimilar-compound-split, and (c) the fully-dissimilar-split datasets. 43
Figure 3.7. Mean (a) MCC and (b) F1-score test results of RF- and SVM-based DTI prediction models constructed via target feature-based modelling approach. 45
Figure 3.8. Test performance results of medium-scale PCM models (on the mDavis dataset) based on RMSE (the scores are reported as 1-RMSE, so higher values represent better performance), Spearman' s rank correlation, MCC and F1-score; (a) each color corresponds to an evaluation metric, and (b) scores are displayed only for the selected representative models (marked with asterisk in the legend). The ranking in the legend is based on the models' performance from best to worst according to their RMSE scores. Shades of red and blue represent conventional descriptors and learned representations, respectively. 48

Figure 3.9. Regression-based test performance results of protein family-specific PCM models (each using a different representation type as input feature vectors) for random-split, dissimilar-compound-split, and fully-dissimilar-split datasets based on (a) median corrected RMSE, and (b) Spearman correlation scores. The models are ranked according to decreasing performance on the fully-dissimilar-split dataset. 52
Figure 3.10. Classification-based test performance results of protein family-specific PCM models (each using a different representation type as input feature vectors) in terms of MCC scores for (a) random-split, (b) dissimilar-compound-split, and (c) fully-dissimilar-split datasets. The models are ranked according to decreasing performance on the fully-dissimilar-split dataset. 53
Figure 3.11. Performance comparison of well-performing conventional descriptor sets and learned representations for three different splits of ion-channel and protease family datasets in terms of; (a) Spearman rank correlation, and (b) median corrected MCC scores. 54
Figure 3.12. Split-based test performance scores of family-specific PCM models in terms of RMSE, Spearman rank correlation, and median corrected MCC metrics. .. 56
Figure 3.13. Clustered heatmaps of different protein featurization approaches for transferase and ion channel families on; (a) the random-split, (b) dissimilar-compound-split, and (c) the fully-dissimilar-split datasets. 59
Figure 3.14. Scatter plots of compound similarities and protein similarities against prediction errors of test data points in (a) random-split, (b) dissimilar-compound-split, and (c) fully-dissimilar-split sets of transferases for transformer-avg models..... 62
Figure 4.1. The workflow of the CROssBAR knowledge graph construction process (Doğan, Atas, et al., 2021) 71
Figure 4.2. Hepatocellular Carcinoma network as a prototype for CROssBAR knowledge graphs..... 77
Figure 4.3. (a) the output knowledge graph of trifluoperazine and gastric cancer query (b) critical signalling pathways and their relation to trifluoperazine and gastric cancer over critical genes/proteins..... 79
Figure 4.4. Large-scale version of COVID-19 knowledge graph 81
Figure 4.5. Simplified version of COVID-19 knowledge graph..... 82
Figure 4.6. CROssBAR knowledge graph diversity analysis use case, intersection graphs between: (a) breast cancer and ovarian cancer, (b) breast cancer and osteosarcoma, (c) ovarian cancer and osteosarcoma, (d) breast cancer, ovarian cancer, and osteosarcoma (triple-wise) queries. Venn diagrams displaying the statistics of shared: (e) nodes, and (f) edges, between KGs of different query terms. 86
Figure 4.7. Bar graphs indicating observed and expected frequencies (overlapping bars with different shades of colors) for each of the 140 selected highly connected/hub terms in 1365 CROssBAR KGs constructed with random term queries. “*” indicate that the corresponding observed frequency is statistically significantly lower compared to the expected frequency according to Fisher’s exact test. 87
Figure 4.8. Pairwise node identity percentage histogram (in log scale) between all KG pair combinations in our 1365 CROssBAR knowledge graphs constructed with random term queries. 88

Figure 4.9. Biological component-wise bar graphs indicating observed frequencies (in vertical axis) of all terms (in horizontal axis by ranking the terms according to decreasing frequencies) that are presented in our 1365 CROssBAR knowledge graphs constructed with random term queries. Dashed lines correspond to mean values of observed frequencies. 90
Figure 4.10. Biological component-wise query runtime histograms of 1365 CROssBAR knowledge graphs constructed with random term queries. 92
Figure 5.1. The schematic representation of the HetCPI framework. (The HGT part of the figure was adapted from Hu et al. (Hu et al., 2020)). 105
Figure 5.2. Bar plots of Spearman and median corrected MCC scores of HetCPI and baseline models with different architecture alternatives on the dissimilar-compound-split of the transferases test dataset. 115
Figure 5.3. Bioactivity distributions of real values, HetCPI predictions, and RF predictions, (a) both collectively and (b) separately, for test data points in the dissimilar-compound-split set of transferases. 118
Figure 5.4. Co-crystalized 3-D structures of (a) PIM1 in complex with a benzofuranone inhibitor and (b) HER3 in complex with bosutinib. 120
Figure 5.5. Histograms of (a) normalized predicted bioactivity scores in log scale, and (b) prediction differences, for the undruggable protein HER3 and the druggable protein PIM1 121
Figure 5.6. Prediction error box plots of compound clusters having bioactivity measurements for PIM1 protein in the dissimilar-compound-split of the transferases training dataset. 122
Figure 3.1. Clustered heatmaps of hydrolases for protein families on (a) the random-split, (b) dissimilar-compound-split, and (c) fully-dissimilar-split datasets. 159

LIST OF ABBREVIATIONS

DTI	Drug-Target Interaction
PCM	Proteochemometric
QSAR	Quantitative Structure-Activity Relationship
VS	Virtual Screening
PRL	Protein Representation Learning
NLP	Natural Language Processing
RF	Random Forest
SVM	Support Vector Machine
t-SNE	t-Distributed Stochastic Neighbor Embedding
MCC	Matthew's Correlation Coefficient
RMSE	Root Mean Square Error
RNN	Recurrent Neural Network
mLSTM	Multiplicative Long/Short Term Memory
TAPE	Tasks Assessing Protein Embeddings
BERT	Bidirectional encoder representations from transformers
ELMo	Embeddings from language models
ECFP	Extended-Connectivity Fingerprint
CROssBAR	Comprehensive Resource of Biomedical Relations with Deep Learning Applications and Knowledge Graph Representations
HCC	Hepatocellular Carcinoma
HGT	Heterogenous Graph Transformer
KG	Knowledge Graph
HPO	Human Phenotype Ontology
GNN	Graph Neural Network
ML	Machine Learning

CHAPTER 1

INTRODUCTION

1. Drug Discovery and Development Process

The discovery and development of new drugs is one of the most crucial and demanding objectives of modern medicine, requiring extensive research efforts to overcome various scientific and regulatory obstacles. It is a multi-stage and long-term process (Figure 1.1) that involves *(i)* the identification and validation of a biomolecular target, *(ii)* finding potential drug candidates that interact with the target, *(iii)* optimizing the chemical structure and properties of the lead compound, *(iv)* preclinical studies including animal trials and *(v)* clinical studies including human trials to test its efficacy and safety, following with *(vi)* FDA approval to be released to the market if successfully passing all stages (Blass, 2015). The average cost of developing a new drug is around $1.8 billion over approximately 13 years. The high failure rate of drug candidates in clinical trials due to their low efficacy and high toxicity levels makes drug discovery an extremely challenging, high-risk, and expensive endeavor (Paul et al., 2010).

The goal of drug design is to develop drugs that can bind to the target with high affinity and specificity, and thereby modulate its activity in a desirable way. As the initial step in the process (i.e., step *i* and *ii*), the identification of interactions between drug candidate molecules (e.g., compounds) and the target (e.g., druggable protein) constitutes the basis of drug discovery. It is achieved by measuring the bioactivity levels of these compounds via screening assays. With the advancements in high-throughput screening technology, it is now possible to scan thousands of compounds simultaneously; but still, it is not feasible to fully analyze a certain portion of the target and compound spaces due to the excessive number of possible protein-compound combinations (Rifaioglu et al., 2019). Thus, there is a need for the design of more efficient screening assay setups to reduce the time and cost required for successfully introducing a drug to the market.

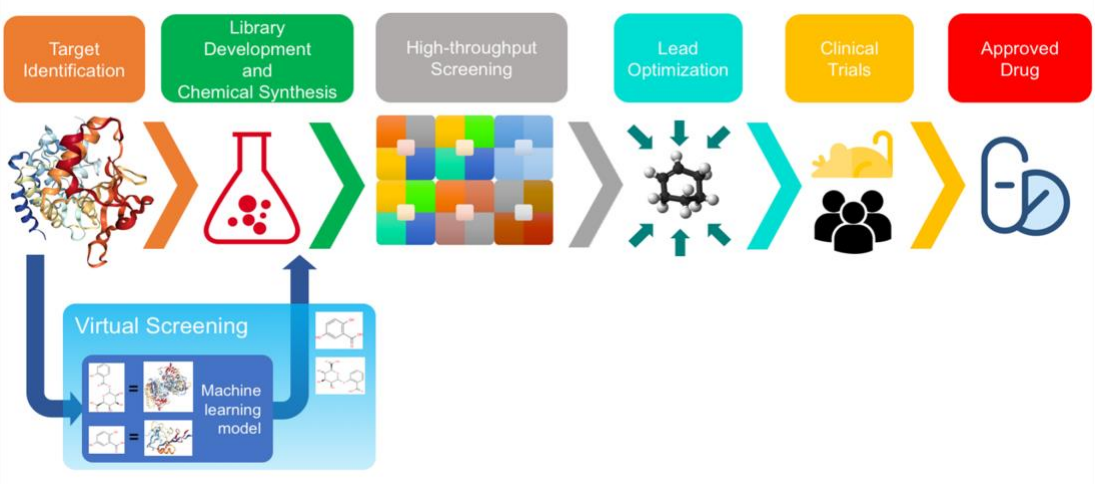


Figure 1.1. The pipeline of drug discovery and development process (Rifaioglu et al., 2019)

2. *In Silico* Prediction of Drug-Target Interactions (DTIs)

2.1. Modelling Approaches in DTI Prediction

The increasing availability of biological data and the development of computational methods offer new opportunities for accelerating drug discovery. In order to aid in designing screening experiments more effectively, computational approaches have become crucial in recent years to identify potential drug candidates worth considering in the experimental stages. The technique used for *in silico* prediction of unknown drug/compound-target interactions (DTIs) is called virtual screening (VS). Conventional VS methods are roughly categorized as ligand-based (e.g., QSAR modeling) and structure-based (e.g., molecular docking), which aim to predict interactions between a set of compounds and a predefined target protein. Ligand-based approaches utilize molecular property-based compound similarities, while structure-based approaches employ 3-D structures of targets and compounds for predicting these interactions (Lavecchia & Di Giovanni, 2013). In these applications, off-target effects are generally overlooked, and other possible targets of the compounds cannot be identified. However, it is known that most of the bioactive compounds act on multiple targets, and identification of off-targets is important, especially for drug repurposing and side-effect identification studies.

Differently from conventional VS methods, proteochemometric (PCM) modeling, as a relatively new approach in this area, overcomes this limitation by incorporating both compound and target features without requiring 3-D structures and dynamic information for DTI prediction. It can predict bioactivity relationships between large sets of compounds and targets under a single system instead of building a separate prediction model for each target protein (Cortés-Ciriano et al., 2015). There are promising studies that state PCM models outperform conventional quantitative structure-activity relationship (QSAR) models in DTI prediction modeling (van Westen et al.,2013; Paricharak et al.,2015).

There are multiple factors affecting the success of a drug candidate, mainly due to the extremely dynamic and complicated structure of biological systems. Considering this fact, network-based approaches utilizing omics data gain importance in DTI prediction with the development of systems biology and network pharmacology. These approaches differ from the abovementioned methods in terms of the input data type. Traditional VS methods and PCM modeling only utilize compound and/or protein knowledge for the inference of bioactivity data. However, network-based approaches integrate heterogeneous biological data, including protein-protein interactions, drug/compound-target protein interactions, and signaling/metabolic pathways, together with high-level concepts such as protein-disease relationships, drug-disease indications, pathway-disease modulations, and phenotypic implications. The direct and indirect relationships in molecular and cellular processes may carry hidden patterns affecting the interactions of proteins and compounds; thus, their involvement in bioactivity modeling has the potential to increase the success rate in drug discovery (Ye et al., 2021).

2.2. Utilization of Artificial Intelligence (AI) in DTI Prediction

The remarkable ability of artificial intelligence (AI) to process vast amounts of data and extract valuable insights has led to its widespread adoption in all steps of drug discovery including target identification and validation (Jeon et al., 2014), small-molecule design and optimization (Olivecrona et al., 2017; Segler et al., 2018), drug sensitivity prediction and biomarker discovery (B. Li et al., 2015), prediction of ADME-Tox properties (Wenzel et al., 2019), as well as prediction of clinical response (E. W. Huang et al., 2020) and drug approvals (Ciray & Doğan, 2022). Both VS and systems-based DTI prediction methods heavily rely on AI-centric modelling techniques, including classical machine learning and cutting-edge deep learning algorithms. Developing an AI-centric DTI prediction model involves several key steps, which are displayed in Figure 1.2.

The first step is to construct/select an appropriate bioactivity dataset, typically derived from publicly available data repositories such as DrugBank (Law et al., 2014), BindinDB (Gilson et al., 2016), ChEMBL (Mendez et al., 2019), and PubChem (Y. Wang et al., 2017). The bioactivity dataset includes information on experimentally validated drug/compound-target interactions along with the corresponding labels for supervised learning, where the system is trained using labeled samples to yield the expected output (Vamathevan et al., 2019). Depending on the type of the prediction task (i.e., classification/regression), these labels can be in binary format as “active (1)” and “inactive (0)” for classification, or real value format corresponding bioactivity measurements (e.g., pKd, pKi, pIC50, etc.) for regression. The dataset must also be preprocessed to filter out low-quality or irrelevant data and normalized for consistency if required. Then, it needs to be split into train/validation/test sets using specific strategies such as random split, stratified split, or temporal split.

The next step is the featurization of the input samples (i.e., proteins and/or compounds) into numerical representations (i.e., fixed-length feature vectors) to be processed by supervised learning algorithms. These feature vectors are traditionally constructed by

applying specific rules or calculations on various molecular characteristics of the samples, including their physicochemical, structural, topological, or functional properties, and are derived from sequences of proteins and line notations of compounds (e.g., SMILES), or their 3-D atomic coordinates (Rifaioglu et al., 2019). Recently, learned embedding approaches have emerged as a promising alternative to traditional methods in obtaining effective feature representations without using any domain-specific knowledge. They employ natural language processing (NLP) algorithms for generating "word embeddings" to extract hidden molecular knowledge directly from raw textual chemical and biological data (i.e., line notations and sequences) (Unsal et al., 2022). Additionally, they leverage the structural information and relationships between nodes (i.e., protein or compound entries) in a graph through graph-based algorithms to generate "graph embeddings". These methods allow for a more comprehensive representation of complex data and have demonstrated considerable potential in a wide range of applications in the chemical and biological domains (Nelson et al., 2019).

After the featurization of the data, selecting an appropriate algorithm is crucial to develop an effective prediction model. In DTI prediction, classical machine learning (ML) algorithms such as support vector machines (SVMs), random forest (RF), and neural networks (NNs) have been widely utilized with successful applications. However, the increasing availability of data and computational power has made deep learning (DL) algorithms, such as deep neural networks (DNNs), convolutional neural networks (CNNs), and recurrent neural networks (RNNs), a more attractive option (H. Chen et al., 2018). Graph learning algorithms, as a special type of DL algorithms, have also been employed in DTI prediction. These algorithms, including graph convolutional networks (GCNs) and graph attention networks (GATs), can handle graph-structured data and enable the integration of graph topology information into the learning process to be used for a variety of graph-related tasks, such as node classification and link prediction (Nguyen et al., 2021).

Once a suitable algorithm is selected relevant to the data and prediction task and the model architecture is designed (for DL models), the next step is to train the model on the prepared data to identify patterns and make predictions. During the training process, the model adjusts its internal parameters to enhance its predictions. In addition, there are external parameters known as "hyperparameters" (e.g., the number of hidden layers in a neural network or the number of trees in a random forest) that must be set before training the model. The adjustment of hyperparameters is an exhaustive process that requires trying various combinations of them, especially for DL models. However, it is crucial for optimizing the model performance, as hyperparameters control the complexity of the model and how it learns from the data (Andonie, 2019).

After determining the best hyperparameter combinations based on the validation set performance and training the model with these hyperparameters on the training set, it is essential to evaluate the model performance on an independent test set to assess its generalizability, consistency, and applicability. Common performance metrics for classification tasks include accuracy, precision, recall, F1-score, AUROC, and MCC

score (Rifaioglu et al., 2019). For regression tasks, RMSE, r^2, Pearson/Spearman's correlation coefficient, and CI are widely used (Cichońska et al., 2021; Rifaioglu et al., 2021).

Each of these steps is critical to building accurate and reliable AI systems for a variety of applications including DTI prediction and should be considered carefully.

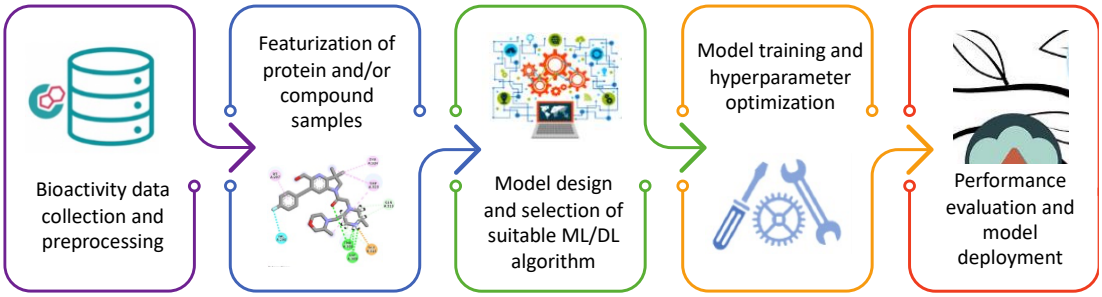


Figure 1.2. The typical steps involved in constructing a supervised learning-based AI model for DTI prediction

1.1. Problem Statement

The application of AI for DTI prediction is an area of research with great potential to accelerate the drug discovery process. However, these methods have some limitations that hinder their applicability in real-world cases and highlight the need for more robust models. One of the major limitations of AI-based models is their reliance on training data, which may be biased or incomplete. In DTI prediction, this issue mainly arises due to low coverage of protein and compound spaces, and poor dataset design, generally leading to over-optimistic performance results. Additionally, AI models have limited ability to generalize beyond the training data, making them fail when encountering new, unseen samples. The multi-layer architecture of DL models allows them to handle vast amounts of data and recognize deeper patterns, which makes them preferable to the conventional ML methods for generalizing the data. However, they still suffer from data-related issues limiting their efficient use. Therefore, introducing gold-standard, large-scale benchmark datasets with high-quality and diversity, along with properly designed, realistic train/test split scenarios will be invaluable for the community to facilitate the development of reliable and applicable DTI prediction models.

Another drawback of current AI models for the DTI prediction problem is their tendency to rely solely on bioactivity data, neglecting the potential benefits of incorporating multi-omics data. This may prevent the full capture of the complex interplay between compounds and the biological systems they interact with directly or indirectly. As a result, more advanced approaches such as graph-based deep learning, and new comprehensive representations are required to unveil non-linear relationships between target proteins and drug candidate compounds. To leverage heterogeneous biomedical data, the integration of distinct open-access data repositories (e.g., UniProt,

ChEMBL, KEGG, OMIM) is essential. These databases contain vast amounts of semantically complementary biomedical data that can offer further insights into human physiology and pathophysiology. However, most are technically disconnected from each other, having only cross-references, which restricts their holistic use. One of the most effective ways to conjugate and represent complex associations between different layers of biomedical data is to convert it into a knowledge graph (KG) structure as a network of interconnected entries. Once achieved, it can be utilized by network-based techniques for comprehensive analysis, and graph learning architectures to develop prediction models with the ultimate goal of proposing novel treatment options.

1.2. Scope and Objectives

Despite considerable progress in recent years, in silico prediction of DTIs remains a challenging problem due to the limitations of applied AI models and difficulties arising from the complex and dynamic nature of this biological issue. The primary objective of this study is to contribute to the successful utilization of AI in drug discovery by developing innovative predictive models that accurately predict bioactivities and overcome the current obstacles in the field. To achieve this, the study focuses on establishing a reliable standard for the design of robust and industry-applicable state-of-the-art models. This standardization involves providing high-quality benchmark datasets that cover different train/test split scenarios with varying difficulty levels, considering the real-world challenges faced by the pharmaceutical industry. The goal is to enhance the comparability and practicality of different approaches, leading to more reliable and reproducible results. Additionally, the study comprehensively addresses the limitations in the current applications of AI for DTI prediction. This is achieved through classical ML models that employ well-established algorithms proven to be effective in DTI prediction. This study also introduces advanced DL approaches that utilize large amounts of heterogeneous biomedical data through cutting-edge graph learning algorithms to overcome the limitations of classical methods. Ultimately, these proposed models aim to be efficient and powerful tools that contribute to the discovery of new drugs and the advancement of medical science. The proposed thesis study can be divided into three subject parts:

In the first part, we investigated the representation capability of various protein featurization techniques to be used for the automated prediction of DTIs. For this, we benchmarked different sets of conventional protein descriptors and cutting-edge learned protein embeddings by developing support vector machine (SVM) and random forests (RF) based bioactivity prediction models using two different modeling approaches: *(i) the target feature-based approach*, in which an individual predictive model is generated for each compound cluster; and *(ii) the proteochemometric (PCM) approach*, in which both compound-target feature pairs are fed to the system. As a significant contribution to the scientific community, we created gold-standard, reference bioactivity datasets on small-, medium-, and large-scale with challenging train/test splits. These datasets are intended to provide a fair basis for comparing the performance of different models and promoting the development of robust DTI prediction systems with real-world translational value in drug discovery.

In the second part, we built the biomedical knowledge graph (KG) construction pipeline for the CROssBAR system. CROssBAR serves as an integrated biological data resource that contains information on various biomedical entities, including genes/proteins, drugs/compounds, disease/phenotype terms, and pathways/biological processes. The KGs are constructed dynamically based on the user query, which allows for extracting relevant information from a large and diverse pool of biomedical data. To ensure the KGs are biologically relevant and informative, we applied specific filters, including overrepresentation analysis, to focus on the primary relationships between the entities in the graph. This filtering approach helps to maintain a reasonable node and edge size, making it easier for researchers to interpret and use. As a use-case study, we constructed COVID-19 KGs using the same methodology to provide a comprehensive and up-to-date resource for researchers who investigate the molecular mechanisms of this pandemic and seek to develop new treatment options.

In the last part, we introduced a novel systems-level approach to predict compound-protein interactions (CPIs) in a comprehensive and accurate manner. The proposed framework, named HetCPI, leverages information from large-scale biomedical knowledge graphs (KGs) constructed by the CROssBAR system. It processes the direct and indirect relationships between proteins and compounds via the heterogeneous graph transformer (HGT) algorithm and generates low-dimensional representations by preserving graph topology for the subsequent CPI prediction task. We trained, optimized, and tested HetCPI on our large-scale benchmark datasets of transferases and membrane-receptors families, and compared it to state-of-the-art models in the literature. The results demonstrate the effectiveness of HetCPI in predicting bioactivities and its potential to aid the discovery of new drug-target interactions.

This study has significant contributions to the field of drug discovery and biomedical research, including the development of gold-standard benchmark datasets, comprehensive evaluation of protein featurization methods, construction of heterogeneous biomedical KGs through the CROssBAR system, and the development of the HetCPI model utilizing large-scale biomedical data for bioactivity prediction. These outputs and contributions are expected to assist experimental and computational work in biomedical science, eventually promoting progress in drug discovery.

1.3. Structure of the Thesis

This thesis comprises six main chapters. The first chapter provides background information on the drug discovery and development process and outlines the computational approaches used for the prediction of DTIs. The problem statement and the scope and objectives of this thesis are also presented in this chapter. In the second chapter, a review of previous work in the literature related to different machine/deep learning-based DTI prediction methods is provided. In the third chapter, we evaluate the representation capabilities of various protein featurization approaches for DTI prediction and construct large-scale protein family-specific benchmark datasets for bioactivity modeling. The fourth chapter describes the CROssBAR biomedical knowledge graph (KG) construction process and use-case studies including COVID-19 KGs. In the fifth chapter, we propose our systems-based DTI prediction method

called HetCPI, employing heterogeneous graph learning algorithms. In the sixth and final chapter, we summarize the main findings of the study and discuss potential future research directions. Overall, this thesis provides a comprehensive perspective on AI-based computational methods for DTI prediction and contributes to the development of new and improved techniques for this prediction task.

CHAPTER 2

LITERATURE REVIEW

Recently, a great number of AI-centric computational approaches have been proposed for DTI prediction. In this chapter, we present the recent work regarding DTI prediction under two sections, including classical ML applications and cutting-edge DL applications.

2.1. Classical Machine Learning Applications in DTI Prediction

Machine learning has become an essential tool in drug discovery, particularly in predicting DTIs. Various classical ML algorithms such as k-nearest neighbors, matrix factorization, support vector machines, random forest, and shallow neural networks have been widely applied to DTI prediction. Regarding the methodological usage of the input properties, ML approaches are broadly divided into two categories within the context of DTI prediction: similarity-based and feature-based methods.

Similarity-based approaches rely on the assumption that similar compounds and/or targets tend to have similar biological activities, so that they utilize similarities to predict their interactions. These similarities are calculated via various metrics such as the Tanimoto coefficient, Euclidean distance, and Cosine similarity, and processed by different algorithms including bipartite local models (Bleakley & Yamanishi, 2009; Buza & Peška, 2017; Mei et al., 2013), regularized least squares (van Laarhoven et al., 2011; van Laarhoven & Marchiori, 2013; Xia et al., 2010), and matrix factorization (Gönen, 2012; Y. Liu et al., 2016; Peska et al., 2017; Zheng et al., 2013). Among similarity-based ML models, SimBoost and KronRLS (also known as CGKronRLS) are two state-of-the-art bioactivity prediction models, which yield competitive results even with cutting-edge DL models on widely used benchmark datasets (Monteiro et al., 2022; Rifaioglu et al., 2021). SimBoost uses gradient-boosting regression trees to predict binding affinities by extracting three types of features: individual features for each drug and target protein, network-based features derived from drug similarity and target similarity matrices, and network-based features derived from DTIs where drug and target nodes are connected to each other via binding affinity values (He et al., 2017).

CGKronRLS is a regularized least squares-based regression model that employs the Kronecker product kernel to combine drug similarity and target similarity kernels into a larger kernel for predicting binding affinities, which enables faster model training. It utilizes 2-D structural similarities for compounds and Smith-Waterman alignment scores for target similarities (Cichonska et al., 2017). Similarity-based models have a

significant drawback in that they are unable to generalize well to new targets or compounds, as the similarity measures have a limited applicability domain. Additionally, they may struggle to handle complex molecular interactions that cannot be captured by simple similarity measures, which may lead to poor performance in predicting interactions for structurally diverse compounds or targets with low sequence homology.

Feature-based ML approaches employ fixed-length numerical feature vectors of drugs and targets as input, unlike similarity-based models that rely on similarity matrices. These feature vectors are generated based on diverse properties of proteins and/or compounds, including their molecular, physicochemical, structural, and functional characteristics. The feature vectors are then processed by various ML algorithms to extract information for the prediction of DTIs or bioactivity values. SVM is one of the most widely used ML algorithms in DTI prediction known for its effectiveness in both regression and classification tasks (Geppert et al., 2009; Mousavian et al., 2016; Ning et al., 2009; Poorinmohammad et al., 2015; Shar et al., 2016; Strömbergsson et al., 2008; Tabei & Yamanishi, 2013; Yabuuchi et al., 2011). It can deal with high-dimensional variables in small data sets and can be used for linear and nonlinear problems to classify data points by setting decision boundaries (L. Zhang et al., 2017).

Mousavian et al. developed an SVM model for DTI prediction using bigram Position Specific Scoring Matrix (PSSM) features for proteins and PubChem fingerprints for drugs. They compared PSSM features with pseudo-amino acid composition (PAAC) as one of the most widely used protein features and demonstrated the high-confidence prediction ability of the PSSM model specifically for enzymes and ion channels datasets. They also investigated the impact of the negative sample selection strategy on the accuracy of predictions, and they observed a reduction in performance when changing the sampling method from random to balanced. This suggests that the choice of sampling method can have a significant impact on performance for the classification-based DTI prediction task and needs to be carefully considered (Mousavian et al., 2016).

RF is another state-of-the-art ML algorithm that has been extensively used for DTI prediction (Bosc et al., 2019; Kumari et al., 2015; Shar et al., 2016; Shi et al., 2019; Singh et al., 2015; Y. Wang et al., 2015; H. Yu et al., 2012). It is an ensemble method based on many decision trees generated from bootstrap samples of the training data and random subsets of the features, with each tree making independent predictions. Therefore, it promotes diversity among the decision trees, which leads to better performance compared to a single decision tree. RF has several advantages over other algorithms, such as being fast, robust against noise and overfitting, able to handle high-dimensional data, and considered one of the most successful ensemble methods (Ballester & Mitchell, 2010; Shar et al., 2016).

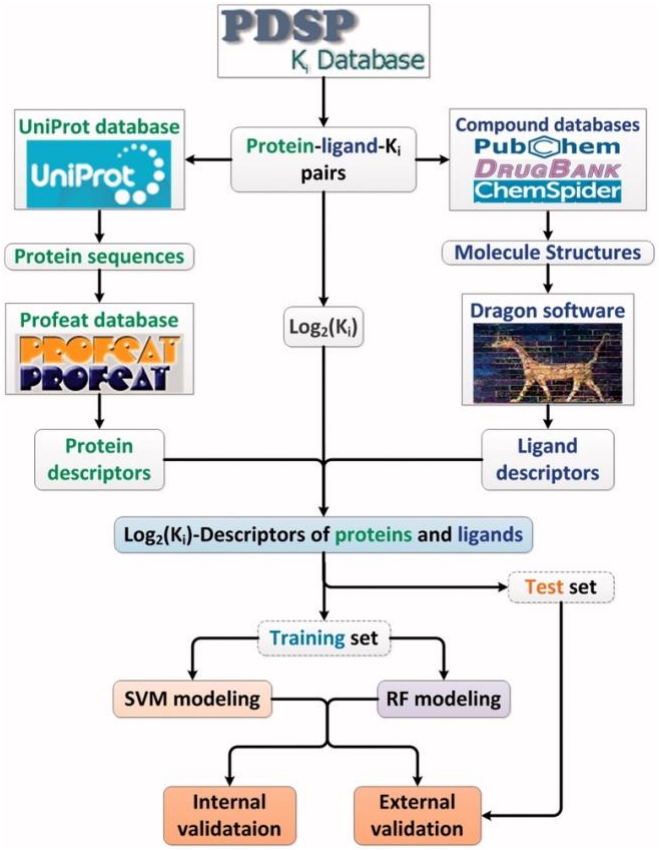


Figure 2.1. The flowchart of the Pred-binding method (Shar et al., 2016)

Wang et al. developed three family-specific RF regression models to predict protein-ligand binding affinities for HIV-1 protease, trypsin, and carbonic anhydrase families, as well as two generic models for diverse protein families. The models were trained on a comprehensive set of features, including protein sequence, binding pocket, ligand structure, and intermolecular interaction. The study revealed that the selected important features differed greatly among the families due to their distinct functions. Moreover, the family-specific models outperformed the generic models. Overall, the findings emphasize the importance of considering the unique functions and features of different protein families in predicting binding affinity (Y. Wang et al., 2015).

Shi et al. proposed a new RF classification model, LRF-DTI, for predicting DTIs. The method utilizes the pseudo-position specific scoring matrix (PsePSSM) and FP2 molecular fingerprinting to extract features from drugs and targets. The extracted features are processed using Lasso to reduce dimensionality and SMOTE to handle unbalanced data. The model was evaluated on four different datasets, including enzymes, ion channels, G-protein-coupled receptors (GPCRs), and nuclear receptors, achieving high prediction accuracies ranging from 94.9% to 98.1%. The authors also demonstrated the method's ability to perform well on new datasets, highlighting its potential for new drug research and target protein development (Shi et al., 2019).

2.2. Cutting-edge Deep Learning Applications in DTI Prediction

Deep learning (DL) is a subfield of ML that uses artificial neural networks (ANNs) with multiple hidden layers to learn hierarchical representations of data. These networks are capable of handling very large, high-dimensional data sets with billions of parameters that pass through nonlinear functions. This ability allows them to model high-level abstractions contained in data, resulting in improved performance, and have quickly surpassed classical ML approaches. In DL architectures, each data processing layer is trained on the features produced from the output of the previous layer. This enables the learning of high-level features without the need for any data preprocessing contrary to classical ML algorithms and reveals non-linear relationships in large and complex data (e.g., biological/biomedical datasets) (Rifaioglu et al., 2019). Successful applications of DL in many areas such as computer vision (Russakovsky et al., 2015), speech recognition (J. Huang & Kingsbury, 2013), and natural language processing (Y. Wu et al., 2016) have led to their widespread use in the field of drug discovery and development, as well. There are various studies developed DTI prediction models utilizing different DL architectures, including deep feedforward neural networks (FNNs) (Koutsoukas et al., 2017; Ma et al., 2015; Ramsundar et al., 2015), convolutional neural networks (CNNs) (Goh et al., 2017; Öztürk et al., 2018, 2019; Wallach et al., 2015), pairwise input neural networks (PINNs) (Rifaioglu et al., 2021; C. Wang et al., 2014), and recurrent neural networks (RNNs) (Abbasi et al., 2020; Karimi et al., 2019; C. Wang et al., 2014; S. Zheng et al., 2020).

As one of the earliest examples of FNNs, Ma et al. developed QSAR models based on single-task and multi-task FNNs for bioactivity prediction. They represented compounds as molecular descriptors based on atom pairs and donor-acceptor pair descriptors, and they used Merck's Kaggle challenge data set and in-house data sets. The authors created several models using different hyperparameters and found that using a single set of hyperparameters performed better than using optimized parameters for different data sets. The FNNs outperformed the RF classifier, and the multi-task FNNs generally performed better than the single-task FNNs. Additionally, the performance of the single-task FNNs increased with the increasing size of the training data sets (Ma et al., 2015). CNNs are specifically designed to work with images or other multidimensional inputs. The key idea behind CNNs is that they use convolutional layers to extract local features from input data. These convolutional layers are typically followed by pooling layers, which help to reduce the dimensionality of the feature maps and make the model more computationally efficient (Rifaioglu et al., 2019).

DeepDTA (Öztürk et al., 2018) and WideDTA (Öztürk et al., 2019) are two popular CNN-based models developed by Öztürk et al. for predicting compound-protein binding affinity. DeepDTA uses drug SMILES and protein sequences as input, which are processed by two distinct stacked CNN blocks that learn latent features of the drugs and targets separately, followed by max-pooling layers. The obtained features are then concatenated and fed to three fully connected layers for prediction. In an attempt to enhance DeepDTA performance, WideDTA uses four different input representations of compounds and proteins, which are SMILES and ligand maximum common substructure (LMCS) for compounds, and aa sequences and protein domains and

motifs (PDM) for proteins. Moreover, it represents compound SMILES and protein aa sequences as sets of words rather than the character-based representation. It applies the same CNN architecture with DeepDTA to extract features from each representation. Both methods have been evaluated on Davis and KIBA benchmark datasets. The results showed that the word-based sequence representation in WideDTA is a promising alternative to the character-based sequence representation used in DeepDTA. However, including PDM information as well as LMCS words did not provide additional useful information for binding affinity prediction.

Long short-term memory neural networks (LSTMs) are a specific type of RNN that are capable of learning long-term dependencies, making them particularly useful for tasks where understanding contextual relationships is critical (Vamathevan et al., 2019). Abbasi et al. proposed DeepCDA (Figure 2.2) as an extension of DeepDTA with the integration of LSTM layers after CNN blocks that is followed by a two-sided attention mechanism to encode the mutual interaction of protein subsequences with compound substructures. It is proposed for accurate prediction of binding affinities, particularly in cases where the test and training data are sampled from different domains with different distributions. The method improves the generalizability of the model by utilizing the domain adaptation technique and achieves promising results compared to other state-of-the-art approaches (Abbasi et al., 2020).

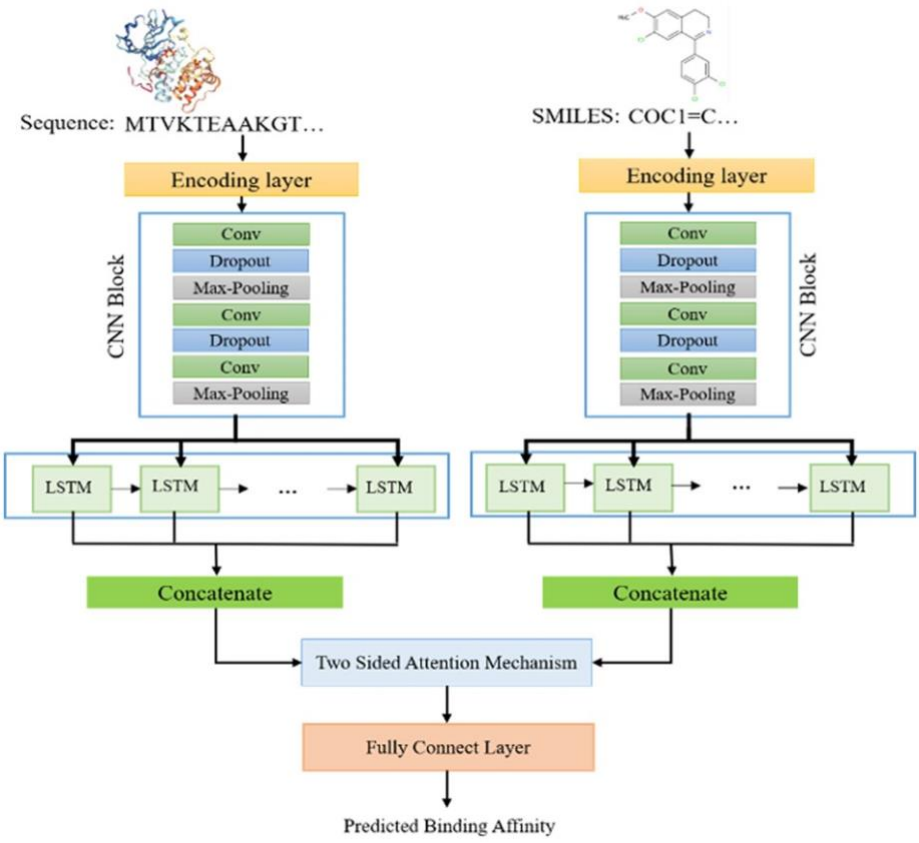


Figure 2.2. The overall view of the DeepCDA framework (Abbasi et al., 2020)

In a recent study by Rifaioglu et al., a PINN-based proteochemometric (PCM) model named MDeePred was proposed for compound–protein binding affinity prediction. The method incorporates multiple types of protein features including sequence, structural, evolutionary, and physicochemical properties as 2-D vectors, and employs ECFP4 fingerprints for compounds. As a PINN architecture, the system consists of a protein input CNN and a compound input FNN in the first stage. The second stage includes concatenating the flattened output of the inception module on the protein side with the last layer of the compound side, followed by two fully connected layers. MDeePred was evaluated on widely used benchmark datasets and showed sufficiently high predictive performance compared with state-of-the-art methods (Rifaioglu et al., 2021).

Graph neural networks (GNNs), a special type of DL architecture, have gained considerable attention in recent years for predicting DTIs due to their ability to capture structural and relational information in data (S. Li et al., 2020; Liao et al., 2022; Lin et al., 2020; Nguyen et al., 2021; Torng & Altman, 2019; Tsubaki et al., 2019; Yan & Liu, 2022; Yang et al., 2022). GNNs are designed to process graph-structured data and transform it into a low-dimensional feature space while preserving the geometric characteristics through representation learning, known as graph embedding (Z. Zhang, Chen, et al., 2022). Different variants of GNNs, such as graph convolutional networks (GCNs), graph attention networks (GATs), and graph recurrent neural networks (GRNNs), have been adopted from well-known DL architectures or mechanisms like CNNs, attention mechanisms, and RNNs. These variants differ in how they update and aggregate node features through message passing between neighboring nodes (Z. Zhang, Cui, et al., 2022). DTI prediction models employ GNNs predominantly for compound and/or protein structures represented as individual molecular graphs.

One of the pioneering and widely recognized applications of GNNs in DTI prediction is the GCN framework developed by Torng and Altman (Torng & Altman, 2019). Their approach utilizes graph-autoencoders to learn general pocket features and encode protein pockets into a fixed-size latent space. The model then employs separate GCN modules for extracting features from protein pocket graphs and 2-D ligand graphs. The interaction layer combines the learned features from both GCNs and feeds them into a classifier for prediction. This approach demonstrates the ability of graph-autoencoders to handle varying pocket sizes and the effectiveness of GCNs in capturing protein-ligand binding interactions, achieving comparable or superior performance than other methods on common benchmark datasets.

A recent study by Liao et al. introduces a new framework called GSAML-DTA for drug-target binding affinity prediction (Liao et al., 2022). GSAML-DTA integrates GATs, GCNs, and a self-attention mechanism to capture structural information from drug and protein graphs. It considers the contribution of individual drug atoms and protein residues to the binding affinity and utilizes mutual information to filter out irrelevant information in the combined representations. The results show that GSAML-DTA outperforms existing methods on widely-used benchmark datasets and provides interpretability by identifying important binding atoms and residues.

In another recent study, Yang et al. present MGraphDTA (Figure 2.3), a deep multi-scale GNN for drug-target affinity (DTA) prediction (Yang et al., 2022). The model

incorporates a dense connection and utilizes a super-deep GNN with 27 graph convolutional layers to capture both local and global compound structures. It also uses a multi-scale CNN to extract target features. Additionally, a novel visual explanation method called gradient-weighted affinity activation mapping (Grad-AAM) is introduced to provide chemical insights for model interpretation. The performance of MGraphDTA is evaluated on seven benchmark datasets and compared to state-of-the-art DL models. The results show significant enhancements in DTA prediction, emphasizing the improved generalization and interpretability of the proposed method. Despite the impressive performance of DL methods in DTI prediction, their lack of interpretability due to the black-box nature of DL remains a challenge. Therefore, the emergence of GNN-based approaches that aim to enhance interpretability is extremely valuable and promising, particularly in biomedical applications.

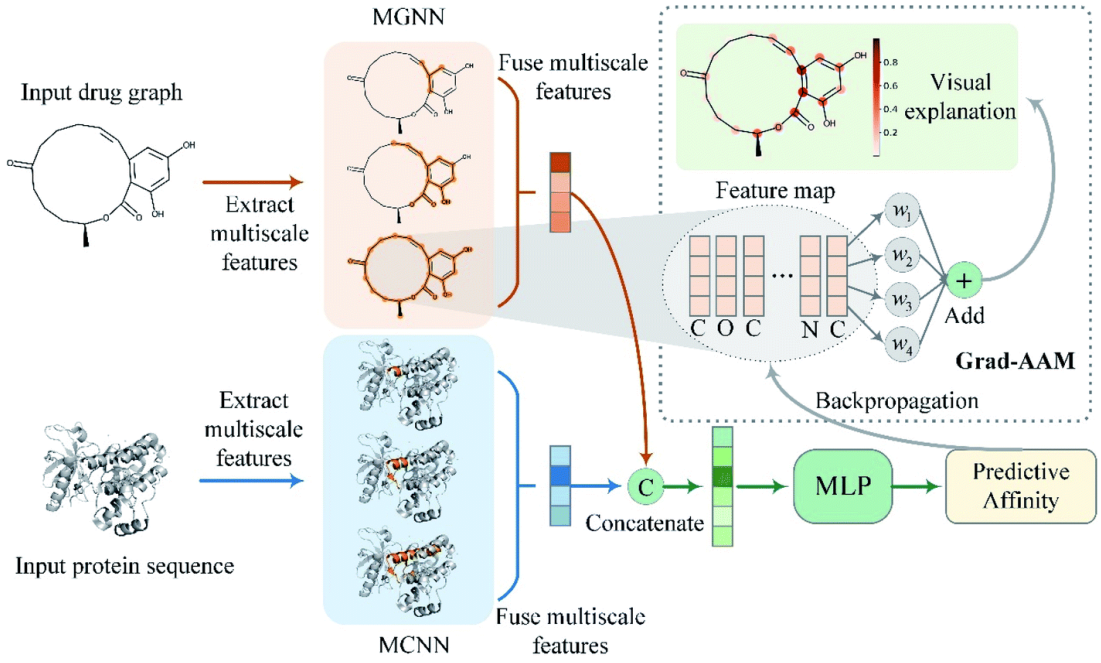


Figure 2.3. Overview of the MGraphDTA model (Yang et al., 2022)

CHAPTER 3

A COMPREHENSIVE BENCHMARK ANALYSIS FOR ML-BASED DRUG-TARGET INTERACTION PREDICTION

3.1. Chapter Overview

The identification of drug/compound-target interactions (DTIs) constitutes the basis of drug discovery, for which computational predictive approaches have been applied. As a relatively new data-driven paradigm, proteochemometric (PCM) modeling utilizes both protein and compound properties as a pair at the input level and processes them via statistical/machine learning. The representation of input samples (i.e., proteins and their ligands) in the form of quantitative feature vectors is crucial for the extraction of interaction-related properties during the artificial learning and subsequent prediction of DTIs. Lately, the representation learning approach, in which input samples are automatically featurized via training and applying a machine/deep learning model, has been utilized in biomedical sciences. In this chapter, we performed a comprehensive investigation of different computational approaches/techniques for protein featurization (including both conventional approaches and the novel learned embeddings), data preparation and exploration, machine learning-based modeling, and performance evaluation with the aim of achieving better data representations and more successful learning in DTI prediction. For this, we first constructed realistic and challenging benchmark datasets on small, medium, and large scales to be used as reliable gold standards for specific DTI modeling tasks. We developed and applied a network analysis-based splitting strategy to divide datasets into structurally different training and test folds. Using these datasets together with various featurization methods, we trained and tested DTI prediction models and evaluated their performance from different angles. Our main findings can be summarized under 3 items: (i) random splitting of datasets into train and test folds leads to near-complete data memorization and produce highly over-optimistic results, as a result, it should be avoided, (ii) learned protein sequence embeddings work well in DTI prediction and offer high potential, even though no information related to protein structures, interactions or biochemical properties is utilized during their generation, and (iii) PCM models tend to learn from compound features and leave out protein features, mostly due to the natural bias in DTI data, indicating the requirement for new and unbiased datasets. We hope this study will aid researchers in designing robust and high-performing data-driven DTI prediction systems that have real-world translational value in drug discovery. The findings of this chapter were published in the Journal of Cheminformatics (https://doi.org/10.1186/s13321-023-00689-w).

3.2. Introduction

For the automated artificial learning of DTIs to be successful, input feature vectors should comprise information about the interaction-related properties of compounds and targets. The better the input data is represented, the better the model can learn and generalize the shared properties among the dataset. Therefore, the featurization of the input samples is crucial to construct models with high predictive performance.

Various types of featurization approaches have been used for representing compounds and proteins. Due to the abundance of ligand-based DTI prediction methods, compound representations are extensively studied in the literature (Cereto-Massagué et al., 2015; Muegge & Mukherjee, 2016; Sawada et al., 2014). Therefore, this chapter focuses on protein representation techniques, a rapidly developing area lately. Sequence-based protein representations, which utilize amino acid sequences as input, are widely preferred in protein-associated predictive tasks since 3-D structural information is not available for many proteins and/or proteoforms. Additionally, the computational intensity of protein structured-based models is usually high. Considering algorithmic approaches, sequence-based protein representations can be grouped as conventional/classical descriptors (or descriptor sets) and learned embeddings. Conventional descriptors are mostly model-driven, meaning that they are generated by applying predefined rules and/or statistical calculations on sequences considering various molecular properties that include physicochemical (Chou, 2005; Ong et al., 2007; G. J. P. van Westen et al., 2013), geometrical (M. Sun et al., 2016; D. Wu et al., 2012) and topological (G. J. P. van Westen et al., 2013) characteristics of amino acids, as well as sequence composition (Ong et al., 2007; Saravanan & Gautham, 2015), semantic similarities (Perlman et al., 2011), functional characteristics/properties (Doğan, 2018; Doğan et al., 2016; Doğan, Güzelcan, et al., 2021; Yamanishi et al., 2011), and evolutionary relationships (Saini et al., 2016; M. Sun et al., 2016) of proteins. Learned protein embeddings (a.k.a. representations) are constructed via data-driven approaches that project protein sequences into high-dimensional vector spaces in the form of continuous feature representations using machine/deep learning algorithms. These protein representation learning (PRL) methods usually borrow their data modeling concepts from the field of natural language processing (NLP), where amino acids in a sequence are treated like words in a sentence/document. Due to this reason, many PRL methods are also called “protein language models”. These models usually process raw protein sequences within unsupervised learning, without any prior knowledge about their physical, chemical, or biological attributes (Unsal et al., 2022). Even though they are trained solely on the information about the arrangement of amino acids in the sequence, these models are still found to be successful in automatically extracting physicochemical properties (Asgari & Mofrad, 2015) and functional characteristics of proteins (Alley et al., 2019). PRL methods have a wide range of applications including the prediction of secondary structure (Alley et al., 2019; Heinzinger et al., 2019; Mirabello & Wallner, 2019; Rao et al., 2019), ligand-target protein interaction (Kim et al., 2021; Öztürk et al., 2019; Rifaioglu et al., 2021), splice junction prediction (Dutta et al., 2018), family classification (Asgari & Mofrad, 2015), protein function (You et al., 2018), remote homology detection (Rao et al., 2019; Strodthoff et al., 2020), and protein engineering/design (Alley et al., 2019; Rao et al., 2019).

For evaluating the effectiveness of different types of protein featurization in different areas of protein informatics, carefully designed benchmark studies are required. In contrast to studies that investigate compound featurization, only a few works are available for benchmarking protein representations. These studies mostly focus on tasks such as protein family prediction (Ong et al., 2007), bioactivity modeling (Ain et al., 2014; van Westen et al., 2013), and predicting biological properties for protein (re)design (Xu et al., 2020). Also, these studies mainly evaluate conventional descriptors rather than novel featurization approaches. As a result, there is an immediate requirement to evaluate cutting-edge protein language models and compare them against well-known conventional descriptors in the context of drug-target interactions for drug discovery/repurposing.

PCM modeling has shown promising results compared to traditional approaches in DTI prediction (Ain et al., 2014; van Westen et al., 2013); however, it is still far away from conquering this problem. One of the reasons behind this (apart from the topic of featurization) is that the mechanism of learning is not well-understood in PCM, unlike ligand-based modeling. In ligand-based methods, the model predicts new interactors for a target protein based on molecular similarities to its known ligands. In PCM, there are two factors, i.e., the compound features and the protein features, and it is not clear to what degree similarities in-between protein samples and in-between compound samples contribute to the artificial learning of their interactions, and whether there is bias in this process. Another problem associated with data-driven DTI prediction is the reporting of over-optimistic performance results due to; *(i)* low coverage on compound and/or target spaces in training datasets, in terms of molecular and biological properties (i.e., limited variance), which prevents models from gaining the ability to generalize, and *(ii)* poorly planned and applied train/test dataset preparation (e.g., splitting data randomly) and model evaluation strategies. Most of the self-proclaimed high-performing DTI prediction models in the literature are not translated well into real-world cases due to these non-realistic assessments. Recently, there have been efforts in terms of applying different dataset-splitting strategies including temporal splitting (Lenselink et al., 2017), non-overlapped sampling (Liang & Yu, 2020; Sawada et al., 2014), cluster-cross-validation (Mayr et al., 2018), and scaffold-based splitting (Z. Wu et al., 2018) to build robust models. The temporal splitting strategy only considers time-dependent data point separation. In the non-overlapped sampling strategy, three different settings are applied: warm start (common drugs and targets are present in both the training and test sets), cold start for drugs (drugs in the training set are unseen in the test set while common proteins are shared in these sets), cold start for proteins (proteins in the training set are not involved in the test set, but common drugs are allowed to be present in both sets) (Ye et al., 2021). This strategy only differentiates samples in terms of identity and does not take similarities between compounds and/or proteins into account. Although cluster-cross-validation and scaffold-based splitting methods prevent the involvement of similar compounds in train and test sets, they do not take target protein similarities into consideration. These strategies are not sufficient for evaluating PCM-based DTI prediction models, in which there are three types of relationships to account for; *(i)* compound-target protein interactions, *(ii)* compound-compound similarities, and *(iii)* protein-protein similarities.

New computational approaches, evaluation strategies, and datasets are required in order to address the aforementioned issues in the data-centric evaluation and prediction of DTIs. With the aim of contributing to the field of data-driven bioactivity modeling for drug discovery and repurposing, here, we performed a rigorous benchmarking study. One of the goals of this chapter is to identify feature types with better representation capabilities to be used in the automated prediction of DTIs. To achieve this, we built prediction models for various sequence-based protein representations. We employed widely used conventional protein descriptors by selecting those that reflect different molecular aspects of proteins. We also utilized state-of-the-art protein representation learning methods (i.e., protein language models). Another goal of the chapter is the preparation of new challenging benchmark datasets with high coverage on both compound and protein spaces, which can also be utilized in future studies. We carefully prepared small-, medium- and large-scale datasets by applying extensive filtering operations and a network-based splitting strategy to acquire realistic and well-balanced datasets. To our knowledge, this data-splitting strategy which considers 3 types of relationships (i.e., drug-target interactions, protein-protein similarities, and compound-compound similarities), is proposed here for the first time. We used these datasets in our protein representation benchmarks. In this chapter, we also evaluated different forms of; (i) DTI modeling techniques, (ii) preliminary and explanatory data exploration approaches, and (iii) model performance evaluation and comparison strategies.

The study in this chapter is summarized in a schematic workflow in Figure 3.1. Firstly, we prepared benchmark DTI prediction datasets by applying filters specific to each data scale and explored them via different data visualization techniques. We then split these datasets into train and test folds using different strategies to reflect the real-world data-centric challenges in drug discovery. For the construction of machine learning models, we implemented target feature-based and PCM modeling approaches, and trained/tested our models under various conditions. All details regarding the construction of datasets, representations and DTI prediction models are provided in the Methods section. In the Results and Discussion section, we evaluated the effectiveness of each protein featurization technique on different benchmarks and modeling approaches and discussed their strengths and weaknesses in comparison to each other. We shared our datasets, results, and source code in a reusable form under the “ProtBENCH” platform at https://github.com/HUBioDataLab/ProtBENCH.

As the first comprehensive benchmark study including both conventional and novel protein representation methods in the context of drug discovery and repurposing, we hope this work will aid researchers in choosing suitable approaches and techniques according to their specific modeling tasks. Furthermore, our newly constructed challenging benchmark datasets can be used as reliable, reference/gold-standard datasets in further studies to design robust DTI prediction models with real-world translational value.

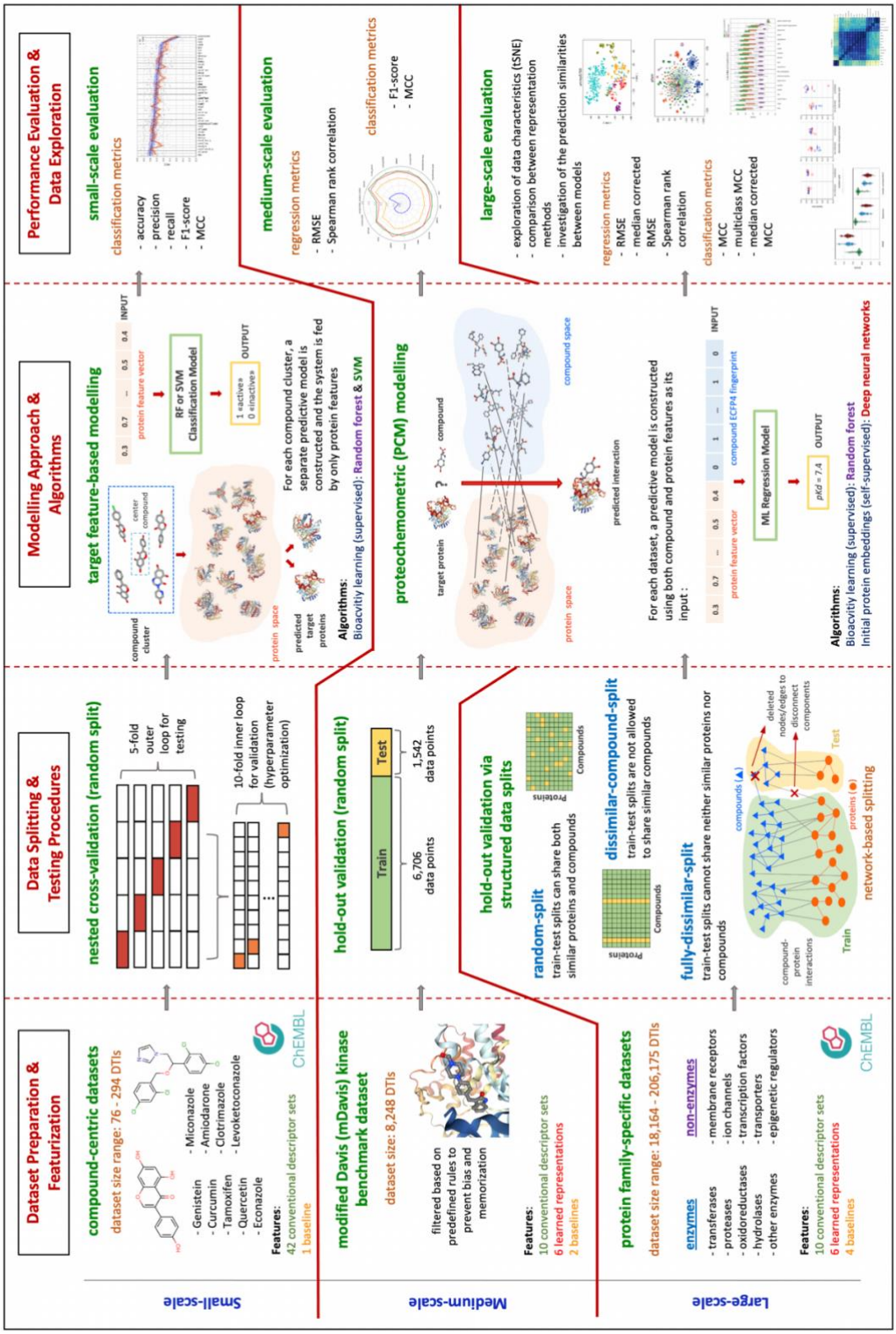


Figure 3.1. The schematic overview of the study in this chapter.

3.3. Materials and Methods

In this section, we first explain the construction of benchmark datasets, with emphasis on the train/test data splitting strategies. Next, we explain featurization techniques used for representing proteins and compounds. Then, we summarize modelling approaches and algorithms employed for DTI prediction, along with additional explorative analysis such as the t-SNE projections. Finally, we mention performance evaluation metrics and the tools/libraries we employed.

3.3.1. Dataset Construction and Splitting

In machine learning applications, two significant factors that affect the generalization capability of models are the dataset content/size and the approach used in splitting data points to train/validation/test folds. We constructed and used three groups of datasets at different scales (i.e., small, medium, and large), each of which have distinctive characteristics.

3.3.1.1. Small-scale: Compound-centric datasets

Here, the aim is to construct datasets of target proteins to be used in DTI prediction models, in which the only input is target feature vectors, and the task is to classify them to their correct ligands. Each dataset is composed of targets of a specific drug/compound as reported in the ChEMBL (v24) database (Gaulton et al., 2017) considering experimentally measured bioactivities. Bioactivity data points with pChEMBL values, i.e., -log(IC50/EC50/Ki/Kd/Potency, …), greater than 5 (equivalent to IC50/EC50/Ki/Kd/Potency <10 uM) are placed in the positives (actives) dataset, and instances with pChEMBL <= 5 are placed in the negatives (inactives) set. In most cases, sizes of these compound centric training datasets were too small to construct robust prediction models. In order to overcome this problem, we first selected compounds with the highest number of active and inactive bioactivity data points, which we called “center compounds". Afterward, we constructed compound clusters around these center compounds by calculating pairwise molecular similarities between each center compound and all other compounds in the ChEMBL database using ECFP4 fingerprints and the Tanimoto coefficient. Compounds that are similar to a center compound with Tanimoto similarity >= 0.3 (as also used in previous studies such as (Jasial et al., 2016)) are placed in the cluster of the corresponding center compound and their bioactivity data (i.e., active and inactive targets) are incorporated into the cluster’s bioactivity dataset. Therefore, nine independent compound centric, single task classification datasets (with center compounds of Curcumin, Tamoxifen, Quercetin, Genistein, Econazole, Levoketoconazole, Amiodarone, Miconazole, Clotrimazole) were constructed, and their dataset sizes (i.e., the number of targets) range from 76 to 294. Statistics of these datasets, including cluster sizes, active and inactive number of targets, are summarized in electronic supplementary information (ESI) Table S1. In order to balance the number of active and inactive targets in each dataset (initially, the number of inactive targets were considerably low), new proteins which are less than 50% similar to positive targets and less than 80% similar to negative targets already existing in the dataset were selected from ChEMBL and added to the negatives dataset. Due to the small size of datasets, separating a hold-out test fold was not feasible. Therefore, a nested cross-validation approach (with 10-fold inner

loop in validation and 5-fold outer loop in testing) was applied during model evaluation. These datasets are used in the small-scale target feature-based analysis described in section 3.4.2.

3.3.1.2. Medium-scale: mDavis kinase dataset

We employed the previously proposed Davis kinase dataset (Davis et al., 2011) for performing benchmark analysis on medium-scale, which is a commonly used benchmark for regression-based DTI prediction. The original train-test instances in the Davis dataset are taken from the study by Ozturk et al. (Öztürk et al., 2018). This dataset includes ~30,000 DTI data points (real-valued bioactivities); however, the activity values of ~20,000 of them are recorded as 10 uM (i.e., pKd = 5). These are the data points correspond to cases in which an activity was not observed when the maximum dose of 10 uM is applied (so the highest dose is incorrectly recorded as the bioactivity value). In order to prevent bias, we removed these instances from both train and test portions of the dataset. For the train portion, three additional filters were applied to avoid data memorization. All bioactivities of a compound or target are discarded if the compound or target:

1. only contains active or inactive data points based on the threshold pKd = 6.2, which is the median bioactivity value of the dataset,

2. has an active-to-inactive ratio > 4 or < 1/4 considering its bioactivity data points,

3. has a bioactivity distribution with standard deviation < 0.3, which means bioactivity values vary within a narrow range.

A successful machine learning model is expected to learn general principles from data rather than memorizing it. The instances fulfilling the conditions above may not contribute to the learning process, as they can be easily predictable regardless of the algorithm or feature set, since they have very similar outcomes. We removed these instances from the dataset; otherwise, the model would perform well just by memorizing the outcome of these cases. After these filtering operations, the finalized set, which we call the modified Davis (mDavis) dataset, contains 6,706 train and 1,542 test data points. This dataset is used in the medium-scale PCM-based analysis described in section 3.4.3.

3.3.1.3. Large-scale: Protein family-specific datasets

With the aim of constructing large-scale gold standard datasets, we applied rigorous filtering operations on the recorded bioactivities of target proteins from different protein families including membrane receptors, ion channels, transporters, transcription factors, epigenetic regulators, and enzymes with five subgroups (i.e., transferases, proteases, hydrolases, oxidoreductases, and other enzymes). Protein family information is taken from th ChEMBL (Gaulton et al., 2017) target protein classification. We excluded classes such as secreted proteins, other categories, and unclassified proteins which have inadequate number of bioactivity data points. Here, we actually mean protein super families; however, these terms are used in different

(but related) contexts in various resources, as a result, we use the term "family" throughout the article for convenience.

For enzymes, subclasses belonging to the same main class were merged based on their EC number annotations. Bioactivity data of these families are retrieved from the ChEMBL (v24) database. Bioactivity data points that satisfy the following criteria, target type: "single protein", standard relation: "=", pChEMBL value: "not null", and assay type: "B" (binding assay) are included in the dataset and the rest are discarded.

For each protein family-based dataset, three types of train-test folds were extracted based on different dataset splitting strategies based on molecular similarities in-between compounds and proteins. For this, we binarized pairwise similarity measurements as "similar" or "non-similar". UniRef50 clusters (Suzek et al., 2015) were used for generating protein similarity matrices, where proteins in the same cluster were accepted as similar to each other (equivalent to a threshold of 50% sequence similarity). Otherwise, proteins were considered dissimilar to each other. For compounds, Tanimoto coefficient-based pairwise similarities were calculated using compound ECFP4 fingerprints and the RDKit library (Landrum, 2016). Compound pairs with a Tanimoto score >= 0.5 were accepted as similar to each other. Otherwise, compounds were considered dissimilar to each other.

Random-split dataset

This dataset is constructed by applying a complete random splitting strategy, so that similar compounds and proteins are presented in both train and test sets. Random splitting is one of the most widely used dataset split strategies in machine learning applications; however, it eases the prediction task due to the sharing of highly similar instances between train and test sets. Thus, models usually display overoptimistic performance results. In our random-split protein family-specific datasets, at least 95% of proteins and 60% of compounds in test sets are found to be similar to the ones in their respective train sets.

Dissimilar-compound-split dataset

This dataset is constructed by applying a strategy that only considers compound similarities while distributing bioactivity data points into train-test splits. Compounds in train and test splits are dissimilar to each other (Tanimoto score < 0.5). Therefore, similar compounds are not allowed to take part in both train and test splits. This strategy makes the prediction task more difficult and realistic compared to random splitting and partly prevents the model from memorizing bioactivities over identical or highly similar compound fingerprints shared between train and test folds.

Fully-dissimilar-split dataset

The aim here is to create train test folds in a way that neither compounds nor proteins are similar to each other between train and test. This dataset is constructed using a network-based splitting strategy to separate bioactivity data points (i.e., compound-target pairs) into disconnected components. Later, each component is either used in training or test splits. Actually, this dataset is extremely challenging for any DTI

prediction method. However, this approach is crucial to evaluate a DTI prediction model's ability to accurately predict new targets and/or new ligands that are truly novel (i.e., there is no bioactivity information for these compounds and target proteins in source databases, moreover, there are no compounds and target proteins significantly similar to these compounds and targets in source bioactivity databases), as this is one of the most crucial expectations from the PCM modeling approach. The steps of the network-based splitting process are provided below:

1) Protein-protein and compound-compound pairwise similarity matrices were constructed independently for each protein family, based on protein family membership information and interacting compounds for those proteins (obtained from ChEMBL bioactivity data points). Similarity values were binarized according to the procedure explained above (i.e., 50% sequence/molecular similarity threshold for both protein and compounds).

2) A heterogeneous network was constructed for each protein family by merging similarity matrices and bioactivity data using the NetworkX Python library (Hagberg et al., 2008), where nodes represent proteins and compounds, and edges represent protein-protein or compound-compound similarities, and compound-protein interaction (bioactivity) relationships. It is ensured that any two components that are disconnected from each other in the network do not share any similarity at all (either directly or indirectly). As a result, all bioactivity data points in a particular component can be placed in the training fold, while the ones in another component can be placed in the test fold. As a result, bioactivity data points (i.e., compound-target pairs) in training and test folds are always guaranteed to be fully dissimilar from each other. In practice, the problem was that nearly all nodes in the network formed a giant connected component, which means that it was not possible to distribute data points to training and test folds over disconnected components.

3) In order to overcome this issue, we preferred to discard some of the nodes (e.g., compounds) and edges (e.g., bioactivity data points) from the dataset to subdivide the giant connected components into smaller pieces. Instead of removing nodes and edges randomly, which may cause the loss of a high number of data points, we employed the Louvain heuristic algorithm (Blondel et al., 2008) to detect communities in the giant component. This algorithm computes the partition of graph nodes by maximizing the network modularity. By discarding bioactivity edges (or in some cases, discarding nodes if the edge of interest is a similarity-based edge between two compounds) between different communities, the number of disconnected components was increased. Finally, bioactivity data points in each component were assigned either to training or test sets in a way that the ratio of the number of training fold data points to the test fold could be held within reasonable values, which still varied considering different protein families (i.e., from the minimum of 8.70% to a maximum of 23.97%).

Discarded data points of the fully-dissimilar-split dataset were also excluded from training-test folds of random-split and dissimilar-compound-split datasets for keeping instances of three sets exactly the same, to yield fully comparable results. The sizes of these datasets (after discarding data points) range from 18,164 to 206,175 depending on the protein family. Detailed split-based statistics are provided in ESI Table S4. These datasets are used in the large-scale PCM-based analysis described in section 3.4.

3.3.2. Types of Featurization for Proteins and Compounds

We converted proteins and compounds into fixed-length numerical feature vectors to be used in our DTI prediction models as input samples. The following sub-sections describes different featurization approaches used in this study.

3.3.2.1. Protein representations

On the basis of sequence-based modeling approaches utilized, we divided this subsection into two categories as conventional protein descriptors and learned protein embeddings. These methods are explained below in terms of their molecular and technical aspects. Names, descriptions, and feature vector dimensions of these descriptors are given in Table 3.1.

Conventional descriptor sets

This category comprises methods that employ model-driven approaches. This is achieved by transforming various molecular properties of proteins, such as sequence composition, evolutionary relationships, functional characteristics, or physicochemical properties of amino acids, into fixed-length numerical feature vectors with the implementation of predefined rules or statistical calculations. Hence, they convert protein sequences into a quantitative and machine-processible format that stores the relevant molecular information. Ten conventional protein descriptor sets used in all 3 of the benchmark analyses of this study are briefly explained below.

- *apaac (amphiphilic pseudo amino acid composition)* represents the amino acid composition of protein sequences without losing the residue order effect by using sequence-order factors. These factors are computed from correlation functions of hydrophobicity and hydrophilicity indices of amino acids. Therefore, apaac keeps the distribution of amphiphilic amino acids along the protein chain. It was proposed by Chou in 2005 and used for the prediction of enzyme subfamily classes (Chou, 2005).

- *ctdd (distribution)* provides distribution patterns of amino acids in terms of the class they belong to considering a particular property. It utilizes 7 types of physicochemical properties including hydrophobicity with 7 different versions, normalized Van der Waals Volume, polarity, polarizability, charge, secondary structures, and solvent accessibility. Each property is divided into 3 classes and 20 amino acids are distributed into these classes based on their values for corresponding property (i.e., helix -EALMQKRH-, strand -VIYCWFT-, and coil -GNPSD- classes for secondary structure property). The distribution patterns are determined according to five different positions (residues) for the corresponding class, which are the first residue, and the residues exactly at the 25%, 50%, 75%, and 100% of the sequence. These positions are divided by the length of the whole protein sequence for the calculation of fractions of each class. This descriptor set was first proposed by Dubchak for protein fold recognition task (Dubchak et al., 1999).

- *ctriad (conjoint triad)* is based on the frequency of triple amino acid combinations in a protein sequence, where amino acids are first converted into a 7-letter reduced alphabet. These seven groups include {A,G,V}, {I,L,F,P}, {Y,M,T,S}, {H,N,Q,W}, {R,K}, {D,E}, and {C}. Amino acid groups are determined according to dipoles and volumes of the amino acid side chains. Ctriad was first used for the prediction of protein-protein interactions by Shen et al. (Shen et al., 2007).

- *dde (dipeptide deviation from expected mean)* is a type of sequence composition descriptor set that relies on the deviation of dipeptide compositions from the expected means. Three parameters, i.e., dipeptide composition (*Dc*), theoretical mean (*Tm*), and theoretical variance (*Tv*), are computed for the construction of the dde descriptor set. Saravanan and Gautham proposed it in 2015 for the use of B-cell epitope prediction (Saravanan & Gautham, 2015).

- *geary* utilizes the distribution of structural and physicochemical properties of amino acids along the sequence. It was first developed by Geary in 1954 (Geary, 1954) as a measure of spatial autocorrelation that uses the square-difference of property values. Li et al. served it as a protein descriptor set via the PROFEAT web server (Z. R. Li et al., 2006). Also, Ong et. al. implemented it for the prediction of protein functional families (Ong et al., 2007).

- *k-sep_pssm (k-separated-bigrams-pssm)* is a column transformation-based descriptor set that computes the bigram transition probabilities between residues in terms of their positional distances from each other, which corresponds to the "k" value. The transition probabilities are calculated from transformations on position-specific scoring matrix (pssm) profiles of proteins. Pssm profiles represent evolutionary conservation of amino acids in a protein sequence, which are derived from multiple sequence alignments of a homolog set of protein sequences. This descriptor set was first proposed in the study of Saini et al. for improving protein fold recognition (Saini et al., 2016). Wang et al. developed the POSSUM tool to calculate a set of PSSM-based descriptors including k-sep_pssm, and they utilized these descriptors for the prediction of bacterial secretion effector proteins (J. Wang et al., 2017).

- *pfam* represents domain profiles of proteins, according to protein domain annotations in the Pfam database (El-Gebali et al., 2019), in the form of binary feature vectors. For each protein, it encodes the presence (1) and absence (0) of a unique list of domains presented in proteins in the corresponding dataset. This descriptor set was employed in the studies of Yamanishi et al. (Yamanishi et al., 2011) and Liu et al. (H. Liu et al., 2015) with the purpose of predicting drug-target interactions.

- *qso (quasi-sequence order)* reflects the indirect effect of the protein sequence order by calculating coupling factors in terms of distances between contiguous residues in the sequence. The distances are determined using different amino acid distance matrices such as the Schneider–Wrede distance matrix (Schneider & Wrede, 1994), which is derived from hydrophobicity, hydrophilicity, polarity, and

side-chain volume properties of amino acids. It was first utilised by Chou for the prediction of protein subcellular locations (Chou, 2000).

- *spmap (subsequence profile map)* is a feature space mapping method that represents sequence composition by calculating the distribution of fixed-length protein subsequence (with a length of 5 residues in the default version) clusters in a protein sequence. Subsequence clusters are generated using BLOSUM62 matrix-based similarities of all possible subsequences in the given protein set, extracted by the sliding windows approach. It was proposed by Sarac et al. for functional classification of proteins (Sarac et al., 2008). Later, spmap was used for GO term (Rifaioglu et al., 2018) and EC number (Dalkiran et al., 2018) prediction. In this study, spmap-based feature vectors were generated using clusters of 5-residue subsequences of ChEMBL target proteins.

- *taap (total amino acid properties)* represents the total sum of corresponding residue values in a protein sequence for the selected properties from the AAIndex database (Kawashima et al., 2008). It was first employed by Gromiha and Suwa for better discrimination of outer membrane proteins (Gromiha & Suwa, 2006). The properties used in our study are normalized average hydrophobicity scale, average flexibility indices, polarizability parameter, free energy of solution in water, residue accessible surface area in tripeptide, residue volume, steric parameter, relative mutability, hydrophilicity value and the side chain volume.

iFeature stand-alone tool (Z. Chen et al., 2018) was employed for the calculation of apaac, ctdd, ctriad, dde, geary and qso feature vectors. Protein domain annotations were retrieved from the Pfam database (El-Gebali et al., 2019) for the construction of pfam feature vectors. Spmap feature vectors were calculated using our in-house algorithm explained above (Sarac et al., 2008). For the construction of k-sep_pssm and taap vectors, POSSUM (J. Wang et al., 2017) and PROFEAT (P. Zhang et al., 2017) web servers were used, respectively.

Learned embeddings

This category comprises protein representations that utilize solely data-driven approaches for the extraction of molecular information from protein sequences. Learned representations are constructed via the process of artificial learning, in which a model is trained on specific unsupervised/self-supervised tasks such as the prediction of the next amino acid in the sequence. For generating such protein representation models, deep neural network-based architectures and design choices that are frequently used in natural language processing (NLP) field are preferred. During the training process, the model takes protein sequences as input, projects them into a high-dimensional vector space, and generate output in the form of fixed-length numerical feature vectors called “embeddings”. These numerical feature vectors can later be used for representing proteins in other predictive tasks (mostly supervised) such as DTI prediction.

Four protein representation learning methods (making 6 models in total, as 2 of these methods have 2 versions each) used in this study are briefly explained below. Names,

descriptions, and feature vector dimensions of these embeddings are given in Table 3.1.

- *unirep* is one of the best-known learned protein representations, which was developed in 2019 by Alley et al. using a variation of recurrent neural networks (RNN) called the multiplicative long-/short-term-memory (mLSTM) architecture (Alley et al., 2019). Alley et al. trained the model on approximately 24 million protein amino acid sequences in the UniRef50 clusters of UniProt, with the objective of predicting the next amino acid in these sequences. They evaluated the representation capability of unirep on various tasks including the prediction of protein stability, semantic similarity, secondary structure, evolutionary and functional information. In our study, we constructed both 1900- and 5700-dimensional unirep protein embeddings (obtained by averaging and summing the output embedding of size 1900x3, respectively) for sequences in our datasets and evaluated them as independent representation methods. The unirep model is available at https://github.com/churchlab/UniRep.

- *transformer* is a deep architecture that utilizes the attention mechanism in a way to allow the extraction of context without depending on the sequential order information in the input samples (Vaswani et al., 2017), and it is the current state-of-the-art in the representation learning and generative modelling of natural languages. As part of the "Tasks Assessing Protein Embeddings (TAPE)" study, Rao et al. developed a transformer-based protein representation learning model using the Bidirectional Encoder Representations from Transformers (BERT) algorithm (Rao et al., 2019). This model was trained on approximately 32 million protein sequence fragments taken from the Pfam domain annotation database (El-Gebali et al., 2019), via masked-token prediction. It was also tested on tasks such as secondary structure prediction, contact prediction, remote homology detection, fluorescence landscape prediction, and stability landscape prediction. For each sequence, the model returns two different versions of representation vectors: (i) averaged, and (ii) pooled, both in 768-dimensions. We used both of these embeddings in our study as independent representation methods. TAPE transformer model is accessible at https://github.com/songlab-cal/tape.

- *protvec* was developed by Asgari and Mofrad (Asgari & Mofrad, 2015) as one of the first models used in the construction of learned protein embeddings. It was trained on 546,790 sequences in the UniProtKB/Swiss-Prot database using the skip-gram modelling approach, in which, given a target residue, the model predicts the surrounding amino acids in the sequence. In protvec, protein sequences were embedded into 100-D vectors of 3-gram subsequences (i.e., 3 consecutively located amino acids) as biological words. For characterizing biophysical and biochemical properties of sequences, these 3-grams were analyzed qualitatively and quantitatively in terms of mass, volume, van der Waals volume, polarity, hydrophobicity, and charge. Protein feature extraction capability of protvec was also evaluated in terms of protein family classification and disordered sequence characterization tasks by representing each protein sequence as the summation of 100-D vectors of its 3-grams. 100-D vector representation of protvec can be retrieved from http://dx.doi.org/10.7910/DVN/JMFHTN.

- *seqvec* utilizes bi-directional language model architecture of the "Embeddings from Language Models (ELMo)" method for extracting features relevant to per-residue and per-protein prediction tasks. Heinzinger et al. developed the seqvec model by training on approximately 33 million UniRef50 sequences with the goal of predicting the next amino acid in the sequence (Heinzinger et al., 2019). The authors evaluated the success of seqvec on the prediction of secondary structures and regions with intrinsic disorder at the residue level, and subcellular localization prediction at the protein level. 1024-dimensional seqvec protein embeddings can be computed using the seqvec data repository at https://github.com/rostlab/SeqVec.

Random feature vectors

We constructed dummy feature vectors (to be used in baseline prediction models) for performance comparison against real representations, with the aim of observing to what degree proteins descriptors are utilized by DTI prediction models. The one for proteins, namely random200, is a descriptor that constructs a feature vector (with the size of 200x1) for each protein sequence containing randomly generated continuous values ranging from 0 to 1 in each dimension. A similar random descriptor has also been constructed for compounds (i.e., 1024x1 sized binary vectors), which is explained below, in section *3.3.2.2*.

3.3.2.2. Compound representations

We employed the (circular) fingerprinting approach, and used Extended-Connectivity Fingerprints (ECFPs) as feature vectors (representations) of compounds. ECFPs are circular topological fingerprints that are widely used for molecular characterization, similarity searching, and structure-activity relationship modeling. ECFPs represent the presence of particular substructures by considering circular atom neighborhoods within a diameter range (Rogers & Hahn, 2010). We constructed 1024-bit ECFP4 fingerprints (corresponding to a radius of 2) using RDKit (Landrum, 2016), for which compound SMILES notations were used as input. As output, a fixed-length binary fingerprint vector was generated for each compound by applying a hash function on its substructures. For the medium- and large-scale PCM models, we also generated 1024-bit "random compound fingerprints" to be used in dummy (baseline) models for evaluating the effect of compound information on DTI prediction performances. To be able to simulate ECFP4 fingerprints more realistically, we adjusted the frequency of ones and zeros in the vectors to 0.1 and 0.9, respectively (similar to real ECFP4 feature vectors in our datasets) by introducing prior probabilities during vector construction.

3.3.3. Modelling Approaches

In order to evaluate different protein featurization methods on DTI prediction, we utilized 2 different modelling approaches: (i) target feature-based modelling, and (ii) proteochemometric (PCM) modelling. Below, we summarized each approach together with the implementation details.

3.3.3.1. Target feature-based modelling

In this modelling approach (which is also known as "*in silico* target-fishing" or "reverse-screening based modelling" in the literature), we trained an independent DTI prediction model for each selected drug/compound cluster (please see section 2.1.1 for more information about the dataset). Feature vectors of proteins that are in the positives and negatives dataset of the compound of interest are given to the model as input for training and testing. Here, the model predicts whether a query protein could be the target of the corresponding compound, via binary classification. Hence, the system input is solely composed of protein features, where compounds are just used as labels.

We generated separate models for each protein descriptor set using support vector machine (SVM) and random forests (RF) classifiers, as these are widely used and well-performing machine learning algorithms. The models are implemented with scikit-learn python library (Pedregosa et al., 2011). For SVM models, "rbf" kernel was applied with optimized C and gamma parameters within ranges of [1,10,100] and [0.001,0.01,0.1,1], respectively. RF models were constructed by adjusting the parameters as; number of trees: 200, and the maximum feature number: the square root of the total number of features. Nested cross-validation (with 10-fold inner loop in validation and 5-fold outer loop in testing) was applied for model evaluation. In the end, we trained and tested 1935 RF and 1935 SVM models (i.e., 43 protein descriptor sets -including random200- for 9 different drug/compound clusters, 5-fold outer loop in nested cross validation: 43*9*5).

3.3.3.2. Proteochemometric (PCM) modelling

We constructed PCM models for both medium-scale and large-scale datasets (please see sections 3.3.1.2 and 3.3.1.3 for more information about these datasets). Here, we only used the RF regression algorithm, since we observed that RF models performed better than SVM models in the previous analysis of target feature-based modelling. For parameters, we adjusted the number of trees to 100 and maximum ratio of features to 0.33 (corresponding to one third of the total number of features).

Here, the task is predicting the actual binding affinity (bioactivity) values of the input samples (i.e., compound-target pairs) in terms of pKd/pChEMBL values. We constructed PCM models for 10 conventional protein descriptor sets, including apaac, ctdd, ctriad, dde, geary, k-sep_pssm, pfam, qso, spmap and taap, and 6 learned representations including protvec, seqvec, transformer-avg, transformer-pool, unirep1900, and unirep5700. Since PCM models are based on compound-target pairs, protein representations were concatenated with 1024 bits ECFP4 representations of compounds to construct the finalized input feature vectors.

We generated two baseline models (to be used in both medium-scale and large-scale analysis) by concatenating random200 protein feature vectors with (i) real ECFP4 fingerprints, and (ii) random compound fingerprints, which are named "random200" and "random200_random-ecfp4" models, respectively. Furthermore, we constructed two additional baseline models to be used in the large-scale analysis, in which the protein features are not utilized at all. In the first one, we used the real ECFP4

fingerprint of the compound in the corresponding compound-target pair to represent the pair (called "only-ecfp4"), and in the second one, we used random compound fingerprints to represent input pairs (called "only-random-ecfp4"). Thus, we trained and tested 18 PCM models for the medium-scale analysis using the mDavis kinase dataset (i.e., models built on 10 conventional descriptor sets, 6 learned representations, and 2 baseline models). For the large-scale analysis, we trained and tested models for 20 featurization types (10 conventional descriptor sets + 6 learned embeddings + 4 baseline models) on 10 protein family-specific datasets each having 3 versions of train-test split folds (explained in section 3.3.1.3 in detail). Therefore, we constructed 600 PCM models (i.e., 3 splits * 10 families * 20 types of featurization).

3.3.4. t-SNE Projection of Protein Representations on Large-Scale Datasets

t-distributed stochastic neighbor embedding (t-SNE) is a non-linear dimensionality reduction technique that is frequently employed for the visualization of high dimensional datasets (Van Der Maaten & Hinton, 2008). For exploratory analysis of protein family-specific large-scale datasets, we applied the t-SNE algorithm on the feature vectors of each protein featurization method and colored nodes according to protein (sub)families and train-test fold data points in two different analyses. For the application of t-SNE, we employed the scikit-learn (Pedregosa et al., 2011) manifold module with default parameters (i.e., 2-D embedding space, perplexity=30, and Euclidean distance metric). We investigated different perplexity values in the range of 40 to 100, and decided that the default value performed sufficiently well for all projections.

Table 3.1. Properties of the selected protein descriptor sets and representations used in our benchmarks.

Name	Approach	Description	Dimension
apaac	Model-driven (physico-chemistry)	Amino acid composition regarding the sequence order correlated factors computed from hydrophobicity and hydrophilicity indices of a.a*	80
ctdd	Model-driven (physico-chemistry)	Chain length-based distribution of a.a for selected physicochemical properties	195
ctriad	Model-driven (physico-chemistry)	Triad frequency of residues classified on dipoles and volumes of aa side chains	343
dde	Model-driven (sequence comp.**)	Dipeptide composition deviation	400
geary	Model-driven (physico-chemistry)	Autocorrelation regarding the distribution of physicochemical properties of a.a	240
k-sep_ pssm	Model-driven (sequence homology)	Column transformation-based position specific scoring matrix (pssm) profiles	400
pfam	Model-driven (functional properties)	Protein domain profiles	38-294 ***
qso	Model-driven (physico-chemistry)	Sequence order effect based on physicochemical distances between coupled residues	100
spmap	Model-driven (sequence comp.*)	Subsequence-based feature map	544
taap	Model-driven (physico-chemistry)	Summation of corresponding residue values for selected physicochemical properties	10
random 200	-	Randomly generated continuous numbers between 0 and 1 with uniform distribution	200
protvec	Data-driven (learned embedding)	Sequence embedding utilizing skip-gram modelling approach	100
seqvec	Data-driven (learned embedding)	Sequence embedding based on bi-directional language model architecture “ELMo”	1024
transformer	Data-driven (learned embedding)	Transformer-architecture based embedding method that utilizes attention mechanism	768
unirep	Data-driven (learned embedding)	Sequence embedding based on mLSTM architecture as a variation of recurrent neural networks	1900 & 5700

* amino acids, ** compositon, *** size varies depending on the dataset, since pfam vectors only include the domains presented in the given protein dataset.

3.3.5. Performance Evaluation

The performance of target feature-based classification models (in small-scale analysis) was evaluated via accuracy, precision, recall, F1-score, and Matthews Correlation Coefficient (MCC) metrics via nested cross-validation. F1-score is the harmonic mean of precision and recall; thus, it takes both false positive and false negative predictions into account. MCC incorporates all true and false predictions into the equation, and it is preferred over both accuracy and F1-score due to its robustness and reliability, especially in the cases of dataset imbalance (Chicco & Jurman, 2020).

The performance of PCM-based regression models (in both medium-scale and large-scale analysis) was evaluated using Root Mean Square Error (RMSE) and Spearman rank correlation (r_s) metrics over the hold-out test sets. RMSE computes the deviation of predictions from the actual values, and lower RMSE scores indicate better model performance. Spearman correlation evaluates the relationship between the ranks of the predicted and actual values. One of the problems related to regression-based prediction models is that, the distribution of predicted values can have a shifted average (i.e., the rank of predictions is in correlation with the true labels; however, the mean/median prediction value is either higher or lower than the true mean). Value-based performance metrics suffer from this problem and report underestimated scores. In order to handle this problem in the large-scale analysis (where the problem is evident), we calculated an additional version of RMSE via median correction, so that the median value of predictions becomes equal to the median of the true value distribution (i.e., the median corrected RMSE score).

We also evaluated the results of PCM-based regression models on the basis of classification, using F1-score and MCC metrics. To achieve this in the medium-scale analysis (on the mDavis dataset), samples were classified as active (1) or inactive (0) based on an activity cut-off value of pKd = 7 (i.e., 100 nM in terms of Kd) using the RF classification algorithm. For the large-scale analysis over protein family-specific datasets, regression-based prediction results were converted into binary class and multiclass formats, as it was not possible to retrain 600 models due to high computational requirements. For the binary class, median pChEMBL values of the data points in the training datasets were used as threshold values to separate actives and inactives from each other (i.e., compound-target pairs with bioactivity values higher than the median value of the dataset are accepted as actives, and the ones equal to or lower than the median are accepted as inactives). We also calculated corrected version of MCC using the procedure explained above for "median corrected RMSE" score, and similarly called this metric the "median corrected MCC". For the calculation of multi-class scores, samples were placed into six different classes based on their true pChEMBL values (class1: <5.0, class2: 5.0 - 5.5, class3: 5.5 - 6.0, class4: 6.0 - 6.5, class5: 6.5 - 7.0, and class5: >=7.0) and calculated average MCC scores over all 6 classes. The reason behind using such a variety of performance metrics was to evaluate models from as many different aspects as possible.

The equations for the basic versions of these metrics are given below:

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP} \quad (1)$$

$$Recall/Sensitivity = \frac{TP}{FN + TP} \quad (2)$$

$$Precision = \frac{TP}{FP + TP} \quad (3)$$

$$F1 - score = 2\ x\ \frac{precision * recall}{precision + recall} \quad (4)$$

$$MCC = \frac{TP\ x\ TN - FP\ x\ FN}{\sqrt{(TP + FP)(TP+FN)(TN+FP)(TN+FN)}} \quad (5)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (6)$$

$$Spearman\ rank\ correlation\ (r_s) = 1 - \frac{6\sum_{i=1}^{n} {D_i}^2}{n(n^2-1)} \quad (7)$$

where D_i = R(y_i) - R($\hat{y}_i$); D_i denotes the difference between ranks of true (y_i) and predicted ($\hat{y}_i$) values of samples with the dataset size *n*. TP, TN, FP, and FN represent the total counts of true positive, true negative, false positive, and false negative predictions, respectively.

In this study, we used Python (v3) programming language, scikit-learn library (Pedregosa et al., 2011) for the t-SNE projection and machine learning applications, NetworkX package (Hagberg et al., 2008) for splitting protein family-specific datasets, RDKit toolkit (Landrum, 2016) for compound featurization and clustering, POSSUM (J. Wang et al., 2017) and PROFEAT (P. Zhang et al., 2017) web tools as well as iFeature stand-alone tool (Z. Chen et al., 2018) for protein featurization, and seaborn (Waskom, 2021) and matplotlib (Hunter, 2007) libraries for the heatmap analysis and data visualization.

3.4. Results and Discussion

In this section, we evaluate and discuss the results of our benchmark experiments. For this, we first carried out a data exploration analysis. Next, we trained DTI prediction models under different settings and measured their performance. At each subsection, we discussed our findings from various aspects to address shortcomings in bioactivity modelling studies.

Here, we employed random forest (RF) as our main machine learning algorithm (along with support vector machine—SVM, in some of the cases) for predicting DTIs. The reasons behind using a classical machine learning algorithm in this benchmark study

rather than more complex deep learning-based architectures is that: (i) RF has been used in this field for a long while and shown to work well on numerous occasions, (ii) deep learning-based complex architectures have already been used in the training stage of learned representations (i.e., protein embeddings); thus, the use of additional complex architecture in the supervised DTI prediction stage could have prevented the observation of the ability of learned representations in extracting ligand interaction-related properties of proteins, and also, hinder the evaluation of model-driven (i.e., conventional descriptor sets) and data-driven (i.e., learned representations) approaches on common ground, and (iii) hyperparameter value selection have a significant effect on the performance of deep learning models. If we had used deep learning models in this benchmark study, the model performances would have been heavily influenced by the specific hyperparameter settings used, and any differences in performances could not be attributed solely to the representation capabilities of the featurization approaches. In this study, the main aim is to fairly compare and evaluate different representation approaches rather than constructing a single DTI prediction model with maximized performance. As a result, we used classical machine learning algorithms, which do not require the same level of hyperparameter tuning as their deep learning-based counterparts.

3.4.1. Exploration of Data Characteristics

In this subsection, we first visualized members of protein family-specific datasets on 2-D via t-SNE projection. Then, we analyzed split-based characteristics of our datasets by plotting pairwise similarity distributions of proteins and compounds, bioactivity distributions of train-test folds, together with their respective t-SNE embeddings.

3.4.1.1. t-SNE projection of protein families

For each protein representation, two independent t-SNE projections (one for the enzyme, and another one for the non-enzyme protein families) were carried out (Figure 3.2a and 3.2b). Projections for 8 protein featurization methods are shown in Figure 2. As displayed in these t-SNE plots, generally, protein families are well clustered in both enzyme and non-enzyme projections, with slightly less apparent clusters in enzymes, probably due to the sharing of enzyme-specific properties between proteins. Also, members of the other-enzymes class are scattered among other clusters as its members do not have distinctive characteristics. Although the majority of protein representations are successful at separating at least some of the families, projections of learned embeddings have clearer clusters in general, which indicates their ability of extracting family-specific features. Considering conventional descriptor sets, homology (i.e., k-sep_pssm) and domain profiles (i.e., pfam) are observed to have more distinctive abilities for the classification of protein families, compared to physicochemistry (e.g., apaac, ctdd, ctriad, geary, qso) and sequence composition (i.e., dde). The t-SNE projection of spmap, being a sequence composition-based descriptor set based on protein subsequence (5-mer) clusters, is similar to the projection of random200 descriptor set. This result indicates that 5-residue subsequences of proteins cannot capture family-specific patterns. Highly distinct from other representations, taap has a projection in the form of an S-shaped curve. Feature vectors of proteins with similar residue content and sequence length are similar to each other (independent from the actual order of amino acids on the sequence) according to the taap descriptor

set, since taap uses the total sum of the amino acid-based property values to represent a protein. Due to the fact that t-SNE aims to preserve local neighborhoods, proteins form a continuous curve similar to time-series data when represented by taap.

3.4.1.2. Split-based characteristics of protein family-specific datasets

Pairwise similarity distributions

To explore protein and compound diversity in our datasets, we evaluated protein-protein and compound-compound pairwise similarities of the members of a selected representative protein family (i.e., transferases), in terms of "train vs. train", "test vs. test", and "train vs. test" dataset comparisons for each split strategy (i.e., random-split, dissimilar-compound-split, and fully-dissimilar-split). For this, we aligned protein sequences using EMBOSS Needle global pairwise sequence alignment tool (Rice et al., 2000) and plotted histograms based on identity values of unique protein pairs in the corresponding datasets. We extracted pairwise compound similarities by calculating Tanimoto coefficient between fingerprint representations using the *simsearch* function of the Chemfp python package (Dalke, 2019). Since it was highly infeasible to calculate pairwise similarities for billions of compound pairs, we randomly sampled 10% of all compounds in the transferases dataset and set the minimum similarity detection threshold as 0.1. Again, we only considered a unique list of compound pairs.

Figure 3.3 displays similarity distributions of pairs of proteins and compounds involved in the transferases dataset, in which the values may be greater than one since the plot is normalized to equalize the total area to one (i.e., the density plot). Having a similarity value in the range of 0 - 0.5 for the majority of protein and compound pairs in all plots demonstrates the high diversity of samples which is a desirable characteristic for computational bioactivity modelling. As displayed in Figure 3.3, similarity distributions only slightly change between different split methods, considering "train vs. train" and "test vs. test" sample similarities, whereas there are significant differences between the samples of "train vs. test", for both compounds and proteins, in terms of different splits. The absence of similarity values greater than 0.5 for compound "train vs. test" pairs in the dissimilar-compound-split dataset, and both protein and compound "train vs. test" pairs in the fully-dissimilar-split dataset validates the similarity-centric characteristics of our datasets. Exceptional pairs of proteins with high similarity values in the fully-dissimilar-split dataset stem from the discrepancies between UniRef50 clusters and our pairwise alignment results, and their number is found to be insignificant (please note that the frequencies are given on logarithmic scale in Figure 3.3). These results validate the capability of our methodology in terms of producing challenging (and presumably realistic) train-test datasets, so that the bioactivity prediction models trained and tested on these datasets hopefully reflect the real-world performances while discovering novel drug candidates and/or new targets.

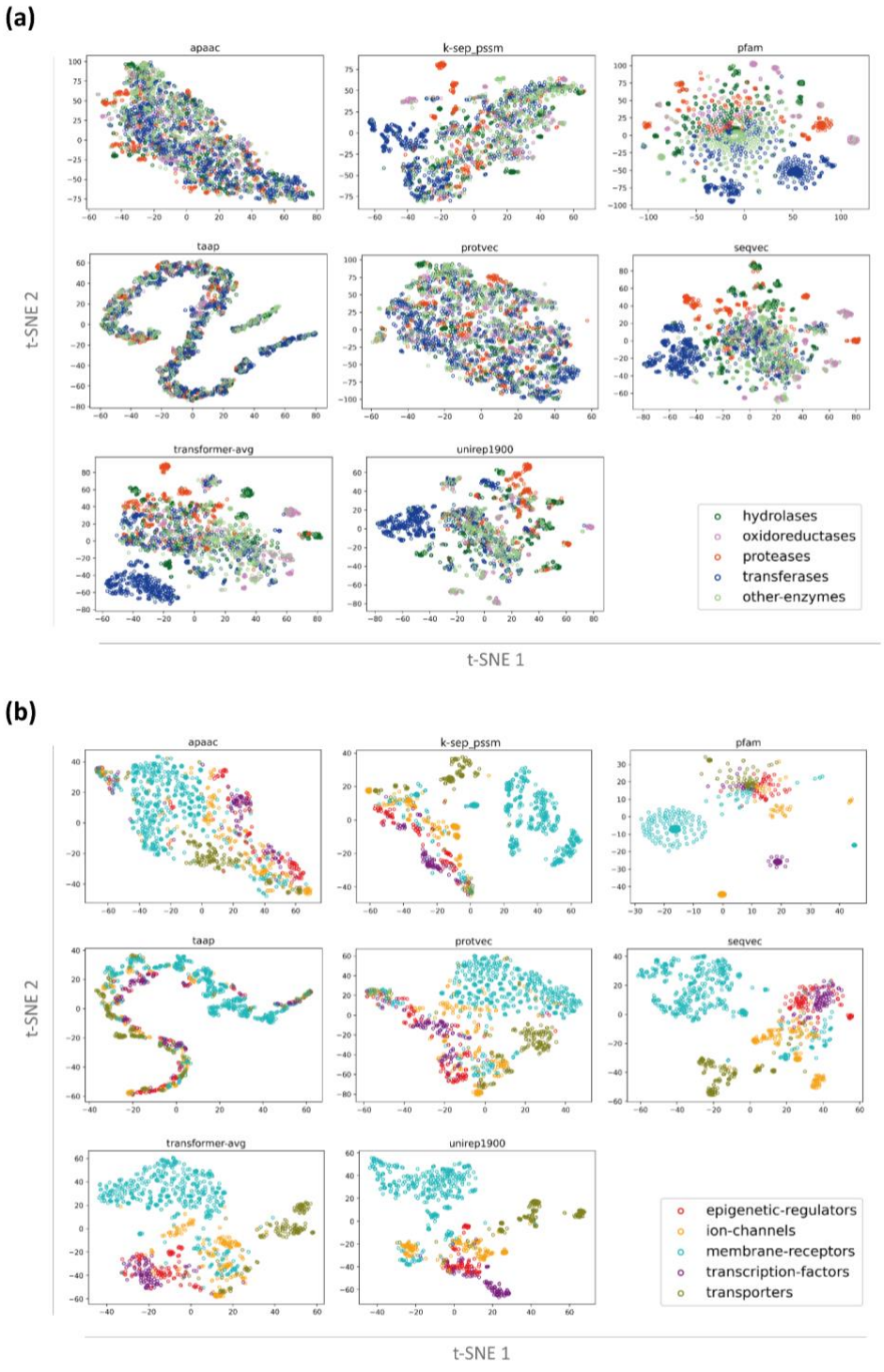


Figure 3.2. t-SNE based visualization of conventional (apaac, k-sep_pssm, pfam, taap) and learned (protvec, seqvec, transformer-avg, unirep1900) protein representations on; (a) enzymes including hydrolases, oxidoreductases, proteases, transferases, and other-enzymes groups, and (b) non-enzyme protein families including epigenetic regulators, ion channels, membrane receptors, transcription factors, and transporters, in different colors.

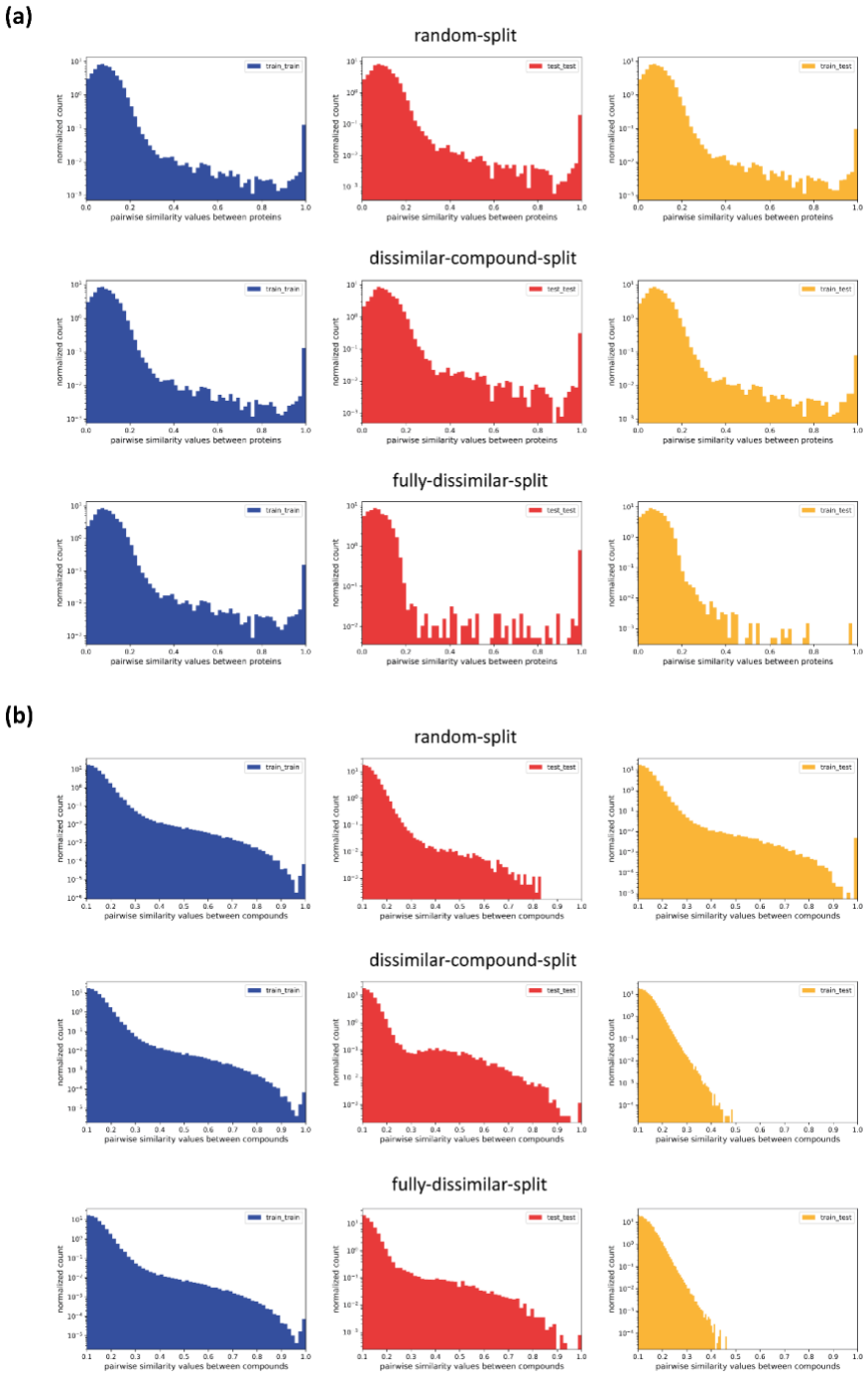


Figure 3.3. Pairwise similarity distributions of **(a)** proteins and **(b)** compounds for "train vs. train", "test vs. test", and "train vs. test" samples in random-split, dissimilar-compound-split, and fully-dissimilar-split of the transferases dataset (shown in the logarithmic scale).

The assessment of the IID assumption

Most of the traditional machine learning algorithms such as RF and SVM operate on the independent and identically distributed data (IID) assumption for the samples in training and test splits. In other words, the values of the variables in a dataset are assumed to be independent of each other and have the same probability distribution. This assumption may be violated if there is a shift in the distribution of the input or output variables between train/test splits, which may affect the performance of the model (Darrell et al., 2015). Therefore, it is important to evaluate the IID assumption while developing a machine learning model.

To explore the IID assumption in terms of output variables (i.e., bioactivity values as target labels), we plotted bioactivity distributions of protein family-specific datasets based on train-test samples of each split. Figure 3.4 displays pChEMBL value-based histograms for transferases, ion channels, and membrane receptors. Median bioactivities vary between 5.7 and 7.1 for different protein families. When comparing bioactivities of train and test sets of each family, it is observed that distributions have similar shapes, regardless of the dataset split strategy. In addition, they generally have very similar mean and median values, although the difference is slightly higher in the fully-dissimilar-split datasets of some families. Having bioactivity distributions that are consistent with each other in training and test folds implies good coverage of bioactivity data and supports the suitability of our large-scale datasets for bioactivity modelling. These results also indicate that a stratified-split strategy is not required for our datasets.

In cases of the presence of a shift in output variables, models require extrapolating beyond the minimum and maximum target values in the training datasets. This may be a limiting factor for regression-based algorithms that can only generate predictions within the boundaries of training output values (Hengl et al., 2018). Therefore, we recommend checking this issue before constructing DTI prediction models.

We also compared the distributions of protein representations and ecfp4 compound fingerprints in-between training and test splits to check the IID assumption for input variables. For protein representations with continuous values, we applied Kolmogorov–Smirnov (KS) test and calculated KS distance scores for each feature (i.e., each dimension in a representation) of train and test samples along with corresponding p-values. Figure 3.5 displays the distributions of these scores for apaac and transformer-avg representations (i.e., feature dimension sizes are 80 and 768, respectively) on three different train/test splits of the transferases family dataset. Although maximum KS distance scores are generally lower for conventional descriptors (i.e., around 0.2) than learned embeddings (i.e., around 0.5), they have similar distributions overall, where the variance is much lower in the random-split dataset compared to dissimilar-compound and fully-dissimilar split sets. There was a significant ($p\text{-value} < 0.01$) shift between the KS distance value distributions of train and test samples for 19 (for fully-dissimilar-split) and 7 (for dissimilar-compound-split) features out of the total 80 features in apaac, and 558 (for fully-dissimilar-split) and 189 (for dissimilar-compound-split) features out of the total 768 transformer-avg features; whereas, none of the variables were significantly shifted in the random-split dataset, considering both representations.

For compound fingerprints, we applied the chi-square test, since they are composed of binary variables rather than continuous ones. We didn't plot the score distributions of the chi-square test since it doesn't provide a direct distance measure. Instead, we evaluated these shifts based on their p-values. Therefore, 743 and 689 of a total of 1024 compound fingerprints were significantly shifted on the fully-dissimilar and dissimilar-compound splits, respectively, whereas this number was 47 for the random split. For significance, we accepted a p-value < 0.001 since the chi-square test is sensitive to sample size, which has the risk of falsely defining significant relationships in the presence of large sample size, as in our case.

The observation of a shift between the KS distance score distributions of models trained on fully-dissimilar and dissimilar-compound splits was not surprising since this is a common issue in real-world drug discovery applications, where the general aim is to seek completely novel small molecules that are bioactive against the targets of interest. It is also one of the reasons why most of the models, well-performing on "easy" datasets (i.e., random split), start to fail in realistic scenarios. It is possible to mitigate the shifting problem by applying preprocessing techniques such as feature dropping or importance weighting (Dharani et al., 2019), especially where the goal is to develop a model using simple descriptors and algorithms based on linear operations.

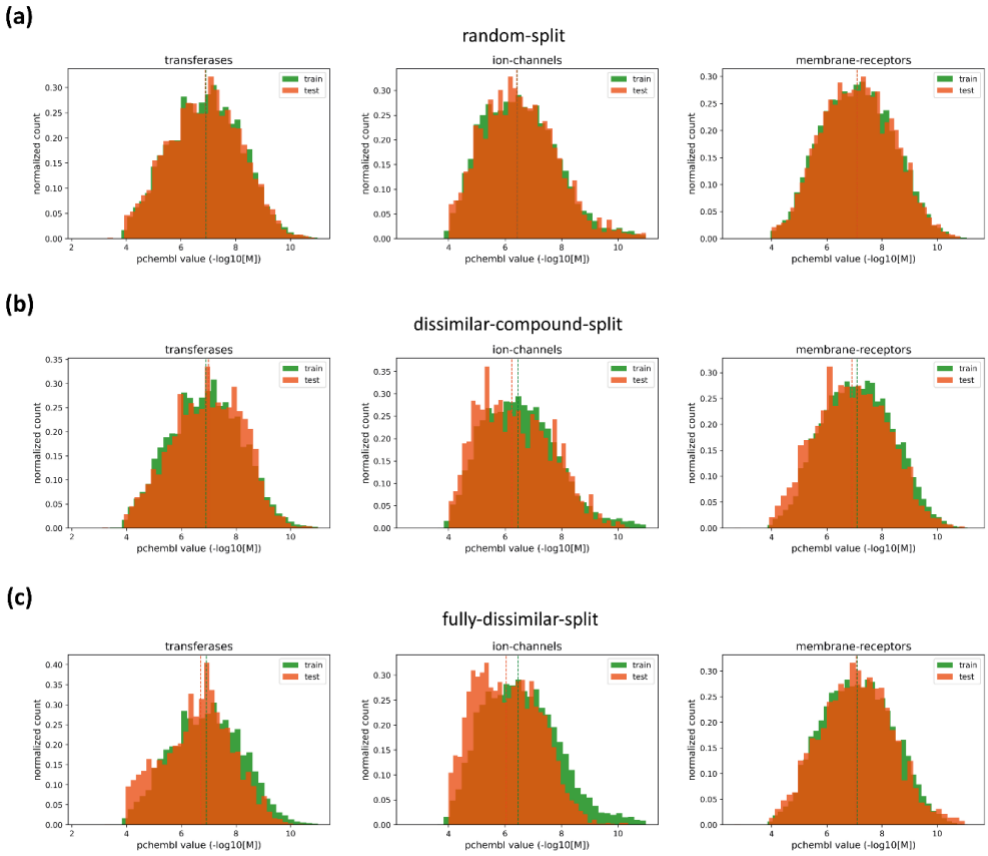


Figure 3.4. Histogram plots displaying bioactivity distributions of transferase, ion channel, and membrane receptor families based on train (green bars) and test (orange bars) samples of; **(a)** random-split, **(b)** dissimilar-compound-split, and **(c)** fully-dissimilar-split datasets, along with their median values shown as vertical dashed lines.

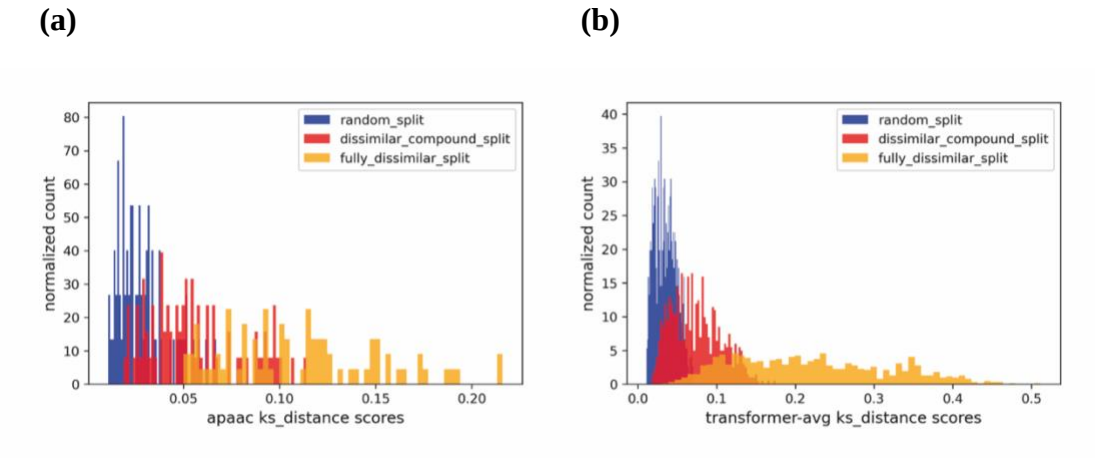


Figure 3.5. KS distance (between train and test samples) score distributions of **(a)** apaac, and **(b)** transformer-avg representations among random, dissimilar-compound, and fully-dissimilar splits in the transferases family proteins.

t-SNE projection of train-test datasets for three splits

In this analysis, we visualized the distribution of bioactivity data points (i.e., compound-protein pairs) on 2-D via t-SNE projections to observe how train and test fold samples are separated from each other under different splitting settings. For each protein family-based dataset, 1,500 data points were randomly selected (from both train and test folds), since the number of training samples dominates test samples in the original datasets. Each bioactivity data point was represented by the concatenation of its protein and compound feature vectors, and used as input to the t-SNE algorithm.

In Figure 3.6, t-SNE plots of transferases and ion channels (i.e., the representative families, as these are two widely utilized target families in drug discovery) are given for k-sep_pssm and unirep1900 representations. Panel a, b, and c correspond to the random-split, dissimilar-compound-split, and the fully-dissimilar-split datasets, respectively. For the random-split dataset, 2-D embeddings of the train and test samples largely overlap, since they share similar proteins and compounds. These overlaps significantly decrease in dissimilar-compound-split dataset and almost disappear in the fully-dissimilar-split dataset, as expected. This analysis can be considered as a visual validation of the implemented splitting strategies, and it provides clues about the difficulty levels of our prediction tasks.

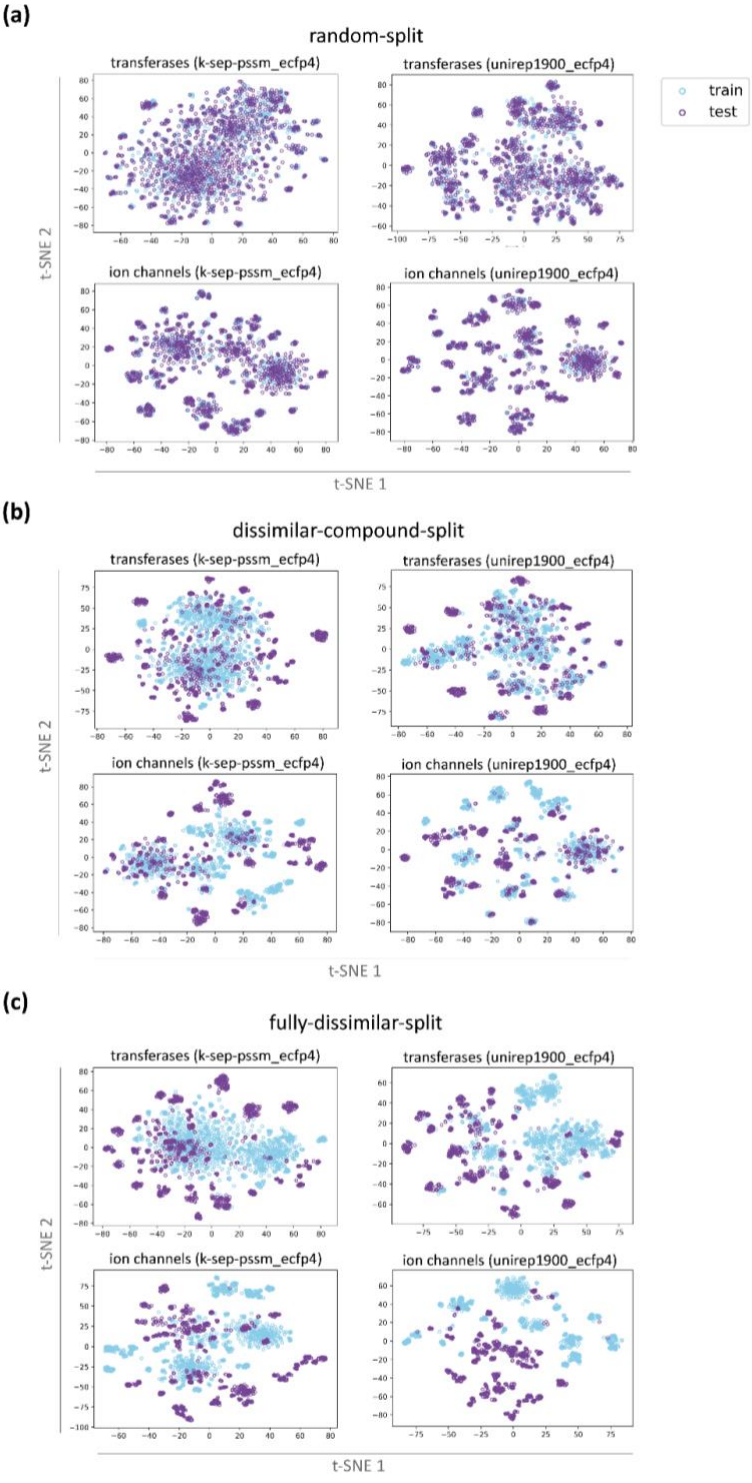


Figure 3.6. t-SNE projections of train-test samples (i.e., compound-protein pairs) of transferase and ion channel families for k-sep_pssm and unirep1900 representations on; **(a)** the random-split, **(b)** dissimilar-compound-split, and **(c)** the fully-dissimilar-split datasets.

3.4.2. Small-Scale Analysis (Target Feature-based Modelling)

There are numerous conventional descriptor sets for proteins in the literature, most of which can be utilized for DTI prediction. Evaluating all descriptor sets on our large-scale datasets would not be feasible considering the computational cost; as a result, we decided to carry out a small-scale analysis to pre-select the descriptor sets that are successful in DTI prediction, and use the selected descriptors in both the medium-scale and large-scale analysis later. Additionally, it was required to determine the supervised learning algorithm to be used for DTI prediction in this study, and due to, again, the computational complexity related issues, we decided to make a performance comparison (between SVM and RF) on these small-scale datasets.

In this analysis, we assessed the success of SVM- and RF-based DTI prediction models, each utilizing one of the 42 conventional protein descriptor sets (including the ones explained in section 3.3.2.1. and additional ones that fall into the same categories as these), and a baseline (i.e., the random200 descriptor). The models are trained and tested on 9 independent compound-centric datasets (i.e., the clusters of Curcumin, Tamoxifen, Quercetin, Genistein, Econazole, Levoketoconazole, Amiodarone, Miconazole, and Clotrimazole) via nested cross-validation using the target feature-based modelling approach (please see section 3.3.3.1). In this approach, the system only employs protein features as input, so it eliminates the effect of compound representations on the model prediction performance, which is expected to provide a suitable setting for an initial comparison of protein representations. Here, the task of each model is the binary classification of input proteins, as active or inactive, against the corresponding compound cluster.

Figure 3.7 displays mean F1-score and MCC values of 9 datasets for each representation model, in which orange and blue colors correspond to SVM and RF models, respectively (all results including accuracy, precision, recall, F1-score, and MCC metrics are given in Appendix A Table 3.3). The ranking of protein descriptor sets on the horizontal axis was done according to decreasing RF model scores. Figure 6 clearly displays that RF models outperform SVM models with a few exceptions such as the pfam model in terms of the MCC score. When model performances are compared in terms of protein representations, pssm-based descriptors perform better than other descriptors in general. These results indicate that evolutionary relationships of proteins carry important knowledge regarding bioactivity/interaction mechanisms. Some of the sequence composition-based descriptors such as dde, tpc, and spmap, and physicochemistry-based descriptors such as apaac and paac, also performed well. Moreover, obtaining scores that are significantly higher than the baseline (i.e., random200), even for the models with the lowest performance, implies that protein representations carry signals/patterns relevant to bioactivity modelling. However, these results cannot be generalized as they cover only a small portion of the bioactivity space; thus, it is important to observe how these models behave when the data scale is changed.

At the end of this analysis, we decided to continue with RF, to be used throughout the study. Also, we selected 10 conventional descriptor sets with both high and low performances, and distinct properties regarding the protein features they incorporated and used them in the following benchmarks (i.e., apaac, ctdd, ctriad, dde, geary, k-

sep_pssm, pfam, qso, spmap and taap). Here, instead of simply selecting the best-performing descriptors, we sought a diverse set of descriptors that are constructed using different types of information (i.e., physicochemistry, sequence homology, etc.). Another criterion was that the selected descriptors should not have similar performance scores (especially when they are based on the same type of information). Therefore, rather than comparing similar approaches with a high probability of yielding similar results on medium- and large-scale analyses, we attempted to acquire a representative set of descriptors, each of which has the potential to reveal a different characteristic presented in target protein sequences.

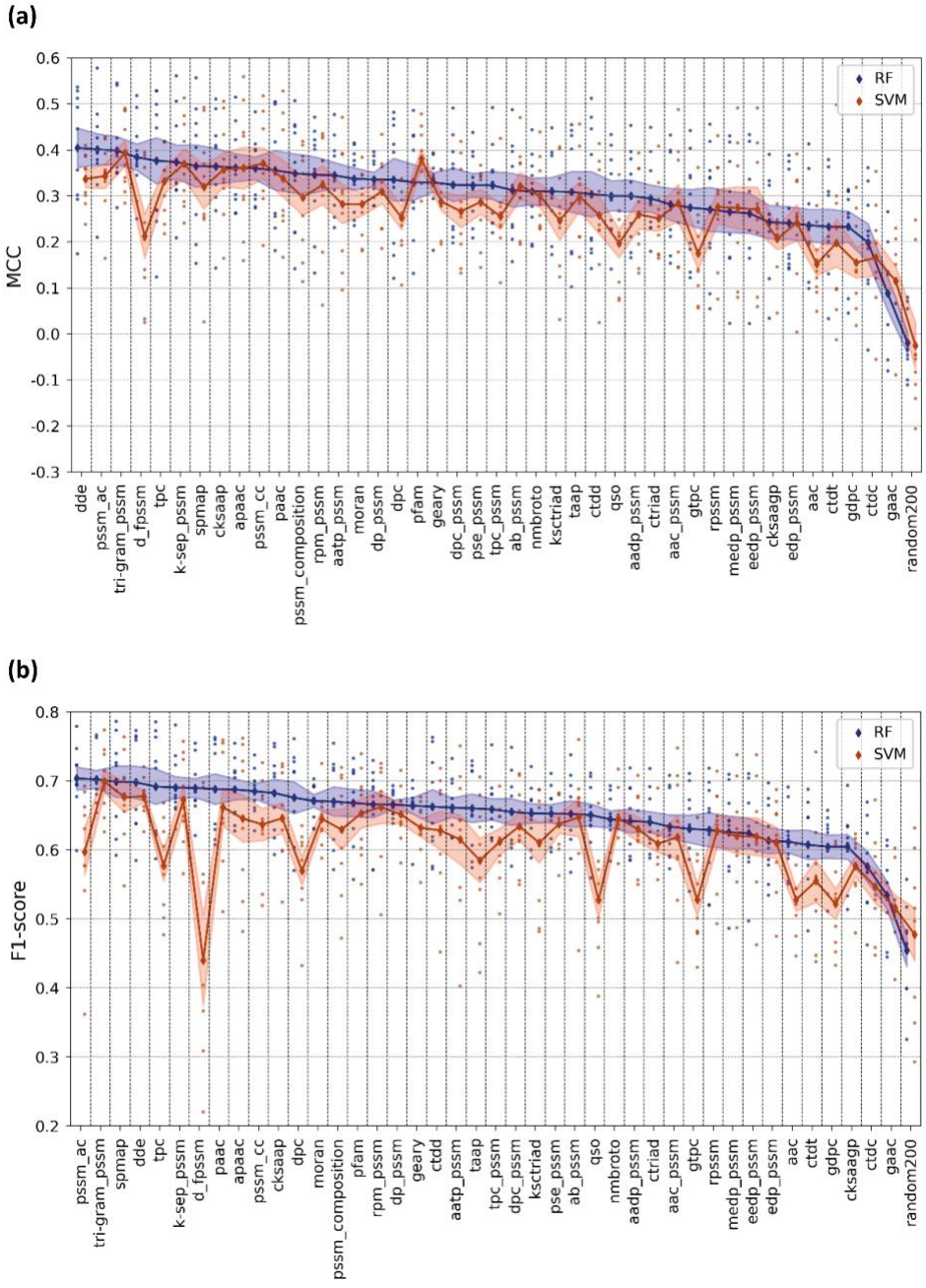


Figure 3.7. Mean **(a)** MCC and **(b)** F1-score test results of RF- and SVM-based DTI prediction models constructed via target feature-based modelling approach.

3.4.3. Medium-Scale Analysis (PCM Modelling)

PCM modelling approach can handle high numbers of training instances, belonging to different compounds and proteins, within a single predictive model, in contrast to ligand- and target feature-based modelling which requires the generation of separate models for each protein or compound (or compound cluster), respectively. Thus, PCM modeling brings the advantage of learning from larger datasets, which is a critical requirement in machine learning, in general. Another advantage of PCM modeling is the joint utilization of compound and protein features to better model their interaction-related properties, without the requirement of scarce and difficult to analyze 3-D structural information, unlike target-based structure modelling approaches. In the following benchmarks, we aimed to evaluate protein representations in terms of PCM modeling, over the problem of regression-based DTI prediction.

Here, we constructed PCM models for 10 selected conventional protein descriptor sets, 6 learned protein embeddings, and 2 baseline models (i.e., random representations, please see "Methods" section) using RF regression algorithm on the mDavis kinase dataset (please see section 3.3.3.2).

Model performance results based on RMSE, Spearman rank correlation, MCC and F1-score (all computed on the hold-out test set of the mDavis dataset) are given in Figure 3.8 (also available in Appendix A Table 3.4). The results indicate that the rankings of models are mostly consistent among both classification and regression metrics with slight differences, excluding pfam. As a domain profile-based descriptor set, pfam is the best performing model in terms of F1-score (0.538) and has a moderately high MCC score (0.41); however, it is also one of the worst performers in terms of RMSE (0.854) and Spearman (0.497) scores. It can be inferred from these results that domain profiles of proteins might not contain sufficient information to make precise bioactivity value predictions, but it can be useful if the aim is just to classify protein-compound pairs as active or inactive (i.e., binary prediction). The results also indicate that the seqvec model displays the best performance for almost all metrics (RMSE: 0.794, Spearman: 0.571, MCC: 0.445, F1-score: 0.53). Apart from seqvec, other learned embeddings also have higher performance scores compared to conventional descriptors in general. Mean Spearman rank correlation and MCC scores of learned representations are 0.530 and 0.417, respectively, whereas the same scores are 0.511 and 0.388 for conventional descriptor sets. Learned embeddings do not utilize any molecular or biological knowledge during their self-supervised training, but still, they are capable of representing proteins that yield high performance DTI prediction. Well performing descriptors in the previous small-scale analysis, k-sep_pssm (homology) and apaac (physicochemistry), also have competitive performance results here (Spearman: 0.545 and 0.532, respectively). On the other hand, dde (Spearman: 0.508) and spmap (Spearman: 0.491) could not yield their high ranks here in the medium-scale analysis (i.e., dde and spmap had the ranks of 1 and 8 on the small-scale, whereas, they ranked 9 and 16 on the medium-scale, respectively). It is possible to state that while homology- and physicochemistry-based descriptors gained from increased dataset size (i.e., for apaac and k-sep_pssm, small-scale analysis mean MCCs are 0.361 and 0.374, respectively, whereas their medium-scale analysis mean MCCs are 0.418 and 0.434), sequence composition could not improve its performance when trained on larger datasets.

Also, there is an overall increase in MCC scores of conventional descriptor sets (excluding dde and spmap) when we compare the results of small- and medium-scale analyses. In addition to the contribution of the increased sample size, this situation can be associated with the involvement of compound features in PCM-based models, which probably led to a better learning over the joint protein-compound interaction space. On the other hand, PCM models here had lower F1 scores than the target feature-based models in the small-scale analysis. In order to calculate MCC and F1-scores for PCM models, we converted real-valued predictions into binary format at the cut-off value pChEMBL = 7, which is also used in other studies as a bioactivity threshold for kinase inhibitors (Cichońska et al., 2021). However, only 27% of the test samples became active at this threshold, causing a class imbalance in the mDavis kinase dataset. Therefore, the decrease in F1-scores on the medium-scale analysis might be related to this issue, since F1-score is sensitive to imbalanced datasets (see "Performance evaluation" section in "Methods"). To further explore the conflict between MCC scores and F1-scores for the small-scale vs. medium-scale comparison, we calculated the mean performances of conventional descriptors on the medium-scale (F1-score: 0.493, MCC: 0.388), and compared them to the results of the same set of descriptors on the small-scale (F1-score: 0.672, MCC: 0.337). Then, we recalculated MCC and F1-scores of the medium-scale models based on the median pKd value of the test set to evaluate the results in such a scenario as if we had a balanced number of positive (i.e., active) and negative (i.e., inactive) samples in the test set. We obtained the mean scores of F1-score: 0.705 and MCC: 0.355 based on the cut-off pKd = 6.21 (the median value). The increase in F1-score, which is even higher than the mean F1-score in the small-scale analysis, together with the fact that there is no significant change in MCC, supports the idea that MCC is the more appropriate option in the presence of the class imbalance problem. It also highlights the importance of selecting suitable evaluation metrics depending on the case at hand.

Finally, the baseline models displayed the lowest performances in this analysis, similar to the results of the target feature-based modelling experiment.

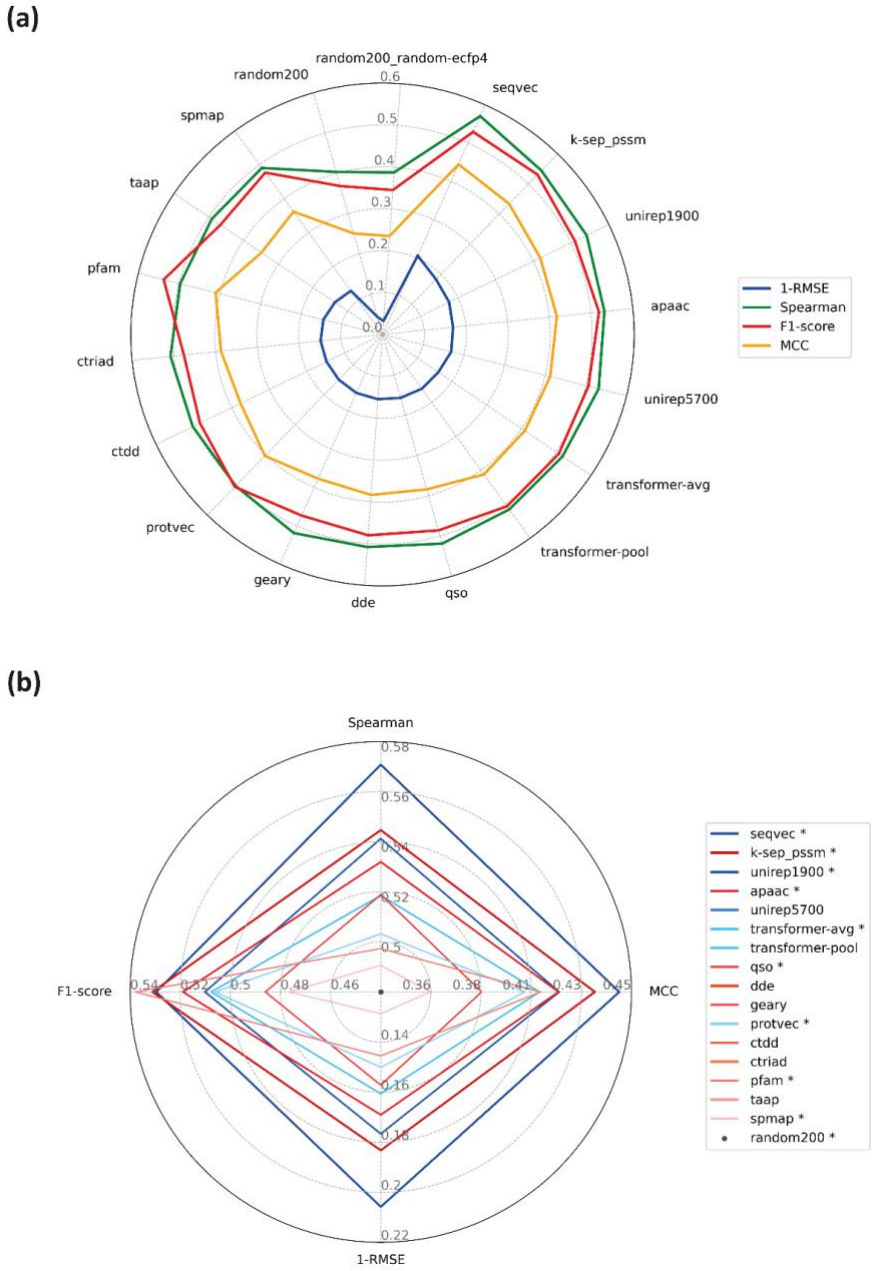


Figure 3.8. Test performance results of medium-scale PCM models (on the mDavis dataset) based on RMSE (the scores are reported as 1-RMSE, so higher values represent better performance), Spearman' s rank correlation, MCC and F1-score; **(a)** each color corresponds to an evaluation metric, and **(b)** scores are displayed only for the selected representative models (marked with asterisk in the legend). The ranking in the legend is based on the models' performance from best to worst according to their RMSE scores. Shades of red and blue represent conventional descriptors and learned representations, respectively.

3.4.4. Large-Scale Analysis (PCM Modelling)

The main goal of this analysis is evaluating protein representations over a highly realistic scenario, especially in terms of discovering new drugs and/or targets, using our carefully prepared large-scale datasets, and to compare their overall performance in machine learning-based DTI prediction. Secondly, we aimed to display how model performances can change dramatically when the same samples are distributed to train and test sets differently, to point out the importance of train-test data split. Furthermore, we evaluated the suitability of various performance metrics under different modeling approaches.

In this analysis, we constructed protein family-specific bioactivity datasets including enzyme (i.e., transferases, hydrolases, oxidoreductases, proteases, and other enzymes) and non-enzyme groups (i.e., membrane receptors, ion channels, transporters, transcription factors, and epigenetic regulators). For each family, three versions of train-test splits with differing difficulty levels were constructed by considering pairwise similarities of proteins and/or compounds (please see section 3.3.1.3 for details). PCM models were trained independently on each of these splits using the same protein representations employed in the previous (medium-scale) analysis. As a result, 600 DTI prediction models were built, trained, and tested in total (please see section 3.3.3.2 for details).

We evaluated model performances from several perspectives using multiple scoring metrics. Median corrected RMSE and Spearman correlation scores are displayed as line plots in Figure 3.9, in which the light colored (transparent) circles indicate individual model performances on each protein family, and the dark colored diamonds represent mean scores averaged over all families. The models are ranked according to descending performance on the fully-dissimilar-split dataset (for both metrics). In Figure 3.10, model performances are provided as box plots over three different forms of the MCC metric. The models are ranked according to descending mean values of median corrected MCC scores for the fully-dissimilar-split and dissimilar-compound-split datasets, and according to multiclass MCC scores for the random-split dataset. Protein family-specific performances are available in Appendix A Table 3.3.

3.4.4.1. Investigation of performance metrics

The intra-family rankings of models are generally consistent with each other among five different metrics (Table 3.5). However, there are some discrepancies between the scores depending on the data split. Considering regression metrics, some of the models trained/tested on the fully-dissimilar-split and dissimilar-compound-split datasets show high performance in terms of RMSE (i.e., low RMSE values), whereas at the same time, they displayed low Spearman correlations, which indicates inconsistency. RMSE is a measure of the difference between predicted and actual values, and is utilized when the goal is to predict continuous values and measure the overall error in predictions. On the other hand, Spearman's rank correlation is a measure of the strength and direction of the relationship between two ranked variables. Spearman's correlation is commonly used when the goal is to determine the degree to which two variables are related. In challenging scenarios (e.g., on the fully-dissimilar-split and dissimilar-compound-split datasets), continuous value-based prediction of

bioactivities (via regression) is unstable and unreliable due to the difficulty of the task. Thus, it would be a better choice to evaluate the success of the models in terms of the correlation and consistency between actual and predicted values using correlation scores (e.g., Spearman's). On the random-split dataset, the prediction task is not considered to be difficult (relative to the other two splits), as a result, the predicted values are expected to be more stable and reliable. Using RMSE metric in this scenario allows us to directly measure the accuracy of the predictions and differentiate the model performances in a more precise manner. As a result, both types of scores can be considered for easy cases (i.e., the random-split dataset). In classification-based assessment, the single-class MCC metric is not as restrictive as the regression or multiclass evaluation metrics since it is less sensitive to deviations in prediction values. However, it may suffer from the shifted mean problem when applied to regression-based PCM models by binarizing bioactivity values. Obtaining MCC values close to 0 (Figure 3.10) despite moderate Spearman correlation scores (Figure 3.9) on challenging datasets is a sign of a systematic shift in model prediction outputs, which we handled by conducting median correction on the real-valued prediction results. In Figure 3.10, it can be observed that median correction provided a significant increase in single-class MCC scores of the fully-dissimilar-split and dissimilar-compound-split datasets. Also, median corrected MCC scores are highly consistent with the Spearman correlation scores (Appendix A Table 3.5). Considering the multiclass MCC metric, prediction scores are around zero for most of the models on challenging split sets. Since this metric expects prediction values to fit narrow intervals, it is more restrictive than the single class-based metrics. However, this seems to be an advantage for evaluating models on the random-split set. As seen in Figure 9a, on the random-split dataset, the variance of the mean multiclass MCC score distribution is greater than the single-class MCC scores (i.e., models are better separated from each other). Furthermore, its ranking is highly consistent with the results of the medium-scale experiments, in which the top performers were learned representations, together with k-sep_pssm and apaac conventional descriptor sets. Thus, it can be inferred that the multiclass MCC metric discerns models better than binary class MCC in the random data split setting, and it partly handles the overfitting problem which frequently occurs on randomly split large-scale datasets.

5.4.6.2. Evaluation of protein representations

Performance results in Figure 3.9 and 3.10 indicate that the representation capability of different protein descriptor sets depends on the protein family and the difficulty level of the split used for training and testing. Also, there is no significant difference between the mean performances of different protein representations for a particular dataset split, with a few exceptions. Considering family-based performance averages, pfam is one of the best representations on the fully-dissimilar-split and dissimilar-compound-split datasets, while it is the lowest performer on the random-split dataset (Figure 3.9 and 3.10). Contrary to pfam, k-sep_pssm is one of the best performers on the random-split and dissimilar-compound-split datasets but the worst one on the fully-dissimilar-split dataset (Figure 3.9 and 3.10), though the performance results on the random-split dataset are very close to each other. As a homology-based descriptor set, k-sep_pssm is expected to capture hidden similarities between evolutionarily related sequences, especially by taking advantage of the presence of highly similar proteins

between the train and test splits. On the other hand, the utilization of protein domain profiles seems to make pfam more suitable for acquiring bioactivity related information from evolutionarily distant sequences, probably due to highly sensitive HMM-based domain/family profile search procedures implemented in Pfam and similar databases. Interestingly, taap, which is a simple descriptor set, is involved in the top-performing PCM models for all dataset splits. However, taap was one of the lowest performers in the small- (among the selected 10 conventional descriptor sets) and medium-scale analyses. Its simplicity is observed to become an advantage with the increase in bioactivity dataset size and complexity. Apart from these, physicochemistry-based descriptors including apaac (in all splits), ctriad (on the fully-dissimilar-split dataset) and qso (on both the fully-dissimilar-split and dissimilar-compound-split datasets), and learned representations perform well in the large-scale analyses. In particular, the top performance results of unirep5700 and transformer-avg on the fully-dissimilar-split dataset demonstrate the potential of protein representation learning methods in the data-driven DTI prediction.

We also conducted protein family-specific evaluations to understand whether different protein representations display similar results across families. In Figure 3.11, we plotted the performance of the models of protease and the ion channel families, in the form of a conventional descriptor set vs. learned representation comparison, using the Spearman and median corrected MCC scores, for all three dataset splits. Ion channels are known for their transmembrane regions and specific ion selectivity, whereas proteases are enzymes involved in catalyzing peptide bond cleavage. For a fair comparison, we selected four well-performing conventional descriptors instead of including all of them, since we have only four different types of learned representations. For this, we involved apaac, k-sep_pssm, pfam, and taap as conventional descriptor sets and protvec, seqvec, transformer-avg, and unirep5700 as learned representations. Figure 3.11 shows that learned representations outperform conventional descriptors in the challenging splits of proteases, considering both metrics. However, the results are the opposite for the ion channel family, on which the conventional descriptor sets performed better. One possible reason for this might be due to distinct structural and functional characteristics of ion channels that can be detected more easily via conventional descriptors, which leverage the physicochemical properties of amino acids, evolutionary information, or domain profiles of proteins. In contrast, learned embeddings may struggle to capture these characteristics, particularly when the dataset sizes are relatively small, as in this case (i.e., around 30K training data points for ion channels, while substantially larger for proteases, with approaximately 85K data points). On the random-split dataset, there is no observable difference between conventional descriptor sets and the learned representations, probably due to the non-discriminative characteristic of this data splitting strategy, which poses non-challenging cases for all models.

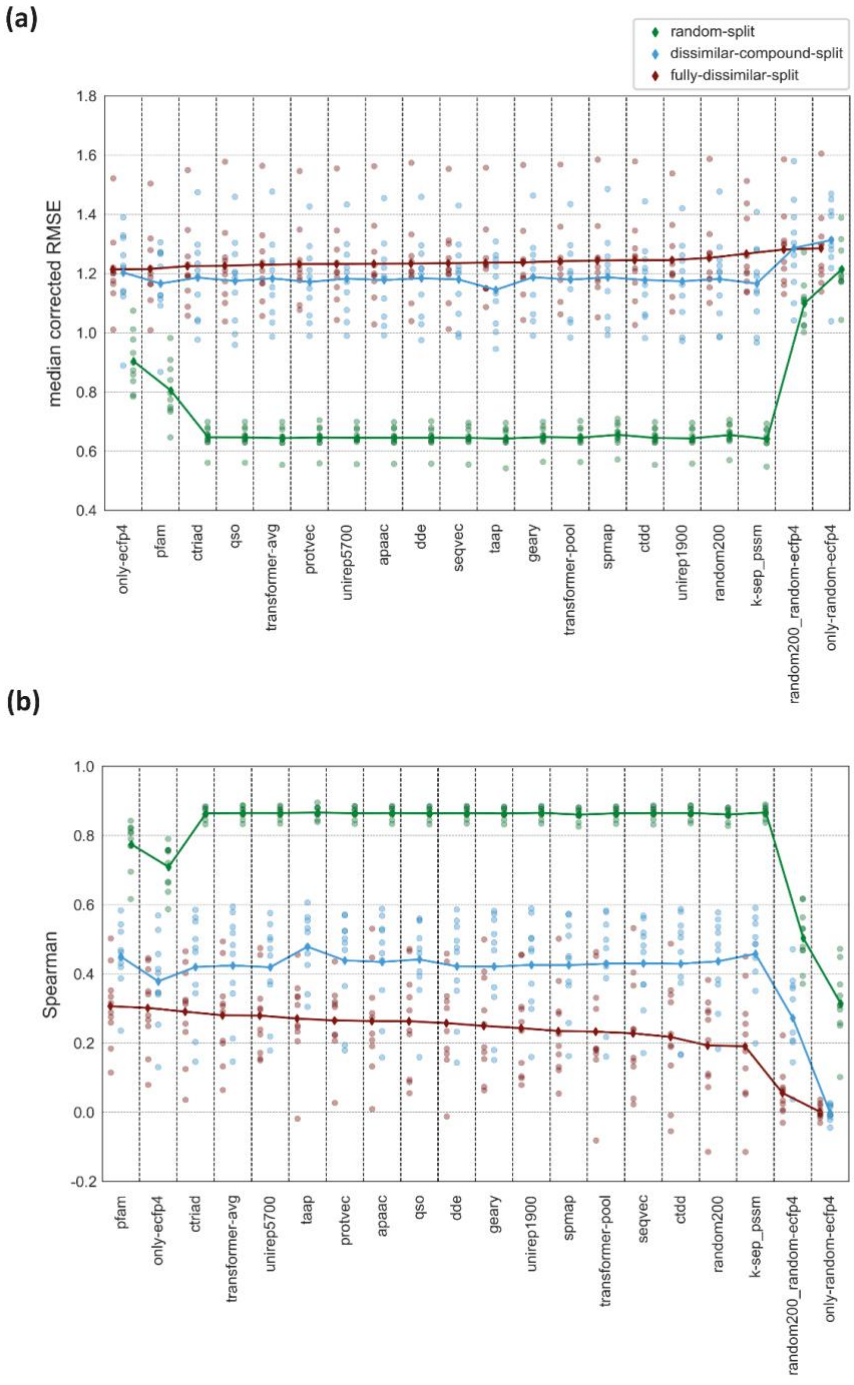


Figure 3.9. Regression-based test performance results of protein family-specific PCM models (each using a different representation type as input feature vectors) for random-split, dissimilar-compound-split, and fully-dissimilar-split datasets based on **(a)** median corrected RMSE, and **(b)** Spearman correlation scores. The models are ranked according to decreasing performance on the fully-dissimilar-split dataset.

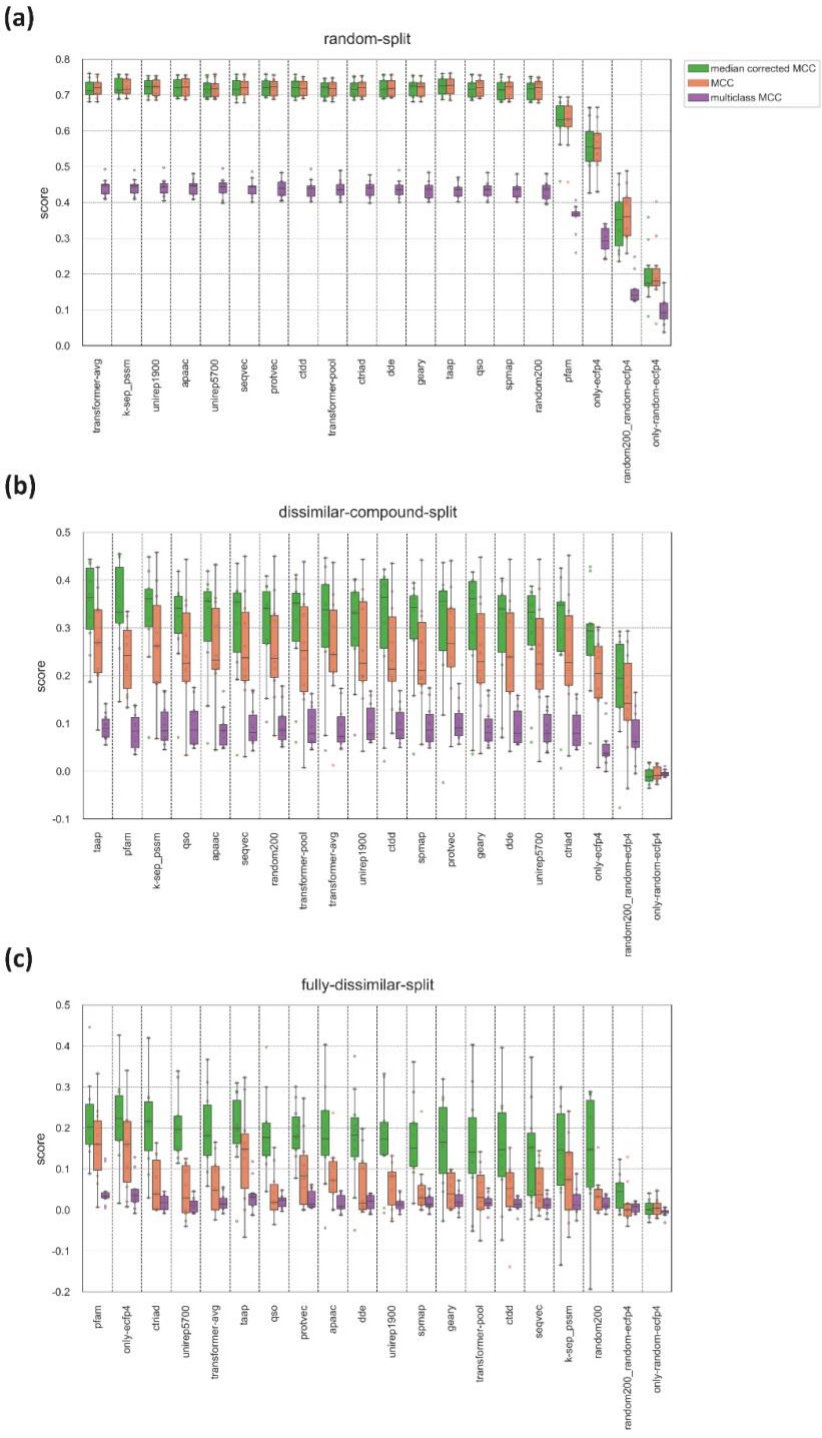


Figure 3.10. Classification-based test performance results of protein family-specific PCM models (each using a different representation type as input feature vectors) in terms of MCC scores for **(a)** random-split, **(b)** dissimilar-compound-split, and **(c)** fully-dissimilar-split datasets. The models are ranked according to decreasing performance on the fully-dissimilar-split dataset.

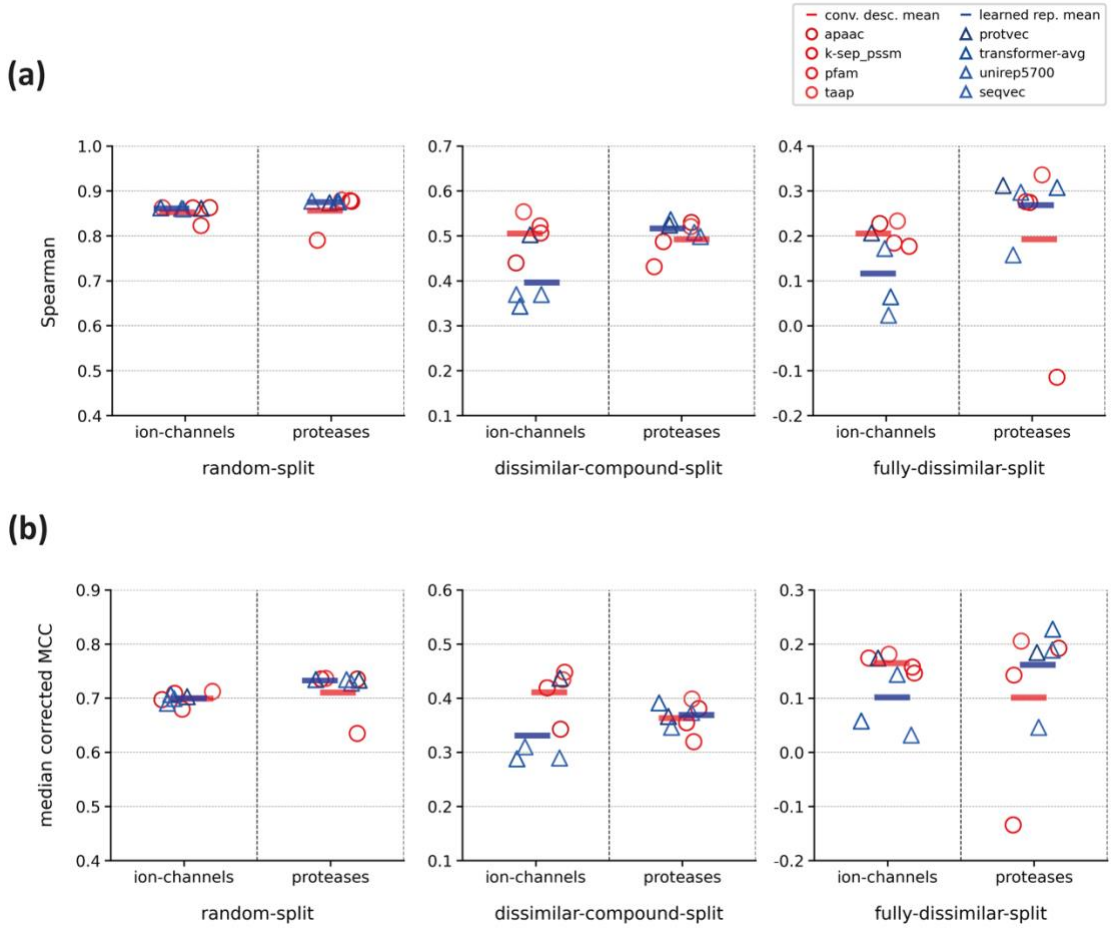


Figure 3.11. Performance comparison of well-performing conventional descriptor sets and learned representations for three different splits of ion-channel and protease family datasets in terms of; **(a)** Spearman rank correlation, and **(b)** median corrected MCC scores.

Results presented in Figure 3.11 are also correlated with the scores on other protein families (Appendix A Table 3.5). For non-enzyme families, the average Spearman's correlation values (based on the representations in Figure 3.11) are 0.29 (cd: conventional descriptors) and 0.26 (le: learned embeddings) in the fully-dissimilar-split, 0.40 (cd) and 0.34 (le) in the dissimilar-compound-split, and 0.84 (cd) and 0.87 (le) in the random-split datasets. For enzyme families, these values are 0.23 (cd) and 0.26 (le) in the fully-dissimilar-split, 0.51 (cd) and 0.52 (le) in the dissimilar-compound-split, and 0.84 (cd) and 0.86 (le) in the random-split datasets. The results show that, in challenging datasets, conventional descriptors perform better on non-enzyme families, while learned embeddings perform better on enzyme families. It suggests that the type of protein representation used can have an impact on the model performance depending on the type/family of protein being studied, possibly due to the intrinsic properties of these protein families. This observation can be useful for developing new strategies to improve model performances. All of the learned representations in our study were obtained from unsupervised deep learning models trained on large datasets including all protein families. Limiting the training datasets

of these methods to specific families (or fine-tuning the pre-trained models on these families) would increase their representation power towards that family.

When taking all these findings into account, we can clearly state that the representation capabilities of different featurization approaches considerably vary among protein families and splitting strategies, even though some common inferences can be made. We believe that, while choosing a featurization approach in DTI prediction, protein family-specific findings should be taken into account, rather than considering the overall (i.e., average) results. Regarding learned representations, re-training (or fine-tuning) the models using a distinct dataset with desired characteristics (e.g., members of a certain family) may be a good choice to better learn the features associated with that group of proteins.

3.4.4.3. Comparison of data splitting strategies

To compare models across three dataset splits, we plotted performance scores by pooling 200 models of each split (including the baseline models) without grouping by families or representation methods. The results are displayed in Figure 3.12 via violin plots. This figure shows a significant decrease in overall performances with the increasing difficulty levels of splits, which is not a surprising outcome. Nevertheless, it highlights the importance of splitting datasets into train/test folds for performance evaluation, with the aim of preventing the reporting of over-optimistic results and yielding a fair assessment of model successes. Figure 3.12 also displays that the model performances are distributed more evenly over the whole range of scores in the fully-dissimilar-split and dissimilar-compound-split datasets, compared to the random-split dataset, in which most of the models produced very similar scores, creating a dense region on the plot. This observation indicates that random splitting has less power in distinguishing different models from each other.

In the fully-dissimilar split, neither similar proteins nor similar compounds are shared between train and test folds. As a result, this dataset is suitable to evaluate the performance of DTI prediction models in terms of predicting novel ligands to understudied targets (or completely new target candidates). Whereas in the dissimilar-compound split, similar proteins are presented in between train and test sets. Nevertheless, it is useful for discovering novel ligands against well-studied target proteins, or proteins for which structurally highly similar and well-studied targets exist.

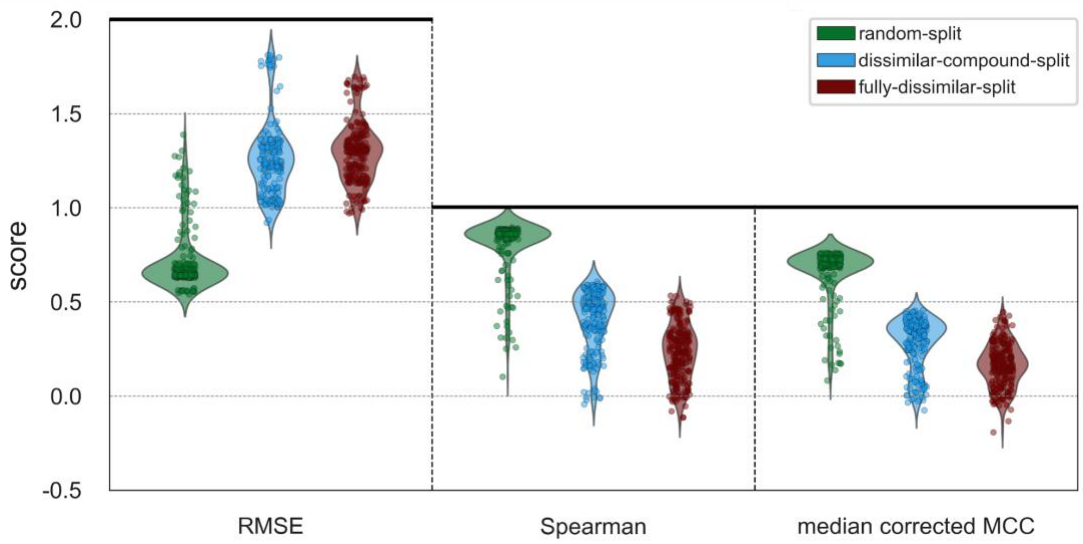


Figure 3.12. Split-based test performance scores of family-specific PCM models in terms of RMSE, Spearman rank correlation, and median corrected MCC metrics.

3.4.4.4. Examination of baseline models

Table 3.2 contains family-based average Spearman scores of the best-performing models and the baseline models, for each dataset split. The models based on randomly generated protein and/or compound representations have lower performance scores on the fully-dissimilar-split dataset, which is mainly due to the absence of identical proteins and compounds (or ones with high similarity) in between train and test samples. One of the baseline models included in this analysis uses only compound representations (i.e., only-ecfp4 model). This model does not utilize a protein vector. As a result, the model learns activities over the compound features only, without any information regarding which protein this compound interacts with. This is different from a conventional ligand-based DTI prediction model, in which target proteins would be used as labels of the input compounds (i.e., as "a target of protein X" or "not a target of protein X"). Here, since the information about proteins is not utilized at all, the model tries to learn interactions blindly and make predictions without knowing which target it is giving predictions for.

The average Spearman correlation score of the best-performing model on the fully-dissimilar-split dataset is around 0.3, which is quite close to the only-ecfp4 model. This indicates that the success obtained by even the best model has mostly originated from the characteristics of compounds (i.e., a certain compound being active no matter which target it has been screened against, or another compound being inactive in most of the experiments). Thus, these results reveal the requirement for; (i) unbiased model training datasets, and (ii) novel/improved featurization techniques, to construct robust DTI prediction models that can be utilized in the pharmaceutical industry, especially under these challenging scenarios.

Model performances are higher on the dissimilar-compound-split dataset compared to the fully-dissimilar-split dataset, due to the inclusion of similar (and identical) proteins

between training and test. Also, models based on completely random vectors (on both the compound and protein sides) have lower performances, expectedly. On both of the challenging datasets, the best model is well differentiated from the random vector-based baseline models. Although the overall mean difference between the best model and the random200 model is considerably low on the dissimilar-compound-split, the differences are distinct when making protein family-specific comparisons rather than taking the average of all families (e.g., for ion channels; the average Spearman score of the top performing models including k-sep_pssm, pfam, taap, and protvec is 0.52, and the Spearman score of random200 model is 0.37). On the dissimilar-compound-split dataset, the random200 model outperformed the only-ecfp4 model by learning the relationship between the bioactivity data points of the same proteins which are shared between training and test. As experimental bioactivity measurements are mainly obtained from target-based assays, the number of bioactivity data points per protein is considerably high, compared to the number of bioactivity data points per compound (Table S3 and S4). Also, in many assays, different derivatives of the same compound are tested, which results in similar bioactivity values. Due to this bias in experimental assays, memorization over protein identity yields falsely successful results, as reflected in the performance of the random200 model on the dissimilar-compound-split dataset (average Spearman score = 0.436).

On the random-split dataset, the best model displays a high success rate (Spearman score: 0.868). However, high performance scores of the baseline models, including those based on randomly generated vectors (e.g., random200), clearly indicate the over-optimistic evaluation, and emphasize the importance of train-test data splitting, once again. These results also demonstrate the importance of baseline model-based investigation in the field of DTI prediction, for a fair and realistic performance evaluation. It is possible to state that, the results reported in previous DTI prediction studies in which (i) the models are only evaluated based on random splitting (including both hold-out testing and fold-based cross-validation), and (ii) there is no proper baseline model comparisons, may be invalid.

Table 3.2. Protein family-based average Spearman scores of the best models and baseline models in each dataset split.

Name of the descriptor set/representation (explanation)	**Fully-dissimilar-split**	**Dissimilar-compound-split**	**Random-split**
Best performing protein representation (compound: ECFP4)	0.363	0.518	0.868
random200 (protein: random continuous vectors, compound: ECFP4)	0.193	0.436	0.861
only-ecfp4 (no protein vector, compound side: ECFP4)	0.302	0.379	0.709
random200-random-ecfp4 (protein: random continuous vectors, compound: random binary vectors)	0.056	0.272	0.504
only-random-ecfp4 (no protein vector, compound side: random binary vectors)	0.002	-0.002	0.315

3.4.4.5. Exploration of the prediction similarities between family-specific PCM models

In this experiment, we plotted heatmaps based on pairwise similarities between the protein family-specific PCM model predictions via calculating their intersections, using a categorization composed of six classes (i.e., pChEMBL value bins of <5, 5.0 to 5.5, 5.5 to 6.0, 6.0 to 6.5, 6.5 to 7.0, and 7.0>=). To calculate the similarity between a pair of models, for each bioactivity data point, we count a similar prediction if both models predict pChEMBL values in the same bin (no matter they are correct or not), otherwise we count a non-similar prediction. We then calculate percent similarity values based on all counts. To emphasize prediction similarity values between model pairs, color scales were arranged so that the darkest color corresponds to the maximum value, and the lightest color was set to 85%, 65%, and 20% similarity for the random-split, dissimilar-compound-split, and the fully-dissimilar-split datasets, respectively.

In Figure 3.13, heatmaps of transferase and ion channel families are given for all three dataset splits (heatmaps for the remaining families are available in Appendix B Figure 3.1). As observed from Figure 3.13, the overall consensus between models decreases with increasing difficulty levels (i.e., the average similarity is over 80% for most of the models in the random-split dataset, while this value drops to 30-60% in the fully-dissimilar-split dataset). Although clusters vary across different splits and protein families, generally the learned embeddings and physicochemistry-based conventional descriptors are clustered among themselves. Considering the fully-dissimilar-split dataset of transferases; the average prediction similarity between the models that utilize learned representations (except protvec) is 60.8%, and among the models that use physicochemistry-based conventional descriptor sets (i.e., qso, apaac, geary, ctriad) is 68.2%, whereas the average prediction similarity between the physicochemistry-based conventional vs. learned representations (considering the same models) is 46.5%. These findings are also parallel to the t-SNE projection results provided in Figure 3.2. Considering the type of utilized information, all learned representations exploit the arrangement of amino acids on the protein sequence. On the other hand, physicochemistry-based descriptors aggregate pre-calculated amino acid-based features to construct protein feature vectors. This difference is also reflected in their prediction similarities. Spmap and random200 representations are often clustered together and have similar t-SNE projections, as well. Finally, models that utilize pfam and taap descriptor sets are quite differentiated from the rest on the random-split and dissimilar-compound-split datasets, which is expected based on their distinct featurization strategies.

The results of this analysis can be used to obtain rational combinations of featurization approaches to better represent proteins in DTI prediction models (e.g., concatenating feature vectors that have a low correct prediction overlap). This may yield a more successful learning of interaction-relevant properties of proteins, and significantly improve the overall model performances.

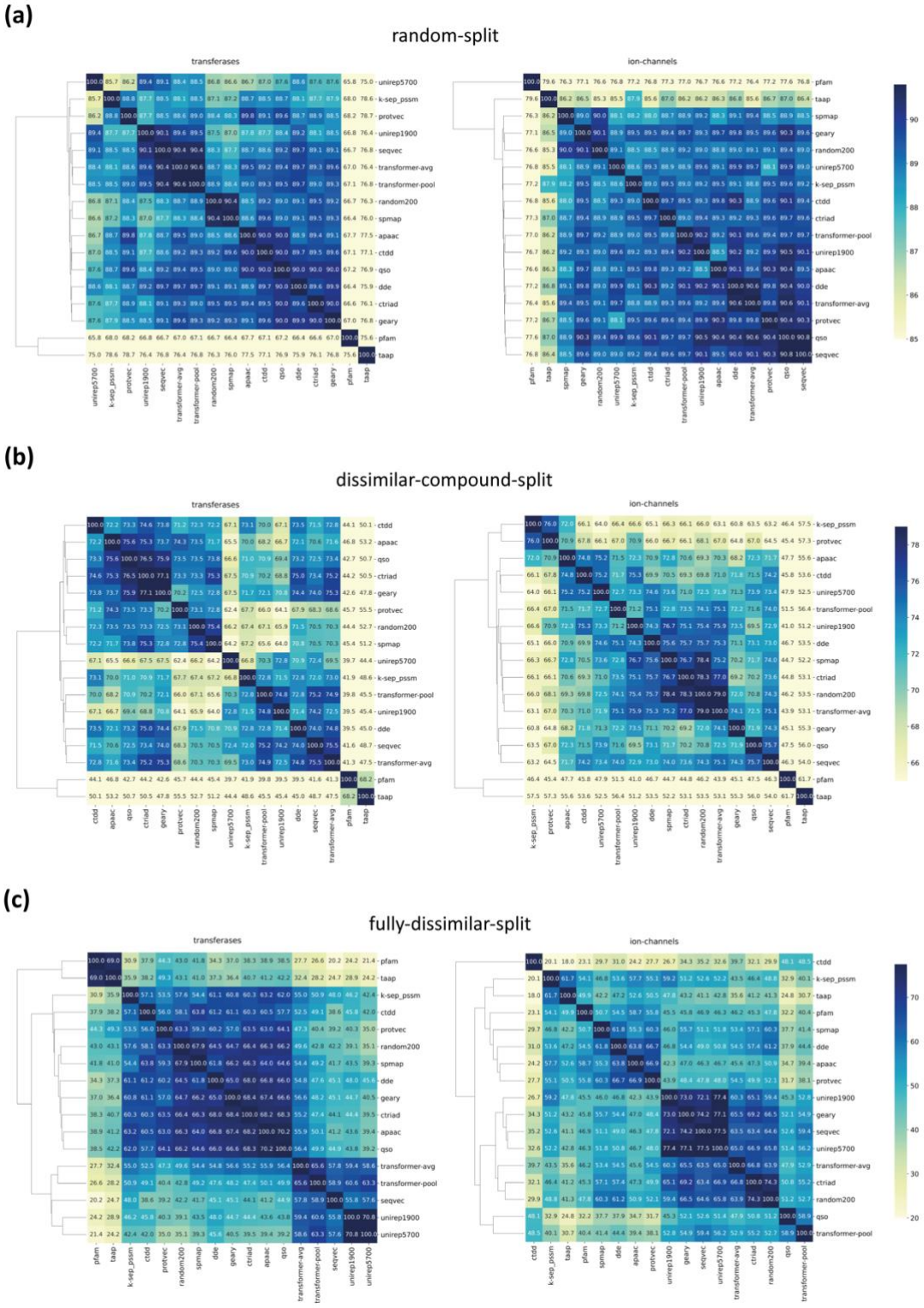


Figure 3.13. Clustered heatmaps of different protein featurization approaches for transferase and ion channel families on; **(a)** the random-split, **(b)** dissimilar-compound-split, and **(c)** the fully-dissimilar-split datasets.

3.4.4.6. *Applicability domain (AD) analysis of family-specific PCM models*

The concept of AD is used to define the boundaries of a model within which is expected to provide accurate and reliable predictions, and to assess its usability. It has been included as an essential requirement for QSAR models by the Organization of Economic Co-operation and Development (OECD). In the scope of QSAR modeling, AD is defined as the chemical structure space in which the model produces reliable predictions (Hanser et al., 2016). It is significant because the reliable predictions of a QSAR model are typically restricted to query compounds that share high structural similarities with the training compounds (Sahigara et al., 2012). In contrast to QSAR models, PCM modeling approach takes both protein and compound space into account and has the potential to reveal complex relationships between them since the model performance is not solely based on the similarity of compounds. Although the concept of AD is not directly applicable to PCM modeling, there have been some efforts to evaluate the AD of PCM models using k-nearest neighbors (k-NN) (Ain et al., 2014; Subramanian et al., 2017) and Gaussian processes (GP) (Cortes-Ciriano et al., 2014).

In this study, we employed the k-NN approach to assess the AD of our models. For this, we first calculated Tanimoto similarities between test and training compounds based on their ecfp4 fingerprints. For each test compound, we calculated the average Tanimoto score of the most similar five training compounds (i.e., 5 nearest neighbors), as described in the study by Subramanian et al. (Subramanian et al., 2017). Then, we applied the same strategy for test proteins using sequence similarities mentioned in the "Pairwise similarity distributions" sub-section. In Figure 3.14, we plotted compound and protein similarities vs. prediction errors for each test datapoint in random-split, dissimilar-compound-split, and fully-dissimilar-split sets of the transferases family dataset for transformer-avg based models.

The figure displays that most of the data points with high similarities of proteins and compounds have low prediction errors, but there is no direct correlation between similarity and error values as usually observed in QSAR models. At each similarity percentage interval, there are data points with low and high prediction errors at varying frequencies, even at extremely low similarities. This confirms the extrapolation ability of the PCM modeling approach. However, the number of data points with higher error increases in challenging datasets, which narrows the applicability domain of the models on these datasets. The average prediction error (e) and similarity values of proteins (p) and compounds (c) based on Figure 3.13 are 0.48 (e), 66% (p), 77% (c) for random-split, 0.92 (e), 64% (p), 35% (c) for dissimilar-compound-split, and 0.94 (e), 23% (p), 33% (c) for the fully-dissimilar-split, respectively. These values also indicate that the changes in the similarity of compounds have a higher impact on the error, compared to proteins. The results were similar in our other models, as well. It is possible to infer from these results that PCM models tend to utilize compound features more than protein features, mostly due to the natural bias in DTI data.

Overall, these results indicate that models can reliably predict a considerable amount of the test dataset (i.e., 88%, 59%, and 61% of test samples are predicted with errors < 1 in random-split, dissimilar-compound-split, and fully-dissimilar-split sets, respectively; Table 3.3). However, it is also possible to state that the applicability is limited in challenging datasets. The shift between the input feature value distributions

can be one of the main reasons behind obtaining a lower performance and a narrower range of applicability for the models trained on fully-dissimilar and dissimilar-compound splits (Figure 3.5). At the same time, this is a natural part of the problems at hand, which are discovering truly novel drugs and/or effectively targeting understudied proteins. While it is possible to improve performances to some extent by applying preprocessing techniques, classical machine learning methods, and available representation approaches are only partially sufficient to handle the DTI prediction problem in realistic scenarios. Therefore, more advanced approaches such as multi-modal deep learning and new comprehensive representations, specifically developed for bioactivity modeling, are required to effectively unveil non-linear relationships between target proteins and drug candidate compounds.

Table 3.3. Prediction error percentages of transformer-avg models with different thresholds on random, dissimilar-compound, and fully-dissimilar splits of transferases family dataset.

	PE > 0.5 (%)	**PE > 1 (%)**	**PE > 1.5 (%)**	**PE > 2 (%)**
Random-split	36.6	11.8	3.6	0.9
Dissimilar-compound-split	69.6	40.8	17.8	6.9
Fully-dissimilar-split	64.9	39.0	22.5	10.7

PE: Prediction error

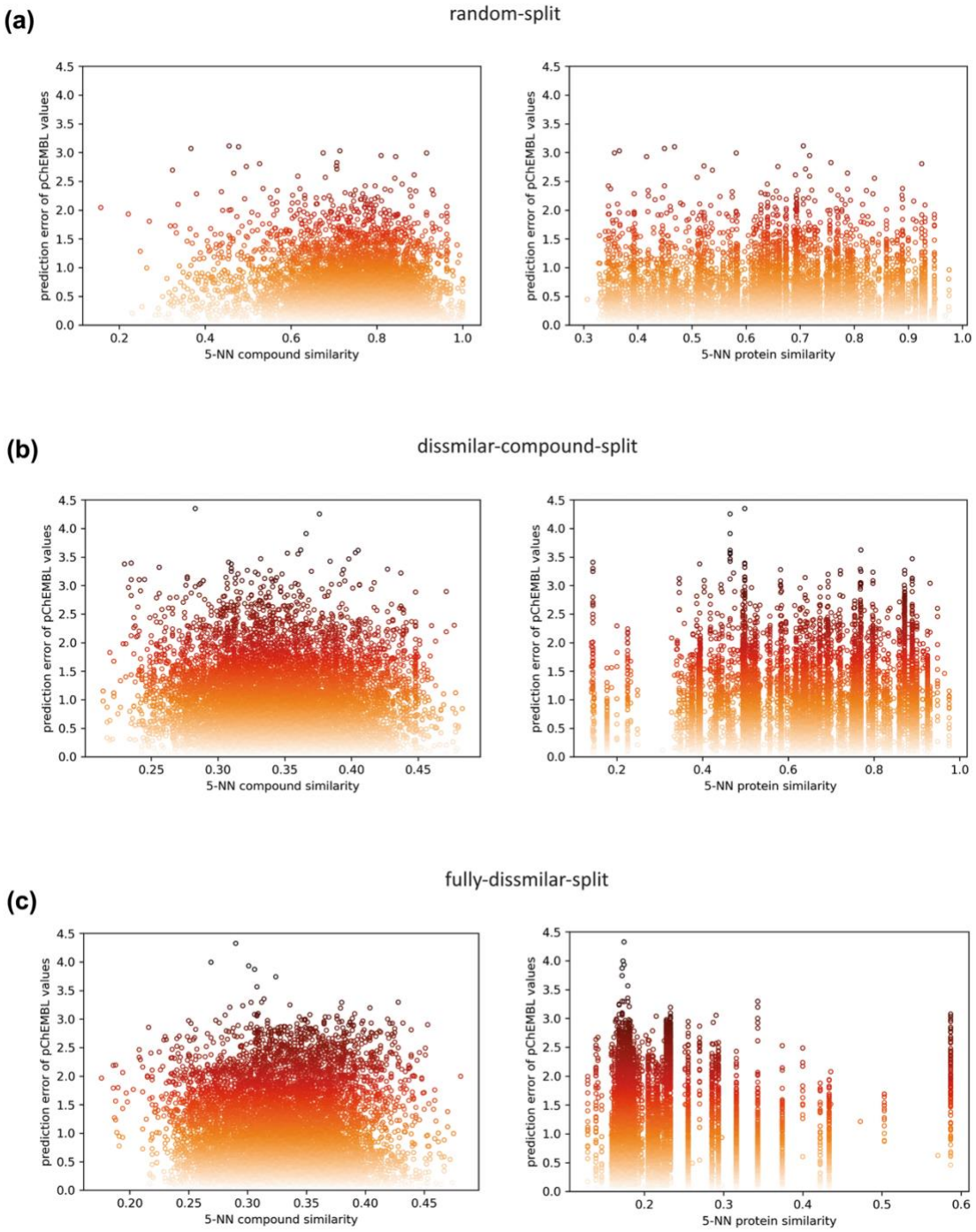


Figure 3.14. Scatter plots of compound similarities and protein similarities against prediction errors of test data points in **(a)** random-split, **(b)** dissimilar-compound-split, and **(c)** fully-dissimilar-split sets of transferases for transformer-avg models

3.5. Conclusion

In this chapter, we performed a rigorous benchmark analysis to investigate; (i) bioactivity datasets at different scales and their splitting into train-test folds, (ii) preliminary and explanatory analysis of data, (iii) different modeling and algorithmic

approaches, (iv) the representation capability of various protein featurization techniques, and (v) robust and fair performance evaluation strategies, for machine learning-based DTI prediction modelling. For this, we built target feature-based and PCM-based models, and trained/tested them on carefully constructed datasets with varying sizes and difficulty levels, using numerous protein representations, and evaluated them from different perspectives. Datasets, results and the source of the study in this chapter is fully shared in our "ProtBENCH" platform at https://github.com/HUBioDataLab/ProtBENCH.

Below, we summarized the major contributions of this chapter to the literature:

(i) We proposed challenging benchmark datasets with high coverage on both compound and protein spaces that can be used as reliable, reference/gold-standard datasets for DTI modelling tasks. These datasets are protein family-specific, and each has three versions in terms of train/test splits for different prediction tasks (i.e., random split for predicting known inhibitors for known targets, dissimilar-compound split for predicting novel inhibitors for known targets, and fully-dissimilar split for predicting new inhibitors for new targets). Thus, they yield fair evaluation of models at multiple difficulty levels and facilitate the prevention of over-optimistic performance results. We evaluated these datasets in the framework of PCM modeling, which is a highly promising data-driven approach for high performance ML-based drug discovery. These datasets can be used in future studies to evaluate newly proposed modeling and/or algorithmic techniques for DTI prediction.

(ii) We employed a network-based strategy for splitting data into train-test folds, by considering both protein-protein and compound-compound pairwise similarities, which is proposed here for the first time, according to our knowledge. This strategy ensures that train and test folds are totally dissimilar from each other with a minimum loss of data points. One of the current limitations in drug development is the problems related to discovering novel molecules that are structurally different from existing drugs and drug candidates. The network-based splitting strategy we applied here forces prediction models to face this limitation by supplying more realistic, hard-to-predict test samples. Hence, it can aid researchers in designing more powerful and robust DTI prediction models that have a real translational value.

(iii) Protein representation learning have a wide range of applications with promising results in different sub-fields of protein science, despite being a relatively new approach. However, the studies regarding their usage in DTI prediction modeling are limited, and there is no comprehensive benchmark study to evaluate their performance against well-known and widely used featurization approaches. Due to this reason, we extended the scope of our study by involving state-of-the-art learned representations and discussed their potential in DTI prediction.

One of the critical observations of this study is the dramatic change in performance scores when the samples are distributed to train and test sets differently, (i.e., scores on datasets with challenging splits are significantly lower compared to the results on randomly split datasets), which highlights the importance of data splitting to conduct

realistic evaluations for drug and/or target discovery. This study also emphasizes the importance of exploratory analysis of datasets and the usage of multiple scoring metrics as well as the inclusion of baseline models for a proper discussion of model successes.

Regarding the performance-based comparison of different protein featurization approaches, it is not possible to put forward an outstanding representation method, as their success largely depends on the dataset and the applied splitting strategy. Although both conventional descriptor sets and learned embeddings have their own strengths and weaknesses depending on the case, competitive results of learned embeddings display their potential widespread utilization in drug discovery and development in the near future. On the other hand, considerably low performance results on challenging datasets (e.g., fully-dissimilar-split) in the overall evaluation revealed the requirement for unbiased bioactivity datasets and further improved protein representation techniques to capture hidden and complex features shared between highly distant homologs.

We hope that the results of this chapter, together with the data-driven approaches proposed, and the benchmark datasets prepared and shared, will aid the ongoing work in computational drug discovery and repurposing.

CHAPTER 4

CROssBAR: GENERATION AND ANALYSIS OF BIOMEDICAL KNOWLEDGE GRAPHS

4.1. Chapter Overview

This chapter was performed as part of the CROssBAR project, which is co-funded by TÜBİTAK-Katip Çelebi & British Council-Newton fund and conducted jointly with METU, Hacettepe University and EMBL-EBI as an international project. CROssBAR (Comprehensive Resource of Biomedical Relations with Deep Learning Applications and Knowledge Graph Representations) is a comprehensive system that integrates large-scale biomedical data from various resources including relations between numerous biomedical entities such as genes/proteins, drugs/compounds, disease/phenotype terms and pathways/biological processes. The main goal of the CROssBAR project is to build an open access, user-friendly and online web-service that obtains user query-specific biologically meaningful modules using integrated data enriched with deep learning-based predictions, and to display them to the user via easy-to-interpret, interactive, and heterogenous biomedical knowledge graphs (KGs), which will be constructed on-the-fly, in real-time. The CROssBAR project comprises multiple modules, each serving a specific purpose for the development of the CROssBAR system: *(i)* the construction of the CROssBAR database, and its API service to serve the integrated biomedical data, *(ii)* the development of deep-learning-based DTI prediction models for large-scale prediction of unknown DTIs, *(iii)* network-based organization and analysis of large-scale biomedical data using knowledge graph representations, *(iv) in vitro* wet-lab experiments to validate the relevance of *in silico* generated knowledge, and *(v)* the establishment of an open access web-service, with on-the-fly generation and visualization of query-based knowledge graphs.

In this chapter, we worked on the knowledge graph (KG) construction procedure of the CROssBAR system (module *iii*). The term KG defines a specialized data representation approach, in which a collection of entities is linked to each other in a semantic context. To determine the data retrieval steps and filters required to generate KGs, we first built a prototype network for hepatocellular carcinoma (HCC) disease with manual processing. Then, all operations applied for the construction of this prototype network were automatized so that a generic underlying query runs in the background when the CROssBAR database is searched, and a KG is generated from the resulting dataset using the CytoScape web-browser plug-in. We also performed extensive analyses to evaluate the diversity, stability, and practicality of CROssBAR

KGs. By applying the same methodology with manual curation to a certain extent, we constructed COVID-19 KGs as a use-case to better understand the molecular mechanisms of this new coronavirus (SARS-CoV-2) pandemic. Finally, we provided an example search from the CROssBAR web-service and roughly evaluated the output KG in a biological manner to exemplify the potential uses of the web-service. The findings of the CROssBAR project were published in Nucleic Acids Research journal (https://doi.org/10.1093/nar/gkab543), and the open-access CROssBAR web-service is available at https://crossbar.kansil.org. Through the web service (https://crossbar.kansil.org/covid_main.php), CROssBAR COVID-19 KGs can be interactively explored, visualized, and downloaded. They are also included in the European COVID-19 Data Portal (https://www.covid19dataportal.org/related-resources).

This chapter focuses exclusively on the aspects I contributed to the CROssBAR project. The comprehensive overview of the entire research and analysis conducted can be reached from our publication. My specific contributions to CROssBAR include; (i) Constructing the prototype Hepatocellular Carcinoma (HCC) network, (ii) Designing and developing a pipeline for automating the query-based KG construction process (iii) Implementing overrepresentation analysis for node filtering (iv) Generating CROssBAR COVID-19 KGs (v) Conducting an in-depth analysis of graph diversity and stability, and (vi) Performing graph construction runtime tests.

4.2. Introduction

The data explosion that originated in the -omics era of biological research necessitated the development of more systemic approaches for the analysis of biomedical data to develop novel and effective treatment approaches. However, different layers of the available data are produced using different technologies and maintained by different organizations, thus the data is scattered across individual computational resources, and the connections between them are not well-established although the entities in these resources are biologically related and complementary to each other. This connectivity problem hinders the effective usage and systemic analysis of multi-omics data for better understanding biological mechanisms. In addition to the connectivity problem, another issue is the incompleteness of the knowledge space (e.g., unknown interactions between ligands and target biomolecules, or missing associations among proteins).

There are some studies in the literature that integrate large-scale biological data and communicate it via textual or visual representations. One of the early applications of these data integration approaches is BioGraph data mining platform (Liekens et al., 2011), which allows for searching biomedical concepts to find relevant functional paths and identify disease gene prioritizations. In Bio4j project (Pareja-Tobes et al., 2015), a graph-based platform was constructed by integrating data from different public resources such as UniProt, Gene Ontology and Expasy to provide an infrastructure for querying and managing protein related information. In project Rephetio (Himmelstein et al., 2017), researchers built the Hetionet resource by combining biomedical data from various sources in a systematic way and storing it in a graph database with the main goal of deducing new drug/compound-disease relations. A similar strategy is employed in the BioGrakn project (Messina, Pribadi, et

al., 2018) to develop a biomedical knowledge graph (KG) utilizing the Grakn database infrastructure. Another system also called BioGraph (Messina, Fiannaca, et al., 2018) gathers gene/protein, function and cancer related miRNA data from several databases and enables users to query the data to generate basic network-based visualizations on returned entities.

While these studies provide useful tools and techniques for the life sciences research community, the majority of them demand complex database queries specific to the language of the related graph database, which can be challenging for researchers with little or no programming experience. Some of them require local installation, do not provide an easily interpretable visualization, or involve only a small portion of the biological resources. Such issues limit their comprehensiveness, functionality and/or practicality that prevent them from becoming widely used tools or services.

In the CROssBAR project, we aimed to address these shortcomings by developing a comprehensive integrated biomedical system enriched with *in silico* predictions and generating informative knowledge graphs based on particular biomedical entities such as genes/proteins, drugs/compounds, biological pathways, diseases/phenotypes, or specific combinations of them. It is a freely available, open-access, and user-friendly online biomedical data integration and representation tool with a coding-free interface designed to be easily used by the life sciences research community.

As a part of the CROssBAR project, we carried out the knowledge graph construction sub-module in this chapter. Knowledge graphs are a way of representing heterogeneous data that can be considered as multi-partite networks involving the relationships (edges) between different types of entities (nodes) in a semantic context. In CROssBAR knowledge graphs (CROssBAR-KGs), nodes represent biological components/terms, and edges represent known or predicted pairwise relationships between these terms. Nodes and edges are directly obtained from the CROssBAR database (CROssBAR-DB) during the construction of KGs. A number of data sources including UniProt, Ensembl, InterPro, IntAct, PubChem, ChEMBL, DrugBank, Reactome, KEGG, Orphanet, OMIM, Experimental Factor Ontology (EFO), Gene Ontology and Human Phenotype Ontology (HPO) are integrated into CROssBAR-DB to provide a broad spectrum of biological information. We first constructed a prototype hepatocellular carcinoma (HCC) disease network with manual processing to designate the data retrieval steps and filters necessary to produce KGs. All procedures used for building this prototype network were then automated; so that, a KG is generated and visualized on-the-fly based on the query term(s) of the user on the CROssBAR web-service while data retrieval and filtration operations are run simultaneously in the background. At each step of the process, an overrepresentation-based enrichment analysis is applied to select the terms that are significantly associated with the growing graph, and to discard the rest. With the aim of examining diversity, stability, and practicality of CROssBAR KGs, some qualitative and quantitative analyses as well as runtime tests were also performed. As a use case of the system, we constructed COVID-19 CROssBAR-KGs for a systemic assessment of the current knowledge about SARS-CoV-2 infection to better understand its molecular mechanisms and to aid the research community for the development of effective treatment strategies. Finally, we provided an example search on the CROssBAR web-service and basically analyzed the output KG in terms of the relation between query drug (trifluoperazine)

and disease (gastric cancer) terms to illustrate possible applications of the CROssBAR system.

The CROssBAR system, which assembles relevant pieces of biological data and allows its comprehensive analysis at the systemic level, can assist experimental and computational work in biomedical research with the ultimate goal of providing novel treatment solutions.

4.3. Materials and Methods

4.3.1. Construction of the Prototype Hepatocellular Carcinoma (HCC) Network

The prototype HCC network was created in 8 main steps:

1) *The selection of HCC related genes/proteins:* 61 HCC-related genes were identified from 4 different biological databases (i.e., KEGG (Kanehisa et al., 2016), OMIM (*OMIM - Online Mendelian Inheritance in Man*, n.d.), OpenTargets (Carvalho-Silva et al., 2019), TCGA (*The Cancer Genome Atlas Program - National Cancer Institute*, n.d.)), some of which were common among the databases. These genes were uploaded into CytoScape network analysis and visualization software (Shannon et al., 2003) and connected to HCC disease node to generate a bi-partite network based on the disease-gene relationships as the initial step.
   - KEGG (Kanehisa et al., 2016) (H00048-Hepatocellular Carcinoma): 20 genes
   - OMIM (*OMIM - Online Mendelian Inheritance in Man*, n.d.) (Phenotype MIM 114550 - Hepatocellular Carcinoma + Hepatoblastoma): 9 genes
   - OpenTargets (Carvalho-Silva et al., 2019) (EFO_0000182 - Hepatocellular Carcinoma): 18 genes were selected with scores higher than 0.2 based on the "genetic associations" column filter.
   - TCGA_HCC (*The Cancer Genome Atlas Program - National Cancer Institute*, n.d.): 34 genes were selected based on expert knowledge.
2) *The involvement of protein-protein interactions:* The protein-protein interactions (PPIs) between HCC-related genes were retrieved from STRING application (Szklarczyk et al., 2015) on CytoScape. Only interactions with a confidence score of 0.95 and above were integrated into the network. Hence, 45 PPIs between 31 proteins were included.
3) *The determination of HCC related pathways and their gene associations:* Signaling pathways associated with HCC disease pathway (hsa05225) in the KEGG database were incorporated into the network as pathway-disease and pathway-gene associations. These signaling pathways were uploaded from KEGGParser application on CytoScape. Apart from these, other KEGG signaling pathways associated with HCC-related genes were also added to the

network using STRING enrichment application on CytoScape with FDR cutoff = 0.05 and with at least 5 enriched genes. Therefore, 66 interactions between 22 genes and 10 pathways were mapped to the network.

4) *The inclusion of other diseases associated with HCC related genes:* Associations between HCC-related genes and other diseases were also identified via STRING enrichment application for diseases on the KEGG database, and integrated into the network as disease-gene associations. From the enrichment results, disease terms with at least 10 enriched genes were included (i.e., 72 interactions between 27 genes and 5 diseases). EFO disease terms were retrieved from GWAS (Genome-Wide Association Studies) Catalog (Buniello et al., 2019). For each EFO term, enrichment score and p-value were calculated based on the ratios of EFO terms in HCC genes and in total GWAS gene set using the formula (1) and (2) in *Methods Section 4.4.4*. EFO terms belonging to "disease" root, and having enrichment score > 20 and p-value < 0.005 were considered. 35 interactions between 20 genes and 7 EFO disease terms were mapped to the network.

5) *The involvement of associations between pathways and diseases:* In addition to associations of genes with pathways and diseases, KEGG database includes disease-pathway associations, as well. Therefore, 26 interactions between 10 pathways and 5 diseases of the network were obtained from KEGG, and integrated into the network.

6) *The determination of associations between HCC related genes and HPO terms:* HPO terms were retrieved from Human Phenotype Ontology database (Köhler et al., 2019). For each HPO term, enrichment score and p-value were calculated based on ratios of HPO terms in HCC genes and in total HPO gene set using the same formula in step 4. Only HPO terms with enrichment score > 65 and p-value < 10^-5 were considered. The top 10 HPO terms, which have not a parent-child relationship with each other, were selected and associated with the corresponding genes. 120 interactions between 22 genes and 10 HPO terms were mapped in total.

7) *The selection of drugs interacting with HCC related proteins:* The drug-target interaction data involving approved and investigational drugs were extracted from DrugBank database (Law et al., 2014) and integrated into the network (i.e., 63 interactions between 21 network proteins and 57 drugs). This integration is essential for the repurposing studies of potential drugs for HCC disease.

8) *The determination of interactions between compounds and HCC related proteins:* Bioactive compound interactions of the proteins in the HCC network were retrieved from ExCAPE dataset (Sun et al., 2017), which includes experimentally measured bioactivity data in PubChem and ChEMBL. Compounds with pXC50 >= 5.0 were labeled as active and pXC50 < 5.0 as inactive. For each compound, the enrichment score was calculated based on the ratios of active & inactive datapoint numbers of compounds in HCC genes and in total ExCAPE gene set (same formula in step 4 and 6). Only compounds with enrichment score > 1 were considered. They were clustered based on Tanimoto

similarities with threshold=0.5, and top 5 compound nodes that are not in the same clusters were selected based on the enrichment scores (with p-value<0.05) for each protein. For predicted compound-protein interactions, DEEPScreen predictions were used, and the same procedure used for the selection of experimentally known interactions from the EXCAPE dataset was applied. Hence, 26 interactions between 11 proteins & 12 compounds and 25 interactions between 5 proteins & 23 compounds were mapped from ExCAPE dataset and DEEPScreen predictions, respectively.

4.3.2. Automating the Query-Based KG Construction Process of CROssBAR

CROssBAR KGs are generated on-the-fly, in real-time, as it is not feasible to pre-calculate them due to the astronomical amount of possible queries. It is accomplished by a series of backend operations that gather the necessary information for the user query term(s) from the CROssBAR database and display it as a KG representation via the CROssBAR web-service. Although there are slight differences and modifications, the KG construction process of CROssBAR is mainly automated using the same procedure applied for the prototype HCC network generation, as described in *Section 4.3.1*.

During the construction of a CROssBAR KG, first, the gene/protein entries that are directly connected to the query term (i.e., core proteins) are fetched (e.g., member genes/proteins of a queried signaling pathway) from "Proteins" collection of the CROssBAR database. Then, neighboring/interacting proteins are retrieved from "IntAct" collection of the database. Before integrating them into the graph, the system calculates enrichment scores for each interacting protein using the equation in *Section 4.3.4.*, and filters out based on the selected cut-off value. The process is followed by the enrichment-based filtering and addition of terms from other biological component types (i.e., diseases, phenotypes, drugs, compounds, and additional biological processes/pathways related to these proteins) along with their relationships; however, this time, both core and neighboring proteins are taken into consideration to retrieve associated terms and to calculate the enrichment scores. The inclusion of pathways in the network is significant since many diseases don't act at a single gene level but on a systemic level. Pathway information is expected to capture these high-resolution relations successfully. In the step of compound addition, structural property-based filtering is also incorporated in the enrichment analysis to select compounds that are as diverse from each other as possible in terms of molecular structures. To achieve this; *(i)* The pairwise molecular similarities between all bioactive compounds in CROssBAR-DB were calculated from circular fingerprints (ECFP4) of compounds using the Tanimoto coefficient. *(ii)* These compounds were clustered based on a predefined similarity cut-off value of 0.5, meaning that each cluster is composed of compounds that are at least 50% similar to each other. *(iii)* The cluster information is pre-calculated and recorded on CROssBAR server. Each time a knowledge graph is being constructed, enrichment score ranked compounds are checked one by one in terms of their cluster membership and if there already is a compound from the same cluster in the graph, the compound in turn is discarded (i.e., not incorporated to the graph). The same clustering-based selection approach is applied for computationally predicted compounds interacting with the proteins in the graph, which are obtained from our in-house developed tool DEEPScreen. Following the finalization of the

compounds, overlaps between drug and compound nodes are checked using the identifier mapping between ChEMBL and DrugBank databases and merged. The node type of the merged node is set as a drug, since drugs are considered more reliable in terms of evidence on their molecular properties and interactions, compared to drug candidate compounds. Evidence-based coloring of bioactivity edges is applied here as well as in the prototype HCC network, where the edge is colored green if its source is DrugBank, blue for ChEMBL and red for DEEPScreen. The graph construction process is finished with the incorporation of edges for additional relation types between non-protein nodes (e.g., drug-disease, disease-pathway, and disease-HPO associations), to further enrich the provided information. If the user performs a heterogenous search that contains multiple terms from different component types, both core and neighboring genes/proteins are independently collected for each non-protein query term, queried gene/protein entries are added to this list (if there are any), and the term collection process is continued using the union of these genes/proteins as the source. This approach enables the exploration of direct and indirect relationships between all query terms.

The pipeline used for the construction of the CROssBAR KGs is given in Figure 4.1. Here, the finalized filtered set of each biological component is shown with a shape surrounded by a black frame. The graph is built using these entities together with their inter-component relations.

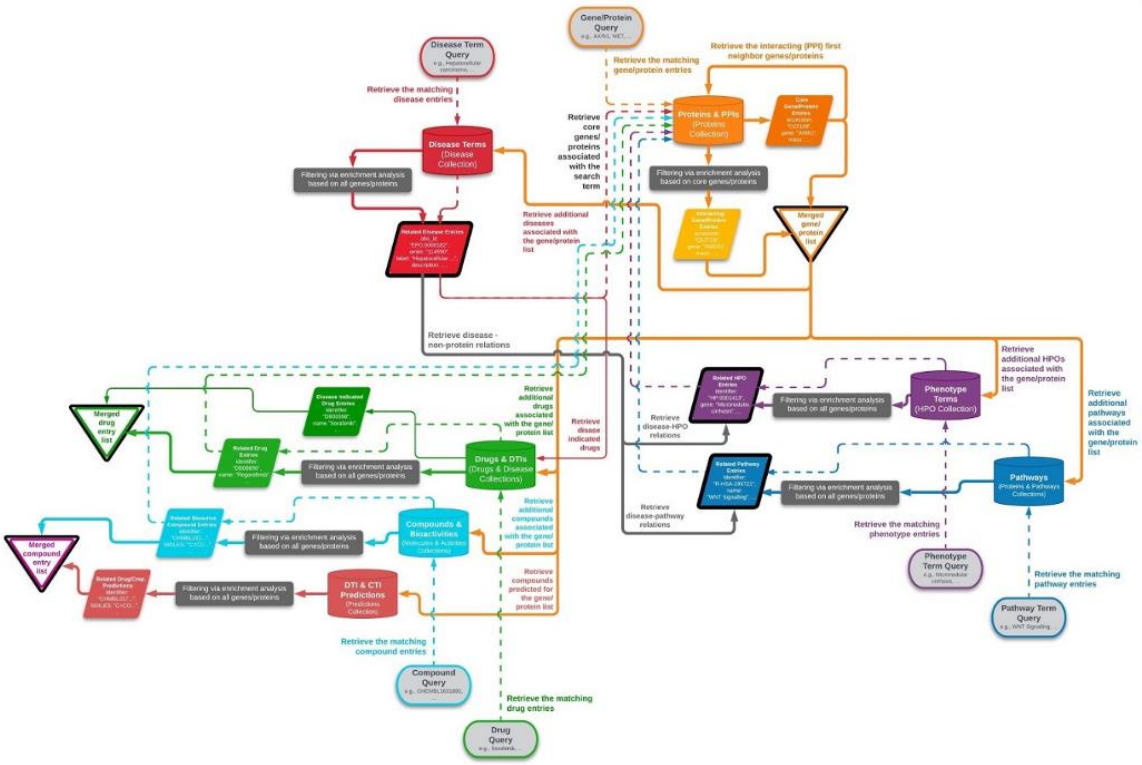


Figure 4.1. The workflow of the CROssBAR knowledge graph construction process (Doğan, Atas, et al., 2021)

4.3.3. Generation of CROssBAR COVID-19 KGs

As a use-case, we constructed the CROssBAR COVID-19 KGs with 2 different versions, *(i)* a large-scale version for comprehensive analysis or a detailed inspection, and *(ii)* a simplified version for fast interpretation. COVID-19 related data could not be pulled to the CROssBAR database during the construction of COVID-19 KGs since the majority of the data has not been integrated into the regular releases of biological databases. Hence, we had to make manual interventions to obtain the data from CROssBAR data resources for the generation of COVID-19 KGs.

4.3.3.1. Large-Scale COVID-19 KG

Construction of the large-scale COVID-19 graph started with acquiring the related EFO disease term named: "COVID-19" (id: MONDO:0100096). The disease term for "Severe acute respiratory syndrome" (id: EFO:0000694) (the original SARS) was also incorporated into the graph since SARS is better annotated compared to COVID-19. The construction process is continued as follows:

1) <u>COVID-19 related genes/proteins and PPIs:</u> COVID-19 related genes/proteins and their interactions were retrieved from the IntAct database's COVID-19 dataset (downloaded in March 2021). Unlike a genetic disease, human genes/proteins represent only a portion of infectious diseases due to host-pathogen molecular interactions. Therefore, we aimed to incorporate SARS-CoV and SARS-CoV-2 genes/proteins besides the host genes/proteins into the graph. Without any filtering, the dataset contained 2,951 gene/protein and metabolite nodes from various organisms and 7,706 edges. Due to high number of genes/proteins in the dataset, there was a risk of incorporating non-specific/irrelevant terms from the other biological components at later steps. To address this risk, several filtering operations were applied on this dataset. First, all non-gene/protein nodes were eliminated and the genes/proteins if the corresponding organism is not human or SARS-CoV/SARS-CoV-2 were discarded. Second, the protein entries that are not reviewed (i.e., not from UniProtKB/Swiss-Prot) except SARS-CoV-2 ORF10 (accession: A0A663DJA2), which currently is an unreviewed protein entry in UniProtKB/TrEMBL, were removed. A portion of the host genes/proteins were also filtered out using interaction-based information, according to their confidence scores reported in IntAct. The edges between host proteins and SARS-CoV and/or SARS-CoV-2 proteins were discarded if the confidence score was less than 0.35. The edges between host proteins in the KG (i.e., neighbouring proteins) were also discarded if their interaction confidence score is less than 0.6. The disconnected components made up of host proteins that were formed due to the edge filtering operation were removed, as well. Orthology relations between SARS-CoV and SARS-CoV-2 genes/proteins were annotated with "is ortholog of" edge type. The interactions of the subunits of large protein complexes such as the NSPs of replicase polyprotein 1ab of SARS-CoV/SARS-CoV-2 were mapped to their corresponding protein complex nodes and the subunit nodes were excluded from the graph. After these operations, the finalized number of genes/proteins is 778 (746 host genes/proteins, and 15 SARS-CoV and 17 SARS-CoV-2 genes/proteins) and the number of edges (i.e., PPIs including both virus-human and human-human associations) is 1,674. After this point, we started collecting new nodes and edges from various biological components based on the overrepresentation analysis and curation.

2) COVID-19 related drugs and compounds: The approved/investigational drug interactions of COVID-19 related proteins were retrieved from DrugBank database, v5.1.6 release. To incorporate only the most relevant drug-target interactions, an overrepresentation analysis was applied with respect to the associations of drugs with target genes/proteins in the KG using the hypergeometric distribution, as described in *Section 4.3.4*. The selected drugs were mapped to their corresponding protein targets in the graph via the edge label of green color, as this represents the highest level of confidence in terms of receptor-ligand interactions. DrugBank also has a COVID-19 specific drug list, which includes a curated list of drugs currently under research for COVID-19 treatment. These drugs were included in the KG as well. Drugs without any known targets (or the targets are known but not presented in the KG), were connected directly to the COVID-19 disease node. Drug repurposing based curated and experimental results from critical SARS-CoV-2 related publications such as Gordon *et al.*[21] were also incorporated with suitable edge labels depending on the data source. Finally, drug-disease relationships based on the reported drug indications on KEGG resource were added. The KG contains well-studied drugs for COVID-19 treatment such as Remdesivir (DB14761), Favipiravir (DB12466), Dexamethasone (DB01234) etc., as well as rather under-studied or non-studied ones (in the context of COVID-19) such as Isosorbide (DB09401) and Rocaglamide (DB15495).

For the retrieval of compound-target interactions based on experimentally measured bioactivities, ChEMBL database (v27) was utilized. We retrieved the ChEMBL bioactivity data points in binding assays, where the targets are human, SARS-CoV and SARS-CoV-2 proteins, and the pChEMBL value is greater than or equal to 5. Overrepresentation analysis was applied to select the most relevant ones. Here, only drugs/compounds with enrichment scores greater than 1 and p-value less than 0.05 were considered. Compounds were clustered based on Tanimoto coefficient based molecular similarities with a threshold of 0.5, and top 5 overrepresented compound nodes, which are in different clusters, were selected for each target protein (if exist) and incorporated into the KG. We also incorporated selected compound-host target protein and compound-SARS-CoV-2 organism interactions from SARS-CoV-2 curated dataset of ChEMBL, including both binding and functional assays. Finally, the edge labels were set accordingly (i.e., blue colored edges).

For computationally predicted drug and compound-target protein interactions, our in-house deep learning based tools DEEPScreen (Rifaioglu et al., 2020) and MDeePred (Rifaioglu et al., 2021) were used. DEEPScreen large-scale prediction run results were scanned and 326 bioactive drug/compound-target interaction predictions for 18 human proteins were incorporated to the KG following the application of overrepresentation analysis, as same with selection of experimental bioactivities from ChEMBL. For *in silico* drug repurposing, both human ACE2 receptor protein and SARS-CoV-2 3C-like proteinase models of MDeePred were used to scan full DrugBank drugs dataset to predict new binders for ACE2 and 3C-like proteinase. In order to avoid the crowding of the KG, only five selected inhibitors for each protein were incorporated. The selected bioactive drug predictions for ACE2 are Eribaxaban (DB06920), 7-Hydroxystaurosporine (DB01933), Becatecarin (DB06362), Ticagrelor (DB08816) and Amcinonide (DB00288); whereas the predictions for the 3C-like protease are Quinfamide (DB12780), Diloxanide furoate (DB14638), Phenyl aminosalicylate

(DB06807), Netarsudil (DB13931) and Amlodipine (DB00381). These predicted interactions were labelled with red colored edges.

We also merged nodes with respect to drug-compound entry correspondences in DrugBank and ChEMBL databases. This way, some of the drug nodes also contain experimental bioassay-based relations (i.e., blue colored edges) and computationally predicted relations (i.e., red colored edges). At the end of these procedures, the total number of drugs (nodes) in the KG is 158 and the total number of drug interactions (edges) is 346. The total number of drug candidate small-molecule compounds is 167 and the total number of compound interactions (edges) is 664. Out of all drug/compound-target interaction edges, 120 correspond to drug development procedures, 382 to experimental bioassays and 508 to deep-learning-based predictions.

3) Pathways of COVID-19 related host genes/proteins: Signaling and metabolic pathway information was taken from Reactome (via CROssBAR database) and KEGG pathways data sources. The most relevant pathways were determined by overrepresentation analysis and mapped to the related genes/proteins in the KG. Some of the incorporated pathways are directly related to SARS-CoV-2 infection such as "Viral mRNA Translation" (R-HSA-192823) or "ISG15 antiviral mechanism" (R-HSA-1169408) and the others are innate pathways of the host (human) such as "Endocytosis" (hsa04144), "Cell cycle" (hsa04110) or "NF-kappa B signaling pathway" (hsa04064). We also incorporated pathway-disease relations (in the sense of pathways that are modulated due to presence of certain diseases) from KEGG database. The finalized number of pathways in the KG is 100 (32 for KEGG and 68 for Reactome, among which there are corresponding terms) and the total number of gene/protein-pathway associations (edges) is 1333 (557 for KEGG and 776 for Reactome).

4) COVID-19 related phenotypic implications: The resource for the phenotype terms is the Human Phenotype Ontology (HPO) database. For each phenotype term that is associated with at least one gene in the KG according to HPO data, enrichment score and *p*-value were calculated via overrepresentation analysis. Phenotype terms that are not in a close parent-child relationship with each other in the HPO direct acyclic graph were selected from the score-ranked HPO term list. HPO also has a curated list of SARS related phenotype terms. These terms were also added into the KG and mapped to "COVID-19" and "Severe acute respiratory syndrome" disease nodes. This way, COVID-19 related phenotypes including symptoms such as Fever (HP:0001945), Myalgia (HP:0003326), Respiratory distress (HP:0002098), Immunodeficiency (HP:0002721) and etc. are included in the graph. The finalized number of phenotype terms in the KG (nodes) is 43 and the number of HPO term-gene/protein associations (edges) is 2427.

5) Other associated diseases of COVID-19 related host genes/proteins: The aim behind this step is collecting the non-infectious (mostly genetic) diseases that utilize the same (or similar) biological mechanisms/processes of human, so that it may indicate potential risks for COVID-19 patients, or potential COVID-19 related repurposing options for drugs that are currently used to treat these diseases. For this, disease terms that are associated with genes/proteins in the COVID-19 KG were collected from the CROssBAR database resources: EFO disease collection (mainly

including OMIM and Orphanet disease entries) and KEGG diseases database. The linkage of proteins and EFO terms was achieved through OMIM ids. The most relevant disease terms were selected based on the results of the overrepresentation analysis. Finally, disease-HPO term relations were also integrated into the KG using the disease association information provided in HPO resource. At the end of this step, diseases such as Small cell lung cancer (H00013), Amyotrophic lateral sclerosis - ALS (H00058), Bruck syndrome (Orphanet:2771), Osteosarcoma (EFO:0000637) etc. have entered the KG. The finalized number of disease terms in the KG is 41 (19 for KEGG and 22 for EFO) and the number of disease-gene/protein associations (edges) is 120 (67 for KEGG and 53 for EFO). There are also 56 HPO term-disease associations including HPO associations of "COVID-19" and "Severe acute respiratory syndrome" disease nodes -integrated in step 4- and other disease terms.

The finalized large-scale COVID-19 KG includes 1,289 nodes (i.e., genes/proteins, drugs/compounds, pathways, diseases/phenotypes) and 6,743 edges (i.e., various types of relations).

4.3.3.2. Simplified COVID-19 KG

For the construction of the simplified COVID-19 KG, the starting point was the COVID-19 associated proteins in the UniProt COVID-19 portal (https://covid-19.uniprot.org/), instead of the IntAct Coronavirus dataset. The remaining steps of building the graph were mainly similar with the large-scale COVID-19 KG except that, additional nodes representing the organisms: human, SARS-CoV and SARS-CoV-2 were placed and connected to the corresponding proteins. The aim here was to prevent the presence of singleton protein nodes due to the reduced number of included gene/proteins and PPIs in the simplified graph. It is also important to note that the simplified version is not just a subset of the large-scale KG. Since the starting point of gene/protein collection were different between two KGs, the resulting graphs have slightly different contents as well. For example, the drugs Siltuximab (DB09036), Pirfenidone (DB04951) are specific to the simplified KG. The simplified COVID-19 KG includes a total of 435 nodes and 1,061 edges.

The Cytoscape network files and overrepresentation analysis results of the KGs are available at CROssBAR project GitHub repository (https://github.com/cansyl/CROssBAR) . The graphs can also be directly visualized and explored interactively via CROssBAR web-service (https://crossbar.kansil.org/covid_main.php).

4.3.4. Node Filtering via Overrepresentation Analysis

Because knowledge graphs are built by incorporating all biological terms that are directly or indirectly associated with query term(s), searches without further filtering would result in huge graphs with tens of thousands of nodes and edges. In this case, it would not be possible to visually perceive a biologically relevant result from the giant network. Moreover, it would not be feasible to construct and interactively display this graph due to the excessive computational demands. To address this problem, we applied a multi-staged overrepresentation-based enrichment analysis during the construction of graphs. In this analysis, an independent enrichment score and its

statistical significance is calculated for each biological entity in the database to be considered as its relevance to the graph that is being constructed. It is performed using a modified version of the hypergeometric test for overrepresentation, which is also known as the one-tailed Fisher's exact test, and is calculated based on the statistics of the associations with gene/protein nodes. For example, the enrichment score ($E_{D,W}$) and its significance ($S_{D,W}$), in terms of *p*-value, for a disease term *D*, in graph *W* is calculated as follows:

$$E_{D,W} = \frac{m_D^2/n_W}{M_D/N} \quad (1)$$

$$S_{D,W} = \sum_{i=m_D}^{n_W} \frac{\binom{M_D}{m_D}\binom{N-M_D}{n_W-m_D}}{\binom{N}{n_W}} \quad (2)$$

where $E_{D,W}$ is the enrichment score calculated for the disease term *D* in graph *W;* m_D^2 represents the square of the number of genes/proteins in graph *W* that are associated with disease *D;* n_W represents the total number of gene/protein nodes in graph *W*; M_D is the total number of genes/proteins (not necessarily in graph *W*) that is associated with disease *D*; and *N* represents the total number of reviewed human gene/protein entries (i.e., UniProtKB/Swiss-Prot entries) in the CROssBAR database that is annotated with any disease entry. $S_{D,W}$ represents the significance (*p*-value) for the disease term *D* of graph *W* calculated in the hypergeometric test. In the formula, m_D and n_W values represent foreground distribution and are calculated on-the-fly since they change depending on the graph. However, M_D and *N* reflect background distribution and are not specific to graphs. Therefore, they were precalculated to decrease computational time required for the construction of CROssBAR KGs.

Considering the enrichment analysis for diseases, while constructing the graph, an enrichment score is calculated for each disease entry in the CROssBAR database and these scores are used to rank disease entries according to their biological relevance to graph W (i.e., in the order of decreasing scores). A cut-off value *k* is employed to include the top *k* relevant disease entries to graph *W*. The default value for *k* is 10, which means that only top-10 relevant diseases will be included. Apart from diseases, the same methodology is used to filter out terms of neighboring genes/proteins, pathways, phenotypes, drugs, and compounds. In the traditional way of calculating an enrichment score, m_D is without square. The reason behind taking the square of m_D is mainly to highlight the term with higher m_D value (i.e., a higher degree) in a case of multiple terms with very similar enrichment scores. Here, significance values are not directly used in the filtering operation, since the main objective is not including only significantly over-represented terms, but just reducing the number of nodes in the graph by filtering out the ones that are least relevant.

4.4. Results and Discussion

4.4.1. Prototype HCC Network

The finalized HCC network includes 185 nodes (i.e., genes/proteins, drugs/compounds, signaling pathways, KEGG (Kanehisa et al., 2016) and EFO diseases, HPO terms) and 478 edges (i.e., interactions). To make the network visually interpretable, each node type was represented with a different color and shape as shown in Figure 4.2. It also has a hierarchical and circular structure, in which gene/protein nodes are located on the innermost circle and drug/compound nodes are on the outermost. Moreover, different colors were assigned to the bioactivity edges to represent their confidence levels; where drug-protein interaction edges are colored green as the highest confidence level while experimentally known and predicted interactions are represented in blue and red, respectively. In addition to the data retrieval and filtering steps implemented for the prototype network, its visualization is also utilized for the generation of automated CROssBAR KGs. The cytoscape file of the prototype HCC network is available in its GitHub repository (https://github.com/cansyl/CROssBAR-Prototype_HCC_Network), which can be downloaded and examined.

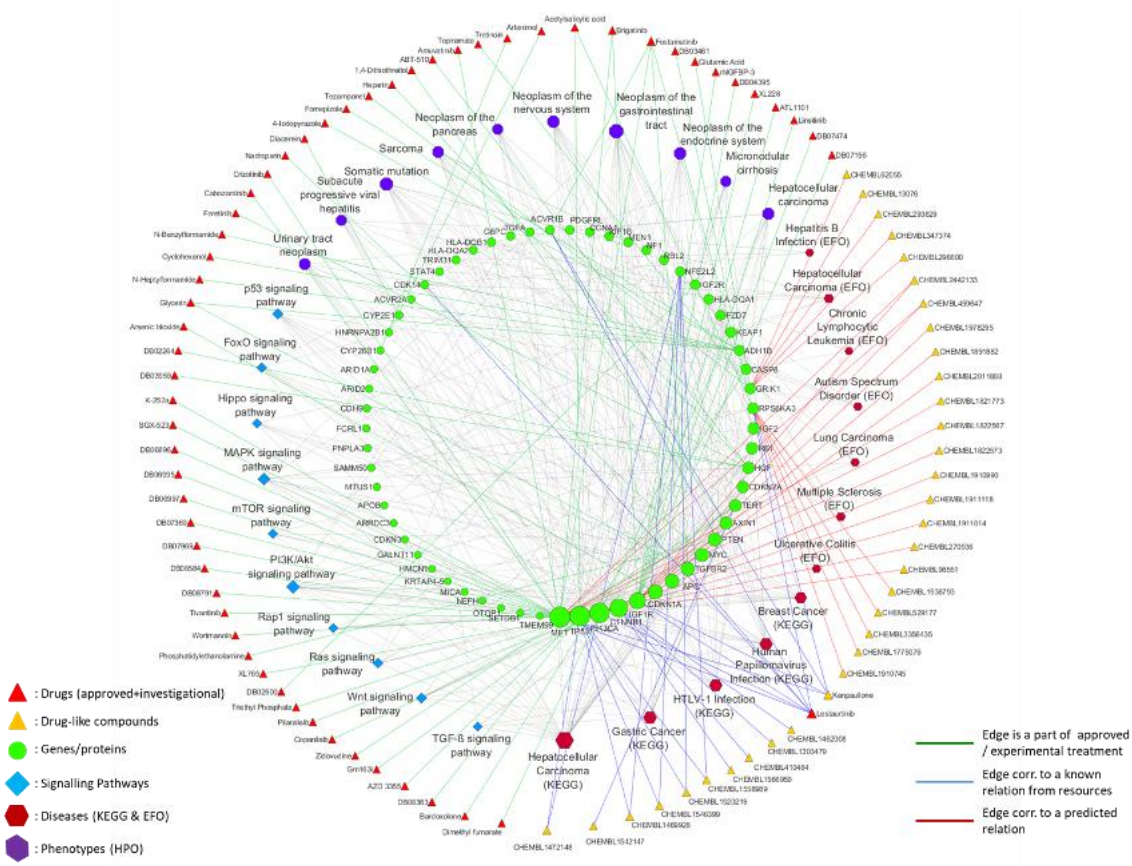


Figure 4.2. Hepatocellular Carcinoma network as a prototype for CROssBAR knowledge graphs

4.4.2. Use-Case Study on CROssBAR Web-Service (Query: TFP + Gastric Cancer)

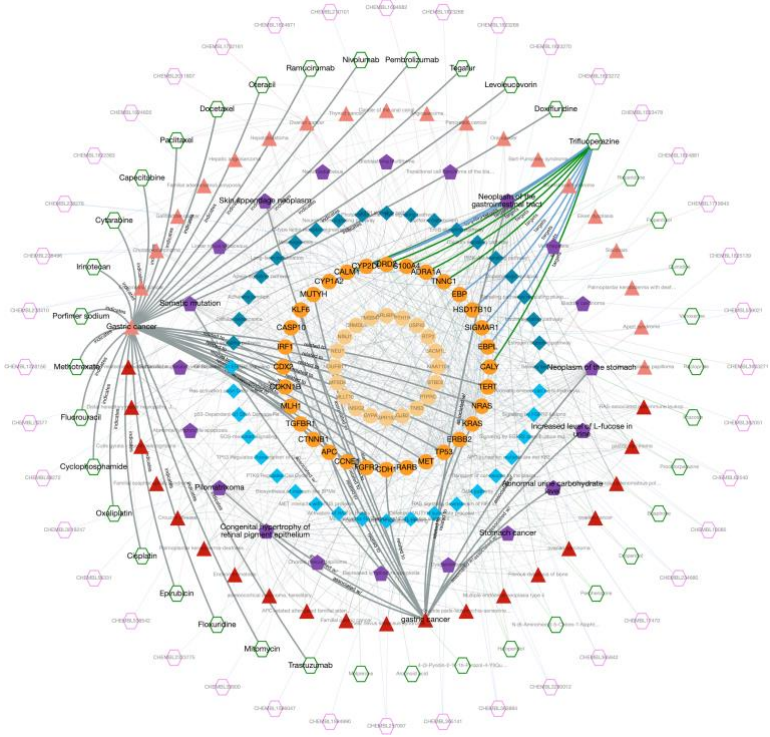
To provide an example about one of the many possible uses of the CROssBAR system, we explore the relation between a drug (trifluoperazine) and a disease (gastric cancer), to make a very quick and rough evaluation on the potential repurposing of this drug towards the disease of interest. Trifluoperazine (TFP) is an approved antipsychotic agent mainly used in the treatment of schizophrenia. There are also many studies showing the combinatorial effect of TFP in enhancing the efficacy of cancer drugs, which achieves this mainly by modulating drug efflux pumps such as P-glycoprotein (Jaszczyszyn et al., 2012). Moreover, it affects various signaling pathways involved in cancer progression so it can increase the apoptotic response induced by other cancer drugs or possess anti-angiogenic properties that can be helpful in preventing metastasis (Feng et al., 2018). As far as we are aware, TFP has no *in vitro*, *in vivo* or clinical studies concerning the treatment of gastric cancer, although there are studies on other types of cancer such as colorectal, pancreatic, and lung, in the literature. Also, there is a study indicating the inverse association between antipsychotic use and the risk of gastric cancer. Thus, this may be a convenient scenario for investigating the relationship between two potentially related biomedical entities, gastric cancer and trifluoperazine. To construct the corresponding knowledge graph, we queried the CROssBAR-WS (https://crossbar.kansil.org) with this drug and disease entries and selected the number of nodes to be incorporated into the graph (from each biomedical component) as 20. The resulting graph is shown in Figure 4.3.

TFP exerts its antipsychotic effect with the blockage of the dopamine D2 receptor. This relation is shown in the graph, where TFP binds to the DRD2 gene/protein node and is associated with the dopaminergic synapse pathway. In the KG, TFP also has other approved targets such as CALM1, ADRA1A, and TNNC1 proteins (approved drug-target interaction edges have green color), and these proteins are associated with the calcium signaling pathway. Moreover, DRD2 and CALM1 are associated with the rap1 signaling pathway, as well. Both calcium and rap1 signaling pathways have other gene/protein associations such as ERBB2, KRAS, and CDH1, which are further associated with gastric cancer disease. In light of these relations, TFP can be explored via additional *in silico* and wet-lab studies, in terms of its potential to become a repurposed agent for the treatment of gastric cancer, which may show its activity on gastric cancer cells via calcium and rap1 signaling pathways (Figure 4.3).

Some of the proteins that are associated with gastric cancer (e.g., KRAS, ERBB2, TP53, etc.) are also related to other cancer disease nodes in the graph such as pancreatic cancer, ovarian cancer, endometrial cancer, and cholangiocarcinoma, which means that TFP may also have a potential against these cancer types, worthy of further exploration. Other antipsychotic or anxiolytic agents such as risperidone, haloperidol, perphenazine, buspirone, droperidol, and prochlorperazine are enriched in the network as well, which bind to DRD2, CALM1 and/or ADRA1A. These drugs may also become alternative repurposed drugs for gastric cancer treatment or other cancers presented in the KG. In addition to the above-mentioned approved drug-target interactions, the graph also includes enriched drugs and drug-like compounds having experimentally measured bioactivities -from ChEMBL- (shown with blue colored edges) or computationally predicted interactions -by our in-house tool DEEPScreen- (shown with red colored edges) against the targets DRD2, ADRA1A, EBP, and

SIGMAR1; which can also be considered for the diseases in the graph. Finally, there are several phenotypic implication terms (from HPO) on the KG, such as the abnormal urine carbohydrate level and the congenital hypertrophy of retinal pigment epithelium, which are associated with gastric cancer disease node and/or gastric cancer-related genes. These phenotypic implications could also be helpful in considering clinical studies.

(a)



(b)

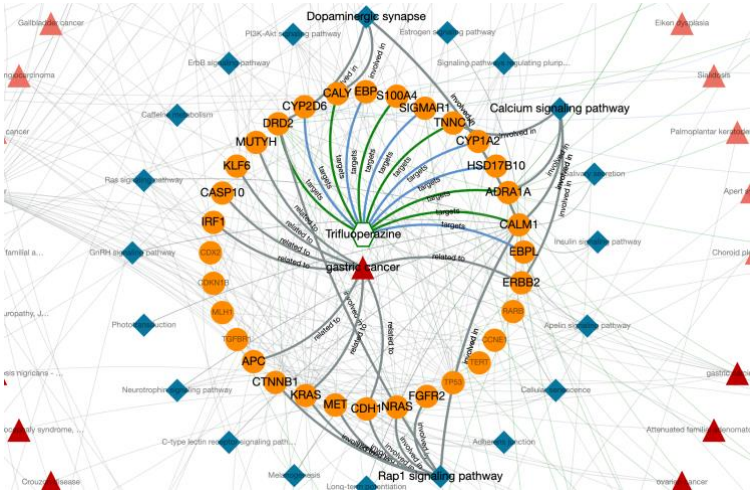


Figure 4.3. (a) the output knowledge graph of trifluoperazine and gastric cancer query **(b)** critical signalling pathways and their relation to trifluoperazine and gastric cancer over critical genes/proteins

4.4.3. Literature-Based Validation of COVID-19 KGs

Starting from the end of 2019, COVID-19 pandemic has ravaged the entire globe and caused immeasurable damage. As of March 2021, the scientific endeavor to develop effective drugs and vaccines is at peak, and a systemic evaluation of the current knowledge about SARS-CoV-2 infection is expected aid researchers in this struggle. To demonstrate the capabilities of CROssBAR, we have constructed two different versions of the COVID-19 knowledge graph, *(i)* a large-scale version including nearly the entirety of the related information on different CROssBAR-integrated data sources, which is ideal for further network and machine learning based analysis or a detailed inspection (Figure 4.4), and *(ii)* a simplified version distilled to include only the most relevant genes/proteins as provided in UniProt-COVID-19 portal (https://covid-19.uniprot.org), which is ideal for fast interpretation (Figure 4.5).

CROssBAR COVID-19 KGs incorporate several drugs that can be utilized for developing novel treatments against SARS-CoV-2. Several of these drugs have already been reported in the COVID-19 literature and included based on this information; however, some of them were completely new. These new drugs have been incorporated to the graph either due to the overrepresentation analysis (based on the COVID-19 related host genes/proteins in the graph) or predicted to interact to with host or SARS-CoV-2 proteins by our deep-learning-based tools DEEPScreen and MDeePred. Here, we demonstrate a short literature-based validation study on the relevance of these new drugs for COVID-19. Table S.6. shows the promising drugs in our knowledge graph together with the source (i.e., whether they entered the graph due to enrichment or predicted by our deep-learning-based systems). It is interesting to observe that some of the drugs in this list are currently under clinical trials against COVID-19. The list includes calcineurin, IL-6 and IL-17a inhibitors such as cyclosporine, tocilizumab, and ixekizumab, which play roles in the immune system and effective against inflammatory diseases. As an immunomodulator agent, interferon beta-1a is also included in the list, inducing the synthesis of antiviral mediators by binding to type I interferon receptors. In addition to these, there are also other type of drugs in the list such as tenectaplase, vazegepant and simvastatin, which have promising clinical study results especially for the prevention of severe pulmonary damages and respiratory failures due to COVID-19. Ascorbic acid (i.e., vitamin c) and epigallocatechin gallate (i.e., phenolic antioxidant) are two examples of natural products that have COVID-19 related clinical studies. Apart from these, other enriched/predicted drugs such as amlodipine, artenimol, lifitegrast, amcinonide and becatecarin have been shown as potential drugs for COVID-19 via *in vitro*, *in vivo* and/or *in silico* studies including machine learning and molecular docking applications, although some of these studies are yet to be peer-reviewed. As a potent inhibitor of NF-κB activation in T-cells, rocaglamide and its derivatives may also be potential drug candidates for the treatment of COVID-19; however, there is no COVID-19 related study about these drugs in the literature yet, except from a study reviewing antiviral activity potential of rocaglamide as a flavagline. It is also important to mention that further research is required to properly assess the potential of these drugs for repurposing against SARS-CoV-2 infection.

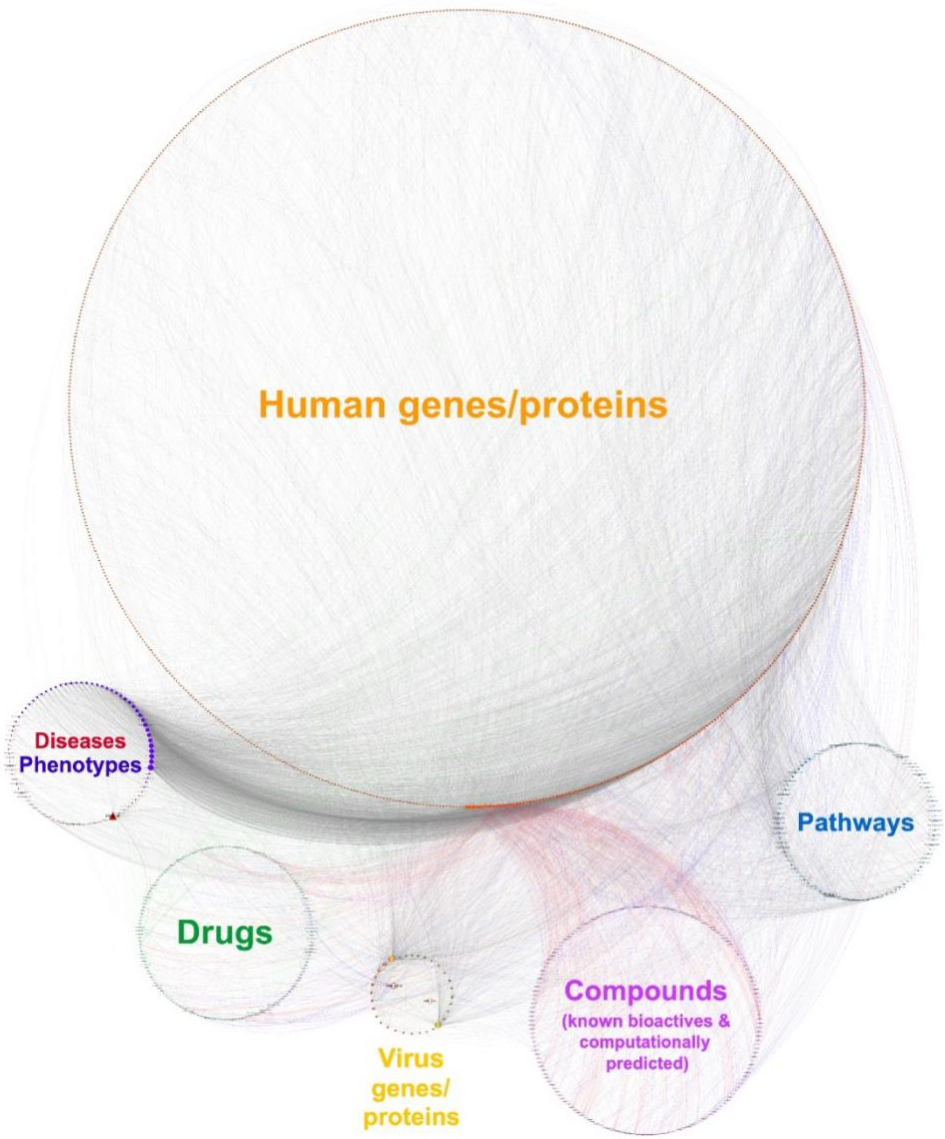


Figure 4.4. Large-scale version of COVID-19 knowledge graph

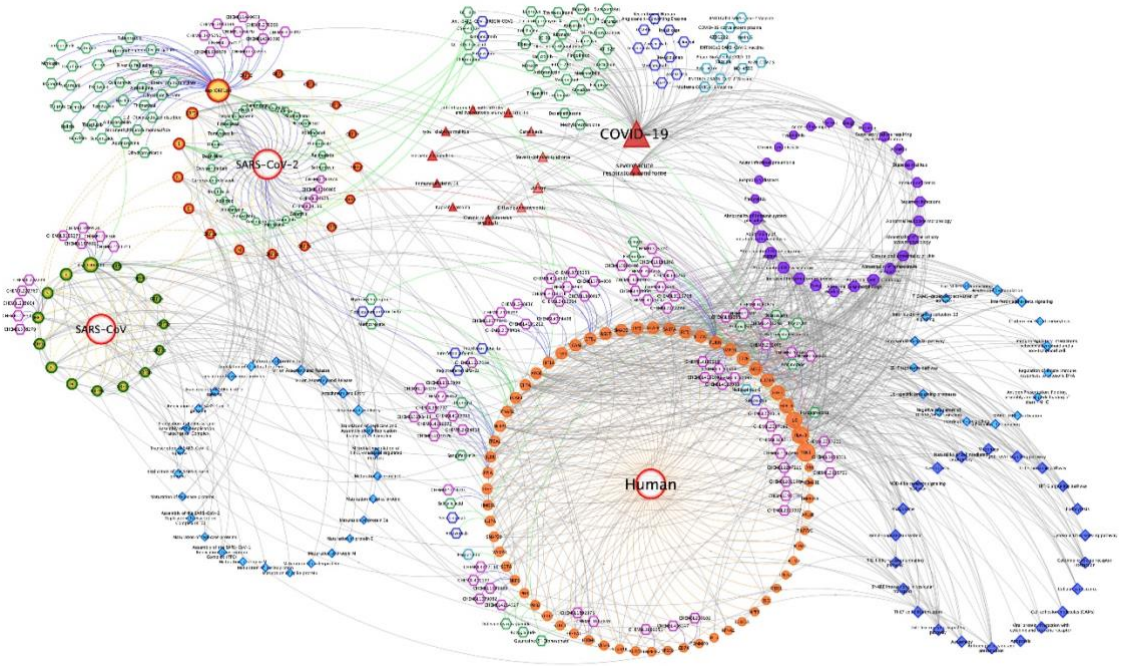


Figure 4.5. Simplified version of COVID-19 knowledge graph

Table 4.1. Literature based information for new potential COVID-19 based repurposing of CROssBAR COVID-19 knowledge graph drugs.

Drug Name	DrugBank ID	Description	Source	Clinical Trial ID	Current State *
Cyclosporine	DB00091	calcineurin inhibitor	DrugBank & Chembl	NCT04392531	Phase 4
Tocilizumab	DB06273	IL-6 inhibitor	DrugBank	NCT04377750	Phase 4
Amlodipine	DB00381	calcium channel blocker	MDeePred	NCT04330300	Phase 4
Siltuximab	DB09036	IL-6 inhibitor	DrugBank	NCT04330638	Phase 3
Prednisolone	DB00860	glucocorticoid steroid	Chembl	NCT04381936	Phase 2-3
Vazegepant	DB15688	calcitonin gene-related peptide (CGRP) receptor antagonist	DeepScreen	NCT04346615	Phase 2-3
Quercetin	DB04216	polyphenolic flavonoid	DrugBank	NCT04377789	Not Applicable
Artenimol	DB11638	artemisinin derivative and antimalarial agent	DrugBank	-	in-silico study
Lifitegrast	DB11611	integrin antagonist	DrugBank	-	in-silico study
Amcinonide	DB00288	corticosteroid	MDeePred	-	in-silico study
Becatecarin	DB06362	diethylaminoethyl analogue of rebeccamycin	MDeePred	-	in-silico study
Quinfamide	DB12780	antiprotozoal agent	MDeePred	-	in-silico study
Rocaglamide	DB15495	eIF4A inhibitor	DrugBank	-	-
Didesmethyl rocaglamide	DB15496	eIF4A inhibitor	DrugBank	-	-

* Some of these drugs have multiple clinical trials concerning COVID-19. In these cases, the one with the latest phase is given.

4.4.4. Analysis of Knowledge Graph Diversity and Stability

Highly studied biomedical entities (e.g., TP53 gene, JAK-STAT signaling pathway, etc.) typically have a high number of recorded relationships in databases. As a result, they frequently appear in biological networks or gene set enrichment analyses. In CROssBAR, the goal is to build knowledge graphs with specialized content for the relevant query term(s), thus, we expect to observe diversity in our KGs. Additionally, we aim to produce stable outputs, which means that searches for terms that are

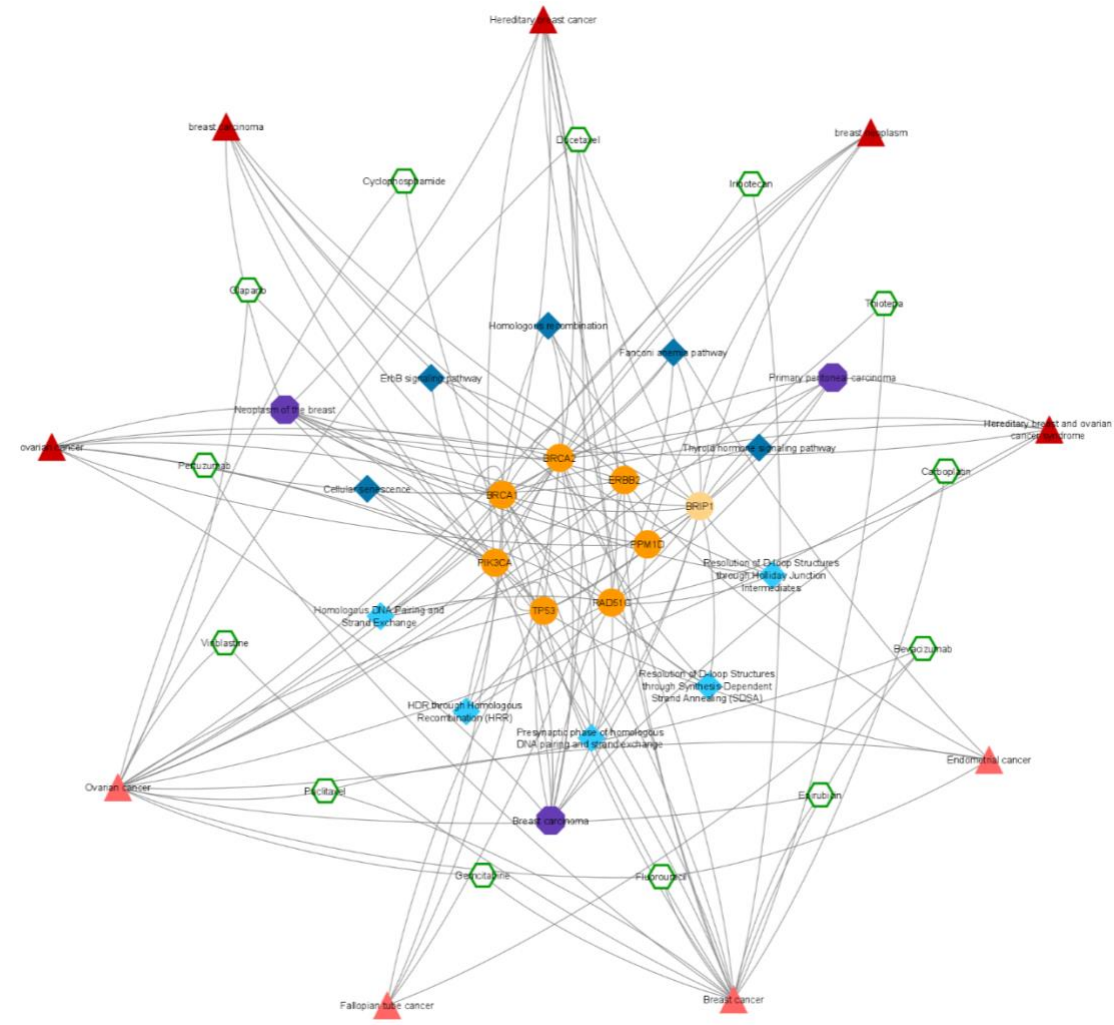
biologically related should produce KGs with similar content in terms of incorporated nodes and edges. To investigate both diversity and stability of CROssBAR KGs, we conducted two experiments; *(i)* a use case analysis and *(ii)* a quantitative analysis.

In the use case analysis, we independently queried three different diseases (types of cancer), the first two of which are similar to each other in terms of the affected biological mechanisms, and the third one is relatively dissimilar from them in the same sense. The selected disease terms are breast cancer, ovarian cancer, and osteosarcoma, respectively. The reason behind selecting another type of cancer as the third disease (instead of, for example, a rare disease, which would be highly unrelated to the first 2 diseases) was to create a rather realistic use case scenario that would allow us to observe the issues related to graph diversity, if there are any. Breast cancer and ovarian cancer are both associated with mutations and/or overexpression/amplification in certain genes (e.g., BRCA1, BRCA2, PIK3C, ERBB2, etc.) and aberrations in related pathways, which exhibits a risk of co-occurrence in women. On the other hand, osteosarcoma, the most common type of primary bone cancer, does not have a known direct relationship with breast or ovarian cancers. Besides, primary osteosarcomas of the breast and ovary are reported as very rare malignancies. Therefore, we expected to observe shared mechanisms/terms between KGs of breast and ovarian cancers, whereas the KG of osteosarcoma was expected to be relatively more diverse.

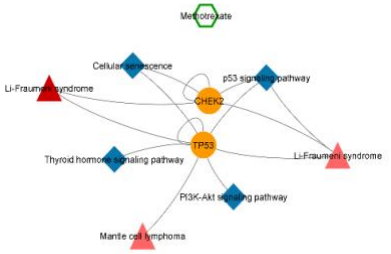
We queried CROssBAR with these disease terms using default parameters (i.e., the number of nodes to be included in KGs for each biological/biomedical component is 10, organism: human, only include reviewed protein entries from the UniProtKB/Swiss-Prot database) to construct the KGs. The resulting graphs are composed of 162 nodes and 563 edges for breast cancer, 123 nodes and 397 edges for ovarian cancer, and 98 nodes and 208 edges for osteosarcoma, and displayed in Figure 4.6. After that, we calculated pairwise and triple-wise intersections between the contents of these three KGs. Graphs that are composed of intersecting nodes and edges are given together with Venn diagram-based statistics in Figure 4.6. We observed that the content-based identity (i.e., presence of the same nodes and edges) between KGs of breast and ovarian cancers is around 30%, whereas the overall identity between breast and osteosarcoma, and between ovarian and osteosarcoma are both around 6%. It is also important to note that both breast and ovarian cancer graphs contain the other disease as a similar disease node. It is observed from Figure 4.6 and b that both BRCA1 and BRCA2 genes are presented in breast-ovarian intersection, in addition to well-known cancer driver genes such as TP53, PIK3CA and ERBB2. Breast cancer and ovarian cancer searches also contain other common associations such as pathways, phenoypes, drugs and other diseases (e.g., ErbB signaling pathway, primary peritoneal carcinoma, paclitaxel, fallopian tube cancer, etc.). Their differences are based on known and predicted bioactive compounds, due to the fact that these are selected from large pools of compounds that have direct relationship to the genes/proteins in the corresponding graph. When we omit ligands and only focus on the biological mechanism related components, the graph identity between breast and ovarian cancers is around 35%. On the other hand, breast/ovarian and osteosarcoma intersection only included 3 nodes and 3 edges that involve the TP53 gene (Figure 4.6), which was expected since TP53 mutations are critical in almost all types of cancer. The results of

use case analysis indicate that CROssBAR constructs diverse and stable graphs specific to the user query term(s).

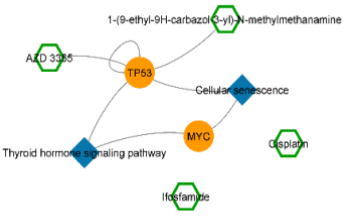
(a)



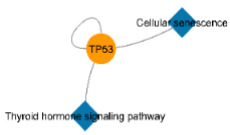
(b)



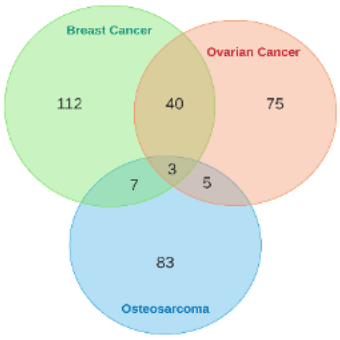
(c)



(d)



(e)



(f)

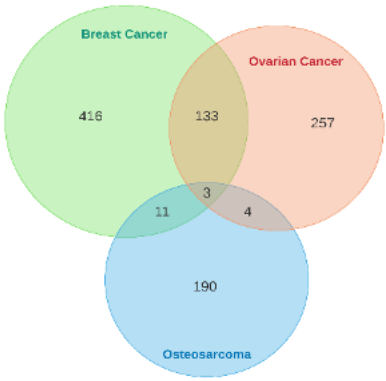


Figure 4.6. CROssBAR knowledge graph diversity analysis use case, intersection graphs between: **(a)** breast cancer and ovarian cancer, **(b)** breast cancer and osteosarcoma, **(c)** ovarian cancer and osteosarcoma, **(d)** breast cancer, ovarian cancer, and osteosarcoma (triple-wise) queries. Venn diagrams displaying the statistics of shared: **(e)** nodes, and **(f)** edges, between KGs of different query terms.

Since the first analysis is only a use case conducted on 3 sample disease queries, we further investigated the matter with a quantitative test on a larger dataset. In our second experiment, we aimed to evaluate whether highly studied, and thus highly connected biological/biomedical entities tend to be presented in our graphs with high frequencies. This would be undesirable as it would mean certain terms usually end up in the graphs no matter what is searched for (i.e., the problem of limited diversity). To test this, we selected 20 terms from each biological/biomedical component (a total of 140 terms) that are among the most connected, by checking the number of their associations (degree) to different genes/proteins in our database. Then, we checked how many times these highly connected terms are presented in CROssBAR KGs. First, to construct these graphs, we queried randomly selected genes/proteins, Reactome and KEGG pathways, EFO and KEGG diseases, HPO terms, drugs and compounds one by one, and in combination with each other, on the CROssBAR web-service, resulting in a total of 1365 KGs. To evaluate whether selected highly connected terms are over-represented in CROssBAR KGs we applied Fisher's exact test independently for each term with the null hypothesis stating that the corresponding term is presented in KGs with an observed frequency (i.e., for a term D, observed frequency is given by; gD/G, where gD is the number of KGs in which term D is presented, and G is the total number of KGs in the analysis) same as its expected frequency based on its general connectivity (i.e., for a term D, expected frequency is given by; t*MD/Msum, where t is the number of terms/nodes in each KG from the same biological component as term D, MD is the total number of genes/proteins that are associated with term D, and Msum is the total number of associations between all genes/proteins and all terms in the same biological component as term D). In the case that the null hypothesis is true, we would conclude that the system is not successful in terms of eliminating promiscuous/hub terms, and they are frequently presented in KGs probably because they are connected to many other terms in the database. On the other hand, a statistically significant deviation from the null hypothesis with an observed frequency of presence in KGs lower than the expected frequency would indicate that these hub terms are not

presented in KGs as it would be expected based on their high connectedness, instead, they are successfully eliminated by our pipeline. We left compounds out of this analysis since their expected frequency values are extremely low due to their high number (e.g., 654051 compounds have at least 1 target association). The results of this analysis are displayed in Figure 4.7 as bar graphs drawn for each of the 140 terms, where observed and expected frequencies are shown within overlapping bars with different colors ("*" indicate that the corresponding observed frequency is significantly lower compared to the expected frequency). These results are also displayed in Table S7 together with contingency table values used in statistical testing and the resulting significance (p-values). It is observed from the results of this analysis that 117 out of 140 highly connected terms are significantly less represented in CROssBAR KGs compared to their expected frequencies. Among these 117 terms, 22 highly connected ones (e.g., 14 HPO terms, 2 Reactome pathways and 6 drugs) were not presented in any KGs at all.

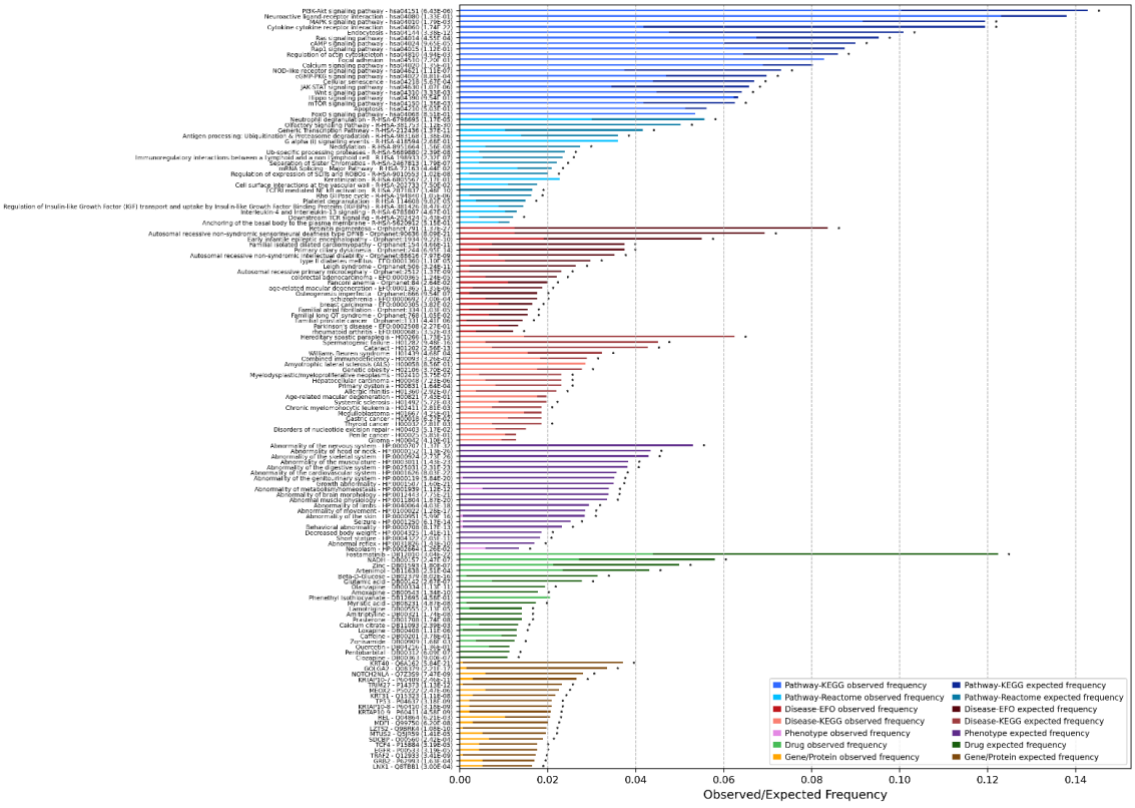


Figure 4.7. Bar graphs indicating observed and expected frequencies (overlapping bars with different shades of colors) for each of the 140 selected highly connected/hub terms in 1365 CROssBAR KGs constructed with random term queries. "*" indicate that the corresponding observed frequency is statistically significantly lower compared to the expected frequency according to Fisher's exact test.

Furthermore, to evaluate the diversity of graphs independent from any set of pre-selected terms, we calculated pairwise node identity percentages between all KG pair combinations (930930 measurements between pairs of 1365 KGs) and drew a histogram of these values in log scale (Figure 4.8). This histogram indicates that the node identity distribution roughly follows a power law distribution in the linear-scale, except for 106 graph pairs with a node identity value of 100%. We investigated these cases and found out that they either belong to query terms from two different source databases that indicate the exact same biological entity (e.g., disease entries from EFO and KEGG databases: "Orphanet:98820: Familial focal epilepsy with variable foci" and "H02214: Familial focal epilepsy with variable foci") or query terms with a close semantic relationship in the respective ontology (e.g., phenotype terms: "HP:0012693: Abnormal thalamic size" and "HP:0012695: Decreased thalamic volume", or Reactome pathways: "R-SSC-937039: IRAK1 recruits IKK complex" and "R-SSC-975144: IRAK1 recruits IKK complex upon TLR7/8 or 9 stimulation"), that have the same gene/protein associations. Therefore, they should not be taken into account. The mean node identity value of the distribution is 0.9% (dashed vertical line in Figure S5), which is significantly lower compared to the identity value observed between KGs of 2 cancer types with similar biological mechanisms (i.e., breast and ovarian cancers with 30% pairwise node identity), even lower than the identity value observed between KGs of 2 dissimilar types of cancer (i.e., breast/ovarian cancers and osteosarcoma with 8% pairwise node identity) in the use case experiment given above. It is also important to note that, approximately 98.8% of the graph pairs have less than 10% node identity values, indicating the high diversity of CROssBAR KGs.

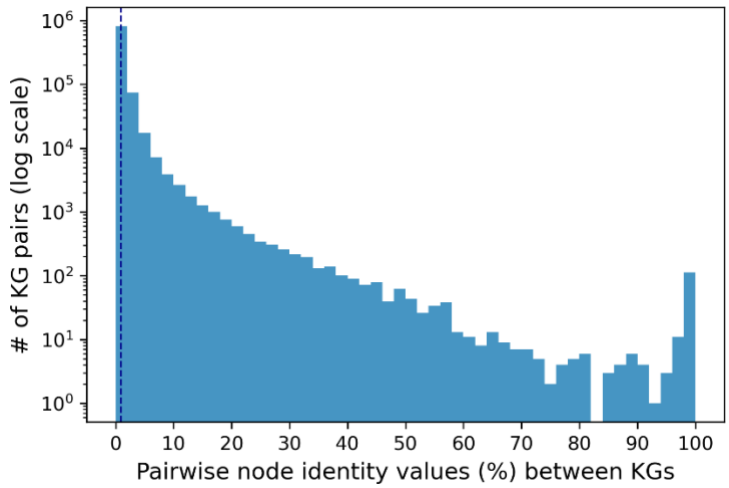


Figure 4.8. Pairwise node identity percentage histogram (in log scale) between all KG pair combinations in our 1365 CROssBAR knowledge graphs constructed with random term queries.

Finally, we calculated observed frequencies of all terms that are presented in our 1365 randomly generated KGs and plotted the results as biological component specific bar graphs, in which different terms are presented on the horizontal axis and their respective observed frequencies are shown on the vertical axis (terms are ranked from the highest observed frequency to the lowest) in Figure 4.9. On each panel, terms from a distinct biological component are shown, together with their mean values as dashed

lines. As shown in Figure 4.9, observed frequency values of even the most frequent terms are considerably low (between 0.012 and 0.055) for all components except KEGG pathways. Moreover, these most frequent terms only constitute a very small portion of the total number of terms in their respective components, which is also indicated by low component-wise mean observed frequency values (dashed lines). For example, the mean frequency value considering core proteins is 0.0025, meaning that, on average a gene/protein is presented in only 1 out of 400 different KGs. Core genes/proteins with the highest observed frequencies are LNMA (P02545), MAPT (P10636) and RAB9A (P51151) with frequency values of 0.057, 0.043 and 0.036, respectively. KEGG pathways are presented in KGs with a mean observed frequency of 0.035 (highest among all biological components), and the most frequent pathways are "Metabolic pathways" (hsa01100), "Neuroactive ligand-receptor interaction" (hsa04080), and "PI3K-Akt signaling pathway" (hsa04151) with 0.227, 0.123 and 0.100, respectively. This was expected since the total number of KEGG signaling pathways are 248 and 10 of them are included in each KG. Also, the one with the highest frequency, "Metabolic pathways", is an umbrella term containing several pathways.

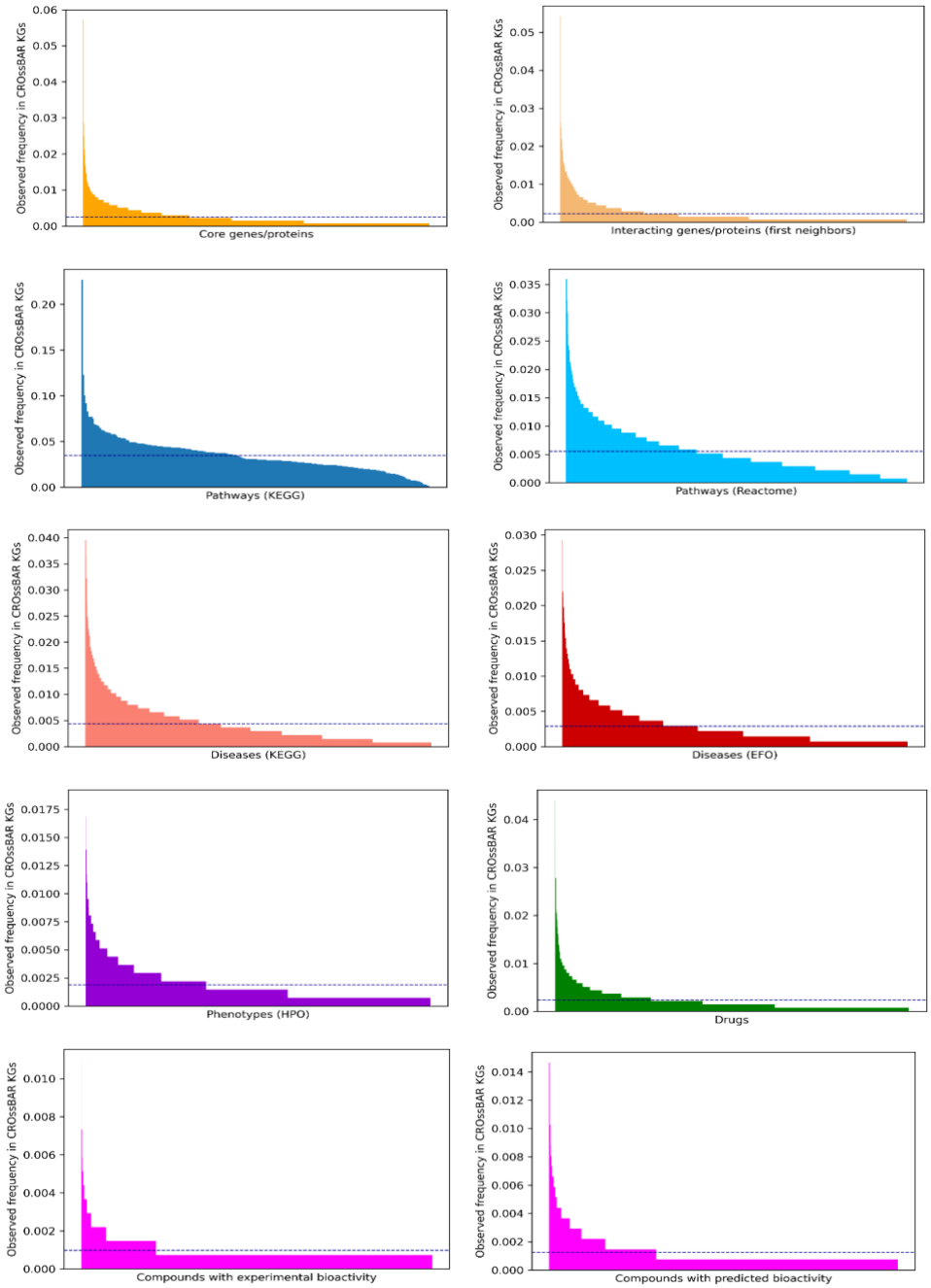


Figure 4.9. Biological component-wise bar graphs indicating observed frequencies (in vertical axis) of all terms (in horizontal axis by ranking the terms according to decreasing frequencies) that are presented in our 1365 CROssBAR knowledge graphs constructed with random term queries. Dashed lines correspond to mean values of observed frequencies.

Biological component-wise bar graphs indicating observed frequencies (in vertical axis) of all terms (in horizontal axis by ranking the terms according to decreasing frequencies) that are presented in our 1365 CROssBAR knowledge graphs constructed with random term queries. Dashed lines correspond to mean values of observed frequencies.

4.4.5. Graph Construction Runtime Tests

Building a CROssBAR knowledge graph is a complex procedure that involves multiple rounds of API queries and quantitative analyses of query results to represent the most relevant nodes and edges as the output. With the aim of observing the practicality of this procedure, we conducted runtime tests by measuring the time (in seconds) that pass from submitting the initial user query to the finalization of the output KG. For this, we utilized the same 1365 random user queries explained in the previous section, which are composed of single term searches of 198 genes/proteins, 92 Reactome pathways, 100 KEGG pathways, 99 EFO diseases, 100 KEGG diseases, 199 HPO terms, 199 drugs and 186 compounds, together with 192 combinatory queries composed of one random term from each component (i.e., gene/protein, pathway, disease, phenotype, drug and compound). The resulting runtimes are shown in Figure 4.10 as histograms, where queries of distinct components are given in different panels, and median times are indicated by vertical dashed lines. As observed from Figure 4.10, runtimes are variable both between the queries of the same component and across different components. Among single term queries, drugs and Reactome pathways constitute the fastest queries with median runtimes of 30 and 29 seconds, respectively, and HPO terms constitute the slowest with 62 seconds. Besides, combinatory term queries took approximately 70 seconds on average. These results indicate that, on average, it is practical to query CROssBAR web-service and generate KGs on-the-fly. It is also important to note that runtimes are approximately linearly correlated with the number of collected core genes/proteins, and thus, querying terms that are associated with high number of genes/proteins (i.e., complex signaling pathways, generic HPO terms, etc.) can take significantly longer compared to mean times shown here. Runtimes also depend on the availability of servers.

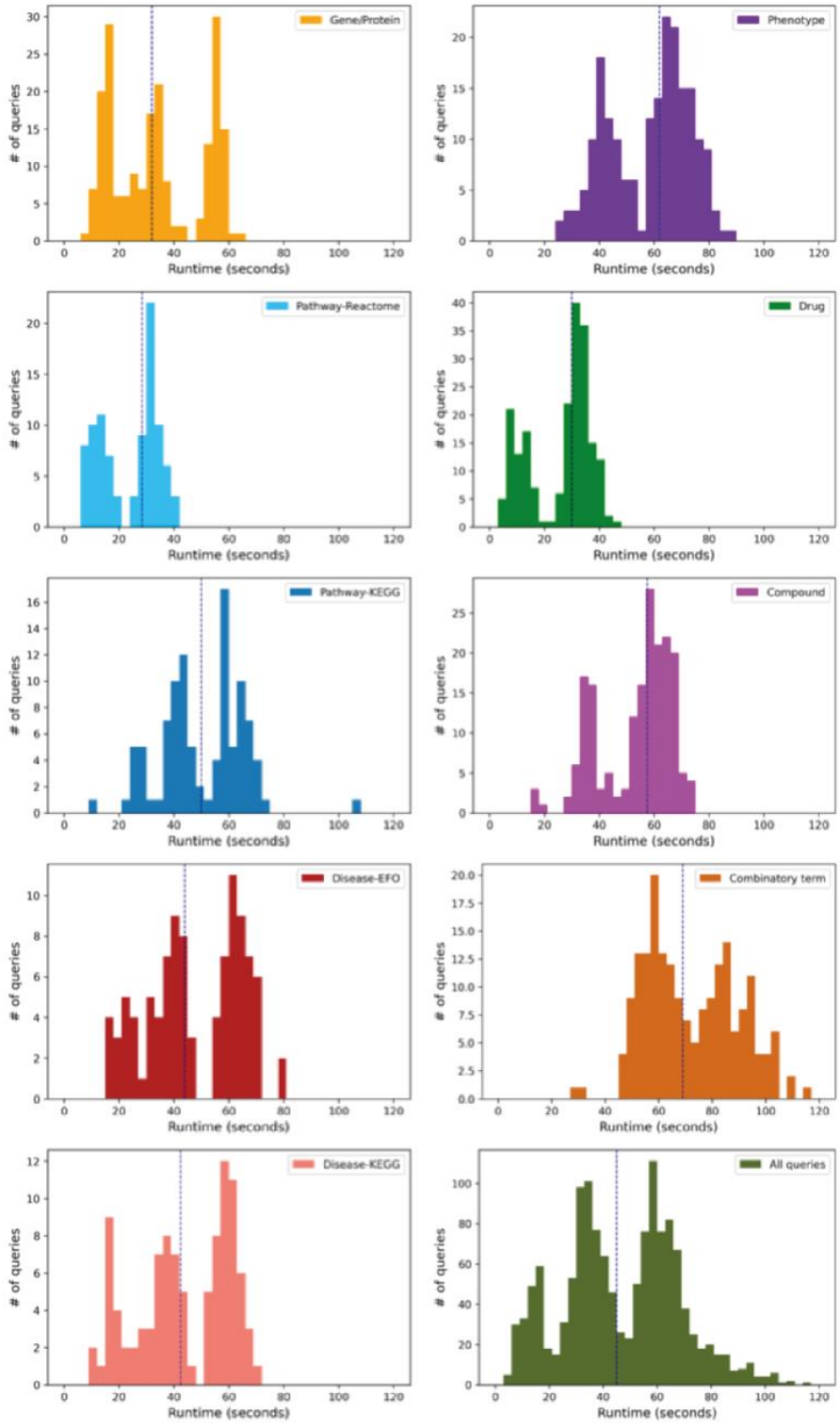


Figure 4.10. Biological component-wise query runtime histograms of 1365 CROssBAR knowledge graphs constructed with random term queries.

4.5. Conclusion

The CROssBAR offers a comprehensive system that integrates and distills large-scale biomedical data from diverse sources, presenting it through heterogeneous knowledge graphs (KGs). The main aim of CROssBAR is to facilitate the interpretation of biological big data by providing a user-friendly interface that presents coherent and easily understandable information.

This chapter specifically focuses on the KG construction module of the CROssBAR. KGs are powerful representations of heterogeneous data, showcasing relationships between different types of entities in a semantic context. CROssBAR KGs enable the exploration of high-level, indirect relationships between specific entities, unveiling hidden connections and facilitating data-driven approaches. The incorporation of overrepresentation analysis enhances the relevance and significance of the included terms, ensuring the graphs represent meaningful relationships between biological components. In CROssBAR, due to the way the overrepresentation analysis is done, specific sub-pathways are incorporated from the Reactome database in most cases, whereas the generic pathway information is incorporated into KGs via KEGG. As a result, pathway information is displayed at different levels of specificity, and thus, not redundant in knowledge graphs. The analyses conducted on the diversity, stability, and feasibility of the constructed KGs demonstrate their reliability and practicality, validating their potential for use in various domains of life sciences research.

We presented a use-case of the system by constructing two COVID-19 KGs. First, the large-scale version, in which nearly the whole of the COVID-19 related data recently accumulated in our source databases are integrated, organized, and presented. Second, the simplified version, where the aim was to provide users with a source that is suitable for quick exploration since the large-scale KG is not easily explorable due to its huge size. We saved the pre-constructed COVID-19 KGs, which are directly accessible and viewable through the links given on our web-service (https://crossbar.kansil.org/covid_main.php). It is also important to note that, due to the content of integrated data resources, CROssBAR heavily contains rare and complex disease data, and mostly leaves infectious diseases out. Nevertheless, the constructed COVID-19 graphs provide rich biomedical information.

Overall, the CROssBAR system empowers researchers by organizing and presenting vast amounts of heterogeneous data, enabling them to gain novel insights, identify potential targets, and expedite the discovery of innovative treatment solutions. With its user-friendly interface and comprehensive data integration, CROssBAR has the potential to revolutionize biomedical research across various domains.

CHAPTER 5

LARGE-SCALE PREDICTION OF DRUG-TARGET INTERACTIONS VIA GRAPH REPRESENTATION LEARNING

5.1. Chapter Overview

Recent developments in data-driven approaches have facilitated the processing and interpretation of vast quantities of biomedical data for drug discovery and development. As a new and practical data structure, heterogeneous knowledge graphs (KGs) have the capacity to represent complex relationships between different layers of biomedical data. In relation to that, graph neural networks (GNNs) have emerged as a novel modelling technique for the inference of graph-based data; however, the majority of GNN algorithms are restricted to homogenous graphs and cannot handle heterogeneous data with multiple types of nodes and edges. Here, we propose a new type of systems-level compound-protein interaction (CPI) representation and subsequent prediction framework called HetCPI, which uses large-scale biomedical KGs obtained from the CROssBAR system as input. To process these biomedical KGs for bioactivity prediction, we employed the heterogeneous graph transformer (HGT) architecture, which handles graph heterogeneity and maintains node- and edge-type dependent representations through its attention mechanism. HetCPI has yielded promising results on challenging protein family-specific benchmark CPI datasets, in comparison to baseline and state-of-the-art methods. HetCPI is anticipated to aid computational drug discovery by leveraging direct and indirect relationships in molecular and cellular processes for bioactivity prediction, thereby accelerating the development of new treatments.

5.2. Introduction

Drug discovery is a complex process involving the identification and optimization of compounds that interact selectively with intended target biomolecules to produce the desired therapeutic effect. Due to the extremely dynamic and complex structure of biological systems, there are numerous factors that influence the outcome of the process. Therefore, computational drug discovery cannot be handled by simple virtual screening alone. On the other hand, taking a systems-based approach that integrates and utilizes direct and indirect relationships in molecular and cellular processes, including protein-protein interactions, drug/compound-protein interactions, and signaling/metabolic pathways, together with high-level concepts such as protein-disease relationships, drug-disease indications, pathway-disease modulations, and phenotypic implications, could increase the success rate of drug discovery.

Advancements in data analysis techniques have facilitated the processing and interpretation of large-scale biomedical data. One of the most promising relationship-centric data types for this purpose is the knowledge graph (KG), which can represent complex associations between different layers of biomedical data. Earlier approaches for constructing biomedical KGs mainly rely on incorporating different biomedical databases in the RDF (Resource Description Framework) format under a unified framework and querying through SPARQL (Antezana et al., 2009; B. Chen et al., 2010). Current construction methods primarily involve extracting information from unstructured text in biomedical literature, often from databases like PubMed (Bakal et al., 2018; Bougiatiotis et al., 2020; S. Yu et al., 2022). These methods involve mining relationships between various biomedical entities from the textual content of scientific articles and then representing them as subject-predicate-object semantic triples, forming the basis of the KG. Another recent approach for biomedical KG construction is based on the integration of diverse biomedical data types from multiple structured sources via cross-references between these sources for mapping. This approach generally maintains the data using a graph database architecture like Neo4j, where entities and their relationships are represented as a network of interconnected nodes and edges. This allows for efficient retrieval and querying of the data using graph-based query languages such as Cypher. Successful applications of these integrative approaches include HetioNet (Himmelstein et al., 2017), BioGrakn (Messina, Pribadi, et al., 2018), CROssBAR (Doğan et al., 2021), Bioteque (Fernández-Torras et al., 2022) and SPOKE (Morris et al., 2023).

Graph neural networks (GNNs) have emerged as a promising modeling technique for the inference of graph-based data by aggregating information from the nodes' neighbors to generate node (or edge) embeddings. However, the majority of GNN methods are restricted to homogeneous graphs or bipartite graphs. Thus, they cannot handle heterogeneous data with multiple types of nodes and edges (Hu et al., 2020). Most GNN applications for DTI prediction rely on the utilization of protein and compound structures as graphs rather than the use of heterogeneous biomedical data (Liao et al., 2022; Torng & Altman, 2019; Yang et al., 2022). Some existing applications incorporating graph-based heterogeneous biomedical data for DTI prediction have primarily focused on drug-drug and protein-protein similarity/interaction networks, as well as protein-drug interaction bipartite graphs, which follow a similar approach to similarity-based ML methods (Thafar et al., 2020, 2022; W. Wang et al., 2022; Yue & He, 2021). However, recent efforts have emerged to benefit from more comprehensive and diverse heterogeneous biomedical graph data, including associations of proteins/genes, drugs/compounds, diseases, side effects, and other relevant information (X.-H. Chen et al., 2023; Jiang et al., 2022; J. Li et al., 2022; Z. Liu et al., 2021; J. Peng et al., 2021; Tian et al., 2022; Wan et al., 2019; Zhou et al., 2021). In the study of Liu et al., a graph autoencoder approach called GADTI was proposed for DTI prediction (Z. Liu et al., 2021). GADTI utilizes a heterogeneous network that integrates diverse datasets related to drugs and targets, including DTI, drug–drug and protein-protein interaction, drug–disease, drug-side effects, and protein–disease association, drug chemical structure similarity, and protein sequence similarity. It combines a GCN and random walk with restart (RWR) in its encoder and employs a DistMult matrix factorization model for the decoder. The success of GADTI is attributed to its ability to aggregate multi-hop neighborhood information while

avoiding over-smoothing. Peng et al. introduce EEG-DTI, an end-to-end learning-based framework for DTI prediction. It can simultaneously optimize the feature extraction process and model parameters for the final prediction task in an end-to-end fashion (J. Peng et al., 2021). EEG-DTI utilizes a heterogeneous network comprising multiple types of biological entities (i.e., drug, protein, disease, and side-effect) and employs multi-layer GCN architecture to handle graph heterogeneity and learn low-dimensional feature representations of drugs and targets for DTI prediction. Compared to existing methods, EEG-DTI demonstrates enhanced performance in DTI prediction. In another study, Ye et al. present KGE_NFM, a unified framework for DTI prediction that combines KG and a recommendation system (Ye et al., 2021). The framework first learns low-dimensional representations for different entities in the KG via DistMult embedding model and then integrates multimodal information using a neural factorization machine (NFM). KGE_NFM was evaluated under realistic scenarios and achieved accurate and robust predictions on different benchmark datasets. In a very recent study by Chen et al., two frameworks, AutoInt_KG and MolGPT_KG, were developed based on the heterogeneity information of transporter-related KG extracted by the RESCAL model to improve drug-transporter prediction and efficient drug design (X.-H. Chen et al., 2023). AutoInt_KG utilizes KG-embeddings and sequence features to predict potential transporters for small molecules via the interaction layer based on the multi-head self-attention mechanism, achieving reliable performance on natural product validation. MolGPT_KG employs KG embeddings and drug SELFIES representations to generate drug-like small molecules targeting specific transporters. These studies highlight the promising use of heterogeneous biomedical data to improve DTI prediction models. By implementing more advanced GNN approaches specifically designed to handle heterogeneous graphs, the effectiveness and potential of these methods can be further enhanced.

In this chapter, we present a new type of systems-level compound-protein interaction (CPI) representation and prediction framework called HetCPI. As its input data, HetCPI utilizes large-scale biomedical KGs constructed by the CROssBAR system (Doğan et al., 2021), which integrates a wide range of publicly available biomedical data sources to build heterogeneous graphs composed of genes/proteins, pathways, diseases, phenotypes, drugs and compounds, together with their multifaceted relationships. HetCPI employs the heterogeneous graph transformer (HGT) architecture (Hu et al., 2020) to learn from highly heterogeneous KGs, extracting node- and edge-type dependent representations via its attention mechanism. This allows HetCPI to effectively capture the complex patterns/webs of biological relationships and generate integrative representations to be used for the subsequent CPI prediction task. To evaluate the performance of HetCPI, we carried out benchmarking experiments on our target protein family-specific bioactivity datasets containing different stratified data splits (see Chapter 3) and compared the results with leading CPI prediction methods from the literature. Furthermore, we conducted a use-case study based on the predictions for druggable and non-druggable protein samples to further evaluate the robustness and reliability of our models. Competitive performance results of HetCPI as well as consistent outcomes in the use-case study indicate the potential of our KG-based approach to improve the accuracy and efficiency of virtual screening, leading to the discovery of new and effective treatments for a wide range of diseases.

5.3. Materials and Methods

5.3.1. Dataset Construction

For the construction of graph-based bioactivity datasets enriched with multiple biological/biomedical relationships, we used the CROssBAR web service. The step-by-step operations below were applied to build these graph structures:

1) A CROssBAR bulk search was performed to construct KGs of all reviewed human proteins in UniProt (i.e., 20,173 protein entries in SwissProt) (The UniProt Consortium, 2021) by querying each protein entry with parameters; predictions=0, num_of_drugs=100, num_of_compounds=100 where other parameters were set as default. Here, we set predictions parameter to 0 to exclude predicted bioactivities due to lower confidence level compared to experimentally measured bioactivities and approved DTIs.
2) For each protein family-specific bioactivity dataset in the first chapter (i.e., epigenetic-regulators, ion-channels, membrane-receptors, transcription-factors, transporters, hydrolases, oxidoreductases, proteases, transferases, and other-enzymes), KGs of proteins belonging to the corresponding family were merged. Therefore, each bioactivity dataset was converted into a KG structure that involves other types of nodes including pathways, phenotypes, diseases apart from proteins and compounds, and relationships such as protein-protein interactions, protein-disease associations, drug-disease indications etc..
3) Proteins not having KGs were removed from bioactivity datasets. Missing compounds or protein-compound interactions involved in filtered datasets but not involved in graphs were merged into graphs with their bioactivity edges.
4) To prevent data leakage, edges of compounds in KGs were removed if they interact with dataset proteins but this interaction is not involved in bioactivity datasets, mainly due to version difference of ChEMBL bioactivity database (Mendez et al., 2019) or assay type filtration. Duplicate nodes and edges in graphs were also removed to prevent redundancy.
5) After constructing KGs based on protein family-specific datasets, bioactivity edges involved in test datasets were removed for each split set (i.e., fully-dissimilar-split, dissimilar-compound-split, random-split). Thus, 3 different train/test split versions of each family-specific KG were generated.
6) Most of the compounds in KGs only relate to one or a few target protein nodes. Therefore, split versions of KGs become disconnected after the removal of test edges. To provide graph connectivity, disconnected compound nodes were reconnected via compound-compound similarity edges. Pairwise compound similarities were calculated using the simsearch function of the Chemfp python package (Dalke, 2019), and compound pairs with a similarity score higher than 0.5 were merged into graphs as edges.

We also introduced a larger KG version as an alternative to this version. In this new version (i.e., integrated CROssBAR KGs), we incorporated all CROssBAR KGs generated in step 1 instead of solely merging KGs of proteins involved in family-

specific datasets. The following steps (steps 3-6) remained the same. Based on preliminary findings, this version became the primary choice for subsequent analyses.

To compare our final models with state-of-the-art models, we used the filtered Davis kinase benchmark dataset, employing the same setup as the MDeePred study (Rifaioglu et al., 2021). The KG trained and tested on the Davis dataset consists of 7,567 train and 1,518 test data points, which is 7,600 and 1,525 in the MDeePred study. The difference is due to the absence of KGs for three proteins (UniProt IDs: Q07785, P62344, and P9WI81) in CROssBAR since they belong to Plasmodium falciparum and Mycobacterium tuberculosis organisms.

Table 5.1 presents node and edge statistics of protein family-specific KG datasets in the first version for each split set. Table 5.2 provides node-type statistics of these KGs, which remain consistent across all splits. Table 5.3 displays the node- and edge-type statistics for the integrated CROssBAR KG version, also employed for three split forms of each protein family, along with the data points from the Davis dataset.

Table 5.1. Node and edge statistics of protein family-specific KG datasets for **(a)** fully-dissimilar-split, **(b)** dissimilar-compound-split, **(c)** random-split strategy

(a)

Fully-dissimilar-split			
Protein family	**Node number**	**Edge number**	**Component size**
epigenetic-regulators	16,620	184,348	141
hydrolases	48,755	481,505	265
ion-channels	31,117	998,961	104
membrane-receptors	106,555	2,505,460	295
other-enzymes	18,277	181,085	51
oxidoreductases	35,621	417,517	71
proteases	55,970	1,162,614	144
transcription-factors	18,712	292,462	72
transferases	120,569	3,255,118	188
transporters	18,555	306,128	128

(b)

Dissimilar-compound-split			
Protein family	**Node number**	**Edge number**	**Component size**
epigenetic-regulators	16,620	184,441	60
hydrolases	48,755	481,449	244
ion-channels	31,117	998,965	96
membrane-receptors	106,555	2,506,142	250
other-enzymes	18,277	179,971	90
oxidoreductases	35,621	416,904	89
proteases	55,970	1,162,145	128
transcription-factors	18,712	292,466	55
transferases	120,569	3,254,123	172
transporters	18,555	306,234	55

(c)

Random-split			
Protein family	**Node number**	**Edge number**	**Component size**
epigenetic-regulators	16,620	184,388	40
hydrolases	48,755	481,542	265
ion-channels	31,117	998,973	51
membrane-receptors	106,555	2,506,136	41
other-enzymes	18,277	179,557	47
oxidoreductases	35,621	416,971	50
proteases	55,970	1,162,239	50
transcription-factors	18,712	292,499	53
transferases	120,569	3,253,553	90
transporters	18,555	306,596	18

Table 5.2. Node-type statistics of protein family-specific KG datasets

Node type	**Hydrolases**	**Proteases**	**Oxidoreductases**	**Transferases**	**Other enzymes**
Protein	246	183	147	583	103
Protein_N	1,688	1,203	924	3,866	784
Drug	1,920	1,896	1,734	3,136	1,122
Compound	40,590	49,456	29,756	105,424	13,917
Pathway	1,138	854	767	1,994	662
kegg_Pathway	234	216	223	244	216
HPO	1,794	1,279	1,160	3,305	840
Disease	678	521	519	1,234	360
kegg_Disease	467	362	391	783	273

Node type	Epigenetic regulators	Ion channels	Membrane receptors	Transcription factors	Transporters
Protein	96	103	257	56	93
Protein_N	803	455	1,157	445	547
Drug	1,070	855	1,961	822	1,003
Compound	12,751	27,940	99,691	16,112	15,061
Pathway	543	468	881	334	478
kegg_Pathway	175	160	203	147	177
HPO	754	708	1,471	435	678
Disease	247	248	538	190	307
kegg_Disease	181	180	396	171	211

Table 5.3. (a) Node- and **(b)** edge-type statistics of the integrated CROssBAR KGs.

(a)

node_type	size
Compound	422,617
Protein	23,419
HPO	8,971
Drug	5,420
Disease	3,815
Pathway	3,184
kegg_Disease	1,879
kegg_Pathway	245
TOTAL	469,550

(b)

edge_type	size
comp_sim	10,256,336
Chembl	629,590
PPI	91,443
HPO	39,466
Pathway	32,706
hpodis	24,259
kegg_path_prot	19,407
Drug	15,342
Disease	7,270
kegg_dis_prot	5,911
kegg_dis_path	1,682
kegg_dis_drug	298
TOTAL	11,123,710

5.3.2. The Representation of Graph Nodes

Graph representation learning algorithms allow us to predict unknown relationships on a graph by incorporating topological structures of graphs via the learning process. However, it requires a knowledge transfer of nodes represented with feature vectors. Although it is possible to randomly initialize feature vectors or to use trivial solutions such as one-hot encodings, utilizing more representative features or embeddings to capture hidden patterns of nodes increases the success rate of graph models. Here, we used the representation approaches below as attributes of each node type:

- *Proteins* were represented by combined vectors of transformer-avg embeddings (vector size: 768), apaac descriptors (vector size: 80), and k-sep_pssm descriptors (vector size: 400) as well-performing protein representations overall based on the results of the benchmark chapter. We also evaluated the performance of a recently promising learned embedding method prott5 (Elnaggar et al., 2021) on filtered human bioactivity datasets of the

transferase, protease, and ion-channel families along with other well-performing conventional descriptors (i.e., apaac and k-sep_pssm) and learned embeddings (i.e., unirep1900 and transformer-avg) with the same benchmark setup (Table 5.8), but it couldn't outperform the others. *Biotech drugs* were also represented by protein embeddings (i.e., transformer-avg embeddings) since they are composed of amino acid sequences.

- *Compounds/Small molecule drugs* were represented by our in-house developed SELFIES embedding approach called "SELFormer" (vector size: 768) (Yüksel et al., 2023). It is a transformer-based NLP model that employs a large-scale pre-training methodology on 2 million molecules in their SELFIES notations to learn flexible and high-quality molecular representations. It performed competitive results with MolBERT (Fabian et al., 2020) and ChemBERTa (Chithrananda & Ramsundar, 2020) embedding approaches on molecular analysis tasks. Alternatively, we constructed models based on the ECFP4 fingerprints (vector size: 1024) of compounds and small molecule drugs.

- *Pathway* representations (vector size: 200) were obtained using TransE embedding method via BioKEEN library (Ali et al., 2019). It utilizes a gene-pathway association network to generate representations. For Reactome pathways, we also used pre-calculated Bioteque KG embeddings (vector size: 128) (Fernández-Torras et al., 2022; Rifaioglu et al., 2021) generated on gene-pathway-disease metapath using a random walk method, which became our final choice.

- *HPO phenotype term* embeddings (vector size: 160) were retrieved from CADA tool (C. Peng et al., 2021). It is a node2vec-based embedding method that uses a gene-phenotype association network as the input graph, which includes disease-level annotations and clinical cases-level annotations.

- *Disease* embeddings (vector size: 100) were obtained using the doc2vec method based on PrimeKG (Chandak et al., 2022) disease definitions. If PrimeKG definition is not available, the description of the disease was taken from its source, and if it has no description, then its name was used as input.

5.3.3. Model Design and Architecture

To generate a reliable and powerful GNN-based CPI prediction model, the selection of input graph data is as critical as the selection of the algorithm, feature vectors, and hyperparameters. Here, we use integrated CROssBAR KGs merged with protein family-specific bioactivity datasets to construct our CPI prediction framework, HetCPI. CROssBAR KGs include only prioritized nodes and edges that are most relevant to the query entry rather than a whole set of biological interactions, which eliminates redundancy and provides clean data that may be more informative for the prediction of CPIs. To process these heterogenous KGs for bioactivity prediction, we used heterogeneous graph transformer (HGT) architecture (Hu et al., 2020) and excluded all bioactivities (i.e., compound-protein edges) from the input graph to prevent data leakage during the message passing procedure, using them only in the prediction part. The HGT model proposed by Hu et al outperforms all the state-of-the-

art GNN baselines for various downstream tasks on the Open Academic Graph with 179 million nodes and 2 billion edges (Hu et al., 2020).

Figure 5.1 displays the schematic workflow of the HetCPI system. The learning process starts with the transfer of knowledge among nodes through their initial attributes, represented by input node feature vectors. This information is fed into the HGT algorithm, which embeds each target node into a lower dimensional space by leveraging meta-relations to handle graph heterogeneity. Once the embeddings are obtained via HGT architecture, the HetCPI system performs an edge regression task by calculating the dot products of the compound and protein embeddings for each pair in the training set, which serve as predicted bioactivity measurements. In addition, the system offers an alternative model design (HetCPI-3FC) where compound and protein representations are concatenated instead of using dot products. These concatenated vectors then pass through three fully-connected layers to generate predictions. To train the model, the loss is computed for the true and predicted bioactivity values of the compound-protein pairs using the mean squared error (MSE) loss function. The weight parameters employed by the HGT architecture are then updated through the backpropagation process, using Adam optimization. This iterative process continues until the loss decreases to a satisfactory level. Following the training phase, the model generates predictions for the samples in the validation and test sets using the embeddings updated based on the final weights. These predictions are then used to evaluate the performance of the model.

We employed HGT architecture primarily due to its ability to handle graph heterogeneity and maintain node- and edge-type dependent representations. It achieves this via heterogeneous mutual attention, heterogeneous message passing, and target-specific aggregation steps that incorporate information from source nodes in order to generate a contextualized representation for each target node. These steps are explained below:

1) Heterogeneous mutual attention: HGT introduces a new mechanism for attention calculation that considers the meta-relations between nodes. Inspired by Transformer architecture (Vaswani et al., 2017), it maps the target node *t* to a Query vector *Q* (4), and the source node *s* to a Key vector *K* (3) with linear projection. Then, instead of directly taking the dot product of the Query and Key vectors like the vanilla Transformer, it uses different weight matrices (W^{ATT}) for each edge type ($\phi(e)$) to calculate the attention matrix for *h* heads, allowing it to capture different semantic relationships between nodes (2). Additionally, a prior tensor (μ) is introduced to represent the general significance of each meta-relation triplet ($<\tau(s),\phi(e),\tau(t)>$) since not all the relationships contribute equally to the target nodes, which is divided by square root of the vector dimension per head ($\sqrt{d}$). It serves as an adaptive scaling factor for attention. To obtain the attention vector of each node pair, multiple attention heads are concatenated. Then, for each target node *t*, the attention vectors are gathered from its neighboring source nodes *N(t)*, and the softmax function is applied to ensure attention weights that sum up to 1 across the source nodes (1). Given a graph with the input node features $H^{(l-1)}$, the *h*-head attention score for each edge (s,e,t) is computed as follows:

$$Attention(s,e,t) = \text{Softmax}_{\forall s \in N(t)} (\|_{i \in [1,h]} ATT - head^i(s,e,t)) \quad (1)$$

$$ATT - head^i(s,e,t) = (K^i(s)W^{ATT}_{\phi(e)}Q^i(t)^\top)\cdot(\mu_{<\tau(s),\phi(e),\tau(t)>} / \sqrt{d}) \quad (2)$$

$$K^i(s) = \text{K-Linear}^i_{\tau(s)}(H^{(l-1)}[s]) \quad (3)$$

$$Q^i(t) = \text{Q-Linear}^i_{\tau(t)}(H^{(l-1)}[t]) \quad (4)$$

2) Heterogeneous message passing: Parallel to the calculation of mutual attention, HGT performs message passing from source nodes to target nodes. Similar to the attention process, this process incorporates the meta relations of edges into the message passing to address the distribution differences of nodes and edges of different types. For each pair of nodes $e = (s,t)$, the HGT calculates a multi-head message by projecting the source node s into message vectors M and using a matrix (W^{MSG}) to incorporate edge dependency (6). Multiple message heads (h) are concatenated to obtain the message for each node pair (5). The message to send on each edge (s,e,t) is computed as follows:

$$Message(s,e,t) = \|_{i\in[1,h]} MSG\text{-}head^i(s,e,t) \quad (5)$$

$$MSG - head^i(s,e,t) = \text{M-Linear}^i_{\tau(s)}(H^{(l-1)}[s])W^{MSG}_{\phi(e)} \quad (6)$$

3) Target-specific aggregation: After the calculation of heterogeneous mutual attention and message passing, the HGT aggregates the information from the source nodes to the target node. The attention vectors obtained from the softmax procedure in the first step ($Attention(s,e,t)$) serve as weights to average the corresponding messages from source nodes ($Message(s,e,t)$) to get the updated vector ($\tilde{H}^{(l)}[t]$). This aggregation process is performed for all target nodes, incorporating information from their neighboring source nodes of different feature distributions (7). The updated vector for each target node is then mapped back to its type-specific distribution using a linear projection followed by a non-linear activation and residual connection (8). The output vector ($H^{(l)}[t]$) of the l-th HGT layer for the target node t is computed as follows:

$$\tilde{H}^{(l)}[t] = \sum_{\forall s\in N(t)} (Attention(s,e,t)\cdot Message(s,e,t)) \quad (7)$$

$$H^{(l)}[t] = \sigma(\text{A-Linear}_{\tau(t)}\tilde{H}^{(l)}[t]) + H^{(l-1)}[t] \quad (8)$$

By stacking multiple HGT blocks for L layers, where L is a small value, the HGT enables each node to reach a large proportion of nodes with different types and relations in the full graph. This results in highly contextualized representations for each node, which can be used for various downstream tasks in heterogeneous networks, such as node classification and link prediction.

The HGT's architecture is designed to leverage the meta-relations to parameterize the weight matrices separately. By distinguishing operators for different relations, the HGT can effectively handle distribution differences in heterogeneous graphs while still achieving parameter sharing. This approach benefits relations with few occurrences by enabling fast adaptation and generalization while maintaining specific characteristics for different relationships through a smaller parameter set.

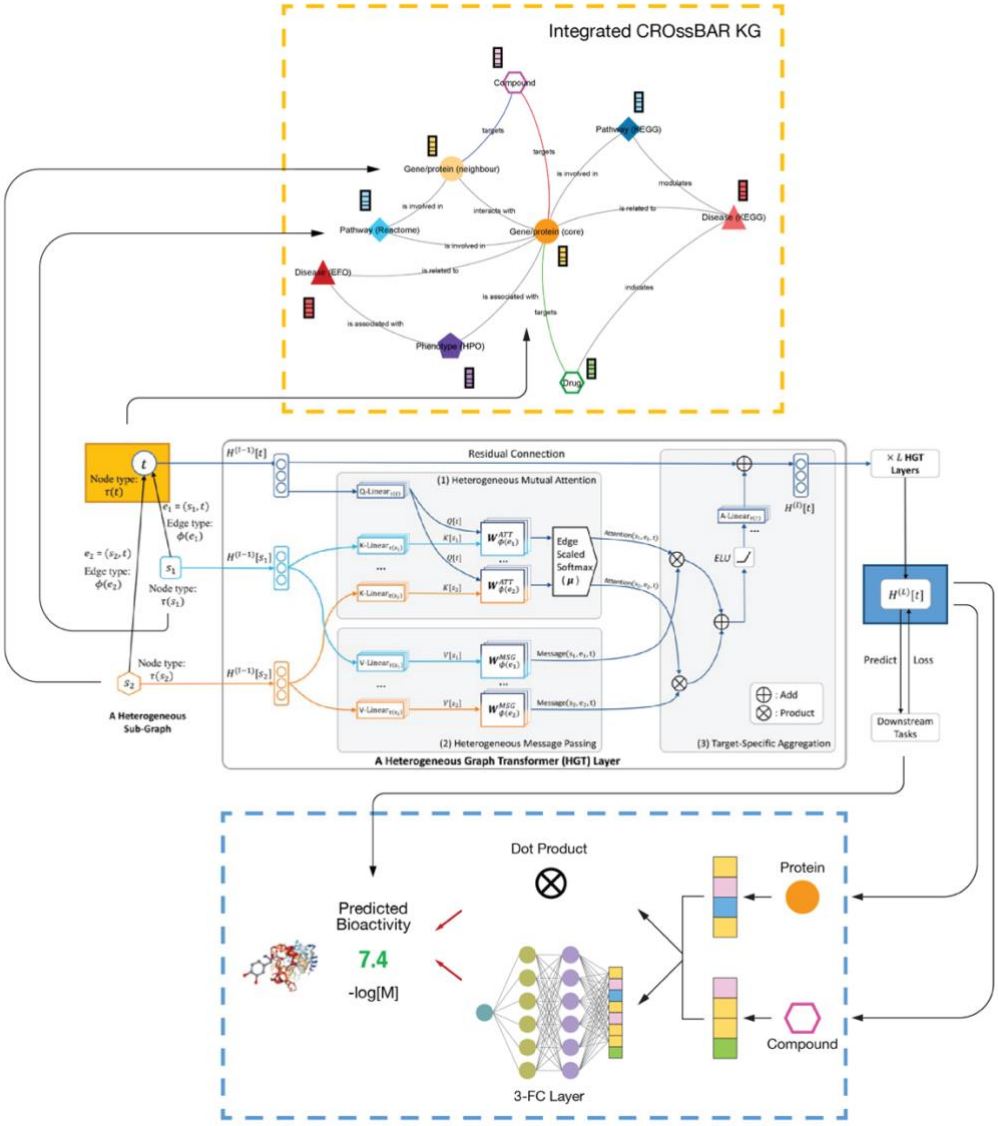


Figure 5.1. The schematic representation of the HetCPI framework. (The HGT part of the figure was adapted from Hu et al. (Hu et al., 2020)).

5.4. Results and Discussion

This section comprises multiple subsections that present a comprehensive evaluation of the HetCPI system across various settings, accompanied by multiple analyses. The first subsection presents the preliminary results obtained from models generated to determine optimal design choices and identify suitable hyperparameter ranges. The second subsection focuses on the performance analysis of the final HetCPI models developed for transferases and membrane receptors, along with a comparison against baseline RF regression models. The third subsection compares HetCPI models with state-of-the-art bioactivity prediction models using the well-known Davis kinase benchmark dataset. In the subsequent subsection, an applicability domain analysis of HetCPI models is performed to assess their usability. The final section presents a use-case study that further supports the reliability and robustness of the HetCPI framework.

5.4.1. Preliminary Results

To conduct these trials, we mainly utilized dissimilar-compound-split datasets of proteases KGs due to a manageable sample size of this protein family, allowing us to perform a substantial number of trials within a reasonable timeframe. We evaluated the performances using the same evaluation metrics in Chapter 3, including RMSE, Spearman, and MCC scores.

5.4.1.1. Hyperparameter search

To search for optimal hyperparameters of models, we split train samples into 80/20 train/validation sets and used the random search method for the hyperparameter ranges given below with epoch number 20:

- *hidden channels:* 8, 16, 32, 64, 128
- *learning rate:* 1e^-5, 0.0001, 0.001, 0.005, 0.01
- *head number:* 1, 2, 4, 8, 16
- *layer number:* 1, 2, 3, 4
- *train batch size:* 128, 256, 512, 1024
- *weight decay:* 0, 1e^-5, 5e^-5, 0.0001, 0.001

Then, we rerun promising models with higher epoch numbers and calculated model performances. Table 5.4 and Table 5.5 display hyperparameters and performance scores of the ten top-performing models ranked by test Spearman scores, respectively. The best model reaches 0.41 for Spearman's correlation and 0.30 for MCC (i.e., median corrected) on the test set while 0.60 (Spearman) and 0.42 (MCC) on the validation set. The score differences between the validation and test set are expected since the compound samples in the test set are not similar to the ones in the train/validation set on the dissimilar-compound-splitting strategy.

Table 5.4. Hyperparameters of top ten models on dissimilar-compound-split of proteases dataset.

Model Name	Epoch Number	Hidden Channels	Learning Rate	Head Number	Layer Number	Batch Size	Weight Decay
model 1	100	32	0.0001	16	1	256	0.00001
model 2	200	32	0.01	8	1	256	0.00005
model 3	200	16	0.0001	1	3	128	0.001
model 4	100	32	0.0001	4	1	128	0.001
model 5	200	128	0.0001	1	1	1024	0.001
model 6	500	32	0.0001	8	2	256	0.00005
model 7	500	128	0.0001	1	1	1024	0.001
model 8	200	16	0.001	4	2	256	0.00005
model 9	200	64	0.0001	8	1	512	0
model 10	200	128	0.0001	2	1	1024	0.00001

Table 5.5. Performance scores of top ten models on dissimilar-compound-split of proteases dataset.

Model Name	Test Loss	Test MCC	Test RMSE	Test Sp.	Train Loss	Valid. Loss	Valid. MCC	Valid. Sp.
model 1	1.612	0.295	1.360	**0.407**	1.209	1.234	0.421	0.597
model 2	1.742	**0.304**	**1.331**	0.382	1.297	1.340	0.411	0.566
model 3	**1.610**	0.277	1.358	0.369	1.194	1.193	0.444	0.617
model 4	1.700	0.215	1.392	0.354	1.215	1.230	0.441	0.602
model 5	1.670	0.242	1.385	0.346	1.277	1.284	0.414	0.581
model 6	1.735	0.254	1.368	0.345	**0.887**	**0.950**	**0.550**	**0.714**
model 7	1.805	0.201	1.442	0.344	1.124	1.154	0.454	0.629
model 8	1.800	0.256	1.420	0.335	1.050	1.144	0.482	0.660
model 9	1.824	0.187	1.425	0.333	1.052	1.139	0.481	0.647
model 10	1.851	0.194	1.442	0.326	1.210	1.332	0.421	0.591

* Median correction was applied for the calculation of MCC and RMSE scores.
* Sp.: Spearman's correlation, Valid.: Validation

5.4.1.2. Different settings for model design

We made several changes to the architecture of the top model in Table 5.5 using the same hyperparameter values to further investigate the impact of various variables on model performance. This time, we didn't split the train set into train/validation folds. The following are the configurations that we assessed:

- *loss function:* mse_loss (default), l1_loss
- *feature scaling:* unscaled (default), standard scale
- *dropout:* applied, not applied (default)
- *neighbor_size in train batches:* -1 (include all neighbors for train edges in the batch, default), 4*2 (include 4 neighbors for 2 levels), 3*3, 4*3
- *activation function:* relu (default), leaky_relu
- *pooling function:* sum (default), mean
- *2 fully-connected layers:* added (fc_2*), not added (default)
- *train/test edges:* excluded from graphs (default), added only train edges (only-tr-edg), added both edges (tr-ts-edg)

Table 5.6 presents the results of the modified models based on the aforementioned arrangements with epoch number 100. We made some inferences from these findings that might be useful in the development of the model architecture. First of all, using standard scaling for features, dropout function for regularization, or leaky relu function for the activation yielded lower performance results than the model with default settings. Involving only train bioactivity edges in the graph instead of removing or adding both train and test edges decreased performance, as well. Other setups such as adding fully-connected layers, selecting different loss and pooling functions, or the neighbor size in train batches gave competitive results with the default model. Hence,

they might be reconsidered in the optimization and finalization phases of the models to enhance performances.

Table 5.6. Performance scores of the models in different design setups

Model Name	Train Loss	Test Loss	Test MCC	Test RMSE	Test Spearman
tr-ts-edg	1.106	1.704	**0.375**	1.296	**0.485**
default	1.184	1.678	0.318	**1.291**	0.452
fc_2*hd-128_l1_loss	0.835	1.057	0.327	1.305	0.450
fc_2*hd-128	1.164	1.694	0.309	1.300	0.444
l1_loss	0.871	1.048	0.349	1.305	0.443
neigh_4*2_l1_loss	0.899	1.049	0.301	1.319	0.433
neigh_4*3_l1_loss_mean	0.897	1.026	0.299	1.297	0.429
l1_loss_mean	0.884	1.044	0.291	1.305	0.427
neigh_3*3	1.246	1.707	0.260	1.308	0.427
tr-ts-edg_l1_loss	0.835	1.076	0.337	1.345	0.427
neigh_3*3_l1_loss_mean	0.899	1.034	0.318	1.306	0.425
neigh_4*2_l1_loss_mean	0.903	1.028	0.309	1.300	0.423
fc_2*hd-512	1.094	1.778	0.294	1.315	0.417
neigh_3*3_l1_loss	0.888	1.053	0.304	1.325	0.417
neigh_4*3_l1_loss	0.887	1.050	0.290	1.325	0.414
leaky-relu_	1.132	1.774	0.278	1.337	0.414
neigh_4*2	1.287	1.753	0.265	1.325	0.408
tr-ts-edg_fc_2*hd-128_l1-loss	0.796	1.102	0.304	1.361	0.383
l1_loss_dropout	0.920	1.267	0.231	1.346	0.374
standard-scale	0.720	2.021	0.303	1.413	0.372
dropout	1.355	2.550	0.219	1.331	0.368
only-tr-edg_l1_loss	0.834	3.069	0.264	2.412	0.361
only-tr-edg	1.107	6.852	0.237	1.431	0.355
standard-scale_l1-loss	0.659	1.162	0.280	1.498	0.328

5.4.1.3. Ablation study based on the removal of different graph components and relations

We performed an ablation study to investigate the effect of different graph node/edge types on learning. After the removal of each component and relation type, we rerun models five times with epoch number 80 and averaged performances. Based on the Spearman test performance results displayed in Table 5.7, the removal of compound-compound similarities (i.e., no_ccs, 0.34), hpo term- and biotech drug-related associations (i.e., no_hpo & no_bd, 0.35 & 0.368), and protein-protein interactions (i.e., no_ppi, 0.393) significantly reduced performance compared to the default model, which includes all interactions (0.455). The single contribution of compound-compound similarities is very high (i.e., cpi_ccs, 0.411) when compared to the baseline

model, which only includes compound-protein interactions (0.34). The absence of disease-related associations (i.e., no_dis, 0.437) caused a slight decrease in performance while the removal of pathway-related associations (i.e., no_path, 0.46) slightly increased the performance. The removal of small-molecule drug-related associations (i.e., no_smd, 0.451) didn't affect performance. Moreover, the single contributions of pathways (i.e., cpi_path, 0.327) and small-molecule drugs (cpi_spi, 0.325) were worse than the baseline model (0.34). However, the involvement of small-molecule drug-related associations along with biotech drugs and pathways slightly increased performance (i.e., no_path, 0.46) compared to the model "cpi_ccs_ppi_hpo" (0.457). One of the reasons for the slight negative effect of pathway associations on performance might be the inclusion of noise through attributes of these nodes if their representation capability is poor. To further explore and handle this situation, we can train models with alternative options of pathway node attributes.

Table 5.7. Test performance scores of the models in the ablation study.

Model Name	**Involved edges**	**MCC**	**RMSE**	**Spearman**
no_path	cpi_ccs_ppi_bpi_spi_hpo_dis	0.317	**1.279**	**0.460**
cpi_ccs_ppi_hpo	cpi_ccs_ppi_hpo	0.316	1.286	0.457
all_included (default)	cpi_ccs_ppi_bpi_spi_hpo_dis_path	0.332	1.282	0.455
no_smd_path	cpi_ccs_ppi_bpi_hpo_dis	0.333	1.290	0.452
no_smd	cpi_ccs_ppi_bpi_hpo_dis_path	**0.348**	1.293	0.451
cpi_ccs_hpo	cpi_ccs_hpo	0.325	1.285	0.443
no_dis	cpi_ccs_ppi_bpi_spi_hpo_path	0.322	1.297	0.437
cpi_ccs_ppi_bpi_hpo	cpi_ccs_ppi_bpi_hpo	0.317	1.301	0.428
cpi_ccs	cpi_ccs	0.275	1.306	0.411
no_ppi	cpi_ccs_bpi_spi_hpo_dis_path	0.240	1.331	0.393
no_bd	cpi_ccs_ppi_spi_hpo_dis_path	0.263	1.341	0.368
no_hpo	cpi_ccs_ppi_bpi_spi_dis_path	0.276	1.371	0.350
no_ccs	cpi_ppi_bpi_spi_hpo_dis_path	0.271	1.356	0.342
only_cpi_included (baseline)	cpi	0.260	1.338	0.340
cpi_path	cpi_path	0.266	1.362	0.327
cpi_spi	cpi_spi	0.271	1.352	0.325

* cpi: compound-protein interactions, ccs: compound-compound similarities
bpi: biotech drug-protein interactions, spi: small-molecule drug-protein interactions
ppi: protein-protein interactions,
smd, bd: nodes/edges belonging to small molecule drugs (smd) and biotech drugs (bd)
path, dis, hpo: all associations belonging to pathway (path), disease (dis), and phenotype (hpo) terms

We also compared these results with the protein representation comparison study mentioned in *Section 5.3.2. (The representation of graph nodes)*. The top-performing graph models above compete with these RF models (Table 5.8), but they need to be tuned to surpass the best-performing RF models.

Table 5.8. Test performance scores of RF regression models on dissimilar-compound-split of human proteases bioactivity dataset.

Model Name	RMSE	Spearman	MCC
apaac	1.276	0.446	0.301
k-sep_pssm	**1.259**	**0.485**	**0.325**
prott5	1.279	0.446	0.299
transformer-avg	1.269	0.466	**0.325**
unirep1900	1.285	0.439	0.304

5.4.2. Utilization of Alternative Node Attributes and Integrated CROssBAR KGs for Model Construction

In this part, we extended our analyses by developing models based on the integrated CROssBAR KGs, which were trained on dissimilar-compound-split datasets of proteases and transferases. These models were then compared with their corresponding family-specific KG versions. We also utilized alternative node attributes for Reactome pathway nodes (i.e., Bioteque embeddings) and proteins (i.e., apaac and k-sep_pssm descriptors). Moreover, we removed all edges between compound-protein pairs. In our case, these edges signify the presence of experimental bioactivity measurements, rather than indicating the actual interaction between the pairs because the prediction task here is an edge regression based on the prediction of real bioactivity measurements. However, the edges between compound-protein pairs involved in CROssBAR KGs represent the presence of bioactivity for corresponding pairs with a pChEMBL threshold higher than 5. Therefore, including these edges for the message-passing process can introduce bias into our predictions.

The hyperparameter values used for the models in this subsection are as follows:

- *hidden channels:* 32 (for transferases), 64 (for proteases),
- *train batch size:* 256 (for family-specific KGs of proteases), 512 (for all the others),
- *weight decay:* 1e^-5 (for proteases), 5e^-5 (for transferases),
- *epoch number:* 25 (for proteases), 60 (for integrated CROssBAR KGs of proteases), 100 (for family-specific KGs of proteases),
- *learning rate:* 0.0001, *head number:* 8, *layer number:* 3.

Based on the findings in Table 5.9, models utilizing integrated CROssBAR KGs consistently outperformed models utilizing family-specific KGs for both protein families. While some models exhibited similar performance across both versions (e.g., tr_btq for transferases), there is a notable difference in general (e.g., Spearman (Sp.) score of tr-ap-ks_btq model is 0.53 and 0.44 for integrated-CROssBAR KGs and for family-specific KGs on transferases, respectively). As an alternative strategy to these modelling approaches, we developed a single model trained on integrated CROssBAR KGs using training data from all families, rather than conducting family-specific training. We evaluated the performance of this model on aggregated test data from all families, as well as by considering predictions for each family independently. Although the overall results on the combined test samples were moderate (Sp.: 0.47), the family-specific calculations yielded poorer scores (transferases: 0.476 (Sp.), proteases: 0.378 (Sp.)) compared to the models based on the family-specific training.

Therefore, we decided not to employ the model trained on all data points from all families.

In terms of node attributes, Bioteque Reactome pathway embeddings seem a better alternative than BioKEEN embeddings (default embedding utilized in the models above), providing a slight but consistent increase in performance (e.g., tr-ap-ks_btq: 0.529 (Sp.), tr-ap-ks: 0.522 (Sp.) on transferases; tr_btq: 0.481 (Sp.), tr: 0.451 (Sp.) on proteases for integrated-CROssBAR KGs). For protein representations, there is no consistent outcome valid for all families. Their effect varies among different protein families, parallel to the results of the ProtBENCH study in Chapter 3. For transferases, combinations of transformer embeddings with apaac and k-sep_pssm descriptors yielded the best performance while utilizing transformer embedding alone achieved higher performance than the others.

Overall, based on these outcomes, we determined several key points for the design choices of our finalized HetCPI system. For input KG, we constructed our models using integrated CROssBAR KGs rather than family-specific KGs and trained independently for each bioactivity dataset instead of merging their data points. For node attributes, we replaced Reactome pathway BioKEEN embeddings with Bioteque embeddings. For protein embeddings, we combined transformer, apaac, and k-sep_pssm representations for transferases.

Table 5.9. Test performance scores of models constructed using different node attribute types and alternative KG versions on dissimilar-compound-split bioactivity datasets of human proteases and transferases.

Model Name	Protein Family	KG	Spearman	Med. Cor. MCC	Med. Cor. RMSE
tr-ap-ks_btq	transferases	integrated-CROssBAR	**0.529**	**0.395**	**1.066**
tr-ap-ks	transferases	integrated-CROssBAR	0.522	0.359	1.076
tr-ap_btq	transferases	integrated-CROssBAR	0.511	0.371	1.091
tr_btq	transferases	family-specific	0.509	0.375	1.077
tr_btq	transferases	integrated-CROssBAR	0.505	0.344	1.073
ap_btq	transferases	integrated-CROssBAR	0.497	0.353	1.095
tr	transferases	integrated-CROssBAR	0.491	0.344	1.080
ap	transferases	integrated-CROssBAR	0.458	0.361	1.115
tr-ap-ks_btq	transferases	family-specific	0.443	0.373	1.128
tr_btq	proteases	integrated-CROssBAR	**0.481**	**0.355**	**1.266**
tr_btq	proteases	family-specific	0.452	0.308	1.280
tr	proteases	integrated-CROssBAR	0.451	0.388	1.284
tr-ap-ks_btq	proteases	integrated-CROssBAR	0.414	0.271	1.327
ap	proteases	integrated-CROssBAR	0.409	0.280	1.313
tr-ap_btq	proteases	integrated-CROssBAR	0.385	0.282	1.323

* tr: transformer, ap: apaac, ks: k-sep_pssm (protein representations), btq: Bioteque Reactome pathway embeddings

5.4.3. Model Performance Evaluation for Different HetCPI Architectures

Transferases are one of the most prominent protein families with a high number of protein members that have a significant role in drug discovery. Therefore, the finalized models of the HetCPI system are specifically designed for transferases. To determine the optimal model architecture for the HetCPI system, we constructed alternative model structures that incorporate three fully-connected (FC) layers. These layers take the compound and protein embeddings generated from the HGT layer as the input. Instead of computing the dot product between these embeddings, they are merged together by concatenation and sequentially passed through the FC layers. We also investigated the contribution of ECFP4 compound fingerprints and compared them with SELFormer embeddings in terms of their effect on model performance. To compare the HetCPI models with baselines, we constructed random forest (RF) regression models since RF is a widely used algorithm in CPI prediction. These RF models utilized the initial protein and compound node attributes of the HetCPI models as input features. Furthermore, we developed additional baseline models to investigate whether the models could effectively learn protein and compound embeddings through heterogeneous graphs using the HGT algorithm. These baseline models were constructed by excluding the HGT module from the HetCPI architecture and processing initial protein and compound embeddings either computing their dot products (DP_SELFormer and DP_ECFP4) or passing through 3-FC layers (3FC_SELFormer and 3FC-ECFP4).

For the optimization of these models, we performed a hyperparameter search on our three different stratified datasets for the transferases (146,677 data points in total, with the train/test ratio of 17:1 -average for three splits-): "fully-dissimilar-split" (predicting new inhibitors for new targets), "dissimilar-compound-split" (predicting novel inhibitors for known targets), and "random-split" (predicting known inhibitors for known targets). We split the training samples into 95/5 train/validation sets with a close range to train/test ratio and used the random search method for the hyperparameter ranges given below with epoch number 5, weight decay 0, and dropout 0.5:

- *hidden channels:* 16, 32, 64, 128
- *learning rate:* 0.0001, 0.0005, 0.001, 0.005
- *head number:* 4, 8, 16, 32
- *layer number:* 1, 2, 3
- *train batch size:* 128, 256, 512, 1024

After determining the hyperparameters, we further fine-tuned the epoch number. The total run time of models ranges from 20 minutes to 5.5 hours, depending on the model architecture and selected hyperparameters. These computations were performed using an NVIDIA RTX A5000 graphic card equipped with 24 GB of GPU memory. The selected hyperparameter sets for each split are as follows (Table 5.10):

Table 5.10. Hyperparameters of the finalized HetCPI models on fully-dissimilar-split (FDS), dissimilar-compound-split (DCS), and random-split (RS) datasets of transferases.

Dataset Split	Epoch Number	Hidden Channels	Learning Rate	Head Number	Layer Number	Batch Size
FDS[1]	20, 40	32	0.0001	8	2	512
FDS[2]	20	32	0.0005	32	2	256
DCS	25, 100, 300	32	0.0001	8	3	512
RS	50, 100, 250, 300, 500	128	0.001	16	1	1024

FDS[1] – 20 epoch: HetCPI_ECFP4, HetCPI_SELFormer

FDS[1] – 40 epoch: DP_ECFP4, HetCPI-3FC_SELFormer, 3FC_SELFormer

FDS[2]: DP_SELFormer, HetCPI-3FC_ECFP4, 3FC_ECFP4

DCS: 3FC_SELFormer (100 epoch), 3FC_ECFP4 (300 epoch), all remaining models (25 epoch)

RS – 50 epoch: HetCPI_ECFP4, HetCPI-3FC_ECFP4

RS – 100 epoch: DP_ECFP4

RS – 250 epoch: HetCPI_SELFormer, HetCPI-3FC_SELFormer

RS – 300 epoch: 3FC_SELFormer, 3FC_ECFP4

RS – 500 epoch: DP_SELFormer

In Figure 5.2, bar plots display model performances based on Spearman and median corrected MCC scores. To obtain error bars and assess the significance of the performance differences between models, we conducted five runs for each model and computed the averages. Although paired t-test is commonly used for model performance comparison, it is not suitable in most cases due to the violation of the independence assumption. To address this, we employed the deep-significance tool, which enables statistical testing for DL models using techniques like Almost Stochastic Order (ASO). ASO compares two score distributions instead of comparing their means and calculates the epsilon minimum (eps_min) value. A model performs better than the other if its eps_min value is less than 0.5, where lower values indicate a more confident result (Ulmer et al., 2022).

As shown in Figure 5.2, HetCPI_SELFormer (i.e., the default HetCPI architecture) consistently outperformed the baseline models across all splits for transferases except for the random split, where RF_ECFP4 model is stochastically dominant over other models (eps_min: 0.0, confidence level: 0.95). In the random-split set, 3FC_ECFP4 baseline model also performed better than HetCPI models, which means that graph-based learning did not provide an advantage over simpler baseline models for easy scenarios. Although the performance of different HetCPI architectures varies across dataset splits, HetCPI_SELFormer outperforms other HetCPI architectures in challenging scenarios (i.e., fully-dissimilar-split and dissimilar-compound-split sets), excluding HetCPI_ECFP4 for the fully-dissimilar-split set. Based on the eps_min score of HetCPI_SELFormer, which is 1 for comparison with HetCPI_ECFP4 Spearman and 0.38 for HetCPI_ECFP4 MCC (median corrected), and 0 for all other comparisons, we can state that HetCPI_SELFormer model is stochastically dominant over other models in fully-dissimilar-split and dissimilar-compound split sets with a confidence level 0.95. Its superior performance compared to DP and 3FC baseline models also provides strong validation for the meaningful learning of representations

through the heterogeneous graph learning approach in challenging scenarios. This indicates that leveraging the rich information encoded in the heterogeneous graph structure leads to improved predictive capabilities.

The usage of ECFP4 compound fingerprints is advantageous for the 3 FC-layered models and the random split set compared to SELFormer embeddings. In fact, SELFormer performed much better in challenging scenarios for the models without 3 FC layers. In the random split, HetCPI models with ECFP4 fingerprints outperformed other HetCPI models but were unable to surpass the baseline RF_ECFP4 and 3FC-ECFP4 model. These results indicate that ECFP4 is more effective at capturing patterns between similar data, while SELFormer is more suitable for identifying deeper patterns that can be helpful in inferring data from structurally different compounds. However, this observation is valid only for HetCPI models and does not provide an advantage in RF models, indicating that advanced algorithms are necessary to fully exploit the potential of SELFormer. The inclusion of 3 FC layers did not lead to an improvement in model performance, which is a usual trend observed in most DL models, as well.

Another significant outcome is the robustness of HetCPI models against mean shifting problem in regression tasks, as they yielded higher scores even without applying median correction (Appendix A Table 5.1). As expected, the challenging dataset splits resulted in relatively lower performance for both the baseline RF and the HetCPI models, compared to the random split scenario. In summary, the impressive performance of HetCPI models in challenging scenarios is encouraging and suggests that these models possess great adaptability to real-world CPI prediction scenarios. This adaptability is a crucial factor for their practical usefulness and applicability.

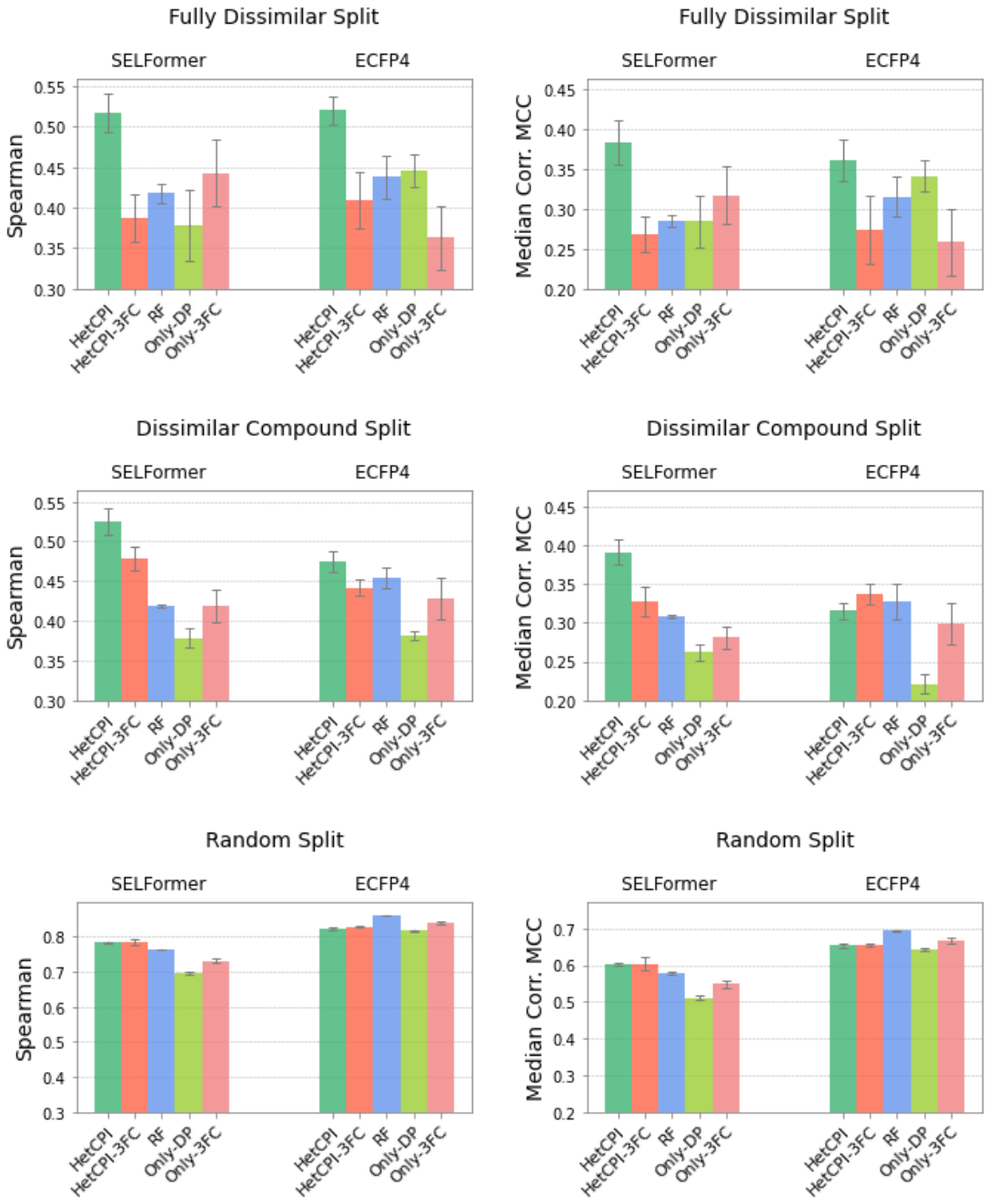


Figure 5.2. Bar plots of Spearman and median corrected MCC scores of HetCPI and baseline models with different architecture alternatives on the dissimilar-compound-split of the transferases test dataset.

5.4.4. Performance Comparison with State-of-the-Art (SOTA) Models

To fairly compare the HetCPI system with the state-of-the-art CPI prediction models in the literature, we used the filtered Davis kinase benchmark dataset (9,125 data points in total, with the train/test ratio of 5:1) with the same settings used in the MDeePred study (Rifaioglu et al., 2021). For the MCC score, instead of taking the median of training bioactivity values as the threshold, three threshold values 1 uM, 100 nM, and 30 nM were used, respectively. In addition to the evaluation metrics above, we

calculated the average area under precision-recall curve (AUPRC) and the concordance index (CI) scores for the comparison of SOTA models. AUPRC measures the trade-off between precision and recall across different decision thresholds by calculating the area under the precision-recall curve using interpolation methods. It ranges from 0 to 1, where 1 indicates perfect precision and recall (Boyd et al., 2013). For computing the average AUPRC score, ten interaction threshold values from the pKd interval [6 M, 8 M] were considered to binarize pKds into true class labels. CI for a set of paired data (i.e., compound-protein pairs) represents the probability that the predictions for two randomly selected pairs are correctly ordered based on their respective labels (i.e., pKd values). It ranges from 0 to 1, where 1 indicates perfect agreement between the predicted and observed rankings, and is used to evaluate the discriminatory power of a predictive model (Gönen & Heller, 2005). The formula used for the calculation of CI is given below:

$$CI = 1/Z \sum_{\gamma i > \gamma j} h(fi - fj), \quad (1)$$

where fi and γi represent the predicted and real binding affinity values for the ith pair, respectively. Z is a normalization constant equal to the number of pairs and h(u) is a step function, which returns 1.0, 0.5 and 0.0 for $u > 0$, $u = 0$ and $u < 0$, respectively.

The selected hyperparameters after model optimization are as follows:

- *hidden channels:* 128,
- *train batch size:* 256
- *weight decay:* 0.001,
- *epoch number:* 800,
- *learning rate:* 0.0001,
- *head number:* 64,
- *layer number:* 2,
- *dropout:* 0.1

Based on the findings presented in Table 5.11, the HetCPI-3FC_ECFP4 model demonstrates strong competitiveness with SOTA models for bioactivity prediction. It performs exceptionally well in CI and MCC (100 nM) scores and achieves scores comparable to the best-performing methods in other metrics. The performance scores of different HetCPI model architectures align with the random-split case of transferases. However, it is important to note that the Davis kinase benchmark dataset is a medium-scale dataset with significantly fewer datapoints compared our family-specific large-scale datasets. Therefore, it may not be representative enough, despite its common usage in DTI prediction studies. Additionally, the dataset is split randomly, which make the prediction task easier and can lead to overoptimistic results due to a lack of generalizability. Consequently, while the HetCPI models yield competitive or even superior performance results compared to other SOTA models, we believe that the Davis dataset and other widely used benchmark datasets for DTI prediction may not provide results that accurately reflect real-world scenarios. Thus, they may not be suitable for addressing this complex biological problem comprehensively.

Overall, the results highlight the potential of HetCPI as a KG-based learning approach. By utilizing the rich structural and contextual information encoded in KGs, HetCPI demonstrates its capability to offer meaningful insights and accurate predictions for compound-protein bioactivities.

Table 5.11. Performance comparison with SOTA models on the filtered Davis dataset. Standard deviations are given in parentheses.

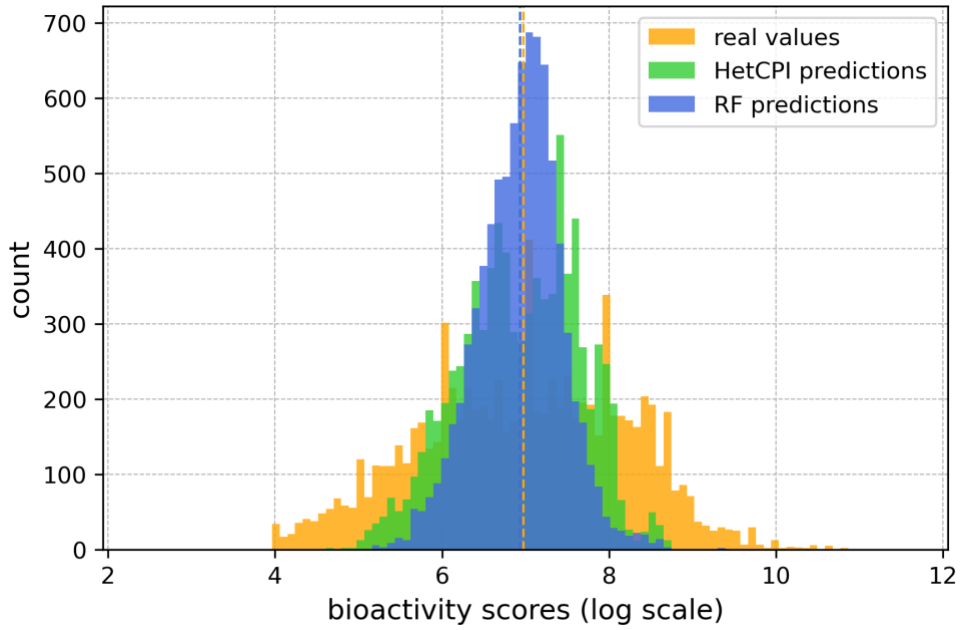
Method	**CI**	**RMSE**	**Spearman**	**Average AUPRC**	**MCC* (1 uM)**	**MCC (100 nM)**	**MCC (30 nM)**
HetCPI-3FC _ECFP4	**0.744 (0.012)**	0.702 (0.017)	0.649 (0.024)	0.776 (0.006)	0.41 (0.055)	**0.59 (0.013)**	0.58 (0.01)
HetCPI-3FC _SELFormer	0.728 (0.008)	0.746 (0.022)	0.615 (0.019)	0.749 (0.007)	0.35 (0.037)	0.561 (0.018)	0.562 (0.024)
HetCPI_ECFP4	0.707 (0.007)	0.895 (0.026)	0.56 (0.024)	0.69 (0.019)	0.347 (0.015)	0.5 (0.022)	0.526 (0.015)
HetCPI _SELFormer	0.706 (0.005)	0.863 (0.019)	0.554 (0.014)	0.699 (0.009)	0.343 (0.038)	0.501 (0.023)	0.517 (0.033)
MGraphDTA	0.74 (0.002)	**0.695 (0.009)**	**0.654 (0.005)**	-	-	-	-
MDeePred	0.733 (0.004)	0.742 (0.009)	0.618 (0.009)	**0.803 (0.006)**	**0.424 (0.014)**	0.572 (0.011)	0.585 (0.01)
CGKronRLS	0.74 (0.003)	0.769 (0.01)	0.643 (0.008)	0.773 (0.01)	0.422 (0.009)	0.564 (0.016)	**0.617 (0.029)**
DeepDTA	0.653 (0.005)	0.931 (0.015)	0.430 (0.013)	0.529 (0.018)	0.229 (0.051)	0.298 (0.04)	0.208 (0.035)

* The MCC classification metric was calculated by binarizing the predictions as active and inactive at the thresholds of 1 uM, 100 nM, and 30 nM.

5.4.5. Exploring the Predictive Power of HetCPI for Extreme Values

Predicting extreme values accurately can be quite challenging when using regression models, as these models typically focus on capturing overall trends and patterns in data. In this subsection, we examined how well the HetCPI model handles these values and evaluated its potential applicability in scenarios where extreme values are particularly important. We compared the predicted bioactivity distributions (i.e., median corrected) of HetCPI and baseline RF model with the real distribution of test data points in the dissimilar-compound-split set of transferases. Figure 5.3 demonstrates that the HetCPI distribution bears a closer resemblance to the actual distribution in comparison to the RF model. In the RF model, most predictions are concentrated around the mean value of the training dataset (i.e., 6.94), resulting in a narrower range between 5 and 9. Although HetCPI also exhibits a similar range of distribution, it is more balanced with a higher number of predictions in ranges of 5-6 and 8-9. However, both models struggle to predict values below 5 and above 9, indicating their limited capability to predict extreme values in challenging scenarios.

(a)



(b)

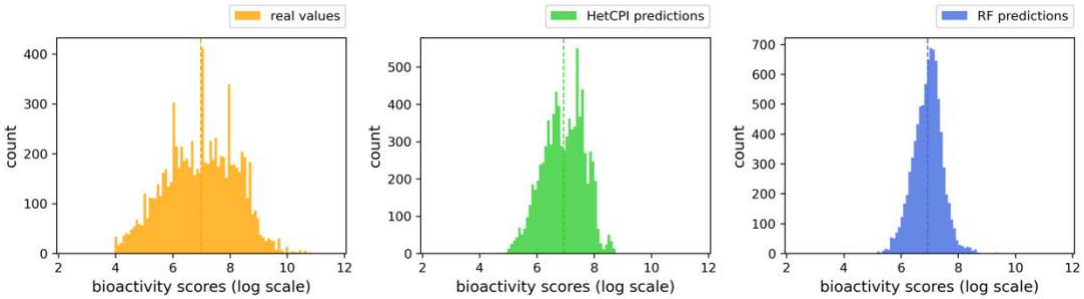


Figure 5.3. Bioactivity distributions of real values, HetCPI predictions, and RF predictions, **(a)** both collectively and **(b)** separately, for test data points in the dissimilar-compound-split set of transferases.

To quantitatively analyze the results shown in Figure 5.3, we calculated the percentage of data falling within specific thresholds and computed Spearman scores for the samples in those ranges. We also determined the percentages of true positives (TPs) for thresholds above 6.97 (i.e., the median value of the training dataset as the active/inactive cut-off), and true negatives (TNs) for thresholds below 6.97. As displayed in Table 5.12, the extreme bioactivity values within the 5-6 and 8-9 ranges correspond to approximately 15-19% of the experimental data, whereas the predicted values for these ranges are about 1-11%. The percentages drop to 3-5% for bioactivities below 5 and above 9 in the experimental data, while predictions have almost zero occurrences in these extreme ranges. Spearman performance scores of HetCPI and RF prediction models are also considerably low for these extreme ranges, where even slight deviations from real values greatly impact performance due to the limited data. These findings reveal the limitations of the models in accurately

predicting extreme values for regression tasks. Nevertheless, the notable percentages of TPs and TNs, especially for HetCPI, indicate their ability to correctly classify extreme values in binary classification. Although precise predictions in challenging cases are not the primary objective, employing data preprocessing techniques such as oversampling strategies can enhance the success rate of predicting extreme values by transforming imbalanced regression distributions into distributions that approximate a more uniform-like distribution.

Table 5.12. Evaluation of HetCPI and RF model predictions for extreme bioactivity values in the dissimilar-compound-split of the transferases test set.

Bioactivity threshold	Real values (%)	HetCPI predictions (%)	RF predictions (%)	Spearman		HetCPI (%) *		RF (%)	
				HetCPI	RF	TP	TN	TP	TN
<5	5.81	0.14	0.0	0.059	0.134	-	85.0	-	80.7
5-6	15.85	10.78	3.89	0.12	0.096	-	78.7	-	70.0
8-9	18.75	5.26	1.75	0.141	0.094	86.0	-	77.3	-
9<	3.15	0.0	0.02	0.066	0.252	85.2	-	78.0	-

*Active/inactive cutoff: 6.97 (median value of the training dataset)

5.4.6. Use-Case Study: Evaluation of Bioactivity Predictions for Druggable and Undruggable Proteins

To further evaluate the robustness and reliability of HetCPI, we conducted a use-case study focusing on the predictions for druggable and undruggable proteins. As an example of a druggable protein, we selected PIM1 (Figure 5.4a). PIM1 is a proto-oncogene with serine/threonine kinase activity that plays a vital role in cell growth, survival, and apoptosis. It exerts its oncogenic effects through the regulation of MYC transcriptional activity, control of cell cycle progression, and inhibition of proapoptotic proteins via phosphorylation. Notably, abnormal elevation of PIM1 is associated with various types of cancer. It is a promising cancer drug target, particularly in prostate cancer (Tursynbay et al., 2016). Figure 5.4a displays its co-crystalized structure with a benzofuranone inhibitor (PDB ID: 5VUB). In our transferases training dataset (i.e., 138,297 data points in the dissimilar-compound-split set), PIM1 has 3,019 data points with a mean bioactivity value of 7.7. The PDB structure in Figure 5.4b belongs to the HER3 protein (PDB ID: 6OP9). HER3, also known as ERBB3, is a pseudo-kinase member of the EGFR family having a significant role in tumor progression and drug resistance. Unlike other kinases, its pseudo-kinase domain activates its partner HER family members, including HER2 and EGFR, through allosteric regulation. Despite its significance as a therapeutic target in various tumors, no HER3-directed therapies have received clinical approval so far (Haikala & Jänne, 2021). It has only 23 data points with a mean bioactivity value of 5.86 in our training dataset, which is very low compared to PIM1.

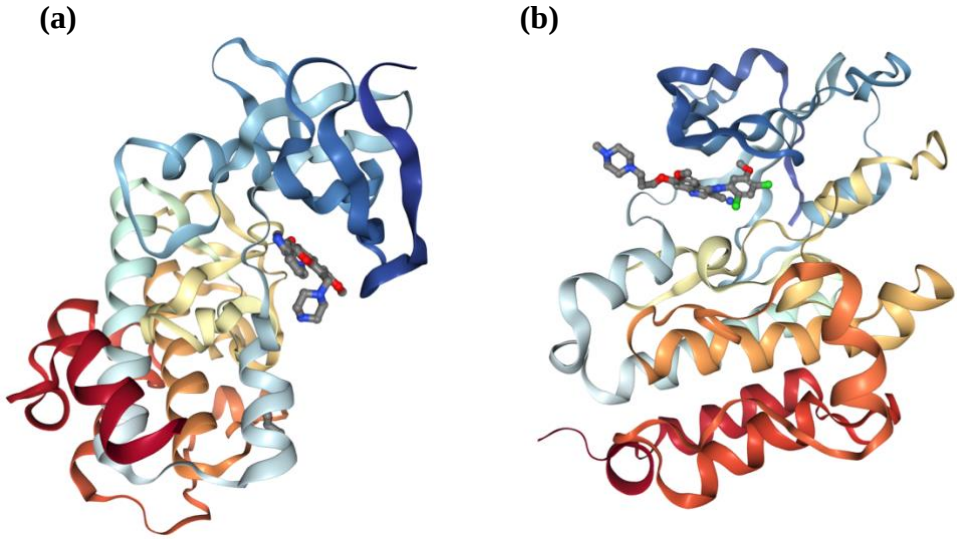


Figure 5.4. Co-crystalized 3-D structures of **(a)** PIM1 in complex with a benzofuranone inhibitor and **(b)** HER3 in complex with bosutinib.

5.4.6.1. Comparing distributions of predicted bioactivities and prediction differences of druggable protein PIM1 and undruggable protein HER3

We assessed the bioactivity predictions of PIM1 and HER3 against 422,617 compounds in our KG and visualized their distributions, with the horizontal axis representing the normalized scores based on the mean bioactivity value of the training dataset (i.e., 6.94 -in log-scale-), and the vertical axis representing the counts (Figure 5.5a). We also plotted the distribution of prediction differences of PIM1 and HER3 for each compound (Figure 5.5b). Figure 5.4a demonstrates a significant difference in the distributions of the two proteins (t-test score: -1518, p-value: 0.0). HER3 exhibits a considerably lower mean score compared to PIM1. Conversely, PIM1 demonstrates the opposite pattern, with a higher mean score. This discrepancy accurately reflects the druggability states of these well-known protein examples and underscores the discriminative capability of the HetCPI system. However, the predictions for HER3 also include a small number of active compounds with bioactivity values higher than the mean bioactivity value of the training dataset (i.e., 1,508 data points, 0.36% of all compounds). While it is possible that some of these predictions may have high errors due to the imperfect nature of prediction models, it is also worth considering the possibility that some of these compounds could represent previously unknown bioactivities of HER3. There have been instances such as KRAS protein, which was initially considered undruggable but later successfully targeted with specific compound types (L. Huang et al., 2021). Therefore, active predictions for HER3 may also hold potential significance and merit further exploration through experimental investigations.

Figure 5.5a also reveals a remarkably similar distribution shape for PIM1 and HER3 proteins, with variations primarily observed in the ranges of the distributions. Although the differences between per-compound predicted bioactivities for two proteins do not follow a constant or uniform pattern (Figure 5.5b), ruling out the possibility of a technical issue or trivial predictions without meaningful learning, its

resemblance to the distributions in Figure 5.5a suggests a correlation between compound features and predictions that consistently yields similar rankings regardless of protein features, which appear to influence the values of bioactivity predictions. This situation may arise due to inherent limitations of the model, such as excessive convergence of protein and compound features due to the graph-based learning process, making it challenging to learn discriminative patterns effectively. Another limitation may be due to the natural bias present in experimental bioactivity data. Screening assays tend to be target-centric, resulting in protein-compound pairs being predominantly diversified by the compounds. Consequently, prediction models lean towards learning primarily from the compound side. To gain deeper insights into this phenomenon, further analyses can be conducted, such as exploring the learned features of the model and incorporating an interpretability module to understand the model behavior. Additionally, strategies like weighting protein features to enhance their influence during the learning process or investigating binding-specific protein characteristics that potentially contribute to the predictions could be explored. These approaches can help uncover the underlying factors driving the observed similarity in bioactivity prediction distribution shapes for these distinct proteins, leading to more robust and accurate prediction models.

(a) **(b)**

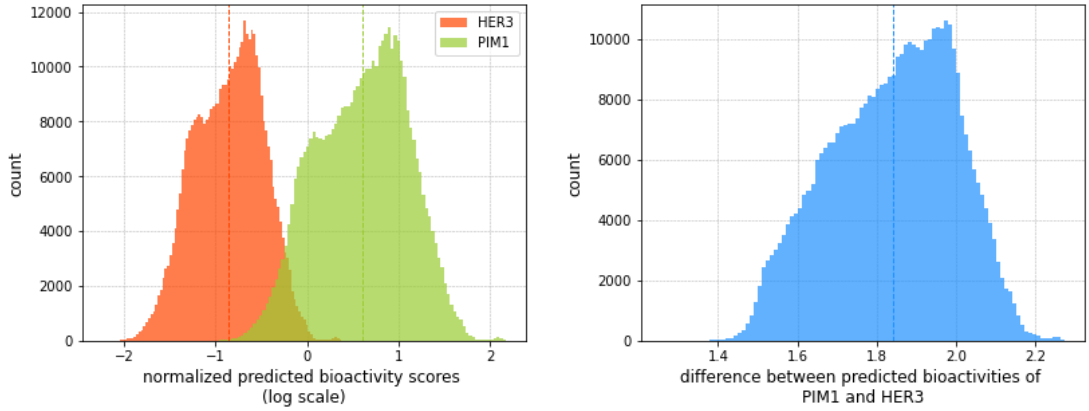


Figure 5.5. Histograms of **(a)** normalized predicted bioactivity scores in log scale, and **(b)** prediction differences, for the undruggable protein HER3 and the druggable protein PIM1.

5.4.6.2. Compound-centric analysis of PIM1 predictions

We performed additional analyses on PIM1 predictions for 3,019 compounds having measured bioactivity values for PIM1 in the training dataset of transferases on dissimilar-compound-split. The model performance on these data samples yielded the following scores: 1.11 (RMSE), 0.75 (Spearman), and 0.36 (MCC) for HetCPI; 0.31 (RMSE), 0.99 (Spearman) and 0.88 (MCC) for baseline RF model. While the number of test data points for PIM1 is limited, we also computed the performance scores for these 36 test data points to assess models' behaviors on both seen and unseen samples. The results for the test data were as follows: 1.53 (RMSE), 0.74 (Spearman), and 0.20 (MCC) for HetCPI; 1.62 (RMSE), 0.47 (Spearman), and 0.31 (MCC) for RF. HetCPI

predictions exhibit closer proximity to the real bioactivity values compared to RF predictions for the test data (i.e., better RMSE and Spearman scores). However, when evaluating the predictions in binary class format, there is only a one-sample difference between the models for 36 data points. This slight difference leads to a higher MCC score for the RF model, highlighting the sensitivity of MCC measurement to even minor variations when dealing with a limited number of samples. Considering both the training and test performance scores of the models for PIM1, we can infer that HetCPI demonstrates more consistent results between the training and test data compared to the baseline RF model. The notably high performance scores of the RF model on the training dataset, followed by a significant drop in performance on the test samples may indicate a potential issue of memorizing the training data.

To explore whether specific groups of compounds exhibit higher or lower performance, we also performed clustering on the 3,019 compounds based on Tanimoto similarities using the Butina clustering function of RDKit, setting a cut-off of 0.5. Figure 5.6 displays the prediction error box plots of clusters with at least five members of 3,019 compounds. To calculate prediction errors, we computed the differences between the predicted and real bioactivity values of each sample, expressed as pChEMBL units. In this plot, blue bars represent clusters of compounds with overestimations, while red bars correspond to clusters with underestimations. Upon analyzing the median prediction errors of compound clusters, it was observed that the majority of clusters exhibited prediction errors within the range of -1 and 1. Additionally, there was a higher number of overestimations compared to underestimations across the clusters.

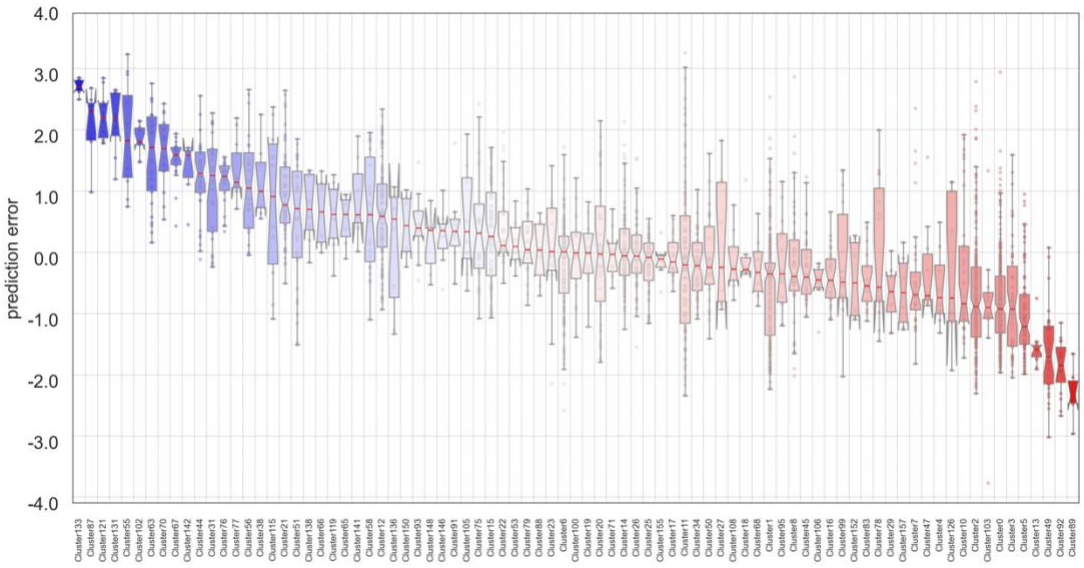
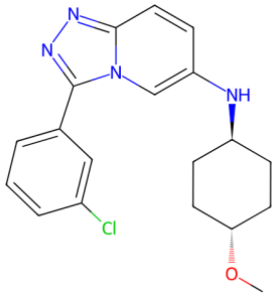
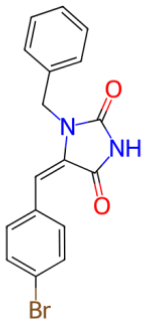


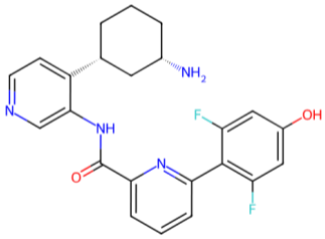
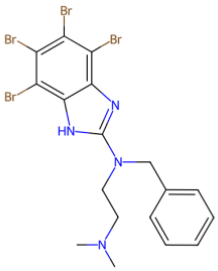
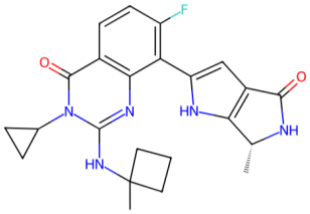
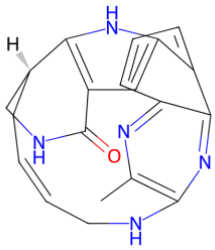
Figure 5.6. Prediction error box plots of compound clusters having bioactivity measurements for PIM1 protein in the dissimilar-compound-split of the transferases training dataset.

To examine these clusters more closely, we selected one representative compound structure from each cluster, including Cluster 87 and 133 (over-estimation), Cluster 17 and 155 (close-estimation), and Cluster 89 and 92 (under-estimation). Table 5.13

displays these compounds along with their cluster groups, and experimental and predicted bioactivities, to exemplify the specific characteristics of their respective clusters and provide a clearer understanding of the over-estimation, close-estimation, and under-estimation patterns observed. One possible explanation for the over-estimated compounds in Cluster 87 and 133 could be the presence of halogens (e.g., Br, Cl, F) in their structures. Halogens are commonly used substituents to enhance the potency of compounds against their protein targets due to their electronegativities (Wilcken et al., 2013). Since these halogen-containing groups are prevalent in many bioactive compounds, the model may have learned to associate these groups with increased bioactivity, resulting in higher predicted values for these compounds. On the other hand, the under-estimated compound groups can be attributed to the limitations of the model in predicting extreme values, as discussed in Section 5.4.5. Although the predicted values of these samples are considerably high, their experimental values are even higher. The structures of these compounds are less familiar, and their high bioactivity values are relatively rare in the training dataset, falling in the extreme range. This scarcity makes it challenging for the model to learn the discriminative properties specific to these compounds and accurately predict their bioactivities.

Table 5.13. Representative compound structures of over-, close- and under-estimated clusters along with experimental and predicted bioactivities.

Over-estimation	Cluster 87 (7 members)	Experimental bioactivity: 6.14 Predicted bioactivity: 8.05	 CHEMBL3911729
Over-estimation	Cluster 133 (7 members)	Experimental bioactivity: 4.24 Predicted bioactivity: 6.96	 CHEMBL522916

Close-estimation	Cluster 17 (18 members)	Experimental bioactivity: 9.0 Predicted bioactivity: 8.79	 CHEMBL3651904
Close-estimation	Cluster 155 (6 members)	Experimental bioactivity: 7.2 Predicted bioactivity: 7.26	 CHEMBL3684971
Under-estimation	Cluster 89 (6 members)	Experimental bioactivity: 10.15 Predicted bioactivity: 7.97	 CHEMBL3944421
Under-estimation	Cluster 92 (15 members)	Experimental bioactivity: 10.22 Predicted bioactivity: 7.64	 CHEMBL3810121

5.5. Conclusion and Future Directions

In this chapter, we introduced HetCPI, a novel framework for compound-protein interaction (CPI) representation and prediction. By leveraging large-scale biomedical knowledge graphs constructed by the CROssBAR system, HetCPI effectively captures the complex relationships between genes/proteins, pathways, diseases, phenotypes, drugs, and compounds. The heterogeneous graph transformer (HGT) architecture is employed to learn from these diverse biomedical relationships and generate integrative representations for CPI prediction that carries rich structural and contextual information encoded in KGs. Our benchmarking experiments on target protein family-

specific bioactivity datasets demonstrate the superiority of HetCPI over baseline models, showcasing its adaptability and potential for realistic CPI prediction scenarios. Furthermore, the competitive performance of HetCPI against state-of-the-art models in bioactivity prediction validates its effectiveness as a KG-based learning approach. The outcomes of the use-case study on the predictions of druggable (PIM1) and undruggable proteins (HER3) with a significant difference in predicted bioactivity score distributions further support its reliability and robustness. Overall, HetCPI represents a significant advancement in the use of KGs and graph neural networks for virtual screening with the potential to contribute to the discovery of new and effective treatments for various diseases. We plan to generate predictions also for other human proteins and openly share HetCPI as a programmatic tool with the research community, which would foster collaboration and enable further exploration and refinement of the method.

There are several promising avenues for further exploration and development. Firstly, expanding the application of HetCPI to additional protein families and disease areas holds great potential. By encompassing a broader range of targets and therapeutic areas, HetCPI can contribute to the discovery of novel drug candidates and facilitate personalized medicine approaches. Secondly, the integration of additional data sources into the KGs could further enhance the predictive power of HetCPI. Integrating new types of nodes and edges, such as cell-line information including drug sensitivity measurements, gene expression profiles, mutation annotations, biological ontologies including GO term annotations for molecular function, biological process, and cellular component, side effects and toxicity profiles of drug candidate compounds, and metabolomics data would provide a richer representation of diverse biological and chemical information. This way, inferences about preclinical and clinical study results can be provided, as well. Furthermore, the incorporation of an iterative active learning process during model training could facilitate more efficient and targeted data acquisition, leading to better model performance.

CHAPTER 6

CONCLUSION

This thesis study addresses the challenges and limitations in computational drug discovery and proposes innovative solutions to enhance the predictive capabilities of AI models in DTI/CPI prediction. The main objective is to contribute to the effective utilization of AI in drug discovery by developing robust and industry-applicable state-of-the-art CPI prediction models.

The study tackles several key problems concerning *in silico* prediction of DTIs. Primarily, it focuses on the limitations of existing AI models, including biases and incompleteness in training data. To overcome these limitations, the study introduces gold-standard protein family-specific benchmark datasets on a large scale that provide high-quality and diverse samples, along with realistic train/test split scenarios. This enables more reliable and reproducible evaluations of different DTI prediction models and promotes the development of accurate and applicable systems. The study offers a comprehensive perspective on the protein aspect of DTI prediction, which is often overshadowed by the focus on ligands, by exploring the representation capability of different protein featurization techniques. It demonstrates the potential of learned protein representations for widespread utilization in bioactivity modeling with competitive performance results compared to classical featurization approaches. Additionally, the study emphasizes the significance of dataset splitting for conducting realistic evaluations for drug and/or target discovery. It uncovers the limited ability of traditional ML algorithms to extrapolate data, as indicated by their lower performance on challenging dataset splits.

Another issue addressed in this thesis is the neglect of multi-layered heterogeneous data in current AI models for DTI prediction. The interactions between ligands and proteins are highly complicated and dependent on molecular and cellular activities. To gain a deeper understanding of this complexity, it is important to consider the meta-relations of proteins and compounds with other biomedical entities, going beyond similarity measurements and molecular properties. To effectively utilize this diverse data, we carried out the knowledge graph (KG) construction procedure of CROssBAR system, which comprises integrating large-scale biological/biomedical data from open-access data repositories and representing them in the form of heterogeneous and computable KGs. CROssBAR KGs provide sub-networks of interconnected entities including genes/proteins, diseases/phenotypes, biological processes/pathways, and

drugs/compounds. These entities are linked through a variety of relationships such as protein-protein interactions, protein-pathway associations, drug-disease indications, and disease-phenotype associations. CROssBAR KGs facilitate in-depth analysis of complex systems data, empowering researchers to address a wide range of biological challenges including the identification of compound-target interactions.

This study also presents a novel systems-level approach called HetCPI, designed for compound-protein interaction (CPI) prediction. By utilizing cutting-edge heterogeneous graph learning algorithms and processing information embedded in the CROssBAR KGs, HetCPI demonstrates remarkable improvements in bioactivity prediction, particularly in challenging scenarios. It highlights the generalizability of HetCPI with improved predictions for previously unseen data, addressing one of the major bottlenecks in drug discovery. This framework successfully extracts hidden knowledge from multi-layered biomedical data and offers potential advancements in drug discovery.

In summary, this thesis study contributes to computational drug discovery and biomedical research by providing gold-standard benchmark datasets, constructing biomedical data representations in the form of KGs, and developing innovative predictive models for compound-protein bioactivities. These contributions aim to enhance the ongoing work in computational drug discovery, ultimately facilitating the discovery of new drugs and advancing medical science.

Despite achieving an improvement in DTI/CPI prediction and making significant contributions to the literature, this study has certain limitations, as well. Firstly, there is still room for enhancing model performance, particularly in challenging cases. This suggests that further refinements and optimizations are possible to achieve more accurate predictions. Secondly, due to technical constraints, it was not feasible to include all available biomedical data sources (e.g., transcriptomic and metabolomic data, clinical information, gene mutations and annotations, protein domain profiles, etc.) leading to potential limitations in dataset coverage and, consequently, influencing overall performance. Data collection is challenging due to factors such as data availability, accessibility, and the necessity for specialized expertise, making it difficult to achieve a fully comprehensive dataset. Lastly, computational requirements pose a constraint on the study. As the complexity of the data increases, it becomes necessary to allocate additional computational resources to effectively search and optimize an extensive range of model parameters. Balancing computational efficiency with data complexity remains a critical consideration for future improvements in the field.

Moving forward, future directions for this study could involve further exploration and expansion of the application of heterogeneous graph-based AI models. One crucial aspect to focus on is improving the interpretability of the proposed methods to unravel the underlying mechanisms and features contributing to the decision-making process of the models. To address the inherent black-box nature of DL, integrating explainability modules into graph learning architectures holds promise and is currently a topic of considerable research attention. This step will enable researchers to

understand the rationale behind model predictions, building trust in the decisions. Another noteworthy area of development in AI is the rapid advancement of large language models (LLMs) and their integration into various fields. As an implementation of LLMs in biomedical research, ChatGSE (Lobentanzer & Saez-Rodriguez, 2023) demonstrates the potential of combining human ingenuity with machine memory through an open and modular conversational platform to enhance biomedical analyses. Developing a similar platform tailored specifically for drug discovery can facilitate the entire drug development process. By integrating safety and toxicity pipelines into this platform, it can offer further filtration of bioactive compounds. This integration would decrease the failure rates of hit compounds in subsequent stages of drug development, contributing to the identification of more efficient and safer drug candidates. Overall, the application of AI in drug discovery holds great potential, and further advancements in the field will undoubtedly contribute to the development of effective and safe therapeutics.

REFERENCES

Abbasi, K., Razzaghi, P., Poso, A., Amanlou, M., Ghasemi, J. B., & Masoudi-Nejad, A. (2020). DeepCDA: deep cross-domain compound–protein affinity prediction through LSTM and convolutional neural networks. *Bioinformatics*, *36*(17), 4633–4642. https://doi.org/10.1093/BIOINFORMATICS/BTAA544

Ain, Q. U., Méndez-Lucio, O., Ciriano, I. C., Malliavin, T., van Westen, G. J. P., & Bender, A. (2014). Modelling ligand selectivity of serine proteases using integrative proteochemometric approaches improves model performance and allows the multi-target dependent interpretation of features. *Integrative Biology*, *6*(11), 1023–1033. https://doi.org/10.1039/C4IB00175C

Ali, M., Hoyt, C. T., Domingo-Fernández, D., Lehmann, J., & Jabeen, H. (2019). BioKEEN: a library for learning and evaluating biological knowledge graph embeddings. *Bioinformatics (Oxford, England)*, *35*(18), 3538–3540. https://doi.org/10.1093/BIOINFORMATICS/BTZ117

Alley, E. C., Khimulya, G., Biswas, S., AlQuraishi, M., & Church, G. M. (2019). Unified rational protein engineering with sequence-based deep representation learning. *Nature Methods*, *16*(12), 1315–1322. https://doi.org/10.1038/s41592-019-0598-1

Andonie, R. (2019). Hyperparameter optimization in learning systems. *Journal of Membrane Computing*, *1*(4), 279–291. https://doi.org/10.1007/S41965-019-00023-0/METRICS

Antezana, E., Blondé, W., Egaña, M., Rutherford, A., Stevens, R., De Baets, B., Mironov, V., & Kuiper, M. (2009). BioGateway: A semantic systems biology tool for the life sciences. *BMC Bioinformatics*, *10*(SUPPL. 10), 1–15. https://doi.org/10.1186/1471-2105-10-S10-S11/FIGURES/7

Asgari, E., & Mofrad, M. R. K. (2015). Continuous distributed representation of biological sequences for deep proteomics and genomics. *PLoS ONE*, *10*(11), 141287. https://doi.org/10.1371/journal.pone.0141287

Bakal, G., Talari, P., Kakani, E. V., & Kavuluru, R. (2018). Exploiting Semantic Patterns over Biomedical Knowledge Graphs for Predicting Treatment and Causative Relations. *Journal of Biomedical Informatics*, *82*, 189. https://doi.org/10.1016/J.JBI.2018.05.003

Ballester, P. J., & Mitchell, J. B. O. (2010). A machine learning approach to predicting protein-ligand binding affinity with applications to molecular docking. *Bioinformatics (Oxford, England)*, *26*(9), 1169. https://doi.org/10.1093/BIOINFORMATICS/BTQ112

Blass, B. E. (2015). Drug Discovery and Development: An Overview of Modern Methods and Principles. *Basic Principles of Drug Discovery and Development*, 1–34. https://doi.org/10.1016/B978-0-12-411508-8.00001-3

Bleakley, K., & Yamanishi, Y. (2009). Supervised prediction of drug–target interactions using bipartite local models. *Bioinformatics*, *25*(18), 2397. https://doi.org/10.1093/BIOINFORMATICS/BTP433

Blondel, V. D., Guillaume, J. L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, *2008*(10), P10008. https://doi.org/10.1088/1742-5468/2008/10/P10008

Bosc, N., Atkinson, F., Felix, E., Gaulton, A., Hersey, A., & Leach, A. R. (2019). Large scale comparison of QSAR and conformal prediction methods and their applications in drug discovery. *Journal of Cheminformatics*, *11*(1), 1–16. https://doi.org/10.1186/S13321-018-0325-4/TABLES/6

Bougiatiotis, K., Aisopos, F., Nentidis, A., Krithara, A., & Paliouras, G. (2020). Drug-Drug Interaction Prediction on a Biomedical Literature Knowledge Graph. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *12299 LNAI*, 122–132. https://doi.org/10.1007/978-3-030-59137-3_12/TABLES/3

Boyd, K., Eng, K. H., & Page, C. D. (2013). Area under the precision-recall curve: Point estimates and confidence intervals. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *8190 LNAI*(PART 3), 451–466. https://doi.org/10.1007/978-3-642-40994-3_29/COVER

Buniello, A., Macarthur, J. A. L., Cerezo, M., Harris, L. W., Hayhurst, J., Malangone, C., McMahon, A., Morales, J., Mountjoy, E., Sollis, E., Suveges, D., Vrousgou, O., Whetzel, P. L., Amode, R., Guillen, J. A., Riat, H. S., Trevanion, S. J., Hall, P., Junkins, H., … Parkinson, H. (2019). The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Research*, *47*(D1), D1005–D1012. https://doi.org/10.1093/nar/gky1120

Buza, K., & Peška, L. (2017). Drug–target interaction prediction with Bipartite Local Models and hubness-aware regression. *Neurocomputing*, *260*, 284–293. https://doi.org/10.1016/J.NEUCOM.2017.04.055

Carvalho-Silva, D., Pierleoni, A., Pignatelli, M., Ong, C., Fumis, L., Karamanis, N., Carmona, M., Faulconbridge, A., Hercules, A., McAuley, E., Miranda, A., Peat, G., Spitzer, M., Barrett, J., Hulcoop, D. G., Papa, E., Koscielny, G., & Dunham, I. (2019). Open Targets Platform: new developments and updates two years on. *Nucleic Acids Research*, *47*(D1), D1056–D1065. https://doi.org/10.1093/nar/gky1133

Cereto-Massagué, A., Ojeda, M. J., Valls, C., Mulero, M., Garcia-Vallvé, S., & Pujadas, G. (2015). Molecular fingerprint similarity search in virtual screening. *Methods*, *71*(C), 58–63. https://doi.org/10.1016/j.ymeth.2014.08.005

Chandak, P., Huang, K., & Zitnik, M. (2022). Building a Knowledge Graph to Enable Precision Medicine. *BioRxiv*, 2022.05.01.489928. https://doi.org/10.1101/2022.05.01.489928

Chen, B., Dong, X., Jiao, D., Wang, H., Zhu, Q., Ding, Y., & Wild, D. J. (2010). Chem2Bio2RDF: A semantic framework for linking and data mining chemogenomic and systems chemical biology data. *BMC Bioinformatics*, *11*(1), 1–13. https://doi.org/10.1186/1471-2105-11-255/TABLES/5

Chen, H., Engkvist, O., Wang, Y., Olivecrona, M., & Blaschke, T. (2018). The rise of deep learning in drug discovery. *Drug Discovery Today*, *23*(6), 1241–1250. https://doi.org/10.1016/J.DRUDIS.2018.01.039

Chen, X.-H., Ruan, Y., Liu, Y.-G., Duan, X.-Y., Jiang, F., Tang, H., Zhang, H.-Y., & Zhang, Q.-Y. (2023). Transporter proteins knowledge graph construction and its application in drug development. *Computational and Structural Biotechnology Journal*, *21*, 2973–2984. https://doi.org/10.1016/J.CSBJ.2023.05.001

Chen, Z., Zhao, P., Li, F., Leier, A., Marquez-Lago, T. T., Wang, Y., Webb, G. I., Smith, A. I., Daly, R. J., Chou, K.-C., & Song, J. (2018). iFeature: a Python package and web server for features extraction and selection from protein and peptide sequences. *Bioinformatics*, *34*(14), 2499–2502. https://doi.org/10.1093/bioinformatics/bty140

Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, *21*(1), 1–13. https://doi.org/10.1186/s12864-019-6413-7

Chithrananda, S., & Ramsundar, B. (2020). ChemBERTa: Large-Scale Self-Supervised Pretraining for Molecular Property Prediction. *ArXiV*.

Chou, K.-C. (2000). Prediction of Protein Subcellular Locations by Incorporating Quasi-Sequence-Order Effect. *Biochemical and Biophysical Research Communications*, *278*(2), 477–483. https://doi.org/10.1006/bbrc.2000.3815

Chou, K.-C. (2005). Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. *Bioinformatics*, *21*(1), 10–19. https://doi.org/10.1093/bioinformatics/bth466

Cichońska, A., Ravikumar, B., Allaway, R. J., Wan, F., Park, S., Isayev, O., Li, S., Mason, M., Lamb, A., Tanoli, Z., Jeon, M., Kim, S., Popova, M., Capuzzi, S., Zeng, J., Dang, K., Koytiger, G., Kang, J., Wells, C. I., … Aittokallio, T. (2021). Crowdsourced mapping of unexplored target space of kinase inhibitors. *Nature Communications 2021 12:1*, *12*(1), 1–18. https://doi.org/10.1038/s41467-021-23165-1

Cichonska, A., Ravikumar, B., Parri, E., Timonen, S., Pahikkala, T., Airola, A., Wennerberg, K., Rousu, J., & Aittokallio, T. (2017). Computational-experimental approach to drug-target interaction mapping: A case study on kinase

inhibitors. *PLOS Computational Biology*, *13*(8), e1005678. https://doi.org/10.1371/JOURNAL.PCBI.1005678

Ciray, F., & Doğan, T. (2022). Machine learning-based prediction of drug approvals using molecular, physicochemical, clinical trial, and patent-related features. *Https://Doi.Org/10.1080/17460441.2023.2153830*, *17*(12), 1425–1441. https://doi.org/10.1080/17460441.2023.2153830

Cortés-Ciriano, I., Ain, Q. U., Subramanian, V., Lenselink, E. B., Méndez-Lucio, O., IJzerman, A. P., Wohlfahrt, G., Prusis, P., Malliavin, T. E., van Westen, G. J. P., & Bender, A. (2015). Polypharmacology modelling using proteochemometrics (PCM): recent methodological developments, applications to target families, and future prospects. *MedChemComm*, *6*(1), 24–50. https://doi.org/10.1039/C4MD00216D

Cortes-Ciriano, I., Van Westen, G. J. P., Lenselink, E. B., Murrell, D. S., Bender, A., & Malliavin, T. (2014). Proteochemometric modeling in a Bayesian framework. *Journal of Cheminformatics*, *6*(1), 1–16. https://doi.org/10.1186/1758-2946-6-35/FIGURES/6

Dalke, A. (2019). The chemfp project. *Journal of Cheminformatics 2019 11:1*, *11*(1), 1–21. https://doi.org/10.1186/S13321-019-0398-8

Dalkiran, A., Rifaioglu, A. S., Martin, M. J., Cetin-Atalay, R., Atalay, V., & Doğan, T. (2018). ECPred: A tool for the prediction of the enzymatic functions of protein sequences based on the EC nomenclature. *BMC Bioinformatics*, *19*(1), 1–13. https://doi.org/10.1186/S12859-018-2368-Y/TABLES/14

Darrell, T., Kloft, M., Pontil, M., Rätsch, G., & Rodner, E. (2015). Machine Learning with Interdependent and Non-identically Distributed Data (Dagstuhl Seminar 15152). *Dagstuhl Reports*. https://doi.org/10.4230/DAGREP.5.4.18

Davis, M. I., Hunt, J. P., Herrgard, S., Ciceri, P., Wodicka, L. M., Pallares, G., Hocker, M., Treiber, D. K., & Zarrinkar, P. P. (2011). Comprehensive analysis of kinase inhibitor selectivity. *Nature Biotechnology*, *29*(11), 1046–1051. https://doi.org/10.1038/nbt.1990

Dharani, G., Nair, N. G., Satpathy, P., & Christopher, J. (2019). Covariate Shift: A Review and Analysis on Classifiers. *2019 Global Conference for Advancement in Technology, GCAT 2019*. https://doi.org/10.1109/GCAT47503.2019.8978471

Doğan, T. (2018). HPO2GO: prediction of human phenotype ontology term associations for proteins using cross ontology annotation co-occurrences. *PeerJ*, *6*(8), e5298. https://doi.org/10.7717/PEERJ.5298

Doğan, T., Atas, H., Joshi, V., Atakan, A., Rifaioglu, A. S., Nalbat, E., Nightingale, A., Saidi, R., Volynkin, V., Zellner, H., Cetin-Atalay, R., Martin, M., & Atalay, V. (2021). CROssBAR: comprehensive resource of biomedical relations with

knowledge graph representations. *Nucleic Acids Research*, *49*(16), e96–e96. https://doi.org/10.1093/NAR/GKAB543

Doğan, T., Güzelcan, E. A., Baumann, M., Koyas, A., Atas, H., Baxendale, I. R., Martin, M., & Cetin-Atalay, R. (2021). Protein domain-based prediction of drug/compound–target interactions and experimental validation on LIM kinases. *PLOS Computational Biology*, *17*(11), e1009171. https://doi.org/10.1371/JOURNAL.PCBI.1009171

Doğan, T., Macdougall, A., Saidi, R., Poggioli, D., Bateman, A., O'Donovan, C., & Martin, M. J. (2016). UniProt-DAAC: domain architecture alignment and classification, a new method for automatic functional annotation in UniProtKB. *Bioinformatics*, *32*(15), 2264. https://doi.org/10.1093/BIOINFORMATICS/BTW114

Dubchak, I., Muchnik, I., Mayor, C., Dralyuk, I., & Kim, S.-H. (1999). Recognition of a protein fold in the context of the SCOP classification. *Proteins: Structure, Function, and Genetics*, *35*(4), 401–407. https://doi.org/10.1002/(SICI)1097-0134(19990601)35:4<401::AID-PROT3>3.0.CO;2-K

Dutta, A., Dubey, T., Singh, K. K., & Anand, A. (2018). SpliceVec: Distributed feature representations for splice junction prediction. *Computational Biology and Chemistry*, *74*, 434–441. https://doi.org/10.1016/J.COMPBIOLCHEM.2018.03.009

El-Gebali, S., Mistry, J., Bateman, A., Eddy, S. R., Luciani, A., Potter, S. C., Qureshi, M., Richardson, L. J., Salazar, G. A., Smart, A., Sonnhammer, E. L. L., Hirsh, L., Paladin, L., Piovesan, D., Tosatto, S. C. E., & Finn, R. D. (2019). The Pfam protein families database in 2019. *Nucleic Acids Research*, *47*(D1), D427–D432. https://doi.org/10.1093/nar/gky995

Elnaggar, A., Heinzinger, M., Dallago, C., Rehawi, G., Wang, Y., Jones, L., Gibbs, T., Feher, T., Angerer, C., Steinegger, M., Bhowmik, D., & Rost, B. (2021). *ProtTrans : Toward Understanding the Language of Life Through Self-Supervised Learning*. *14*(8), 1–16. https://doi.org/10.1109/TPAMI.2021.3095381

Fabian, B., Nov, L. G., & Meyers, J. (2020). Molecular representation learning with language models and domain-relevant auxiliary tasks. *ArXiv*.

Feng, Z., Xia, Y., Gao, T., Xu, F., Lei, Q., Peng, C., Yang, Y., Xue, Q., Hu, X., Wang, Q., Wang, R., Ran, Z., Zeng, Z., Yang, N., Xie, Z., & Yu, L. (2018). The antipsychotic agent trifluoperazine hydrochloride suppresses triple-negative breast cancer tumor growth and brain metastasis by inducing G0/G1 arrest and apoptosis. *Cell Death & Disease 2018 9:10*, *9*(10), 1–15. https://doi.org/10.1038/s41419-018-1046-3

Fernández-Torras, A., Duran-Frigola, M., Bertoni, M., Locatelli, M., & Aloy, P. (2022). Integrating and formatting biomedical data as pre-calculated knowledge

graph embeddings in the Bioteque. *Nature Communications 2022 13:1*, *13*(1), 1–18. https://doi.org/10.1038/s41467-022-33026-0

Gaulton, A., Hersey, A., Nowotka, M., Bento, A. P., Chambers, J., Mendez, D., Mutowo, P., Atkinson, F., Bellis, L. J., Cibrián-Uhalte, E., Davies, M., Dedman, N., Karlsson, A., Magariños, M. P., Overington, J. P., Papadatos, G., Smit, I., & Leach, A. R. (2017). The ChEMBL database in 2017. *Nucleic Acids Research*, *45*(D1), D945–D954. https://doi.org/10.1093/nar/gkw1074

Geary, R. C. (1954). The Contiguity Ratio and Statistical Mapping. *The Incorporated Statistician*, *5*(3), 115–146.

Geppert, H., Humrich, J., Stumpfe, D., Gärtner, T., & Bajorath, J. (2009). Ligand prediction from protein sequence and small molecule information using support vector machines and fingerprint descriptors. *Journal of Chemical Information and Modeling*, *49*(4), 767–779. https://doi.org/10.1021/CI900004A/SUPPL_FILE/CI900004A_SI_001.PDF

Gilson, M. K., Liu, T., Baitaluk, M., Nicola, G., Hwang, L., & Chong, J. (2016). BindingDB in 2015: A public database for medicinal chemistry, computational chemistry and systems pharmacology. *Nucleic Acids Research*, *44*(D1), D1045–D1053. https://doi.org/10.1093/nar/gkv1072

Goh, G. B., Siegel, C., Vishnu, A., Hodas, N. O., & Baker, N. (2017). Chemception: A Deep Neural Network with Minimal Chemistry Knowledge Matches the Performance of Expert-developed QSAR/QSPR Models. *ArXiv*. https://arxiv.org/abs/1706.06689v1

Gönen, M. (2012). Predicting drug–target interactions from chemical and genomic kernels using Bayesian matrix factorization. *Bioinformatics*, *28*(18), 2304–2310. https://doi.org/10.1093/BIOINFORMATICS/BTS360

Gönen, M., & Heller, G. (2005). Concordance probability and discriminatory power in proportional hazards regression. *Biometrika*, *92*(4), 965–970. https://doi.org/10.1093/BIOMET/92.4.965

Gromiha, M. M., & Suwa, M. (2006). Influence of amino acid properties for discriminating outer membrane proteins at better accuracy. *Biochimica et Biophysica Acta - Proteins and Proteomics*, *1764*(9), 1493–1497. https://doi.org/10.1016/j.bbapap.2006.07.005

Hagberg, A., Swart, P., & S Chult, D. (2008). Exploring Network Structure, Dynamics, and Function using NetworkX. *Proceedings of the 7th Python in Science Conference (SciPy 2008)*. http://conference.scipy.org/proceedings/SciPy2008/paper_2

Haikala, H. M., & Jänne, P. A. (2021). 30 years of HER3: From basic biology to therapeutic interventions. *Clinical Cancer Research : An Official Journal of the*

American Association for Cancer Research, *27*(13), 3528. https://doi.org/10.1158/1078-0432.CCR-20-4465

Hanser, T., Barber, C., Marchaland, J. F., & Werner, S. (2016). Applicability domain: towards a more formal definition. *SAR QSAR Environ Res.*, *27*(11), 893–909. https://doi.org/10.1080/1062936X.2016.1250229

He, T., Heidemeyer, M., Ban, F., Cherkasov, A., & Ester, M. (2017). SimBoost: a read-across approach for predicting drug-target binding affinities using gradient boosting machines. *Journal of Cheminformatics*, *9*(1), 1–14. https://doi.org/10.1186/S13321-017-0209-Z/FIGURES/8

Heinzinger, M., Elnaggar, A., Wang, Y., Dallago, C., Nechaev, D., Matthes, F., & Rost, B. (2019). Modeling aspects of the language of life through transfer-learning protein sequences. *BMC Bioinformatics*, *20*(1), 723. https://doi.org/10.1186/s12859-019-3220-8

Hengl, T., Nussbaum, M., Wright, M. N., Heuvelink, G. B. M., & Gräler, B. (2018). Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. *PeerJ*, *2018*(8), e5518. https://doi.org/10.7717/PEERJ.5518/SUPP-1

Himmelstein, D. S., Lizee, A., Hessler, C., Brueggeman, L., Chen, S. L., Hadley, D., Green, A., Khankhanian, P., & Baranzini, S. E. (2017). Systematic integration of biomedical knowledge prioritizes drugs for repurposing. *ELife*, *6*. https://doi.org/10.7554/ELIFE.26726

Hu, Z., Dong, Y., Wang, K., & Sun, Y. (2020). Heterogeneous Graph Transformer. *The Web Conference 2020 - Proceedings of the World Wide Web Conference, WWW 2020*, 2704–2710. https://doi.org/10.1145/3366423.3380027

Huang, E. W., Bhope, A., Lim, J., Sinha, S., & Emad, A. (2020). Tissue-guided LASSO for prediction of clinical drug response using preclinical samples. *PLoS Computational Biology*, *16*(1). https://doi.org/10.1371/JOURNAL.PCBI.1007607

Huang, J., & Kingsbury, B. (2013). Audio-visual deep learning for noise robust speech recognition. *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 7596–7599. https://doi.org/10.1109/ICASSP.2013.6639140

Huang, L., Guo, Z., Wang, F., & Fu, L. (2021). KRAS mutation: from undruggable to druggable in cancer. *Signal Transduction and Targeted Therapy 2021 6:1*, *6*(1), 1–20. https://doi.org/10.1038/s41392-021-00780-4

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science and Engineering*, *9*(3), 90–95. https://doi.org/10.1109/MCSE.2007.55

Jasial, S., Hu, Y., Vogt, M., & Bajorath, J. (2016). Activity-relevant similarity values for fingerprints and implications for similarity searching. *F1000Research*, *5*, 591. https://doi.org/10.12688/f1000research.8357.2

Jaszczyszyn, A., Gąsiorowski, K., Świątek, P., Malinka, W., Cieślik-Boczula, K., Petrus, J., & Czarnik-Matusewicz, B. (2012). Chemical structure of phenothiazines and their biological activity. *Pharmacological Reports : PR*, *64*(1), 16–23. https://doi.org/10.1016/S1734-1140(12)70726-0

Jeon, J., Nim, S., Teyra, J., Datti, A., Wrana, J. L., Sidhu, S. S., Moffat, J., & Kim, P. M. (2014). A systematic approach to identify novel cancer drug targets using machine learning, inhibitor design and high-throughput screening. *Genome Medicine*, *6*(7). https://doi.org/10.1186/S13073-014-0057-7

Jiang, L., Sun, J., Wang, Y., Ning, Q., Luo, N., & Yin, M. (2022). Identifying drug–target interactions via heterogeneous graph attention networks combined with cross-modal similarities. *Briefings in Bioinformatics*, *23*(2). https://doi.org/10.1093/BIB/BBAC016

Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M., & Tanabe, M. (2016). KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Research*, *44*(D1), D457–D462. https://doi.org/10.1093/nar/gkv1070

Karimi, M., Wu, D., Wang, Z., & Shen, Y. (2019). DeepAffinity: interpretable deep learning of compound–protein affinity through unified recurrent and convolutional neural networks. *Bioinformatics*, *35*(18), 3329–3338. https://doi.org/10.1093/BIOINFORMATICS/BTZ111

Kawashima, S., Pokarowski, P., Pokarowska, M., Kolinski, A., Katayama, T., & Kanehisa, M. (2008). AAindex: amino acid index database, progress report 2008. *Nucleic Acids Research*, *36*, D202–D205. https://doi.org/10.1093/nar/gkm998

Kim, P. T., Winter, R., & Clevert, D. A. (2021). Unsupervised representation learning for proteochemometric modeling. *International Journal of Molecular Sciences*, *22*(23), 12882. https://doi.org/10.3390/IJMS222312882/S1

Köhler, S., Carmody, L., Vasilevsky, N., Jacobsen, J. O. B., Danis, D., Gourdine, J.-P., Gargano, M., Harris, N. L., Matentzoglu, N., McMurry, J. A., Osumi-Sutherland, D., Cipriani, V., Balhoff, J. P., Conlin, T., Blau, H., Baynam, G., Palmer, R., Gratian, D., Dawkins, H., … Robinson, P. N. (2019). Expansion of the Human Phenotype Ontology (HPO) knowledge base and resources. *Nucleic Acids Research*, *47*(D1), D1018–D1027. https://doi.org/10.1093/nar/gky1105

Koutsoukas, A., Monaghan, K. J., Li, X., & Huan, J. (2017). Deep-learning: Investigating deep neural networks hyper-parameters and comparison of performance to shallow methods for modeling bioactivity data. *Journal of Cheminformatics*, *9*(1), 1–13. https://doi.org/10.1186/S13321-017-0226-Y/FIGURES/5

Kumari, P., Nath, A., & Chaube, R. (2015). Identification of human drug targets using machine-learning algorithms. *Computers in Biology and Medicine*, *56*, 175–181. https://doi.org/10.1016/J.COMPBIOMED.2014.11.008

Landrum, G. (2016). *RDKit: Open-Source Cheminformatics Software*. . http://www.rdkit.org/

Lavecchia, A., & Di Giovanni, C. (2013). Virtual screening strategies in drug discovery: a critical review. *Current Medicinal Chemistry*, *20*(23), 2839–2860.

Law, V., Knox, C., Djoumbou, Y., Jewison, T., Guo, A. C., Liu, Y., MacIejewski, A., Arndt, D., Wilson, M., Neveu, V., Tang, A., Gabriel, G., Ly, C., Adamjee, S., Dame, Z. T., Han, B., Zhou, Y., & Wishart, D. S. (2014). DrugBank 4.0: Shedding new light on drug metabolism. *Nucleic Acids Research*, *42*(D1), 1091–1097. https://doi.org/10.1093/nar/gkt1068

Lenselink, E. B., Ten Dijke, N., Bongers, B., Papadatos, G., Van Vlijmen, H. W. T., Kowalczyk, W., Ijzerman, A. P., & Van Westen, G. J. P. (2017). Beyond the hype: deep neural networks outperform established methods using a ChEMBL bioactivity benchmark set. *Journal of Cheminformatics*, *9*(1), 45. https://doi.org/10.1186/s13321-017-0232-0

Li, B., Shin, H., Gulbekyan, G., Pustovalova, O., Nikolsky, Y., Hope, A., Bessarabova, M., Schu, M., Kolpakova-Hart, E., Merberg, D., Dorner, A., & Trepicchio, W. L. (2015). Development of a Drug-Response Modeling Framework to Identify Cell Line Derived Translational Biomarkers That Can Predict Treatment Outcome to Erlotinib or Sorafenib. *PLoS ONE*, *10*(6). https://doi.org/10.1371/JOURNAL.PONE.0130700

Li, J., Wang, J., Lv, H., Zhang, Z., & Wang, Z. (2022). IMCHGAN: Inductive Matrix Completion With Heterogeneous Graph Attention Networks for Drug-Target Interactions Prediction. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *19*(2), 655–665. https://doi.org/10.1109/TCBB.2021.3088614

Li, S., Wan, F., Shu, H., Jiang, T., Zhao, D., & Zeng, J. (2020). MONN: A Multi-objective Neural Network for Predicting Compound-Protein Interactions and Affinities. *Cell Systems*, *10*(4), 308-322.e11. https://doi.org/10.1016/J.CELS.2020.03.002

Li, Z. R., Lin, H. H., Han, L. Y., Jiang, L., Chen, X., & Chen, Y. Z. (2006). PROFEAT: a web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence. *Nucleic Acids Research*, *34*, W32–W37. https://doi.org/10.1093/nar/gkr284

Liang, S., & Yu, H. (2020). Revealing new therapeutic opportunities through drug target prediction: A class imbalance-tolerant machine learning approach. *Bioinformatics*, *36*(16), 4490–4497. https://doi.org/10.1093/bioinformatics/btaa495

Liao, J., Chen, H., Wei, L., & Wei, L. (2022). GSAML-DTA: An interpretable drug-target binding affinity prediction model based on graph neural networks with self-attention mechanism and mutual information. *Computers in Biology and Medicine*, *150*, 106145. https://doi.org/10.1016/J.COMPBIOMED.2022.106145

Liekens, A. M. L., De Knijf, J., Daelemans, W., Goethals, B., De Rijk, P., & Del-Favero, J. (2011). BioGraph: unsupervised biomedical knowledge discovery via automated hypothesis generation. *Genome Biology*, *12*(6). https://doi.org/10.1186/GB-2011-12-6-R57

Lin, X., Zhao, K., Xiao, T., Quan, Z., Wang, Z. J., & Yu, P. S. (2020). DeepGS: Deep Representation Learning of Graphs and Sequences for Drug-Target Binding Affinity Prediction. *Frontiers in Artificial Intelligence and Applications*, *325*, 1301–1308. https://doi.org/10.3233/FAIA200232

Liu, H., Sun, J., Guan, J., Zheng, J., & Zhou, S. (2015). Improving compound-protein interaction prediction by building up highly credible negative samples. *Bioinformatics*, *31*(12), i221–i229. https://doi.org/10.1093/bioinformatics/btv256

Liu, Y., Wu, M., Miao, C., Zhao, P., & Li, X. L. (2016). Neighborhood Regularized Logistic Matrix Factorization for Drug-Target Interaction Prediction. *PLOS Computational Biology*, *12*(2), e1004760. https://doi.org/10.1371/JOURNAL.PCBI.1004760

Liu, Z., Chen, Q., Lan, W., Pan, H., Hao, X., & Pan, S. (2021). GADTI: Graph Autoencoder Approach for DTI Prediction From Heterogeneous Network. *Frontiers in Genetics*, *12*, 650821. https://doi.org/10.3389/FGENE.2021.650821

Lobentanzer, S., & Saez-Rodriguez, J. (2023). A Platform for the Biomedical Application of Large Language Models. *ArXiv*. https://arxiv.org/abs/2305.06488v2

Ma, J., Sheridan, R. P., Liaw, A., Dahl, G. E., & Svetnik, V. (2015). Deep neural nets as a method for quantitative structure-activity relationships. *Journal of Chemical Information and Modeling*, *55*(2), 263–274. https://doi.org/10.1021/CI500747N/SUPPL_FILE/CI500747N_SI_003.PDF

Mayr, A., Klambauer, G., Unterthiner, T., Steijaert, M., Wegner, J. K., Ceulemans, H., Clevert, D. A., & Hochreiter, S. (2018). Large-scale comparison of machine learning methods for drug target prediction on ChEMBL. *Chemical Science*, *9*(24), 5441–5451. https://doi.org/10.1039/c8sc00148k

Mei, J. P., Kwoh, C. K., Yang, P., Li, X. L., & Zheng, J. (2013). Drug–target interaction prediction by learning from local information and neighbors. *Bioinformatics*, *29*(2), 238–245. https://doi.org/10.1093/BIOINFORMATICS/BTS670

Mendez, D., Gaulton, A., Bento, A. P., Chambers, J., De Veij, M., Félix, E., Magariños, M. P., Mosquera, J. F., Mutowo, P., Nowotka, M., Gordillo-Marañón, M., Hunter, F., Junco, L., Mugumbate, G., Rodriguez-Lopez, M., Atkinson, F., Bosc, N., Radoux, C. J., Segura-Cabrera, A., … Leach, A. R. (2019). ChEMBL: towards direct deposition of bioassay data. *Nucleic Acids Research*, *47*(D1), D930–D940. https://doi.org/10.1093/NAR/GKY1075

Messina, A., Fiannaca, A., la Paglia, L., la Rosa, M., & Urso, A. (2018). BioGraph: A web application and a graph database for querying and analyzing bioinformatics resources. *BMC Systems Biology*, *12*(5), 75–89. https://doi.org/10.1186/S12918-018-0616-4/TABLES/3

Messina, A., Pribadi, H., Stichbury, J., Bucci, M., Klarman, S., & Urso, A. (2018). BioGrakn: A knowledge graph-based semantic database for biomedical sciences. *Advances in Intelligent Systems and Computing*, *611*, 299–309. https://doi.org/10.1007/978-3-319-61566-0_28/COVER

Mirabello, C., & Wallner, B. (2019). rawMSA: End-to-end Deep Learning using raw Multiple Sequence Alignments. *PLoS ONE*, *14*(8), e0220182. https://doi.org/10.1371/JOURNAL.PONE.0220182

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

Morris, J. H., Soman, K., Akbas, R. E., Zhou, X., Smith, B., Meng, E. C., Huang, C. C., Cerono, G., Schenk, G., Rizk-Jackson, A., Harroud, A., Sanders, L., Costes, S. V., Bharat, K., Chakraborty, A., Pico, A. R., Mardirossian, T., Keiser, M., Tang, A., … Baranzini, S. E. (2023). The scalable precision medicine open knowledge engine (SPOKE): a massive knowledge graph of biomedical information. *Bioinformatics*, *39*(2). https://doi.org/10.1093/BIOINFORMATICS/BTAD080

Mousavian, Z., Khakabimamaghani, S., Kavousi, K., & Masoudi-Nejad, A. (2016). Drug–target interaction prediction from PSSM based evolutionary information. *Journal of Pharmacological and Toxicological Methods*, *78*, 42–51. https://doi.org/10.1016/J.VASCN.2015.11.002

Muegge, I., & Mukherjee, P. (2016). An overview of molecular fingerprint similarity search in virtual screening. *Expert Opinion on Drug Discovery*, *11*(2), 137–148. https://doi.org/10.1517/17460441.2016.1117070

Nelson, W., Zitnik, M., Wang, B., Leskovec, J., Goldenberg, A., & Sharan, R. (2019). To embed or not: Network embedding as a paradigm in computational biology. *Frontiers in Genetics*, *10*(MAY), 381. https://doi.org/10.3389/FGENE.2019.00381/BIBTEX

Nguyen, T., Le, H., Quinn, T. P., Nguyen, T., Le, T. D., & Venkatesh, S. (2021). GraphDTA: predicting drug–target binding affinity with graph neural networks. *Bioinformatics*, *37*(8), 1140–1147. https://doi.org/10.1093/BIOINFORMATICS/BTAA921

Ning, X., Rangwala, H., & Karypis, G. (2009). Multi-assay-based structure-activity relationship models: Improving structure-activity relationship models by incorporating activity information from related targets. *Journal of Chemical*

Information and Modeling, *49*(11), 2444–2456. https://doi.org/10.1021/CI900182Q/SUPPL_FILE/CI900182Q_SI_001.HTM

Olivecrona, M., Blaschke, T., Engkvist, O., & Chen, H. (2017). Molecular de-novo design through deep reinforcement learning. *Journal of Cheminformatics*, *9*(1). https://doi.org/10.1186/S13321-017-0235-X

OMIM - Online Mendelian Inheritance in Man. (n.d.). Retrieved May 25, 2019, from https://www.omim.org/

Ong, S. A., Lin, H. H., Chen, Y. Z., Li, Z. R., & Cao, Z. (2007). Efficacy of different protein descriptors in predicting protein functional families. *BMC Bioinformatics*, *8*, 300. https://doi.org/10.1186/1471-2105-8-300

Öztürk, H., Özgür, A., & Ozkirimli, E. (2018). DeepDTA: deep drug–target binding affinity prediction. *Bioinformatics*, *34*(17), i821–i829. https://doi.org/10.1093/BIOINFORMATICS/BTY593

Öztürk, H., Ozkirimli, E., & Özgür, A. (2019). WideDTA: prediction of drug-target binding affinity. *ArXiv*, 1902.04166. https://arxiv.org/abs/1902.04166v1

Pareja-Tobes, P., Tobes, R., Manrique, M., Pareja, E., & Pareja-Tobes, E. (2015). Bio4j: a high-performance cloud-enabled graph-based data platform. *BioRxiv*, 016758. https://doi.org/10.1101/016758

Paul, S. M., Mytelka, D. S., Dunwiddie, C. T., Persinger, C. C., Munos, B. H., Lindborg, S. R., & Schacht, A. L. (2010). How to improve R&D productivity: the pharmaceutical industry's grand challenge. *Nature Reviews. Drug Discovery*, *9*(3), 203–214. https://doi.org/10.1038/nrd3078

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Brucher, M., & Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*(85), 2825–2830.

Peng, C., Dieck, S., Schmid, A., Ahmad, A., Knaus, A., Wenzel, M., Mehnert, L., Zirn, B., Haack, T., Ossowski, S., Wagner, M., Brunet, T., Ehmke, N., Danyel, M., Rosnev, S., Kamphans, T., Nadav, G., Fleischer, N., Fröhlich, H., & Krawitz, P. (2021). CADA: phenotype-driven gene prioritization based on a case-enriched knowledge graph. *NAR Genomics and Bioinformatics*, *3*(3). https://doi.org/10.1093/NARGAB/LQAB078

Peng, J., Wang, Y., Guan, J., Li, J., Han, R., Hao, J., Wei, Z., & Shang, X. (2021). An end-to-end heterogeneous graph representation learning-based framework for drug–target interaction prediction. *Briefings in Bioinformatics*, *22*(5), 1–9. https://doi.org/10.1093/BIB/BBAA430

Perlman, L., Gottlieb, A., Atias, N., Ruppin, E., & Sharan, R. (2011). Combining drug and gene similarity measures for drug-target elucidation. *Journal of Computational Biology*, *18*(2), 133–145. https://doi.org/10.1089/cmb.2010.0213

Peska, L., Buza, K., & Koller, J. (2017). Drug-target interaction prediction: A Bayesian ranking approach. *Computer Methods and Programs in Biomedicine*, *152*, 15–21. https://doi.org/10.1016/J.CMPB.2017.09.003

Poorinmohammad, N., Mohabatkar, H., Behbahani, M., & Biria, D. (2015). Computational prediction of anti HIV-1 peptides and in vitro evaluation of anti HIV-1 activity of HIV-1 P24-derived peptides. *Journal of Peptide Science*, *21*(1), 10–16. https://doi.org/10.1002/PSC.2712

Ramsundar, B., Kearnes, S., Riley, P., Webster, D., Konerding, D., Pande, V., & Edu, P. (2015). *Massively Multitask Networks for Drug Discovery*. https://arxiv.org/abs/1502.02072v1

Rao, R., Bhattacharya, N., Thomas, N., Duan, Y., Chen, X., Canny, J., Abbeel, P., & Song, Y. S. (2019, June 19). Evaluating Protein Transfer Learning with TAPE. *33rd Conference on Neural Information Processing Systems*.

Rice, P., Longden, I., & Bleasby, A. (2000). EMBOSS: The European Molecular Biology Open Software Suite. *Trends in Genetics*, *16*(6), 276–277. https://doi.org/10.1016/S0168-9525(00)02024-2

Rifaioglu, A. S., Atas, H., Martin, M. J., Cetin-Atalay, R., Atalay, V., & Doğan, T. (2019). Recent applications of deep learning and machine intelligence on in silico drug discovery: methods, tools and databases. *Briefings in Bioinformatics*, *20*(5), 1878–1912. https://doi.org/10.1093/bib/bby061

Rifaioglu, A. S., Cetin-Atalay, R., Dogan, T., Martin, M., & Atalay, M. V. (2021). MDeePred: Novel Multi-Channel Protein Featurization for Deep Learning based Binding Affinity Prediction in Drug Discovery. *Bioinformatics*, *37*(5), 693–704.

Rifaioglu, A. S., Doğan, T., Saraç, Ö. S., Ersahin, T., Saidi, R., Atalay, M. V., Martin, M. J., & Cetin-Atalay, R. (2018). Large-scale automated function prediction of protein sequences and an experimental case study validation on PTEN transcript variants. *Proteins: Structure, Function, and Bioinformatics*, *86*(2), 135–151. https://doi.org/10.1002/PROT.25416

Rifaioglu, A. S., Nalbat, E., Atalay, V., Martin, M. J., Cetin-Atalay, R., & Doğan, T. (2020). DEEPScreen: high performance drug–target interaction prediction with convolutional neural networks using 2-D structural compound representations. *Chemical Science*, *11*(9), 2531–2557. https://doi.org/10.1039/C9SC03414E

Rogers, D., & Hahn, M. (2010). Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling*, *50*(5), 742–754. https://doi.org/10.1021/ci100050t

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., & Fei-Fei, L. (2015). ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, *115*(3), 211–252. https://doi.org/10.1007/s11263-015-0816-y

Sahigara, F., Mansouri, K., Ballabio, D., Mauri, A., Consonni, V., & Todeschini, R. (2012). Comparison of Different Approaches to Define the Applicability Domain of QSAR Models. *Molecules*, *17*(5), 4791. https://doi.org/10.3390/MOLECULES17054791

Saini, H., Raicar, G., Lal, S., Dehzangi, A., Imoto, S., & Sharma, A. (2016). Protein Fold Recognition Using Genetic Algorithm Optimized Voting Scheme and Profile Bigram. *Journal of Software*, *11*(8), 756–767. https://doi.org/10.17706/jsw.11.8.756-767

Sarac, O. S., Gürsoy-Yüzügüllü, O., Cetin-Atalay, R., & Atalay, V. (2008). Subsequence-based feature map for protein function classification. *Computational Biology and Chemistry*, *32*(2), 122–130. https://doi.org/10.1016/j.compbiolchem.2007.11.004

Saravanan, V., & Gautham, N. (2015). Harnessing Computational Biology for Exact Linear B-Cell Epitope Prediction: A Novel Amino Acid Composition-Based Feature Descriptor. *OMICS: A Journal of Integrative Biology*, *19*(10), 648–658. https://doi.org/10.1089/omi.2015.0095

Sawada, R., Kotera, M., & Yamanishi, Y. (2014). Benchmarking a Wide Range of Chemical Descriptors for Drug-Target Interaction Prediction Using a Chemogenomic Approach. *Molecular Informatics*, *33*(11–12), 719–731. https://doi.org/10.1002/minf.201400066

Schneider, G., & Wrede, P. (1994). The Rational Design of Amino Acid Sequences by Artificial Neural Networks and Simulated Molecular Evolution: De Novo Design of an Idealized Leader Peptidase Cleavage Site. *Biophysical Journal*, *66*, 335–344.

Segler, M. H. S., Preuss, M., & Waller, M. P. (2018). Planning chemical syntheses with deep neural networks and symbolic AI. *Nature*, *555*(7698), 604–610. https://doi.org/10.1038/NATURE25978

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., & Ideker, T. (2003). Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Research*, *13*(11), 2498–2504. https://doi.org/10.1101/gr.1239303

Shar, P. A., Tao, W., Gao, S., Huang, C., Li, B., Zhang, W., Shahen, M., Zheng, C., Bai, Y., & Wang, Y. (2016). Pred-binding: large-scale protein–ligand binding affinity prediction. *Journal of Enzyme Inhibition and Medicinal Chemistry*, *31*(6), 1443–1450. https://doi.org/10.3109/14756366.2016.1144594/SUPPL_FILE/IENZ_A_1144594_SM1461.ZIP

Shen, J., Zhang, J., Luo, X., Zhu, W., Yu, K., Chen, K., Li, Y., & Jiang, H. (2007). Predicting protein-protein interactions based only on sequences information.

Proceedings of the National Academy of Sciences of the United States of America, *104*(11), 4337–4341. https://doi.org/10.1073/pnas.0607879104

Shi, H., Liu, S., Chen, J., Li, X., Ma, Q., & Yu, B. (2019). Predicting drug-target interactions using Lasso with random forest based on evolutionary information and chemical structure. *Genomics*, *111*(6), 1839–1852. https://doi.org/10.1016/J.YGENO.2018.12.007

Singh, H., Singh, S., Singla, D., Agarwal, S. M., & Raghava, G. P. S. (2015). QSAR based model for discriminating EGFR inhibitors and non-inhibitors using Random forest. *Biology Direct*, *10*(1), 1–12. https://doi.org/10.1186/S13062-015-0046-9/TABLES/4

Strodthoff, N., Wagner, P., Wenzel, M., & Samek, W. (2020). UDSMProt: universal deep sequence models for protein classification. *Bioinformatics*, *36*(8), 2401. https://doi.org/10.1093/BIOINFORMATICS/BTAA003

Strömbergsson, H., Daniluk, P., Kryshtafovych, A., Fidelis, K., Wikberg, J. E. S., Kleywegt, G. J., & Hvidsten, T. R. (2008). Interaction model based on local protein substructures generalizes to the entire structural enzyme-ligand space. *Journal of Chemical Information and Modeling*, *48*(11), 2278–2288. https://doi.org/10.1021/CI800200E/SUPPL_FILE/CI800200E_SI_001.ZIP

Subramanian, V., Ain, Q. U., Henno, H., Pietilä, L. O., Fuchs, J. E., Prusis, P., Bender, A., & Wohlfahrt, G. (2017). 3D proteochemometrics: using three-dimensional information of proteins and ligands to address aspects of the selectivity of serine proteases. *MedChemComm*, *8*(5), 1037. https://doi.org/10.1039/C6MD00701E

Sun, J., Jeliazkova, N., Chupakhin, V., Golib-Dzib, J.-F., Engkvist, O., Carlsson, L., Wegner, J., Ceulemans, H., Georgiev, I., Jeliazkov, V., Kochev, N., Ashby, T. J., & Chen, H. (2017). ExCAPE-DB: an integrated large scale dataset facilitating Big Data analysis in chemogenomics. *Journal of Cheminformatics*, *9*(1), 17. https://doi.org/10.1186/s13321-017-0203-5

Sun, M., Wang, X., Zou, C., He, Z., Liu, W., & Li, H. (2016). Accurate prediction of RNA-binding protein residues with two discriminative structural descriptors. *BMC Bioinformatics*, *17*(1), 231. https://doi.org/10.1186/s12859-016-1110-x

Suzek, B. E., Wang, Y., Huang, H., McGarvey, P. B., & Wu, C. H. (2015). UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics*, *31*(6), 926–932. https://doi.org/10.1093/bioinformatics/btu739

Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., Simonovic, M., Roth, A., Santos, A., Tsafou, K. P., Kuhn, M., Bork, P., Jensen, L. J., & von Mering, C. (2015). STRING v10: protein–protein interaction networks, integrated over the tree of life. *Nucleic Acids Research*, *43*(D1), D447–D452. https://doi.org/10.1093/nar/gku1003

Tabei, Y., & Yamanishi, Y. (2013). Scalable prediction of compound-protein interactions using minwise hashing. *BMC Systems Biology*, *7*(6), 1–13. https://doi.org/10.1186/1752-0509-7-S6-S3/FIGURES/13

Thafar, M. A., Alshahrani, M., Albaradei, S., Gojobori, T., Essack, M., & Gao, X. (2022). Affinity2Vec: drug-target binding affinity prediction through representation learning, graph mining, and machine learning. *Scientific Reports 2022 12:1*, *12*(1), 1–18. https://doi.org/10.1038/s41598-022-08787-9

Thafar, M. A., Thafar, M. A., Olayan, R. S., Olayan, R. S., Ashoor, H., Ashoor, H., Albaradei, S., Albaradei, S., Bajic, V. B., Gao, X., Gojobori, T., & Essack, M. (2020). DTiGEMS+: Drug-target interaction prediction using graph embedding, graph mining, and similarity-based techniques. *Journal of Cheminformatics*, *12*(1), 1–17. https://doi.org/10.1186/S13321-020-00447-2/TABLES/5

The Cancer Genome Atlas Program - National Cancer Institute. (n.d.). Retrieved May 25, 2019, from https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga

The UniProt Consortium. (2021). UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Research*, *49*. https://doi.org/10.1093/nar/gkaa1100

Tian, Z., Peng, X., Fang, H., Zhang, W., Dai, Q., & Ye, Y. (2022). MHADTI: predicting drug–target interactions via multiview heterogeneous information network embedding with hierarchical attention mechanisms. *Briefings in Bioinformatics*, *23*(6). https://doi.org/10.1093/BIB/BBAC434

Torng, W., & Altman, R. B. (2019). Graph Convolutional Neural Networks for Predicting Drug-Target Interactions. *Journal of Chemical Information and Modeling*. https://doi.org/10.1021/ACS.JCIM.9B00628/ASSET/IMAGES/LARGE/CI9B00628_0012.JPEG

Tsubaki, M., Tomii, K., & Sese, J. (2019). Compound–protein interaction prediction with end-to-end learning of neural networks for graphs and sequences. *Bioinformatics*, *35*(2), 309–318. https://doi.org/10.1093/BIOINFORMATICS/BTY535

Tursynbay, Y., Zhang, J., Li, Z., Tokay, T., Zhumadilov, Z., Wu, D., & Xie, Y. (2016). Pim-1 kinase as cancer drug target: An update. *Biomedical Reports*, *4*(2), 140. https://doi.org/10.3892/BR.2015.561

Ulmer, D., Hardmeier, C., & Frellsen, J. (2022). *deep-significance - Easy and Meaningful Statistical Significance Testing in the Age of Neural Networks*. https://arxiv.org/abs/2204.06815v1

Unsal, S., Atas, H., Albayrak, M., Turhan, K., Acar, A. C., & Doğan, T. (2022). Learning Functional Properties of Proteins with Language Models. *Nature Machine Intelligence*, .

Vamathevan, J., Clark, D., Czodrowski, P., Dunham, I., Ferran, E., Lee, G., Li, B., Madabhushi, A., Shah, P., Spitzer, M., & Zhao, S. (2019). Applications of machine learning in drug discovery and development. *Nature Reviews Drug Discovery 2019 18:6*, *18*(6), 463–477. https://doi.org/10.1038/s41573-019-0024-5

Van Der Maaten, L., & Hinton, G. (2008). Visualizing Data using t-SNE. *Journal of Machine Learning Research*, *9*, 2579–2605.

van Laarhoven, T., & Marchiori, E. (2013). Predicting Drug-Target Interactions for New Drug Compounds Using a Weighted Nearest Neighbor Profile. *PLOS ONE*, *8*(6), e66952. https://doi.org/10.1371/JOURNAL.PONE.0066952

van Laarhoven, T., Nabuurs, S. B., & Marchiori, E. (2011). Gaussian interaction profile kernels for predicting drug–target interaction. *Bioinformatics*, *27*(21), 3036–3043. https://doi.org/10.1093/BIOINFORMATICS/BTR500

van Westen, G. J. P., Swier, R. F., Cortes-Ciriano, I., Wegner, J. K., Overington, J. P., Jzerman, A. P. I., Van Vlijmen, H. W. T., & Bender, A. (2013). Benchmarking of protein descriptor sets in proteochemometric modeling (part 1): Modeling performance of 13 amino acid descriptor sets. *Journal of Cheminformatics*, *5*(1), 41. https://doi.org/10.1186/1758-2946-5-41

van Westen, G. J., Swier, R. F., Cortes-Ciriano, I., Wegner, J. K., Overington, J. P., Ijzerman, A. P., van Vlijmen, H. W., & Bender, A. (2013). Benchmarking of protein descriptor sets in proteochemometric modeling (part 2): modeling performance of 13 amino acid descriptor sets. *Journal of Cheminformatics*, *5*(1), 42. https://doi.org/10.1186/1758-2946-5-42

Vaswani, A., Brain, G., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention Is All You Need. *31st Conference on Neural Information Processing Systems*.

Wallach, I., Dzamba, M., & Heifets, A. (2015). AtomNet: A Deep Convolutional Neural Network for Bioactivity Prediction in Structure-based Drug Discovery. *ArXiv*. https://arxiv.org/abs/1510.02855v1

Wan, F., Hong, L., Xiao, A., Jiang, T., & Zeng, J. (2019). NeoDTI: neural integration of neighbor information from a heterogeneous network for discovering new drug–target interactions. *Bioinformatics*, *35*(1), 104–111. https://doi.org/10.1093/BIOINFORMATICS/BTY543

Wang, C., Liu, J., Luo, F., Tan, Y., Deng, Z., & Hu, Q. N. (2014). Pairwise input neural network for target-ligand interaction prediction. *Proceedings - 2014 IEEE International Conference on Bioinformatics and Biomedicine, IEEE BIBM 2014*, 67–70. https://doi.org/10.1109/BIBM.2014.6999129

Wang, J., Yang, B., Revote, J., Leier, A., Marquez-Lago, T. T., Webb, G., Song, J., Chou, K.-C., & Lithgow, T. (2017). POSSUM: a bioinformatics toolkit for

generating numerical sequence feature descriptors based on PSSM profiles. *Bioinformatics*, *33*(17), 2756–2758. https://doi.org/10.1093/bioinformatics/btx302

Wang, W., Liang, S., Yu, M., Liu, D., Zhang, H. J., Wang, X. F., & Zhou, Y. (2022). GCHN-DTI: Predicting drug-target interactions by graph convolution on heterogeneous networks. *Methods*, *206*, 101–107. https://doi.org/10.1016/J.YMETH.2022.08.016

Wang, Y., Bryant, S. H., Cheng, T., Wang, J., Gindulyte, A., Shoemaker, B. A., Thiessen, P. A., He, S., & Zhang, J. (2017). PubChem BioAssay: 2017 update. *Nucleic Acids Research*, *45*. https://doi.org/10.1093/nar/gkw1118

Wang, Y., Guo, Y., Kuang, Q., Pu, X., Ji, Y., Zhang, Z., & Li, M. (2015). A comparative study of family-specific protein-ligand complex affinity prediction based on random forest approach. *Journal of Computer-Aided Molecular Design*, *29*(4), 349–360. https://doi.org/10.1007/S10822-014-9827-Y/TABLES/4

Waskom, M. (2021). seaborn: statistical data visualization. *Journal of Open Source Software*, *6*(60), 3021. https://doi.org/10.21105/joss.03021

Wenzel, J., Matter, H., & Schmidt, F. (2019). Predictive Multitask Deep Neural Network Models for ADME-Tox Properties: Learning from Large Data Sets. *Journal of Chemical Information and Modeling*, *59*(3), 1253–1268. https://doi.org/10.1021/ACS.JCIM.8B00785/SUPPL_FILE/CI8B00785_SI_001.ZIP

Wilcken, R., Zimmermann, M. O., Lange, A., Joerger, A. C., & Boeckler, F. M. (2013). Principles and applications of halogen bonding in medicinal chemistry and chemical biology. *Journal of Medicinal Chemistry*, *56*(4), 1363–1388. https://doi.org/10.1021/JM3012068/ASSET/IMAGES/LARGE/JM-2012-012068_0010.JPEG

Wu, D., Huang, Q., Zhang, Y., Zhang, Q., Liu, Q., Gao, J., Cao, Z., & Zhu, R. (2012). Screening of selective histone deacetylase inhibitors by proteochemometric modeling. *BMC Bioinformatics*, *13*, 212. https://doi.org/10.1186/1471-2105-13-212

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Kaiser, Ł., Gouws, S., Kato, Y., Kudo, T., Kazawa, H., … Dean, J. (2016). *Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation*.

Wu, Z., Ramsundar, B., Feinberg, E. N., Gomes, J., Geniesse, C., Pappu, A. S., Leswing, K., & Pande, V. (2018). MoleculeNet: a benchmark for molecular machine learning. *Chemical Science*, *9*(2), 513–530. https://doi.org/10.1039/C7SC02664A

Xia, Z., Wu, L.-Y., Zhou, X., & Wong, S. T. (2010). Semi-supervised drug-protein interaction prediction from heterogeneous biological spaces. *BMC Systems Biology 2010 4:2*, *4*(2), 1–16. https://doi.org/10.1186/1752-0509-4-S2-S6

Xu, Y., Verma, D., Sheridan, R. P., Liaw, A., Ma, J., Marshall, N. M., McIntosh, J., Sherer, E. C., Svetnik, V., & Johnston, J. M. (2020). Deep Dive into Machine Learning Models for Protein Engineering. *Journal of Chemical Information and Modeling*, *60*(6), 2773–2790. https://doi.org/10.1021/acs.jcim.0c00073

Yabuuchi, H., Niijima, S., Takematsu, H., Ida, T., Hirokawa, T., Hara, T., Ogawa, T., Minowa, Y., Tsujimoto, G., & Okuno, Y. (2011). Analysis of multiple compound–protein interactions reveals novel bioactive molecules. *Molecular Systems Biology*, *7*(1), 472. https://doi.org/10.1038/MSB.2011.5

Yamanishi, Y., Pauwels, E., Saigo, H., & Stoven, V. (2011). Extracting sets of chemical substructures and protein domains governing drug-target interactions. *Journal of Chemical Information and Modeling*, *51*(5), 1183–1194. https://doi.org/10.1021/ci100476q

Yan, X., & Liu, Y. (2022). Graph–sequence attention and transformer for predicting drug–target affinity. *RSC Advances*, *12*(45), 29525–29534. https://doi.org/10.1039/D2RA05566J

Yang, Z., Zhong, W., Zhao, L., & Yu-Chian Chen, C. (2022). MGraphDTA: deep multiscale graph neural network for explainable drug–target binding affinity prediction. *Chemical Science*, *13*(3), 816–833. https://doi.org/10.1039/D1SC05180F

Ye, Q., Hsieh, C. Y., Yang, Z., Kang, Y., Chen, J., Cao, D., He, S., & Hou, T. (2021). A unified drug–target interaction prediction framework based on knowledge graph and recommendation system. *Nature Communications 2021*, *12*(1), 1–12. https://doi.org/10.1038/s41467-021-27137-3

You, R., Huang, X., & Zhu, S. (2018). DeepText2GO: Improving large-scale protein function prediction with deep semantic text representation. *Methods*, *145*, 82–90. https://doi.org/10.1016/j.ymeth.2018.05.026

Yu, H., Chen, J., Xu, X., Li, Y., Zhao, H., Fang, Y., Li, X., Zhou, W., Wang, W., & Wang, Y. (2012). A systematic prediction of multiple drug-target interactions from chemical, genomic, and pharmacological data. *PloS One*, *7*(5). https://doi.org/10.1371/JOURNAL.PONE.0037608

Yu, S., Yuan, Z., Xia, J., Luo, S., Ying, H., Zeng, S., Ren, J., Yuan, H., Zhao, Z., Lin, Y., Lu, K., Wang, J., Xie, Y., & Shum, H.-Y. (2022). *BIOS: An Algorithmically Generated Biomedical Knowledge Graph*. https://arxiv.org/abs/2203.09975v2

Yue, Y., & He, S. (2021). DTI-HeNE: a novel method for drug-target interaction prediction based on heterogeneous network embedding. *BMC Bioinformatics*, *22*(1). https://doi.org/10.1186/S12859-021-04327-W

Yüksel, A., Ulusoy, E., Ünlü, A., Deniz, G., & Doğan, T. (2023). *SELFormer: Molecular Representation Learning via SELFIES Language Models*. https://arxiv.org/abs/2304.04662v1

Zhang, L., Tan, J., Han, D., & Zhu, H. (2017). From machine learning to deep learning: progress in machine intelligence for rational drug discovery. *Drug Discovery Today*, *22*(11), 1680–1685. https://doi.org/10.1016/J.DRUDIS.2017.08.010

Zhang, P., Tao, L., Zeng, X., Qin, C., Chen, S. Y., Zhu, F., Yang, S. Y., Li, Z. R., Chen, W. P., & Chen, Y. Z. (2017). PROFEAT Update: A Protein Features Web Server with Added Facility to Compute Network Descriptors for Studying Omics-Derived Networks. *Journal of Molecular Biology*, *429*(3), 416–425. https://doi.org/10.1016/j.jmb.2016.10.013

Zhang, Z., Chen, L., Zhong, F., Wang, D., Jiang, J., Zhang, S., Jiang, H., Zheng, M., & Li, X. (2022). Graph neural network approaches for drug-target interactions. *Current Opinion in Structural Biology*, *73*, 102327. https://doi.org/10.1016/J.SBI.2021.102327

Zhang, Z., Cui, P., & Zhu, W. (2022). Deep Learning on Graphs: A Survey. *IEEE Transactions on Knowledge and Data Engineering*, *34*(1), 249–270. https://doi.org/10.1109/TKDE.2020.2981333

Zheng, S., Li, Y., Chen, S., Xu, J., & Yang, Y. (2020). Predicting drug–protein interaction using quasi-visual question answering system. *Nature Machine Intelligence 2020 2:2*, *2*(2), 134–140. https://doi.org/10.1038/s42256-020-0152-y

Zheng, X., Ding, H., Mamitsuka, H., & Zhu, S. (2013). Collaborative matrix factorization with multiple similarities for predicting drug-Target interactions. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, *Part F128815*, 1025–1033. https://doi.org/10.1145/2487575.2487670

Zhou, D., Xu, Z., Li, W., Xie, X., & Peng, S. (2021). MultiDTI: drug–target interaction prediction based on multi-modal representation learning to bridge the gap between new chemical entities and known heterogeneous network. *Bioinformatics*, *37*(23), 4485–4492. https://doi.org/10.1093/BIOINFORMATICS/BTAB473

APPENDICES

APPENDIX A

CHAPTER 3: PERFORMANCE RESULTS OF BENCHMARK ANALYSIS

Table 3.1. Model performance scores (in terms of MCC) in the small-scale analysis (on the compound-centric datasets) for; **(a)** random forest, and **(b)** SVM models. The 3 best performances for each dataset are shown in bold font.

(a)

Model	ChEMBL id (only the numeric part) of the center compound of each compound cluster									Mean	Standard error
	44	**50**	**83**	**91**	**104**	**633**	**808**	**116438**	**295698**		
aac	0.283	0.112	0.359	0.238	0.413	0.199	0.227	0.099	0.195	0.236	0.035
aac_pssm	0.226	0.278	0.395	0.432	0.170	0.212	0.399	0.153	0.270	0.282	0.035
aadp_pssm	0.166	0.211	0.396	0.453	0.177	0.347	0.330	0.240	0.380	0.300	0.035
aatp_pssm	0.231	0.312	0.432	0.477	0.332	0.280	0.398	0.254	0.387	0.345	0.028
ab_pssm	0.237	0.231	0.397	0.408	0.157	0.323	0.487	0.206	0.364	0.312	0.037
apaac	0.263	0.262	0.463	0.346	0.303	0.514	0.422	**0.293**	0.384	0.361	0.030
cksaagp	0.243	0.034	0.390	0.333	0.222	0.323	0.261	0.120	0.260	0.243	0.037
cksaap	0.324	0.268	0.510	0.410	0.297	0.489	0.405	0.143	0.431	0.364	0.039
ctdc	0.117	-0.035	0.361	0.213	0.332	0.280	0.223	0.185	0.109	0.199	0.041
ctdd	0.176	0.209	0.446	0.204	0.142	0.512	0.471	0.185	0.390	0.304	0.049
ctdt	0.212	0.033	0.309	0.201	0.363	0.169	0.379	0.130	0.301	0.233	0.038
ctriad	0.231	0.179	0.355	0.230	0.243	0.449	0.285	0.280	0.396	0.294	0.029
d_fpssm	**0.364**	0.288	0.429	0.371	0.398	0.518	0.492	0.250	0.345	0.384	0.029
dde	0.302	0.293	**0.528**	**0.492**	0.356	0.512	**0.536**	0.174	0.446	**0.404**	0.043
dp_pssm	0.279	**0.376**	0.369	0.387	0.297	0.316	0.325	0.278	0.396	0.336	0.016
dpc	0.318	0.170	0.467	0.482	0.197	0.453	0.395	0.121	0.412	0.335	0.046
dpc_pssm	0.227	0.210	0.396	0.491	0.198	0.369	0.360	0.233	0.433	0.324	0.036
edp_pssm	0.144	0.172	0.375	0.388	0.164	0.141	0.358	0.152	0.273	0.241	0.036
eedp_pssm	0.186	0.225	0.419	0.433	0.164	0.310	0.336	0.022	0.262	0.262	0.043
gaac	0.089	-0.081	0.266	0.020	0.088	0.200	0.135	-0.057	0.129	0.088	0.038
gdpc	0.314	0.045	0.311	0.292	0.253	0.310	0.262	0.076	0.232	0.233	0.034
geary	0.337	0.182	0.390	0.394	0.294	0.339	0.360	**0.345**	0.322	0.329	0.021

gtpc	0.304	0.134	0.344	0.198	0.122	0.442	0.339	0.184	0.404	0.275	0.039
k-sep_pssm	0.330	0.312	**0.561**	0.326	**0.452**	0.506	0.313	0.209	0.355	0.374	0.037
ksctriad	0.295	0.190	0.394	0.268	0.316	0.448	0.315	0.172	0.387	0.310	0.031
medp_pssm	0.167	0.164	0.417	0.431	0.216	0.309	0.375	0.023	0.285	0.265	0.045
moran	0.325	0.236	0.367	0.391	0.264	0.412	0.415	0.247	0.377	0.337	0.024
nmbroto	0.254	0.190	0.365	0.444	0.277	0.327	0.344	0.195	0.391	0.310	0.029
paac	0.230	0.181	0.502	0.417	0.255	0.505	0.500	0.262	0.347	0.356	0.043
pfam	0.325	0.286	0.486	0.432	0.430	0.203	0.273	0.132	0.399	0.329	0.039
pse_pssm	0.209	0.267	0.438	0.327	0.386	0.228	0.340	**0.314**	0.396	0.323	0.026
pssm_ac	**0.371**	**0.379**	0.393	0.426	0.324	**0.578**	0.451	0.210	**0.478**	**0.401**	0.034
pssm_cc	0.293	0.276	0.401	0.413	0.302	0.483	0.400	0.274	0.398	0.360	0.025
pssm_composition	0.220	0.221	0.455	0.451	0.333	0.373	**0.507**	0.142	0.437	0.349	0.043
qso	0.229	0.247	0.410	0.224	0.378	0.433	0.291	0.141	0.348	0.300	0.033
random200	0.071	0.080	-0.035	-0.033	-0.046	-0.100	-0.055	-0.111	0.054	-0.019	0.024
rpm_pssm	0.289	0.132	0.401	0.471	**0.440**	0.410	0.358	0.189	0.422	0.346	0.039
rpssm	0.055	0.197	0.355	0.456	0.187	0.340	0.418	0.087	0.338	0.270	0.048
spmap	0.338	0.259	0.431	0.405	0.194	**0.556**	0.489	0.153	**0.457**	0.365	0.046
taap	0.174	0.179	0.455	0.283	**0.431**	0.442	0.402	0.102	0.309	0.309	0.044
tpc	0.245	0.268	0.476	**0.499**	0.270	0.529	**0.528**	0.124	**0.448**	0.376	0.050
tpc_pssm	0.220	0.263	0.330	**0.492**	0.352	0.241	0.356	0.241	0.412	0.323	0.030
tri-gram_pssm	**0.350**	**0.349**	**0.541**	0.362	0.426	**0.545**	0.386	0.274	0.352	**0.398**	0.030

(b)

Model	ChEMBL id (only the numeric part) of the center compound of each compound cluster									Mean	Standard error
	44	50	83	91	104	633	808	116438	295698		
aac	0.122	0.049	0.183	0.102	0.243	0.279	0.091	0.161	0.136	0.152	0.025
aac_pssm	0.178	0.146	0.396	**0.488**	0.285	0.306	0.378	0.144	0.242	0.285	0.040
aadp_pssm	0.195	0.142	0.360	0.287	0.218	0.254	0.394	0.132	0.359	0.260	0.032
aatp_pssm	0.096	0.180	0.390	0.392	0.311	0.305	0.471	0.132	0.262	0.282	0.042
ab_pssm	0.296	0.311	0.385	0.372	0.201	0.470	0.370	0.195	0.295	0.322	0.030
apaac	0.159	0.292	**0.508**	0.430	0.358	**0.559**	0.438	0.200	0.300	0.360	0.045
cksaagp	0.215	0.045	0.238	0.248	0.240	0.216	0.146	0.229	0.283	0.207	0.024
cksaap	0.329	0.260	0.336	0.380	0.433	0.316	**0.491**	0.164	**0.506**	0.357	0.036
ctdc	0.197	-0.055	0.280	0.225	0.353	0.158	0.134	0.079	0.121	0.166	0.039
ctdd	0.151	0.226	0.375	0.357	0.225	0.310	0.414	0.237	0.024	0.258	0.041
ctdt	0.229	-0.013	0.233	0.099	**0.498**	0.300	0.284	0.096	0.053	0.198	0.052
ctriad	0.171	0.174	0.293	0.215	0.408	0.329	0.233	0.164	0.271	0.251	0.027
d_fpssm	0.205	0.123	0.242	0.388	0.025	0.140	0.394	0.033	0.362	0.212	0.048
dde	0.307	0.301	0.360	0.283	0.389	0.386	0.358	0.238	0.409	0.337	0.019
dp_pssm	0.228	**0.434**	0.370	0.363	0.229	0.329	0.335	0.186	0.313	0.310	0.027
dpc	0.202	0.106	0.308	0.291	0.236	0.333	0.263	0.193	0.337	0.252	0.025
dpc_pssm	0.195	0.142	0.418	0.287	0.218	0.254	0.394	0.132	0.358	0.266	0.035
edp_pssm	0.131	0.253	0.391	0.331	0.215	0.256	0.262	0.004	0.334	0.242	0.039
eedp_pssm	0.220	0.248	0.373	0.415	0.080	0.266	**0.491**	0.059	0.283	0.271	0.048
gaac	0.151	0.067	0.172	0.120	-0.089	0.230	0.111	0.065	0.206	0.115	0.032
gdpc	0.088	0.136	0.273	0.319	0.055	0.134	0.018	0.139	0.234	0.155	0.034
geary	0.277	0.201	0.357	0.327	0.308	0.288	0.280	0.195	0.358	0.288	0.020
gtpc	0.140	0.056	0.135	0.060	0.105	0.212	0.243	**0.259**	0.362	0.175	0.034
k-sep_pssm	0.277	0.347	**0.513**	0.288	0.418	0.474	0.456	0.241	0.312	**0.370**	0.033
ksctriad	0.124	0.231	0.406	0.277	0.031	0.402	0.338	0.148	0.262	0.247	0.043
medp_pssm	0.198	0.270	0.365	0.415	0.080	0.280	**0.491**	0.091	0.271	0.274	0.046

moran	0.264	0.215	0.345	0.347	0.181	0.315	0.384	0.184	0.303	0.282	0.025
nmbroto	0.231	0.228	0.354	0.289	0.439	0.282	0.386	0.236	0.258	0.300	0.025
paac	0.257	0.268	0.480	0.314	0.246	**0.528**	0.443	0.167	0.338	0.338	0.040
pfam	0.288	0.328	0.478	**0.452**	0.327	0.353	0.432	**0.367**	0.398	**0.380**	0.021
pse_pssm	0.248	0.280	0.367	0.334	0.144	0.348	0.342	0.176	0.343	0.287	0.027
pssm_ac	0.331	**0.353**	0.290	0.392	0.351	0.423	0.357	0.164	**0.428**	0.343	0.027
pssm_cc	0.294	**0.368**	0.484	0.311	**0.517**	0.446	0.373	0.192	0.349	**0.370**	0.034
pssm_composition	0.154	0.204	0.386	0.411	0.257	0.419	0.376	0.114	0.348	0.297	0.039
qso	0.220	0.078	0.249	0.211	0.346	0.268	0.073	0.213	0.113	0.197	0.031
random200	-0.044	-0.054	-0.110	-0.206	-0.046	-0.083	0.248	-0.140	0.205	-0.026	0.051
rpm_pssm	0.331	0.177	0.376	0.426	0.300	0.430	0.435	0.062	0.377	0.324	0.042
rpssm	0.133	0.312	0.260	**0.431**	0.249	0.305	0.361	0.076	0.361	0.276	0.038
spmap	**0.342**	0.260	0.486	0.254	0.249	**0.492**	0.394	0.026	0.370	0.319	0.048
taap	0.211	0.284	0.456	0.325	0.210	0.408	0.337	0.167	0.289	0.299	0.032
tpc	**0.356**	0.218	0.417	0.354	0.300	0.419	0.269	0.239	0.418	0.332	0.026
tpc_pssm	0.111	0.260	0.285	0.232	0.237	0.276	0.264	0.209	**0.425**	0.256	0.027
tri-gram_pssm	**0.340**	0.309	**0.487**	0.368	**0.444**	0.485	0.490	**0.284**	0.331	**0.393**	0.028

Table 3.2. Model performance scores in the medium-scale analysis (on the mDavis dataset). The best performance for each metric is shown in bold font.

Model	**RMSE**	**Spearman**	**F1-score**	**MCC**
seqvec	**0.794**	**0.571**	0.530	**0.445**
k-sep_pssm	0.817	0.545	**0.531**	0.434
unirep1900	0.823	0.541	0.510	0.418
apaac	0.831	0.532	0.519	0.418
unirep5700	0.831	0.531	0.506	0.412
transformer-avg	0.839	0.519	0.508	0.410
transformer-pool	0.840	0.515	0.506	0.412
qso	0.843	0.519	0.486	0.384
dde	0.845	0.508	0.480	0.384
geary	0.847	0.519	0.473	0.377
protvec	0.850	0.503	0.506	0.403
ctdd	0.851	0.503	0.484	0.376
ctriad	0.852	0.508	0.476	0.387
pfam	0.854	0.497	0.538	0.410
taap	0.863	0.492	0.467	0.349
spmap	0.871	0.491	0.477	0.362
random200	0.957	0.403	0.368	0.251
random200_random-ecfp4	0.968	0.388	0.346	0.235

Table 3.3. Model performance scores (in terms of the median corrected MCC) in the large-scale analysis on the protein family specific datasets of; **(a)** the random-split, **(b)** dissimilar-compound-split, and **(c)** the fully-dissimilar-split. The 3 best performances for each protein family are shown in bold font (ran200_ran-ecfp4: random200_ random-ecfp4, only-ran-ecfp4: only-random-ecfp4).

(a)

Random-split	epigenetic-regulators	hydrolases	ion-channels	membrane-receptors	other-enzymes	oxidoreductases	proteases	transcription-factors	transferases	transporters
apaac	**0.745**	0.755	0.697	0.689	0.754	0.692	**0.735**	0.714	0.696	0.728
ctdd	0.741	0.747	0.700	0.686	**0.757**	**0.694**	0.730	0.711	0.694	**0.732**
ctriad	0.734	0.749	0.701	0.686	0.752	**0.694**	0.731	0.706	0.694	0.726
dde	0.741	**0.756**	0.703	0.689	0.754	0.692	**0.735**	0.709	0.691	0.722
geary	0.733	**0.754**	0.701	**0.694**	0.752	0.681	**0.735**	**0.721**	0.696	0.728
k-sep_pssm	**0.757**	0.749	**0.709**	0.688	0.754	0.690	**0.735**	0.706	**0.704**	0.720
pfam	0.678	0.694	0.679	0.458	0.609	0.561	0.635	0.645	0.628	0.622
qso	0.734	**0.757**	0.700	0.685	0.754	0.690	0.733	0.704	0.691	0.728
random200	0.728	0.751	0.687	0.680	0.746	0.685	0.734	0.709	0.687	0.726
spmap	0.737	0.748	0.697	0.682	**0.757**	0.680	0.728	0.709	0.683	0.720
taap	**0.760**	0.747	**0.712**	0.687	0.750	0.693	**0.736**	**0.721**	0.700	**0.730**
protvec	0.741	0.742	0.703	**0.693**	**0.758**	**0.696**	0.733	0.714	0.696	0.726
seqvec	**0.745**	0.749	0.699	0.690	**0.757**	0.678	0.728	0.709	0.700	0.724
transformer-avg	0.736	0.748	**0.707**	**0.691**	**0.760**	0.681	0.734	0.701	0.702	0.718
transformer-pool	0.734	0.746	0.695	0.689	0.741	0.684	0.733	0.714	0.694	**0.730**
unirep1900	0.744	0.745	0.703	0.686	0.753	**0.696**	0.731	**0.716**	**0.703**	0.728
unirep5700	0.729	0.749	0.690	0.688	0.755	0.690	0.734	0.706	**0.705**	0.726
only-ecfp4	0.591	0.665	0.643	0.426	0.600	0.514	0.519	0.534	0.576	0.503
ran200_ran-ecfp4	0.382	0.481	0.400	0.256	0.449	0.401	0.320	0.265	0.319	0.235
only-ran-ecfp4	0.296	0.175	0.082	0.165	0.358	0.137	0.189	0.171	0.224	0.173

(b)

Dissimilar-compound-split	epigenetic-regulators	hydrolases	ion-channels	membrane-receptors	other-enzymes	oxidoreductases	proteases	transcription-factors	transferases	transporters
apaac	0.137	0.355	0.342	0.249	0.419	0.391	0.381	0.058	**0.358**	0.362
ctdd	0.021	**0.407**	0.311	0.241	0.386	0.423	**0.391**	0.048	0.342	0.405
ctriad	0.045	0.351	0.276	0.243	0.354	0.405	0.353	0.006	0.346	**0.425**
dde	0.089	0.371	0.327	0.223	0.359	0.398	0.341	0.071	0.340	0.403
geary	0.044	0.375	0.291	0.242	0.397	0.400	**0.394**	0.036	**0.347**	0.417
k-sep_pssm	0.239	**0.382**	0.419	**0.298**	0.381	**0.449**	0.354	0.071	0.318	0.368
pfam	**0.455**	0.329	**0.448**	0.146	**0.452**	0.339	0.319	**0.257**	0.308	0.366
qso	0.247	0.373	0.338	**0.278**	0.345	0.356	0.369	0.071	0.324	**0.419**
random200	0.152	**0.386**	0.273	0.266	0.348	0.409	0.341	0.103	0.341	0.388
spmap	0.158	0.351	0.289	**0.274**	0.361	0.395	0.383	0.036	0.335	0.369
taap	**0.289**	0.371	**0.434**	0.243	**0.443**	0.360	**0.398**	**0.187**	0.322	**0.438**
protvec	-0.024	0.348	**0.437**	0.222	0.363	0.381	0.366	**0.118**	0.344	0.390
seqvec	0.192	0.349	0.310	0.228	0.374	**0.435**	0.373	0.033	**0.359**	0.378
transformer-avg	0.075	0.349	0.288	0.250	**0.447**	**0.428**	**0.391**	0.043	0.328	0.390

transformer-pool	0.104	0.348	0.364	0.258	0.411	0.402	0.357	0.061	0.316	0.376
unirep1900	0.161	0.334	0.277	0.256	0.402	0.400	0.376	0.076	0.329	0.370
unirep5700	0.061	0.350	0.289	0.255	0.387	0.374	0.346	0.090	0.320	0.376
only-ecfp4	**0.428**	0.244	0.243	0.168	0.419	0.281	0.307	0.058	0.306	0.309
ran200_ran-ecfp4	-0.076	0.254	0.293	0.133	0.210	0.270	0.138	0.083	0.178	0.284
only-ran-ecfp4	0.004	-0.020	0.018	0.001	-0.028	-0.015	-0.008	0.016	-0.020	-0.035

(c)

Fully-dissimilar-split	**epigenetic-regulators**	**hydrolases**	**ion-channels**	**membrane-receptors**	**other-enzymes**	**oxidoreductases**	**proteases**	**transcription-factors**	**transferases**	**transporters**
apaac	0.403	0.156	0.146	0.243	0.129	-0.044	0.192	0.063	0.300	**0.240**
ctdd	0.396	0.132	-0.074	0.253	0.162	0.074	0.207	0.101	0.247	-0.017
ctriad	**0.420**	**0.203**	0.086	0.220	0.125	0.030	0.212	**0.238**	0.273	**0.276**
dde	0.375	0.170	0.124	0.206	**0.232**	-0.050	0.195	0.150	0.295	0.029
geary	0.319	0.160	0.155	0.195	0.172	0.044	**0.268**	0.066	0.275	-0.027
k-sep_pssm	0.252	0.181	0.157	0.137	0.043	0.052	-0.134	0.086	0.300	**0.297**
pfam	**0.446**	**0.208**	**0.174**	**0.270**	0.221	0.088	0.142	0.156	0.301	0.198
qso	0.397	0.111	0.044	0.166	0.187	0.141	0.215	0.129	0.300	0.202
random200	0.289	0.040	0.146	0.226	**0.282**	0.070	0.149	0.051	0.284	-0.194
spmap	0.361	0.103	**0.182**	0.213	0.114	0.015	0.209	0.091	0.287	0.118
taap	0.289	0.155	**0.181**	**0.286**	0.208	-0.028	0.205	0.129	0.310	0.194
protvec	0.275	0.146	**0.174**	0.235	0.131	0.077	0.184	**0.160**	0.301	0.204
seqvec	0.372	0.154	0.032	0.155	0.199	0.018	0.046	0.150	0.276	-0.023
transformer-avg	0.367	0.129	0.058	**0.265**	0.176	0.092	**0.227**	0.144	0.311	0.187
transformer-pool	0.403	0.148	-0.052	0.244	0.079	**0.133**	0.170	0.117	**0.313**	-0.040
unirep1900	0.325	0.186	0.143	0.217	0.205	**0.132**	0.159	-0.008	**0.332**	0.004

Table 5.4. Comparison of test performance scores for different architecture alternatives of HetCPI models on the transferases bioactivity dataset. The best performance for each split is shown in bold font.

Method	**Data split**	**RMSE**	**Med. Cor. RMSE**	**Spearman**	**MCC**	**Med. Cor. MCC**
RF_SELFormer	FDS*	1.147	1.125	0.418	0.056	0.285
RF_ECFP4	FDS	1.225	1.118	0.438	0.010	0.315
DP_SELFormer	FDS	1.205	1.163	0.378	0.206	0.284
DP_ECFP4	FDS	1.158	1.167	0.446	0.325	0.341
3FC_SELFormer	FDS	**1.143**	**1.106**	0.443	0.243	0.317
3FC_ECFP4	FDS	1.207	1.184	0.363	0.261	0.258
HetCPI_SELFormer	FDS	1.149	1.110	0.517	**0.340**	**0.383**
HetCPI_ECFP4	FDS	1.208	1.191	**0.520**	0.333	0.361
HetCPI-3FC_SELFormer	FDS	1.191	1.228	0.387	0.253	0.268
HetCPI-3FC_ECFP4	FDS	1.166	1.182	0.409	0.263	0.274

RF_SELFormer	DCS*	1.161	1.123	0.419	0.235	0.308
RF_ECFP4	DCS	1.238	1.105	0.454	0.154	0.328
DP_SELFormer	DCS	1.193	1.151	0.378	0.163	0.262
DP_ECFP4	DCS	1.202	1.180	0.382	0.210	0.221
3FC_SELFormer	DCS	1.187	1.134	0.419	0.232	0.281
3FC_ECFP4	DCS	1.312	1.145	0.429	0.199	0.299
HetCPI_SELFormer	DCS	**1.092**	**1.071**	**0.525**	**0.365**	**0.391**
HetCPI_ECFP4	DCS	1.222	1.205	0.475	0.307	0.315
HetCPI-3FC_SELFormer	DCS	1.184	1.103	0.479	0.284	0.327
HetCPI-3FC_ECFP4	DCS	1.311	1.195	0.442	0.292	0.337
RF_SELFormer	RS*	0.828	0.831	0.765	0.576	0.576
RF_ECFP4	RS	**0.643**	**0.648**	**0.861**	**0.695**	**0.693**
DP_SELFormer	RS	0.909	0.904	0.696	0.509	0.511
DP_ECFP4	RS	0.748	0.749	0.817	0.646	0.642
3FC_SELFormer	RS	0.915	0.854	0.732	0.524	0.548
3FC_ECFP4	RS	0.682	0.683	0.840	0.665	0.668
HetCPI_SELFormer	RS	0.785	0.787	0.783	0.605	0.604
HetCPI_ECFP4	RS	0.730	0.728	0.823	0.652	0.653
HetCPI-3FC_SELFormer	RS	0.796	0.777	0.785	0.592	0.603
HetCPI-3FC_ECFP4	RS	0.722	0.712	0.829	0.648	0.654

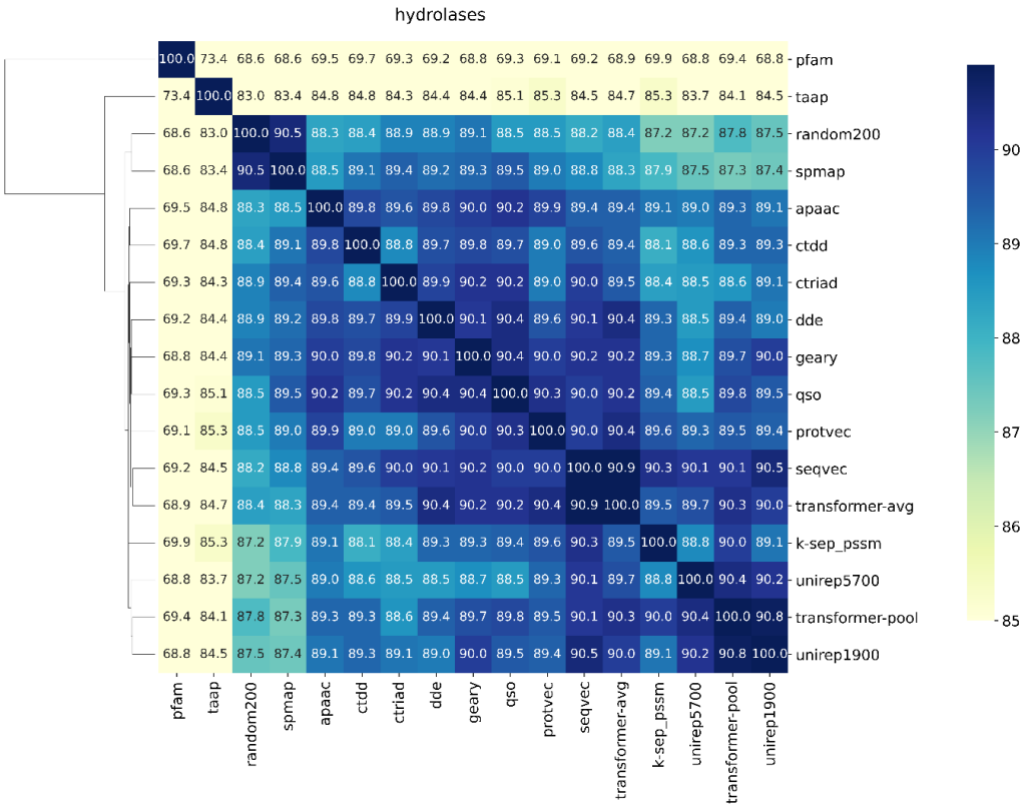
* FDS: fully-dissimilar-split, DCS: dissimilar-compound-split, and RS: random-split datasets.

APPENDIX B

CHAPTER 3: CLUSTERED HEATMAPS OF PROTEIN FAMILY-SPECIFIC PCM MODELS

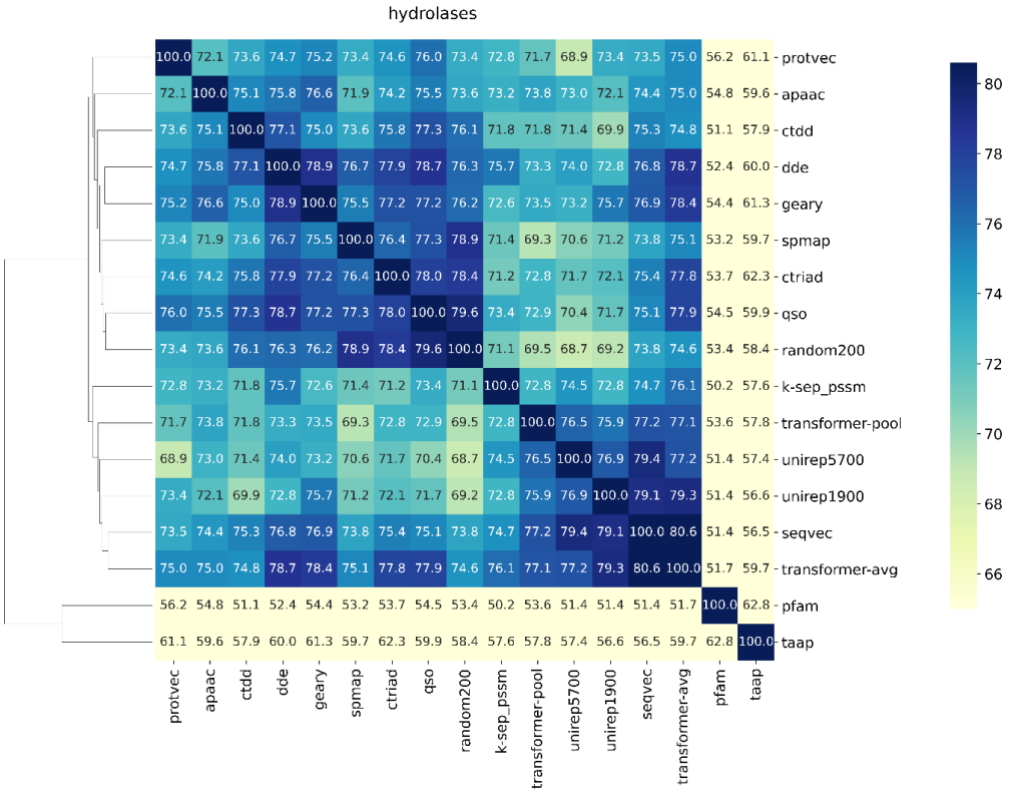
(a)

Random-split



(b)

Dissimilar-compound-split



(c)

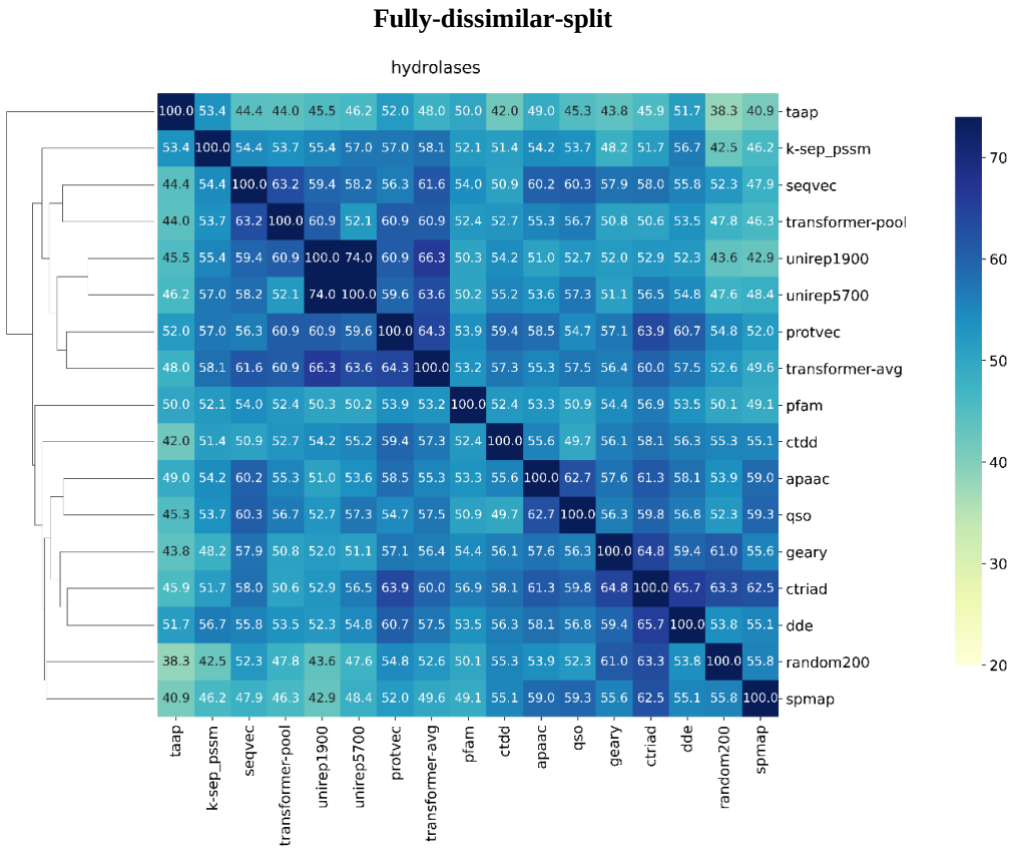


Figure 3.1. Clustered heatmaps of hydrolases for protein families on **(a)** the random-split, **(b)** dissimilar-compound-split, and **(c)** fully-dissimilar-split datasets.

CURRICULUM VITAE

PERSONAL INFORMATION

Surname, Name: Ataş Güvenilir, Heval

Nationality: Turkish

EDUCATION

Degree	Institution	Year of Graduation
PhD	METU Medical Informatics	2023 (expected)
MS	METU Biotechnology	2016
BS	METU Molecular Biology and Genetics	2014
High School	Kabatas High School	2009

WORK EXPERIENCE

Year	Place	Enrollment
2020-Present	Biological Data Science Lab., Hacettepe University	Project Researcher
2017-2020	Cancer Systems Biology lab., METU	Project Researcher

SCHOLARSHIPS

2021-2023: TÜBİTAK-ARDEB-2247A (120C123)

2018-2022: YÖK 100/2000

2016-2020: TÜBİTAK - BİDEB 2211E

2014-2016: TÜBİTAK - BİDEB 2210E

2009-2014: TÜBİTAK - BİDEB 2205

FOREIGN LANGUAGES

Advanced English

PUBLICATIONS

Atas Guvenilir, H., & Doğan, T. (2023). How to approach machine learning-based prediction of drug/compound–target interactions. Journal of Cheminformatics, 15(1), 1–36. https://doi.org/10.1186/S13321-023-00689-W

Unsal, S., Atas, H., Albayrak, M., Turhan, K., Acar, A. C., & Doğan, T. (2022). Learning functional properties of proteins with language models. Nature Machine Intelligence, 4(3), 227–245. https://doi.org/10.1038/s42256-022-00457-9

Cetin-Atalay, R., Kahraman, D. C., Nalbat, E., Rifaioglu, A. S., Atakan, A., Donmez, A., Atas, H., … Doğan, T. (2021). Data Centric Molecular Analysis and Evaluation of Hepatocellular Carcinoma Therapeutics Using Machine Intelligence-Based Tools. Journal of Gastrointestinal Cancer, 52(4), 1266–1276. https://doi.org/10.1007/S12029-021-00768-X

Doğan, T., Güzelcan, E. A., Baumann, M., Koyas, A., Atas, H., Baxendale, I. R., … Cetin-Atalay, R. (2021). Protein domain-based prediction of drug/compound–target interactions and experimental validation on LIM kinases. PLOS Computational Biology, 17(11), e1009171. https://doi.org/10.1371/JOURNAL.PCBI.1009171

Doğan, T., Atas, H., Joshi, V., Atakan, A., Rifaioglu, A. S., Nalbat, E., … Atalay, V. (2021). CROssBAR: comprehensive resource of biomedical relations with knowledge graph representations. Nucleic Acids Research, 49(16), e96–e96. https://doi.org/10.1093/NAR/GKAB543

Cichońska, A., Ravikumar, B., Allaway, R. J., Wan, F., Park, S., Isayev, O., … The IDG-DREAM Drug-Kinase Binding Prediction Challenge Consortium, Aittokallio, T. (2021). Crowdsourced mapping of unexplored target space of kinase inhibitors. Nature Communications, 12(1), 1–18. https://doi.org/10.1038/s41467-021-23165-1

Rifaioglu, A. S., Atas, H., Martin, M. J., Cetin-Atalay, R., Atalay, V., & Doğan, T. (2018). Recent applications of deep learning and machine intelligence on in silico drug discovery: Methods, tools and databases. Briefings in Bioinformatics. https://doi.org/10.1093/bib/bby061

Atas, H., Tuncbag, N., & Doğan, T. (2018). Phylogenetic and Other Conservation-Based Approaches to Predict Protein Functional Sites. Methods in Molecular Biology Computational Drug Discovery and Design, 51-69. https://doi.org/10.1007/978-1-4939-7756-7_4

Durusu, İ. Z., Hüsnügil, H. H., Ataş, H., Biber, A., Gerekçi, S., Güleç, E. A., & Özen, C. (2017). Anti-cancer effect of clofazimine as a single agent and in combination with cisplatin on U266 multiple myeloma cell line. Leukemia Research, 55, 33–40. https://doi.org/10.1016/j.leukres.2017.01.019