



Swansea University  
Prifysgol Abertawe

SWANSEA UNIVERSITY

# Data mining patterns of receptor-drug interactions across vast biological and chemical space

*James Arthur Witts*

Submitted to Swansea University in  
fulfilment of the requirements for the Degree of  
Doctor of Philosophy

2019

Copyright: The author, James A. Witts, 2020.

## Abstract

The aim of the project was to determine how machine learning tools can assist in the process of drug discovery and *in silico* screening. As the development costs and attrition rates for candidate compounds can be high, a method of predicting likelihood for approval or for therapeutic promise with machine learning and high performance computing tools could be of great benefit to medical researchers, biotech, chemical and pharmaceutical industries. The first phase of the project was to determine if knowledge of the recorded *in vitro* protein interactions of particular compounds (listed in DrugBank and ToxCast) would be sufficient to determine whether a candidate compound could be designated as having a good (i.e approved drug) or a bad (i.e toxic) profile. The learning models assessed showed promise in correctly designating candidate compounds based on a small number of proteins used for pharmacological profiling, with over 90% overall profiling prediction accuracy in the best case, however the vast majority of interactions between compounds and proteins are unknown and so a predictive approach is needed. The second phase of the project was to provide a method for predicting these hitherto unknown interactions, by predicting protein-compound interaction pairs through clustering techniques. Several clustering methods based on protein and compound similarity measurement techniques were investigated and found that when tested on blind *in vitro* interactions, approximately half on average were detected successfully, highlighting the promising potential to strengthen the predictive profiling models. The third and final phase of the project focused on the development of a compound and protein target prediction interface, TargetPredict (<http://proteins.swan.ac.uk/cheminf/>), which incorporates the data and methodologies presented throughout the whole project into a single centralised source, to provide the sector for the first time with a unified tool, avoiding the onerousness of current approaches that either require the use of multiple often incompatible websites or require extensive coding experience.

## Declaration

This work has not previously been accepted in substance for any degree and is not being concurrently submitted in candidature for any degree.

Signed: \_\_\_\_\_

Date: \_\_\_\_\_

## Statement 1

This thesis is the result of my own investigations, except where otherwise stated. Where correction services have been used, the extent and nature of the correction is clearly marked in a footnote(s).

Other sources are acknowledged by footnotes giving explicit references. A bibliography is appended.

Signed: \_\_\_\_\_

Date: \_\_\_\_\_

## Statement 2

I hereby give consent for my thesis, if accepted, to be available for photocopying and for inter-library loan, and for the title and summary to be made available to outside organisations.

Signed: \_\_\_\_\_

Date: \_\_\_\_\_

# Contents

<b>Acknowledgements</b>	<b>vii</b>
<b>List of Figures</b>	<b>viii</b>
<b>List of Tables</b>	<b>xi</b>
<b>Table of Abbreviations</b>	<b>xv</b>
<b>1 Introduction and Background</b>	<b>1</b>
1.1 Introduction . . . . .	1
1.1.1 Aims and objectives . . . . .	1
1.1.1.1 Objective 1 . . . . .	1
1.1.1.2 Objective 2 . . . . .	2
1.1.1.3 Objective 3 . . . . .	2
1.2 Drug Regulation Process . . . . .	3
1.2.1 UK Drug Regulation History . . . . .	4
1.2.2 UK Drug Approval Process . . . . .	5
1.2.3 FDA Drug Regulation History . . . . .	7
1.2.4 FDA Approval Process . . . . .	8
1.2.5 Drug Assessment Phases . . . . .	9
1.2.5.1 Pre-clinical Phases . . . . .	10
1.2.5.2 Clinical Trials . . . . .	10
1.2.6 Drug Development Cost and Attrition . . . . .	11
1.2.7 Approaches to reducing attrition . . . . .	13
1.3 Compound and Protein Database Repositories . . . . .	16
1.3.1 DrugBank, T3DB and HMDB . . . . .	16
1.3.2 ToxCast . . . . .	18
1.3.3 ChEMBL . . . . .	19
1.3.4 Matador . . . . .	19
1.3.5 CTDBase . . . . .	20
1.3.6 BindingDB . . . . .	21
1.3.7 UniProt . . . . .	21
1.3.8 PubChem . . . . .	22
1.3.9 KEGG . . . . .	23
1.4 Program Language and Tool Analysis . . . . .	24



1.4.1	R . . . . .	24
1.4.2	Python . . . . .	26
1.4.3	Matlab/Octave . . . . .	26
1.4.4	MySQL . . . . .	27
1.4.5	WEKA . . . . .	27
1.4.6	OpenBabel . . . . .	28
1.4.7	HPC Wales/Supercomputing Wales . . . . .	28
1.5	Summary . . . . .	29
1.6	References . . . . .	29
<b>2</b>	<b>DrugReferenceDatabase - Compiling interactions across various drug and protein repositories</b>	<b>36</b>
2.1	Introduction . . . . .	36
2.2	Drug and Protein Repository Analysis . . . . .	36
2.2.1	Main Database Sources . . . . .	37
2.2.1.1	DrugBank . . . . .	37
2.2.1.2	ToxCast . . . . .	39
2.2.1.3	PubChem Compounds . . . . .	42
2.2.2	Additional Database Sources . . . . .	45
2.2.2.1	T3DB . . . . .	45
2.2.2.2	Matador . . . . .	46
2.2.2.3	BindingDB . . . . .	46
2.2.2.4	ChEMBL . . . . .	47
2.2.2.5	UniProt FASTA Files . . . . .	49
2.2.2.6	CTDBase . . . . .	50
2.2.2.7	PubChem BioAssay . . . . .	53
2.3	Database Design and Implementation . . . . .	55
2.3.1	Implementing the initial design . . . . .	57
2.3.1.1	Step 1: Creating the Detailed Record Tables . . . . .	57
2.3.1.2	Step 2: Creating the Link Tables . . . . .	58
2.3.1.3	Step 3: Transfer to MySQL . . . . .	59
2.3.1.4	Limitations . . . . .	60
2.3.2	Implementing the revised design . . . . .	61
2.3.2.1	Creation of Hit, Miss and Conflict Tables . . . . .	62
2.3.2.2	Analysis of Interactions and Conflicts . . . . .	63
2.3.2.3	Resolving Conflicts . . . . .	65
2.3.2.4	Limitations . . . . .	68
2.4	Discussion . . . . .	70
2.4.1	Further Work . . . . .	71
2.5	References . . . . .	72
<b>3</b>	<b>Using machine learning approaches for profiling drug-protein interactions</b>	<b>74</b>
3.1	Introduction . . . . .	74
3.1.1	Background . . . . .	74

3.2	Methodology . . . . .	77
3.2.1	Software and Hardware . . . . .	78
3.3	Active Pairs Only . . . . .	79
3.3.1	Initial Findings . . . . .	81
3.3.2	Classifier Results . . . . .	83
3.3.3	Discussion . . . . .	84
3.4	Active and Inactive Pairs . . . . .	88
3.4.1	Initial Findings . . . . .	89
3.4.2	Classifier Results . . . . .	91
3.4.3	Discussion . . . . .	92
3.5	Chemical Properties . . . . .	93
3.5.1	Initial Findings . . . . .	93
3.5.2	Classifier Results . . . . .	95
3.5.3	Discussion . . . . .	97
3.6	Blind Testing . . . . .	98
3.6.1	Classifier Results . . . . .	99
3.6.2	Discussion . . . . .	103
3.7	Conclusions . . . . .	103
3.8	References . . . . .	104
<b>4</b>	<b>Refinement of drug profiling approaches by incorporation of exper-</b>	
	<b>imental assay results</b>	<b>108</b>
4.1	Introduction . . . . .	108
4.2	Reassessing the DrugReferenceDatabase . . . . .	109
4.2.1	Potency Values . . . . .	109
4.2.2	Repository Parameter Review . . . . .	111
4.2.2.1	DrugBank . . . . .	111
4.2.2.2	ToxCast . . . . .	111
4.2.2.3	BindingDB . . . . .	112
4.2.2.4	ChEMBL . . . . .	113
4.2.2.5	Repositories not considered . . . . .	114
4.2.3	Initial Findings . . . . .	115
4.3	Revised Analysis . . . . .	117
4.3.1	Active Interactions Only . . . . .	118
4.3.1.1	Discussion . . . . .	121
4.3.2	Active Interactions with Chemical Properties . . . . .	122
4.3.2.1	Discussion . . . . .	125
4.4	Database Schema Redesign . . . . .	129
4.5	Conclusions and Further Work . . . . .	132
4.6	References . . . . .	132
<b>5</b>	<b>Protein-compound interaction prediction based on compound and</b>	
	<b>protein similarity</b>	<b>134</b>
5.1	Introduction . . . . .	134
5.2	Similarity Comparison Review . . . . .	135

---

5.2.1	Initial Approach . . . . .	135
5.2.2	Alternative Similarity Clustering Methods . . . . .	137
5.2.2.1	WNN-GIP . . . . .	137
5.2.2.2	NetLapRLS . . . . .	137
5.2.2.3	BLM-NII . . . . .	138
5.2.2.4	MSCMF . . . . .	138
5.2.2.5	KBMF2K . . . . .	139
5.2.2.6	SELF-BLM . . . . .	139
5.2.2.7	NRLMF and PyDTI . . . . .	140
5.2.3	Alternative Compound Comparison Methods . . . . .	140
5.2.4	Alternative Protein Comparison Methods . . . . .	141
5.3	In Silico Docking Pipeline . . . . .	143
5.4	Methodology . . . . .	145
5.5	Results . . . . .	147
5.5.1	AUC and AUPR Performance . . . . .	147
5.5.2	Prediction of New Interactions . . . . .	148
5.5.3	Verification of New Interactions . . . . .	150
5.5.3.1	Initial Findings . . . . .	150
5.5.3.2	Docking Results . . . . .	154
5.6	Conclusions . . . . .	159
5.6.1	Further Work . . . . .	160
5.7	References . . . . .	161
<b>6</b>	<b>TargetPredict: Design and Implementation of an Interface for Drug Profiling and Interaction Similarity Clustering</b>	<b>163</b>
6.1	Introduction . . . . .	163
6.2	Requirements . . . . .	163
6.3	Constructing the Interface . . . . .	164
6.4	Interface Design . . . . .	165
6.4.1	Compound Target Browser . . . . .	166
6.4.2	Compound/Protein Similarity . . . . .	168
6.4.2.1	Nearest Neighbours . . . . .	169
6.4.2.2	Interaction Prediction . . . . .	170
6.5	Implementation . . . . .	170
6.5.1	Compound Browser . . . . .	170
6.5.2	Nearest Neighbours Predictions . . . . .	173
6.5.3	Interaction Prediction . . . . .	175
6.6	Discussion . . . . .	178
6.6.1	Further Work . . . . .	180
6.7	References . . . . .	181
<b>7</b>	<b>Discussion</b>	<b>183</b>
7.1	Introduction . . . . .	183
7.2	Objective Review . . . . .	183
7.2.1	Objective 1 Review . . . . .	183

7.2.2 Objective 2 Review . . . . .	185
7.2.3 Objective 3 Review . . . . .	187
7.3 Further Work . . . . .	189
7.4 Concluding Remarks . . . . .	190
7.5 References . . . . .	190
<b>Appendix</b>	<b>194</b>
<b>Bibliography</b>	<b>195</b>

# Acknowledgements

I would like to thank my supervisor Dr Jonathan Mullins for providing me not only with the ability for undertaking this thesis, but also for providing continuous support throughout the last four years. His assistance was vital for reaching this final point, and I am extremely grateful.

I would also like to thank Karl Austin-Muttitt, John Walshe, William Edwards and Sam West, who have provided valuable advice during my first steps into this new venture, and for the friendship that has formed from these years of hard work.

Finally, I would like to give thanks to my family, who have not only provided infinite patience and meals, but also comfort when times were hard.

# List of Figures

1.1	Graphical representation of objective 2 . . . . .	3
1.2	Graphical representation of objective 3 . . . . .	3
1.3	A Thalidomide sample packet that was distributed to women in early pregnancy . . . . .	4
1.4	Packaging for a variety of Thalidomide used from 2006 . . . . .	5
1.5	Overview of the SMC and NICE Appraisal process . . . . .	7
1.6	Overview of the drug development and approval process for FDA approval . . . . .	9
1.7	Illustration of initial process to generate <i>in silico</i> compound-protein interactions . . . . .	14
1.8	Screenshot of DrugBank website displaying information on the drug Abacavir . . . . .	17
1.9	CTDBase Data curation process . . . . .	20
1.10	Screenshot of UniProt website displaying information on the Androgen Receptor protein . . . . .	22
2.1	Screenshot of DrugBank Protein Target Identifiers . . . . .	38
2.2	Screenshot of the DrugBank Interaction set after pre-processing was performed . . . . .	39
2.3	Screenshot of a segment of the ToxCast <i>in vitro</i> interaction summary matrix . . . . .	40
2.4	Interaction Set with ToxCast drugs added . . . . .	41
2.5	A view of the contents of a PubChem SDF for aspirin . . . . .	44
2.6	Bash Command to rename batch SDF files generated by OpenBabel to PubChem Compound IDs . . . . .	44
2.7	A screenshot of T3DB Compounds . . . . .	46
2.8	A screenshot of T3DB Targets . . . . .	46
2.9	A screenshot of the Matador interaction set . . . . .	47
2.10	A screenshot of a segment of the BindingDB interactions set and the fields which were used for the purposes of finding interactions of interest . . . . .	47
2.11	SQL Query used to gather PubChem Substance references of ChEMBL compounds . . . . .	48
2.12	SQL Query used to gather UniProt interactions of ChEMBL compounds that contain a reference to PubChem . . . . .	48
2.13	UniProt results when filtered to the human organism only . . . . .	49
2.14	Text entry of the protein Androgen receptor . . . . .	50

2.15	Screenshot of the CTD Batch Query interface . . . . .	52
2.16	Screenshot of the results of querying diseases linked to human proteins	52
2.17	Screenshot of a segment of the CTDBase Interactions . . . . .	53
2.18	A screenshot of BioAssay active results of D(2) dopamine receptor (Uniprot Accession Code P14416) . . . . .	55
2.19	A screenshot of BioAssay inactive results of D(2) dopamine receptor (Uniprot Accession Code P14416) . . . . .	55
2.20	Diagram demonstrating a simplified conceptual overview of the database and the relationships required to be established . . . . .	56
2.21	Entity Relationship Diagram of the first revision of the database . . .	57
2.22	Graphical example of creating a detailed record table from UniProt Accession codes . . . . .	58
2.23	Graphical example of creating a link table between drugs and protein targets . . . . .	59
2.24	Example SQL Query to obtain all targets from the database . . . . .	59
2.25	Entity relationship diagram of changes taken to document interactions	61
2.26	Graphical example of creating the database's active interaction table	63
2.27	Modification to the SQL query to obtain activity comments from ChEMBL . . . . .	66
2.28	Entry for a non-live PubChem record CID 653, associated with the Matador Database under the chemical name DTC . . . . .	69
2.29	Bar chart demonstrating number of non-live PubChem Compound IDs introduced from data obtained from extracting the compound repositories . . . . .	70
3.1	Illustrative example of an interaction matrix . . . . .	79
3.2	Graph of the distribution of "Good" (DrugBank) and "Bad" (ToxCast) profile compounds based on active interactions only . . . . .	80
3.3	Example query used to obtain active compound-protein pairings for "Good-Profile" drugs . . . . .	80
3.4	Illustrated example of the process used to convert interactions found in the database to a format suitable for use in WEKA . . . . .	81
3.5	Segment of the J48 Classification tree applied to Panel 44 interactions, before class balancing . . . . .	86
3.6	Segment of the J48 Classification tree applied to Panel 44 interactions, after class balancing . . . . .	87
3.7	Illustrated example of the revised process used to convert active and inactive compound-protein pairs found in the database to a format suitable for use within WEKA . . . . .	88
3.8	Query used to obtain inactive compound-protein pairings for "Good- Profile" drugs . . . . .	89
3.9	Graph of the number of attributes within each panel after incorporating chemical properties . . . . .	94
3.10	Graph of the distribution of "Good" (DrugBank) and "Bad" (ToxCast) profile compounds after parsing for chemical properties . . . . .	95

3.11	Graph comparing results of the J48 classifier models when applied to the training set and blind testing set. . . . .	100
4.1	Example of documentation of a target in DrugBank (Mu-type opioid receptor on the drug Oxycodone) . . . . .	111
4.2	Example of the UniProt targets associated with ToxCast assays, where some assays contain multiple proteins . . . . .	112
4.3	Revised Query of the ChEMBL database to retrieve activity values from proteins of interest . . . . .	113
4.4	JRIP ruleset for Panel 331 on the thresholded and unbalanced dataset	122
4.5	JRIP Ruleset for Panel 331 on the unbalanced protein and chemical property dataset . . . . .	126
4.6	Segment of the J48 Classification tree on the unbalanced Panel 331 dataset after thresholding was applied . . . . .	128
4.7	A revised Entity Relationship diagram of the DrugReferenceDatabase incorporating assay results to determine targets of interest . . . . .	131
5.1	Predicting interaction profiles via similarity clustering . . . . .	136
5.2	High level summary of the <i>in silico</i> pipeline . . . . .	144
5.3	Simulated binding of PubChem Compound ID 11982778 (YM218) with UniProt ID P11229 (ACM1) . . . . .	156
5.4	Simulated binding of PubChem CompoundID 3744660 (36673-16-2) with UniProt ID Q06187 (BTK_HUMAN) . . . . .	157
5.5	Plot of the Empirical Distribution Function between the top 10 and bottom 10 predictions of the NRLMF method . . . . .	158
5.6	Assay concentration values against the <i>in silico</i> pipeline binding energies	159
6.1	Simple illustration of the design of the interface . . . . .	166
6.2	Simple illustration of the design of the Compound Browser Tab . . .	167
6.3	Design of the Nearest Neighbour Tab . . . . .	168
6.4	Design of the Interaction Prediction Tab . . . . .	169
6.5	Example of a reactive variable in the Interface server code . . . . .	171
6.6	Screenshot of the Compound Browser when ChEMBL is selected . . .	172
6.7	Screenshot of the Compound Browser when ToxCast is selected . . .	172
6.8	Screenshot of the Nearest Neighbour Predictions tab with a custom compound . . . . .	174
6.9	Screenshot of the Interaction Prediction Tab with a custom compound	176
6.10	Screenshot of PubChem's Similarity Search Platform . . . . .	179
6.11	Screenshot of the DrugBank Chemical Structure Search platform . . .	179



## List of Tables

1.1	Drug Success Rates . . . . .	11
1.2	Clinical Phase Costs . . . . .	12
1.3	UniProt Entry Codes for Bowes et. al panel for pharmacological profiling	15
2.1	Fields of interest used for pre-processing the DrugBank protein interaction sets . . . . .	38
2.2	Fields of interest that were used in the compound reference tables in ToxCast . . . . .	40
2.3	Fields of interest that were used in the protein assay table in ToxCast	40
2.4	DrugBank records which were manually annotated to PubChem . . .	43
2.5	List of Compartments used for gathering tissue groups for human protein text files gathered via UniProt . . . . .	50
2.6	Fields of interest within the CTDBase files . . . . .	53
2.7	The arguments passed to PubChem to retrieve BioAssay interactions or inactive bindings for a single UniProt accession code . . . . .	54
2.8	5 target and inactive pair conflicts found within DrugBank and ToxCast	60
2.9	Summary of all active and inactive results found in the repositories considered based on the DrugBank and ToxCast protein panels . . . .	63
2.10	Summary of all conflicts for the DrugBank and ToxCast protein panels detected in the repositories considered . . . . .	64
2.11	Distribution of the number of repositories which flagged a discovered protein-compound conflict as active when an inactive pairing is indicated within another repository . . . . .	64
2.12	Table investigating conflicts which have been flagged as active by one repository and inactive by one repository . . . . .	65
2.13	Summary of search terms used to gather active and inactive protein-compound pairs from ChEMBL . . . . .	66
2.14	Summary of all active and inactive results found in the repositories considered based on the DrugBank and ToxCast protein panels, after revisions . . . . .	67
2.15	Summary of all conflicts for the DrugBank and ToxCast protein panels detected in the repositories considered, after revisions . . . . .	67
2.16	Distribution of the number of repositories which flagged a discovered protein-compound conflict as active when an inactive pairing is indicated within another repository, after revisions . . . . .	68

2.17	Revised numbers of conflicts which have been flagged as active by one repository and active by one other repository following consideration of activity comments from ChEMBL . . . . .	68
3.1	Caption describing the classifiers used within WEKA for all experiments	78
3.2	Top 15 interacting proteins in Panel 44 . . . . .	82
3.3	Top 15 interacting proteins in Panel 331 . . . . .	82
3.4	Top 15 interacting proteins in the Pharmacology Panel . . . . .	83
3.5	Classifier results for active interactions only where attributes were unmodified in weightings . . . . .	84
3.6	Classifier results for active interactions only where class balancing was performed . . . . .	84
3.7	Top 15 proteins with inactive results in Panel 44 . . . . .	89
3.8	Top 15 proteins with inactive results in Panel 331 . . . . .	90
3.9	Top 15 proteins with inactive results within the Pharmacology Panel	90
3.10	Classifier results when inactive results (excluding inconclusive results) are incorporated, and where attributes were unmodified in weightings	91
3.11	Classifier results when inactive results (excluding inconclusive results) are incorporated, and class balancing was performed . . . . .	92
3.12	Classifier results when inactive results (including inconclusive results) are incorporated, and where attributes were unmodified in weightings	92
3.13	Classifier results when inactive results (including inconclusive results) are incorporated, and class balancing was performed . . . . .	92
3.14	List of five compounds which could not be parsed through the Mordred library . . . . .	94
3.15	Classifier results using only chemical property attributes unmodified in weightings . . . . .	96
3.16	Classifier results using only chemical property attributes where class balancing was performed . . . . .	96
3.17	Classifier results using chemical properties in conjunction with protein active flags where attributes were unmodified in weightings . . . . .	97
3.18	Classifier results using chemical properties in conjunction with protein active flags where class balancing was performed . . . . .	97
3.19	Distribution of classes on the blind testing set . . . . .	98
3.20	Blind testing classifier results on active interactions only where attributes were unmodified in weightings . . . . .	101
3.21	Blind testing classifier results on active interactions only where class balancing was performed . . . . .	101
3.22	Blind testing classifier results when inactive results (including inconclusive results) are incorporated, and where attributes were unmodified in weightings . . . . .	101
3.23	Blind testing classifier results when inactive results (including inconclusive results) are incorporated, and class balancing was performed .	101
3.24	Blind testing classifier results using only chemical property attributes unmodified in weightings . . . . .	102

3.25	Blind testing classifier results using only chemical property attributes where class balancing was performed . . . . .	102
3.26	Blind testing classifier results using chemical properties in conjunction with protein active flags where attributes unmodified in weightings . .	102
3.27	Blind testing classifier results using chemical properties in conjunction with protein active flags where class balancing was performed . . . . .	102
4.1	Definitions of some common measures of potency . . . . .	110
4.2	Summary of all active results found before thresholding . . . . .	116
4.3	Summary of all active results found after thresholding to assay concentration levels under or equal to 10 micromolar . . . . .	116
4.4	Top 15 interacting proteins in Panel 44 after thresholding was applied	116
4.5	Top 15 interacting proteins in Panel 331 after thresholding was applied	117
4.6	Top 15 interacting proteins in the Pharmacology Panel after thresholding was applied . . . . .	117
4.7	Distribution of compounds for the training and testing sets used for the thresholding analysis . . . . .	118
4.8	Classifier results for interactions thresholded under 10 micromolar potencies of Ki, Kd, EC50, IC50, and AC50, unmodified in weightings	120
4.9	Classifier results on interactions thresholded under 10 micromolar potencies of Ki, Kd, EC50, IC50, and AC50, where class balancing was performed . . . . .	120
4.10	Details of proteins referenced by the JRIP ruleset . . . . .	122
4.11	Classifier results on interactions thresholded under 10 micromolar potencies of Ki, Kd, EC50, IC50, and AC50, in addition to chemical properties with no modifications to weights . . . . .	124
4.12	Classifier results on interactions thresholded under 10 micromolar potencies of Ki, Kd, EC50, IC50, and AC50, in addition to chemical properties, where class balancing was performed . . . . .	124
4.13	Details of proteins referenced by the full J48 decision tree in Figure 4.6	127
5.1	Performance of the methods tested by Yamanishi et al . . . . .	137
5.2	Description of the elements used in the PubChem Fingerprint System	141
5.3	The statistics of the thresholded interaction datasets used for the clustering analyses . . . . .	145
5.4	AUC and AUPR values of datasets using the BLAST for assessing protein similarity . . . . .	147
5.5	AUC and AUPR Values of datasets using the Smith and Waterman (SW) method for assessing protein similarity . . . . .	148
5.6	Number of inactive results detected in the top 1,000 new interacting pairs for the NRLMF and NetLapRLS methods . . . . .	149
5.7	Top 10 and Bottom 10 of the 1,000 predictions made by the CMF method . . . . .	151
5.8	Top 10 and Bottom 10 of the 1,000 predictions made by the NetLapRLS method . . . . .	152

5.9	Top 10 and Bottom 10 of the 1,000 predictions made by the NRLMF method . . . . .	153
5.10	Assay results of the predictions classed as inactive . . . . .	153
5.11	Details of the ToxCast assays of the predictions found to be inactive through thresholding . . . . .	154
5.12	Binding energies for the top 10 and bottom 10 of the 1,000 predictions made using the CMF method . . . . .	155
5.13	Binding energies for the top 10 and bottom 10 of the 1,000 predictions made using the NetLapRLS method . . . . .	155
5.14	Binding energies for the top 10 and bottom 10 of the 1,000 predictions made using the NRLMF method . . . . .	156

# Table of Abbreviations

<b>2D</b> .....	Two-dimensional
<b>3D</b> .....	Three-dimensional
<b>AC50</b> .....	Half-maximal Activity Concentration
<b>AI</b> .....	Artificial Intelligence
<b>API</b> .....	Application Programming Interface
<b>AUC</b> .....	The Area Under a receiver operating characteristic Curve
<b>AUPR</b> .....	The Area Under a Precision-Recall Curve
<b>BLAST</b> .....	Basic Local Alignment Search Tool
<b>BLMNII</b> .....	Bipartite Local Model with Neighbour-based Interaction profile Inferring
<b>BLOSUM</b> .....	Blocks Substitution Matrix
<b>CDER</b> .....	Centre for Drug Evaluation and Research
<b>CDRScan</b> .....	Cancer Drug Response Profile Scan
<b>CMF</b> .....	Collaborative Matrix Factorization model
<b>CPU</b> .....	Central Processing Unit
<b>CSD</b> .....	The United Kingdom's Committee on the Safety of Drugs
<b>CSForest</b> .....	Cost Sensitive Forest
<b>CSV</b> .....	Comma-separated values
<b>CTDBase</b> .....	Comparative Toxicogenomics Database
<b>EC50</b> .....	Median Effective Concentration required to induce a 50
<b>ECDF</b> .....	Empirical Cumulative Distribution Function
<b>EMA</b> .....	European Medicines Agency
<b>EMBL</b> .....	European Molecular Biology Laboratory
<b>EMBOSS</b> .....	European Molecular Biology Open Software Suite
<b>ESSSR</b> .....	Extended Smallest Set of Smallest Rings
<b>EU</b> .....	European Union
<b>FASTA</b> .....	Fast Alignment
<b>FDA</b> .....	The United States Food and Drug Administration

<b>FK</b> .....	Foreign Key
<b>FTP</b> .....	File Transfer Protocol
<b>GCP</b> .....	Good Clinical Practices
<b>GLP</b> .....	Good Laboratory Practices
<b>GMP</b> .....	Good Manufacturing Practices
<b>HMDB</b> .....	Human Metabolome Database
<b>HPC</b> .....	High Performance Computing
<b>IC50</b> .....	Inhibit Cellular Proliferation by 50
<b>ICH</b> .....	International Conference on Harmonization
<b>IND</b> .....	Investigational New Drug
<b>JRip</b> .....	Java Repeated Incremental Pruning to produce error reduction
<b>JSON</b> .....	JavaScript Object Notation
<b>KBMF2K</b> .....	Kernelized Bayesian Matrix Factorization with Twin Kernels
<b>Kcal/mol</b> .....	Kilocalorie per mole
<b>Kd</b> .....	Dissociation Constant
<b>KEGG</b> .....	Kyoto Encyclopaedia of Genes and Genomes
<b>Ki</b> .....	Inhibition Constant
<b>MACCS</b> .....	Molecular Access System
<b>MHRA</b> .....	The United Kingdom's Medicines and Healthcare products Regulatory Agency
<b>MySQL</b> .....	"My" Structured Query Language
<b>NA</b> .....	Not Applicable
<b>NetLapRLS</b> .....	Network Laplacian Regularized Least Square
<b>NIH</b> .....	National Institute for Health
<b>NRLMF</b> .....	Neighbourhood Regularized Logistic Matrix Factorization
<b>OPQ</b> .....	Office of Pharmaceutical Quality
<b>PART</b> .....	Partial decision Trees
<b>PDB</b> .....	Protein Data Bank
<b>PHP</b> .....	PHP: Hypertext Preprocessor
<b>PK</b> .....	Primary Key
<b>PPV</b> .....	Positive Predictive Value
<b>PSOVina</b> .....	Particle Swarm Optimization in Vina
<b>Pub</b> .....	PubChem
<b>PUGREST</b> .....	Power User Gateway Representational State Transfer
<b>PyDTI</b> .....	Python library for Drug Target Interaction prediction

<b>QSAR</b> .....	Quantitative Structure Activity Relationships
<b>REPTree</b> .....	Reduced Error Pruning Tree
<b>SDF</b> .....	Structure Data Format
<b>SIMCOMP</b> .....	The Similar Compound system
<b>SMARTS</b> .....	Simplified Molecular-Input Line-Entry System Arbitrary Target Specification
<b>SMILES</b> .....	Simplified Molecular-Input Line-Entry System
<b>SQL</b> .....	Structured Query Language
<b>SVM</b> .....	Support Vector Machines
<b>SW</b> .....	Smith and Waterman
<b>T3DB</b> .....	Toxin and Toxin Target Database
<b>Tox21</b> .....	Toxicology in the 21st Century
<b>TSV</b> .....	Tab-separated values
<b>UI</b> .....	User Interface
<b>UK</b> .....	United Kingdom
<b>URL</b> .....	Uniform Resource Locator
<b>USA</b> .....	United States of America
<b>WEKA</b> .....	Waikato Environment for Knowledge Analysis
<b>WNN-GIP</b> .....	Weighted Nearest Neighbours Gaussian Interaction Profile
<b>XAMPP</b> .....	Cross-platform Apache, MySQL, PHP and Perl
<b>XML</b> .....	Extensible Markup Language
<b>ZeroR</b> .....	Zero Rule classifier

# Chapter 1

## Introduction and Background

### 1.1 Introduction

The purpose of this study was to investigate if data mining tools and techniques were valuable for use in the fields of drug development and classification. A drug as defined by the Oxford Online Dictionary is a 'medicine or other substance which has a physiological effect when ingested or introduced into the body' [1]. These physiological effects can either be beneficial or detrimental in nature, and can vary due to a number of factors such as the method of application or how much of a drug is consumed (in quantity or over a period of time).

#### 1.1.1 Aims and objectives

The aim of the project was to **investigate and evaluate if a machine learning approach could provide a benefit to specific areas of drug discovery and drug development**. This aim was split into three objectives.

##### 1.1.1.1 Objective 1

**Investigate the history and techniques of computational drug discovery and development.** The concepts, procedures and techniques applied within the drug industry must be appraised before attempts could be made at classification and clustering through machine learning. This objective included an analysis of the data present within publicly available chemical and drug repositories such as DrugBank [2] and ToxCast [3]. Also included were the steps needed to be performed in order to make the repositories suitable for data mining operations.



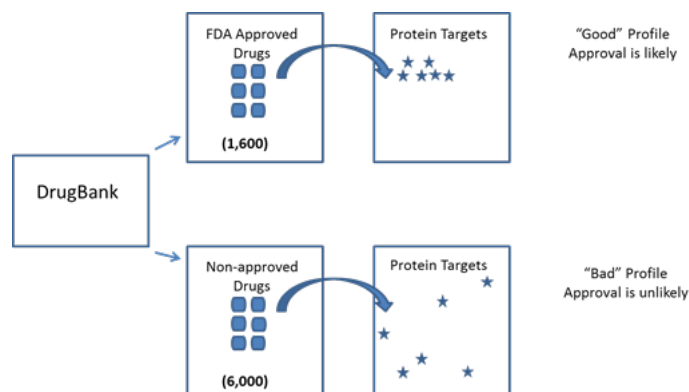
### 1.1.1.2 Objective 2

**Investigate and evaluate machine learning techniques applied to *in-silico* drug screening.** Once appropriate research had been undertaken to understand the structure of drug and chemical repositories, a suitable architecture was required to store and aggregate the information across multiple databases. Once this information had been obtained, an investigation was undertaken to determine if a machine learning tool could be used to construct a classification model which would be able to discriminate between "good" (i.e approved) and "bad" (i.e potentially toxic) compound profiles. The objective considered classification model construction on a number of fronts such as the documentation of protein interactions with drug compounds or the physical properties of drugs such as molecular weight and charge. Figure 1.1 provides a graphical representation of this objective.

The deliverables for this objective were an analysis and evaluation of the classification models generated, and a comparison of model accuracy based on the variations of attributes within the training data (i.e filtering to specific collections of proteins such as the Bowes *et al.* Panel 44 set [4]) or sources of data (i.e the use of simulated protein interactions with ToxCast/DrugBank compounds).

### 1.1.1.3 Objective 3

**Investigate and evaluate machine learning techniques for the purpose of clustering drug compounds or protein targets.** This objective builds upon the work of objective 2 through investigating if specific regions of compound (chemical) or protein (biological) space could be clustered into related groups. One area of exploration was investigating if a machine learning model could predict clusters of compounds based on their documented protein targets. Another area focused work on the opposite inference, to investigate if protein targets can be clustered based on the chemical properties of a drug compound. This objective addressed a fundamental question of discerning patterns in how compounds and proteins in chemical and biological space respectively interact. Figure 1.2 provides a graphical representation of this objective.



Challenge : Discrimination of "good" vs. "bad" profiles – emphasis on Yes / No for a given compound.

Figure 1.1: Graphical representation of objective 2

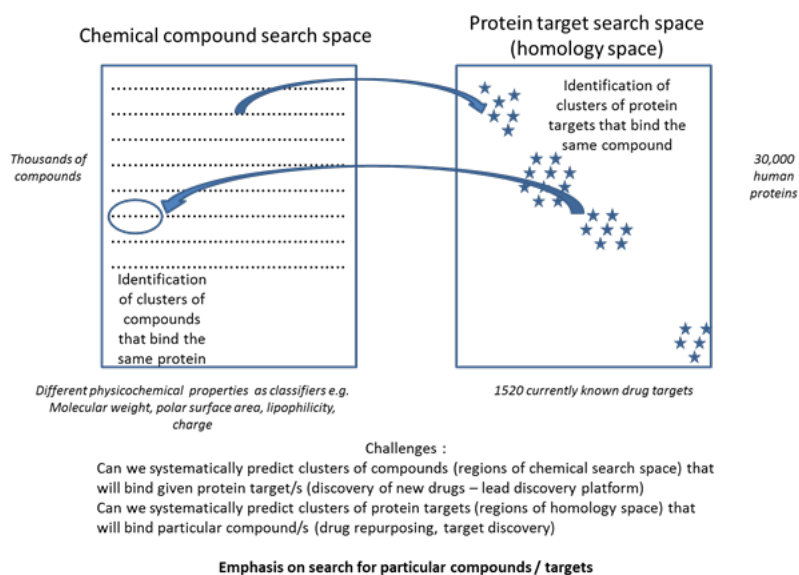


Figure 1.2: Graphical representation of objective 3

## 1.2 Drug Regulation Process

In order to undertake an effective analysis of compound and protein details, an appraisal of the current processes and techniques used within the pharmaceutical industry was required. Generally, in order for a drug to be considered approved by a regulator the benefits it provides to a patient must outweigh the risks associated with its use. In the United Kingdom, in order for a drug to become approved for use and marketed it must either have a license from the Medicines and Healthcare products Regulatory Agency (MHRA) for UK sale, or from the European Medicines Agency

(EMA) for sale across European Union member states. Within the United states, drug safety regulation is controlled by the Food and Drug Administration (FDA).

### 1.2.1 UK Drug Regulation History

Effective drug regulation within the UK started due an incident over a medicine known as thalidomide during the late 1950s. Smithells *et al.* describes that the chemical had been marketed and prescribed for a wide variety of purposes such as asthma and migraines. However, thalidomide's use in the treatment of symptoms during early pregnancy [5] resulted in it's withdrawal when its use had been connected to birth defects in children, an effect which had not been recorded at the time when distributed.

Vargesson states that it was unclear whether or not the disaster could have been prevented at that time, as the appropriate testing procedure at the time had been carried out for the drug [6]. However, Vargesson also highlighted factors which might have increased use of the drug, through minimal packaging information as shown in Figure 1.3, through to the distribution of sample drugs to doctors and physicians to distribute to patients who had suffered morning sickness symptoms. The episode demonstrated that a flaw existed in the testing of drugs in that different species had differing reactions and responses to drugs (mice were subsequently shown to have a much lower adverse reaction to thalidomide), and procedures were put into place to incorporate multiple species *in vivo* and extensive *in vitro* testing during the drug screening process. While a form of thalidomide is still in use as a treatment to diseases such as leprosy, its dangers are now clearly highlighted to patients as shown in Figure 1.4.



**Figure 1.3:** A Thalidomide sample packet that was distributed to women in early pregnancy [6]

The MHRA states that as a result of this event, the Committee on the Safety of Drugs (CSD) had been conceived to prevent such an incident from occurring again, which in turn led to the formation of the Yellow Card Scheme, the Medicines Act



**Figure 1.4:** Packaging for a variety of Thalidomide used from 2006

1968, and the formation of MHRA through mergers in 2003 [7].

The Stationary Office provides a history of the Yellow Card Scheme, and its initial naming as the Register of Adverse Reaction to Drugs from the CSD, which provided an ability for doctors and dentists to report side effects encountered by patients from administering drugs [8]. In its first year of use, the report described that up to 100 of the yellow coloured forms had been submitted to the CSD every week. As of 2014, almost 750,000 yellow card reports for adverse drug reactions had been submitted, and the scheme as of 2005 has been expanded to allow patients to report adverse effects (to which up to approximately 1,750 patient reports are submitted every year) [9].

The Medicines Act stipulates that in order for a person to sell or distribute medicine, a licence must be obtained from the government [10]. The act also defined three main categories of drugs with appropriate restrictions of distribution: general sales list medicines, pharmacy only medicines and prescription only medicines, where the latter categories are reserved for drugs with an increased level of risk, or potential for misuse. Examples of misuse include the use of antibiotics for common illnesses, as well as the use of opioids, sedatives and stimulants for either recreational purposes, or for managing addictions.

### 1.2.2 UK Drug Approval Process

In order to obtain a license, a company must demonstrate that the drug is safe for human use. There are four types of application to obtain a license in the UK [11]:

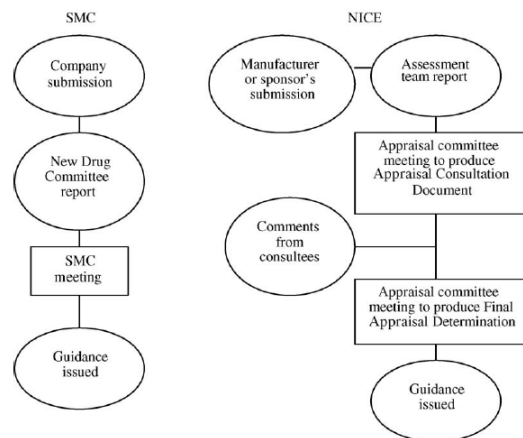
- A decentralised process in order to market a medicine within the UK and specific European Union member states

- A mutual recognition process if approval has been obtained within an EU member state and a company desires to expand marketing to other countries
- A national process in order to market a medicine solely within the UK
- A centralised process with the European Medicines Agency in order to market a medicine throughout the entire European Union

The Medicines and Healthcare Products Regulatory Agency (MHRA) can provide guidance to developers of medicines before submission, which can include matters of advertising and labelling, clinical and non-clinical matters for a fee. Once a license has been obtained for marketing, organisations on behalf of National Health Service (NHS) trusts provide guidance as to whether the drug is economically viable as well as provide reasonable quality of care for patients. There are different organisations for each nation in the UK which provide this guidance, which are:

- The National Institute for Health and Care Excellence (NICE) for English and Welsh NHS Trusts [12]
- The Scottish Medicines Consortium (SMC) for NHS Scotland [13]
- The Department of Health, Social Services and Public Safety for Northern Ireland [14]
- The All Wales Medicines Strategy Group (AWMSG) for Welsh NHS Trusts [15]

The scale of advice provided by these authorities can vary. In a report by Cairns comparing the policies between the SMC and NICE, the author states that the SMC has greater emphasis on delivering reviews rapidly to reduce inefficiency, while NICE includes additional opportunities for stakeholders (patients, consultants etc.) to make comments [16]. Figure 1.5 describes a simplified overview of the appraisal process by Cairns. In the event that the NHS does not approve of a drug's economic viability, a company still has the option for providing the drug for private sale provided the license from the MHRA is still valid.



**Figure 1.5:** Overview of the SMC and NICE Appraisal process [16]

### 1.2.3 FDA Drug Regulation History

The FDA was known by its current name in 1930, but its existence began with the foundation of the 1906 Pure Foods and Drugs Act. The law affirmed that the then Bureau of Chemistry had the ability to regulate food and drugs to prevent the sale of adulterated and misbranded products [17]. Before the act had been implemented, some state governments within the United States were already performing some means of regulation within the food dairy industry, which included penalizing industries under charges of adulteration if they failed to indicate if impurities were present within their products.

Interest in expanding state law federally increased during the 20th century due to increased costs of regulation in interstate trade, in addition to concerns regarding imperfections of enforcement within certain states. Although Law argued that food and dairy quality improved through the act's foundation, a lack of ability to create legally binding statutes and small penalties for violations meant that enforcement by the FDA may not have been that effective until the foundation to the Food, Drug and Cosmetics Act 1938.

This view is shared by Kinch *et al.*, who explored the failure of an anti-infective variant of sulfanilamide which was initially discovered in the late 1930s [18]. Kinch *et al.* explained that numerous variations were developed due to a need to combat bacterial infections, and due to inconsistent quality control procedures, a variant known as elixir sulfanilamide killed more than 100 people in 1937 due to poisoning (many of whom were children which led to public outrage). Borchers *et al.* details the additions and amendments made in the Food, Drug and Cosmetics Act which was

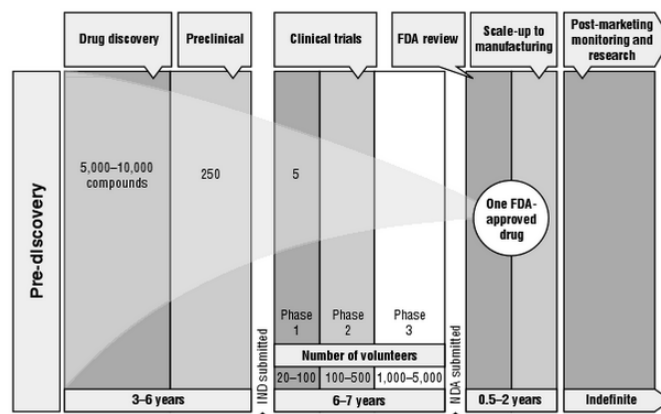
formed following the incident [19]:

- A drug required scientific safety testing before it could be considered for marketing which must be proven by the manufacturer
- Cosmetic and therapeutic devices were now under regulation by the FDA
- Proof of fraud was no longer required to challenge false claims made for drugs
- Addition of poisonous substances was prohibited except when required for production or were otherwise unavoidable
- The establishment of food standards in order to promote honesty and fair dealing in the interest of consumers
- Expansion of legal capacities of the FDA

Although the law itself has been revised during the decades since its conception, Borchers *et al.* affirm that the law's requirement of scientific testing for drugs was a first time approach, and became one of the major factors in the shaping of the modern pharmaceutical industry. It was this, coupled with the large demand for the antibiotic penicillin during World War 2 which also provided the ability for academics, government scientists, drug companies and medical practitioners to collaborate with one another.

#### 1.2.4 FDA Approval Process

In order for a drug to be considered suitable for use by the FDA, the drug developer must take a number of steps to prove that the drug performs as described and is safe for its described use. Lipsky *et al.* states that the first step for a promising pre-clinical drug is to file an application for an Investigational New Drug (IND) to the FDA [20]. The report continues to explain that once the IND application has been approved, the drug company must perform clinical trial phases of increasing sample sizes to investigate factors such as dosage amounts, safety and drug effectiveness. On compilation of this evidence, a New Drug Application is filed for which the FDA will either approve or reject the drug, or request the company to perform further study before a decision is made. Figure 1.6 provides a graphical representation of this process [21]. A typical drug following this process would take approximately 10-15 years to develop following the FDA pipeline.



**Figure 1.6:** Overview of the drug development and approval process for FDA approval [21]

### 1.2.5 Drug Assessment Phases

Although the exact procedures vary in getting a drug to market, there are some broad similarities in the procedures of assessing a drug's safety. There are two main phases: a pre-clinical testing phase in a lab environment and clinical trials which involve testing a drug's response on human subjects. As a candidate drug progresses along trials over the course of its development the scale and complexity of the tests increase.

During all phases of assessment, the drug company must also follow good practice procedures to ensure that the output remains to a high standard. These standards can be summarised to three key practices, the first of which is Good Laboratory Practices (GLP) [22], which details the conditions and processes which must be followed for a clinical or non-clinical trial to be performed. The second practice, Good Clinical Practice (GCP) [23], are a set of principles created by the International Conference on Harmonization (ICH) that provide instructions on the design and methodology of the study, and the output data which is to be expected within the assessment. The final practice, Good Manufacturing Practice (GMP) [24], relates to the regulation of the design and development of the drug in manufacture, which includes the monitoring and quality control of production facilities.

In terms the US regulation to the above practices, the US Code of Federal Regulations Title 21 Part 58 [25] states that the FDA requires researchers of companies to undertake a good level of practice in undertaking this phase, which include standards such as study conduct and reporting, operating procedures and quality assurance procedures to ensure processes are followed correctly. The FDA's Center for Drug Evaluation and Research (CDER)'s Office of Pharmaceutical Quality (OPQ) monitors and ensures standards are followed during a drug's lifecycle within the FDA



[26]. In terms of UK regulation, the MHRA and DHSC have published guidance companies should follow with respect to GLP, GMP and GCP [27].

#### 1.2.5.1 Pre-clinical Phases

Whitmore describes the pre-clinical phase as the establishment of safety required for a drug which has not had a history of clinical use before human testing can be performed [28]. In order to establish this level of safety, a company must demonstrate sufficient testing to prove that a compound of interest does not cause an undesired effect, or potential to cause serious harm. One of the main realms of testing is known as *in vitro* testing which are laboratory experiments performed on cell receptor or enzyme systems within a laboratory environment. Other formats of testing include *in vivo* testing (experimentation on animals) as well as *in-silico* testing (experimentation within a simulated virtual environment).

After tests have been conducted on pre-clinical phases, companies will then decide if a drug has sufficient merit to proceed to human testing. In terms of success rates to clinical trials, Paul *et al.* reported that the chances of progression from preclinical to clinical trials was 69%, based on research and development data from 13 pharmaceutical companies [29].

#### 1.2.5.2 Clinical Trials

If a drug proceeds to human testing, and has received approval on an IND application, the drug company must begin the process of conducting trials to assess the effectiveness and dosage limits for the medicine in question. The National Institute of Health (NIH) provides a definition for the phases of trials involved [30]:

- Phase I Trials - A test performed on a small group of people (under 100) to assess drug safety and record potential side effects
- Phase II Trials - A test administered to a larger group (several hundred) to further assess drug safety and effectiveness
- Phase III Trials - A test administered to a large group (several thousand) of people to confirm the findings of previous phases, monitor any recorded side effects and to compare results against standard or equivalent treatments
- Phase IV Trials - A phase conducted after a drug has been marketed and approved by a regulator in order to track its safety, risks and benefits and optimal usage amounts

In terms of difficulty of completing the trials, Table 1.1 reveals the likelihood of approval rates in the various stages of FDA clinical trials from a study of 835 drugs companies between 2003-2011 [31]. The table reveals that while companies may have a good chance of generating promising data for progression in the first phase, the chance decreases by almost 30% from phase II to III. This could be caused from new issues revealed from the increased scale of testing, as the likelihood of further issues and potential dangers increases with even larger scale group tests in further phases.

<b>Phase Progress</b>	<b>Success Rate (%)</b>
Phase I to Phase II	64.5%
Phase II to Phase III	32.4%
Phase III to NDA Submission	60.1%
NDA Submission to Approval	83.2%

**Table 1.1:** Drug Success Rates [31]

### 1.2.6 Drug Development Cost and Attrition

In order for a drug to be considered as suitable for use by the general public, there are a number of guidelines which must be followed which can vary according to the procedures of the regulators of different countries. However, Lesk provides an overview of characteristics of a suitable drug [32]:

- It must be safe to use in that the drug does not cause a severe detrimental effect to patients
- It must be effective in its design in the process of diagnosis, treatment or prevention of a disease
- It must be stable enough in order for the drug to perform its actions and not remain in a patient's system for longer than necessary
- Its contents must be available to produce in sufficient quantities either naturally or synthetically
- Its concept should be novel so that the drug would not be duplicated by other pharmaceutical companies which would result in a loss of profits necessary to recover development costs

As chemical compounds could fail to meet these characteristics or fail a drug regulator's guidelines, there is a rate of attrition present within drug development

where only a certain number of drug candidates are suitable for further testing and production. Additional issues, such as a lack of efficacy, an unforeseen toxicity or a difficult pharmacological profile may also have an impact on the viability for further development. This attrition rate can be considered to be quite high; one study by Morgan *et al.* estimated that the success rate of drugs that were able to enter clinical trials from primary *in vitro* work were between 11.7% and 24% across 5 studies of drug approval rates [33].

The costs associated with the process of drug development can vary, with different studies reporting varying figures and estimates of cost due to the lack of specific financial information provided by pharmaceutical companies. One report by Morgan *et al.* found that the development cost could vary between \$92 million USD to \$883.6 million USD, explaining that the cause for the variation found during their assessment of a number of drug cost evaluation studies was due to differences in methodology, sources and timescales. Another report by Sertkaya *et al.* found that development costs were found to be more within the region of between \$1.3 billion USD to \$1.7 billion USD [34], however this report also provided additional detail in terms of the estimated costs for each clinical trial phase, listed in Table 1.2. The report also highlighted that some of the main contributors for expenditures were due to procedure costs (15-22% of the total project cost), staff funding (11-29%) and site monitoring procedures (9-14%).

Clinical Phase Level	Estimated Cost (\$ million USD)
Phase I	1.4 to 6.6
Phase II	7 to 19.6
Phase III	11.5 to 52.9

**Table 1.2:** Clinical Phase Costs [34]

Another assessment of drug costs was in an analysis performed by Paul *et al.*, which viewed the drug development cost from a perspective of out of pocket expenses for clinical trial studies, and for capitalized cost, the rate of return of investors for funding the research [29]. Their study found that a drug launch would incur approximately \$873 million USD of out of pocket expenses, and a capitalized cost of \$1.7 billion USD, however the estimate did not include investments involved for exploratory discovery or for expenses incurred post launch of the drug. Despite these variations in estimated costs, there appears to be a large amount of expense involved in the process of drug development, which can result in costly losses to a company in the event that a candidate drug fails to launch due to particular issues.

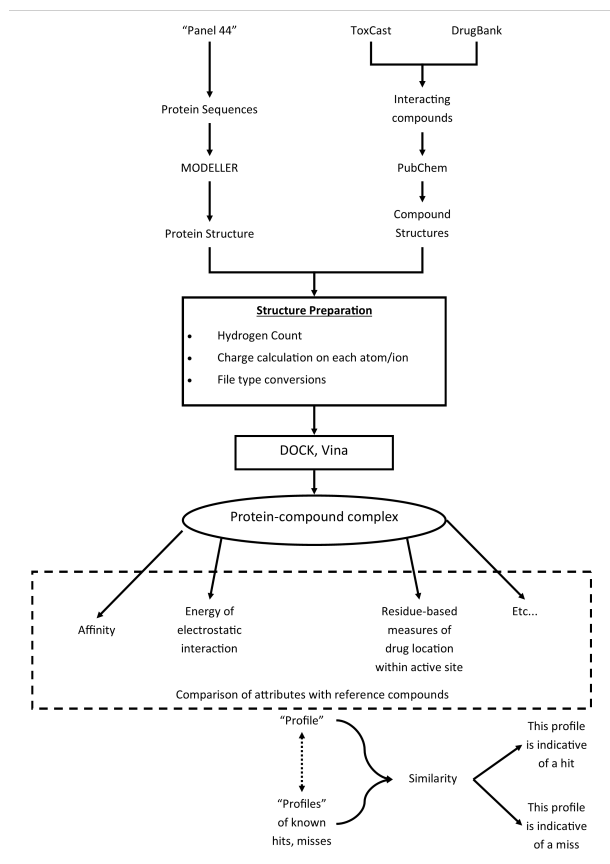
### 1.2.7 Approaches to reducing attrition

One approach which has been adopted to reduce the impact of drug attrition is through *in vitro* pharmacological profiling, a technique which is explained by Bowes *et al.* as identifying undesirable compounds through screening compounds against a broad range of protein targets identified in previous studies [4]. Their paper explains that the use of profiling can identify early stage hazards which could disrupt or halt a drug development project, as well as reduce costs in safety assessments of compounds where little or no off-target activities are found. In their study they analysed the range of protein targets screened by four pharmaceutical companies, and identified a minimal panel of 44 *in vitro* targets which provide an early indication of a hazardous compound. This panel is commonly referred to as Panel 44, however two targets in the paper make reference to two UniProt entry codes, resulting in a searchspace of 46 UniProt entry codes. This panel of proteins can be found on Table 1.3.

Another way of reducing drug attrition is through the process of replacement or part replacement by *in silico* screening, a technique of using computer simulations and programs to determine the likelihood of a protein becoming a target for a drug. The main advantage of implementing *in silico* methods is that it provides a means of expansion beyond what would be considered practical for an *in vitro* study, in turn providing further information as to how a candidate drug could behave when used. One study performed by Ramsundar *et al.* made use of a multitask network for the purposes of virtual screening, and had found that the network had made accurate predictions of drug interactions against differing diseases, but the network required a considerable amount of infrastructure and time to execute. According to Google, the results generated by the report were based on 37.8 million data points over 200 biological processes, and required over 50 million CPU hours to compute.

Another approach into this field is from a study conducted by Austin-Muttitt, which made use of 3D protein and compound structure information in conjunction with a collection of simulator software tools to output a set of binding properties [36]. An example of the initial process that was used for this study is shown in Figure 1.7. These properties were then compared in terms of similarity to known protein-drug interactions to calculate the likelihood that a link exists between a candidate drug and a protein. The disadvantage of employing such a pipeline however is that access to 3D structures of proteins and compounds is required to make use of the pipeline to its full potential, which can be a limiting factor as some areas of chemical and proteomic space do not possess accurate models currently.

There have also been studies with machine learning for predicting problematic and



**Figure 1.7:** Illustration of initial process to generate *in silico* compound-protein interactions [36]

beneficial drugs. One study by Pereira et al. investigated the effectiveness of machine learning algorithms in predicting toxicity of a compound [37]. In their investigation they found that the use of molecular descriptors had generated high degrees of accuracy for prediction of toxicity. Another study by Chen et al. investigated if clustering algorithms could assist in identifying candidate drugs to treat Hepatitis C, and had discovered 20 compounds of interest where a prior indication for treatment had not been described [38].

Another set of approaches which have made use of machine learning techniques is through the process of using similarity of drugs and proteins and clustering to determine appropriate targets. The first study into this area was by Yamanishi *et al.*, which made use of a two stage process to predict interactions [39]. The first phase was to construct a model based on a set of known compound-protein interactions, in addition to scales of similarity of compounds and proteins against one another. The second phase then involved the inclusion of compounds and proteins which were not present in the model to assess the model's reliability in predicting known interactions. The study had found that when tested, the models had

managed to predict the interaction profile of most of the test compounds, providing a potentially suitable alternative to large scale screening between compounds and proteins. The additional advantage to the use of this model is that there are fewer requirements to process candidates, needing only the SMILES string code and amino acid sequence for compounds and proteins respectively. Since the paper’s publication, other studies have been published which make use of similar principles, but with either different methodologies for constructing the interaction prediction model, or differing techniques of calculating the similarity of drugs and proteins.

There have also been machine learning approaches applied to the prediction of structure of proteins. Bruno *et al.* had achieved an accuracy rating of 94% with identifying protein images into specific categories of clear, precipitate, crystal, and other (where a protein was found to be significantly different to the other three categories) [40]. This high degree of accuracy via a neural network, coupled with the large amount (500,000) of image training data provides promise for large scale virtual screening as an application.

UniProt ID	Protein Entry Name
P04150	GCR_HUMAN
P06239	LCK_HUMAN
P07550	ADRB2_HUMAN
P08172	ACM2_HUMAN
P08588	ADRB1_HUMAN
P08908	5HT1A_HUMAN
P08913	ADA2A_HUMAN
P10275	ANDR_HUMAN
P11229	ACM1_HUMAN
P14416	DRD2_HUMAN
P14867	GBRA1_HUMAN
P20309	ACM3_HUMAN
P21397	AOFA_HUMAN
P21554	CNR1_HUMAN
P21728	DRD1_HUMAN
P22303	ACES_HUMAN
P23219	PGH1_HUMAN
P23975	SC6A2_HUMAN
P25021	HRH2_HUMAN
P25101	EDNRA_HUMAN
P28222	5HT1B_HUMAN
P28223	5HT2A_HUMAN
P29274	AA2AR_HUMAN
P31645	SC6A4_HUMAN
P32238	CCKAR_HUMAN
P34972	CNR2_HUMAN
P35348	ADA1A_HUMAN
P35354	PGH2_HUMAN
P35367	HRH1_HUMAN
P35372	OPRM_HUMAN
P37288	V1AR_HUMAN
P41143	OPRD_HUMAN
P41145	OPRK_HUMAN
P41595	5HT2B_HUMAN
P46098	5HT3A_HUMAN
Q01959	SC6A3_HUMAN
Q05586	NMDZ1_HUMAN
Q08499	PDE4D_HUMAN
Q12809	KCNH2_HUMAN
Q13936	CAC1C_HUMAN
Q14432	PDE3A_HUMAN
Q14524	SCN5A_HUMAN
P15382	KCNE1_HUMAN
P51787	KCNQ1_HUMAN
P02708	ACHA_HUMAN
P43681	ACHA4_HUMAN

**Table 1.3:** UniProt Entry Codes for Bowes et. al panel for pharmacological profiling [4]. Proteins in bold separated by lines indicate two UniProt entry codes grouped into single targets by the panel

## 1.3 Compound and Protein Database Repositories

This section describes the main database repositories which were considered for gathering information on drugs, compounds, proteins and their interactions with one another for use within the project. These sources are listed as follows:

- DrugBank and The Toxin and Toxin-target database (T3DB)
- ToxCast
- PubChem
- ChEMBL
- Matador
- CTDBase
- BindingDB
- UniProt
- KEGG
- Human Metabolome Database (HMDB)

To determine a potential interaction between a compound and a protein, the source needed to specify that a link existed between them, in addition to providing the means to be linked across other sources. This meant that assays would be selected if a link could be established to UniProt (and for the scope of this study, linked with the human species), and compounds would be selected if a link could be established towards a single source, which was PubChem. Initially, all links between proteins and compounds found from these repositories were merged into the largest dataset possible, to maximise the amount of information captured. Further refinements were then made to the selection of assay types and results in Chapters 4 and 5.

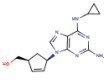
### 1.3.1 DrugBank, T3DB and HMDB

The DrugBank project is an online database which was created by the University of Alberta, which allows users to view details on a variety of drugs which are held in

various categories such as experimental, withdrawn, illicit as well as FDA-approved [2]. The database itself links protein sequences found within databases such as Swiss-Prot and UniProt with data that is found within medicinal and chemical reference handbooks. With the latter being in hard copy formats, most of the information contained within the DrugBank database has been manually curated from a number of sources over a number of years, and its transfer into a computer format has made the website valuable to obtain information on drug and protein interactions.

Each drug record held on the website (an example of which is shown on Figure 1.8) contains information such as chemical properties, treatment information and documented interactions with proteins. The website was primarily used to identify drugs which were assigned to the FDA-approved group, which were then used as a basis for beneficial protein interaction profiles ("Good" Profile compounds). In addition to this, drugs which were assigned to the experimental drugs were gathered for use as a potential test set. The website's structure allows users to download these types of specific sections without the need for downloading the whole database, which reduced the requirements for additional filtering and pre-processing. The website is free to use at the time of writing, with no requirement for licensing providing projects that make use of the database are for non-commercial applications.

The screenshot shows the DrugBank website interface for the drug Abacavir. The browser address bar shows 'www.drugbank.ca/drugs/DB01048'. The page features a navigation menu with tabs for Identification, Taxonomy, Pharmacology, ADMET, Pharmacoeconomics, Properties, Spectra, References, Interactions, and 0 Comments. Below the navigation, there are buttons for 'Targets (1)', 'Enzymes (3)', and 'Biointeractions (4)', along with a 'Show Drugs with Similar Structures' button. A blue banner promotes the 'Get DrugBank to go!' app for iOS and Android, with a 'Sign up to get early access' button. The main content area is a table with the following information:

Identification	
Name	Abacavir
Accession Number	DB01048 (APRD00216)
Type	Small Molecule
Groups	Approved, Investigational
Description	Abacavir (ABC) is a powerful nucleoside analog reverse transcriptase inhibitor (NRTI) used to treat HIV and AIDS. [Wikipedia] Chemically, it is a synthetic carbocyclic nucleoside and is the enantiomer with 1S, 4R absolute configuration on the cyclopentene ring. In vivo, abacavir sulfate dissociates to its free base, abacavir.
Structure	

At the bottom of the structure section, there are buttons for 'MOL', 'SDF', '3D-SDF', 'PDB', 'SMILES', 'InChI', and 'View 3D Structure'.

**Figure 1.8:** Screenshot of DrugBank website displaying information on the drug Abacavir [41]

Other databases that were also developed by the University of Alberta and were also used in the project were the Toxin and Toxin Target Database (T3DB) [42], and the The Human Metabolome Database (HMDB) [43], which was structured



in a similar way to the DrugBank platform. T3DB contained protein interaction information on toxins, which provided an ideal platform for identifying compounds which were potentially harmful. HMDB's focus was providing information of small molecule metabolites within the human body, which also included interactions with particular proteins. This also provided an ideal platform for identifying compounds largely considered to be harmless. Both T3DB and HMDB databases were free to use for non-commercial applications.

### 1.3.2 ToxCast

ToxCast is the database resulting from a toxic effect screening programme developed by the United States Environment Protection Agency (EPA), which contained results from mostly *in vitro* assays between compounds and proteins to observe for changes which can suggest a potentially toxic effect has occurred [3]. The program primarily focuses on compounds which have been highlighted as a concern by regulators but have limited information on potential health effects. These are then passed through a number of high throughput assays, before the results are quality control checked and released to the public.

There is no interactive website platform for viewing the information within ToxCast; instead the information is split into a number of packages which are freely available to download online for commercial and non-commercial use [44]. The database also contains a MySQL database dump and programming packages to access all the high throughput information processed by ToxCast.

One feature which ToxCast contains that differs from other database sources is that the assay summary files document cases where a chemical has been noted to have no interaction with a protein, which is a useful data point for flagging potential false positive cases. Assays can also be filtered by a specific source, providing an additional means for further filtering. One of the sources of interest is Tox21, an EPA programme that has amongst its goals to "Prioritize specific compounds for more extensive toxicological evaluation" [45]. The assays used by Tox21 are cell-based, meaning the assays in question attempt to quantify a response of a compound to a test organism which in turn can refer to more than one protein target. Another source of interest is NovaScreen, a commercial panel for preclinical drug development and was used for the Panel 331 report [46]. NovaScreen's assays are biochemical, which document binding activity of a compound to a biological molecule and is generally focused on one protein target.

These features, coupled with the sample types of potentially harmful compounds

made the ToxCast repository an ideal candidate for use in obtaining interaction profiles of possibly harmful compounds.

### 1.3.3 ChEMBL

ChEMBL is a database developed by the European Bioinformatics Institute which contains information on bioactive drug-like small molecules, and their interactions with proteins [47]. The main motivations behind the ChEMBL project were to provide a centralized format on sources of bioactivity which were either difficult or labour intensive to access due to differing formats, and to provide a freely available platform for interactions which at the time would only have been available through the use of commercial products. The interactions themselves are based upon peer reviewed scientific journals, and the information contained within the database can either be accessed on an interactive website or downloaded in its entirety for bulk access.

To access the activity information contained within ChEMBL, a database dump in the MySQL format was downloaded, which allowed the execution of queries to filter the database to the required interactions of certain proteins. Cross references to PubChem and UniProt also assisted in linking the compounds and proteins contained within ChEMBL to other databases.

### 1.3.4 Matador

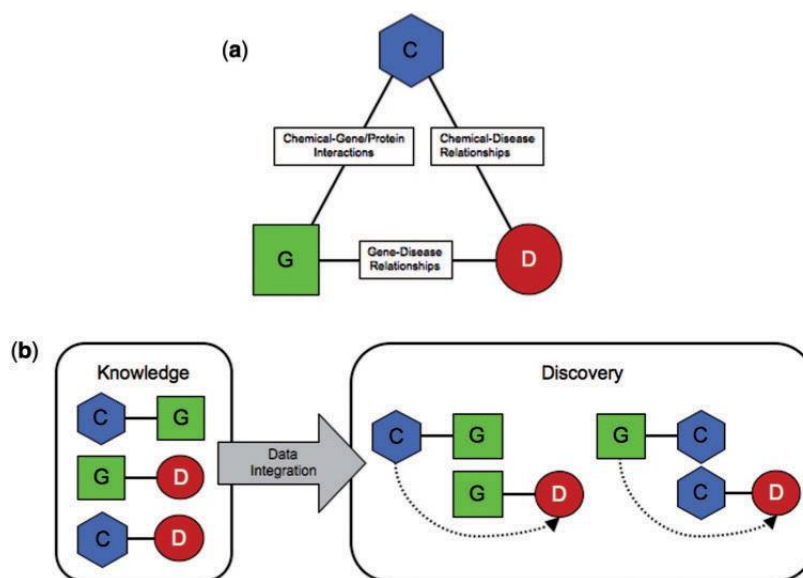
The Manually Annotated Targets and Drugs Online Resource (Matador for short) is a repository of drug-protein interactions which had been annotated from abstracts within PubMed [48] or Online Mendelian Inheritance in Man (OMIM) entries [49], through the means of text mining techniques and manual automation [50]. The Matador platform is a subset of the SuperTarget database where text mining of interactions revealed the ability to determine the type of binding between drugs and protein targets. The activities stored within this database primarily use PubChem and UniProt identifiers, which made the recorded interactions easy to merge with other databases. The website is also free to use, provided the information is not used for commercial purposes in which case a license is required from the database developers.

### 1.3.5 CTDBase

The Comparative Toxicogenomics Database is a database developed by the MDI Biological Laboratory, designed to provide links between chemicals, genes and diseases [51]. In a description of the database, Davis *et al.* describe the three main differences between CTDBase and the other data sources:

- The database's primary focus is on environmental chemicals
- The database integrates sources from literature and information inferred from other interactions listed within the database ( e.g. gene A is associated with disease B because gene A has a curated interaction with chemical C, and chemical C has a curated association with disease B)
- The database can provide a resource for generating hypotheses on chemical actions and diseases which may not be otherwise apparent.

Figure 1.9 illustrates an overall model of the structure of the database, and the process followed to create additional discovery links based on knowledge found within the literature which is gathered by text mining software and then verified by biocurators. The database contains an interactive web platform which can be used to query for individual chemicals/genes/diseases, however a bulk query platform exists to obtain results from multiple queries.



**Figure 1.9:** CTDBase Data curation process [51]. (a) reflects the types of information gathered by curators. (b) reflects how connections are to be inferred based on prior knowledge found within the database

Individual sections of the database are also available for download for programmatic access. CTDBase is listed as free to access, provided the database designers have been notified for what purposes the information is used for. Although the database drug and protein contents contain no direct reference to PubChem or Uniprot, platforms exist to link entities to one another, making CTDBase an ideal platform to discover additional drug-protein interactions as well as a link to proteins and diseases for additional filtering or protein panel creation.

### 1.3.6 BindingDB

The Binding database is a website that was developed by the Skaggs School of Pharmacy and Pharmaceutical Sciences based at the University of California [52]. Initially released in 2000, Gilson *et al.* explains that one of the unique aspects of the database in its data collection methods was the capture of data from the US Patent system, which is absent from scientific literature. As of 2015 the database claimed to contain approximately 35,000 data points from recent US patents, however the database also considers and gathers information from other databases and sources of scientific literature not covered by other research group efforts. The interactions stored within are mainly focused on protein targets which either contain a three dimensional structure deposited in the Protein Data Bank, or that can be modelled accurately.

The database as of July 2018 contains over approximately 150,000 compounds, and 1,361 target proteins, all of which is freely accessible either through individual searches through the website, or through downloading specific sections from the database for bulk access. This coupled with PubChem and UniProt references contained within the database also make BindingDB a useful source for downloading additional drug-protein target information.

### 1.3.7 UniProt

UniProt is described by the Uniprot Consortium as a freely accessible database which contains manually annotated and reviewed information on proteins [53]. Designed as a platform to organise an increasing number of sequenced proteins, the consortium states that the UniProt platform consists of a number of key parts which satisfy different requirements. The main section, UniProtKB/Swiss-Prot, is the section that contains manually curated information on over half a million sequences. Another section, UniProtKB/TrEMBL, contains information on proteins which are either

unreviewed or have been user submitted, of which a significantly larger number of sequences exists (80 million). Platforms also exist of sets of protein sequences which are non-redundant based on varying levels of sequence similarity (50%, 90% and 100% known as UniRef50, UniRef90 and UniRef100 respectively). Each record held on the website (an example of which is shown in Figure 1.10) contains information such as the protein entry name, gene, species as well as information on Amino Acid Sequences and tissue group associations. The website also contains a platform which can convert UniProt accession codes into identifiers used by other database systems, and vice versa [54]. As a platform which is referenced and contains many references of databases, the UniProt database was considered the most suitable reference for documenting the human proteins featured within the project.

The screenshot shows the UniProtKB entry for P10275 (ANDR\_HUMAN). The page is titled 'UniProtKB - P10275 (ANDR\_HUMAN)'. The main content area displays the following information:

- Protein:** Androgen receptor
- Gene:** AR
- Organism:** Homo sapiens (Human)
- Status:** Reviewed - Annotation score: ★★★★★ - Experimental evidence at protein level<sup>1</sup>
- Function<sup>1</sup>:** Steroid hormone receptors are ligand-activated transcription factors that regulate eukaryotic gene expression and affect cellular proliferation and differentiation in target tissues. Transcription factor activity is modulated by bound coactivator and corepressor proteins. Transcription activation is down-regulated by NROB2. Activated, but not phosphorylated, by HIPK3 and ZIPK/DAPK3. 8 Publications
- Miscellaneous:** Isoform 3 and isoform 4 lack the C-terminal ligand-binding domain and may therefore constitutively activate the transcription of a specific set of genes independently of steroid hormones. 1 Publication
- Enzyme regulation<sup>1</sup>:** AIM-100 (4-amino-5,6-biaryl-turo[2,3-d]pyrimidine) suppresses TNK2-mediated phosphorylation at Tyr-269. Inhibits the binding of the Tyr-269 phosphorylated form to androgen-responsive enhancers (AREs) and its transcriptional activity. 1 Publication
- Sites:** A table listing binding sites for Androgen at positions 706, 753, and 878. Each entry includes a description, actions (Add BLAST), and a graphical view.
- Regions:** A table listing DNA binding (500-622), Zinc finger (500-580), and Zinc finger (596-620). Each entry includes a description, PROSITE-ProRule annotation, actions (Add BLAST), and a graphical view.

**Figure 1.10:** Screenshot of UniProt website displaying information on the Androgen Receptor protein [55]

### 1.3.8 PubChem

PubChem is a freely accessible database which contains information on chemical compounds, and is managed by the National Center for Biotechnology Information [56]. Initially launched to the public in 2004, there are 3 main components of the PubChem platform, setup as individual databases: these are PubChem BioAssay, PubChem Substance and PubChem Compound. The BioAssay database contains information on the parameters and results of biological activity testing, while the Substance

database contains chemical descriptions and properties of chemicals. Finally, the PubChem Compound segment provides unique chemical structures of the content found within the PubChem Substance database. Each compound in the combined platform contains elements such as the 2D/3D structure, as well as chemical and physical property information. Each compound also contains the results of all protein screening test performed against it, stating whether or not a pair is active (a target) or inactive (where no activity has been documented). Some of the interactions also contain a reference to a UniProt accession code, making it possible to cross reference to other database interaction profiles. The PubChem platform also contains an ability to cross reference compounds to other compound databases. Coupled with PubChem references being used in the other databases assessed, the PubChem platform provided an ideal central reference to link to the other studied databases.

While a web platform exists for conducting individual queries, PubChem contains multiple means of access for conducting bulk queries. Users can either download the entirety of the database in bulk through PubChem's file transfer protocol server, or through the use of online job submission platforms or application programming interfaces to programmatically filter the database to the required information. Although some of the databases assessed by the project already contained chemical and physical property information, the records provided by PubChem contained records which were more recent and of higher quality. For example, chemical structure files contained within the DrugBank platform were only provided in a 2D format when accessed; however, PubChem compound records of the same chemical are provided in 3D co-ordinates.

### 1.3.9 KEGG

The Kyoto Encyclopedia of Genes and Genomes is a collection of databases which are used to document diseases, genes, drugs and compounds [57]. Initiated in 1995 at Kyoto University, Ogata *et al.* describe that the KEGG platform was designed as an effort to link sets of genes with a network of interacting molecules in cells [58]. They continue to describe the main database design, consisting of three segments:

- A pathway segment for representing additional details on interacting molecules
- A gene segment for documenting fully sequenced and partial genomes,
- A ligand segment for documenting all chemical compounds.

A segment of the database interactions contained within KEGG has been widely referenced as a baseline for considering machine learning methods with compounds interacting with specific protein panels by Yamanishi *et al.* [39]. This set, referred to as the "gold standard", provides a list of compound interactions by protein type, which are enzymes, nuclear receptors, GPCRs, and ion channels. However, while the interactions presented within these sets can be accessed and documented, higher throughput access to the KEGG database beyond individual queries requires the purchase of a subscription. The yearly end-user cost at the time of writing was \$2000 a year. With the availability of other freely downloadable databases, this database was not considered for this project beyond the information already present in the "gold standard" compound-protein interaction matrices.

## 1.4 Program Language and Tool Analysis

This section will provide a summary of the tools which were considered and used for retrieving and accessing the information used in the project. Where a tool has been used specifically for the work described in certain chapters, these tools will be specified within each section.

### 1.4.1 R

The statistical programming language R is a freely available tool which was developed and supported by the R Foundation for Statistical Computing [59]. Initially developed in 1993 from a mixture of the S programming language and the Scheme programming language, R provided support for utilising a number of statistical and visualisation techniques, as well as providing the ability for users to create customised functions and libraries through the use of packages to expand its functionality. A number of these libraries and extensions were used in conjunction with the R language throughout the project to process information and results, which are listed below:

- RStudio - an integrated development environment which provided a means to organise and run multiple scripts. Also provided a means to easily preview data objects in session for debugging [60].
- ReadR - a file reader library which provides an extension and additional functionality to R's existing file reader [61].

- Biostrings - a package which provides support for reading and manipulating compound and protein structure files [62].
- ChemmineR - a cheminformatics package for analysing small molecule data within R.
- RCurl - a package which allows R to download and manipulate information available on the Internet [63].
- RMySQL - a package which allows the access and execution of database queries via the MySQL query language [64].
- JSONLite - a package which allows the reading and access of files in the JSON format [65].
- RCDK - a package which allows the access and manipulation of compound structure files within the R environment [66].
- fingerprint - a package used in conjunction with RCDK to generate compound similarity matrices based on chemical fingerprinting, a process which generates a string of flags to quickly compare compounds [67].
- foreign - a package which allows conversion of items within R to files suitable for use within other programs such as WEKA [68].
- reshape2 - a package which provides R with the ability to easily transform a list of similarity values into a similarity matrix [69].
- doParallel and foreach - packages which provide R to perform a large number of similar tasks in parallel and make use of multiple CPU cores [70] [71].

The R language contains two main methods of execution: either through line by line execution, or through scripts which contain a series of instructions. Both methods save the output in a workspace which can either be accessed and altered directly, or saved to reduce the amount of processing needed to review results. Workspaces also provide an ideal space to debug potential faults that arise during development of the project.



### 1.4.2 Python

Python is a freely available programming language that was first distributed in 1991. Like R, Python is an interpreted language which processes individual commands or scripts of commands without the need for a compiler to create an executable, which also allows for Python projects to be cross platform compatible. Designed as a general purpose programming language to improve readability and simplicity of coding solutions, Python's functionality can be expanded through the installation of packages that allow the language to perform more specialized tasks such as bioinformatics and machine learning. One package, referred as PyDTI, makes use of the programming language is a library which performs prediction of new interactions between drugs and proteins using differing clustering techniques, which was used in the project implementation in Chapter 5 [72]. Another package, referred to as "Mordred" [73], parses through compound structure files to generate a list of chemical properties which are listed within the Appendix.

While Python could have been used to assist in compatibility with use of the clustering library, there were also some feature differences between the two languages; for example R's natural ability to save and restore a workspace is not possible with Python unless external packages are installed and used, which would have increased the difficulty of debugging and analysing some variables, so R was chosen over Python as the main language for the project. Python scripts could still be used in conjunction with R when needed however, as a Python script could still be called via command prompts using R. This reduced the project's development time of replicating existing clustering code as well the time for learning a new language set.

### 1.4.3 Matlab/Octave

The statistical programming language Matlab is a platform which was designed for use by engineers and scientists. With programming primarily focused on matrix-like structures, the language is ideally suited for solving and computing mathematical problems, however access to Matlab requires the purchase of a yearly subscription (some academic institutes include Matlab in their suites of software for students however). While open source alternatives to Matlab exist such as Octave, its functionality is limited to Matlab, which can also cause compatibility issues when code is used on both platforms. To maintain a wide audience of researchers that could replicate the project implementation, this language was not considered.

#### 1.4.4 MySQL

In order to reduce the difficulty of filtering the information stored within the protein and drug repositories, a database management system was necessary to maintain the data in a consistent format. These systems could then be queried to filter and generate the required information based on particular user parameters. One such management system was MySQL, which is a freely available open source database management system developed by the Swedish company MYSQL AB (now part of Oracle Corporation) in 1995. The MySQL system provides users with a system that could be easily and freely distributed across multiple platforms without concerns over licensing. One of the advantages of using MySQL is that the system is normally featured in combined software bundles known as AMPs (Apache-MySQL-PHP), which allow for easy setup of local web and database servers by combining the components together into a single package.

One of these software bundles which was used by the project was XAMPP [74], an open source package which is compatible with multiple operating systems, and an ideal platform for setup and replication of the project implementation. This also allowed the generation of a local prototype quickly for generating a website to demonstrate the project method and results.

#### 1.4.5 WEKA

The Waikato Environment for Knowledge Analysis (WEKA) is an application which contains a collection of machine learning algorithms used for solving classification problems [75]. Built on the Java programming language, the application also contains a number of visualisation and analysis tools which assist users with managing and pre-processing datasets, as well as providing a platform to easily save and re-use classification models. As a Java application, WEKA is cross-platform compatible, and can be either be run via a graphical user interface or a command prompt terminal similar to OpenBabel for programmatic access. Users can either perform individual classification experiments using WEKA's Explorer platform, or make use of the Experimental platform to perform multiple classification experiments at once on a dataset using equal model training and testing conditions. A package manager is also included for installing user defined classifier or pre-processing methods.

As WEKA is freely available to use, the program was considered as a convenient option for use in assessing drug profiling from protein interactions, as well as for replication of the implementation and results. In order to process information via the

WEKA platform, the program requires that data is loaded in a specific file format known as ARFF which specified the type of information that is stored (i.e if a column contains set values/classes, is a string or numeric). However, the R programming language contained a library which was suitable for creating this type of file format [68].

#### **1.4.6 OpenBabel**

OpenBabel is a program which provides the ability to process multiple chemical data languages [76], with one of its main features being the ability to convert files into multiple formats. The program also has the ability to process grouped chemical structure files into individual files for quicker access via database queries. As the PubChem platform returned a collection of compound structures as a single file, the OpenBabel platform was considered an ideal method of splitting these entries into individual records.

OpenBabel contains two main methods of operation: users can either make use of a Graphical User Interface to perform simple individual queries, or make use of a command prompt terminal for more thorough access of larger datasets. In addition to this, packages exist within R which can interact with the program and generate results in a format more suitable for further analysis [77]. The program and the packages associated with it are open source and free to use, with no requirements for licensing.

#### **1.4.7 HPC Wales/Supercomputing Wales**

To assist in the reduction of drug attrition, this project was conceived as part of a wider project which focused on the implementation and scaling of ligand (drug) docking algorithms on the high performance computing platform HPC Wales [78], which worked upon computed structures of the entire human proteome with particular application to drug discovery.

The HPC Wales project was a platform which allowed academic users to perform multiple programming operations in parallel on high specification hardware. To accomplish this, the platform design was based on a queueing system, where users submitted a set of tasks in a job which was then processed when the user defined requirements of resources were met. Jobs were then executed on the platform until a conclusion is reached, or a user or organisation defined time limit is reached in which case the processing is terminated. To request the use of the platform, academic

users were required to submit the type of work that was needed to be performed, the software required and the scale of the resources that were needed to accomplish the task. The project requisitioned the resources of HPC Wales to perform high scale *in silico* tests, as well as clustering operations which scaled orders of magnitude beyond what a desktop environment could accomplish in a reasonable timeframe.

Based at multiple Universities across Wales, the programme was restructured in 2017/2018 in the formation of Supercomputing Wales [79], which upgraded the specifications of the hardware and changed its target towards academic institutions and users, however support is still provided to commercial enterprises.

## 1.5 Summary

In this chapter, the project outlined a brief background of the rationale to the project, and the tasks that were to be accomplished in providing a potential alternative means to pharmacological profiling and drug target prediction. The development pipeline and attrition rates for drug development can be high due to the costs associated with clinical screening tests, and methods exist to potentially provide ways to reduce these risks and costs. This chapter also reviewed some of the database repositories available on the internet which provide information on drugs and protein targets, as well as tools which are used to process and assess them and to replicate alternative methods of pharmacological profiling and *in silico* screening. The following chapters describe the methods and tools in more detail and their application in the accomplishment of the specific tasks required by the project aims and objectives.

## 1.6 References

- [1] Oxford University Press. (2015). “Drug - Definition of Drug in English from the Oxford dictionary,” [Online]. Available: <http://www.oxforddictionaries.com/definition/english/drug> (visited on 10/29/2015).
- [2] Wishart, D. S., Knox, C., Guo, A. C., Sharavista, S., Hassanali, M., Stothard, P., Chang, Z., and Woolsey, J., “DrugBank: A comprehensive resource for in silico drug discovery and exploration,” *Nucleic Acids Research*, vol. 34, pp. D668–D672, Jan. 2006.
- [3] Dix, D. J., Houck, K. A., Martin, M. T., Richard, A. M., Setzer, R. W., and Kavlock, R. J., “The ToxCast program for prioritizing toxicity testing of environmental chemicals,” *Toxicological Sciences*, vol. 95, no. 1, pp. 5–12, 2007.

- [4] Bowes, J., Brown, A. J., Hamon, J., Jarolimek, W., Sridhar, A., Waldron, G., and Whitebread, S., “Reducing safety-related drug attrition: The use of in vitro pharmacological profiling,” *Nature Reviews*, vol. 11, no. 12, pp. 909–922, Dec. 2012.
- [5] Smithells, R. W. and Newman, C. G. H., “Recognition of thalidomide defects,” *Journal of medical genetics*, vol. 29, no. 10, pp. 716–723, 1992.
- [6] Vargesson, N., “Thalidomide-induced teratogenesis: History and mechanisms,” *Birth Defects Research Part C: Embryo Today: Reviews*, vol. 105, no. 2, pp. 140–156, 2015.
- [7] The Medicines and Healthcare Products Regulatory Agency, *Medicines & medical devices regulation: What you need to know*, <http://www.mhra.gov.uk/home/groups/comms-ic/documents/websiteresources/con2031677.pdf>, 2008.
- [8] The Stationary Office, *Report of an independent review of access to the Yellow Card scheme*, <http://www.mhra.gov.uk/home/groups/comms-ic/documents/websiteresources/con2015008.pdf>, Apr. 2004.
- [9] Medicines and Healthcare products Regulatory Agency. (2016). “Yellow Card scheme looks to the future at 50th anniversary forum,” [Online]. Available: <https://www.gov.uk/government/news/yellow-card-scheme-looks-to-the-future-at-50th-anniversary-forum> (visited on 02/05/2016).
- [10] HM Government. (1968). “Medicines Act 1968: Chapter 67,” [Online]. Available: [http://www.legislation.gov.uk/ukpga/1968/67/pdfs/ukpga\\_19680067\\_en.pdf](http://www.legislation.gov.uk/ukpga/1968/67/pdfs/ukpga_19680067_en.pdf) (visited on 02/04/2016).
- [11] HM Government. (2016). “Apply for a license to market a medicine in the UK,” [Online]. Available: <https://www.gov.uk/guidance/apply-for-a-licence-to-market-a-medicine-in-the-uk> (visited on 02/04/2016).
- [12] National Institute for Health and Care Excellence. (2016). “Who we are,” [Online]. Available: <https://www.nice.org.uk/about/who-we-are> (visited on 02/04/2016).
- [13] Scottish Medicines Consortium. (2016). “What we do,” [Online]. Available: [https://www.scottishmedicines.org.uk/About\\_SMC/What\\_we\\_do](https://www.scottishmedicines.org.uk/About_SMC/What_we_do) (visited on 02/04/2016).
- [14] Department of Health, Social Services and Public Safety. (2018). “Legislation covering medicines,” [Online]. Available: <https://www.health-ni.gov.uk/articles/legislation-covering-medicines> (visited on 08/14/2018).
- [15] All Wales Medicines Strategy Group. (2016). “About us,” [Online]. Available: [http://www.awmsg.org/awmsg\\_about\\_us.html](http://www.awmsg.org/awmsg_about_us.html) (visited on 02/04/2016).
- [16] Cairns, J., “Providing guidance to the NHS: The Scottish Medicines Consortium and the National institute for Clinical Excellence compared,” *Health Policy*, vol. 76, no. 2, pp. 134–143, 2006.

- [17] Law, M. T., “How do regulators regulate? Enforcement of the Pure Food and Drugs Act, 1907-38,” *Journal of Law, Economics, and Organization*, vol. 22, no. 2, pp. 459–489, 2006.
- [18] Kinch, M., Patridge, E., Plummer, M., and Hoyer, D., “An analysis of FDA-approved drugs for infectious disease: Antibacterial agents,” *Drug Discovery Today*, vol. 19, no. 9, pp. 1283–1287, Sep. 2014.
- [19] Borchers, A. T., Hagie, F., Keen, C. L., and Gershwin, M. E., “The history and contemporary challenges of the US Food and Drug Administration,” *Clinical Therapeutics*, vol. 29, no. 1, pp. 1–16, 2007.
- [20] Lipsky, M. S. and Sharp, L. K., “From idea to market: The drug approval process,” *The Journal of the American Board of Family Practice*, vol. 14, no. 5, pp. 362–367, 2001.
- [21] Whitmore, E., *Development of FDA-regulated medical products: a translational approach*, 2nd ed. Milwaukee, WI, USA: ASQ Quality Press, Jan. 2012, ch. 1, pp. 1–17.
- [22] Helder, T., “Chapter 12 introduction: Good laboratory practice,” in *Good Clinical, Laboratory and Manufacturing Practices: Techniques for the QA Professional*, The Royal Society of Chemistry, 2007, pp. 171–181, ISBN: 978-0-85404-834-2. DOI: 10.1039/9781847557728-00169.
- [23] Talbot, D. and Downes, N., “Chapter 1 introduction: Good clinical practice,” in *Good Clinical, Laboratory and Manufacturing Practices: Techniques for the QA Professional*, The Royal Society of Chemistry, 2007, pp. 3–11, ISBN: 978-0-85404-834-2. DOI: 10.1039/9781847557728-00001.
- [24] Dolman, J., “Chapter 26 introduction: Good manufacturing practice,” in *Good Clinical, Laboratory and Manufacturing Practices: Techniques for the QA Professional*, The Royal Society of Chemistry, 2007, pp. 371–385, ISBN: 978-0-85404-834-2. DOI: 10.1039/9781847557728-00369.
- [25] United States Food and Drug Administration. (2020). “eCFR - Code of Federal Regulations Title 21 Part 58: Good Laboratory Practice for Nonclinical Laboratory Studies,” [Online]. Available: <https://www.ecfr.gov/cgi-bin/text-idx?SID=9b179b470add6a7c8afd29fba0948dcd&mc=true&node=pt21.1.58&rgn=div5> (visited on 05/11/2020).
- [26] United States Food and Drug Administration. (Feb. 10, 2020). “Office of Pharmaceutical Quality — FDA,” [Online]. Available: <https://www.fda.gov/about-fda/center-drug-evaluation-and-research-cder/office-pharmaceutical-quality> (visited on 05/11/2020).
- [27] HM Government. (2020). “Medicines, medical devices and blood regulation and safety: Good practice, inspections and enforcement,” [Online]. Available: <https://www.gov.uk/topic/medicines-medical-devices-blood/good-practice> (visited on 05/11/2020).

- [28] Whitmore, E., *Development of FDA-regulated medical products: a translational approach*, 2nd ed. Milwaukee, WI, USA: ASQ Quality Press, Jan. 2012, ch. 6, pp. 63–76.
- [29] Paul, S. M., Mytelka, D. S., Dunwiddie, C. T., Persinger, C. C., Munos, B. H., Lindborg, S. R., and Schacht, A. L., “How to improve r&d productivity: The pharmaceutical industry’s grand challenge,” *Nature reviews Drug discovery*, vol. 9, no. 3, p. 203, 2010.
- [30] National Institutes of Health. (2016). “Glossary of common terms,” [Online]. Available: <http://www.nih.gov/health-information/nih-clinical-research-trials-you/glossary-common-terms> (visited on 01/14/2016).
- [31] Hay, M., Thomas, D. W., Craighead, J. L., Economides, C., and Rosenthal, J., “Clinical development success rates for investigational drugs,” *Nature Biotechnology*, vol. 32, no. 1, pp. 40–51, Jan. 2014.
- [32] Lesk, A. M., *Introduction to Bioinformatics*, 4th ed. Oxford University Press, 2014, ch. 6, pp. 264–265.
- [33] Morgan, S., Grootendorst, P., Lexchin, J., Cunningham, C., and Greyson, D., “The cost of drug development: A systematic review,” *Health Policy*, vol. 100, no. 1, pp. 4–17, 2011.
- [34] Sertkaya, A., Wong, H.-H., Jessup, A., and Beleche, T., “Key cost drivers of pharmaceutical clinical trials in the united states,” *Clinical Trials*, vol. 13, no. 2, pp. 117–126, 2016.
- [35] Ramsundar, B., Kearnes, S., Riley, P., Webster, D., Konerding, D., and Pande, V., “Massively multitask networks for drug discovery,” *arXiv preprint arXiv:1502.02072*, 2015.
- [36] Austin-Muttitt, K., “Modelling competition binding assays: A step towards in silico pharmacological profiling tools,” Report submitted to Swansea University, 2019.
- [37] Pereira, M., Costa, V. S., Camacho, R., Fonseca, N. A., Simões, C., and Brito, R. M., “Comparative study of classification algorithms using molecular descriptors in toxicological databases,” in *4th Brazilian Symposium on Bioinformatics: July 29-31, 2009; Porto Alegre, Brazil*, Guimarães, K. S., Panchenko, A., and Przytycka, T. M., Eds., Springer Berlin Heidelberg, 2009, pp. 121–132.
- [38] Chen, L., Lu, J., Huang, T., Yin, J., Wei, L., and Cai, Y.-D., “Finding candidate drugs for hepatitis c based on chemical-chemical and chemical-protein interactions,” *PLoS One*, vol. 9, no. 9, pp. 1–6, 2014.
- [39] Yamanishi, Y., Araki, M., Gutteridge, A., Honda, W., and Kanehisa, M., “Prediction of drug–target interaction networks from the integration of chemical and genomic spaces,” *Bioinformatics*, vol. 24, no. 13, pp. i232–i240, 2008.

- [40] Bruno, A. E., Charbonneau, P., Newman, J., Snell, E. H., So, D. R., Vanhoucke, V., Watkins, C. J., Williams, S., and Wilson, J., “Classification of crystallization outcomes using deep convolutional neural networks,” *PLOS one*, vol. 13, no. 6, e0198883, 2018.
- [41] Wishart, D. S., Knox, C., Guo, A. C., Sharavista, S., Hassanali, M., Stothard, P., Chang, Z., and Woolsey, J. (2016). “DrugBank: Abacavir,” [Online]. Available: <http://www.drugbank.ca/drugs/DB01048> (visited on 05/06/2016).
- [42] Wishart, D., Arndt, D., Pon, A., Sajed, T., Guo, A. C., Djoumbou, Y., Knox, C., Wilson, M., Liang, Y., Grant, J., *et al.*, “T3db: The toxic exposome database,” *Nucleic acids research*, vol. 43, no. D1, pp. D928–D934, 2014.
- [43] Wishart, D. S., Feunang, Y. D., Marcu, A., Guo, A. C., Liang, K., Vázquez-Fresno, R., Sajed, T., Johnson, D., Li, C., Karu, N., *et al.*, “Hmdb 4.0: The human metabolome database for 2018,” *Nucleic acids research*, vol. 46, no. D1, pp. D608–D617, 2017.
- [44] United States Environment Protection Agency. (2018). “Toxicity ForeCaster (ToxCast) Data — Safer Chemicals Research — US EPA,” [Online]. Available: <https://www.epa.gov/chemical-research/toxicity-forecaster-toxcas-ttm-data> (visited on 08/09/2018).
- [45] Tice, R. R., Austin, C. P., Kavlock, R. J., and Bucher, J. R., “Improving the human hazard characterization of chemicals: A tox21 update,” *Environmental health perspectives*, vol. 121, no. 7, pp. 756–765, 2013.
- [46] Sipes, N. S., Martin, M. T., Kothiyia, P., Reif, D. M., Judson, R. S., Richard, A. M., Houck, K. A., Dix, D. J., Kavlock, R. J., and Knudsen, T. B., “Profiling 976 toxcast chemicals across 331 enzymatic and receptor signaling assays,” *Chemical research in toxicology*, vol. 26, no. 6, pp. 878–895, 2013.
- [47] Gaulton, A., Hersey, A., Nowotka, M., Bento, A. P., Chambers, J., Mendez, D., Motow, P., Atkinson, F., Bellis, L. J., Cibrián-Uhalte, E., *et al.*, “The chembl database in 2017,” *Nucleic acids research*, vol. 45, no. D1, pp. D945–D954, 2016.
- [48] National Center for Biotechnology Information. (2018). “Home - PubMed - NCBI,” [Online]. Available: <https://www.ncbi.nlm.nih.gov/pubmed> (visited on 08/16/2018).
- [49] McKusick, V. A., *Mendelian inheritance in man: a catalog of human genes and genetic disorders*. JHU Press, 1998, vol. 1.
- [50] Günther, S., Kuhn, M., Dunkel, M., Campillos, M., Senger, C., Petsalaki, E., Ahmed, J., Urdiales, E. G., Gewiess, A., Jensen, L. J., *et al.*, “Supertarget and matador: Resources for exploring drug-target relationships,” *Nucleic acids research*, vol. 36, no. suppl\_1, pp. D919–D922, 2007.
- [51] Davis, A. P., Murphy, C. G., Saraceni-Richards, C. A., Rosenstein, M. C., Wieggers, T. C., and Mattingly, C. J., “Comparative toxicogenomics database: A knowledgebase and discovery tool for chemical–gene–disease networks,” *Nucleic acids research*, vol. 37, no. suppl\_1, pp. D786–D792, 2008.



- [52] Gilson, M. K., Liu, T., Baitaluk, M., Nicola, G., Hwang, L., and Chong, J., “Bindingdb in 2015: A public database for medicinal chemistry, computational chemistry and systems pharmacology,” *Nucleic acids research*, vol. 44, no. D1, pp. D1045–D1053, 2015.
- [53] The UniProt Consortium, “UniProt: A hub for protein information,” *Nucleic Acids Research*, vol. 43, no. D1, pp. D204–D212, Jan. 2015.
- [54] The Uniprot Consortium. (2018). “Retrieve/ID mapping,” [Online]. Available: <https://www.uniprot.org/uploadlists/> (visited on 08/10/2018).
- [55] The UniProt Consortium. (2018). “AR - Androgen receptor - Homo sapiens (Human) - AR gene & protein,” [Online]. Available: <https://www.uniprot.org/uniprot/P10275> (visited on 08/01/2018).
- [56] Kim, S., Thiessen, P. A., Bolton, E. E., Chen, J., Fu, G., Gindulyte, A., Han, L., He, J., He, S., Shoemaker, B. A., Wang, J., Yu, B., Zhang, J., and Bryant, S. H., “PubChem Substance and Compound databases,” *Nucleic Acids Research*, vol. 44, no. D1, pp. D1202–D1213, Jan. 2016.
- [57] Kanehisa, M., Goto, S., Sato, Y., Furumichi, M., and Tanabe, M., “Kegg for integration and interpretation of large-scale molecular data sets,” *Nucleic acids research*, vol. 40, no. D1, pp. D109–D114, 2011.
- [58] Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H., and Kanehisa, M., “Kegg: Kyoto encyclopedia of genes and genomes,” *Nucleic acids research*, vol. 27, no. 1, pp. 29–34, 1999.
- [59] R Core Team, *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria, 2015. [Online]. Available: <https://www.R-project.org>.
- [60] RStudio Team, *Rstudio: Integrated development environment for r*, RStudio, Inc., Boston, MA, 2015. [Online]. Available: <http://www.rstudio.com/>.
- [61] Wickham, H., Hester, J., and Francois, R., *Readr: Read rectangular text data*, R package version 1.1.1, 2017. [Online]. Available: <https://CRAN.R-project.org/package=readr>.
- [62] Pagès, H., Aboyoun, P., Gentleman, R., and DebRoy, S., *Biostrings: String objects representing biological sequences, and matching algorithms*, R package version 2.44.2, 2017.
- [63] Lang, D. T. and the CRAN Team, *RCurl: General Network (HTTP/FTP/...) Client Interface for R*, 2016. [Online]. Available: <https://CRAN.R-project.org/package=RCurl>.
- [64] Ooms, J., James, D., DebRoy, S., Wickham, H., and Horner, J., *Rmysql: Database interface and 'mysql' driver for r*, R package version 0.10.13, 2017. [Online]. Available: <https://CRAN.R-project.org/package=RMySQL>.
- [65] Ooms, J., “The jsonlite package: A practical and consistent mapping between json data and r objects,” *arXiv:1403.2805 [stat.CO]*, 2014. [Online]. Available: <https://arxiv.org/abs/1403.2805>.

- [66] Guha, R., “Chemical informatics functionality in r,” *Journal of Statistical Software*, vol. 18, no. 6, 2007.
- [67] Guha, R., *Fingerprint: Functions to operate on binary fingerprint data*, R package version 3.5.7, 2018. [Online]. Available: <https://CRAN.R-project.org/package=fingerprint>.
- [68] R Core Team, *Foreign: Read data stored by 'minitab', 's', 'sas', 'spss', 'stata', 'systat', 'weka', 'dbase', ...* R package version 0.8-69, 2017. [Online]. Available: <https://CRAN.R-project.org/package=foreign>.
- [69] Wickham, H., “Reshaping data with the reshape package,” *Journal of Statistical Software*, vol. 21, no. 12, pp. 1–20, 2007. [Online]. Available: <http://www.jstatsoft.org/v21/i12/>.
- [70] Microsoft Corporation and Weston, S., *Doparallel: Foreach parallel adaptor for the 'parallel' package*, R package version 1.0.14, 2018. [Online]. Available: <https://CRAN.R-project.org/package=doParallel>.
- [71] Microsoft Corporation and Weston, S., *Foreach: Provides foreach looping construct for r*, R package version 1.4.4, 2017. [Online]. Available: <https://CRAN.R-project.org/package=foreach>.
- [72] Liu, Y., Wu, M., Miao, C., Zhao, P., and Li, X.-L., “Neighborhood regularized logistic matrix factorization for drug-target interaction prediction,” *PLoS computational biology*, vol. 12, no. 2, e1004760, 2016.
- [73] Moriwaki, H., Tian, Y.-S., Kawashita, N., and Takagi, T., “Mordred: A molecular descriptor calculator,” *Journal of cheminformatics*, vol. 10, no. 1, p. 4, 2018.
- [74] Apache Friends. (2018). “XAMPP Installers and Downloads for Apache Friends,” [Online]. Available: <https://www.apachefriends.org/index.html> (visited on 08/14/2018).
- [75] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H., “The weka data mining software: An update,” *ACM SIGKDD Explorations Newsletter*, vol. 11, no. 1, pp. 10–18, 2009.
- [76] O’Boyle, N. M., Banck, M., James, C. A., Morley, C., Vandermeersch, T., and Hutchison, G. R., “Open babel: An open chemical toolbox,” *Journal of cheminformatics*, vol. 3, no. 1, p. 33, 2011.
- [77] Horan, K. and Girke, T., *Chemmineob: R interface to a subset of openbabel functionalities*, R package version 1.14.0, 2017. [Online]. Available: <https://github.com/girke-lab/ChemmineOB>.
- [78] HPC Wales. (2016). “About,” [Online]. Available: <http://www.hpcwales.co.uk/> (visited on 01/06/2016).
- [79] Supercomputing Wales. (2018). “About,” [Online]. Available: <https://www.supercomputing.wales/about/> (visited on 08/08/2018).

## Chapter 2

# DrugReferenceDatabase - Compiling interactions across various drug and protein repositories

### 2.1 Introduction

On researching the availability of interactions between drugs and proteins, it was found that each source had stored their information in various formats, which increased the difficulty to access and pre-process individual repository files with specific filters. The purpose of this chapter is to document the steps that were taken to extract the information from databases detailed in Chapter 1 with the intention of creating a database which would then aggregate and centralize all of the drug and human protein interaction information in a consistent format. Issues encountered during dataset pre-processing, and database design and development will also be discussed throughout this chapter, and the steps which were taken to resolve them.

### 2.2 Drug and Protein Repository Analysis

Before the requirements of the database could be considered, an analysis of the foundations of each database source was necessary to discover which elements could be documented, and which elements could be incorporated into the database. This section will include a description of the features from each individual database, and the steps needed to pre-process the databases to gather the information required. The sources will be split into two segments: the first being main database sources

which were the initial repositories used for the compilation of a drug profiling tool, and the second being additional repositories which were considered for investigating further interactions based on a centralized database field entry.

## 2.2.1 Main Database Sources

### 2.2.1.1 DrugBank

The DrugBank database contains a number of downloadable files which can be accessed by users. According to the database documentation [1], protein interactions for drugs can be accessed via two groups, the first being the type of interaction that exists between the drugs and proteins. These types are listed below:

- Target - a protein to which a drug binds and causes an alteration of normal function and a desirable effect.
- Enzyme - a protein which catalyzes chemical reactions
- Transporter - a protein which shuttles ions or molecules across membranes, either into or out of cells
- Carrier - a protein which binds to drugs and carries them to cell transporters.

In addition to these binding types, each group contains a sub-group known as pharmacologically active proteins, which list proteins directly related to the mechanism of action of at least one drug. Once a binding type has been selected, a user can then download interactions of proteins based on drug category groups. An individual drug can be assigned to one or more drug groups, which are described as follows:

- Approved - A drug which has been considered approved by one or more drug regulation body. These drugs are typically approved by the FDA
- Experimental - A compound which has been shown experimentally to bind to specific proteins
- Withdrawn - A drug which has been withdrawn by at least one regulator.
- Illicit - A drug which has been scheduled by at least one regulator
- Investigational - A drug which is currently undergoing a drug approval process.

All interaction types (Target, Enzyme, Transporters and Carriers) for approved drugs were downloaded for the purposes of profiling, and experimental drugs for the purposes of using these interactions as a potential test set for clustering operations. The DrugBank website also contained a Structure Data Format (SDF) file for each drug category group, of which all drugs listed in a group were populated into a single SDF file. The version of DrugBank that was downloaded was Version 4.3, which was released from June 22, 2015.

Figure 2.1 displays a window of protein targets with approved drugs, of which Table 2.1 describes the data contained within each column that was used for the project. All protein identifier files follow the same format. The first step was to filter proteins to the human species only, of which the csv documented it either as the string "Human" or the Latin name "Homo sapiens". There was also a number of records which were discovered to be human proteins but had no species information listed, in which case the file was manually amended after the species was verified via UniProt.

C	D	E	F	G	H	I	J	K	L	
Gene Name	GenBank Protein ID	GenBank Gene ID	UniProt ID	Uniprot Title	PDB ID	GeneCard ID	GenAtlas ID	HGNC ID	Species	Drug IDs
ftsI	1574687.L42023		P45059	FTSI_HAEIN					Haemophilus influenzae (strain	DB00303
HDC	32109.X54297		P19113	DCHS_HUMAN		HDC	HDC	HGNC:4855	Human	DB00117
NOS2	292242.L09210		P35228	NOS2_HUMAN		NOS2A	NOS2A	HGNC:7873	Human	DB00125; DB00155; DB01110; DB01
PYGL	183353.M14636		P06737	PYGL_HUMAN		PYGL	PYGL	HGNC:9725	Human	DB00131
AMT	391721.D13811		P48728	GCST_HUMAN		AMT	AMT	HGNC:473	Human	DB00116
CACNA1I	5565888.AF129133		Q9P0X4	CAC1I_HUMAN		CACNA1I	CACNA1I	HGNC:1396	Human	DB00568; DB00617; DB00661; DB0C
ADORA1	256195.S45235		P30542	AA1R_HUMAN		ADORA1	ADORA1	HGNC:262	Human	DB00201; DB00277; DB00640; DB0C
ABL1	28237.X16416		P00519	ABL1_HUMAN		ABL1	ABL1	HGNC:76	Human	DB00171; DB00619; DB01254; DB04
FCER1A	31318.X06949		P12319	FCERA_HUMAN		FCER1A	FCER1A	HGNC:3609	Human	DB00043; DB00895
F8	182818.M14113		P00451	FA8_HUMAN		F8	F8	HGNC:3546	Human	DB00055; DB00100
PTGS1	387018.M31822		P23219	PGH1_HUMAN		PTGS1	PTGS1	HGNC:9604	Human	DB00154; DB00159; DB00244; DB0C
ADRBK2	312395.X69117		P35626	ARBK2_HUMAN		ADRBK2	ADRBK2	HGNC:290	Human	DB00171
rpsD	42798.X02543		P0A7V8	RS4_ECOLI					Escherichia coli (strain K12)	DB00254; DB00256; DB00453; DB0C
DRD1	30397.X55760		P21728	DRD1_HUMAN		DRD1	DRD1	HGNC:3020	Human	DB00246; DB00248; DB00268; DB0C
TMP1	7537304.J04230		P12461	TYSY_CANAL					Yeast	DB01099
SLC13A1	12620132.AF260824		Q9BZW2	S13A1_HUMAN		SLC13A1	SLC13A1	HGNC:10916	Human	DB00139
FLT4	297050.X69878		P35916	VGFR3_HUMAN		FLT4	FLT4	HGNC:3767	Human	DB00398; DB01268; DB06589; DB0C
SCNN1B	1004271.X87159		P51168	SCNNB_HUMAN		SCNN1B	SCNN1B	HGNC:10600	Human	DB00384; DB00594
COL1A1	1419998.774616		B034E9	COL1A1_HUMAN		COL1A1	COL1A1	HGNC:2187	Human	DB00048

**Figure 2.1:** Screenshot of the DrugBank Protein Target Identifiers file. ID references the UniProt accession code, whereas the Species and Drug IDs fields reference the species attached to the protein and the drugs that are targets of the protein in question respectively

Field Name	Description
ID	The UniProt accession code for the protein
Species	The species associated with the protein
Drug IDs	The drugs associated with the protein, referenced by their ID on DrugBank. Stored in a semi-colon delimited format

**Table 2.1:** Fields of interest used for pre-processing the DrugBank protein interaction sets

Once the file had been filtered, it was then necessary to transform the aggregated protein interactions into aggregated drug interactions for the purposes of drug profiling. This would generate a separate interaction profile for each interaction type. The next step was to then merge the interaction types together into a single entity, providing a complete interaction profile for all drugs. Figure 2.2 displays a sample table of the filtered results after pre-processing, where the protein interactions were split

by binding type (targets, enzymes, transporters and carriers), and also listed by Drug Classification (set as 'Good-Profile' for FDA approved drugs and 'Exp-Profile' for experimental drugs). While this information would have been sufficient for use in building a profiling model in conjunction with the ToxCast set and for report formatting, these aggregate results needed to be split into individual protein-drug interactions for the process of compiling a database table of interactions.

A	B	C	D	E
DrugID	TargProtID	EnzProtID	TranProtID	CarrProtID
DB00001	P00734	NA	NA	NA
DB00002	P12314;P00533;O75015;P09871;P00736;P02745;P02746;P02747;P08637;P12318;P31994;P31995	NA	NA	NA
DB00004	P14784;P01589;P31785	NA	NA	NA
DB00005	P01375;P12314;O75015;P20333;P01374;P09871;P00736;P02745;P02746;P02747;P08637;P12318;P31994;P31995	P35354	NA	NA
DB00006	P00734	P05164	NA	NA
DB00007	P30968	NA	NA	NA
DB00008	P48551;P17181	NA	NA	NA
DB00009	P00747;P05121;P02671;Q03405	NA	NA	NA
DB00010	Q02643	NA	NA	NA
DB00011	P48551;P17181	NA	NA	NA
DB00012	P19235	NA	NA	NA
DB00013	P00747;P05121;Q03405;P00748;P98164;P05120;P00750;Q9Y5V6;P05154;P14543	NA	NA	NA
DB00014	P22888;P30968	NA	NA	NA
DB00015	P00747;P05121;P02671;Q03405	NA	NA	NA
DB00016	P19235	NA	NA	NA
DB00017	P30968	NA	NA	NA
DB00018	P48551;P17181	NA	NA	NA
DB00019	P08246;Q99082	NA	NA	NA
DB00020	P15509;P26951;P13727;P32927;P34741	NA	NA	NA
DB00021	P47872	NA	NA	NA

**Figure 2.2:** Screenshot of the DrugBank Interaction set after pre-processing was performed. Each protein is grouped into the binding type file it was discovered in for a particular drug, with each protein separated by a semi-colon (;).

### 2.2.1.2 ToxCast

The ToxCast database contains a summary file which is a matrix of compounds screened against a set of assays to determine if toxic effects can be detected. From the downloads section [2], two compressed archive files were download: one contained chemical information files, while the other contained a summary of the interaction results as well as details of the protein assays used. The version specified from the files is version 2.0, which was available on October 2015. From these summary files, the following were used for extracting compound, protein and interaction information:

- hitc\_Matrix\_151020.csv - A file which provides an interaction matrix
- DSSTox\_ToxCastRelease\_20151019.xlsx - A file providing detailed information on the chemicals used for interaction testing
- Chemical\_Summary\_151020.csv - A file providing summarised information on the compounds used for testing.
- Assay\_Summary\_151020.csv - A file providing details on the protein assays used for testing.

A screenshot of the summary matrix file can be seen on Figure 2.3. There are a number of values present within the matrix according to the documentation [3]: a value of 1 demonstrated that a binding had been found between a compound and an assay, whereas a 0 value indicated that no binding had been found. NA values represented pairs which had not yet been tested by the database, and a -1 value indicated that insufficient information existed from testing to draw a conclusion. The keys found on the outside of the interaction matrix reference the more detailed protein testing panels and compound detail tables. The fields used in the protein table are listed in Table 2.3, while the compound details are split into two tables, of which the main fields used are referenced in Table 2.2 to gather the compound information required for further processing.

A	B		C		D		E	
	ACEA_T47D_80hr_Negative	ACEA_T47D_80hr_Positive	APR_HepG2_CellCycleArrest_1h_dn	APR_HepG2_CellCycleArrest_1h_up	APR_HepG2			
100005		0	0NA	NA	NA			NA
1000051	NA	NA	NA	NA	NA			NA
10001135	NA	NA	NA	NA	NA			NA
100016		1	0NA	NA	NA			NA
100027		0	0NA	NA	NA			NA
10004441	NA	NA	NA	NA	NA			NA
100061	NA	NA	NA	NA	NA			NA
100107	NA	NA	NA	NA	NA			NA
100118	NA	NA	NA	NA	NA			NA
100141	NA	NA	NA	NA	NA			NA
100152	NA	NA	NA	NA	NA			NA

**Figure 2.3:** Screenshot of a segment of the ToxCast *in vitro* interaction summary matrix. Values of 1 demonstrated that a binding had found between a compound and protein, whereas values of 0 indicated no binding was found. NA values represented pairs which were inconclusive in binding or had not been tested.

Field Name	Description
chid	The ToxCast reference for a compound
code	The Chemistry abstracts service code referenced by the summary matrix. Each code begins with a C.
Substance.Name	The name of the compound
Structure.SMILES	The SMILES string of the compound

**Table 2.2:** Fields of interest that were used in the compound reference tables in ToxCast

Field Name	Description
assay_component_endpoint_name	The name of the protein assay tested in the summary matrix
organism_col_name	The organism associated with the protein assay
technological_target_uniprot_accession_number	The measured UniProt accession codes used in the assay
intended_target_uniprot_accession_number	The objective UniProt accession codes used in the assay

**Table 2.3:** Fields of interest that were used in the protein assay table in ToxCast

The first step in filtering was extracting all assays which contained only human organisms. Like DrugBank, there were some erroneous records present within the file which required manual correction. One protein, listed by UniProt accession code 'P04386' (Regulatory protein GAL4), was associated with baker's yeast instead of

human as specified by the assay information file. This was corrected manually in the source file. Once the required proteins had been found, the interaction matrix was then filtered down to assay panels linked to human organisms, and all binding and non-binding protein-compound pairs were documented by stepping through the matrix.

While the drug reference table contained information on individual drugs, there were certain proteins in the protein reference table which consisted of one or more UniProt protein identifiers. If an assay contained more than one protein it was delimited by a '|' character. It was therefore necessary to detail all the individual protein interactions in the event a multiple protein assay was found to be a target or non-binding to a compound. The protein assay reference table also contained information on technological targets, which according to the documentation [4] specified proteins that were measured in the assay in addition to the objective protein targets. These were also included in processing the interactions in ToxCast, to ensure as much information as possible was gathered to determine a link between a compound and a protein within the dataset.

Once the process had been completed, compounds found within ToxCast known to interact with a human assay were appended to the interactions found within DrugBank. These compounds had been labelled under an additional "DrugClass" category known as "Bad-Profile", to differentiate between FDA approved drugs ("Good-Profile") in the DrugBank repository. Columns to detail inactive protein-compound pairs were also documented in an additional column. Figure 2.4 provides a view of the dataset after addition of the ToxCast drugs. Note that as ToxCast did not categorise the type of interaction that had took place, all ToxCast interactions were listed under the TargetProt field along with DrugBank's targets.

A	B	C	D	E	F	G	H	I	J	K	L	M
DrugID	TargProtID	EnzProtID	TranProtID	CarrProtID	DrugClass	NoOfTargs	NoOfEnz	NoOfTran	NoOfCarr	MissProtID	NoOfMisses	
C100005	Q07869;O76074;Q069	NA	NA	NA	Bad-Profile	4	0	0	0	P03372;P059	129	
C1000051	NA	NA	NA	NA	Bad-Profile	0	0	0	0	Q16236;P10	15	
C10001135	P04637	NA	NA	NA	Bad-Profile	1	0	0	0	Q16236;P10	13	
C100016	P11509;P33261;Q969	NA	NA	NA	Bad-Profile	7	0	0	0	Q71U36;P8	116	
C100027	P03956;P29350;P102	NA	NA	NA	Bad-Profile	4	0	0	0	P03372;Q71	116	
C10004441	NA	NA	NA	NA	Bad-Profile	0	0	0	0	P03372;Q16	15	
C100061	P03372	NA	NA	NA	Bad-Profile	1	0	0	0	P05412;P01	71	
C100107	P10276;P10826;P139	NA	NA	NA	Bad-Profile	6	0	0	0	P05412;P01	66	
C100118	P04637	NA	NA	NA	Bad-Profile	1	0	0	0	Q16236;P10	14	
C100141	Q16236	NA	NA	NA	Bad-Profile	1	0	0	0	P10275;P35	14	
C100152	P03372;P04637	NA	NA	NA	Bad-Profile	2	0	0	0	Q16236;P10	13	
C10016203	P35869;O75469;P107	NA	NA	NA	Bad-Profile	5	0	0	0	Q71U36;P8	111	
C100185	Q43889;P10276;P107	NA	NA	NA	Bad-Profile	35	0	0	0	P05412;P01	70	
C100196	P04637	NA	NA	NA	Bad-Profile	1	0	0	0	Q16236;P10	14	
C100209	P04637	NA	NA	NA	Bad-Profile	1	0	0	0	P05412;P01	71	
C100210	P03372;Q92731	NA	NA	NA	Bad-Profile	2	0	0	0	P05412;P01	72	
C100221	Q16236;P35869;P113	NA	NA	NA	Bad-Profile	3	0	0	0	P10275;Q96	12	
C10022283	NA	NA	NA	NA	Bad-Profile	0	0	0	0	P05412;P01	68	
C10022318	NA	NA	NA	NA	Bad-Profile	0	0	0	0	Q16236;P10	15	
C10023548	P35869;P11511;P033	NA	NA	NA	Bad-Profile	3	0	0	0	P10275;P04	4	

Figure 2.4: Interaction Set with ToxCast drugs added



### 2.2.1.3 PubChem Compounds

In order to discover potential links between compounds within the repositories, it was necessary to find a database which was cross referenced by the majority of repositories, and also contained additional information which would have benefited the process of compiling information for the database. The PubChem database was found to be a platform which contained more detailed structure information files than the repositories themselves, as well as providing more detailed chemical property information. There are two main platforms on the PubChem website with regard to compounds: the Substance platform, which provides repository supplied information, and the Compound platform, which attempts to normalize a collection of substances into a unique compound record, and also provides chemical and structural information.

While compounds and substances can be viewed on an individual basis via the website, a systematic approach was necessary to successfully link the DrugBank and ToxCast repositories to the PubChem Compound platform which had no references beyond a SMILES structure, chemical name, and in some instances an SDF file. To accomplish this task, the PUGREST API was used [5] which was a system that made use of URL links to download filtered database information to users. This API was primarily used to find PubChem IDs on DrugBank and ToxCast, through two methods of searching:

1. First, attempting to identify the compound via the SMILES string, and
2. If no record is found via the SMILES string, then attempting to find a compound by its name.

For example, to search for a compound which contained the SMILES string of CCCC (Butane), the following link would need to be constructed: <https://pubchem.ncbi.nlm.nih.gov/rest/pug/compound/smiles/CCCC/cids/txt>. This would generate a text file of results to PubChem entries which would match the SMILES string, which at the time of writing would reference the compound number of 7843. A search for a compound by name such as butane would result in the following URL: <https://pubchem.ncbi.nlm.nih.gov/rest/pug/compound/name/butane/cids/txt>. Although named searches would have increased the likelihood of more than one result being generated, it was considered necessary for some searches as some compound SMILES strings generated errors when queried. In the event that either search methods generated more than one result, the first entry from the result generated by the API was used for linking.

A script was generated to iterate through all unique compounds of DrugBank and ToxCast following the above pattern of searching. On DrugBank, this was accomplished through finding the SMILES strings for any drug which had a structure; if any query had failed, the DrugBank structure found within its database would have been used. On ToxCast, the SMILES string in its compound reference table was used, with the chemical name as a fallback if the SMILES query had failed. On accomplishing these queries, it was found that only a small number of compounds could not be linked to PubChem. In the case of DrugBank, 1,399 out of 1,592 approved drugs found a reference to the PubChem Compound repository. 8 drugs were manually corrected to find a reference to PubChem Compound, which are detailed in Table 2.4. In total, 8,753 compounds from ToxCast and 1,407 drugs from the DrugBank approved group contained a link to the PubChem Compound repository.

DrugBank ID	DrugBank Name	PubChem Compound ID
DB00115	Cyanocobalamin	44176380
DB00475	Chlordiazepoxide	2712
DB00707	Porfimer sodium	57166
DB00895	Benzylpenicilloyl Polylysine	45266800
DB00995	Auranofin	24199313
DB01590	Everolimus	6442177
DB06290	Simeprevir	24873435
DB06439	Tyloxapol	70789242

**Table 2.4:** DrugBank records which were manually annotated to PubChem

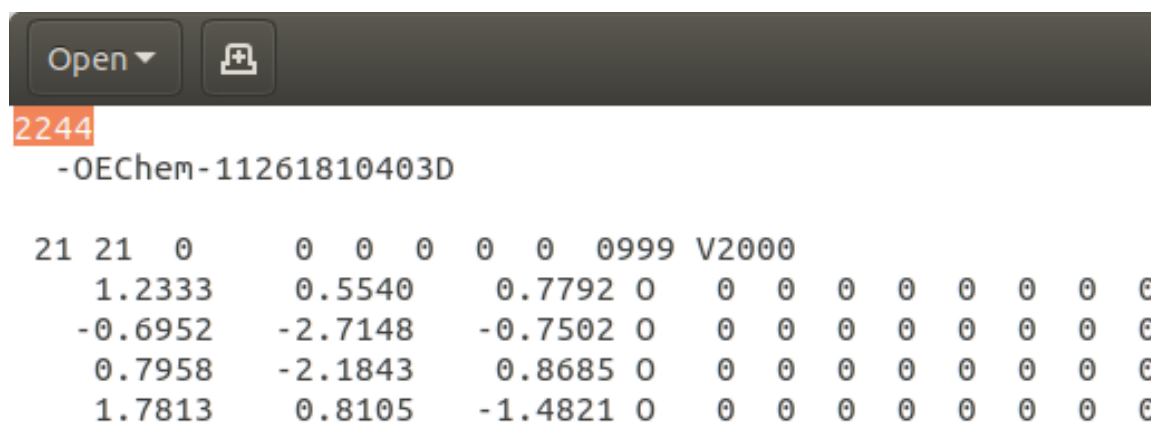
On development of a script using the PUGREST API interface, an alternative platform was found known as the PubChem Identifier Exchange Service [6]. This platform allowed users to provide multiple sets of information at once to perform queries instead of performing queries compound by compound. This could have been used on SMILES and drug synonyms to discover PubChem entries, but was not used as the searches had already been concluded using the PUGREST API. Instead, the exchange service had been used to discover the "Parents" of PubChem compounds, which are the foundation structure of PubChem compounds. These items were more suited to performing docking operations as they were more likely to contain 3D formats in their structure fields. While the exchange service provided an ideal platform for performing queries in bulk, the API allowed a platform prototype to conduct specific queries on custom compounds within a platform prototype.

Another web service which was used on PubChem was the Structure Download Service [7], which allowed users to download SDF files in bulk. On the provision of valid PubChem ID values, the website provides a single SDF file containing all the compounds queried provided they exist within the database. The platform could be

set to download files in either 2D or 3D co-ordinate format, however only compounds with a 3D format would be presented in the result file if the latter option was selected. As the file generated was a single SDF file, it was necessary to split it to reduce the time needed to process individual compound records. This was performed through a two-stage process:

1. Splitting the SDF file via OpenBabel, then
2. Applying a consistent storage and naming convention to the individual SDF files

While OpenBabel allowed users to split a compound into individual records, it did not name the split compounds by their PubChem compound ID, instead applying a simple counter to the split SDF files. On further investigation of the individual SDF files (as shown in Figure 2.5), the first line on each SDF was the PubChem compound ID. Through the use of a bash command (Figure 2.6), it was possible to apply a renaming command on all SDFs within a directory using the first line in each file.



**Figure 2.5:** A view of the contents of a PubChem SDF for aspirin. The first line of each SDF within PubChem makes reference to its Compound ID within the PubChem Compound system

```
for f in *.sdf; do mv "$f" "$(head -n1 "$f").sdf"; done;
```

**Figure 2.6:** Bash Command to rename batch SDF files generated by OpenBabel to PubChem Compound IDs

## 2.2.2 Additional Database Sources

After processing the DrugBank and ToxCast datasets, there were 2,305 human proteins present which are defined as proteins of interest. As this represented approximately 10% of the human genome, it was considered to be of suitable training and testing scale for the purposes of formulating a prototype within the project timescale, while also reducing the amount of resources and storage needed to process the whole human proteome from additional repositories. This section describes the process of extracting additional interaction information from these repositories, as well as any information considered beneficial for the purposes of filtering and further development of the database in the future.

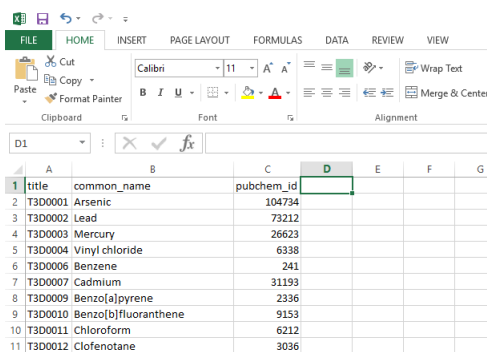
### 2.2.2.1 T3DB

Although the T3DB repository was designed and built by the same group that developed the DrugBank platform, the interaction files are stored in a somewhat different format. In order to obtain interactions from T3DB, two files were needed:

- The Toxin Data Field File, which contained information on drugs assessed by the repository and contained a reference to the PubChem Compound platform
- The Toxin-Target Mechanisms of Action file, which detailed the individual interactions between T3DB drugs and Uniprot accession codes

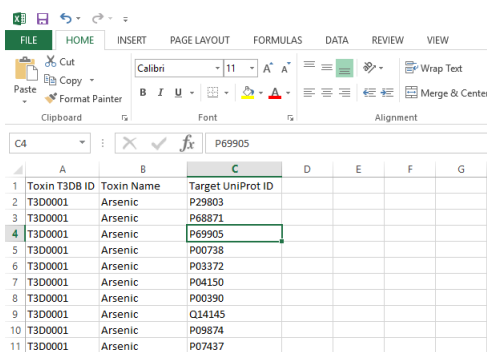
Both files were available in either a CSV or JSON format for access [8], which were downloaded in September 2017. When attempts were made to load the Toxin Data field file in a CSV format, there were errors present within the file which prevented complete loading within R. Further investigation of this error had found that the file contained text abstracts which contained commas and were not escaped correctly with speech quotes, which in turn caused the parser to generate more columns than necessary for certain compounds and fail. When the JSON format was used, all compounds had parsed successfully as the parser did not rely on commas to separate columns. Figure 2.7 shows the Toxin Data Field file parsed and truncated to columns of interest.

Figure 2.8 shows a truncated sample table of T3DB's mechanism of action file. For the purposes of the project, all compounds which contained a reference to PubChem and interacted with the protein targets detailed in ToxCast and DrugBank were considered and incorporated in the database.



	A	B	C	D	E	F	G
1	title	common_name	pubchem_id				
2	T3D0001	Arsenic	104734				
3	T3D0002	Lead	73212				
4	T3D0003	Mercury	26623				
5	T3D0004	Vinyl chloride	6338				
6	T3D0006	Benzene	241				
7	T3D0007	Cadmium	31193				
8	T3D0009	Benzo[a]pyrene	2336				
9	T3D0010	Benzo[b]fluoranthene	9153				
10	T3D0011	Chloroform	6212				
11	T3D0012	Clofenotane	3036				

Figure 2.7: A screenshot of T3DB Compounds



	A	B	C	D	E	F	G
1	Toxin T3DB ID	Toxin Name	Target UniProt ID				
2	T3D0001	Arsenic	P29803				
3	T3D0001	Arsenic	P68871				
4	T3D0001	Arsenic	P69905				
5	T3D0001	Arsenic	P00738				
6	T3D0001	Arsenic	P03372				
7	T3D0001	Arsenic	P04150				
8	T3D0001	Arsenic	P00390				
9	T3D0001	Arsenic	Q14145				
10	T3D0001	Arsenic	P09874				
11	T3D0001	Arsenic	P07437				

Figure 2.8: A screenshot of T3DB Targets

### 2.2.2.2 Matador

The interactions found within Matador are downloadable via a single tab separated file [9], of which a screenshot of the data downloaded from 27th September, 2017 is shown in Figure 2.9. The fields of interest within the Matador file were the 'chemical\_id' and 'uniprot\_id' fields which contained the PubChem Compound ID and UniProt accession codes, respectively. As the file contained aggregated interactions by compound, the only steps needed in preprocessing the dataset was to split the space delimited protein accession codes into individual interactions, and then to filter these codes to proteins of interest.

### 2.2.2.3 BindingDB

While users can download the entirety of the BindingDB database for access, the database curators have also split the interactions element of the database into tab separated files. These files contain references to other databases, one of which is PubChem. Figure 2.10 displays a screenshot of the main segments of the BindingDB interaction dataset downloaded from 1st July 2017, which provides interactions of PubChem compound IDs against UniProt accession codes.

A	B	C	D	E	F	
chemical_id	chemical_name	atc	protein_id	protein_name	mesh_id	uniprot_id
11954269	everolimus	L04AA18	9606.ENSP00000354587	FRAP1		Q9Y4I3 Q96QW8_HUMAN Q96QG3 Q6LE87
11954225	gold sodium thiomalate	M01CB01	9606.ENSP00000255040	APCS	D000209	P02743
11954225	gold sodium thiomalate	M01CB01	9606.ENSP00000273550	FTH1	D000209	Q3S5W1 P02794 Q6NZ44_HUMAN
11954225	gold sodium thiomalate	M01CB01	9606.ENSP00000336829	FGG	D000209	P04470 P04469 P02679 Q9UC63_HUMAN Q9
11954225	gold sodium thiomalate	M01CB01	9606.ENSP00000348068	SERPINA1	D000209	P01009 Q9UCM3_HUMAN Q9UCE6_HUMAN
11954225	gold sodium thiomalate	M01CB01	9606.ENSP00000277480	LCN2	D000209	Q5SYW0_HUMAN Q5SYV9_HUMAN P80188
11954225	gold sodium thiomalate	M01CB01	9606.ENSP00000349451	HPR	D000209	P00739 P00738 P00737 Q9ULB0 Q92659 Q92
11954225	gold sodium thiomalate	M01CB01	9606.ENSP00000349711	ORM1	D000209	P02763 Q8TC16 Q6IB74_HUMAN Q5U067_H
11954225	gold sodium thiomalate	M01CB01	9606.ENSP00000256733	ENSP00000256733	D000209	
11954225	gold sodium thiomalate	M01CB01	9606.ENSP00000278207	SAA3P	D000209	P22614 O95735
11954225	gold sodium thiomalate	M01CB01	9606.ENSP00000351029	HLA-A	D000209	P06338 P05534 P04439 P01892 P01891 O981
11954225	gold sodium thiomalate	M01CB01	9606.ENSP00000354291	ENSP00000354291	D000209	Q6NVW5_HUMAN
11954225	gold sodium thiomalate	M01CB01	9606.ENSP00000278222	SAA4	D000209	P35542 Q6FHJ4_HUMAN
11954225	gold sodium thiomalate	M01CB01	D000209	D000209		
11954225	gold sodium thiomalate	M01CB01	D002097	D002097		
11954225	gold sodium thiomalate	M01CB01	9606.ENSP00000217407	LBP	D000209	O43438 P18428 Q9UD66 Q9H403 Q92672 Q8
11954225	gold sodium thiomalate	M01CB01	9606.ENSP00000261981	SERPINA3	D000209	Q59GP9 Q13703 P01011 Q9UNU9 Q96DW8 C
11954225	gold sodium thiomalate	M01CB01	9606.ENSP00000291440	CPAMD8	D000209	Q9ULDT_HUMAN Q8NC09_HUMAN Q8I233
11954225	gold sodium thiomalate	M01CB01	9606.ENSP00000295807	ALP	D000209	Q9I170 Q9I143 Q9I171 Q9I157 Q9I147

Figure 2.9: A screenshot of the Matador interaction set

A	B	C	D
BindingDB Reactant_set_id	PubChem CID	UniProt (SwissProt) Entry Name of Target Chain	UniProt (SwissProt) Primary ID of Target Chain
53983	644735	RUNX1_HUMAN	Q01196
53984	649963	RUNX1_HUMAN	Q01196
53985	652912	RUNX1_HUMAN	Q01196
53986	5295075	RUNX1_HUMAN	Q01196
53987	656318	RUNX1_HUMAN	Q01196
53988	657445	RUNX1_HUMAN	Q01196
53989	657862	RUNX1_HUMAN	Q01196
53990	659293	RUNX1_HUMAN	Q01196
53991	659456	RUNX1_HUMAN	Q01196
53992	660761	RUNX1_HUMAN	Q01196
53993	661575	RUNX1_HUMAN	Q01196
53994	662757	RUNX1_HUMAN	Q01196
53995	662970	RUNX1_HUMAN	Q01196
53996	663645	RUNX1_HUMAN	Q01196
53997	664843	RUNX1_HUMAN	Q01196
53998	666073	RUNX1_HUMAN	Q01196
53999	666651	RUNX1_HUMAN	Q01196
54000	666800	RUNX1_HUMAN	Q01196
54001	666864	RUNX1_HUMAN	Q01196

Figure 2.10: A screenshot of a segment of the BindingDB interactions set and the fields which were used for the purposes of finding interactions of interest

According to the documentation, interactions are listed individually, though in the case of UniProt accession codes there is a column (UniProt Secondary IDs) which lists deprecated or obsolete protein codes at the time of the dataset's creation. As the project was working on information obtained from 2015 onwards, this field was also considered in documenting potential interactions in the event an obsolete protein had been recorded. The secondary list was delimited by commas, which required splitting so that all protein codes were considered.

#### 2.2.2.4 ChEMBL

In order to extract interactions from ChEMBL, it was necessary to download a MySQL dump of the database to perform the necessary queries [10]. The version of ChEMBL used was version 23, downloaded on September 2017. While cross references to PubChem existed within this version, the PubChem references were solely to the PubChem Substance platform. In order to obtain the interactions of interest, a two-step query was performed:

1. Query the ChEMBL platform to determine which ChEMBL compounds contained a reference to a PubChem Substance
2. Query the ChEMBL platform to determine which ChEMBL compounds had proteins listed in DrugBank/ToxCast listed as targets

These queries were separated to reduce the amount of duplicated records that would have been generated in the interactions from different substance references. The query for obtaining substance references is shown on Figure 2.11. With the retrieval of all substances from the first query, PubChem's identifier exchange service was used to convert the database's stored substances to the PubChem Compound platform. In most cases, multiple substance references involved single compounds, however in the case of 680 ChEMBL records there were entries which referred to multiple compounds. These compounds were processed directly by their ChEMBL ID in the Identifier Exchange Service to determine the correct individual record for the database.

```
SELECT DISTINCT md.molregno, md.chembl_id, cr.src_compound_id
FROM activities act JOIN molecule_dictionary md
ON act.molregno = md.molregno JOIN compound_structures cs
ON md.molregno = cs.molregno JOIN compound_records cr
ON cr.molregno = act.molregno JOIN source src
ON src.src_id = cr.src_id AND src.src_id = '7';
```

**Figure 2.11:** SQL Query used to gather PubChem Substance references of ChEMBL compounds

An example query used to extract all PubChem interactions with UniProt proteins is shown on Figure 2.12. To save on space, only a few example proteins were included in the figure, however this query was executed to find all interactions involving the human protein accession codes found in DrugBank and ToxCast.

```
SELECT DISTINCT mol.molregno, mol.chembl_id, comp.accession
FROM component_sequences as comp, target_components as tc,
compound_records cr, source src, assays as assay,
activities as act, molecule_dictionary as mol
WHERE comp.accession IN ("A5X5Y0", "A8MPY1", "A9UF02", "000141")
AND tc.component_id = comp.component_id AND assay.tid = tc.tid
AND act.assay_id = assay.assay_id AND mol.molregno = act.molregno
AND cr.molregno = act.molregno AND src.src_id = cr.src_id
AND src.src_id = '7' ORDER BY comp.accession;
```

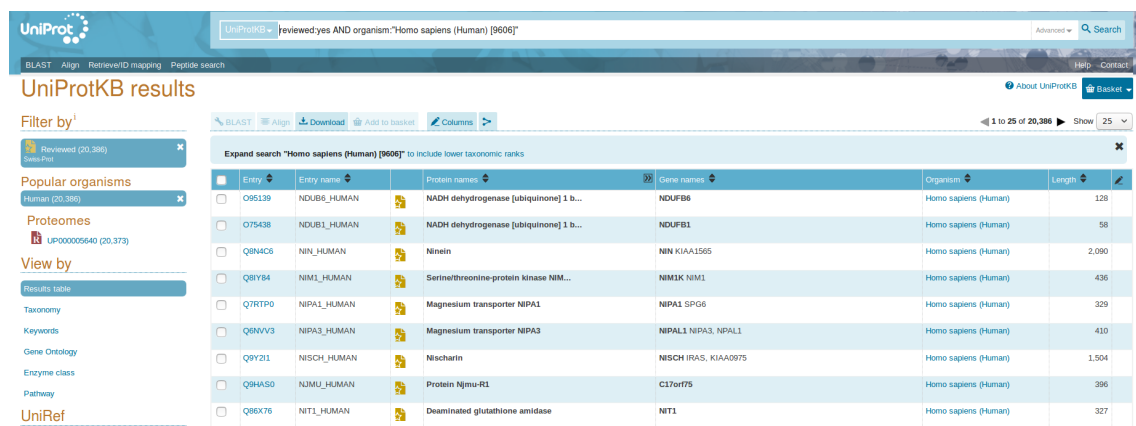
**Figure 2.12:** SQL Query used to gather UniProt interactions of ChEMBL compounds that contain a reference to PubChem

### 2.2.2.5 UniProt FASTA Files

While most compound databases contained references to the UniProt website, it was necessary to use some of the UniProt website's functions to gather additional information, such as retrieving all accession codes from the human genome and to obtain tissue groups associated with each protein. Searching the database would also have provided the ability to search for potential obsolete records which had been gathered during the extraction process.

To retrieve all human proteins, the website provided an interactive view of a download facility for proteins which have been filtered. As Figure 2.13 shows, it was possible to gather the codes by simply selecting "Human" from the organism tab and then selecting the download button to retrieve the accession codes in a readable file format. The viewer could also be altered to select specific fields, which included the amino acid string. The download option on the viewer provided users with the ability to download the filtered results, as well as text files of the individual proteins grouped together to provide information which was not present in the column selection window. Figure 2.14 displays an example of a protein text output which provided information on the tissue groups that were associated with that protein. While the project was mainly focused on the extraction of the proteins present in DrugBank and ToxCast, all documented human proteins within UniProt were extracted to assist in the process of integrating the whole human proteome for future iterations of the database.

In order to effectively group proteins into similar tissue groups, a compartment list of 47 tissue groups was generated from popular tissue groups within the protein text files. These tissue names are listed within Table 2.5, of which 61,777 protein/tissue associations were found when querying the 20,224 human proteins found.



Entry	Entry name	Protein name	Gene name	Organism	Length
Q65139	NDUB6_HUMAN	NADH dehydrogenase [ubiquinone] 1 b...	NDUF66	Homo sapiens (Human)	128
Q75438	NDUB1_HUMAN	NADH dehydrogenase [ubiquinone] 1 b...	NDUF61	Homo sapiens (Human)	58
Q8W4C6	NIN_HUMAN	Ninein	NIN KIAA1565	Homo sapiens (Human)	2,090
Q8Y184	NIM1_HUMAN	Serine/threonine-protein kinase NM...	NIM1K NIM1	Homo sapiens (Human)	436
Q7RT90	NIP1_HUMAN	Magnesium transporter NIP1	NIP1 SPG6	Homo sapiens (Human)	329
Q8NVV3	NIP3_HUMAN	Magnesium transporter NIP3	NIPAL1 NIP3, NIPAL1	Homo sapiens (Human)	410
Q8Y211	NISCH_HUMAN	Nischarin	NISCH IRAS, KIAA0975	Homo sapiens (Human)	1,504
Q8HAS0	NJMU_HUMAN	Protein Njmu-R1	CLT6r75	Homo sapiens (Human)	396
Q86X76	NT1_HUMAN	Deaminated glutathione amidase	NT1	Homo sapiens (Human)	327

Figure 2.13: UniProt results when filtered to the human organism only [11]



```

ID ANDR_HUMAN Reviewed; 920 AA.
AC P10275; A0A0B4J1T2; A2R0W2; B1AKD7; C0JKD3; C0JKD4; E7EVX6; Q9UD95;
DT 01-JUL-1989, integrated into UniProtKB/Swiss-Prot.
DT 16-MAR-2016, sequence version 3.
DT 18-JUL-2018, entry version 264.
DE RecName: Full=Androgen receptor;
DE AltName: Full=Dihydrotestosterone receptor;
DE AltName: Full=Nuclear receptor subfamily 3 group C member 4;
GN Name=AR; Synonyms=DHTR, NR3C4;
OS Homo sapiens (Human).
OC Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
OC Catarrhini; Hominidae; Homo.
OX NCBI_TaxID=9606;
RN [1]
RP NUCLEOTIDE SEQUENCE [MRNA] (ISOFORM 1).
RX PubMed=3216866; DOI=10.1210/mend-2-12-1265;
RA Lubahn D.B., Joseph D.R., Sar M., Tan J., Higgs H.N., Larson R.E.,
RA French F.S., Wilson E.M.;
RT "The human androgen receptor: complementary deoxyribonucleic acid
RT cloning, sequence analysis and gene expression in prostate.";
RL Mol. Endocrinol. 2:1265-1275(1988).
RN [2]
RP NUCLEOTIDE SEQUENCE [MRNA] (ISOFORM 1).
3X TISSUE SPECIES
RX PubMed=3174628; DOI=10.1073/pnas.85.19.7211;
RA Chang C., Kokontis J., Liao S.;
RT "Structural analysis of complementary DNA and amino acid sequences of
RT human and rat androgen receptors.";
RL Proc. Natl. Acad. Sci. U.S.A. 85:7211-7215(1988).
RN [3]
RP NUCLEOTIDE SEQUENCE [MRNA] (ISOFORM 1).
RC TISSUE=Prostate;
RX PubMed=2911578; DOI=10.1073/pnas.86.1.327;
RA Tilley W.D., Marcelli M., Wilson J.D., McPhaul M.J.;
RT "Characterization and expression of a cDNA encoding the human androgen
RT receptor.";
RL Proc. Natl. Acad. Sci. U.S.A. 86:327-331(1989).
RN [4]

```

**Figure 2.14:** Text entry of the protein Androgen receptor (ANDR\_HUMAN), Uniprot accession code P10275 [12]. Note the highlighted text displays a tissue field associated with the protein

Adipose Tissue	Mammary Gland
Adrenal Gland	Mesenchyma - Stem Cell
Amnion	Mesothelium
Brain	Muscle - Connective Tissue
Blood Vessels	Nasopharynx
Blood Circulation	Nervous System
Bone	Ovary
Cartilage	Pancreas
Carcinoma	Parathyroid Gland
Cervix	Penis
Colon	Placenta
Ear	Prostate
Embryo - Fetus	Skin
Endothelium	Smooth Muscle
Epithelium	Synovium
Esophagus	Stomach
Eye	Testis
Fibroblast	Thyroid
Hair	Tooth
Heart	Umbilicus
Intestine	Urinary Bladder
Kidney	Uterus
Liver	Vagina
Lung	

**Table 2.5:** List of Compartments used for gathering tissue groups for human protein text files gathered via UniProt

### 2.2.2.6 CTDBase

The structure of CTDBase allowed users to either download segments of interactions or associations [13], or to construct queries to generate filtered results on batch queries [14]. The areas of interest from this database were Chemical-gene interactions and the gene disease associations, however there were some practicality issues in

downloading these sections of the database. While the chemical-gene interaction table was of a fairly small size (22MB compressed from the set downloaded at February 2016), the gene disease association table was over 1GB in size compressed, which meant it would have been more efficient to use the batch query function using the protein accession codes gathered from DrugBank and ToxCast to obtain a set of disease associations. This batch query platform, shown in Figure 2.15, allowed users to obtain either directly referenced or assumed associations between specific proteins and diseases amongst other connections, however the values that the platform accepts for search terms did not allow UniProt accession codes. The lack of cross-referencing is also apparent with the chemicals, with no PubChem compound values present in either the interaction set or CTDBase's chemical vocabulary set.

To resolve the cross-referencing issues with proteins, CTDBase's gene vocabulary list was downloaded, which contained cross references to UniProt accession codes. Once these CTD protein ID values were obtained, the chemical-gene interaction dataset could then be filtered to compounds which only interacted with proteins of interest. The ID values also permitted a search for gene-disease associations, provided that the numerical values were specified in the platform to the gene reference values. For example, a gene ID with the value of 2 would need to have the search term 'gene:2' in the search field. This search was primarily focused on curated gene-disease associations only. To resolve the cross referencing issues with compounds, PubChem Compound IDs were passed to the identity exchange service which directed to database references within CTDBase.

Figures 2.16 and 2.17 display screenshots of some of the results generated by the batch query and part of the interaction dataset, whereas Table 2.6 displays the fields which were of interest to document within the database.

**1 Select your input type**

- Chemicals (MeSH® names, synonyms, or IDs, or CAS RNs) ?
- Diseases (MeSH or OMIM names, synonyms, or IDs) ?
- Genes (NCBI symbols or IDs) ?
- Gene Ontology terms (GO names, synonyms, or IDs) ?
- Pathways (KEGG or REACTOME names or IDs) ?
- References (PubMed® IDs or DOIs) ?

**2 Provide query terms (up to 4,000)**

Return-, tab- or |-delimited

Or upload a tab-separated file:

Browse... No file selected.

Identifiers column: 1

**3 Choose data to download**

**Data**

**Chemical-gene interactions** ?

Curated

**Chemical associations** ?

Curated

**Gene associations** ?

Curated

**Disease associations** ?

All

Curated

Inferred

**Pathway associations** ?

Inferred

Enriched (recommended)

**Gene Ontology associations** ?

Enriched (recommended)

All

**Format**

TSV (tab-separated values)

CSV (comma-separated values)

JSON

XML

**Download** **Clear**

See also: [Data Downloads](#) to obtain complete data sets.

**Figure 2.15:** Screenshot of the CTD Batch Query interface [14]. Users can select associations which have a direct reference (Curated) or from links generated by prior knowledge within the database (Inferred)

Input	DiseaseName	DiseaseID	GeneSymbol	GeneID	DiseaseCategories
gene:1 [No associated data found]					
gene:10	Breast Neoplasms	MESH:D001943	NAT2		10 Cancer Skin disease
gene:10	Coronary Artery Disease	MESH:D003324	NAT2		10 Cardiovascular disease
gene:10	Dermatitis, Occupational	MESH:D009783	NAT2		10 Occupational disease Skin disease
gene:10	Drug Hypersensitivity	MESH:D004342	NAT2		10 Immune system disease
gene:10	Drug-Induced Liver Injury	MESH:D056486	NAT2		10 Digestive system disease
gene:10	Drug-Related Side Effects and Adverse Reactions	MESH:D064420	NAT2		10
gene:10	Leukopenia	MESH:D007970	NAT2		10 Blood disease
gene:10	Liver Neoplasms	MESH:D008113	NAT2		10 Cancer Digestive system disease
gene:10	Mesothelioma, Malignant	MESH:D062839	NAT2		10 Cancer Respiratory tract disease
gene:10	Neural Tube Defects	MESH:D009436	NAT2		10 Congenital abnormality Nervous system disease
gene:10	Precursor Cell Lymphoblastic Leukemia-Lymphoma	MESH:D054198	NAT2		10 Cancer Immune system disease Lymphatic disease
gene:10	Prostatic Neoplasms	MESH:D011471	NAT2		10 Cancer Urogenital disease (male)
gene:10	Urinary Bladder Neoplasms	MESH:D001749	NAT2		10 Cancer Urogenital disease (female) Urogenital disease (male)
gene:100	Autistic Disorder	MESH:D001321	ADA		100 Mental disorder

**Figure 2.16:** Screenshot of the results of querying diseases linked to human proteins

A	B	C	D	E	F	G	H	
ChemicalName	ChemicalID	CasRN	GeneSymbol	GeneID	GeneForms	Organism	OrganismID	Interaction
10074-G5	C534883		MAX	4149	protein			10074-G5 affects the folding of
10074-G5	C534883		MAX	4149	protein			10074-G5 inhibits the reaction [I
10074-G5	C534883		MYC	4609	protein	Homo sapiens		9606 10074-G5 gene[log results in decre
10074-G5	C534883		MYC	4609	protein	Homo sapiens		9606 10074-G5 results in decreased i
10074-G5	C534883		MYC	4609	protein	Homo sapiens		9606 10074-G5 results in decreased i
10074-G5	C534883		MYC	4609	protein			10074-G5 affects the folding of
10074-G5	C534883		MYC	4609	protein			10074-G5 inhibits the reaction [I
10.10-bis(4-pyridinylmethyl)-9(10H)-anthracenone	C112297		FOS	2353	protein	Mus musculus		10090 10.10-bis(4-pyridinylmethyl)-9(1
10.10-bis(4-pyridinylmethyl)-9(10H)-anthracenone	C112297		KCNQ1	3784	protein			10.10-bis(4-pyridinylmethyl)-9(1
10.10-bis(4-pyridinylmethyl)-9(10H)-anthracenone	C112297		KCNQ2	3785	protein	Mus musculus		10090 10.10-bis(4-pyridinylmethyl)-9(1
10.10-bis(4-pyridinylmethyl)-9(10H)-anthracenone	C112297		KCNQ2	3785	protein			10.10-bis(4-pyridinylmethyl)-9(1
10.10-bis(4-pyridinylmethyl)-9(10H)-anthracenone	C112297		KCNQ3	3786	protein			10.10-bis(4-pyridinylmethyl)-9(1
10.11-dihydro-10.11-dihydroxy-5H-dibenzazepine-5-carboxamide	C004822	35079-97-1	EPHX1	2052	gene	Homo sapiens		9606 [EPHX1 gene SNP affects the r
10.11-dihydro-10.11-dihydroxy-5H-dibenzazepine-5-carboxamide	C004822	35079-97-1	EPHX1	2052	protein	Homo sapiens		9606 [EPHX1 protein results in increa
10.11-dihydro-10.11-dihydroxycarbamazepine	C039775		ABCB1	5243	protein	Homo sapiens		9606 ABCB1 protein results in increa

**Figure 2.17:** Screenshot of a segment of the CTDBase Interactions

Field Name	Description
GeneID	The Entrez Gene ID identifier
DiseaseName	The name of the disease identified to be connected to a gene
ChemicalID	CTDBase's reference of a compound

**Table 2.6:** Fields of interest within the CTDBase files. GeneID references are possible to be cross-referenced to UniProt proteins via the UniProt website [15]. ChemicalID references are possible to be cross-referenced via PubChem's Identifier Exchange Service [6].

### 2.2.2.7 PubChem BioAssay

In addition to storing information on compounds, PubChem also features a platform which stored interactions between compounds and proteins. Known as Bioassay, this platform also contains non-interacting compound-protein pairs, which was considered a valuable additional source of additional information as only ToxCast contained information on non-interacting pairs. Two methods of access existed for obtaining the BioAssay interaction data: either through downloading and filtering the entirety of the BioAssay platform by downloading individual assay XML files via the FTP [16], or through accessing specific interactions via a URL. As the former method involved processing approximately 1 million bioassays to determine which contained UniProt accession codes, it was considered more practical to make use of programmatic URL access to download interactions on the 2,305 human proteins required for documentation in the project.

The method of access for BioAssay however was more complex in comparison to using the PUGREST API. For example, to query all interactions associated with a UniProt accession code, a user would need to construct the following link <https://pubchem.ncbi.nlm.nih.gov/assay/pcget.cgi?> with the additional arguments detailed in Table 2.7, with each argument separated by an '&' symbol.

On submission of a query, a tab delimited file was returned which contained all the active or inactive bindings listed within the BioAssay platform for that particular protein, an example of which is shown on Figure 2.18. These results tables were filtered to remove instances where no PubChem Compound ID was referenced. A limitation of

Argument	Description	Example
task	The type of information required from PubChem	task=bioactivity_sameseq
protacxn	The UniProt protein accession code	protacxn=P10275
start	At what position in the interaction list to start the download	start=1
limit	At what position in the interaction list to limit the download to	limit=1000000
actvty	What type of bindings to download (active or inactive)	actvty=active

**Table 2.7:** The arguments passed to PubChem to retrieve BioAssay interactions or inactive bindings for a single UniProt accession code

the scraping of these results is that the specific assay experiment details were not able to be extracted, however a method exists within the PubChem website to trace back to the assay for a particular compound and protein. For example, to extract the assay results generated between the Epidermal Growth Factor Receptor (UniProt Accession Code P00533) and the compound N-(3-Bromophenyl)-6,7-diethoxyquinazolin-4-amine (PubChem Compound ID 2857), entering the following URL <https://pubchem.ncbi.nlm.nih.gov/target/protein/P00533#cid=2857> will generate a page of all experiments which have taken place between that protein and compound. While impractical for bulk gathering, it is a suitable method for investigating individual experiments of interest from the BioAssay platform.

Another potential issue discovered during the download process was that of querying for inactive compound protein pairs, some compound protein pairs made reference to a RefSeq accession code as shown in Figure 2.19. These codes were being used in the same way as UniProt accession codes by the BioAssay system when the results table was compiled. While converting these codes into UniProt accession codes was a trivial matter, the different codes can lead to issues in tracking individual results within BioAssay's XML files, as no reference is made to the RefSeq accession codes.

activity	acname	acvalue	cid	acc
Active	EC50	2.00E-05	44356251	P14416
Active	Kd	3.00E-05	53248209	P14416
Active	EC50	3.16E-05	76325149	P14416
Active	Ki	4.00E-05	37459	P14416
Active	Ki	4.00E-05	208951	P14416
Active	EC50	5.01E-05	76325155	P14416
Active	Ki	5.75E-05	25069121	P14416
Active	Ki	6.00E-05	59227	P14416
Active	Ki	7.00E-05	53248676	P14416
Active	Ki	8.60E-05	57267	P14416
Active	Ki	9.70E-05	10021692	P14416
Active	Ki	0.0001	208951	P14416
Active	Ki	0.00011	11957529	P14416
Active	Ki	0.00012	5265	P14416
Active	Ki	0.00013	11595240	P14416
Active	Ki	0.000133	119146	P14416

**Figure 2.18:** A screenshot of BioAssay active results of D(2) dopamine receptor (Uniprot Accession Code P14416)

activity	acname	acvalue	cid	acc
Inactive	Potency	0.0112	12456	NP_000786
Inactive	Potency	0.055963	57377248	NP_000786
Inactive	Potency	0.111661	57383410	NP_000786
Inactive	Potency	0.140572	57383407	NP_000786
Inactive	Potency	0.1986	5702225	NP_000786
Inactive	Potency	0.353102	16007841	NP_000786
Inactive	Potency	1.57725	17171662	NP_000786
Inactive	Potency	5.87	3917	NP_000786
Inactive	Potency	19.8574	16758194	NP_000786
Inactive	IC50	20	72737723	NP_000786
Inactive	IC50	20	72737738	NP_000786
Inactive	Potency	22.2803	3218	NP_000786
Inactive	Potency	1.00E+06	57377249	NP_000786
Inactive				NP_000786
Inactive				6 NP_000786
Inactive				19 NP_000786

**Figure 2.19:** A screenshot of BioAssay inactive results of D(2) dopamine receptor (Uniprot Accession Code P14416)

## 2.3 Database Design and Implementation

In order to satisfy the requirements of the project, the database design needed to possess the following capabilities:

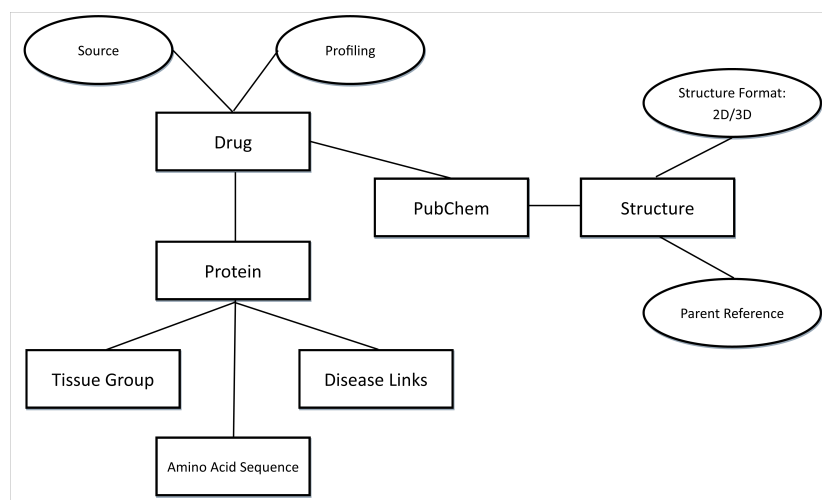
- To document all compounds and drugs and their modes of action (active or inactive) with human proteins on the database repositories considered
- To differentiate between any compounds found to be initially classified as "Good" (DrugBank approved drugs) or "Bad" (ToxCast screened compounds).
- To cross-reference all drug and compound sources within a centralised database (PubChem)
- To cross-reference all human proteins within a centralised database (UniProt)

- To highlight which databases contained information relating to specific *in vitro* protein-compound interactions
- To document the location of the structures of all compounds, if available
- To document the amino acid sequence of all proteins, if available

In addition to the above, the following features were also considered to be desirable for the purposes of additional filtering:

- To document the tissue groups associated with each protein
- To document all diseases associated with each protein

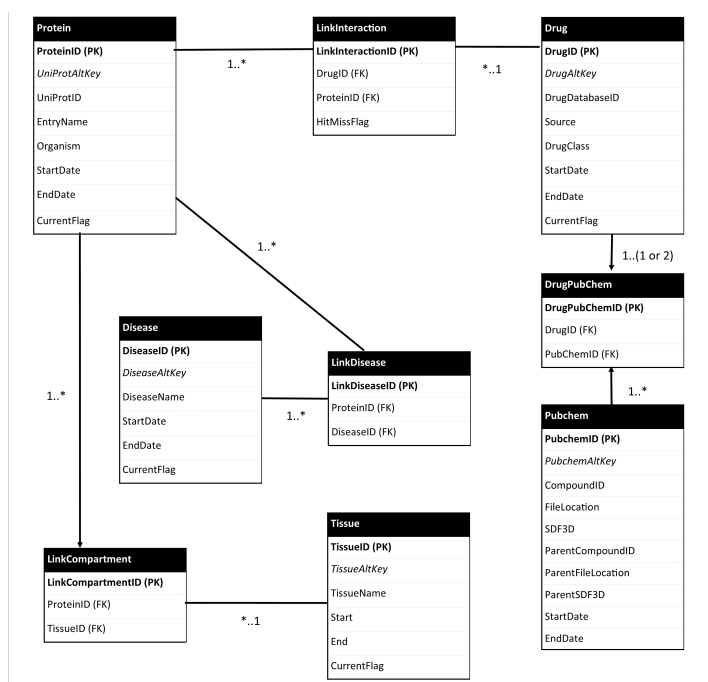
Figure 2.20 illustrates the relationships that needed to be created within the database in order to satisfy the above requirements. Drugs from the database sources needed the ability to link to protein targets, as well as the ability to link to PubChem compound IDs in order to properly organise the databases and remove potential duplicates from experiments, thus enhancing the integrity of the database. Proteins would need to have information on their amino acid sequences, and links to potential diseases and associated tissue groups.



**Figure 2.20:** Diagram demonstrating a simplified conceptual overview of the database and the relationships required to be established

To reduce the amount of duplicated data that would exist from creating these links, the database design focused on link tables which would connect various elements with one another through IDs. The Entity Relationship Diagram that incorporated these link tables is shown in Figure 2.21, where link tables make use of the primary

keys of the detailed drug, protein, tissue and disease tables to reduce duplication. The design also incorporated a historical entry system; in the event that a drug or protein had changed in terms of database identifiers, the database could keep a historical record of changes without resulting in a large number of changes to the database. Alternate keys would then ensure that a user could gather all entries from an individual record from the inception of the database, while the primary keys in conjunction with the start and end dates can provide a snapshot of a record over a certain period of time.



**Figure 2.21:** Entity Relationship Diagram of the first revision of the database, where ToxCast and DrugBank were the initial main database sources. This design allowed users to quickly discover areas which were shared by both DrugBank and ToxCast via cross-referencing to PubChem. PK refers to Primary Key, FK refers to Foreign Key

### 2.3.1 Implementing the initial design

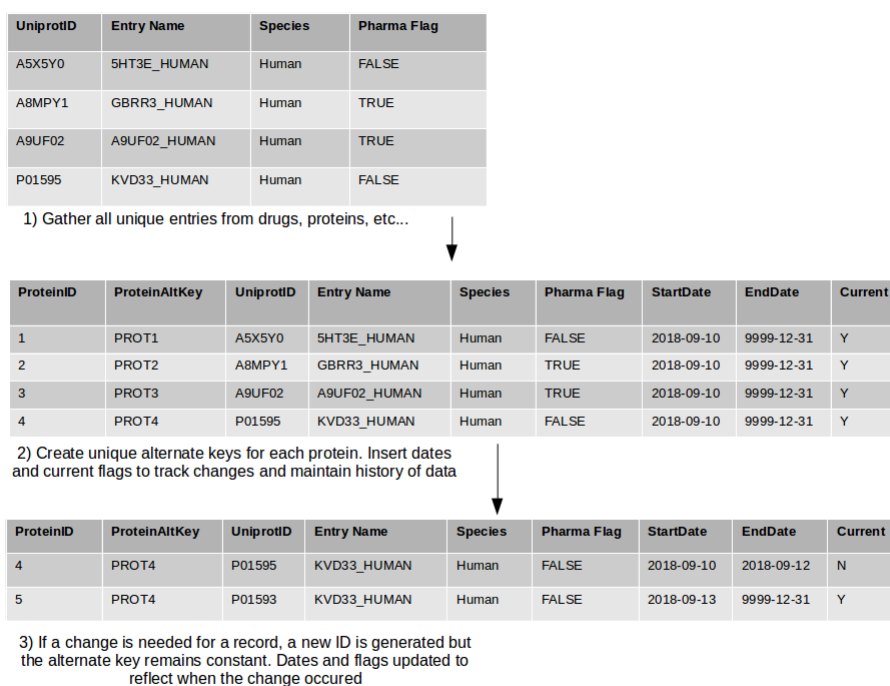
To incorporate the design, the interactions, compounds and proteins gathered from the repository analysis needed to be transformed to populate the tables. This involved a multiple step process using R scripts to compile the necessary database tables.

#### 2.3.1.1 Step 1: Creating the Detailed Record Tables

For the first step, it was necessary to transform all found drugs and proteins into individual unique records. Once all unique entries had been found, an alternate



key was generated for each new record which in conjunction with a date field would allow the database to keep track of any changes that were needed on the records, such as a change of an entry name. These were generated with a combination of an auto-counter and a string of characters, normally associated with the table itself (i.e the alternate key for a protein would be constructed as PROT1, PROT2, and so on for each new unique record). The table's primary key would continually increment and be used in the link tables so the user would have the ability to view interactions from a certain point in time in the databases' history. Figure 2.22 provides a graphical summary of this process.

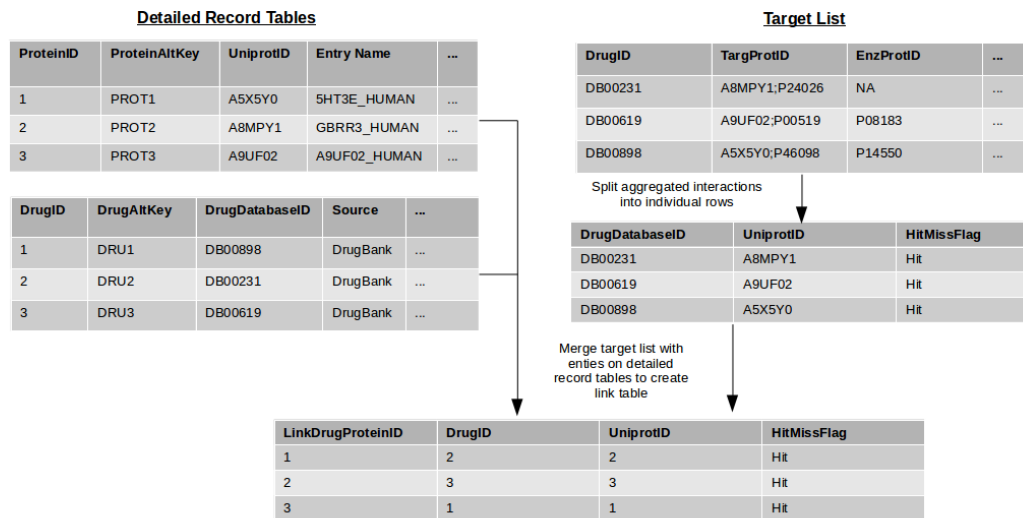


**Figure 2.22:** Graphical example of creating a detailed record table from UniProt Accession codes

### 2.3.1.2 Step 2: Creating the Link Tables

Once the detailed record tables had been compiled, it was then possible to join areas of the database with one another via link tables, which made reference to the primary keys generated in the detailed record tables. This operation reduced the amount of duplicated data needed within the database as SQL queries could be used to retrieve the duplicated information. This mainly involved merging the already compiled interaction, disease and PubChem link tables with the detailed record tables, and

then filtering the sets so that only the reference keys remained with the necessary additional table information (for example, the presence or absence of a hit or miss relating to a potential interaction in the link table between drugs and proteins). With the use of SQL INNER JOIN statements, specific information on the detailed record tables could then be referenced when a query is made on a link table. Figure 2.23 provides an illustrated example of creating link tables, while Figure 2.24 provides an example of a type of SQL query performed on the initial design to access specific details of drugs and proteins from the population of listed interactions.



**Figure 2.23:** Graphical example of creating a link table between drugs and protein targets

```
SELECT d.DrugDatabaseID, p.UniprotID, ldp.HitMissFlag
FROM linkdrugprotein ldp
INNER JOIN drug d on d.DrugID=ldp.DrugID
INNER JOIN protein p on p.ProteinID=ldp.ProteinID
WHERE ldp.HitMissFlag="Hit"
```

**Figure 2.24:** Example SQL Query to obtain all targets from the database

### 2.3.1.3 Step 3: Transfer to MySQL

After the compilation of the link and detailed record tables, a MySQL database was created using the XAMPP platform to populate the database with tables. Initially,

the files were imported via PhpMyAdmin which would create the tables and headers based on the information found in the files imported, however the platform was not suitable for some of the larger detailed record tables. This process was then adjusted to creating empty tables with the necessary fields, and then using MySQL's command line interface to populate the tables from an external CSV file.

### 2.3.1.4 Limitations

While the initial design was suited for documenting the targets found within DrugBank and ToxCast, it was found that conflicts existed between some repositories. As an example, the database contained one inactive pairing between ToxCast Drug C99661 (Valproic Acid) and UniProt Accession code O00570 (Transcription factor SOX-1). When cross-referenced with the PubChem database to Compound ID 3121, it was found that this pairing was also listed as a target-drug interaction in the database within DrugBank, under DrugBank reference DB00313. Table 2.8 shows some of the 51 conflicts that were found between listed DrugBank targets and ToxCast inactive pairs. With the initial design of Table 2.21, it would have been difficult to determine whether a pairing conflict was ultimately validated as a target or inactive, or to highlight if an issue existed with a particular pairing. With the planned incorporation of additional repositories within the database, the design needed to be altered.

ToxCast Reference	DrugBank Reference	Name	PubChem Compound ID	Uniprot ID	Uniprot Entry Name
C99661	DB00313	Valproic Acid	3121	O00570	SOX1_HUMAN
C58559	DB00277	Theophylline	2153	O76074	PDE5A_HUMAN
C797637	DB00367	Levonorgestrel	13109	P11511	CP19A_HUMAN
C298464	DB00564	Carbamazepine	2554	P33261	CP2CJ_HUMAN
C73314	DB01065	Melatonin	896	Q92753	RORB_HUMAN

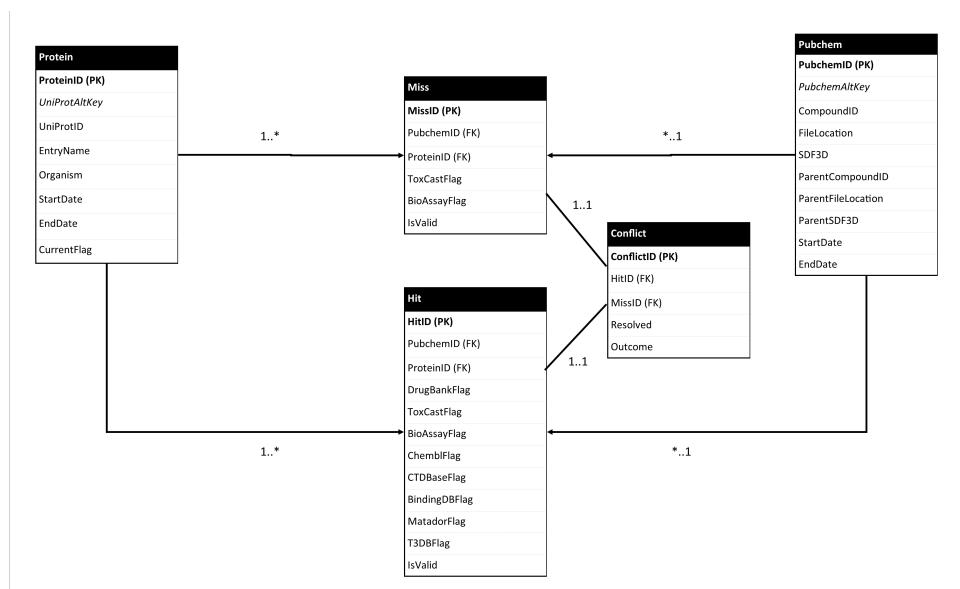
**Table 2.8:** 5 target and inactive pair conflicts found within DrugBank and ToxCast

Another issue these conflicts raised was the potential for an entry to exist within multiple repositories, causing a potential conflict in training a profiling tool where a protein-drug pairing existed to inform both a "Good" and "Bad" profile if the repository's database reference was solely used as a means of identification. The database in its initial design would also have separated protein interactions with the same compound if the database sources were different, resulting in potentially incomplete compound-protein interaction profiles.

### 2.3.2 Implementing the revised design

To address the concerns raised in the limitations of the initial design, a revision was necessary to some of the tables within the database. Figure 2.25 displays the impact of the revised design on the interaction table, with the key changes documented below:

- Targets and inactive compound-protein pairs were listed separately, as not all databases allowed the documentation of inactive sites and would have introduced a degree of confusion and irrelevant data for non-interactive bindings.
- A conflict table to specify whether a conflict has since been identified and confirmed as a target or non-interactive pair, either through a literature/database update or through other means
- Validity flags for targets and inactive compound-protein pairs, to isolate potential conflicts from the main data set
- Flags to indicate in which repository an interaction pair was documented. This provided the ability to assess which repositories caused the most conflicts with one another



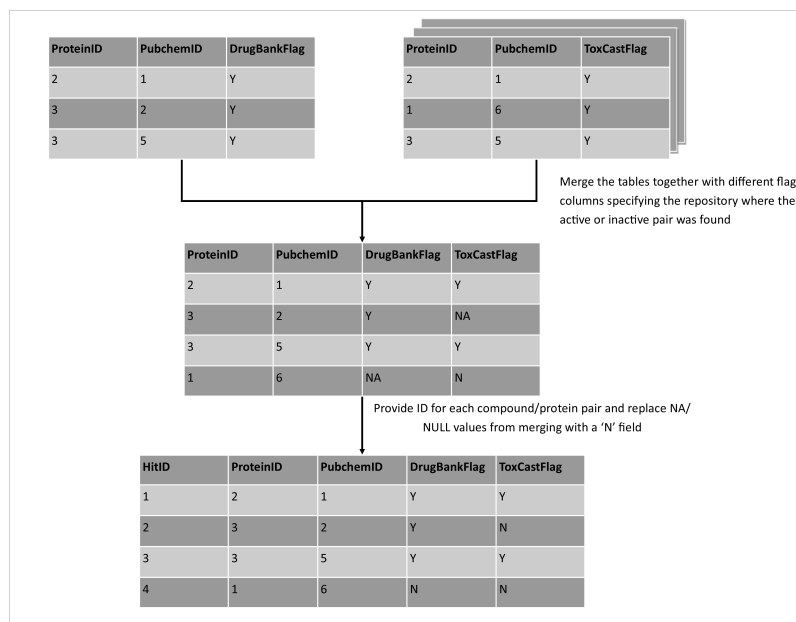
**Figure 2.25:** Entity relationship diagram of changes taken to document interactions. This replaced the LinkInteraction element found in Figure 2.21, and interactions were linked directly via the PubChem table instead of via a repository's internal database reference.

These changes provided additional flexibility to the database structure, allowing users not only the ability to filter interactions and inactive sites to specific databases, but also to investigate, isolate and potentially resolve conflicts that exist between repositories. The focus on PubChem compounds as a centralized source also allowed the combination of protein interaction profiles for a single compound, where certain proteins were considered in one database but not in another.

### **2.3.2.1 Creation of Hit, Miss and Conflict Tables**

For the changes to be implemented, a number of modifications were needed to the process of compiling interaction link tables, in particular for the active and inactive compound-protein pairs which now needed to specify in which database the instance of activity or inactivity was documented. To accomplish this, interaction files for each repository were compiled and given an individual 'Yes' flag for that particular repository. When all the repository interactions were merged with the reference keys generated in the detailed record tables, a list would be generated with the 'Yes' flags combined in an individual protein-compound interaction, and 'NA' values for the remainder of fields where a flag was not found. These were easily replaced with a 'No' flag to clearly distinguish that the repository had not documented the interaction in question. Finally, a validity flag was inserted to specify whether or not an instance of activity or inactivity has been determined to be valid, removing its potential erroneous impact on queries while still preserving the information as to which repository the information was gathered from within the database. Figure 2.26 displays a graphical representation of compiling a segment of the active interaction table as an example.

In order to document the conflicts, both tables were searched for matching protein and compound reference key pairs. Once these had been found, a new table was compiled of specified pairs which had been found within the hit and miss tables, along with additional fields to specify if the conflict had been resolved, and which outcome had been decided. Until this conclusion had been reached or a solution had been found to these conflicts, all pairs found within this conflict table were to be considered as invalid results, and not used as part of a training set for the drug profiling experiments within Chapter 3.



**Figure 2.26:** Graphical example of creating the database's active interaction table

### 2.3.2.2 Analysis of Interactions and Conflicts

On compilation of the revised database design, an analysis was conducted on the amount of information which had been compiled from each repository. Table 2.9 provides a summary of all active and inactive results that had been recorded from the repositories that were considered in the project, based on the parameters of searching for the human proteins which were present within DrugBank or ToxCast. The overall entry in this table provides a summary of the numbers of all unique records within the database, of which ChEMBL and PubChem BioAssay provide the majority of active and inactive results.

Repository	Number of Active Results	Number of Inactive Results	Number of Compounds	Number of Proteins
DrugBank	13,658	N/A	4,276	2,070
ChEMBL	1,340,634	N/A	357,298	1,239
CTDBase	204,945	N/A	9,383	2,257
BindingDB	31,067	N/A	23,739	234
Matador	8,182	N/A	781	1,089
T3DB	29,476	N/A	3,117	1,353
ToxCast	69,754	315,612	8,719	276
PubChem BioAssay	344,229	24,956,727	360,602	1,087
<i>Overall</i>	<i>1,868,271</i>	<i>25,220,443</i>	<i>373,661</i>	<i>2,305</i>

**Table 2.9:** Summary of all active and inactive results found in the repositories considered based on the DrugBank and ToxCast protein panels

Table 2.10 provides a summary of where conflicts had resided, of which a large proportion of conflicts had been drawn from ChEMBL, BioAssay, and BindingDB.

These conflicts were further assessed to determine how many repositories with active interactions had been flagged as "Yes" per conflict. As shown in Table 2.11, it was shown that 90% of the conflicts found only had one repository flagged as "Yes" within the Hit Table. These one hit flag only conflicts were further investigated to determine which miss repository generated the conflict, and as shown in Table 2.12 it was found that the main source of these conflicts were caused either by ChEMBL active pairs against BioAssay-listed inactive pairs, or from BioAssay listing both an active and inactive pairing.

Repository	Actives Flagged as Conflicts	Actives Flagged as Conflicts (%)	Inactives Flagged as Conflicts	Inactives Flagged as Conflicts (%)
DrugBank	532	3.9	N/A	N/A
ChEMBL	319,964	23.87	N/A	N/A
CTDBase	12,088	5.9	N/A	N/A
BindingDB	14,795	47.62	N/A	N/A
Matador	306	3.74	N/A	N/A
T3DB	2,054	6.97	N/A	N/A
ToxCast	10,516	15.08	15,860	5.03
PubChem BioAssay	180,835	52.53	471,811	1.89
<i>Overall</i>	<i>482,588</i>	<i>25.83</i>	<i>482,588</i>	<i>1.91</i>

**Table 2.10:** Summary of all conflicts for the DrugBank and ToxCast protein panels detected in the repositories considered

Number of Repositories flagged as "Yes" in Conflict	Number of Conflicts	Proportion of Conflicts Found (%)
1	435,165	90.17
2	37,461	7.76
3	9,174	1.9
4	518	0.1
5 to 7	270	0.06

**Table 2.11:** Distribution of the number of repositories which flagged a discovered protein-compound conflict as active when an inactive pairing is indicated within another repository

Hit Repository	Miss Repository	Number of Conflicts Found
ChEMBL	PubChem BioAssay	271,256
PubChem BioAssay	PubChem BioAssay	138,494
ToxCast	PubChem BioAssay	5,818
ChEMBL	ToxCast	5,009
CTDBase	PubChem BioAssay	4,024
CTDBase	ToxCast	3,796
BindingDB	PubChem BioAssay	1,035
T3DB	ToxCast	720
PubChem BioAssay	ToxCast	318
ToxCast	ToxCast	94
T3DB	PubChem BioAssay	62
Matador	PubChem BioAssay	38
DrugBank	PubChem BioAssay	28
Matador	ToxCast	14
DrugBank	ToxCast	4
BindingDB	ToxCast	0

**Table 2.12:** Table investigating conflicts which have been flagged as active by one repository and inactive by one repository

### 2.3.2.3 Resolving Conflicts

On further investigation of these conflicts, three issues had been identified. The first issue with regard to BioAssay internal conflicts is that multiple assays were on occasion present within the repository which list differing results for a specific compound-protein pairing. In order to resolve these conflicts further investigation would be necessary into the testing parameters and types of assay performed. A solution to resolve these conflicts would likely involve parsing through the XML assay files which are available on PubChem's FTP server.

With regard to the second issue of ChEMBL conflicts, on further analysis of the database schema it had been found that activities listed within this repository contained comments which specified the type of binding, of which some activities listed contained inactive records that had not been documented correctly. There had also been instances of inconclusive interactions which had been incorrectly listed as targets. Figure 2.27 provides an example of the modified query to obtain these comments, of which 1,419 differing types of comments had been listed for the 2,305 protein codes queried. As there was little consistency to the formatting of the activity comments listed within the repository, a collection of activity types was used to filter the ChEMBL activity set, which documented the majority of interactions that had been found.

Table 2.13 provides a summary of the comment search terms that had been used to gather the majority of the active and inactive compound-protein pairings in the query results, however further work would be needed to document all interactions present and to filter comments considered to be unclear. On obtaining the activities



from ChEMBL, it was found that instead of 1.3 million active protein/compound instances, there were only 195,350 instances that were considered as active, while 398,879 instances were considered inactive. Approximately 721,000 instances were considered as inconclusive and required removal from the database.

```
SELECT DISTINCT mol.molregno,mol.chembl_id,comp.accession,
act.activity_comment
FROM component_sequences as comp, target_components as tc,
compound_records cr, source src, assays as assay,
activities as act, molecule_dictionary as mol
WHERE comp.accession IN ("A5X5Y0","A8MPY1","A9UF02","000141")
AND tc.component_id = comp.component_id AND assay.tid = tc.tid
AND act.assay_id = assay.assay_id AND mol.molregno = act.molregno
AND cr.molregno = act.molregno AND src.src_id = cr.src_id
AND src.src_id = '7' ORDER BY comp.accession;
```

**Figure 2.27:** Modification to the SQL query to obtain activity comments from ChEMBL

Active Search Terms	Inactive Search Terms
active	inactive
inhibitor	not active
substrate	no inhibition
antagonist	

**Table 2.13:** Summary of search terms used to gather active and inactive protein-compound pairs from ChEMBL

The third issue found was the high proportion of BindingDB conflicts. The cause of these conflicts was due to the source of the interactions found in the file downloaded from BindingDB. On the downloads section of BindingDB, users could download either the entirety of the database or segments where the source of interaction was from differing repositories. In the initial analysis, only the interactions from PubChem had been retrieved of which a high proportion of conflicts are present. To correct this issue, the entire BindingDB interaction file was downloaded, which not only increased the number of interactions from the repository, but also the number of compounds.

Once these changes had been made to the ways of extracting information from BindingDB and ChEMBL, the total figures for actives, inactives, and conflicts were revised, found in Table 2.14 and Table 2.15 respectively. This also included changes to the search space within BioAssay to include all compounds found within the proteins queried. While the number of active records had been reduced from the corrections to ChEMBL, the amount of conflicts that were present within the database was also reduced as can be seen on Table 2.16 and Table 2.17. In the revised summary, the

number of conflicts between PubChem BioAssay and ChEMBL had been reduced from 271,256 conflicts (the largest amount of conflicts before the revision) to 432 conflicts. The largest number of conflicts now resided entirely on interactions only documented by the PubChem BioAssay platform, with 154,259 conflicts detected.

Although the revised design had reduced the amount of conflicts that had been detected in the initial revision, a prior consideration of assay types from the various sources could have further reduced the amount of conflicts present, such as through the removal of assays which would be incompatible by the majority of other repository assay types.

Repository	Number of Active Results	Number of Inactive Results	Number of Compounds	Number of Proteins
DrugBank	13,658	N/A	4,276	2,070
ChEMBL	193,647	395,314	235,316	864
CTDBase	204,945	N/A	9,383	2,257
BindingDB	670,933	N/A	406,927	1,213
Matador	8,182	N/A	781	1,089
T3DB	29,476	N/A	3,117	1,353
ToxCast	69,754	315,612	8,719	276
PubChem BioAssay	731,059	30,212,401	953,481	1,261
<i>Overall</i>	<i>1,509,574</i>	<i>30,665,860</i>	<i>1,157,529</i>	<i>2,305</i>

**Table 2.14:** Summary of all active and inactive results found in the repositories considered based on the DrugBank and ToxCast protein panels, after revisions

Repository	Actives Flagged as Conflicts	Actives Flagged as Conflicts (%)	Inactives Flagged as Conflicts	Inactives Flagged as Conflicts (%)
DrugBank	1,062	7.78	N/A	N/A
ChEMBL	30,749	15.88	30,881	7.81
CTDBase	14,818	7.23	N/A	N/A
BindingDB	27,749	4.14	N/A	N/A
Matador	580	7.09	N/A	N/A
T3DB	2,362	8.01	N/A	N/A
ToxCast	11,122	15.94	9,253	2.93
PubChem BioAssay	200,705	27.45	220,913	0.73
<i>Overall</i>	<i>236,628</i>	<i>15.68</i>	<i>236,628</i>	<i>0.77</i>

**Table 2.15:** Summary of all conflicts for the DrugBank and ToxCast protein panels detected in the repositories considered, after revisions

Number of Repositories flagged as "Yes" in Conflict	Number of Conflicts	Proportion of Conflicts Found (%)
1	193,357	81.71
2	35,661	15.05
3	6,648	2.81
4	622	0.26
5 to 7	390	0.16

**Table 2.16:** Distribution of the number of repositories which flagged a discovered protein-compound conflict as active when an inactive pairing is indicated within another repository, after revisions

Hit Repository	Miss Repository	Number of Conflicts Found
PubChem BioAssay	PubChem BioAssay	154,259
ToxCast	PubChem BioAssay	6,441
CTDBase	PubChem BioAssay	4,187
CTDBase	ToxCast	3,964
BindingDB	ChEMBL	3,591
BindingDB	PubChem BioAssay	3,181
CTDBase	ChEMBL	2,061
ChEMBL	ChEMBL	1,637
T3DB	ToxCast	731
PubChem BioAssay	ToxCast	444
ChEMBL	PubChem BioAssay	353
ToxCast	ChEMBL	259
DrugBank	ChEMBL	165
ChEMBL	ToxCast	113
ToxCast	ToxCast	108
Matador	ChEMBL	83
PubChem BioAssay	ChEMBL	79
T3DB	PubChem BioAssay	59
Matador	PubChem BioAssay	40
DrugBank	PubChem BioAssay	33
T3DB	ChEMBL	17
Matador	ToxCast	14
BindingDB	ToxCast	12
DrugBank	ToxCast	4

**Table 2.17:** Revised numbers of conflicts which have been flagged as active by one repository and active by one other repository following consideration of activity comments from ChEMBL

### 2.3.2.4 Limitations

One of the limitations noted by the DrugReferenceDatabase is that in some instances PubChem records retrieved via the PUGREST platform returned results which had at some point been specified as a non-live record. A non-live record in this instance referred to a compound which had either not yet been released publicly or had been declared later as obsolete or found to be incorrect. The presence of these non-live records had been highlighted when attempting to process parent compounds via PubChem's identifier exchange service, where certain compound ID entries had not returned a result. Figure 2.28 displays an example of a non-live record which would only be directly retrievable by entering the ID directly into the URL. Of the 373,692 PubChem entries documented within the database, 62

compound entries had been identified as potential non-live records. As Figure 2.29 shows, ChEMBL had been the primary source of non-live records being introduced into the DrugReferenceDatabase, with only a small proportion being introduced from querying BioAssay, ToxCast and DrugBank.

Another limitation of the database is that the interactions documented do not specify the conditions of the assay that led to the interaction being classified, which in turn led to the large number of conflicts detected. Further consideration of the types of assay that had been performed would have reduced the amount of conflicts, as well as provide the means of filtering for specific assays of interest whilst compiling a list of interactions.

2-Hydroxy-3-[(2-hydroxy-4-oxochromen-3-yl)methyl]chromen-4-one

PubChem CID: 653

Molecular Formula:  $C_{19}H_{12}O_6$

InChI Key: KSKRYQVHJQRUNC-UHFFFAOYSA-N

NOTE: NON-LIVE RECORD. See the related substances for more information.

PUBCHEM > COMPOUND > 2-HYDROXY-3-[(2-HYDROXY-4-OXOCHROMEN-3-YL)METHYL]C...

Create Date: 2005-03-25

Contents

- 1 2D Structure
- 2 Names and Identifiers
  - 2.1 Computed Descriptors
    - 2.1.1 IUPAC Name
    - 2.1.2 InChI
    - 2.1.3 InChI Key
    - 2.1.4 Canonical SMILES
    - 2.1.5 Isomeric SMILES
  - 2.2 Molecular Formula
  - 2.3 Status

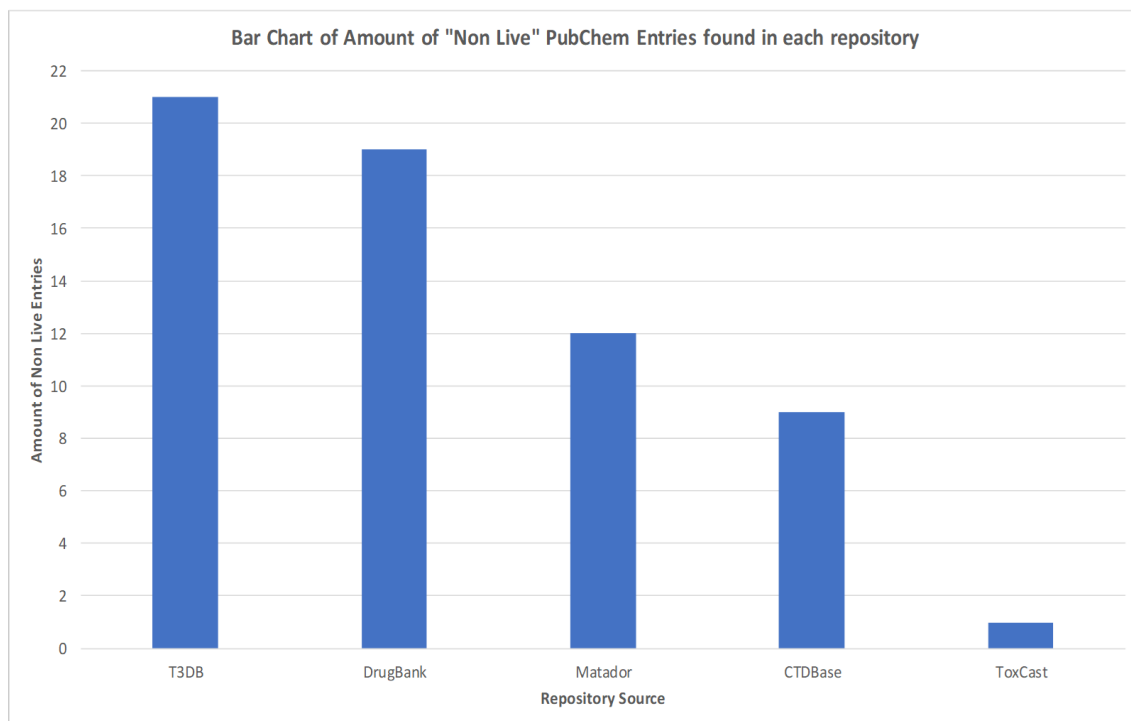
1 2D Structure

Search Download Get Image

Magnify

**Figure 2.28:** Entry for a non-live PubChem record CID 653, associated with the Matador Database under the chemical name DTC

As the PUGREST API and ChEMBL platforms have the capacity to return non-live records, these entries have remained in the database but have been documented for potential flagging as invalid hits and misses until a solution has been determined through an update by the PubChem team or through further filtering of the API's results to determine whether a result is a correct record before integrating further species or all human protein accession codes. Further work must also be undertaken to determine if these non-live entries already exist within the database under a different live PubChem Compound ID, to investigate the potential for merging cross-references.



**Figure 2.29:** Bar chart demonstrating number of non-live PubChem Compound IDs introduced from data obtained from extracting the compound repositories

## 2.4 Discussion

On implementation of the revised design, the database contained interactions between compounds and human proteins from a significant number of repositories detailing a vast collection of functional data. Over 1.1 million PubChem entries are stored with interactions of the 2,305 proteins contained within DrugBank and ToxCast, where over 1.5 million instances of interactions and over 30 million instances of non-interactions are stored. The database also contains references to repositories that use their own identifiers, of which over 250,000 cross-references are present. For ease of installation, the complete database has been exported into an individual SQL file which can be obtained from the project appendix.

Despite the vast amount of information that is currently contained within this database, the information captured and gathered by the repositories considered only represents a very small scale of the protein-drug search space, and further work is necessary in order to refine issues encountered during the data gathering phase. For example, to document the protein interaction profiles of all compounds found with the 2,305 proteins would require approximately 2 billion instances of active or inactive pairings to be stored. This would be exponentially increased if the database was to

consider the whole human genome (proteome) which includes over 20,000 UniProt accession codes. The information captured by the database is also subject to further scrutiny and information retrieval, as the testing methods of each repository could deliver differing results and in turn generate conflicts.

### **2.4.1 Further Work**

From the numbers of conflicts which have been detected between the repositories, in particular with BioAssay itself, it is clear that further information beyond records of "bare" interaction is needed. A possible approach would be to consider the experimental parameters for tests between compounds and proteins to further assess which tests would be the most appropriate to use as a classification of activity. However, as each repository has different experimental parameters and methods of storing them, a further revision of the database's design would be needed to store these parameters correctly and efficiently for comparison and analysis purposes. A means to accomplish this would have been to assess the type of assay, such as in the case of ToxCast as described in Chapter 1 where assays were categorised into biochemical and cell-based assays.

Another area for further consideration is the planned changes for PubChem which were ongoing at the time of writing. One of these changes to the platform which was under testing in beta at the time of writing included a bulk search function, which allowed users to pass either compound, substance or BioAssay entry fields via a CSV file to generate a downloadable list of information within PubChem. This bulk search functionality has the potential to remove the need to use the Identifier Exchange Service or the PUGREST API for both small and large scale queries; however use of the bulk search platform raised similar issues to the API and Exchange service in large scale queries, where the website would time out and cancel a query attempt if result compilation exceeded a timeout value assigned by the PubChem website. Another change announced by the PubChem team was the retirement and eventual decommissioning of the PubChem BioAssay Tools platform from November 1, 2018 [17]. This would likely cause an impact on the current method of web based scraping of BioAssay interactions based on UniProt accession codes, however XML file scraping would likely need to be applied in further data gathering attempts going forward to obtain the experimental parameters of the assays in bulk from the BioAssay platform.

## 2.5 References

- [1] Wishart, D. S., Knox, C., Guo, A. C., Sharavista, S., Hassanali, M., Stothard, P., Chang, Z., and Woolsey, J. (2016). “Drugbank: Documentation and sources,” [Online]. Available: <http://www.drugbank.ca/documentation> (visited on 05/06/2016).
- [2] United States Environment Protection Agency. (2018). “Toxicity ForeCaster (ToxCast) Data — Safer Chemicals Research — US EPA,” [Online]. Available: <https://www.epa.gov/chemical-research/toxicity-forecaster-toxcas-ttm-data> (visited on 08/09/2018).
- [3] United States Environment Protection Agency. (2015). “US EPA TOXCAST DATA RELEASE OCTOBER 2015 - Summary Files,” [Online]. Available: [ftp://newftp.epa.gov/Computational\\_Toxicology\\_Data/High\\_Throughput\\_Screening\\_Data/Previous\\_Data/ToxCast\\_Data\\_Release\\_Oct\\_2015/Summary\\_Files/README\\_INVITRODB\\_V2\\_SUMMARY.pdf](ftp://newftp.epa.gov/Computational_Toxicology_Data/High_Throughput_Screening_Data/Previous_Data/ToxCast_Data_Release_Oct_2015/Summary_Files/README_INVITRODB_V2_SUMMARY.pdf) (visited on 12/04/2018).
- [4] Phuong, J., Truong, L., Sipes, N., Connors, K., Houck, K., Judson, R., and Martin, M. (2015). “ToxCast Assay Annotation Version 1.0 Data User Guide,” [Online]. Available: [https://www.epa.gov/sites/production/files/2015-08/documents/toxcast\\_assay\\_annotation\\_data\\_users\\_guide\\_20141021.pdf](https://www.epa.gov/sites/production/files/2015-08/documents/toxcast_assay_annotation_data_users_guide_20141021.pdf) (visited on 12/04/2018).
- [5] Kim, S., Thiessen, P. A., Bolton, E. E., and Bryant, S. H., “PUG-SOAP and PUG-REST: Web services for programmatic access to chemical information in PubChem,” *Nucleic Acids Research*, vol. 43, no. W1, W605–W611, Jul. 2015.
- [6] National Center for Biotechnology Information. (2018). “PubChem Identifier Exchange Service,” [Online]. Available: <https://pubchemdocs.ncbi.nlm.nih.gov/identifier-exchange-service> (visited on 12/04/2018).
- [7] National Center for Biotechnology Information. (2018). “Downloading PubChem Data,” [Online]. Available: <https://pubchemdocs.ncbi.nlm.nih.gov/downloads> (visited on 12/04/2018).
- [8] Wishart, D., Arndt, D., Pon, A., Sajed, T., Guo, A. C., Djoumbou, Y., Knox, C., Wilson, M., Liang, Y., Grant, J., *et al.* (2018). “T3DB: Downloads,” [Online]. Available: <http://www.t3db.ca/downloads> (visited on 12/12/2018).

- [9] Günther, S., Kuhn, M., Dunkel, M., Campillos, M., Senger, C., Petsalaki, E., Ahmed, J., Urdiales, E. G., Gewiess, A., Jensen, L. J., *et al.* (2018). “MATADOR,” [Online]. Available: <http://matador.embl.de/> (visited on 12/12/2018).
- [10] Gaulton, A., Hersey, A., Nowotka, M., Bento, A. P., Chambers, J., Mendez, D., Motowolo, P., Atkinson, F., Bellis, L. J., Cibrián-Uhalte, E., *et al.* (2017). “ChEMBLdb Releases,” [Online]. Available: [ftp://ftp.ebi.ac.uk/pub/databases/chembl/ChEMBLdb/releases/chembl\\_23/](ftp://ftp.ebi.ac.uk/pub/databases/chembl/ChEMBLdb/releases/chembl_23/) (visited on 12/12/2018).
- [11] The Uniprot Consortium. (2018). “UniProtKB,” [Online]. Available: <https://www.uniprot.org/uniprot/> (visited on 12/12/2018).
- [12] The Uniprot Consortium. (2018). “Androgen receptor,” [Online]. Available: <https://www.uniprot.org/uniprot/P10275.txt> (visited on 12/12/2018).
- [13] Davis, A. P., Murphy, C. G., Saraceni-Richards, C. A., Rosenstein, M. C., Wiegers, T. C., and Mattingly, C. J. (2018). “Data Downloads — CTD,” [Online]. Available: <http://ctdbase.org/downloads/> (visited on 12/12/2018).
- [14] Davis, A. P., Murphy, C. G., Saraceni-Richards, C. A., Rosenstein, M. C., Wiegers, T. C., and Mattingly, C. J. (2018). “Batch Query — CTD,” [Online]. Available: <http://ctdbase.org/tools/batchQuery.go> (visited on 12/12/2018).
- [15] The Uniprot Consortium. (2018). “Retrieve/ID mapping,” [Online]. Available: <https://www.uniprot.org/uploadlists/> (visited on 08/10/2018).
- [16] National Center for Biotechnology Information. (2018). “BioAssay Download FTP,” [Online]. Available: <ftp://ftp.ncbi.nlm.nih.gov/pubchem/Bioassay/> (visited on 12/12/2018).
- [17] National Center for Biotechnology Information. (Apr. 27, 2018). “PubChem BioAssay Tools to be replaced — PubChem Blog,” [Online]. Available: <https://pubchemblog.ncbi.nlm.nih.gov/2018/04/27/pubchem-bioassay-tools-to-be-replaced/> (visited on 07/29/2019).



## Chapter 3

# Using machine learning approaches for profiling drug-protein interactions

### 3.1 Introduction

On conclusion of the implementation of the DrugReferenceDatabase, it was possible to quickly extract via database queries the customised drug-protein interaction panels necessary to conduct investigations of computational drug profiling and interaction prediction. This chapter will focus on the construction of classification models to predict the profile of candidate compounds, as to whether or not their properties and their interaction with proteins (not considering dosage) would be sufficient to label the compound as a "Good" (i.e FDA-like) or a "Bad" (i.e ToxCast associated) profile. In order to accomplish this goal, the chapter provides an overview of recent similar studies within the field of drug discovery which made use of machine learning tools and conduct a review of the tools and classification algorithms that are intended to be used for constructing this classification model. The chapter then provides a description and discussion of the testing procedures used in the process of training and testing these profiling models, followed by a summary of findings and areas for further work.

#### 3.1.1 Background

Machine learning tools have been used in a number of studies in the drug discovery process in a variety of approaches. They have also been increasingly adopted by industries within the drug development process; in a review of artificial intelligence processes by Mak *et al.*, a number of partnerships have been formed between AI

companies and pharmaceutical companies [1]. Some examples of pharmaceutical industry involvement highlighted within the review include work by Bayer in the process of predicting treatment response to certain cancer types, and AstraZeneca in evaluating targets against neurodegenerative diseases. While the review found that no developed drugs had yet made use of AI approaches, it highlighted that the possibility was increasingly likely within 2 to 3 years.

Machine learning methods have also become increasingly portable with improvements in processing power. In a review of machine learning methods in pharmaceutical research by Ekins, the review had found some models such as drug-drug interaction prediction and target prediction could be implemented on mobile phone platforms [2]. Mobile phone platforms have also been used in increasing frequency in crowd-sourcing, with projects on the BOINC platform providing the means for users to contribute various types of hardware to large scale computing problems, including mobile phones via an app [3] [4]. While these approaches described in this section explore different questions to the core question being addressed in this work, their discussion highlights the depth of information that can be applied to a classification problem.

In terms of recent academic studies in drug discovery and machine learning, one good example is a study conducted by Scheeder *et al.* which assessed the use of machine learning tools through using small molecule drug images [5]. This study reported that there were a number of fields of interest through the use of images, from method of action and toxicity prediction to drug similarity clustering, however such techniques required further efforts from research groups in sharing of datasets to verify results. A similar study by Jimenez-Carretero *et al.* also proposed and applied an automated means of using microscopy images for the purposes of toxicity screening [6]. They found that the models trained with this method had predicted toxicity of drugs with a different mechanism of molecular action than that present in the training set, which was later found to be confirmed through cell assays. Another toxicity prediction platform approach reported similar promising results of accuracy, with the eToxPred platform being able to identify as much as 72% of toxic compounds tested [7].

Another example but in a narrower field of application by Esteva *et al.* made similar use of image classification to determine differing levels of skin cancer progression [8]. Within this study they had found that the developed model had performed to an equivalent level to a dermatologist in classifying cancer types, and although they had acknowledged that a dermatologist would use additional means beyond visual identification of skin, further work to complement the classifier within a professional

setting could provide benefits.

Other studies made use of machine learning in drug discovery for a different application through identifying organ targets through the assessment of *in vitro* pharmacological profiles [9]. In the report by Remez *et al.*, an approach was developed to create a profile of the impact of drugs on organs within the human body (split into 47 individual entities) as an alternative to conducting *in vivo* studies. They found that in some cases organ toxicity was related to a drug's primary targets, however in most cases additional information was required to identify potential organ hazards, in which *in silico* target predictions assisted in complementing existing *in vitro* interaction profiles. Another example which makes use of panels for predictions include a technique named CDRScan[10]; a response prediction algorithm which attempts to identify ideal candidates for cancer treatment. The study in question makes use of predictive values of compounds' activity against cancer cell lines, for which a panel of 567 genes strongly associated with cancer pathology were used in the process of predicting potential interactions. The report concluded that a high degree of accuracy had been obtained, and proposed 14 oncology and 23 non-oncology drugs with potential anti-cancer properties.

Other recent drug discovery approaches include to treatment options for specific population groups. In one study by Aljumah *et al.*, models were constructed of treatment preferences between young and old patient groups in the management of diabetes, and found that in older age groups in Saudi Arabia that drug treatment was more appropriate, while for younger age groups diet and weight control were of higher importance.[11].

Machine learning also provides a case for improving rates of diagnosis for conditions. Another study related to diabetes constructed training models in WEKA to predict the diagnosis of patients from suffering the type 2 variant of the condition [12]. That study had found the model, which was based on patient's health properties such as blood pressure and body mass index, had achieved a result of 95.2% accuracy in assessing positive and negative cases of type 2 diabetes.

Other approaches also make use of the same data sources as described in Chapter 2. One example is a report by Rodríguez-Pérez *et al.* which made use of the PubChem BioAssay platform for generating protein-compound matrices for machine learning [13]. In the study they found that in most cases machine learning methods making use of compound-protein matrices were complicated by issues of data imbalance, where most screenings which took place are classified as inactive. In addition to this, another problem which was highlighted in the study was the lack of information which

could be contributed by industry findings due to it being withheld from the public. This finding was reinforced by the data gathering exercise that was conducted for proteins in Chapter 2. The researchers found that with the BioAssay platform it was possible to construct a complete active/inactive compound-protein interaction matrix from certain assays, however as stated in the report by Vogt *et al.* discussing the method of interaction, no quality checks were performed for potential conflicts which have been highlighted in this study's attempts of extracting interactions from the BioAssay platform itself [14].

In terms of studies which make use of interaction matrices for the purposes of machine learning, some papers make reference to methods of predicting interactions through the use of drug and protein structural similarity [15] [16]. These kind of methods however are beyond the scope of this chapter, instead these individual approaches are considered in more detail in Chapter 5. The approach presented within this chapter differs from other studies in that it makes use of only *in vitro* interaction information or known chemical properties to make predictions.

## 3.2 Methodology

In order to ensure the data gathered was explored efficiently in generating a profiling model, three approaches were considered in generating a data set. The first approach involves the use of the protein panel discussed in Chapter 1 from the report by Bowes *et al.* [17]; referred to as Panel 44, this panel contained 44 targets to identify potentially hazardous compounds. As two targets within this panel refer to two protein entries, the Panel is defined by 46 UniProt accession codes. The second approach refers to another protein panel commonly referred to as Panel 331 [18], of which 144 human proteins were present. The third approach considered proteins had been listed as pharmacologically active in DrugBank, of which 668 proteins were present within the DrugReferenceDatabase. This protein panel was referred to as the Pharmacology Panel. While the database contained more information to generate a greater searchspace for profiling, the use of these protein panels was considered to be plausible for the generation of a potential profiling tool. The UniProt Accession codes attached to each of these panels are included in the Appendix.

With the generated datasets, a selection of features and attributes were then used to investigate their impact on profiling accuracy. The first approach would consider a simple activity flag matrix between compounds and proteins. The second approach expands on this matrix with the inclusion of inactive compound protein pairs and

inconclusive results. The final approach implements chemical property attributes, both as individual entities and in conjunction with the active protein compound pairs.

### 3.2.1 Software and Hardware

To ensure that a variety of classification algorithms were freely available to allow appropriate reproducibility for this study, the classification software WEKA had been used, which is a Java based application that provides a collection of different types of classifier which were free to use. While most classifiers are based on a similar principle of building a set of rules based on the attributes presented within a dataset, their methods of execution and presentation of models can differ. For example, while models generated by the J48 and REPTree algorithms would generate binary decision trees for classifying candidate drugs, rule-based classifiers such as JRip would construct models which base decisions on 1 or more attributes to reach a decision of a single instance. Table 3.1 provides a summary of the types of classifiers which were used in the classification experiments. A variety of classifier types were used to assess performance, however additional classifiers are also available from WEKA's package manager.

<b>Classifier Name</b>	<b>Classifier Type</b>
DecisionTable [19]	Ruleset
JRip [20]	Ruleset
PART [21]	Ruleset
J48 [22]	Decision Tree
RandomForest [23]	Decision Tree
RandomTree [23]	Decision Tree
REPTree [24]	Decision Tree
Logistic [25]	Functional
Naive Bayes [26]	Bayes

**Table 3.1:** Caption describing the classifiers used within WEKA for all experiments

All tests were performed with default settings defined by WEKA, of which version 3.8.3 was used. All classifiers were evaluated by 10 fold cross validation, which distributes the dataset 10 times into a 90%/10% training/test split to provide an average performance of the classification results, and so that all instances within the dataset are considered for evaluation purposes. The environment which was used to train and test the classifier models was a HP EliteBook 745 G3 laptop, which contained an AMD processor, a 256GB SSD drive and 8GB of RAM.

### 3.3 Active Pairs Only

In order to construct a dataset which was suitable for use within a classifier, it was necessary to format a list of interactions into a matrix of a similar style to the dataset within ToxCast. Figure 3.1 illustrates a graphical example of a small scale interaction matrix, where each row in the matrix comprises of a compound which had interactions present within one of the protein panels listed. Columns would comprise of proteins of interest, and individual cell values within this matrix would represent one of the following:

- 0 - Indicating that no interaction had been found between a compound and a protein
- 1 - Indicating that an interaction had been found between a compound and a protein

CompoundID	P04150	P15382	P14867	P25101	DrugClass
1234	0	0	1	0	Bad-Profile
2745	1	0	0	0	Good-Profile
3125	0	1	0	0	Good-Profile
3761	0	0	0	1	Bad-Profile

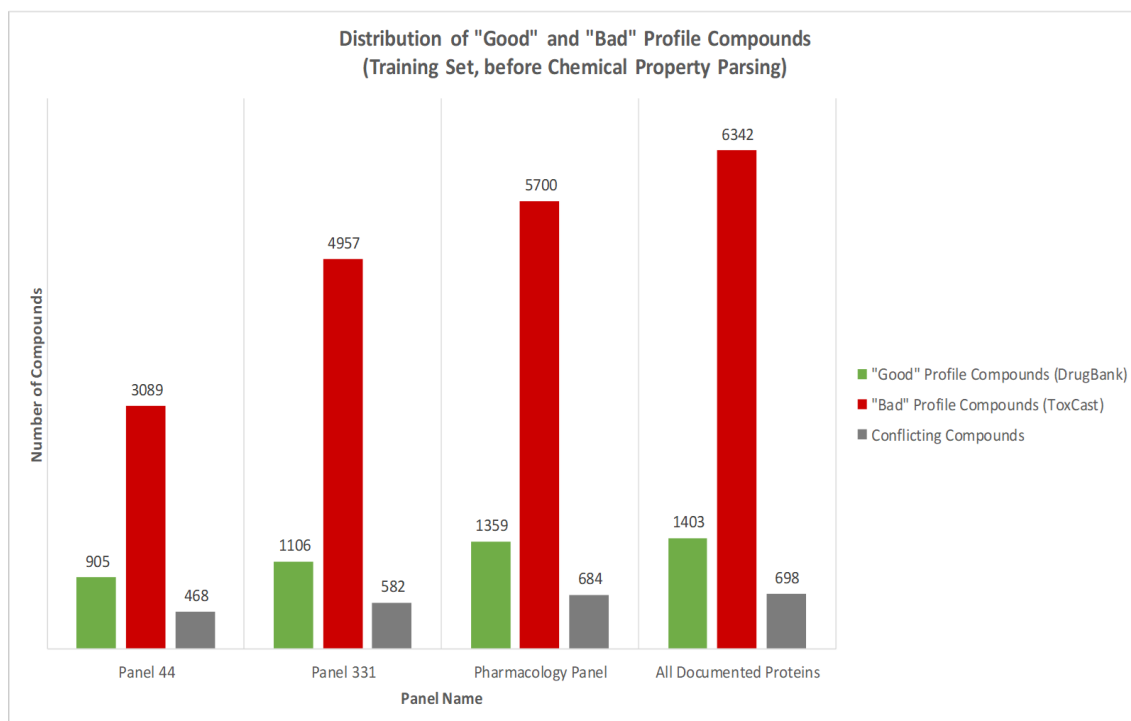
**Figure 3.1:** Illustrative example of an interaction matrix

Figure 3.3 provides an example of a truncated query which was used by the database to retrieve the interacting compounds of interest. In order to ensure that as many interactions as possible were retrieved, all repositories documented within the database were considered in the retrieval. Any interactions which were flagged as a conflict as described in Chapter 2 were not included in the training or test sets to ensure that the models used the most correct information possible.

Figure 3.2 provides a summary of the numbers of compounds that were retrieved from the three main panel approaches, which would be the numbers considered for all approaches as they contain at least one interaction. The fourth column in the table details compounds which are present within both databases (i.e compounds which contain interaction information from both DrugBank and ToxCast as described in Chapter 2). These were initially removed from the model to provide the best chance of

demonstrating a differentiation between the DrugBank and ToxCast sources, however this presented the test with a heavier class imbalance between "Good" and "Bad" profiles.

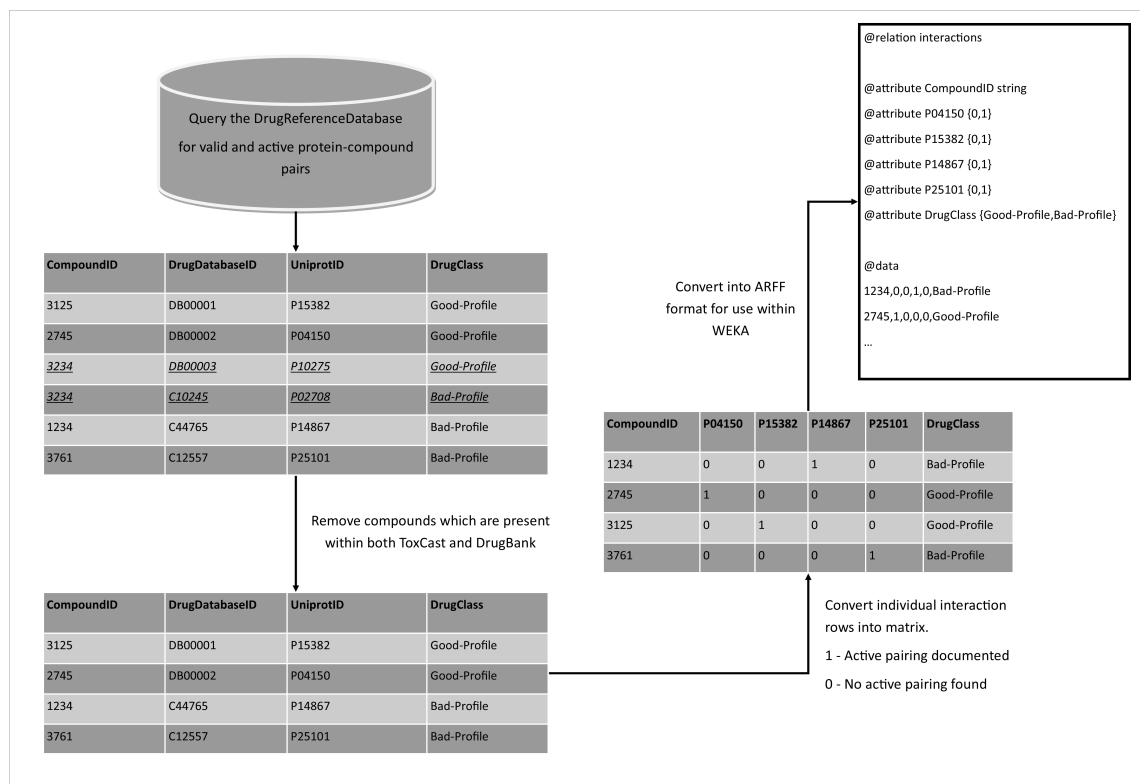
Once the lists had been retrieved and pre-processed, the task of conversion to a format suitable for WEKA could commence. Figure 3.4 provides an illustrated example of this process, where the database's individual interactions are converted into a matrix format. In addition to this, WEKA has a specific format known as ARFF which provides header information to describe each attribute within the dataset. Attributes which can be considered within WEKA include numeric, string and nominal (i.e class) values. R provided a library package known as "foreign" to easily generate the header file of the matrix for this purpose [27].



**Figure 3.2:** Summary of distribution of "Good" (DrugBank) and "Bad" (ToxCast) profile compounds based on active interactions only. The number of conflicting compounds refers to compounds which were present within both DrugBank and ToxCast and which were excluded from the training

```
SELECT p.CompoundID,u.UniprotID FROM hit h
INNER JOIN pubchem p ON p.PubchemID=h.PubchemID
INNER JOIN protein u ON u.ProteinID=h.ProteinID
WHERE d.DrugClass="Good-Profile" AND h.IsValid="TRUE"
AND u.UniprotID IN ('P10275','P04150',...)
```

**Figure 3.3:** Example query used to obtain active compound-protein pairings for "Good-Profile" drugs



**Figure 3.4:** Illustrated example of the process used to convert interactions found in the database to a format suitable for use in WEKA

### 3.3.1 Initial Findings

From initial analysis, it was clear that there was an imbalance present between both classes within all sets, which could cause an impact on the performance of the classifiers. In addition, there was also a clear imbalance in the reported protein screenings involving several proteins. Table 3.2, 3.3 and 3.4 provides a summary of the 15 most screened proteins in terms of active results for Panel 44, Panel 331 and Pharmacology panel respectively after potential conflicts had been removed. Between all panels, there is a greater degree of interactions present with the "Bad" ToxCast compounds in comparison to the "Good" DrugBank compounds, in particular with a handful of proteins which have significantly higher amounts of interaction in comparison to the rest of the panel. This is likely due to the higher reported incidence of interactions of the "Bad" profile class, in addition to the way that the ToxCast interaction matrix had been parsed, with certain assays involving more than one protein. Despite this, there appeared to be some distinction present between both classes for highly screened proteins to potentially determine a difference in interaction profiles. For example, in the case of Panel 44, each class in the top 15 interactions



had 5 proteins which were not present within the opposing class, whilst for Panel 331 and the Pharmacological panel the top 15 proteins in each class were different.

In order to prevent issues regarding the class imbalance, further adjustments were made to the classes in terms of their weighting, which is the impact an instance's class will have within the dataset. This was accomplished through the use of WEKA's ClassBalancer pre-processing command, which adjusted the weighting of attributes so that the smaller population of "Good" profile compounds were overall considered equally to the "Bad" profile compounds by the classifiers. The weight adjusted values were considered in conjunction with the dataset unmodified with all panels to assess their impact on overall performance.

Panel 44					
Good Profile Compounds (DrugBank)			Bad Profile Compounds (ToxCast)		
UniProt ID	Entry Name	No. of Interactions	UniProt ID	Entry Name	No. of Interactions
Q12809	KCNH2_HUMAN	75	<b>P10275</b>	<b>ANDR_HUMAN</b>	<b>1,449</b>
<b>P11229</b>	<b>ACM1_HUMAN</b>	<b>74</b>	<b>P04150</b>	<b>GCR_HUMAN</b>	<b>1,034</b>
P08172	ACM2_HUMAN	72	<b>P22303</b>	<b>ACES_HUMAN</b>	<b>288</b>
P35367	HRH1_HUMAN	66	Q01959	SC6A3_HUMAN	245
<b>P35348</b>	<b>ADA1A_HUMAN</b>	<b>64</b>	P35354	PGH2_HUMAN	233
P08913	ADA2A_HUMAN	60	P23975	SC6A2_HUMAN	226
P20309	ACM3_HUMAN	60	Q12809	KCNH2_HUMAN	150
P23975	SC6A2_HUMAN	58	P20309	ACM3_HUMAN	141
P28223	5HT2A_HUMAN	58	P08913	ADA2A_HUMAN	134
<b>P31645</b>	<b>SC6A4_HUMAN</b>	<b>53</b>	P08172	ACM2_HUMAN	130
P14416	DRD2_HUMAN	52	P28223	5HT2A_HUMAN	121
<b>P08908</b>	<b>5HT1A_HUMAN</b>	<b>51</b>	<b>P35372</b>	<b>OPRM_HUMAN</b>	<b>119</b>
Q01959	SC6A3_HUMAN	50	P14416	DRD2_HUMAN	117
<b>P07550</b>	<b>ADRB2_HUMAN</b>	<b>47</b>	<b>P21728</b>	<b>DRD1_HUMAN</b>	<b>111</b>
P35354	PGH2_HUMAN	46	P35367	HRH1_HUMAN	107

**Table 3.2:** Top 15 interacting proteins in Panel 44. Proteins which are not present in the opposite class' top 15 interactions are highlighted in bold font.

Panel 331					
Good Profile Compounds (DrugBank)			Bad Profile Compounds (ToxCast)		
UniProt ID	Entry Name	No. of Interactions	UniProt ID	Entry Name	No. of Interactions
P10635	CP2D6_HUMAN	101	O75469	NR1I2_HUMAN	2054
P08684	CP3A4_HUMAN	88	P37231	PPARG_HUMAN	1532
P42574	CASP3_HUMAN	75	P10275	ANDR_HUMAN	1449
Q12809	KCNH2_HUMAN	75	Q96RI1	NR1H4_HUMAN	1409
P11229	ACM1_HUMAN	74	P04150	GCR_HUMAN	1034
P08172	ACM2_HUMAN	72	P11511	CP19A_HUMAN	938
P35367	HRH1_HUMAN	66	P10826	RARB_HUMAN	581
P05177	CP1A2_HUMAN	63	Q14994	NR1I3_HUMAN	551
P08913	ADA2A_HUMAN	60	P12931	SRC_HUMAN	518
P20813	CP2B6_HUMAN	60	P05177	CP1A2_HUMAN	500
P23975	SC6A2_HUMAN	58	P04798	CP1A1_HUMAN	467
P28223	5HT2A_HUMAN	58	P03956	MMP1_HUMAN	465
P31645	SC6A4_HUMAN	53	P14780	MMP9_HUMAN	453
P11712	CP2C9_HUMAN	52	Q07869	PPARA_HUMAN	451
P14416	DRD2_HUMAN	52	P33261	CP2CJ_HUMAN	370

**Table 3.3:** Top 15 interacting proteins in Panel 331

Pharmacology Panel					
Good Profile Compounds (DrugBank)			Bad Profile Compounds (ToxCast)		
UniProt ID	Entry Name	No. of Interactions	UniProt ID	Entry Name	No. of Interactions
P10635	CP2D6_HUMAN	101	O75469	NR1I2_HUMAN	2054
P01375	TNFA_HUMAN	93	P37231	PPARG_HUMAN	1532
P08183	MDR1_HUMAN	92	P10275	ANDR_HUMAN	1449
P08684	CP3A4_HUMAN	88	Q03181	PPARD_HUMAN	1208
Q12809	KCNH2_HUMAN	75	P04150	GCR_HUMAN	1034
P11229	ACM1_HUMAN	74	P11511	CP19A_HUMAN	938
P05231	IL6_HUMAN	72	P11473	VDR_HUMAN	814
P08172	ACM2_HUMAN	72	P19838	NFKB1_HUMAN	748
P10415	BCL2_HUMAN	70	P13500	CCL2_HUMAN	701
P35367	HRH1_HUMAN	66	P28702	RXRB_HUMAN	633
P35348	ADA1A_HUMAN	64	P05412	JUN_HUMAN	582
Q92887	MRP2_HUMAN	63	P10826	RARB_HUMAN	581
P01584	IL1B_HUMAN	60	Q03405	UPAR_HUMAN	556
P08913	ADA2A_HUMAN	60	P13631	RARG_HUMAN	540
P20309	ACM3_HUMAN	60	P05121	PAI1_HUMAN	526

**Table 3.4:** Top 15 interacting proteins in the Pharmacology Panel

### 3.3.2 Classifier Results

Table 3.5 displays the results of the classifiers against the panels specified where weights were unmodified. The first classifier in the table (ZeroR) simply classifies all instances as the majority class within the training fold, so classifiers which have better accuracy levels than ZeroR are generating decisions which are better than simply assuming all instances are the training dataset’s majority class. In this instance, most classifiers did not exceed overall accuracy levels significantly, and accuracy levels for classifying ”Good” profile compounds were poor. In terms of panel selection having an impact on accuracy, the increase of scale in search space had a positive impact on performance. Table 3.6 displays the result of the classifiers when class balancing was performed, which showed that while ”Good” profile accuracy had improved with this alteration, ”bad” profile accuracy had suffered in turn. Increases in the scale of protein search space had also increased accuracy levels with the weighting change implemented.

Classifier Algorithm	Accuracy (%)								
	Panel 44			Panel 331			Pharma Panel		
	Good Prof.	Bad Prof.	Overall	Good Prof.	Bad Prof.	Overall	Good Prof.	Bad Prof.	Overall
ZeroR (Baseline)	0	100	85.71	0	100	89.3	0	100	88.14
DecisionTable	27.5	96.9	87.02	18.7	98.5	89.94	26.2	97.8	89.32
JRip	33.4	96.3	87.28	30.9	97.4	90.32	34.2	96.9	89.49
PART	30.4	95.9	86.56	42.9	96.2	90.51	52.1	94.9	89.81
J48	31.6	96.3	87.02	30	97.2	90	38.7	96	89.16
RandomForest	28.1	96	86.27	39.3	97.3	91.12	36.1	97.6	90.32
RandomTree	25.4	95.1	85.15	37.2	95.6	89.37	37.9	95.2	88.39
REPTree	30	96.3	86.82	30.5	97.4	90.26	41.2	96.3	89.79
Logistic	29.1	96.6	86.95	38.4	97.1	90.8	47.1	93.9	88.33
NaiveBayes	41.9	91.5	84.43	47.3	92.7	87.83	44.7	92.8	87.14

**Table 3.5:** Classifier results for active interactions only where attributes were unmodified in weightings

Classifier Algorithm	Accuracy (%)								
	Panel 44			Panel 331			Pharma Panel		
	Good Prof.	Bad Prof.	Overall	Good Prof.	Bad Prof.	Overall	Good Prof.	Bad Prof.	Overall
ZeroR (Baseline)	29.5	70	64.19	50.4	50.1	50.09	50.4	50	50.08
DecisionTable	87.2	68.9	71.52	87.8	75.7	77	91.3	67	69.92
JRip	84.9	72.3	74.13	89.3	76.7	78.04	90.8	71.4	73.73
PART	73	79.4	78.45	74.6	84.8	83.75	61.8	92.1	88.49
J48	81	75.5	76.29	75.4	83.1	82.24	64.3	90.3	87.23
RandomForest	68.6	81.5	79.63	74.6	85.6	84.45	64	92.7	89.3
RandomTree	60.9	81.2	78.29	66	85.7	83.59	51.9	91.2	86.56
REPTree	82.2	75.2	76.23	88	76.7	77.89	84.7	78.1	78.84
Logistic	81.2	75.4	76.19	78.4	84.6	83.96	63.6	89.1	86.1
NaiveBayes	71.2	79.7	78.52	80.5	81.4	81.3	64.9	89.3	86.43

**Table 3.6:** Classifier results for active interactions only where class balancing was performed

### 3.3.3 Discussion

From the results gathered from the protein panels, it was found that the class imbalance had some impact on overall classification accuracy, requiring the need for implementing some element of class distribution adjustment in order to retrieve more satisfactory classification levels for "Good" profile compounds. To illustrate an example of the impact of balancing on the classifier models, Figure 3.5 displays an example model of the J48 classifier generated from the complete unmodified dataset within Panel44, while Figure 3.6 displays the model with the ClassBalancer implemented on the dataset. On the class decision branches, the number on the left indicates how many instances were classified correctly, while the one on the right indicates how many instances were classed incorrectly by the model. It can be noted that while the tops of both trees place similar emphasis on the heavily screened proteins listed earlier, the balanced tree places additional emphasis on interactions taking place more widely within the panel which led to the increase in accuracy.

While the classifiers did not perform as well with regard to "Good" profile compounds, it was pleasing to find that a large majority of the "Bad" profile class

had been classified on models which made use of a number of proteins for potential profiling, with some classifiers correctly predicting 80+% of the compounds on the cross validation testing. With some compound crossover being found between DrugBank and ToxCast, it is expected that there may be some instances where certain compounds could be considered beneficial in one context but can also have the potential to cause harm in another. Further consideration should therefore have been made between the modes of action and dosage to further refine the classifications. However, there is some reasonable basis to the categorisations, as they distinguish between two general groupings of compounds – those that are largely non-toxic at typical therapeutic dosage levels and those that are largely not.

This would require further analysis in determining which beneficial compounds could be considered too harmful; one area of promise would be in the use of different repositories to determine beneficial and harmful candidate drugs not considered by the model and to be considered later in the chapter after all avenues of modification have been explored.

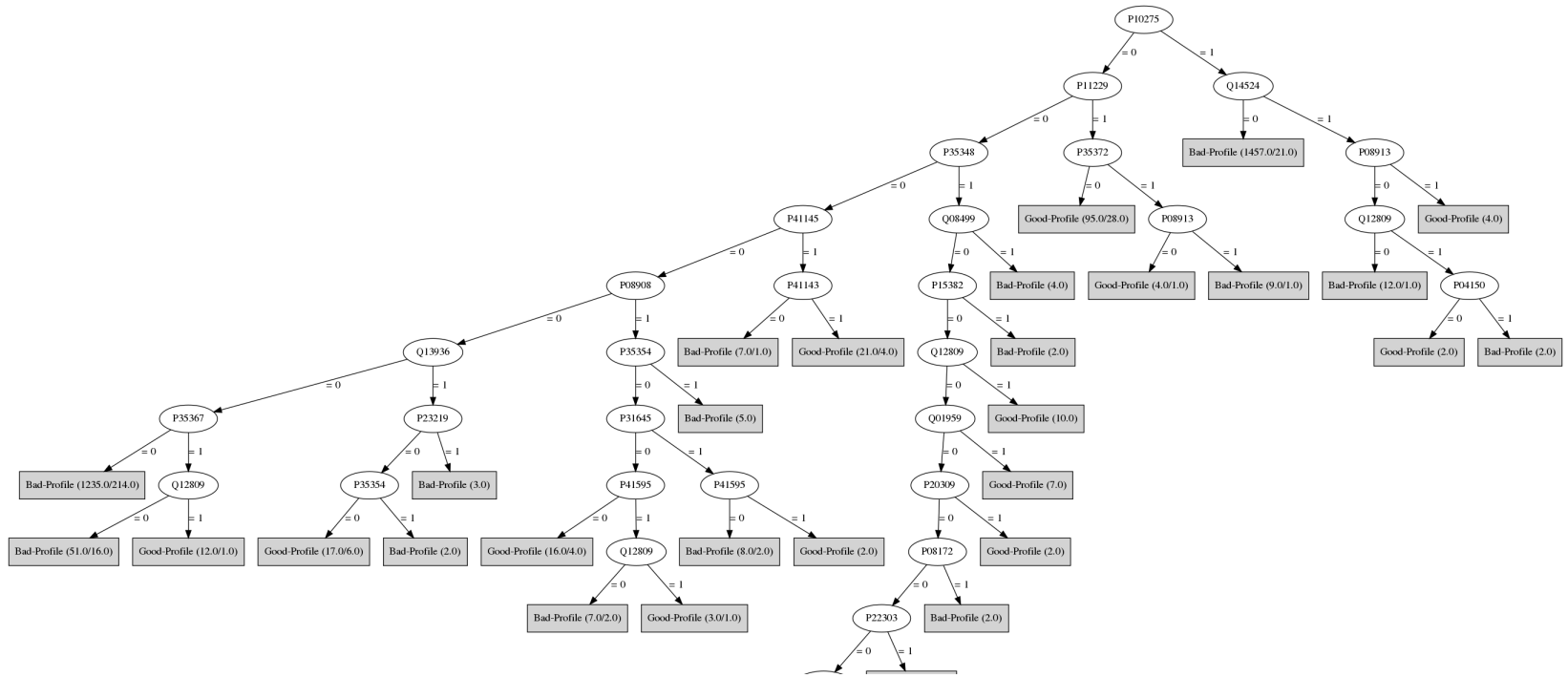


Figure 3.5: Segment of the J48 Classification tree applied to Panel 44 interactions, before class balancing

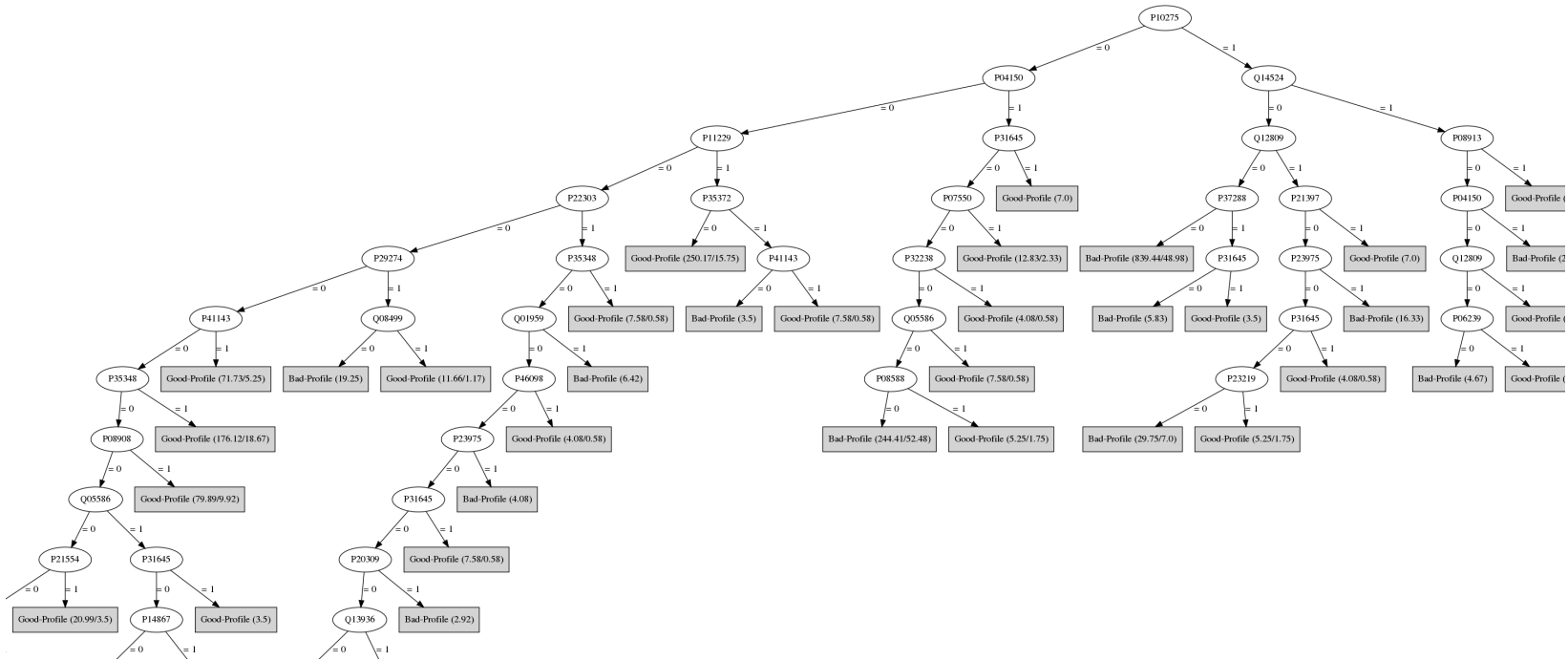
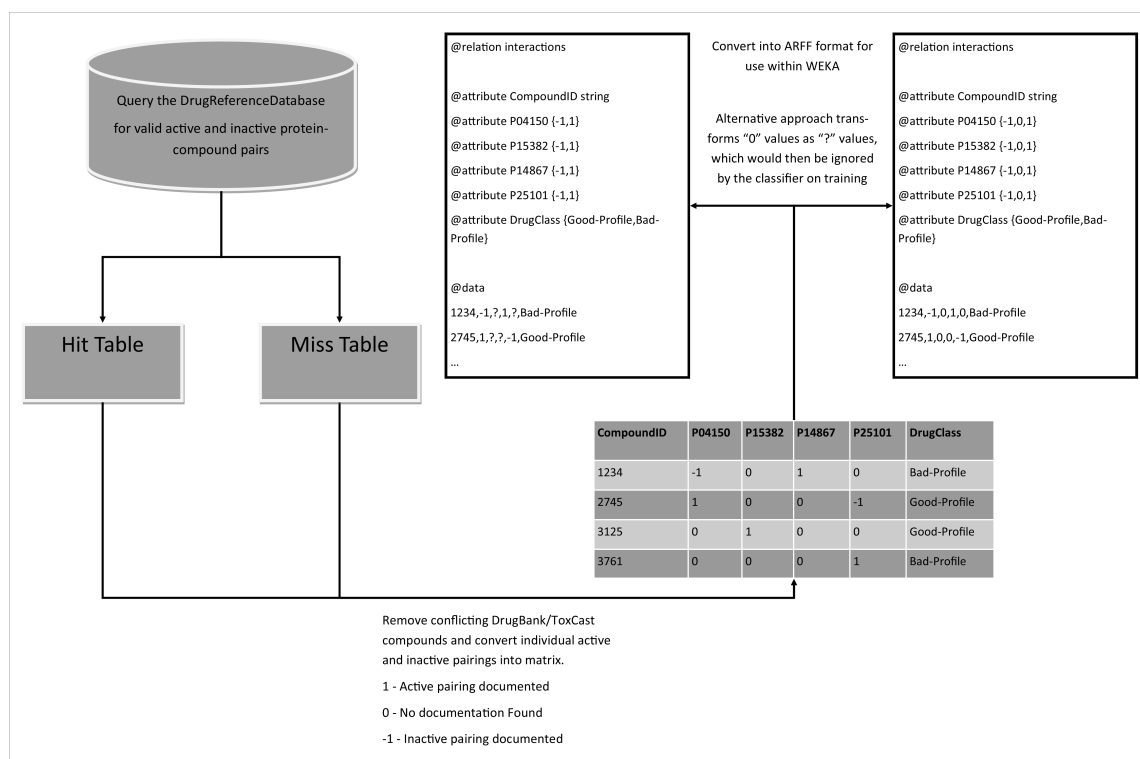


Figure 3.6: Segment of the J48 Classification tree applied to Panel 44 interactions, after class balancing

### 3.4 Active and Inactive Pairs

To consider inactive compound protein pairings for this part of the analysis, it was necessary to perform some modifications to the construction process of the interaction matrix. In this instance, two types of modification were performed and assessed: in the first modification, the matrix assigned a '-1' nominal value to areas of the matrix where an inactive pairing had been found. This would leave the '0' nominal value as instances where no active or inactive confirmation had been documented in order to be considered by the classifier. The second modification would involve replacing the '0' nominal value for undocumented pairs as a non-applicable value so they would not be considered as a value within the classifier's training or testing. As the vast majority of the interaction search space being queried is missing, it was expected that the overall performance of this second modification would be poor. Figure 3.7 displays an illustrated example of the changes that were made to accommodate both modifications, while Figure 3.8 provides an example of the query that had been used to retrieve inactive protein-compound pairs from the DrugReferenceDatabase.



**Figure 3.7:** Illustrated example of the revised process used to convert active and inactive compound-protein pairs found in the database to a format suitable for use within WEKA

```

SELECT p.CompoundID,u.UniprotID FROM miss m
INNER JOIN pubchem p ON p.PubchemID=m.PubchemID
INNER JOIN protein u ON u.ProteinID=m.ProteinID
WHERE d.DrugClass="Good-Profile" AND m.IsValid="TRUE"
AND u.UniprotID IN ('P10275','P04150',...)

```

**Figure 3.8:** Query used to obtain inactive compound-protein pairings for "Good-Profile" drugs

### 3.4.1 Initial Findings

Tables 3.7, 3.8 and 3.9 provide summaries of the top 15 proteins which were considered to be inactive in the panels used. The distributions highlighted in these tables highlight that an equivalent or greater number of inactive results exist within the searchspace, however even with the inclusion of these results there would still be a large proportion of the interaction matrix which would be considered as inconclusive. This in turn would lead either to classifiers being trained on a large proportion of missing information, or classifiers which would likely mainly focus on inconclusive results. In the former case, the classifiers would likely assume that the majority class would be the most accurate decision path, while the latter would likely have issues on compounds which have been screened more heavily and contain more active and inactive results than those found in the training model.

Panel 44					
Good Profile Compounds (DrugBank)			Bad Profile Compounds (ToxCast)		
UniProt ID	Entry Name	No. of Interactions	UniProt ID	Entry Name	No. of Interactions
P21728	DRD1_HUMAN	85	P04150	GCR_HUMAN	1467
P37288	V1AR_HUMAN	78	P10275	ANDR_HUMAN	890
P41143	OPRD_HUMAN	70	P35372	OPRM_HUMAN	644
P35372	OPRM_HUMAN	69	P41143	OPRD_HUMAN	638
P41145	OPRK_HUMAN	69	P37288	V1AR_HUMAN	615
P32238	CCKAR_HUMAN	61	P14416	DRD2_HUMAN	611
P25101	EDNRA_HUMAN	60	P11229	ACM1_HUMAN	577
P06239	LCK_HUMAN	58	P21728	DRD1_HUMAN	495
P21554	CNR1_HUMAN	58	Q12809	KCNH2_HUMAN	421
P14416	DRD2_HUMAN	57	P41145	OPRK_HUMAN	375
P29274	AA2AR_HUMAN	57	P51787	KCNQ1_HUMAN	335
P21397	AOFA_HUMAN	56	P08908	5HT1A_HUMAN	142
P22303	ACES_HUMAN	52	P31645	SC6A4_HUMAN	138
P23219	PGH1_HUMAN	52	P28223	5HT2A_HUMAN	137
P04150	GCR_HUMAN	51	P08588	ADRB1_HUMAN	136

**Table 3.7:** Top 15 proteins with inactive results in Panel 44



Panel 331					
Good Profile Compounds (DrugBank)			Bad Profile Compounds (ToxCast)		
UniProt ID	Entry Name	No. of Interactions	UniProt ID	Entry Name	No. of Interactions
P29466	CASP1_HUMAN	125	P04150	GCR_HUMAN	3170
P28482	MK01_HUMAN	113	P11511	CP19A_HUMAN	2822
P21728	DRD1_HUMAN	102	P37231	PPARG_HUMAN	2629
P37288	V1AR_HUMAN	102	P10276	RARA_HUMAN	2575
P35372	OPRM_HUMAN	92	Q96RI1	NR1H4_HUMAN	2544
P49146	NPY2R_HUMAN	86	P10275	ANDR_HUMAN	2485
P25929	NPY1R_HUMAN	85	P10827	THA_HUMAN	2361
P11712	CP2C9_HUMAN	83	Q07869	PPARA_HUMAN	2033
P08246	ELNE_HUMAN	79	P10826	RARB_HUMAN	2020
P33261	CP2CJ_HUMAN	79	Q14994	NR1I3_HUMAN	1994
P08575	PTPRC_HUMAN	78	P03372	ESR1_HUMAN	1763
P14416	DRD2_HUMAN	78	P29466	CASP1_HUMAN	1119
P25101	EDNRA_HUMAN	78	P28482	MK01_HUMAN	1074
P08912	ACM5_HUMAN	77	P35968	VGFR2_HUMAN	1054
P21452	NK2R_HUMAN	77	P12931	SRC_HUMAN	1031

**Table 3.8:** Top 15 proteins with inactive results in Panel 331

Pharmacology Panel					
Good Profile Compounds (DrugBank)			Bad Profile Compounds (ToxCast)		
UniProt ID	Entry Name	No. of Interactions	UniProt ID	Entry Name	No. of Interactions
P13569	CFTR_HUMAN	233	P16473	TSHR_HUMAN	3876
P11473	VDR_HUMAN	208	P04150	GCR_HUMAN	3775
P10828	THB_HUMAN	141	P19838	NFKB1_HUMAN	3703
P10253	LYAG_HUMAN	137	P19793	RXRA_HUMAN	3422
P28482	MK01_HUMAN	134	P11511	CP19A_HUMAN	3404
P21728	DRD1_HUMAN	121	Q92731	ESR2_HUMAN	3401
P37288	V1AR_HUMAN	115	Q03181	PPARD_HUMAN	3291
P41143	OPRD_HUMAN	104	P37231	PPARG_HUMAN	3266
P35372	OPRM_HUMAN	103	P10276	RARA_HUMAN	3095
P41145	OPRK_HUMAN	101	P10275	ANDR_HUMAN	3082
P11712	CP2C9_HUMAN	97	P05412	JUN_HUMAN	3025
P14416	DRD2_HUMAN	92	P62508	ERR3_HUMAN	2650
P33261	CP2CJ_HUMAN	92	P11473	VDR_HUMAN	2610
P08246	ELNE_HUMAN	88	P10827	THA_HUMAN	2495
P08912	ACM5_HUMAN	88	P03372	ESR1_HUMAN	2269

**Table 3.9:** Top 15 proteins with inactive results within the Pharmacology Panel

### 3.4.2 Classifier Results

Tables 3.10 and 3.11 show the results of the classifiers when inactive results were considered and unknown pairings were excluded from consideration. The results show that the removal of the unknown variables had a negative impact on the overall accuracy of the results in the unmodified analysis, and in most cases classifiers had assumed the same majority only model as *ZeroR*. In the instance of the DecisionTable rule, which achieved far higher accuracy levels, the models generated resorted to using a very small selection of highly screened proteins within the panel to determine a compound’s potential profile, and so this is likely to be overfitted. While class balancing improved overall accuracy levels, these metrics did not exceed the levels that were obtained in Table 3.6, where class balancing was applied on active interactions where attributes were unmodified in weightings.

For the second approach, the results of which are found in Tables 3.12 and 3.13, inconclusive values were given a separate nominal class in conjunction with the introduction of inactive classes. The addition of inactive pairs provided an improvement in overall accuracy in comparison to Tables 3.5 and 3.6 in both unmodified and profile weight balanced cases, however these models are again likely to be overfitted to the dataset. Further assessment of some of the classifier models confirmed that some comparisons relied on inconclusive results to make decisions, which would likely cause issues for compounds which have been more heavily screened and contain more active and inactive results than the profiles used to train the models. While a similar metric was applied during the Active Only experiment, it was not possible to obtain a complete screening profile for the compounds, and it was reasonable to group all instances of a protein containing no active or target results against a compound as a class in itself during the Active Only experiments.

Classifier Algorithm	Accuracy (%)								
	Panel 44			Panel 331			Pharma Panel		
	Good Prof.	Bad Prof.	Overall	Good Prof.	Bad Prof.	Overall	Good Prof.	Bad Prof.	Overall
ZeroR (Baseline)	0	100	85.71	0	100	89.3	0	100	88.14
DecisionTable	83.3	98.1	95.95	94.7	99.6	99.08	95.7	99.8	99.32
JRip	13.5	98.1	86	5.3	99.2	89.14	0	0	0
PART	0	100	85.71	0	100	89.3	0	100	88.14
J48	0	100	85.71	0	100	89.3	0	100	88.14
RandomForest	0	100	85.71	0	100	89.3	0	100	88.14
RandomTree	0	100	85.71	0	100	89.3	0	100	88.14
REPTree	0	100	85.71	0	100	89.3	0.6	100	88.19
Logistic	38.9	97	88.72	35.3	97.5	90.83	37.6	97.4	90.3
NaiveBayes	21.3	96	85.28	30.2	90.1	83.69	40	92	85.87

**Table 3.10:** Classifier results when inactive results (excluding inconclusive results) are incorporated, and where attributes were unmodified in weightings

Classifier Algorithm	Accuracy (%)								
	Panel 44			Panel 331			Pharma Panel		
	Good Prof.	Bad Prof.	Overall	Good Prof.	Bad Prof.	Overall	Good Prof.	Bad Prof.	Overall
ZeroR (Baseline)	29.5	70	64.19	50.4	50.1	50.09	50.4	50	50.08
DecisionTable	93.6	93.9	93.88	96.8	98.1	98	98.4	98.9	98.81
JRip	78.9	80.4	80.15	86.6	86.9	86.87	83.9	90.5	89.74
PART	65	65.7	65.57	59.5	72.7	71.26	74.8	78.2	77.77
J48	50.3	64.6	62.59	63.4	64.6	64.44	50.5	78.8	75.43
RandomForest	63.6	70.9	69.88	63.6	70.9	69.88	74.7	73.8	73.92
RandomTree	61.6	69.9	68.71	76.1	71.2	71.69	73.3	71.2	71.46
REPTree	62.9	66.5	65.99	77.3	71.9	72.5	78.7	76.6	76.81
Logistic	91.8	68.3	71.62	83.6	76.5	77.28	84.9	78	78.83
NaiveBayes	62.9	55.9	56.9	80.9	87.9	87.12	79.1	91	89.6

**Table 3.11:** Classifier results when inactive results (excluding inconclusive results) are incorporated, and class balancing was performed

Classifier Algorithm	Accuracy (%)								
	Panel 44			Panel 331			Pharma Panel		
	Good Prof.	Bad Prof.	Overall	Good Prof.	Bad Prof.	Overall	Good Prof.	Bad Prof.	Overall
ZeroR (Baseline)	0	100	85.71	0	100	89.3	0	100	88.14
DecisionTable	83.3	98.1	95.95	94.8	99.6	99.1	95.7	99.8	99.31
JRip	91.1	97.4	96.5	96.4	99.2	98.92	97.3	99.4	99.19
PART	87.4	97.7	96.24	94.8	99.3	98.96	96.1	99.8	99.37
J48	87.2	97.7	96.21	93.1	99.5	98.78	92.4	99.7	98.88
RandomForest	87.4	98.4	96.83	96.2	99.9	99.49	94.8	100	99.35
RandomTree	83.1	97.4	95.32	92.2	99.2	98.49	90.5	98.4	97.47
REPTree	88.1	97.8	96.4	95.6	99.6	99.14	96.3	99.7	99.26
Logistic	83.5	97.5	95.52	91.8	97.6	96.96	78.4	96.1	93.96
NaiveBayes	86.7	91.6	90.91	98.9	96.7	96.92	98.2	93.6	94.17

**Table 3.12:** Classifier results when inactive results (including inconclusive results) are incorporated, and where attributes were unmodified in weightings

Classifier Algorithm	Accuracy (%)								
	Panel 44			Panel 331			Pharma Panel		
	Good Prof.	Bad Prof.	Overall	Good Prof.	Bad Prof.	Overall	Good Prof.	Bad Prof.	Overall
ZeroR (Baseline)	29.5	70	64.19	50.4	50.1	50.09	50	50.4	50.08
DecisionTable	93.6	93.9	93.88	96.8	98.2	98.04	98.4	98.9	98.81
JRip	93.6	92.3	92.48	98.3	98.9	98.82	97.5	99.3	99.07
PART	92.9	93.6	93.49	96.6	99.2	98.92	97.5	99.5	99.28
J48	93.1	94.2	94.08	96.9	98.9	98.67	96.7	99.2	98.93
RandomForest	93.1	95.1	94.8	98.3	99.8	99.61	98.7	99.7	99.61
RandomTree	83.3	94.1	92.58	92.9	99	98.35	91	98.6	97.66
REPTree	94.3	93.8	93.85	96.9	98.9	98.71	98.1	99.3	99.12
Logistic	91.8	93.2	93	94.8	98.9	98.49	94.7	99.3	98.77
NaiveBayes	91.8	90.5	90.71	98.9	96.2	96.49	98.5	92.5	93.18

**Table 3.13:** Classifier results when inactive results (including inconclusive results) are incorporated, and class balancing was performed

### 3.4.3 Discussion

While the inclusion of inactive results showed promise in providing an improvement in accuracy of profiling, it is unlikely that such accuracy levels would be obtained when tested in a real world situation. It is a more likely scenario that despite the inclusion of various repositories in the searching for interactions, certain compounds will only be assessed in association with certain proteins, which appears to be providing a distinction between repositories instead of providing a protein interaction model

of promise for further development. As it would be impractical for this study to conduct complete experimental panel screening of the compounds used for training, *in silico* work would likely be needed to fill in gaps in an interaction matrix to ensure that the most complete interaction profile possible can be provided for training the classifiers, or to provide additional information from more easily extracted resources to differentiate between the "Good" and "Bad" compounds assessed.

## 3.5 Chemical Properties

In addition to considering a compound's protein interaction profile, another area which was considered for profiling was a compound's chemical composition. Two avenues of approach were considered for evaluating models in this section: the first involved the use of only the chemical property flags present within the compound's SDF file that could be practically implemented by a classifier, while the second approach would use these properties in conjunction with the protein interaction flags from the previous analyses. The second approach considered both the active targets only analysis in addition to the active and inactive targets analysis to assess the impact that the addition of chemical properties would have on the model accuracy levels.

To obtain the chemical properties, the python external library "Mordred" [28] was used to parse through the training SDF files in order to generate a collection of properties. As the DrugReferenceDatabase contained references to compound SDF files, Mordred provided an automated means of extracting property information, in addition to providing greater levels of detail than that provided by the use of PubChem's PUGREST platform.

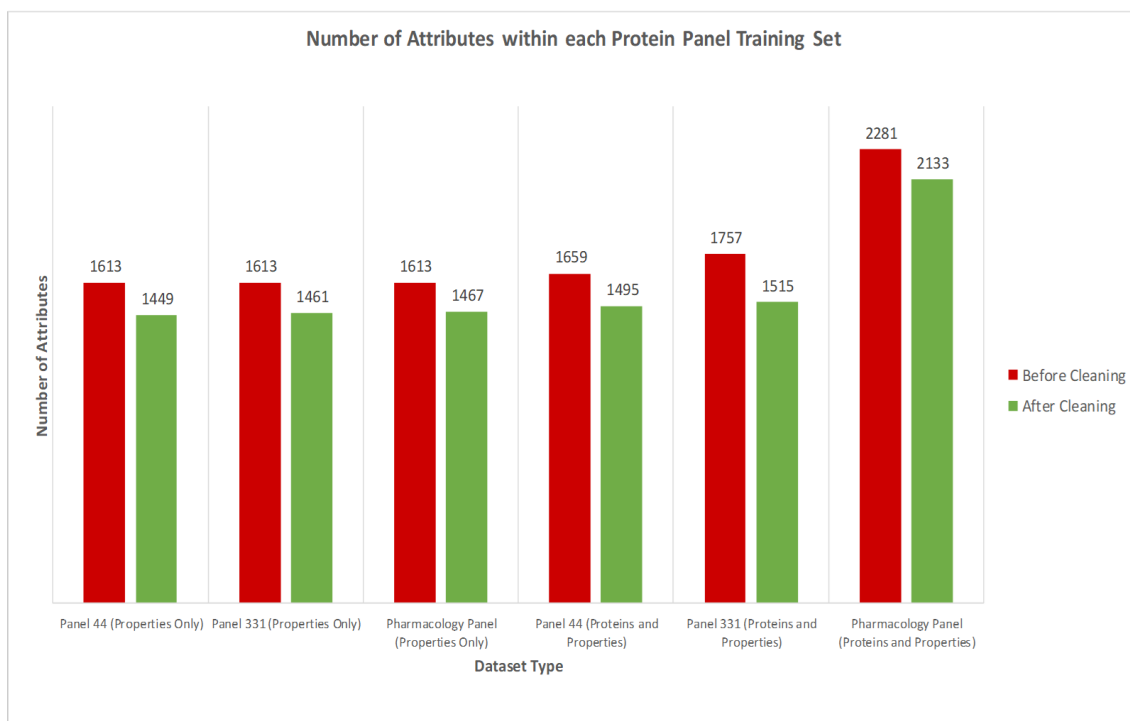
### 3.5.1 Initial Findings

On running the script on all documented compounds, it was found that in some instances a compound caused the "Mordred" descriptor script to crash as a result of a failure to parse the SDF when extracting certain properties. This affected only 5 compounds of the 6,349 compounds queried which had an SDF, and so these were removed from this section of the analysis. The compounds in question are documented in Table 3.14, along with the potential impact on the panels caused by their removal. Once the script had parsed the structures successfully, 1,613 chemical attributes were generated from the set which were suitable for use within a classifier. These

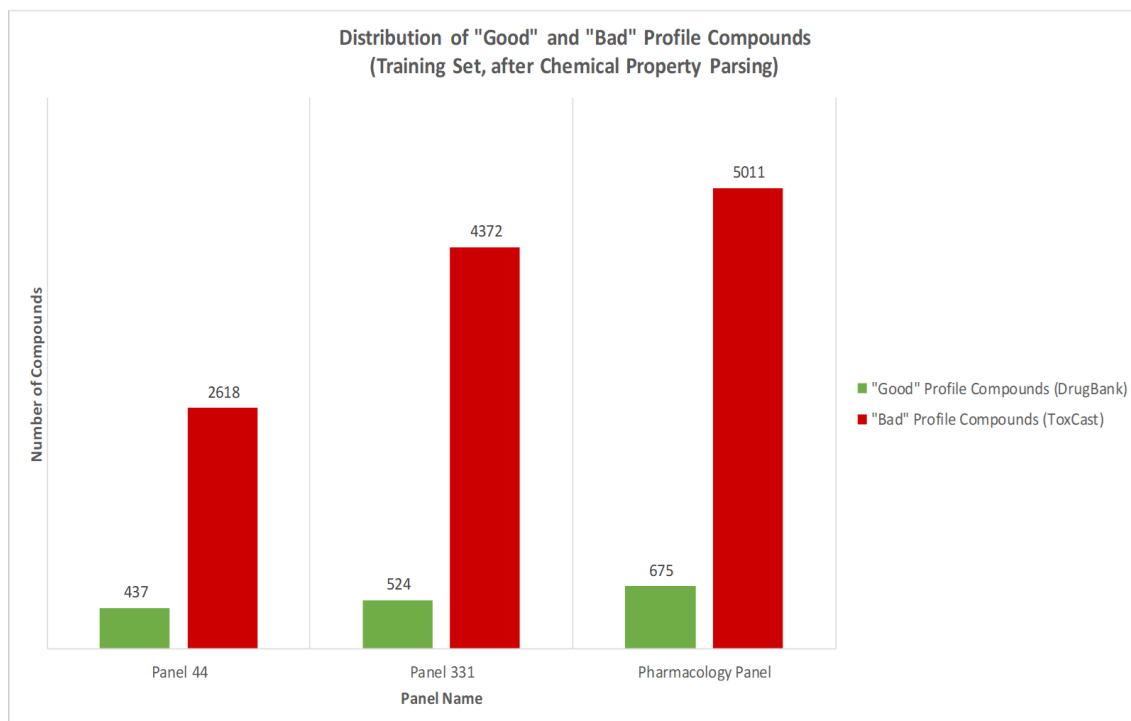
attributes are listed within the Appendix. In cases where attributes with either little or no variation were removed (via WEKA's RemoveUseless pre-processing filter), the actual number of attributes would range between 1,400-1,500 attributes depending on the scale of the panel screened; specific attribute numbers are shown in Figure 3.9. Figure 3.10 provides a revised distribution of classes for the training set after chemical properties have been incorporated.

PubChem Compound ID	Compound Name	Impact on Panel 44	Impact on Panel 331	Impact on Pharma Panel
62390	Nickelocene	No	No	Yes
79154	Chromium	No	Yes	Yes
82917	Dichlorovanadocene	Yes	Yes	Yes
92884	Cobaltocene	Yes	No	Yes
299728	Triphenylbismuth dichloride	Yes	Yes	Yes

**Table 3.14:** List of five compounds which could not be parsed through the Mordred library



**Figure 3.9:** Graph of the number of attributes within each panel after incorporating chemical properties



**Figure 3.10:** Summary of distribution of "Good" (DrugBank) and "Bad" (ToxCast) profile compounds after parsing for chemical properties. Conflicting compounds between classes have already been removed for this graph.

### 3.5.2 Classifier Results

Tables 3.15 and 3.16 provide the results of profiling classifiers based solely on the chemical properties, with which performance metrics were improved slightly in comparison to using active interactions only. One element of interest from the results was that with the NaiveBayes classifier, the consideration of properties had caused a "flip" in the pattern of class accuracy in comparison to the other classifiers, and had performed well in classifying "Good" profile instances but toiled in misclassifying over half of the compounds considered to be a "Bad" profile. Class balancing once again revealed similar findings to before, with "Good" profile prediction accuracy levels improving but with a trade-off of slightly poorer "Bad" profile prediction accuracy.

Tables 3.17 and 3.18 provide the results for the classifiers when chemical properties are considered in conjunction with the protein active flags specified in the "Active Pairs Only" experiment. These attributes when combined had provided an increase in accuracy levels for a majority of classifiers in comparison to considering each method individually. For example, on the unweighted REPTree classifier and Panel 44, the profile prediction accuracy levels for "Good" and "Bad" profile compounds were 41.9% and 94.7% respectively on just using chemical properties, and 30% and

96.3% with the protein activities only. With chemical properties combined with the interaction profiles, the profile prediction accuracy levels increased to 59.5% and 94.5% for "Good" and "Bad" profile compounds respectively. This provides an increase of between 17.6% to 29.5% on "Good" profile prediction accuracy, with a negligible decrease in "Bad" profile prediction accuracy. This highlights promise that attributes from each set are being used to reach better conclusions.

In terms of Panel performance, larger sized panels again did not necessarily provide an increase in profiling performance, with the Pharmacology panel not providing as much of a performance increase in comparison to the smaller panels of Panel 44 and Panel 331. For example with the unbalanced REPTree classifier and the Pharmacology Panel, the increase of "Good" profile prediction accuracy was between 7.4% to 20.2%, indicating that the greater amount of time to train and evaluate the larger sets may not be beneficial.

Classifier Algorithm	Accuracy (%)								
	Panel 44			Panel 331			Pharma Panel		
	Good Prof.	Bad Prof.	Overall	Good Prof.	Bad Prof.	Overall	Good Prof.	Bad Prof.	Overall
ZeroR (Baseline)	0	100	85.7	0	100	89.3	0	100	88.13
DecisionTable	33.4	95.5	86.64	22.3	97.7	89.67	18.7	97.7	88.29
JRip	51.5	91	85.33	41	94.4	88.73	44.3	93	87.21
PART	45.5	94.8	87.79	39.5	96	89.99	37.5	96.1	89.18
J48	49.7	92.2	86.12	50	94.3	89.52	50.7	93.1	88.02
RandomForest	19.5	97.7	86.55	14.3	98.7	89.67	14.1	98.3	88.34
RandomTree	37.1	90.1	82.52	31.7	93.2	86.6	34.4	91.8	85
REPTree	41.9	94.7	87.1	38.2	96.2	89.99	28.4	97.2	89.01
Logistic	51.3	84	79.28	45.8	89.5	84.84	44.6	88.3	83.13
NaiveBayes	86.3	44.7	50.64	83.2	58.3	60.95	73	66.7	67.43

**Table 3.15:** Classifier results using only chemical property attributes unmodified in weightings

Classifier Algorithm	Accuracy (%)								
	Panel 44			Panel 331			Pharma Panel		
	Good Prof.	Bad Prof.	Overall	Good Prof.	Bad Prof.	Overall	Good Prof.	Bad Prof.	Overall
ZeroR (Baseline)	20.1	80.1	71.48	39.7	60	57.84	39.7	60	57.6
DecisionTable	79.9	82	81.73	77.7	80.6	80.29	79.9	80.7	80.62
JRip	77.8	80.4	80.07	77.3	83.5	82.84	76.6	81.5	80.9
PART	74.1	83.4	82.09	75.2	85.5	84.44	74.7	84	82.87
J48	67	85.3	82.65	61.5	88	85.13	62.1	86.4	83.49
RandomForest	50.3	93.5	87.37	44.8	95	89.6	45.5	93.8	88.04
RandomTree	36.4	90.2	82.52	31.7	92.5	86	31.3	91.5	84.35
REPTree	78.7	82.7	82.16	80.3	80.7	80.7	75.9	83	82.18
Logistic	49.4	86.6	81.31	45.6	89.3	84.64	45	88.2	83.1
NaiveBayes	86.5	44.6	50.57	83.6	58	60.7	73.5	66.5	67.34

**Table 3.16:** Classifier results using only chemical property attributes where class balancing was performed

Classifier Algorithm	Accuracy (%)								
	Panel 44			Panel 331			Pharma Panel		
	Good Prof.	Bad Prof.	Overall	Good Prof.	Bad Prof.	Overall	Good Prof.	Bad Prof.	Overall
ZeroR (Baseline)	0	100	85.7	0	100	89.3	0	100	88.13
DecisionTable	65.4	96.8	92.31	35.1	97.4	90.71	35.3	97.4	90.03
JRip	67.7	90.9	87.59	62.4	93.5	90.13	49.9	94	88.76
PART	61.1	93.7	89.03	57.1	96.3	92.1	54.5	96.1	91.19
J48	59.5	93.3	88.45	49.6	95.8	90.87	46.1	95.7	89.85
RandomForest	19.7	97.9	86.71	14.5	98.9	89.83	15.3	98.6	88.71
RandomTree	39.1	91	83.57	30.3	93.3	86.58	33.2	92.3	85.3
REPTree	59.5	94.5	89.53	52.7	96	91.32	48.6	95.5	89.98
Logistic	54.9	85.8	81.41	53.6	90.3	86.34	57	88.5	84.75
NaiveBayes	86.5	44.8	50.8	83.8	58.7	61.42	73.6	67.3	68.08

**Table 3.17:** Classifier results using chemical properties in conjunction with protein active flags where attributes were unmodified in weightings

Classifier Algorithm	Accuracy (%)								
	Panel 44			Panel 331			Pharma Panel		
	Good Prof.	Bad Prof.	Overall	Good Prof.	Bad Prof.	Overall	Good Prof.	Bad Prof.	Overall
ZeroR (Baseline)	20.1	80.1	71.49	39.7	60	57.84	39.7	60	57.6
DecisionTable	82.6	81.8	81.93	77.5	83.9	83.23	79.3	81.6	81.34
JRip	80.3	84.4	83.83	76.7	86.9	85.8	76	85.6	84.44
PART	79.9	87.5	86.38	72.5	90.8	88.87	69.8	92.3	89.66
J48	71.2	88.6	86.12	68.3	91.7	89.24	68.1	90.7	87.99
RandomForest	50.3	93.8	87.56	47.9	95.8	90.67	47.9	95	89.38
RandomTree	38	89.8	82.42	33.4	93.5	87.07	36.1	92.5	85.82
REPTree	80.1	85.4	84.68	78.1	87	86.07	81.8	83.9	83.66
Logistic	50.3	87.7	82.32	46.9	91.5	86.76	51.7	90.6	86
NaiveBayes	87	44.7	50.77	83.8	58.5	61.23	73.8	67.2	67.99

**Table 3.18:** Classifier results using chemical properties in conjunction with protein active flags where class balancing was performed

### 3.5.3 Discussion

The process of using chemical properties in conjunction with protein indicator flags has provided potentially promising models for consideration, but there are a few issues that needed to be resolved with regard to the chemical parsing in order for the approach to be considered feasible for processing potential candidates on a wider scale and automation of the process. The Mordred parser's code would likely require further investigation to highlight and potentially skip compounds which would prove to be problematic, however considering the very small impact the problematic compounds had on the overall number of compounds considered, a fix was not implemented.



## 3.6 Blind Testing

While testing was performed by cross validation on the datasets compiled, some of the test results generated made it necessary to incorporate an additional dataset to determine whether or not the protein interactions and chemical properties gathered thus far hold potential promise for a profiling model, or if further work was required in the processing of active and inactive protein pair. To accomplish this, compounds and interactions were gathered from two repositories, which were as follows:

- T3DB [29], to incorporate potentially harmful compounds which had not been considered by either ToxCast or DrugBank. These would be labelled as the test's "Bad" profile compounds.
- The Human Metabolome Database (HMDB) [30], a repository which contains protein interaction information of metabolites present within the human body and are thus not considered as harmful. These would be labelled as the test's "Good" profile compounds.

While T3DB compounds were available for access via the DrugReferenceDatabase, the HMDB repository needed to be processed and converted to equivalent PubChem compounds to be considered for both the protein and chemical property comparison. As a large number of metabolites were available from HMDB, test compound sets from this repository were generated and filtered to have a similar size in terms of molecular weight to the majority of compounds which had been trained. As a final step of processing the test set, any compounds which were found to be present within the training sets by virtue of being listed in more than one source were removed. Table 3.19 displays the number of compounds that were available for testing within each panel after processing and filtering had been performed.

Drug Class	Number of Compounds		
	Panel 44	Panel 331	Pharma Panel
Good Profile	272	389	797
Bad Profile	732	986	1,706

**Table 3.19:** Distribution of classes on the blind testing set

As the repositories used for testing had no information related to inactive results, any inactive results were gathered via the DrugReferenceDatabase through querying the relevant PubChem compounds that were present in other repositories. These results in conjunction with the active results and chemical property attribute test

sets could then be passed to WEKA, which provided the ability to easily assess classification accuracy on the trained models built in the previous tests.

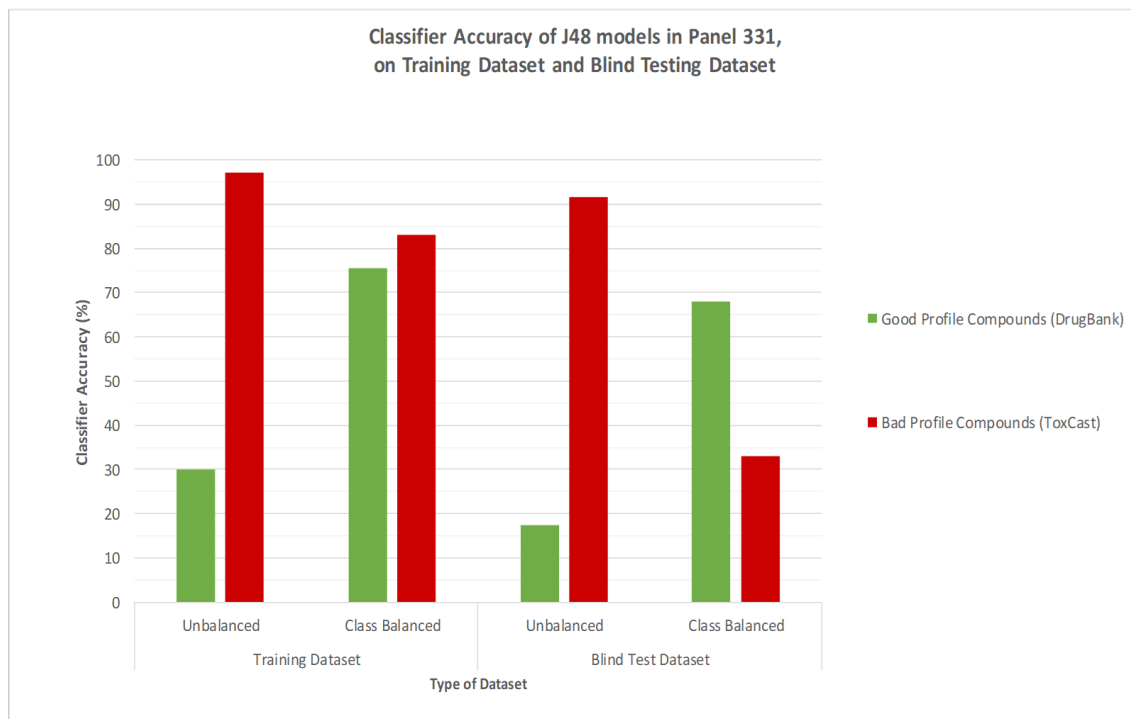
### 3.6.1 Classifier Results

Table 3.20 and 3.21 provide the classifier results of the active only classifier models against all panels studied. Accuracy levels from the blind tests were poor, with only a small proportion of the test Good Profile compounds being correctly classified. While the models generated through weight balancing improved the Good Profile accuracy level, the performance degradation on Bad Profile compounds was to a far greater extent than observed with the training data set. Figure 3.11 provides an example of the extent of the accuracy drop to the J48 classifier when applied to the training and blind testing set. These testing results, in conjunction with the results of the training data indicate that the interaction profiles of compounds in both classes were currently too similar in its current form to generate a clear distinction, despite the removal of conflicting compounds.

Table 3.22 and 3.23 provide the classifier results of models where inactive compound-protein results were separated from results which were considered to be inconclusive. Surprisingly it was found that the training model's rule sets were in a large disagreement with the test data. On further investigation of the distribution of the training and test sets, it was found that the distribution of inactive results were different between both sets; while the training set had a majority of inactive results placed within the "Bad" Profile group, a majority of inactive results in the test set were actually within the "Good" Profile group. This led to the models on the training set associating inactive results with "Bad" Profile compounds, which in turn resulted in the test set misclassifying to such a degree that inverse decisions were taking place. This further indicates that there is not enough interaction information present to generate a clear distinction between both classes of compounds. Balancing generated similar findings, but with a small improvement in Good Profile classification accuracy.

Tables 3.24 and 3.25 detail the classifier results of the compound property only experiment. While accuracy levels were still poorer than had been obtained from cross-validation with the training set, there was a smaller improvement in overall accuracy in comparison to the protein activity analyses. This highlights some promise in a distinction being present between both classes in terms of chemical properties. However, when these properties were used in conjunction with protein activity flags, the classification results shown in Tables 3.26 and 3.27 show that no significant differences in accuracy levels were detected. This is in contradiction to the findings

from the cross-validation analysis using the training set, which indicates that further work is likely needed in terms of refinement of the classifier rulesets on protein interactions.



**Figure 3.11:** Graph comparing results of the J48 classifier models when applied to the training set and blind testing set. Note that the actions taken by class balancing has generated a larger difference in class accuracy when the model had been applied to the blind testing set.

Classifier Algorithm	Accuracy (%)								
	Panel 44			Panel 331			Pharma Panel		
	Good Prof.	Bad Prof.	Overall	Good Prof.	Bad Prof.	Overall	Good Prof.	Bad Prof.	Overall
ZeroR	0	100	72.91	0	100	71.71	0	100	68.16
DecTable	14.7	91.4	70.62	21.1	96.2	74.98	14.1	97.3	70.8
Jrip	17.3	94	73.2	29.3	95.3	76.65	17.8	93.8	69.64
PART	13.6	83.2	64.34	34.4	88.4	73.16	28.7	83.5	66.04
J48	16.2	88.1	68.63	17.5	91.6	70.62	19.3	93.4	69.8
RandomForest	18.8	94.8	74.2	24.2	91.7	72.58	39.9	86.6	71.75
RandomTree	19.5	93.9	73.71	26.2	89.8	71.78	42.7	88.6	73.99
REPTree	17.3	94.3	73.41	25.4	91.5	72.8	17.2	87.2	64.92
Logistic	15.1	95.6	73.8	29.6	91.6	74.04	33.1	86.2	69.32
NaiveBayes	24.3	90.4	72.51	31.4	85.5	70.18	18.9	80.2	60.73

**Table 3.20:** Blind testing classifier results on active interactions only where attributes were unmodified in weightings

Classifier Algorithm	Accuracy (%)								
	Panel 44			Panel 331			Pharma Panel		
	Good Prof.	Bad Prof.	Overall	Good Prof.	Bad Prof.	Overall	Good Prof.	Bad Prof.	Overall
ZeroR	100	0	27.09	0	100	71.71	100	0	31.84
DecTable	82.4	25.1	40.64	80.7	20.4	37.45	86.6	13.5	36.8
Jrip	80.5	38.1	49.6	77.1	18.4	34.98	85.4	18.6	39.87
PART	70.6	45.9	52.59	58.9	37.5	43.56	55.6	74.5	68.48
J48	75.7	45.1	53.39	67.9	33.1	42.91	57.2	63.8	61.73
RandomForest	68	48.2	53.59	62.2	38.4	45.16	54.3	75	68.44
RandomTree	66.9	48.2	53.29	55.5	38.1	43.05	47.1	75.2	66.24
REPTree	75.4	39.1	48.9	76.3	16.3	33.31	82.3	36.3	50.98
Logistic	79.4	37.6	48.9	74.6	22.9	37.53	64.5	59	60.73
NaiveBayes	50.7	65.6	61.55	79.4	29.1	43.35	31.4	71.7	58.89

**Table 3.21:** Blind testing classifier results on active interactions only where class balancing was performed

Classifier Algorithm	Accuracy (%)								
	Panel 44			Panel 331			Pharma Panel		
	Good Prof.	Bad Prof.	Overall	Good Prof.	Bad Prof.	Overall	Good Prof.	Bad Prof.	Overall
ZeroR	0	100	72.91	0	100	71.71	0	100	68.15
DecTable	24.6	26.1	25.7	27	13.2	17.09	55.5	9.6	24.2
Jrip	27.6	26.6	26.89	30.6	12.1	17.31	58.1	7.7	23.73
PART	25	25.4	25.3	25.7	16.6	19.2	55.7	9.6	24.25
J48	29.4	31.6	30.98	27.2	17	19.93	54	11.3	24.85
RandomForest	24.6	28.3	27.29	26.5	16.1	19.05	53.6	11.8	25.13
RandomTree	33.5	26.9	28.69	30.3	16.1	20.15	55.8	21.8	32.64
REPTree	30.5	23.4	25.3	28.3	13.5	17.67	56.6	10.2	24.97
Logistic	30.5	30.7	30.68	34.7	14.7	20.36	58.1	17.7	30.56
NaiveBayes	64.3	35.9	43.63	47.3	8.9	19.78	65.7	5.8	24.89

**Table 3.22:** Blind testing classifier results when inactive results (including inconclusive results) are incorporated, and where attributes were unmodified in weightings

Classifier Algorithm	Accuracy (%)								
	Panel 44			Panel 331			Pharma Panel		
	Good Prof.	Bad Prof.	Overall	Good Prof.	Bad Prof.	Overall	Good Prof.	Bad Prof.	Overall
ZeroR	100	0	27.09	0	100	71.71	100	0	31.84
DecTable	34.9	10.2	16.93	30.1	11.4	16.65	58.3	7.9	23.93
Jrip	37.5	9.4	17.03	32.1	10.6	16.73	56	7.9	23.21
PART	32.7	11.7	17.43	27.8	13	17.16	56.3	8.1	23.49
J48	32	12.3	17.63	28.5	14	18.11	57.6	7.9	23.69
RandomForest	30.5	14.9	19.12	28.8	12.7	17.24	56.5	8.3	23.65
RandomTree	32.7	16.8	21.12	29	19.2	21.96	57.8	11.4	26.17
REPTree	31.3	12.7	17.73	29.6	12.9	17.6	57.5	7.7	23.57
Logistic	36.8	12.2	18.82	29.3	15.6	19.49	58	12	26.64
NaiveBayes	65.4	23.4	34.76	46.5	8.7	19.42	65.4	5.7	24.69

**Table 3.23:** Blind testing classifier results when inactive results (including inconclusive results) are incorporated, and class balancing was performed

Classifier Algorithm	Accuracy (%)								
	Panel 44			Panel 331			Pharma Panel		
	Good Prof.	Bad Prof.	Overall	Good Prof.	Bad Prof.	Overall	Good Prof.	Bad Prof.	Overall
ZeroR	0	100	72.91	0	100	71.71	0	100	68.16
DecTable	30.1	95.1	77.49	13.9	97.5	73.82	11.3	96.5	69.4
Jrip	31.3	93	76.29	29.8	91.9	74.33	21.3	92.5	69.84
PART	33.1	93.4	77.09	19	95.7	74.04	15.8	95.8	70.32
J48	35.7	89.9	75.2	21.9	92.5	72.51	15.7	95	69.72
RandomForest	16.5	97.5	75.6	9.8	98.7	73.53	5.3	98.5	68.84
RandomTree	37.9	85.9	72.91	20.6	93.1	72.58	18.3	91.6	68.28
REPTree	36.8	93.2	77.89	19	93.9	72.73	15.2	95.5	69.92
Logistic	40.4	74.3	65.14	36.5	86.6	72.44	30	84	66.8
NaiveBayes	86.4	66.3	71.71	79.4	76.3	77.16	56.2	80.7	72.91

**Table 3.24:** Blind testing classifier results using only chemical property attributes unmodified in weightings

Classifier Algorithm	Accuracy (%)								
	Panel 44			Panel 331			Pharma Panel		
	Good Prof.	Bad Prof.	Overall	Good Prof.	Bad Prof.	Overall	Good Prof.	Bad Prof.	Overall
ZeroR	0	100	72.91	100	0	28.29	100	0	31.84
DecTable	52.6	82.4	74.3	43.7	85.6	73.75	42.3	84.8	71.23
Jrip	62.5	74.2	71.02	48.3	84.8	74.47	41.3	73.7	63.36
PART	59.2	78.3	73.11	58.1	81.4	74.84	48.6	82.1	71.43
J48	54	79.2	72.41	42.4	87.8	74.98	38.1	81.9	68
RandomForest	37.1	91.4	76.69	30.1	92.3	74.69	18.1	93.1	69.2
RandomTree	34.9	91.5	76.2	30.8	87.9	71.78	20.5	91.6	68.96
REPTree	59.6	65.8	64.14	45	86.3	74.62	59.1	78.7	72.47
Logistic	38.2	72.8	63.45	36.2	86.8	72.51	30.4	84.2	67.08
NaiveBayes	86.4	66.3	71.71	79.4	76	76.95	57	80.2	72.79

**Table 3.25:** Blind testing classifier results using only chemical property attributes where class balancing was performed

Classifier Algorithm	Accuracy (%)								
	Panel 44			Panel 331			Pharma Panel		
	Good Prof.	Bad Prof.	Overall	Good Prof.	Bad Prof.	Overall	Good Prof.	Bad Prof.	Overall
ZeroR	0	100	72.91	0	100	71.71	0	100	68.16
DecTable	33.1	92.8	76.59	19	91.2	70.76	8.5	96.6	68.56
Jrip	41.5	86.6	74.4	36.2	86.2	72.07	19.7	91.7	68.76
PART	43.4	87.8	75.8	39.8	88.9	75.05	30.4	88.5	69.96
J48	49.3	85	75.3	33.7	90.6	74.47	18.7	90.5	67.64
RandomForest	16.9	97.7	75.8	10.3	98.6	73.6	4.1	98.3	68.32
RandomTree	32.4	90.4	74.7	28.5	91.6	73.75	17.6	94.4	69.92
REPTree	29.8	92.1	75.2	19.3	94.7	73.38	24	92.8	70.88
Logistic	46	71.2	64.34	49.1	60.3	57.16	37.4	67.4	57.81
NaiveBayes	86.4	66.3	71.71	79.9	76.4	77.38	56	80.3	72.55

**Table 3.26:** Blind testing classifier results using chemical properties in conjunction with protein active flags where attributes unmodified in weightings

Classifier Algorithm	Accuracy (%)								
	Panel 44			Panel 331			Pharma Panel		
	Good Prof.	Bad Prof.	Overall	Good Prof.	Bad Prof.	Overall	Good Prof.	Bad Prof.	Overall
ZeroR	0	100	72.91	100	0	28.29	100	0	31.84
DecTable	67.6	47.5	52.99	60.9	80.9	75.27	48.8	80	70.08
Jrip	69.5	72.8	71.91	63.5	51.7	55.05	47.6	80.4	69.92
PART	73.2	66.7	68.43	68.6	83.9	79.56	49.4	84.2	73.15
J48	60.3	74.3	70.52	56.3	82.8	75.27	53.1	80.9	72.07
RandomForest	37.5	89.9	75.7	30.1	91.5	74.11	19.3	92.4	69.12
RandomTree	37.1	89.5	75.3	20.3	90.6	70.69	16.1	92.6	68.24
REPTree	56.3	74.9	69.82	62	76.9	72.65	46.2	84.8	72.51
Logistic	46	79.9	70.71	52.7	78.8	71.42	39.1	77.3	65.16
NaiveBayes	86.4	65.8	71.41	79.9	76.2	77.24	56.7	80.1	72.67

**Table 3.27:** Blind testing classifier results using chemical properties in conjunction with protein active flags where class balancing was performed

### 3.6.2 Discussion

Overall, the model decisions on the blind test set provided some informative steps forward in terms of further work. As revealed in the initial findings of the protein activity sets, only a small number of proteins in the panel studies had been revealed to be highly flagged as active or inactive. These highly screened proteins in turn could heavily impact the outcome of the models, therefore further work would be needed to balance out the number of highly screened proteins while in turn not causing a significant impact on the scale of compounds being parsed for training and testing.

One solution to resolve this problem is to change the interpretation of an active "hit" and an inactive "miss" from the repositories where it is possible to do so. Currently a large amount of conflicting information is present in the DrugReferenceDatabase which has not been accounted for in these tests, and the process of solely obtaining links between compounds and proteins is now proven to be insufficient for a profiling tool. Further database revisions would be needed to attempt to gather assay result values and units where available, so that queries can be performed to extract different sets of definitions for active and inactive compound protein pairings. This could in theory reduce the number of conflicts which are currently present within PubChem BioAssay and ChEMBL, which in turn could provide a cleaner interaction matrix and overall higher classification accuracy levels. As these changes require an alteration to the database schema and the process of gathering the information from the repositories, the tasks of generating this database and the results from this process will be discussed fully in the next chapter.

## 3.7 Conclusions

In this chapter, the interactions and chemical properties obtained from various repositories were used to build classification models which attempted to differentiate between "Good" Profile (i.e FDA-like) and "Bad" Profile (i.e ToxCast) compounds. During the process of filtering the pharmacology protein panel, it had been found that a small number of proteins were being screened to far higher levels than the rest of the panel, which in turn impacted upon the models generated by the classifiers which focused on these highly screened proteins. While initial classification results based on cross-validation of the training sets provided some promise of a reasonable distinction between the two classes of compounds, results obtained on consideration of "unknown compounds" not included in the previous training sets revealed that no proper discrimination could be achieved at this stage.

While these results were somewhat disappointing, there are a number of avenues which can be explored to further refine the amount of information which is being used to build and evaluate these models, while incorporating as much information from the considered repositories as possible. One of these areas is that dosage was not taken into consideration against the broadness of DrugBank and ToxCast. For example, while an FDA-approved drug could be considered beneficial if taken at the correct dosage, higher levels of dosage could generate a harmful response and not be considered as beneficial. Furthermore, some compounds featured in ToxCast may not have caused harm by direct interaction with specific proteins but might have been assigned as toxins due to the interactions or effects of downstream metabolites. Other compounds are assigned as toxins due to cellular and tissue effects that cannot be attributed to interactions with proteins at all.

Assignments of interaction may also be further refined by consideration of assay types, as documented within ToxCast and discussed in Chapter 1 of the availability of specified biochemical and cell-based assays. Detailed examination of the sources and types of assay from all sources could provide further clues for compound classification, as well as provide cleaner output in reducing the amount of conflicts that were discovered by ensuring assays are compatible.

With implementation of these avenues, in conjunction with more extensive protein screening from future revisions of the repositories studied, this profiling approach could provide better integration of compound and protein properties, in turn generating a tool to assist in reliably predicting potential candidates for drug discovery and development and early detection of toxicity.

## 3.8 References

- [1] Mak, K.-K. and Pichika, M. R., “Artificial intelligence in drug development: Present status and future prospects,” *Drug discovery today*, 2018.
- [2] Ekins, S., “The next era: Deep learning in pharmaceutical research,” *Pharmaceutical research*, vol. 33, no. 11, pp. 2594–2603, 2016.
- [3] Anderson, D. P., “BOINC: A system for public-resource computing and storage,” in *Grid Computing, 2004. Proceedings. Fifth IEEE/ACM International Workshop on*, IEEE, 2004, pp. 4–10.

- [4] Funai, C., Tapparello, C., Ba, H., Karaoglu, B., and Heinzelman, W., “Extending volunteer computing through mobile ad hoc networking,” in *Global Communications Conference (GLOBECOM), 2014 IEEE*, IEEE, 2014, pp. 32–38.
- [5] Scheeder, C., Heigwer, F., and Boutros, M., “Machine learning and image-based profiling in drug discovery,” *Current opinion in systems biology*, 2018.
- [6] Jimenez-Carretero, D., Abrishami, V., Fernández-de-Manuel, L., Palacios, I., Quílez-Álvarez, A., Díez-Sánchez, A., Pozo, M. A. del, and Montoya, M. C., “Tox<sub>-</sub>(r) cnn: Deep learning-based nuclei profiling tool for drug toxicity screening,” *PLoS computational biology*, vol. 14, no. 11, e1006238, 2018.
- [7] Pu, L., Naderi, M., Liu, T., Wu, H.-C., Mukhopadhyay, S., and Brylinski, M., “E toxpred: A machine learning-based approach to estimate the toxicity of drug candidates,” *BMC Pharmacology and Toxicology*, vol. 20, no. 1, p. 2, 2019.
- [8] Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., and Thrun, S., “Dermatologist-level classification of skin cancer with deep neural networks,” *Nature*, vol. 542, no. 7639, p. 115, 2017.
- [9] Remez, N., Garcia-Serna, R., Vidal, D., and Mestres, J., “The in vitro pharmacological profile of drugs as a proxy indicator of potential in vivo organ toxicities,” *Chemical research in toxicology*, vol. 29, no. 4, pp. 637–648, 2016.
- [10] Chang, Y., Park, H., Yang, H.-J., Lee, S., Lee, K.-Y., Kim, T. S., Jung, J., and Shin, J.-M., “Cancer drug response profile scan (cdrscan): A deep learning model that predicts drug effectiveness from cancer genomic signature,” *Scientific reports*, vol. 8, no. 1, p. 8857, 2018.
- [11] Aljumah, A. A., Ahamad, M. G., and Siddiqui, M. K., “Application of data mining: Diabetes health care in young and old patients,” *Journal of King Saud University-Computer and Information Sciences*, vol. 25, no. 2, pp. 127–136, 2013.
- [12] Wu, H., Yang, S., Huang, Z., He, J., and Wang, X., “Type 2 diabetes mellitus prediction model based on data mining,” *Informatics in Medicine Unlocked*, vol. 10, pp. 100–107, 2018.
- [13] Rodríguez-Pérez, R., Miyao, T., Jasial, S., Vogt, M., and Bajorath, J., “Prediction of compound profiling matrices using machine learning,” *ACS Omega*, vol. 3, no. 4, pp. 4713–4723, 2018.



- [14] Vogt, M., Jasial, S., and Bajorath, J., “Extracting compound profiling matrices from screening data,” *ACS Omega*, vol. 3, no. 4, pp. 4706–4712, 2018.
- [15] Ding, H., Takigawa, I., Mamitsuka, H., and Zhu, S., “Similarity-based machine learning methods for predicting drug–target interactions: A brief review,” *Briefings in bioinformatics*, vol. 15, no. 5, pp. 734–747, 2013.
- [16] Chen, X., Yan, C. C., Zhang, X., Zhang, X., Dai, F., Yin, J., and Zhang, Y., “Drug–target interaction prediction: Databases, web servers and computational models,” *Briefings in bioinformatics*, vol. 17, no. 4, pp. 696–712, 2015.
- [17] Bowes, J., Brown, A. J., Hamon, J., Jarolimek, W., Sridhar, A., Waldron, G., and Whitebread, S., “Reducing safety-related drug attrition: The use of in vitro pharmacological profiling,” *Nature Reviews*, vol. 11, no. 12, pp. 909–922, Dec. 2012.
- [18] Sipes, N. S., Martin, M. T., Kothiya, P., Reif, D. M., Judson, R. S., Richard, A. M., Houck, K. A., Dix, D. J., Kavlock, R. J., and Knudsen, T. B., “Profiling 976 toxcast chemicals across 331 enzymatic and receptor signaling assays,” *Chemical research in toxicology*, vol. 26, no. 6, pp. 878–895, 2013.
- [19] Kohavi, R., “The power of decision tables,” in *European conference on machine learning*, Springer, 1995, pp. 174–189.
- [20] Cohen, W. W., “Fast effective rule induction,” in *Machine Learning Proceedings 1995*, Elsevier, 1995, pp. 115–123.
- [21] Frank, E. and Witten, I. H., “Generating accurate rule sets without global optimization,” in *Proceedings of the Fifteenth International Conference on Machine Learning*, Morgan Kaufmann Publishers Inc., 1998, pp. 144–151.
- [22] Quinlan, J. R., *C4.5: Programs for Machine Learning*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1993, ISBN: 1-55860-238-0.
- [23] Breiman, L., “Random forests,” *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.
- [24] Quinlan, J. R., “Simplifying decision trees,” *International journal of man-machine studies*, vol. 27, no. 3, pp. 221–234, 1987.
- [25] Le Cessie, S. and Van Houwelingen, J. C., “Ridge estimators in logistic regression,” *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, vol. 41, no. 1, pp. 191–201, 1992.

- [26] John, G. H. and Langley, P., “Estimating continuous distributions in bayesian classifiers,” in *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*, Morgan Kaufmann Publishers Inc., 1995, pp. 338–345.
- [27] R Core Team, *Foreign: Read data stored by 'minitab', 's', 'sas', 'spss', 'stata', 'systat', 'weka', 'dbase', ...* R package version 0.8-69, 2017. [Online]. Available: <https://CRAN.R-project.org/package=foreign>.
- [28] Moriwaki, H., Tian, Y.-S., Kawashita, N., and Takagi, T., “Mordred: A molecular descriptor calculator,” *Journal of cheminformatics*, vol. 10, no. 1, p. 4, 2018.
- [29] Wishart, D., Arndt, D., Pon, A., Sajed, T., Guo, A. C., Djoumbou, Y., Knox, C., Wilson, M., Liang, Y., Grant, J., *et al.*, “T3db: The toxic exposome database,” *Nucleic acids research*, vol. 43, no. D1, pp. D928–D934, 2014.
- [30] Wishart, D. S., Feunang, Y. D., Marcu, A., Guo, A. C., Liang, K., Vázquez-Fresno, R., Sajed, T., Johnson, D., Li, C., Karu, N., *et al.*, “Hmdb 4.0: The human metabolome database for 2018,” *Nucleic acids research*, vol. 46, no. D1, pp. D608–D617, 2017.

## Chapter 4

# Refinement of drug profiling approaches by incorporation of experimental assay results

### 4.1 Introduction

In the previous chapter, classifiers were assessed on their suitability as a potential tool for profiling promising candidate compounds, based on a mixture of documented interactions (without considering dosage), and compound properties. While initial results showed promise on cross-validation, the classifier models when assessed on different interaction repositories had shown that a clear differentiation between a potentially beneficial compound and a potentially toxic compound was not possible based on the information gathered. The purpose of this chapter is to describe a different approach to determine interactions through the consideration of assay concentration values and outcomes, and in turn compare the results found in the HMDB/T3DB repositories to investigate if narrowing the interaction searchspace would deliver performance improvements. The chapter first approached the compound concentrations and test types which the repositories shared in common, in addition to determining the values used to consider a protein compound pair as active. The classifiers were then retrained on the revised dataset and assessed against the test repositories, and their results compared with those of Chapter 3. Finally, the chapter concluded with a discussion of the design of a revised schema which could incorporate the customised thresholds efficiently.

## 4.2 Reassessing the DrugReferenceDatabase

In the iteration of the DrugReferenceDatabase discussed in Chapter 2, links between compounds and proteins were gathered from various repositories based on keywords, or from repository's documentation stating that a pair between a compound and protein was defined as active or inactive. As the repositories could have varying definitions of active and inactive results, an attempt was made to remove instances where pairs were classified as both active and inactive in order to deliver confident declarations of activities to the profiling classifiers. These attempts however had uncovered a large amount of conflicts in certain repositories, which in conjunction with poor accuracy levels of non-conflicting interactions meant that further work was required in order to reveal a clearer pattern between "Good" and "Bad" profile candidate drugs.

One avenue which was not explored in the DrugReferenceDatabase attempt was to consider the experimental result values which have been documented in certain studied repositories (BindingDB, ChEMBL, PubChem BioAssay, ToxCast). These parameters were not documented in the first iterations of the database in order to gather as many links between proteins and compounds as possible for the classifier, and due to potential compatibility issues with repositories such as DrugBank where no experimental results were documented, only target links. To ensure that as much relevant information as possible was gathered for a thresholding attempt, the repositories needed to be reassessed to determine which elements would be suitable for a revised design of the database which is focused on experimental parameters and their values. In addition to this, a review of the literature was also needed in order to determine what values and scales would be used to define active and inactive results for a compound-protein pairing. The benefit of such a process being undertaken would allow the consideration of pairings which did not contain a keyword, but instead possessed an experimental value which would be indicative of an active or inactive result. Once these steps were undertaken the protein-specific experiments described in Chapter 3 could then be reassessed to investigate whether or not these measures presented an improvement in profiling accuracy.

### 4.2.1 Potency Values

Beyond listing a link between a compound and a target, some repositories provide additional tabulated information on the conditions and results which took place during the experiment. One of these attributes which is commonly shared by the repositories

Expression of Potency	Measure
EC50/AC50	The molar concentration of an agonist that produces 50% of the maximal possible effect of that agonist.
IC50	The molar concentration of an antagonist that reduces the response to an agonist by 50%
Ki	The negative logarithm to the equilibrium dissociation constant of a compound determined directly in a binding assay using a labeled form of the compound
Kd	The negative logarithm to the equilibrium dissociation constant of a compound determined in inhibition studies

**Table 4.1:** Definitions of some common measures of potency [4]

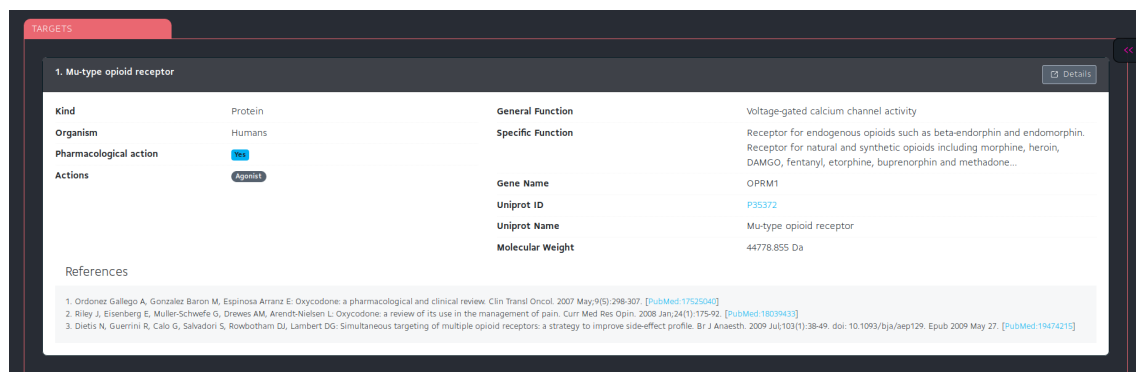
studied by this project is the scale of potency, which is defined by Waldman as an expression of the activity of a drug in terms of the concentration or amount of the drug required to produce a defined effect [1]; higher levels of potency would therefore require lower concentrations of a compound. Table 4.1 provides a summary of some typical experimental measures which are used to define certain expressions of potency, typically measured in mols (the amount of molecules of a compound).

In terms of the use of these concentration values to determine activity, a strong active pairing could be determined if the concentration result documented in an assay is very small. The exact concentration value to use to determine a compound protein pair as active can vary between applications: in one discussion by Parmentier *et al.*, a potent IC50 concentration is defined by concentrations under 1 micromolar, while marginal and weak potencies are defined as between 1 and 10 micromolar and above 10 micromolar respectively [2]. Another report by Hughes *et al.* also describes relevant potency concentrations, and how compound screening assays for hit discovery are typically run at concentrations between 1 to 10 micromolar [3]. As the aim of this chapter was to reduce the number of potentially weak cases of interactions being considered in the DrugReferenceDatabase, it was considered reasonable to classify assays with documented potencies of under 10 micromolar as representing a potential interaction between a compound and a protein. While this is a very low concentration to consider, all potency values would be documented so that variable thresholds for activity could be applied later if needed. This potency extraction process, coupled with some alterations to the data extraction process of certain repositories would lead to more confident identification of protein-compound interactions, and in turn to more focused and accurate classifier models.

## 4.2.2 Repository Parameter Review

### 4.2.2.1 DrugBank

In DrugBank's documentation, each protein-drug association is listed under one of four categories: targets, enzymes, transporters and carriers, with each documented link containing at least one reference. Figure 4.1 provides an example of a documented target for the drug Oxycodone (DrugBank reference DB00497), and the type of parameters which are available from the pairing. While the repository contains no experimental parameters to consider for thresholding, these targets were considered to be a reliable source for activity given that only a small number (202) of conflicts against inactive compound-protein pairs had been found, the majority of which (165) were present on the ChEMBL repository which contained experimental parameters and could be filtered to a further degree.



The screenshot shows the DrugBank interface for the target '1. Mu-type opioid receptor'. The page is titled 'TARGETS' and includes a 'Details' button. The main content is organized into several sections:

- Kind:** Protein
- Organism:** Humans
- Pharmacological action:** Yes (indicated by a blue 'Yes' button)
- Actions:** Agonist (indicated by a blue 'Agonist' button)
- General Function:** Voltage-gated calcium channel activity
- Specific Function:** Receptor for endogenous opioids such as beta-endorphin and endomorphin. Receptor for natural and synthetic opioids including morphine, heroin, DAMGO, fentanyl, etorphine, buprenorphin and methadone...
- Gene Name:** OPRM1
- Uniprot ID:** P35372
- Uniprot Name:** Mu-type opioid receptor
- Molecular Weight:** 44778.855 Da

References are listed at the bottom:

1. Ordonez Gallego A, Gonzalez Baron M, Espinosa Arranz E. Oxycodone: a pharmacological and clinical review. *Clin Transl Oncol*. 2007 May;9(5):298-307. [PubMed:17525040]
2. Riley J, Eisenberg E, Muller-Schwefe G, Drewes AM, Arendt-Nielsen L. Oxycodone: a review of its use in the management of pain. *Curr Med Res Opin*. 2008 Jan;24(1):175-92. [PubMed:18039433]
3. Diets N, Guerini R, Calo G, Salvadori S, Rowbotham DJ, Lambert DG. Simultaneous targeting of multiple opioid receptors: a strategy to improve side-effect profile. *Br J Anaesth*. 2009 Jul;103(1):38-49. doi: 10.1093/bja/aep129. Epub 2009 May 27. [PubMed:19474215]

**Figure 4.1:** Example of documentation of a target in DrugBank (Mu-type opioid receptor on the drug Oxycodone) [5]

### 4.2.2.2 ToxCast

In the ToxCast repository, interaction results were documented in a matrix environment, where compounds displayed an active, inactive or inconclusive/unknown value against a range of assays. In the initial approach when this repository was considered for the DrugReferenceDatabase, this matrix was parsed to consider a potential link between a compound and a protein, even if an assay that was flagged as '1' inside the matrix contained more than one protein in the assay. This was performed in order to gather as much protein information as possible from ToxCast. This process however had introduced some "noise" in the form of conflicts, in addition to some highly screened proteins sourced from ToxCast having a presence in assays of multiple

proteins. Figure 4.2 displays an example of the data files where one assay assesses a single protein, while another assesses multiple proteins.

In terms of experimental parameters for ToxCast, another matrix file is present in the dataset files which provides AC50 values. The values present in the matrix are measured in micromolar, where a value is documented between a compound and assay if it has been tested and defined as Active inside the hit matrix. In the event that a pair has been listed as inactive, the matrix provides a default constant value of 1 mol, which is represented as 1 million micromolar in the matrix. These results therefore indicate that in some cases, a label of activity from ToxCast may not necessarily reflect a strong interaction when compared with the other repositories. All *in vitro* assays with a single UniProt entry code were considered for assessment of a potential interaction with a compound.

	AM	BI	BJ	BK	BL	BM
1	assay_component_endpoint_name	intended_target_uniprot_accession_number	intended_target_organism_id	intended_target_track_status	technological_target_gene	technological_target_entrez_gene
5	ACEA_T47D_80hr_Positive	P03372		1 live	NA	NA
20	APR_HepG2_p53Act_1h_dn	P04637		1 live		329
21	APR_HepG2_p53Act_1h_up	P04637		1 live		329
40	APR_HepG2_p53Act_24h_dn	P04637		1 live		329
41	APR_HepG2_p53Act_24h_up	P04637		1 live		329
60	APR_HepG2_p53Act_72h_dn	P04637		1 live		329
61	APR_HepG2_p53Act_72h_up	P04637		1 live		329
64	ATG_Atr_CIS_up	P35869		1 live		21
66	ATG_AP_1_CIS_up	P05412 P01100	1 1	live live	183 134	3725 2353
68	ATG_AP_2_CIS_up	P05549 Q92481 Q726R9	1 1 1	live live live	314 315 440	7020 7021 83741
70	ATG_BRE_CIS_up	Q15797		1 live		196
72	ATG_C_EBP_CIS_up	P17676		1 live		58
76	ATG_CRE_CIS_up	O43889		1 live		366
78	ATG_DR4_LXR_CIS_up	Q13133 P55055	1 1	live live	363 332	10062 7376
80	ATG_DR5_CIS_up	P10276 P10826 P13631	1 1 1	live live live	279 280 281	5914 5915 5916
82	ATG_F_Box_CIS_up	P22415		1 live		333

**Figure 4.2:** Example of the UniProt targets associated with ToxCast assays, where some assays contain multiple proteins

#### 4.2.2.3 BindingDB

For the BindingDB repository, there is a list of binding affinity values for each compound and protein target. The assay values listed for each instance which were related to potency are  $K_i$ ,  $IC_{50}$ ,  $K_d$  and  $EC_{50}$ , and concentrations for each of these potencies were measured in nanomolar. Like ToxCast, BindingDB does contain compound results against multiple protein assays; to ensure that a direct link between a compound and protein could be made, any results gathered were filtered to single protein assays.

In addition to the assay values, BindingDB also provides a source field which specified from which area the repository had obtained the information. In most cases the source of the interaction had originated from US Patent documentation as per the project's goal and description, however there were a number of sources which

made reference to either ChEMBL or PubChem BioAssay entries, highlighting the potential for duplication of experimental results whilst querying for interactions.

#### 4.2.2.4 ChEMBL

In order to extract experimental parameters from ChEMBL, some alterations were required to the ChEMBL database query to expand on the interaction results already gathered. Figure 4.3 displays the altered query used to extract these parameters, where one of the main changes was to reference proteins directly through their ChEMBL code instead of the UniProt accession code (conversion of UniProt IDs to ChEMBL IDs was carried out via UniProt's cross reference platform). This reduced the database query's complexity by reducing the number of joins needed, while in turn ensuring that only single protein assays were selected. Assay results which were obtained from the query related to potency were AC50, EC50, IC50, Kd, Ki, and Potency, and all concentrations were listed in nanomolar. In some cases the experiment values listed relations other than equals (for example concentrations listed as more than or less than a certain value); to ensure the most accurate concentrations were gathered from ChEMBL, only precise measurements were considered for thresholding.

On further analysis of the results from this modified query, it was found that the majority of activity labels documented in ChEMBL from Chapter 2 were only related to interactions that were sourced from PubChem BioAssay entries. This would explain the cause of the removal of such a large number of ChEMBL records from Chapter 2 when these activity labels were considered, and that a large area of the ChEMBL searchspace remained unexplored by the classifiers.

```
SELECT m.chembl_id AS compound_chembl_id, s.canonical_smiles, r.compound_key,
NVL(TO_CHAR(d.pubmed_id),d.doi) AS pubmed_id_or_doi,
a.description AS assay_description, act.standard_type,
act.standard_relation, act.standard_value, act.standard_units,
act.activity_comment
FROM compound_structures s, molecule_dictionary m, compound_records r, docs d, activities act,
assays a, target_dictionary t
WHERE s.molregno (+) = m.molregno AND m.molregno = r.molregno
AND r.record_id = act.record_id AND r.doc_id = d.doc_id
AND act.assay_id = a.assay_id AND a.tid = t.tid
AND t.chembl_id IN ("CHEMBL2109242","CHEMBL2343",...)
```

**Figure 4.3:** Revised Query of the ChEMBL database to retrieve activity values from proteins of interest [6] [7]



#### 4.2.2.5 Repositories not considered

As the purpose of the chapter was to determine interactions which had the potential to be thresholded for generating new classifier models, there were repositories which were studied in Chapter 2 that were not considered in this chapter. These were CTDBase and Matador. In the case of CTDBase, compound-protein pairs were defined by a delimited list of keywords which specify the kind of interaction which takes place (such as "affects binding" or "increases activity"). However, as these recorded interactions did not comprise any experimental values it was difficult to determine which were likely candidates for strong interactions taking place. With the large number of keywords present in this dataset, attempts to extract thresholded interactions from these keywords would have been impractical for this analysis, and also had the potential to introduce a large degree of noise.

In the case of Matador, compound protein pairs are scored using two columns: these are defined as "protein\_score" which is a scale of confidence for the interaction, and "mesh\_score" which according to the documentation is scored according to Medical Subject Headings, where interactions derived from such sources receive lower scores. Matador then uses the maximum of these two scores to determine a "Matador Score". On further analysis of the values present it was decided not to include the Matador interactions as it was difficult to determine the calculations which form the basis of these scores. This, coupled with the lack of variety on the "protein\_score" field (where interactions were either scored as 95% or 0%) and lack of concentration-based measures meant that the interactions in Matador would have been difficult to filter to the other repositories.

Finally, while it was possible to obtain experimental values from PubChem BioAssay as was highlighted during the discussion of Chapter 2, a number of issues were present in the raw results which were difficult to troubleshoot without further parsing of the raw XML files. These issues were the lack of a BioAssay ID to determine the experimental conditions which took place between a compound and protein, and a large number of instances being classified as "PubChem Standard Value" where unit measurements and types were not specified. As the raw XML files would require a considerable degree of storage space and time to extract reliable instances of chemical properties, it was decided that BioAssay would not be parsed or considered usable for this analysis, and so instead as an alternative use was made of the values which had been provided from BindingDB and ChEMBL which had listed BioAssay as the source of interaction.

### 4.2.3 Initial Findings

After obtaining the experimental results from all repositories which possessed comparable values, the datasets were pre-processed to remove values which could be considered as not applicable to the analysis. These included potentially erroneous instances (molar concentrations documented as negative where no logarithm is defined), and instances where the molar concentration was greater than 1 mol (concentrations requiring a considerable amount of the compound tested to detect an interaction). Table 4.2 provides a summary of the number of interactions found from each repository after pre-processing. This table also includes the DrugBank repository where drugs classified as "Approved" are considered.

When compared with the summaries from Chapter 2, there have been decreases across all repositories. One area of particular note is the scale of activity results, where there has been a decrease of results from approximately 32 million to approximately 2 million. Another area of interest is the scale of reduction of compounds, where numbers have decreased from approximately 1.1 million to 600 thousand, likely caused by the removal of the BioAssay platform due to the difficulty in determining reliable assay results. Changes to the processing of ToxCast has also led to a considerable reduction in size from approximately 380 thousand results and 8,700 compounds, highlighting the high frequency of multi-protein assays in the dataset.

Table 4.3 provides a summary of the amount of active compound-protein results once the activity levels were filtered to under or equal to 10 micromolar. While ToxCast results had been reduced to a large degree from applying the threshold, both ChEMBL and BindingDB had a large proportion of results which met the thresholding condition, indicating that a reasonable searchspace existed. In terms of activities related to individual proteins, the large imbalance between the top scoring proteins and the rest of the panel has also been reduced. Tables 4.4, 4.5 and 4.6 feature the top 15 proteins in the panels studied, which reveal that the proteins are now more evenly distributed than before thresholding was performed. For example, in Panel 44 the protein accession codes P10275 and P04150 had 1,449 and 1,034 compounds respectively linked to these proteins as active. After thresholding and further pre-processing was performed, these counts were reduced to 209 and 112 compounds respectively.

Repository Source	Number of Results considered Active	Number of Compounds	Number of Proteins
ChEMBL	1,159,623	510,168	1,206
BindingDB	706,390	368,192	1,198
ToxCast	215,453	2,189	228
<i>DrugBank</i>	<i>9,480</i>	<i>1,407</i>	<i>1,469</i>
<i>Overall</i>	<i>2,090,946</i>	<i>604,173</i>	<i>1,868</i>

**Table 4.2:** Summary of all active results found before thresholding. DrugBank is listed in italics as a repository which does not have assay results but clearly specifies targets

Repository Source	Number of Results considered Active	Number of Compounds	Number of Proteins
ChEMBL	785,244	396,121	1,133
BindingDB	627,994	333,629	1,128
ToxCast	11,175	1,288	211
<i>DrugBank</i>	<i>9,480</i>	<i>1,407</i>	<i>1,469</i>
<i>Overall</i>	<i>1,433,843</i>	<i>482,345</i>	<i>1,847</i>

**Table 4.3:** Summary of all active results found after thresholding to assay concentration levels under or equal to 10 micromolar. DrugBank is listed in italics as a repository which does not have assay results but clearly specifies targets

Panel 44					
Good Profile Compounds (DrugBank)			Bad Profile Compounds (ToxCast)		
UniProt ID	Entry Name	No. of Interactions	UniProt ID	Entry Name	No. of Interactions
P08913	ADA2A_HUMAN	122	<b>P10275</b>	<b>ANDR_HUMAN</b>	<b>209</b>
<b>P28223</b>	<b>5HT2A_HUMAN</b>	<b>117</b>	<b>P04150</b>	<b>GCR_HUMAN</b>	<b>112</b>
P11229	ACM1_HUMAN	112	Q01959	SC6A3_HUMAN	104
P08172	ACM2_HUMAN	106	P23975	SC6A2_HUMAN	88
P35367	HRH1_HUMAN	104	<b>P35372</b>	<b>OPRM_HUMAN</b>	<b>57</b>
<b>P35348</b>	<b>ADA1A_HUMAN</b>	<b>102</b>	P08172	ACM2_HUMAN	52
P23975	SC6A2_HUMAN	99	<b>P21728</b>	<b>DRD1_HUMAN</b>	<b>44</b>
P14416	DRD2_HUMAN	97	<b>P22303</b>	<b>ACES_HUMAN</b>	<b>44</b>
P20309	ACM3_HUMAN	95	P11229	ACM1_HUMAN	41
<b>P41595</b>	<b>5HT2B_HUMAN</b>	<b>94</b>	P14416	DRD2_HUMAN	41
Q12809	KCNH2_HUMAN	92	P08913	ADA2A_HUMAN	40
<b>P31645</b>	<b>SC6A4_HUMAN</b>	<b>87</b>	P20309	ACM3_HUMAN	39
<b>P08908</b>	<b>5HT1A_HUMAN</b>	<b>81</b>	P35367	HRH1_HUMAN	32
Q01959	SC6A3_HUMAN	77	Q12809	KCNH2_HUMAN	31
<b>P07550</b>	<b>ADRB2_HUMAN</b>	<b>75</b>	<b>P29274</b>	<b>AA2AR_HUMAN</b>	<b>29</b>

**Table 4.4:** Top 15 interacting proteins in Panel 44 after thresholding was applied. Proteins which are not present in the opposite class' top 15 interactions are highlighted in bold font.

Panel 331					
Good Profile Compounds (DrugBank)			Bad Profile Compounds (ToxCast)		
UniProt ID	Entry Name	No. of Interactions	UniProt ID	Entry Name	No. of Interactions
P08684	CP3A4_HUMAN	196	<b>P03372</b>	<b>ESR1_HUMAN</b>	<b>349</b>
<b>P10635</b>	<b>CP2D6_HUMAN</b>	<b>193</b>	P33261	CP2CJ_HUMAN	277
P05177	CP1A2_HUMAN	142	<b>P10275</b>	<b>ANDR_HUMAN</b>	<b>209</b>
P33261	CP2CJ_HUMAN	141	<b>P03956</b>	<b>MMP1_HUMAN</b>	<b>190</b>
P11712	CP2C9_HUMAN	136	P11712	CP2C9_HUMAN	167
<b>P08913</b>	<b>ADA2A_HUMAN</b>	<b>122</b>	<b>O75469</b>	<b>NR1I2_HUMAN</b>	<b>147</b>
<b>P28223</b>	<b>5HT2A_HUMAN</b>	<b>117</b>	<b>P37231</b>	<b>PPARG_HUMAN</b>	<b>145</b>
<b>P11229</b>	<b>ACM1_HUMAN</b>	<b>112</b>	P05177	CP1A2_HUMAN	137
<b>P08172</b>	<b>ACM2_HUMAN</b>	<b>106</b>	<b>P11511</b>	<b>CP19A_HUMAN</b>	<b>136</b>
<b>P35367</b>	<b>HRH1_HUMAN</b>	<b>104</b>	<b>Q96RI1</b>	<b>NR1H4_HUMAN</b>	<b>121</b>
<b>P23975</b>	<b>SC6A2_HUMAN</b>	<b>99</b>	P08684	CP3A4_HUMAN	120
P20813	CP2B6_HUMAN	98	<b>P04150</b>	<b>GCR_HUMAN</b>	<b>112</b>
<b>P14416</b>	<b>DRD2_HUMAN</b>	<b>97</b>	<b>Q01959</b>	<b>SC6A3_HUMAN</b>	<b>104</b>
<b>Q12809</b>	<b>KCNH2_HUMAN</b>	<b>92</b>	P20813	CP2B6_HUMAN	91
<b>P18825</b>	<b>ADA2C_HUMAN</b>	<b>89</b>	<b>P35968</b>	<b>VGFR2_HUMAN</b>	<b>89</b>

**Table 4.5:** Top 15 interacting proteins in Panel 331 after thresholding was applied. Proteins which are not present in the opposite class' top 15 interactions are highlighted in bold font.

Pharmacology Panel					
Good Profile Compounds (DrugBank)			Bad Profile Compounds (ToxCast)		
UniProt ID	Entry Name	No. of Interactions	UniProt ID	Entry Name	No. of Interactions
P08684	CP3A4_HUMAN	196	<b>P03372</b>	<b>ESR1_HUMAN</b>	<b>349</b>
<b>P10635</b>	<b>CP2D6_HUMAN</b>	<b>193</b>	P33261	CP2CJ_HUMAN	277
P33261	CP2CJ_HUMAN	141	<b>P05121</b>	<b>PAI1_HUMAN</b>	<b>268</b>
P11712	CP2C9_HUMAN	136	<b>Q03405</b>	<b>UPAR_HUMAN</b>	<b>236</b>
<b>P08913</b>	<b>ADA2A_HUMAN</b>	<b>122</b>	<b>P13500</b>	<b>CCL2_HUMAN</b>	<b>222</b>
<b>P28223</b>	<b>5HT2A_HUMAN</b>	<b>117</b>	<b>P10275</b>	<b>ANDR_HUMAN</b>	<b>209</b>
<b>P08183</b>	<b>MDR1_HUMAN</b>	<b>114</b>	<b>P03956</b>	<b>MMP1_HUMAN</b>	<b>190</b>
<b>P11229</b>	<b>ACM1_HUMAN</b>	<b>112</b>	P11712	CP2C9_HUMAN	167
<b>P08172</b>	<b>ACM2_HUMAN</b>	<b>106</b>	<b>O75469</b>	<b>NR1I2_HUMAN</b>	<b>147</b>
<b>P35367</b>	<b>HRH1_HUMAN</b>	<b>104</b>	<b>P37231</b>	<b>PPARG_HUMAN</b>	<b>145</b>
<b>P35348</b>	<b>ADA1A_HUMAN</b>	<b>102</b>	<b>P11511</b>	<b>CP19A_HUMAN</b>	<b>136</b>
<b>P28335</b>	<b>5HT2C_HUMAN</b>	<b>101</b>	<b>P00750</b>	<b>TPA_HUMAN</b>	<b>124</b>
<b>P18089</b>	<b>ADA2B_HUMAN</b>	<b>100</b>	P08684	CP3A4_HUMAN	120
<b>P23975</b>	<b>SC6A2_HUMAN</b>	<b>99</b>	<b>P00749</b>	<b>UROK_HUMAN</b>	<b>115</b>
<b>P20813</b>	<b>CP2B6_HUMAN</b>	<b>98</b>	<b>P04150</b>	<b>GCR_HUMAN</b>	<b>112</b>

**Table 4.6:** Top 15 interacting proteins in the Pharmacology Panel after thresholding was applied. Proteins which are not present in the opposite class' top 15 interactions are highlighted in bold font.

## 4.3 Revised Analysis

With the interactions found thus far, the experiments which made use of protein activity flags could be reassessed to determine if an improvement in accuracy could be obtained from the test interactions found from HMDB and T3DB as the "Good" Profile and "Bad" Profile candidates respectively. The experiments selected for this chapter were the "Active Only" and "Proteins and Property", following the decision that data that did not meet the threshold would be grouped together with potential

inconclusive results. Interactions and compounds found in HMDB and T3DB were kept separate from the other repositories to ensure the training data had no influence on testing. Furthermore, the weight restriction on the HMDB repository was removed to observe the impact of the classifier performance on all metabolites which had interactions with the panels studied.

Table 4.7 provides a summary of the numbers of compounds used for both training and testing on both analysis sets across all panels studied. In terms of class distribution of the training set, the imbalance that was present in Chapter 3 has now been reduced, with the skew now towards towards "Good" profile compounds. While this imbalance is to a greater degree on testing sets now that all metabolites are included from HMDB, this would have no impact on the training sets used to build the classifier models.

The analysis process was to first train classifiers on the thresholded interactions from ToxCast, ChEMBL and BindingDB (including the targets from DrugBank approved drugs), which were then tested directly on the HMDB and T3DB compounds and targets. Results were presented in a similar format to Chapter 3, where unbalanced and balanced training sets (Using the ClassBalancer filter) would be assessed.

Panel	Number of Compounds			
	Training		Testing	
	Good Profile	Bad Profile	Good Profile	Bad Profile
Panel 44 (Active Only)	714	430	122	713
Panel 44 (With Chemical Properties)	714	430	122	707
Panel 331 (Active Only)	862	983	1,954	1,018
Panel 331 (With Chemical Properties)	862	983	1,954	1,016
Pharma Panel (Active Only)	1,224	1,071	18,562	1,851
Pharma Panel (With Chemical Properties)	1,224	1,071	18,562	1,843

**Table 4.7:** Distribution of compounds for the training and testing sets used for the thresholding analysis

### 4.3.1 Active Interactions Only

Table 4.8 displays the classifier results for the test set when no class balancing has been performed. Overall performance has been improved on two panels with thresholding implemented, with Panel 331 in particular demonstrating a considerable improvement in classification accuracy. Further investigation of this Panel's results highlighted that while performance for "Bad" profile drugs had decreased from levels of 95% to 70%, "Good" profile drugs had increased to degrees of almost 50% with some classifiers, indicating that some clear interaction patterns are being found in Panel 331. While the larger Pharmacology panel provided some improvement in overall

performance in comparison to Chapter 3's results, it was not to the extent of that for Panel 331, indicating the same findings as Chapter 3 - that the provision of additional information would not necessarily provide higher degrees of accuracy. Panel 44 was the only dataset which did not provide a significant improvement in performance, with most instances being classified as "Good" profile. Another feature of the test results was in that in all cases, high accuracy performance was found with the test metabolites regardless of weighting restrictions, demonstrating that the compounds found in HMDB shared similar interaction profiles to the FDA-approved DrugBank compounds.

Attempts at weight balancing the training set as shown in Table 4.9 had produced mixed results in comparison to the test result findings in Chapter 3. While some classifiers such as the DecisionTable had produced improvements in prediction of the minority "Bad" profile class across all panels, other classifiers such as Naive Bayes did not produce an accuracy improvement in comparison to the unbalanced training set models. Panel 331 continued to be the highest accuracy protein panel after balancing, with Panel 44 still providing the worst overall accuracy despite small improvements in "Bad" profile classification.

Classifier Algorithm	Accuracy (%)								
	Panel 44			Panel 331			Pharma Panel		
	Good Prof.	Bad Prof.	Overall	Good Prof.	Bad Prof.	Overall	Good Prof.	Bad Prof.	Overall
ZeroR	(0) 100	(100) 0	(72.91) 14.61	(0) 0	(100) 100	(71.71) 34.25	(0) 100	(100) 0	(68.16) 90.93
DecTable	(14.7) 96.7	(91.4) 19.2	(70.62) 30.54	(21.1) 77.2	(96.2) 70.7	(74.98) 74.97	(14.1) 99.9	(97.3) 45.9	(70.80) 95.02
Jrip	(17.3) 94.3	(94.0) 37.9	(73.20) 46.11	(29.3) 96.5	(95.3) 63.3	(76.65) 85.09	(17.8) 99.6	(93.8) 35.7	(69.64) 93.78
PART	(13.6) 94.3	(83.2) 26.9	(64.34) 36.77	(34.4) 76.6	(88.4) 72.1	(73.16) 75.03	(28.7) 99.5	(83.5) 38.5	(66.04) 93.97
J48	(16.2) 97.5	(88.1) 10.1	(68.63) 22.87	(17.5) 95.2	(91.6) 67.5	(70.62) 85.73	(19.3) 99.6	(93.4) 35.9	(69.80) 93.8
RandomForest	(18.8) 93.4	(94.8) 26.5	(74.2) 36.29	(24.2) 97.7	(91.7) 69.2	(72.58) 87.95	(39.9) 99.5	(86.6) 36.1	(71.75) 93.78
RandomTree	(19.5) 90.2	(93.9) 26.4	(73.71) 35.69	(26.2) 97.4	(89.8) 71.7	(71.78) 88.59	(42.7) 99.5	(88.6) 30.9	(73.99) 93.28
REPTree	(17.3) 92.6	(94.3) 27.3	(73.41) 36.89	(25.4) 95.8	(91.5) 72.3	(72.8) 87.72	(17.2) 99.5	(87.2) 35.8	(64.92) 93.74
Logistic	(15.1) 90.2	(95.6) 30	(73.80) 38.8	(29.6) 76.8	(91.6) 64.4	(74.04) 72.54	(33.1) 99.5	(86.2) 36.9	(69.32) 93.78
NaiveBayes	(24.3) 91.8	(90.4) 41.5	(72.51) 48.86	(31.4) 76.3	(85.5) 73.1	(70.18) 75.17	(18.9) 98.5	(80.2) 54.5	(60.73) 94.47

**Table 4.8:** Classifier results on interactions thresholded under 10 micromolar potencies of Ki, Kd, EC50, IC50, and AC50, unmodified in weightings. Results from the blind testing on Chapter 3 are shown in brackets

Classifier Algorithm	Accuracy (%)								
	Panel 44			Panel 331			Pharma Panel		
	Good Prof.	Bad Prof.	Overall	Good Prof.	Bad Prof.	Overall	Good Prof.	Bad Prof.	Overall
ZeroR	(100) 100	(0) 0	(27.09) 14.61	(0) 100	(100) 0	(71.71) 65.75	(100) 0	(0) 100	(31.84) 9.07
DecTable	(82.4) 92.6	(25.1) 37.9	(40.64) 45.87	(80.7) 97.4	(20.4) 69.4	(37.45) 87.82	(86.6) 99.7	(13.5) 49.9	(36.80) 95.21
Jrip	(80.5) 92.6	(38.1) 39.1	(49.60) 46.95	(77.1) 95.4	(18.4) 67.1	(34.98) 85.73	(85.4) 99.5	(18.6) 34	(39.87) 93.58
PART	(70.6) 88.5	(45.9) 42.8	(52.59) 49.46	(58.9) 77.9	(37.5) 72	(43.56) 75.91	(55.6) 99.5	(74.5) 37.2	(68.48) 93.82
J48	(75.7) 86.1	(45.1) 42.4	(53.39) 48.74	(67.9) 97.2	(33.1) 66.5	(42.91) 86.68	(57.2) 99.6	(63.8) 35.9	(61.73) 93.8
RandomForest	(68.0) 91.8	(48.2) 37.3	(53.59) 45.27	(62.2) 97.7	(38.4) 69.1	(45.16) 87.92	(54.3) 99.5	(75.0) 25.7	(68.44) 92.84
RandomTree	(66.9) 89.3	(48.2) 39.4	(53.29) 46.71	(55.5) 96	(38.1) 63.5	(43.05) 84.83	(47.1) 99.5	(75.2) 45.1	(66.24) 94.54
REPTree	(75.4) 93.4	(39.1) 38.6	(48.90) 46.59	(76.3) 96.9	(16.3) 70.3	(33.31) 87.82	(82.3) 99.5	(36.3) 37.8	(50.98) 93.88
Logistic	(79.4) 88.5	(37.6) 41.7	(48.90) 48.5	(74.6) 77.3	(22.9) 62.1	(37.53) 72.07	(64.5) 99.3	(59) 34.4	(60.73) 93.41
NaiveBayes	(50.7) 89.3	(65.6) 41.9	(61.55) 48.86	(79.4) 76.5	(29.1) 73	(43.35) 75.27	(31.4) 97.7	(71.7) 55.1	(58.89) 93.84

**Table 4.9:** Classifier results on interactions thresholded under 10 micromolar potencies of Ki, Kd, EC50, IC50, and AC50, where class balancing was performed. Results from the blind testing on Chapter 3 are shown in brackets

#### 4.3.1.1 Discussion

Although thresholding experimental assay values had considerably reduced the size of the data set being considered by the classifiers, the results where only protein interactions are considered has so far delivered potentially promising patterns, with higher accuracy levels being reported across most classifiers and protein panels. The modifications made have also generated models of a similar scale in terms of decisions and rules to those found in Chapter 3, which indicates that there is still sufficient information present after the threshold was applied to make a distinction between protein interaction profiles of compounds. For example, in the case of Panel 331 and the J48 tree classifier, the interaction link dataset had generated a tree of 65 decision leaves (areas leading to a classification of a compound instance), of which 129 attribute branches are present (all areas involving comparison of an attribute). The thresholded protein dataset had generated a similarly sized decision tree of 53 decision leaves, with 105 attribute branches.

While some classifiers reveal some interesting protein activity patterns and accuracy levels, there are some rulesets and decision trees which highlight that there could still be insufficient levels of coverage for protein-compound activity to come to sound conclusions. One classifier of interest found during the model assessment was the JRIP classifier on the unbalanced Panel 331 dataset: 5 rules were generated by this classifier, which only considered a mixture of activity flags for 9 out of the 144 proteins present in the panel and so was based on a highly restricted subset of proteins. This small ruleset had led to accuracy levels of 96.5% and 63.3% for "Good" and "Bad" profile test compounds respectively, and as the ruleset in Figure 4.4 shows, almost all of the tested compounds in HMDB would be satisfied being assessed by the interactions (or lack thereof) of these 9 proteins. The entry names attached to these proteins referenced by this ruleset in question are detailed in Table 4.10. While this may be considered effective and applicable based on the current information stored in the repositories, there is concern that more widely screened compounds would not be classified correctly with the use of these classifiers despite incorporation of thresholding.



```

JRIP rules:
=====
Rule 1) (P03372 = 0) and (P03956 = 0) and (P10635 = 1) and (P34969 = 0) and (P28223 = 1)
=> DrugClass=Good-Profile (43.0/0.0)
Rule 2) (P03372 = 0) and (P03956 = 0) and (P10275 = 0) and (P10635 = 1)
=> DrugClass=Good-Profile (176.0/29.0)
Rule 3) (P03372 = 0) and (P03956 = 0) and (P33261 = 0) and (P10275 = 0) and (P37231 = 0)
=> DrugClass=Good-Profile (835.0/266.0)
Rule 4) (P11509 = 1) => DrugClass=Good-Profile (10.0/0.0)
Rule 5) (Default if all rules fail) => DrugClass=Bad-Profile (781.0/93.0)

```

**Figure 4.4:** JRIP ruleset for Panel 331 on the thresholded and unbalanced dataset

UniProt Accession ID	Entry Code	Protein name
P03372	ESR1_HUMAN	Estrogen receptor
P03956	MMP1_HUMAN	Interstitial collagenase
P10635	CP2D6_HUMAN	Cytochrome P450 2D6
P34969	5HT7R_HUMAN	5-hydroxytryptamine receptor 7
P28223	5HT2A_HUMAN	5-hydroxytryptamine receptor 2A
P33261	CP2CJ_HUMAN	Cytochrome P450 2C19
P37231	PPARG_HUMAN	Peroxisome proliferator-activated receptor gamma
P11509	CP2A6_HUMAN	Cytochrome P450 2A6
P10275	ANDR_HUMAN	Androgen Receptor

**Table 4.10:** Details of proteins referenced by the JRIP ruleset

### 4.3.2 Active Interactions with Chemical Properties

Table 4.11 displays the results for the classifiers without class balancing implemented, when the thresholded interactions are combined with the chemical properties. The table also provides the test results obtained from Chapter 3, shown in brackets. From analysis of these results it appears as though the thresholding has had more of a positive impact when used in conjunction with the chemical properties, with high accuracy levels reported for most classifiers and panels. Panel 331 continues to provide the highest levels of improvement in comparison to other panels, while Panel 44 provides a greater degree of improvement when chemical properties are considered. A result of particular interest from all results was from the J48 decision tree and Panel 331, which reported a very high level of accuracy when assessed against the test compounds ("Good" profile accuracy of 97.7%, "Bad" profile accuracy of 85.5%). This is a considerable improvement on the chapter 3 "Good" profile accuracy rate of 33.7% from J48, but with a tradeoff of a decrease in "Bad" profile accuracy which was previously at 90.6%. Notwithstanding the slight fall in "Bad" profile accuracy, the attainment of an overall average classification accuracy of 93.5% is an exceptional outcome.

With class balancing implemented, the results shown in Table 4.12 also show an overall improvement in performance for most classifiers when compared to the Chapter 3 results. The act of class balancing however had a varied impact on

performance depending on the protein panel and classifier used. In one example with the Decision Table classifier, performance from Panel 44 and the Pharmacology Panel show worse performance overall compared to the findings of Chapter 3. In the case of Panel 331 however, overall performance improved by approximately 14% compared to the findings of Chapter 3 when class balancing was applied. This result is also an improvement over the unbalanced Panel 331 dataset, which performed poorly to the other panels when compared to the findings of Chapter 3. Another classifier of interest from further analysis of the balanced dataset include the RandomForest classifier, which reported high level of accuracy on the Pharmacology panel, with an overall performance improvement of approximately 26%. This classifier however had performed poorly in Panel 44 and Panel 331 when compared to the Chapter 3 results.

Classifier Algorithm	Accuracy (%)								
	Panel 44			Panel 331			Pharma Panel		
	Good Prof.	Bad Prof.	Overall	Good Prof.	Bad Prof.	Overall	Good Prof.	Bad Prof.	Overall
ZeroR	(0) 100	(100) 0	(72.91) 14.72	(0) 0	(100) 100	(71.71) 34.21	(0) 100	(100) 0	(68.16) 90.97
DecTable	(33.1) 69.7	(92.8) 39.3	(76.59) 43.79	(19) 10.6	(91.2) 89.9	(70.76) 37.74	(8.5) 62.1	(96.6) 46.6	(68.56) 60.74
Jrip	(41.5) 84.4	(86.6) 58.3	(74.4) 62.12	(36.2) 74.1	(86.2) 77.2	(72.07) 75.15	(19.7) 15.3	(91.7) 59.3	(68.76) 19.28
PART	(43.4) 70.5	(87.8) 58.8	(75.8) 60.55	(39.8) 78.5	(88.9) 77.9	(75.05) 78.28	(30.4) 94.6	(88.5) 64.7	(69.96) 91.94
J48	(49.3) 83.6	(85) 61.7	(75.3) 64.9	(33.7) 97.7	(90.6) 85.5	(74.47) 93.54	(18.7) 53.8	(90.5) 61.3	(67.64) 54.48
RandomForest	(16.9) 66.4	(97.7) 59	(75.8) 60.07	(10.3) 24.4	(98.6) 88.7	(73.6) 46.36	(4.1) 43.4	(98.3) 80	(68.32) 46.67
RandomTree	(32.4) 74.6	(90.4) 49.1	(74.7) 52.83	(28.5) 75.3	(91.6) 78.9	(73.75) 76.57	(17.6) 26.2	(94.4) 60.4	(69.92) 29.27
REPTree	(29.8) 75.4	(92.1) 49.1	(75.2) 52.96	(19.3) 92.7	(94.7) 81.5	(73.38) 88.89	(24) 38.3	(92.8) 65	(70.88) 40.67
Logistic	(46) 71.3	(71.2) 43.7	(64.34) 47.77	(49.1) 59.7	(60.3) 56.4	(57.16) 58.55	(37.4) 10.2	(67.4) 55.8	(57.81) 14.34
NaiveBayes	(86.4) 65.6	(66.3) 69	(71.71) 68.52	(79.9) 97.1	(76.4) 80.6	(77.38) 91.45	(56) 97.8	(80.3) 85.2	(72.55) 96.66

**Table 4.11:** Classifier results on interactions thresholded under 10 micromolar potencies of Ki, Kd, EC50, IC50, and AC50, in addition to chemical properties with no modifications to weights. Results from the blind testing on Chapter 3 are shown in brackets

Classifier Algorithm	Accuracy (%)								
	Panel 44			Panel 331			Pharma Panel		
	Good Prof.	Bad Prof.	Overall	Good Prof.	Bad Prof.	Overall	Good Prof.	Bad Prof.	Overall
ZeroR	(0) 100	(100) 0	(72.91) 14.72	(100) 100	(0) 0	(28.29) 65.79	(100) 0	(0) 100	(31.84) 9.03
DecTable	(67.6) 68	(47.5) 33	(52.99) 38.12	(60.9) 96.4	(80.9) 76.5	(75.27) 89.56	(48.8) 16.3	(80) 81.1	(70.08) 22.15
Jrip	(69.5) 78.7	(72.8) 57.1	(71.91) 60.31	(63.5) 59.6	(51.7) 79.3	(55.05) 66.33	(47.6) 91.2	(80.4) 82	(69.92) 90.4
PART	(73.2) 91.8	(66.7) 59.7	(68.43) 64.42	(68.6) 98.5	(83.9) 79.9	(79.56) 92.12	(49.4) 99.1	(84.2) 66.2	(73.15) 96.12
J48	(60.3) 91	(74.3) 52.8	(70.52) 58.38	(56.3) 98	(82.8) 82.9	(75.27) 92.83	(53.1) 55.7	(80.9) 57.5	(72.07) 55.9
RandomForest	(37.5) 61.5	(89.9) 70.2	(75.7) 68.88	(30.1) 20.1	(91.5) 87.3	(74.11) 43.09	(19.3) 97.1	(92.4) 81	(69.12) 95.65
RandomTree	(37.1) 57.4	(89.5) 55.9	(75.3) 56.09	(20.3) 49.9	(90.6) 78	(70.69) 59.53	(16.1) 96.9	(92.6) 68.6	(68.24) 94.33
REPTree	(56.3) 77	(74.9) 64.9	(69.82) 66.71	(62) 77.5	(76.9) 86.8	(72.65) 80.67	(46.2) 16.1	(84.8) 74.1	(72.51) 21.38
Logistic	(46) 72.1	(79.9) 47.7	(70.71) 51.27	(52.7) 75.1	(78.8) 58.5	(71.42) 69.43	(39.1) 21.8	(77.3) 54.4	(65.16) 24.75
NaiveBayes	(86.4) 65.6	(65.8) 69	(71.41) 68.52	(79.9) 97.1	(76.2) 80.6	(77.24) 91.45	(56.7) 97.8	(80.1) 85.1	(72.67) 96.65

**Table 4.12:** Classifier results on interactions thresholded under 10 micromolar potencies of Ki, Kd, EC50, IC50, and AC50, in addition to chemical properties, where class balancing was performed. Results from the blind testing on Chapter 3 are shown in brackets

### 4.3.2.1 Discussion

Whilst the previous investigation into using chemical properties in conjunction with protein interaction flags had led to somewhat disappointing results, the results generated when thresholding was applied has led to some interesting findings being made. Performance on classifying test compounds appears to vary according to particular classifiers and the panel assessed, and performance improvements are not necessarily connected to increases in search space size.

One area of interest on further assessment of the classifier models was that in the case of the JRIP rule based classifier, it appears as though the chemical property rules were only assigned despite the presence of protein activity flags, but these rulesets in turn have delivered high levels of accuracy when tested. Figure 4.5 displays an example of the JRIP ruleset on the unbalanced chemical and protein flag dataset for Panel 331. In comparison to the classifier results obtained from Chapter 3 (36.2% "Good" Profile accuracy, 86.2% "Bad" profile accuracy), the thresholded dataset had generated better performance for "Good" profile drugs (74.1% accuracy) with a small decrease in "Bad" profile accuracy (77.2% accuracy). While the additional compounds from Chapter 3 should in theory have generated a more accurate classifier, it appears as though the reduction in compounds from thresholding interactions has led to more distinct patterns being detected when chemical properties have been considered. While no proteins were present in this particular example's ruleset, it should be assumed that a candidate compound being tested should have at least one interaction recorded in the panel as per the training set requirements, in order to be considered for assessment by the model. Full details of the chemical properties specified in the ruleset can be found in Mordred's online documentation[8].

While some classifiers only made use of chemical properties, protein flags have been used in addition to chemical properties with some models which generated high levels of accuracy. One of these models is the J48 decision tree with the unbalanced Panel 331 dataset, where Figure 4.6 displays a segment of the decision tree in question. This tree had managed to successfully classify 97.7% of the "Good" profile test instances, and 85.5% of the "Bad" profile test instances, while containing a mixture of both protein flag and chemical property decisions in the tree. Table 4.13 highlights the proteins which were present in the complete sized tree, which contained 84 leaves and contained 167 decision points.

```
JRIP rules:
=====
Rule 1) (nBase >= 1) and (ETA_beta_ns >= 1.5)
=> DrugClass=Good-Profile (531.0/53.0)
Rule 2) (SpMAD_Dt >= 18.859751) and (P03372 = 0) and (ETA_dEpsilon_B <= 0.058233)
=> DrugClass=Good-Profile (118.0/12.0)
Rule 3) (MID_N >= 2.061687) and (SlogP_VSA2 >= 29.486992) and (Mm <= 0.676512)
=> DrugClass=Good-Profile (71.0/27.0)
Rule 4) (MID_N >= 3.797155) and (ZMIC2 <= 30.557898) and (MAXd0 >= 10.708321)
=> DrugClass=Good-Profile (79.0/17.0)
Rule 5) (IC4 >= 4.7213) and (MID_0 <= 5.481618) and (AATSOi <= 161.054881)
=> DrugClass=Good-Profile (54.0/16.0)
Rule 6) (TopoPSA(NO) >= 76.72) and (ZMIC3 <= 27.602677) and (ETA_epsilon_5 <= 0.801852)
=> DrugClass=Good-Profile (14.0/0.0)
Rule 7) (MINsssCH <= -0.488589) and (MATS4p <= -0.138851)
=> DrugClass=Good-Profile (15.0/4.0)
Rule 8) (AXp-0d <= 0.730514) and (ATSC2i >= -0.212614) and (AATSC4dv <= -0.580355)
=> DrugClass=Good-Profile (29.0/9.0)
Rule 9) (ATSC0s >= 82.873946) and (GATS7c <= 0.867391) and (MATS1i <= -0.215065)
=> DrugClass=Good-Profile (9.0/1.0)
Rule 10) (TIC3 >= 205.195441) and (BCUTc-1l >= -0.364677)
=> DrugClass=Good-Profile (17.0/6.0)
Rule 11) (AXp-6d <= 0.045192) and (MATS2s >= 0.101228) and (AXp-0d <= 0.726224)
=> DrugClass=Good-Profile (15.0/4.0)
Rule 12) (SlogP_VSA1 >= 5.316789) and (ATSC3d <= -11.252078) and (AATS6m <= 44.859391)
=> DrugClass=Good-Profile (11.0/4.0)
Rule 13) (SM1_Dzpe >= 1.357252) and (MID_h <= 11.275799) and (MATS1c <= -0.522564)
=> DrugClass=Good-Profile (15.0/6.0)
Rule 14) (TopoPSA >= 76.74) and (AATSC3c >= 0.004934) and (GATS5c <= 0.758399)
=> DrugClass=Good-Profile (9.0/3.0)
Rule 15) => DrugClass=Bad-Profile (858.0/37.0)
```

**Figure 4.5:** JRIP Ruleset for Panel 331 on the unbalanced protein and chemical property dataset

UniProt Accession ID	Entry Code	Protein name	Times referenced in tree
P11509	CP2A6.HUMAN	Cytochrome P450 2A6	6
P10275	ANDR.HUMAN	Androgen receptor	4
P10635	CP2D6.HUMAN	Cytochrome P450 2D6	4
O15111	IKKA.HUMAN	Inhibitor of nuclear factor kappa-B kinase subunit alpha	2
P03372	ESR1.HUMAN	Estrogen receptor	2
P03956	MMP1.HUMAN	Interstitial collagenase	2
P04798	CP1A1.HUMAN	Cytochrome P450 1A1	2
P05177	CP1A2.HUMAN	Cytochrome P450 1A2	2
P06401	PRGR.HUMAN	Progesterone receptor	2
P07948	LYN.HUMAN	Tyrosine-protein kinase Lyn	2
P08246	ELNE.HUMAN	Neutrophil elastase	2
P08253	MMP2.HUMAN	72 kDa type IV collagenase	2
P11511	CP19A.HUMAN	Aromatase	2
P11712	CP2C9.HUMAN	Cytochrome P450 2C9	2
P29074	PTN4.HUMAN	Tyrosine-protein phosphatase non-receptor type 4	2
P30536	TSPO.HUMAN	Translocator protein	2
P33261	CP2CJ.HUMAN	Cytochrome P450 2C19	2
P35968	VGFR2.HUMAN	Vascular endothelial growth factor receptor 2	2
P37231	PPARG.HUMAN	Peroxisome proliferator-activated receptor gamma	2
P45452	MMP13.HUMAN	Collagenase 3	2
P47898	5HT5A.HUMAN	5-hydroxytryptamine receptor 5A	2
P48730	KC1D.HUMAN	Casein kinase I isoform delta	2
P49137	MAPK2.HUMAN	MAP kinase-activated protein kinase 2	2
P49146	NPY2R.HUMAN	Neuropeptide Y receptor type 2	2
P51452	DUS3.HUMAN	Dual specificity protein phosphatase 3	2
Q07869	PPARA.HUMAN	Peroxisome proliferator-activated receptor alpha	2
Q14994	NR1I3.HUMAN	Nuclear receptor subfamily 1 group I member 3	2
Q96R11	NR1H4.HUMAN	Bile acid receptor	2
Q9HCS2	CP4FC.HUMAN	Cytochrome P450 4F12	2
Q9UBN7	HDAC6.HUMAN	Histone deacetylase 6	2

**Table 4.13:** Details of proteins referenced by the full J48 decision tree in Figure 4.6



## 4.4 Database Schema Redesign

Although the thresholding works on similar information to that used in Chapters 2 and 3, the process of documenting the assay experimental values and the revised definition of activity between a compound and protein meant that a database schema redesign was necessary in order to effectively store and query the information needed quickly and efficiently.

The schema defined in Chapter 2 made use of three main components for the defining the type of action which occurs between a compound and a protein: a hit table, which specified the repositories where an active classification was made; a miss table, which specified the repositories where an inactive classification was made; and finally a conflict table which specified pairs which were present in both tables for exclusion. As the definition of what threshold potency might constitute an active "hit" might vary from user to user, a database that considers assay experiment values would need to document all instances in one table. Figure 4.7 displays an example of a schema design prototype which could be used in order to store the information required for applying customised threshold levels. The table AssayResults reference directly to a UniProt protein accession code and PubChem compound ID, in addition to defining the repository which specified the result and the type and measurement method used. To ensure that some means of tracing the interaction to the source can be undertaken, cross reference database tables for drugs and proteins were needed to specify the IDs used in the other repositories.

While this database schema would be suitable for documenting the interactions discussed in this chapter, there are a number of areas for improvement which require further investigation before a database prototype could be considered for implement. One issue raised by reassessment of the repositories is that there may be instances where duplication may be present; whilst this issue was minimised somewhat with repository flags in the initial DrugReferenceDatabase design, the combination of all results into one table can lead to examples where multiple instances of the same compound protein pair experiment could be duplicated. Another area for improvement would be finding an efficient means of documenting the particular assay ID of a repository; whilst an addition of this to the schema would make the process of tracing the interaction to the source more efficient, there are some repositories which make no reference to a particular assay which in turn would make documentation complex. There are also instances which make no reference towards quantitative assay results, but would still be considered as relevant and potentially reliable sources of information,



such as the DrugBank approved target interactions. The altered schema would need to have the ability to incorporate and exclude these results efficiently when further repositories are considered. Finally, the schema should be able to incorporate and differentiate between what is considered as a single protein and multiple protein assay to ensure that users would be able to differentiate between different assay results when needed.

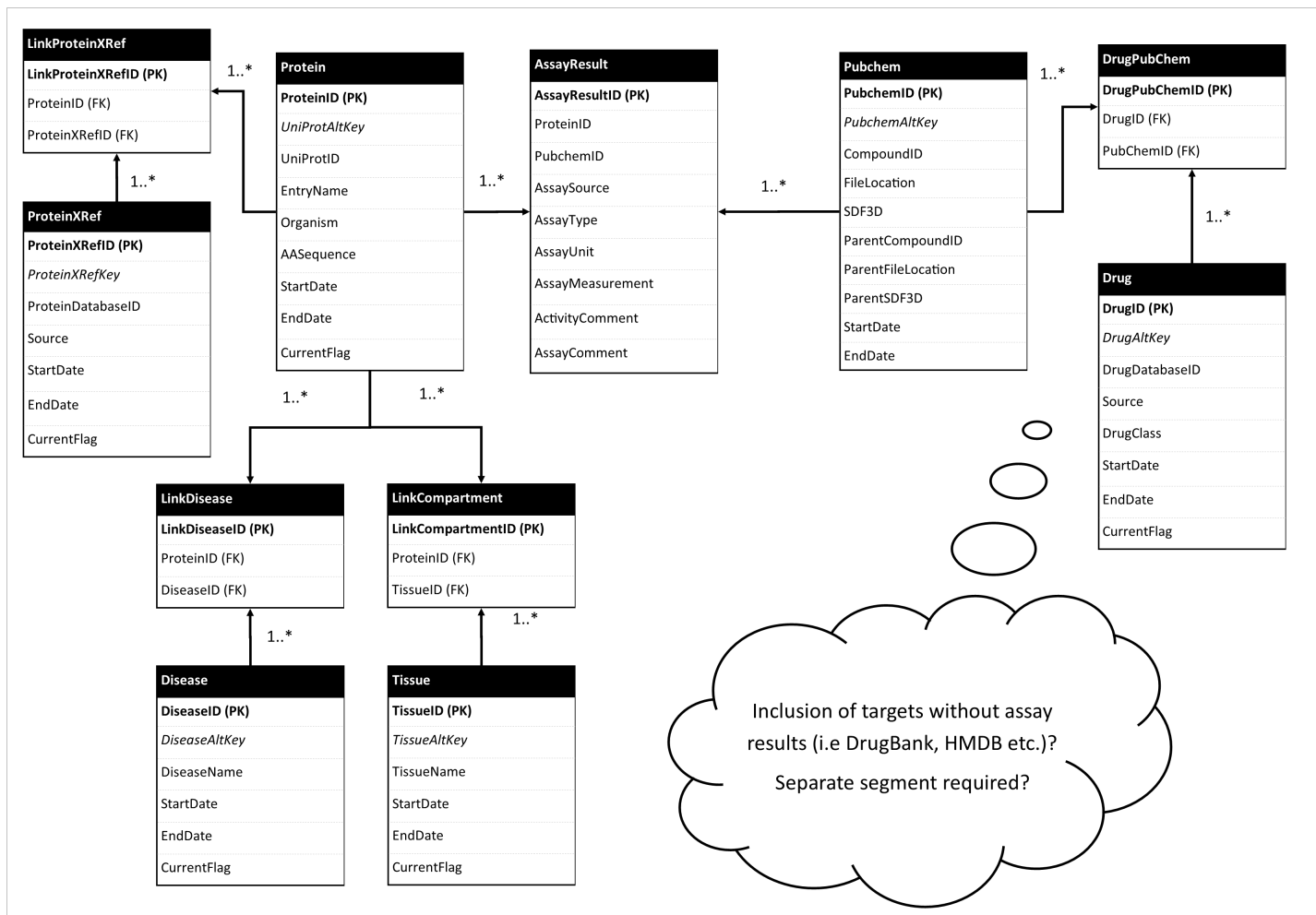


Figure 4.7: A revised Entity Relationship diagram of the DrugReferenceDatabase incorporating assay results to determine targets of interest

## 4.5 Conclusions and Further Work

Overall, applying quantitative assay experimental results and filtering the numbers have generated some substantial enhancements for the classifiers used for profiling. Despite these improvements however, there is still some further work which could be undertaken to improve the classifier's potential. One of these areas for instance would be further experimentation with differing concentration thresholds of Ki and IC50 etc. . The work detailed in this chapter only explored the potential of interaction being assigned when concentrations were under 10 micromolar, but further analyses could also be undertaken with differing values. These could include but are not limited to applying concentrations which could be considered as defining borderline activity, and also concentrations which could be considered as clearly inactive. Further observation should also have been applied to the descriptions and experimental conditions supplied from each repository to ensure that results provided from a particular repository were compatible with others.

Other areas for potential include looking beyond molar concentration values to assess levels of activity. On some of repositories considered such as BindingDB and ChEMBL there are a number of additional result parameters stored such as the temperature of the assay, or the pH level of the solution which might be suitable as additional parameters. Further considerations could also have been made with regard to dosage levels if documented by the assay, to further validate if a compound is correctly classified as beneficial or harmful.

## 4.6 References

- [1] Waldman, S. A., "Does potency predict clinical efficacy? illustration through an antihistamine model," *Annals of Allergy, Asthma & Immunology*, vol. 89, no. 1, pp. 7–12, 2002.
- [2] Parmentier, Y., Bossant, M.-J., Bertrand, M., and Walther, B., "5.10 - in vitro studies of drug metabolism," in *Comprehensive Medicinal Chemistry II*, Taylor, J. B. and Triggle, D. J., Eds., Oxford: Elsevier, 2007, pp. 231–257, ISBN: 978-0-08-045044-5.
- [3] Hughes, J. P., Rees, S., Kalindjian, S. B., and Philpott, K. L., "Principles of early drug discovery," *British journal of pharmacology*, vol. 162, no. 6, pp. 1239–1249, 2011.

- [4] Neubig, R. R., Spedding, M., Kenakin, T., and Christopoulos, A., “International union of pharmacology committee on receptor nomenclature and drug classification. xxxviii. update on terms and symbols in quantitative pharmacology,” *Pharmacological reviews*, vol. 55, no. 4, pp. 597–606, 2003.
- [5] Wishart, D. S., Knox, C., Guo, A. C., Sharavista, S., Hassanali, M., Stothard, P., Chang, Z., and Woolsey, J. (2019). “Oxycodone - drugbank,” [Online]. Available: <https://www.drugbank.ca/drugs/DB00497> (visited on 05/17/2019).
- [6] Gaulton, A., Hersey, A., Nowotka, M., Bento, A. P., Chambers, J., Mendez, D., Motow, P., Atkinson, F., Bellis, L. J., Cibrián-Uhalte, E., *et al.* (2019). “Schema questions and sql examples - chembl interface documentation,” [Online]. Available: <https://chembl.gitbook.io/chembl-interface-documentation/frequently-asked-questions/schema-questions-and-sql-examples#retrieve-compound-activity-details-for-a-target> (visited on 05/17/2019).
- [7] Davies, M., Nowotka, M., Papadatos, G., Dedman, N., Gaulton, A., Atkinson, F., Bellis, L., and Overington, J. P., “ChEMBL web services: Streamlining access to drug discovery data and utilities,” *Nucleic acids research*, vol. 43, no. W1, W612–W620, 2015.
- [8] Moriwaki, H., Tian, Y.-S., Kawashita, N., and Takagi, T. (2018). “Descriptor list — mordred 1.1.2a1 documentation,” [Online]. Available: <http://mordred-descriptor.github.io/documentation/master/descriptors.html>.

## Chapter 5

# Protein-compound interaction prediction based on compound and protein similarity

### 5.1 Introduction

In previous chapters, only database *in vitro* interactions were considered, where a repository specified either through target listings or through recorded assay results that activity existed between a compound and a protein. While the information provided by extracting interactions this way led to interesting protein interaction profiles being found, it was also appreciated that in most circumstances there was still a great deal of the related compound and protein search space for which knowledge of interactions is undocumented and inconclusive. The purpose of this chapter is to describe the consideration and implementation of some of the *in silico* techniques that could be used to provide knowledge of potential interactions. The first method in question involves the use of similarity measurement techniques with the rationale that a similar compound or protein should share a similar interaction profile if they were believed to be similar in structure. Another approach which will be used in conjunction with the similarity prediction technique is the application of a docking pipeline which can predict an interaction between a compound and a protein through comparing binding strengths and properties to those of known hits and misses. This method will be used to investigate high scoring pairs to determine if the similarity clustering methods holds merit.

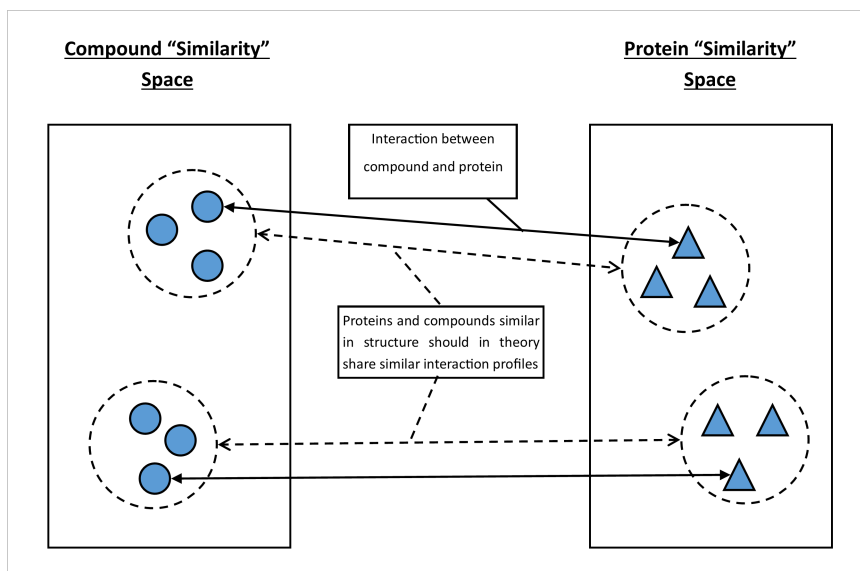
## 5.2 Similarity Comparison Review

When discussing chemical groupings, Cronin describes that similar objects have a tendency to share similar properties, and that when this method is applied to chemicals, a compound which has a presence within a group of well documented compounds could have its features and activities determined [1]. This term is defined as "Read-Across", in that toxicology, activity and properties could be read across in a selected group. One of the techniques used to accomplish this is via Quantitative Structure Activity Relationships, defined by Cronin as 'a mathematical model that relates (usually statistically) the activity or potency of a series of chemicals to physio-chemical properties or descriptors of chemicals'.

One study which made use of drug and protein similarities in such a fashion to determine potential interactions was reported by Yamanishi *et al.*, who designed and assessed a number of statistical methods to make predictions relating to four protein panels, split according to four types of protein: enzymes, ion channels, G-protein-coupled receptors (GPCRs) and nuclear receptors [2]. The report makes use of three main sets of information to make these predictions: the first being the chemical structure similarity of compounds, the second being the amino acid sequence similarity of proteins, and the third piece of information being the binary interaction matrix of proteins and compounds, where a value of 1 indicated an active compound-protein pairing and a value 0 was indicated all other cases. An illustration of this method in principle is shown in Figure 5.1, in that if a compound shares a high degree of similarity with another compound, there could be a high likelihood that their interaction profiles would be compatible. The same would also apply to proteins, where similarly sequenced structures could also in theory have a similar interaction profile.

### 5.2.1 Initial Approach

Following compilation of the similarities and interaction matrices, Yamanishi *et al.* assessed three different prediction techniques. The first and most naive of methods, the nearest profile method, made use of the assumption specified above, where the interaction profile of a new compound or protein would follow a proportion of the interaction profile of a compound or protein which was most similar. The second technique provides an expansion of the first technique by considering other nearest compounds, providing a weighted interaction profile. The interaction profile of a new compound could then be determined from the most similar compounds. The final



**Figure 5.1:** Predicting interaction profiles via similarity clustering. It could be assumed that proteins and compounds which are similar in structure should in general share similar interaction profiles.

method described was the bipartite graph, which attempted to detect interaction patterns between chemical and protein space to create a pharmacological space. This generates a matrix of confidence values to specify the likelihood of an interaction between a compound and a protein, and after application of a threshold provides the interaction profile for a new compound or protein. To assess these methods, the interaction sets of the 4 protein panels were split into training and test sets via 10-fold cross validation to determine if the methods had successfully predicted the correct interaction profile. To determine accuracy levels at various thresholds, receiver operating curves were used from which an average performance value could be determined by measuring the area under the curve (AUC).

On further analysis of the results outlined in Table 5.1, the bipartite graphs were revealed to have the highest performance of the methods, with the enzyme protein panel being able to successfully identify 57.4% of the activities tested when targets were defined as the top 1% of prediction scores. While other protein panels did not perform as well in predicting all the activities in the test set, those that it did predict when thresholded were mostly correct when assessing the positive predictive value (PPV).

Data	Method	AUC	Sensitivity	Specificity	PPV
Enzyme	Nearest profile	0.767	0.538	0.995	0.532
	Weighted profile	0.812	0.386	0.993	0.384
	Bipartite graph learning	0.904	0.574	0.995	0.570
Ion channel	Nearest profile	0.751	0.166	0.995	0.576
	Weighted profile	0.811	0.239	0.998	0.826
	Bipartite graph learning	0.851	0.271	0.999	0.936
GPCR	Nearest profile	0.729	0.156	0.994	0.474
	Weighted profile	0.739	0.146	0.994	0.444
	Bipartite graph learning	0.899	0.234	0.996	0.681
Nuclear receptor	Nearest profile	0.710	0.073	0.993	0.440
	Weighted profile	0.626	0.114	0.998	0.818
	Bipartite graph learning	0.843	0.148	0.999	0.954

**Table 5.1:** Performance of the methods tested by Yamanishi et al [2]

## 5.2.2 Alternative Similarity Clustering Methods

While the work of Yamanishi and colleagues is from 2008, there have been other reports which have made use of the dataset presented by Yamanishi *et al.*, either to test their own methods of target prediction or to provide alternative methods of calculating similarity of compounds and proteins. This section will provide an overview of some recent advances.

### 5.2.2.1 WNN-GIP

The Weighted Nearest Neighbour Gaussian Interaction Profile (WNN-GIP) method was presented in a report from 2013 by Van Laarhoven *et al.* [3]. This method combined two elements: the first element of a Gaussian Interaction Profile (GIP) constructs kernel matrices of the similarity and interaction matrices, which are then used with a new compound's interaction profile to provide a series of scores. The limitation of this feature used on its own is that a candidate drug would require at least one known interaction. This limitation could however be circumvented through the use of the weighted nearest neighbours technique (WNN), which would build an interaction profile based on similarly structured compounds and then pass this profile as input to the GIP kernels.

### 5.2.2.2 NetLapRLS

The NetLapRLS method was a technique proposed by Xia *et al.* in 2010, which composed of two elements [4]. The first and main element of the method was the Laplacian Regularized Least Square (LapRLS) technique, which creates one score matrix for drugs and one score matrix for proteins. These scores would then have a regularized term applied with their respective similarity matrices. The final step of this process would then be to average the separated score matrices to obtain a



final score matrix of predictions. The Net, or network element of the process is a modification of the parameters of this process, which generates a matrix that directly incorporates the known interaction matrix into the regularization function. The method in question has been described as computationally efficient as all scores can be calculated at the same time in comparison to methods which require multiple calculations or iterations.

### 5.2.2.3 BLM-NII

The BLM-NII method, as described by Mei *et al.* combined the bipartite graph learning model (BLM) of Yamanishi *et al.* with a method referred to as neighbour-based interaction profile inferring (NII) [5]. The rationale behind an expansion was to compensate for a weakness of the bipartite learning model's requirement for candidate drugs having some form of interaction profile, where candidates with no interaction profile would have weak predictions from a lack of information.

In principle, the BLM-NII routine behaves similarly to the WNN process, where neighbour interaction profiles are used as part of the final prediction, however there are two main differences. The first difference is that the interaction prediction profile is used to assist the labelling of interactions inside the interaction matrix before training the model, while the WNN method uses the prediction profile directly as final prediction scores. The second difference is that of efficiency, where the NII element is an extension of the BLM model and is only triggered by new drug and target candidates, whereas the WNN method is applied towards all drug and target candidates. While the authors state that their method had performed better overall to Yamanishi's method [2] and GIP [6], Van Laarhoven *et al.* had stated in his WNN-GIP study that the evaluation technique of selecting test compounds or proteins with only one interaction would produce a bias in results [3].

### 5.2.2.4 MSCMF

The multiple similarities collaborative matrix factorization model (MSCMF) is a technique that had been proposed by Zheng *et al.* in 2013, based on the premise that while many methods had been considered in the similarity prediction field, most had only allowed a single similarity matrix as a form of input [7]. This method improves upon this by allowing the use of more than one similarity matrix for drugs and proteins. These similarities are then weighted and projected into a low rank matrix to select the best similarity candidates as input for predicting targets. Predictions

are then calculated through the inner products of these low rank matrices. While this technique provides the ability for users to experiment with the combination of different kinds of similarity measurement, this is offset by the increased computation time needed to compile the similarities in addition to weighting and combining these matrices into predicted interactions. It is possible however to execute predictions via this algorithm with a single interaction matrix for both proteins and drugs.

#### **5.2.2.5 KBMF2K**

Another method which was studied with similarity interaction prediction was KBMF2k, which made use of a Bayesian formula in conjunction with dimensionality reduction and matrix factorization to make predictions in association with the similarity matrices [8]. Essentially the process creates a projection of a low dimensional pharmacological space of both drugs and proteins through a combination of two inputs: the first being a kernel matrix of the drug/protein similarity matrix, and the second being a parameter matrix of drugs and proteins. Once the low dimensional spaces have been compiled (one for drugs and one for proteins), these are then used as parameters to compile a final interaction matrix.

While this method was shown to have an overall improvement in target prediction in comparison to Yamanishi's method, it was noted to require a large computation time due to the calculations required to compile the protein/drug parameter matrices and the final prediction matrix. In addition to this, source code for the KBMF2k method required Matlab software which is a commercial application and was more restricted to those compiled via an open source language.

#### **5.2.2.6 SELF-BLM**

Although each study discussed so far has provided their own set of results which have shown promise in determining interactions from the Yamanishi dataset, the main limitation of most methods described is that they do not have the ability to consider inactive compound-protein pairs which can be present within the interaction searchspace. This presented a potential issue where new interactions could have a high likelihood of interaction but have experimental evidence (or a lack of it) indicating the opposite. While this issue could be circumvented with manual filtering of interactions classified as inactive, performance could potentially be improved if these elements were considered as input for the model's predictions. One such example of this technique being used is from a report in 2017 by Keum *et al.*, which presented a

technique known as SELF-BLM [9]. This technique clustered proteins and compounds into similar groups, and then assessed existing interactions; if all drugs inside a cluster did not interact with an individual protein or with the other proteins in its cluster, these drugs were labelled as negative for interaction with that particular protein.

### 5.2.2.7 NRLMF and PyDTI

While all of the discussed methods have provided promising findings, most have either not included source code to easily replicate the experiments, or have evaluated the methods under different evaluation conditions which made the process of assessment with customised sets difficult. However, one recent study in 2016 by Liu *et al.* presented a python library which combined some of the clustering methods assessed in a single python library [10]. Referred to as PyDTI, the library allowed users to run most of the reviewed techniques (WNN-GIP, BLMNII, NetlapRLS and CMF) in similar testing conditions. The KBMF2k method is also available via this library, but this requires a Matlab environment in order to function.

In addition to the provision of this library, Liu also included his own technique at target prediction with the library: Neighborhood Regularized Logistic Matrix Factorization (NRLMF). In summary, the NRLMF method functions similarly to KBMF2k, but attempts to assign higher importance and weighting on areas in the interaction matrix that are labelled as confirmed. This algorithm and library were expanded in another study to further improve on its prediction performance on compounds with little recorded activity levels [11].

### 5.2.3 Alternative Compound Comparison Methods

In addition to the variety of methods available, there have also been approaches which consider alternate approaches to compiling the similarity matrices. To determine the similarity of the compound on the Yamanishi et al dataset, a tool known as SIMCOMP was used which provided a similarity score based on substructures which were shared between two tested compounds. To determine protein similarity, amino acid sequences were compared by the Smith and Waterman algorithm [12], which provided high scores for matching amino acid strings while penalising mismatches in sequence and length. One alternative compound similarity method was assessed by Öztürk *et al.*, which proposed the use of comparing compounds via their SMILES codes, which presents a simplified structure of a compound in the form of a string of characters [13]. The report explores 13 avenues of comparison with the SMILES

Section	Bit Positions	Description
Hierarchic Element Counts	0 to 114	Test for the presence or count of individual chemical atoms represented by their atomic symbol
Rings in a canonic ESSSR ring set	115 to 262	Test for the presence or count of the Extended Smallest Set of Smallest Rings, which is a ring which does not share three consecutive atoms with any other ring in a chemical's structure
Simple atom pairs	263 to 326	Test for the presence of patterns of bonded atom pairs
Simple atom nearest neighbours	327 to 415	Test for the presence of atom nearest neighbour patterns regardless of bond order or count, but where bond aromaticity is significant
Detailed atom neighbourhoods	416 to 459	Test for the presence of detailed atom neighbourhood patterns regardless of count, but where bond orders are specific, and the bond aromaticity matches both single and double bond
Simple SMARTS patterns	460 to 712	Test for the presence of simple SMARTS patterns
Complex SMARTS patterns	713 to 880	Test for the presence of complex SMARTS patterns

**Table 5.2:** Description of the elements used in the PubChem Fingerprint System

string, of which some simple examples include edit distance (comparing the amount of edit operations needed to have one compound match another), the longest common subsequence (finding the longest SMILES element which is compatible with both compounds) and fingerprinting (the setup of binary flags for common patterns and comparing the flags triggered between two compounds). When these comparison methods were explored by the WNN-GIP algorithm, Öztürk *et al.* found equivalent performance with the SIMCOMP compound comparison, but noted that the use of SMILES comparison techniques was more efficient in computation time. The supplementary information of the report includes Java source code to compile the similarity matrices.

In terms of fingerprinting, there are various types available each containing different features and lengths. One example of a fingerprint format is the Pubchem Fingerprint [14], which contains 881 flags represented as binary bits. The PubChem system splits these bits into 7 sections, which are detailed in Table 5.2. Another example of a fingerprinting system is the MACCS format, which is a shorter fingerprint format of 166 bits and mainly detects interesting chemical features in a compound's substructural patterns (SMARTS). To calculate the similarity between two compound fingerprints, the Tanimoto coefficient is typically used.

#### 5.2.4 Alternative Protein Comparison Methods

For comparing the similarity of proteins, all methods make use of the Smith and Waterman algorithm, which is a process developed in 1981 to perform local sequence alignment [12]. The motivation behind the algorithm was an alternative to comparing the entire protein sequence, instead focusing on the comparison of regions of varying

length to detect regions of the protein which are similar. This is accomplished in several stages. The first stage initializes a substitution matrix, which typically provides scores on matching amino acid sequences and penalties for mismatching amino acids (though in some cases a reduced score may be given if the amino acid is similar in nature). The scoring matrix can either be user defined, or make use of typically used substitution scores such as the Blocks Substitution Matrix (BLOSUM) which can be altered depending on the expected scale of similarity present within the dataset (BLOSUM62 for instance is used for scoring proteins which could typically have less than 62% similarity). In addition to substitution matrix scoring, a gap penalty scheme is defined which determines the penalty of a score when a segment comparison requires opening or extending gaps. Once these scoring values have been defined, a scoring matrix is constructed where the dimensions are the length of the two protein sequences. The matrix is then traversed from left to right, top to bottom to compare amino acid sequences and considering outcomes of substitutions or adding gaps. The highest score occupies the scoring matrix (defaulting to 0 if only negative values are present), and the source of this score is specified so a traceback can be performed of a similar region which is shared between two amino acid strings. The final score is then determined from the highest score detected in the matrix.

To convert the scores into similarities, equation 5.1 is followed, where  $\text{Sim}(A,B)$  is the similarity measurement between proteins A and B, and where SW represents the Smith and Waterman algorithm score output. While this method is effective at detecting similar regions on comparisons of variable length proteins, the procedure of calculating similarities can be computationally expensive on either large collections of proteins or proteins which have a long amino acid sequence.

$$\text{Sim}(A, B) = \frac{SW(A, B)}{\sqrt{SW(A, A)}\sqrt{SW(B, B)}} \quad (5.1)$$

An alternative process which can be used to measure similarity is the Basic Local Alignment Search Tool (BLAST) [15]. This open source heuristic algorithm is developed by NCBI and returns the most similar matches from a protein database provided by a user. This in turn generates a partially filled similarity matrix of high scoring matches. Compared to the Smith and Waterman algorithm, BLAST is considered to be more efficient in terms of computation time and resource usage, however at a cost of thoroughness as BLAST will not document or score patterns which are difficult to detect. The algorithm is available for use either via the web platform, or through installation of a program which can be called via command line.

## 5.3 In Silico Docking Pipeline

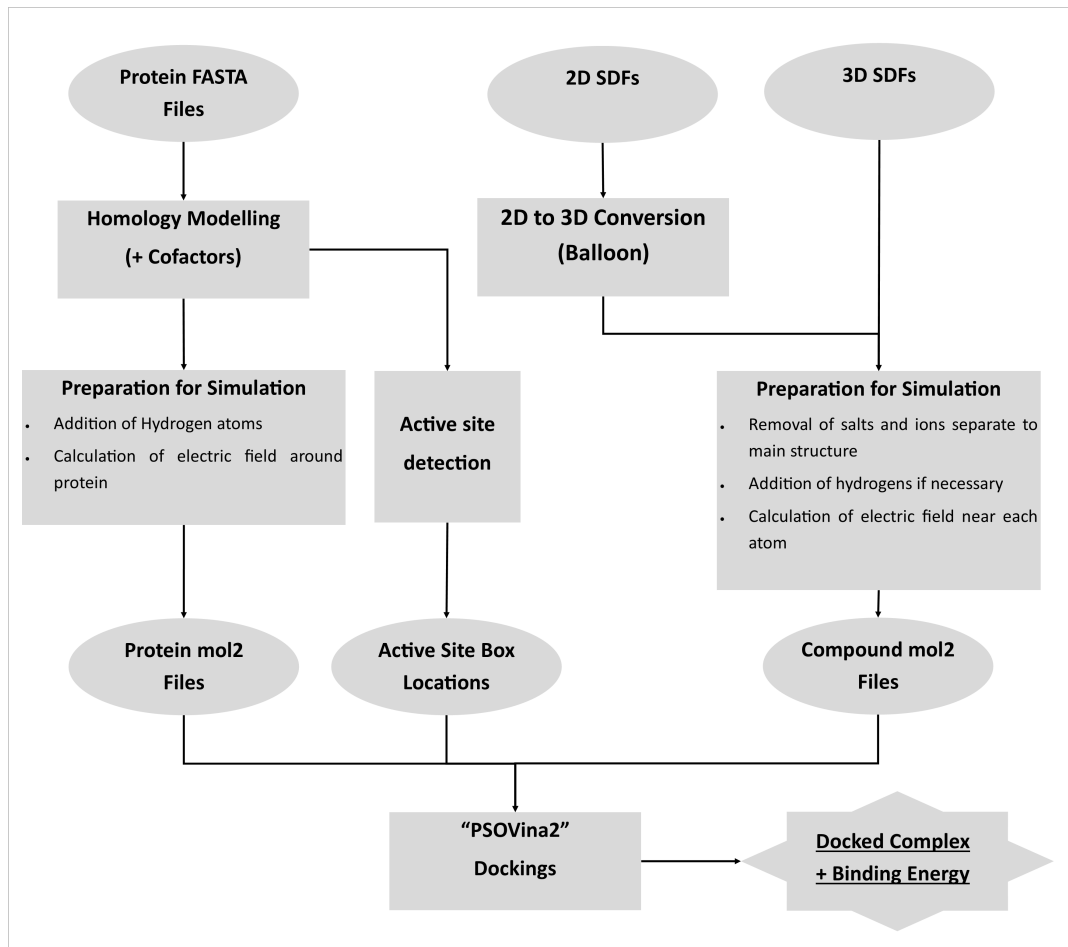
Another method of determining inconclusive activities between compounds and proteins is to make use of molecular docking pipelines to detect if a predicted binding between a protein and compound is similar in properties to those which are considered to be active. One process developed by Austin-Muttitt parsed through a selection of compounds and proteins through a selection of docking programs [16]. This process would first be run on compound-protein pairs with a known activity in order to build a profile of outputs which could be most associated with an active classification. Unknown pairings could then be run and compared with these reference profiles to generate a prediction of likelihood of activity. A summary of the *in silico* process used is shown in Figure 5.2.

In order for the pipeline to function, the process would require a protein in FASTA format, and a compound structure (in either a 3D or 2D format). These inputs would then follow a pre-processing stage which would provide the correct format for docking operations to be performed. For proteins, homology modelling is performed which attempts a conversion of the amino acid sequence into a 3D structure based on similar structures in the Protein Data Bank (PDB). The co-factor process attached to homology modelling attempts to either search for cofactor or prosthetic group molecules (for example, haems and flavins) if they exist within the protein, or attempts to search for and generate cofactors if they do not exist through querying similar protein structures. This is followed by other preparation operations to ensure that all necessary inputs are present for the docking algorithm, and the identification of sites in the protein which would be ideal attachment sites for compounds.

In terms of compounds, ideally the structures should be in a 3D format. This is not always possible however, either due to the complexity of the compound or the protein in question or due to the absence of a model. There is a process available however which attempts to transform 2D structures into 3D space, however the results in turn may not be as reliable as validated 3D structures. Both 2D and 3D structures must also follow additional preparation steps, such as the removal of ions and structures which are separate from the main compound structure to reduce any scope for error at the docking stage.

Once the structures have been pre-processed and the active sites identified, these are then used as inputs for the PSOVina2 program, which generates a combined protein-compound structure, and predicts the binding energies of the complex. These values can then be compared with other binding energies to determine if the results

generated are indicative of an active or inactive compound-protein pairing.



**Figure 5.2:** High level summary of the *in silico* pipeline [16]. Protein and compound structures are pre-processed into a suitable format for docking operations. Ideal binding sites are also identified on the protein. Once pre-processing is complete, PSOVina2 generates a combined protein-compound structure, and predicts the binding energy of the complex.

## 5.4 Methodology

With all methods that used similarity comparison methods making use of the Yamanishi dataset, the purpose of the work outlined in this chapter was to assess how the similarly documented techniques would perform on customised datasets. The intention of this was to assess the performance and present potential new targets from the three protein panels assessed in previous chapters. A three stage process was considered: the first stage being the investigation of Area under the Receiver Operating Characteristic Curve (AUC) and the Area under the Precision Recall Curve (AUPR), using the protein panels assessed during the profiling tool experiments. Typically for this type of analysis where there is only a small number of instances available for the class of interest (active compound-protein pairings), the AUPR would be considered the most appropriate measure of performance; Van Laarhoven *et al.* state that in comparison to AUC, the AUPR metric punishes the presence of false positives, which would be considered costly considering the lack of true interactions present in a typical compound-protein interaction search space. In addition to the Panels used in the profiling chapters, a protein set which combines the proteins from all panels was also considered. This combined protein panel was to ensure that all protein similarities were considered in a single environment. 10-fold cross validation was performed for testing these methods, where random drug-target pairings (approximately 10% of the interaction matrix) are labelled as inconclusive in the interaction matrix to assess the scores that would be generated. Table 5.3 provides a summary of the number of compounds, proteins and interactions present from each protein panel for this analysis.

Panel	Number of Active Results	Number of Compounds	Number of Proteins	Total Searchspace	Proportion of Targets (%)
Panel44	3,482	1,144	46	52,624	6.617
Panel331	8,203	1,845	142	261,990	3.131
PharmaPanel	13,552	2,295	571	1,310,445	1.034
All	14,899	2,338	636	1,486,968	1.001

**Table 5.3:** The statistics of the thresholded interaction datasets used for the clustering analyses

In addition to assessing the difference in performance between panels, different methods of calculating similarities of proteins and drugs were also considered. In terms of compound similarity, the algorithms will be assessed on the PubChem and MACCS fingerprint formats measured on the Tanimoto coefficient. Whilst the SIMCOMP method used in the Yamanishi studies is the main similarity measurement method used in previous studies, the technique is only accessible for compounds



which contain a KEGG database reference. The use of fingerprinting would therefore ensure as many compounds as possible would be considered for predictions.

In terms of protein similarity, the algorithms will be assessed using the Smith and Waterman algorithm and BLAST algorithm. As no parameters were present for gap opening penalties or the substitution method used for the Yamanishi similarity matrix, protein similarity matrices were compiled with parameters which were deemed suitable for a majority of comparisons on EMBOSS Water [17], which is a platform which calculates individual pairwise similarities and makes use of the Smith and Waterman algorithm. These default values were BLOSUM62 for the substitution matrix, and gap open and gap extension penalties of 10 and 0.5 respectively. For BLAST protein similarity queries, default values were applied and the percentage identities were used to compile a matrix of similarities. In the event that there was more than one entry for scoring from BLAST (caused by the splitting of large protein sequences of for multiple comparisons), the final similarity score would be the average of all entries. To prevent potential division by zero errors on the BLAST similarity matrix, a small default value (0.01) was assigned to protein comparisons which did not return a result via BLAST. This was typically in keeping with the similarity values returned for dissimilar proteins calculated via the Smith and Waterman method.

The second phase of the analysis considers the top 1,000 interaction predictions on the panels and methods which were shown to be the most promising in predicting targets in terms of AUC and AUPR performance metrics. As the methods available via the PyDTI library do not consider inactive compound-protein pairs in their predictions, any prediction made will be compared against known inactive pairs to assess their overall presence in the predictions. For the purpose of this analysis, inactive compound-protein pairs would be assay potencies which were documented above 10 micromolar.

The final stage of the process is to use the highest scoring new interactions documented through docking simulations to determine whether the newly predicted interactions hold any promise. Whilst other studies have verified new interactions from future versions of repositories, there were either no updates or an insufficient amount of additional data on the repositories assessed to assist with the verification process here. The *in silico* pipeline on the other hand would consider all of the inconclusive interactions which were not considered inactive, and provide a better overview of potentially promising targets.

## 5.5 Results

### 5.5.1 AUC and AUPR Performance

Tables 5.4 and 5.5 provide the AUC values of the various combinations of panels, prediction methods and similarity measurement techniques. The first finding from these results is that in terms of similarity measurement, there appears to be no significant difference in performance measurements, which indicates that the quicker and more efficient similarity measurement techniques could be just as effective as the more detailed comparison methods. Secondly, the number of proteins within a panel appears to have some degree of impact on performance, depending on the clustering method that is applied. For example, with the CMF method using BLAST and MACCS similarity metrics performance had typically improved with the larger sized panels, with an AUC and AUPR performance metric of 0.901 and 0.475 respectively for the Pharmacology Panel, while methods such as WNN-GIP under the same conditions did not perform as well in terms of accuracy, with an AUC and AUPR performance metric of 0.578 and 0.016 respectively.

Overall from assessment of the AUPR values, the most promising techniques for further evaluation with all the panels assessed would be the NetLapRLS and NRLMF methods. Whilst these methods show that only about half of the blind targets tested were detected, the AUPR values exceed the proportion of targets that would be detected solely by chance, which would be the proportion of active interactions present within a Panel's interaction matrix.

Panel	Similarity Method and Technique (AUC)									
	BLAST and PubChem					BLAST and MACCS				
	BLMNII	WNN-GIP	NetLapRLS	NRLMF	CMF	BLMNII	WNN-GIP	NetLapRLS	NRLMF	CMF
Panel 44	0.888654	0.809544	0.843829	0.920768	0.608573	0.889338	0.817318	0.84371	0.926787	0.624856
Panel 331	0.902232	0.756118	0.895371	0.893767	0.794667	0.90164	0.719048	0.895398	0.896903	0.837919
Pharma Panel	0.92846	0.569696	0.952315	0.923112	0.898545	0.930801	0.577776	0.952522	0.923865	0.901112
All	0.923389	0.548911	0.949937	0.916686	0.896142	0.930514	0.562718	0.950086	0.917315	0.897169
Panel	Similarity Method and Technique (AUPR)									
	BLAST and PubChem					BLAST and MACCS				
	BLMNII	WNN-GIP	NetLapRLS	NRLMF	CMF	BLMNII	WNN-GIP	NetLapRLS	NRLMF	CMF
Panel 44	0.372323	0.341152	0.515327	0.625003	0.121649	0.36554	0.340736	0.515523	0.627035	0.126813
Panel 331	0.31349	0.160590	0.434463	0.422046	0.280216	0.308407	0.128566	0.434511	0.424384	0.352034
Pharma Panel	0.278411	0.016069	0.50186	0.494855	0.471786	0.286922	0.016601	0.501955	0.493723	0.475117
All	0.262354	0.012777	0.485837	0.465314	0.46002	0.286209	0.013402	0.485907	0.465187	0.460665

**Table 5.4:** AUC and AUPR values of datasets using the BLAST for assessing protein similarity

Panel	Similarity Method and Technique (AUC)									
	SW and PubChem					SW and MACCS				
	BLMNII	WNN-GIP	NetLapRLS	NRLMF	CMF	BLMNII	WNN-GIP	NetLapRLS	NRLMF	CMF
Panel 44	0.88914	0.821760	0.843828	0.920224	0.614128	0.899602	0.836698	0.843702	0.926263	0.625839
Panel 331	0.905905	0.787639	0.895347	0.893444	0.813731	0.905299	0.783345	0.895374	0.896574	0.840536
Pharma Panel	0.941587	0.816201	0.952216	0.923102	0.896723	0.943427	0.830181	0.952421	0.923879	0.900114
All	0.937656	0.822222	0.949767	0.916658	0.894634	0.944773	0.830561	0.949916	0.917384	0.895914
Panel	Similarity Method and Technique (AUPR)									
	SW and PubChem					SW and MACCS				
	BLMNII	WNN-GIP	NetLapRLS	NRLMF	CMF	BLMNII	WNN-GIP	NetLapRLS	NRLMF	CMF
Panel 44	0.374415	0.352173	0.515336	0.622474	0.119878	0.367246	0.377115	0.515527	0.624457	0.12668
Panel 331	0.319472	0.216278	0.434428	0.423553	0.309744	0.314619	0.195570	0.434476	0.425764	0.354777
Pharma Panel	0.300181	0.170556	0.501704	0.496097	0.467857	0.308962	0.185169	0.501801	0.494985	0.471301
All	0.291216	0.167899	0.485685	0.466793	0.454673	0.318713	0.168581	0.485756	0.466688	0.455549

**Table 5.5:** AUC and AUPR Values of datasets using the Smith and Waterman (SW) method for assessing protein similarity

### 5.5.2 Prediction of New Interactions

Table 5.6 shows the number of inactive compound-protein pairs that were found in the top 1,000 potential new interactions, for which the rate of false positives detected ranged from 12% to 32% depending on the dataset used. In general Panel 44 generated the fewest false positives in its predictions, with the larger protein panels predicting a somewhat higher but similar level of false positives to each other. With the results filtered to the top 100 ranked predictions, the false positive rates change to a range of 6% to 38%, with Panel 44 still providing the lowest level of false positives. Whilst this analysis provides a preliminary indication of the reliability of the algorithm's top predictions, it should be pointed out that the inactive results were all instances where the recorded potency of Ki/AC50/IC50 etc. was more than 10 micromolar, but where there was nevertheless some form of recorded interaction, and that a large proportion of the interaction space is therefore still considered to be not fully conclusive. The results of this type of analysis would therefore be subject to alteration if the threshold is adjusted or further interactions are introduced from updated repositories. On further assessment of the other methods, CMF's predictions were found to be of note after detecting the lowest number of inactive interactions, with a range of 1% to 13% of new interactions being identified as inactive records. The data produced in this approach have also been incorporated with the new interaction predictions by the *in silico* pipeline.

Method	Data Set	Inactives Flagged	Inactives Flagged (Top 100)
NRLMF	Panel44SWMaccs	127 (12.7%)	8 (8%)
	Panel44BlastMaccs	130 (13.0%)	8 (8%)
	Panel44BlastPub	134 (13.4%)	6 (6%)
	Panel44SWPub	137 (13.7%)	6 (6%)
	PanelPharmaBlastMaccs	242 (24.2%)	23 (23%)
	PanelPharmaSWMaccs	243 (24.3%)	23 (23%)
	PanelPharmaBlastPub	246 (24.6%)	24 (24%)
	PanelPharmaSWPub	249 (24.9%)	24 (24%)
	AllSWMaccs	276 (27.6%)	32 (32%)
	AllBlastMaccs	277 (27.7%)	32 (32%)
	AllBlastPub	285 (28.5%)	37 (37%)
	AllSWPub	285 (28.5%)	37 (37%)
	Panel331BlastMaccs	321 (32.1%)	33 (33%)
	Panel331SWMaccs	323 (32.3%)	34 (34%)
	Panel331BlastPub	326 (32.6%)	36 (36%)
	Panel331SWPub	327 (32.7%)	35 (35%)
NetLapRLS	Panel44BlastPub	144 (14.4%)	15 (15%)
	Panel44SWPub	144 (14.4%)	15 (15%)
	Panel44BlastMaccs	145 (14.5%)	15 (15%)
	Panel44SWMaccs	145 (14.5%)	15 (15%)
	PanelPharmaSWPub	229 (22.9%)	21 (21%)
	PanelPharmaBlastPub	230 (23.0%)	21 (21%)
	PanelPharmaBlastMaccs	233 (23.3%)	21 (21%)
	PanelPharmaSWMaccs	233 (23.3%)	21 (21%)
	AllBlastMaccs	237 (23.7%)	26 (26%)
	AllBlastPub	237 (23.7%)	26 (26%)
	AllSWMaccs	237 (23.7%)	26 (26%)
	AllSWPub	237 (23.7%)	26 (26%)
	Panel331BlastMaccs	300 (30.0%)	38 (38%)
	Panel331BlastPub	300 (30.0%)	38 (38%)
	Panel331SWMaccs	300 (30.0%)	38 (38%)
	Panel331SWPub	300 (30.0%)	38 (38%)
CMF	Panel44BlastMaccs	7 (0.7%)	0 (0%)
	AllBlastPub	13 (1.3%)	0 (0%)
	AllBlastMaccs	15 (1.5%)	1 (1%)
	PanelPharmaBlastPub	23 (2.3%)	5 (5%)
	PanelPharmaSWMaccs	24 (2.4%)	1 (1%)
	Panel44SWPub	25 (2.5%)	4 (4%)
	AllSWPub	27 (2.7%)	6 (6%)
	Panel331SWPub	32 (3.2%)	6 (6%)
	PanelPharmaSWPub	43 (4.3%)	7 (7%)
	PanelPharmaBlastMaccs	46 (4.6%)	5 (5%)
	AllSWMaccs	60 (6.0%)	10 (10%)
	Panel331BlastPub	62 (6.2%)	3 (3%)
	Panel44BlastPub	71 (7.1%)	12 (12%)
	Panel44SWMaccs	86 (8.6%)	7 (7%)
	Panel331SWMaccs	98 (9.8%)	13 (13%)
	Panel331BlastMaccs	108 (10.8%)	9 (9%)

**Table 5.6:** Number of inactive results detected in the top 1,000 new interacting pairs for the NRLMF and NetLapRLS methods. SW is the abbreviated term for Smith and Waterman, while Pub is the abbreviated term for the PubChem fingerprint method

### 5.5.3 Verification of New Interactions

To verify the results and attempt to confirm whether or not the top interaction predictions were likely to be genuine targets, an *in silico* method was used on the data produced by the best performing clustering methods (NRLMF and NetLapRLS), in conjunction with the predictions on the CMF method which flagged the least amount of false positives on investigating the new interactions. As the *in silico* method required a significant amount of time to process dockings, a subset of the new interactions from each of the three methods were selected for assessment. The selection of the top 10 and bottom 10 results from the 1,000 set generated a suitable amount of results for assessment as a small scale case study of the *in silico* method, as well as testing whether or not there is a significant difference present between high scoring and low scoring interaction predictions. In addition to this, predictions within this subset which had been classed as inactive from assessment of the repository interactions was also considered to determine if a difference in binding might be detected to verify their determination as inactive.

To ensure that as diverse a set of compounds and proteins as possible were considered for assessing the predictions *in silico*, the combined protein panel was used for this purpose, with the BLAST algorithm for comparing proteins and the MACCS fingerprint format for comparing compounds.

#### 5.5.3.1 Initial Findings

Tables 5.7, 5.8, and 5.9 provide a summary of the top and bottom ranked dockings for each method assessed, with indicators highlighting whether or not the interaction was found to be inactive (outside the 10 micromolar potency threshold) and if the SDF structure for the compound was in a 3D format and optimal for the *in silico* pipeline. For predictions of the top 10 and bottom 10 groups being labelled as inactive via thresholding, CMF flagged 1 prediction as inactive from the bottom 10 and none from the top 10, whereas NetLapRLS and NRLMF methods flagged 4 (1 from the top 10, 3 from bottom 10) and 7 (6 from top 10 and 1 from bottom 10) predictions as inactive respectively. These results were on the face of it, broadly disappointing. However, closer scrutiny of the predictions revealed some possible explanations. A particularly striking finding is that NRLMF identified 6 of its top 10 predictions as inactive, though on further investigation of these predictions, it was found that most of the Ki/Kd/IC50 etc. potencies returned were very close to the 10 micromolar threshold, and in some cases had been labelled as active by ToxCast. Table 5.10

provide a summary of the inactive predictions, and Table 5.11 provide details of the ToxCast assays connected to the predictions. Of the 12 predictions which were labelled as inactive in the groups assessed, 20 assay results were returned which provided micromolar concentration measurements of AC50, IC50, Ki. In some cases multiple assay results were returned on an inactive protein-compound pair when queried, such as the PubChem Compound ID 60196404 (DSSTox.CID.27353), which returned 6 different concentration values for the UniProt accession code P03372 (Estrogen Receptor).

Overall however, the predictions generated have shown a disagreement on the *in vitro* assay results. On comparing all of the top 10 predictions (where inactive results should not be expected), 7 out of 30 predictions were found to have been inactive through existing assay results, and were therefore potentially incorrect predictions. On comparing all of the bottom 10 predictions (where inactive results should be expected), 25 out of 30 predictions were found not to have an inactive assay result, and were also potentially incorrect predictions. Therefore, on initial findings, only 28 out of 60 predictions appear to be correct. This indicated that either a modification would be needed to the active threshold level, or additional data points for the top and bottom groupings would be needed.

Ranking	CMF Predictions					
	Compound ID	Compound Name	UniProt ID	Gene Code	Inactive?	3D?
Top 10	71349	Lanreotide	P31644	GBRA5	No	No
	86160	Spiroxamine	P31644	GBRA5	No	Yes
	451668	Decitabine	P31644	GBRA5	No	Yes
	8364	2-Ethylhexyl Salicylate	P31644	GBRA5	No	Yes
	5311128	Goserelin	P31644	GBRA5	No	No
	19700	Brilliant Blue FCF	P08913	ADA2A	No	No
	11097730	Cyclanilide	P31644	GBRA5	No	Yes
	2585	Carvedilol	P31644	GBRA5	No	Yes
	4054	Memantine	P31644	GBRA5	No	Yes
	66494	Phenolphthalin	P08913	ADA2A	No	Yes
Bottom 10	1923	8-Hydroxyquinoline	P08913	ADA2A	No	Yes
	5978	Vincristine	P31644	GBRA5	No	No
	86160	Spiroxamine	P22303	ACES	No	Yes
	25015677	Ro3280	P08913	ADA2A	Yes	Yes
	4506	Nitrazepam	P08913	ADA2A	No	Yes
	40839	Vindesine	P36544	ACHA7	No	No
	71349	Lanreotide	Q12923	PTN13	No	No
	33746	Ketazolam	P05023	AT1A1	No	Yes
	9854073	Cabazitaxel	P08913	ADA2A	No	No
	19700	Brilliant Blue FCF	Q9HB55	CP343	No	No

**Table 5.7:** Top 10 and Bottom 10 of the 1,000 predictions made by the CMF Method. Column Compound ID refers to the ID within the PubChem Compound database. Column Inactive refers to whether the prediction has been detected by the repository as above the 10 micromolar active threshold. Column 3D refers to whether or not the compound in this prediction has an SDF model in a 3D format.

Ranking	NetLapRLS Predictions					
	Compound ID	Compound Name	UniProt ID	Gene Code	Inactive?	3D?
<b>Top 10</b>	3369	Fludiazepam	P48169	GBRA4	No	Yes
	3369	Fludiazepam	Q16445	GBRA6	No	Yes
	3033621	Cinolazepam	P48169	GBRA4	No	Yes
	3033621	Cinolazepam	Q16445	GBRA6	No	Yes
	702	Ethanol	P18507	GBRG2	No	Yes
	2811	Clotiazepam	P48169	GBRA4	No	Yes
	2811	Clotiazepam	Q16445	GBRA6	No	Yes
	3369	Fludiazepam	Q9UN88	GBRT	No	Yes
	3033621	Cinolazepam	Q9UN88	GBRT	No	Yes
	2157	Amiodarone	P10635	CP2D6	Yes	Yes
<b>Bottom 10</b>	47811	Pergolide	P08173	ACM4	No	Yes
	3034368	Mancozeb	P05177	CP1A2	Yes	No
	941651	Thiothixene	P18089	ADA2B	No	Yes
	5533	Trazodone	P08172	ACM2	No	Yes
	60196404	DSSTox_CID_27353	P03372	ESR1	Yes	No
	11954293	SCHEMBL6029138	P20309	ACM3	No	No
	3000715	Thiopental	Q9UN88	GBRT	No	Yes
	5353853	Oxiconazole	P35367	HRH1	No	Yes
	63062	Afimoxifene	P10635	CP2D6	No	Yes
	4485	Nifedipine	P10635	CP2D6	Yes	Yes

**Table 5.8:** Top 10 and Bottom 10 of the 1,000 predictions made by the NetLapRLS Method. Column Compound ID refers to the ID within the PubChem Compound database. Column Inactive refers to whether the prediction has been detected by the repository as above the 10 micromolar active threshold. Column 3D refers to whether or not the compound in this prediction has an SDF model in a 3D format.

Ranking	NRLMF Predictions					
	Compound ID	Compound Name	UniProt ID	Gene Code	Inactive?	3D?
Top 10	156328	Besonprodil	P14416	DRD2	Yes	Yes
	11982778	DSSTox_CID.28150	P11229	ACM1	No	No
	4170	Metolazone	P00915	CAH1	Yes	Yes
	9849616	Surinabant	Q9HCS2	CP4FC	No	Yes
	11057	Crystal violet	P50406	5HT6R	Yes	No
	2520	Verapamil	P10635	CP2D6	Yes	Yes
	15096	Tributyltin chloride	P18825	ADA2C	Yes	No
	3198	Econazole	P18825	ADA2C	No	Yes
	37175	Imazalil	Q01959	SC6A3	Yes	Yes
	5280961	Genistein	P50406	5HT6R	No	Yes
Bottom 10	9909677	DSSTox_CID.27347	P25101	EDNRA	No	Yes
	4037	Meclofenamic acid	P16473	TSHR	No	Yes
	60196408	DSSTox_CID.27379	P37288	V1AR	Yes	No
	4184	Mianserin	Q99720	SGMR1	No	Yes
	3562	Halothane	P42261	GRIA1	No	Yes
	6128	Androstenedione	P08172	ACM2	No	Yes
	3744660	DSSTox_CID.24928	Q06187	BTK	No	No
	4940	Propiomazine	P18825	ADA2C	No	Yes
	9872438	DSSTox_CID.27368	P34969	5HT7R	No	Yes
	5957	Adenosine triphosphate	P15692	VEGFA	No	Yes

**Table 5.9:** Top 10 and Bottom 10 of the 1,000 predictions made by the NRLMF Method. Column Compound ID refers to the ID within the PubChem Compound database. Column Inactive refers to if the prediction has been detected by the repository as above the 10 micromolar active threshold. Column 3D refers to whether or not the compound in this prediction has an SDF model in a 3D format.

CompoundID	UniProtID	Assay Type	Assay Value	Standard Units	Source
25015677	P08913	ac50	20.747	$\mu\text{m}$	ToxCast
3034368	P05177	ac50	1000000	$\mu\text{m}$	ToxCast
60196404	P03372	ac50	1000000	$\mu\text{m}$	ToxCast
60196404	P03372	ac50	51.358	$\mu\text{m}$	ToxCast
60196404	P03372	ac50	16.598	$\mu\text{m}$	ToxCast
60196404	P03372	ac50	32.096	$\mu\text{m}$	ToxCast
60196404	P03372	ac50	68.607	$\mu\text{m}$	ToxCast
60196404	P03372	ac50	59.345	$\mu\text{m}$	ToxCast
156328	P14416	ac50	16.221	$\mu\text{m}$	ToxCast
11057	P50406	ac50	17.759	$\mu\text{m}$	ToxCast
15096	P18825	ac50	1000000	$\mu\text{m}$	ToxCast
37175	Q01959	ac50	12.304	$\mu\text{m}$	ToxCast
60196408	P37288	ac50	1000000	$\mu\text{m}$	ToxCast
4170	P00915	ki	54	$\mu\text{m}$	BindingDB
2520	P10635	ic50	43.3	$\mu\text{m}$	BindingDB
2157	P10635	ac50*	31.623	$\mu\text{m}$	ChEMBL
4485	P10635	ac50*	39.811	$\mu\text{m}$	ChEMBL
2520	P10635	ic50	43.3	$\mu\text{m}$	ChEMBL

**Table 5.10:** Assay results of the predictions classed as inactive. Column Compound ID refers to the ID within the PubChem Compound database. \*These compounds had been listed as 'potency' under ChEMBL, but refer to AC50 assay types from PubChem BioAssay



CompoundID	ToxCast Assay Name	UniProtID	Assay Value (AC50 $\mu\text{m}$ )
25015677	NVS_GPCR_hAdra2A	P08913	20.747
3034368	NVS_ADME_hCYP1A2_Activator	P05177	1000000
3034368	NVS_ADME_hCYP1A2	P05177	1000000
60196404	OT_ERa_ERE_GFP_0480	P03372	1000000
60196404	ATG_ERa_TRANS_up	P03372	1000000
60196404	OT_ERa_ERaERa_0480	P03372	1000000
60196404	TOX21_ERa_BLA_Antagonist_ratio	P03372	51.358
60196404	OT_ERa_ERE_GFP_0120	P03372	16.598
60196404	ACEA_T47D_80hr_Positive	P03372	1000000
60196404	TOX21_ERa_LUC_BG1_Agonist	P03372	1000000
60196404	NVS_NR_hER	P03372	32.096
60196404	TOX21_ERa_BLA_Agonist_ratio	P03372	68.607
60196404	ATG_ERE_CIS_up	P03372	1000000
60196404	TOX21_ERa_LUC_BG1_Antagonist	P03372	59.345
156328	NVS_GPCR_hDRD2s	P14416	16.221
11057	NVS_GPCR_h5HT6	P50406	17.759
15096	NVS_GPCR_hAdra2C	P18825	1000000
37175	NVS_TR_hDAT	Q01959	12.304
60196408	NVS_GPCR_hV1A	P37288	1000000

**Table 5.11:** Details of the ToxCast assays of the predictions found to be inactive through thresholding

### 5.5.3.2 Docking Results

Table 5.12, Table 5.13 and Table 5.14 detail the binding energies which were output by the *in silico* pipeline from the CMF, NRLMF and NetLapRLS methods, where lower binding energies are indicative of stronger levels of binding and in turn increased likelihood of activity. Values which are labelled as NA were docking operations which were not scored, either due to a model being rendered incorrectly or an erroneous binding site being defined. Of the 60 total docking operations passed to the *in silico* pipeline, 59 operations returned a binding energy value.

As an illustrative example, Figure 5.3 provides an example of a protein-compound pair which has generated a high binding energy value from the results generated. In this example, the docking has placed the compound closely to the protein structure, with the compound having many contact points with the protein (highlighted by dashed red lines). On the other hand, Figure 5.4 shows a protein-compound pair which has provided a lower binding energy value, approximately half the strength of the association of that on Figure 5.3. In this instance, the compound has fewer contact points, and is located towards the periphery of the protein structure which in turn has contributed to the generation of a lower binding energy value.

CompoundID	UniProtID	Is Inactive?	Docking Binding Energy (kcal/mol)
Top 10			
71349	P31644	No	-5.40
86160	P31644	No	-4.44
451668	P31644	No	-5.06
8364	P31644	No	-4.03
5311128	P31644	No	-2.28
19700	P08913	No	-8.61
11097730	P31644	No	-4.96
2585	P31644	No	-4.68
4054	P31644	No	-5.57
66494	P08913	No	-8.70
Bottom 10			
1923	P08913	No	-6.46
5978	P31644	No	-4.97
86160	P22303	No	-6.30
25015677	P08913	Yes	-9.30
4506	P08913	No	-8.08
40839	P36544	No	-5.26
71349	Q12923	No	-6.22
33746	P05023	No	-7.32
9854073	P08913	No	-7.56
19700	Q9HB55	No	-7.48

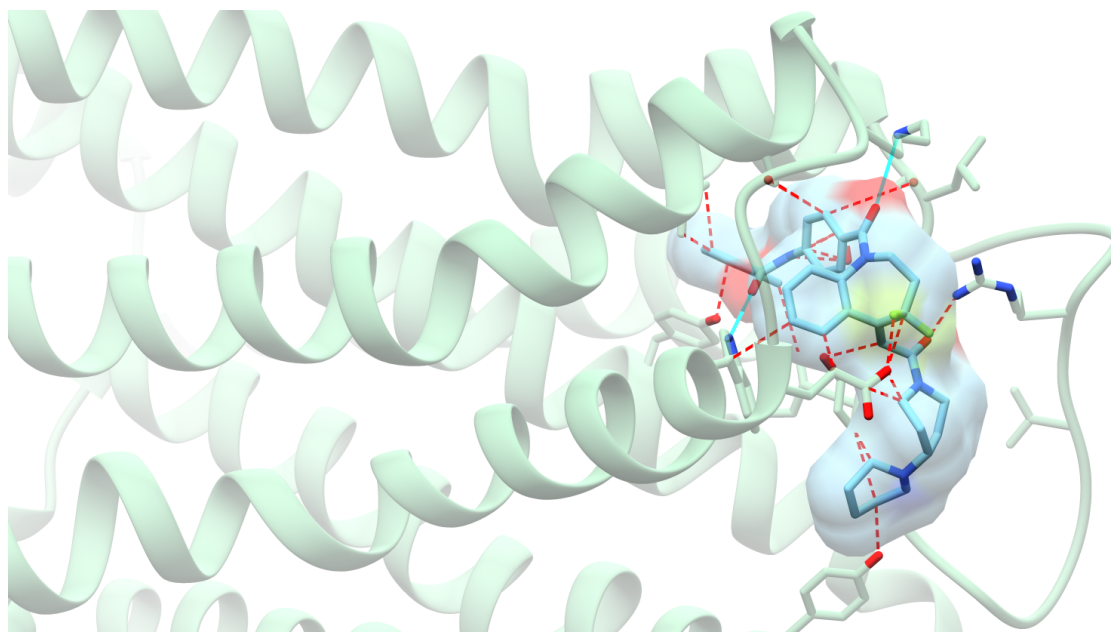
**Table 5.12:** Binding energies for the top 10 and bottom 10 of the 1,000 predictions made using the CMF method.

CompoundID	UniProtID	Is Inactive?	Docking Binding Energy (kcal/mol)
Top 10			
3369	P48169	No	-6.45
3369	Q16445	No	-6.54
3033621	P48169	No	-6.09
3033621	Q16445	No	-6.60
702	P18507	No	-2.49
2811	P48169	No	-5.94
2811	Q16445	No	-5.76
3369	Q9UN88	No	-6.62
3033621	Q9UN88	No	-6.36
2157	P10635	Yes	-7.34
Bottom 10			
47811	P08173	No	-7.19
3034368	P05177	Yes	-4.45
941651	P18089	No	-8.31
5533	P08172	No	-8.04
60196404	P03372	Yes	-5.68
11954293	P20309	No	-9.83
3000715	Q9UN88	No	-4.80
5353853	P35367	No	-9.62
63062	P10635	No	-7.24
4485	P10635	Yes	-6.26

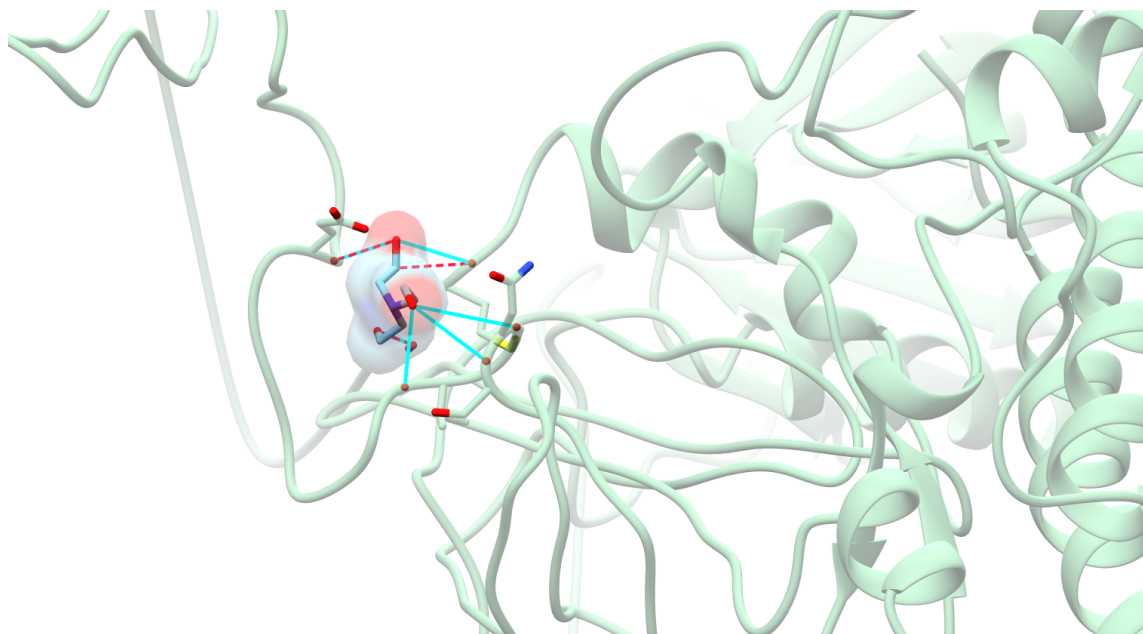
**Table 5.13:** Binding energies for the top 10 and bottom 10 of the 1,000 predictions made using the NetLapRLS method

CompoundID	UniProtID	Is Inactive?	Docking Binding Energy (kcal/mol)
Top 10			
156328	P14416	Yes	-8.74
11982778	P11229	No	-8.87
4170	P00915	Yes	-7.92
9849616	Q9HCS2	No	-6.73
11057	P50406	Yes	-7.92
2520	P10635	Yes	-6.68
15096	P18825	Yes	NA
3198	P18825	No	-5.84
37175	Q01959	Yes	-6.81
5280961	P50406	No	-7.98
Bottom 10			
9909677	P25101	No	-5.73
4037	P16473	No	-5.24
60196408	P37288	Yes	-6.89
4184	Q99720	No	-8.56
3562	P42261	No	-4.16
6128	P08172	No	-8.19
3744660	Q06187	No	-4.19
4940	P18825	No	-5.63
9872438	P34969	No	-9.55
5957	P15692	No	-5.65

**Table 5.14:** Binding energies for the top 10 and bottom 10 of the 1,000 predictions made using the NRLMF method. Binding energies labelled NA indicate a docking which was not considered due to incomplete output.



**Figure 5.3:** Simulated binding of PubChem Compound ID 11982778 (YM218) with UniProt ID P11229 (ACM1\_HUMAN), with a binding energy of -8.87 kcal/mol. Hydrogen bonds between the compound and protein are shown as blue solid lines. Contact points between the compound and protein are shown as dashed red lines.



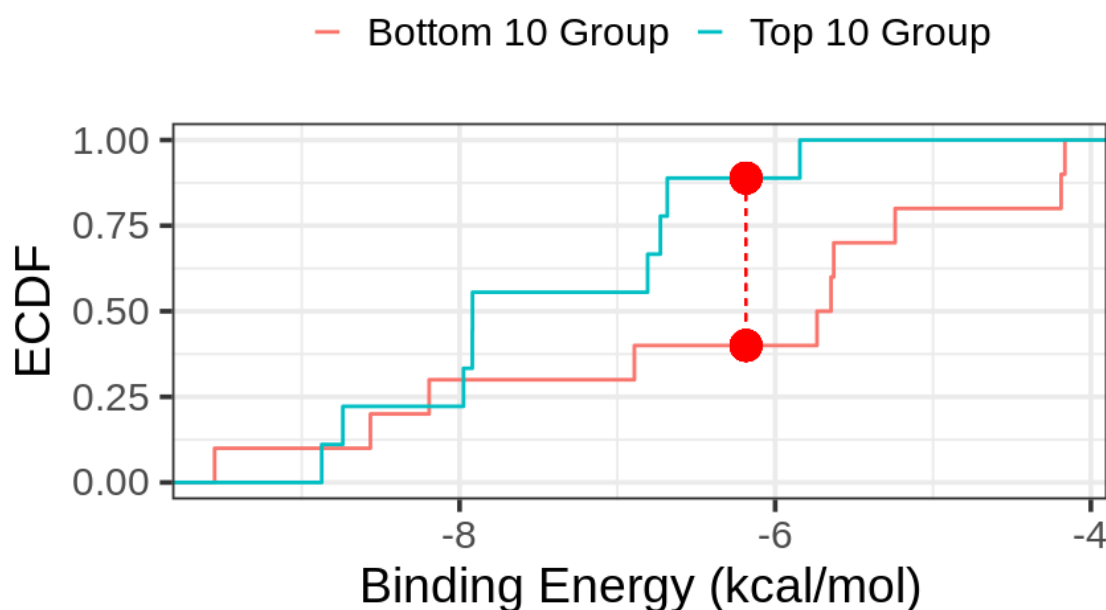
**Figure 5.4:** Simulated binding of PubChem CompoundID 3744660 (36673-16-2) with UniProt ID Q06187 (BTK\_HUMAN), with a binding energy of  $-4.19$  kcal/mol. Hydrogen bonds between the compound and protein are shown as blue solid lines. Contact points between the compound and protein are shown as dashed red lines.

To determine if the top prediction's binding energies were significantly stronger than the scores on the bottom of the 1,000 predictions, a two sample Kolmogorov-Smirnov test was performed [18]. This testing method calculates the distribution function of two groups of values, and measures the largest gap between the two groups. On the alternative hypothesis that the distribution function in the top group exceeds that of the bottom group, a low  $p$  value would indicate the layout of the top group is stronger in binding energy than the bottom group. These tests indicated that in the case of CMF and NetLapRLS, the top scores were not significantly stronger in binding energies, with  $p$  values of 0.9048 and 0.6703 respectively. In the case of NRLMF however, the scores on the top prediction brackets were considered significantly stronger, with a  $p$  value of 0.033. Figure 5.5 displays the distribution function plot of the NRLMF prediction brackets. This highlights that in the case of one of methods, the predictions made by the clustering method were compatible with those of the *in silico* pipeline.

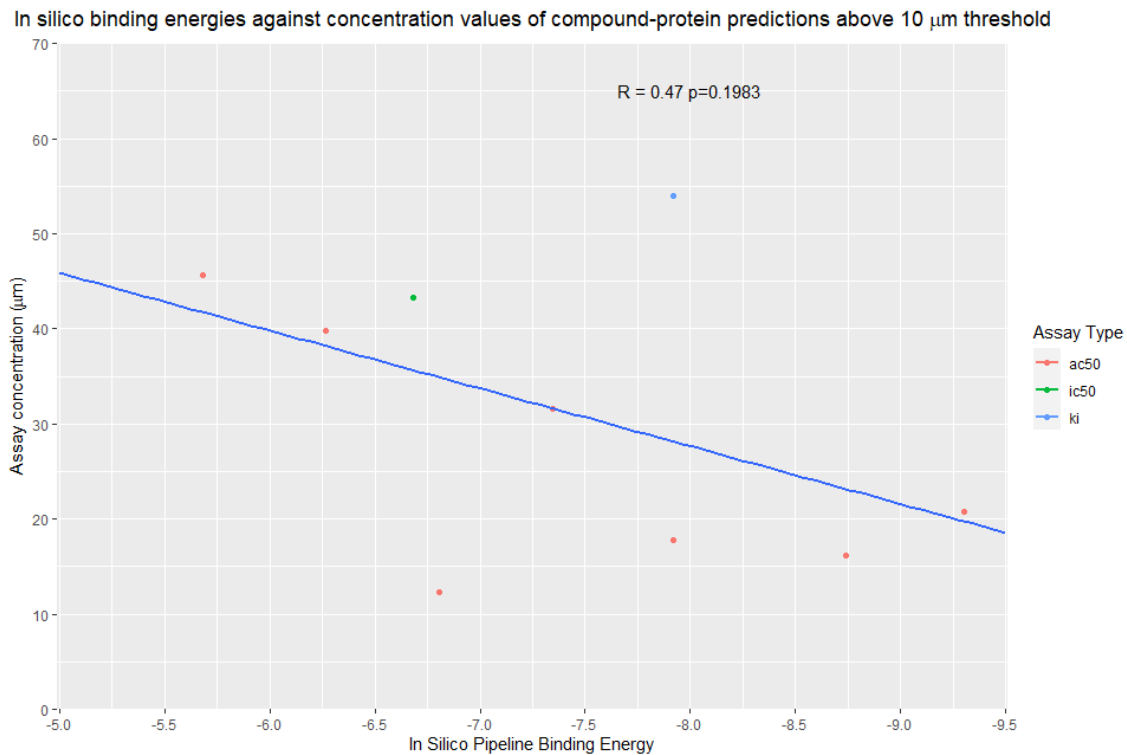
In terms of predictions determined to be inactive from 10 micromolar thresholding, binding energies generated from the *in silico* platform were compared to the *in vitro* assay results. In the event that a compound-protein pair had more than one assay type measurement, the results were averaged for visualisation in a scatter plot shown in Figure 5.6, which exclude 4 ToxCast assay results which had a recorded assay result

of 1 mol. There is some indication that in general as reported potency of Ki/Kd/IC50 etc. decreases, the binding energy value obtained by the *in silico* pipeline increases, however this is not the case with the Ki assay measurements, and the number of datapoints are insufficient to indicate the correlation is statistically significant with a  $p$  value of 0.1983. Additional datapoints and separation of assay types would be needed in future work to increase and balance representation of the assay types, and to investigate if modifications would be needed to thresholding values applied.

## K-S Test: NLRMF Predictions



**Figure 5.5:** Plot of the Empirical Distribution Function between the top 10 and bottom 10 predictions of the NLRMF method. The red dots and line signifies the maximum distance between the distribution of both groups.



**Figure 5.6:** Assay concentration values against the *in silico* pipeline binding energies

## 5.6 Conclusions

In this chapter, an assessment was made of methods which predicted compound-protein interactions through the clustering of similar structures. Whilst studies into this field have evaluated their techniques on a standard interaction set, no approaches have considered use on a panel beyond the Yamanishi dataset. A three-phased approach was undertaken: the first phase was to assess which combination of methods, comparison measurements and protein panels were shown to have to most promise in predicting interactions. It was found that of the 5 clustering methods assessed, two methods (NRLMF and NetLapRLS) were shown to be the most promising candidates with the protein panels used, which when tested on blinded *in vitro* interactions, approximately half on average were detected successfully.

The second phase of the analysis was to determine how many of the top scoring interactions from these clustering methods were determined to be inactive through searching assay results which were found to be above a 10 micromolar concentration result. Of the top 1,000 interactions from the NRLMF and NetLapRLS methods, between 12% to 32% of interactions were determined to be inactive, depending on the

protein panel and comparison metrics used. Of the methods which found the least number of inactive interactions, CMF was found to have flagged the fewest inactive interactions, with a range of 1% to 13%.

The final phase of the analysis was to determine if the predictions made via the CMF, NetLapRLS and NRLMF methods could be used in conjunction with an *in silico* docking pipeline, to verify whether the predictions could be considered accurate and whether there was a distinction between the top (10) and bottom (10) ends of the top 1,000 predictions. Although the NetLapRLS and CMF methods did not show a difference in binding energies between the top and bottom groups, the NRLMF method's top predictions were found to have significantly stronger binding energies in comparison to the bottom predictions. This highlights that there is potential for similarity clustering and the *in silico* pipeline to be used in combination in a complementary way.

### 5.6.1 Further Work

Although the findings of the clustering platform show promise as a method to complement the drug development process, further work and scoping would be needed to expand on this case study. One of the main limitations of the clustering techniques for example is that compound-protein pairs which are considered to be inactive are not considered as input, and although approaches exist which carry out these methods, code to replicate the process is not freely available i.e. to use to replicate the analysis alongside other methods. Further investigation would therefore be needed to assess if the approaches can be adapted to incorporate inactive interactions, and to construct an environment where all methods can be tested in similar experimental conditions.

Another area for exploration includes consideration of other compound and protein comparison methods; whilst the findings demonstrated no significant difference between the comparison techniques assessed, this represented only a small selection of a wide variety of comparison methods available. In addition to this, a mixture of comparison methods could also be considered to generate an average of similarities (a consensus approach) so that a more substantial number of properties might be considered for comparison.

Finally, a point for further consideration is the scale of the *in silico* operations. Although the case study's small scale revealed that predictions of one of the clustering methods promisingly provided higher binding energies, a wider study would further verify the findings, as well as provide additional data points to gauge inactive predictions against assay values. This would in turn determine the extent of modification

that might be needed to the 10 micromolar threshold, and to assess any impacts on performance due to these modifications.

## 5.7 References

- [1] Cronin, M. T. D., “Chapter 1 an introduction to chemical grouping, categories and read-across to predict toxicity,” in *Chemical Toxicity Prediction: Category Formation and Read-Across*, The Royal Society of Chemistry, 2013, pp. 1–29, ISBN: 978-1-84973-384-7. DOI: 10.1039/9781849734400-00001.
- [2] Yamanishi, Y., Araki, M., Gutteridge, A., Honda, W., and Kanehisa, M., “Prediction of drug–target interaction networks from the integration of chemical and genomic spaces,” *Bioinformatics*, vol. 24, no. 13, pp. i232–i240, 2008.
- [3] Van Laarhoven, T. and Marchiori, E., “Predicting drug-target interactions for new drug compounds using a weighted nearest neighbor profile,” *PloS one*, vol. 8, no. 6, e66952, 2013.
- [4] Xia, Z., Wu, L.-Y., Zhou, X., and Wong, S. T., “Semi-supervised drug-protein interaction prediction from heterogeneous biological spaces,” in *BMC systems biology*, BioMed Central, vol. 4, 2010, S6.
- [5] Mei, J.-P., Kwoh, C.-K., Yang, P., Li, X.-L., and Zheng, J., “Drug–target interaction prediction by learning from local information and neighbors,” *Bioinformatics*, vol. 29, no. 2, pp. 238–245, 2012.
- [6] Laarhoven, T. van, Nabuurs, S. B., and Marchiori, E., “Gaussian interaction profile kernels for predicting drug–target interaction,” *Bioinformatics*, vol. 27, no. 21, pp. 3036–3043, 2011.
- [7] Zheng, X., Ding, H., Mamitsuka, H., and Zhu, S., “Collaborative matrix factorization with multiple similarities for predicting drug-target interactions,” in *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM, 2013, pp. 1025–1033.
- [8] Gönen, M., “Predicting drug–target interactions from chemical and genomic kernels using bayesian matrix factorization,” *Bioinformatics*, vol. 28, no. 18, pp. 2304–2310, 2012.
- [9] Keum, J. and Nam, H., “Self-blmm: Prediction of drug-target interactions via self-training svm,” *PloS one*, vol. 12, no. 2, e0171839, 2017.



- [10] Liu, Y., Wu, M., Miao, C., Zhao, P., and Li, X.-L., “Neighborhood regularized logistic matrix factorization for drug-target interaction prediction,” *PLoS computational biology*, vol. 12, no. 2, e1004760, 2016.
- [11] Ban, T., Ohue, M., and Akiyama, Y., “Nrlmf $\beta$ : Beta-distribution-rescored neighborhood regularized logistic matrix factorization for improving the performance of drug-target interaction prediction,” *Biochemistry and biophysics reports*, vol. 18, p. 100615, 2019.
- [12] Smith, T. F., Waterman, M. S., *et al.*, “Identification of common molecular subsequences,” *Journal of molecular biology*, vol. 147, no. 1, pp. 195–197, 1981.
- [13] Öztürk, H., Ozkirimli, E., and Özgür, A., “A comparative study of smiles-based compound similarity functions for drug-target interaction prediction,” *BMC bioinformatics*, vol. 17, no. 1, p. 128, 2016.
- [14] National Center for Biotechnology Information. (2009). “Pubchem substructure fingerprint,” [Online]. Available: [ftp://ftp.ncbi.nlm.nih.gov/pubchem/specifications/pubchem\\_fingerprints.txt](ftp://ftp.ncbi.nlm.nih.gov/pubchem/specifications/pubchem_fingerprints.txt).
- [15] Boratyn, G. M., Camacho, C., Cooper, P. S., Coulouris, G., Fong, A., Ma, N., Madden, T. L., Matten, W. T., McGinnis, S. D., Merezhuk, Y., *et al.*, “BLAST: A more efficient report with usability improvements,” *Nucleic acids research*, vol. 41, no. W1, W29–W33, 2013.
- [16] Austin-Muttitt, K., “Modelling competition binding assays: A step towards in silico pharmacological profiling tools,” Report submitted to Swansea University, 2019.
- [17] Madeira, F., Park, Y. m., Lee, J., Buso, N., Gur, T., Madhusoodanan, N., Basutkar, P., Tivey, A. R. N., Potter, S. C., Finn, R. D., and Lopez, R., “The EMBL-EBI search and sequence analysis tools APIs in 2019,” *Nucleic Acids Research*, vol. 47, no. W1, W636–W641, Apr. 2019, ISSN: 0305-1048. DOI: 10.1093/nar/gkz268.
- [18] Marsaglia, G., Tsang, W. W., Wang, J., *et al.*, “Evaluating kolmogorov’s distribution,” *Journal of Statistical Software*, vol. 8, no. 18, pp. 1–4, 2003.

## Chapter 6

# TargetPredict: Design and Implementation of an Interface for Drug Profiling and Interaction Similarity Clustering

### 6.1 Introduction

Although the work thus far has revealed promise for establishing new approaches for compound profiling, one of the current limitations with most methods is a lack of ability to easily reproduce the analyses and although some methods have provided source code of workings, those with limited technical expertise in the programming languages use may struggle to make use of them. The aim of this chapter is to present the design and implementation of a small-scale prototype which attempts to make available reproducible techniques and methods to a wider community. The interface itself is available online at <http://proteins.swan.ac.uk/cheminf>.

### 6.2 Requirements

For the interface to be implemented successfully, a number of requirements needed to be satisfied. This section will detail the main requirements, as well as a number of desirable requirements. As this interface is a small-scale prototype, only the information gathered from the work described in Chapters 4 and 5 was used.

**Traversal of the compiled data.** In order for a user to be able to determine the information that was employed in this project, the interface should ideally have the means to allow a user to navigate and search for specific targets and assay result

values. Specific search terms (such as assay types, assay thresholding values and protein accession codes) and a table download facility would allow a user to construct small customised datasets without programming knowledge.

**Calculation of nearest neighbours.** As part of the attempts towards discovering interactions for a candidate compound, a method of quickly detecting similar structures and the display of their documented interaction profiles would provide a user with a selection of potential candidates for further investigation. The interface must allow a user to upload a user-specified compound or protein, and view the nearest neighbours present in the interface's similarity matrix. For the purposes of compound similarity, the MACCS and PubChem fingerprint formats were used. For proteins, the Smith and Waterman and BLAST methods were used.

**Prediction of interactions.** To incorporate the similarity clustering methods used in Chapter 5, the interface needs to have the ability to compile a customised similarity matrix relating to a user-selected protein or compound. The interface must then have the ability to communicate with the Python PyDTI library and process predicted interactions with the custom protein or compound, and determine which are likely to be correct predictions

**Profiling of custom compounds.** To incorporate the profiling methods implemented in Chapters 3 and 4, the interface should provide the ability to potentially classify a user-selected compound. This can be accomplished by identifying the interaction profile of similarly structured compounds and identifying those which are present in DrugBank [1] or ToxCast [2].

To assist in the satisfaction of these requirements, the R package "Shiny" was used, which allows users of the package to construct webpages and in turn make use of the datasets and functions in R to generate customised datasets for presentation [3]. Not only would the use of this package reduce the amount of time needed to manually construct a webpage design via HTML, it also allows the use of existing scripts to construct and present datasets based on user input on the page, in turn further reducing the amount of time needed to compile additional scripts in languages different to those used throughout the project.

## 6.3 Constructing the Interface

To construct the interface, it was necessary to compile and store the datasets required for the interface to query and manipulate in an easily accessible format. Websites constructed with the "Shiny" package are composed of two elements: a "UI" element

which presents the tables, tabs and user input elements, and a "Server" element which performs the appropriate actions when certain actions are triggered by the user. Before a "Shiny" app is implemented however, a user can load and perform start up functions which can then be accessed by all users of the app. On small scale datasets it is possible for a workspace to be saved in a format that can be restored in a single command by R, and this was used to compile the necessary elements for the server's use. For the interface to function in line with the requirements, the following elements were required:

**Interaction and Similarity Matrices.** With an interaction and similarity matrix stored in the app's environment, it is possible for the app to quickly append a custom compound or protein into the matrix when required. The storage of the matrices also allows the app to compile temporary datasets for use by the PyDTI python library to make predictions by means of the clustering techniques, outlined in Chapter 5.

**Fingerprint sets.** To compare custom compounds against those present in the interaction space, the storage of the fingerprint sets of each compound would allow comparisons to be executed quickly, while reducing the amount of space needed to store SDF coordinates of a compound. The "Rcdk" library referenced in Chapter 5 allows fingerprint sets of compounds to be saved inside an R workspace for this purpose.

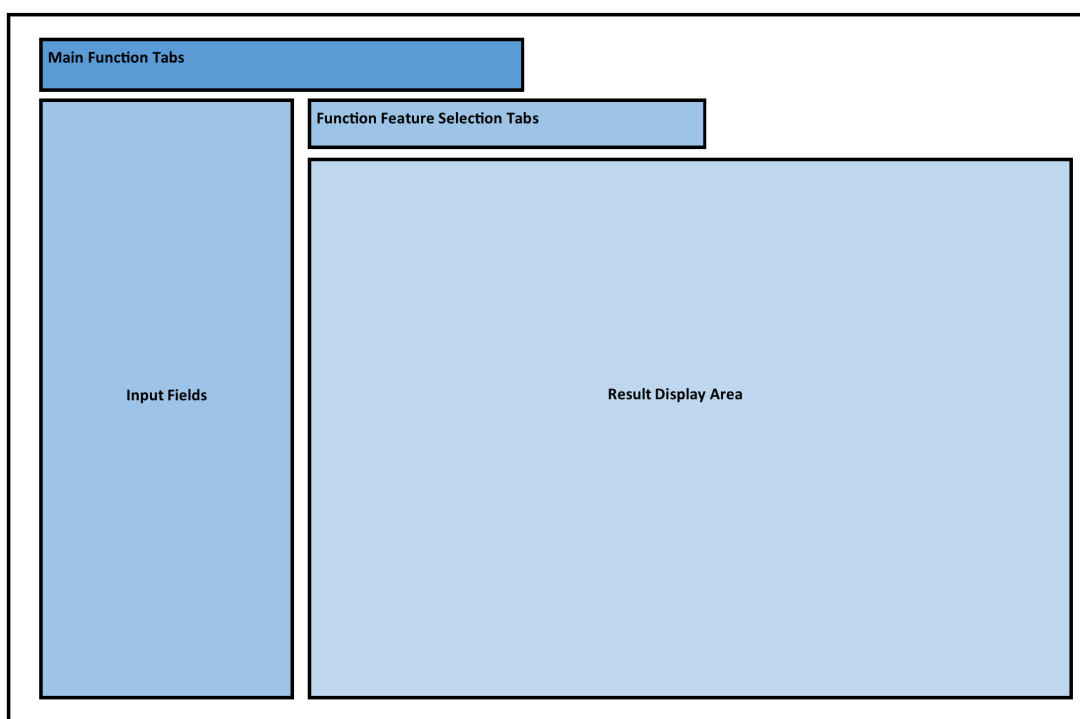
**Protein FASTA files.** The storage of protein FASTA files will allow a custom protein to be quickly compared against those present in the interaction space.

**Assay/Interaction Result sets.** The storage of the assay results (from ToxCast, BindingDB and ChEMBL) and target lists (from DrugBank) will allow the app to filter these datasets based on user input.

## 6.4 Interface Design

Figure 6.1 provides an overview of the planned design for the interface. The main page design consists of two main elements: a user input segment which allows the user to provide input and custom structures, and a result segment which displays the appropriate result tables according to the features and inputs the user has selected. To navigate the interface, the user will have a number of tabs to select: the top tab will provide the ability for the user to traverse to separate tools of the interface, such as one tab for the assay and target browser and one tab for the similarity clustering. The tabs placed above the result segment of the page will specify individual features

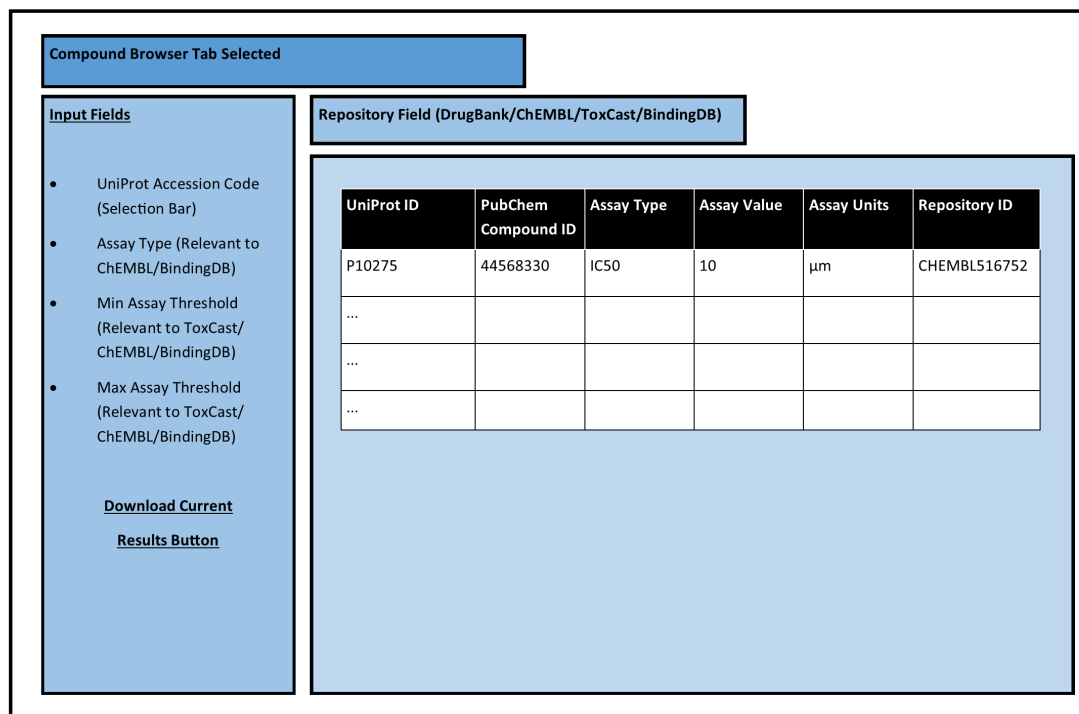
of the current tool selected by the user; for example if the user has selected the interaction browser, they will be able to cycle through the individual databases for results of individual proteins. On the other hand with the similarity clustering, the user's options would be to either cycle through nearest neighbour comparisons or interaction predictions of a custom compound/protein.



**Figure 6.1:** Simple illustration of the design of the interface

### 6.4.1 Compound Target Browser

Figure 6.2 provides an overview of the design for the Compound Target Browser tab. In this tab, the user will be able to browse through the information gathered from DrugBank, ToxCast, ChEMBL and BindingDB. For the purposes of this prototype, the information presented would be limited to the information described in Chapters 4 and 5, and be limited to interactions of proteins present in Panel 44, Panel 331 or the Pharmacology Panel. This provides a reasonably sized dataset for browsing and downloading whilst keeping the server load time to a minimum.



**Figure 6.2:** Simple illustration of the design of the Compound Browser Tab

To provide the user with the ability to filter the information presented, user input bars are provided on the left-hand side, which can vary depending on the repository selected by the user. For example, information from DrugBank is limited to displaying target links between compounds and proteins, thus the input field would be limited to solely selecting the UniProt accession code. With other datasets however information is given by assay result values, therefore additional input fields would be needed to select items such as the assay type performed, and threshold values for the minimum and maximum assay result value to be displayed.

Another feature included with the design is the ability for users to access specific repositories with the references provided by the browser (such as providing a URL for a specific DrugBank compound). This could be accomplished by simply applying a URL reference tag ( $\langle a \rangle$ ) to the correct repository using the reference on the table of results. Furthermore, the user should have the ability to download an easily accessible version of the results which have been displayed on the interface.

## 6.4.2 Compound/Protein Similarity

Figure 6.3 and 6.4 provide design overviews of the Nearest Neighbour and Interaction Prediction Tabs if the Compound/Protein similarity tab is selected. In these tabs, the user will be able to upload user-specified proteins and compounds in order to make predictions on possible interactions and in turn investigate its structural similarity to structures present in the searchspace. The nearest neighbours tab would display a user-specified compound/protein's nearest neighbours and display their interaction profile, whilst the interaction prediction tab would focus on the clustering methods described in Chapter 5 to build an interaction profile of a custom compound or protein.

PubChem Compound ID	Similarity Score	Compound Present in DrugBank?	Compound Present in ToxCast?	Interacts With
1676	0.91	Yes	No	P30542 P29275 ....
...				
...				
...				

Figure 6.3: Design of the Nearest Neighbour Tab

UniProtID	Raw Cluster Score	Projected Cluster Score
P43119	3.73	0.99
...		
...		
...		

Figure 6.4: Design of the Interaction Prediction Tab

#### 6.4.2.1 Nearest Neighbours

To calculate the nearest neighbours for a user-supplied compound or protein, the user needs the ability to upload their own structures to the app. Once uploaded, the server loads the file into the appropriate format and performs the similarity operations. As there was a selection of comparison techniques used for the investigation in Chapter 5, the design of the app would provide the ability for the user to select a comparison technique. This would mean selecting either the PubChem or MACCS fingerprint formats for compounds, and either the Smith and Waterman or BLAST algorithms for protein similarity.

Whilst information on similar compounds and proteins is informative, other information can be retrieved to provide further benefit to a user. In the case of compounds for example, a compound with high similarity can have an interaction profile extracted from the repositories to provide the user with likely protein candidates of interaction. This can also be further expanded to include whether or not the nearest neighbour compound contains appropriate information in certain repositories. For example, a compound with high similarity linked to DrugBank could be connected to an existing approved drug, whereas a compound with high similarity linked to



ToxCast has been connected to toxicity screening and testing. Whilst such a format is more limited than that provided by the classification models developed during Chapters 3 and 4, this information can still provide useful information as to what extent a candidate compound settles in certain repositories.

### 6.4.2.2 Interaction Prediction

While nearest neighbours provide interaction information on compounds and proteins similar to the user-supplied entry uploaded by the user, this information is limited to individual compounds, which can make aggregation of a complete interaction profile difficult. The clustering methods investigated in Chapter 5 consider all compounds and proteins to output a prediction profile of a compound/protein, and the method's prediction of likelihood of an interaction taking place.

The process of preparing a compound or protein for interaction is similar to that of the Nearest Neighbours in terms of uploading a structure and selecting a comparison method, however there are a number of additions to the input design to incorporate the prediction pipeline. One of these additions is the selection of a clustering method, such as WNN-GIP, NRLMF and so on. Another addition is the selection of other similarity comparison techniques to conduct further analyses, such as the selection of the protein comparison method if a user is uploading a user-specified compound, and vice versa.

## 6.5 Implementation

### 6.5.1 Compound Browser

To implement the Compound Browser, the interface was required to generate result tables based on user input. To accomplish this, the R Package "Shiny" makes use of reactive variables, which perform actions on the server in the event that a variable on an input field has been changed. For the purposes of the Compound Browser, that would mean that in the event the UniProt accession code or assay specific values had changed, this would trigger a variable change event on the server and then generate a new result table for the user. An example of a reactive variable is shown on Figure 6.5, which alters the DrugBank target list according to the UniProt accession code selected by the user. In addition to this, the target list is modified to provide URLs of elements which are accessible via their repository's websites, allowing a user to extract further information if needed.

To download a table, "Shiny" has a DownloadHandler element which triggers a download event on the user's browser, using the results table that is generated and currently visible on the user's page. To ensure the table is easily accessible for analysis, URLs are not integrated on downloaded tables. Figure 6.6 displays a screenshot of the completed Compound Browser, showing all the input fields that can be selected. In the event the user enters information that does not exist in the selected repository, an empty result table will be generated. To minimise the potential for empty tables to be generated, the selection fields for each repository have been restricted to the fields that are only present in that repository. For example, in the case of ToxCast assay results, values are restricted to AC50 assay results, which means that no Assay Type filter has been included as shown in Figure 6.7. Another example is with regard to UniProt Accession codes, of which the options available to filter the result tables to are restricted to the proteins present in the user's currently selected repository.

```
filteredDB <- reactive({
  #If input is empty return empty table
  if(is.null(input$proteinDBInput))
  {
    return(NULL)
  }
  DrugBank %>%
  #DrugBank ID
  mutate('DrugBank ID' = paste0('<a href="https://www.drugbank.ca/drugs/',
    DrugDatabaseID,' " target="_blank">',DrugDatabaseID,'</a>',sep="")) %>%
  #UniProt ID
  mutate('Uniprot Accession Code' = paste0('<a href="https://www.uniprot.org/uniprot/',
    UniprotID,'" target="_blank">',UniprotID,'</a>',sep="")) %>%
  #PubChem ID
  mutate('PubChem Compound ID' = paste0('<a href="https://pubchem.ncbi.nlm.nih.gov/compound/',
    CompoundID,'" target="_blank">',CompoundID,'</a>',sep="")) %>%
  #Search Term
  filter(UniprotID == input$proteinDBInput) %>%
  #Select Fields for Display
  select('DrugBank ID','Uniprot Accession Code','PubChem Compound ID')
})
```

**Figure 6.5:** Example of a reactive variable in the Interface server code. This example generates a target list from DrugBank whenever the user selects a UniProt Accession code from a selection bar

Compound Interaction    Compound/Protein Similarity    Help/About

### Compound Interaction Search

UniProt Accession Code  
P10275 (ANDR\_HUMAN)

Assay Type

ic50  
 ki  
 ec50  
 kd  
 potency  
 ac50

Min um Assay Value  
10

Max um Threshold Value (Capped at 1 mol)  
100

[Download](#)

DrugBank    ToxCast    **ChEMBL**    BindingDB

ChEMBL ID	UniProt Accession Code	PubChem Compound ID	Standard_Type	Assay Value	Standard
CHEMBL516752	P10275	44568330	ic50	10	um
CHEMBL1830770	P10275	11551147	ic50	10	um
CHEMBL3695597	P10275	58115334	ic50	10	um
CHEMBL224278	P10275	44422052	ic50	10	um
CHEMBL1835807	P10275	56589543	ic50	10	um
CHEMBL454138	P10275	10026128	ic50	10	um
CHEMBL1086086	P10275	46831697	ic50	10	um
CHEMBL3658381	P10275	59612698	ic50	10	um
CHEMBL188261	P10275	44397094	ic50	10.4	um
CHEMBL3622575	P10275	70806890	ic50	10.5	um
CHEMBL1372017	P10275	673702	ic50	10.6	um
CHEMBL2346971	P10275	44125830	ic50	10.7152	um
CHEMBL3622589	P10275	122191906	ic50	10.9	um

**Figure 6.6:** Screenshot of the Compound Browser when ChEMBL is selected. The results in the table displayed on the right are from results selected by the input fields entered on the left. Input fields on the left are generated according to the repository selected; as the selected ChEMBL repository contains a variety of assay results, assay types and threshold value fields are generated appropriately.

Compound Interaction    Compound/Protein Similarity    Help/About

### Compound Interaction Search

UniProt Accession Code  
P10275 (ANDR\_HUMAN)

Min um Assay Value  
10

Max um Threshold Value (Capped at 1 mol)  
100

[Download](#)

DrugBank    **ToxCast**    ChEMBL    BindingDB

UniProt Accession Code	PubChem Compound ID	Standard_Type	Assay Value	Standard_Units
P10275	62311	ac50	10.0627	um
P10275	8571	ac50	10.2362	um
P10275	33565	ac50	10.4029	um
P10275	6	ac50	10.4068	um
P10275	86173	ac50	10.4381	um
P10275	44235958	ac50	10.4383	um
P10275	8478	ac50	10.5385	um
P10275	60196435	ac50	10.5424	um
P10275	8606	ac50	10.5578	um
P10275	44235958	ac50	10.5701	um
P10275	8607	ac50	10.5778	um
P10275	25015677	ac50	10.5932	um
P10275	5757	ac50	10.6069	um

**Figure 6.7:** Screenshot of the Compound Browser when ToxCast is selected. As ToxCast is limited to AC50 assay results, the input fields on the left do not provide an option for filtering to specific assay types as is the case for BindingDB and ChEMBL.

### 6.5.2 Nearest Neighbours Predictions

For nearest neighbours of compounds and proteins to be processed, the server needed the ability for users to upload files to the server where the app was resided. To allow file uploads, "Shiny" provides developers with a FileInput field, on which a file upload will store a file in a temporary directory attached to the user. When a user disconnects from the server by closing the interface window, the temporary directory is deleted to free server space. These file input fields can be restricted to certain file extensions to prevent accidental upload of erroneous files. In the case of uploading a user-supplied compound, comparisons can be made with either the compound's structure file or the SMILES code, meaning that the file input field for a user-supplied compound is restricted to SDF files or text files containing a SMILES code. For user-supplied proteins, the file input field was restricted to FASTA format files. To initiate a nearest neighbour comparison event, the process of triggering the server is slightly different to that of the Compound Browser as all inputs are required to be selected before computation can commence. Instead of using a reactive variable to respond to a change in an input field, "Shiny" allows a developer to associate processes with the execution of an event with the interface, such as the user clicking a button. This means that a button can be used to execute the nearest neighbour comparison based on the input the user has selected.

With regard to the event of comparing nearest neighbours, the app required the ability to call on external programmes, as not all comparison methods were able to be executed from inside R's environment. Whilst compound comparison methods and the Smith and Waterman algorithm for proteins could be contained in the R workspace for calculation, the BLAST algorithm for comparing proteins could only be accessed via command prompt and also required an external file for querying a custom protein against a set. The process needed to satisfy the requirements of BLAST was two-fold: first, on provision of a user-specified protein, an external file was created and provided with a temporary id which cannot be used by any other user until the operation was completed. Once the temporary file is compiled, R can make use of a function called "system", which allows a developer to call scripts and pass arguments outside the R environment. To reduce the need for storing a temporary result file, the "system" command could also store output which was generated by the command line. Once a user has ended their session by closing the browser window with the interface, any files which have been uploaded by the user are deleted from the system. Files which are also compiled for use by other systems (i.e. Python's PyDTI library) are also removed once the script has concluded.

Figure 6.8 displays an example of the nearest neighbour tab when a custom compound has been assigned. To provide information beyond the similarity results as described in the Interface design, the elements in the Compound Browser were filtered to gather results and targets of interest for a similar compound. To provide additional information on a custom protein, the result will specify the nearest neighbour's presence in either Panel 44, Panel 331 or the Pharmacology panel. Whilst the target lists present in DrugBank could be confirmed as an active protein-compound pairing, the definition of an active protein-compound pairing could depend on the assay results from ChEMBL, ToxCast and BindingDB. For the purposes of the prototype, a number of conditions were applied to the information from the Compound Browser to reduce the overall size and processing time needed to discover the nearest neighbours and perform the interaction predictions via clustering. The conditions used are listed below:

- Compounds must originate from either DrugBank or ToxCast, however interactions can be referenced from ChEMBL and BindingDB
- Interactions are limited to the proteins contained in Panel 44, Panel 331 and the Pharmacology Panel.
- An interaction is defined as a compound-protein interaction listed in DrugBank, or as an assay result present in ToxCast, ChEMBL or BindingDB provided a molar concentration has been documented as less than 10 micromolar

The screenshot shows a web interface for 'Compound/Protein Similarity Target Prediction'. On the left, there is a form for uploading a custom compound. The 'Custom Candidate Type' is set to 'Custom Compound'. The 'Upload Custom Compound (.sdf or .smi)' section shows a file named 'TestComp3.sdf' has been uploaded. The 'Fingerprint Comparison Method' is set to 'PubChem'. Below the form are buttons for 'Calculate Nearest Neighbours' and 'Download Nearest Neighbours'. On the right, there is a table with two tabs: 'Nearest Neighbours' (selected) and 'Interaction Prediction'. The table has four columns: 'PubChem Compound ID', 'Similarity Score', 'Compound Present In DrugBank?', and 'Compound Present In ToxCast?'. The table contains three rows of data:

PubChem Compound ID	Similarity Score	Compound Present In DrugBank?	Compound Present In ToxCast?
1676	0.91	TRUE	FALSE
4740	0.90	TRUE	FALSE
3182	0.87	TRUE	FALSE

**Figure 6.8:** Screenshot of the Nearest Neighbour Predictions tab with a custom compound. The table not only provides a list of the most similar compounds, but also provides their interaction profile and specifies whether or not the compound is present in DrugBank or ToxCast

### 6.5.3 Interaction Prediction

To predict interactions as per the analysis of Chapter 5, the interface required a means to store similarity and interaction matrices with the custom compound or protein incorporated into the set. This required expansion of the nearest neighbour calculations is managed through appending the results to the similarity matrix requiring modification, and assigning a custom ID to the user-supplied compound or protein uploaded. Once compiled, the modified similarity matrix is then be written to temporary memory on the server along with the interaction matrix (containing a blank interaction profile for the custom compound or protein), and a similarity matrix where a custom structure has not been provided (so if a custom compound was provided the user would select either a Smith and Waterman or BLAST similarity methods for proteins). For the purposes of the prototype, the similarity and interaction matrices against which custom structures could be compared against was restricted to one panel, which was the combination of Panel 44, Panel 331 and the Pharmacology Panel. This combined matrix of all proteins provided the widest possible searchspace for demonstrating the features of the interface whilst reducing its visual complexity.

Once these user-supplied structures had been written, the PyDTI library could be called to access this temporary panel and make predictions. As the PyDTI library was built to output the top predictions from each method, an addition was made to the python library to filter new interactions to the custom compound or protein supplied by the user, instead of filtering the list of new predictions to a certain number. Although there are R libraries available which can call and store python functions and variables in the R workspace [4], the library is restricted to use of a single CPU core. This impacts the library's performance, as the numeric analysis platform Numpy used by the PyDTI library can make use of multiple cores for their calculations if they are present on the hardware.

Figure 6.9 gives an example view of the prediction interface when a custom compound has been supplied. The first column reflects the raw cluster score which was provided by the library clustering method. The scale and range of values can vary according to the function, so to provide some consistency for all methods, a normalization function was applied to distribute all the predictions to a 0 to 1 scale, where 1 indicates an extremely likely interaction, and 0 indicates an extremely unlikely interaction. The PyDTI authors had implemented a function which normalized the predictions made by their own method, NRLMF, for which equation 6.1 was used, where  $x$  indicates the prediction scores of the method.

The equation used by NRLMF however was not practical to use for methods where

extremely small cluster values were generated. For example, the NetLapRLS method generated small raw cluster values for all predictions made, and applying the above formula would have led to a projected value of 0.5 for all predictions, which would make differentiating high scoring and low scoring predictions difficult. Therefore an alteration was made to the normalization formula which was applied to all methods except NRLMF, shown in equation 6.2, where  $\bar{x}$  is the mean of the predictions, and  $|x|$  is the standard deviation of the predictions. This alters the normalization according to the overall distribution of the method's predictions, in turn providing an easier scale of comparison of high scoring and low scoring predictions for a particular custom compound or protein.

$$Norm(x) = \frac{1}{1 + e^{-x}} \quad (6.1)$$

$$Norm(x) = \frac{1}{1 + e^{-(x-\bar{x})/|x|}} \quad (6.2)$$

Compound Interaction | Compound/Protein Similarity | Help/About

Compound/Protein Similarity Target Prediction

Custom Candidate Type: Custom Compound

Upload Custom Compound (.sdf or .smi): TestComp3.sdf (Upload complete)

Fingerprint Comparison Method: PubChem

Target Prediction Clustering Method: NetLapRLS (selected)

Protein Comparison Method: BLAST

Uniprot Accession Code	Raw Cluster Score	Projected Cluster Score
P33261	1.62e-07	0.57
P08684	1.38e-07	0.54
P03372	1.34e-07	0.54
P05177	1.23e-07	0.53
P11712	1.23e-07	0.53
P10635	1.08e-07	0.51
P05121	9.8e-08	0.5
Q4U2R8	9.64e-08	0.5
Q03405	9.05e-08	0.5
P10275	8.34e-08	0.49
P13500	8.27e-08	0.49
P37231	7.8e-08	0.48
P35367	7.47e-08	0.48
O75469	7.11e-08	0.48

**Figure 6.9:** Screenshot of the Interaction Prediction Tab with a custom compound. This table provides the score of the cluster method, as well as a score which attempts to normalize the score to a 0 to 1 scale.

Although the interface successfully conveyed predictions, an issue was detected when multiple users tested the interface. In the event that a reactive variable has been triggered and requires a long computation time, the "Shiny" package can halt other processes. This meant that if a new user triggered an interaction prediction, the interface would freeze for other users until the interaction prediction function had been completed. This issue had not been detected during the testing of the Compound Browser and Nearest Neighbour tabs due to the short processing time

needed to generate the required results for users.

To circumvent this issue, a data analysis company FellStat proposed a solution which made use of a package referred to as "future" [5] [6]. This package creates a separate process within R to calculate the predictions, and assigns a "promise" to the interface that the result table will be compiled in the future. This "promise" would prevent the server from halting other processes in the event a time consuming task is currently in progress. When the separate process has been completed, the package would then generate a results table for display on the website. A limitation to the use of a separate process however is that tracking the progress of the prediction is difficult if the majority of the process is outside the R environment.



## 6.6 Discussion

The implementation of a user interface has provided the ability for users to quickly and easily view and utilise elements of the analyses undertaken throughout this work, in turn avoiding the requirement for coding experience (needed to make use of source code examples), or the use of multiple websites. The platform not only holds a considerable wealth of aggregated interaction information, but it also provides users with the means of uploading their own compound and protein structures and in return receiving valuable target predictions, as well as the interaction profiles of similarly structured candidates.

In comparison to other publicly available interfaces, the prototype presented provides additional functionality and comprehensiveness, however there are some elements that can be further improved upon to strengthen the platform's features. For example, in the case of the PubChem-informed screen as shown in Figure 6.10, a similarity search can be performed but the results do not indicate interactions associated with similar compounds unless individual records are accessed. Furthermore the precise degree of similarity is not clearly displayed to indicate how similar a structure is to the compound used in the query, but a threshold can be applied to limit the results to compounds beyond a certain degree of similarity (which by default is set to 90% by the platform). The DrugBank-informed screen shown in Figure 6.11 also provides a facility which allows users to upload and compare user-supplied compounds, and in turn has provided further filtering options such as weight and drug category, however the method of designation of similarity is not specified and the platform at present does not allow the comparison of protein sequences.

In terms of other clustering similarity projects which have provided an interface, one project referred to as DrugE-Rank allowed users to perform multiple similarity clustering operations on user-supplied compounds and proteins at once, but did not provide a means of displaying the raw value output from each method, or a means of specifying the type of comparison to be performed on a user-supplied compound or protein [7]. The DrugE-Rank interface did however allow users to supply a DrugBank identifier or UniProt accession code in lieu of a structure, which could be incorporated as a means of input for future revisions of the interface developed in this work.

Although this demo only contains a small search space, the implementation process has established the foundations for further development, and in turn the process has highlighted areas for further expansion and development to further widen the scope, functionality and the presentation of these promising approaches and findings

to the wider public. An online version of the interface prototype is available at <http://proteins.swan.ac.uk/cheminf>.

The screenshot shows the PubChem Similarity Search Platform interface. At the top, there is a search bar with the text "CID2244 structure" and a search icon. Below the search bar, there are tabs for "Identity (1)", "Similarity (>1,000)", "Substructure (>1,000)", and "Superstructure (>1,000)". A "TANIMOTO THRESHOLD" slider is set to 90%, with a range from 1% to 100%. Below the slider, there is a "Percentage of the database searched: 20%" indicator and a "Search All" checkbox. The results section shows "1,000 results (incomplete)" and a "SORT BY" dropdown set to "Relevance". Two results are displayed: "Aspirin; ACETYSALICYLIC ACID; 50-78-2; 2-Acetoxybenzoic Acid; 2-(Acetoxy)benzoic Acid; ..." and "Methyl Salicylate; Methyl 2-hydroxybenzoate; 119-36-8; Wintergreen Oil; Gaultheria Oil; ...". Each result includes its Compound CID, molecular formula (MF), molecular weight (MW), InChIKey, IUPAC Name, and Create Date. On the right side, there are "ACTIONS ON RESULTS WITH ID TYPE" for "CID - Compounds", including "Push to Entrez", "Save for Later", and "Linked Data Sets".

**Figure 6.10:** Screenshot of PubChem's Similarity Search Platform [8]. Structures within PubChem can be compared to other structures within the repository, and filtered to a certain threshold, however further information is limited to the individual record pages.

The screenshot shows the DrugBank Chemical Structure Search platform interface. At the top, there is a navigation bar with "DRUGBANK" and links for "Browse", "Search", "Downloads", and "Commercial Data". Below the navigation bar, there is a "Chemical Structure Search" section. The main interface is divided into "STRUCTURE SEARCH" and "MOLECULAR WEIGHT" tabs. The "STRUCTURE SEARCH" tab is active, showing a "Search by structure" section with a chemical structure editor. The editor contains a complex chemical structure with a piperazine ring, a sulfonamide group, and a pyridine ring. To the right of the editor, there are "Search Options" including "Similarity", "Substructure", and "Exact" radio buttons. The "Similarity" option is selected. Below the radio buttons, there are input fields for "Similarity threshold" (set to 0.7), "Minimum Weight" (set to e.g. 100), and "Maximum Weight" (set to e.g. 200). There is also a "Maximum Results" dropdown set to 100. At the bottom, there are checkboxes for "Drug Types (default all):" including "Approved", "Vet approved", "Nutraceutical", "Illicit", "Investigational", and "Experimental". A "Search" button and a "Load example" button are also present.

**Figure 6.11:** Screenshot of the DrugBank Chemical Structure Search platform [9]. Structures can either be drawn or uploaded to the platform and compared against other drugs within DrugBank.

### 6.6.1 Further Work

While the interface provides a user with the means of compiling an interaction profile for a user-supplied compound or protein, there are several areas that may be identified for further improvement. One such example is to improve the Compound Browser to allow the querying of more than one protein accession code, allowing users to construct specialised interaction panels based on results in certain repositories. This could be further expanded on the interface design to allow users to create specialised interaction and similarity panels for use in the PyDTI clustering methods.

Another area for further implementation is the incorporation of the the *in silico* pipeline used in Chapter 5, which at present requires a degree of manual preparation for structures. Although the classifier models used in Chapters 3 and 4 could in theory be applied to predicted interactions above a certain threshold score, further work is needed in verifying the findings made by the interaction predictions to ensure that genuine targets are being identified before the profiling classifier models can be implemented. An R package known as "RWeka" allows the use of WEKA classifiers in the R workspace [10], meaning it would be possible to train and evaluate classifiers on predefined interaction sets.

Another area of note is that while the current format and use of saved workspaces is practical for small scale datasets, a larger platform would require some alteration to the design and infrastructure to reduce the amount of resources used by a system, and to reduce load and use time by users. Ideally a database platform would need to be implemented to allow users to save custom panels and results to the server to reduce the requirement for users running the same experiment repeatedly or to have to download and filter results manually. Further investigation of database security methods would also be necessary for a database platform to be implemented on a public platform.

Finally, one limitation of the "Shiny" package is that when no connection has been established by a user, the server may close down the R workspace to reduce load on the deployed server. This leads to increased load times if the interface has not been accessed in some time due to the workspace being reloaded. A configuration option is however available in the "Shiny" server setup files that allows a user to alter or disable this timeout. A script could be implemented to allow the server to be restarted once this shutdown threshold has been reached and to perform certain "housekeeping" tasks, such as checking for updated workspaces and files and the removal of temporary files.

## 6.7 References

- [1] Wishart, D. S., Knox, C., Guo, A. C., Sharavista, S., Hassanali, M., Stothard, P., Chang, Z., and Woolsey, J., “DrugBank: A comprehensive resource for in silico drug discovery and exploration,” *Nucleic Acids Research*, vol. 34, pp. D668–D672, Jan. 2006.
- [2] Sipes, N. S., Martin, M. T., Kothiya, P., Reif, D. M., Judson, R. S., Richard, A. M., Houck, K. A., Dix, D. J., Kavlock, R. J., and Knudsen, T. B., “Profiling 976 toxcast chemicals across 331 enzymatic and receptor signaling assays,” *Chemical research in toxicology*, vol. 26, no. 6, pp. 878–895, 2013.
- [3] Chang, W., Cheng, J., Allaire, J., Xie, Y., and McPherson, J., *Shiny: Web application framework for r*, R package version 1.3.2, 2019. [Online]. Available: <https://CRAN.R-project.org/package=shiny>.
- [4] Ushey, K., Allaire, J., and Tang, Y., *Reticulate: Interface to 'python'*, R package version 1.13, 2019. [Online]. Available: <https://CRAN.R-project.org/package=reticulate>.
- [5] FellStar. (Jul. 30, 2018). “Long running tasks with shiny: Challenges and solutions,” [Online]. Available: <http://blog.fellstat.com/?p=407> (visited on 08/15/2019).
- [6] Bengtsson, H., *Future: Unified parallel and distributed processing in r for everyone*, R package version 1.14.0, 2019. [Online]. Available: <https://CRAN.R-project.org/package=future>.
- [7] Yuan, Q., Gao, J., Wu, D., Zhang, S., Mamitsuka, H., and Zhu, S., “Druge-rank: Improving drug–target interaction prediction of new candidate drugs or targets by ensemble learning to rank,” *Bioinformatics*, vol. 32, no. 12, pp. i18–i27, 2016.
- [8] National Center for Biotechnology Information. (2019). “PubChem Similarity Search,” [Online]. Available: <https://pubchem.ncbi.nlm.nih.gov/#query=CID2244%20structure&tab=similarity> (visited on 09/20/2019).
- [9] Wishart, D. S., Knox, C., Guo, A. C., Sharavista, S., Hassanali, M., Stothard, P., Chang, Z., and Woolsey, J. (2019). “Chemical Structure Search - DrugBank,” [Online]. Available: [https://www.drugbank.ca/structures/search/small\\_molecule\\_drugs/structure](https://www.drugbank.ca/structures/search/small_molecule_drugs/structure) (visited on 09/24/2019).

- [10] Hornik, K., Buchta, C., and Zeileis, A., “Open-source machine learning: R meets Weka,” *Computational Statistics*, vol. 24, no. 2, pp. 225–232, 2009. DOI: 10.1007/s00180-008-0119-7.

# Chapter 7

## Discussion

### 7.1 Introduction

Overall, this study has investigated and utilised a number of data-based avenues in the compound discovery and *in silico* screening fields, handling not only a sizeable aggregated collection of interactions across several repositories, but also a collection of experimental assay results which can be further analysed, filtered and adapted for further studies. The purpose of this final discussion is to reflect upon the knowledgebase and literature exploited in the project, and to specify the limitations and further areas for improvement for future studies within this area.

### 7.2 Objective Review

As part of the discussion process, the objectives presented in Chapter 1 are revisited and progress reviewed. This section will summarise the key findings of the work described in each chapter and discuss the overall advances made.

#### 7.2.1 Objective 1 Review

The first objective of the project was to **explore the history and techniques of computational drug discovery and development**. Through a review of the literature, it was confirmed that the drug development process can be costly, both in terms of the outlay money required [1], and in terms of attrition rates where drug development had been abandoned due to increased risks or hazards detected during testing [2]. This formed one of the drivers for this study, to investigate and develop approaches that could potentially complement the drug development process and in turn reduce attrition rates.

This led to a review of publicly available drug and protein repositories, of which a number were used throughout the project. One of these was DrugBank [3], which contained sufficient information to form a good foundation for identifying compound-protein pairings, as well as providing a source of FDA-approved drugs which could be used as a reference for what could be considered a beneficial or "good" interaction profile for candidate compounds. The DrugBank repository has since expanded and improved, with the 5th revision [4] not only providing cross references to other repositories (circumventing the need to run cross-referencing operations when aggregating information from different sources), but also through the release of a platform referred to as DrugBankPlus [5], which is a commercial platform directed towards providing pre-prepared formats of DrugBank suitable for machine learning uses. These improvements highlight the efforts repository holders are making towards the amenability of their sources for machine learning approaches.

The ToxCast repository on the other hand provided an ideal reference for interaction profiles which could be considered harmful [6], as the repository's purpose was to measure toxicity of compounds against a panel of proteins. This was then used as a reference for what could be considered a "bad" interaction profile for candidate compounds. This repository has also been received a major update, with Version 4 of the database released in August 2019 [7]. Other repositories such as ChEMBL [8] and BindingDB [9] supplemented the interaction profiles of compounds present within ToxCast and DrugBank, through documenting potency assay results ( $K_i$ ,  $K_d$ ,  $IC_{50}$  etc.) that could be filtered to define what constitutes an active compound-protein pairing. ChEMBL in turn has also received an update since information was last extracted [10], whilst BindingDB is updated on a monthly basis and warrants continual revisiting for addition of new data points.

On reflection of the repositories utilised however, we had perhaps attempted to gather too much information in the process of building the DrugReferenceDatabase, which attempted to aggregate all interactions found into a single accessible point instead of querying multiple sources. Some repositories which were used in initial approaches were later considered to be incompatible with other repositories going forward. One such repository was the PubChem BioAssay platform [11], which contained a large number of conflicting information elements in assignment of activity of particular compound-protein pairs. The retrieval method used in the project to access assay results of particular proteins generated inconsistent potency types and unit measurements, information which was needed to determine which compound-protein pairings could be classified as active or inactive. Further investigation of the

platform discovered that it would be necessary to parse through Bioassay's XML files to obtain these results, which was considered impractical due to the scale of information which is present within the files. There has however been development of a number of approaches and platforms that attempt to retrieve the information from BioAssay's XML files, such as a package in R developed by Backman *et al.* [12], which captures information relating to small molecules recorded in the BioAssay platform into the R workspace.

Other repositories found to be unsuitable for further interaction pre-processing towards the later stages of the project also included CTDBase [13], which did not specify clearly the type or strength of the interactions which took place between proteins and targets, that made the process of filtering the results impossible for thresholding. This database however provided a valuable link to proteins involved in specific diseases in the implementation of a potential filter for searching proteins and compounds implicated in specific diseases for further investigation. Issues were also present for the Matador database, where the confidence metrics and scores were difficult to use to determine activity strength of a compound-protein pairing, and thus was incompatible with thresholded assay results.

Despite these obstacles, the DrugReferenceDatabase platform that emerged from the investigation of repository sources has generated a substantial combined repository for multiple compound-protein links to be extracted, in addition to establishing an initial schema which could be further expanded to accommodate more assay results in future revisions, and populate additional datapoints from future releases of repositories. Further work into considering modes of action, dosage and assay types would also improve the quality of links that have been generated, however investigation would be necessary to determine if these types of filters would be compatible across all repositories utilised in this thesis.

### 7.2.2 Objective 2 Review

The second objective of the project was to **investigate and evaluate machine learning techniques applied to *in silico* drug screening**. The main goal of this objective was to determine if this approach could be applied to determine whether a candidate drug could be designated as having a good (i.e approved drug from DrugBank) or a bad (i.e sourced from ToxCast) profile. To allow results to be easily reproduced, the WEKA system had been used which provided a collection of classifiers suitable for constructing profiler models [14].

The initial approach had considered the non-conflicting active and inactive



compound-protein pairs from the DrugReferenceDatabase, in addition to consideration of chemical properties obtained through a chemical structure parser known as "Mordred" [15]. These interaction training sets were focused on three panels which have a history in pharmacological profiling: Panel 44 [16], Panel 331 [17], and the DrugBank Pharmacology Panel. The classifier models trained on this information had generated somewhat disappointing results, with models being unable to distinguish test candidates introduced from HMDB, which contained interactions of human metabolites, and T3DB, which contained interactions involving known toxic compounds. It was concluded on further investigation of these results that the poor prediction accuracy levels could have been caused by the high degree of conflicting definitions of activity from referencing multiple repositories.

These findings led to a separate approach from that of the interactions of the DrugReferenceDatabase, through the consideration and thresholding of experimental assay results and potencies (such as  $K_i$ ,  $IC_{50}$  etc.). While the available number of applicable interactions had decreased considerably from this approach, prediction levels had considerably improved. One result of particular note was the use of the J48 decision tree classifier with Panel 331 interactions. When these interactions were used in conjunction with compound chemical properties, the generated model when tested on HMDB and T3DB compounds provided an overall accuracy level of 93.5%, which is an exceptional outcome and a great improvement over the initial approach.

Although the analyses overall provided a promising first step in the development of a profiling tool, there are still a number of avenues that could be explored through further work around this objective. One such approach is to further manipulate a potency thresholding value that had been used to define activity of an assay result. The approach in this work made use of a 10 micromolar threshold in accordance with a study by Hughes *et al.* which assessed common definitions of potency [18], however alterations to this threshold value might improve profiling accuracy from further clarification of interacting compound-protein pairs.

Another avenue to be considered is the selection process for selecting all compounds under certain sources to be "Good" (DrugBank and HMDB compounds) or "Bad" (ToxCast and T3DB compound) references for the classifiers. This was a generalisation of course as most drug compounds are toxic at higher concentrations, some indeed have tolerable toxicity at therapeutic concentration; and many compounds regarded as toxins can be tolerable at low concentration. Further consideration should therefore have been made between the modes of action and dosage to further refine the classifications. However, there is some reasonable basis to the categorisations, as they

distinguish between two general groupings of compounds – those that are largely non-toxic at typical therapeutic dosage levels and those that are largely not. Further refinements in the selection process would however have provided the means of introducing additional training/testing candidates for classifiers from other sources.

Another area for further consideration is the consideration of alternative tools and classifiers that could be used on interaction datasets. In the case of WEKA, further classifiers could be introduced via the integrated Package Manager, such as CSForest which is a decision tree which is specifically designed to work on imbalanced datasets [19]. In terms of tools, the eToxPred platform by Pu *et al.* discussed during the exploration of machine learning approaches had achieved a high degree of accuracy in determining compound toxicity [20]. This tool could be used to assist in the reduction of conflicting information, such as determining if compounds residing in both DrugBank and ToxCast should be assigned to either a "Good" or "Bad" profile class, in addition to consideration of dosage. This in turn could provide additional training or test points to enrich the analysis. Furthermore there are other types of classifier that could be explored, such as in a survey by Stephenson *et al.* which highlighted that Deep Learning and Neural Network approaches appear to have increasing use in recent publications [21].

### 7.2.3 Objective 3 Review

The third objective of the project was to **investigate and evaluate machine learning techniques for the purpose of clustering similar drug compounds or protein targets**. A review of the literature revealed that two avenues were highly applicable to this aim: the first was to predict interactions via structural similarity clustering, through a "Read-Across" principle where a compound or protein belonging to a highly documented group of objects would share similar properties, typically applied with a mathematical model known as a QSAR (Quantitative Structure Activity Relationships) [22]. The second avenue involved the implementation of *in silico* docking pipelines, to simulate binding energies and locations between a compound and protein to determine the likelihood of activity between them. The similarity approaches considered made use of a Python library called PyDTI [23], which was developed by Liu *et al.*, but further similarity approaches could in turn be considered in further approaches. For example, as noted in the review in Chapter 4, a report by Ban *et al.* has provided an expansion to the similarity clustering method of NRLMF [23], which attempts to improve on target prediction performance for compounds with little recorded protein activity [24]. Furthermore, a report by

Kurgan *et al.* has surveyed and documented 35 approaches of similarity clustering, some of which contain source code which could be used for replication and further assessment and analysis [25].

Of the similarity clustering methods available within PyDTI, two (NRLMF [23] and NetLapRLS [26]) had managed to successfully predict approximately half of tested interactions, which were random *in vitro* interactions that had been "blinded" (removed as a target) from the interaction search space. These models were then assessed on new interaction predictions, however analysis of these predictions revealed two anomalies. The first anomaly was that a fairly large proportion (approximately 27%) of the high ranking positive predictions had been identified as inactive based on referencing protein-compound assay results which did not meet the 10 micromolar potency threshold, whilst low ranking predictions did not highlight as many inactive results as expected. This further supports the idea that a re-evaluation of the threshold may be necessary in a future approach to observe the impact additional data points might have on widening the activity search space. When other models had been investigated for potential false positives, the CMF method [27] had flagged the least number of inactive assay results, which warranted further investigation of *in silico* docking for these predictions.

The second anomaly was the binding energies generated by an *in silico* docking pipeline [28], that had been used to attempt to identify if high and low scoring compound-protein predictions were appropriately active or inactive. On analysis of the *in silico* pipeline results, only one method (NRLMF) had shown agreement between the pipeline's high binding energies (indicative of a likely active protein-compound pairing), and the clustering method's top predictions. This finding highlights some promise in combining structural clustering and *in silico* docking in unison, however a further approach might be considered to further verify top scoring interactions. Van Laarhoven *et al.* for instance in his similarity clustering approach assessed new interaction predictions against updated revisions of compound-protein interaction repositories [29], and found that approximately 21% of the top 100 predictions made via WNN-GIP were found in an updated revision. This technique could also be applied in future approaches now that several major revisions have taken place on a good proportion of the repositories assessed.

Overall, the analyses had revealed a number of promising avenues, and in order to provide users with the ability to make use of the similarity clustering pipeline, a web interface platform prototype was implemented and released. The platform, named TargetPredict, not only allows users to browse and download the assay results

gathered from the various repositories throughout the project, but also provides the ability to generate predicted interaction profiles of user-selected proteins and compounds. This interface, with further modifications and expansion of features and interaction search space, could have the potential to provide users who have limited or no programming experience to make use of the rich data incorporated in a single combined platform, in an area of research which is currently limited by the need to engage with separate platforms across multiple websites.

### 7.3 Further Work

In terms of further work for the wider project approach, it should be emphasised that although a significant amount of information has been integrated, the results that have been presented thus far are based only on a small proportion of the overall interaction search space, and that a large area still remains unexplored. As future revisions of repositories will provide additional datapoints, it will become necessary to implement a revised database schema, in order to store and utilise the additional information efficiently. Future revisions of the database would in turn require adoption of suitable security measures, to ensure the database and server is secure when used in conjunction with the TargetPredict platform.

Another area for potential future incorporation is to consider the proteins of other species in the analyses. The current retrieval approaches have either been restricted to a single species, or to protein panels composed to be relevant to pharmacological profiling or toxicity prediction. With additional time, the analyses can be expanded to investigate and potentially include other species commonly used in *in vitro* testing, which would be useful for application in the pharmaceutical industry. The introduction of these species could also introduce proteins that share sequence similarity to certain human protein sequences, and in turn enrich the similarity clustering pipeline.

Finally, the design and implementation of a queueing and logging system should be considered, so that potentially time consuming tasks can be performed via the interface. As some of the tasks performed can take a long period of time to execute even on a small dataset, the increase of the search space would further increase the execution time and make the current requirement of a user remaining on the interface platform impractical. A queueing system would reduce the resource load on the hardware where the interface is implemented, as well as provide the ability for a user to submit a job to the TargetPredict platform and return to retrieve results after a period of time has elapsed, or upon notification to the user. Further investigation

would be necessary into the use of an email system to assist with the notification process.

## 7.4 Concluding Remarks

Overall, the research has generated two notable computational tools which have aggregated and compiled a vast amount of information across multiple sources into one consistent source, and provides the community with access to a wealth of standardised information in a single platform that may be applied to important biological and medical questions without requiring a degree of coding experience.

This wealth of information provides a powerful source of input for predictions when used in conjunction with the promising findings of similarity clustering and compound profiling, and the developments made so far have provided a foundation for further development and expansion with the potential of providing a valuable tool to complement the compound discovery pipelines. The interface is freely available for use at <http://proteins.swan.ac.uk/cheminf/>.

## 7.5 References

- [1] Sertkaya, A., Wong, H.-H., Jessup, A., and Beleche, T., “Key cost drivers of pharmaceutical clinical trials in the united states,” *Clinical Trials*, vol. 13, no. 2, pp. 117–126, 2016.
- [2] Hay, M., Thomas, D. W., Craighead, J. L., Economides, C., and Rosenthal, J., “Clinical development success rates for investigational drugs,” *Nature Biotechnology*, vol. 32, no. 1, pp. 40–51, Jan. 2014.
- [3] Wishart, D. S., Knox, C., Guo, A. C., Sharavista, S., Hassanali, M., Stothard, P., Chang, Z., and Woolsey, J., “DrugBank: A comprehensive resource for in silico drug discovery and exploration,” *Nucleic Acids Research*, vol. 34, pp. D668–D672, Jan. 2006.
- [4] Wishart, D. S., Feunang, Y. D., Guo, A. C., Lo, E. J., Marcu, A., Grant, J. R., Sajed, T., Johnson, D., Li, C., Sayeeda, Z., *et al.*, “Drugbank 5.0: A major update to the drugbank database for 2018,” *Nucleic acids research*, vol. 46, no. D1, pp. D1074–D1082, 2017.

- [5] OMx Personal Health Analytics. (Sep. 19, 2019). “DrugBankPlus — Scientific Drug Data,” [Online]. Available: <https://www.drugbankplus.com/> (visited on 09/19/2019).
- [6] Dix, D. J., Houck, K. A., Martin, M. T., Richard, A. M., Setzer, R. W., and Kavlock, R. J., “The ToxCast program for prioritizing toxicity testing of environmental chemicals,” *Toxicological Sciences*, vol. 95, no. 1, pp. 5–12, 2007.
- [7] United States Environment Protection Agency. (2019). “ToxCast and Tox21 Summary Files,” [Online]. Available: [https://epa.figshare.com/articles/ToxCast\\_and\\_Tox21\\_Summary\\_Files/6062479/4](https://epa.figshare.com/articles/ToxCast_and_Tox21_Summary_Files/6062479/4).
- [8] Gaulton, A., Hersey, A., Nowotka, M., Bento, A. P., Chambers, J., Mendez, D., Mutowo, P., Atkinson, F., Bellis, L. J., Cibrián-Uhalte, E., *et al.*, “The chembl database in 2017,” *Nucleic acids research*, vol. 45, no. D1, pp. D945–D954, 2016.
- [9] Gilson, M. K., Liu, T., Baitaluk, M., Nicola, G., Hwang, L., and Chong, J., “Bindingdb in 2015: A public database for medicinal chemistry, computational chemistry and systems pharmacology,” *Nucleic acids research*, vol. 44, no. D1, pp. D1045–D1053, 2015.
- [10] Mendez, D., Gaulton, A., Bento, A. P., Chambers, J., De Veij, M., Félix, E., Magariños, M. P., Mosquera, J. F., Mutowo, P., Nowotka, M., *et al.*, “ChEMBL: Towards direct deposition of bioassay data,” *Nucleic acids research*, vol. 47, no. D1, pp. D930–D940, 2018.
- [11] National Center for Biotechnology Information. (2018). “BioAssay Download FTP,” [Online]. Available: <ftp://ftp.ncbi.nlm.nih.gov/pubchem/Bioassay/> (visited on 12/12/2018).
- [12] Backman, T. W. H. and Girke, T., “Bioassayr: Cross-target analysis of small molecule bioactivity,” *Journal of chemical information and modeling*, vol. 56, no. 7, pp. 1237–1242, 2016.
- [13] Davis, A. P., Murphy, C. G., Saraceni-Richards, C. A., Rosenstein, M. C., Wiegers, T. C., and Mattingly, C. J., “Comparative toxicogenomics database: A knowledgebase and discovery tool for chemical–gene–disease networks,” *Nucleic acids research*, vol. 37, no. suppl\_1, pp. D786–D792, 2008.
- [14] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H., “The weka data mining software: An update,” *ACM SIGKDD Explorations Newsletter*, vol. 11, no. 1, pp. 10–18, 2009.

- [15] Moriwaki, H., Tian, Y.-S., Kawashita, N., and Takagi, T., “Mordred: A molecular descriptor calculator,” *Journal of cheminformatics*, vol. 10, no. 1, p. 4, 2018.
- [16] Bowes, J., Brown, A. J., Hamon, J., Jarolimek, W., Sridhar, A., Waldron, G., and Whitebread, S., “Reducing safety-related drug attrition: The use of in vitro pharmacological profiling,” *Nature Reviews*, vol. 11, no. 12, pp. 909–922, Dec. 2012.
- [17] Sipes, N. S., Martin, M. T., Kothiya, P., Reif, D. M., Judson, R. S., Richard, A. M., Houck, K. A., Dix, D. J., Kavlock, R. J., and Knudsen, T. B., “Profiling 976 toxcast chemicals across 331 enzymatic and receptor signaling assays,” *Chemical research in toxicology*, vol. 26, no. 6, pp. 878–895, 2013.
- [18] Hughes, J. P., Rees, S., Kalindjian, S. B., and Philpott, K. L., “Principles of early drug discovery,” *British journal of pharmacology*, vol. 162, no. 6, pp. 1239–1249, 2011.
- [19] Siers, M. J. and Islam, M. Z., “Software defect prediction using a cost sensitive decision forest and voting, and a potential solution to the class imbalance problem,” *Information Systems*, vol. 51, pp. 62–71, 2015.
- [20] Pu, L., Naderi, M., Liu, T., Wu, H.-C., Mukhopadhyay, S., and Brylinski, M., “E toxpred: A machine learning-based approach to estimate the toxicity of drug candidates,” *BMC Pharmacology and Toxicology*, vol. 20, no. 1, p. 2, 2019.
- [21] Stephenson, N., Shane, E., Chase, J., Rowland, J., Ries, D., Justice, N., Zhang, J., Chan, L., and Cao, R., “Survey of machine learning techniques in drug discovery,” *Current drug metabolism*, vol. 20, no. 3, pp. 185–193, 2019.
- [22] Cronin, M. T. D., “Chapter 1 an introduction to chemical grouping, categories and read-across to predict toxicity,” in *Chemical Toxicity Prediction: Category Formation and Read-Across*, The Royal Society of Chemistry, 2013, pp. 1–29, ISBN: 978-1-84973-384-7. DOI: 10.1039/9781849734400-00001.
- [23] Liu, Y., Wu, M., Miao, C., Zhao, P., and Li, X.-L., “Neighborhood regularized logistic matrix factorization for drug-target interaction prediction,” *PLoS computational biology*, vol. 12, no. 2, e1004760, 2016.
- [24] Ban, T., Ohue, M., and Akiyama, Y., “Nrlmf $\beta$ : Beta-distribution-rescored neighborhood regularized logistic matrix factorization for improving the performance of drug–target interaction prediction,” *Biochemistry and biophysics reports*, vol. 18, p. 100615, 2019.

- [25] Kurgan, L. and Wang, C., “Survey of similarity-based prediction of drug-protein interactions,” *Current medicinal chemistry*, Nov. 2018, ISSN: 0929-8673. DOI: 10.2174/0929867325666181101115314. [Online]. Available: <https://doi.org/10.2174/0929867325666181101115314>.
- [26] Xia, Z., Wu, L.-Y., Zhou, X., and Wong, S. T., “Semi-supervised drug-protein interaction prediction from heterogeneous biological spaces,” in *BMC systems biology*, BioMed Central, vol. 4, 2010, S6.
- [27] Zheng, X., Ding, H., Mamitsuka, H., and Zhu, S., “Collaborative matrix factorization with multiple similarities for predicting drug-target interactions,” in *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM, 2013, pp. 1025–1033.
- [28] Austin-Muttitt, K., “Modelling competition binding assays: A step towards in silico pharmacological profiling tools,” Report submitted to Swansea University, 2019.
- [29] Van Laarhoven, T. and Marchiori, E., “Predicting drug-target interactions for new drug compounds using a weighted nearest neighbor profile,” *PloS one*, vol. 8, no. 6, e66952, 2013.



# Appendix

The SQL database of the DrugReferenceDatabase, raw results from the findings of Chapters 3, 4 and 5, and the protein panels assessed are available on the flash drive attached to this thesis. Example structures and proteins are also included for use in the TargetPredict platform.

## Bibliography

- Aljumah, A. A., Ahamad, M. G., and Siddiqui, M. K., “Application of data mining: Diabetes health care in young and old patients,” *Journal of King Saud University-Computer and Information Sciences*, vol. 25, no. 2, pp. 127–136, 2013.
- All Wales Medicines Strategy Group. (2016). “About us,” [Online]. Available: [http://www.awmsg.org/awmsg\\_about\\_us.html](http://www.awmsg.org/awmsg_about_us.html) (visited on 02/04/2016).
- Anderson, D. P., “BOINC: A system for public-resource computing and storage,” in *Grid Computing, 2004. Proceedings. Fifth IEEE/ACM International Workshop on*, IEEE, 2004, pp. 4–10.
- Apache Friends. (2018). “XAMPP Installers and Downloads for Apache Friends,” [Online]. Available: <https://www.apachefriends.org/index.html> (visited on 08/14/2018).
- Austin-Muttitt, K., “Modelling competition binding assays: A step towards in silico pharmacological profiling tools,” Report submitted to Swansea University, 2019.
- Backman, T. W. H. and Girke, T., “Bioassayr: Cross-target analysis of small molecule bioactivity,” *Journal of chemical information and modeling*, vol. 56, no. 7, pp. 1237–1242, 2016.
- Ban, T., Ohue, M., and Akiyama, Y., “Nrlmf $\beta$ : Beta-distribution-rescored neighborhood regularized logistic matrix factorization for improving the performance of drug–target interaction prediction,” *Biochemistry and biophysics reports*, vol. 18, p. 100615, 2019.
- Bengtsson, H., *Future: Unified parallel and distributed processing in r for everyone*, R package version 1.14.0, 2019. [Online]. Available: <https://CRAN.R-project.org/package=future>.

- Boratyn, G. M., Camacho, C., Cooper, P. S., Coulouris, G., Fong, A., Ma, N., Madden, T. L., Matten, W. T., McGinnis, S. D., Merezhuk, Y., *et al.*, “BLAST: A more efficient report with usability improvements,” *Nucleic acids research*, vol. 41, no. W1, W29–W33, 2013.
- Borchers, A. T., Hagie, F., Keen, C. L., and Gershwin, M. E., “The history and contemporary challenges of the US Food and Drug Administration,” *Clinical Therapeutics*, vol. 29, no. 1, pp. 1–16, 2007.
- Bowes, J., Brown, A. J., Hamon, J., Jarolimek, W., Sridhar, A., Waldron, G., and Whitebread, S., “Reducing safety-related drug attrition: The use of in vitro pharmacological profiling,” *Nature Reviews*, vol. 11, no. 12, pp. 909–922, Dec. 2012.
- Breiman, L., “Random forests,” *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.
- Bruno, A. E., Charbonneau, P., Newman, J., Snell, E. H., So, D. R., Vanhoucke, V., Watkins, C. J., Williams, S., and Wilson, J., “Classification of crystallization outcomes using deep convolutional neural networks,” *PLOS one*, vol. 13, no. 6, e0198883, 2018.
- Cairns, J., “Providing guidance to the NHS: The Scottish Medicines Consortium and the National institute for Clinical Excellence compared,” *Health Policy*, vol. 76, no. 2, pp. 134–143, 2006.
- Chang, W., Cheng, J., Allaire, J., Xie, Y., and McPherson, J., *Shiny: Web application framework for r*, R package version 1.3.2, 2019. [Online]. Available: <https://CRAN.R-project.org/package=shiny>.
- Chang, Y., Park, H., Yang, H.-J., Lee, S., Lee, K.-Y., Kim, T. S., Jung, J., and Shin, J.-M., “Cancer drug response profile scan (cdrscan): A deep learning model that predicts drug effectiveness from cancer genomic signature,” *Scientific reports*, vol. 8, no. 1, p. 8857, 2018.
- Chen, L., Lu, J., Huang, T., Yin, J., Wei, L., and Cai, Y.-D., “Finding candidate drugs for hepatitis c based on chemical-chemical and chemical-protein interactions,” *PLoS One*, vol. 9, no. 9, pp. 1–6, 2014.
- Chen, X., Yan, C. C., Zhang, X., Zhang, X., Dai, F., Yin, J., and Zhang, Y., “Drug–target interaction prediction: Databases, web servers and computational models,” *Briefings in bioinformatics*, vol. 17, no. 4, pp. 696–712, 2015.

- Cohen, W. W., “Fast effective rule induction,” in *Machine Learning Proceedings 1995*, Elsevier, 1995, pp. 115–123.
- Cronin, M. T. D., “Chapter 1 an introduction to chemical grouping, categories and read-across to predict toxicity,” in *Chemical Toxicity Prediction: Category Formation and Read-Across*, The Royal Society of Chemistry, 2013, pp. 1–29, ISBN: 978-1-84973-384-7. DOI: 10.1039/9781849734400-00001.
- Davies, M., Nowotka, M., Papadatos, G., Dedman, N., Gaulton, A., Atkinson, F., Bellis, L., and Overington, J. P., “ChEMBL web services: Streamlining access to drug discovery data and utilities,” *Nucleic acids research*, vol. 43, no. W1, W612–W620, 2015.
- Davis, A. P., Grondin, C. J., Johnson, R. J., Sciaky, D., King, B. L., McMorran, R., Wieggers, J., Wieggers, T. C., and Mattingly, C. J., “The comparative toxicogenomics database: Update 2017,” *Nucleic acids research*, vol. 45, no. D1, pp. D972–D978, 2016.
- Davis, A. P., Murphy, C. G., Saraceni-Richards, C. A., Rosenstein, M. C., Wieggers, T. C., and Mattingly, C. J., “Comparative toxicogenomics database: A knowledgebase and discovery tool for chemical–gene–disease networks,” *Nucleic acids research*, vol. 37, no. suppl\_1, pp. D786–D792, 2008.
- Davis, A. P., Murphy, C. G., Saraceni-Richards, C. A., Rosenstein, M. C., Wieggers, T. C., and Mattingly, C. J. (2018). “Batch Query — CTD,” [Online]. Available: <http://ctdbase.org/tools/batchQuery.go> (visited on 12/12/2018).
- Davis, A. P., Murphy, C. G., Saraceni-Richards, C. A., Rosenstein, M. C., Wieggers, T. C., and Mattingly, C. J. (2018). “Data Downloads — CTD,” [Online]. Available: <http://ctdbase.org/downloads/> (visited on 12/12/2018).
- Department of Health, Social Services and Public Safety. (2018). “Legislation covering medicines,” [Online]. Available: <https://www.health-ni.gov.uk/articles/legislation-covering-medicines> (visited on 08/14/2018).
- Ding, H., Takigawa, I., Mamitsuka, H., and Zhu, S., “Similarity-based machine learning methods for predicting drug–target interactions: A brief review,” *Briefings in bioinformatics*, vol. 15, no. 5, pp. 734–747, 2013.
- Dix, D. J., Houck, K. A., Martin, M. T., Richard, A. M., Setzer, R. W., and Kavlock, R. J., “The ToxCast program for prioritizing toxicity testing of environmental chemicals,” *Toxicological Sciences*, vol. 95, no. 1, pp. 5–12, 2007.

- Dolman, J., “Chapter 26 introduction: Good manufacturing practice,” in *Good Clinical, Laboratory and Manufacturing Practices: Techniques for the QA Professional*, The Royal Society of Chemistry, 2007, pp. 371–385, ISBN: 978-0-85404-834-2. DOI: 10.1039/9781847557728-00369.
- Ekins, S., “The next era: Deep learning in pharmaceutical research,” *Pharmaceutical research*, vol. 33, no. 11, pp. 2594–2603, 2016.
- Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., and Thrun, S., “Dermatologist-level classification of skin cancer with deep neural networks,” *Nature*, vol. 542, no. 7639, p. 115, 2017.
- Eswar, N., Webb, B., Marti-Renom, M. A., Madhusudhan, M. S., Eramian, D., Shen, M.-y., Pieper, U., and Sali, A., “Comparative protein structure modeling using modeller,” *Current Protocols in Bioinformatics*, pp. 5–6, 2014.
- FellStar. (Jul. 30, 2018). “Long running tasks with shiny: Challenges and solutions,” [Online]. Available: <http://blog.fellstat.com/?p=407> (visited on 08/15/2019).
- Frank, E. and Witten, I. H., “Generating accurate rule sets without global optimization,” in *Proceedings of the Fifteenth International Conference on Machine Learning*, Morgan Kaufmann Publishers Inc., 1998, pp. 144–151.
- Funai, C., Tapparello, C., Ba, H., Karaoglu, B., and Heinzelman, W., “Extending volunteer computing through mobile ad hoc networking,” in *Global Communications Conference (GLOBECOM), 2014 IEEE*, IEEE, 2014, pp. 32–38.
- Gaulton, A., Hersey, A., Nowotka, M., Bento, A. P., Chambers, J., Mendez, D., Mutowo, P., Atkinson, F., Bellis, L. J., Cibrián-Uhalte, E., *et al.*, “The chembl database in 2017,” *Nucleic acids research*, vol. 45, no. D1, pp. D945–D954, 2016.
- Gaulton, A., Hersey, A., Nowotka, M., Bento, A. P., Chambers, J., Mendez, D., Mutowo, P., Atkinson, F., Bellis, L. J., Cibrián-Uhalte, E., *et al.* (2017). “ChEMBLdb Releases,” [Online]. Available: [ftp://ftp.ebi.ac.uk/pub/databases/chembl/ChEMBLdb/releases/chembl\\_23/](ftp://ftp.ebi.ac.uk/pub/databases/chembl/ChEMBLdb/releases/chembl_23/) (visited on 12/12/2018).
- Gaulton, A., Hersey, A., Nowotka, M., Bento, A. P., Chambers, J., Mendez, D., Mutowo, P., Atkinson, F., Bellis, L. J., Cibrián-Uhalte, E., *et al.* (2019). “Schema questions and sql examples - chembl interface documentation,” [Online]. Available: <https://chembl.gitbook.io/chembl-interface-documentat>

ion/frequently-asked-questions/schema-questions-and-sql-examples#retrieve-compound-activity-details-for-a-target (visited on 05/17/2019).

- Gilson, M. K., Liu, T., Baitaluk, M., Nicola, G., Hwang, L., and Chong, J., “Bindingdb in 2015: A public database for medicinal chemistry, computational chemistry and systems pharmacology,” *Nucleic acids research*, vol. 44, no. D1, pp. D1045–D1053, 2015.
- Gönen, M., “Predicting drug–target interactions from chemical and genomic kernels using bayesian matrix factorization,” *Bioinformatics*, vol. 28, no. 18, pp. 2304–2310, 2012.
- Guha, R., “Chemical informatics functionality in r,” *Journal of Statistical Software*, vol. 18, no. 6, 2007.
- Guha, R., *Fingerprint: Functions to operate on binary fingerprint data*, R package version 3.5.7, 2018. [Online]. Available: <https://CRAN.R-project.org/package=fingerprint>.
- Günther, S., Kuhn, M., Dunkel, M., Campillos, M., Senger, C., Petsalaki, E., Ahmed, J., Urdiales, E. G., Gewiess, A., Jensen, L. J., *et al.*, “Supertarget and matador: Resources for exploring drug-target relationships,” *Nucleic acids research*, vol. 36, no. suppl\_1, pp. D919–D922, 2007.
- Günther, S., Kuhn, M., Dunkel, M., Campillos, M., Senger, C., Petsalaki, E., Ahmed, J., Urdiales, E. G., Gewiess, A., Jensen, L. J., *et al.* (2018). “MATADOR,” [Online]. Available: <http://matador.embl.de/> (visited on 12/12/2018).
- Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H., “The weka data mining software: An update,” *ACM SIGKDD Explorations Newsletter*, vol. 11, no. 1, pp. 10–18, 2009.
- Hay, M., Thomas, D. W., Craighead, J. L., Economides, C., and Rosenthal, J., “Clinical development success rates for investigational drugs,” *Nature Biotechnology*, vol. 32, no. 1, pp. 40–51, Jan. 2014.
- Helder, T., “Chapter 12 introduction: Good laboratory practice,” in *Good Clinical, Laboratory and Manufacturing Practices: Techniques for the QA Professional*, The Royal Society of Chemistry, 2007, pp. 171–181, ISBN: 978-0-85404-834-2. DOI: 10.1039/9781847557728-00169.

- HM Government. (1968). “Medicines Act 1968: Chapter 67,” [Online]. Available: [http://www.legislation.gov.uk/ukpga/1968/67/pdfs/ukpga\\_19680067\\_en.pdf](http://www.legislation.gov.uk/ukpga/1968/67/pdfs/ukpga_19680067_en.pdf) (visited on 02/04/2016).
- HM Government. (2016). “Apply for a license to market a medicine in the UK,” [Online]. Available: <https://www.gov.uk/guidance/apply-for-a-licence-to-market-a-medicine-in-the-uk> (visited on 02/04/2016).
- HM Government. (2020). “Medicines, medical devices and blood regulation and safety: Good practice, inspections and enforcement,” [Online]. Available: <https://www.gov.uk/topic/medicines-medical-devices-blood/good-practice> (visited on 05/11/2020).
- Horan, K., Cao, Y., Backman, T., and Girke, T., “ChemmineR: a compound mining framework for R,” *Bioinformatics*, vol. 24, no. 15, pp. 1733–1734, 15 2008.
- Horan, K. and Girke, T., *Chemmineob: R interface to a subset of openbabel functionalities*, R package version 1.14.0, 2017. [Online]. Available: <https://github.com/girke-lab/ChemmineOB>.
- Hornik, K., Buchta, C., and Zeileis, A., “Open-source machine learning: R meets Weka,” *Computational Statistics*, vol. 24, no. 2, pp. 225–232, 2009. DOI: 10.1007/s00180-008-0119-7.
- HPC Wales. (2016). “About,” [Online]. Available: <http://www.hpcwales.co.uk/> (visited on 01/06/2016).
- Hughes, J. P., Rees, S., Kalindjian, S. B., and Philpott, K. L., “Principles of early drug discovery,” *British journal of pharmacology*, vol. 162, no. 6, pp. 1239–1249, 2011.
- Jimenez-Carretero, D., Abrishami, V., Fernández-de-Manuel, L., Palacios, I., Quílez-Álvarez, A., Díez-Sánchez, A., Pozo, M. A. del, and Montoya, M. C., “Tox<sub>r</sub> cnn: Deep learning-based nuclei profiling tool for drug toxicity screening,” *PLoS computational biology*, vol. 14, no. 11, e1006238, 2018.
- John, G. H. and Langley, P., “Estimating continuous distributions in bayesian classifiers,” in *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*, Morgan Kaufmann Publishers Inc., 1995, pp. 338–345.
- Kanehisa, M., Goto, S., Sato, Y., Furumichi, M., and Tanabe, M., “Kegg for integration and interpretation of large-scale molecular data sets,” *Nucleic acids research*, vol. 40, no. D1, pp. D109–D114, 2011.

- Keum, J. and Nam, H., “Self-blm: Prediction of drug-target interactions via self-training svm,” *PloS one*, vol. 12, no. 2, e0171839, 2017.
- Kim, S., Thiessen, P. A., Bolton, E. E., and Bryant, S. H., “PUG-SOAP and PUG-REST: Web services for programmatic access to chemical information in PubChem,” *Nucleic Acids Research*, vol. 43, no. W1, W605–W611, Jul. 2015.
- Kim, S., Thiessen, P. A., Bolton, E. E., Chen, J., Fu, G., Gindulyte, A., Han, L., He, J., He, S., Shoemaker, B. A., Wang, J., Yu, B., Zhang, J., and Bryant, S. H., “PubChem Substance and Compound databases,” *Nucleic Acids Research*, vol. 44, no. D1, pp. D1202–D1213, Jan. 2016.
- Kinch, M., Patridge, E., Plummer, M., and Hoyer, D., “An analysis of FDA-approved drugs for infectious disease: Antibacterial agents,” *Drug Discovery Today*, vol. 19, no. 9, pp. 1283–1287, Sep. 2014.
- Kohavi, R., “The power of decision tables,” in *European conference on machine learning*, Springer, 1995, pp. 174–189.
- Kurgan, L. and Wang, C., “Survey of similarity-based prediction of drug-protein interactions,” *Current medicinal chemistry*, Nov. 2018, ISSN: 0929-8673. DOI: 10.2174/0929867325666181101115314. [Online]. Available: <https://doi.org/10.2174/0929867325666181101115314>.
- Laarhoven, T. van, Nabuurs, S. B., and Marchiori, E., “Gaussian interaction profile kernels for predicting drug–target interaction,” *Bioinformatics*, vol. 27, no. 21, pp. 3036–3043, 2011.
- Lang, D. T. and the CRAN Team, *RCurl: General Network (HTTP/FTP/...) Client Interface for R*, 2016. [Online]. Available: <https://CRAN.R-project.org/package=RCurl>.
- Lang, P. T., Brozell, S. R., Mukherjee, S., Pettersen, E. F., Meng, E. C., Thomas, V., Rizzo, R. C., Case, D. A., James, T. L., and Kuntz, I. D., “DOCK 6: Combining techniques to model RNA-small molecule complexes,” *RNA*, vol. 15, no. 6, pp. 1219–1230, 2009.
- Law, M. T., “How do regulators regulate? Enforcement of the Pure Food and Drugs Act, 1907-38,” *Journal of Law, Economics, and Organization*, vol. 22, no. 2, pp. 459–489, 2006.
- Le Cessie, S. and Van Houwelingen, J. C., “Ridge estimators in logistic regression,” *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, vol. 41, no. 1, pp. 191–201, 1992.



- Lesk, A. M., *Introduction to Bioinformatics*, 4th ed. Oxford University Press, 2014, ch. 6, pp. 264–265.
- Lipsky, M. S. and Sharp, L. K., “From idea to market: The drug approval process,” *The Journal of the American Board of Family Practice*, vol. 14, no. 5, pp. 362–367, 2001.
- Liu, Y., Wu, M., Miao, C., Zhao, P., and Li, X.-L., “Neighborhood regularized logistic matrix factorization for drug-target interaction prediction,” *PLoS computational biology*, vol. 12, no. 2, e1004760, 2016.
- Madeira, F., Park, Y. m., Lee, J., Buso, N., Gur, T., Madhusoodanan, N., Basutkar, P., Tivey, A. R. N., Potter, S. C., Finn, R. D., and Lopez, R., “The EMBL-EBI search and sequence analysis tools APIs in 2019,” *Nucleic Acids Research*, vol. 47, no. W1, W636–W641, Apr. 2019, ISSN: 0305-1048. DOI: 10.1093/nar/gkz268.
- Mak, K.-K. and Pichika, M. R., “Artificial intelligence in drug development: Present status and future prospects,” *Drug discovery today*, 2018.
- Marsaglia, G., Tsang, W. W., Wang, J., *et al.*, “Evaluating kolmogorov’s distribution,” *Journal of Statistical Software*, vol. 8, no. 18, pp. 1–4, 2003.
- McKusick, V. A., *Mendelian inheritance in man: a catalog of human genes and genetic disorders*. JHU Press, 1998, vol. 1.
- Medicines and Healthcare products Regulatory Agency. (2016). “Yellow Card scheme looks to the future at 50th anniversary forum,” [Online]. Available: <https://www.gov.uk/government/news/yellow-card-scheme-looks-to-the-future-at-50th-anniversary-forum> (visited on 02/05/2016).
- Mei, J.-P., Kwoh, C.-K., Yang, P., Li, X.-L., and Zheng, J., “Drug–target interaction prediction by learning from local information and neighbors,” *Bioinformatics*, vol. 29, no. 2, pp. 238–245, 2012.
- Mendez, D., Gaulton, A., Bento, A. P., Chambers, J., De Veij, M., Félix, E., Magariños, M. P., Mosquera, J. F., Mutowo, P., Nowotka, M., *et al.*, “ChEMBL: Towards direct deposition of bioassay data,” *Nucleic acids research*, vol. 47, no. D1, pp. D930–D940, 2018.
- Microsoft Corporation and Weston, S., *Foreach: Provides foreach looping construct for r*, R package version 1.4.4, 2017. [Online]. Available: <https://CRAN.R-project.org/package=foreach>.

- Microsoft Corporation and Weston, S., *Doparallel: Foreach parallel adaptor for the 'parallel' package*, R package version 1.0.14, 2018. [Online]. Available: <https://CRAN.R-project.org/package=doParallel>.
- Morgan, S., Grootendorst, P., Lexchin, J., Cunningham, C., and Greyson, D., “The cost of drug development: A systematic review,” *Health Policy*, vol. 100, no. 1, pp. 4–17, 2011.
- Moriwaki, H., Tian, Y.-S., Kawashita, N., and Takagi, T. (2018). “Descriptor list — mordred 1.1.2a1 documentation,” [Online]. Available: <http://mordred-descriptor.github.io/documentation/master/descriptors.html>.
- Moriwaki, H., Tian, Y.-S., Kawashita, N., and Takagi, T., “Mordred: A molecular descriptor calculator,” *Journal of cheminformatics*, vol. 10, no. 1, p. 4, 2018.
- National Center for Biotechnology Information. (2009). “Pubchem substructure fingerprint,” [Online]. Available: [ftp://ftp.ncbi.nlm.nih.gov/pubchem/specifications/pubchem\\_fingerprints.txt](ftp://ftp.ncbi.nlm.nih.gov/pubchem/specifications/pubchem_fingerprints.txt).
- National Center for Biotechnology Information. (2016). “PubChem Help,” [Online]. Available: <https://pubchem.ncbi.nlm.nih.gov/help.html> (visited on 05/06/2016).
- National Center for Biotechnology Information. (2018). “BioAssay Download FTP,” [Online]. Available: <ftp://ftp.ncbi.nlm.nih.gov/pubchem/Bioassay/> (visited on 12/12/2018).
- National Center for Biotechnology Information. (2018). “Downloading PubChem Data,” [Online]. Available: <https://pubchemdocs.ncbi.nlm.nih.gov/downloads> (visited on 12/04/2018).
- National Center for Biotechnology Information. (2018). “Home - PubMed - NCBI,” [Online]. Available: <https://www.ncbi.nlm.nih.gov/pubmed> (visited on 08/16/2018).
- National Center for Biotechnology Information. (Apr. 27, 2018). “PubChem BioAssay Tools to be replaced — PubChem Blog,” [Online]. Available: <https://pubchemblog.ncbi.nlm.nih.gov/2018/04/27/pubchem-bioassay-tools-to-be-replaced/> (visited on 07/29/2019).
- National Center for Biotechnology Information. (2018). “PubChem Identifier Exchange Service,” [Online]. Available: <https://pubchemdocs.ncbi.nlm.nih.gov/identifier-exchange-service> (visited on 12/04/2018).

- National Center for Biotechnology Information. (2019). “PubChem Similarity Search,” [Online]. Available: <https://pubchem.ncbi.nlm.nih.gov/#query=CID2244%20structure&tab=similarity> (visited on 09/20/2019).
- National Institute for Health and Care Excellence. (2016). “Who we are,” [Online]. Available: <https://www.nice.org.uk/about/who-we-are> (visited on 02/04/2016).
- National Institutes of Health. (2016). “Glossary of common terms,” [Online]. Available: <http://www.nih.gov/health-information/nih-clinical-research-trials-you/glossary-common-terms> (visited on 01/14/2016).
- Neubig, R. R., Spedding, M., Kenakin, T., and Christopoulos, A., “International union of pharmacology committee on receptor nomenclature and drug classification. xxxviii. update on terms and symbols in quantitative pharmacology,” *Pharmacological reviews*, vol. 55, no. 4, pp. 597–606, 2003.
- Neudert, G. and Klebe, G., “DSX: A knowledge-based scoring function for the assessment of protein–ligand complexes,” *Journal of Chemical Information and Modeling*, vol. 51, no. 10, pp. 2731–2745, 2011.
- O’Boyle, N. M., Banck, M., James, C. A., Morley, C., Vandermeersch, T., and Hutchison, G. R., “Open babel: An open chemical toolbox,” *Journal of cheminformatics*, vol. 3, no. 1, p. 33, 2011.
- Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H., and Kanehisa, M., “Kegg: Kyoto encyclopedia of genes and genomes,” *Nucleic acids research*, vol. 27, no. 1, pp. 29–34, 1999.
- OMx Personal Health Analytics. (Sep. 19, 2019). “DrugBankPlus — Scientific Drug Data,” [Online]. Available: <https://www.drugbankplus.com/> (visited on 09/19/2019).
- Ooms, J., “The jsonlite package: A practical and consistent mapping between json data and r objects,” *arXiv:1403.2805 [stat.CO]*, 2014. [Online]. Available: <https://arxiv.org/abs/1403.2805>.
- Ooms, J., James, D., DebRoy, S., Wickham, H., and Horner, J., *Rmysql: Database interface and ‘mysql’ driver for r*, R package version 0.10.13, 2017. [Online]. Available: <https://CRAN.R-project.org/package=RMySQL>.
- Oxford University Press. (2015). “Drug - Definition of Drug in English from the Oxford dictionary,” [Online]. Available: <http://www.oxforddictionaries.com/definition/english/drug> (visited on 10/29/2015).

- Öztürk, H., Ozkirimli, E., and Özgür, A., “A comparative study of smiles-based compound similarity functions for drug-target interaction prediction,” *BMC bioinformatics*, vol. 17, no. 1, p. 128, 2016.
- Pagès, H., Aboyoun, P., Gentleman, R., and DebRoy, S., *Biostrings: String objects representing biological sequences, and matching algorithms*, R package version 2.44.2, 2017.
- Pantelev, J., Gao, H., and Jia, L., “Recent applications of machine learning in medicinal chemistry,” *Bioorganic & medicinal chemistry letters*, 2018.
- Parmentier, Y., Bossant, M.-J., Bertrand, M., and Walther, B., “5.10 - in vitro studies of drug metabolism,” in *Comprehensive Medicinal Chemistry II*, Taylor, J. B. and Triggle, D. J., Eds., Oxford: Elsevier, 2007, pp. 231–257, ISBN: 978-0-08-045044-5.
- Paul, S. M., Mytelka, D. S., Dunwiddie, C. T., Persinger, C. C., Munos, B. H., Lindborg, S. R., and Schacht, A. L., “How to improve r&d productivity: The pharmaceutical industry’s grand challenge,” *Nature reviews Drug discovery*, vol. 9, no. 3, p. 203, 2010.
- Pereira, M., Costa, V. S., Camacho, R., Fonseca, N. A., Simões, C., and Brito, R. M., “Comparative study of classification algorithms using molecular descriptors in toxicological databases,” in *4th Brazilian Symposium on Bioinformatics: July 29-31, 2009; Porto Alegre, Brazil*, Guimaraes, K. S., Panchenko, A., and Przytycka, T. M., Eds., Springer Berlin Heidelberg, 2009, pp. 121–132.
- PhRMA, *2015 Biopharmaceutical Research Industry Profile*. Washington, DC, USA: Pharmaceutical Research and Manufacturers of America (PhRMA), 2015.
- Phuong, J., Truong, L., Sipes, N., Connors, K., Houck, K., Judson, R., and Martin, M. (2015). “ToxCast Assay Annotation Version 1.0 Data User Guide,” [Online]. Available: [https://www.epa.gov/sites/production/files/2015-08/documents/toxcast\\_assay\\_annotation\\_data\\_users\\_guide\\_20141021.pdf](https://www.epa.gov/sites/production/files/2015-08/documents/toxcast_assay_annotation_data_users_guide_20141021.pdf) (visited on 12/04/2018).
- Pu, L., Naderi, M., Liu, T., Wu, H.-C., Mukhopadhyay, S., and Brylinski, M., “E toxpred: A machine learning-based approach to estimate the toxicity of drug candidates,” *BMC Pharmacology and Toxicology*, vol. 20, no. 1, p. 2, 2019.
- Quinlan, J. R., “Simplifying decision trees,” *International journal of man-machine studies*, vol. 27, no. 3, pp. 221–234, 1987.

- Quinlan, J. R., *C4.5: Programs for Machine Learning*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1993, ISBN: 1-55860-238-0.
- R Core Team, *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria, 2015. [Online]. Available: <https://www.R-project.org>.
- R Core Team, *Foreign: Read data stored by 'minitab', 's', 'sas', 'spss', 'stata', 'systat', 'weka', 'dbase', ...* R package version 0.8-69, 2017. [Online]. Available: <https://CRAN.R-project.org/package=foreign>.
- Ramsundar, B., Kearnes, S., Riley, P., Webster, D., Konerding, D., and Pande, V., “Massively multitask networks for drug discovery,” *arXiv preprint arXiv:1502.02072*, 2015.
- Remez, N., Garcia-Serna, R., Vidal, D., and Mestres, J., “The in vitro pharmacological profile of drugs as a proxy indicator of potential in vivo organ toxicities,” *Chemical research in toxicology*, vol. 29, no. 4, pp. 637–648, 2016.
- Riley, P., Webster, D., and Ramsundar, B. (2018). “Google AI Blog: Large-Scale Machine Learning for Drug Discovery,” [Online]. Available: <https://ai.googleblog.com/2015/03/large-scale-machine-learning-for-drug.html> (visited on 08/08/2018).
- Rodríguez-Pérez, R., Miyao, T., Jasial, S., Vogt, M., and Bajorath, J., “Prediction of compound profiling matrices using machine learning,” *ACS Omega*, vol. 3, no. 4, pp. 4713–4723, 2018.
- RStudio Team, *Rstudio: Integrated development environment for r*, RStudio, Inc., Boston, MA, 2015. [Online]. Available: <http://www.rstudio.com/>.
- Scheeder, C., Heigwer, F., and Boutros, M., “Machine learning and image-based profiling in drug discovery,” *Current opinion in systems biology*, 2018.
- Scottish Medicines Consortium. (2016). “What we do,” [Online]. Available: [https://www.scottishmedicines.org.uk/About\\_SMC/What\\_we\\_do](https://www.scottishmedicines.org.uk/About_SMC/What_we_do) (visited on 02/04/2016).
- Sertkaya, A., Wong, H.-H., Jessup, A., and Beleche, T., “Key cost drivers of pharmaceutical clinical trials in the united states,” *Clinical Trials*, vol. 13, no. 2, pp. 117–126, 2016.
- Siers, M. J. and Islam, M. Z., “Software defect prediction using a cost sensitive decision forest and voting, and a potential solution to the class imbalance problem,” *Information Systems*, vol. 51, pp. 62–71, 2015.

- Sipes, N. S., Martin, M. T., Kothiya, P., Reif, D. M., Judson, R. S., Richard, A. M., Houck, K. A., Dix, D. J., Kavlock, R. J., and Knudsen, T. B., “Profiling 976 toxcast chemicals across 331 enzymatic and receptor signaling assays,” *Chemical research in toxicology*, vol. 26, no. 6, pp. 878–895, 2013.
- Smith, T. F., Waterman, M. S., *et al.*, “Identification of common molecular subsequences,” *Journal of molecular biology*, vol. 147, no. 1, pp. 195–197, 1981.
- Smithells, R. W. and Newman, C. G. H., “Recognition of thalidomide defects,” *Journal of medical genetics*, vol. 29, no. 10, pp. 716–723, 1992.
- Stephenson, N., Shane, E., Chase, J., Rowland, J., Ries, D., Justice, N., Zhang, J., Chan, L., and Cao, R., “Survey of machine learning techniques in drug discovery,” *Current drug metabolism*, vol. 20, no. 3, pp. 185–193, 2019.
- Supercomputing Wales. (2018). “About,” [Online]. Available: <https://www.supercomputing.wales/about/> (visited on 08/08/2018).
- Talbot, D. and Downes, N., “Chapter 1 introduction: Good clinical practice,” in *Good Clinical, Laboratory and Manufacturing Practices: Techniques for the QA Professional*, The Royal Society of Chemistry, 2007, pp. 3–11, ISBN: 978-0-85404-834-2. DOI: 10.1039/9781847557728-00001.
- The Medicines and Healthcare Products Regulatory Agency, *Medicines & medical devices regulation: What you need to know*, <http://www.mhra.gov.uk/home/groups/comms-ic/documents/websiteresources/con2031677.pdf>, 2008.
- The Stationary Office, *Report of an independent review of access to the Yellow Card scheme*, <http://www.mhra.gov.uk/home/groups/comms-ic/documents/websiteresources/con2015008.pdf>, Apr. 2004.
- The UniProt Consortium, “UniProt: A hub for protein information,” *Nucleic Acids Research*, vol. 43, no. D1, pp. D204–D212, Jan. 2015.
- The UniProt Consortium. (2018). “AR - Androgen receptor - Homo sapiens (Human) - AR gene & protein,” [Online]. Available: <https://www.uniprot.org/uniprot/P10275> (visited on 08/01/2018).
- The Uniprot Consortium. (2018). “Androgden receptor,” [Online]. Available: <https://www.uniprot.org/uniprot/P10275.txt> (visited on 12/12/2018).
- The Uniprot Consortium. (2018). “Retrieve/ID mapping,” [Online]. Available: <https://www.uniprot.org/uploadlists/> (visited on 08/10/2018).

- The Uniprot Consortium. (2018). “UniProtKB,” [Online]. Available: <https://www.uniprot.org/uniprot/> (visited on 12/12/2018).
- Tice, R. R., Austin, C. P., Kavlock, R. J., and Bucher, J. R., “Improving the human hazard characterization of chemicals: A tox21 update,” *Environmental health perspectives*, vol. 121, no. 7, pp. 756–765, 2013.
- Trott, O. and Olson, A. J., “AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading,” *Journal of Computational Chemistry*, vol. 31, no. 2, pp. 455–461, 2010.
- United States Environment Protection Agency. (2015). “US EPA TOXCAST DATA RELEASE OCTOBER 2015 - Summary Files,” [Online]. Available: [ftp://newftp.epa.gov/Computational\\_Toxicology\\_Data/High\\_Throughput\\_Screening\\_Data/Previous\\_Data/ToxCast\\_Data\\_Release\\_Oct\\_2015/Summary\\_Files/README\\_INVITRODB\\_V2\\_SUMMARY.pdf](ftp://newftp.epa.gov/Computational_Toxicology_Data/High_Throughput_Screening_Data/Previous_Data/ToxCast_Data_Release_Oct_2015/Summary_Files/README_INVITRODB_V2_SUMMARY.pdf) (visited on 12/04/2018).
- United States Environment Protection Agency. (2018). “Toxicity ForeCaster (ToxCast) Data — Safer Chemicals Research — US EPA,” [Online]. Available: <https://www.epa.gov/chemical-research/toxicity-forecaster-toxcasttm-data> (visited on 08/09/2018).
- United States Environment Protection Agency. (2019). “ToxCast and Tox21 Summary Files,” [Online]. Available: [https://epa.figshare.com/articles/ToxCast\\_and\\_Tox21\\_Summary\\_Files/6062479/4](https://epa.figshare.com/articles/ToxCast_and_Tox21_Summary_Files/6062479/4).
- United States Food and Drug Administration. (2020). “eCFR - Code of Federal Regulations Title 21 Part 58: Good Laboratory Practice for Nonclinical Laboratory Studies,” [Online]. Available: <https://www.ecfr.gov/cgi-bin/text-idx?SID=9b179b470add6a7c8afd29fba0948dcd&mc=true&node=pt21.1.58&rgn=div5> (visited on 05/11/2020).
- United States Food and Drug Administration. (Feb. 10, 2020). “Office of Pharmaceutical Quality — FDA,” [Online]. Available: <https://www.fda.gov/about-fda/center-drug-evaluation-and-research-cder/office-pharmaceutical-quality> (visited on 05/11/2020).
- Ushey, K., Allaire, J., and Tang, Y., *Reticulate: Interface to 'python'*, R package version 1.13, 2019. [Online]. Available: <https://CRAN.R-project.org/package=reticulate>.

- Van Laarhoven, T. and Marchiori, E., “Predicting drug-target interactions for new drug compounds using a weighted nearest neighbor profile,” *PloS one*, vol. 8, no. 6, e66952, 2013.
- Vargesson, N., “Thalidomide-induced teratogenesis: History and mechanisms,” *Birth Defects Research Part C: Embryo Today: Reviews*, vol. 105, no. 2, pp. 140–156, 2015.
- Vogt, M., Jasial, S., and Bajorath, J., “Extracting compound profiling matrices from screening data,” *ACS Omega*, vol. 3, no. 4, pp. 4706–4712, 2018.
- Volkamer, A., Griewel, A., Grombacher, T., and Rarey, M., “Analyzing the topology of active sites: On the prediction of pockets and subpockets,” *Journal of Chemical Information and Modeling*, vol. 50, no. 11, pp. 2041–2052, 2010.
- Waldman, S. A., “Does potency predict clinical efficacy? illustration through an antihistamine model,” *Annals of Allergy, Asthma & Immunology*, vol. 89, no. 1, pp. 7–12, 2002.
- Whitmore, E., *Development of FDA-regulated medical products: a translational approach*, 2nd ed. Milwaukee, WI, USA: ASQ Quality Press, Jan. 2012, ch. 1, pp. 1–17.
- Whitmore, E., *Development of FDA-regulated medical products: a translational approach*, 2nd ed. Milwaukee, WI, USA: ASQ Quality Press, Jan. 2012, ch. 6, pp. 63–76.
- Wickham, H., “Reshaping data with the reshape package,” *Journal of Statistical Software*, vol. 21, no. 12, pp. 1–20, 2007. [Online]. Available: <http://www.jstatsoft.org/v21/i12/>.
- Wickham, H., François, R., Henry, L., and Müller, K., *Dplyr: A grammar of data manipulation*, R package version 0.8.1, 2019. [Online]. Available: <https://CRAN.R-project.org/package=dplyr>.
- Wickham, H., Hester, J., and François, R., *Readr: Read rectangular text data*, R package version 1.1.1, 2017. [Online]. Available: <https://CRAN.R-project.org/package=readr>.
- Wishart, D., Arndt, D., Pon, A., Sajed, T., Guo, A. C., Djoumbou, Y., Knox, C., Wilson, M., Liang, Y., Grant, J., *et al.*, “T3db: The toxic exposome database,” *Nucleic acids research*, vol. 43, no. D1, pp. D928–D934, 2014.



- Wishart, D., Arndt, D., Pon, A., Sajed, T., Guo, A. C., Djoumbou, Y., Knox, C., Wilson, M., Liang, Y., Grant, J., *et al.* (2018). “T3DB: Downloads,” [Online]. Available: <http://www.t3db.ca/downloads> (visited on 12/12/2018).
- Wishart, D. S., Feunang, Y. D., Guo, A. C., Lo, E. J., Marcu, A., Grant, J. R., Sajed, T., Johnson, D., Li, C., Sayeeda, Z., *et al.*, “Drugbank 5.0: A major update to the drugbank database for 2018,” *Nucleic acids research*, vol. 46, no. D1, pp. D1074–D1082, 2017.
- Wishart, D. S., Feunang, Y. D., Marcu, A., Guo, A. C., Liang, K., Vázquez-Fresno, R., Sajed, T., Johnson, D., Li, C., Karu, N., *et al.*, “Hmdb 4.0: The human metabolome database for 2018,” *Nucleic acids research*, vol. 46, no. D1, pp. D608–D617, 2017.
- Wishart, D. S., Knox, C., Guo, A. C., Sharavista, S., Hassanali, M., Stothard, P., Chang, Z., and Woolsey, J., “DrugBank: A comprehensive resource for in silico drug discovery and exploration,” *Nucleic Acids Research*, vol. 34, pp. D668–D672, Jan. 2006.
- Wishart, D. S., Knox, C., Guo, A. C., Sharavista, S., Hassanali, M., Stothard, P., Chang, Z., and Woolsey, J. (2016). “DrugBank: Abacavir,” [Online]. Available: <http://www.drugbank.ca/drugs/DB01048> (visited on 05/06/2016).
- Wishart, D. S., Knox, C., Guo, A. C., Sharavista, S., Hassanali, M., Stothard, P., Chang, Z., and Woolsey, J. (2016). “Drugbank: Documentation and sources,” [Online]. Available: <http://www.drugbank.ca/documentation> (visited on 05/06/2016).
- Wishart, D. S., Knox, C., Guo, A. C., Sharavista, S., Hassanali, M., Stothard, P., Chang, Z., and Woolsey, J. (2019). “Chemical Structure Search - DrugBank,” [Online]. Available: [https://www.drugbank.ca/structures/search/small\\_molecule\\_drugs/structure](https://www.drugbank.ca/structures/search/small_molecule_drugs/structure) (visited on 09/24/2019).
- Wishart, D. S., Knox, C., Guo, A. C., Sharavista, S., Hassanali, M., Stothard, P., Chang, Z., and Woolsey, J. (2019). “Oxycodone - drugbank,” [Online]. Available: <https://www.drugbank.ca/drugs/DB00497> (visited on 05/17/2019).
- Wu, H., Yang, S., Huang, Z., He, J., and Wang, X., “Type 2 diabetes mellitus prediction model based on data mining,” *Informatics in Medicine Unlocked*, vol. 10, pp. 100–107, 2018.

- Xia, Z., Wu, L.-Y., Zhou, X., and Wong, S. T., “Semi-supervised drug-protein interaction prediction from heterogeneous biological spaces,” in *BMC systems biology*, BioMed Central, vol. 4, 2010, S6.
- Yamanishi, Y., Araki, M., Gutteridge, A., Honda, W., and Kanehisa, M., “Prediction of drug–target interaction networks from the integration of chemical and genomic spaces,” *Bioinformatics*, vol. 24, no. 13, pp. i232–i240, 2008.
- Yang, J., Yan, R., Roy, A., Xu, D., Poisson, J., and Zhang, Y., “The I-TASSER Suite: Protein structure and function prediction,” *Nature Methods*, vol. 12, no. 1, pp. 7–8, 2015.
- Yuan, Q., Gao, J., Wu, D., Zhang, S., Mamitsuka, H., and Zhu, S., “Druge-rank: Improving drug–target interaction prediction of new candidate drugs or targets by ensemble learning to rank,” *Bioinformatics*, vol. 32, no. 12, pp. i18–i27, 2016.
- Zheng, X., Ding, H., Mamitsuka, H., and Zhu, S., “Collaborative matrix factorization with multiple similarities for predicting drug-target interactions,” in *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM, 2013, pp. 1025–1033.