Analysis and Prediction of Selectivity in Biological and Chemical Space



Anne F. Nierobisch

Kellogg College

A Thesis Submitted For The Degree of Master of Research

Hilary Term 2020

Dedication

To Phil, the calm yang to my turbulent yin: the only person I know who can stand in front of a burning car in the middle of nowhere, with no phone connection and zombies attacking, and utter, calmly and with complete conviction, 'It could be worse'.

Acknowledgement

I would like to thank all of my four supervisors: Dr. Lisa Sach-Peltason, Dr. Mathias Leddin, Professor Charlotte M. Deane, Professor Garrett M. Morris - your support, guidance and understanding have been immeasurable in helping me to become a better scientist. I'd also like to thank Dr. Jitao David Zhang and Dr. Samuel Croset for their time and feedback.

Secondly, I would like to shout a big thank you to OPIG - a research group unequaled in Oxford. Thank you for the constant supply of cookies and the witty quips. Awesome people tend to cluster in OPIG and I can't thank you all enough for my time with you. To Fergus I and Fergus II - thank you for your support in learning und understanding machine learning. To Dan and Fergus II, I cannot thank you enough for proofreading my thesis. To Susan, Catherine and Connor - thanks for being rays of sunshine in my daily OPIG life. To James - you were the best house and OPIG mate I could have ever wished for, and I will always think back fondly to our daily coffee dates and our pizza nights at home.

Thirdly, I wish to express my deepest thanks to Julia, Alistair, Rodrigo, Iain and Sara. Dearest Julia, I'm grateful for your support and continuous friendship. Dear Alistair and Rodrigo, I am thankful that you listened to me. Dear Iain and Sara, I will never forget the lessons you taught me.

Lastly, to Phil - you know why.

Abstract

Small molecules can be designed to interact with a specific protein or proteins, and, by extension, to modulate the biological pathways in which the respective protein is involved. Small molecules can thus influence a disease phenotype on a cellular, tissue, or organism level. However, some compounds, so-called 'promiscuous' compounds, modulate the phenotype by binding to a wide range of on- and off-targets. In the case of drugs, this binding to off-targets can be responsible for side effects, as the perturbation of other biological pathways can lead to severe or even lethal side effects. 'Selective' compounds target only one protein or a small set of closely related targets, the 'on-targets', at a given bioactivity range. Contrary to promiscuous small molecules, they should have a low number of off-targets.

The selectivity or promiscuity of a compound is dependent on both protein and ligand structural features. I am the first to explore the inclusion of individual residue movement as a descriptor in random forest models. Using simple 2D flexibility descriptors, I show that I can identify flexible regions, depending on the data set. However these flexibility values are too uniform and coarse to be used as descriptors in the random forest algorithm to identify up discriminating features.

I also show that the current 2D random forest models and the available kinase data are sufficient to correctly predict the activity of selective compound-target pairs depending on previous data.

Lastly, I investigated the use of 3D protein-ligand interaction features in filtering out selective (and promiscuous) compounds and the use of interaction grids to classify candidate compounds in different binding affinity ranges. Both promiscuous ligands and selective compounds active against closely related targets tend to have a conserved binding mode across their kinase targets. Additionally, binding features cannot be used to separate strong or weak binders for a given target ; however, a separation based on a 3D grid might be promising.

However, from a biological perspective, the often unknown involvement of a target in other pathways ('pathway crosstalk') might add confounding biological readouts and thus unexpected side effects, even if the compound is selective against the target of

interest. Hence, I explored drug targets from a pathway-centric perspective. I show that the overall number of pathways associated with all drug targets of a drug does not correlate with its success or its withdrawal. Lastly, I show that - depending on target class - drug targets are generally associated with a wider range of pathways than 'undrugged' proteins, highlighting either a research bias towards drug targets or possibly that drug targets are chosen due to a higher involvement in pathway crosstalk. I also found that pathways could be targeted for the same disease area using multiple associated targets, thus highlighting a potential for finding novel drug targets outside the traditional 'target class'-centric paradigm.

In conclusion: in order to study selectivity, I have focused on a widely studied target class, namely kinases. My results indicate promise in using 3D-based selectivity models to derive kinase inhibitors, especially combined with the use of a standardized nomenclature across kinase binding pockets. Secondly, I also show the necessity of using a more 'pathway-centric' exploration of drug targets instead of the current, traditional 'target class'-centric approaches.

Declaration

I declare that no parts of this thesis or its research herein have been reproduced or accepted for another award or degree or diploma at any other university or learning institution. This thesis contains no other person's work except where stated in the text.

Anne F. Nierobisch

11th June 2020

Contents

Dedication i

Acknowledgement ii

Abstract iii

Declaration v

1 Introduction 1

1.1 Motivation 1

1.2 Pathway Crosstalk: The Biological Analogue to Compound Selectivity and Promiscuity 2

1.3 Drug Pharmacology: Deconvoluting The Selectivity or Promiscuity of a Given Compound 4

1.3.1 The Molecular Basis of Selectivity and Promiscuity 7

1.3.2 Identifying Protein and Ligand Features Responsible For Selectivity or Promiscuity 9

1.3.3 Tool Compounds: Selective Compounds To Explore Biological Questions 13

1.4 Project Aims and Thesis Structure 15

2 The Connection between Drugs, Drug Targets and Pathways: an Extension of The Drug-Target Paradigm 19

2.1 Introduction 19

2.2 Methods 19

2.3 Results 23

2.4 Conclusion 35

3 Investigation of whether the Addition of Residue-Specific Movement leads to Improvement in Random Forest Models 39

3.1 Introduction 39

3.2 Simple Metrics For Describing The Flexibility of a Residue 39

3.2.1 B-Factor 39

3.2.2 The Root-Mean-Square Deviation (RMSD) 40

3.3 Data Sets . . . 41
3.3.1 The Fulle Kinase Data Set: a Large Data Set Of Kinase Inhibitors Collected For Modelling Kinase Activity . . . 41
3.3.2 The KLIFS Database: A Database Detailing Kinase Target-Ligand Interactions . . . 43
3.4 Random Forests . . . 44
3.4.1 Random Forest Models . . . 45
3.4.2 Metrics used in Classification . . . 50
3.5 Methods . . . 53
3.5.1 Description of Amino Acid Sequences of Proteins . . . 53
3.5.2 Defining the Binding Pocket of the Kinases . . . 54
3.5.3 Calculation of the Per-Residue RMSD Values . . . 54
3.5.4 Calculation Of The Normalized B-Factor . . . 55
3.5.5 Building the Random Forest Models . . . 56
3.6 Results . . . 60
3.6.1 Exploration of the B-Factor as a Measure for Flexibility of a Residue at a Given Position . . . 60
3.6.2 Exploration of the RMSD as a Measure for the Flexibility of a Residue at a Given Position . . . 68
3.7 Conclusion and Future Directions . . . 76

4 Prediction of Selective Compounds Using 2D-Kinase Random Forest Models 80

4.1 Introduction . . . 80
4.2 Methods . . . 80
4.2.1 A Manually Curated Data Set of Selective Compounds for Testing RF Models . . . 80
4.2.2 Building The Random Forest Models . . . 84
4.3 Results . . . 86
4.3.1 Models For Predicting Selectivity . . . 86
4.4 Conclusions and Future Directions . . . 95

5 Using 3D-Kinase Models to Classify Candidate compounds in Different Binding Affinity Ranges 97

5.1 Introduction 97
5.2 Methods 98
5.2.1 Annotation of Each Kinase PDB Structure with Binding Affinity Data 98
5.2.2 Calculation of the Intermolecular Interactions between a Ligand and its Kinase Target 98
5.2.3 Using Random Forest Models to Test Whether Ligand-Kinase Interactions Can Be Used to Discriminate between Binders in Different Binding Affinity Ranges 99
5.2.4 Comparison of the Shape and Binding Interactions of a Promiscuous and a Selective Compound 100
5.2.5 Building the CRANkS Interaction Maps 101
5.3 Results 105
5.3.1 Case Study : Comparison of the Binding Shapes of a Promiscuous and a Selective Kinase Ligand 105
5.3.2 Case Study: Comparison of the Binding Interactions of a Promiscuous and a Selective Kinase Ligand 111
5.3.3 Investigating the Use of Binding Interactions To Differentiate Between Strong and Weak Binders on a Per-Target Basis 121
5.3.4 Use of CRANkS Interaction Maps to Differentiate Between Strong and Weak Binders 127
5.4 Conclusion and Future Directions 129

6 Concluding Remarks **133**

References **138**

A Appendix **154**

List of Figures

1.1 Illustration of the use of target and compound descriptors in Proteochemometric Modelling (PCM). 10

2.1 Pathway count for approved drugs, withdrawn drugs and investigational drugs. 24

2.2 For each drug, the number of its unique targets is plotted versus the number of its associated unique pathways. The Hexagonal binning plot shows the number of drug targets versus the number of pathways for each drug 26

2.3 Distributions of pathway counts for each protein within the groups of 'undrugged' proteins and drug targets 28

2.4 Pathways contain a wide range of target classes. 30

2.5 Target classes of drug targets mapped to exemplarily pathways show a broad distribution in associated pathways. 31

2.6 Connection between drug targets, disease areas and a set of 'representative' pathways. 33

2.7 Within the targets classes 'Enzyme' and 'Kinase', disease areas can be varied according to pathways. 34

3.1 Illustration of the bilobal kinase structure. 43

3.2 Illustration of a decision tree and its implementation in the Random Forest (RF) Algorithm. 46

3.3 Illustration of the building and the internal validation (using *out of bag* samples) of the Random Forest (RF) Algorithm. 48

3.4 Illustration of k-fold cross validation (CV) method. 48

3.5 Illustration of the Confusion Matrix 51

3.6 Illustration of the ROC Curve. 52

3.7 Heatmap of the normalized B-Factors for the 85 binding residues of the TKL kinases. 61

3.8 The relative feature importance of the different features in the four RF models for the target class TK. 68

3.9 The relative feature importance of the different features in the four RF models for the target class TK 73

3.10 Distribution of the per-residue RMSD values (in Å) for a given target in the target class AGC. 74
3.11 The RMSD values (in Å) for the 85 KLIFS-defined binding pocket residue positions for the target class CAMK, plotted according to the KLIFS nomenclature. 75
4.1 Kinase targets of selective kinase compounds collected from literature. . 83
4.2 Relative feature importance of the different kinase models. 89
5.1 Calculating the protein-ligand interaction counter 104
5.2 Visualisation of filtering out promiscuous compounds using interaction maps and/or shape maps. 106
5.3 Distribution of pairwise SuCOS scores for promiscuous kinase ligands . 108
5.4 Distribution of pairwise SuCOS scores for the selective ligand Tocafitinib (MI1), selective against JAK1, JAK2, JAK3. 110
5.5 Sidechain and Backbone Interactions of Staurosporine across Kinase targets. 113
5.6 Interactions of Staurosporine in the kinases ALK, PDK1 and GSK3B. . 114
5.7 Sidechain Interactions of the selective Ligand MI1 (CP-690550) with 4 Kinase Targets: JAK1, JAK2, JAK3 and TYK2. 117
5.8 Backbone chain interactions of the selective ligand MI1 (CP-690550) with its 4 Kinase targets: JAK1, JAK2, JAK3 and TYK2. 118
5.9 Ligand-Target interactions between MI1 (CP-690550 or Tofacitinib) and its targets JAK1, JAK2, JAK3 and TYK2. 119
5.10 Visualization of how protein-ligand interaction maps could ultimately be used to classify compounds as selective or promiscuous. 122
5.11 Testing the potential of the CRANkS interaction maps to classify compounds in different binding affinity ranges. 127
5.12 Re-docked conformers of the ligand SBC into its orginal target CDK2 (PDB code: 2BKZ), together with the interaction grid. 129
A.1 Drug targets, drug targets associated with only investigational drugs, proteins with bioactivity data and proteins without bioactivity data within 24 representative pathways. 154
A.2 Distribution of the number of pathways for the groups ‘investigational’, ‘withdrawn’ and ‘approved’ drugs in Figure 2.1. 155

A.3 Heatmap showing the number of proteins common between any two pathways within my set of 'representative' pathways. 156
A.4 24 'representative pathways' clustered according to their contained proteins. 157
A.5 Close-up of the clustering of the 2190 Pathways from the Reactome Pathway Database according to proteins involved in multiple pathways. . . . 158
A.6 Close-up of the chosen 408 pathways after cut-off, clustered according to proteins common between them. 159
A.7 Exemplary distributions of the pairwise correlation coefficients for the different input features in the RF Models. 160
A.8 Number of protein kinases with a specific number of ligands. 165
A.9 The Distribution of the number of crystal structures annotated with binding affinities for kinase in the DFG-in and αC helix in conformations. . 166
A.10 The Distribution of the binding affinities values (pKI, pKd, pIC_{50}) for all crystal structures for a given target in the DFG-in conformation, αC helix in conformation. 167
A.11 Distribution of the RMSD values for the pairwise alignment of promiscuous kinase ligands. 169
A.12 Distribution of the RMSD for the pairwise alignment of the selective kinase inhibitor Tofacitinib (MI1). 171

List of Tables

1.1 Illustration of selective and promiscuous compounds within and across target classes. 6

2.1 Relabelling of target classes in ChEMBL23. 21

2.2 Results of the Dunn post-hoc tests for testing the statistical significance of the number of pathways for drugs 25

2.3 Results of the Pearson correlation tests, the Kendall τ and the distance correlation used for testing the statistical significance of the correlation between the number of unique targets and unique pathways for a given drug. 27

2.4 p-values between the distributions of 'undrugged' proteins and drug targets in Figure 2.3 derived by using a Poisson Distribution Regression. . 28

3.1 The number of targets used for RMSD comparison for the DFG-in- and out conformation, separated by kinase subclasses. 55

3.2 Number of the active/inactive data points, targets and compounds used in the RF models to test the influence of the B-Factors. 58

3.3 Number of the features for each RF model used to test the influence of the B-Factors. 58

3.4 Number of the active/inactive data points, targets and compounds used in the RF models to test the influence of the RMSD values. 59

3.5 Number of the features for each RF model used to test the influence of the RMSD values. 59

3.6 The different parameters of the random forest models. 60

3.7 Evaluation metrices of RF models for the TK kinase group, trained on a data set excluding all selective compounds. 64

3.8 Evaluation metrices of RF models for the CAMK kinase group, trained on a data set excluding all selective compounds. 65

3.9 Evaluation metrices of RF models for the AGC kinase group, trained on a data set excluding all selective compounds. 66

3.10 Evaluation metrices of RF models for the GMGC kinase group, trained on a data set excluding all selective compounds. 67

3.11 Evaluation metrices of RF models for the TK kinase group, trained on a data set excluding all selective compounds. 70

3.12 Evaluation metrices of RF models for the CAMK kinase group, trained on a data set excluding all selective compounds. 71
3.13 Evaluation metrices of RF models for the GMGC kinase group, trained on a data set excluding all selective compounds. 72
4.1 Number of the active/inactive data points, targets and compounds in each model. 85
4.2 Evaluation metrics of RF models for the different groups of kinases, trained on a data set excluding all selective compounds. 87
4.3 Comparison of the actual and predicted activity of target-compound combinations for selective and two 'historic' compounds. 91
4.4 Predicted activity of a selective compounds against all targets in the data set. 92
4.5 Predictions for selective compounds not present in the data set provided by Fulle *et al.* 94
5.1 Data and performance of models at a threshold of 6.3 pK-value (500 nM). 124
5.2 Data and performance of models at a threshold of 7 pK-value (10 nM). 125
5.3 Data and performance of models at a threshold of 8 pK-value (100 nM). 126
5.4 Interaction Novelty Scores for re-docked ligands derived by using CRANkS interaction maps. 128
5.5 Interaction novelty scores for cross-docked ligands derived by using CRANkS interaction maps. 128
A.1 57 selective kinase compounds and 2 historic compounds used to test my models in terms of selectivity prediction. 163

1 Introduction

1.1 Motivation

Ever since its beginning, humanity has been plagued with diseases, most prominently cancer (Mukherjee, 2010); the Bible also describes instances of diabetes and autonomic neuropathy (Mathew and Pandian, 2010). The key question in fighting diseases is to establish the relation between the disease type, the physiological processes in the body and the modulation of the protein(s) and its (their) associated pathways in order to derive safe and effective drugs to prevent disease onset and progression (Waring et al., 2015). In the following, a protein will be referred to as a target, which can be modulated using compounds (also referred to as inhibitors); a drug is a compound which has been approved for disease prevention and/or treatment in the human body.

From a holistic perspective, a drug or a compound can cause changes in the normal biological processes of a human body, thereby influencing its phenotype. It can also cause changes in a phenotypic assay (Schneider and Klabunde, 2013). The causes for these changes traced back to two key factors:

1. Within pathways, protein targets are connected with other targets and entities. Unknown connections between the entities in a pathway can result in a confusing biological read-out for a compound or unexpected side effects in patients for a given drug. This phenomenon is referred to as 'pathway crosstalk' and is further discussed in section 1.2.

2. A given compound can either interact with a multitude of targets or a few, closely related targets; this compound is then denoted as 'promiscuous' or 'selective' respectively. The selectivity or promiscuity of a compound depends on structural factors: its chemical moieties must be complementary to the binding residues found in the target's binding pocket and be able to adopt a shape suitable for binding. Furthermore, the binding pocket itself is flexible and single residues, 'gate-keeper residues', can prevent or facilitate binding of a specific compound. These concepts are further illustrated in Section 1.3.

These two factors lead to the research questions addressed in my thesis, discussed in Section 1.4. My journey has been supported by the explosive growth in databases: namely the Protein Database (PDB) (Bernstein et al., 1977), the Pathway databases Reactome (Fabregat et al., 2017) and KEGG PATHWAY (Kanehisa et al., 2016), and, lastly, the availability of selective inhibitors (Arrowsmith et al., 2015) and large kinase data sets for machine learning (Sorgenfrei et al., 2018).

1.2 Pathway Crosstalk: The Biological Analogue to Compound Selectivity and Promiscuity

To understand human physiology, disease onset and progression, the human body has been hierarchically organised in levels, from molecular interactions at the level of targets and pathways, to the cell and tissue-level, to the organs and organ systems, to the entire human organism as the so-called 'phenotype' (Berger and Iyengar, 2011; Schneider and Klabunde, 2013). 'Systems Biology' is a holistic approach to model biological phenomena based on large datasets ('omics' data) for these levels (Berg, 2014). The human proteome of the body has been extensively characterised (Thul and Lindskog, 2017). Proteins – potential drug targets – can be expressed in differing concentrations at different points in time in the cells within organs, with some protein-coding genes expressed ubiquitously ('housekeeping proteome') (Uhlen et al., 2015).

Finally, the level of research interest in proteins varies according to their perceived importance in biological function and/or disease involvement, as shown for genes (Haynes et al., 2018). Within a cell, proteins do not act in isolation, but are organised both spatially and temporally with other molecules in sets of interactions to elicit specific biological processes within the cell. These responses subsequently modulate the phenotype. They are 'grouped' in cell signalling pathways according to their (perceived) role in biological processes by manual curation (Thul and Lindskog, 2017). A biological pathway is a cascade of interactions between the molecules in a cell, including but not limited to proteins, RNA, and DNA and chemical substrates. These actions lead to a measurable product or change in a cell. Pathway databases include the Reactome Pathway Database (Fabregat et al., 2017), KEGG PATHWAY Database (Kanehisa et al., 2016), PathCards (Belinky et al., 2015) and the crowd-sourced pathway database WikiPathways (Pico et al., 2008; Kutmon et al., 2015; Kelder et al., 2011).

Pathway databases can differ in their their definition of pathways and their associated proteins. Depending on the focus of the database and the curator(s), pathways may be 'incomplete' between pathway databases, partly because researchers and curators focus on different aspects of pathways (Chowdhury and Sarkar, 2015). Thus, pathway information can be distributed in different amounts across various pathway databases. So far, there is no standard pathway assembly and naming convention for pathway databases (Chowdhury and Sarkar, 2015). A key factor in disease onset, development and progression and also in drug resistance is pathway crosstalk, in which the activation of one pathway ensures the downstream response of a second pathway (Sam et al., 2016). Different methods to calculate crosstalk have been developed and different crosstalk databases established, including OmniPath (Türei et al., 2016), XtalkDB (Sam et al., 2016). Many of these interactions might be highly context-specific (Fabregat et al., 2017) and perturbations of these functions are involved in disease processes (Kanehisa et al., 2016). Drug targets might be involved in many cellular events and pathways, which plays a key role in deriving their therapeutic potential. For example, the Glycogen Synthase Kinase-3 (GSK-3) is involved in the Wnt signalling pathway, the insulin pathway, and in pathways important to the neuronal development and transcription. A perturbation of its function thus results in a variety of system diseases, including Alzheimer's Disease, diabetes and/or psychiatric disorders (Saraswati et al., 2018). This exemplifies the fine balance between drug safety and efficacy when choosing a drug target within biological systems.

Properties of ideal drug targets, according to Gashaw *et al.*, include having a function in the disease phenotype, an experimentally determined three-dimensional (3D) structure (or a close homolog thereof), and favourable, predicted or experimentally-derived, side effects and phenotypic data. The number of side effects can be expected to rise with the number of targets modulated by the drug (Gashaw et al., 2011). Thus, ideally, all targets, which are potentially modulated by a drug candidate have to be carefully investigated by trained biologists in terms of their connections to other pathways.

Additionally, Bull *et al.* used machine-learning approaches to determine which features could distinguish between a data set of drug targets and current 'non-targets'. Their results showed that drug targets are mostly non-polar, membrane-bound proteins, with long *in vivo* half-lives of the drug targets, which indicates good experimental accessibility

(Bull and Doig, 2015).

These approaches fail to take the position of a drug target in a pathway and its connections to other biological pathways into account. Yet, due to the (often unknown or unexplored) cascade-like connections of entities in pathways, any systematic exploration of drug targets is highly challenging.

1.3 Drug Pharmacology: Deconvoluting The Selectivity or Promiscuity of a Given Compound

The vast 'chemical space' encompasses all potential small organic molecules. The number of small, 'drug-like' molecules with up to 30 carbon, nitrogen, oxygen, and sulfur atoms has been estimated to be 10^{60} organic molecules (Bohacek et al., 1996). Yet, for the pharmaceutical researcher, only areas with biologically active ('bioactive') compounds are used as starting points for drug development. These areas are can be referred to as 'biological space' or 'biologically relevant' space. These molecules will be used in target-driven drug discovery as starting points. Methods utilised include similarity, shape and pharmacophore searching methods (Bohacek et al., 1996; Ai et al., 2015) for novel compounds, scaffold-hopping approaches to derive novel chemotypes (Hu et al., 2016; Sun et al., 2017) and/or the development of quantitative structure-activity relationships (QSAR) (Bradley et al., 2014; Cruz-Monteagudo et al., 2014) that explore the interplay between the 3D structure of a chemical and its biological activity towards a specific target. These traditional approaches could lead to bias in drug development; for example, it has been reported that compounds in industrial drug screens are biased towards molecules that proteins have evolved to recognise (Hert et al., 2009).

Thus, identifying active compounds against a specific target already presents a challenge. However, determining whether a compound is selective or promiscuous against a given target panel (active or inactive against a given target panel) is the main challenge in current drug pharmacology. The available experimental data (and computational predictions) can allow an estimation about whether that compound is selecive or promiscuous. For example, the tool compound FM-381 is highly selective for one target (the kinase JAK3), while the rigid small molecule staurosporine is promiscuous as it binds to multiple kinase targets (Wang et al., 2016). Computational predictions are

often used to guide experimental testing and propose hypotheses about compound and protein features that are most important in determining the selectivity or promiscuity of a compound (Ain et al., 2014; Liggi et al., 2013).

Whether a compound is selective or promiscuous depends on the available data; for example, the molecule staurosporine used to be a selective compound for probing the Protein Kinase C (PKC) until 2003, when further testing of staurosporine revealed that it is actually a promiscuous compound that binds to a multitude of targets (Wang et al., 2016). Bajorath *et al.* postulated that the ultimate goal of chemogenomics is to predict all missing target-ligand interactions so as to gain a global overview of the ligand-target matrix (Vogt et al., 2018).

However, compound selectivity or promiscuity has to be defined within the context of the so-called 'protein target classes', defined by Barnash *et al* as a 'family of structurally and/or functionally related proteins' (Barnash et al., 2017). In these classes, high-resolution 3D-structures of one member could ideally be used for analyzing drug design opportunities for other members. Examples of promiscuity or selectivity as target class-dependent properties are given in Table 1.1.

Promiscuous compounds have been referred to as 'frequent hitters', as they generate a positive (binding) readout in multiple diverse biological assays, adding another layer of complexity to biological readouts (see Section 1.2). These screening hits can also be sometimes false positive assay signals such as Pan-Assay Interference Compounds (PAINS), which appear to be active under assay conditions due to non-specific binding (*e.g.* chelation, autofluorescence, aggregation, or degradation) (Baell and Walters, 2014; Baell and Holloway, 2010; Jasial et al., 2017). While methods for filtering them out by presence of substructure exist, Jasial *et al.* showed that PAINS compounds can be consistently active or inactive, despite their PAINS substructure, indicating that the structural frame of a PAINS substructure can play a role in its effects (Jasial et al., 2017). Compared to these false positives, truly 'promiscuous' compounds bind to different proteins.

Knowing the selectivity or promiscuity of a compound in the context of drug design is particularly important for drug polypharmacology, which refers to a drug achieving a specific effect by binding to a primary target ('on-target') and multiple secondary targets

('off-targets'). Knowledge about all the targets of a drug is necessary for predicting, explaining and/or avoiding side effects.

Interestingly, drugs have been shown to be more promiscuous than bioactive compounds (Jasial et al., 2016). Bajorath *et al.* explained these findings by a more extensive testing of drugs due to safety concerns or a preference towards promiscuous compounds in drug research due to a heightened therapeutic efficacy ('promiscuity enrichment model') (Jasial et al., 2016).

Table 1.1: Illustration of selective and promiscuous compounds within and across target classes, using four exemplary situations. For simplicity's sake, only 'Kinases', 'GPCR' and 'Transporter' are used as the three representative target classes and targets. Experimentally verified binding or nonbinding is symbolized by a green checkmark or red cross respectively; a target-compound combination with no known activity data is annotated with a black question mark. A 'globally promiscuous' compound can bind to a variety of targets from other target classes ('global polypharmacology'); while compound promiscuity within a target family as is referred to as 'local polypharmacology' (Wang et al., 2016). Selective compounds to be used as chemical probes or tool compounds are required to possess, among other qualities, a >30-fold selectivity to its target(s) relative to the other sequence-related targets of the same target family. Also, they should have been successfully profiled against an 'industry standard' panel of pharmaceutically important off-targets (Arrowsmith et al., 2015). Lastly, the knowledge about each target-compound interaction is limited; thus, the classification of a compound as 'promiscuous' or 'selective' often depends on the available data for this particular compound at a given point in time.

	Target Class						
	Kinase			GPCR		Transporter	
Targets / Compounds	JAK1	JAK2	MET	A2A	BK2	SERT	NET
'Globally' promiscuous compound	✓	✓	✓	✓	✓	×	✓
'Locally' promiscuous compound	✓	✓	✓	×	×	×	×
Selective compound	✓	✓	×	×	×	×	×
'Normal' situation for a compound	✓	?	×	?	✓	?	✓

1.3.1 The Molecular Basis of Selectivity and Promiscuity

To understand and predict the selectivity of proteins and compounds, one must understand the concept of both protein and ligand selectivity at the molecular level, namely the three-dimensional binding of a ligand in a specific conformation to a protein in a specific conformation. Small changes in binding mode can render a compound selective or promiscuous towards a given target panel (Huggins et al., 2012). The binding of a ligand to a given protein target, and, by extension, the promiscuity or selectivity of a ligand, depends, to varying degrees, on the following key factors: (i) complementary features of the ligand and target and the ability of both (ii) the ligand and (iii) protein to adopt a conformation favourable for binding. The physicochemical properties of the compound are governed by its chemical moieties and the shape it can adopt. The functionality of a binding pocket is determined by its amino acid residues, its shape and its location in the protein (Stank et al., 2016), and in some cases, the presence of one or more co-factors.

1. The target and its ligand must have complementary features for binding to occur

Binding between a ligand and its targets occurs when the binding residues of the binding pocket and the chemical moities of the compound must have complementary features for binding. The number and types of interaction determine the strength of binding and hence the binding affinity between protein and ligand.

2. The compound has to adopt a conformation favourable for binding

Both the compound and the amino acid residues in the binding pocket have to be able to adopt conformations which promote binding. Various hypotheses about the relationship between the molecular properties of a compound and its promiscuity or selectivity have been put forward; notably, increased hydrophobicity is correlated with compound promiscuity, although Shoichet *et al.* warned that this could be an artifact due to hydrophobic drug aggregation in high throughput screening (Seidler et al., 2003). Haupt *et al.*, however, inferred that for promiscuous drugs, compound properties such as conformational flexibility, molecular weight, and hydrophobicity are not important for promiscuity (Haupt et al., 2013). Sturm *et al.* demonstrated that promiscuous ligands can adapt to different protein environments by conformation change or using different

chemical features to interact with the protein and emphasised that ligand flexibility can be used to explain the differences in target binding sites (Sturm et al., 2012). Sturm *et al.* also identified chemotypes capable of binding to a variety of protein environments.

However, rigid molecules such as Staurosporine have been reported to be promiscuous, due to their ability to mimic the binding mode of ATP, the actual substrate, despite the lack of similarity. Given that Staurosporine is rigid, this highlights that the flexibility of compounds is only one factor contributing to the selectivity (or promiscuity) of a compound (Tanramluk et al., 2009). This discrepancy in findings could be explained by the fact that some drugs tend to be designed as more rigid entities, to avoid or reduce conformational entropy loss upon binding (Haupt et al., 2013).

3. The protein has to adapt a shape for binding and provide residues for binding

Sturm *et al.* and Haupt *et al.* both demonstrated that structural similarity of the binding site plays an important role in target promiscuity (Haupt et al., 2013; Sturm et al., 2012). In addition, Sturm *et al.* demonstrated that similar binding sites do not have to share evolutionary relationships, such as amino acid sequence or overall fold conservation (Sturm et al., 2012). In an extensive study about the binding pocket similarity of all highly resolved protein structures, Duran-Frigola *et al.* showed that similar cavities were found among proteins with no apparent relationship. Although sequence-related proteins tend to share cavities, proteins with close sequence similarity could show divergence in the binding site. Additionally, most structures have generally between one to five cavities, a potentially rich opportunity for selective targeting of proteins. They suggested that structure-based polypharmacology could be used to fine-tune compound selectivity and promiscuity (Duran-Frigola and Aloy, 2013). Their findings agree with the approach by Volkamer *et al.*, who collected a data set of binding pockets from kinases structures, defined on- and off-targets based on their target of interest, and fused these pockets into a 'combined' pocket. This combined pocket could then be used to derive a 'difference pocket' with the following areas: the 'common core', the 'target-specific', and 'off-target specific areas'. Compounds can thus be developed with complementary features for the target-specific part. This approach offers an easy visual interpretation of selectivity criteria based only on target features (Volkamer et al., 2016). Lastly, Gao *et al.* showed that the structural space of protein pockets is likely completely explored and can be

described in terms of about 1,000 pocket shapes (Gao and Skolnick, 2013). They also noted that many protein binding pockets are promiscuous, binding a variety of ligands.

Lastly, the environment in the protein has to be complementary to the compound features. The shape and volume of the protein pocket needs to correspond to (or exceed) the shape of the putative binder (Stank et al., 2016). The structural flexbility of a binding pocket, or 'protein dynamics', should allow a protein to adapt to its binding partner. For example, it has been reported that the co-crystallization of one specific receptor with different ligands can result in a variety of observed pocket shapes (Stank et al., 2016).

Flexible regions in the binding pocket or flexible residues, such as the 'gatekeeper' residues, can promote or hinder binding (Tanramluk et al., 2009). Stank *et al.* classified five different classes of protein pocket dynamics for the better assessment of the druggability of a protein pocket (Stank et al., 2016); Najmanovich *et al.* analysed sidechain flexibility in target-compound binding, specifically to what extent binding pocket residues change their conformation upon binding of a small ligand. Najmanovich *et al.* found that in 85% of their data set (with 1,333 paired protein structures, each pair consisting of the complexed, holo, and uncomplexed, apo, protein structure), up to three residues undergo a change in residue flexibility. Based on their analyses, they proposed the following scale for residue flexibility: Lys > Arg, Gln, Met > Glu, Ile, Leu > Asn, Thr, Val, Tyr, Ser, His, Asp > Cys, Trp, Phe. Najmanovich *et al.* emphasize that lysine sidechains in binding pockets are flexible 25 times more often than the phenylalanine and tryptophan sidechains (Najmanovich et al., 2000). Thus, the flexibility of residues and/or flexible regions in the binding pocket may play a corresponding role to compound flexibility as they too can prevent or promote compound-target binding.

1.3.2 Identifying Protein and Ligand Features Responsible For Selectivity or Promiscuity

Knowing the selectivity (or promiscuity) of a compound across a set of biological targets can be invaluable in drug discovery, as this knowledge may help to explain unwanted side effects. In contrast to the high experimental cost of screening a compound across a target panel, *in silico* methods, such as machine learning (Giblin et al., 2018) or target prediction tools (Mervin et al., 2015; Keiser et al., 2007) provide a quick and inexpensive

way to assess the selectivity of a compound.

To the best of my knowledge, however, no work has been published on specifically deriving chemical probes using machine learning.

However, Proteochemometric Modelling (PCM) has been developed to predict and explain compound selectivity. As an extension of QSAR, PCM employs both ligand and target features simultaneously to predict the bioactivity of compounds (Figure 1.1). As PCM assesses which compound and target features are necessary for compound selectivity and promiscuity, it enables the deconvolution of both the target and the chemical spaces in parallel (Cortés-Ciriano et al., 2015). For an extensive review of PCM, see Corts-Ciriano *et al.*, 2015.

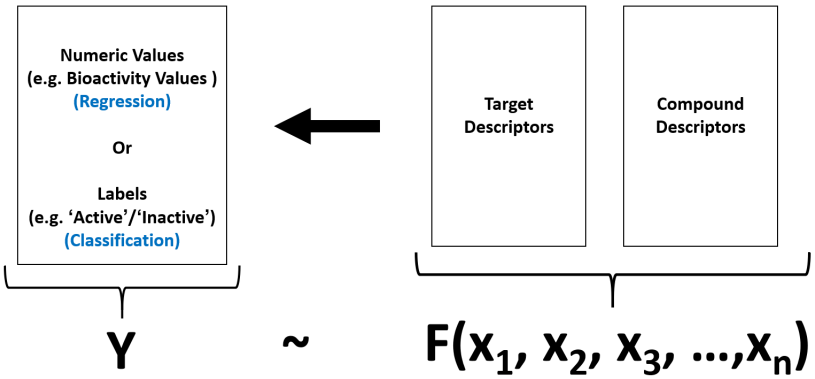


Figure 1.1: Illustration of the use of target and compound descriptors in Proteochemometric Modelling (PCM). In PCM, a compound-target pair is encoded by a combination of protein and compound descriptors. This target-compound combination is associated with either a numerical value (*e.g.* binding affinity, bioactivity value) in the case of a regression model, or a categorical label (*e.g.* 'active', 'inactive') in the case of a classification model, respectively. This matrix is used as the training set for the model of interest as detailed in Section 3.4.1.

PCM has been used for selectivity prediction in different target classes, such as kinases (Sorgenfrei et al., 2018), proteases (Ain et al., 2014), cycooxygenases (Cortés-Ciriano et al., 2015) and adenosine receptors (Van Westen et al., 2012). It has been used for both classification (Giblin et al., 2018) and regression models (Ain et al., 2014), using a variety of machine learning approaches, such as Random Forest or Support Vector Machines (Cortes-Ciriano et al., 2015).

Cortès-Ciriano *et al.* quantified the contribution of chemical substructures on decreasing or increasing the bioactivity value of a substructure towards the targets COX-1 or

COX-2 using a PCM model built on 11 mammalian cyclooxygenases, 3,228 distinct compounds and 4,937 datapoints (13.9% matrix completeness). Specific compound features describing the substructures in the training set were 'virtually removed' by iteratively setting each substructure to zero in the chosen compound descriptor. The model was retrained using these 'changed' compound descriptors, and the difference in bioactivity with a substructure 'present' or 'absent' in the compounds was then calculated and thus the influence of a substructure on the bioactivity was assessed, allowing the retrieval of specific substructures that could increase or decrease the bioactivity of a compound towards a specific COX isoenzyme type (Cortes-Ciriano et al., 2015). Ain *et al.* extended Cortès-Ciriano's work by systematically virtually removing both the features corresponding to the chemical substructure and amino acid residues of the protein (Ain et al., 2014); again allowing the retrieval of important substructures and residues for bioactivity prediction. Subramanian *et al.* described the protein binding pockets with knowledge-based and Schrödinger's WaterMap–derived fields in their PCM models, which allowed a quick visual identification of the main affinity and selectivity hot spots (Subramanian et al., 2013).

While PCM has been extensively used for selectivity prediction, its use is limited by the available data. As I show in Section 2, public data sources are biased towards specific target classes, thus restricting the availability of data and the application of potential models (Cortés-Ciriano et al., 2015). Additionally, the data can be skewed in terms of reported bioactivity values, resulting in an 'imbalanced class distribution'. Active data are more prevalent than inactive data, and thus could bias the machine learning algorithm (Cortés-Ciriano et al., 2015). To account for a lack of data, van Westen *et al.* used bioactivity data from orthologuous proteins, closely related proteins stemming from the same gene found in different species (Van Westen et al., 2012). As orthologous proteins plausibly have similar biological functions and share similar bioactivity profiles, the mapping between species is supposed to extend the available chemical and bioactivity space (Mervin et al., 2018). By combining bioactivity data from human and rat adenosine receptors to predict selective compounds against target human adenosine receptors, van Westen *et al.* concluded that models built only on human data perform worse than models built on both rat and human data (Van Westen et al., 2012). Similarly, Mervin *et al.* analysed the chemical space provided by orthologue

data and noted its complementarity to the current human chemical space in terms of chemical novelty and diversity, concluding that the inclusion of data from orthologous targets has to be decided on a per target basis and parameter/model basis (Mervin et al., 2018).

Similarly, compound selectivity can be inferred by predicting the target for a compound. Examples for current target prediction tools include PIDGIN by Mervin *et al.* (Mervin et al., 2015), SEA (Similarity Ensemble Approach) by Keiser *et al.* (Keiser et al., 2007) and the SwissTargetPrediction Tool (Gfeller et al., 2014). Paricharak *et al.* combined both PCM and target prediction for predicting the bioactivity values and the polypharmacology of a compound and discovered novel Plasmodium falciparum DHFR inhibitors. Paricharak *et al.* especially emphasized that target prediction methods predict the likelihood that a compound binds to the target in question, while PCM modelling aims to predict a numerical value in the form of binding affinity or bioactivity (Paricharak et al., 2015). Both methods are thus highly complementary.

Similarly, the Similarity Ensemble Approach (SEA) groups proteins using the chemical similarity of their known ligands. Although biological relationships were found only based on chemical features, Keiser *et al.* could experimentally verify their (often surprising) results, thus validating their approach. This method thus allows the investigation of the relationship between structurally dissimilar proteins and subsequently the elucidation of compound promiscuity (Keiser et al., 2007). The Target Interference GEneratoR (TIGER), by Schneider *et al.*, is also based on ligand similarity: the query compound is converted into a topologiocal pharmacophore descriptiors, which is then compared to the pharmacophore representations of reference molecules (with associated targets) using self-organizing maps. These reference compounds have been clustered into sets of locally similar molecules. The most similar compounds to the query in terms of topological pharmacophore descriptiors are then used to calculate similarity scores and subsequently assign potential drug targets to the query compound. TIGER was able to identify targets which other prediction models, such as SPIDER and SEA, could not. However, as experimental verification showed that some target hypotheses were incorrect, Schneider *et al.* also concluded that TIGER seems to work best at the target family level. This again highlights the difficulty of developing a hypothesis for a compound's selectivity or promiscuity within and accross target families (Schneider and

Schneider, 2017).

Alternatively, the question of binding site similarity can also be adressed in terms of protein binding pocket shape and interactions. For example, Tanramluk *et al.* compared the binding sites for ATP and its competitive inhibitor, staurosporine, accross 80 kinases and showed, by grouping structures with similar inhibitor binding profiles, that the shape of the binding pocket can contribute to inhibitor selectivity (Tanramluk et al., 2009). So far, to the best of my knowledge, while target binding site similarity has been compared extensively, and multiple methods have been developed for the comparison of binding pockets, target binding pocket similarity has not been used to predict compounds for binding (Ehrt et al., 2018; Kellenberger et al., 2008).

1.3.3 Tool Compounds: Selective Compounds To Explore Biological Questions

The selectivity of a given compound will have consequences on the entire biological system, as its binding to target(s) can cause a wide range of pathways to be activated. The activation of multiple pathways by the same compound points to shared targets between the pathways in question (Hancock et al., 2010).

Various measurements have been developed to quantify the selectivity of a particular compound towards a target, *e.g.* the Gini selectivity metric (or Gini coefficient) (Bosc et al., 2017), the selectivity score (Karaman et al., 2008), and the window score (WS) and ranking score (RS) (Bosc et al., 2017). Hancock *et al.* screened 32 compounds across a panel of representative cell signaling pathways and assessed the selectivity of compounds on pathways with a compound-pathway selectivity score. With this multi-pathway profiling approach, they successfully characterised the on- and off-pathway activities of compounds (Hancock et al., 2010).

Deconvoluting the effect of a promiscuous compound on a biological system can be difficult, as there can be biological effects ('readouts') overlaid over one another. Thus, to explore a biological system with small molecules, researchers use so-called 'tool compounds', also known as 'chemical probes'. These small molecules are ideally both highly potent (< 100 nM) and highly selective against the target of interest (> 30-fold difference in binding affinity relative to other sequence-related proteins). As tool compounds have

a definite mode of action and thus selectively modulate the function of the protein, they can help elucidate mechanistic and phenotypic questions about the role of the target in the biological system, and are the biochemical equivalent of the genetic approaches CRISPR and RNAi. Ideally, each protein should be characterized by a second probe with distinct structural features (orthogonal) to the first one (Arrowsmith et al., 2015).

Based on these criteria, tool compounds have been collected in "crowd-sourcing" initiatives, such as the Chemical Probes Portal ('www.chemicalprobes.org') (Arrowsmith et al., 2015). As an extension, Butler *et al.* mined ChEMBL for compounds that fit these criteria (Butler et al., 2017). This data-driven selection can result in the false classification of tool compounds due to historical or commercial biases, and subsequent incomplete knowledge (Antolin et al., 2017) In response, Wang *et al.* developed the 'Tool Score' (TS) to assess the selectivity of a compound towards the target of interest ('on-target'). The TS is a data-driven model based on the confidence in the available bioactivity data ('strength') and selectivity features learnt from machine learning ('selectivity'). It is used to select the 'optimal selective' tool compound for a specific target in phenotypic screening based on the available bioactivity data. Their approach allows the use of heterogenous bioactivity data consisting of different measurement and assay types. Subsequently, different values for 'strength' are assigned according to the data origin, such as ChEMBL, DrugBank or in-house datasets (Wang et al., 2016). In terms of selectivity, training on a dataset of 507 compounds showed that the best features to distinguish between 'good' tool compounds and unselective molecules were (a) the difference in spread of bioactivity values between the on- and the off-targets, (b) the significance of that spread as assessed by a Wilcoxon test and (c) a high number of datapoints available for the on-target molecule in specific concentration bins (Wang et al., 2016). These features were incorporated into model-independent metrics, which, when multiplied with the 'strength', yields TS. Compounds ranked highly by TS exhibited on-target potency and selectivity. However, as outlined in Section 1.2, the phenotypic response of a compound is not only a function of its target(s), but also of the position and connection of said target(s) in the pathway. Subsequently, if the targeted protein was upstream of multiple pathways, then compounds with high TS do not correlate with high pathway selectivity (Wang et al., 2016).

This example emphasises that both the targets of the compound and their position and

connection in the pathway have to be considered in order to understand biological read-outs from assays. The classification of a compound as a 'tool compound' depends on its known bioactivity data. As mentioned earlier, before 2003, Staurosporine was considered to be a selective compound for probing Protein Kinase C (PKC), until further testing showed the heightened promiscuity of Staurosporine against multiple kinases (Wang et al., 2016). Subsequently, Staurosporine has been listed as a 'historical' compound in the Chemical Probes Portal (Butler et al., 2017).

Wang *et al.* emphasised explicitly that their TS is based on currently available bioactivity data, and thus would miss 'good' tool compounds if a poor coverage of bioactivity data was available (Wang et al., 2016). Finally, from an experimental perspective, the selectivity of a compound is concentration-dependent, which needs to be reflected in the experimental set-up of the phenotypic system. For example, nilotinib is selective for BCR-ABL at an effective concentration of 10 nM or lower ('on target biology'), whereas at effective concentrations above 1 mM nilotinib also binds to other targets. Subsequently, the observed phenotype can then no longer be reliably traced back to on-target effects (Wang et al., 2016).

To the best of my knowledge, no work has been published on deriving tool compounds using machine learning. Generating bioactivity values by machine learning might not generate accurate enough data to augment the currently available data. Examples of current approaches to derive protein and compound features that account for selectivity were discussed in Section 1.3.2.

1.4 Project Aims and Thesis Structure

I have presented the two main challenges in deriving selective compounds as potential drugs: the crosstalk between pathways and the structural basis of protein-compound interaction, where small changes in features could change the selectivity or promiscuity of a compound. This thesis is separated into four parts corresponding to the different project aims :

1. In Section 2, I show that the success of a drug does not depend on the number of its pathways, but their nature. Depending on target class, drug targets are generally associated with a wider range of pathways than 'undrugged' proteins,

highlighting either a research bias towards drug targets or, possibly, that drug targets are chosen due to a higher involvement in pathway crosstalk. With these analyses, drug targets were investigated from a pathway-centric perspective, analogously to Bajorath *et al.'s* classification of promiscuous compounds. Even if a compound is selective against a specific target, its involvement in other pathways ('pathway crosstalk') could cause other, confounding biological readouts on top of the expected readout. By using drugs and drug targets as a proxy, I investigated whether withdrawn and approved drugs differ significantly to gain an understanding of the importance of pathway behaviour.

2. In Section 3, I show that residue movement, calculated with the simplest descriptors B-Factor and RMSD, cannot be used to improve PCM models. To the best of my knowledge, residue movement has not been incorporated into a PCM model. According to the analyses by Najmanovich *et al.*, residues show different movement according to their chemical features. However, Najmanovich did not place this movement into the context of the position of a specific residue. To close this gap, I analysed the different the movement of residues at equivalent positions across kinase binding pockets.

3. In Section 4, I successfully managed to predict the target-compound activities of known selective compounds, using a manually collected dataset of kinase inhibitors as an external test set, However, if data was not previously available for these compounds, the PCM models failed to classify the activity/inactivity of the target-compound combinations in 42.1% of the cases correctly, highlighting the incomplete coverage of chemical space as limits of PCM models.

4. In Section 5, I performed an exploratory investigation of the use of binding interactions and shapes derived from selective and promiscuous compounds to analyse whether they contain notable features suitable for building ligand-structure-based models to filter databases for selective or promiscuous compounds. I also explored whether strong and weak binders can be separated using interactions using random forest models or the CRANkS maps, a knowledge-based 3D target-specific interaction map. I show that promiscuous and selective compounds tend to have conserved binding modes across targets; however, any signal from binding interac-

tions is inconclusive to develop a selectivity filter or a classifier to separate strong and weak binders.

2 The Connection between Drugs, Drug Targets and Pathways: an Extension of The Drug-Target Paradigm

2.1 Introduction

As discussed in Section 1.2, protein targets can be hierachically organized in pathways according to their biological connection and function (Berger and Iyengar, 2011; Schneider and Klabunde, 2013). However, so far, the position of a drug target in a pathway and its connections to other biological pathways has not been explored in a systematic fashion. To close this gap, I present a novel, complementary view of the difference between proteins known to bind drugs ('drug targets') and proteins not known to bind drugs (so called 'undrugged' proteins) in terms of their pathway connections.

I show that the approval or withdrawal of a drug is independent of its number of targeted pathways. I also show that drug targets are generally associated with more pathways than 'undrugged' proteins. As this disregards the nature of pathways, I also analysed the connection between pathways, targets and diseases. My results indicate that disease involvement is independent of the actual target class of the drug target. Hence, target selection should focus on the location of the target within the pathway and the involvement of the target in other pathways. This suggests that we should move away from the paradigm of target-class focused drug development (Barnash et al., 2017).

2.2 Methods

Collection of drug targets

After downloading DrugBank Version 5.1.0 (Wishart et al., 2017), only drug targets with the species annotation 'homo sapiens' or 'human' and drugs with drug type 'small molecule' associated with a compound SMILES string were kept. Drugs labelled 'nutraceutical' and 'vet approved' were discarded and the remaining drugs relabelled: drugs both annotated with 'approved' and 'withdrawn' were relabelled 'withdrawn_approved', while drugs labelled with either 'approved' or 'withdrawn' among their annotations were relabelled 'approved' and 'withdrawn' respectively. The drug label 'investigational' was kept. These molecules were furthered filtered to exclude metals and highly reactive,

non-target specific molecules such as Kalium ("K") and Lithium ions"[Li+]". Said filtering was based on manual curation. Ultimately, within the 408 pathways selected (see below), 1560 human drug targets and 1688 'approved', 457 'investigational', 40 'withdrawn' and 89 'withdrawn_approved' small molecule drugs were collected. The last two groups were collectively treated as 'withdrawn'.

Pathway Selection

All 2190 pathways from Reactome 10 (downloaded 08/05/2018) were clustered according to the proteins associated with each pathway in a dendrogram, please see Figure A.5 in the Appendix, using the Python package Scipy version 1.1.0 with the metric 'Jaccard' and the method 'complete'. The Jaccard index or Jaccard similarity coefficient is commonly used to quantify the similarity between two clusters A and B and is defined as the size of the intersection of these sets divided by the size of their union (Fletcher and Islam, 2018):

$$J(A,B) = \frac{|A \bigcap B|}{|A \bigcup B|}. \tag{2.1}$$

Scipy version 1.1.0 uses the Jaccard distance for clustering:

$$d_J(A,B) = 1 - J(A,B). \tag{2.2}$$

The tree at height 12 (truncate_mode = level, level=45) after visual inspection to achieve a balance between the different levels of the pathway hierachy (Figure A.6, Appendix). This resulted in 11,104 unique proteins within 408 Pathways. A set of 'representative' pathways were selected in consultation with one of my collaborators at Roche, Basel, Dr. Mathias Leddin, according to current biological knowledge, research interest and importance.

Quantifying pathways by protein classes and drugs

Using DrugBank, 1560 proteins of the 11,104 proteins were identified as drug targets ('drug targets'). From the remaining proteins, 1605 were classified as 'researched' and 7939 as 'unresearched', depending on whether there was at least one bioactivity measurement for this protein in ChEMBL (Version 23), henceforth referred to as 'ChEMBL23'. To ensure accuracy, I defined the assay type as "binding" ("B") and the confidence score

Table 2.1: Relabelling of target classes in ChEMBL23 to provide a high-level nomenclature of target classes. The labels are taken directly from ChEMBL23; 'G protein-coupled receptor' is abbreviated to 'GPCR' for simplicity.

New Label	Original Label
'GPCR'	'Membrane receptor', 'Family A GPCR'
	'Family C GPCR','Family B GPCR'
	'Frizzled family GPCR' , 'Taste family GPCR'
'Enzyme'	'Enzyme with no other identifier', 'Cytochrome P450'
	'Phosphodiesterase', 'Protease',
'Other'	'Auxiliary transport protein', 'Secreted protein'
	'Surface antigen', 'Other nuclear protein'
	'Other membrane protein', 'Structural protein'
	'Other membrane protein', 'Adhesion'
	'Other cytosolic protein'
'Other GPCR'	'Toll-like and IL-1 receptors'
'Kinase'	'Kinase',
'Ligand-gated ion channel'	'Ligand-gated ion channel'
'Voltage-gated ion channel'	'Voltage-gated ion channel'
'Other Ion Channel'	'Other Ion Channel'
'Unclassified Protein'	'Unclassified Protein'
'No TargetClass available'	'No target class in ChEMBL23'

as either 8 (Homologous single protein target assigned) or 9 (direct single protein target assigned). Additionally, the first and second target class identifier for each protein was extracted from ChEMBL23 and remapped as listed in Table 2.1.

Using the association between targets and pathways, the number of pathways that each target was associated with were counted and the distributions plotted using the Python package Seaborn. Drugs were linked to pathways using the drug target-pathway association in Reactome. A Kruskal-Wallis test was performed with the Python scipy.stats module (version 1.1.0) (Jones et al.), a Dunn post-hoc test with the Python package *scipy* (Version 1.2.0) (Jones et al.). The two groups were non-normally distributed samples (See Figure A.2). Additionally, it was assumed that the number of pathways for approved, withdrawn and investigational drugs were independent of each other, as each drug interacts with a distinctive set of drug targets (and thus a distinctive number of pathways). As the rank-based nonparametric Kruskal-Wallis can be used to test whether there is statistically significant difference between two or more groups of an independent, non-normally distributed variables, it was used to test for statistical sig-

nificance between the distribution of pathways and drugs (Ostertagova et al., 2014). The Dunn post-hoc test was used as it is also a non-parametric test and can be used regardless of how the measures are distributed in each sample group (Dinno, 2015). The difference between the distributions of pathways for drug targets was assessed using a poisson regression with the R package *stats* (version 3.5.0)(Team, 2006). The correlation between the number of unique drug targets and unique pathways for 'approved', 'investigational' and 'withdrawn' drugs was statistically assessed using the Pearson correlation, Kendall's τ coefficient and the distance correlation correlation coefficient. The Pearson correlation coefficient measures the linear relationship between two variables (Schober et al., 2018). Kendall's τ coefficient or Kendall rank correlation coefficient measures the ordinal association between two vectors (Brossart et al., 2018); it is thus a non-parametric measure of rank correlation. Both coefficients are bounded between +1 and -1, with higher values denoting a positive or negative correlation between the variables of interest and 0 denoting no relationship at all. They were both calculated using the *stats* package from *scipy* (Version 1.2.0) (Jones et al.).

Lastly, the distance correlation coefficient measures independence between two vectors, without assuming a linear relationship between the variables. A distance correlation of zero implies that the analysed vectors are independent. It has been used to detect nonlinear associations which were undetectable by the Pearson correlation coefficient in a variety of fields, including astrophysics (Edelmann et al., 2019; Martínez-Gómez et al., 2013). This coefficient is bounded between +1 and -1, with higher values denoting a strong dependent relationship between the vectors of interest and 0 denoting independent vectors. The distance correlation coefficient and the corresponding p-value were calculated using an adapted script by and Wladston *et al.* (Filho, 2015) and *scipy* (Version 1.2.0) (Jones et al.). For each of the representative pathways, the drug targets were collected. Their target class annotation was extracted from ChEMBL23 and remapped as described previously.

Mapping ATC Codes to Drug Targets within Pathways

WHO Anatomical Therapeutic Chemical Classification System code, or ATC code, denote for drugs the organ or system on which the active ingredient acts, the associated therapeutic area and the drug's chemical properties (Santos et al., 2016). The ATC Codes for all drugs, regardless of label, were collected from DrugBank. Their SMILES strings were converted into Morgan Fingerprints (radius 2, nbits 2048), and their pairwise Tanimoto similarity calculated using RDKit (Landrum, 2020). Targets were mapped with a disease area, if at least two drugs with a Tanimoto similarity of 0.3 or less and the same ATC code could be associated with said target. Additionally, following Santos *et al.* (Santos et al., 2016), only drug-target combination with an experimentally validated mechanism of action (annotation: "known" in DrugBank) were selected. These stringent conditions resulted in 33 targets. Targets were mapped to their pathway(s) using the target-pathway association from Reactome. Target classes for the drug targets were extracted from ChEMBL23.

2.3 Results

Previously, drug polypharmacology has been defined on the target level (Jasial et al., 2016; Haupt et al., 2013). The present study extends this view to the pathway level by quantifying the extent of pathway polypharmacology (i.e. the number of associated pathways) of 'investigational', 'approved' and 'withdrawn' drugs; the results are shown in Figure 2.1. Figure 2.1 indicates a difference between the different distributions. As they were non-normally distributed, I used the Kruskal-Wallis test and showed a statistically significant difference between the distributions ($p = 1.7 \cdot 10^{-32}$, α=0.05); the Dunn Test as a Post-hoc Test showed a statistically significant difference between all groups with α=0.05, see also Table 2.2).

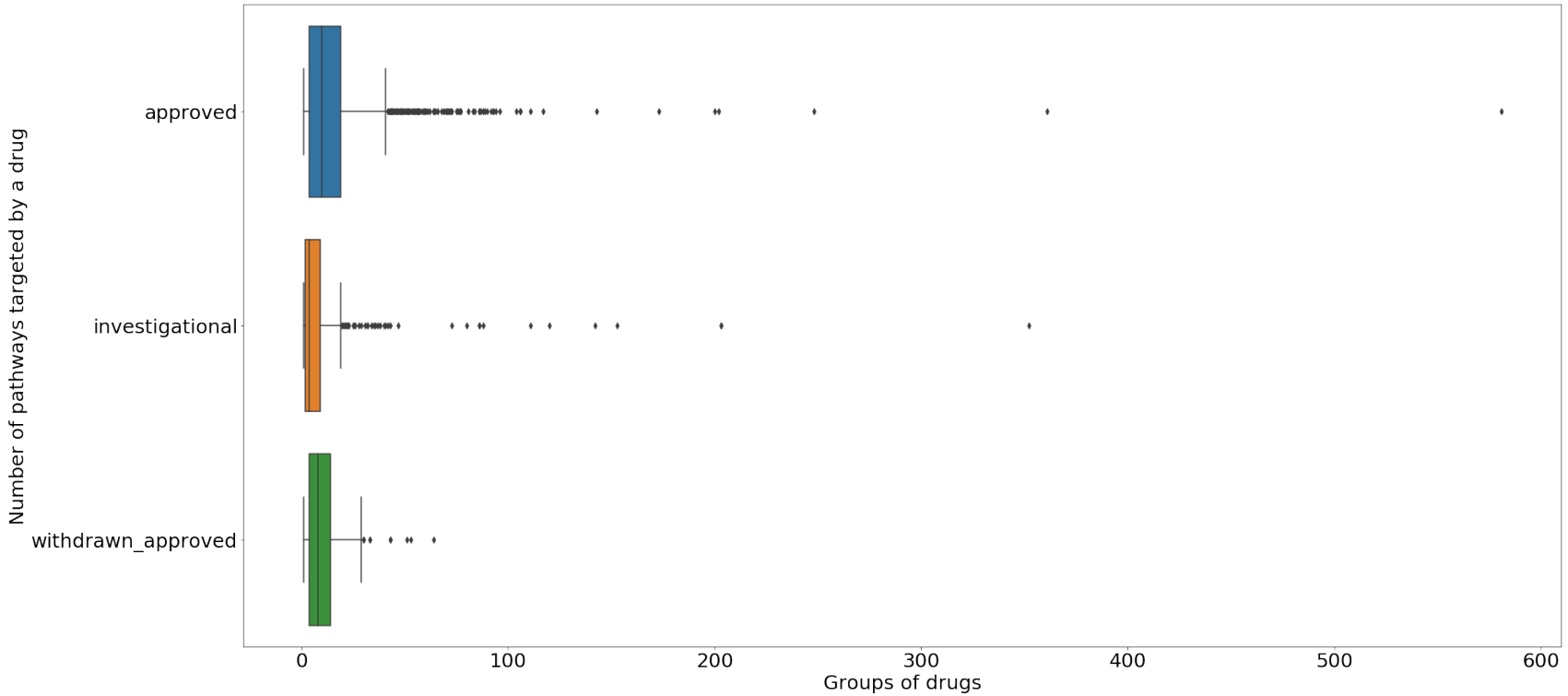


Figure 2.1: Pathway count for approved drugs, withdrawn drugs and investigational drugs. Drugs were collected with their respective targets from Drugbank and classified as specified in Section 2.2. Each drug been plotted with the number of associated pathways, derived via the drug-drug target-pathway association. Approved, withdrawn (here labelled: withdrawn_approved) and investigational drugs are coloured blue, green and orange, respectively. Approved drugs show a wider distribution in terms of targeted pathways than withdrawn and investigational drugs; according to the results of the Dunn post-hoc tests, the difference in terms of distribution is statistically significant (see Table 2.2). This highlights that approved drugs interact with a wider range of pathways and/or a bias towards the exploration of drug targets of approved drugs.

Table 2.2: Results of the Dunn post-hoc tests for testing the statistical significance of the number of pathways for drugs (Figure 2.1). All p-values were considered significant under my chosen threshold of α=0.05.

Group of drugs	Approved	Investigational	Withdrawn
Approved	1.00	$4.70 \cdot 10^{-33}$	$2.37 \cdot 10^{-2}$
Investigational	$4.77 \cdot 0^{-33}$	1.00	$2.30 \cdot 10^{-5}$
Withdrawn	$2.37 \cdot 10^{-2}$	$2.30 \cdot 10^{-5}$	1.00

Drug polypharmacology is correlated with pathway polypharmacology

Next, I analysed the number of pathways versus the number of targets for each drug to quantify the relationship between polypharmacology on the target and pathway levels. For this purpose, I separated drugs into 'approved', 'investigational' and 'withdrawn' drugs as detailed in Section 2.2. Then, as shown in Figure 2.2, The number of associated drug targets with a drug and the pathways associated with a drug were plotted for visual inspection. Additionally, the Pearson correlation, Kendall's τ coefficient, also called Kendall's rank correlation coefficient, and the distance correlation were used to assess whether an increase in unique drug targets is correlated with an increase in the number of associated, unique pathways for a given drug class (see Section 2.2 for further details). The results are listed in Table 2.3.

Both Figure 2.2 and Table 2.3 indicate a correlation between the number of unique drug targets and pathways for drugs in the groups 'approved', 'investigational' and 'withdrawn' drugs. To verify whether the correlation was statistically significant and to quantify its strength, three measures were calculated: the Pearson correlation coefficient to measure the strength of the linear relationship, Kendall's τ to measure the rank correlation and the distance correlation coefficient to measure the (in)dependence of the two variables. The chosen threshold for the p-value was α=0.05. As shown in Table 2.3, for all three groups, Kendall's τ and the Pearson correlation coefficient show a strong correlation over 0.7 with p-values below the α-threshold, indicating a linear relationship between the number of drug targets and pathways for a given group of drugs. This is further supported by the distance correlation coefficient; its high values and p-values below α=0.05 indicate a strong dependent relationship between the number of (unique) drug targets and pathways.

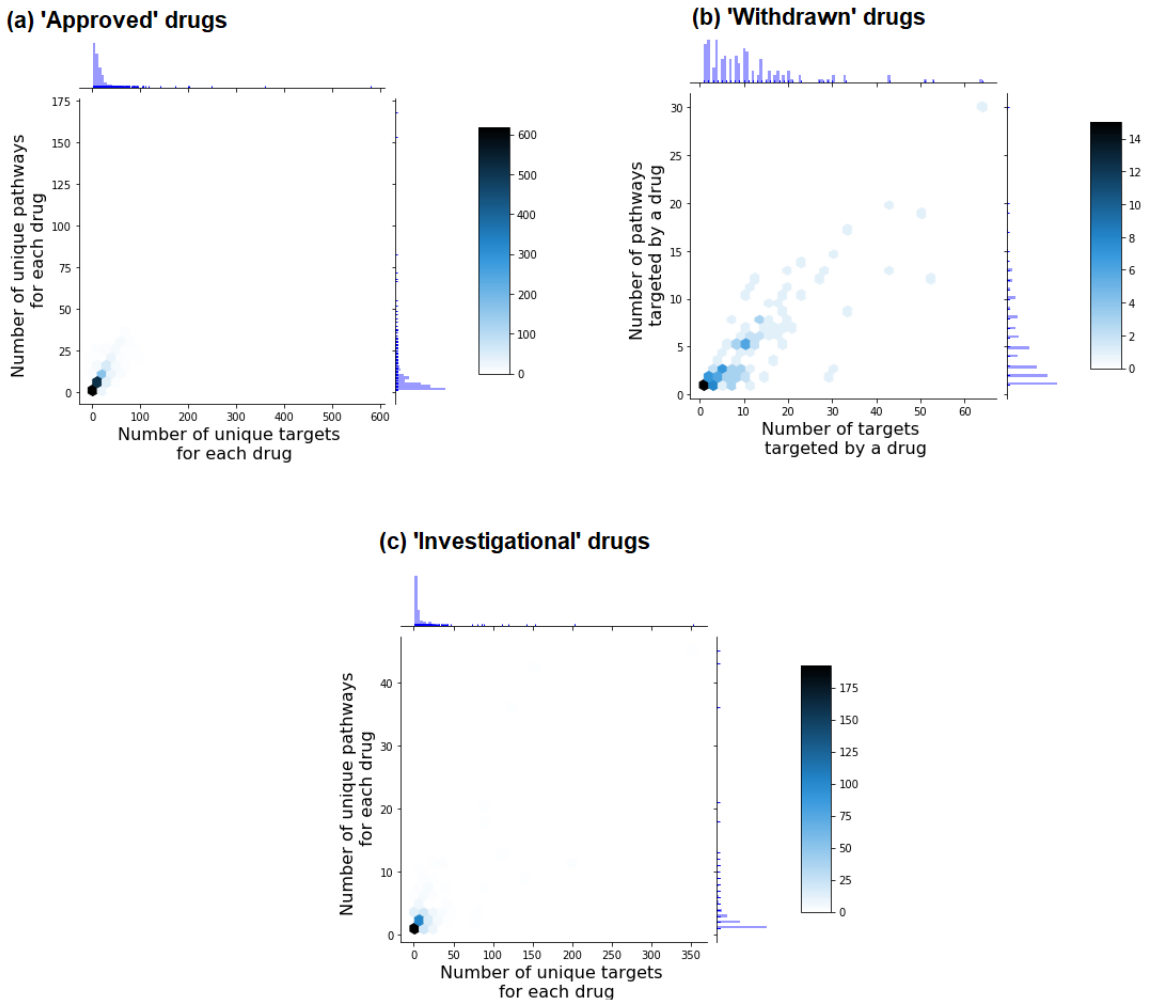


Figure 2.2: Drugs were classified as 'active', 'investigational' and 'withdrawn' drugs. For each drug, the number of its targets is plotted versus the number of its associated pathways. A hexagonal binning plot is used to denote the number of drugs. The number of unique drug target rises linearly with the number of unique pathways; this has been verified by calculating the Pearson coefficient, the Kendall τ coefficent and the Distance correlation coefficient (see Table 2.3).

Table 2.3: Results of the Pearson correlation tests, the Kendall τ and the distance correlation used for testing the statistical significance of the correlation between the number of unique targets and unique pathways for a given drug. . The p-value is indicated in brackets behind the corresponding correlation coefficient. For a graphical representation, see Figure 2.2). Each p-values is significant under my chosen threshold of α=0.05.

Metrics	Approved Drugs	Withdrawn Drugs	Investigational drugs
Pearson Correlation	0.81 (0.00)	$1.21 \cdot 10^{-37}$)	0.78 ($1.00 \cdot 10^{-89}$)
Kendall's τ	0.74 (0.00)	0.70 (0.00)	0.53 ($2.08 \cdot 10^{-44}$)
Distance correlation	0.83 (0.00)	0.82 (0.00)	0.69 (0.00)

Then, to estimate the involvement in drug targets in biological processes, I investigated whether current drug targets are involved in more pathways than proteins, which are not associated with any drug ('undrugged' proteins). I accounted for research interest according to target class by stratifying drug targets and 'undrugged' proteins according to their ChEMBL target class annotation as outlined in Section 2.2. Target class is defined as a family of structurally and/or functionally related proteins (Barnash et al., 2017) and some classes, *e.g.* the G-protein-coupled receptors, account for more human drug targets. Thus, they are of higher research interest than other target classes (Santos et al., 2016).

From 11,104 proteins in 408 pathways, I had previously selected 1560 drug targets. I separated the remaining proteins into 'undrugged' proteins and proteins with no available bioactivity data based on whether bioactivity information was available for these proteins in ChEMBL23 (Gaulton et al., 2017). Then, I classified 1605 proteins as 'undrugged', and 7939 proteins as 'unresearched'. Pathways are not fully explored in terms of bioactivity data (also see Appendix, Figure A.4). The distribution of pathway counts for each protein is displayed in Figure 2.3, indicating significant differences between drug targets and 'undrugged' proteins. I excluded proteins without bioactivity data from my analyses, as my research focused on the difference between drug targets and 'undrugged' proteins.

All distributions except 'Other GPCR', 'Voltage gated Ion Channel' and 'Other Ion Channel' showed a statistically significant difference under the Poisson Distribution Regression with a significance level of α of 0.05 (Table 2.4). Figure 2.3 and Table 2.4 indicate that drug targets are generally implicated in more pathways than 'undrugged' proteins, and thus show a greater amount of known pathway crosstalk. Pathway crosstalk

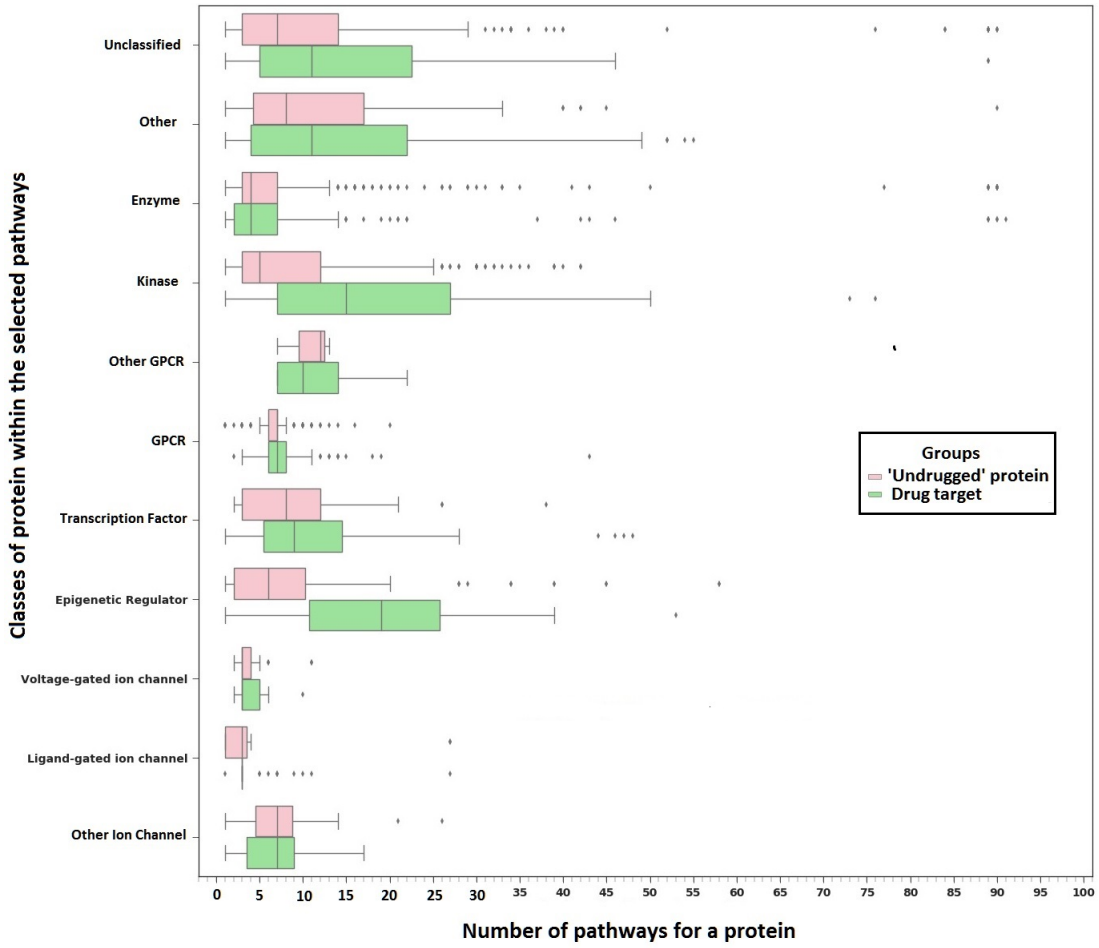


Figure 2.3: Distributions of pathway counts for each protein within the groups of 'undrugged' proteins and drug targets. Drug targets are coloured green, while 'undrugged' proteins are coloured pink. Drug targets generally show a broader pathway distribution than 'undrugged' proteins.

Table 2.4: *p*-values between the distributions of 'undrugged' proteins and drug targets in Figure 2.3 derived by using a Poisson Distribution Regression. Any statistically insignificant differences between the distributions are annotated in red, with the significance threshold α set at 0.05. There is a statistically significant difference between 'undrugged' proteins and drug targets for almost all of the pathway distributions, regardless of target class.

Target Class	*p*-value
Unclassified	$1.04 \cdot 10^{-12}$
Other	$4.56 \cdot 10^{-8}$
Enzyme	$2.0 \cdot 10^{-16}$
Kinase	$2.0 \cdot 10^{-16}$
Other GPCR	$5.91 \cdot 10^{-1}$
GPCR	$1.86 \cdot 10^{-3}$
Transcription Factor	$3.59 \cdot 10^{-5}$
Epigenetic Regulator	$2.0 \cdot 10^{-16}$
Voltage gated Ion channel	$6.5 \cdot 10^{-1}$
Ligand Ion Channel	$1.89 \cdot 10^{-4}$
Other Ion Channel	$1.95 \cdot 10^{-1}$

between my set of 'representative pathways' is exemplified in Figure A.1 and Figure A.4, in the Appendix.

Drug target classes vary widely in pathways

As stated previously, target classes are important for structure-based drug design. Some target classes, *e.g.* kinases, have been shown to be of traditional higher research interest Gaulton et al. (2017). To gain a complementary view of target classes from a biological perspective, I show the ChEMBL target classes of drug targets, if available, within 24 representative pathways (Figure 2.4), and the complementary view from the perspective of drug target classes for three selected pathways (Figure 2.5).

The target classes of drug targets within a pathway can vary widely between pathways. For example, 'MAPK1/MAPK3 signalling' has a wider range of associated target classes (also probably due to its higher number of protein targets), whereas the pathways 'CTLA4 inhibitory signaling' and 'Regulation of Hypoxia-inducible Factor (HIF) by oxygen' show a lower target class variety in terms of their drug targets.

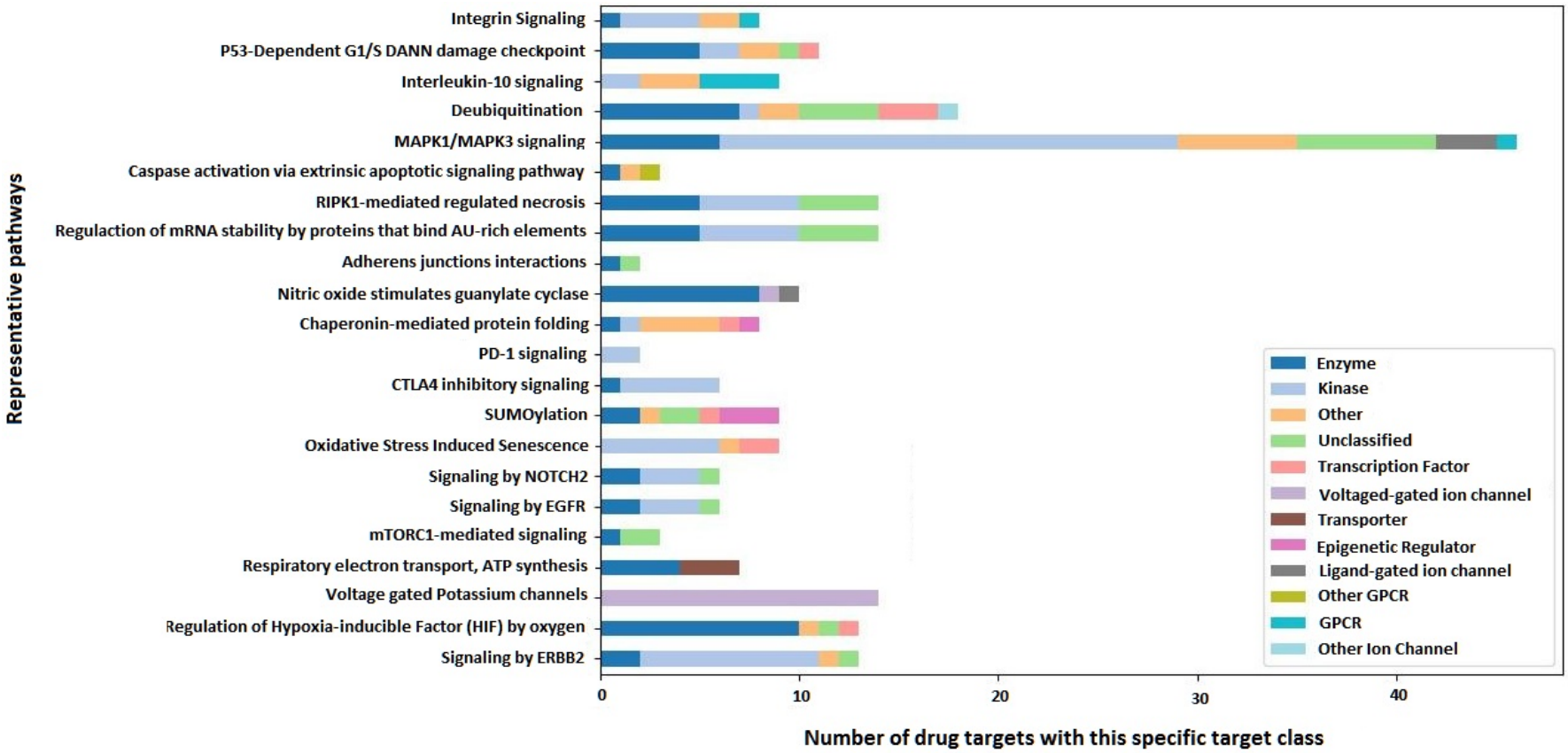


Figure 2.4: Pathways contain a wide range of target classes. Within each pathway, the target classes of drug targets were mapped (see legend). The histogram bars shows target count for the specific target classes. The drug targets can belong to multiple target classes in pathways such as MAPK1/MAPK3 Signaling.

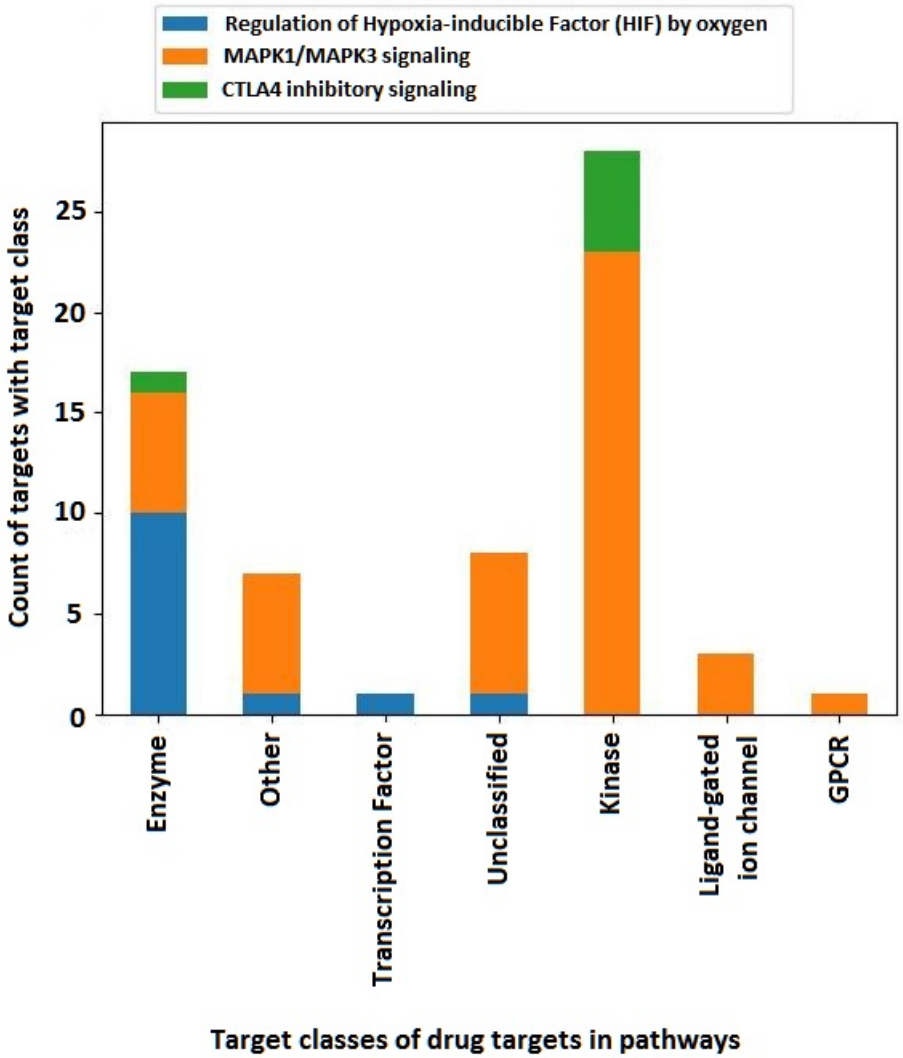


Figure 2.5: Target classes of drug targets mapped to exemplarily pathways show a broad distribution in associated pathways. By highlighting the target classes for three pathways, 'MAPK1/MAPK3 signalling' (orange), 'CTLA4 inhibitory signaling' (green) and 'Regulation of Hypoxia-inducible Factor (HIF) by oxygen' (blue), and show that drug targets in the MAPK1/MAPK3 signalling pathway have a broad distribution of associated target classes, while the 'CTLA4 signaling pathways' is only associated with 'Enzyme' and 'Kinase'. Targeting proteins according to target class could result in hitting targets in a multitude of other pathways at once. Drug targets and pathways are targeted for a single or multiple disease indication.

To extend my previous analyses, such as mechanism-indication pairs, in the biological context of pathways, I visualized both current drug targets and their associated disease areas within pathways. When available, I used one or multiple of the fourteen ATC Codes. Based on previous work by Santos *et al.* (Santos et al., 2016), I linked targets to the disease areas by connecting the ATC code from a drug to its efficacy target and then mapped the disease area from the target onto the pathway (Figure 2.6). I accounted for drug polypharmacology, where another drug target might cause the effect of a drug, and

structurally similar drugs with the following strategy: the ATC code was only mapped if at least two drugs with the same ATC code and a Tanimoto similarity equal to or less than 0.3 were available for a target. These requirements are more stringent than the previous analyses by Santos *et al* (Santos et al., 2016). resulting in 33 targets and thus a more sparse data matrix.

Figure 2.6 highlights that – within my representative pathways and despite the sparse data – targets in pathways, such as the Deubiquination signalling pathway, have been targeted for the same disease indications. Also, targets, such as the Epidermal Growth Factor, UniProt ID: P00533, have been targeted for a single indication, despite being present in multiple pathways. Finally, I show that there is a focused research interest on some drug targets, such as the Prostaglandin G/H Synthase 2, UniProt ID: P35354.

By connecting target classes and disease areas, Santos *et al.* showed that target classes are generally implicated in a wide variety of disease areas. The only exception were kinases, which seem to be only implicated in cancer (Santos et al., 2016). Using ATC codes and drug targets classes if available, I extended Santos' analysis by linking drug target classes, disease areas and my representative pathways (Figure 2.7). Due to my stringent criteria, only the target classes 'Enzyme' and 'Kinase' had enough data for a representative view within my pathways. I show that 'Kinases' are – unsurprisingly – only associated with cancer (Santos et al., 2016), while there is no discernible connection between drug targets, their pathways and the target class 'Enzyme'. Given their great biochemical diversity, that is not unsurprising.

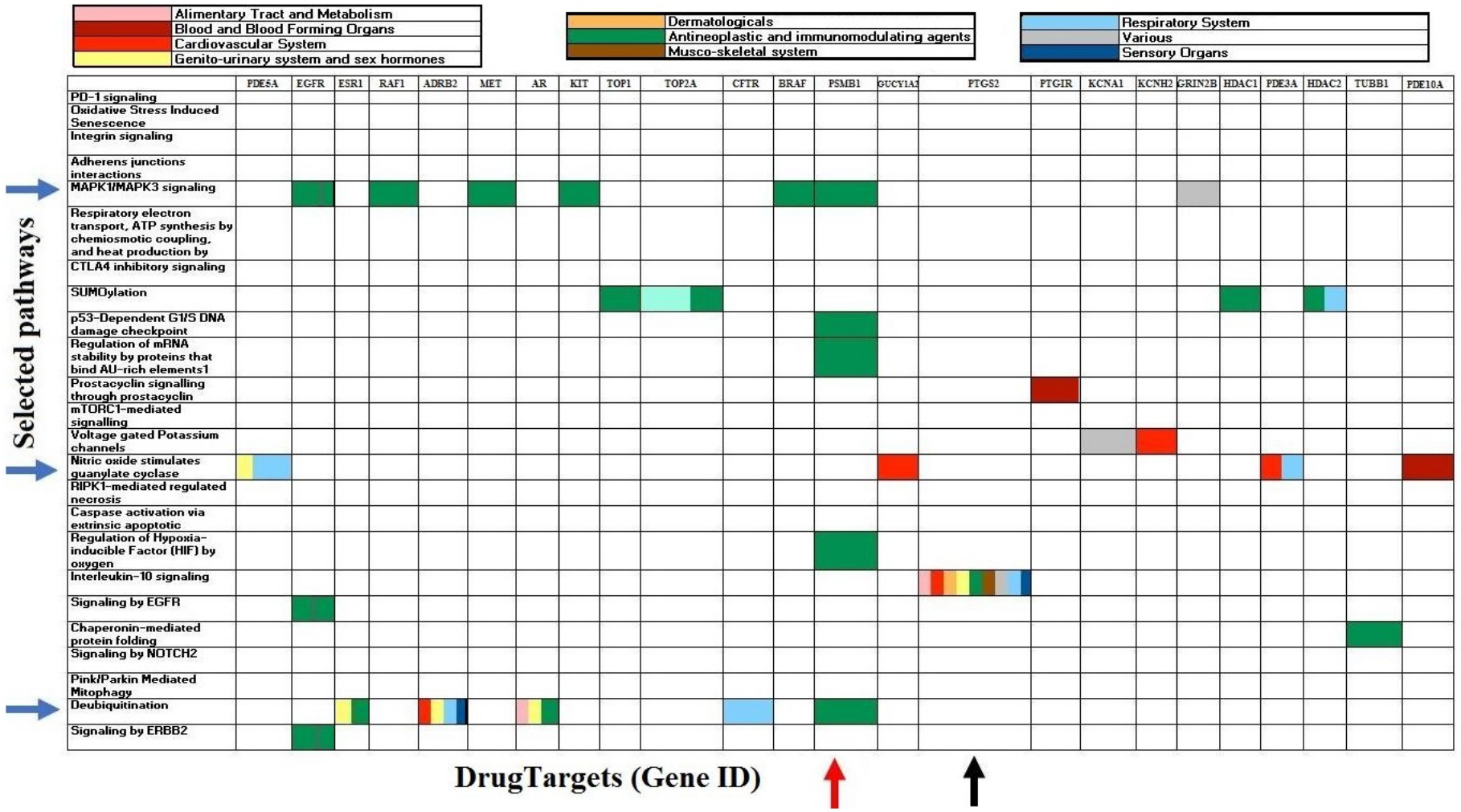


Figure 2.6: Connection between drug targets, disease areas and a set of 'representative' pathways. The colours denoted the disease areas of the ATC Codes (see legend). Within pathways, targets have been targeted small molecule drugs for a wide spectrum of diseases. Despite the sparse data matrix, Targets, *e.g.* the Epidermal Growth Factor, UniProt ID: P00533, (red arrow), are present in multiple pathways and have been primarily targeted for one disease indication. Other targets, *e.g.* Prostaglandin G/H Synthase 2, UniProt ID: P35354, (black arrow) has been targeted for multiple disease indications according to my criteria. The 'MAPK1/MAPK3 signalling pathway', the 'Deubiquination pathway', the 'Nitric oxide stimulates guanylate cyclase pathway' and the 'SUMOylation pathway' (blue arrows) all contain targets, which are targeted for the same diseases in the respective pathways. This highlights that pathways are targeted for specific diseases via specific protein targets of interest. Across pathways, target classes are connected with one disease area or multiple disease areas.

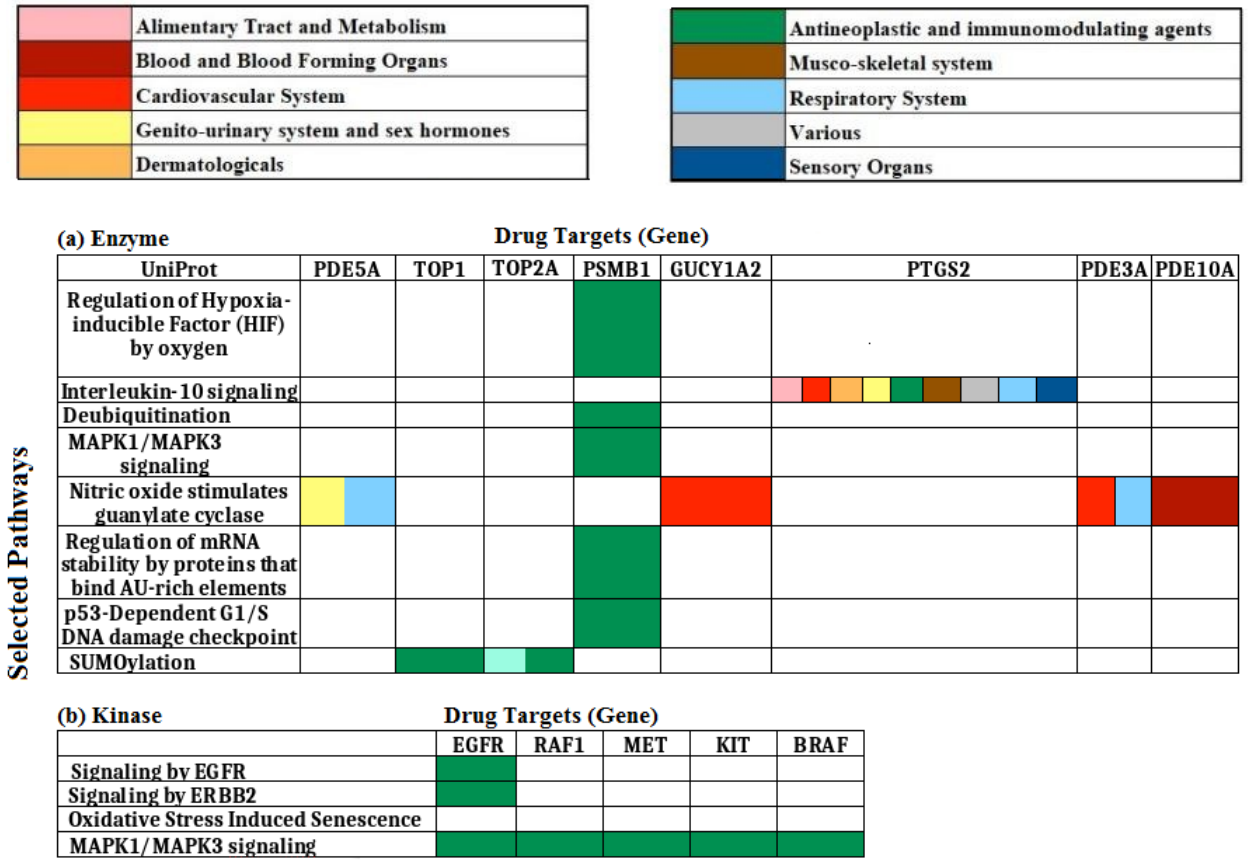


Figure 2.7: Within the targets classes 'Enzyme' and 'Kinase', disease areas can be varied according to pathways. Due to my stringent criteria and the set of representative pathways, I can only show the connection between target class, disease area, target and pathways for the target class (a) 'Enzyme' and (b) 'Kinases' within my representative 24 pathways. For (a) 'Enzyme' there is no discernible connection between target class and pathway membership for targets. For (b) 'Kinases', there was either no mapping available or the mapping was exclusively Antineoplastic and Immunomodulating agents, highlighting a bias towards cancer.

2.4 Conclusion

I extended previous analyses of drug polypharmacology by looking at the connection between the drug, its targets and the associated pathway(s). One of the key limitations of my analyses is a 'representative' selection of pathways. In Reactome, the pathways are organized in hierarchies of events, which can contain a varying number of pathways at different levels of granularity (Figure A.5, in the Appendix). As the level of pathway investigation is based on similarity clustering of protein presence/absence, entities in pathways may be repeated. This difference in pathway granularity can be considered an indication for research interest, but might also represent a bias in drug target research (Jasial et al., 2016). Additionally, my analysis is only focused on the Reactome Pathway Database; however, pathways in different databases can differ in terms of grouped entities due to research focus, manual curation bias (Belinky et al., 2015; Chowdhury and Sarkar, 2015)and variation in nomenclature due to lack of naming convention (Chowdhury and Sarkar, 2015). Oher entities, such as DNA, RNA and metabolites, and the variation of protein content between tissue types and organs in pathways are also not accounted for (Uhlen et al., 2015).

Firstly, drugs labelled 'approved' show the highest spread in terms of individual pathway count. There is a significant difference in terms of pathway count for 'approved', 'withdrawn' and 'investigational drugs'. This highlights that approved drugs might be targeting more pathways than drugs in the other groups, However, the high involvement of 'approved' drugs in a wide range of pathways might be attributable to a heightened research interest in their function, and safety/efficacy concerns during clinical trials. Additionally, the biological role of a pathway and/or the position of the drug target within the pathway might be more important for future drug success than the number of targeted pathways.

To distinguish potential drug targets from undrugged proteins and thus improve drug discovery, human drug targets have been analysed in terms of target classes (Santos et al., 2016; Rask-Andersen et al., 2011). Extending the definition for successful drug targets (Gashaw et al., 2011), I argue that three characteristics are important for a successful drug target: its involvement, if any, to other pathways, its position in the pathway, and the biological function of the associated pathways. Under the assumption

that pathways can be used as proxies for biological processes, I show that for most drug target classes, proteins that are targeted by commercial drugs tend to occur in a larger number of pathways than other proteins, which can be explained as follows: disease relevant proteins may play a role in a wide variety of biological functions, as is the case for GSK-3. These findings could also reflect the tendency of the pharmaceutical industry to develop drugs against well-studied proteins, and to further generate data around these targets and their action in the biological context. This result is the biological equivalent of the observation that bioactive compounds appear to be less promiscuous than drugs (Jasial et al., 2016). However, I used the number of pathways as approximations for biological processes, as pathways have a non-linear, graph-like nature with different reactions, *e.g.* association and dissociation. Due to this inherent complexity, I cannot account for the biological complexity inherent in the position of protein in the pathway(s) or their association or dissociation in protein complexes.

To investigate the connection between drug, drug/'researched proteins' and the underlying biology, a network-based approach might also have been used. Here, the interactions relationship between different biological entities is represented as a network, with the nodes representing proteins, genes or other biological entities and the edges their relationship, *e.g.* a physical or functional relationship based on 'omics'-data. A set of genes of interest can then be overlayed on top of biological network and see whether they cluster in one region of the network to explore underlying biological processes Chasman et al. (2016); Zhang and Itan (2019); Goh and Wong (2016). While pathways ca be defined as well-studied processes of biochemical reactions and events of regulation and signaling, derived by consensus expert curation, networks are genome- or proteome-wide interactions simplified in networks (Creixell et al., 2015). These networks represent a more mechanical and independent view on the interactions than pathways. However, in my analyses, I focused on pathways to ascertain whether drug (or drug targets or 'researched proteins') are implicated in more biological processes and used pathways as a proxy for biological processes. However, I also could have used a functional network of protein and drug targets to investigate the involvement of drug targets and 'researched proteins' in biological processes and diseases. The involvements in biological processes could have been ascertained using GO terms (The Gene Ontology Consortium, 2018), which (ideally) assign each protein a set of hierarchically organized vocabulary providing

information about their involvement in biological processes ('biological process'), molecular functions ('molecular function') and their location in the cell ('cellular component') (The Gene Ontology Consortium, 2018). However, by using the ATC codes of drugs to assign disease involvement to drug targets, I provided a complementary approach, based on the current medical knowledge (Santos et al., 2016). This focus allowed to connect targets in pathways to current drug targets for specific disease indications.

Connecting the protein target classes to pathways or ATC codes to drug targets, and then to pathways, also highlight opportunities to modulate a pathway of interest via a number of diverse protein target classes, underlining the importance of taking a pathway-centric approach to target selection in drug discovery. By mapping target classes for drug targets onto pathways, I showed that some pathways have a very varied target class composition, while other pathways, such as the 'Voltage gated potassium Channel' only have one drug target class present. As 'Enzymes' and 'Kinases' are major drug target classes, their prevalence in the pathways are not surprising (Hopkins and Groom, 2002). Lesser known target classes, such as 'Unclassified' proteins, appear as drug targets in pathways. Thus, focusing only on prominent target classes in the search for drug targets in a pathway might result in false leads.

Previously, drug targets have been connected to disease areas (Koscielny et al., 2016) and arranged in mechanism-indication pairs (Shih et al., 2017). My mapping, however, shows a more holistic view of current drug targets, their connection to pathways, and their disease areas as indicated by the ATC code. However, my data matrix is sparse, as I used very stringent criteria to account for drug polypharmacology and structurally similar drugs instead of a 'simpler' one-to-one mapping (Santos et al., 2016). These stringent criteria also bias my research against more explored drugs and drug targets and highlights that pathway space is not well explored for drug targets and disease indications. Based on this sparse data, I show that drug targets (*e.g.* Epidermal Growth Factor Receptor, UniProt ID:P00533) can be present in multiple pathways and be targeted for only one disease, such as cancer. This indicates that drug targets can achieve their effect through their involvement in multiple pathways. Additionally, one target can dominate one pathway and be targeted for multiple disease areas (*e.g.* Prostaglandin G/H Synthase 2, UniProt ID: P35354). This might again represent a research bias, as its proven track record as a drug target might lead to novel drug candidates (with different

indications) developed against it. This could be an addition to the current criteria for drug target selection (Gashaw et al., 2011).

The 'Deubiquination Pathway' contains three targets for the Antineoplastic and immunomodulating system and respiratory system, and the 'Nitric oxide stimulates guanylate cyclase Pathway' contain several targets targeted for the same disease areas, such as cancer and genito-urinary system and sex hormones. This highlights that drug targets in one pathway can be targeted for the same disease indication; novel opportunities for drug targets might thus be discovered by analysing the other targets in that pathway. The actual selection of drug targets in a pathway will probably depend on its location in the pathway and any involvement in other pathways.

Lastly, I extended the connection between target classes and disease areas (Santos et al., 2016); due to my stringent selection criteria and my sample of 24 selective pathways, my analyses are constricted to two target classes 'Kinase' and 'Enzyme'. The target class 'Kinases' is only associated with the disease cancer within my set of representative pathways, which is in agreement with the results by Santos *et al.* (Santos et al., 2016). I could find no discernible association between target class, pathway and disease area for target class 'Enzyme'. I hypothesize that the disease indication of a target depends on its membership to a pathway – and thus a biological process – and is independent of any target class membership. In conjunction with Figure 2.4, the analysis of target class distribution within pathways, I argue that target selection should focus on the location of the target within the pathway and the involvement of the target in other pathways, irrespective of the actual target class. Novel drug target opportunities might be missed by insisting on a target class specific approach. However, target class approaches should be used to construct target class-specific compound libraries and to understand novel, unexplored targets on a structural level (Barnash et al., 2017).

3 Investigation of whether the Addition of Residue-Specific Movement leads to Improvement in Random Forest Models

3.1 Introduction

While pharmacophoric and flexibility features have been extensively used to describe compounds in Proteochemometric Models (PCM) (see Section 1.3.2), so far, proteins have only been described in terms of their amino acid sequence and as a three-dimensional knowledge-based and Watermap-derived three-dimensional binding pocket (Subramanian et al., 2017) (see Section 3.5.1).

As discussed in Section 1.3.1, residues within a given binding pocket can be flexible and thus potentially contribute to the binding or non-binding of a given ligand, influencing the selectivity of the ligand. As the movement of a given residue has so far not been considered as a feature in current PCM models, I tested the inclusion of two simple flexibility descriptors for a residue at a given position - the B-Factor and the RMSD value - in PCM models. I chose the kinases due to their importance in a variety of diseases, such as diabetes (Fountas et al., 2015), cancer (Cicenas et al., 2018), cardiovascular diseases (Shahin et al., 2017) and central nervous system (CNS) disorders (Chico et al., 2009) and the high number of publicly available kinase bioactivity data, structures and selective kinase inhibitors, providing an extensive dataset for testing.

3.2 Simple Metrics For Describing The Flexibility of a Residue

3.2.1 B-Factor

In a protein crystal, the protein's atoms move, even at very low temperatures. During the experimental measurement of a protein crystal, the scattered X-ray intensity for a specific atom is attenuated if the atom is displaced from its mean position in the protein crystal. The atomic movements of the atoms, or displacement from the mean position, diminish the scattered X-ray intensity (Carugo, 2018).

These oscillation amplitudes of every atoms around its equilibrium positions in the protein crystal are quantified with the so-called B-factor, Temperature or Debye-Waller

factor. This factor can be calculated from the atomic displacement **u**, and is proportional to the mean square displacement of the atom from its equilibrium position, using Equation 3.1 (Carugo, 2018). The units are in Å^2.

$$B = 8 \cdot \frac{\pi^2}{3} \cdot \langle u^2. \rangle \tag{3.1}$$

While the B-Factor can be interpreted as the flexibility of a specific atom at a certain position; it can also contain experimental error, such as insufficient resolution and subsequent poor modelling. Subsequently, it has to be normalized using the resolution of the protein as - to the best of my knowledge - first described by Parthasarathy and Murthy (Parthasarathy and Murthy, 2008). As structures are refined without any upper limit on the atomic B-factors, some atoms might have very large B-factors (even above 100 Å^2), depending on the crystallographer; hence the fitting procedure has a large influence on the final B-Factor value (Carugo, 2018).

3.2.2 The Root-Mean-Square Deviation (RMSD)

The root-mean-square deviation (RMSD) is a similarity measure widely used in the analysis of macromolecular structures and dynamics, *e.g.* to quantify the quality of a modelled protein structure by comparing it to a native structure or to track dynamic changes in Molecular Dynamics (MD) simulations (Sargsyan et al., 2017; Kufareva and Abagyan, 2012).

The RMSD allows the quantification of the average distance between the atoms of two or more superimposed protein structures by superimposing the structures and subsequently averaging the distance between each corresponding i-th atom pair. Ideally, symmetric evidence is taken into account. The RMSD is calculated using Equation 3.2; **v** and **w** represent the Cartesian coordinates (x,y,z) for each atom of two or more structures that are being compared (Kufareva and Abagyan, 2012).

$$\begin{aligned} RMSD(\mathbf{v}, \mathbf{w}) &= \sqrt{\frac{1}{N} \sum_{i=1}^{n} \mid \mathbf{v}_i - \mathbf{w}_i \mid^2}, \\ &= \sqrt{\frac{1}{N} \sum_{i=1}^{n} ((\mathbf{v}_{ix} - \mathbf{w}_{ix})^2 + (\mathbf{v}_{iy} - \mathbf{w}_{iy})^2 + (\mathbf{v}_{iz} - \mathbf{w}_{iz})^2)}. \end{aligned} \tag{3.2}$$

The lower the RMSD, the higher the similarity between the compared structures. Typically, the protein backbone atoms are used for alignment and calculation of the RMSD; further detail about calculating the RMSD for my analyses can be found in Section 3.5.3. However, the RMSD can be distorted by the conformational differences between structures. Calculating the RMSD between two identical structures which differ in the position of a flexible terminus or loop can result in a large RMSD value as an effective superpostion is impossible (Kufareva and Abagyan, 2012). Secondly, Carugo *et al.* analysed the dependence of the RMSD on the crystallographic resolution of the superimposed protein pairs, showing that - on average — smaller RMSD values are associated with protein structure pairs at a better resolution, while protein pairs at differing resolutions tend to have increased RMSD values.

3.3 Data Sets

3.3.1 The Fulle Kinase Data Set: a Large Data Set Of Kinase Inhibitors Collected For Modelling Kinase Activity

As of 2014, the known kinases are classified either as 'conventional' and 'atypical' protein kinases. The 'conventional' kinases are separated based on their sequence similarity into the eight subfamilies AGC,CAMK, CK1, CMGC, STE, TK, TKL and Other. All members of the atypical protein kinases have little or no sequence similarity to any other member from the protein kinase family and are thus organised in a class of their own (P. J. van Linden et al., 2013). Fulle *et al.* provides a comprehensive table of 536 kinases online (Fulle, 2020); for each human kinase, Fulle *et al.* provided the four different kinase identifiers usually found in the literature (xName , Manning Name , HGNC Name and the trivial kinase target name), the group, family and subfamily the kinase belongs to, and, finally the corresponding UniProt identifier.

As outlined previously in Section 1.3.2, PCM models have been successfully used to predict the (in)activity or bioactivities of target-compound pairs. To profile the entire kinome, including cancer-related mutations, Sorgenfrei *et al.* collected an extensive kinase data set of kinases from a variety of heterogeneous data sources: a prior kinase dataset collected by Merget *et al.*, which included proprietary data from Merck (Merget et al., 2016), a kinase profiling data set by Christmann-Franck *et al.* (Christmann-Franck et al., 2016) and ChEMBL Version 22 (Sorgenfrei et al., 2018). This data is freely

available online and set consists of 305 kinases, 48,953 compounds and 1,324,732 data points (Fulle, 2019). After preparing the dataset as detailed in Section 3.5.5, it consisted of 261 proteins, 45,854 compounds and a bioactivity matrix of 268,144 datapoints; this discrepancy in number is due to the missing proprietary data by Merck *et al.* To the best of my knowledge, this data set is the most comprehensive kinase data set currently available.

However, this dataset consists of bioactivity data collected using different measurement types. For example, the data set by Merget *et al.* consists of the 'Published Kinase Inhibitor Set' (PKIS) data set by Glaxosmith-Kline, consisting of pIC_{50}-values (Merget et al., 2016) and the Tang data set, which comprises the Davis Study, the Metz Study and the Anastassiadis Study measured using an ATP-dependent competition binding assays, a kinase activity and catalytic activities respectively (Tang et al., 2014). To make these measurements comparable, Tang introduced the KIBA score (Tang et al., 2014); Merget *et al.* converted said KIBA scores back to the negative logarithm to the power of ten to derive the molar concentration. I binarized my data into 'active' and 'inactive' target-compound combinations using a cut-off value of 6.3 (corresponding to an IC_{50} value of 500 nM) , which was described as an adequate threshold to separate non-binders and hits/compounds with weaker off-target interactions by Sorgenfrei *et al.* (Merget et al., 2016). Similarly, I also binarized the data set provided by Fulle *et al.* (see Section 3.5.5).

3.3.2 The KLIFS Database: A Database Detailing Kinase Target-Ligand Interactions

The catalytic domain of kinase structure has a bilobal structure, consisting of the so-called smaller N-terminal lobe and the bigger C-terminal lobe, denoted N-lobe and C-lobe respectively. Connected by a flexible hinge region, these two lobes form a deep cleft with the evolutionarily conserved ATP-binding pocket in the middle, to which the the co-factor ATP binds (Xu et al., 2011).

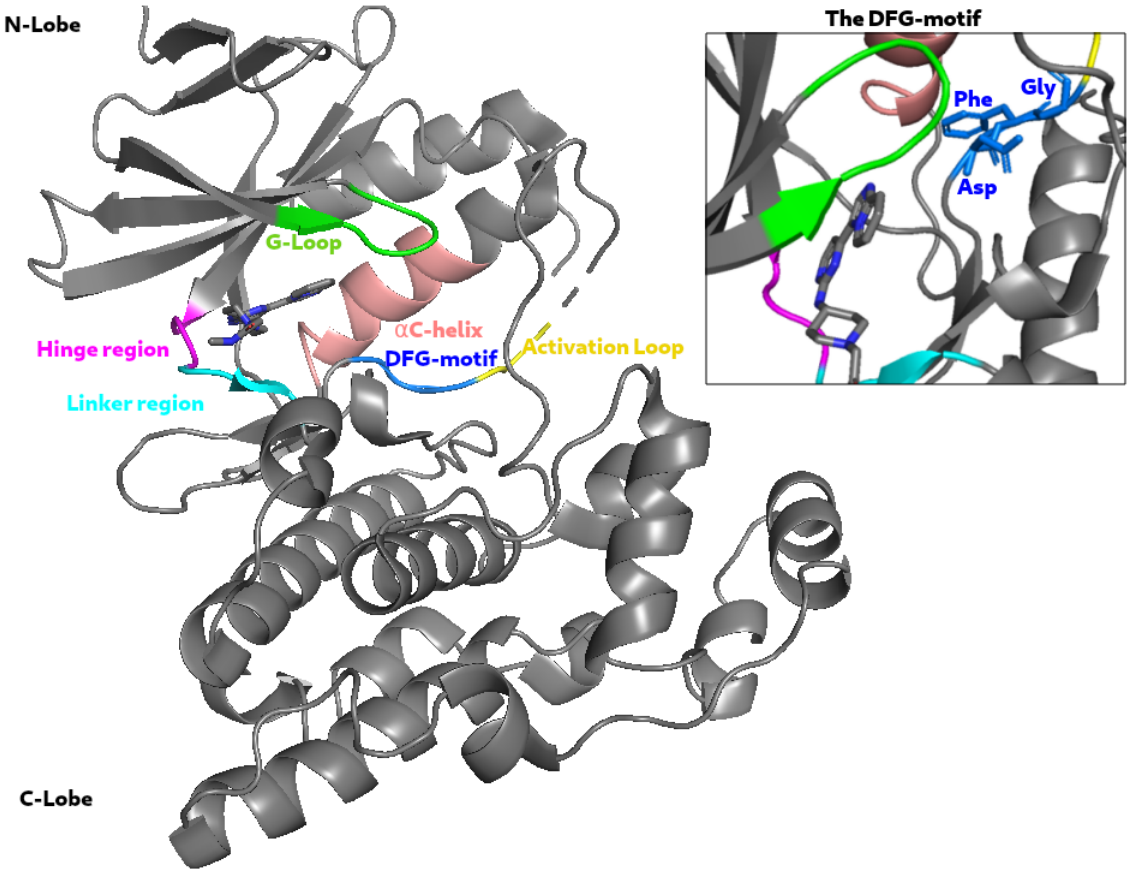


Figure 3.1: Illustration of the bilobal kinase structure. A kinase has a bilobal structure, consisting of a N-Lobe and a C-Lobe, connected by a hinge region. The three residues Asp, Phe and Gly form the DFG-motif, which determines the activation or inactivation of kinases. Both the conformations of the DFG-motif and the αC helix are used to classify kinase inhibitors. Other important regions include the flexible activation loop and the flexible glycine-rich loop (G-loop). The kinase shown is MAPK10 (JNK3), PDB code: 3CGO; the annotation follows the KLIFS annotation for 3CGO and the key regions for JNK3 (Mishra and Günther, 2018).

The catalytic residues and the activation loop are located in the cleft; as shown in Figure 3.1, the N-terminal of the flexible activation-loop contains a conserved three-residue

motif 'Asp-Phe-Gly', the so-called DFG motif. The conformational state of this motif is the determining factor for the activation/inactivation of kinases. In the 'active' or DFG-in conformation, the phenylalanine side (Phe) chain of the DFG motif occupies the ATP binding pocket and the Aspartate (Asp) is pointing to the outside of the pocket; in the 'inactive' conformation or 'DFG-out' conformation, the residues swap their position ('DFG-flip'). In this 'inactive' conformation, the Aspartate side-chain is now pointed into the ATP pocket and the Phenyl sidechain is rotated out of the ATP-binding pocket (Xu et al., 2011). This 'DFG-flip' controls the the kinase catalytic activity (Xu et al., 2011).

Kinase inhibitors tend to bind to different parts of the kinase binding pocket (Bhullar et al., 2017). To provide detailed information about the type of inhibitor, the type of kinase conformation and their 3D interaction for determining the key factors of kinase-ligand binding and selectivity, Kooistra *et al.* extracted over 2,900 kinase structures of human and mouse protein kinases from the PDB (Kooistra et al., 2015). The binding pockets were determined and annotated according to the conformation of the DFG-motif (DFG-in, DFG-out, DFG-outlike) and the αC helix (out, in, outlike); similarly, inhibitors were classified as either Type I, Type I 1/2, Type II, Type III or Type IV (Bhullar et al., 2017).

A total of 85 binding pocket residues were defined for each kinase. Using molecular interaction fingerprints of these binding site residues with each respective ligand, the different kinase-ligand interaction features were mapped onto the amino acid sequence, allowing the identification of family- or group-specific interaction features.

This information, together with bioactivity information, is publicly available at the web-based database KLIFS (*Kinase-Ligand Interaction Fingerprints and Structures*) (P. J. van Linden et al., 2013; Kooistra et al., 2015).

3.4 Random Forests

In recent years, the amount of available medical, biological and bioactivity data have continuously, sometimes even exponentially, increased, resulting in novel opportunities for explaining biological phenomena or finding pattern in bioactivity data (Jiang and Liu, 2015; Hu and Bajorath, 2017). Due to their size and/or complexitiy, new technologies

and methods for efficient storage, management and analysis of larger data sets have been developed (Tetko et al., 2016; Padgavankar and Gupta, 2014). Similarly, to integrate large data sets from different sources, analyse and extract the key features and draw conclusions for a given research question, machine learning and data visualization have increased in importance (Heureux et al., 2017).

In my analyses, I used classification models to predict compounds as either 'active' or 'inactive' as Fulle's data set contains heterogeneous bioactivity data (Section 3.5). I used the supervised Random Forest algorithm due to its speed and robustnessss to both overfitting and noise. I assessed the quality of the predictions using the metrics sensitivity, specificity, Matthew Correlation Coefficient and Receiver-operating characteristic curve. For further information please refer to Sections 3.4.1 and 3.4.2 respectively.

For further information about the different kind of machine learning (ML) models as well as an extensive overview over the currently used algorithms, I refer the interested readers to Dey *et al.*, 2016 and Kotsiantis *et al.*, 2006.

3.4.1 Random Forest Models

The basic unit of the random forest algorithm is the 'decision tree'. A small change in data can lead to a large change in the structure of a single decision tree (and hence the predicted label). The ensemble model Random Forest (RF) uses a large, user-specified number of decision trees for a robust aggregated prediction, as illustrated in Figure 3.2.

When constructing a RF model, the entire available data set is split into a training and a test set, most commonly using a 80/20% split; the RF is then constructed as detailed in Figure 3.3. Two methods are used for getting an unbiased estimate of the RF model: the internal *Out-of-Bag* Validation and the external validation method k-fold cross validation, detailed in Figure 3.3 and 3.4 respectively.

When building the trees, the subset of features available at each node is usually evaluated using the Gini impurity. It denotes the probability of incorrectly classifying a randomly chosen sample in a node, if the sample was labeled by the distribution of class labels from the data set in the node. At each node, the feature cut-of value with the highest Gini impurity will then be used as the threshold to split data at that node. The process

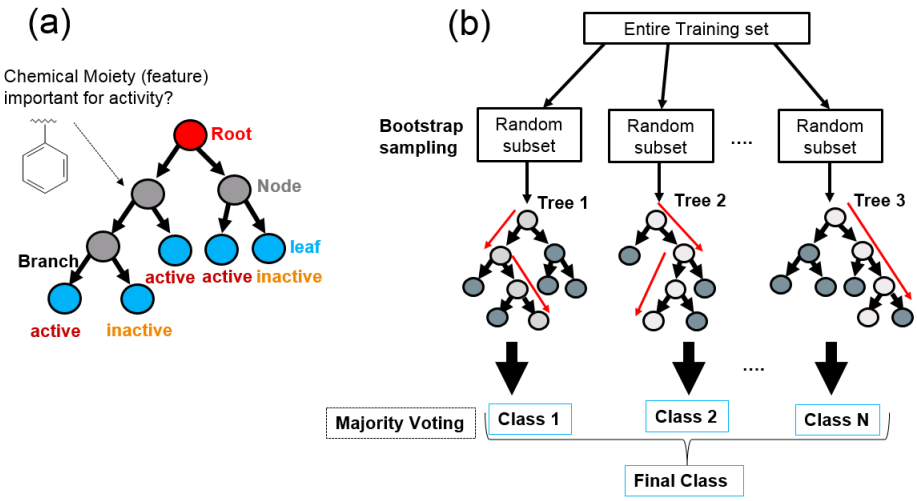


Figure 3.2: Illustration of a decision tree and how decision trees are implemented in a RF algorithm. **(a)** In a decision tree, each feature (node) is assessed in terms of its importance for the overall prediction of the label. After training, the path from the root (top level node) to the final leaf via the branches (outcome of each individual assessment) is used to make decisions for each item in the test set based on the rules at each node, with the nodes being split according to the features considered most important according to the selected importance criterion. The tree with the lowest cross-validated error will be used as the final tree (Kingsford and Salzberg, 2008). **(b)** In a RF model, a random subset of features (bootstrap sample) is selected to build each decision tree. As every tree is different, the paths from the root to the final leaves used for making decisions will be different for every tree (red arrow). The results are derived by consensus prediction in a classification model.

will repeat until all features have been used to construct this tree for this particular subset of the data. The Gini impurity I_G is is defined in Equation 3.3; J stands for the number of classes in the data set, n for the node and p_i is the fraction of items labeled with class i in the data set:

$$I_G(n) = 1 - \sum_{i=1}^{J} p_i^2. \tag{3.3}$$

The RF algorithm has a wide array of parameters, such as the number of trees for a RF model, which can be used to finetune the predictive power, the complexity, speed and quality of the RF model. These so-called hyper-parameters cannot be learnt by the algorithm and must be set before training. Different combinations of hyper-parameters should be explored to find the optimal values for training the RF algorithm on a particular kind of problem and data set (Chicco, 2017; Pedregosa et al., 2011).

Using a RF model offers two major advantages: firstly, the importance of each feature

can be quantified as a value between 0 and 1 using the Scikit-learn package; the higher the value for the feature, the greater the importance for the overall model. Secondly, overfitting is prevented as random subsets of the original features are used to construct the trees, making the RF algorithm robust against noise.

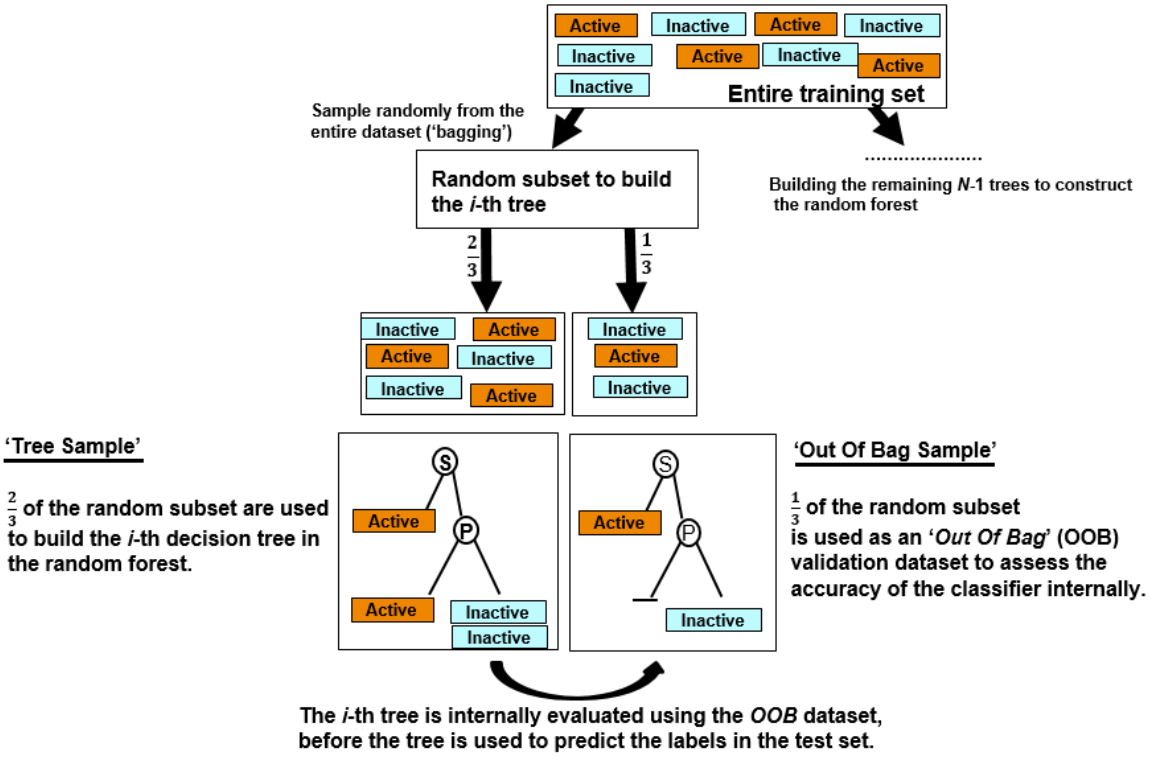


Figure 3.3: Illustration of the building and the internal validation (using *out of bag* samples) of the Random Forest (RF) Algorithm. Subsamples are randomly selected from the overall training set ('Bagging' or 'Bootstrap Aggregation') to construct each of the N trees specified by the user. Two thirds are used to construct the decision tree, whereas the third one ('*Out Of Bag* Sample' or OOB Sample) is used to evaluate how well the tree correctly predicts previously unseen labels as an internal estimate of the error rate of the classifier

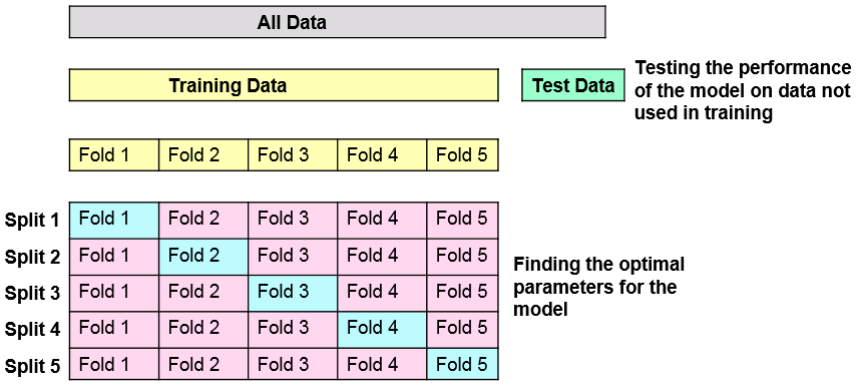


Figure 3.4: Illustration of k-fold cross validation (CV) method. The data is random shuffled and separated into a training (yellow) and a test set (green). The training set is further separated into k equal folds. The model is then trained on *k-1* folds (pink) and tested on the kth fold (blue). This is repeated k times to derive the optimal parameters for the model. The test set is then used to gauge the performance of the model on previously unseen data. Importantly, each fold should contain roughly the same amount from each class in the data set to avoid an 'imbalanced data set'.

However, in a given data set, the number of data points available for one class ('majority class') can outnumber those in another class ('minority class') drastically. In these 'imbalanced' data sets, the classifier then tends to be biased towards the majority class, as it is constructed to minimize the overall error. A bootstrap sample might contain few (or no) examples of the minority class; the resulting tree thus predicts the minority class poorly (Chen and Breiman, 2004). Three strategies exist to address this issue: downsampling of the majority class, oversampling of the minority class and adjusting the class weights for the majority and the minority 'class during the building of the RF model.

Downsampling and oversampling artificially alters the class distribution so that classes are used equally in building the tress in Random Forests. As such, they increase the discriminative power between the classes. Recent reviews suggest that down-sampling the majority class performs better than over-sampling the minority class when tree-based classifiers are used. However, Chao *et al.* also warn that downsampling the majority class might result in a loss of information as this data is not used for building the trees (Chen and Breiman, 2004). Lastly, to account for the bias of RF classifiers towards the majority class, a penalty can be placed on mmisclassifying the minority class by giving the minority class a higher weight (higher missclassification cost). These class weights are incorporated by the RF algorithm as follows: Firstly, class weights are used to weight the Gini criterion for finding splits. Secondly, the class weights are used in each terminal node. Here, the class prediction for each terminal node is decided by 'weighted majority vote' - the weighted vote for each class is the weight for this class multiplied by the number of items in this class at the terminal node. The final prediction for each class in a RF forest is then determined by aggregating the vote from each tree (Chen and Breiman, 2004). While the use of weights may disguise variation in the dataset, Chao *et al.* compared the performance of a RF model with a downsampled majority class and a RF model with a weighted minority class and reported only minor differences, depending on the chosen performance metrics (Chen and Breiman, 2004).

Lastly, the Matthews Correlation Coefficicent (MCC), Equation 3.11, can be used to judge the performance of a classifier on an imbalanced data set, as discussed in Section 3.4.2.

3.4.2 Metrics used in Classification

If the labels of a compound are *correctly* predicted as 'active' or 'inactive', the predicted label is denoted as a True Positive (TP) or a True Negative respectively; if a compound has been *incorrectly* classified as 'active' or 'inactive', the predicted label for the compound is labeled as a False Positive (FP) or a False Negative (FN) respectively. These four outcomes of a classification algorithm are usually presented in a so-called 'contingency matrix' or 'confusion matrix' as illustrated in Figure 3.5; they are used to quantify the diagnostic ability of a RF model as detailed in Equations 3.4-3.7.

The True Positive Rate (TPR), also known as 'recall', 'hit rate', 'sensitivity' or 'probability of detection', is defined as the ratio of the number of positive events correctly classified as such (TP) to the total number of all events classified as positive (P), as shown in Equation 3.4.

$$TPR = \frac{TP}{P} = \frac{TP}{TP + FN}. \tag{3.4}$$

TPR is thus a measure of what proportion of actual positives was identified correctly. Precision details the relevance of these results; it highlights how many of the items identified as positive were labelled correctly.

$$Precision = \frac{TP}{FP + TP}. \tag{3.5}$$

The True Negative Rate (TNR), 'selectivity' or 'specificity', is defined as the ratio of all events correctly classified as negative to the total events predicted to be negative (N), see Equation 3.6:

$$TNR = 1 - FPR = \frac{TN}{N} = \frac{TN}{FP + TN}. \tag{3.6}$$

The False Positive Rate (FPR) is defined as the ratio of all events incorrectly classified as positive (FP) to the total number of all negatives. FPR and TPR are connected by Equation 3.7:

$$FPR = 1 - TNR = \frac{FP}{N} = \frac{FP}{FP + TN}. \tag{3.7}$$

TPR quantifies the avoidance of False Negatives (FN); TNR quantifies the avoidance of False Positives (FP). Lastly, the accuracy is defined as the proportion of correctly retrieved results (TP, TN) among the total number of predicted items:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}. \tag{3.8}$$

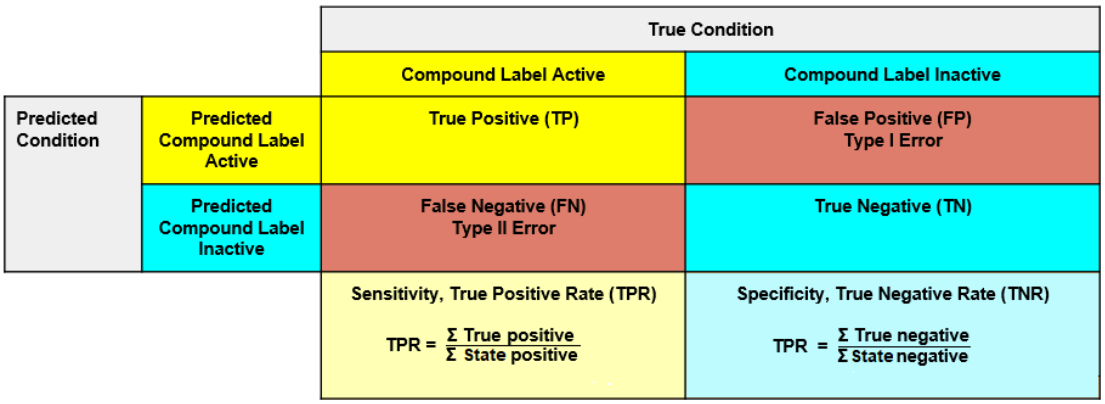


Figure 3.5: The confusion matrix illustrates the four outcomes of a simple binary classification model. A compound label can be correctly or incorrectly predicted. In terms of binary classification, TP denotes a correct prediction ('hit'), TN a correct rejection (TN), FP a false prediction and FN a false rejection respecively. The number of predictions per class is used to quantify the ability of a classifier, see Equations 3.4-3.7. This matrix is colour-coded: True positive (TP) and the True Positive Rate (TPR) are shaded in yellow; the negative equivalents are shaded in blue. Both the False Negatives (FN) and the False Positives (FP) are shaded in red. This matrix is based on Patel *et al.* (Patel, 2018).

Sensitivity and specificity are measures to assess how well the model avoids predicting false negatives and false positives respectively. The trade-off between these measures is visualized by plotting the TPR as a function of the FPR in the so-called Receiver Operating Characteristic curve (ROC) curve, discussed in Figure 3.6.

The Area Under the ROC Curve (AUC) quantifies the overall performance; it quantifies the ability of the model to distinguish between two populations, *e.g.* between active and inactive compounds in activity prediction. As illustrated in Figure 3.6, the higher the AUC, the better the model separates the different classes correctly. The AUC is

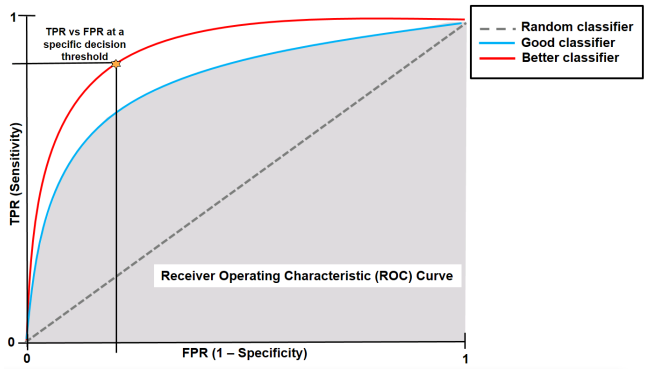


Figure 3.6: Plotting the FPR against the TPR results in a Receiver Operating Characteristic (ROC) curve, which can used to assess the diagnostic quality of a machine learning model. The ROC curve helps visualizing the trade-off between specificity and sensitivity at any chosen threshold T, as every step of the curve denotes the percentage of FPR and TPR (yellow star) for a given T. A FPR of zero would represent no falsely classified labels, whereas a FPR of 1 would show that every compound has been falsely classified. The diagonal line (grey dashes) denotes a completely random classifier. The Area Under the ROC Curve (AUC), shaded in grey, quantifies the average performance over all thresholds T. The higher the value for the AUC, the better the performance, for example, the red curve is a better performing classifier than one that produced the blue curve.

bounded between 0 and 1. A value of 0.5 would represent a random classifier, whereas any value worse than 0.5 denotes a classification worse than random and any value above 0.5 represents a classification better than random. An AUC of 1 would correspond to a perfect classifier (Powers, 2012). The AUC is a poor metric in the case of an imbalanced data set as it is an overoptimistic estimator of the ability of the classifier on the majority class. The F_1 score (balanced F-score, F-measure) and the average precision (AP) instead of a precision-recall curve have been commonly used to assess and summarize model performance for the positive class in these cases.

$$F_1 = 2 \cdot \frac{precision \cdot recall}{(precision + recall)}, \tag{3.9}$$

The AP summarizes a recall-precision plot by being the weighted mean of achieved precisions, Ps at every threshold n, weighted by the increase in recall, R_n, from the previous threshold, R_{n-1} (Scikit-learn developers, 2020a).

$$AP = \Sigma_n (R_n - R_{n-1}) \cdot P_n, \tag{3.10}$$

However, Chicco *et al.* proposed the the Matthews correlation coefficient (MCC; see Equation 3.11) as an improved balanced prediction metric for imbalanced data sets.

While the F_1 score varies if the class is swapped (positive class renamed negative and *vice versa*), the MCC remains invariant and is high only if all four categories in the confusion matrix have been predicted correctly.

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(FN + FP)(TN + FN)}}. \tag{3.11}$$

The MCC is bounded between 1 and -1; -1 represents a complete disagreement between the predicted and the observed values (worst value), 0 a classification that is no better than random and a value of +1 would represent a perfect classifier (Chicco, 2017).

3.5 Methods

3.5.1 Description of Amino Acid Sequences of Proteins

The quality of a RF model depends on the accurate description of its input data. In ML, proteins are usually described by the sequence and properties of their amino acids (AA). After aligning the sequences of all proteins, *e.g.* with CLUSTAL Omega (Mervin et al., 2018), descriptors for the AA sequences of the binding sites are calculated and used as input features. Various protein descriptors have been developed to describe the properties of the AAs. The most commonly used protein descriptors in machine learning are the Z-scales 5 by Sandberg *et al.* (Ain et al., 2014; Giblin et al., 2018; Cortés-Ciriano et al., 2015). Sandberg *et al.* applied PCA on a matrix of 26 physicochemical AA properties, such as experimental chromatography, and calculated variables, to describe AA properties, *e.g.* the volume of the sidechains and the octanol-water partition coefficient. Using PCA, the Z-scales 3 and 5 were derived by keeping three and five principal components respectively. Sandberg *et al.* interpreted the first scale (z1) as related to lipophilicity, the second scale (z2) as a summarization of bulk properties and z3 as a measure of polarity. The final Z-scales, z4 and z5, relate to properties such as electronegativity, heat of formation and electrophilicity (Sandberg et al., 1998). For an extensive review of the different AA descriptors, see van Westen et al., 2013a and van Westen et al, 2013b.

Currently, each AA would be described by the same unique descriptor values, regardless of its position in the protein. For example, a leucine would be described by the same values regardless of whether the leucine was a gate-keeping residue or positioned outside

the binding pocket. However, this does not reflect reality as the amino acids at a given position influence one another in terms of binding; changing one acid potentially changes protein-compound interactions, the formation of helices (secondary structure) or the pracking of secondary structural elements (tertiary structure), (Barley et al., 2018).

3.5.2 Defining the Binding Pocket of the Kinases

As outlined in Section 3.3.2, the KLIF database was constructed to give a unified view of the catalytic site across kinase subfamilies, using the currently available experimental X-ray structures in the PDB. It also allows classification of protein structures having the DFG-In or DFG-out motif, with a further subclassification into the following categories according to the position of the αC helix: 'in', 'out' and 'out-like'. Van Linden *et al.* also defined 85 binding pocket residues for each kinase in their database. I downloaded these 85 residues, together with other annotations, from KLIFS, on the 15.02.2019. The residue numbers for each residue in each binding pocket and the amino acid sequence for each binding pocket were then used to extract the coordinates of the binding pocket residue and the corresponding B-factors (see Sections 3.5.3 and 3.2.1). The amino acid sequence was converted into numerical values using the Z-scales 5 of Sandberg *et al.* (Sandberg et al., 1998).

3.5.3 Calculation of the Per-Residue RMSD Values

The binding pocket residues for each protein were collected as described in Section 3.5.2. The pairwise RMSD between the pairs of each binding pocket residue was calculated as follows, using Biopython Version 1.72 (Cock et al., 2009) and Python (Version 3.6.5): The DFG-in and DFG-out conformation annotations for the available kinase X-ray structures were downloaded in October 2019 and were used to separate the X-ray structures into DFG-in and DFG-out conformational datasets. All pairwise combinations of the structures were derived and the binding pocket extracted using KLIFS.

For each binding pocket residue, a window of five residues, consisting of the two residues before and after the binding residue and the binding residue itself, was constructed, if possible. Then, these five residues were superimposed for binding pocket residue

and structure based on an adapted script by Dr. West, Code 1. This script is based on the Superimposer function from the Bio.PDB Package (Hamelryck and Manderick, 2003); in this context, all atoms of the five selected residues were superimposed and rotated/translated for an overall minimization of the RMSD between the atoms. Then, the RMSD values between the whole residue, the backbone atoms and the C_α atoms were calculated and saved. After the RMSD between all binding pocket residue pairs had been calculated, the highest RMSD value was selected for a residue if multiple RMSD values had been calculated. The RMSD values of the backbone However, if different binding pocket residues were present in different experimental structures for the same target, the number of actual binding pocket residues could be higher than the usual 85 binding residues. Diverging structures were subsequently discarded for the analysis. Uncertainties in the structures were not accounted for due to time constraints; this is further discussed in Section 3.7. The number of the remaining structures is presented in Table 3.1.

Distinguishing between symmetric equivalent atoms of the same atom type for a given residue, such as the Carbon-atoms of the aromatic ring of Tyrosine, is not possible in X-ray crystallographic structures, as the final structure (and hence the final labelling) represents the 'best guess' of the experimenter. In the context of my analysis, I assumed that any error introduced by such symmetric atoms is negligible.

Table 3.1: The number of targets used for RMSD comparison for the DFG-in- and out conformation, separated by kinase subclasses. For each target, I calculated the RMSD for the same residue between different combination of protein structures. The highest RMSD per residue was selected for a given target.

Class	DFG-in Conformation	DFG-out conformation
Name	# Targets	# Targets
CAMK	12	5
TKL	11	1
STE	11	3
CMGC	17	5

3.5.4 Calculation Of The Normalized B-Factor

The structure data set was collected as detailed in the previous section; experimental protein structures for a target were collected if they have a resolution better than or equal to 2.5 Å and are derived by X-ray crystallography. The B-Factors of the atoms

constituting the peptide backbone of the protein ('C', 'CA', 'N' and 'O' respectively) were used to calculate the mean B-Factor μ_{bb} and the standard deviation σ_{bb} for each target. The B-Factors x_i for each atom of each target was normalized using Equation 3.12. Then, the B-factors of all atoms for a residue were collected, summarized and then divided by the number of atoms belonging to that residue to account for the fact that residues with longer chains were more likely to be flexible.

$$x_{normalized,i} = \frac{x_i - \mu_{bb}}{\sigma_{bb}}. \tag{3.12}$$

I extracted the final B-Factors for these binding pocket residues for each PDB file of a given target and standardized them using the KLIFS nomenclature. Finally, the B-Factor for each binding pocket residue was averaged across the different structures for the same target to get an average B-Factor value for that position. If a binding pocket residue differed from other binding pocket residues at that position, that structure was discarded from the analysis.

3.5.5 Building the Random Forest Models

The targets for each kinase subclass were extracted from the data set of Fulle *et al.*, by using all kinases (target identifier 'xname') from the 'group' identifier provided by Fulle *et al.* (Fulle, 2020). The amino acid residues for the specific target were extracted from UniProt (March 2019) and converted into the appropriate Z-scales 5 ((Sandberg et al., 1998)). Each compound in the data set by Sorgenfrei *et al.* was standardized using MolVs (Version 0.1.1) (Swain, 2019) and converted into SMILES strings. RDkit (Version 2018.03.3) was used to calculate the Morgan Fingerprints (MFP) with a radius of 2 and nbits 1024, again with RDKit (Version 2018.03.3) (Landrum, 2020). To reduce the number of features, highly correlated features were removed from the training and test set. Columns with highly co B-Factors and RMSD values with a correlation threshold of equal or over 0.7 were removed. This threshold was chosen due to the originally low number of features; also, the pairwise correlation threshold for a strong collinearity is suggested to be starting at 0.7 (Garcia Asuero et al., 2006; Akoglu, 2018; Dormann et al., 2013). Correlated columns of amino acid residues and compound features were dropped from the input data set, if the correlation between columns exceeded 0.5. (Garcia Asuero et al., 2006; Akoglu, 2018; Dormann et al., 2013). Here, due to the originally higher

number of features, a stricter, more conservative collinearity threshold of 0.5 was chosen; this threshold corresponds to moderate (positive or negative) collinearity (Garcia Asuero et al., 2006; Akoglu, 2018; Dormann et al., 2013). Additionally, correlation histograms for the B-Factors, RMSD values, the amino acid and the compound features were plotted and their distribution visually inspected. Examples for correlation histograms for the features listed above are provided in Figure A.7a, in the Appendix.

The SciKit-Learn function *StratifiedKFold*(Scikit-learn developers, 2020b) was used to split each data set into five equal folds with the same percentage of active and inactive compound-target combinations. The seed of the random number generator was set to 42 for reproducibility; 'shuffle' was set to 'True' to guarantee that the samples were randomly shuffled before being split into the 5 folds. The number of total datapoints, targets and compounds per target classes for the testing of the influence of the B-Factors and the RMSD values are shown in Table 3.2 and (3.4) respectively.The number of features for testing the influence of the B-Factor and the RMSD-values in RF models are shown in Table 3.3 and Table 3.5 respectively. For each target class, the models were trained with the eight conditions listed in Table 3.6 to get an approximation of the influence of the parameters on the model performance.

The SciKit-Learn classifier *RandomForestClassifier* was used to perform the binary classification of the models. The seed for the random number generator was set to 12345 for reproducibility; the parameter *oob_score* was set to true to get an internal validation of the model performance. To avoid training on a imbalanced data set and the resulting disadvantages (Section (3.4.1)), the parameter 'class weight' of the SciKit-Learn classifier *RandomForestClassifier* was used. The class weight, w_j, uses the the class labels to automatically adjust weights to be inversely proportional to the class frequencies in the input data, thus ensuring that the majority and minority classes are weighted equally in our model. The number of observations is denoted by n, the number of observations in class j by n_j and the total number of classes by k.

$$w_j = \frac{n}{kn_j}. \tag{3.13}$$

Table 3.2: Number of the active/inactive data points, targets and compounds used in the RF models to test the influence of the B-Factors. For each kinase group, a data set to be used for model building and training and a test set were constructed as described previously.

Kinase	# Targets	# Compounds	Data points		
		# Overall	# active	# inactive	
AGC	20	5,072	18,652	3,692	14,960
CAMK	15	2,773	14,272	1,318	12,954
STE	15	2,379	8,434	1,149	7,285
TK	26	6,546	14,272	6,394	22,203
TKL	10	2,841	4,331	1,147	3,184
GMGC	29	6,463	28,412	5,732	22,680

Table 3.3: Number of the features for each RF model used to test the influence of the B-Factors. Only kinase groups with more than five B-Factor as features were analysed further in RF models.

Kinase	Features		
Group	MFP	AAS	B-Factor
AGC	911	204	16
CAMK	911	194	7
STE	907	194	2
TK	926	266	28
TKL	893	198	2
GMGC	935	200	33

Table 3.4: Number of the active/inactive data points, targets and compounds used in the RF models to test the influence of the RMSD values. For each kinase group, a data set to be used for model building and training and a test set were constructed as described previously.

Kinase	# Targets	# Compounds	Data points		
		# Overall	# active	# inactive	
AGC	11	4,897	1,627	2,890	15,517
CAMK	10	5,884	12,212	3,768	15,980
STE	10	2,659	9,339	1,179	8,160
TK	26	13,523	44,290	12,736	31,554
TKL	8	2,861	5,008	975	4,033
GMGC	16	9,045	23,045	6,543	16,502

Table 3.5: Number of the features for each RF model used to test the influence of the RMSD values. Only kinase groups with more than five RMSD values as feature were analysed further in RF models.

Kinase	Features		
Group	MFP	AAS	RMSD value
AGC	911	204	4
CAMK	911	194	7
STE	907	194	3
TK	926	266	21
TKL	893	198	2
GMGC	935	200	11

Table 3.6: The condition of the eight different models used to explore the influence of parameters on the predictability of compounds. Class weight was 'balanced' and the depth of a tree was 20.

Model	Number of trees	Criterion for looking for the best split
1	1000	sqrt
2	1000	log2
3	500	sqrt
4	500	log2
5	250	sqrt
6	250	log2
7	100	log2
8	100	sqrt

3.6 Results

3.6.1 Exploration of the B-Factor as a Measure for Flexibility of a Residue at a Given Position

As the B-Factor was considered only a very rudimentary and superficial approximation or residue flexibility, see Section 3.2.1, it was only minimally explored. After calculating the normalized B-Factors for every kinase subfamily, I first analysed whether the normalized B-factor contained any kind of useful information before proceeding further. I thus plotted the normalized B-Factor for each KLIFS residue position for all targets, as shown in Figure 3.7 for the kinase group TKL. Interestingly, despite the fact that B-Factors "absorb" experimental error, residues in the regions 4-10, 81-83 and 84-85 show higher normalized B-Factors. The regions correspond to the flexible glycine-rich (G-loop), DFG-motif and activation loop, respectively, according to the KLIFS nomenclature for kinase binding pocket residues (P. J. van Linden et al., 2013). These results seemed tentatively promising for use in a random forest model, as they indicated that flexible residues (or regions) can be identified using the normalized B-Factor.

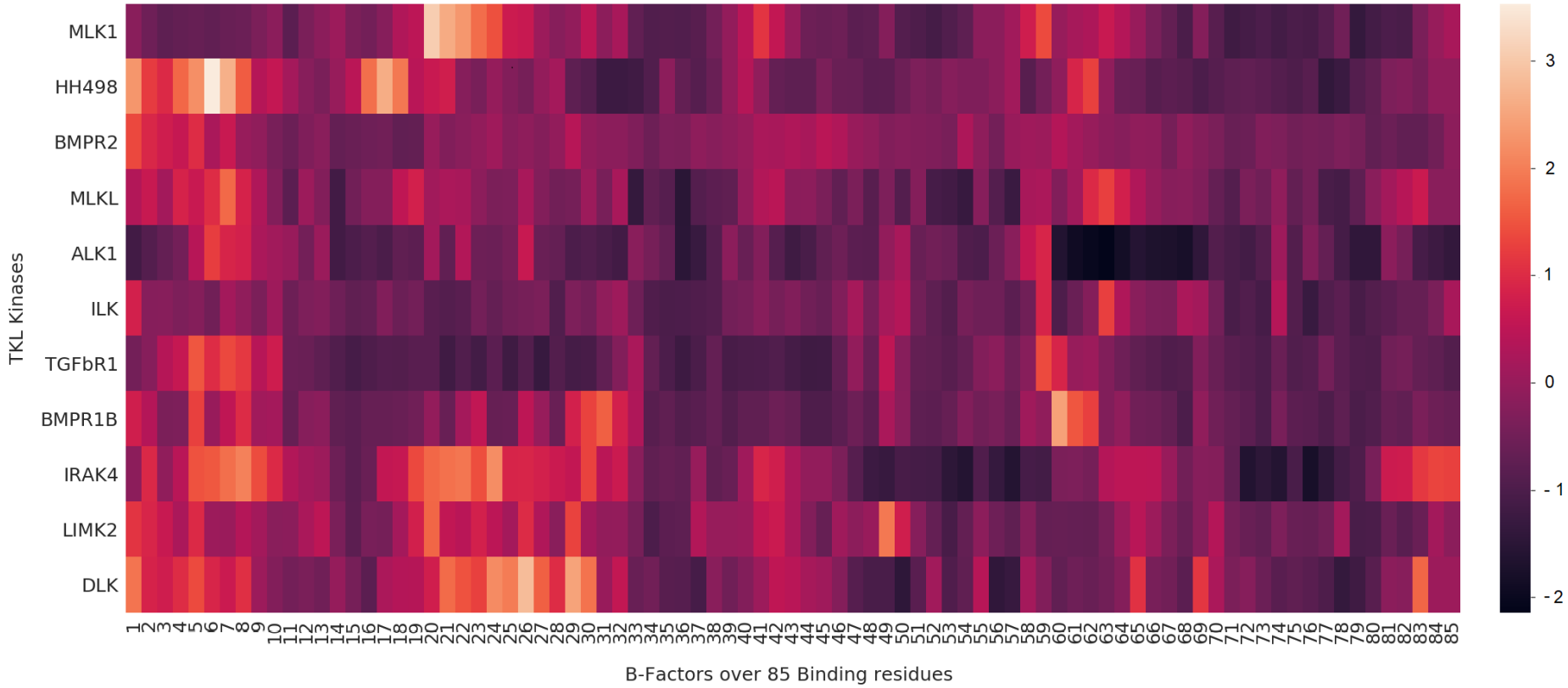


Figure 3.7: Heatmap of the normalized B-Factors for the 85 binding residues of the TKL kinases. The averaged, normalized B-Factor for each KLIFS binding pocket residue for a given target was plotted against the respective KLIFS residue. Interestingly, regions with higher normalized B-Factors correspond to flexible regions, such as the Glycine-rich loop, the DFG-motif and the Activation loop (KLIFS residues 4-9, 81-83 and 84-85, respectively).

To explore whether the normalized B-Factor can be used as an approximation for position-dependent flexibility of a specific residue in random forest models, I built four kinds of models using the features listed in Section 3.5.5 to predict the activity of the protein-ligand pairs in my data set:

- Model (1) contains only amino acid descriptors and Morgan fingerprints (MFP) to establish the baseline performance.
- Model (2) contains amino acid descriptors, MFP and normalized B-Factor values to assess whether adding B-Factors to the amino acid descriptors improved the baseline model (1).
- Model (3) tests the importance of the B-Factors as replacement for the amino acid descriptors by using only MFP and normalized B-Factors;
- Model (4) is the reference model to (3). As such, it contains MFP and a randomly generated label for each target ('target label').

When correlated normalized B-Factors were removed with a threshold of 0.7, see Section 3.5.5 for further details. Between 2-33 normalized B-Factors remained as features, depending on the specific kinase target class (see also Table 3.3 for further information). To test their influence on improving the current RF models, four models were built as described above if more than five normalized B-Factor remained as features. Hence, models were only built for the kinase classes 'AGC', 'CAMK', 'TK' and 'GMGC'. The number of active and inactive data points for the different folds, see Table 3.2, already highlights that the data sets are imbalanced towards inactive data points; this class imbalance was taken into account during training, see also Section 3.5.5. The results for all PCM models for each kinase group are listed below in Table 3.7 - Table 3.10.

The results for CAMK highlight that due to the low number of positive examples the sensitivity - the correct prediction of positive samples - is low at ~ 0.60, and performs worse than random, while the specificity - the correct prediction of negative samples - is at ~ 0.97. The high specificity corresponds to a high AUC values around 0.93; similarly, the MCC is around 0.60, indicating good prediction rate of my models. The results for the other kinase groups follow the same trend, which highlights the influence of the number of positive and negative samples.

The best performing RF models are the ones incorporating only amino acid features and Morgan Fingerprints, in terms of performance metrices (see Table 3.7 - Table 3.10). However, RF models incorporating both RMSD values and Morgan Fingerprints and models incorporating RMSD values, amino acid descriptors and Morgan Fingerprints have similar AUCs and MCCs, indicating that the B-Factors might contain information valuable for describing the kinase binding pockets of the respective targets. However, models (4), containing a target label to describe the kinase targets and Morgan Fingerprints, also show relatively high performance values, indicating that the chemical features describing the kinases might be mainly responsible for the high performance. Additionally, the AUCs and MCCS for Models (2) and (3) are often comparable to those of Model (1), which contains only Morgan Fingerprints and Amino acid descriptors.

Hence, to test this hypothesis, I also summed up the feature importances of each descriptor in each model. As exemplified for the target class TK, Figure 3.8, a random target label contributes very little to the model on its own. Interestingly, however, both the amino acid descriptors and the RMSD values contribute more than the target label and within in the same range, indicating that the RSMD values contains specific information about the binding pocket residues. However, when both amino acid descriptors and RMSD values are used, the amino acid descriptors contribute more to the overall RF model. In all four cases, the Morgan Fingerprints were the most important features(80%). This implies that our current models might be biased in terms of a well-explored chemical kinase space. This results in the Morgan Fingerprints being the most important features for a well-performing RF model. For this reason, these kinase models might not be adequate representations to investigate the influence and importance of B-Factors as descriptors to improve RF models. Future experiments should thus place an emphasis on exploring and considering chemical similarity in these RF models. Alternatively, a different choice of protein family with a less explored chemical space might have been a more appropriate choice for testing the influence of 'flexibility descriptors' in a RF model.

Table 3.7: Evaluation metrics of PCM models for the TK kinase group, trained on a data set excluding all selective compounds. For each RF model trained on the target, this table list the sensititivity, the specificity, the area under the ROC curve (AUC), the precision and the Matthews Correlation Coefficient (MCC) (Section 3.4.2) as well as the Out-of-bag performance and the Out-of-bag error (Section 1.3.2). The number of active or inactive data points in training- and test set and the number of features are detailed in Table 3.2.

Model	Parameter		Sensitivity	Specificity	AUC	Precision	MCC	OOB	OOB error
	# Trees	Best Split							
Model (1) MPF/ AAS	100	log2	0.75 +/- 0.010	0.92 +/- 0.004	0.93 +/- 0.003	0.74 +/- 0.009	0.67 +/- 0.005	0.88 +/- 0.002	0.12 +/- 0.002
	100	sqrt	0.75 +/- 0.007	0.89 +/- 0.007	0.91 +/- 0.004	0.67 +/- 0.015	0.61 +/- 0.012	0.86 +/- 0.001	0.14 +/- 0.001
	250	sqrt	0.75 +/- 0.008	0.92 +/- 0.004	0.93 +/- 0.003	0.74 +/- 0.008	0.67 +/- 0.006	0.88 +/- 0.002	0.12 +/- 0.002
	250	log2	0.75 +/- 0.010	0.89 +/- 0.006	0.91 +/- 0.003	0.67 +/- 0.013	0.62 +/- 0.013	0.86 +/- 0.002	0.14 +/- 0.002
	500	log2	0.75 +/- 0.010	0.90 +/- 0.006	0.91 +/- 0.003	0.67 +/- 0.013	0.62 +/- 0.013	0.86 +/- 0.002	0.14 +/- 0.002
	500	sqrt	0.75 +/- 0.008	0.93 +/- 0.003	0.93 +/- 0.003	0.74 +/- 0.007	0.67 +/- 0.005	0.88 +/- 0.001	0.12 +/- 0.001
	1000	log2	0.75 +/- 0.007	0.90 +/- 0.006	0.91 +/- 0.003	0.68 +/- 0.013	0.62 +/- 0.011	0.86 +/- 0.001	0.14 +/- 0.001
	1000	sqrt	0.75 +/- 0.007	0.93 +/- 0.003	0.93 +/- 0.003	0.74 +/- 0.008	0.67 +/- 0.007	0.88 +/- 0.002	0.12 +/- 0.002
Model (2) MPF/ B-Factor/ AAS	100	log2	0.73 +/- 0.007	0.93 +/- 0.005	0.93 +/- 0.003	0.75 +/- 0.013	0.67 +/- 0.011	0.88 +/- 0.001	0.12 +/- 0.001
	100	sqrt	0.73 +/- 0.010	0.92 +/- 0.005	0.92 +/- 0.002	0.73 +/- 0.009	0.65 +/- 0.004	0.87 +/- 0.001	0.13 +/- 0.001
	250	sqrt	0.73 +/- 0.010	0.93 +/- 0.004	0.93 +/- 0.002	0.75 +/- 0.012	0.67 +/- 0.010	0.89 +/- 0.001	0.11 +/- 0.001
	250	log2	0.73 +/- 0.005	0.92 +/- 0.005	0.92 +/- 0.003	0.73 +/- 0.011	0.65 +/- 0.007	0.88 +/- 0.001	0.12 +/- 0.001
	500	log2	0.73 +/- 0.006	0.92 +/- 0.005	0.92 +/- 0.003	0.74 +/- 0.012	0.66 +/- 0.009	0.88 +/- 0.002	0.12 +/- 0.002
	500	sqrt	0.73 +/- 0.010	0.93 +/- 0.004	0.93 +/- 0.003	0.75 +/- 0.010	0.67 +/- 0.009	0.89 +/- 0.001	0.11 +/- 0.001
	1000	log2	0.73 +/- 0.007	0.93 +/- 0.005	0.92 +/- 0.002	0.74 +/- 0.013	0.65 +/- 0.009	0.88 +/- 0.002	0.12 +/- 0.002
	1000	sqrt	0.73 +/- 0.009	0.93 +/- 0.004	0.93 +/- 0.002	0.76 +/- 0.010	0.67 +/- 0.010	0.89 +/- 0.001	0.11 +/- 0.001
Model (3) MFP/ B-Factor	100	log2	0.69 +/- 0.012	0.91 +/- 0.002	0.89 +/- 0.006	0.68 +/- 0.004	0.59 +/- 0.008	0.85 +/- 0.002	0.15 +/- 0.002
	100	sqrt	0.68 +/- 0.011	0.91 +/- 0.001	0.88 +/- 0.006	0.67 +/- 0.005	0.58 +/- 0.010	0.85 +/- 0.001	0.15 +/- 0.001
	250	sqrt	0.69 +/- 0.010	0.91 +/- 0.002	0.89 +/- 0.006	0.68 +/- 0.006	0.60 +/- 0.009	0.86 +/- 0.002	0.14 +/- 0.002
	250	log2	0.68 +/- 0.011	0.91 +/- 0.001	0.88 +/- 0.005	0.68 +/- 0.004	0.59 +/- 0.008	0.85 +/- 0.002	0.15 +/- 0.002
	500	log2	0.68 +/- 0.010	0.91 +/- 0.001	0.88 +/- 0.005	0.68 +/- 0.005	0.58 +/- 0.009	0.85 +/- 0.002	0.15 +/- 0.002
	500	sqrt	0.69 +/- 0.011	0.91 +/- 0.001	0.89 +/- 0.006	0.68 +/- 0.003	0.59 +/- 0.007	0.86 +/- 0.002	0.14 +/- 0.002
	1000	log2	0.68 +/- 0.010	0.91 +/- 0.001	0.88 +/- 0.005	0.68 +/- 0.005	0.59 +/- 0.009	0.85 +/- 0.002	0.15 +/- 0.002
	1000	sqrt	0.69 +/- 0.011	0.91 +/- 0.002	0.89 +/- 0.006	0.69 +/- 0.005	0.60 +/- 0.008	0.86 +/- 0.002	0.14 +/- 0.002
Model (4) Random Targetlabel	100	log2	0.69 +/- 0.012	0.91 +/- 0.002	0.89 +/- 0.006	0.68 +/- 0.004	0.59 +/- 0.008	0.85 +/- 0.002	0.15 +/- 0.002
	100	sqrt	0.68 +/- 0.011	0.91 +/- 0.001	0.88 +/- 0.006	0.67 +/- 0.005	0.58 +/- 0.010	0.85 +/- 0.001	0.15 +/- 0.001
	250	sqrt	0.69 +/- 0.010	0.91 +/- 0.002	0.89 +/- 0.006	0.68 +/- 0.006	0.60 +/- 0.009	0.86 +/- 0.002	0.14 +/- 0.002
	250	log2	0.68 +/- 0.011	0.91 +/- 0.001	0.88 +/- 0.005	0.68 +/- 0.004	0.59 +/- 0.008	0.85 +/- 0.002	0.15 +/- 0.002
	500	log2	0.68 +/- 0.010	0.91 +/- 0.001	0.88 +/- 0.005	0.68 +/- 0.005	0.58 +/- 0.009	0.85 +/- 0.002	0.15 +/- 0.002
	500	sqrt	0.69 +/- 0.011	0.91 +/- 0.001	0.89 +/- 0.006	0.68 +/- 0.003	0.59 +/- 0.007	0.86 +/- 0.002	0.14 +/- 0.002
	1000	log2	0.68 +/- 0.010	0.91 +/- 0.001	0.88 +/- 0.005	0.68 +/- 0.005	0.59 +/- 0.009	0.85 +/- 0.002	0.15 +/- 0.002
	1000	sqrt	0.69 +/- 0.011	0.91 +/- 0.002	0.89 +/- 0.006	0.69 +/- 0.005	0.60 +/- 0.008	0.86 +/- 0.002	0.14 +/- 0.002

Table 3.8: Evaluation metrics of PCM models for the CAMK kinase group, trained on a data set excluding all selective compounds. For each RF model trained on the target, this table list the sensititivity, the specificity, the area under the ROC curve (AUC), the precision and the Matthews Correlation Coefficient (MCC) (Section 3.4.2) as well as the Out-of-bag performance and the Out-of-bag error (Section 1.3.2). The number of active or inactive data points in training- and test set and the number of features are detailed in Table 3.2.

Model	Parameter		Sensitivity	Specificity	AUC	Precision	MCC	OOB	OOB error
	# Trees	Best Split							
Model (1) MPF/ AAS	100	log2	0.61 +/- 0.033	0.98 +/- 0.003	0.94 +/- 0.009	0.74 +/- 0.031	0.64 +/- 0.033	0.94 +/- 0.002	0.06 +/- 0.002
	100	sqrt	0.60 +/- 0.020	0.97 +/- 0.003	0.92 +/- 0.011	0.69 +/- 0.026	0.61 +/- 0.024	0.94 +/- 0.002	0.06 +/- 0.002
	250	sqrt	0.61 +/- 0.032	0.98 +/- 0.002	0.94 +/- 0.008	0.74 +/- 0.022	0.64 +/- 0.028	0.94 +/- 0.001	0.06 +/- 0.001
	250	log2	0.61 +/- 0.029	0.97 +/- 0.003	0.93 +/- 0.011	0.70 +/- 0.030	0.62 +/- 0.027	0.94 +/- 0.001	0.06 +/- 0.001
	500	log2	0.60 +/- 0.024	0.97 +/- 0.003	0.93 +/- 0.011	0.70 +/- 0.029	0.62 +/- 0.025	0.94 +/- 0.002	0.06 +/- 0.002
	500	sqrt	0.61 +/- 0.030	0.98 +/- 0.002	0.94 +/- 0.008	0.74 +/- 0.024	0.65 +/- 0.027	0.94 +/- 0.001	0.06 +/- 0.001
	1000	log2	0.60 +/- 0.020	0.97 +/- 0.003	0.93 +/- 0.010	0.70 +/- 0.026	0.62 +/- 0.022	0.94 +/- 0.001	0.06 +/- 0.001
	1000	sqrt	0.61 +/- 0.027	0.98 +/- 0.002	0.94 +/- 0.008	0.74 +/- 0.024	0.64 +/- 0.026	0.94 +/- 0.001	0.06 +/- 0.001
Model (2) MPF/ B-Factor/ AAS	100	log2	0.58 +/- 0.031	0.98 +/- 0.002	0.93 +/- 0.009	0.72 +/- 0.014	0.61 +/- 0.016	0.94 +/- 0.001	0.06 +/- 0.001
	100	sqrt	0.57 +/- 0.027	0.97 +/- 0.003	0.90 +/- 0.009	0.65 +/- 0.024	0.57 +/- 0.023	0.93 +/- 0.001	0.07 +/- 0.001
	250	sqrt	0.58 +/- 0.028	0.98 +/- 0.004	0.93 +/- 0.008	0.73 +/- 0.026	0.62 +/- 0.016	0.94 +/- 0.001	0.06 +/- 0.001
	250	log2	0.56 +/- 0.033	0.97 +/- 0.002	0.90 +/- 0.008	0.65 +/- 0.017	0.57 +/- 0.026	0.93 +/- 0.001	0.07 +/- 0.001
	500	log2	0.56 +/- 0.036	0.97 +/- 0.002	0.91 +/- 0.008	0.65 +/- 0.014	0.57 +/- 0.024	0.93 +/- 0.001	0.07 +/- 0.001
	500	sqrt	0.58 +/- 0.028	0.98 +/- 0.003	0.93 +/- 0.008	0.73 +/- 0.020	0.62 +/- 0.014	0.94 +/- 0.000	0.06 +/- 0.000
	1000	log2	0.56 +/- 0.029	0.97 +/- 0.002	0.91 +/- 0.008	0.66 +/- 0.016	0.57 +/- 0.022	0.93 +/- 0.001	0.07 +/- 0.001
	1000	sqrt	0.58 +/- 0.027	0.98 +/- 0.003	0.93 +/- 0.007	0.74 +/- 0.023	0.62 +/- 0.015	0.94 +/- 0.001	0.06 +/- 0.001
Model (3) MFP/ B-Factor	100	log2	0.59 +/- 0.021	0.97 +/- 0.001	0.92 +/- 0.009	0.64 +/- 0.010	0.58 +/- 0.015	0.93 +/- 0.002	0.07 +/- 0.002
	100	sqrt	0.58 +/- 0.013	0.98 +/- 0.003	0.93 +/- 0.008	0.72 +/- 0.020	0.61 +/- 0.009	0.94 +/- 0.003	0.06 +/- 0.003
	250	sqrt	0.58 +/- 0.014	0.98 +/- 0.003	0.93 +/- 0.008	0.71 +/- 0.022	0.61 +/- 0.014	0.94 +/- 0.001	0.06 +/- 0.001
	250	log2	0.59 +/- 0.020	0.97 +/- 0.003	0.92 +/- 0.008	0.65 +/- 0.021	0.58 +/- 0.016	0.93 +/- 0.002	0.07 +/- 0.002
	500	log2	0.59 +/- 0.023	0.97 +/- 0.003	0.92 +/- 0.008	0.66 +/- 0.022	0.58 +/- 0.020	0.93 +/- 0.002	0.07 +/- 0.002
	500	sqrt	0.59 +/- 0.009	0.98 +/- 0.002	0.93 +/- 0.008	0.72 +/- 0.019	0.62 +/- 0.009	0.94 +/- 0.001	0.06 +/- 0.001
	1000	log2	0.59 +/- 0.020	0.97 +/- 0.003	0.92 +/- 0.008	0.66 +/- 0.022	0.58 +/- 0.019	0.93 +/- 0.001	0.07 +/- 0.001
	1000	sqrt	0.59 +/- 0.013	0.98 +/- 0.003	0.93 +/- 0.007	0.72 +/- 0.023	0.62 +/- 0.014	0.94 +/- 0.001	0.06 +/- 0.001
Model (4) Random Targetlabel	100	log2	0.58 +/- 0.036	0.96 +/- 0.006	0.91 +/- 0.015	0.60 +/- 0.045	0.55 +/- 0.042	0.92 +/- 0.003	0.08 +/- 0.003
	100	sqrt	0.55 +/- 0.034	0.96 +/- 0.006	0.90 +/- 0.016	0.59 +/- 0.045	0.53 +/- 0.041	0.92 +/- 0.003	0.08 +/- 0.003
	250	sqrt	0.58 +/- 0.033	0.96 +/- 0.004	0.91 +/- 0.015	0.61 +/- 0.038	0.55 +/- 0.038	0.92 +/- 0.003	0.08 +/- 0.003
	250	log2	0.56 +/- 0.028	0.96 +/- 0.005	0.90 +/- 0.017	0.60 +/- 0.040	0.53 +/- 0.035	0.92 +/- 0.003	0.08 +/- 0.003
	500	log2	0.55 +/- 0.037	0.96 +/- 0.005	0.91 +/- 0.016	0.59 +/- 0.043	0.53 +/- 0.042	0.92 +/- 0.003	0.08 +/- 0.003
	500	sqrt	0.55 +/- 0.012	0.95 +/- 0.003	0.90 +/- 0.008	0.57 +/- 0.020	0.51 +/- 0.017	0.92 +/- 0.001	0.08 +/- 0.001
	1000	log2	0.59 +/- 0.020	0.97 +/- 0.003	0.92 +/- 0.008	0.66 +/- 0.022	0.58 +/- 0.019	0.93 +/- 0.001	0.07 +/- 0.001
	1000	sqrt	0.57 +/- 0.020	0.96 +/- 0.005	0.91 +/- 0.013	0.60 +/- 0.038	0.54 +/- 0.031	0.93 +/- 0.003	0.07 +/- 0.003

Table 3.9: Evaluation metrics of PCM models for the AGC kinase group, trained on a data set excluding all selective compounds. For each RF model trained on the target, this table list the sensititivity, the specificity, the area under the ROC curve (AUC), the precision and the Matthews Correlation Coefficient (MCC) (Section 3.4.2) as well as the Out-of-bag performance and the Out-of-bag error (Section 1.3.2). The number of active or inactive data points in training- and test set and the number of features are detailed in Table 3.2.

Model	Parameter		Sensitivity	Specificity	AUC	Precision	MCC	OOB	OOB error
	# Trees	Best Split							
Model (1) MPF/ AAS	100	log2	0.80 +/- 0.026	0.95 +/- 0.004	0.96 +/- 0.005	0.81 +/- 0.014	0.76 +/- 0.021	0.92 +/- 0.001	0.08 +/- 0.001
	100	sqrt	0.81 +/- 0.022	0.95 +/- 0.004	0.96 +/- 0.005	0.81 +/- 0.014	0.76 +/- 0.018	0.92 +/- 0.001	0.08 +/- 0.001
	250	log2	0.80 +/- 0.024	0.95 +/- 0.005	0.95 +/- 0.005	0.79 +/- 0.014	0.74 +/- 0.011	0.92 +/- 0.002	0.08 +/- 0.002
	250	sqrt	0.81 +/- 0.025	0.95 +/- 0.004	0.96 +/- 0.005	0.81 +/- 0.013	0.76 +/- 0.019	0.92 +/- 0.002	0.08 +/- 0.002
	500	log2	0.80 +/- 0.027	0.95 +/- 0.006	0.95 +/- 0.004	0.79 +/- 0.015	0.75 +/- 0.014	0.92 +/- 0.001	0.08 +/- 0.001
	500	sqrt	0.81 +/- 0.025	0.95 +/- 0.004	0.96 +/- 0.005	0.81 +/- 0.013	0.76 +/- 0.019	0.92 +/- 0.002	0.08 +/- 0.002
	1000	log2	0.80 +/- 0.026	0.95 +/- 0.006	0.95 +/- 0.005	0.79 +/- 0.015	0.74 +/- 0.013	0.92 +/- 0.001	0.08 +/- 0.001
	1000	sqrt	0.81 +/- 0.022	0.95 +/- 0.004	0.96 +/- 0.005	0.81 +/- 0.014	0.76 +/- 0.018	0.92 +/- 0.001	0.08 +/- 0.001
Model (2) MPF/ B-Factor/ AAS	100	log2	0.78 +/- 0.013	0.95 +/- 0.002	0.95 +/- 0.004	0.81 +/- 0.007	0.75 +/- 0.009	0.92 +/- 0.001	0.08 +/- 0.001
	100	sqrt	0.78 +/- 0.014	0.95 +/- 0.002	0.95 +/- 0.005	0.79 +/- 0.005	0.73 +/- 0.009	0.91 +/- 0.002	0.09 +/- 0.002
	250	sqrt	0.78 +/- 0.013	0.96 +/- 0.003	0.95 +/- 0.004	0.82 +/- 0.011	0.75 +/- 0.012	0.92 +/- 0.001	0.08 +/- 0.001
	250	log2	0.79 +/- 0.015	0.95 +/- 0.002	0.95 +/- 0.005	0.79 +/- 0.006	0.74 +/- 0.008	0.92 +/- 0.001	0.08 +/- 0.001
	500	log2	0.79 +/- 0.016	0.95 +/- 0.001	0.95 +/- 0.005	0.79 +/- 0.004	0.74 +/- 0.011	0.92 +/- 0.002	0.08 +/- 0.002
	500	sqrt	0.79 +/- 0.016	0.95 +/- 0.001	0.95 +/- 0.005	0.79 +/- 0.004	0.74 +/- 0.011	0.92 +/- 0.002	0.08 +/- 0.002
	1000	log2	0.79 +/- 0.015	0.95 +/- 0.002	0.95 +/- 0.005	0.79 +/- 0.007	0.74 +/- 0.011	0.92 +/- 0.001	0.08 +/- 0.001
	1000	sqrt	0.78 +/- 0.018	0.96 +/- 0.002	0.95 +/- 0.004	0.81 +/- 0.008	0.75 +/- 0.013	0.92 +/- 0.001	0.08 +/- 0.001
Model (3) MFP/ B-Factor	100	log2	0.78 +/- 0.017	0.96 +/- 0.003	0.95 +/- 0.004	0.81 +/- 0.009	0.74 +/- 0.010	0.92 +/- 0.001	0.08 +/- 0.001
	100	sqrt	0.78 +/- 0.017	0.95 +/- 0.002	0.95 +/- 0.004	0.79 +/- 0.007	0.73 +/- 0.012	0.91 +/- 0.001	0.09 +/- 0.001
	250	sqrt	0.78 +/- 0.016	0.96 +/- 0.003	0.95 +/- 0.004	0.81 +/- 0.009	0.75 +/- 0.010	0.92 +/- 0.001	0.08 +/- 0.001
	250	log2	0.78 +/- 0.015	0.95 +/- 0.003	0.95 +/- 0.005	0.79 +/- 0.009	0.73 +/- 0.012	0.92 +/- 0.001	0.08 +/- 0.001
	500	log2	0.78 +/- 0.016	0.95 +/- 0.002	0.95 +/- 0.004	0.79 +/- 0.008	0.73 +/- 0.012	0.92 +/- 0.001	0.08 +/- 0.001
	500	sqrt	0.78 +/- 0.017	0.96 +/- 0.003	0.96 +/- 0.004	0.82 +/- 0.008	0.75 +/- 0.010	0.92 +/- 0.000	0.08 +/- 0.000
	1000	log2	0.78 +/- 0.018	0.95 +/- 0.003	0.95 +/- 0.005	0.79 +/- 0.009	0.74 +/- 0.012	0.92 +/- 0.001	0.08 +/- 0.001
	1000	sqrt	0.78 +/- 0.017	0.96 +/- 0.002	0.96 +/- 0.004	0.81 +/- 0.007	0.75 +/- 0.011	0.92 +/- 0.001	0.08 +/- 0.001
Model (4) Random Targetlabel	100	log2	0.77 +/- 0.019	0.93 +/- 0.004	0.94 +/- 0.005	0.74 +/- 0.012	0.69 +/- 0.016	0.90 +/- 0.002	0.10 +/- 0.002
	100	sqrt	0.76 +/- 0.015	0.93 +/- 0.005	0.94 +/- 0.005	0.74 +/- 0.014	0.69 +/- 0.012	0.90 +/- 0.002	0.10 +/- 0.002
	250	sqrt	0.77 +/- 0.019	0.93 +/- 0.004	0.95 +/- 0.005	0.74 +/- 0.012	0.69 +/- 0.015	0.90 +/- 0.002	0.10 +/- 0.002
	250	log2	0.77 +/- 0.017	0.93 +/- 0.005	0.94 +/- 0.005	0.74 +/- 0.013	0.69 +/- 0.012	0.90 +/- 0.002	0.10 +/- 0.002
	500	log2	0.77 +/- 0.015	0.93 +/- 0.004	0.94 +/- 0.005	0.74 +/- 0.012	0.69 +/- 0.012	0.90 +/- 0.001	0.10 +/- 0.001
	500	sqrt	0.77 +/- 0.020	0.93 +/- 0.005	0.95 +/- 0.005	0.74 +/- 0.013	0.69 +/- 0.015	0.90 +/- 0.002	0.10 +/- 0.002
	1000	log2	0.77 +/- 0.016	0.93 +/- 0.003	0.94 +/- 0.005	0.74 +/- 0.010	0.69 +/- 0.014	0.90 +/- 0.002	0.10 +/- 0.002
	1000	sqrt	0.77 +/- 0.019	0.93 +/- 0.005	0.95 +/- 0.005	0.74 +/- 0.014	0.69 +/- 0.016	0.90 +/- 0.001	0.10 +/- 0.001

Table 3.10: Evaluation metrics of PCM models for the GMGC kinase group, trained on a data set excluding all selective compounds. For each RF model trained on the target, this table list the sensititivity, the specificity, the area under the ROC curve (AUC), the precision and the Matthews Correlation Coefficient (MCC) (Section 3.4.2) as well as the Out-of-bag performance and the Out-of-bag error (Section 1.3.2). The number of active or inactive data points in training- and test set and the number of features are detailed in Table 3.2.

Model	Parameter		Sensitivity	Specificity	AUC	Precision	MCC	OOB	OOB error
	# Trees	Best Split							
Model (1) MPF/ AAS	100	log2	0.70 +/- 0.018	0.94 +/- 0.003	0.92 +/- 0.004	0.74 +/- 0.013	0.65 +/- 0.018	0.89 +/- 0.002	0.11 +/- 0.002
	100	sqrt	0.69 +/- 0.014	0.93 +/- 0.005	0.91 +/- 0.005	0.71 +/- 0.016	0.63 +/- 0.017	0.88 +/- 0.002	0.12 +/- 0.002
	250	log2	0.69 +/- 0.014	0.93 +/- 0.004	0.91 +/- 0.006	0.71 +/- 0.013	0.63 +/- 0.015	0.88 +/- 0.002	0.12 +/- 0.002
	250	sqrt	0.70 +/- 0.018	0.94 +/- 0.003	0.92 +/- 0.005	0.74 +/- 0.013	0.65 +/- 0.018	0.89 +/- 0.002	0.11 +/- 0.002
	500	log2	0.69 +/- 0.015	0.93 +/- 0.004	0.91 +/- 0.006	0.71 +/- 0.014	0.63 +/- 0.016	0.88 +/- 0.002	0.12 +/- 0.002
	500	sqrt	0.70 +/- 0.019	0.94 +/- 0.003	0.92 +/- 0.005	0.74 +/- 0.013	0.65 +/- 0.018	0.89 +/- 0.002	0.11 +/- 0.002
	1000	log2	0.69 +/- 0.015	0.93 +/- 0.005	0.91 +/- 0.006	0.72 +/- 0.016	0.63 +/- 0.017	0.88 +/- 0.002	0.12 +/- 0.002
	1000	sqrt	0.70 +/- 0.018	0.94 +/- 0.004	0.92 +/- 0.005	0.74 +/- 0.014	0.66 +/- 0.019	0.89 +/- 0.002	0.11 +/- 0.002
Model (2) MPF/ B-Factor/ AAS	100	log2	0.69 +/- 0.018	0.93 +/- 0.006	0.91 +/- 0.006	0.71 +/- 0.019	0.62 +/- 0.018	0.88 +/- 0.002	0.12 +/- 0.002
	100	sqrt	0.71 +/- 0.004	0.90 +/- 0.008	0.89 +/- 0.006	0.64 +/- 0.018	0.59 +/- 0.014	0.86 +/- 0.002	0.14 +/- 0.002
	250	sqrt	0.69 +/- 0.015	0.93 +/- 0.006	0.91 +/- 0.005	0.71 +/- 0.018	0.63 +/- 0.015	0.88 +/- 0.001	0.12 +/- 0.001
	250	log2	0.71 +/- 0.007	0.90 +/- 0.007	0.89 +/- 0.006	0.65 +/- 0.017	0.60 +/- 0.012	0.86 +/- 0.002	0.14 +/- 0.002
	500	log2	0.56 +/- 0.036	0.97 +/- 0.002	0.91 +/- 0.008	0.65 +/- 0.014	0.57 +/- 0.024	0.93 +/- 0.001	0.07 +/- 0.001
	500	sqrt	0.71 +/- 0.009	0.90 +/- 0.008	0.89 +/- 0.006	0.65 +/- 0.017	0.60 +/- 0.013	0.86 +/- 0.002	0.14 +/- 0.002
	1000	log2	0.71 +/- 0.007	0.91 +/- 0.006	0.89 +/- 0.006	0.65 +/- 0.014	0.60 +/- 0.011	0.86 +/- 0.002	0.14 +/- 0.002
	1000	sqrt	0.69 +/- 0.012	0.93 +/- 0.006	0.91 +/- 0.005	0.72 +/- 0.017	0.63 +/- 0.014	0.88 +/- 0.001	0.12 +/- 0.001
Model (3) MFP/ B-Factor	100	log2	0.68 +/- 0.011	0.93 +/- 0.008	0.91 +/- 0.005	0.72 +/- 0.023	0.63 +/- 0.016	0.88 +/- 0.001	0.12 +/- 0.001
	100	sqrt	0.70 +/- 0.008	0.91 +/- 0.007	0.90 +/- 0.005	0.66 +/- 0.016	0.59 +/- 0.013	0.86 +/- 0.003	0.14 +/- 0.003
	250	sqrt	0.68 +/- 0.010	0.93 +/- 0.007	0.91 +/- 0.005	0.72 +/- 0.023	0.63 +/- 0.018	0.88 +/- 0.001	0.12 +/- 0.001
	250	log2	0.70 +/- 0.009	0.91 +/- 0.007	0.90 +/- 0.005	0.67 +/- 0.017	0.60 +/- 0.014	0.86 +/- 0.003	0.14 +/- 0.003
	500	log2	0.70 +/- 0.018	0.91 +/- 0.007	0.90 +/- 0.005	0.67 +/- 0.017	0.60 +/- 0.016	0.87 +/- 0.004	0.13 +/- 0.004
	500	sqrt	0.68 +/- 0.013	0.93 +/- 0.005	0.91 +/- 0.005	0.72 +/- 0.014	0.63 +/- 0.013	0.88 +/- 0.000	0.12 +/- 0.000
	1000	log2	0.70 +/- 0.007	0.91 +/- 0.006	0.90 +/- 0.005	0.67 +/- 0.014	0.61 +/- 0.010	0.87 +/- 0.002	0.13 +/- 0.002
	1000	sqrt	0.68 +/- 0.009	0.93 +/- 0.005	0.91 +/- 0.005	0.72 +/- 0.017	0.63 +/- 0.013	0.88 +/- 0.001	0.12 +/- 0.001
Model (4) Random Targetlabel	100	log2	0.63 +/- 0.020	0.91 +/- 0.007	0.88 +/- 0.004	0.65 +/- 0.014	0.55 +/- 0.012	0.85 +/- 0.001	0.15 +/- 0.001
	100	sqrt	0.63 +/- 0.016	0.91 +/- 0.008	0.87 +/- 0.003	0.64 +/- 0.017	0.55 +/- 0.010	0.85 +/- 0.001	0.15 +/- 0.001
	250	sqrt	0.63 +/- 0.017	0.91 +/- 0.006	0.88 +/- 0.004	0.65 +/- 0.013	0.55 +/- 0.011	0.86 +/- 0.002	0.14 +/- 0.002
	250	log2	0.63 +/- 0.013	0.91 +/- 0.007	0.87 +/- 0.004	0.64 +/- 0.014	0.54 +/- 0.007	0.85 +/- 0.001	0.15 +/- 0.001
	500	log2	0.63 +/- 0.019	0.91 +/- 0.006	0.87 +/- 0.004	0.64 +/- 0.013	0.55 +/- 0.011	0.86 +/- 0.002	0.14 +/- 0.002
	500	sqrt	0.64 +/- 0.017	0.91 +/- 0.006	0.88 +/- 0.004	0.65 +/- 0.013	0.55 +/- 0.010	0.86 +/- 0.002	0.14 +/- 0.002
	1000	log2	0.63 +/- 0.019	0.91 +/- 0.007	0.87 +/- 0.004	0.64 +/- 0.013	0.55 +/- 0.008	0.86 +/- 0.002	0.14 +/- 0.002
	1000	sqrt	0.64 +/- 0.018	0.91 +/- 0.006	0.88 +/- 0.004	0.65 +/- 0.011	0.55 +/- 0.010	0.86 +/- 0.002	0.14 +/- 0.002

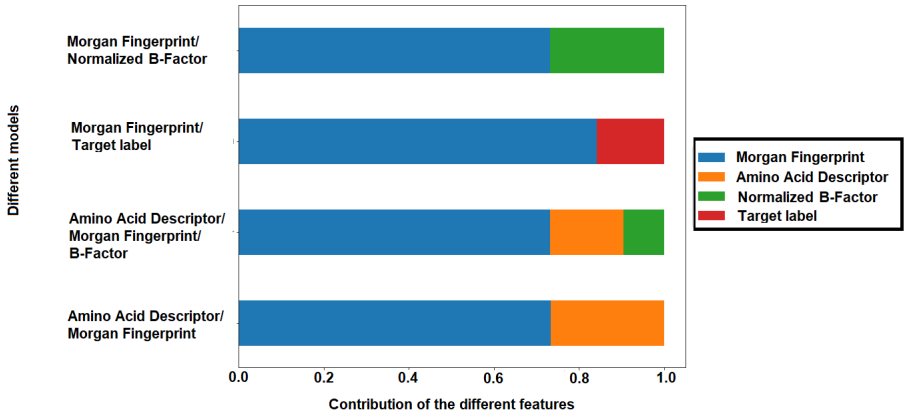


Figure 3.8: The relative feature importance of the different features in the four RF models for the target class TK. Each plot shows the contribution of the different features (MFP, AAS, B-Factor) in my four models. Interestingly, in terms of overall feature importance, the normalized B-Factor is comparable to the amino acid descriptors, indicating that some information about the binding site is contained in these features. However, when used together with amino acid descriptor and Morgan Fingerprints, the normalized B-Factors contribute less than the amino acid descriptors to the overall model. Lastly, this graph also shows that the Morgan Fingerprints have the highest feature importance with up to 80%. In each case, the features sum up to 1.

3.6.2 Exploration of the RMSD as a Measure for the Flexibility of a Residue at a Given Position

After calculating the per-residue RMSD values as described in Section 3.5.3, I assessed their potential for being an approximation for 'residue flexibility' by plotting the RMSD values for all available targets in a given kinase subclass and the RMSD values versus the respective binding pocket residue, denoted by the KLIFS nomenclature.

As shown in Figure 3.10 for the target class AGC, most of the RMSD values for a given target class are concentrated in the range of 0.1-0.5 Å. The corresponding violin plots for the other target classes show the same trend (not shown). For each target class, highly correlated RMSD values were removed with a threshold of 0.7, as detailed in Section 3.5.5. The corresponding values are presented in Table 3.5.

Four RF models were built for each target class analogously to Section 3.6.1, if the number of available RMSD features exceeded at least five features (see Table 3.4). Hence, only RF models for the target classes 'GMGC', 'AGC' and 'CAMK' were built. The results of the RMSD can be interpreted analogously to the previous section, Section 3.6.1. Generally, Models (1), RF models incorporating only amino acid features and Morgan Fingerprints, perform best in terms of performance metrices (see Table 3.11 - Table 3.13).

However, RF models incoporating both RMSD values and Morgan Fingerprints and models incorporating RMSD values, amino acid descriptors and Morgan Fingerprints perform comparably, indicating that the RMSD values might be used as a replacement for amino acid descriptors describe the target binding pocket of kinases. They also perform comparably in terms of MCCs and AUCs to models (1). However, given the relatively high performance of Models (4), models incorporating a target label to describe the kinase targets and Morgan Fingerprints, these results are not surprising and indicate that the chemical features describing the kinases might be responsible for the high performance.

Hence, to test this hypothesis and to investigate the relative influence of the addition of the RMSD values in the different models, I also summed up the feature importances of each descriptor in each model. As exemplified for the target class TK, Figure 3.9, a random target label contributes very little to the model on its own. Interestingly, however, both the amino acid descriptors and the RMSD values contribute more than the target label and within in the same range, indicating that the RSMD values contains specific information about the binding pocket residues. However, when both amino acid descriptors and RMSD values are used, the amino acid descriptors contribute more to the overall RF model. In all four cases, the Morgan Fingerprints were the most important features (80%). This implies that our current models might be biased in terms of a well-explored chemical kinase space. Future experiments should thus place an emphasis on exploring and considering chemical similarity in these RF models. Alternatively, a different choice of protein family with a less explored chemical space might have been a more appropriate choice for testing the influence of 'flexibility descriptors' in a RF model.

Table 3.11: Evaluation metrics of PCM models for the TK kinase group, trained on a data set excluding all selective compounds. For each RF model trained on the target, this table list the sensititivity, the specificity, the area under the ROC curve (AUC), the precision and the Matthews Correlation Coefficient (MCC) (Section 3.4.2) as well as the Out-of-bag performance and the Out-of-bag error (Section 1.3.2). The number of active or inactive data points in training- and test set and the number of features are detailed in Table 3.4.

Model	Parameter		Sensitivity	Specificity	AUC	Precision	MCC	OOB	OOB error
	# Trees	Best Split							
Model (1) MPF/ AAS	100	log2	0.82 +/- 0.011	0.94 +/- 0.002	0.96 +/- 0.002	0.84 +/- 0.004	0.77 +/- 0.007	0.90 +/- 0.001	0.10 +/- 0.001
	100	sqrt	0.79 +/- 0.008	0.93 +/- 0.002	0.94 +/- 0.003	0.83 +/- 0.004	0.74 +/- 0.006	0.89 +/- 0.001	0.11 +/- 0.001
	250	sqrt	0.82 +/- 0.010	0.94 +/- 0.003	0.96 +/- 0.002	0.85 +/- 0.005	0.77 +/- 0.007	0.90 +/- 0.001	0.10 +/- 0.001
	250	log2	0.79 +/- 0.008	0.93 +/- 0.002	0.95 +/- 0.002	0.83 +/- 0.004	0.74 +/- 0.007	0.89 +/- 0.001	0.11 +/- 0.001
	500	log2	0.79 +/- 0.010	0.94 +/- 0.002	0.95 +/- 0.003	0.83 +/- 0.004	0.74 +/- 0.008	0.89 +/- 0.001	0.11 +/- 0.001
	500	sqrt	0.82 +/- 0.010	0.94 +/- 0.002	0.96 +/- 0.002	0.85 +/- 0.004	0.77 +/- 0.007	0.90 +/- 0.001	0.10 +/- 0.001
	1000	log2	0.80 +/- 0.010	0.94 +/- 0.001	0.95 +/- 0.003	0.83 +/- 0.003	0.74 +/- 0.008	0.89 +/- 0.001	0.11 +/- 0.001
	1000	sqrt	0.82 +/- 0.009	0.94 +/- 0.003	0.96 +/- 0.002	0.85 +/- 0.006	0.77 +/- 0.008	0.90 +/- 0.001	0.10 +/- 0.001
Model (2) MPF/ RMSD/ AAS	100	log2	0.81 +/- 0.006	0.94 +/- 0.002	0.96 +/- 0.003	0.85 +/- 0.004	0.76 +/- 0.005	0.90 +/- 0.001	0.10 +/- 0.001
	100	sqrt	0.79 +/- 0.009	0.93 +/- 0.002	0.94 +/- 0.003	0.82 +/- 0.004	0.73 +/- 0.007	0.89 +/- 0.001	0.11 +/- 0.001
	250	sqrt	0.82 +/- 0.006	0.94 +/- 0.001	0.96 +/- 0.003	0.85 +/- 0.003	0.76 +/- 0.005	0.90 +/- 0.001	0.10 +/- 0.001
	250	log2	0.79 +/- 0.008	0.93 +/- 0.001	0.94 +/- 0.003	0.83 +/- 0.002	0.74 +/- 0.005	0.89 +/- 0.001	0.11 +/- 0.001
	500	log2	0.79 +/- 0.007	0.93 +/- 0.001	0.94 +/- 0.003	0.83 +/- 0.003	0.74 +/- 0.005	0.89 +/- 0.001	0.11 +/- 0.001
	500	sqrt	0.81 +/- 0.008	0.94 +/- 0.002	0.96 +/- 0.002	0.85 +/- 0.004	0.76 +/- 0.007	0.90 +/- 0.001	0.10 +/- 0.001
	1000	log2	0.79 +/- 0.007	0.93 +/- 0.001	0.94 +/- 0.003	0.83 +/- 0.002	0.74 +/- 0.005	0.89 +/- 0.001	0.11 +/- 0.001
	1000	sqrt	0.81 +/- 0.008	0.94 +/- 0.002	0.96 +/- 0.002	0.85 +/- 0.004	0.76 +/- 0.006	0.90 +/- 0.001	0.10 +/- 0.001
Model (3) MFP/ RMSD	100	log2	0.83 +/- 0.013	0.93 +/- 0.003	0.96 +/- 0.002	0.83 +/- 0.005	0.76 +/- 0.008	0.90 +/- 0.001	0.10 +/- 0.001
	100	sqrt	0.81 +/- 0.010	0.93 +/- 0.003	0.95 +/- 0.002	0.82 +/- 0.008	0.74 +/- 0.011	0.89 +/- 0.001	0.11 +/- 0.001
	250	sqrt	0.83 +/- 0.012	0.94 +/- 0.004	0.96 +/- 0.002	0.84 +/- 0.007	0.77 +/- 0.008	0.90 +/- 0.001	0.10 +/- 0.001
	250	log2	0.81 +/- 0.010	0.93 +/- 0.003	0.95 +/- 0.002	0.83 +/- 0.006	0.75 +/- 0.009	0.90 +/- 0.001	0.10 +/- 0.001
	500	log2	0.81 +/- 0.01	0.93 +/- 0.00	0.95 +/- 0.00	0.83 +/- 0.01	0.75 +/- 0.01	0.90 +/- 0.00	0.10 +/- 0.00
	500	sqrt	0.83 +/- 0.01	0.93 +/- 0.00	0.96 +/- 0.00	0.84 +/- 0.01	0.77 +/- 0.01	0.90 +/- 0.001	0.10 +/- 0.001
	1000	log2	0.83 +/- 0.01	0.93 +/- 0.00	0.96 +/- 0.00	0.83 +/- 0.01	0.76 +/- 0.01	0.90 +/- 0.001	0.10 +/- 0.001
	1000	sqrt	0.83 +/- 0.01	0.94 +/- 0.00	0.96 +/- 0.00	0.84 +/- 0.01	0.77 +/- 0.01	0.90 +/- 0.001	0.10 +/- 0.001
Model (4) Random Targetlabel	100	log2	0.84 +/- 0.008	0.91 +/- 0.006	0.94 +/- 0.003	0.78 +/- 0.010	0.73 +/- 0.008	0.88 +/- 0.002	0.12 +/- 0.002
	100	sqrt	0.82 +/- 0.009	0.91 +/- 0.006	0.94 +/- 0.003	0.78 +/- 0.010	0.72 +/- 0.009	0.88 +/- 0.001	0.12 +/- 0.001
	250	sqrt	0.84 +/- 0.011	0.91 +/- 0.005	0.94 +/- 0.003	0.78 +/- 0.009	0.73 +/- 0.009	0.89 +/- 0.002	0.11 +/- 0.002
	250	log2	0.82 +/- 0.014	0.91 +/- 0.006	0.94 +/- 0.004	0.78 +/- 0.010	0.72 +/- 0.011	0.88 +/- 0.001	0.12 +/- 0.001
	500	log2	0.82 +/- 0.013	0.91 +/- 0.007	0.94 +/- 0.004	0.78 +/- 0.012	0.72 +/- 0.012	0.88 +/- 0.001	0.12 +/- 0.001
	500	sqrt	0.84 +/- 0.010	0.91 +/- 0.005	0.94 +/- 0.003	0.78 +/- 0.009	0.73 +/- 0.009	0.89 +/- 0.001	0.11 +/- 0.001
	1000	log2	0.82 +/- 0.011	0.91 +/- 0.006	0.94 +/- 0.003	0.78 +/- 0.011	0.72 +/- 0.010	0.88 +/- 0.001	0.12 +/- 0.001
	1000	sqrt	0.84 +/- 0.010	0.91 +/- 0.006	0.94 +/- 0.003	0.78 +/- 0.010	0.73 +/- 0.009	0.89 +/- 0.002	0.11 +/- 0.002

Table 3.12: Evaluation metrics of PCM models for the CAMK kinase group, trained on a data set excluding all selective compounds. For each RF model trained on the target, this table list the sensititivity, the specificity, the area under the ROC curve (AUC), the precision and the Matthews Correlation Coefficient (MCC) (Section 3.4.2) as well as the Out-of-bag performance and the Out-of-bag error (Section 1.3.2). The number of active or inactive data points in training- and test set and the number of features are detailed in Table 3.4.

Model	Parameter		Sensitivity	Specificity	AUC	Precision	MCC	OOB	OOB error
	# Trees	Best Split							
Model (1) MPF/ AAS	100	log2	0.79 +/- 0.017	0.97 +/- 0.005	0.96 +/- 0.004	0.90 +/- 0.014	0.80 +/- 0.007	0.93 +/- 0.001	0.07 +/- 0.001
	100	sqrt	0.77 +/- 0.022	0.97 +/- 0.006	0.95 +/- 0.005	0.89 +/- 0.019	0.78 +/- 0.005	0.92 +/- 0.001	0.08 +/- 0.001
	250	sqrt	0.79 +/- 0.020	0.97 +/- 0.004	0.96 +/- 0.005	0.90 +/- 0.012	0.80 +/- 0.009	0.93 +/- 0.001	0.07 +/- 0.001
	250	log2	0.77 +/- 0.024	0.97 +/- 0.006	0.95 +/- 0.005	0.89 +/- 0.016	0.78 +/- 0.007	0.92 +/- 0.000	0.08 +/- 0.000
	500	log2	0.77 +/- 0.025	0.97 +/- 0.005	0.96 +/- 0.006	0.90 +/- 0.014	0.78 +/- 0.010	0.92 +/- 0.001	0.08 +/- 0.001
	500	sqrt	0.79 +/- 0.018	0.97 +/- 0.005	0.96 +/- 0.005	0.90 +/- 0.014	0.80 +/- 0.006	0.93 +/- 0.001	0.07 +/- 0.001
	1000	log2	0.77 +/- 0.023	0.97 +/- 0.005	0.95 +/- 0.005	0.89 +/- 0.013	0.78 +/- 0.011	0.92 +/- 0.001	0.08 +/- 0.001
	1000	sqrt	0.79 +/- 0.018	0.97 +/- 0.006	0.96 +/- 0.005	0.90 +/- 0.019	0.80 +/- 0.008	0.93 +/- 0.001	0.07 +/- 0.001
Model (2) MPF/ RMSD/ AAS	100	log2	0.79 +/- 0.018	0.97 +/- 0.005	0.96 +/- 0.004	0.91 +/- 0.015	0.80 +/- 0.007	0.93 +/- 0.001	0.07 +/- 0.001
	100	sqrt	0.78 +/- 0.022	0.97 +/- 0.005	0.95 +/- 0.004	0.89 +/- 0.015	0.78 +/- 0.008	0.92 +/- 0.001	0.08 +/- 0.001
	250	sqrt	0.78 +/- 0.018	0.97 +/- 0.006	0.95 +/- 0.004	0.89 +/- 0.019	0.79 +/- 0.007	0.92 +/- 0.001	0.08 +/- 0.001
	250	log2	0.77 +/- 0.022	0.97 +/- 0.006	0.95 +/- 0.006	0.88 +/- 0.016	0.77 +/- 0.007	0.92 +/- 0.001	0.08 +/- 0.001
	500	log2	0.78 +/- 0.015	0.97 +/- 0.006	0.95 +/- 0.004	0.89 +/- 0.019	0.79 +/- 0.005	0.93 +/- 0.001	0.07 +/- 0.001
	500	sqrt	0.79 +/- 0.012	0.97 +/- 0.004	0.96 +/- 0.004	0.90 +/- 0.014	0.80 +/- 0.006	0.93 +/- 0.001	0.07 +/- 0.001
	1000	log2	0.78 +/- 0.015	0.97 +/- 0.005	0.95 +/- 0.004	0.90 +/- 0.016	0.79 +/- 0.003	0.93 +/- 0.001	0.07 +/- 0.001
	1000	sqrt	0.79 +/- 0.013	0.97 +/- 0.004	0.96 +/- 0.004	0.90 +/- 0.014	0.80 +/- 0.006	0.93 +/- 0.001	0.07 +/- 0.001
Model (3) MFP/ RMSD	100	log2	0.78 +/- 0.018	0.97 +/- 0.005	0.95 +/- 0.004	0.89 +/- 0.015	0.79 +/- 0.003	0.93 +/- 0.001	0.07 +/- 0.001
	100	sqrt	0.77 +/- 0.019	0.97 +/- 0.005	0.95 +/- 0.007	0.88 +/- 0.013	0.77 +/- 0.008	0.92 +/- 0.001	0.08 +/- 0.001
	250	sqrt	0.78 +/- 0.018	0.97 +/- 0.006	0.96 +/- 0.005	0.89 +/- 0.017	0.79 +/- 0.005	0.93 +/- 0.001	0.07 +/- 0.001
	250	log2	0.77 +/- 0.022	0.97 +/- 0.006	0.95 +/- 0.006	0.88 +/- 0.016	0.77 +/- 0.007	0.92 +/- 0.001	0.08 +/- 0.001
	500	log2	0.77 +/- 0.020	0.97 +/- 0.006	0.95 +/- 0.006	0.88 +/- 0.016	0.77 +/- 0.005	0.92 +/- 0.001	0.08 +/- 0.001
	500	sqrt	0.78 +/- 0.018	0.97 +/- 0.005	0.96 +/- 0.004	0.90 +/- 0.016	0.79 +/- 0.005	0.93 +/- 0.002	0.07 +/- 0.002
	1000	log2	0.77 +/- 0.017	0.97 +/- 0.006	0.95 +/- 0.006	0.88 +/- 0.017	0.77 +/- 0.005	0.92 +/- 0.001	0.08 +/- 0.001
	1000	sqrt	0.78 +/- 0.016	0.97 +/- 0.006	0.96 +/- 0.004	0.90 +/- 0.017	0.79 +/- 0.003	0.93 +/- 0.001	0.07 +/- 0.001
Model (4) Random Targetlabel	100	log2	0.75 +/- 0.029	0.95 +/- 0.008	0.94 +/- 0.008	0.83 +/- 0.018	0.73 +/- 0.013	0.91 +/- 0.001	0.09 +/- 0.001
	100	sqrt	0.76 +/- 0.024	0.96 +/- 0.009	0.94 +/- 0.009	0.84 +/- 0.025	0.74 +/- 0.013	0.91 +/- 0.001	0.09 +/- 0.001
	250	log2	0.75 +/- 0.033	0.96 +/- 0.008	0.94 +/- 0.008	0.84 +/- 0.018	0.74 +/- 0.013	0.91 +/- 0.002	0.09 +/- 0.002
	250	sqrt	0.77 +/- 0.026	0.96 +/- 0.009	0.94 +/- 0.008	0.84 +/- 0.024	0.75 +/- 0.014	0.91 +/- 0.001	0.09 +/- 0.001
	500	log2	0.75 +/- 0.027	0.96 +/- 0.008	0.94 +/- 0.008	0.84 +/- 0.021	0.74 +/- 0.011	0.91 +/- 0.001	0.09 +/- 0.001
	500	sqrt	0.76 +/- 0.027	0.96 +/- 0.009	0.94 +/- 0.008	0.84 +/- 0.023	0.74 +/- 0.015	0.91 +/- 0.001	0.09 +/- 0.001
	1000	log2	0.75 +/- 0.028	0.96 +/- 0.007	0.94 +/- 0.008	0.84 +/- 0.017	0.74 +/- 0.013	0.91 +/- 0.001	0.09 +/- 0.001
	1000	sqrt	0.76 +/- 0.025	0.96 +/- 0.008	0.94 +/- 0.008	0.84 +/- 0.021	0.74 +/- 0.014	0.91 +/- 0.001	0.09 +/- 0.001

Table 3.13: Evaluation metrics of PCM models for the GMGC kinase group, trained on a data set excluding all selective compounds. For each RF model trained on the target, this table list the sensititivity, the specificity, the area under the ROC curve (AUC), the precision and the Matthews Correlation Coefficient (MCC) (Section 3.4.2) as well as the Out-of-bag performance and the Out-of-bag error (Section 1.3.2). The number of active or inactive data points in training- and test set and the number of features are detailed in Table 3.4.

Model	Parameter		Sensitivity	Specificity	AUC	Precision	MCC	OOB	OOB error
	# Trees	Best Split							
Model (1) MPF/ AAS	100	log2	0.77 +/- 0.006	0.92 +/- 0.005	0.92 +/- 0.003	0.78 +/- 0.009	0.69 +/- 0.008	0.87 +/- 0.002	0.13 +/- 0.002
	100	sqrt	0.76 +/- 0.008	0.89 +/- 0.006	0.91 +/- 0.002	0.74 +/- 0.010	0.65 +/- 0.006	0.85 +/- 0.003	0.15 +/- 0.003
	250	sqrt	0.77 +/- 0.008	0.92 +/- 0.005	0.92 +/- 0.003	0.79 +/- 0.010	0.69 +/- 0.008	0.87 +/- 0.001	0.13 +/- 0.001
	250	log2	0.77 +/- 0.010	0.89 +/- 0.006	0.91 +/- 0.003	0.74 +/- 0.010	0.66 +/- 0.009	0.86 +/- 0.001	0.14 +/- 0.001
	500	log2	0.76 +/- 0.009	0.90 +/- 0.006	0.91 +/- 0.003	0.74 +/- 0.010	0.65 +/- 0.006	0.86 +/- 0.001	0.14 +/- 0.001
	500	sqrt	0.77 +/- 0.006	0.92 +/- 0.005	0.93 +/- 0.003	0.79 +/- 0.010	0.69 +/- 0.009	0.88 +/- 0.001	0.12 +/- 0.001
	1000	log2	0.77 +/- 0.011	0.90 +/- 0.005	0.91 +/- 0.003	0.75 +/- 0.008	0.66 +/- 0.007	0.86 +/- 0.001	0.14 +/- 0.001
	1000	sqrt	0.77 +/- 0.008	0.92 +/- 0.005	0.93 +/- 0.003	0.79 +/- 0.010	0.69 +/- 0.008	0.88 +/- 0.002	0.12 +/- 0.002
Model (2) MPF/ RMSD/ AAS	100	log2	0.76 +/- 0.004	0.92 +/- 0.006	0.92 +/- 0.003	0.79 +/- 0.012	0.69 +/- 0.010	0.87 +/- 0.002	0.13 +/- 0.002
	100	sqrt	0.76 +/- 0.006	0.88 +/- 0.004	0.90 +/- 0.002	0.72 +/- 0.006	0.64 +/- 0.002	0.84 +/- 0.002	0.16 +/- 0.00
	250	sqrt	0.76 +/- 0.008	0.92 +/- 0.005	0.92 +/- 0.003	0.79 +/- 0.010	0.69 +/- 0.009	0.88 +/- 0.002	0.12 +/- 0.002
	250	log2	0.76 +/- 0.007	0.89 +/- 0.005	0.90 +/- 0.002	0.73 +/- 0.008	0.64 +/- 0.005	0.85 +/- 0.001	0.15 +/- 0.001
	500	log2	0.76 +/- 0.007	0.89 +/- 0.005	0.90 +/- 0.002	0.73 +/- 0.008	0.64 +/- 0.003	0.85 +/- 0.001	0.15 +/- 0.001
	500	sqrt	0.76 +/- 0.009	0.92 +/- 0.004	0.92 +/- 0.003	0.79 +/- 0.008	0.69 +/- 0.009	0.88 +/- 0.002	0.12 +/- 0.002
	1000	log2	0.76 +/- 0.009	0.89 +/- 0.005	0.91 +/- 0.002	0.73 +/- 0.006	0.64 +/- 0.001	0.85 +/- 0.002	0.15 +/- 0.002
	1000	sqrt	0.76 +/- 0.008	0.92 +/- 0.005	0.92 +/- 0.003	0.79 +/- 0.010	0.69 +/- 0.009	0.88 +/- 0.002	0.12 +/- 0.002
Model (3) MFP/ RMSD	100	log2	0.76 +/- 0.009	0.92 +/- 0.004	0.92 +/- 0.002	0.78 +/- 0.008	0.68 +/- 0.007	0.87 +/- 0.001	0.13 +/- 0.001
	100	sqrt	0.75 +/- 0.012	0.89 +/- 0.007	0.91 +/- 0.002	0.74 +/- 0.012	0.64 +/- 0.010	0.85 +/- 0.003	0.15 +/- 0.003
	250	sqrt	0.76 +/- 0.009	0.92 +/- 0.005	0.92 +/- 0.002	0.78 +/- 0.009	0.68 +/- 0.009	0.87 +/- 0.001	0.13 +/- 0.001
	250	log2	0.76 +/- 0.013	0.90 +/- 0.008	0.91 +/- 0.002	0.74 +/- 0.011	0.65 +/- 0.007	0.85 +/- 0.002	0.15 +/- 0.002
	500	log2	0.76 +/- 0.012	0.90 +/- 0.006	0.91 +/- 0.002	0.74 +/- 0.009	0.65 +/- 0.007	0.86 +/- 0.001	0.14 +/- 0.001
	500	sqrt	0.76 +/- 0.009	0.92 +/- 0.004	0.92 +/- 0.002	0.79 +/- 0.008	0.69 +/- 0.008	0.87 +/- 0.001	0.13 +/- 0.001
	1000	log2	0.76 +/- 0.012	0.90 +/- 0.006	0.91 +/- 0.002	0.74 +/- 0.009	0.65 +/- 0.006	0.86 +/- 0.001	0.14 +/- 0.001
	1000	sqrt	0.76 +/- 0.009	0.92 +/- 0.004	0.92 +/- 0.002	0.79 +/- 0.007	0.69 +/- 0.006	0.87 +/- 0.001	0.13 +/- 0.001
Model (4) Random Targetlabel	log2	100	0.72 +/- 0.013	0.90 +/- 0.009	0.90 +/- 0.004	0.74 +/- 0.018	0.63 +/- 0.016	0.85 +/- 0.001	0.15 +/- 0.001
	100	sqrt	0.72 +/- 0.008	0.89 +/- 0.007	0.89 +/- 0.003	0.73 +/- 0.013	0.61 +/- 0.009	0.84 +/- 0.002	0.16 +/- 0.002
	250	sqrt	0.73 +/- 0.010	0.90 +/- 0.009	0.90 +/- 0.003	0.75 +/- 0.016	0.63 +/- 0.012	0.85 +/- 0.001	0.15 +/- 0.001
	250	log2	0.72 +/- 0.008	0.89 +/- 0.008	0.89 +/- 0.004	0.73 +/- 0.014	0.61 +/- 0.010	0.84 +/- 0.003	0.16 +/- 0.003
	500	log2	0.72 +/- 0.009	0.89 +/- 0.008	0.89 +/- 0.004	0.73 +/- 0.014	0.62 +/- 0.011	0.85 +/- 0.003	0.15 +/- 0.003
	500	sqrt	0.73 +/- 0.011	0.90 +/- 0.009	0.90 +/- 0.003	0.75 +/- 0.015	0.64 +/- 0.009	0.85 +/- 0.001	0.15 +/- 0.001
	1000	log2	0.72 +/- 0.009	0.89 +/- 0.008	0.89 +/- 0.004	0.73 +/- 0.014	0.62 +/- 0.011	0.85 +/- 0.003	0.15 +/- 0.003
	1000	sqrt	0.73 +/- 0.011	0.90 +/- 0.009	0.90 +/- 0.003	0.75 +/- 0.015	0.64 +/- 0.011	0.85 +/- 0.001	0.15 +/- 0.001

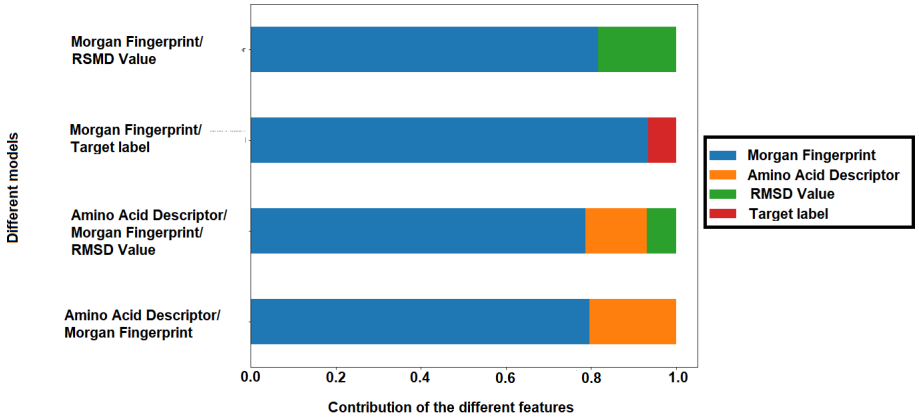


Figure 3.9: The relative feature importance of the different features in the four RF models for the target class TK. Each plot shows the contribution of the different features (MFP, AAS, RMSD) in my four models. As expected, the target label contributes little to the overall model performance. Interestingly, however, the RMSD values contribute in the range of the amino acid descriptors, indicating that information about the binding pocket is contained within the RMSD values. However, if the RMSD is used together with amino acid descriptor and Morgan Fingerprints, the information contained in its features is not as important as the one contained in the amino acid descriptors. Lastly, the Morgan Fingerprints have the highest feature importance, indicating that they are mostly used by the RF models. In each case, the features sum up to 1.

Lastly, I investigated whether the 'flexibility' of a residue at a given position, as approximated by the RMSD, can be connected to known structural features of a given kinase protein. Hence, I plotted the RMSD values for all 85 binding residues (using the KLIFS nomenclature) versus the highest RMSD value for each residue and target. As shown in Figure 3.11 for the target class 'AGC', unsurprisingly, most residues are associated with a low RMSD value within 0.1-0.5 Å; this trend can be observed for all other target classes as well (not shown).

However, interestingly, for the targets MELK, PIM1 and CHK1, there is a region corresponding to high RMSD values starting from residue 4 until residue 10, which belong to the G-loop, a flexible loop in protein kinases, responsible for binding (P. J. van Linden et al., 2013). Also, the targets MELK and AMPKa2 have a region of high RMSD values around KLIFS residue number 80-85, corresponding to the DFG-motif and the activation-loop.

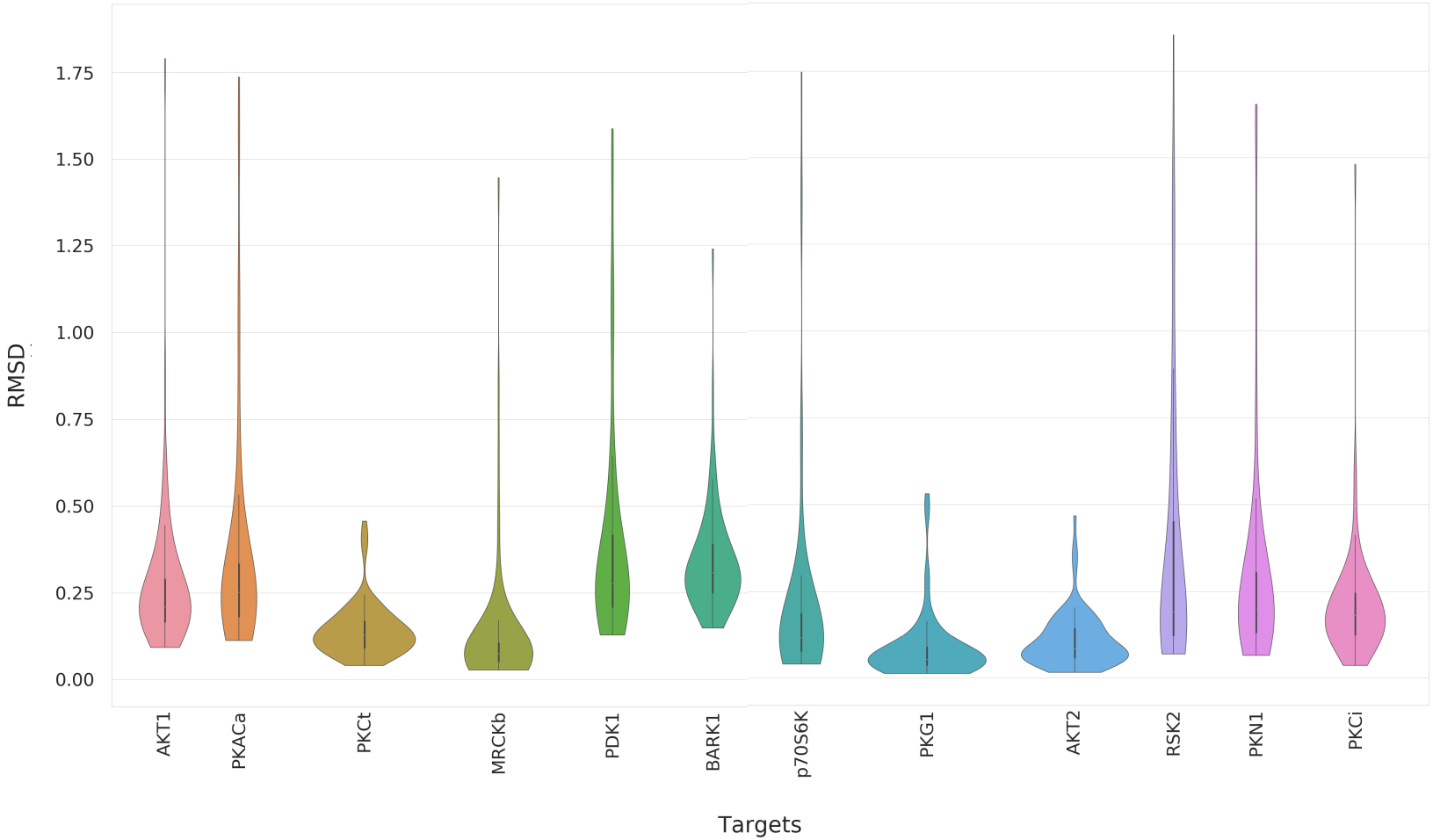


Figure 3.10: Distribution of the RMSD values for a given target in the target class AGC. For all AGC targets, for which I could calculate the RMSD values as described in Section 3.5.3, I plotted the distribution for all 85 binding residues. Interestingly, as shown here for the target class 'AGC', most RMSD values fall within the range of 0.1 to 0.5 Å.

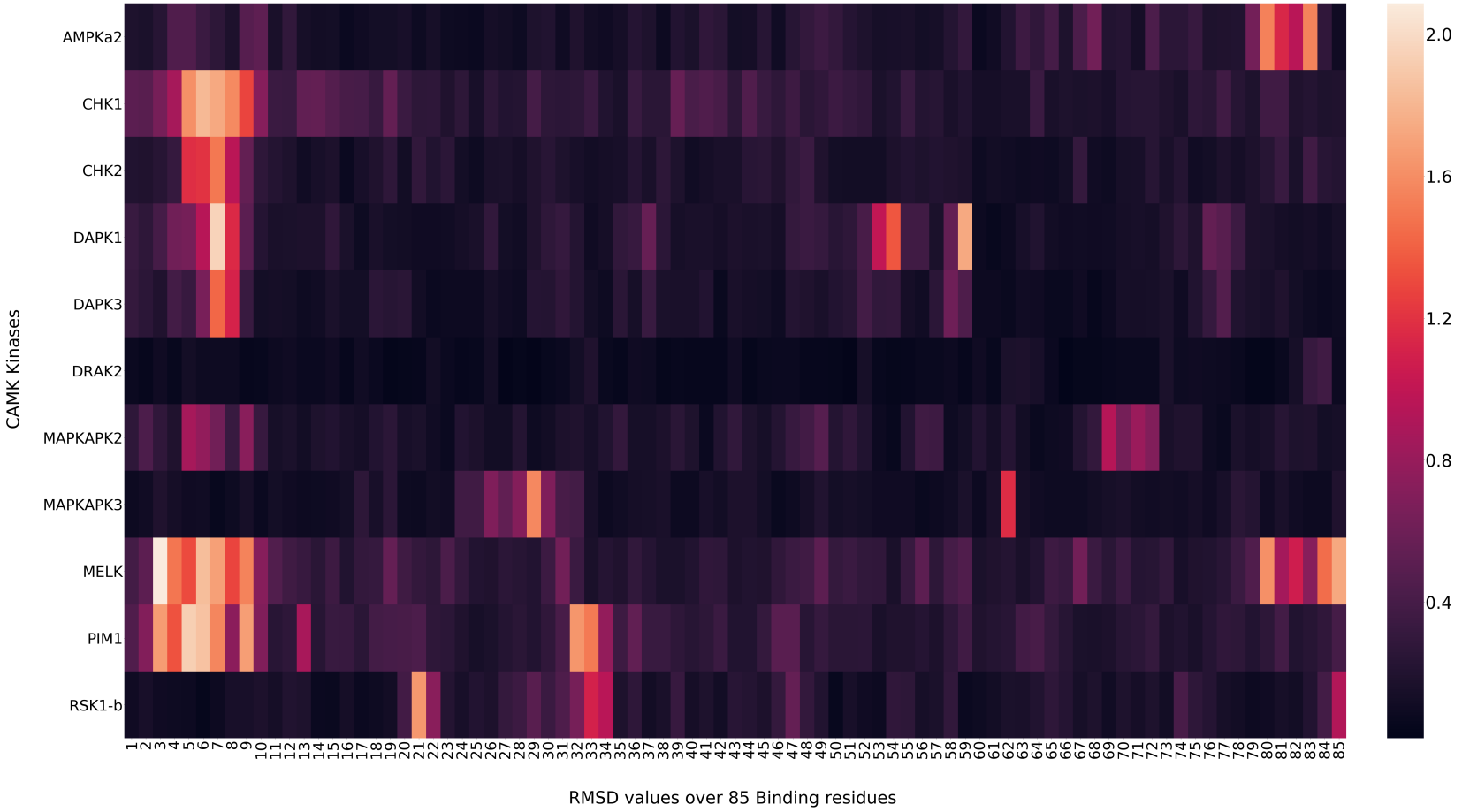


Figure 3.11: The RMSD values (in Å) for the 85 KLIFS-defined binding pocket residue positions for the target class CAMK, plotted according to the KLIFS nomenclature. Unsurprisingly, CAMK targets show a similar trend as AGC targets (see Figure 3.10); most residues are associated with a low RMSD. Regions with high RMSD values (see colorbar) correspond to flexible regions, such as the glycine-rich G-loop (KLIFS residues 4-9) and the DFG-motif and the activation loop (KLIFS residues 80-85 (P. J. van Linden et al., 2013)).

3.7 Conclusion and Future Directions

In this section, I explored the use of simple flexibility descriptors, per-residue normalized B-Factor and per-Residue RMSD values, for each binding pocket residue in RF models as complimentary protein features, given that the flexibility of a given residue can depend on the position of the residue in the binding pocket.

In all the different kinase groups, the models including Morgan Fingerprints, RMSD and normalized B-Factor values performed comparably, and often are often comparable to the models which include Morgan Fingerprints and amino acid descriptors. However, when used together with amino acid descriptors and Morgan Fingerprints, the simple flexibility descriptors were considered less important in terms of feature importance than the amino acid descriptors.
Most importantly, however, the Morgan Fingerprints were considered the most important features in a given RF model. Hence, the kinase chemical space is well-explored and can be used to build well-performing RF models with minor contribution from target descriptors, *e.g.* target labels. However, this indicates that I might not have seen a huge performance increase given that the models are already well-performing. In this case, future experiments might focus on different protein families. Also, the chemical similarity of compounds could be used as a way to built 'less well-performing' RF models, which could then be used to test the influence of the simple and more complex flexibility descriptors. Similarly, different conclusions might be reached if other 'residue flexibility' descriptors and/or other machine learning models, such as support vector machines, are used.

It could be argued that the entire kinase-ligand space for a given target has not been comprehensively sampled, and hence using the calculation of the simple flexibility descriptors is limited by the available data set. However, using the KLIFS-based, per residue RMSD values, the flexible regions (G-loop, activation loop, DFG-motif) within a kinase protein binding pocket could be identified, indicating that flexible regions of a protein can be distinguished. However, Figure 3.11 highlights that the three different regions cannot always be identified, as they also show low per-residue RMSD values for regions I expected to be flexible. I conclude that the strength of this signal depends on the available crystal structures for each target, thus the calculation of per-residue

flexibility values, such as the RMSD values, is limited by the available structure data.

There are multiple caveats with my current analyses. Firstly, uncertainties in specific atom positions in a crystal structure could have been quantified using occupancy and B-Factor values and subsequently accounted for by filtering out these structures. The normalized B-Factor for a given binding pocket residue was averaged to account for experimental uncertainty, while the highest RMSD value was selected for a given binding pocket residue was selected. Hence, the normalized B-Factors depend on the experimental uncertainty. Also, the normalized B-Factor was calculated on a per-atom basis. For each residue, the normalized B-Factors were then summarized and divided by the number of atoms; however, this could result in artificially lower normalized B-Factor values for residues with a higher number of atoms. Secondly, a further, more fine-grained division of the crystal structures into the αC helix in and out conformations was not performed as the RMSD and the normalized B-Factor were calculated on a per-residue basis, independently of the conformation, and the highest residue movement value for a given residue was desired. However, as shown in this section, while flexible regions were identified, the single per residue flexibility appears to be too coarse for the Random Forests investigated here.

Lastly, the B-Factor heavily depends on the refinement and the model and parameters the crystallographer chooses for building the model of the protein into the electron map, measured and calculated from the X-ray experiments. For each crystallographic structural model, the crystallographer has to consider the trade-offs of building a detailed crystallographic structure versus the danger of describing more than is justified by the experimental data (Merritt, 2012). The B-Factor thus also has a degree of experimental uncertainty about the 'true position' of the respective atom associated with it. Secondly, crystal contacts and the protein-ligand interaction also influence the size of a B-Factor. Crystal contacts are intermolecular or interchain interactions between proteins in a crystal and thus an artefact of crystallization (Liu et al., 2014). These interactions 'fixate' interaction chains in position. Similarly, protein-ligand interactions also 'lock' the position of interacting residue atoms and their corresponding ligand atoms in position. Their B-Factors thus tend to have lower values (Majewski et al., 2019).

Due to these considerations, the usefulness of the B-Factor as a measure for residue

movement was only briefly explored. However, it is interesting that flexible regions in kinases could still be identified using the normalized B-Factor, see Figure 3.7.

Future directions might include using molecular dynamics to gain the full potential range of motion of a residue at a given position. Future experiments could also explore the inclusion of residue flexibility descriptors of a residue in an interaction grid, such as CRANkS maps (Patel, 2018), as further discussed in Section 5.4.

4 Prediction of Selective Compounds Using 2D-Kinase Random Forest Models

4.1 Introduction

Determining the selectivity or promiscuity of a compound is not a trivial task. While Sorgenfrei *et al.* used their PCM kinase model to profile compounds as selective, achieving an average AUC of 0.89 ± 0.06 (Sorgenfrei et al., 2018), they did not use their models to predict known selective kinase inhibitors.

To test the limits of current PCM methods in predicting selective compounds, I collected a data set of 59 selective compounds from public sources as a test set. RF models were trained on the data points in the data set provided by Fulle *et al.*. To avoid data leakage between training and test set, selective compounds present in Fulle's data set were deleted from the training set and assigned to the test set. For each selective compound, its activity against the targets in the data set was calculated. If experimental data was available to test the validity of these predictions, then the sensitivity and specificity was calculated. The novelty of this research is that I explicitly test how well current kinase PCM models can predict our selective kinase inhibitors.

4.2 Methods

4.2.1 A Manually Curated Data Set of Selective Compounds for Testing RF Models

The amount of available experimental data for a given compound plays a significant role in determining whether a compound can be considered selective. As outlined in Section 1.3, Staurosporine is a historic compound: originally considered a selective compound, new data revealed its promiscuous nature (Wang et al., 2016). Hence, scientists began to develop and test selective compounds for exploring biological systems; the compounds were deposited in portals such as the Structural Genomics Consortium Oxford (SGC) and the Chemical Probes Portal. In view of these considerations, I manually collected selective kinase inhibitors from three sources:

1. The Chemical Probes Portal (CPP) (Chemical Probes Portal, 2020) is a crowd-sourced, web-based portal designed to host information about small-molecule chemical probes with a corresponding rating by experts about their potential to act as probe compounds. Users are free to submit a potential probe with supporting evidence; this evidence will then be evaluated by the experts based on criteria such as potency, selectivity and the mechanism of compound-target interaction. Four categories, the so-called SAB ratings, exist for rating a compound as a chemical probe, ranging from one to four stars: 'Not recommended as a probe for this target' (one star), 'Insufficient validation data to recommend' (two stars), 'Best available probe for this target or a high quality probe that is a useful orthogonal tool' (three stars), 'Recommended as a probe for this target' (four stars). Only probes with a SAB rating of three to four are endorsed as chemical probes by the experts to explore a biological system. Similarly, comments about the best experimental practice, such as the recommended use and concentration of the tool compound in cellular or animal system, are provided (Arrowsmith et al., 2015).

2. The Structural Genomics Consortium (SGC) (Jones et al., 2014; Structural Genomics Consortium, 2020) is a not-for-profit organization which is focused on generating knowledge and reagents that can be used in new drug discovery programs. This includes the determination of 3D protein structures on a large scale and leading the development of chemical probes. The identity of their chemical probes, together with all available experimental data, can be freely downloaded from their website. Their probes meet the following criteria: i*n vitro* IC_{50} or $K_d < 100$ nM, $>$ 30-fold selectivity over proteins in the same family and significant on-target cellular activity at 1 mM.

3. The problem in deriving selective compounds is the amount of missing experimental target-compound combination data. As drugs tend to be well-characterized in terms of their on- and off-targets, Gregori-Puigjane *et al.* used this data to classify drugs as selective and suggest further use as tool compounds. Using stringent criteria, they found 44 selective drugs for different target classes including the target class GPCR, which has relatively few tool compounds available. Hence, using drugs as tool compounds could broaden the application of currently availables tool compounds (Gregori-Puigjane et al., 2012). Gregori-Puigjane *et al.* reported no

kinase selective drugs; however, Wang *et al.* used their tool score to further derive selective kinase inhibitors using the data collected from drugs (Wang et al., 2016). Manual verification showed that at least two of these inhibitors, Gefinitib and Dasatinib, are annotated as 'historic compounds' at The Chemical Probes Portal; please see Table A.1, Appendix, for further information.

From these databases and the publication by Wang *et al.*, all compound-target combinations were extracted and manually verified in term of kinase target class and selectivity annotation in literature. SMILES were standardized using MolVS (Version 0.1.1) (Swain, 2019) and converted into canonical SMILES using RDkit (Version 2018.03.3) (Landrum, 2020). Duplicated entries between the collected data sets were defined as the same target-compound combination and subsequently discarded. As kinase targets can be annotated using a different nomenclature, kinase targets were standardized using the HGNC nomenclature based on the mapping by Fulle et al. (Fulle, 2020). In total, I collected 57 'selective kinase inhibitors' for 69 targets; they are listed with their respective targets (and target classes) in Table A.1 in the Appendix. The SGC and the CPP overlapped only in two selective compounds, namely NVS-PAK1-1 (kinase target: PAK1) and FM-381 (kinase target: JAK3). Both probes had a STAR rating of 4 for use in cells due to the strength of their associated publications. Interestingly, different selective compounds were available across databases to target the same protein, *e.g.* DDR1-IN-1 (database:CPP) and DDR-TRK-1 (database: SGC), which target DDR1, and SGC-CLK-1, newly available starting from January 2020, T3-CLK, MU1210 (SGC) and KH-CB19 (CCP), which target the CLK1 kinase. This highlights the research interest in specific kinases and the complementarity of the two databases. However, the SGC places a strict emphasis on experimental validation, thus increasing the probability of their probes being 'truly selective'. I also caution that missing or false experimental data can cause a potential chemical probe to be miss-classified as 'selective'.

The kinome space covered by the collected chemical probes is visualised in Figure 4.1.

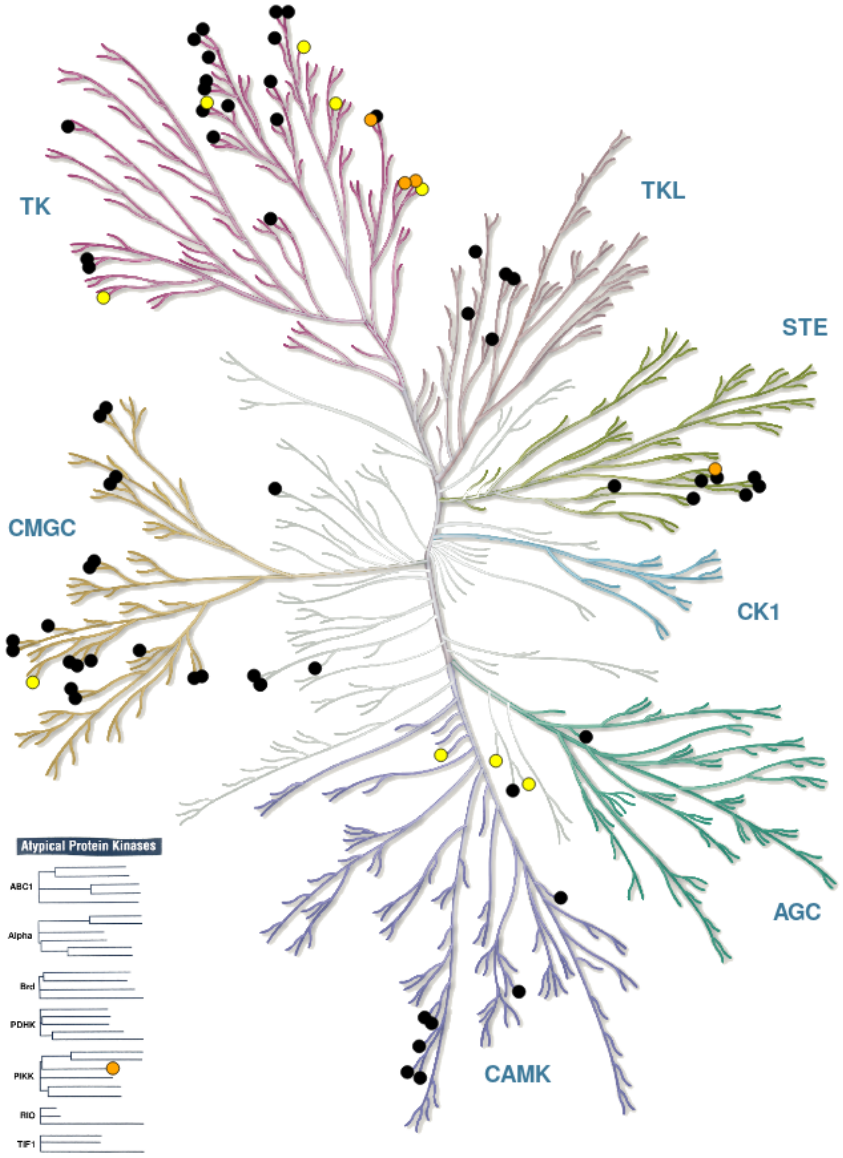


Figure 4.1: Kinase targets of selective kinase compounds collected from literature. From literature, 57 selective kinase inhibitors targeting 69 protein kinases were collected and duplicate target compound combinations removed using the xname nomenclature for the kinases and the canonical SMILES for the selective compounds. The targets of all these inhibitors are plotted according to the number of selective compounds found for each of them. Targets with one compound are annotated black, targets with two compounds are colored yellow and targets with three compounds orange. The only targets not represented on the kinome tree are BCR-ABL (selective compound: GNF-5) and BRAF V600E (selective compound: GDC-0879), see also Table A.1, in the Appendix, as their identifiers could not be resolved using the list of Human Kinases provided by Fulle *et al.*. However, the BCR-ABL gene, also called Philadelphia chromosome', is the result of a reciprocal chromosomal translocation between the ABL1 gene (on chromosome 9) and the breakpoint cluster region (BCR) gene (on chromosome 22). As a driver of uncontrolled cell division, it has been implicated in different types of leukemia (Rossari et al., 2018). The ABL1 gene itself is present on the kinome tree (TK kinase family). This illustration was reproduced courtesy of Cell Signaling Technology, Inc. (www.cellsignal.com).

4.2.2 Building The Random Forest Models

RF models were built for the kinase subclasses TK, TKL, AGC, CAMK, STE and CMGC. This was done with all available data using the conditions detailed in Section 3.5.5 to investigate the improvement when including more data. The number of data points are listed in Table 4.1. Again, the number of inactive data points always exceeds the number of active data points in a rough ratio of either 4:1 or 3:1, highlighting that the model could be biased towards predicting inactive data points. As detailed in Section 3.5.5, I considered the distribution of my data sets when building the models. The performance of the models was evaluated using a variety of evaluation metrics, as listed in Table 4.2.

For the combined data set of all kinases, I used the canonical SMILES of my data set of selective compounds to extract all active/inactive compound-target combinations for a selective compound. I also extracted all compound-target combinations for Gefitinib and Dasatinib as examples of historic compounds. A total of 19 chemical probes had no bioactivity information associated with them, but corresponding targets in my training data set. To get an understanding of how well my model predicts selective compounds for which I had no prior information in the data set, I manually constructed a validation set of these 19 selective compounds and their target(s). This test set was constructed as follows: the SMILES of the selective compounds (Table A.1) were standardized using MolVS (Version 0.1.1) (Swain, 2019), converted into their corresponding canonical SMILES using RDkit (Version 2018.03.3)(Landrum, 2020) and the MFPs (Radius:2, nbits:1024) were calculated for each compound. The amino acid descriptors for the targets of these selective compounds had already been calculated and were merged with the MFPs; these compound-target pairs were then annotated as 'active'.

Table 4.1: Number of the (active or inactive data) points, targets and compounds for each kinase group. For each model, I constructed a data set to be used for model building and training and a test set of selective compounds. Training and test set have no overlap. As the test set includes both active and inactive data points, a given selective compound can be present in test sets for multiple kinase groups.

Kinase	**Training set**					**Test set**				
Group	**# Targets**	**# Compounds**	**# Data Points**	**# Inactive**	**# Active**	**# Targets**	**# Compounds**	**# Data points**	**# Inactive**	**# Active**
All	261	46,857	265,888	205,061	60,827	245	30	2,256	2,108	148
AGC	34	6,204	36,259	30,801	5,458	32	22	327	321	6
CAMK	24	6,059	28,605	23,689	4916	23	15	237	231	6
CK1	9	2,058	7,208	6,636	572	9	21	74	73	1
GMGC	9	11,214	42,025	31,528	10,497	35	22	74	73	1
TK	62	21,186	94,533	68,413	26,120	62	26	328	314	14
TKL	24	5,132	12,418	9,230	3,188	22	21	161	143	18
STE	26	2,933	16,1397	14,265	1,874	24	21	195	190	5

4.3 Results

4.3.1 Models For Predicting Selectivity

The results of the RF models are listed in Table 4.2. For all my RF models for the different kinase data sets, I was able to achieve high AUC values in the range of 0.89-0.96 and high MCC values in the range of 0.7 for my different kinase models. Unsurprisingly, I achieve higher specificity than sensitivity rates, highlighting the influence of the greater number of inactive compound-target data points. In all cases, models using the 'log2' split performed slightly worse than models using the 'sqrt' split, while the number of trees did not have any influence on the performance. The high ROC AUC values around 0.95 indicates a good model; these values are in agreement with the values derived by Sorgenfrei *et al.* who reported average AUC values around 0.9 using a five-fold cross-validated model with their entire data set.

The discrepancy between their reported AUC values and mine can be explained as follows (a) Sorgenfrei *et al.* used a different data set of 305 kinases versus my data set of 261 kinases, and (b) I used different descriptors than Sorgenfrei *et al.*. While Sorgenfrei *et al.* used MFPs with a radius 2 and unspecified number of nbits and Z-scales 3 in their models, I used MFP descriptors with 1024 nbits and a radius of 2, with the kinase amino acid descriptors using Z-scale 5 values.

To account for potential overfitting, I also evaluated my PCM models using the MCC as a balanced measure of the performance of my models on all four classes of the confusion matrix. Using the best-performing models, the MCC values range from 0.63-0.75, indicating that the AUC values might be too optimistic. Unfortunately, Sorgenfrei *et al.* reported no MCC values.

The analysis of the relative feature importance for the different kinase models in Figure 4.2 shows that the amino acid descriptors have their highest contribution in a RF model trained on data from all kinase classes (almost 40%), highlighting the importance of the kinase amino acids in differentiating between the different kinases.

Table 4.2: Evaluation metrics of RF models for the different groups of kinases, trained on a data set excluding all selective compounds. For each RF model trained on the target, this table list the sensititivity, the specificity, the area under the curve (AUC), the precision and the Matthew Correlation Coefficient (MCC) as well as the Out-of-bag (OOB) performance and the Out-of-bag (OOB) error for a combination of different parameters.

Group	Parameter		Metrics							Features		# Data points	
	# Trees	Split	Sensitivity	Specificity	AUC	Precision	MCC	OOB	OOB error	# AAS	# MFP	Training	Testing
All	100	log2	0.79 +/- 0.002	0.92 +/- 0.002	0.93 +/- 0.001	0.74 +/- 0.006	0.70 +/- 0.004	0.89 +/- 0.001	0.11 +/- 0.001	610	1,024	265,888	2256
	100	sqrt	0.79 +/- 0.003	0.85 +/- 0.005	0.90 +/- 0.002	0.62 +/- 0.008	0.60 +/- 0.007	0.84 +/- 0.002	0.16 +/- 0.002	610	1,024	265,888	2256
	250	log2	0.79 +/- 0.004	0.86 +/- 0.002	0.90 +/- 0.002	0.62 +/- 0.003	0.60 +/- 0.004	0.84 +/- 0.001	0.16 +/- 0.001	610	1,024	265,888	2256
	250	sqrt	0.80 +/- 0.002	0.92 +/- 0.002	0.93 +/- 0.001	0.74 +/- 0.005	0.70 +/- 0.003	0.89 +/- 0.000	0.11 +/- 0.000	610	1,024	265,888	2256
	550	log2	0.79 +/- 0.004	0.86 +/- 0.002	0.90 +/- 0.001	0.62 +/- 0.004	0.60 +/- 0.005	0.84 +/- 0.001	0.16 +/- 0.001	610	1,024	265,888	2256
	550	sqrt	0.80 +/- 0.001	0.92 +/- 0.002	0.93 +/- 0.001	0.74 +/- 0.004	0.70 +/- 0.003	0.89 +/- 0.000	0.11 +/- 0.000	610	1,024	265,888	2256
	1000	log2	0.79 +/- 0.003	0.86 +/- 0.002	0.90 +/- 0.001	0.62 +/- 0.003	0.60 +/- 0.004	0.84 +/- 0.001	0.16 +/- 0.001	610	1,024	265,888	2256
	1000	sqrt	0.80 +/- 0.002	0.92 +/- 0.002	0.93 +/- 0.001	0.74 +/- 0.004	0.70 +/- 0.003	0.89 +/- 0.001	0.11 +/- 0.001	610	1,024	265,888	2256
CAMK	100	log2	0.73 +/- 0.011	0.97 +/- 0.003	0.95 +/- 0.004	0.83 +/- 0.015	0.75 +/- 0.014	0.94 +/- 0.001	0.06 +/- 0.001	610	1,024	28,605	237
	100	sqrt	0.72 +/- 0.015	0.95 +/- 0.003	0.93 +/- 0.006	0.72 +/- 0.014	0.67 +/- 0.017	0.91 +/- 0.002	0.09 +/- 0.002	610	1,024	28,605	237
	250	log2	0.72 +/- 0.012	0.95 +/- 0.004	0.93 +/- 0.006	0.72 +/- 0.018	0.67 +/- 0.018	0.91 +/- 0.001	0.09 +/- 0.001	610	1,024	28,605	237
	250	sqrt	0.74 +/- 0.008	0.97 +/- 0.003	0.95 +/- 0.004	0.83 +/- 0.015	0.75 +/- 0.013	0.94 +/- 0.001	0.06 +/- 0.001	610	1,024	28,605	237
	500	log2	0.72 +/- 0.013	0.95 +/- 0.003	0.93 +/- 0.006	0.72 +/- 0.014	0.67 +/- 0.016	0.92 +/- 0.001	0.08 +/- 0.001	610	1024	28,605	237
	500	sqrt	0.74 +/- 0.010	0.97 +/- 0.002	0.95 +/- 0.004	0.83 +/- 0.014	0.75 +/- 0.012	0.94 +/- 0.001	0.06 +/- 0.001	610	1024	28,605	237
	1000	log2	0.72 +/- 0.014	0.95 +/- 0.003	0.93 +/- 0.006	0.73 +/- 0.015	0.67 +/- 0.017	0.92 +/- 0.001	0.08 +/- 0.001	610	1024	28,605	237
	1000	sqrt	0.74 +/- 0.011	0.97 +/- 0.002	0.95 +/- 0.004	0.83 +/- 0.014	0.75 +/- 0.012	0.94 +/- 0.001	0.06 +/- 0.001	610	1024	28,605	237
AGC	100	sqrt	0.80 +/- 0.008	0.93 +/- 0.005	0.93 +/- 0.004	0.75 +/- 0.014	0.71 +/- 0.010	0.90 +/- 0.002	0.10 +/- 0.002	610	1,024	36,259	327
	100	log2	0.80 +/- 0.010	0.95 +/- 0.004	0.95 +/- 0.004	0.80 +/- 0.013	0.75 +/- 0.007	0.92 +/- 0.001	0.08 +/- 0.001	610	1,024	36,259	327
	250	sqrt	0.80 +/- 0.009	0.95 +/- 0.004	0.95 +/- 0.004	0.81 +/- 0.013	0.75 +/- 0.006	0.92 +/- 0.001	0.08 +/- 0.001	610	1,024	36,259	327
	250	log2	0.80 +/- 0.006	0.93 +/- 0.005	0.93 +/- 0.005	0.75 +/- 0.015	0.72 +/- 0.011	0.90 +/- 0.002	0.10 +/- 0.002	610	1,024	36,259	327
	500	sqrt	0.80 +/- 0.009	0.95 +/- 0.005	0.95 +/- 0.004	0.81 +/- 0.014	0.76 +/- 0.005	0.92 +/- 0.001	0.08 +/- 0.001	610	1,024	36,259	327
	500	log2	0.80 +/- 0.008	0.93 +/- 0.005	0.93 +/- 0.004	0.75 +/- 0.013	0.72 +/- 0.009	0.91 +/- 0.002	0.09 +/- 0.002	610	1,024	36,259	327
	1000	sqrt	0.80 +/- 0.009	0.95 +/- 0.005	0.95 +/- 0.004	0.81 +/- 0.014	0.76 +/- 0.004	0.92 +/- 0.001	0.08 +/- 0.001	610	1,024	36,259	327
	1000	log2	0.80 +/- 0.006	0.93 +/- 0.004	0.93 +/- 0.004	0.75 +/- 0.012	0.72 +/- 0.009	0.91 +/- 0.002	0.09 +/- 0.002	610	1,024	36,259	327
CK1	100	sqrt	0.59 +/- 0.055	0.97 +/- 0.004	0.89 +/- 0.014	0.61 +/- 0.028	0.57 +/- 0.038	0.93 +/- 0.003	0.07 +/- 0.003	610	1,024	7,208	74
	100	log2	0.59 +/- 0.035	0.98 +/- 0.002	0.90 +/- 0.013	0.76 +/- 0.022	0.64 +/- 0.027	0.95 +/- 0.002	0.05 +/- 0.002	610	,1024	7,208	74
	250	sqrt	0.59 +/- 0.018	0.98 +/- 0.003	0.91 +/- 0.013	0.77 +/- 0.034	0.65 +/- 0.025	0.95 +/- 0.002	0.05 +/- 0.002	610	1,024	7,208	74
	250	log2	0.59 +/- 0.037	0.97 +/- 0.002	0.89 +/- 0.014	0.63 +/- 0.009	0.58 +/- 0.021	0.94 +/- 0.003	0.06 +/- 0.003	610	1,024	7,208	74
	550	sqrt	0.58 +/- 0.022	0.98 +/- 0.002	0.91 +/- 0.013	0.75 +/- 0.026	0.63 +/- 0.021	0.95 +/- 0.002	0.05 +/- 0.002	610	1,024	7208	74
	550	log2	0.60 +/- 0.036	0.97 +/- 0.002	0.89 +/- 0.013	0.64 +/- 0.011	0.58 +/- 0.019	0.94 +/- 0.003	0.06 +/- 0.003	610	1,024	7,208	74
	1000	sqrt	0.58 +/- 0.025	0.98 +/- 0.001	0.91 +/- 0.014	0.75 +/- 0.022	0.64 +/- 0.024	0.95 +/- 0.002	0.05 +/- 0.002	610	1,024	7,208	74
	1000	log2	0.60 +/- 0.030	0.97 +/- 0.004	0.89 +/- 0.013	0.64 +/- 0.028	0.58 +/- 0.017	0.94 +/- 0.003	0.06 +/- 0.003	610	1,024	7,208	74

continued

Group	Parameter		Metrics							Features		# Data points	
	# Trees	Split	Sensitivity	Specificity	AUC	Precision	MCC	OOB	OOB error	# AAS	# MFP	Training	Testing
TKL	100	log2	0.76 +/- 0.010	0.96 +/- 0.007	0.94 +/- 0.004	0.87 +/- 0.020	0.75 +/- 0.017	0.91 +/- 0.001	0.09 +/- 0.001	610	1024	12,418	161
	100	sqrt	0.75 +/- 0.008	0.95 +/- 0.009	0.92 +/- 0.007	0.85 +/- 0.025	0.73 +/- 0.017	0.90 +/- 0.001	0.10 +/- 0.001	610	1024	12,418	161
	250	log2	0.75 +/- 0.010	0.95 +/- 0.009	0.92 +/- 0.006	0.85 +/- 0.025	0.74 +/- 0.020	0.90 +/- 0.002	0.10 +/- 0.002	610	1024	12,418	161
	250	sqrt	0.76 +/- 0.009	0.96 +/- 0.007	0.95 +/- 0.003	0.87 +/- 0.021	0.75 +/- 0.018	0.91 +/- 0.001	0.09 +/- 0.001	610	1024	12,418	161
	500	log2	0.75 +/- 0.009	0.95 +/- 0.009	0.92 +/- 0.006	0.84 +/- 0.026	0.73 +/- 0.019	0.90 +/- 0.002	0.10 +/- 0.002	610	1024	12,418	161
	500	sqrt	0.76 +/- 0.012	0.96 +/- 0.007	0.95 +/- 0.003	0.87 +/- 0.021	0.75 +/- 0.020	0.91 +/- 0.001	0.09 +/- 0.001	610	1024	12,418	161
	1000	log2	0.75 +/- 0.010	0.95 +/- 0.009	0.92 +/- 0.006	0.84 +/- 0.025	0.73 +/- 0.018	0.90 +/- 0.001	0.10 +/- 0.001	610	1024	12418	161
	1000	sqrt	0.76 +/- 0.010	0.96 +/- 0.007	0.95 +/- 0.003	0.87 +/- 0.021	0.75 +/- 0.019	0.91 +/- 0.001	0.09 +/- 0.001	610	1024	12,418	161
TK	100	log2	0.81 +/- 0.005	0.93 +/- 0.004	0.95 +/- 0.001	0.81 +/- 0.007	0.73 +/- 0.006	0.89 +/- 0.001	0.11 +/- 0.001	610	1024	94,533	328
	100	sqrt	0.79 +/- 0.006	0.89 +/- 0.003	0.92 +/- 0.001	0.74 +/- 0.004	0.67 +/- 0.001	0.86 +/- 0.002	0.14 +/- 0.002	610	1024	94,533	328
	250	log2	0.79 +/- 0.007	0.89 +/- 0.004	0.92 +/- 0.002	0.74 +/- 0.005	0.67 +/- 0.002	0.86 +/- 0.001	0.14 +/- 0.001	610	1024	94,533	328
	250	sqrt	0.81 +/- 0.005	0.93 +/- 0.003	0.95 +/- 0.001	0.81 +/- 0.007	0.74 +/- 0.005	0.89 +/- 0.001	0.11 +/- 0.001	610	1024	94,533	328
	500	log2	0.79 +/- 0.006	0.89 +/- 0.003	0.92 +/- 0.002	0.74 +/- 0.004	0.67 +/- 0.002	0.87 +/- 0.001	0.13 +/- 0.001	610	1024	94,533	328
	500	sqrt	0.81 +/- 0.003	0.93 +/- 0.003	0.95 +/- 0.001	0.81 +/- 0.007	0.74 +/- 0.005	0.89 +/- 0.001	0.11 +/- 0.001	610	1024	94,533	328
	1000	log2	0.79 +/- 0.006	0.89 +/- 0.003	0.92 +/- 0.002	0.74 +/- 0.004	0.67 +/- 0.003	0.87 +/- 0.001	0.13 +/- 0.001	610	1024	94,533	328
	1000	sqrt	0.81 +/- 0.003	0.93 +/- 0.004	0.95 +/- 0.001	0.81 +/- 0.008	0.74 +/- 0.005	0.89 +/- 0.001	0.11 +/- 0.001	610	1024	94,533	328
STE	100	log2	0.62 +/- 0.016	0.97 +/- 0.004	0.93 +/- 0.007	0.72 +/- 0.028	0.63 +/- 0.023	0.93 +/- 0.003	0.07 +/- 0.003	610	1024	161,397	195
	100	sqrt	0.62 +/- 0.012	0.94 +/- 0.006	0.90 +/- 0.004	0.57 +/- 0.021	0.54 +/- 0.013	0.90 +/- 0.002	0.10 +/- 0.002	610	1024	161,397	195
	250	log2	0.62 +/- 0.012	0.94 +/- 0.004	0.90 +/- 0.005	0.57 +/- 0.017	0.54 +/- 0.012	0.90 +/- 0.002	0.10 +/- 0.002	610	1024	161,397	195
	250	sqrt	0.62 +/- 0.021	0.97 +/- 0.004	0.93 +/- 0.007	0.72 +/- 0.033	0.63 +/- 0.028	0.93 +/- 0.003	0.07 +/- 0.003	610	1024	161,397	195
	500	log2	0.63 +/- 0.012	0.94 +/- 0.004	0.91 +/- 0.005	0.58 +/- 0.016	0.55 +/- 0.010	0.90 +/- 0.002	0.10 +/- 0.002	610	1024	161397	195
	500	sqrt	0.62 +/- 0.019	0.97 +/- 0.004	0.94 +/- 0.007	0.73 +/- 0.027	0.63 +/- 0.021	0.93 +/- 0.003	0.07 +/- 0.003	610	1024	161,397	195
	1000	log2	0.63 +/- 0.014	0.94 +/- 0.004	0.91 +/- 0.005	0.58 +/- 0.019	0.55 +/- 0.017	0.90 +/- 0.003	0.10 +/- 0.003	610	1024	161,397	195
	1000	sqrt	0.63 +/- 0.023	0.97 +/- 0.004	0.94 +/- 0.007	0.73 +/- 0.025	0.64 +/- 0.021	0.93 +/- 0.002	0.07 +/- 0.002	610	1024	161,397	195

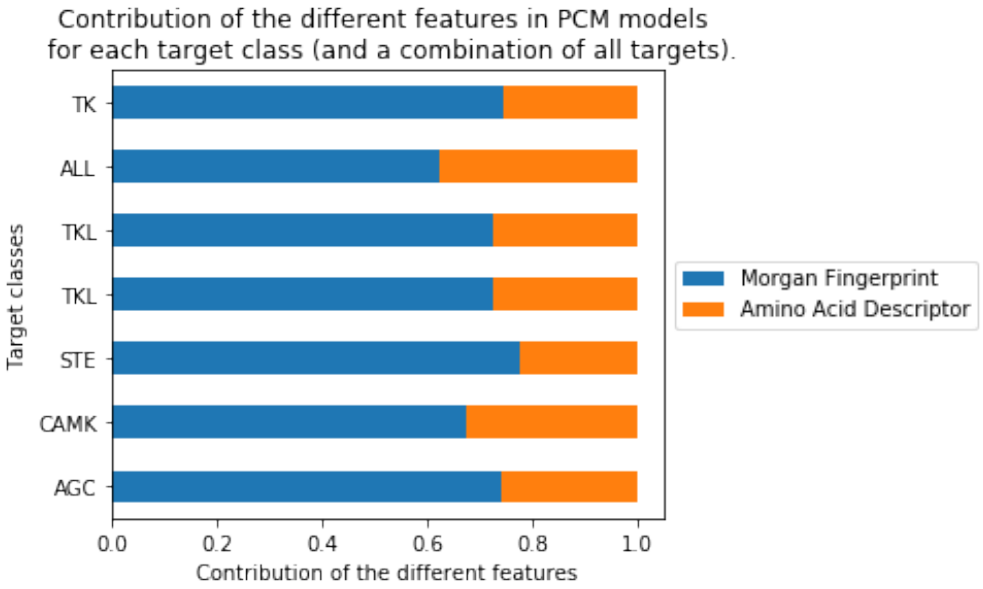


Figure 4.2: The relative feature importance for the Morgan Fingerprints and the amino acid descriptors across models built for various target classes and all kinase target classes ('ALL'). The importance of features for each kinase model was summed up to 1 and plotted against the respective subclasses.

I used the best performing RF model for all kinases in Table 4.2 (with 1000 trees and the square root of the number of features chosen for each decision tree as the best criterion for the best split of features) to predict the activity/inactivity for each compound-target pair in my test set of literature-derived selective compounds (Section 4.2.1). The predictions were then compared with the available experimental data to assess how well selective kinase compounds can be predicted.

Table 4.3 highlights a major caveat in identifying selective compounds: HG-9-91-01, an ATP-competitive inhibitor, has been rated with a SABS rating of three in the Chemical Probes database and is thus recommended as a probe in cells against the SIK1, SIK2 and SIK3 kinases (Chemical Probe Portals Reviewers, 2020) (with reported IC_{50} values of 0.92 nM, 6.6 nM and 9.6 nM for SIK1, SIK2 and SIK3 (MedChemExpress, 2020) respectively) within a concentration range of 100-500 nM. However, due to the limited selectivity data available, it was cautioned against relying on the results of this probe due to unexpected off-target effects. Since the SIK targets (SIK1, SIK2, SIK3) are not included in the data set by Fulle *et al.*, I had no examples for these targets in my training set to verify the prediction of HG-9-91-01, against them. Similarly, for example, due

to missing data, I could not test the prediction of the compound GDC-0879 against the kinase target BRAF V600E, a version of the serine/threonine protein kinase BRAF mutated by substituting glutamic acid for valine at position 600 (Ascierto et al., 2012).

As shown, HG-9-91-01 has been correctly predicted to be active against the TK targets BRK and BTK; in the literature $IC_{50} < 80$ nM and IC_{50} values of < 60 nM have been reported for BTK and BRK respectively (Kinexus BioInformatics Corporation, 2019b,a). As these values are within the recommended range of use in Chemical Probes, this compound would be classified as non-selective within that range. HG-9-91-01 is thus selective within different bioactivity ranges.

However, my current model has an activity cut-off of 6.3 pIC_{50}-units (corresponding to an IC_{50}-value of 500 nM); a threshold I used to identify both potential actives and weaker off-target interactions (Sorgenfrei et al., 2018). As my binary classification cannot predict the bioactivity of compound-target combinations within different bioactivity ranges, a multi-classification model with different bioactivity ranges would have been more appropriate for predicting selective compounds within different bioactivity ranges and will be discussed later.

Table 4.4 shows the number of correctly and incorrectly predicted target-compound combinations, targets, and three measures for the success of my model: the sensitivity and specificity for correctly predicted active and inactive compounds, and the overall prediction rate as a measure of how successful my model was overall in predicting the activity of a compound. Overall, my model successfully predicted the activity of selective compounds in the test set against all its targets present in the data set. However, without bioactivity information for the compounds CCT244747, CGI1746, CCT251545 and NVS-PAK1-1, these prediction can not be verified as indicated by the "-" symbol in the "Specificity" column in Table 4.4.

To test how well my model predicts previously unseen compounds known in literature, I manually constructed a second validation set for 19 selective compounds not found in the data set by Fulle *et al.*, see Section 4.2 for more details. My RF model correctly predicted their activity against their targets for 11 compounds (57.9%) and incorrectly for 8 compounds (42.1%) as shown in Table 4.5. This highlights that despite good prediction rates for previously seen compounds, my model fails at predicting the activity of unseen

Table 4.3: Comparison of the actual and the predicted activity of target-compound combinations for selective and two 'historic' compounds. To illustrate the performance of a kinase RF model in predicting selective compounds, I plotted for each selective compound the performance of predicting its activity against its target, if data was available in the data set. Each compound is listed with its commonly used compound identifiers and its target(s).
The performance metrics are noted within the respective cells; additionally, the cells are color-coded according. Green squares indicate a correctly predicted active target/compound combination (TP); blue a correctly predicted inactive target/compound combination (TN). Similarly, orange and yellow denote false actives (FP) or false inactives (FN). Lastly, grey denotes a target-compound pair with no known bioactivity within my data set. The two 'historic' compounds Gefinitib and Dasatinib are denoted in blue in the 'Selective Compound' column.

Selective Compound	Kinase Target(s)	ABL1	ALK	BRK	BTK	CDK8	EGFR	FGFR1	FGFR2	FGFR3	FGFR4	FLT3	GSK3B	IGF1R	INSR	JAK1	JAK2	MAP2K1	MET	PAK1
AC220	FLT3	FP	TN	TN	TN	TN	TN	TN	TN	TN	TN	TP	TN	TN	TN		FP	TN	FP	TN
BGJ-398	FGFR1, FGFR2, FGFR3, FGFR4	FP	TN		TN		FP	TP	TP	TP	TP	FP	TN	TN	TN		FP		TN	
BMS-911543	JAK2															TP	TP			
CCT244747	CHEK1																			
CCT251545	CDK8, CDK19												TN							
CGI1746	BTK				TP															
CH5424802	ALK		TP																	
CHIR-99021	GSK3A, GSK3B	TN	TN	TN			TN	TN		TN		TN	TP	TN	TN		FP		TN	TN
Dasatinib	ABL1	TP	FP	FN	TP	TN	FN	TN	TN	TN	TN	FP	TN	TN	TN		FP	TN	FP	TN
Erlotinib	EGFR	FN	TN	TN	TN	TN	TP	TN	TN	TN	TN	TN	TN	TN	TN		TN	TN	TN	TN
Filgotinib	JAK1											TP				TP	FP			
GNE7915	LRRK2	TN		TN	TN			TN					TN	FP	FN			TN		
Gefitinib	EGFR	FP	TN	TN	TN	TN	TP	TN	TN	TN	TN	FP	TN	TN	TN		TN	TN	TN	TN
HG-9-91-01	SIK1, SIK2, SIK3	TP		FN	TP			TP					TN		TN			TN		
Lapatinib	EGFR	TN	TN	TN	TN	TN	TP	TN	TN	TN	TN	TN	TN	TN	TN		TN	TN	TN	TN
Linsitinib	IGF1R, INSR, INSRR	TN	TN		TN		TN	TN		TN		TN	TN	TP	TP		FP		TN	TN
NVS-PAK1-1	PAK1, PAK1																			FN
Nilotinib	ABL1	TP	TN	TN	TN	TN	TN	TN	TN	TN	TN	FP	TN	TN	TN		FP	TN	TN	TN
SGX-523	MET	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	FP	TN	TN	TN		FP	TN	TP	TN
Selumetinib	MAP2K1	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	FP	TN	TN	TN		FP	FN	TN	TN
VX-745	MAPK14	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN		TN	TN	TN	TN

Table 4.4: Predicted activity of a selective compounds against all targets in the data set. For each selective compound, I predicted its activity against all targets. If experimental target-compound information was available, I used this information to verify my prediction for this target-compound pair. For each compound, I calculated the number of false positive (FP), false negative (FN), true positive (TP) and true negative (TN), its number of correctly and incorrectly predicted targets (# Correct targets,#Incorrect Targets), its sensitivity, specificity and overall prediction rate. Missing bioactivity information is indicated by the "-" symbol.

Target(s)	Selective Compound	FN	FP	TN	TP	#targets	# Correct Targets	# Incorrect Targets	Overall Prediction	Sensitivity	Specificity
FLT3	AC220	0	14	200	9	223	209	14	0.94	1	0.93
FGFR1, FGFR2, FGFR3, FGFR4	BGJ-398	0	15	40	4	59	44	15	0.75	1	0.73
JAK2	BMS-911543	0	0	0	4	4	4	0	1.0	1	-
CHEK1	CCT244747	0	0	0	1	1	1	0	1.0	1	-
CDK8, CDK19	CCT251545	0	0	1	0	1	1	0	1.00	-	1
BTK	CGI1746	0	0	0	1	1	1	0	1.00	1	-
ALK	CH5424802	0	1	0	1	2	1	1	0.50	1	0
GSK3A, GSK3B	CHIR-99021	0	2	93	2	97	95	2	0.98	1	0.98
ABL1	Dasatinib	33	19	166	15	233	181	52	0.78	0.31	0.9
EGFR	Erlotinib	5	3	217	3	228	220	8	0.96	0.38	0.99
JAK1	Filgotinib	0	2	1	3	6	4	2	0.67	1	0.33
LRRK2	GNE7915	3	8	75	0	86	75	11	0.872	0	0.9
EGFR	Gefitinib	2	0	83	2	87	85	2	0.978	0.5	1
SIK1, SIK2, SIK3	HG-9-91-01	12	4	60	5	81	65	16	0.80	0.29	0.94
EGFR	Lapatinib	0	0	226	3	229	229	0	1.00	1	1
IGF1R, INSR, INSRR	Linsitinib	1	2	92	3	98	95	3	0.97	0.75	0.98
PAK1, PAK1	NVS-PAK1-1	1	0	0	0	1	0	1	0.00	0	-
ABL1	Nilotinib	12	10	170	10	202	180	22	0.89	0.45	0.94
MET	SGX-523	0	11	122	1	134	123	11	0.92	1	0.92
MAP2K1	Selumetinib	1	12	121	0	134	121	13	0.90	0	0.91
MAPK14	VX-745	3	3	201	1	208	202	6	0.97	0.25	0.99

selective compounds in 40% of the cases. This illuminates that the chemical space for these compounds is potentially not well-explored. However, unsurprisingly, compounds with related targets, *e.g.* Tofacitinib against the JAK kinases or G-5555 against the PAK kinases were predicted either all correctly or all incorrectly. These results highlight that closely-related protein targets might use the same amino acid residues for binding, as exemplified in Figures (5.8) -(5.7) for the selective compound CP-690550. This observation has been referred to as the 'inverse similarity principle' - related proteins are likely to interact with similar compounds (Cortés-Ciriano et al., 2015). If multiple targets are more similar in their key binding moieties and experimental evidence of binding for one compound-target pair is available, then this information can be extrapolate the potential for binding of said compound to related target. Subsequently, in order to predict selectivity, the information from the target space and the chemical space have to be equally considered. Table 4.5 thus highlights that not (or no) sufficient binding data of chemically similar compounds against the target(s) of interest or closely related targets was available for the classifier to correctly learn and predict the activity of the selective compound in question.

Table 4.5: Predictions for selective compounds not present in the data set provided by Fulle *et al.* For 19 selective compounds not present in the data, corresponding targets were found. To test whether the available information could help to identify the activity of unseen compounds, these compounds were manually annotated as active and used as a validation set in the model.

Target	Compound Identifier	Predicted Value
FGFR4	BLU9931	Incorrect
JAK3	FM-381	Correct
PAK1	G-5555	Incorrect
PAK3	G-5555	Incorrect
PDK1	GSK2334470	Correct
RIPK2	GSK583	Incorrect
CLK1	KH-CB19	Incorrect
CLK4	KH-CB19	Incorrect
DYRK1A	KH-CB19	Incorrect
IGF1R	NVP-AEW541	Correct
PAK1	NVS-PAK1-1	Incorrect
SYK	P505-15	Correct
CDK4	PD0332991	Correct
CDK6	PD0332991	Correct
PAK1	PF-3758309	Correct
PAK4	PF-3758309	Correct
PAK5	PF-3758309	Correct
PAK6	PF-3758309	Correct
MAPKAPK2	PF3644022	Correct
JAK1	Ruxolitinib	Correct
JAK2	Ruxolitinib	Correct
BIKE	SGC-AAK1-1	Incorrect
GAK	SGC-GAK-1	Correct
LIMK1	TH-257	Incorrect
LIMK2	TH-257	Correct
CDK7	THZ1	Incorrect
JAK1	Tofacitinib	Correct
JAK2	Tofacitinib	Correct
JAK3	Tofacitinib	Correct
FLT3	UNC2025	Correct

4.4 Conclusions and Future Directions

Using my models, I successfully predicted active and inactive compound-target combinations for the majority of the literature-derived selective kinase inhibitors. However, my models highlight the following problems in terms of selectivity prediction:

- As shown by the grey cells in Table 4.3, the bioactivity values for some of my chosen selective compounds were missing in the collected data set by Fulle *et al.*, highlighting that the publicly available data sets for selective kinase compounds are lacking specific compound-target combinations or data found in other databases or online sources. For example, while the chemical probe HG-9-91-01 has the PubChem ID 78357808, I was unable to find a ChEMBL identifier for HG-9-91-01, highlighting that bioactivity data is often dispersed in different data portals with different annotations.

- Similarly, the prediction of HG-9-91-01 highlights the importance of predicting the target-compound bioactivity. HG-9-91-01 is selective towards the SIK targets in the concentration range of 10 nM; outside this range it also binds to other protein kinases such as BTK and BRK, in the concentration range of 100 nM (Chemical Probe Portals Reviewers, 2020; Kinexus BioInformatics Corporation, 2019b,a). However, the Chemical Probes Portal recommended to use HG-9-91-01 as a selective probe in the range 100-500 nM against the SIK kinases. Ideally, a random forest model should be able to predict the target-compound pairs in a fine-grained range of bioactivity values. My data is composed of different data types and also biased towards inactive and medium-active compound-target combinations. While I considered building a multiclassification model consisting of three categories ('active','medium-active' and 'inactive'), ultimately, I decided against it due to the low number of active target-compound examples for the active bins.

Future work could include a similarity analysis of selective compounds and the other compounds in the training set, to explore how selectivity prediction is limited by the explored chemical space. However, in my opinion, using 2D RF models to derive selective compounds is also limited by the distribution of the available bioactivity data. Ideally a broad distribution across different bioactivity ranges is needed to infer the selectivity of a compound at different concentration thresholds. For example, using my models, I

was able to identify the non-selectivity, i.e. promiscuity, of the chemical probe HG-9-91-01 within the bioactivity range corresponding to an IC_{50}-value of 500 nM, which was annotated as selective in the Chemical Probes portal, but which my model highlighted as 'not selective'.

To circumvent the problem of missing compound-target activity data and lack of bioactivity values for given chemotypes in random forest models, I investigated the use of 3D-structural models based on protein-ligand interactions as described in Chapter 5.

5 Using 3D-Kinase Models to Classify Candidate compounds in Different Binding Affinity Ranges

5.1 Introduction

In this chapter, I describe an exploratory investigation into two strategies by which intermolecular features derived from protein-ligand complexes can be used for identifying selective or filtering out promiscuous compounds.

Firstly, candidate compounds might be classified as selective or promiscuous using a structure-based filter which incorporates known interactions of selective or promiscuous ligands, using, for example, the non-superpositional, ligand-structure-based method Ligity (Ebejer et al., 2019). I performed an exploratory comparison of the binding shapes and interactions of promiscuous compounds and selective compounds, using Staurosporine (STU) and Tofacitinib (MI1) as case studies, to test whether such methods might be useful in finding selective or promiscuous kinase compounds. Additionally, I used theshape-chemical feature overlap comparison method SuCOS (Leung et al.) to compare shape-feature similarity between the same ligand bound to different proteins, and show that ligands tend to have a conserved binding mode across kinase targets. The comparison of binding interactions for selective and promiscuous compounds shows no clear trend of distinguishing features, highlighting that ligand-based models would need to be built on a case-by-case basis.

Secondly, candidate compounds might be classified into different binding affinity ranges using known protein-ligand interactions; their predicted binding affinities can be used to classify a compound as selective or promiscuous. I show that binding interactions alone cannot be used to separate ligands into strong or weak binders at different thresholds; this can be explained by the small amount of structural data available, the lack of specific key interactions, or the need to count the number of interactions at given parts of a kinase subpocket.

Thirdly, I investigated the use of the so-called CRANkS (Compound Ranking for Active Novel Scaffolds) interaction grids, 3D crystal-derived interaction maps, developed by Patel *et al.*, to see if they could be used to classify candidate compound into different binding affinity ranges (Patel, 2018). Protein-ligand X-ray crystal structures are used to

generate grids which describe the spatial distribution of the features of the ligand or the protein-ligand interactions. Then, all conformers of the putative binders are generated and scored based on their overlap with the CRANkS interaction grids. Use of these grids would allow the classification of compounds based solely on whether they fulfill known binding interactions with a given kinase, independently on any chemical features, and with immediate visualization of key binding features. My test set of cross-docked and re-docked ligands managed to pick up the respective interactions in the interaction grids; however, future work is needed to make the CRANkS maps a valuable tool for selectivity or promiscuity prediction.

5.2 Methods

5.2.1 Annotation of Each Kinase PDB Structure with Binding Affinity Data

The KLIFS database was downloaded on 14.10.2019; only kinases with the DFG-in motif and the αC helix in were extracted. Each PDB structure was annotated with its corresponding bioactivity using the PDBbind Database version 2018, which provides curated experimental binding affinity data for 16,151 protein-ligand complexes found in the PDB. These binding affinity data consist of the dissociation constant K_d, inhibition constant K_i and the half-maximal inhibitory concentration IC_{50}. If multiple binding affinity values are available for the same protein-ligand complex, the complex is annotated with the following priority: K_i values over K_d over IC_{50} values. Following Boyles *et al.*, the pK_i, pK_d and pIC_{50} values was used interchangeably (Boyles et al., 2019). If multiple annotated PDB structures for the same ligand-target complex were available, the complex with the best resolution was selected.

5.2.2 Calculation of the Intermolecular Interactions between a Ligand and its Kinase Target

Interactions between a protein and a bound ligand depends on their specific pairwise atomic interactions, which can be classified on the basis of atom type, distance and angle constraints. Both PLIP (Salentin et al., 2015) and Arpeggio (Jubb et al., 2016) are freely available tools to characterize protein-ligand interactions on an atomic level. To analyse the kinase-ligand interactions, Arpeggio (version 1.9.2) was used due to the wide range of calculated interactions and ease of use. The structural data was

prepared using the provided script clean_pdb.py (Jubb et al., 2016). The interaction types calculated by Arpeggio were as follows: hydrogen bond, weak hydrogen bond, halogen bond, metal complex, ionic, aromatic, hydrophobic, polar and weak polar (Jubb et al., 2016) To enable a comparison across different crystal structures in Section 5.3.2, each binding residue was standardized using the 85 binding pocket residue positions provided by KLIFS for each binding pocket. The extracted interactions were separated into 'backbone' and 'sidechain' interactions by denoting all interactions involving atoms labelled 'N', 'C', 'CA' and 'O' as 'backbone atoms' and all other atoms as 'sidechain atoms'. The type of interaction for each residue was counted and assigned as either a backbone or a sidechain interaction; to avoid double counting, each interaction was only counted once if the same interaction was reported multiple times for the same residue.

5.2.3 Using Random Forest Models to Test Whether Ligand-Kinase Interactions Can Be Used to Discriminate between Binders in Different Binding Affinity Ranges

After each PDB file was annotated with its binding affinity value, each compound-target complex was annotated with either 'weak binder' or 'strong binder', if the binding affinity was less or more than 8 pK value (10 nM, typically used to denote an active compound), 7 pK value (100 nM) or 6.3 pK value, corresping to 500 nM, the previous threshold for active and nonactive compounds (Sorgenfrei et al., 2018). The resulting categories were used as labels.

A total of 18 targets with at least 18 crystal structures were annotated with their binding affinity values and are listed in Table 5.1. This low number of observation was chosen as the majority of kinase targets could not be associated with a high amount of annotated crystal structures, shown in Figure A.9, in the Appendix. The distribution of the binding affinity values (in terms of their pK values) for each target is shown in Figure A.10, in the Appendix.

The interactions were calculated using Arpeggio as described previously; the presence or absence of a given compound-target interaction was annotated with 1 or 0, respectively, and used as input features for the RF models. Features were labeled as the concatenation of the specific KLIFS binding residue and the kind of interaction it mediates, e.g. 'aromatic_residue21', and the label for the binding affinity range was used as attribute

to be predicted by the classifier.

A 5-fold RF classifier was trained with 80% of the data for training and 20% for testing. The *StratifiedShuffleSplit* function from sklearn.model_selection (scikit-learn version 0.22.1) was used to preserve the distribution of the data in five reshuffled samples. The seed for the random number generator was set to 123. The minimum number of instances in leaves was set to 3 using the function *min_samples_split*; the criterion for determining the best splits was the 'gini' impurity. Performance was measured using MCC, accuracy, F_1 Score and precision-recall (set to 'average'). I set the *class_weight* to 'balanced' to account for the class imbalance, see Table 5.1-Table 5.3.

5.2.4 Comparison of the Shape and Binding Interactions of a Promiscuous and a Selective Compound

To compare the ligand shape of the same ligand bound across different kinase targets, all available X-ray structures of kinases with the conformation DFG-in, αC helix in were filtered for ligands binding to more than five different kinase targets. If multiple PDB files for the same target-ligand combination were available, the PDB structure with the best resolution was selected. For each structure, the binding pocket in which the ligand appears was extracted. Similarly, the selective compound Tofacitinib or CP-690550 (PDB residue name MI1) was selected as an example for a selective ligand.
The SuCOS score (Leung et al.) was used to calculate the binding shape between two ligands; it is calculated partly from both the shape and the chemical features overlap between two ligands aligned in their binding pockets. The structures were pairwise aligned using PyMOL's *align* function (setting *cycles* as 20), using PyMOL version 1.8 (Schrödinger). To estimate the influence of the target alignment on the calculation of the SuCOS score, two alignment were calculated for each ligand-kinase pair with each target being the reference protein in the alignment; the lowest SuCOS score was then extracted (see Figure 5.3 - 5.4). The RMSD values are shown in Figure A.12, Appendix. SuCOS scores from alignments with RMSD values equal or greater than 1 Å were excluded from the analyses. This value was arbitrarily chosen to guarantee a very conservative threshold, especially given the high sequence similarity of the kinase proteins. It is based on the fact that 2 Åis the resolution for a high-quality structure and has been used as a

measure whether docking has been successful for a given protein-ligand complex (Leung et al.) and whether the homology modelling has been successful for a sequence similarity lower than 30–50% (Kryshtafovych and Fidelis, 2009).

Arpeggio was used to calculate nine interaction types for Staurosporine and Tofacitinib as case studies for a promiscuous and selective compound with their respective kinase targets; again, if multiple PDB files for the same target-ligand combination were available, the PDB structure with the best resolution was selected. Common interaction for each KLIFS residue across kinases were counted based on the KLIS nomenclature. The interactions calculated by Arpeggio coincide with the interactions provided by the KLIFS database and MOE ligand interaction calculator, shown in Figure 5.5 and 5.6.

5.2.5 Building the CRANkS Interaction Maps

The CRANkS interaction maps list (a) location and element type of each atoms of the bound ligands, (b) the pharmacophoric features of these ligands, and (c) their protein-ligand interactions (Patel, 2018). The interactions were calculated and stored as a vector between two grid points as detailed in the following. They were built using adapted scripts provided by Patel *et al* (Patel, 2018); the steps are detailed below.

Overview

All binding sites for the conformation of DFG-in and the αC helix in were collected from KLIFS for the CDK2 due to the amount of available data and the distribution of its binding affinity values, see Figure A.10, Appendix. If binding affinity values were available for these conformations, the X-ray structures were annotated as strongly or weakly binding using a threshold of 8 pK-units (100 nM), corresponding to 10 nM, the commonly-used threshold for activity in industry (Sach-Peltason, 2018).

Three PDB structures for one target were randomly chosen as reference; other structures were aligned using all atoms and the *align* function in PyMOL version 1.8 (Schrödinger). Alignments with a RMSD value higher than 0.6 Å were removed. This cut-off value was chosen as a compromise, as it allows for having a high number of structures (and thus having a high number of exemplary interactions), while still maintaining a very conservative threshold to guarantee a high overlap between the kinase interactions which form the grid. The binding interactions were calculated using PLIP (Salentin et al., 2015)

and saved for each grid using Patel's adapted scripts. Ultimately, the final CRANkS interaction grid consists of every single ligand-interaction saved for these particular conformations for the target within the specific binding affinity range. The ligands of the references were removed and then either re-docked or randomly cross-docked into the corresponding kinase targets and their interactions with the respective grid scored as detailed in the following section '3. Scoring of the Candidate Compound'.

1. Preparation of the candidates for docking

The ligand SMILES strings of the candidate compounds were standardized and converted to canonical SMILES using MolVS (Version 0.1.1) and RDkit (Version 2018.03.3). OpenBabel was used to add hydrogens for a pH of 7.4 using the flag '-p' and partial charges added using the force field Merck Molecular Force Field 94 and finally subjected to the docking procedure used by Ahinko *et al.* (Ahinko et al., 2019). Lastly, OBABEL was used to generate 100 3D ligand conformers for each compound using the CONFAB method, which generates all diverse low-energy conformers for a given molecule. This format contains information about the torsions and the assigned partial charges for each atom and is the input for the docking with smina (Koes et al., 2013).

2. Docking

Smina was used to dock the previously prepared candidate conformers into the prepared binding pocket (Koes et al., 2013). First, the binding pockets for the chosen structures were prepared using PyMOL Version 1.8 (Schrödinger). The binding pockets were manually checked for any issues, such as missing side chains, within the limits of the binding box. Water molecules were removed and hydrogen atoms added to polar atoms. The ligand was removed from the binding pocket and the binding pocket saved and converted into the PDBQT format using OBABEL. The flags '–autobox_ligand' and '–autobox_add' were used to calculate a box inside the binding pocket for docking. The box is based on the known ligand for the target and was constructed by creating a box of 8 Å extra space around the initial ligand in each dimension. Lastly, the the flag '–exhaustiveness' was set to 16 instead of the default 8, so that smina has more time to find the best scoring binding mode of the ligand.

3. Scoring of the Candidate Compound

The candidate compounds were scored using the script written by Patel *et al* (Patel, 2018). The CRANkS interaction map was calculated as described previously; then for each docked molecule, the protein-ligand interactions are scored against the interaction grid in turn.

The calculation of this so-called interaction score, or interaction novelty score, involves the use of vectors and a vector field (the previously calculated grid map). As shown in Figure 5.1, an interaction is defined as a unit vector starting from the pharmacophoric

feature on the grid point of the ligand, to its end point, the corresponding pharmacophoric feature on the grid map of the protein.
In this context, a denotes a vector with one particular starting point and interaction, b_1 and b_2 are all vectors with the same type of interaction and starting point as a. Hence, Patel *et al.* referred to these vectors as 'interaction counters'.

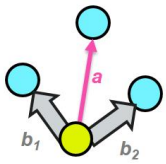


Figure 5.1: The interaction counters b_1 and b_2 for a protein-ligand interaction The vectors b_1 and b_2 (grey arrows) are defined as having the same kind of interaction type as the protein grid points (cyan) and starting point at the compound (yellow dot) as a (pink arrow). The vectors a and b_1, b_2 These vectors, also called interaction counters, are used to calculate the interaction score S_I as a measure for the novelty of observed interactions. This figure is adapted from Patel *et al.* (Patel, 2018).

These interaction counters are then used to calculate the interaction score S_I for a particular conformer as follows: After all interaction counters with the same kind of interactions and starting point as a are found, the dot product of each b-interaction counter with a is calculated. The dot product assigns the position of two vectors a number. It becomes 0 and 1 in case of perpendicular or parallel vectors respectively, see also Figure 5.1. Using the dot product, it is possible to compare the similarity in direction of the interaction counters b_1 or b_2 to interaction a.

The count N_b for each interaction counter is normalised by the maximum count N_{Max}, which is the number of interaction counters found for that type of interaction. Multiplication with the dot product of the interaction counters allows the contribution of each interaction counter to be weighted in terms of similarity of the direction of interaction counter to the direction of the interaction.

Finally, the result is subtracted from 1 to denote how many known interactions have been found (see discussion below). Equation 5.1 describes the calculation of the interaction score, S_I, for each conformer (Patel, 2018):

$$S_I = 1 - \Sigma_1^{n_i} \hat{a}.\hat{b} \times \frac{N_b}{N_{max}}. \tag{5.1}$$

Ultimately, the interaction score S_I for each docked candidate conformer is calculated as

the average of the score for all protein-ligand interactions found for this complex. Then, as the candidate compound is able to form multiple conformers, the overall interaction score is calculated as the average of the interaction scores for all conformers.

The interaction score S_I is bounded between 0 and 1. It can be interpreted as follows: a lower interaction score indicates that more interactions previously seen in the interaction grids are formed between the interaction points and the candidate compounds. Hence, these interactions should be more important for binding. If a compound has a high interaction score, then it should form more novel interactions compared to the ones used to form the grid (Patel, 2018).

5.3 Results

5.3.1 Case Study : Comparison of the Binding Shapes of a Promiscuous and a Selective Kinase Ligand

In this section, I carried out an exploratory analysis of the binding mode and the binding interactions for a promiscuous and a selective ligand for understanding selectivity or promiscuity on a molecular level. To the best of my knowledge, the use of the similarity of binding modes and binding interactions for promiscuous and selective kinase inhibitors has not been explored across kinases; similarly, I believe that this is the first use of standardized KLIFS nomenclature for a direct comparison of binding features across kinases.
I primarily used Staurosporine and MI1 as exemplars of promiscuous and selective ligands respectively. MI1 is a selective ligand against JAK1, JAK2, JAK3 and TYK2 kinases, which allows the comparison of binding interactions for a selective ligand across the same protein isoforms.

First, I used the shape-chemical feature overlap comparison method SuCOS (Leung et al.) (Section 5.2) to compare the shape-chemical features of the same promiscuous ligand across binding pockets from different kinase targets to identify whether promiscuous kinase compounds for a given inhibitor class could be filtered out by comparing candidate compounds to a specific interaction shape.

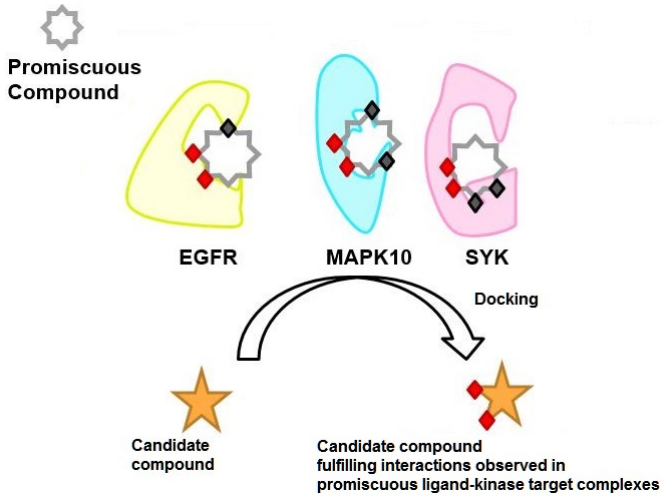


Figure 5.2: Visualisation of a ligand-structure-based method based on kinase-ligand interactions and binding shape. The interactions of a promiscuous ligand (shown as a grey star) are analysed for each target with available crystal structures. Interactions observed and not observed in multiple kinases are shown as red and grey squares respectively. A candidate compound (yellow star) could be compared in terms of its putative interaction with the known, observed interactions, *e.g.* by using docking. A candidate compound which has the same interactions as a promiscuous compound (red squares) might then be labelled as promiscuous. Labelling a compound as selective is more complicated as its interactions and/or shape only occur for its target(s).

As outlined in Section 5.2, I found 11 promiscuous ligands. Each ligand was then pairwise aligned using the binding pockets of their co-complexed targets and compared in terms of their shape and chemical features similarity. The results are in Figure 5.3. A SuCOS score of 1 represents a 100% similarity in terms of overlap, whereas 0% represents no overlap of the two ligands. I show that most promiscuous ligands cluster around 0.6-0.9, indicating high structural and chemical similarity in terms of binding mode. This behaviour is independent of target alignment, as the differences caused in proteins by pairwise alignment are less than 0.2 Å, shown in Figure A.12, in the Appendix. As the calculation of SuCOS depends on the alignment, ligands which could not be well aligned have a lower SuCOS score. As the RMSD of the alignment is generally skewed towards higher RMSD values, see Figure A.12, the SuCOS scores might be generally higher, depending on the ligand.

Especially interesting is the case of the promiscuous (universal) kinase inhibitor Staurosporine (chemical component code: STU), which competes with ATP. The ATP binding site is highly conserved, thus rendering the development of a selective inhibitor difficult. The SuCOS scores of Staurosporine cluster around 0.7-0.8, indicating a high

shape and chemical overlap. Tanramluk *et al.* compared the protein binding sites for Staurosporine and ATP, highlighting that Staurosporine mimics ATP in spite of the low structural similarity and that the shape of the pocket can contribute to inhibitor selectivity. They highlighted that the γ-lactam from Staurosporine and the adenine from ATP recognize similar binding pocket atoms despite the low structural similarity. Hence, the shift to high SuCOS scores of Staurosporine are thus not surprising as it closely mimics ATP in an ATP-conserved binding pocket.
The kinase substrates AMP (Adenosine Monosphosphate), ATP (Adenosine-5′-Triphosphate), ACP (Phosphomethylphosphonic acid adenylate ester), ANP (Phosphoaminophosphonic acid-adenylate ester), and ADP (Adenosine-5′-Diphosphate) show a distribution towards high SuCOS values, highlighting the evolutionary relationship between the substrates, their binding moieties, the shape and arrangement of pharmacophoric features they adopt across binding pockets. Other ligands, such as 1N1 and 3NG, also show high SuCOS scores in general for the same conformation (DFG-in, αC helix in).

Similarly, the SuCOS distribution for the selective compound Tofacitinib is shown in Figure 5.4; Tofacitinib appears to adopt the same binding mode when bound to four different kinase targets.

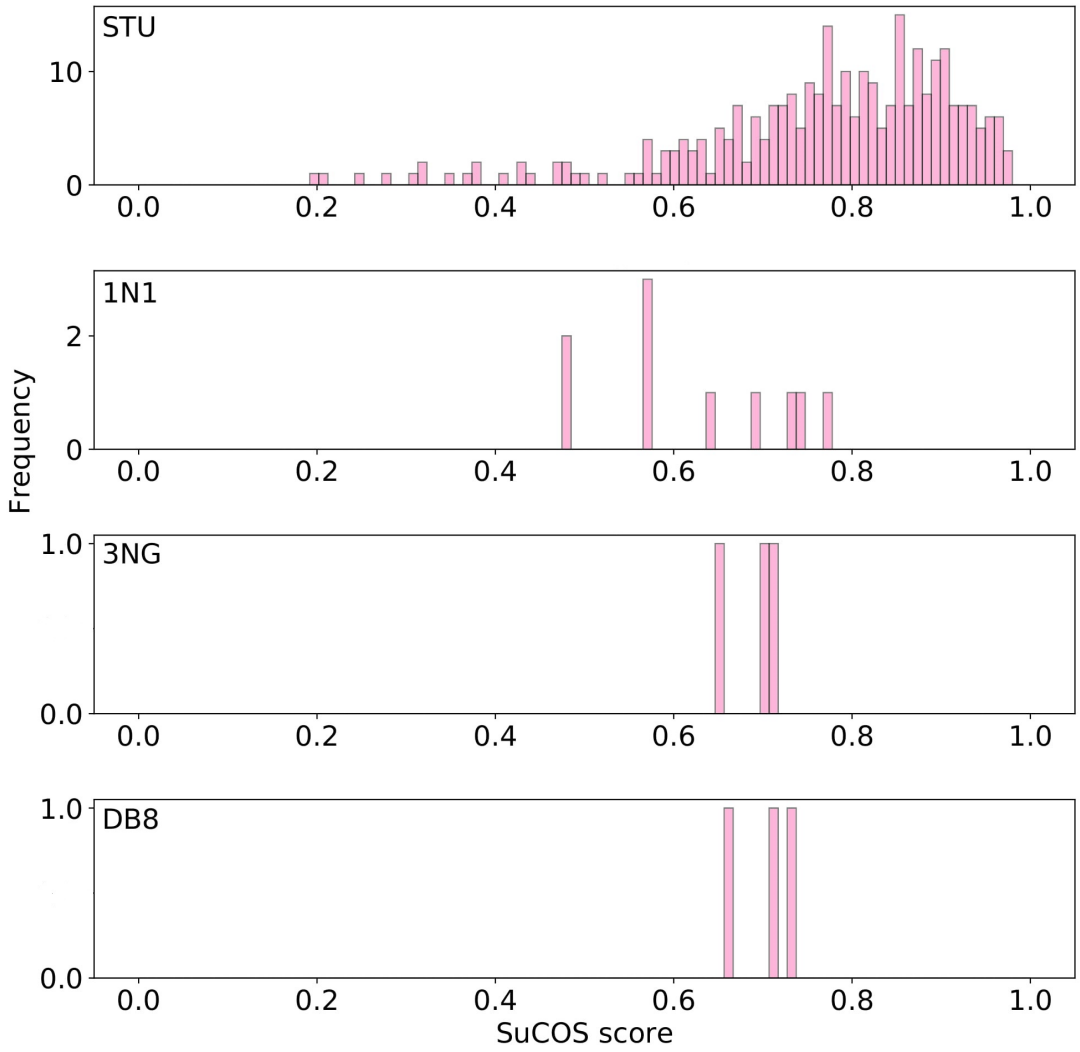
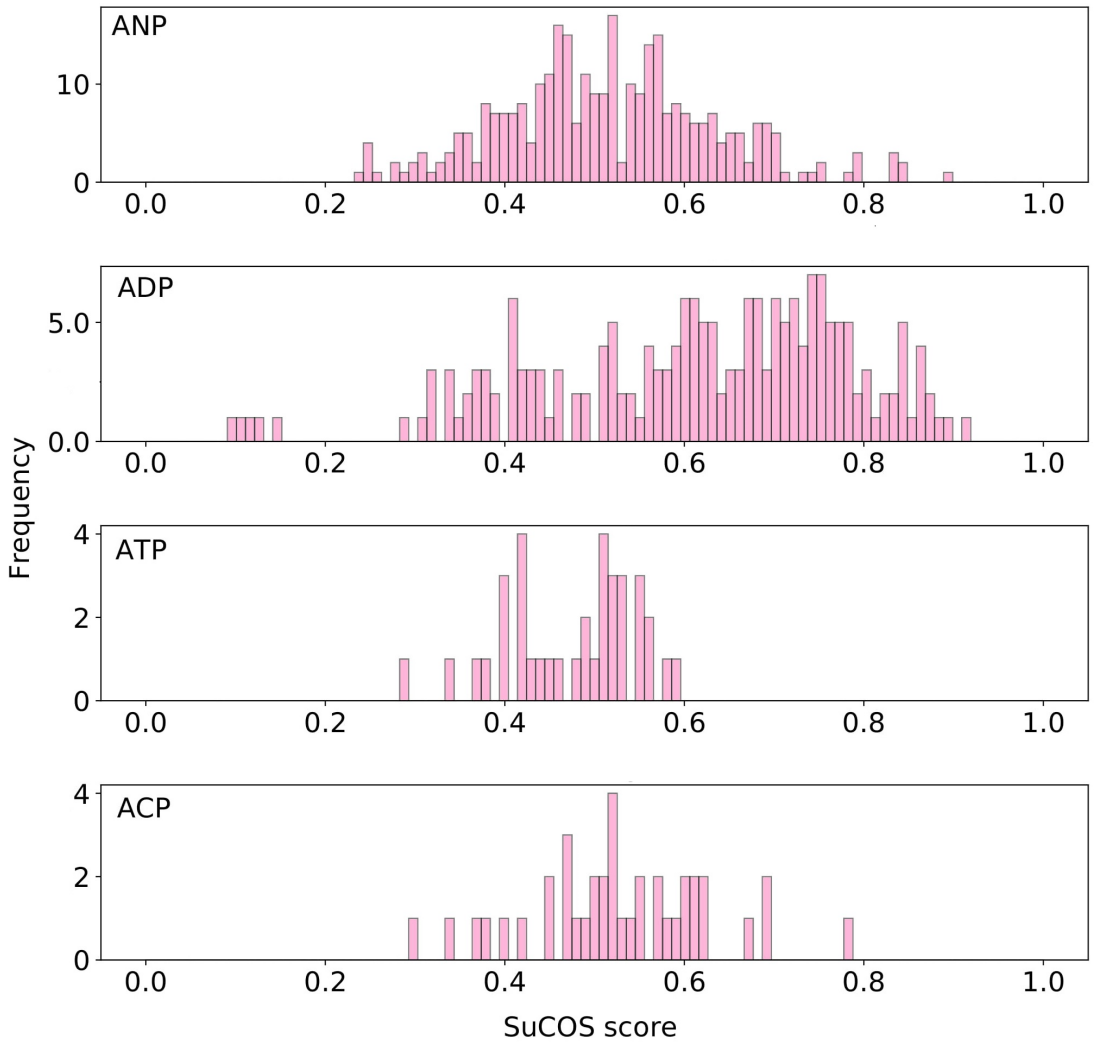
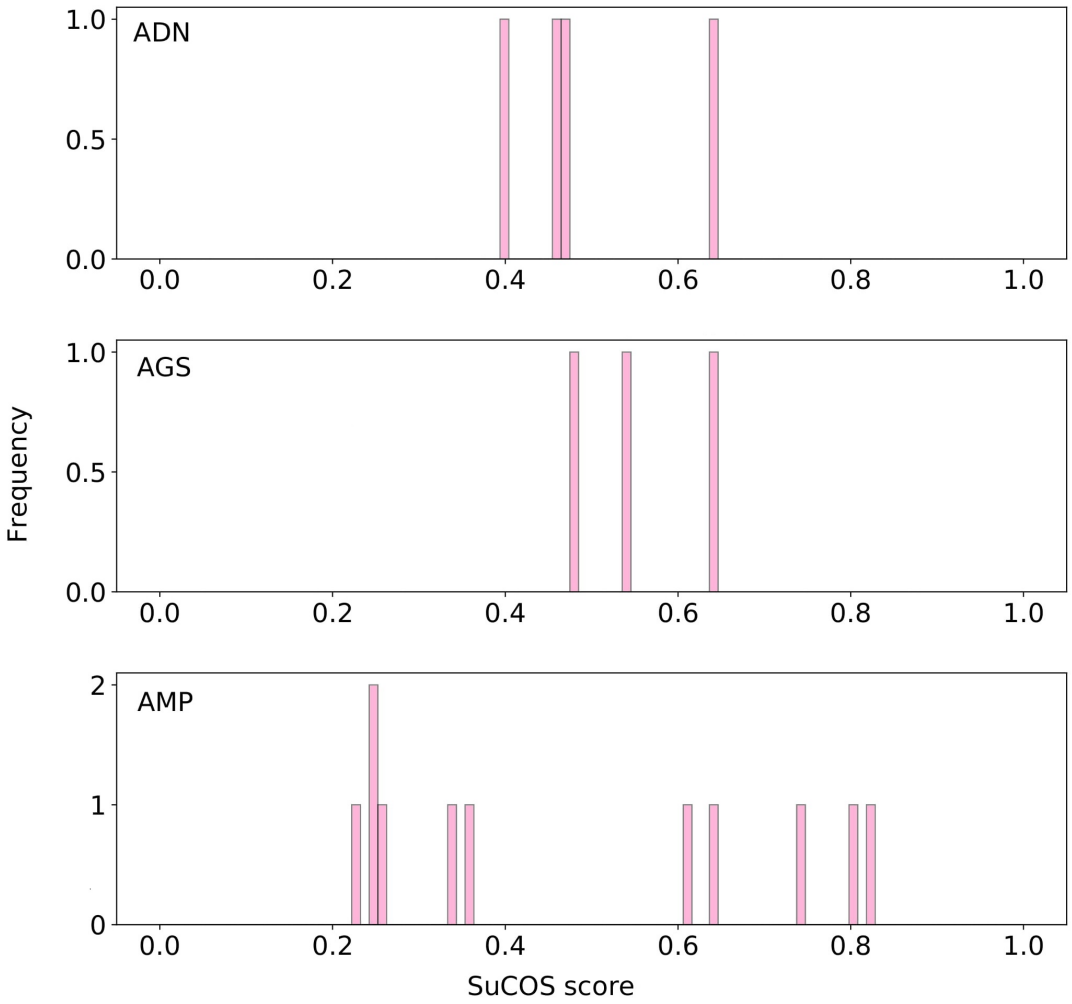


Figure 5.3: Distribution of pairwise SuCOS scores for kinase ligands detected in more than five kinase targets. Ligands bound to more than five different kinase targets were annotated and analysed in terms of their binding shape similarity. Their crystal structures were pairwise aligned, the RMSD calculated for each alignment and kinase pairs with a RMSD equal or higher than 1 were discarded (Figure A.12, in the Appendix). Using the SuCOS score (Leung et al.), their pairwise shape and pharmacophoric overlap was calculated to determine whether promiscuous ligands have a tendency to form similar shapes across different kinases. The SuCOS scores are plotted versus their frequency; the ligand code is in the top left corner.



Distribution of pairwise SuCOS scores for kinase ligands detected in more than five kinase targets (Continued).



[

]**Distribution of pairwise SuCOS scores for kinase ligands detected in more than five kinase targets (Continued).**

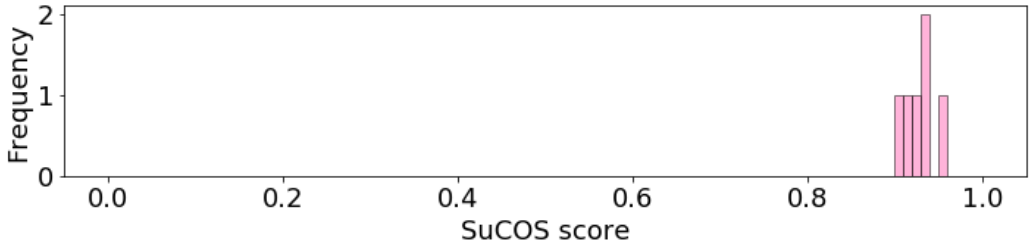


Figure 5.4: Distribution of pairwise SuCOS scores for the selective ligand Tocafitinib or CP-690550 (MI1), selective against JAK1, JAK2, JAK3. The crystal structures for MI1 against JAK1, JAK2, JAK3 as well as TRYK2 were aligned and the SuCOS score calculated as detailed in Section 5.2.4. The PDB codes are 3EYG, 3LXK, 3FUP and 3LXN respectively.

5.3.2 Case Study: Comparison of the Binding Interactions of a Promiscuous and a Selective Kinase Ligand

When comparing the shape-feature overlap of the same selective and promiscuous ligand bound to its kinase target panel, a distribution towards high SuCOS scores is generally observed. This indicates that the ligand shares a conserved binding mode across kinase targets. I also analysed the binding interactions of Staurosporine (Ligand code: STU) as an example of a well-studied promiscuous compound, and Tofacitinib (PDB residue name Code: MI1) as an example of a well-studied selective compound in terms of their backbone and sidechain interactions across their target panel of crystal structures. The results are in Figures 5.5, 5.7 and 5.8 respectively.

Based on my results in Section 5.3.3, I hypothesized that the binding affinity ranges of kinase-ligand complexes cannot be directly connected to specific key residues, but instead are an interplay between the number and location of different kind of interactions. This agrees with the results by Tanramluk *et al.*; they hypothesized that the three residues, the gatekeeper and the size of the gatekeeper interplay to lock the γ-lactam of the Staurosporine in the correct position for binding, allowing the optimal steric interactions with the indolocarbazole of staurosporine to be made. These residues are on the opposite side of the N- and C-lobe (see Section 3.3.2); their interactions influences how close the two lobes come and thus the optimal interaction of Staurosporine with the kinases.
Additionally, Tanramluk *et al.* showed that the number of hydrogen and ionic interactions that Staurosporine makes along the methyl amino (N4) and the methoxy group (O3) groups of staurosporine with the kinases correlates with the major differences in the binding affinities.

As this shows the importance of the number of interactions at specific residues, I analysed whether promiscuous compounds could have a higher percentage of (unspecific) backbone interactions or a higher number of specific residue interactions, which could account for their promiscuous nature and which would allow to build interaction shape models to filter out promiscuous compounds. Staurosporine (STU) was used as a case study, see Figure 5.5.

As shown in Figure 5.5, Stauroporine forms mainly nonpolar interactions (van der Waals, strong and weak hydrogen bonds). These results agree with the results by Tanramluk (Tanramluk et al., 2009) who showed that Staurosporine has very few strong electrostatic features within the binding pocket; its interactions are mainly being dominated by steric non-polar groups (Tanramluk et al., 2009). However, these interactions are mediated via different protein atoms. For example, van der Waals Interactions are mainly mediated by backbone interactions before and around the DFG-motif (KLIFS residues 80-83), indicating that Staurosporine is able to bind more tightly to the backbone itself, whereas strong and weak hydrogen bonds are formed by sidechain atoms. The high number of weak polar bonds in the hinge region (46-48) corresponds to the formation of hydrogen-bonds between the amide and the carbonyl-oxygen of the γ-lactam-ring of Staurosporine, *e.g.* the interactions of the amide bond with Glycine (Gly 1197) and Methionine (Met 1199) of ALK (PDB Code: 3LCS, Figure 5.6a.

Aromatic and Ionic interactions only occur between sidechains, unsurprisingly. Interestingly, however, aromatic interactions are mostly formed at the hinge-region; ionic interactions mostly in the linker region. These interactions can be explained by the 3D-shape of this region: the hinge region is much closer to the γ-lactam ring of the Staurosporine, allowing for a closer aromatic face-to-face or aromatic face-to-edge interaction between an aromatic binding pocket residue and the aromatic ring system of Staurosporine, while the long-range and stronger ionic interaction can be mediated by residues from the linker region further away.

As shown in Figure 5.5, PDPK1 (PDK1) (PDB code: 1OKY) and ALK (PDB code: 3LCS) have the negatively charged sidechains glutamic acid and aspartic acid for ionic interactions available at residue 52 (linker-region), while GSK3B (PDB code: 1Q3D) has a threonine as an uncharged sidechain, which is not available for binding. This highlights that the difference in binding site composition is directly reflected in the protein-ligand interactions. Also, the aromatic binding interactions at residue 48 are only detected in 6 structures out of 30; either because the aromatic ring was not in a position to form π-π stacking interactions or because the target does not contain an aromatic amino acid residue at this particular position.

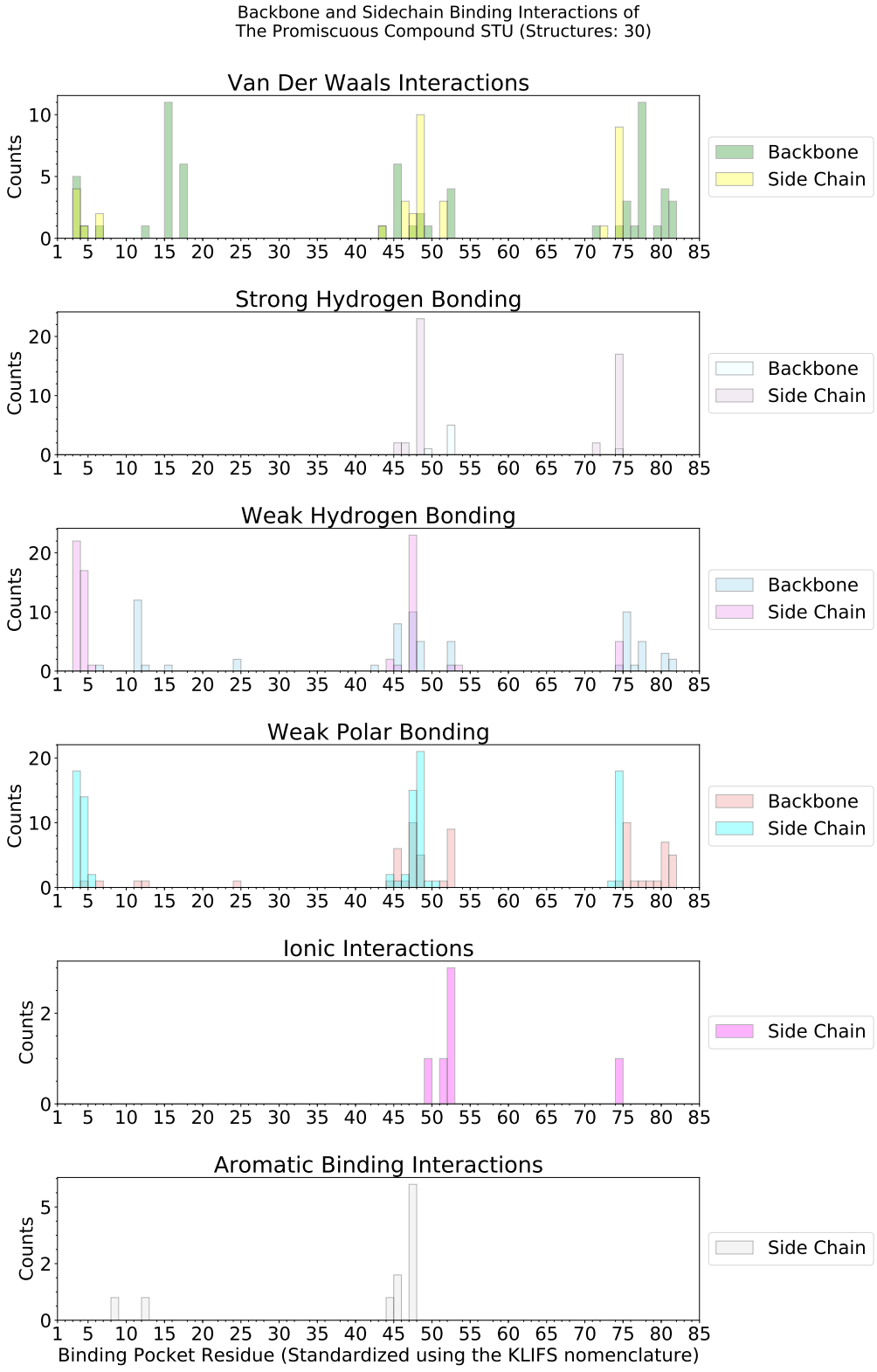


Figure 5.5: Sidechain and Backbone Interactions of Staurosporine across 30 Kinase targets. Across 30 kinases, different interactions between Staurosporine and 85 binding pocket residues were characterized and summed per KLIFS binding pocket residude. The standardization of residues using the KLIFS database allowed for a comparison of binding pocket interactions across kinases. For example, this figure highlights the importance of backbone hydrogen bonds in comparison to sidechain aromatic interactions at KLIFS residue 48.

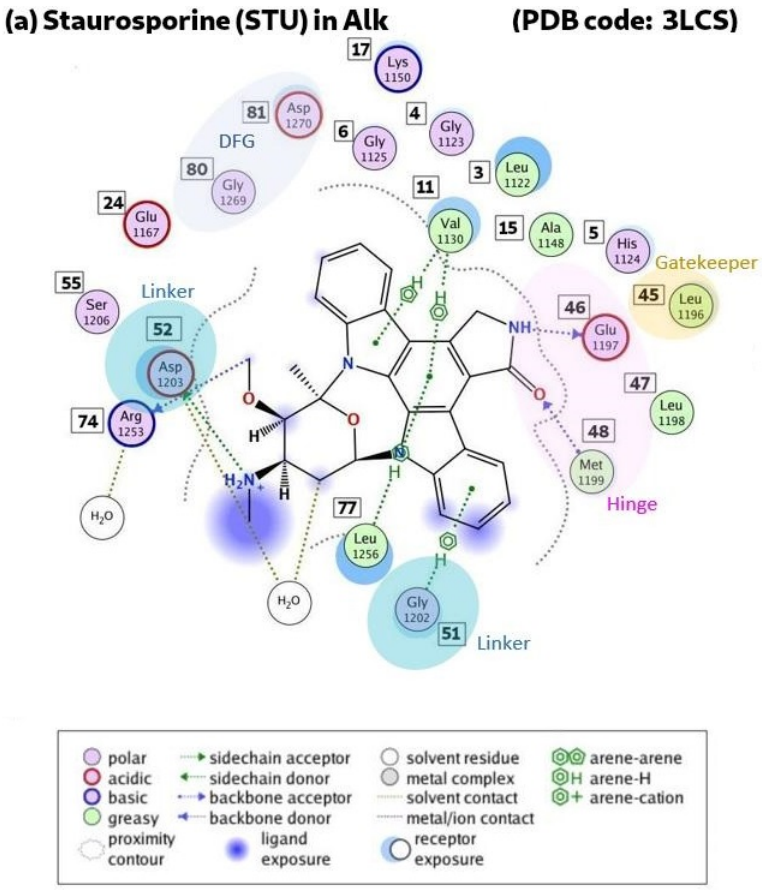


Figure 5.6: Interactions of Staurosporine in the kinases ALK, PDK1 and GSK3B, using MOE (Vilar et al., 2008). The PDB codes are 3CLS, 1OKY and 1Q3D respectively. Each KLIFS residue position is indicated at the left of the respective residue in a gray square. This representation complements Figure 5.5 and allows for a visual inspection of important kinase-ligand interactions. As shown in Figure 5.5, Staurosporine uses the KLIFS residues 46-48 (Hinge-region) for mediating donor and acceptor interactions with the γ-lactam-ring of Staurosporine in all three complexes, indicating their importance. However, the binding interaction at KLIFS residue 52 is only present for PDPK1 (PDK1) and ALK, since GSK3B has an uncharged threonine at that position. This highlights the influence of the binding pocket on binding and shows that, while KLIFS residues 46-48 appear to be important for binding, KLIFS residue 52 might not be.

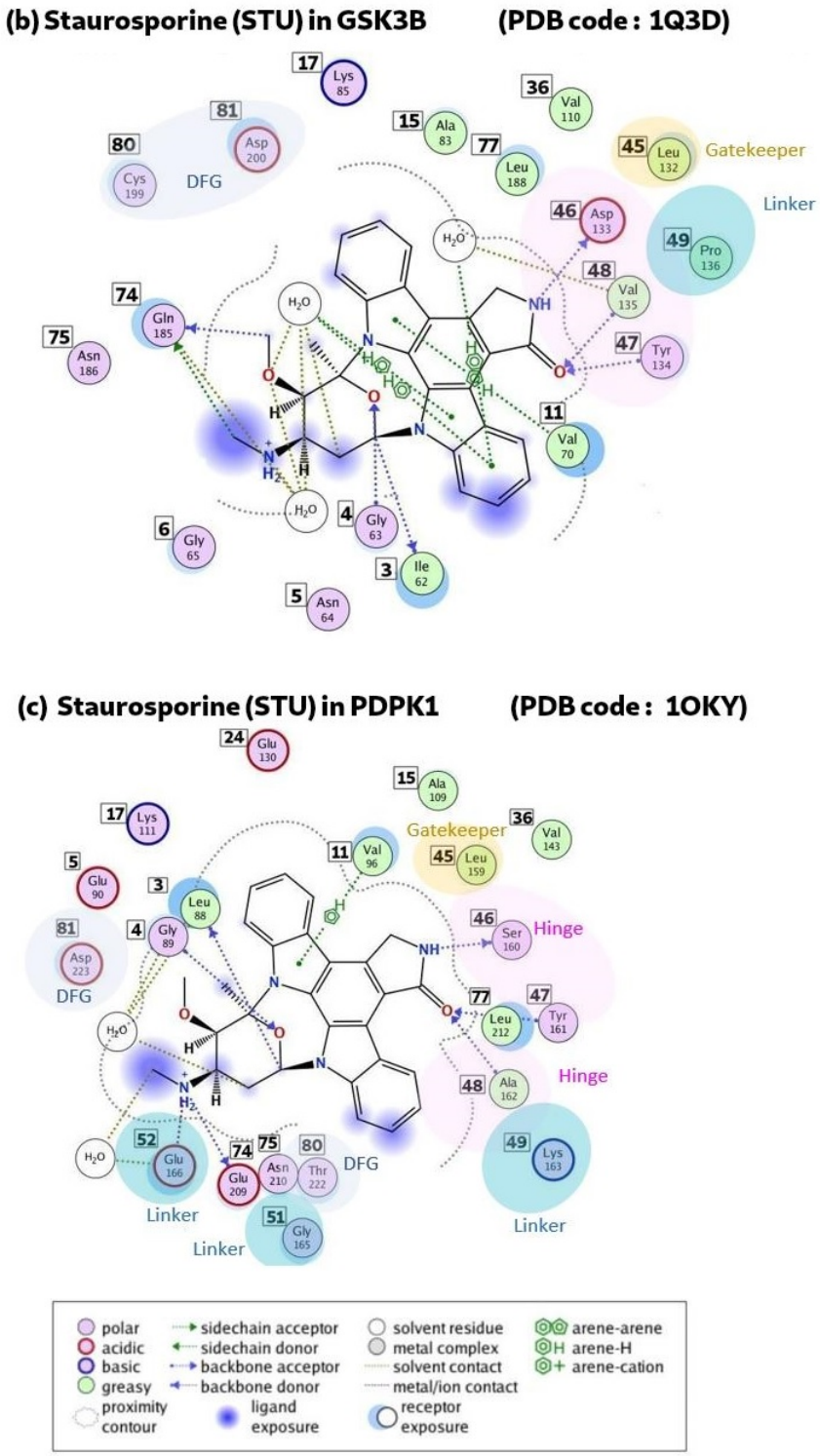


Figure 5.6: Interactions of Staurosporine in the kinases ALK, PDK1 and GSK3B, using MOE (Vilar et al., 2008) (Continued).

As shown in Figure 5.7 and Figure 5.8, weak and strong polar interactions at residue 46 and 48 are also present in the binding interactions of the selective ligand Tofacitinib

(Ligand code: MI1), which is selective for the kinases JAK1, JAK2 and JAK3 (PDB codes 3EYG, 3FUP, 3LXK) in the range of 1-120 nM. It also inhibits TYK2 at higher doses (PDB code 3LXN) (Portal). Analysing the binding interactions of these targets allows both an iso-form specific comparison of binding residues and the comparison between binding pocket residues for a selective kinase compound and a promiscuous kinase compound.

Nearly all isoforms share weak and strong polar backbone interactions at residue 46 and 48, corresponding to donor and acceptor interactions of the nitrogen atom of the pyrrolo(2,3-d)pyrimidine ring of MI1 with corresponding residues in the hinge region. For example, JAK1 and MI1 use Leu-959 and Glu-957 to form backbone donor and acceptor interactions in the hinge region (Figure 5.9). Similarly, aromatic interactions are formed at KLIFS residue 48, in the hinge region, or the selective ligands have weak polar interactions at residue 81, analogously to the interactions of Staurosporine. However, other interactions clearly differ: for example, these selective kinase ligands do not have a ionic interaction with KLIFS residues in the linker region or backbone mediated van der Waals interactions at KLIFS residue 77, see Figure 5.5.

While the hinge region interactions of Tofacitinib are non-specific and correspond to the interactions seen in Staurosporine, there is no easy way to differentiate selective and promiscuous compounds based on interactions. I postulate that the shape of the binding pocket, the number of interactions formed at a given position (Tanramluk et al., 2009) and the availability of the same binding residues across target subfamilies will determine the binding affinity and subsequently, the selectivity or promiscuity of a ligand; taking these factors into account, further steps and associated problems will be discussed in Section 5.4.

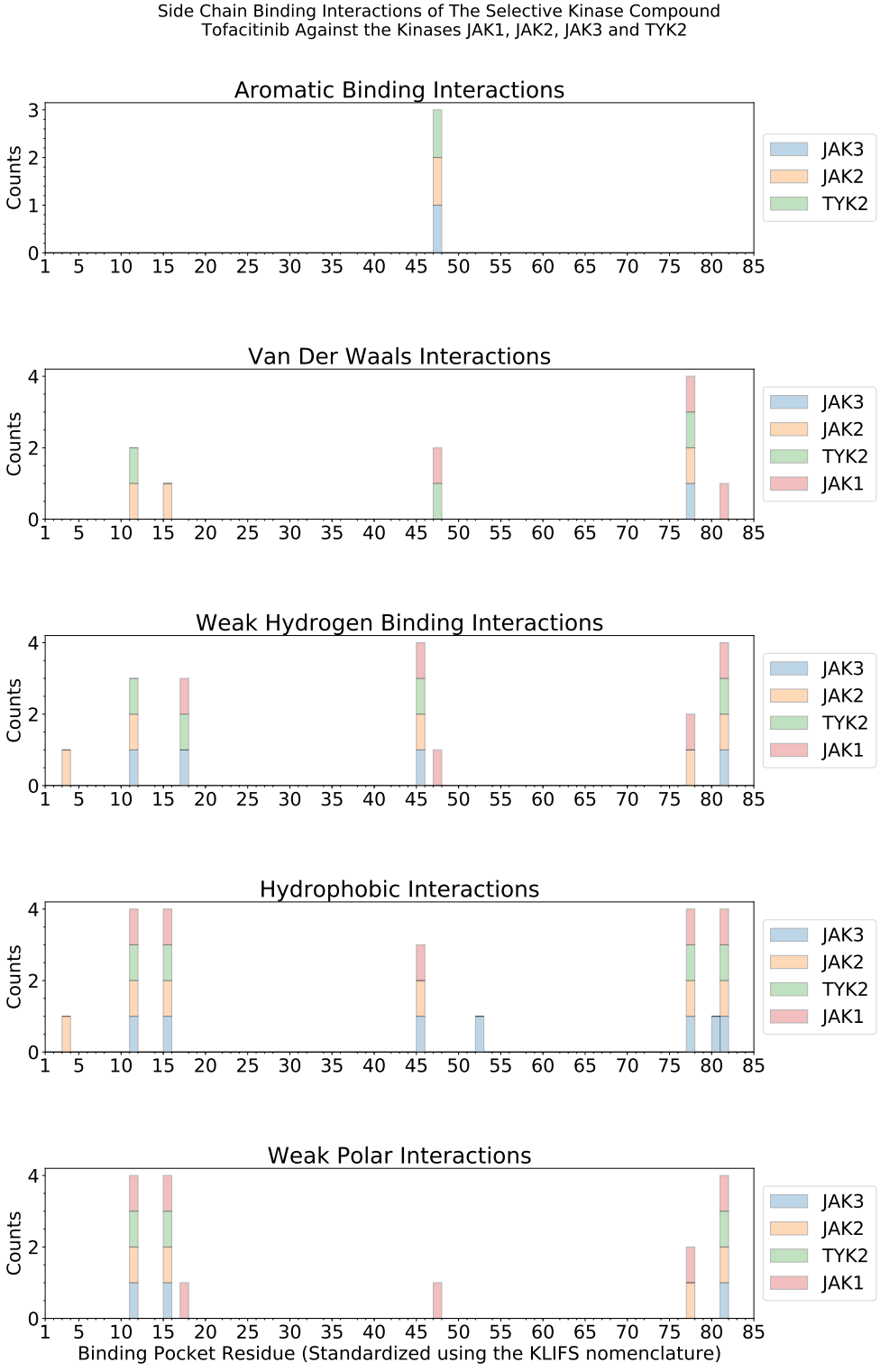


Figure 5.7: Sidechain interactions of the selective ligand MI1 (CP-690550) with 4 Kinase targets: JAK1, JAK2, JAK3 and TYK2; the binding pocket residues are standardized according to the KLIFS nomenclature. Specific interactions are common for (nearly) all complexes with MI1, such as an aromatic interaction at KLIFS residue 48 or van der Waals Interactions with KLIFS residue 77.

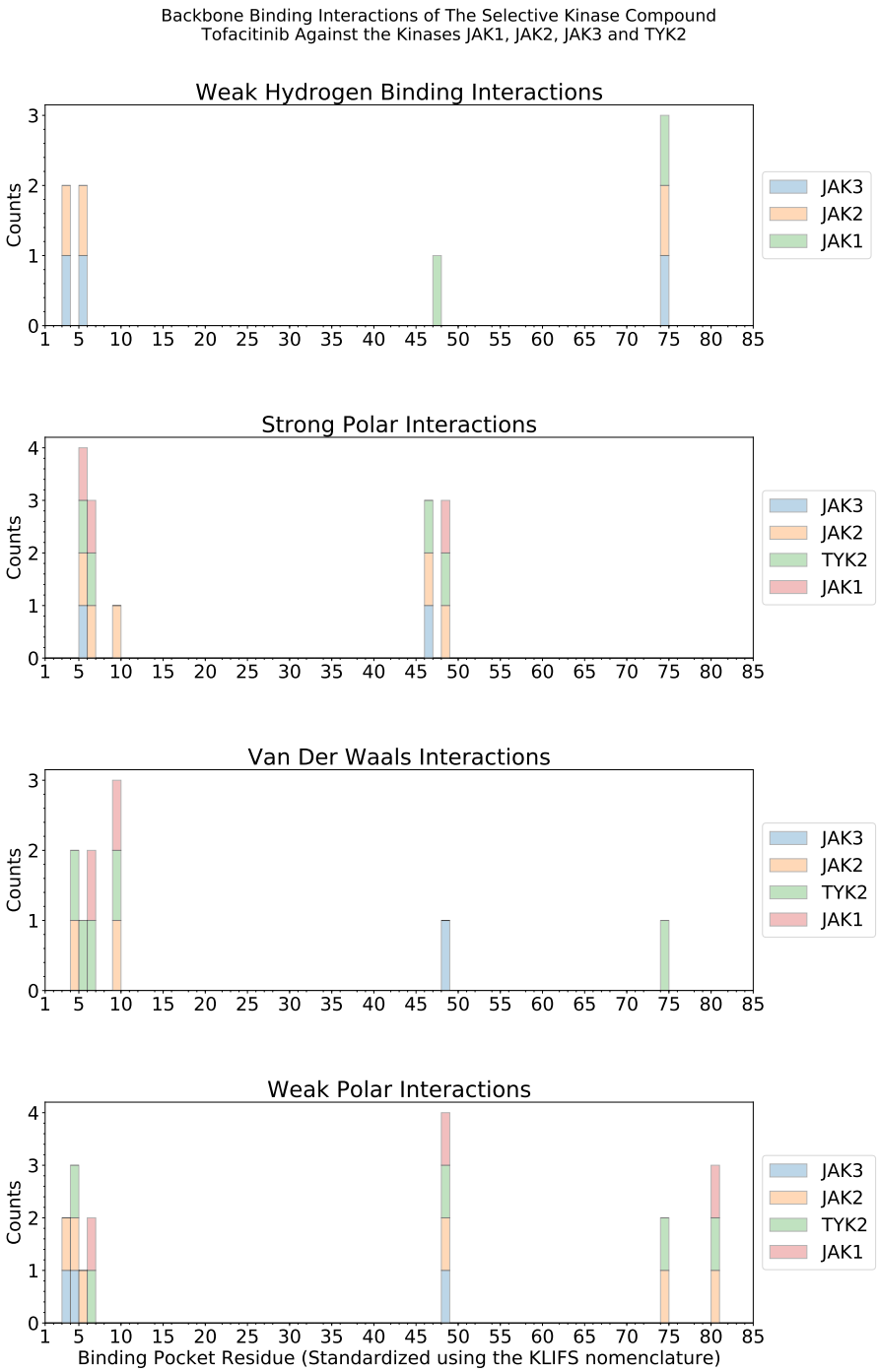


Figure 5.8: Backbone interactions of the selective ligand MI1 with its 4 Kinase targets: JAK1, JAK2, JAK3 and TYK2; the binding pocket residues are standardized according to the KLIFS nomenclature. Interestingly, MI1 forms strong and weak polar interactions with the KLIFS residues 46 and 48 (Hinge-region) in nearly all its complexes. Similarly, Staurosporine forms hydrogen donor and acceptor interactions with the same residues. This highlights the difficulty of differentiating between selective and promiscuous ligands, when using non-specific backbone interactions.

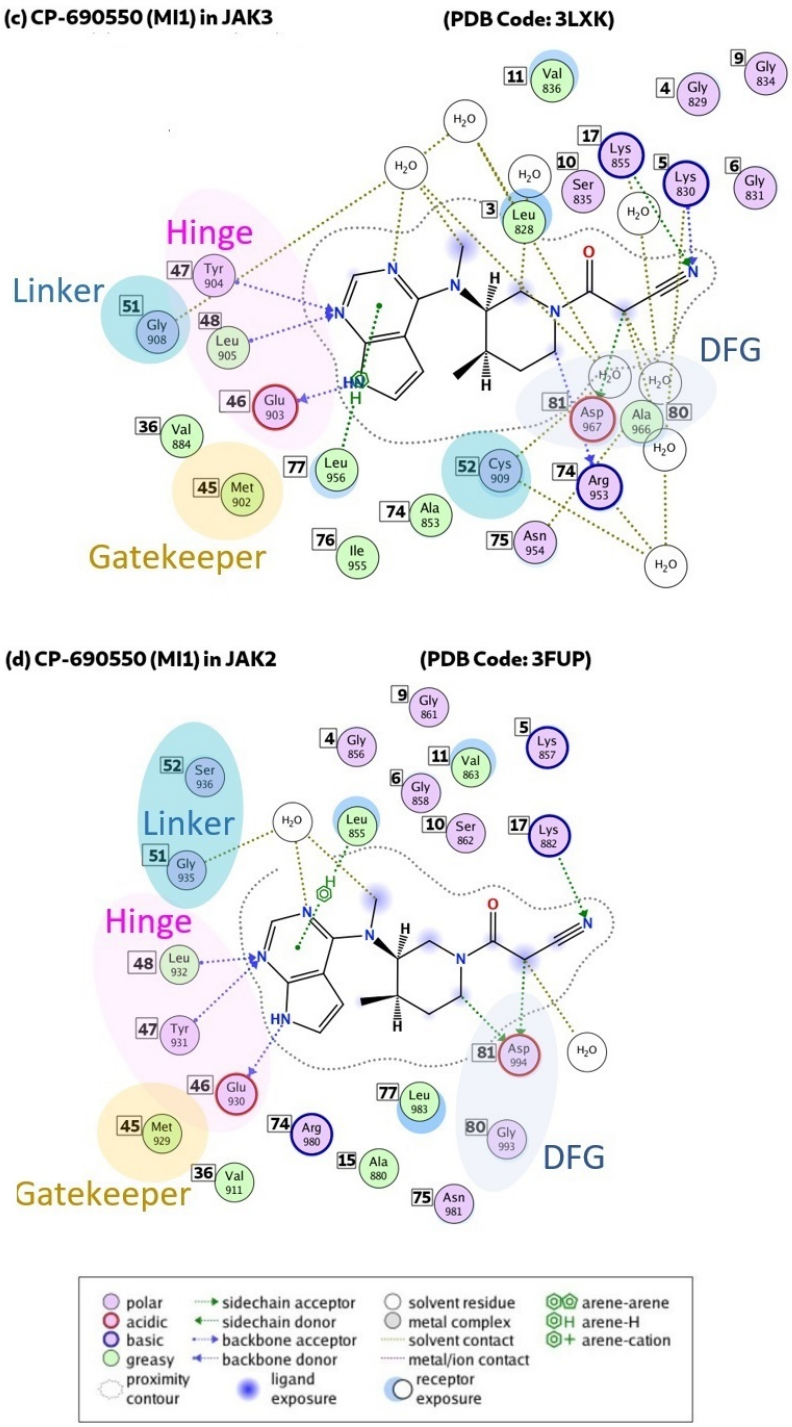


Figure 5.9: Ligand-Target interactions between MI1 (CP-690550 or Tofacitinib) and its targets JAK1, JAK2, JAK3 and TYK2, calculated with MOE (Vilar et al., 2008). This image complements Figures (5.8)-(5.7) by allowing a visual inspection of the found interactions. Backbone interactions with the KLIFS residues 46, 47 and 48 (Hinge-region) are detected for all complexes with MI1, highlighting that these residues are necessary for binding for both promiscuous and selective compounds (see also Figure 5.6). Similarly, key interactions such as sidechain donor interactions with KLIFS residue 17 and 81 are formed in all four complexes. This type of interactions is not observed in Staurosporine complexes (see Figures 5.6 and 5.5), indicating that the binding or nonbinding to specific sidechain residues, key interactions, might be used to filter out selective or promiscuous ligands.

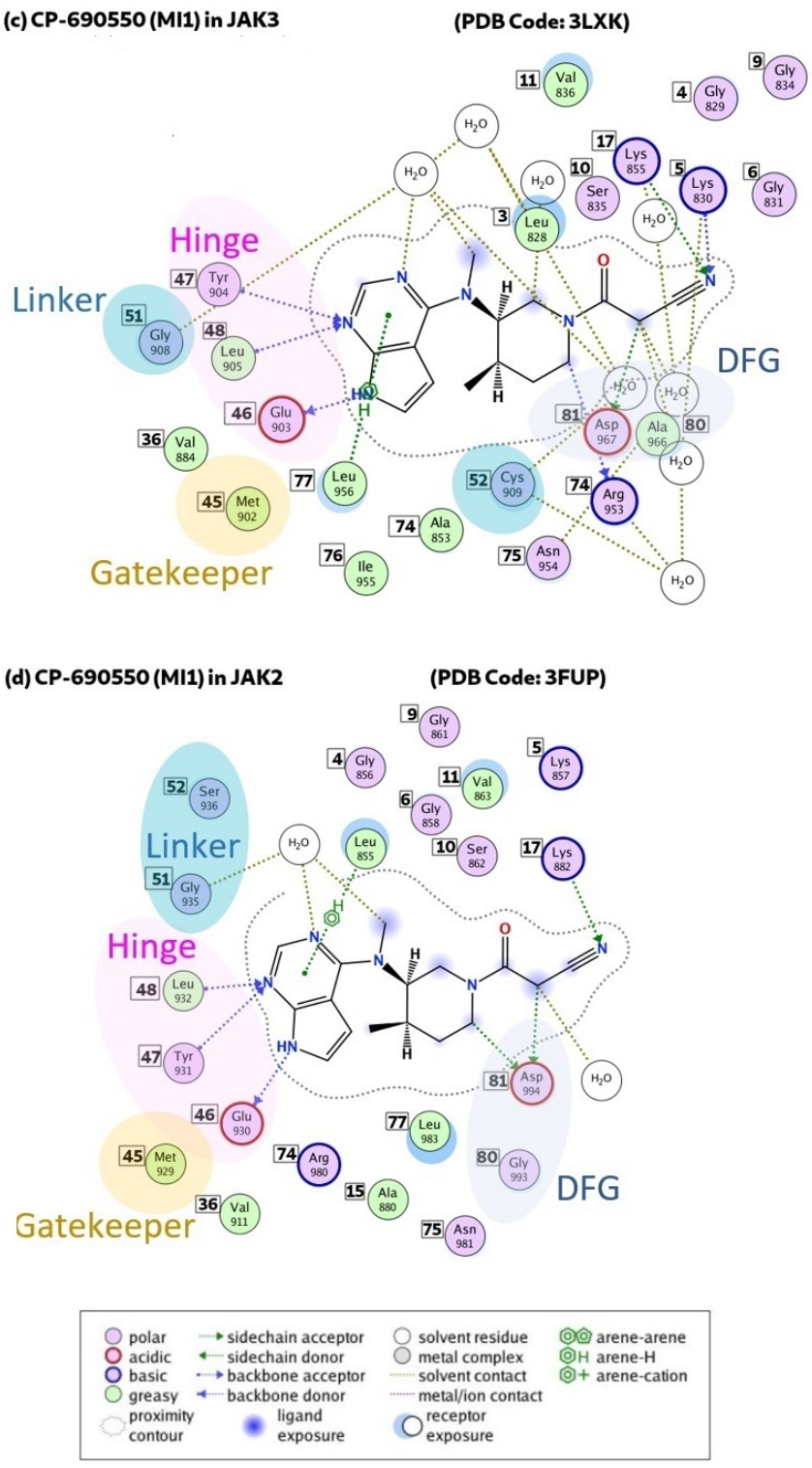


Figure 5.9: **Ligand-Target interactions between MI1 (CP-690550 or Tofacitinib) and its targets JAK1, JAK2, JAK3 and TYK2, calculated with MOE (Vilar et al., 2008) (Continued).**

5.3.3 Investigating the Use of Binding Interactions To Differentiate Between Strong and Weak Binders on a Per-Target Basis

Patel *et al.* developed the CRANkS algorithm to score candidate compounds for a target based on their novelty compared to known binders of the target (Patel, 2018). To construct the CRANkS interaction grids, protein-ligand X-ray crystal structures are used to generate grids which describe the spatial distribution of the features of the ligand or the protein-ligand interactions. All conformers of the putative binders are then generated and scored based on their overlap with the CRANkS interaction grids by comparing the potential interactions of the compounds with the previously observed interactions in X-ray structures. Using a novelty score for the chemical moities mediating the interactions, Patel *et al.* were able to find novel active scaffolds from data sets of active and inactive compounds for a given target. The use of these CRANKS interaction maps offer both a visual interpretation and a scoring of candidate compounds independent of chemical features.

However, if X-ray crystal structures for a given target are not available, then representative interactions for a given kinase target may not be accurately sampled. Hence, if a candidate compound fulfills binding interactions for one target, but not across the target panel, then this compound might either be selective or these specific interactions have not (yet) been observed for the other kinases targets.

In the context of my analyses, I explored the use of these maps to classify kinase candidate compounds as selective or promiscuous (Figure 5.10). In this section, I first investigated whether it was possible to classify strong or weak kinase binders on a per-target basis using binding interactions as features in RF models. Then I tested whether the 3D-structural maps could help in discriminating between strong and weak binders (see Section 5.3.4).

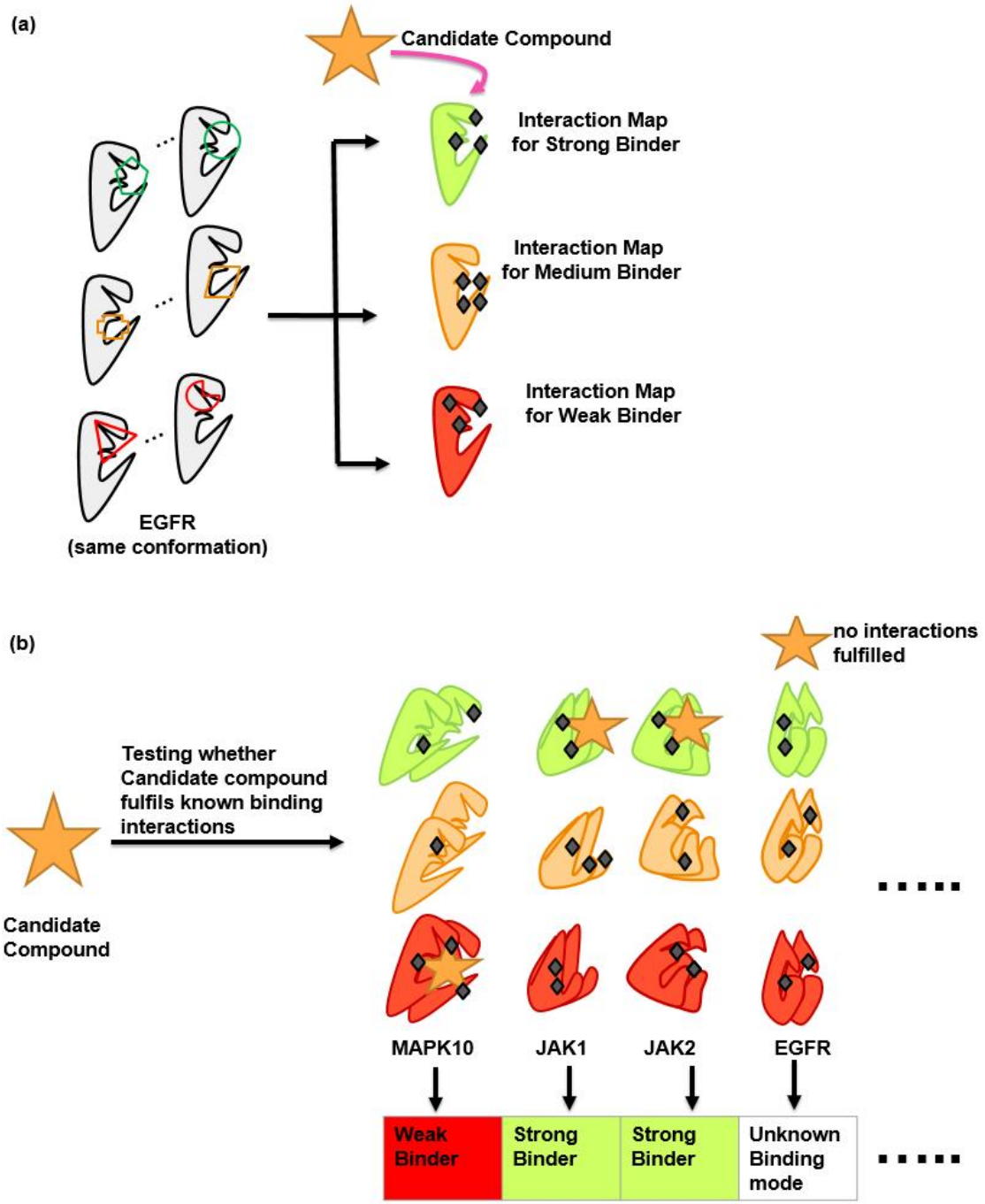


Figure 5.10: Visualization of the use of CRANkS interaction maps to classify a compound as selective or promiscuous. **(a)** All crystal structures with the same conformation (in terms of DFG-motif and αC helix) for a target of interest are annotated with the binding affinities for the respective ligand-protein complex, if available. They are then separated into strong, weak and medium binders, shown in green, yellow and red respectively, and the interaction map for the known protein-ligand interactions for each class are calculated.
(b) The putative interactions of a given candidate compound (yellow star) are then compared with the known interactions for each target and its binding affinity ranges. This results in target-specific affinity predictions for a given compound across a kinase panel, which can be summed up in a selectivity score for the candidate compound.

I analysed whether simple protein-ligand interaction features could be used to separate ligands into strong and weak binders to understand whether kinase-ligand interactions maps could be used to classify candidate compounds into different binding classes. As detailed in Section 5.2, three RF models for 18 kinases were built using thresholds of 8,7, 6.3 pK units (corresponding to 10 nM, 100 nM and 500 nM respectively). The result are in Table 5.1 - 5.3.

As shown, the ratio of strong to weak binders is target-dependent and can be imbalanced towards either strong or weak binders. I accounted for class imbalance during training and calculated four metrices, see Section 3.4.2. The interpretation of the results is now target-dependent.

For example, the MCC for PDPK1 in Table 5.2 at 0.01 indicates a random estimation of the classifier, as does the achieved accuracy and the average precision-recall. The F_1-score is slightly higher at 0.77; however, this metric does not include the number of samples correctly predicted to be negative. In combination with the other metrices, I consider the results of the F_1-score to be over-optimistic. Similarly, CDK2 is imbalanced towards weak binders; its scores can also be interpreted as random.

In conclusion, I primarily used the MCC to interpret the results as it produces a high score only if the classifier obtains good results for all four categories of the confusion matrix and thus a realistic estimation of the prediction ability of the classifier. The Matthews correlation coefficients however range around 0-.004, indicating random results. Meaningful conclusions about using binding interaction features to separate strong and non-binders cannot be drawn based on these performance metrics.

While there may be key interactions separating ligands at different thresholds, due to the low number of available samples, my classifier might not be able to distinguish between them. Subsequently, I tested whether the CRANkS maps by Patel *et al.* could be used to classify ligands as strong or weakly binding in the next section, Section 5.3.4 (Patel, 2018). My rationale was based on Tamranluk *et al.*'s results which correlated the number of formed hydrogen bonds with increased binding affinity. I thus hypothesized that a classification model based on interactions in 3D space would be better at separating strong and weak binders.

Table 5.1: Data and performance of models at a threshold of 6.3 pK-value (500 nM). This table lists the number of crystal structures, binding interaction features, data points in training and test sets and the number of strong and weak binders, as well as the four metrics used to quantify the performance: the Matthews Correlation Coefficient (MCC), the Accuracy (Acc.), Average Precision-Recall and F_1 Score.

Kinase	# Structures	# Strong Binders	# Weak Binders	# Data Points in Test Set	# Data Points in Training Set	Features	Acc.	MCC	Average Precision -Recall	F_1 Score
ALK	25	24	1	-	-	55	-	-	-	-
AURKA (AurA)	50	46	4	10	40	76	0.86	-0.04	0.90	0.92
CDK2	61	45	16	13	48	57	0.72	0.20	0.81	0.85
CHEK1 (CHK1)	87	63	24	18	69	111	0.59	0.19	0.77	0.79
CHEK2 (CHK2)	33	21	12	7	26	86	0.74	0.49	0.79	0.86
CSNK2A1 (CK2a1)	31	24	7	7	24	49	0.6	-0.21	0.68	0.77
EGFR	42	39	3	9	33	61	0.82	-0.08	0.88	0.9
EPHA2 (EphA2)	21	20	1	-	-	55	-	-	-	-
GSK3B	36	36	0	-	-	51	-	-	-	-
IRAK4	20	16	4	4	16	72	0.7	-0.07	0.74	0.81
JAK1	18	18	0	-	-	47	-	-	-	-
JAK2	54	50	4	11	43	79	0.82	0.03	0.92	0.91
JAK3	23	23	0	-	-	66	-	-	-	-
MAPK1 (Erk2)	37	31	6	8	29	56	0.875	0.30	0.92	0.92
MAPK10 (JNK3)	33	30	3	7	26	97	0.86	0	0.86	0.90
PDPK1 (PDK1)	37	24	13	8	29	49	0.475	-0.06	0.63	0.74
PIK3CG (p110g)	61	52	9	13	48	55	0.78	-0.01	0.85	0.88
PIM1	82	65	17	17	64	74	0.66	-0.05	0.82	0.83
Pik3cd (p110d)	31	30	1	-	-	49	-	-	-	-
SYK	43	41	2	9	34	73	0.91	0	1	0.98
TTK	42	36	6	9	33	62	0.87	-0.03	0.89	0.91

Table 5.2: Data and performance of models at a threshold of 7 *pK*-value (10 nM). This table lists the number of crystal structures, binding interaction features, data points in training and test sets and the number of strong and weak binders, as well as the four metrics used to quantify the performance: the Matthews Correlation Coefficient (MCC), the Accuracy (Acc.), Average Precision-Recall and F_1 Score.

Kinase	# Structures	# Strong Binders	# Weak Binders	# Data Points in Test Set	# Data Points in Training Set	Features	Acc.	MCC	Average Precision -Recall	F_1 Score
ALK	25	23	2	5	20	67	1	1	1	1
AURKA (AurA)	50	38	12	10	40	76	0.5	-0.06	0.79	0.75
CDK2	61	21	40	13	48	57	0.57	0.07	0.35	0.79
CHEK1 (CHK1)	87	46	41	18	69	111	0.56	0.10	0.59	0.77
CHEK2 (CHK2)	33	16	17	7	26	86	0.57	0.17	0.56	0.77
CSNK2A1 (CK2a1)	31	17	14	7	24	49	0.6	0.25	0.65	0.79
EGFR	42	32	10	9	33	61	0.62	0.06	0.79	0.82
EPHA2 (EphA2)	21	14	7	5	16	50	0.56	0.10	0.63	0.77
GSK3B	36	31	5	8	28	51	0.85	0.23	0.91	0.92
IRAK4	20	14	6	4	16	72	0.65	0	0.77	0.81
JAK1	18	17	1	-	-	74	-	-	-	-
JAK2	54	46	8	11	43	79	0.73	-0.05	0.81	0.85
JAK3	23	22	1	-	-	111	-	-	-	-
MAPK1 (Erk2)	37	26	11	8	29	56	0.83	0.55	0.90	0.91
MAPK10 (JNK3)	33	15	18	7	26	97	0.46	-0.08	0.48	0.71
PDPK1 (PDK1)	37	18	19	8	29	49	0.55	0.01	0.58	0.77
PIK3CG (p110g)	61	41	20	13	48	55	0.66	0.28	0.77	0.83
PIM1	82	46	36	17	64	74	0.59	0.09	0.62	0.78
Pik3cd (p110d)	31	26	5	7	24	39	0.74	-0.12	0.84	0.86
SYK	43	34	9	9	34	73	0.53	-0.3	0.73	0.77
TTK	42	29	13	9	33	62	0.47	-0.19	0.65	0.72

Table 5.3: Data and performance of models at a threshold of 8 pK-value (100 nM). This table lists the number of crystal structures, binding interaction features, data points in training and test sets and the number of strong and weak binders, as well as the four metrics used to quantify the performance: the Matthews Correlation Coefficient (MCC), the Accuracy (Acc.), Average Precision-Recall and F_1 Score.

Kinase	# Structures	# Strong Binders	# Weak Binders	# Data Points in Test Set	# Data Points in Training Set	Features	Acc.	MCC	Average Precision -Recall	F_1 Score
ALK	25	20	5	5	20	67	0.72	-0.082	0.79	0.83
AURKA (AurA)	50	14	36	10	40	76	0.68	0.01	0.40	0.81
CDK2	61	7	54	13	48	57	0.88	0.15	0.26	0.94
CHEK1 (CHK1)	87	27	60	18	69	111	0.57	0.05	0.37	0.76
CHEK2 (CHK2)	33	8	25	7	26	86	0.89	0.75	0.72	0.94
CSNK2A1 (CK2a1)	31	9	22	7	24	49	0.71	0.49	0.49	0.86
EGFR	42	11	31	9	33	61	0.56	0.01	0.28	0.79
EPHA2 (EphA2)	21	4	17	5	16	50	0.76	-0.05	0.20	0.84
GSK3B	36	18	18	8	28	51	0.63	0.26	0.68	0.80
IRAK4	20	5	15	4	16	72	0.55	-0.12	0.40	0.76
JAK1	18	12	6	4	13	47	0.55	-0.13	0.74	0.77
JAK2	54	26	28	11	43	79	0.53	0.05	0.51	0.76
JAK3	23	12	11	5	18	66	0.40	-0.23	0.59	0.69
MAPK1 (Erk2)	37	19	18	8	29	56	0.53	0.05	0.53	0.74
MAPK10 (JNK3)	33	7	26	7	26	97	0.86	0.17	0.31	0.91
PDPK1 (PDK1)	37	8	29	8	29	49	0.68	-0.13	0.25	0.80
PIK3CG (p110g)	61	21	40	13	48	55	0.60	0.19	0.39	0.81
PIM1	82	24	58	17	64	74	0.54	0.01	0.35	0.77
Pik3cd (p110d)	31	17	14	7	24	39	0.77	0.59	0.77	0.88
SYK	43	15	28	9	34	73	0.62	0.22	0.48	0.79
TTK	42	13	29	9	33	62	0.62	-0.01	0.34	0.77

5.3.4 Use of CRANkS Interaction Maps to Differentiate Between Strong and Weak Binders

As shown in Section 5.3.3, the random forest models tested here did not pick up a high enough signal to separate strong and weak binders. While this might be due to the low numbers of examples, it also suggests that there are no strong key interactions that can be picked up consistently by the random forest models. Tanramluk *et al.* suggested that the number of hydrogen interactions formed in a given subpocket correlates with the binding affinity of Staurosporine. As these binding interactions depend on the 3D-spatial interactions between target and ligand, I tested whether the CRANkS interaction maps could also be used to classify candidate compounds in specific binding affinity ranges using the kinase CDK2. This process is further detailed in Figure 5.12 and in Section 5.2.

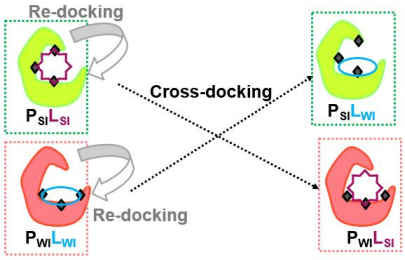


Figure 5.11: Testing the potential of the CRANkS interaction maps to classify compounds in different binding affinity ranges.. To test the potential of CRANkS interaction grids to classify candidate compounds into the respective binding affinity ranges, kinase ligand-target complexes for the kinase CDK2 were separated into strong and weak binders (green and red respectively). Randomnly, the ligand-complexes were aligned, the interaction maps calculated (dashed box) and the ligand extracted. The ligand was re-docked into its original target or cross-docked into the other kinase and its interaction novelty score S_I with the interaction grid of the respective kinase calculated, see also Equation 5.1 and Section 5.2.5.

As detailed in Section 5.2, I expected lower CRANkS interaction novelty scores for re-docked ligands than for cross-docked ligands. Since the redocked ligands have been used to built the CRANkS interaction grids, their calculated interactions should already be present in the interaction grid and less novel; as such, they are more likely to be active. A low interaction novelty score thus indicates that interactions have been formed which have already been observed, while higher interaction novelty scores correspond to novel interactions being formed. As shown in Table 5.4 and Table 5.5, the interaction scores for re-docked and cross-docked ligands are highly similar. Additionally, the derived

novelty interaction scores for the re-docked ligands appear to be very high, indicating that novel interactions in addition to those already used to build the grid were picked up for re-docked ligands. This discrepancy can be explained as follows: Patel *et al.* used constrained docking with Matched Molecular Pairs instead of docking with Smina, potentially leading to the detecting of previously unseen interactions for the re-docked ligand, see Figure 5.12. Additionally, the ligand might not overlap well with the interaction grid; however, manual inspection showed that this was not the case. Further work to explore this discrepancy is needed as detailed in Section 5.4.

Table 5.4: Interaction novelty scores for re-docked ligands derived by using CRANkS interaction maps for randomly selected protein-ligand combinations. Crystal structures for a target were denoted as either strongly or weakly binding, depending on the binding affinity threshold of 8 pK-units (100 nM); their cognitive ligands were extracted and re-docked into the original target. The CRANkS interaction map was calculated using the original protein-ligand complex. I hypothesized that the re-docked ligand should have a high overlap with the grid and thus a low interaction novelty score.

Residue name of re-docked ligand (PDB Code)	Reference structure	Binding affinity	Interaction novelty Score	# Structures
SBC (2BKZ)	2BKZ	strong binder	0.90	4
529 (2BPM)	2BPM	strong binder	0.98	5
QQ2 (2IW6)	2IW6	weak binder	0.95	46
4SP (2IW8)	2IW8	weak binder	1.00	45
RFZ (3MY5)	3MY5	weak binder	0.89	46

Table 5.5: Interaction novelty scores for cross-docked ligands derived by using CRANkS interaction maps for randomly selected protein-ligand combinations. Crystal structures for a target were denoted as either strongly or weakly binding, depending on the binding affinity threshold of 8 pK-units (100 nM); their cognitive ligands were extracted and cross-docked into the interaction maps of the of the same target. The interactions formed between the strong/weak binders and the interaction maps for weak/strong binders should find novel interactions, indicated by a higher interaction novelty score.

Residue name of cross-docked ligand (PDB Code)	Reference structure	Binding affinity	Interaction novelty Score	# Structures
RFZ (3MY5)	2BKZ	weak binder	1.00	4
4SP (2IW8)	2BKZ	weak binder	0.94	4
4SP (2IW8)	2BPM	weak binder	1.00	5
RFZ (3MY5)	2BPM	weak binder	1.00	5
529 (2BPM)	2IW8	strong binder	0.97	45
SBC (2BKZ)	2IW8	strong binder	1.00	45
529 (2BPM)	3MY5	strong binder	0.88	46
SBC (2BKZ)	3MY5	strong binder	0.95	46

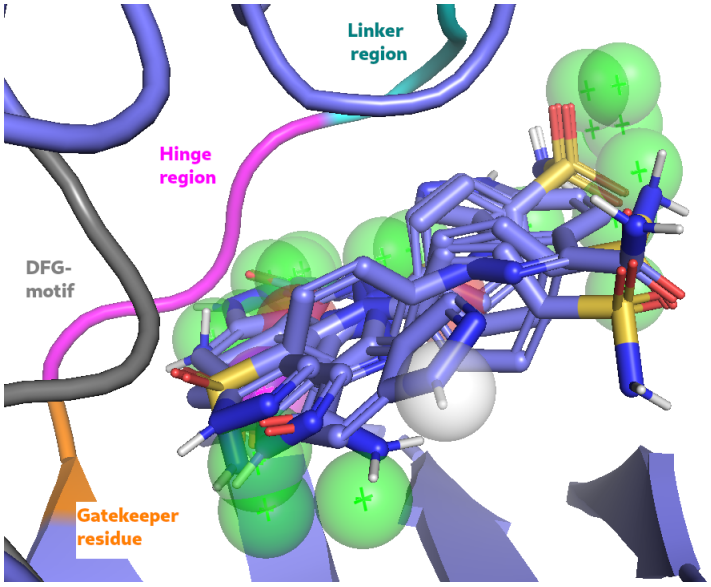


Figure 5.12: An example set of re-docked conformers of the ligand SBC with its orginal target CDK2 (PDB code: 2BKZ), together with the interaction grid.. SBC was re-docked into the original binding pocket. The green, red and white balls denote the PLIP interactions hydrogen bonds, hydrophobic interaction, and salt-bridges respectively (Salentin et al., 2015). These interactions are used to calculate the interaction novelty score S_I as detailed in Equation (5.1) and Section 5.2.5.

5.4 Conclusion and Future Directions

By comparing the binding shapes and chemical overlap of promiscuous and selective compounds, I showed that promiscuous and selective compounds tend to bind to their respective targets with a similar binding mode. By comparing the residue-specific sidechain and backbone interactions of a promiscuous and a selective kinase compound (Staurosporine, Tofacitinib, respectively) across kinase targets — using the KLIFS nomenclature to standardize binding pocket residues — I highlighted the difficulty of connecting specific residues with promiscuity or selectivity.

In future work, the binding interactions for selective and promiscuous compounds could be connected to a sequence alignment of the binding pocket residues as defined by KLIFS. This way, the interactions can be connected to the type of residue at a given common binding pocket position to construct 3D compound interaction shapes which could be used to filter out promiscuous compounds or selective compounds for specific kinases or kinase families. To prioritize binders, Ligity could be used. This method is ligand-structure-based and non-superpositional; it compares triangles or tetrahedra of pharmacophoric interaction points (PIPs) for ligands with PIPs derived from key

interaction features in protein-ligand complexes (Ebejer et al., 2019). As such, Ligitiy is a complementary method to CRANkS. To broaden the scope of this method, compounds which are deemed promiscuous based on their bioactivity values could be docked into the respective crystal structures of their targets and 'promiscuity shapes' based on these predicted binding modes could be constructed.

Secondly, kinase ligands and their interaction points with the kinase could be converted into their pharmacophoric features, then annotated with binding affinity and used to construct a distance model similarly to Tanramluk *et al.*. This distance model could be further divided into different subparts to identify the type or number of features in a given subpocket according to their correlation with binding affinity (Tanramluk *et al.*).

I also showed that strong and weak binders cannot be separated based solely on their absence or presence of protein-ligand binding interactions using RF models. While this can be attributed to the low amount of structural data, this might also agree with the observation that the binding affinity of Staurosporine depends on the number of interactions formed in a particular kinase subpocket. However, this highlights that other types of interaction features might be used in a ML model for a given target to differentiate between different binding affinity ranges, such as the number of interactions in a given subpocket (Tanramluk et al., 2009).

I also analysed the use of the CRANkS interaction maps to classify ligands within binding affinity ranges, as this would allow the classification of compounds based solely on interaction features, independently of any chemical features, and would also allow immediate visualization of key binding features. First tests showed that re-docked and cross-docked ligands differ only subtly in terms of their CRANkS interactions scores, more future work is needed. Firstly, different methods of docking, such as matched molecular pairs or a smaller search box, should be investigated. Similarly, the number of kinase targets needed to construct a grid, their alignment and the similarity of the ligands should be further investigated for their use in deriving kinase-specific interaction maps (Patel, 2018). Particularly, the choice of threshold for judging the quality of alignment is promising as my chosen threshold was a compromise to have a high number of structures for a given grid (thus guaranteeing a high number of examples to calculate interactions from), while still having a very conservative threshold. If successful, weighted CRANkS

Interaction maps could be use to score interactions according to their importance, or be used additionally to model the flexibility of residues to allow for visual interpretation of the importance of a given interaction, a given chemical moiety and/or a given residue. Similarly, AutoCRANkS, a combination of AutoDock and CRANkS could be adopted for a docking procedure to filter out either potentially promiscuous or selective candidate ligands for a given kinase target (Patel, 2018).

However, there is the challenge of small amount of available structural and binding affinity data. For example, only 11 crystal structures with binding affinity values were available for Staurosporine, if the PDBBind Database was used for binding affinity annotation, and the kinase conformation was DFG-in, αC helix in.
This challenge can be circumvented as follows: firstly, databases other than PDBbind can be used to annotate ligands in crystal structures with binding affinity, *e.g.* the Fulle Dataset (Section 3.3.1). Then, if crystal structures are not available for compounds, compounds with an experimental binding affinity values could be docked into the binding pockets to derive potential 3D-binding poses. Here, emphasis should be placed on docking promiscuous compounds, especially if the compounds are promiscuous over a wide range of kinase targets from different target families so as to gain an understanding of which residues, which binding shapes and which pharmacophoric features are fundamental for filtering out promiscuous compounds.

6 Concluding Remarks

In this thesis, I have investigated the concept 'selectivity' both from a biological pathway and a protein-ligand interaction perspective.

To the best of my knowledge, I am the first to analyze drug targets and drugs from a pathway-centric perspective. My work showed that drug targets are associated with a wider range of pathways than 'undrugged' proteins, thus highlighting that protein targets with a high number of associated pathways should be tested for the nature of their pathway connections, not the overall number. Secondly, results indicate that a pathway can be modulated using a variety of different protein target classes; thus the focus on prominent target classes (the current 'target class'-centric approach in drug development) might result in false leads. Additionally, by connecting drug targets, and pathways and disease indications (via the ATC codes of the drugs), I was able to highlight that that drug targets can achieve their effect through their involvement in multiple pathways. Also, one target can dominate one pathway and be targeted for multiple disease areas (*e.g.* Prostaglandin G/H Synthase 2, UniProt ID: P35354). However, this might again represent a research bias as novel drugs might be developed against a proven drug target.

However, these analyses were complicated by several challenges. Firstly, the representation of the connection between proteins and entities in a pathway is an 'approximation' of a biological system based on the 'best guesses' of the annotating biologist. Additionally, biases in research on targets (and pathways) will lead to both well-explored targets and targets with no experimental information ('local gaps in knowledge'). Hence, research bias will also be present in my work and will distort any conclusions about the relationship between (drug) proteins, their drugs/compounds and the corresponding pathways. Similarly, the definition of a pathway can differ between different databases.

Additionally, pathways are organised both in a hierachical system in graphs. As such,they are challenging to simplify. While Bajorath *et al.* was able to quantify the polypharmacology of a drug and a non-drug compound by counting their respective targets, counting the number of implicated pathways is less straightforward as the hierachical nature of pathways allows for one protein to be contained in multiple subpathways and the assignment of a target to a pathway is based on the expertise of the curator.

Future work should focus on the position of drug targets in pathways and the nature of the associated pathways, in particular whether drug targets are preferably found downstream, and targets for withdrawn drugs upstream in a pathway, as well as whether specific (sub)pathways are more prone to be connected with targets associated with side effects.

Secondly, I have investigated the prediction of selective compounds (and associated limitations) from a 2D-data perspective and a 3D-structural perspective. The ability to predict whether a compound will be selective against a target (or promiscuous against a target panel) is important for both predicting the on- and off-target effects of a drug and designing chemical probes to understand biological systems. I show that the current 2D-models can predict the activity of selective compound-target pairs depending on input data.

Two aspects were particularly challenging in this analysis. First, the available data was heavily biased towards inactive compound-target combinations, which resulted fewer examples for active compound-target combinations when training RF models. Secondly, the available data consisted of different data types and differing experimental conditions. This heterogeneity rendered a standardization of compound-target activities difficult. Additionally, collecting a dataset of selective compounds was challenging due to the different focus of deriving chemical probes in the two different databases, chemicalprobes and SGC. Also, downloading additional experimental data for these compounds was not possible due to the set-up of the databases.
Lastly, to truly predict selective compounds, the binding affinity data should be separated into a higher number bins of activity bins in RF models (or ML set-ups in general) to account for the change in selectivity for a compound at different concentrations. However, due to the heterogeneity and uneven distribution of the dataset, training a multi-classification model for this purpose was not possible.

I also analysed the use of simple flexibility descriptor to investigate whether the movement of a residue can be used to improve a machine learning model. I chose the target class 'kinase' as a test set due to the availability of structural and bioactivity data and their involvement in diseases such as cancer and Parkinson's. Interestingly, flexible regions (G-loop, activation loop, DFG-motif) within a kinase protein binding pocket can

be identified. After building RF models with and without flexibility descriptors as features, I could show that flexibility descriptors for a kinase binding pocket contribute in the range of amino acid descriptors. However, their inclusion does not offer a significant improvement over current state of the art models. As my kinase RF models show already high performance metrices, it might be that there is no further room for improvement. Thus, in future experiments, the focus should be on constructing poorer performing models as baseline models to identify improvements. Lastly, other flexibility measures, *e.g.* Molecular Dynamics Simulation to explore gain the full potential range of motion of a residue at a given position, in combination with other machine learning models, such as Support Vector Machines, might offer a definite improvement and might be explored in the future.

There are two main caveats with this analysis. The normalized B-Factor and the RMSD values depend on the alignment of the proteins and the calculation, thus introducing a systematic error. The normalized B-Factors includes experimental uncertainty as it depends on the refinement process chosen by the crystallographer. Additionally, crystal contacts and protein-ligand interaction 'lock' the position of interacting residue atoms and their corresponding ligand atoms in position, resulting in lower B-Factors. Hence, other methods to ascertain the movement of a residue at a given position, such as Molecular Dynamics.

Lastly, I investigated whether kinase structural data could be used to construct binding interaction maps in different binding affinity ranges, against which a candidate compound could be screened; however neither interaction maps nor random forest models could pick up discriminating features.
Here, I had the advantage of having a high number of high-resolution kinase structures with binding affinity data. However, in terms of high resolution structures for one ligand bound to different kinases with associated binding affinities, I was severely limited. For example, only 11 crystal structures with binding affinity values were available for Staurosporine after data collection. Hence, it was not possible to systematically compare the binding information of a promiscuous compound across kinases in terms of their binding affinity values. More information might be available by using other databases for annotation; however, the information contained in these databases might be inconsistent in terms of binding activity, experimental conditions and measurement types. Addition-

ally, they might not cover a wide variety of different chemotypes. To circumvent these challenges, future research might focus on docking structurally distinct compounds with a confirmed binding affinity to derive potential 3D-binding poses.

I showed that a 2D interaction pattern in a kinase binding pocket, combined with a RF model, does not allow the separation of strong and weak binders based on their absence or presence of protein-ligand binding interactions. However, this could indicate that the number and types of interactions in a certain kinase subpocket is more important than the actual interaction types alone. Future research could focus on dividing kinase binding pockets into multiple subpockets to use the number and kind of each interaction in each binding pocket as predictive features for binding strength, similarly to Tanramluk *et al* who showed that the binding affinity of Staurosporine depends on the number of interactions formed in a particular kinase subpocket.

To further investigate whether a 3D approach might yield better insights into separating strong and weak binders, I used the CRANkS interaction maps to classify ligands within binding affinity ranges. Using interaction grids sorely based on pharmacophoric representaton of known interactions would allow the classification of ligands sorely on the basis of interaction features, independently of chemical features, and offer an immediate visual inspection of important binding interactions for a structural biologist. If successful, they could have been used in scoring the importance of interactions and been combined with residue flexibility and mutation information to compare different kinases in terms of interactions. Mutations cannot only alter the function of a protein, but also change the selectivity of a compound towards closely-related targets, as the presence/absence of specific (key) interaction can increase or decrease the binding activity of a compound for this target. Hence, using 3D-approaches, it would be interesting to compare kinases directly in terms of the presence/absence of key binding interactions for a given compound compared to related proteins. This is particularly interesting in view of the evolutionary relationship between kinases, highlighted by the high level of conservation between and within species, orthologues and paralogues using the DYRK kinases (Aranda et al., 2010). Manning *et al.* successfully used sequence similarity to cluster kinases into different kinase families (Manning et al., 2002). Additionally, the 3D analysis of protein-ligand kinase interaction Understanding how subtle differences in kinase structure and sequence can be connected to chemical features of a compound

to derive selective compounds (or understand promiscuity) is particularly interesting in view of the high sequence similarity within kinase families, which renders the prediction or design of a selective drug difficult. In this context, future work might include the use of AutoCRANkS, a combination of AutoDock and CRANkS, for a docking procedure in which potentially promiscuous or selective candidate ligands for a given kinase target could be filtered out based on their individual interactions (Patel, 2018).

However, my first tests showed that there is only a slight difference between the CRANkS interactions scores of re-docked and cross-docked ligands, which might be due to the type of docking or the construction of the grid. However, I used a very conservative threshold for the alignment to guarantee a good overlay of the protein-ligand interactions which are used for building the interaction grids. Hence, a further investigation of different docking methods, the use of matched molecular pairs (Patel, 2018) or the use of a smaller search box, might be future starting points to further explore the use of CRANkS maps in specifying individual interactions responsible for the promiscuity or selectivity of a compound towards a given target panel.

However, I showed that compounds selective against closely related targets and promiscuous ligands tend to have a conserved binding mode across targets. This is unsurprising in view of the evolutionary relationship between kinases and their substrates, highlighting the challenge of deriving selective compounds against such closely related targets.

In summary, the 3D-structural route to derive selective compounds (or filter out promiscuous compounds) should be fully explored by the following strategies: features that differentiate between strong, medium and weak binders should be used to separate compounds according to predicted interactions (*e.g.* using docking). Alternatively, structural 3D-motifs of promiscuous compounds could be used in a structure-based ligand filter to filter out potentially promiscuous or selective candidate compounds early on in the drug discovery process, such as Ligity (Patel, 2018) or AutoCRANkS (Patel, 2018). However, analyses will be limited by the lack of binding available data.

References

M. Ahinko, S. Kurkinen, S. Niinivehmas, O. Pentikäinen, and P. Postila. A practical perspective: The effect of ligand conformers on the negative image-based screening. *International Journal of Molecular Sciences*, 20(11):2779, 2019.

G. Ai, C. Tian, D. Deng, G. Fida, H. Chen, Y. Ma, L. Ding, and Y. Gu. A combination of 2d similarity search, pharmacophore, and molecular docking techniques for the identification of vascular endothelial growth factor receptor-2 inhibitors. *Anti-Cancer Drugs*, 26(4):399–409, 2015.

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

H. Akoglu. User's guide to correlation coefficients. *Turkish Journal of Emergency Medicine*, 18:91–93, 2018.

A. Antolin, J. E. Tym, A. Komianou, I. Collins, P. Workman, and B. Al-Lazikani. Objective, quantitative, data-driven assessment of chemical probes. *Cell Chemical Biology*, 25(2):194–205, 2017.

S. Aranda, A. Laguna, and S. Luna. Dyrk family of protein kinases: Evolutionary relationships, biochemical properties, and functional roles. *FASEB Journal*, 25:449–62, 11 2010.

C. H. Arrowsmith, J. E. Audia, C. Austin, J. Baell, J. Bennett, J. Blagg, C. Bountra, P. E. Brennan, P. J. Brown, M. E. Bunnage, C. Buser-Doepner, R. M. Campbell, A. J. Carter, P. Cohen, R. A. Copeland, B. Cravatt, J. L. Dahlin, D. Dhanak, A. M. Edwards, M. Frederiksen, S. V. Frye, N. Gray, C. E. Grimshaw, D. Hepworth, T. Howe, K. V. M. Huber, J. Jin, S. Knapp, J. D. Kotz, R. G. Kruger, D. Lowe, M. M. Mader, B. Marsden, A. Mueller-Fahrnow, S. Müller, R. C. O'Hagan, J. P. Overington, D. R. Owen, S. H. Rosenberg, R. Ross, B. Roth, M. Schapira, S. L. Schreiber, B. Shoichet, M. Sundström, G. Superti-Furga, J. Taunton, L. Toledo-Sherman, C. Walpole, M. A.

Walters, T. M. Willson, P. Workman, R. N. Young, and W. J. Zuercher. The promise and peril of chemical probes. *Nature Chemical Biology*, 11(8):536–541, 2015.

P. Ascierto, J. Kirkwood, J.-J. Grob, E. Simeone, A. Grimaldi, M. Maio, G. Palmieri, A. Testori, and F. Marincola. The role of braf v600 mutation in melanoma. *Journal of Translational Medicine*, 10:85, 2012.

J. Baell and M. A. Walters. Chemistry: Chemical con artists foil drug discovery. *Nature*, 513(7519):481–483, 2014.

J. B. Baell and G. A. Holloway. New Substructure Filters for Removal of Pan Assay Interference Compounds (PAINS) from Screening Libraries and for Their Exclusion in Bioassays. *Journal of Medicinal Chemistry*, 53:2719–2740, 2010.

M. Barley, N. Turner, and R. Goodacre. Improved descriptors for the QSAR modeling of peptides and proteins. *Journal of Chemical Information and Modeling*, 58(2):234–243, 2018.

K. Barnash, L. I James, and S. V Frye. Target class drug discovery. *Nature Chemical Biology*, 13:1053–1056, 2017.

F. Belinky, N. Nativ, G. Stelzer, S. Zimmerman, T. I. Stein, M. Safran, and D. Lancet. Pathcards: multi-source consolidation of human biological pathways. *Database : The Journal of Biological Databases and Curation*, 2015, 2015.

E. L. Berg. Systems biology in drug discovery and development. *Drug Discovery Today*, 19(2):113–125, 2014.

S. Berger and R. Iyengar. Role of systems pharmacology in understanding drug adverse events. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, 3:129–35, 2011.

F. Bernstein, T. Koetzle, G. Williams, E. Meyer, M. Brice, J. Rodgers, O. Kennard, T. Shimanouchi, and M. Tasumi. The protein data bank: A computer-based archival file for macromolecular structures. *Journal of Molecular Biology*, 112:535—542, 1977.

K. Bhullar, N. Orrego-Lagarón, E. Mcgowan, I. Parmar, A. Jha, B. Hubbard, and H. P. V. Rupasinghe. Kinase-targeted cancer therapies: Progress, challenges and future directions. *Molecular Cancer*, 17(3):1–44, 2017.

R. Bohacek, C. McMartin, and W. Guida. The art and practice of structure-based drug design: A molecular modeling perspective. *Medicinal Research Reviews*, 16:3 – 50, 1996.

N. Bosc, C. Meyer, and P. Bonnet. The use of novel selectivity metrics in kinase research. *BMC Bioinformatics*, 18(1):17, 2017.

F. Boyles, C. Deane, and G. Morris. Learning from the ligand: Using ligand-based features to improve binding affinity prediction. *Bioinformatics*, 36(3):758–764, 2019.

A. R. Bradley, I. D. Wall, D. V. S. Green, C. M. Deane, and B. D. Marsden. Oommppaa: A tool to aid directed synthesis by the combined analysis of activity and structural data. *Journal of Chemical Information and Modeling*, 54(10):2636–2646, 2014.

D. Brossart, V. Laird, and T. Armstrong. Interpreting kendall's tau and tau-u for single-case experimental designs. *Cogent Psychology*, 5:1–26, 2018.

S. C. Bull and A. J. Doig. Properties of protein drug target classes. *PlOS ONE*, 10(3): 535—542, 2015.

K. V. Butler, I. A. Macdonald, N. A. Hathaway, and J. Jin. Report and Application of a Tool Compound Data Set. *Journal of Chemical Information and Modeling*, 57:2699 2706, 2017.

O. Carugo. How large B-factors can be in protein crystal structures. *BMC Bioinformatics*, 19(1):1–9, 2018.

D. Chasman, A. Fotuhi Siahpirani, and S. Roy. Network-based approaches for analysis of complex biological systems. *Current Opinion In Biotechnology*, 39:157–166, 2016.

Chemical Probe Portals Reviewers. Hg-9-91-01, 2020. Available at `https://www.chemicalprobes.org/hg-9-91-01`; Last accessed on 01.01.2020.

Chemical Probes Portal. Increasing the quality and reproducibility of biomedical research using chemical probes, 2020. Available at `https://www.chemicalprobes.org/`, Last accessed on 01.01.2020.

C. Chen and L. Breiman. Using random forest to learn imbalanced data. *Conference Proceedings, University of California, Berkeley*, pages 1–12, 2004.

D. Chicco. Ten quick tips for machine learning in computational biology. *BioData Mining*, 10:35, 2017.

L. Chico, L. Van Eldik, and D. Watterson. Targeting protein kinases in central nervous system disorders. *Nature Reviews: Drug Discovery*, 8:892–909, 2009.

S. Chowdhury and R. R. Sarkar. Comparison of human cell signaling pathway databases—evolution, drawbacks and challenges. *Database*, 28, 2015.

S. Christmann-Franck, G. J.P. van Westen, G. Papadatos, F. Beltran Escudie, A. Roberts, J. Overington, and D. Domine. An unprecedentedly large-scale kinase inhibitor set enabling the accurate prediction of compound-kinase activities: a way towards selective promiscuity by design? *Journal of Chemical Information and Modeling*, 56:1654–1675, 2016.

J. Cicenas, E. Zalyte, A. Bairoch, and P. Gaudet. Kinases and cancer. *Cancers*, 10(3)): 63, 2018.

P. Cock, T. Antao, J. Chang, B. Chapman, C. Cox, A. Dalke, I. Friedberg, T. Hamelryck, F. Kauff, B. Wilczynski, and M. de Hoon. Biopython: Freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics (Oxford, England)*, 25:1422–3, 2009.

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

I. Cortes-Ciriano, D. S. Murrell, G. J. Van Westen, A. Bender, and T. E. Malliavin. Prediction of the potency of mammalian cyclooxygenase inhibitors with ensemble proteochemometric modeling. *Journal of Cheminformatics*, 7(1):1–18, 2015.

P. Creixell, J. Reimand, S. Haider, G. Wu, T. Shibata, M. Vazquez, V. Mustonen, A. Gonzalez-Perez, J. Pearson, C. Sander, B. Raphael, D. Marks, F. Ouellette, A. Valencia, G. Bader, P. Boutros, J. Stuart, R. Linding, N. López-Bigas, and t. Consequences. Pathway and network analysis of cancer genomes. *Nature Methods*, 12: 615–621, 2015.

M. Cruz-Monteagudo, J. Medina-Franco, Y. Perez Castillo, O. Nicolotti, N. Cordeiro, and F. Borges. Activity cliffs in drug discovery: Dr jekyll or mr hyde? *Drug Discovery Today*, 19(8):1069–1080, 2014.

A. Dinno. Nonparametric pairwise multiple comparisons in independent groups using dunn's test. *Stata Journal*, 15(1):292–300, 2015.

C. Dormann, J. Elith, S. Bacher, C. Buchmann, G. Carl, G. Carré, J. García Márquez, B. Gruber, B. Lafourcade, P. Leitão, T. Münkemüller, C. Mcclean, P. Osborne, B. Reineking, B. Schröder, A. Skidmore, D. Zurell, and S. Lautenback. Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography*, 36:27–46, 2013.

M. Duran-Frigola and P. Aloy. Analysis of Chemical and Biological Features Yields Mechanistic Insights into Drug Side Effects. *Chemistry & Biology*, 20(4):594–603, 2013.

J.-P. Ebejer, P. Finn, W. K. Wong, C. Deane, and G. Morris. Ligity: A non-superpositional, knowledge-based approach to virtual screening. *Journal of chemical information and modeling*, 59, 05 2019.

D. Edelmann, K. Fokianos, and M. Pitsillou. An updated literature review of distance correlation and its applications to time series. *International Statistical Review*, 87: 237–262, 2019.

C. Ehrt, T. Brinkjost, and O. Koch. A benchmark driven guide to binding site comparison: An exhaustive evaluation using tailor-made data sets (prospeccts). *PLOS Computational Biology*, 14(11):1–50, 2018.

A. Fabregat, S. Jupe, L. Matthews, K. Sidiropoulos, M. Gillespie, P. Garapati, R. Haw, B. Jassal, F. Korninger, B. May, M. Milacic, C. Duenas Roca, K. Rothfels, C. Sevilla, V. Shamovsky, S. Shorser, T. Varusai, G. Viteri, J. Weiser, and P. D'Eustachio. The Reactome Pathway Knowledgebase. *Nucleic Acids Research*, 46(D1):D481–7, 2017.

W. Filho. Distcorr.py. `https://gist.github.com/wladston/c931b1495184fbb99bec`, 2015.

S. Fletcher and M. Islam. Comparing sets of patterns with the jaccard index. *Australasian Journal of Information Systems*, 22:1–16, 2018.

A. Fountas, L. Diamantopoulos, and A. Tsatsoulis. Tyrosine kinase inhibitors and diabetes: A novel treatment paradigm? *Trends in Endocrinology And Metabolism: TEM*, 26, 2015.

S. Fulle. Profiling and prediction of kinase inhibitors, 2019. Available at `https://github.com/Team-SKI/Publications/tree/master/Profiling_prediction_of_kinase_inhibitors`, Lastly accessed on 14.10.2019.

S. Fulle. List of human kinases, 2020. Available at `http://kinhub.org/kinases.html`; Last accessed on 01.01.2020.

M. Gao and J. Skolnick. A Comprehensive Survey of Small-Molecule Binding Pockets in Proteins. *PLoS Computational Biology*, 9(10):e1003302, 2013.

A. Garcia Asuero, A. Sayago, and G. González. The correlation coefficient: An overview. *Critical Reviews in Analytical Chemistry*, 36:41–59, 2006.

I. Gashaw, P. Ellinghaus, A. Sommer, and K. Asadullah. What makes a good drug target? *Drug Discovery Today*, 16:1037–43, 2011.

A. Gaulton, A. Hersey, M. L. Nowotka, A. Patricia Bento, J. Chambers, D. Mendez, P. Mutowo, F. Atkinson, L. J. Bellis, E. Cibrian-Uhalte, M. Davies, N. Dedman, A. Karlsson, M. P. Magarinos, J. P. Overington, G. Papadatos, I. Smit, and A. R. Leach. The ChEMBL database in 2017. *Nucleic Acids Research*, 45(D1):D945–D954, 2017.

D. Gfeller, A. Grosdidier, M. Wirth, A. Daina, O. Michielin, and V. Zoete. Swisstargetprediction: A web server for target prediction of bioactive small molecules. *Nucleic Acids Research*, 42, 05 2014.

K. Giblin, S. Hughes, H. Boyd, P. Hansson, and A. Bender. Prospectively Validated Proteochemometric Models for the Prediction of Small Molecule Binding to Bromodomain Proteins. *Journal of Chemical Information and Modeling*, 58(9):1870–1888, 2018.

W. Goh and L. Wong. Integrating networks and proteomics: Moving forward. *Trends in Biotechnology*, 34:051–959, 2016.

E. Gregori-Puigjane, V. Setola, J. Hert, B. A. Crews, J. J. Irwin, E. Lounkine, L. Mar-

nett, B. L. Roth, and B. K. Shoichet. Identifying mechanism-of-action targets for drugs and probes. *Proceedings of the National Academy of Sciences*, 109(28):11178–11183, 2012.

T. Hamelryck and B. Manderick. Pdb file parser and structure class implemented in python. *Bioinformatics*, 19:2308–10, 2003.

M. K. Hancock, C. S. Lebakken, J. Wang, and K. Bi. Multi-pathway cellular analysis of compound selectivity. *Molecular BioSystems*, 6(10):1834–1843, 2010.

V. J. Haupt, S. Daminelli, and M. Schroeder. Drug Promiscuity in PDB: Protein Binding Site Similarity Is Key. *PLOS ONE*, 8(6):1–15, 2013.

W. A. Haynes, A. Tomczak, and P. Khatri. Gene annotation bias impedes biomedical research. *Scientific Reports*, 8(1):1362, 2018.

J. Hert, J. J. Irwin, C. Laggner, M. J. Keiser, and B. K. Shoichet. Quantifying biogenic bias in screening libraries. *Nature Chemical Biology*, 5(7):479–483, 2009.

A. Heureux, K. Grolinger, and M. Capretz. Machine learning with big data: Challenges and approaches. *IEEE Access*, 5:7776–7797, 2017.

A. L. Hopkins and C. R. Groom. The druggable genome. *Nature Reviews Drug Discovery*, 1(9):727–730, 2002.

Y. Hu and J. Bajorath. Entering the 'big data' era in medicinal chemistry: molecular promiscuity analysis revisited. *Future Science OA*, 3:FSO179, 2017.

Y. Hu, D. Stumpfe, and J. Bajorath. Recent advances in scaffold hopping. *Journal of Medicinal Chemistry*, 60(4):1238–1246, 2016.

D. Huggins, W. Sherman, and B. Tidor. Rational approaches to improving selectivity in drug design. *Journal of Medicinal Chemistry*, 55:1424–44, 02 2012.

S. Jasial, Y. Hu, and J. Bajorath. Determining the degree of promiscuity of extensively assayed compounds. *PLOS ONE*, 11(4):1–15, 2016.

S. Jasial, Y. Hu, and J. Bajorath. How Frequently Are Pan-Assay Interference Compounds Active? Large-Scale Analysis of Screening Data Reveals Diverse Activity Profiles, Low Global Hit Frequency, and Many Consistently Inactive Compounds. *Journal of Medicinal Chemistry*, 60(9):3879–3886, 2017.

P. Jiang and S. Liu. Big data mining yields novel insights on cancer. *Nature Genetics*, 47:103–4, 2015.

E. Jones, T. Oliphant, and P. Peterson. Scipy: Open source scientific tools for python. Available at `http://www.scipy.org`, Lastly accessed on 20.01.2020.

M. M. Jones, S. Castle-Clarke, D. Brooker, E. N. Nason, F. H. Huzair, and J. Chataway. The Structural Genomics Consortium. *RAND Health Quarterly*, 4(3):1–12, 2014.

H. Jubb, A. Higueruelo, B. Ochoa-Montaño, W. Pitt, D. Ascher, and T. Blundell. Arpeggio: A web server for calculating and visualising interatomic interactions in protein structures. *Journal of Molecular Biology*, 429, 12 2016.

M. Kanehisa, M. Furumichi, M. Tanabe, Y. Sato, and K. Morishima. Kegg: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Research*, 45(D1): D353–D361, 2016.

M. W. Karaman, S. Herrgard, D. K. Treiber, P. Gallant, C. E. Atteridge, B. T. Campbell, K. W. Chan, P. Ciceri, M. I. Davis, P. T. Edeen, R. Faraoni, M. Floyd, J. P. Hunt, D. J. Lockhart, Z. V. Milanov, M. J. Morrison, G. Pallares, H. K. Patel, S. Pritchard, L. M. Wodicka, and P. P. Zarrinkar. A quantitative analysis of kinase inhibitor selectivity. *Nature Biotechnology*, 26(1):127–132, 2008.

M. J. Keiser, B. L. Roth, B. N. Armbruster, P. Ernsberger, J. J. Irwin, and B. K. Shoichet. Relating protein pharmacology by ligand chemistry. *Nature Biotechnology*, 25(2):197–206, 2007.

T. Kelder, M. Van Iersel, K. Hanspers, M. Kutmon, B. Conklin, C. Evelo, and A. Pico. Wikipathways: Building research communities on biological pathways. *Nucleic Acids Research*, 40:D1301–7, 2011.

E. Kellenberger, C. Schalon, and D. Rognan. How to measure the similarity between protein ligand-binding sites? *Current Computer - Aided Drug Design*, 4:209–220, 2008.

Kinexus BioInformatics Corporation. Kinasenet - human protein kinase knowledgebase - brk, 2019a. Accessible at `http://www.kinasenet.ca/showProtein?query=Q13882`; Last accessed on 22.09.2019.

Kinexus BioInformatics Corporation. Kinasenet - human protein kinase knowledgebase - btk, 2019b. Accessible at http://www.kinasenet.ca/showProtein?query=Q06187; Last accessed on 22.09.2019.

C. Kingsford and S. Salzberg. What are decision trees? *Nature Biotechnology*, 26: 1011–3, 2008.

D. R. Koes, M. P. Baumgartner, and C. J. Camacho. Lessons learned in empirical scoring with smina from the csar 2011 benchmarking exercise. *Journal of Chemical Information and Modeling*, 53(8):1893–1904, 2013.

A. Kooistra, G. Kanev, O. P.J. van Linden, R. Leurs, I. J.P. de Esch, and C. de Graaf. KLIFS a structural kinase-ligand interaction database. *Nucleic Acids Research*, 44: :D365–71, 10 2015.

G. Koscielny, P. An, D. Carvalho-Silva, J. Cham, L. Fumis, R. Gasparyan, S. Hasan, N. Karamanis, M. Maguire, E. Papa, A. Pierleoni, M. Pignatelli, T. Platt, F. Rowland, P. Wankar, A. Patrícia Bento, T. Burdett, A. Fabregat, S. Forbes, and I. Dunham. Open targets: A platform for therapeutic target identification and validation. *Nucleic Acids Research*, 45, 2016.

A. Kryshtafovych and K. Fidelis. Protein structure prediction and model quality assessment. *Drug Discovery Today*, 14:386–93, 2009.

I. Kufareva and R. Abagyan. Methods of protein structure comparison. *Methods in Molecular Biology*, 857:231–57, 01 2012.

M. Kutmon, A. Riutta, N. Nunes, K. Hanspers, E. L. Willighagen, A. Bohler, J. Mélius, A. Waagmeester, S. R. Sinha, R. Miller, and et al. Wikipathways: capturing the full diversity of pathway knowledge. *Nucleic Acids Research*, 44(D1), 2015.

G. Landrum. Rdkit: Open-source cheminformatics, 2020. Available at http://www.rdkit.org, Last accessed on 02.01.2020.

S. Leung, M. Bodkin, F. Delft, P. Brennan, and G. Morris. Sucos is better than rmsd for evaluating fragment elaboration and docking poses. *ChemRxiv*, https://chemrxiv.org/articles/SuCOS_is_Better_than_RMSD_for_Evaluating_Fragment_Elaboration_and_Docking_Poses/8100203/1, Deposited on 10.05.2019.

S. Liggi, G. Drakakis, A. Hendry, K. M. Hanson, S. C. Brewerton, G. Wheeler, M. Bodkin, D. A. Evans, and A. Bender. Extensions to in silico bioactivity predictions using pathway annotations and differential pharmacology analysis: Application to xenopus laevis phenotypic readouts. *Molecular Informatics*, 32:1009–1024, 2013.

Q. Liu, Z. Li, and J. Li. Use b-factor related features for accurate classification between protein binding interfaces and crystal packing contacts. *BMC Bioinformatics*, 15: 1–11, 2014.

M. Majewski, S. Ruiz Carmona, and X. Barril. An investigation of structural stability in protein-ligand complexes reveals the balance between order and disorder. *Communications Chemistry*, 2:1–8, 2019.

G. Manning, D. B. Whyte, R. Martinez, T. Hunter, and S. Sudarsanam. The protein kinase complement of the human genome. *Science*, 298(5600):1912–1934, 2002.

E. Martínez-Gómez, M. Richards, and D. Richards. Distance correlation methods for discovering associations in large astrophysical databases. *The Astrophysical Journal*, 781:1–11, 2013.

S. Mathew and J. Pandian. Newer insights to the neurological diseases among biblical characters of old testament. *Annals of Indian Academy of Neurology*, 13:164–6, 07 2010.

MedChemExpress. Hg-9-91-01, 2020. Available at `https://www.medchemexpress.com/HG-9-91-01.html`; Last accessed on 01.01.2020.

B. Merget, S. Turk, S. Eid, F. Rippmann, and S. Fulle. Profiling prediction of kinase inhibitors: Toward the virtual assay. *Journal of Medicinal Chemistry*, 60:474–485, 2016.

E. Merritt. To b or not to b: a question of resolution? *Acta Crystallographica: Section D, Biological Crystallography*, 68:468–77, 2012.

L. H. Mervin, A. M. Afzal, G. Drakakis, R. Lewis, O. Engkvist, and A. Bender. Target prediction utilising negative bioactivity data covering large chemical space. *Journal of Cheminformatics*, 7(1):1–16, 2015.

L. H. Mervin, K. C. Bulusu, L. Kalash, A. M. Afzal, F. Svensson, M. A. Firth, I. Barrett,

O. Engkvist, and A. Bender. Orthologue chemical space and its influence on target prediction. *Bioinformatics*, 34(1):72–79, 2018.

P. Mishra and S. Günther. New insights into the structural dynamics of the kinase jnk3. *Scientific Reports*, 8(1):9435, 2018.

S. Mukherjee. *The Emperor of All Maladies: A Biography of Cancer*. Scribner, New York, 2010.

R. Najmanovich, J. Kuttner, V. Sobolev, and M. Edelman. Side-chain flexibility in proteins upon ligand binding. *Proteins: Structure, Function, and Bioinformatics*, 39: 261–268, 2000.

E. Ostertagova, O. Ostertag, and J. Kováč. Methodology and application of the kruskal-wallis test. *Applied Mechanics and Materials*, 611:115–120, 2014.

O. P. J. van Linden, A. Kooistra, R. Leurs, I. J. P. de Esch, and C. de Graaf. KLIFS: A knowledge-based structural database to navigate kinase-ligand interaction space. *Journal of Medicinal Chemistry*, 57:249–277, 2013.

M. Padgavankar and S. Gupta. Big data storage and challenges. *International Journal of Computer Science and Information Technologies*, 3:2218–2223, 2014.

S. Paricharak, I. Cortés-Ciriano, A. P. Ijzerman, T. E. Malliavin, and A. Bender. Proteochemometric modelling coupled to in silico target prediction: An integrated approach for the simultaneous prediction of polypharmacology and binding affinity/potency of small molecules. *Journal of Cheminformatics*, 7(1):1–11, 2015.

S. Parthasarathy and M. Murthy. Analysis of temperature factor distribution in high-resolution protein structures. *Protein Science*, 6:2561–7, 2008.

H. J. Patel. *Prioritisation of active compounds that are novel scaffolds using a data-driven approach*. PhD thesis, University of Oxford, 2018.

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

A. Pico, T. Kelder, M. Van Iersel, K. Hanspers, B. Conklin, and C. Evelo. Wikipathways: Pathway editing for the people. *PLOS Biology*, 6:e184, 2008.

C. P. Portal. Tofacitinib. Available at `https://www.chemicalprobes.org/tofacitinib/`, Last accessed on 27.02.2020.

D. Powers. The problem of area under the curve. *International Conference on Information Science and Technology (ICIST) IEEE*, pages 567–573, 2012.

M. Rask-Andersen, M. Almén, and H. Schiöth. Trends in the exploitation of novel drug targets. *Nature reviews Drug Discovery*, 10:579–90, 2011.

F. Rossari, F. Minutolo, and E. Orciuolo. Past, present, and future of bcr-abl inhibitors: From chemical development to clinical efficacy. *Journal of Hematology Oncology*, 11: 1–14, 2018.

L. Sach-Peltason. personal communication, 2018.

S. Salentin, S. Schreiber, V. Haupt, M. Adasme, and M. Schroeder. Plip: Fully automated protein-ligand interaction profiler. *Nucleic Acids Research*, 43, 04 2015.

S. A. Sam, J. Teel, A. N. Tegge, A. Bharadwaj, and T. Murali. Xtalkdb: a database of signaling pathway crosstalk. *Nucleic Acids Research*, 45(D1):432–439, 2016.

M. Sandberg, L. Eriksson, J. Jonsson, M. Sjöström, and S. Wold. New chemical descriptors relevant for the design of biologically active peptides. A multivariate characterization of 87 amino acids. *Journal of Medicinal Chemistry*, 41(14):2481–2491, 1998.

R. Santos, O. Ursu, A. Gaulton, A. Patrícia Bento, R. S. Donadi, C. Bologa, A. Karlsson, B. Al-Lazikani, A. Hersey, T. Oprea, and J. Overington. A comprehensive map of molecular drug targets. *Nature Reviews Drug Discovery*, 16:19–34, 2016.

A. P. Saraswati, S. A. Hussaini, N. H. Krishna, B. N. Babu, and A. Kamal. Glycogen synthase kinase-3 and its inhibitors: Potential target for various therapeutic conditions. *European Journal of Medicinal Chemistry*, 144:843–858, 2018.

K. Sargsyan, G. Cédric, and C. Lim. How molecular size impacts RMSD applications in molecular dynamics simulations. *Journal of Chemical Theory and Computation*, 13 (4):1518–1524, 2017.

H.-C. Schneider and T. Klabunde. Understanding drugs and diseases by systems biology? *Bioorganic Medicinal Chemistry Letters*, 23(5):1168–1176, 2013.

P. Schneider and G. Schneider. Unveiling the target promiscuity of pharmacologically active compounds in silico. *Angewandte Chemie International Edition*, 56:11520–11524, 2017.

P. Schober, C. Boer, and L. Schwarte. Correlation coefficients: Appropriate use and interpretation. *Anesthesia Analgesia*, 126:1763–1768, 2018.

L. Schrödinger. The PyMOL Molecular Graphics System, version 1.8.

Scikit-learn developers. Precision-recall, 2020a. Available at `https://scikit-learn.org/stable/auto_examples/model_selection/plot_precision_recall.html`; Last accessed on 01.01.2020.

Scikit-learn developers. Sklearn.model_selection.StratifiedKFold, 2020b. Available at `https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.StratifiedKFold.html`; Last accessed on 01.01.2020.

J. Seidler, S. L. Mcgovern, T. N. Doman, and B. K. Shoichet. Identification and Prediction of Promiscuous Aggregating Inhibitors among Known Drugs. *Journal of Medicinal Chemistry*, 9(46):4477–4486, 2003.

R. Shahin, O. Shaheen, F. El-Dahiyat, M. Habash, and S. Saffour. Research advances in kinase enzymes and inhibitors for cardiovascular disease treatment. *Future Science OA*, 3(4):FSO204, 2017.

H.-P. Shih, X. Zhang, and A. Aronov. Drug discovery effectiveness from the standpoint of therapeutic mechanisms and indications. *Nature Reviews Drug Discovery*, 17(1): 19–33, 2017.

F. A. Sorgenfrei, S. Fulle, and B. Merget. Kinome-Wide Profiling Prediction of Small Molecules. *ChemMedChem*, 13(6):495–499, 2018.

A. Stank, D. Kokh, J. Fuller, and R. Wade. Protein binding pocket dynamics. *Accounts of Chemical Research*, 49(5):809–815, 2016.

Structural Genomics Consortium. Chemical probes, 2020. Available at `https://www.thesgc.org/chemical-probes`, Last accessed on 01.01.2020.

N. Sturm, J. Desaphy, R. J. Quinn, D. Rognan, and E. Kellenberger. Structural insights into the molecular basis of the ligand promiscuity. *Journal of Chemical Information and Modeling*, 52(9):2410–2421, 2012.

V. Subramanian, P. Prusis, L. O. Pietilä, H. Xhaard, and G. Wohlfahrt. Visually interpretable models of kinase selectivity related features derived from field-based proteochemometrics. *Journal of Chemical Information and Modeling*, 53(11):3021–3030, 2013.

V. Subramanian, Q. Ain, H. Henno, L.-O. Pietilä, J. E Fuchs, P. Prūsis, A. Bender, and G. Wohlfahrt. Implementation of 3D proteochemometrics: Using three-dimensional information of proteins and ligands to address aspects of the selectivity of serine proteases. *Medicinal Chemistry Communication*, 8(5):1037–1045, 2017.

J. Sun, L. Carlsson, E. Ahlberg, U. Norinder, O. Engkvist, and H. Chen. Applying Mondrian Cross-Conformal Prediction to Estimate Prediction Confidence on Large Imbalanced Bioactivity Data Sets. *Journal of Chemical Information and Modeling*, 57(7):1591–1598, 2017.

M. Swain. MolVS: Molecule validation and standardization, 2019. Available at `https://molvs.readthedocs.io/en/latest/`, Last accessed on 12.12.2019.

J. Tang, A. Szwajda, S. Kumar Shakyawar, T. Xu, P. Hintsanen, K. Wennerberg, and T. Aittokallio. Making sense of large-scale kinase inhibitor bioactivity data sets: A comparative and integrative analysis. *Journal Of Chemical Information And Modeling*, 54(3):735–43, 2014.

D. Tanramluk, A. Schreyer, W. R. Pitt, and T. L. Blundell. On the origins of enzyme inhibitor selectivity and promiscuity: A case study of protein kinase binding to staurosporine. *Chemical Biology and Drug Design*, 74(1):16–24, 2009.

R. Team. A language and environment for statistical computing. *Computing*, 1, 2006.

I. Tetko, O. Engkvist, U. Koch, J.-L. Reymond, and H. Chen. Bigchem: Challenges and opportunities for big data analysis in chemistry. *Molecular Informatics*, 35(11-12): 615–621, 2016.

The Gene Ontology Consortium. The gene ontology resource: 20 years and still going strong. *Nucleic Acids Research*, 47:D330–D338, 2018.

P. Thul and C. Lindskog. The human protein atlas - a spatial map of the human proteome. *Protein Science*, 27:233–244, 2017.

D. Türei, T. Korcsmáros, and J. Saez-Rodriguez. Omnipath: guidelines and gateway for literature-curated signaling pathway resources. *Nature Methods*, 13(12):966–967, 2016.

M. Uhlen, L. Fagerberg, B. M Hallström, C. Lindskog, P. Oksvold, A. Mardinoglu, Sivertsson, C. Kampf, E. Sjöstedt, A. Asplund, I. Olsson, K. Edlund, E. Lundberg, S. Navani, C. Al-Khalili Szigyarto, J. Odeberg, D. Djureinovic, J. Ottosson Takanen, S. Hober, and F. Pontén. Proteomics. tissue-based map of the human proteome. *Science*, 347(6220):1260419, 2015.

G. J. Van Westen, O. O. Van Den Hoven, R. Van Der Pijl, T. Mulder-Krieger, H. De Vries, J. K. Wegner, A. P. IJzerman, H. W. Van Vlijmen, and A. Bender. Identifying Novel Adenosine Receptor Ligands by Simultaneous Proteochemometric Modeling of Rat and Human Bioactivity Data. *Journal of Medicinal Chemistry*, 55(16):7010–7020, 2012.

S. Vilar, G. Cozza, and S. Moro. Medicinal chemistry and the molecular operating environment (MOE): Application of qsar and molecular docking to drug discovery. *Current topics in Medicinal Chemistry*, 8(18):1555–72, 2008.

M. Vogt, S. Jasial, and J. Bajorath. Computationally derived compound profiling matrices. *Future Science OA*, 4(8):FSO327, 2018.

A. Volkamer, S. Eid, S. Turk, F. Rippmann, and S. Fulle. Identification and Visualization of Kinase-Specific Subpockets. *Journal of Chemical Information and Modeling*, 56(2): 335–346, 2016.

Y. Wang, A. Cornett, F. J. King, Y. Mao, F. Nigsch, C. G. Paris, G. McAllister, and J. L. Jenkins. Evidence-Based and Quantitative Prioritization of Tool Compounds in Phenotypic Drug Discovery. *Cell Chemical Biology*, 23(7):862–874, 2016.

M. Waring, J. Arrowsmith, A. Leach, P. Leeson, S. Mandrell, R. Owen, G. Pairaudeau, W. Pennie, S. Pickett, J. Wang, O. Wallace, and A. Weir. An analysis of the attrition of drug candidates from four major pharmaceutical companies. *Nature Reviews. Drug Discovery*, 14:475–86, 2015.

D. S. Wishart, Y. D. Feunang, A. C. Guo, E. J. Lo, A. Marcu, J. R. Grant, T. Sajed, D. Johnson, C. Li, Z. Sayeeda, and et al. Drugbank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Research*, 46(D1):1074–1082, 2017.

M. Xu, L. Yu, B. Wan, L. Yu, and Q. Huang. Predicting inactive conformations of protein kinases using active structures: Conformational selection of type-II inhibitors. *PlOS ONE*, 6(7):e22644, 2011.

P. Zhang and Y. Itan. Biological network approaches and applications in rare disease studies. *Genes*, 10:1–15, 2019.

A Appendix

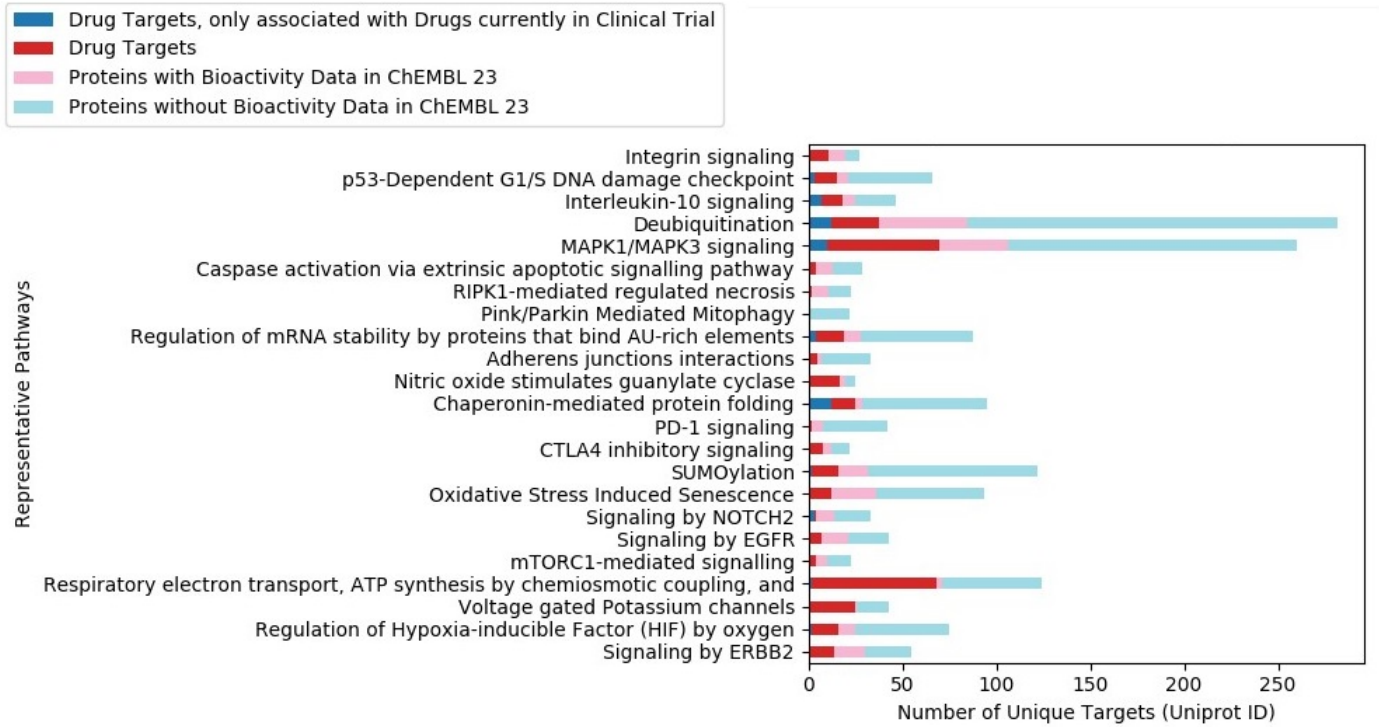


Figure A.1: Drug targets, drug targets associated with only investigational drugs, proteins with bioactivity data and proteins without bioactivity data within 24 representative pathways. Pathways are explored in terms of drug targets, novel drug targets and experimental proteins to various degrees. Novel research interest is focused on the pathways 'Deubiquination', 'MAPK1/MAPK3 Sig- nalling', 'Chaperonin-mediated protein folding' and 'Signalling by Notch2'. I note the differing numbers of proteins without any bioactivity data present in ChEMBL23. This indicates that these proteins might not be experimentally accessible and/or of little research interest.

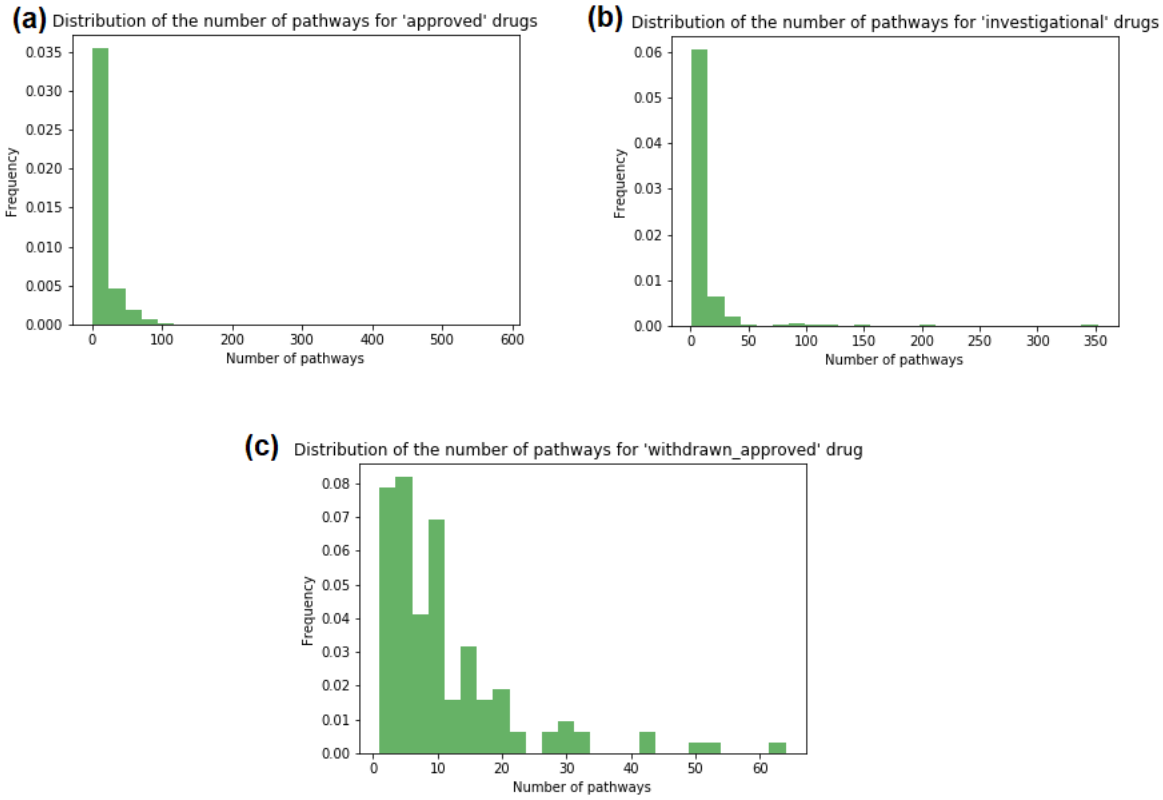


Figure A.2: Distribution of the number of pathways for the groups 'investigational', 'withdrawn' and'approved' drugs in Figure 2.1. The distribution of the number of pathways for (a) 'approved' drugs,(b)'investigational' drugs and (c) 'withdrawn' drugs, denoted as'withdrawn_approved' drugs, is plotted. As all three distributions show a non-normal distribution, the non-parametric Kruskal-Wallis test was chosen to determine whether there was a statistically significant difference between the distributions.

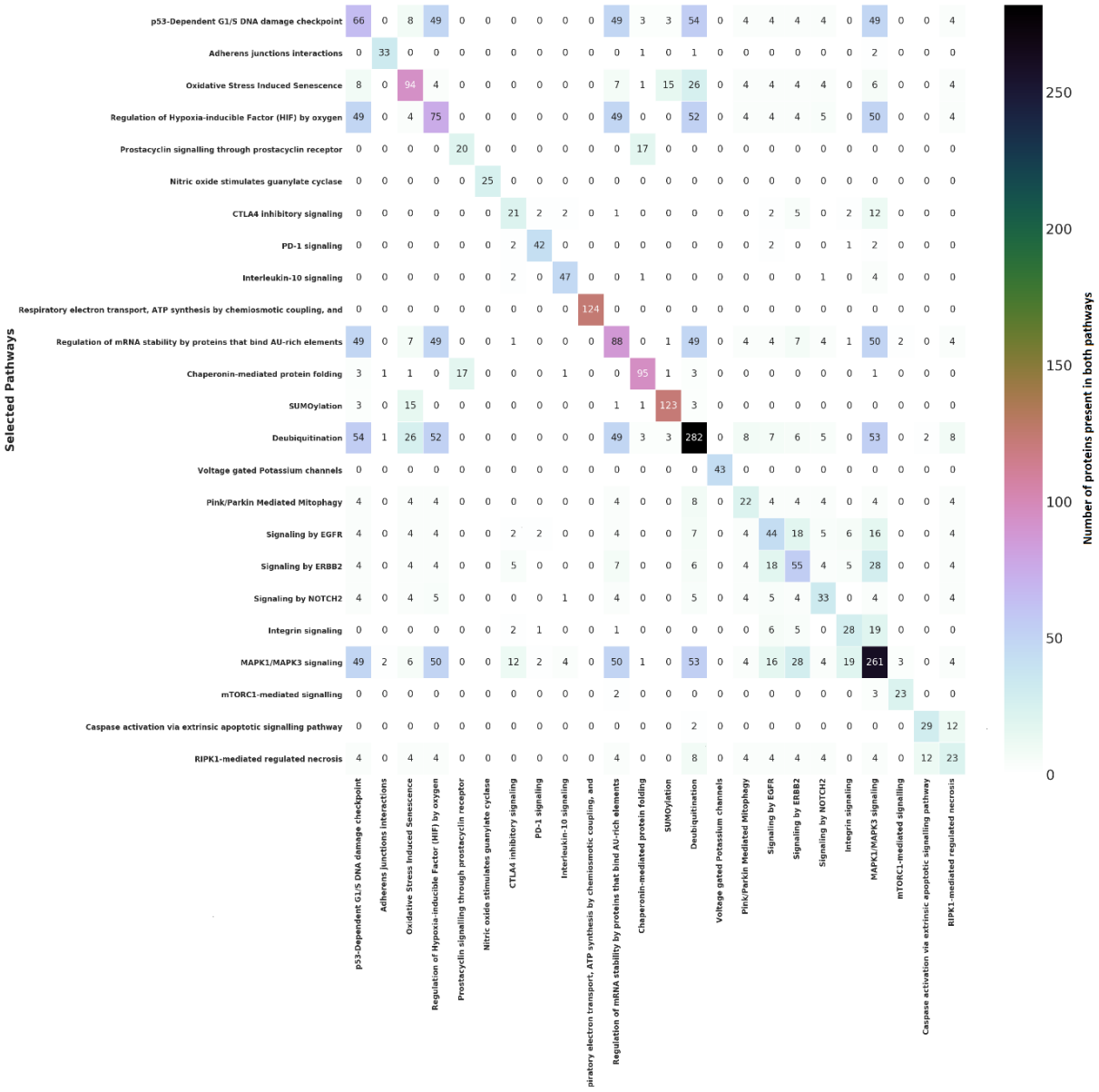


Figure A.3: Heatmap showing the number of proteins common between any two pathways within my set of'representative' pathways. The pathways 'P53 dependent G1/S DNA damage checkpoint', 'Deubiquitination', 'Regulation of mRNA stability by proteins that bind AU-rich elements', 'MAPK1/MAPK3 Signaling' and 'Regulation of Hypoxia-inducible factor (HIF)' show the highest numbers in terms of associated proteins, as they share 44 proteasome subunits. This is exemplified in Figure A.4.

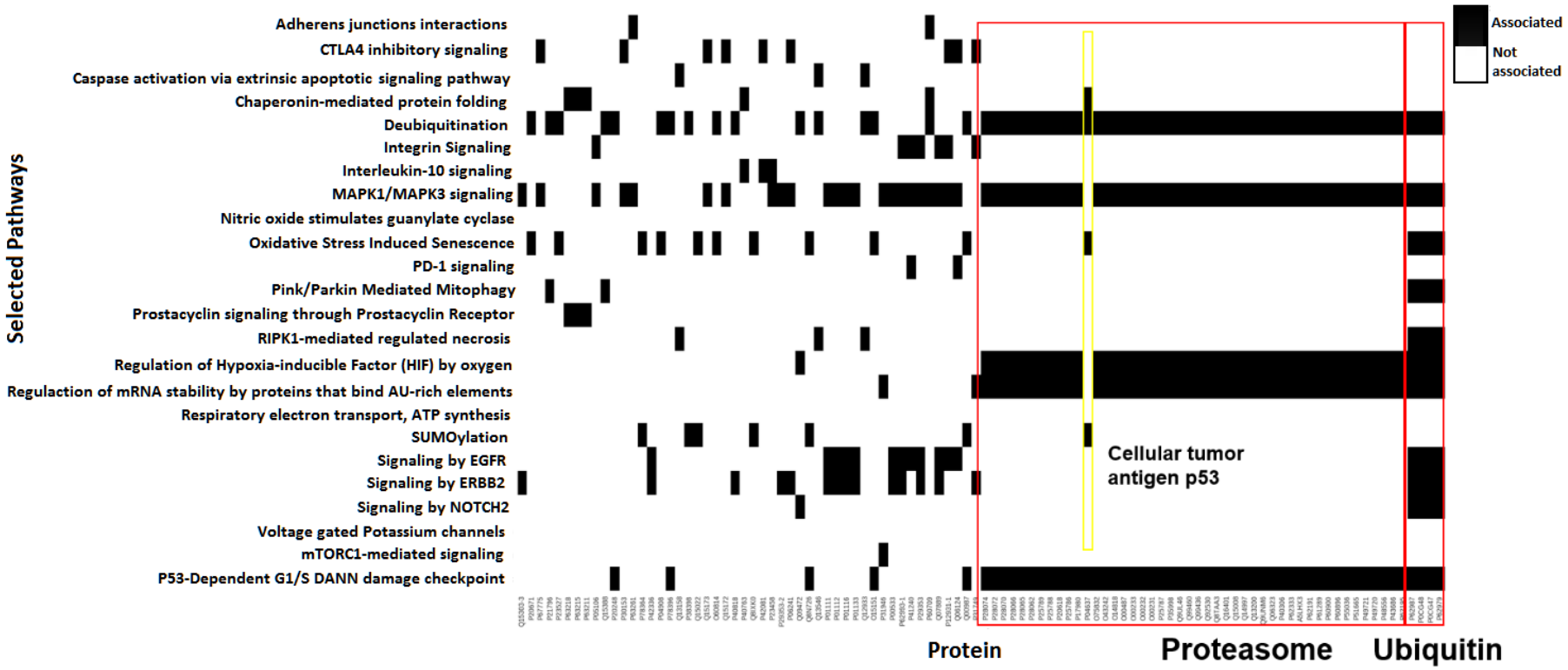


Figure A.4: 24 'representative pathways' clustered according to their contained proteins. Black and white denote the presence and absence of a protein in a pathway respectively. The x-axis labels are not shown due to an otherwise confusing overlap. By clustering my set of 'representative' pathways according to proteins, which are involved in multiple of said pathways, I show that some pathways cluster more closely together. For example, pathways with highly similar biological functions, such as the pathways. 'Signaling by EFGR', 'Signaling by ERBB2' and 'MAPK1/MAPK3 signaling', which are all implicated in signal transduction, share common proteins necessary for their function(s). For clustering, Jaccard-Index (Metric: 'Jaccard') and the clustering method 'single-linkage clustering' (Method: 'Single') was used as detailed in Section 2.2. Heatmap of the 100 proteins (UniProt ID) with the highest numbers of associated pathways within the 24 selected pathways. Black and white denote the presence and absence of a protein in a pathway respectively, as indicated in the legend. The bold red rectangle highlights the different proteasome subunits and thus the prominence of the proteasome subunits in potentially mediating pathway crosstalk between the set of selected 24 selected pathways. It is highly plausible that the proteasome subunits differ in their temporal and spatial assembly into complexes in the different pathways. (The red and the green rectangle denote the presence of the cellular tumor antigen p53 and ubiquitin molecules respectively.)

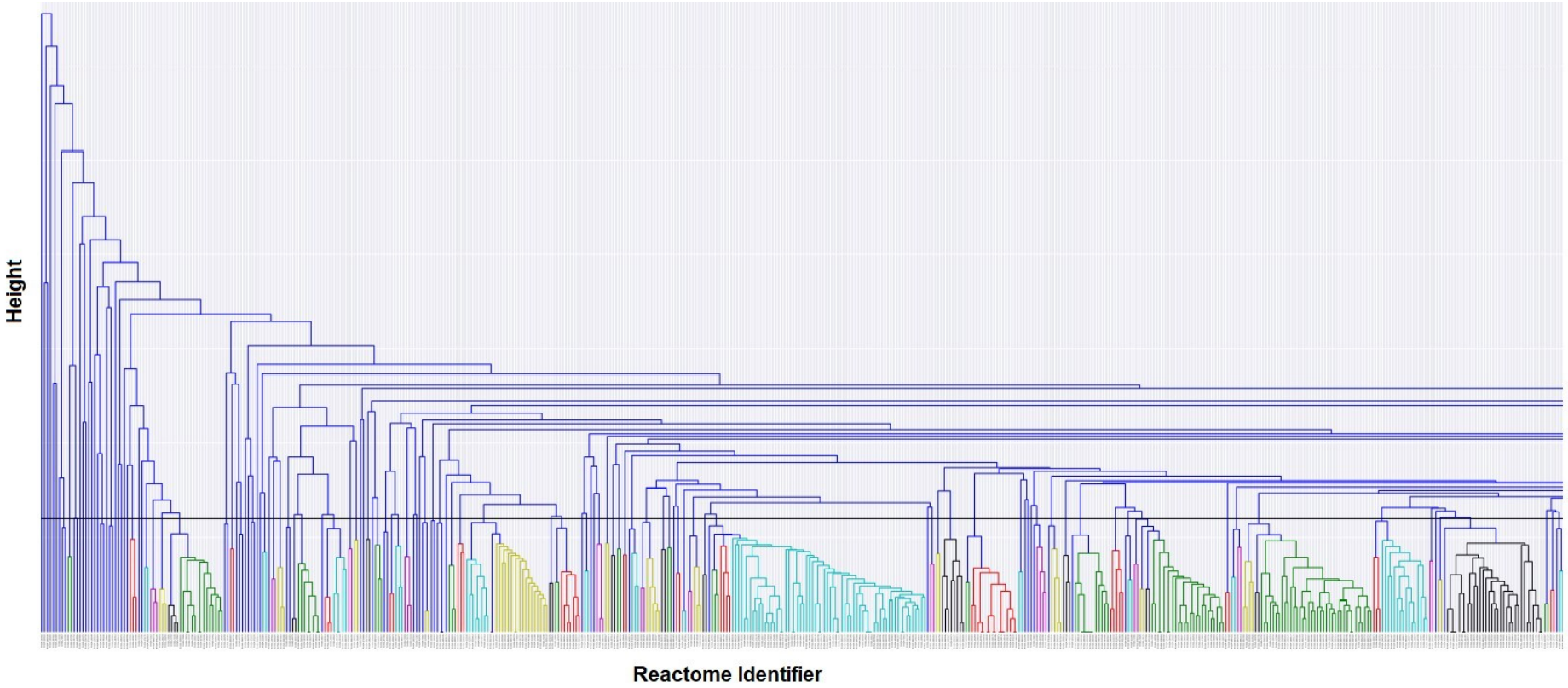


Figure A.5: Close-up of the clustering of the 2190 Pathways from the Reactome Pathway Database according to proteins involved in multiple pathways. Colours represent clusters introduced by the clustering algorithm ('complete linkage', while similarity was assessed by the Jaccard-Index, see also Section 2.2). The different pathway levels in the clusters have different levels of resolutions. These levels reflect the hierarchical nature of Reactome, where pathways can be comprised of multiple 'subpathways'. The definition of pathways depends on the curator and the research interest focused on the pathways and the entities within. This highlights the difficulty in assembling a set of pathways which represents the entire pathway space without being biased towards more well-researched path- ways. Subsequently, pathways were plotted according to proteins common between them and a cut-off height of 12 (black line) was chosen by visual inspection to strike a balance between the different levels of pathway hierarchies.

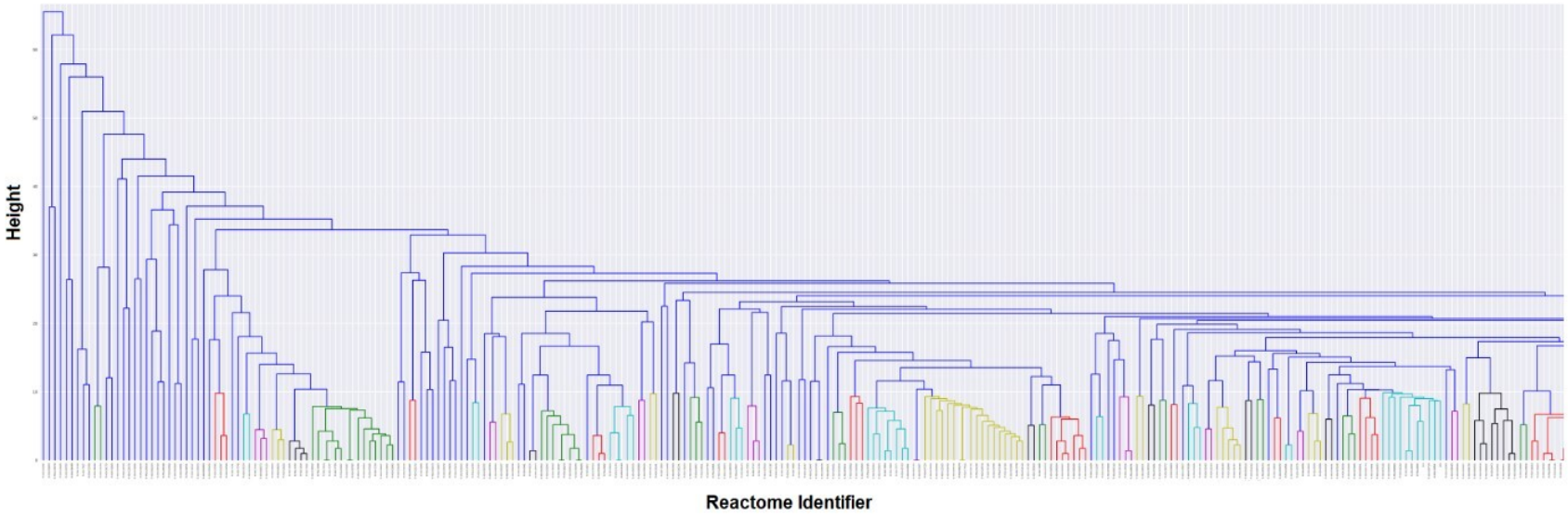


Figure A.6: Close-up of the chosen 408 pathways after cut-off, clustered according to proteins common be- tween them. Colours represent clusters introduced by the algorithm (see Section 2.2). Pathways are still hierarchically clustered according to their similarity, which would ensure repetition of a pathway count. Yet, this also reflects the research interests focused on proteins, and thus indicates again a research bias towards specific proteins and pathways.

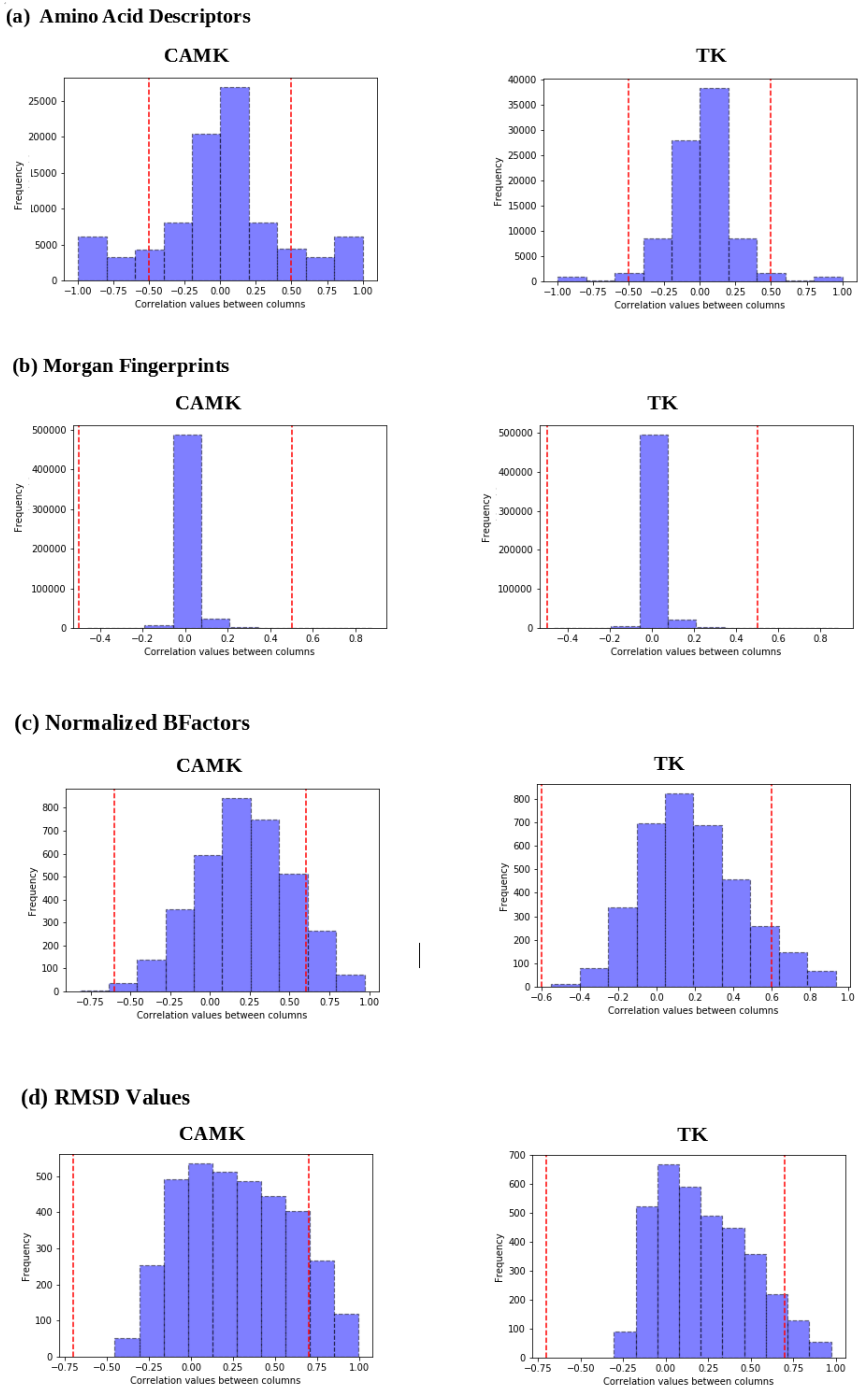


Figure A.7: Exemplary distributions of the pairwise correlation coefficients between the different feature columns. Exemplarily, the histogram of the pairwise correlation coefficient between columns for different features to construct the RF models in Section 3 is plotted to show how the removal of correlated features influences the overall number of specific features. The histogram show the following distributions: (a) Amino acid descriptors, (b) Morgan Fingerprints, (c) normalized BFactors and (d) RMSD values. For the RMSD values and the B-Factors, a threshold of 0.7 was chosen; for the amino acid descriptors and the Morgan Fingerprints a threshold of 0.5 as detailed in Section 3.5.5. The respective thresholds are plotted in red; each histogram is annotated with the corresponding target class.

Listing 1: This short Pseudo code details how the RMSD between the same binding pocket residues for the same target was calculated. Five residues of two protein structures (for the same target) were pairwise aligned and the RMSD of the middle binding pocket residue calculated based on a script by Dr. West. For simplicity's sake, the code is commented as indicated by #. The original and complete Python code can be provided upon request.

```
# Data: All PDB structures, derived from KLIFS, resolution less or equal
# than 2.5 Angstrom) for every target in  a given target family are in the
# subfolder target;  the parent_folder is targetfamily_folder.

# Create all combinations of PDB structures for a given target
# for every target in targetfamily_folder:
        create (dict for a given target within the target family)
        append (PDB codes to dict, if the PDB structure is of
        human origin)
        del (empty dict);
        del (dict, if less than 2 structures had been collected;
        del (PDB identifers with no DFG-In motif)
        save dict

    create (empty DataFrame)
    for every PDB identifier in dict:
        create (pairwise combinations using the PDB identifiers)
        save (pairwise combination in DataFrame; each PDB
        identifier is saved in column 1 and 2 respectively.)

# Collect the residue names and numbers of the KLIFS binding pocket
# residues for each PDB structure; verify whether a sliding
# window of five residues can be constructed
    for each rows in DataFrame:
        Extract PDB identifier
        Open corresponding KLIFS PDB structure file # using BioPandas
        Collect [residue names and residue numbers for a given
        KLIFS kinase binding pocket] # using BioPandas
        Save to list_1

        Collect [binding pocket residues for PDB structure 2] identically
        Save to list_2



        calc [intersection of binding pocket residues in list_1 and list_2]
```

```
        for each binding pocket residue in [intersection]:
            if (two structures have two identical binding pocket residues
               before and after the binding pocket residue
               and an identical binding pocket residue):
                     save in list
            else:
                    continue without saving binding residue in list

# Calculate the RMSD of the binding pocket residues,
# based on an adapted script by Dr. West
    for the binding pocket residues in list:
        load and define (corresponding reference and target structures)
        define atoms_superimpose = [reference and target atoms to
        calculate the RMSD for]
        define atoms_RSMD = [reference and target atoms to superimpose]
        superimpose [atoms_superimpose] # using Bio.PDB.Superimposer()
        calc RMSD of [atoms_RSMD] of the superimposed structure
        save RMSD
```

Table A.1: 57 selective and 2 historic kinase compounds used to test my models in terms of selectivity prediction. I manually collected selective compounds from three data sources: The Chemical Probe Website (CCP), the Structural Genomics Consortium, Oxford (SGC) and the Publication by Wang *et al.* as outlined in Section 4.2.1.

Trivial Name of selective compound	Protein Target	Target Class	Data-base	Canonical smile
CCT251545	CDK8	CMGC	CPP	Cn1cc(-c2ccc(-c3cncc(Cl)c3N3CCC4(CCNC4=O)CC3)cc2)cn1
CCT251545	CDK19	CMGC	CPP	Cn1cc(-c2ccc(-c3cncc(Cl)c3N3CCC4(CCNC4=O)CC3)cc2)cn1
SCH772984	MAPK1	CMGC	CPP	O=C(Nc1ccc2[nH]nc(-c3ccncc3)c2c1)[C@@H]1CCN(CC(=O)N2CCN(c3ccc(-c4ncccn4)cc3)CC2)C1
SCH772984	MAPK3	CMGC	CPP	O=C(Nc1ccc2[nH]nc(-c3ccncc3)c2c1)[C@@H]1CCN(CC(=O)N2CCN(c3ccc(-c4ncccn4)cc3)CC2)C1
GSK2334470	PDK1	AGC	CPP	CNc1nc(-c2ccc3c(N)n[nH]c3c2)cc(N2C[C@@H](C(=O)NC3CCCCC3)CC[C@H]2C)n1
Linsitinib	IGF1R	TK	CPP	CC1(O)CC(c2nc(-c3ccc4ccc(-c5ccccc5)nc4c3)c3c(N)nccn23)C1
Linsitinib	INSR	TK	CPP	CC1(O)CC(c2nc(-c3ccc4ccc(-c5ccccc5)nc4c3)c3c(N)nccn23)C1
Linsitinib	INSRR	TK	CPP	CC1(O)CC(c2nc(-c3ccc4ccc(-c5ccccc5)nc4c3)c3c(N)nccn23)C1
AX15836	MAPK7	CMGC	CPP	CCOc1cc(C(=O)N2CCC(N3CCN(C)CC3)CC2)ccc1Nc1ncc2c(n1)N(S(C)(=O)=O)c1ccccc1C(=O)N2C
CGI1746	BTK	TK	CPP	Cc1c(NC(=O)c2ccc(C(C)(C)C)cc2)cccc1-c1cn(C)c(=O)c(Nc2ccc(C(=O)N3CCOCC3)cc2)n1
PD0332991	CDK4	CMGC	CPP	CC(=O)c1c(C)c2cnc(Nc3ccc(N4CCNCC4)cn3)nc2n(C2CCCC2)c1=O.O=S(=O)(O)CCO
PD0332991	CDK6	CMGC	CPP	CC(=O)c1c(C)c2cnc(Nc3ccc(N4CCNCC4)cn3)nc2n(C2CCCC2)c1=O.O=S(=O)(O)CCO
BYL-719	PIK3CA	Class I	CPP	Cc1nc(NC(=O)N2CCC[C@H]2C(N)=O)sc1-c1ccnc(C(C)(C)C(F)(F)F)c1
CH5424802	ALK	TK	CPP	CCc1cc2c(cc1N1CCC(N3CCOCC3)CC1)C(C)(C)c1[nH]c3cc(CN)ccc3c1C2=O
FM-381	JAK3	TK	CPP	CN(C)C(=O)/C(CN)=C/c1ccc(-c2nc3cnc4[nH]ccc4c3n2C2CCCCC2)o1
VX-745	MAPK14	CMGC	CPP	O=c1ncn2nc(Sc3ccc(F)cc3F)ccc2c1-c1c(Cl)cccc1Cl
MK-5108	AURKA	Other	CPP	O=C(O)[C@]1(Cc2cccc(Nc3nccs3)n2)CC[C@@H](Oc2cccc(Cl)c2F)CC1
Ruxolitinib	JAK1	TK	CPP	NCC[C@H](C1CCCC1)n1cc(-c2ncnc3[nH]ccc23)cn1
Ruxolitinib	JAK2	TK	CPP	NCC[C@H](C1CCCC1)n1cc(-c2ncnc3[nH]ccc23)cn1
CP-724714	ERBB2	TK	CPP	COCC(=O)NC/C=C/c1ccc2ncnc(Nc3ccc(Oc4ccc(C)nc4)c(C)c3)c2c1
NVS-PAK1-1	PAK1	STE	CPP	CC(C)NC(=O)N1CC[C@H](NC2=Nc3cc(F)ccc3N(CC(F)F)c3ccc(Cl)cc32)C1
SGX-523	MET	TK	CPP	Cn1cc(-c2ccc3nnc(Sc4ccc5ncccc5c4)n3n2)cn1
AMG900	AURKA	Other	CPP	Cc1csc(-c2nnc(Nc3ccc(Oc4nccc4-c4ccnc(N)n4)cc3)c3ccccc23)c1
AMG900	AURKB	Other	CPP	Cc1csc(-c2nnc(Nc3ccc(Oc4nccc4-c4ccnc(N)n4)cc3)c3ccccc23)c1
AMG900	AURKC	Other	CPP	Cc1csc(-c2nnc(Nc3ccc(Oc4nccc4-c4ccnc(N)n4)cc3)c3ccccc23)c1
BLU9931	FGFR4	TK	CPP	C=CC(=O)Nc1cccc(C)c1Nc1ncc2cc(-c3c(Cl)c(OC)cc(OC)c3Cl)ccc2n1
CCT241533	CHEK2	CAMK	CPP	COc1cc2nc(-c3cc(F)ccc3O)nc(N[C@@H]3CNC[C@H]3C(C)(C)O)c2cc1OC.Cl
Skepinone-L	MAPK14	nan	CPP	O=C1c2ccc(Nc3ccc(F)cc3F)cc2CCc2ccc(OC[C@H](O)CO)cc21
GDC-0879	BRAF V600E	TKL	CPP	OCCn1cc(-c2ccc3c(c2)CC/C3=NØ)c(-c2ccncc2)n1
AZD1152	AURKB	Other	CPP	CCN(CCCOc1ccc2c(Nc3cc(CC(=O)Nc4cccc(F)c4)n[nH]3)ncnc2c1)CCOP(=O)(O)O
BGJ-398	FGFR1	TK	CPP	CCN1CCN(c2ccc(Nc3cc(N(C)C(=O)Nc4c(Cl)c(OC)cc(OC)c4Cl)ncn3)cc2)CC1
BGJ-398	FGFR2	TK	CPP	CCN1CCN(c2ccc(Nc3cc(N(C)C(=O)Nc4c(Cl)c(OC)cc(OC)c4Cl)ncn3)cc2)CC1
BGJ-398	FGFR3	TK	CPP	CCN1CCN(c2ccc(Nc3cc(N(C)C(=O)Nc4c(Cl)c(OC)cc(OC)c4Cl)ncn3)cc2)CC1
BGJ-398	FGFR4	TK	CPP	CCN1CCN(c2ccc(Nc3cc(N(C)C(=O)Nc4c(Cl)c(OC)cc(OC)c4Cl)ncn3)cc2)CC1
G-5555	PAK1	STE	CPP	CNc1ncc2cc(-c3ccc(-c4cccc(C)n4)cc3Cl)c(=O)n(C[C@H]3OC[C@H](N)CO3)c2n1
G-5555	PAK2	STE	CPP	CNc1ncc2cc(-c3ccc(-c4cccc(C)n4)cc3Cl)c(=O)n(C[C@H]3OC[C@H](N)CO3)c2n1
G-5555	PAK3	STE	CPP	CNc1ncc2cc(-c3ccc(-c4cccc(C)n4)cc3Cl)c(=O)n(C[C@H]3OC[C@H](N)CO3)c2n1
PF-477736	CHEK1	CAMK	CPP	Cn1cc(-c2nc3cc(NC(=O)[C@H](N)C4CCCCC4)cc4c3-c2c[nH][nH]c4=O)cn1
AZD2014	MTOR	Atypical	CPP	CNC(=O)c1cccc(-c2ccc3c(N4CCOC[C@@H]4C)nc(N4CCOC[C@@H]4C)nc3n2)c1
GSK583	RIPK2	AGC	CPP	CC(C)(C)S(=O)(=O)c1ccc2nccc(Nc3n[nH]c4ccc(F)cc34)c2c1

Trivial Name of selective compound	Protein Target	Target Class	Data-base	Canonical smile
Tofacitinib	JAK1	TK	CPP	C[C@@H]1CCN(C(=O)CCN)C[C@@H]1N(C)c1ncnc2[nH]ccc12
Tofacitinib	JAK2	TK	CPP	C[C@@H]1CCN(C(=O)CCN)C[C@@H]1N(C)c1ncnc2[nH]ccc12
Tofacitinib	JAK3	TK	CPP	C[C@@H]1CCN(C(=O)CCN)C[C@@H]1N(C)c1ncnc2[nH]ccc12
CHIR-99021	GSK3A	CMGC	CPP	Cc1c[nH]c(-c2cnc(NCCNc3ccc(CN)cn3)nc2-c2ccc(Cl)cc2Cl)n1
CHIR-99021	GSK3B	CMGC	CPP	Cc1c[nH]c(-c2cnc(NCCNc3ccc(CN)cn3)nc2-c2ccc(Cl)cc2Cl)n1
KH-CB19	CLK1	CMGC	CPP	CCOC(=O)c1c(/C(CN)=C)c2ccc(Cl)c(Cl)c2n1C
KH-CB19	CLK4	CMGC	CPP	CCOC(=O)c1c(/C(CN)=C)c2ccc(Cl)c(Cl)c2n1C
KH-CB19	DYRK1A	CMGC	CPP	CCOC(=O)c1c(/C(CN)=C)c2ccc(Cl)c(Cl)c2n1C
P505-15	SYK	TK	CPP	NC(=O)c1cnc(N[C@@H]2CCCC[C@@H]2N)nc1Nc1cccc(-n2nccn2)c1
BMS-911543	JAK2	TK	CPP	CCn1c(C(=O)N(C2CC2)C2CC2)cc2c3c(ncn3C)c(Nc3cc(C)n(C)n3)nc21
GNF-5	BCR-ABL	TK	CPP	O=C(NCCO)c1cccc(-c2cc(Nc3ccc(OC(F)(F)F)cc3)ncn2)c1
Filgotinib	JAK1	TK	CPP	O=C(Nc1nc2cccc(-c3ccc(CN4CCS(=O)(=O)CC4)cc3)n2n1)C1CC1
THZ1	CDK7	CMGC	CPP	CN(C)C/C=C/C(=O)Nc1ccc(C(=O)Nc2cccc(Nc3ncc(Cl)c(-c4c[nH]c5ccccc45)n3)c2)cc1
CCT244747	CHEK1	CAMK	CPP	COc1cc(Nc2cnc(CN)c(OC(C)CN(C)C)n2)ncc1-c1cnn(C)c1
Selumetinib	MAP2K1	STE	CPP	Cn1cnc2c(F)c(Nc3ccc(Br)cc3Cl)c(C(=O)NOCCO)cc21
AC220	FLT3	TK	CPP	CC(C)(C)c1cc(NC(=O)Nc2ccc(-c3cn4c(n3)sc3cc(OCCN5CCOCC5)ccc34)cc2)no1
Rapamycin	MTOR	Atypical	CPP	CO[C@H]1C[C@@H]2CC[C@@H](C)[C@@](O)(O2)C(=O)C(=O)N2CCCC[C@H]2C(=O)O[C@H]([C@H](C)C[C@@H]2CC[C@@H](O)[C@H](OC)C2)CC(=O)[C@H](C)/C=C()[C@@H](O)[C@@H](OC)C(=O)[C@H](C)C[C@H](C)C=C/=CC=C1C
PF3644022	MAPKAPK2	CAMK	CPP	Cc1ccc(-c2ccc3c(cce4sc5c(c43)NC[C@@H](C)NC5=O)n2)cn1
JNK-IN-8	MAPK8	CMGC	CPP	Cc1cc(NC(=O)c2cccc(NC(=O)/C=C/CN(C)C)c2)ccc1Nc1nccc(-c2cccnc2)n1
JNK-IN-8	MAPK9	CMGC	CPP	Cc1cc(NC(=O)c2cccc(NC(=O)/C=C/CN(C)C)c2)ccc1Nc1nccc(-c2cccnc2)n1
JNK-IN-8	MAPK10	CMGC	CPP	Cc1cc(NC(=O)c2cccc(NC(=O)/C=C/CN(C)C)c2)ccc1Nc1nccc(-c2cccnc2)n1
AMG-18	ERN1	Other	CPP	Cc1ccc2c(NS(=O)(=O)c3ccccc3Cl)cccc2c1Oc1nccc1-c1ccnc(N[C@@H]2CCCNC2)n1
NVP-AEW541	IGF1R	TK	CPP	Nc1ncnc2c1c(-c1cccc(OCc3ccccc3)c1)cn2C1CC(CN2CCC2)C1
PF-3758309	PAK1	STE	CPP	Cc1nc(Nc2n[nH]c3c2CN(C(=O)N[C@H](CN(C)C)c2ccccc2)C3(C)C)c2sccc2n1
PF-3758309	PAK4	STE	CPP	Cc1nc(Nc2n[nH]c3c2CN(C(=O)N[C@H](CN(C)C)c2ccccc2)C3(C)C)c2sccc2n1
PF-3758309	PAK5	STE	CPP	Cc1nc(Nc2n[nH]c3c2CN(C(=O)N[C@H](CN(C)C)c2ccccc2)C3(C)C)c2sccc2n1
PF-3758309	PAK6	STE	CPP	Cc1nc(Nc2n[nH]c3c2CN(C(=O)N[C@H](CN(C)C)c2ccccc2)C3(C)C)c2sccc2n1
GNE7915	LRRK2	TKL	CPP	CCNc1nc(Nc2cc(F)c(C(=O)N3CCOCC3)cc2OC)ncc1C(F)(F)F
UNC2025	MERTK	TK	CPP	CCCCNc1ncc2c(-c3ccc(CN4CCN(C)CC4)cc3)cn([C@H]3CC[C@H](O)CC3)c2n1
UNC2025	FLT3	TK	CPP	CCCCNc1ncc2c(-c3ccc(CN4CCN(C)CC4)cc3)cn([C@H]3CC[C@H](O)CC3)c2n1
WZ4003	NUAK1	CAMK	CPP	CCC(=O)Nc1cccc(Oc2nc(Nc3ccc(N4CCN(C)CC4)cc3OC)ncc2Cl)c1
WZ4003	NUAK2	CAMK	CPP	CCC(=O)Nc1cccc(Oc2nc(Nc3ccc(N4CCN(C)CC4)cc3OC)ncc2Cl)c1
eCF309	MTOR	Atypical	CPP	CCOC(Cn1nc(-c2ccc3oc(N)nc3c2)c2c(N)ncnc21)OCC
BIX-02188	MAP2K5	STE	CPP	CN(C)Cc1cccc(N/C(=C2(=O)Nc3cc(C(N)=O)ccc32)c2ccccc2)c1
HG-9-91-01	SIK1	CAMK	CPP	COc1ccc(N(C(=O)Nc2c(C)cccc2C)c2cc(Nc3ccc(N4CCN(C)CC4)cc3)ncn2)c(OC)c1
HG-9-91-01	SIK2	CAMK	CPP	COc1ccc(N(C(=O)Nc2c(C)cccc2C)c2cc(Nc3ccc(N4CCN(C)CC4)cc3)ncn2)c(OC)c1
HG-9-91-01	SIK3	CAMK	CPP	COc1ccc(N(C(=O)Nc2c(C)cccc2C)c2cc(Nc3ccc(N4CCN(C)CC4)cc3)ncn2)c(OC)c1
TH-257	LIMK1	TKL	SGC, Oxford	CCCCN(Cc1ccccc1)C(=O)c1ccc(S(=O)(=O)Nc2ccccc2)cc1
TH-257	LIMK2	TKL	SGC, Oxford	CCCCN(Cc1ccccc1)C(=O)c1ccc(S(=O)(=O)Nc2ccccc2)cc1
SGC-AAK1-1	BIKE	Other	SGC, Oxford	CCN(CC)S(=O)(=O)Nc1cccc(-c2ccc3c(NC(=O)C4CC4)n[nH]c3c2)c1
SGC-GAK-1	GAK	Other	SGC, Oxford	COc1cc(Nc2ccnc3ccc(Br)cc23)cc(OC)c1OC
DDR-TRK-1	DDR1	TK	SGC, Oxford	Cc1cn(-c2cc(NC(=O)c3ccc4c(c3)CN(c3cncnc3)C[C@@H]4C)cc(C(F)(F)F)c2)cn1
DDR-TRK-1	DDR1	TK	SGC, Oxford	Cc1cn(-c2cc(NC(=O)c3ccc4c(c3)CN(c3cncnc3)C[C@@H]4C)cc(C(F)(F)F)c2)cn1
NVS-PAK1-1	PAK1	STE	SGC, Oxford	CC(C)NC(=O)N1CCC(NC2=Nc3cc(F)ccc3N(CC(F)F)c3ccc(Cl)cc32)C1
Lapatinib	EGFR	TK	Tool compound, (Wang et al., 2016) *et al.*	CS(=O)(=O)CCNCc1ccc(-c2ccc3ncnc(Nc4ccc(OCc5cccc(F)c5)c(Cl)c4)c3c2)o1
Dasatinib	ABL1	TK	Historic Compound (CCP), (Wang et al., 2016)	Cc1nc(Nc2ncc(C(=O)Nc3c(C)cccc3Cl)s2)cc(N2CCN(CCO)CC2)n1
Nilotinib	ABL1	TK	CCP, (Wang et al., 2016)	Cc1cn(-c2cc(NC(=O)c3ccc(C)c(Nc4nccc(-c5cccnc5)n4)c3)cc(C(F)(F)F)c2)cn1
Gefitinib	EGFR	TK	Historic Compound (CCP), (Wang et al., 2016)	COc1cc2ncnc(Nc3ccc(F)c(Cl)c3)c2cc1OCCCN1CCOCC1
Erlotinib	EGFR	TK	Tool compound, (Wang et al., 2016)	CCc1cccc(Nc2ncnc3cc(OCCOC)c(OCCOC)cc23)c1

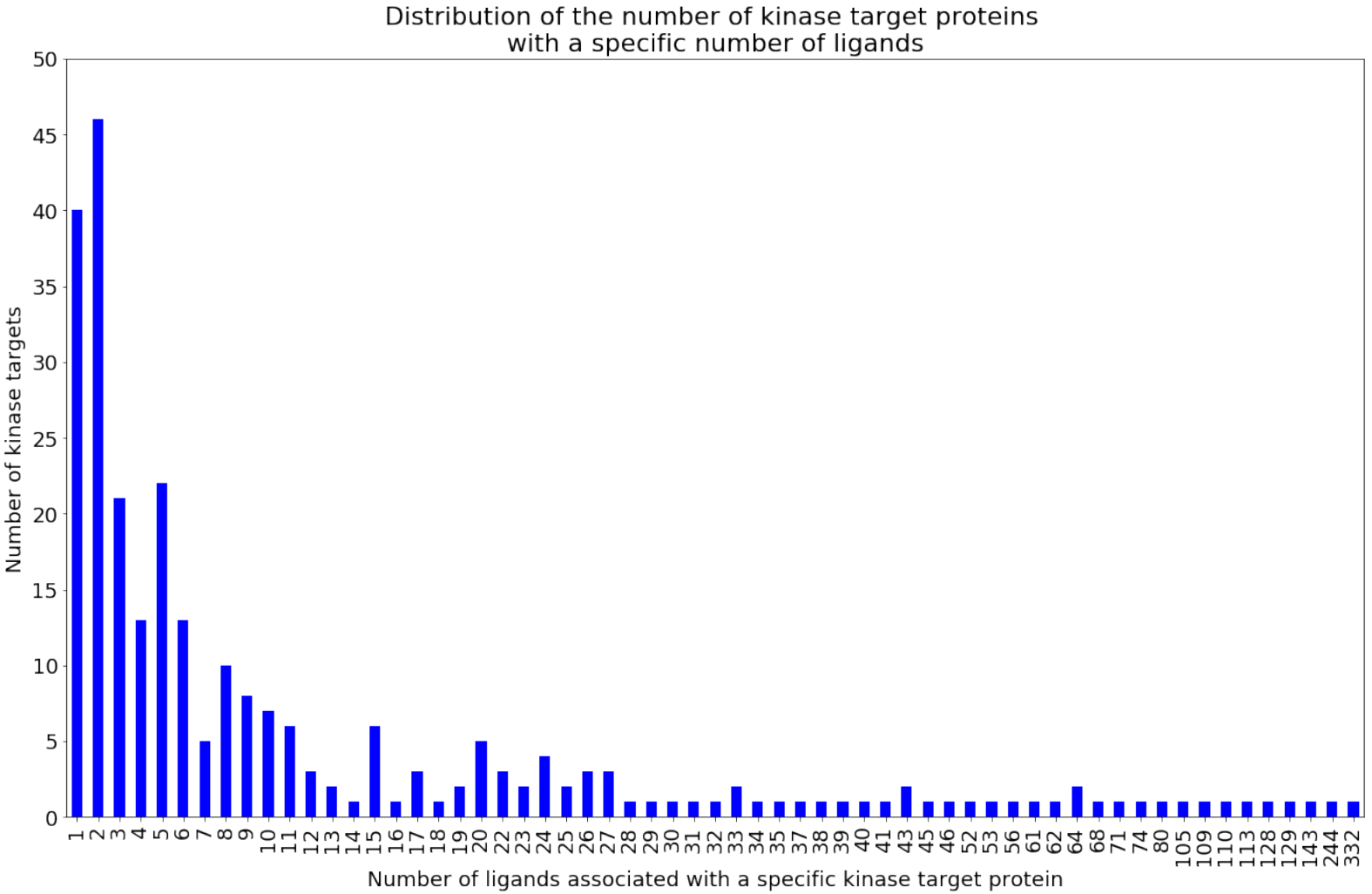


Figure A.8: Number of protein kinases with a specific number of ligands. Using the PDB (Version February 2019), I identified the number of protein-ligand interactions for a given protein kinases. While 40 protein kinase targets could be connected to one ligand (and thus one ligand-target interaction) under my threshold of 2.5 Å, 90 kinase targets could be connected to 10 ligands (and ten ligand-target interactions) and more.

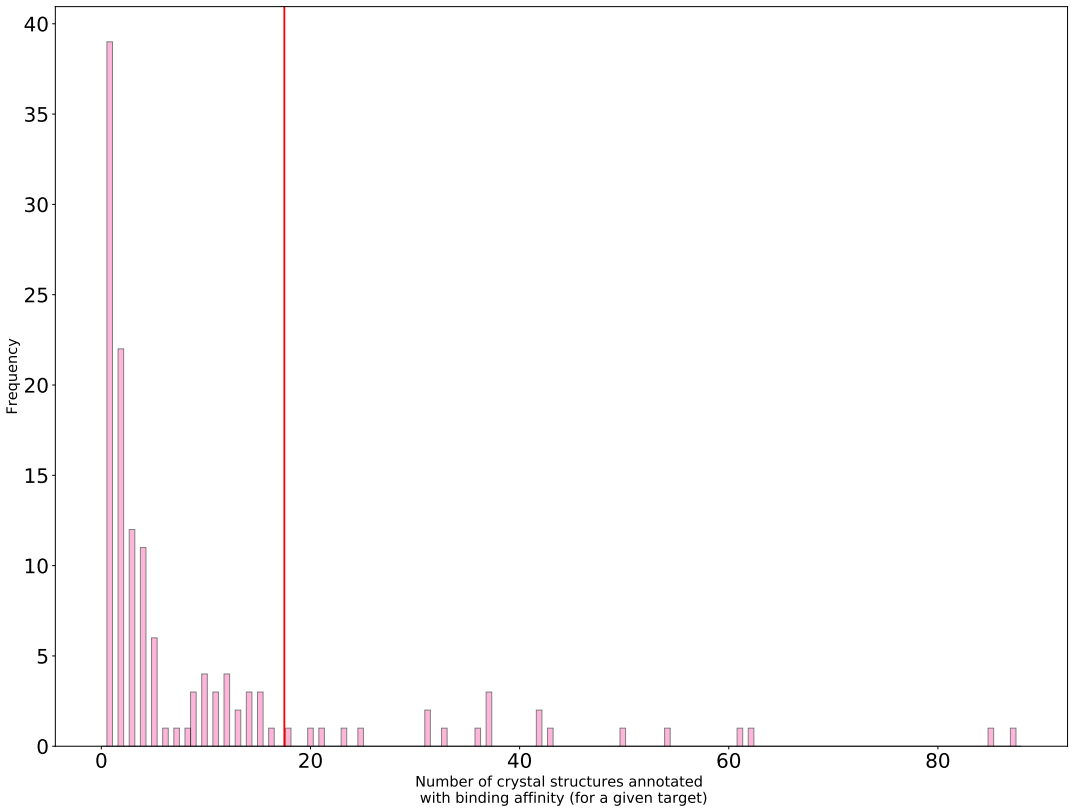


Figure A.9: The Distribution of number of crystal structures annotated with binding affinities for kinase in the DFG-in and αC helix in conformations. All PDB structures in the DFG-in Conformation, αC helix in conformation were annotated with the binding affinity value for the ligand, if available. PDBbind version 2018 was used; pIC_{50}, pK_d, pK_I) were used interchangeably and denoted as pK values, consistent with Boyles *et al.*(Boyles et al., 2019). If multiple ligand-kinase target complexes for the same ligand were available, the complex with the best resolution was selected. The cut-off was set at 18 annotated PDB structures for a given target.

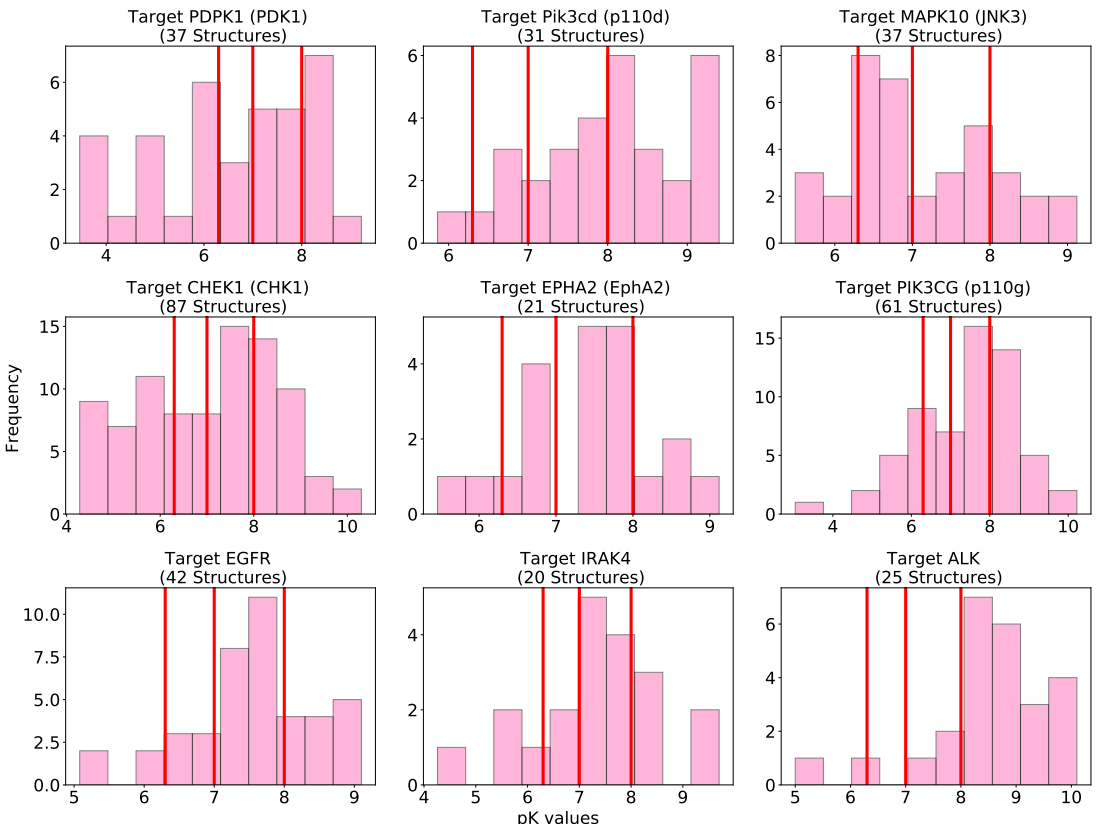


Figure A.10: The Distribution of Binding affinities values (pKI, pKd, pIC_{50}) for crystal structures for a given target in the DFG-in conformation, αC helix in conformation. All PDB files in the DFG-in conformation, αC helix in conformation were annotated with the binding affinity value for the ligand, if available. PDBBind (Version 2018) was used; pIC_{50}, pK_d, pK_I) were used interchangeably and denoted as pK values, consistent with Boyles *et al.*(Boyles et al., 2019). Each subplot is annotated with the HGNC name (also the xName, if differing from the HGNC name) and the number of structures which could be reliably annotated with binding affinities for this target. Three red lines indicate our thresholds for strong, medium and weak binders at 8, 7 and 6.3 respectively, corresponding to 10nM, 100nM and 500nM each.

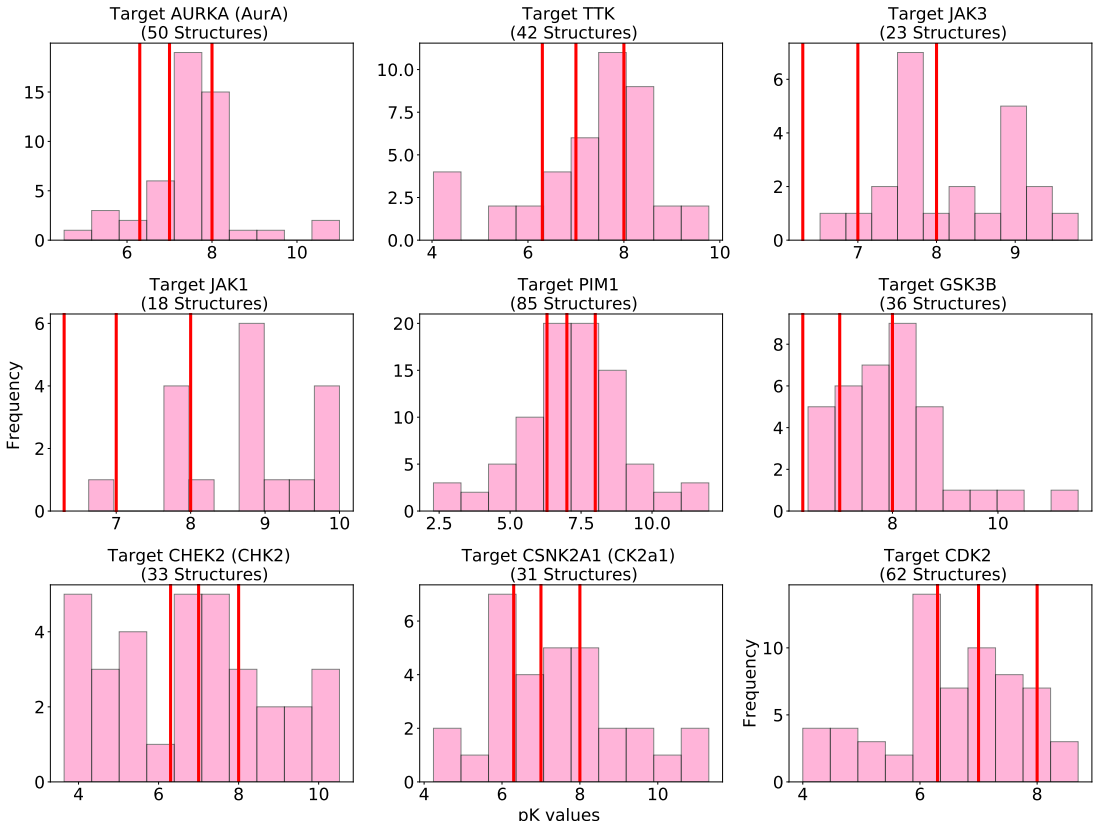


Figure A.10: The Distribution of Binding affinities values (pKI, pKd, pIC_{50}) for crystal structures for a given target in the DFG-in conformation, αC helix in conformation (Continued).

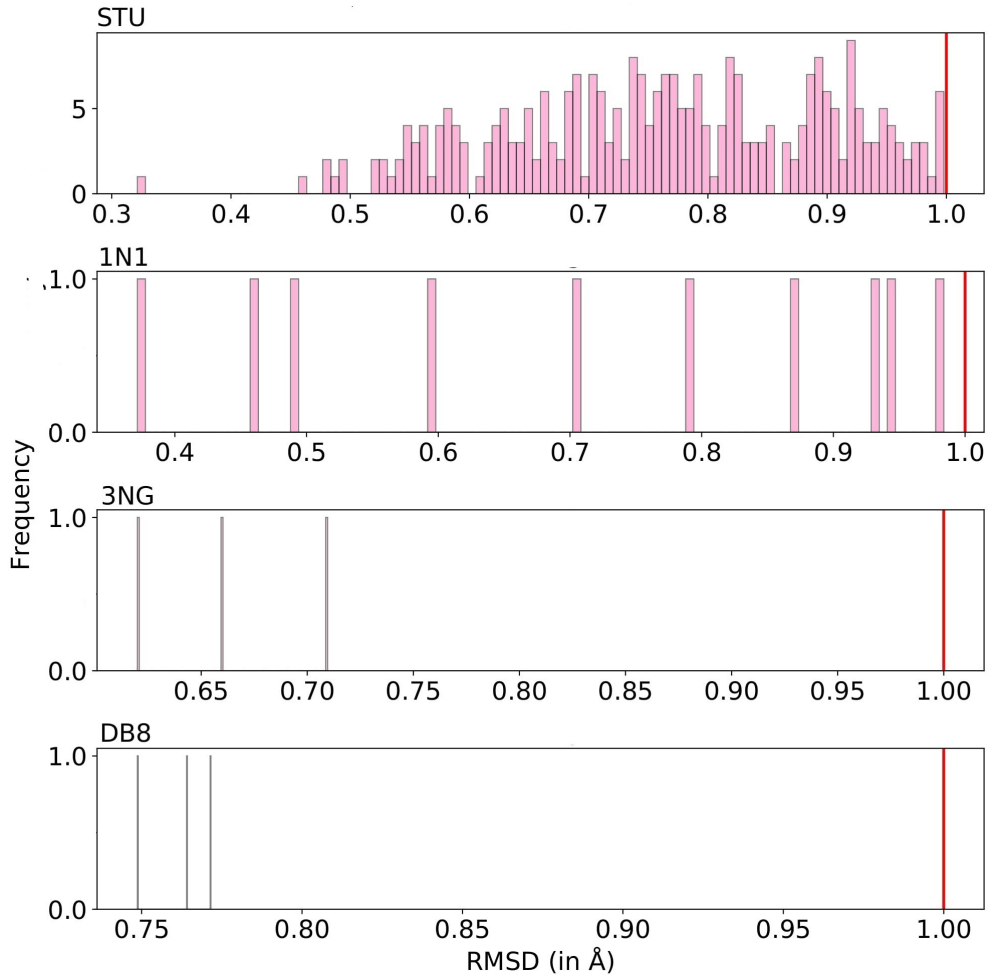


Figure A.11: Distribution of the RMSD values for the pairwise alignment of promiscuous kinase ligands. The crystal structures for the binding pocket of promiscuous compounds were pairwise aligned and the RMSD calculated to judge the validity of the derived SuCOS score.

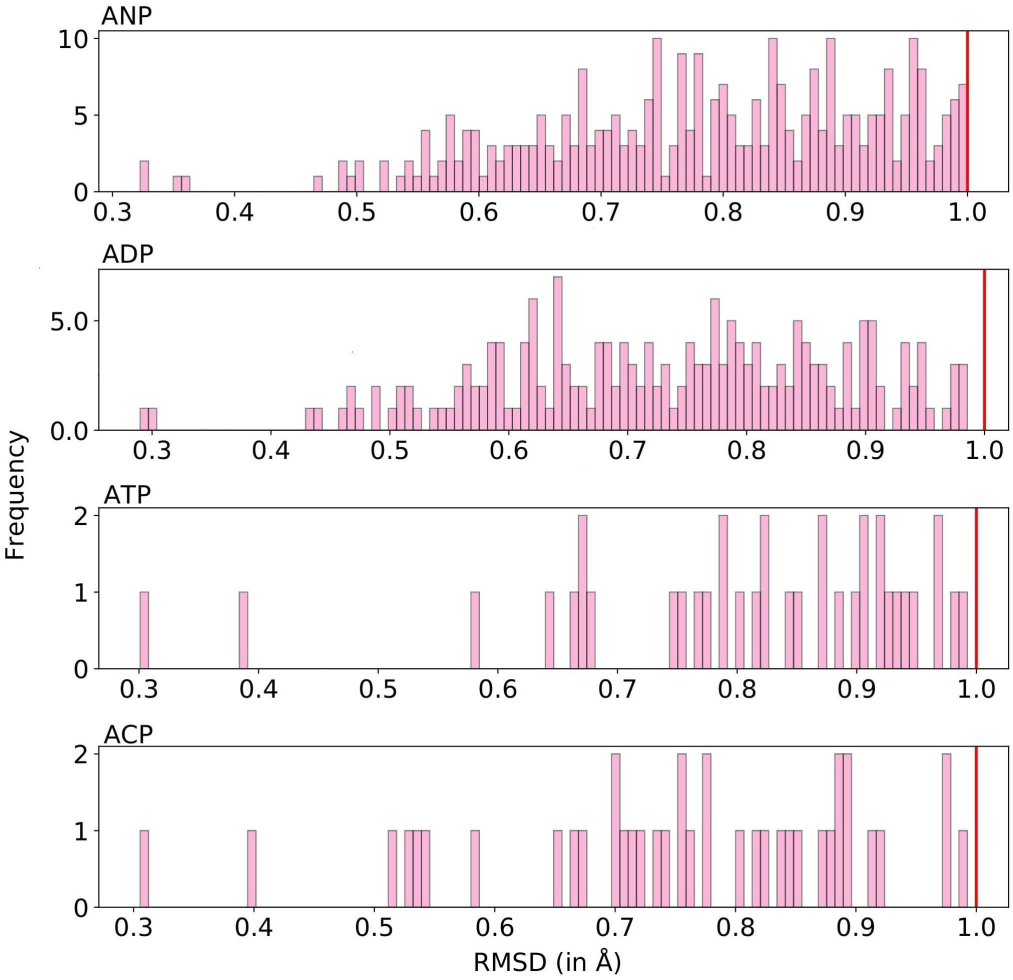


Figure A.11: **Distribution of the RMSD values for the pairwise alignment of promiscuous kinase ligands (Continued).**

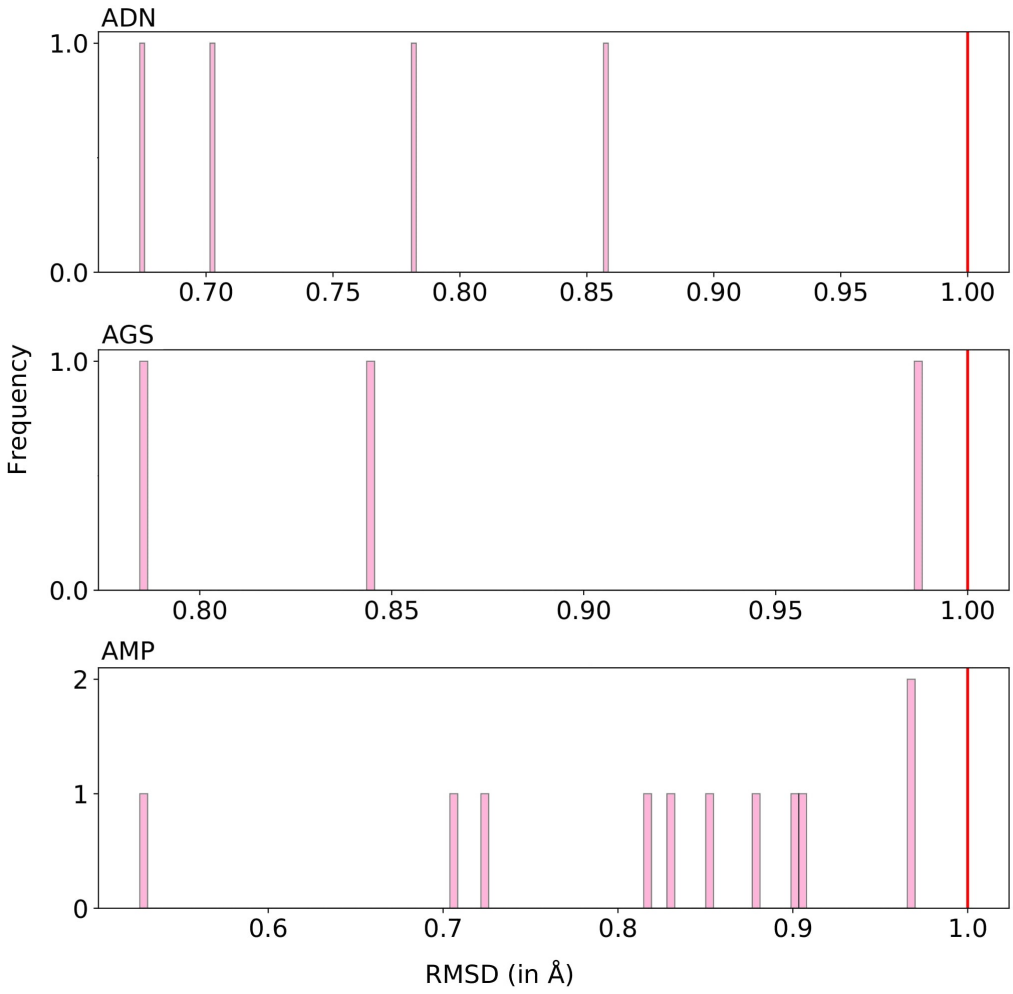


Figure A.11: Distribution of the RMSD values for the pairwise alignment of promiscuous kinase ligands (Continued).

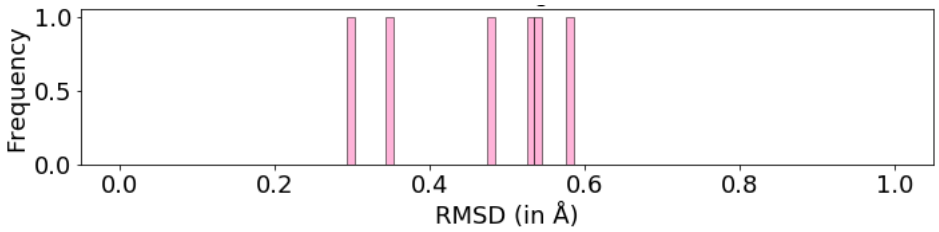


Figure A.12: Distribution of the RMSD for the pairwise alignment of the selective kinase inhibitor Tofacitinib (MI1). The crystal structures for the binding pocket of promiscuous compounds were pairwise aligned and the RMSD calculated to judge the validity of the derived SuCOS score.