Using Heterogeneous Information Sources for Understanding and Predicting Biological Effects of Compounds

Maria-Anna Trapotsi

Hughes Hall
University of Cambridge

December 2021

This thesis is submitted for the degree of Doctor of Philosophy.

Declaration

This thesis is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text.

This thesis is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. I further state that no substantial part of my dissertation has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University of similar institution except as declared in the Preface and specified in the text.

This thesis does not exceed the prescribed word limit for the Physics and Chemistry Degree Committee.

Work done in collaboration:
Chapter 2 in collaboration with Dr Lewis Mervin, who extracted the negative bioactivity data (inactives) and helped me to train the models by incorporating the Probabilistic Random Forest function.

Summary

Using Heterogeneous Information Sources for Understanding and Predicting Biological Effects of Compounds

Maria-Anna Trapotsi

Understanding a compound's biological effects such as its Mechanism of Action (MoA) and safety profile is a challenging task in drug discovery process. However, this understanding can facilitate drug discovery process and provide an early warning for potential risks. Biological effects understanding has been significantly facilitated by the advances in Machine Learning (ML), bioinformatic approaches and the increasing deposition of high throughput data in public databases. There are different types of information/data which can be used and as the volume of this data increases, so too does their potential to deepen our understanding. Therefore, key questions remain around which ML methodologies and which data types to use. In this thesis, the aim was to provide answers to two questions about which data and methods to use for compounds' MoA understanding and how to explore the safety profile of new data modalities such as PROteolysis TArgeting Chimeras (PROTACs).

In the first chapter, "*Probabilistic Random Forest improves bioactivity predictions close to the classification threshold by taking into account experimental uncertainty"*, a novel algorithm was evaluated and benchmarked. A limiting factor in target prediction for MoA understanding is the experimental variability in bioactivity data, which are used to train target prediction models. By applying this novel algorithm, which is a modification of the long-established Random Forest (RF), and comparing it with the classic RF, a benefit was identified in the prediction of compounds which are close to the classification threshold.

The next chapter, "*Comparison of Structural Chemical and Cell Morphology Information for Multitask Bioactivity Predictions"*, provided insights in which type of compound information is more useful in target prediction across 224 targets. The comparison was performed using cell morphology information (in the form of CellProfiler features) from a Cell Painting assay and chemical structure information in the form of Extended Connectivity Fingerprints. The comparison revealed that there were targets better predicted by cell morphology information such as the β-catenin

and other better predicted by chemical structure information such as proteins belonging to the G-protein-Coupled Receptor 1 family.

The final chapter, "*Mitochondrial Toxicity Prediction using Cell Painting Assay on a PROTACs dataset"*, explored the successful profiling of a novel data modality (PROTACs) with the Cell Painting assay and evaluated whether this profiling can be used in the understanding of the safety of those novel compounds. Cell morphology features (in the form of CellProfiler features) successfully predicted mitochondrial toxicity in a PROTACs dataset. This work resulted in the first ML model to predict PROTACs' mitochondrial toxicity using Cell Painting-based features and expanded our knowledge for PROTACs' safety profile prediction.

In summary, the work described in this thesis has furthered the field of *in-silico* target deconvolution and PROTACs' mitochondrial toxicity prediction. Firstly, the work showed that there is benefit of using Probabilistic Random Forest when there is a degree of experimental uncertainty in bioactivity data close to the classification threshold. In addition, this work highlighted targets, where the use of compounds' cell morphology information was beneficial for target prediction and finally showed that PROTACs' cell morphology information can be used for mitochondrial toxicity prediction.

Acknowledgements

I would like to thank and express my gratitude to my academic supervisor, Dr Andreas Bender and my industrial supervisors Dr Ola Engkvist, Dr Ian Barrett and Dr Lewis Mervin at AstraZeneca for their valuable guidance and support throughout my PhD. I would also like to thank all members of the Bender Group, the Quantitative Biology and High Throughput screening teams in AstraZeneca. I would also like to thank Dr Avid Afzal, Dr Kathryn Giblin and Layla Hosseini-Gerami for their helpful discussions and support. The BBSRC and AstraZeneca are also thanked for funding.

Last but by no means least, I owe everlasting thanks to my family, to my mother Georgia Trapotsi, my father George Trapotsis, my brother Leonidas Trapotsis and my sister-in-law Georgia Kostopoulou. This would not have been possible without you all!

List of Publications

Trapotsi, M. A., Barrett, I., Engkvist, O., & Bender, A. (2019). Bioinformatic Approaches in the Understanding of Mechanism of Action (MoA). Target Discovery and Validation: Methods and Strategies for Drug Discovery, 323-363.

Trapotsi, M. A., Mervin, L. H., Afzal, A. M., Sturm, N., Engkvist, O., Barrett, I. P., & Bender, A. (2021). Comparison of Chemical Structure and Cell Morphology Information for Multitask Bioactivity Predictions. Journal of Chemical Information and Modeling, 61(3), 1444-1456.

Trapotsi, M. A., Hosseini-Gerami L., & Bender, A. (2022). Computational Analyses of Mechanism of Action (MoA): Data, Methods and Integration. RSC Chemical Biology.

Mervin, L. H., **Trapotsi, M. A.**, Afzal, A. M., Barrett, I. P., Bender, A., & Engkvist O. (2021). “Probabilistic Random Forest Improves Bioactivity Predictions Close to the Classification Threshold by Taking into Account Experimental Uncertainty”. Journal of Cheminformatics. 13, 62.

Trapotsi, M. A., Mouchet, E., Williams, G., Monteverde, T., Juhani, K., Turkki, R., Miljković, F., Martinsson, M., Mervin, L.H., Müllers, E., Engkvist, O., Barrett, I. P., Bender, A. & Moreau, K. (2022). Cell morphological profiling enables high-throughput screening for PROteolysis TArgeting Chimera (PROTAC) phenotypic signature. bioRxiv.

Table of Contents

1. Chapter 1: Introduction

1.1. The Drug Discovery and Development Process

The drug discovery and development process is lengthy, complex and characterised by high risk and cost[1]. There are multiple stages in this process with some of the traditional phases shown in Figure 1.1. In order a compound to reach the end goal of this process, which is the registration and Food Drug Administration (FDA) approval, must exhibit both a pharmacological effect (i.e., to be efficacious), and an acceptable pharmacokinetic and safety profile. It has been estimated that the lack of efficacy and safety are the two main reasons for failure in Phase II and III clinical trials[2,3]. In 2011-2012, there were 105 Phase II and III failures with a reported reason and the reason was 56% of times related to the lack of efficacy and 28% related to safety (safety includes those failures that were due to an insufficient therapeutic index)[4]. Similarly, in 2013-2015, there were 174 Phase II and III failures with a reported reason and 56% of times the reason was related to lack of efficacy and 28% being related to safety[5].

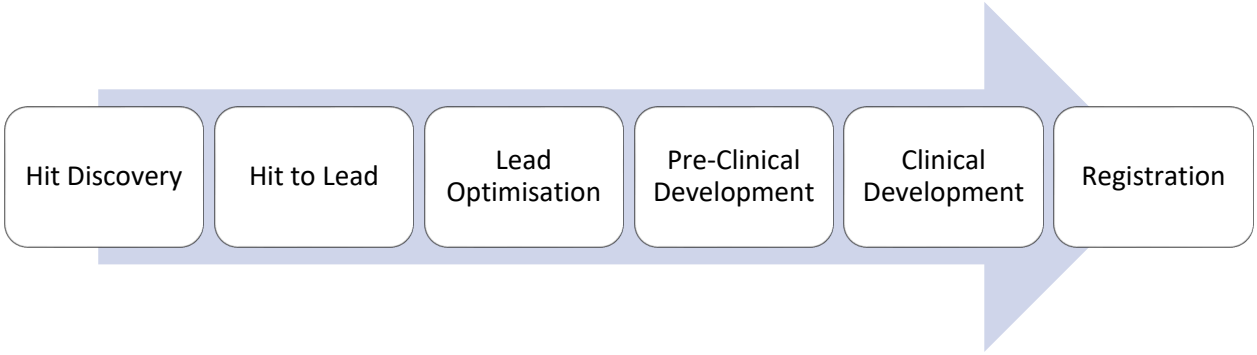


Figure 1.1: Traditional phases of drug discovery and development process (adapted from Jenkinson et al.[6])

Therefore, it is important to find ways to reduce the risk in drug discovery process and one way is to early investigate a compound's Mechanism of Action (MoA) and safety profile by using bioinformatic approaches and Machine Learning (ML) methodologies given a well-specified question with abundant and/or high-quality data[1]. This could be performed during the hit discovery and hit to lead phases by identifying which compounds will act on a target of interest and additionally during lead optimisation to identify potential risks and optimise chemical structures to avoid them. By identifying a compound's MoA, it is possible to understand how a compound produces its therapeutic effect but also to identify any off-target interactions, which could result in

safety issues (Figure 1.2), which will result in failure in the later stages of drug discovery and development pipeline.

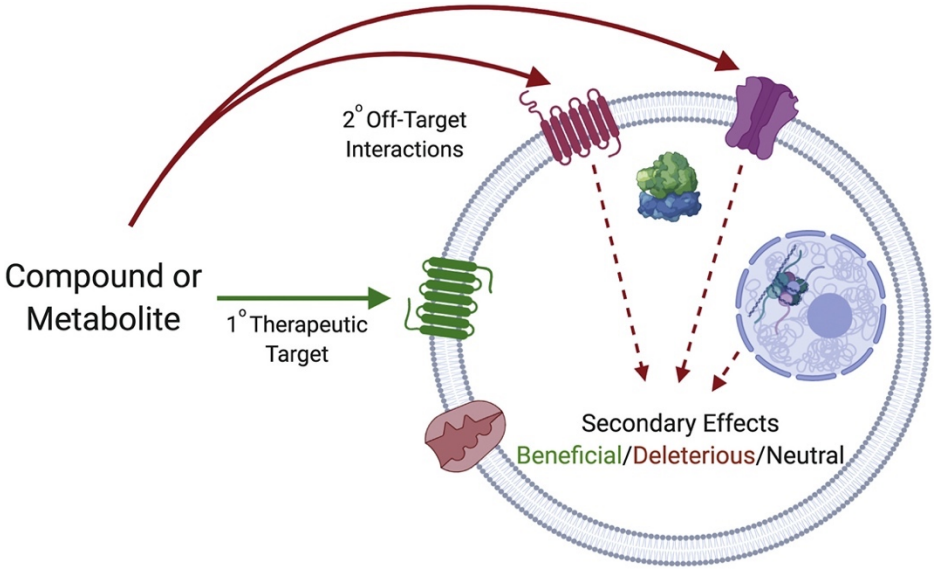


Figure 1.2: Schematic illustration of the potential drugs' interaction with primary and secondary pharmacological targets which may result in different effects. Primary target interactions show the drug's MoA and result in the desirable therapeutic effect, whereas secondary target interactions may cause undesirable effects. Reprinted from Jenkinson et al.[6] with permission from Elsevier.

1.2. The concept of Mechanism of Action in drug discovery

Understanding the 'Mechanism of Action' of therapeutic compounds is a key challenge during the drug discovery pipeline. MoA is the term used to describe the biological interaction through which a molecule produces its pharmacological effect[78]. The term 'Mechanism of Action' is also used interchangeably with the term 'Mode of action' with the main difference that the latter usually refers to the functional or anatomical changes at a cellular level induced by exposure to a substance, whereas Mechanism of Action includes specific targets or pathways modulated by the compound[9]. Understanding the biological mechanism through which a compound induces its pharmacological effect is important for many reasons, including the identification of toxicity or potential side-effects, or for rationalisation of a phenotypic effect to provide more confidence in a lead compound prior to clinical trial[10].

Being aware of a compound's MoA offers advantages, but it is not a requirement to get Food and Drug Administration (FDA) approval if the drug shows safety and enough

efficacy[11] (though Phase II testing may be shortened or skipped if the MoA is well understood[12]). It is estimated that between 10% to 20% of the currently approved drugs have no known target or clear MoA[13]. There are drugs with unknown, partially known, or uncertain MoA which have been used for decades. For example, acetylsalicylic acid (or aspirin) is a non-selective cyclooxygenase inhibitor and preparations containing this active ingredient were used for centuries for the treatment of fever or pain before the discovery that it inhibits cyclooxygenase activity in 1971[14]. Lithium, a drug used for the management of bipolar disorder, does not have a known MoA. Despite the suggestions that lithium is acting through multiple MoAs, such as direct inhibition of glycogen synthase kinase or inhibition of inositol monophospatase and others, the exact MoA through which lithium stabilises mood remains unclear[15]. Another example is the widely used drug Metformin – used in the management of type 2 diabetes – that entered clinical trials in the 1980s[16], but the drug's function is still unclear, other than some proposals such as AMP-activated protein kinase (AMPK) inhibition[17].

Despite the examples above, not knowing a drug's MoA holds more risks in the later stages of drug discovery and development process than knowing it. One example of a drug entering clinical trial with unknown MoA, which led to unwanted consequences is the case of Dimebon. It was initially used as an over the counter antihistamine in Russia since 1983[11]. Years later the same drug entered clinical trials as a potential therapeutic agent for Alzheimer's disease due to two studies in 2001[18] and 2003[19]. The former claimed that the drug enhanced the cognition in a) a rat model of the disease and b) in 14 individuals with this disease. The latter reported that the drug blocks mitochondrial dysfunction triggered by a fragment of amyloid-β (the neurotoxic molecule that build up in the brain in Alzheimer's disease). However, Dimebon failed to affect cognition in a large follow-up Phase III study, and this was attributed to the lack of understanding of its MoA. Further studies showed that actually Dimebon increased amyloid-β levels in a mouse model of Alzheimer's disease[20] and another study identified inhibition of histamine H_1 and serotonin 5-HT_6 receptors as the main biological mechanisms of Dimebon[21]. The proposed MoA explained the positive effects on cognition seen in the smaller-scale trials, but ultimately Dimebon did not stabilise mitochondria as first hypothesised. If this proposed mechanism was

elucidated before the clinical trials, then the failure of Dimebon could have been prevented.

The story of Dimebon underlines the importance of MoA studies in the development of new drugs and especially before they enter the “uncertain word” of clinical trials[11]. However, mechanistic studies are challenging because MoA can be defined on multiple levels of biology as illustrated in Figure 1.3. When a compound’s MoA is elucidated, it is usually defined as the “target(s)”, that the compound interacts with, which as a definition is complex for two main reasons. Firstly, the term “target” can refer to different types of molecular targets such as a protein or RNA molecule and others. Secondly the “target” definition is a relatively ‘shallow’ level of detail. The reason is that after target engagement a number of signalling proteins can be differentially regulated through cellular signal transduction, leading to changes in transcription, translation, metabolism and cell morphology[22] (Figure 1). Following the modulation of protein(s) by direct pharmacological action, cellular signalling proteins propagate signals *via* protein phosphorylation[23], catalysed by enzymes called kinases. These signalling cascades form pathways, which lead to a cellular response through the modified activity of so-called ‘effector’ proteins[24]. Signalling pathways can also interact with each other *via* ‘cross-talk’, forming networks and a coordinated cellular response[25]. Thus, a compound’s MoA can be defined on the systems-level in terms of the pathways that are modulated (signalling proteins), network perturbation, or by changes brought about to the cellular response (effector proteins) - and to further complicate things, the precise response will vary in different cells and tissues due to different patterns of protein expression[26].

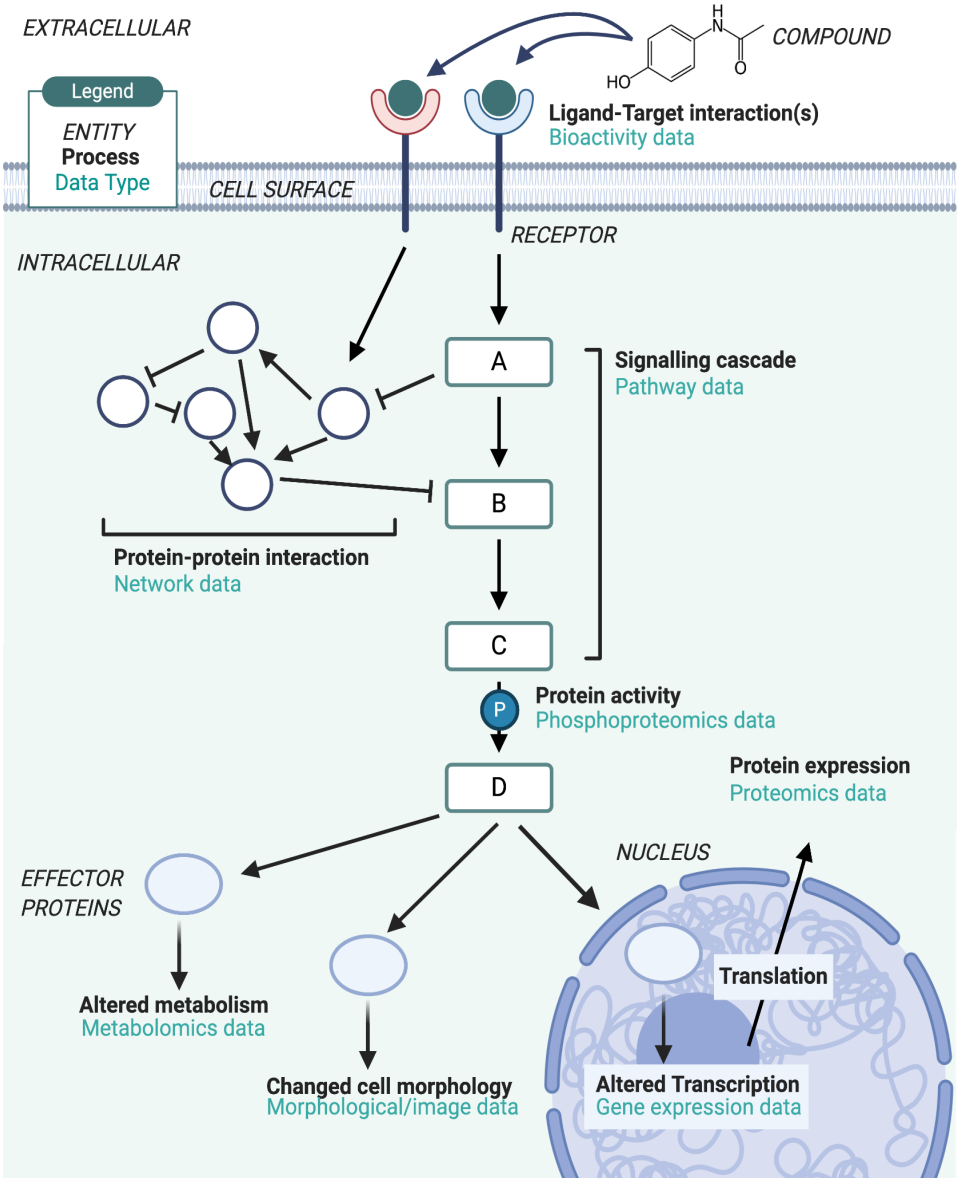


Figure 1.3: Overview of the different types of data/information used in MoA studies and the various levels that MoA can be defined on (Trapotsi et al.[27]).

Therefore, there are different angles to view the MoA of action definition. One could be to view it from a systems-level angle, and another could be to view it from the direct target prediction angle. The latter is a popular view and MoA has been extensively investigated on a protein-target level by predicting targets based on bioactivity data[28] with the ultimate goal to elucidate the MoA of drugs and possible off-target effects[29]. The rationale behind these methods relies on the assumption that structurally similar compounds are more likely to exhibit similar properties[30]. However, the premise is not always valid because of the complex biological processes[31] that occur during a biological dysfunction (i.e. disease) and also because compounds exhibit a broad

range of activity that could be beyond the bioactivity effect (Figure 1.3). The dysfunction can occur in different levels of the biological system such as the expression of genes, biological pathways and proteins.

1.3. Levels of biology for Mechanism of Action definition and different data types used

These different levels of biology which can define a compound's MoA can be captured and measured with different types of data, such as transcriptomics, cell morphology and metabolomics data (Figure 1.3), all of which provide a different aspect of the bioactivity of a compound. Additional information which catalogues known human pathways and networks can also be useful as supplementary prior knowledge to contextualise different types of data - for example, by relating differentially expressed genes to the pathways they participate in.

To better understand the MoA of compounds the use of a combination of different types of biological data can be very enlightening, in particular since the insight gained from different types of information can differ greatly. For example, two structurally similar compounds, the antidiabetic drugs rosiglitazone and troglitazone, exhibit a very different side effect profile due to their different MoAs[32]. Both compounds belong to thiazolidinediones class and treat insulin resistance in type 2 diabetes mellitus and the latter was withdrawn from the market because of hepatotoxicity while the former was developed as an alternative, which has been linked with cardiovascular diseases. A recent study docked the two compounds into predicted binding sites of more than 67,000 protein structures[33]. Targets of troglitazone such as 3-oxo-5-beta-steroid 4-dehydrogenase, neutrophil collagenase and others could explain why troglitazone causes hepatotoxicity. Results for rosiglitazone discerned its interaction with members of the matrix metalloproteinase family, which could lead to cancer and neurovegetative disorders. The concerning cardiovascular side-effects of rosiglitazone were also explained by relevant targets. Therefore, it is important to use different types of data to understand compounds' MoA more comprehensively and a summary table of these levels is shown below (Table 1.1) and are discussed in more detail below.

Table 1.1: Data types commonly used in MoA analysis, the level of biology represented, and some advantages and disadvantages of the data.

Level of Biology	Data Type	MoA Biology Represented	Advantages	Disadvantages
Direct drug-target interaction	Bioactivity	Compound-target binding and functional effects (e.g., activation, inhibition)	Relatively easy and cheap to measure (high-throughput screening or HTS)	Target binding *in vitro* is not necessarily indicative of target engagement *in vivo* due to e.g. ADME/PK effects[34] Does not inform about specific changes in cell signalling following target binding Not all target-ligand interactions are efficacy (MoA) related (i.e., could be side-effect related
Gene level	Transcriptomics	Changes in gene expression arising from modulated signalling (and transcription factor activity)	High-throughput techniques developed for large data acquisition[35] Provides a 'snapshot' of cellular changes in	High level of noise in data arising from fluctuations in biological activity[36]

			signalling following compound administration An array of standard analysis methods has been developed	Assumes gene expression is static, rather than a dynamic process[37] Does not necessarily translate to protein expression due to e.g. post-translational effects[38]
Proteome level	Proteomics	Changes in protein abundance arising from modulated signalling induced by a compound (transcription, translation, protein degradation)	Extends upon transcriptomics data by capturing changes in post-translational regulation	Data generation is costly and cumbersome[39] High biological variability/low reproducibility as well as significant technical variability[39] 'Missing value problem'[40]
Metabolome level	Metabolomics	Changes in metabolite abundance arising from modulated signalling induced by a compound (and metabolic enzymes)	Contains downstream products of transcriptomic and proteomic processes[41]	Data generation is costly and cumbersome, requiring multiple technical methods to capture the entire metabolome[42]

			Can also identify potential toxicity[42]	High biological variability/low reproducibility as well as technical variability due to e.g. long sample runs[43] Lack of comprehensive metabolite annotation and ability to relate to other biochemical components (e.g. enzymes)[44]
Phosphoproteome level	Phosphoproteomics	Changes in protein phosphorylation (protein signalling) induced by a compound	Captures the signalling proteins modulated, thus the specific biological pathways relevant to MoA. Links higher-level' bioactivity data and 'lower-level' e.g., transcriptomics data, enabling a 'systems-view'	Phosphorylation site annotation is not trivial and functional relevance is often unclear[46] Time-consuming assays limiting data availability[47] High biological variability/low reproducibility, as well as technical variability arising from MS instruments[45]

			High-throughput assays in development[45]	
Phenotype level	Phenomics (Cell images)	Changes in cellular morphology (e.g., size and shape of organelles) arising from modulated signalling (and changes in cytoskeletal protein activity)	High-throughput imaging techniques developed for large data acquisition[48] Feature extraction software and methods are evolving[49,50]	May not produce a meaningful signal if the compound is not able to alter cell morphology[51] Features are often highly correlated and biologically ambiguous[52] Requires orthogonal data to be able to relate changes to modulated genes/proteins[48] Phenotypic effects may be subtle and hence the biological signal

				can be overwhelmed by sources of technical variation[48]
Biological pathway level	Biological Pathway Information	Describes cascades of molecular interactions which have a defined entry point, signalling mediators, and cellular effect	Enables groups of genes/proteins to be characterised in terms of shared biological functions for ease of interpretation[53]	Static representation of a dynamic process[54] Interactions between pathways often not considered[53,55] Curation bias- well-studied processes more comprehensive and detailed, and overrepresented in pathway databases[56]

1.3.1. Direct drug-target interaction

One way to define a compound's MoA is to identify the targets that a compound binds to. High throughput experiments are used to evaluate such interactions with the main disadvantage that they are time-consuming and expensive. Therefore, computational methods such as *in-silico* protein target deconvolution is a well-established approach that offers an alternative solution to infer target-ligand interactions by utilizing known bioactivity data/information[57,58].

1.3.1.1. Bioactivity data

Compound-target activity, or 'bioactivity' data translates target binding into a numerical value, usually in terms of a concentration where target activity is seen (or % of some functional effect such as target inhibition) (Table 1.1). In more detail, the biological activity is quantified by using a dose-response relationship and a measurement/experiment is conducted at varying doses to evaluate what the drug actually does against what the drug is and how much is present. This observation has resulted in different ways of measuring the biological activity and hence different equations and definitions exist for understanding and quantifying biological activity. The most commonly used are the: dissociation constant (K_d), inhibition constant (K_i), half maximal inhibitory concentration (IC_{50}) and half maximal effective concentration (EC_{50}).

Unliganded receptor (R) has a probability of being bound by a ligand that is proportional to the ligand concentration (L) and the probability of binding. When the ligand binds to the receptor the receptor with the bound ligand (RL) has a fixed probability of relinquishing it within an equivalent period[59] (Equation 1.1). The K_d measures the equilibrium between the ligand-protein complex and the dissociated components or in other words how much a large object dissociates reversibly into smaller constituents (Equation 1.2).

$$RL \overrightarrow{\leftarrow} R + L \qquad \text{(Equation 1.1)}$$

$$K_d = \frac{[RL]}{[R][L]} \qquad \text{(Equation 1.2)}$$

There is also another measure; named inhibition constant (K_i), which is representing a dissociation constant but more narrowly for the binding of an inhibitor (ligand) to a protein or enzyme and shows the affinity of an inhibitor (Equation 1.3).

$$K_i = \frac{IC_{50}}{1+\frac{[S]}{K_m}} \quad \text{(Equation 1.3)},$$

where [S] is the substrate concentration, K_m is the ligand concentration (without an inhibitor) at the half maximal velocity of the reaction. IC_{50} corresponds to the concentration required to cause an inhibitory effect by 50% and can be also calculated by the equation 1.3 if rearranged. The K_i is usually used when the binding constant is measured through inhibition kinetics, while the K_d is preferred when the binding is measure in a more direct way (e.g., fluorescence quenching).

The IC_{50} and EC_{50} are usually used to measure the potency of a compound, which is a way to quantity the bioactivity of drugs. The EC_{50} corresponds to the half maximal effective concentration and IC_{50} corresponds to the half-maximal inhibitory concentration and can be described by various equations or could be derived from concentration-effect curves[60] as shown in Figure 1.4. The definition of EC_{50} was created because of the need to compare molecules to each other for their effects on a biological component. The IC_{50} is only measured in the context of enzymatic activity, or of a receptor, whereas the EC_{50} can be measured for complex cellular processes[60]. Measurements such as Ki, EC_{50}, IC_{50} etc. are different from one another in practice, but different affinity and potency measures are being merged in order to maximise the amount of usable bioactivity data derived from varied sources, such as for modelling, and one study found that this only moderately increases noise in the combined data[61].

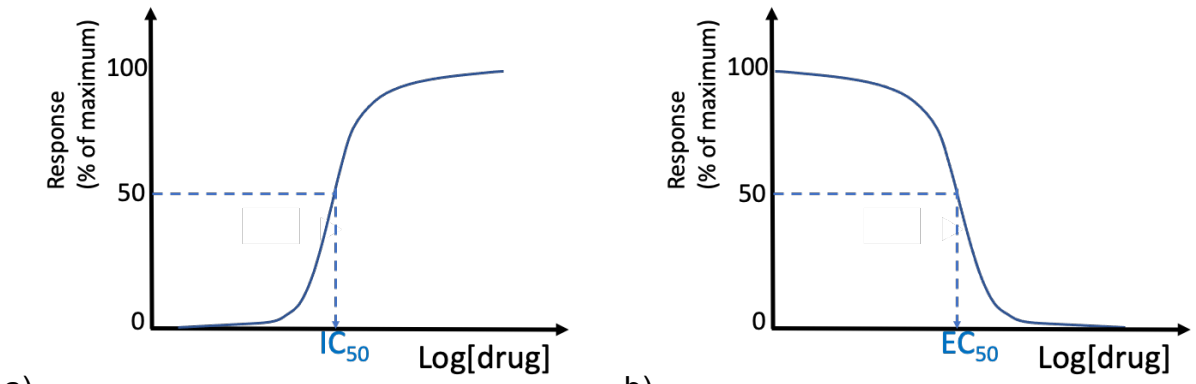


Figure 1.4: Concentration-effect curve of the intensity of the response against the logarithm of the drug concentration, and position of the a) IC_{50} and b) EC_{50}.

The metrics described above are a way to quantify bioactivity and hence this type of data is highly valuable in MoA studies as it can be used to predict targets for orphan compounds[62], or to inform about drug repurposing opportunities[63]. High-throughput screening (HTS) technologies have been developed which enable the screening of thousands of molecules against panels of compound targets, thus large-scale databases of bioactivity measurements are available[64–67]. Although, *in vitro* target binding is not necessarily indicative of target engagement *in vivo*, due to how compounds are absorbed, metabolised, metabolised and excreted (ADME) in a biological system, governed by the compound's pharmacokinetic (PK) properties[34]. This is indeed relevant to any type of biological data measured *in vitro*, but attempts have been made to consider this in bioactivity data by utilising experimental properties such as maximal blood concentration (CMax) and plasma protein binding (PPB)[68]. Furthermore, this can be considered a relatively 'shallow' level of data, since it does not inform about any changes in the many cellular signalling pathways which can be modulated following target binding, and hence the relationship between binding and a functional effect of interest needs to be determined. Additionally, target binding may not necessarily be indicative of MoA, as the so-called 'promiscuity' of some compounds means that they may bind to many 'off-targets'.

Bioactivity data can be accessed publicly in databases such as ChEMBL[65,66], PubChem[67], ExCAPE[64] and BindingDB[69] (Table 1.2). ChEMBL contains more positive/active data points because data are derived from literature, compared to

PubChem, where there is a plethora of negative bioactivity data from HTS experiments. Another database is the ExCAPE (Exascale Compound Activity Prediction Engine) database, which is an integrated version of ChEMBL (version 20) data and PubChem data (extracted in January 2016). One parameter that should be taken into account is the chemical space coverage of the chemical structures in the bioactivity databases. Despite the fact that millions of chemical data have been deposited in such databases (e.g. more than 15 million bioactivity data points for ~2 million compounds, including compound interaction data against ~8,000 protein targets in ChEMBL) and the exponential increase of such data because of the application of parallel and combinatorial synthesis approaches, the available data corresponds only to a small part of chemical space of all possible molecules[70]. This is a parameter that should be considered when extracting and using data for projects that focus on 'poorly explored' areas of chemical space.

Table 1.2: Main sources of bioactivity data, their size/coverage and additional comments

Source	Size/Coverage As of July 2020.	Comments
ChEMBL	2M compounds 1.2M assays 13K targets	Extraction of compound, assay and bioactivity information from journal articles is performed manually by curators
PubChem	103M compounds 253M substances 1.1M assays 95K proteins	Data provided by more than 350 contributors including university labs, government agencies, and pharma companies. Data include siRNAs, miRNAs, carbohydrates, lipids, peptides, and other substances
DrugBank	13.5K drugs 5K proteins	Contains data about FDA-approved drugs as well as experimental drugs going through the FDA approval process

ExCAPE	~1M compounds 1,667 targets	Over 70 million SAR data points extracted from PubChem and ChEMBL and merged in one database across 3 species (human, rat and mouse).
BindingDB	312,146 compounds 1,858 targets 5,928 assays	Database for binding measurements, focusing on the interactions of protein considered to be drug-targets with small, drug-like molecules.

In addition to the chemical space coverage another concern is the quality of such data and the experimental variability. In reality, the maximum achievable accuracy of *in-silico* models depends on the quality of the experimental data (i.e. when models approximate experimental error)[71]. Firstly, in order to better understand the deviation of activity values across the different protein targets to be modelled, one must first explore the experimental variability of bioactivity data in chemogenomic repositories. One such study of public bioactivity data was performed by Kramer et al.[72], who analysed the biological activity data deposited in ChEMBL[73] (version 12) for reproducibility (i.e., the experimental uncertainty of independent measurements). The experimental uncertainty was estimated to yield a mean error of 0.44 pK_i units, a standard deviation of 0.54 pK_i units, and a median error of 0.34 pK_i units. The maximum possible squared Pearson correlation coefficient (R^2) on large data sets was estimated to be 0.81. Further, the heterogenous use of public biochemical IC_{50} data was shown to be problematic, because they are assay specific and comparable only under certain conditions[61]. This phenomenon is particularly relevant for large scale datasets used in target prediction, since it is not feasible to check each data entry manually and it is commonplace to mix available IC_{50} values from public databases even if assay information is not reported. In a similar manner, Kalliokoski et al.[61], analysed the types of errors, redundancy and variability in ChEMBL. IC_{50} variability was assessed comparing all pairs of independent IC_{50} measurements on identical protein-ligand systems. The standard deviation of pIC_{50} data (equal to 0.68) was only 25% larger than the standard deviation of K_i data, suggesting that mixing IC_{50} data

from different assays without knowledge of assay conditions adds a moderate amount of noise to the overall data. The standard deviation of public ChEMBL IC_{50} data, as expected, resulted greater than the standard deviation of in-house intra-laboratory/inter-day IC_{50} data, which showed a standard deviation of pIC_{50} values equal to 0.22 and 0.17 for two different drug-target combinations. Augmenting mixed public IC_{50} data by public K_i data was not found to deteriorate the quality of the mixed IC_{50} data if the K_i is corrected by an offset. For the ChEMBL database, a K_i-IC_{50} conversion factor of 2 was suggested.

Another study reported a median discordance (margin between pXC_{50} values) of 0.48 between laboratory measurements for proteins within the same organism, and 0.42 after discriminating between assay type[74]. Further aggregation of bioactivities observed in human and related (orthologue) biological systems (a common practise during data assimilation to increase data quantity[75,76]), also increased the median standard deviation to 0.51, respectively. Experimental error has also been analysed for proprietary datasets and a recent AstraZeneca study focused on a systematic evaluation of biological assay variability of all biological assays between 2005 and 2014[77]. The authors found less than a two-fold difference in the average experimental uncertainty, where EC_{50} and IC_{50} measurements tend to have lower standard deviations (with a standard deviation above 0.5), compared to K_d and K_i measurements. Novartis analysed randomly picked (repeatedly measured) samples of typical assay endpoints over several years and calculated a standard pIC_{50} deviation of ~0.2 log units[61]. Hence, experimental error is also observed within the same laboratories.

Another factor affecting the deviation of results in bioactivity data is the inconsistent mining and preparation of data for structure-activity modelling. For example, Fourches, et al.[78]. emphasised the need for standardised chemical data curation strategies (e.g., curation of chemical structures and biological data) that should be followed at the onset of any molecular modelling investigation to avoid discrepancies. Another study highlighted the importance of data selection and extraction, and proposed the combined application of various query parameters available to any user of the ChEMBL database and other selection criteria (such as common compound

promiscuity) to harmonize data retrieval[79]. Moreover, discrepancies between bioactivity data in public databases could arise from errors in the data curation and Tiikkainen et al.[80] raised awareness on the frequencies and types of errors in bioactivity data. Error rates for three large bioactivity databases, namely ChEMBL (version 14), Liceptor (version 2012_03) and WOMBAT (version 2012.01) were calculated. The authors observed that the ligand structures showed the highest probability of being discrepant followed by the protein target, activity value, and finally the activity type. Errors in activity values mainly arose due to unit conversion issues (e.g., micromolar affinities curated as nanomolar) and the activity type (e.g., IC_{50}, K_i, etc.) are usually clearly stated in the source articles. Hence, curation-related errors increase the possibilities of non-systematic error in public bioactivity datasets and consequently increase uncertainty for ligand-target annotations. The possibility of experimental annotation error should also be accounted for during modelling, which is the topic of chapter 2 of this thesis by using a methodology called Probabilistic Random Forest (PRF).

In *in-silico* direct drug-target interaction prediction, the bioactivity data are used as an endpoint and therefore a type of information is required for the compounds. Information can be a compound high throughput profiling method or calculated from the compounds' chemical structure. Chemical structure information is usually used because of the assumption that structurally similar compounds should show a similar bioactivity profile. For example, the structure of a chemical compound can be represented mathematically by molecular descriptors, which can be based on physiochemical properties (e.g. measured experimentally) or calculated from compounds' chemical structure (e.g. Morgan Fingerprints)[81]. A widely used example are the circular topological fingerprints such as the Extended Connectivity Fingerprints (ECFPs)[81]. They are widely very popular molecular type of information in throughput drug discovery especially for similarity searching and Structure Activity Relationships (SAR) modelling. However, chemical structure information might not be appropriate or enough to inform of a compound's bioactivity or response to biochemical assays. An example of such a case is the presence of 'Activity Cliffs', where only small transformations to similar structure compounds result in a large difference in potency and bioactivity profiles[82]. Indeed, it has been shown that only 30% of compounds with

high similarity to an active compound are themselves active at the same target[83]. This highlights the need for additional compound representation beyond chemical structure. Examples of such descriptors are the expression response of the 978 LINCS "landmark genes"[84] or cell morphology changes in the form of microscopy images or calculated features[85].

1.3.2. Gene level

Transcriptomics refers to the study of transcriptome/mRNA transcripts that are produced by the genome in a specific cell line or under specific circumstances[86], and serves as an intermediate source of information for the understanding of the effect of a compound in a biological system on the gene level. The information derived from the quantity of mRNA transcripts can inform about the effect of compound perturbations on protein control at the mRNA level. The differential gene expression profile (the difference between the neutral control and treated samples) of a compound can be considered as an extra indicator of its MoA because it provides a holistic image of the cell upon perturbation or compound treatment. However, there are disadvantages related to gene expression profiles of compounds. Gene expression might not be meaningful when considered on its own or a gene expression signature may be noisy, and these two factors should be taken into consideration[87]. Moreover, it is important to mention that different genes can be regulated at different time points. For example, in a study investigating two structurally similar compounds (vincristine and vindesine with Tanimoto coefficient equal to 0.91), these two compounds were tested on the same cell line (A549) and dose (10 μM) but different perturbation times (6 and 24h). The two compounds had similar gene expression profiles when the perturbation time was 24h but different profiles when the perturbation time was 6h[31]. The reason is that genes are regulated at different time points and in more detail the compounds' target, which is the TOPoisomerase 2 Alpha (TOP2A), was downregulated for both drugs when the cell is treated for 24h but on the other hand the target is not affected at 6h.

The genome-wide expression profiling of transcriptional responses upon compound perturbation has gained an increased interest in the exploration of the MoA of bioactive compounds[88]. The reason is that compounds with comparative biological properties

could share a commonality in their MoA and thus compounds with similar gene expression profiles can have a similar MoA[89,90]. Hence, in recent years compound-induced gene expression repositories have been created and evolved and gene expression data have become available in public databases for comparison. There are two databases that contain transcriptomic profiles of cultivated cell lines treated with thousands of chemicals and are widely used as reference datasets. These two databases are the Connectivity Map (CMap)[91] and the Library of Integrated Network-Based Cellular Signatures (LINCS)[35].

1.3.3. (Phospho)Proteome level

Proteomics data measures changes in protein abundance (due to modulation in translation or degradation) arising from compound-induced protein signaling[92]. Proteomics data is complementary to transcriptomic data as it informs about cellular processes following transcription, such as translation and post-translational modifications. By studying interrelationships of protein expressions and modifications following a drug treatment, important insights of a compound's MoA, toxicity and side effect profile can be identified[93]. Therefore, the knowledge about which proteins are differentially expressed due to a compound treatment can inform researchers about the proteins which are key to its mechanistic action. Due to technological limitations (LC-MS/MS measurements can take several days or even weeks to run), data generation is costly and time consuming, and leads to biological variability between replicate measurements (due to e.g., decay in performance of columns over the course of a long experiment)[39]. Another limitation of proteomics data is that not all proteins are quantified in all experiments (missing value problem), though this can be addressed by using data derived from multiple assays to obtain a larger coverage of the proteome[94] or through imputation[40].

Phosphoproteomics data captures changes in the phosphoproteome; the phosphorylation states of signalling proteins (Table 1.1). Cellular signalling is mediated by protein phosphorylation on serine, threonine and tyrosine residues[23], thus by understanding the changes in phosphorylation states of signalling proteins following compound treatment we can infer potential pathways modulated by the compound, beyond information that is visible on the transcriptional and translational

level alone. Phosphoproteomics data is particularly useful in -omics studies as it allows us to build up a "systems-level" view of compound mechanism of action by filling in the gaps downstream of target binding and upstream of changes to effector proteins (e.g., transcription factors, which is reflected in transcriptomics data). One limitation of phosphoproteomics data is that the annotation of phosphorylation sites is not trivial due to for example the presence of multiple serine, threonine and tyrosine residues in one peptide[46]. To address this limitation, services such as PhosphoSitePlus®[95] have been developed which map phosphorylation sites to proteins, and provide biological context through disease and pathway annotations. Furthermore, phosphoproteomic profiling of compounds is time consuming and expensive[47] - though this has been addressed with the P100 assay, which measures only 100 phosphorylated peptides from cellular proteins and thus serves as a reduced representation of the phosphoproteome[45,47]. Furthermore, much like changes to transcription, metabolism and translation, phosphorylation changes are highly variable, and there is added technical variability arising from MS instruments, hence replicate experiments are necessary to ensure the reliability of the data[45].

1.3.4. Metabolome level

Metabolomics data captures the presence of metabolites (small molecules < 1500 Da), and thus primarily captures perturbations to metabolic enzyme activity induced by a compound as a "functional readout of the physiological state"[96] (Table 1.1). Changes on the mRNA level (transcription), lead to changes in translation and protein expression (proteomics), including the expression of enzymes involved in metabolism, thus metabolomics is a complementary source of data which can be integrated with other data types to gain a deeper understanding of MoA on a systems-level[97,98]. Furthermore, as some metabolites are considered to be toxic, metabolomic data can inform about potential off-target effects of a compound to infer its potential safety, or to understand the metabolic pathways perturbed by the compound[42]. Similarly to proteomics data, the main drawback of metabolomics data is that experimental methods are subject to technological limitations - for example multiple methods are required to capture the entire metabolome[41], and difficulties in metabolite deconvolution due to differential fragmentation matters in mass spectrometry measurements[99] as well as a lack of comprehensive metabolite annotation[44], this is

known as the ‘greatest bottleneck’ of metabolomics data interpretation[100]. Again, the metabolome is highly variable and thus must be accounted for by performing replicate experiments - and untargeted approaches performed in different labs have shown wide variation due to experimental variation arising from long sample runs[43,101].

1.3.5. Phenotype (Cell Morphology) level

Cell image or cell morphology data captures the morphological changes which occur when a chemical compound is applied on cell cultures and therefore reveal the phenotypic effect of compounds on the cells[102] (Table 1.1). Such data can show any cell morphological characteristics upon compound perturbation and hence readouts have a general nature, being particularly popular in MoA and toxicology research[103]. This level of information is not specifically tailored as a landmark approach to understand MoA as other levels, which use fix endpoints such as the direct drug target interactions. However, changes in the cell morphology could be a useful information for MoA analysis.

The high throughput imaging has been benefited from the availability of the technology of robotic sample preparation and appropriate microscopy equipment and additionally the recent availability of large libraries of genetic and chemical perturoations[104]. Hence, the availability of image-based data has resulted in an increasing interest for this type of data. Image-based data is used in pharmacological screening to depict cell morphological characteristics or as the phenotype of the cell that is studied[104]. Pharmaceutical companies are using high throughput image data to triage compounds to be used in drug discovery projects[105] and additionally there are public databases, where image data are stored.

Recently, new assays have been developed for large data acquisition, such as the Cell Painting assay (Figure 1.5), which is a multiplex cytological profiling assay that measures diverse cellular states upon compound perturbation[106]. In more detail a set of well characterised fluorescent morphological labels are used to stain seven organelles and cell components[106]. These are the following: nucleus, endoplasmic reticulum, nucleoli, Golgi apparatus, plasma membrane, F-actin and mitochondria. Compounds are applied on Human Bone Osteosarcoma Epithelial Cells (U2OS) and

morphological features are calculated from the images with the opensource CellProfiler software. Automated feature extraction methods have been under much development recently, such as neural networks[50] and segmentation programs such as Columbus and CellProfiler[107].

A variety of image-based datasets have been developed and deposited in public repositories such as the Broad Bioimage Benchmark Collection (BBBC) developed by the Broad Institute[108] and other databases such as the 'Cell Image Library' and the Image Data Resource (IDR)[109]. A large dataset of 30,616 compounds was released in the GigaScience database by Bray et al.[104] , including a variety of perturbations (drugs, natural products, small probe molecules, diversity-oriented synthesis compounds) and numerical image-based features/descriptors. Furthermore, there is a joint effort from Imaging Platform at the Broad Institute of MIT and Harvard with 12 industry and non-profit partners with the aim to release a large reference collection of image data with 1 billion cells responding to over 140,000 small molecules and genetic perturbations, which will greatly benefit researchers seeking access to this data type[110]. Moreover, Recursion Pharmaceuticals is focusing on combining high-content phenotypic screening with machine learning for emerging opportunities in target discovery, and hit identification, releasing their datasets in the public domain[111].

One of the main disadvantages of image-based data is that not all compounds are able to change cellular morphology[51]. Therefore, it is important to select compounds for downstream analysis that are 'active' on the image assay - i.e., compound's image-based profiles that are significantly different from the neutral control wells. This process involves arbitrary cut-offs to define how different a compound is to the control wells and distance-based metrics such as Euclidean distance[112]. In addition, when curating image-based data it is important to evaluate potential intra or inter plate effects as well as the reproducibility between replicate measurements[113]. This is particularly relevant for morphological end-points because phenotypic effects may be subtle, hence the effect of technical variation may overwhelm any biological signal in the data[48]. Furthermore, cell morphological features can often reflect technical properties of the image rather than biological characteristics of the cell, and there is high redundancy among morphological features[52].

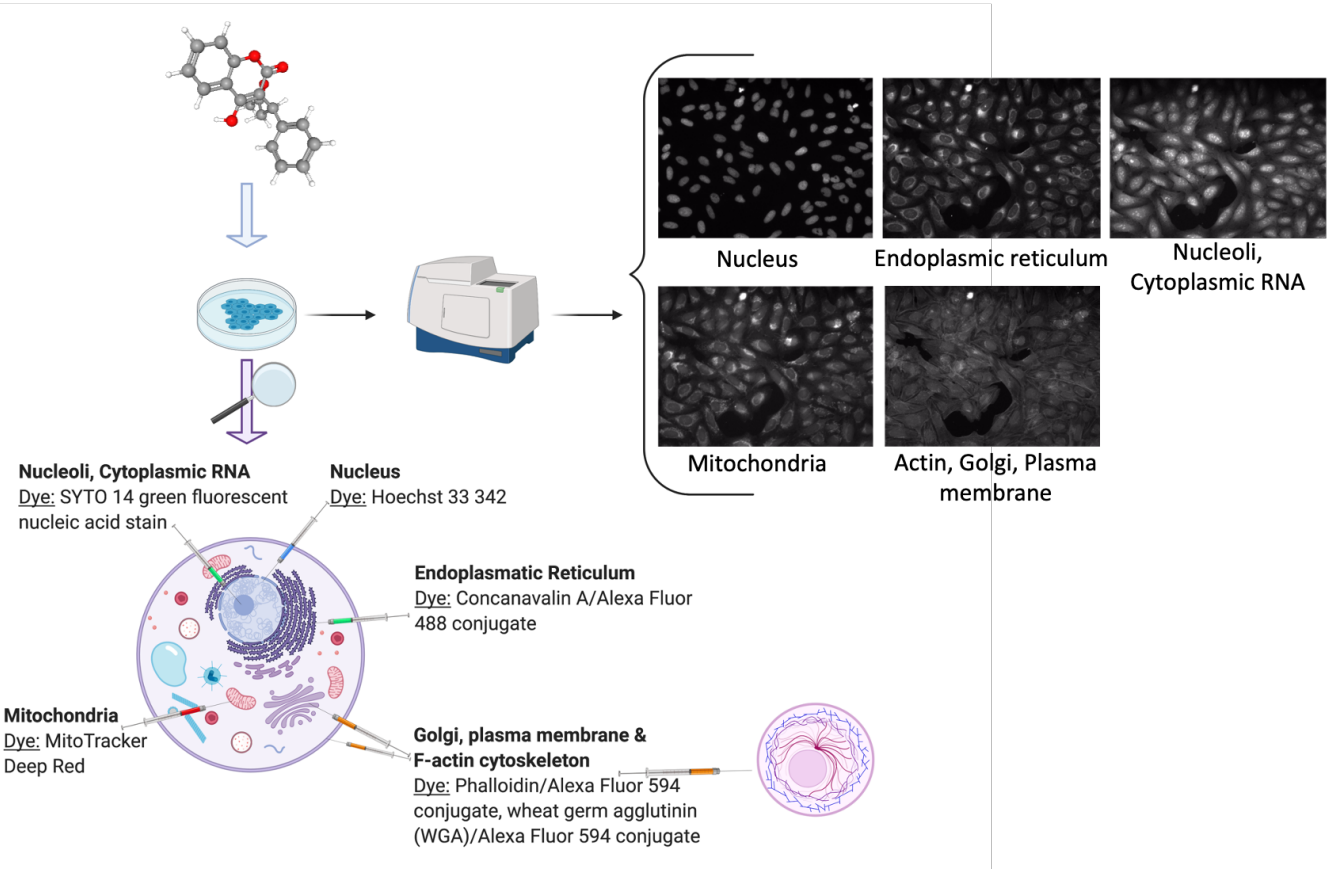


Figure 1.5: Schematic description of the Cell Painting assay demonstrated with the Warfarin compound. Created with BioRender using cell images from the Image Data Resource (IDR0036) (Trapotsi et al.[27]).

1.3.6. Biological pathway level

Biological pathway data is often used to supplement compound-based data such as transcriptomics, (phospho)proteomics and others. This type of data is a source of prior knowledge, which enables biological interpretation[53] and is useful for MoA studies as it links genes/proteins to biological processes and are thus easily interpretable by bench biologists. For example, if a compound induces differential expression of a set of genes known to participate in a certain pathway, then it can be inferred that this pathway is involved in the compound's MoA. Pathway data outlays cascades of molecular interactions which have a defined entry point and cellular effect, for example the JAK-STAT pathway which begins with the modulation of JAK and ends with apoptosis and cell cycle progression[114].

Pathways are outlining cellular molecular interactions, but they are simplified and aim to capture a particular cellular process. This raises questions about how

representative such pathways truly are of the processes they aim to recapitulate, as active entities in a pathway are highly dependent on cell type and context, and they additionally act in a dynamic fashion, while pathways are usually represented as static, standalone processes[54]. Nevertheless, for convenience and ease of interpretation pathways are represented as a 'snapshot' at a given time as governed by the information source the data is mined from, so this must be kept in mind when generating hypotheses using pathway annotations. Additionally, no information on their interactions is taken into account - pathways do not function independently in biological systems[55] therefore these interactions are being catalogued in the public domain to address this shortcoming[115]. Finally, curation bias is also present in pathway data - well-studied processes have more complete or detailed annotations and are also more over-represented in databases, hence again leading to bias in downstream data analysis[56].

Different sources of pathway data (Table 1.3) have been previously reviewed by Chowdhury et al.[54], where each source was comprehensively analysed for researchers to choose the most suitable database based on their needs. For example, Reactome[116] and WikiPathways[117,118] are useful for pathway data sharing due to the way the data is formatted and readable in third-party programs. Pathway data are contained in a number of databases (Table 1.3), and include KEGG[119,120] (mainly metabolic pathways), Reactome[116] (manually curated), WikiPathways[117,118] (collaborative database), HumanCyc[121] (mainly metabolic pathways, but also annotated with gene essentiality and other protein features), Pathway Commons[122] and BioSystems[123] (integration and standardisation of several databases). As well as pathway databases, the Gene Ontology (GO)[124] annotates biological processes, molecular functions and cellular components with their associated proteins - these terms are organised as a hierarchy and are used in much the same way as pathway data in mechanism of action analysis. GO terms are often considered to be highly redundant[125] (multiple terms describing the same or similar process), leading to the development of specific tools for "trimming" GO annotations such as REViGO[126] and GOATOOLS[127].

Table 1.3: Main sources of pathway data, their size/coverage and additional comments.

Source	Size/Coverage As of July 2020	Comments
Reactome	2,423 human pathways 13,248 reactions 10,923 proteins 1,869 small molecules	Reactome is manually curated and peer-reviewed, pathways are arranged in hierarchy under 27 high-level headings such as “Cell Cycle” and “Metabolism”[116]
KEGG	537 human pathways 11,274 drugs	Mainly metabolic pathways, but also contains signal transduction and disease pathways[119,120]
WikiPathways	1,185 human pathways	Open and collaborative platform for curation of pathways by the biology community[117,118]
GO	28,923 Biological Processes (BP) 11,136 Molecular Functions (MF) 4,185 Cellular Processes (CC), across 4,643 species	Not strictly pathways but processes, follows ontology[124]
NCBI BioSystems	3,077 human pathways	Contains records from several source databases (Kegg, BioCyc, Reactome, NCI’s Pathway Interaction Database, Wikipathways and GO), allowing for easy integration with other NCBI databases[123]
HumanCyc	(Last updated 2017) 314 pathways	Subset of BioCyc for Homo Sapiens - metabolic pathways curated from

	2887 reactions 20,830 genes 1,929 compounds	publications and integrated with other databases such as gene essentiality, regulatory networks, protein features, and GO annotations. Subscription required to access most of HumanCyc and BioCyc in general beyond a limited period of free use[121]
Pathway Commons	5,772 pathways	Collects pathway and interaction data (22 different databases) and represents them in the BioPAX standard that aims to enable integration, exchange, visualization and analysis of biological pathway data[122]

Furthermore, a final key limitation of pathway databases is the discrepancies found between pathway databases due to differences in data curation. An example of such differences for mTOR signalling pathway from three different data sources is shown in Figure 1.6. As we can observe there is no perfect overlap between the three data sources and in this specific case the pathway information retrieved from Reactome is a fraction of the information retrieved from KEGG or Wikipathways. Thus, tools such as PathMe[128] can be used to interrogate these differences and to extract consensus pathways or choose the most comprehensive or appropriate annotation database.

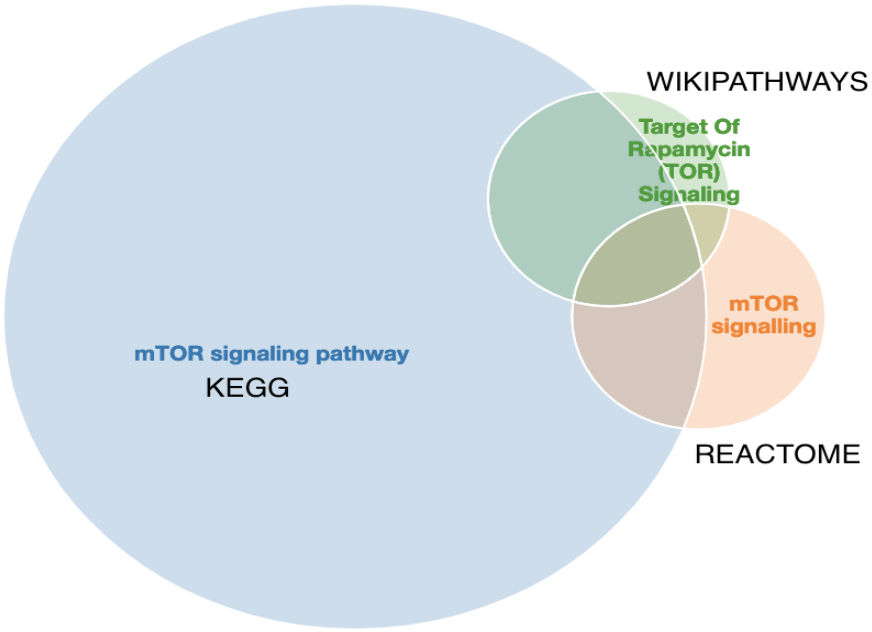


Figure 1.6: The merged mTOR signalling pathway from KEGG (blue), Reactome (orange) and Wikipathways (green) visualised in PathME viewer. The intersection sizes represent the number of entities in common vs. the number of entities in each pathway. For the same pathway, the information from 3 different sources varies. Visualisation created with PathMe Viewer (Trapotsi et al.[27]).

1.4. Methods and their Applications in Mechanism of Action Elucidation

There are various types of methodologies, which can be applied to elucidate compounds MoA, from pathway analysis methods to unsupervised and supervised machine learning. These methodologies and their applications differ in their prerequisites (e.g., data required, limitations in annotations, and computational complexity) and will be reviewed in this section.

1.4.1. Unsupervised Machine Learning

1.4.1.1. Clustering

Clustering methods are commonly used in bioinformatics projects and serve as the first step in data analysis in order to identify groups of samples that are may be related or interacting[129]. Grouping of data into clusters is based on similarity or distance-based metrics (e.g., k-means clustering) or based on data density (e.g., DBSCAN). Clustering is usually used to analyse unstructured and high-dimensional data such as gene expression, chemical and image-based data in order to better understand biological processes on various biological levels[130]. The most popular clustering algorithms are grouped into 3 different categories: hierarchical clustering (HC),

centroid-based clustering (CC) and density-based clustering (DB). Hierarchical clustering seeks to build a hierarchy of clusters, and can either be agglomerative ("bottom-up", each observation is assigned its own cluster and pairs of clusters are merged) or divisive ("top-down", all observations start in one cluster and splits are performed recursively)[131]. Centroid-based clustering represents clusters by a central vector, and in the case of k-means clustering finds *k* 'cluster centres' and assigns samples to the nearest cluster centre, such that the squared distances from the cluster are minimised. In density-based clustering, clusters are defined by areas of higher density than the remainder of the data set, while samples in sparse areas are considered to be noise or outliers. Two parameters must be set for density-based clustering, namely; the minimum number of data points needed to determine a single cluster, and how far away any one point can be from another point in the same cluster. A more recent type of clustering uses classic clustering methodologies combined with Deep learning approaches. Deep Neural Networks (DNNs) can be efficient in transforming mappings from a high-dimensional data space into a lower-dimensional feature space, which theoretically can lead to improved clustering results and an extensive review on such methods has recently been published by Karim et al.[130].

The major benefit of clustering applications for MoA elucidation is that it is relatively fast (in particular centroid-based and density-based clustering, while hierarchical clustering is more time-complex)[132] and are useful when compounds are annotated with their MoA. If compounds that share the same MoA cluster together (in a particular biological space), query compounds can be interrogated for their cluster identity and thus MoA. Such a study which aimed to evaluate whether the clustering of compounds using Cell painting profiles similarity would group compounds with similar annotated protein targets or chemical structure[106]. The authors applied hierarchical clustering on 75 active compounds (on the Cell Painting assay), and they ranked the clusters' enrichment of annotation terms. The enrichment calculation was performed by applying a permutation testing by measuring the fraction of random clusters of the same size that had at least the same number of compounds annotated with the term in question. The construction of random clusters for the permutation testing was performed by drawing from a uniform distribution over the compounds. The clusters that were enriched for annotation terms were convincing mechanistic groups, such as

a cluster that contained modulators of tubulin, which displayed large multinucleated cells with fused nucleoli. Another enriched cluster contained compounds modulators of neuronal receptors, with a different phenotypic response (compared to other enriched clusters), which included enhanced Golgi staining and some cells with fused nucleoli. This observation provided evidence that image data is a source of information that is closely linked to the MoA of a compound.

In a more recent approach, image-based data were also used but this time from a live-cell imaging assay in order to catalogue the morphological phenotypes of 1,008 reference compounds, which were well annotated and their MoA was known[51]. The image data were profiled at four concentrations and in a panel of 15 reporter cell lines, which expressed 12 different fluorescently labelled proteins from their endogenous chromosomal loci, which enables the monitoring of cellular organelle morphology and the activity of various signalling pathways. Features were calculated with Columbus image analysis software and a set of representative features was selected by applying a modified version of the minimum redundancy and maximum relevance (mRMR) algorithm, and phenotypically active compounds were selected based on their Euclidean distance from the controls (DMSO). Hence 57% of the compounds were active in at least 1 cell line. Image features were projected onto a 2-Dimensional t-SNE plot, which showed that there were compounds with the same MoA annotation, which formed clusters like in the case of kinesin inhibitors and some others that did not. Therefore, the coherence of the MoA clusters (with a minimum of 3 members) was evaluated by ranking all compounds' image features (in a single cell line) based on their Pearson correlations and calculating the AUC-ROC value for the categorization of having the same MoA annotation or not. Results showed that 41 out of 83 MoA categories could be accurately distinguished with a ROC-AUC $\geq$0.9. Therefore, there were cases, where image features were successful into clustering compounds with the same MoA and such models can be useful for MoA hypothesis generation.

In addition, another property of clustering analysis is that it can be carried out on one or multiple levels of data, thus clusters can be compared in different spaces (to understand if, for example compounds which cluster together based on their

phosphoproteomic response act similarly on the transcriptional level). Such a comparison was recently performed by Way et al.[133], who compared the ability of L1000 and Cell Painting-based profiles to cluster compounds with the same MoA annotation. They profiled 1,327 molecules with the two assays at six treatment doses on A549 lung cancer cells. The rationale behind this comparison was that these two assays capture some shared and some complementary information in mapping cell state. They applied a clustering analysis using k-means clustering across a range of cluster numbers between k=2 to k=40 and three goodness of fit heuristics, namely, the Silhouette, Davies Bouldin and Bayesian Information Criterion (BIC) scores. The Silhouette indicated how separable are the clusters, the Davies Bouldin quantified the ratio of within-cluster distances to between cluster distances when comparing each cluster to their most similar neighbouring cluster and the BIC was used as a measure of cluster likelihood and predictability. The clustering was performed on the 350 Principal Component space of each assay profiles for each of the six treatment doses. Results showed that Cell Painting can produce more distinct clusters compared to L1000. This observation means that Cell Painting produces more diverse cell states at least under the experimental conditions used in their analysis.

One limitation of the studies above is that in order to cluster the compounds and evaluate them, a MoA annotation is required, and this might not be possible if the analysis is conducted with compounds, which do not have such annotations. Hence, Patel-Murray et al.[134] proposed a multi-omics (transcriptomics, proteomics and metabolomics, as well as epigenomics) approach which does not require reference compounds or large databases of experimental data in related systems, and thus can be applied to the study of agents with uncharacterized MoAs and to rare or understudied diseases[134]. The aim of this study was to better understand compounds' MoA and their beneficial effects in models of Huntington's Disease (HD). To reveal similarities between compounds profiles, clustering was performed in different spaces separately. In gene expression space, the profiles formed only one distinct group, whereas two distinct groups were observed in the metabolite profiling data. Interestingly, the compounds clustered together did not have the most similar chemical structures and also compound pairs with high connectivity scores did not cluster together in the omics data space. This observation highlighted that the assumption of

“compounds’ with similar profiles should share similar properties” is not always true and it depends on the type of -omics data used. To reveal the MoAs for the compounds in the clusters, they applied an interpretable ML algorithm, which mapped each type of the molecular data to a network of molecular interactions. In conclusion, they identified and experimentally validated Huntington’s Disease-relevant MoAs. Therefore, the value of an approach that combines multi-omics with an interpretable ML method to determine previously unknown MoAs is highlighted, even in the absence of a comparable reference.

1.4.1.2. Group Factor Analysis

The increasing need in MoA studies to explore multiple biological layers in parallel spanning the genome, transcriptome, metabolome, proteome and cell image - space has paved the way for the development of methodologies that can perform integrative analyses[135]. An example of such approaches is Group Factor Analysis GFA), which is a dimension reduction technique aiming to explain correlations in a set of data and relate variables to each other[136].

GFA is a method that can search for relationships between different types of data such as chemical descriptors and biological processes[137]. GFA captures relationships (statistical dependencies) by explaining a set of data sets (‘views’) by a reduced (low-dimensional) representation called factors or components[136]. An implementation of GFA developed specifically for factor analysis of multiple types of biological data is Multi-omics Factor Analysis (MOFA), which is proposed as an improvement of previous factor analysis methodologies by enabling analysis of sparse datasets, computational scalability to larger datasets and non-Gaussian data modalities, such as binary readouts[135].

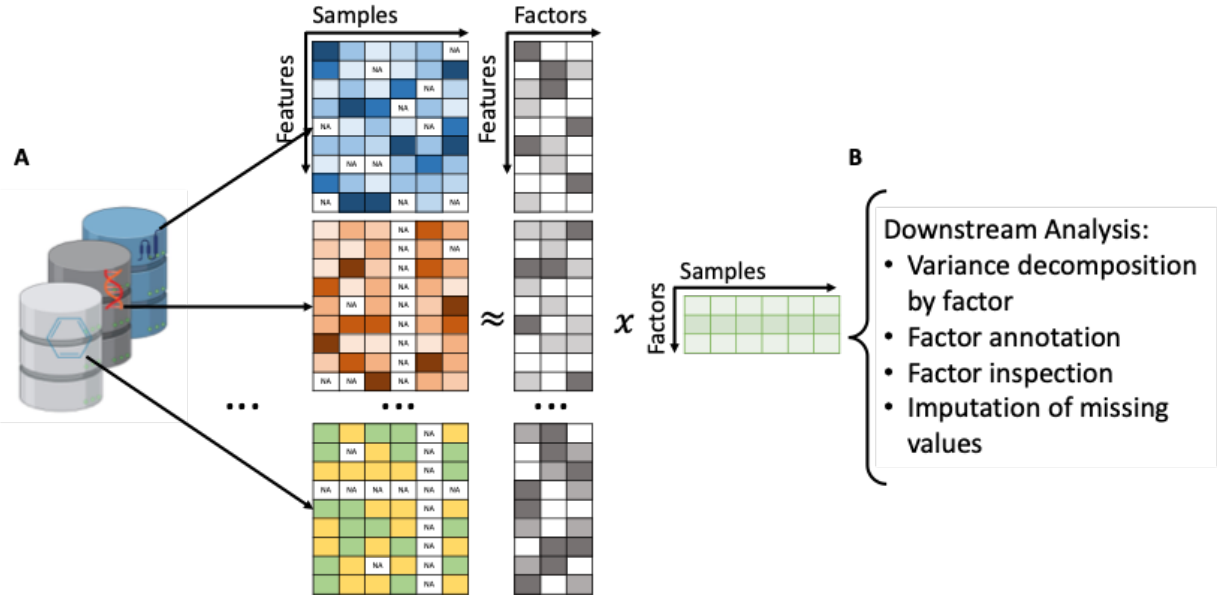


Figure 1.7: a) Demonstration of model overview. MOFA takes a number of data matrices as input from different data modalities and decomposes these matrices into a matrix of factors for each sample and weight matrices, one for each data modality. b) Downstream analysis of MOFA including variance decomposition, assessing proportion of variance explained by each factor in each data modality, inspection of factors and imputation of missing values (adapted from Argelaguet et al.[135]).

MOFA, given a set of data modalities, infers interpretable low-dimensional factors (Figure 1.7a), using group factor analysis. These factors or components capture the major sources of variation across the data and hence enable the identification of continuous gradients or discrete subgroups within the samples. In addition, MOFA can explore to what extent each factor is unique to a single data modality or is manifested in multiple modalities, revealing shared axes of variation between different omics layers. Once the MOFA model is trained the option for downstream analysis (Figure 1.7b) includes visualisation, clustering and classification of samples in factor space.

Group factor analysis methods offer the advantage to integrate multiple data types which enables a data-driven, systems-level analysis of compound MoA, but there are some limitations associated with such methods. Key challenges are the requirement of multiple parameters to be determined, computationally demanding cross validation, manual parameter tuning, and prior information may be required for interpretation of results, such as annotations[138]. In addition, the factors learned from factor analysis can often be difficult to interpret, but methods such as MOFA overcome this limitation

through automated annotation of factors using enrichment analysis, and identification of outlier samples.

Group factor analysis with -omics data has been applied to the elucidation of compound MoA using both transcriptomics data and 3D compound structural descriptors[139]. In this study, the authors identified 11 components that linked 3D compound descriptors with specific gene expression responses, which is highly informative for understanding how compound-structure affects biological response. Through using this method, they were also able to find previously unknown and unexplored shared mechanisms of action, for example through the identification of an 'HSP90 inhibition' factor which contained a prostaglandin analogue (PGJ2) and HSP90 inhibitors (such as geldanamycin), which were found to share similar responses on the gene expression level despite being dissimilar on the compound structure level. In addition, they found that this methodology performed better than the connectivity mapping approach for finding novel shared mechanisms of action between compounds. This methodology is thus particularly useful for drug repurposing. Moreover, MOFA+ has been applied by Rivello et al.[140] into the MoA understanding of Ibrutinib[140]. The scope of their analysis was to map changes in (phospho)protein levels and the associated gene expression profiles with high-throughput single-cell multi-omics profiles and further understand compounds' MoA. They presented QuRIE-seq (Quantification of RNA and Intracellular Epitopes by sequencing) and used multi-factor omics analysis (MOFA+) to map signal transduction over multiple timescales. Results indicated that QuRIE-seq can trace the activation of the B-cell receptor pathway at the minute and hour timescale and provided insights into the mechanism of action of an inhibitory drug, Ibrutinib.

1.4.2. Supervised Machine Learning

Supervised machine learning (ML) methods use labelled data to optimise a function which connects features (e.g., gene expression or compound structure descriptors) to an endpoint (e.g., the activity of a compound at a particular target). There are numerous supervised ML methodologies, which are applied in various stages of the drug discovery pipeline and can improve discovery and decision making for research questions when data is available[1]. From the perspective of understanding compound

MoA, supervised ML has extensively been used in target prediction of primary drug targets (using bioactivity data as the endpoint modelled) and of potential off-target interactions. The knowledge of off-target interactions can inform at an early stage about preclinical and toxic events and consequently potentially minimise the risk of failure in the later drug discovery stages, after which significant money and time has been spent[141].

After the selection of appropriate compound features, supervised ML is carried out by training a model (fitting a function linking the descriptors to the endpoint) and then testing it on a held-out test set to understand how well the model performs with new 'unseen' data, with an optional validation set used to optimise various hyperparameters of the models. Cross-validation (CV) is a useful strategy for smaller data sets, as it splits the data into 'k' folds (where k is the number of folds defined *a priori*) which are subsequently split into multiple training and test sets. There are various methods to split the data into k-folds; for example, a stratified split is used to preserve the percentage of samples for each class in each fold or a group-based split is applied to group compounds based on a property/characteristic (e.g., chemical scaffold) and compounds with the same characteristic will either be present in the train or test set in each fold. It must be kept in mind, however, that different types of split strategies give very different results. For example, in a comparative study between different CV methodologies, the scaffold (group)- based CV was found to be pessimistic, the random selection of compounds in train and test set was overoptimistic and the time series split in addition to random selection was suggested as a most realistic CV approach[142].

1.4.2.1. Decision Trees

There is a variety of algorithms that can be used to train models. One fundamental type of algorithms are the Decision Trees (DTs) that can be used for both regression and classification tasks and hence are usually described as Classification And Regression Trees (CART) and they work by mapping observations to target values[143]. They can predict an endpoint (i.e., a molecular property) based on a set of input variables (descriptors/features)[144]. The input data are in the form of $(x, y) = [(x_1, x_2, \dots, x_m), y]$, where m is the number of features and y represents the target

value. In a DT, there are three types of nodes: a) a root node, b) the internal nodes, and c) the leaf nodes as shown in Figure 1.8. Leaf nodes are also known as terminal nodes. A schematic example of how a DT works is shown in Figure 1.8 and illustrates a classification problem and thus the DT classifies the compounds on a target property y_1 or y_2. The classification of the test compounds is based on the leaf/terminal node that they reach after going through a series of questions. For example, according to the DT shown in Figure 1.8, a test compound will be assigned with the y_1 if it displays a certain condition for molecular descriptor A. If it does not fulfil that condition, then the molecular Descriptor B is examined. If the molecular descriptor B has a value less than 1, then the test compound will be assigned with the target property y_1 or if it has a value greater or equal to 1, then the test compound will be assigned with the target property y_2. A DT works by systematically subdividing the information within a training data (in the root and internal nodes) based on rules and there are various algorithms to define these rules[145]. One of them is the recursive binary splitting. According to that algorithm, different split points are tried and evaluated with a cost function (or sometimes called objective function). The cost function that is used for regression models is expressing the sum squared residuals and is the following:

$$\sum_{i=1}^{n}(y_i - prediction_i)^2 \qquad \text{(Equation 1.4)},$$

where i is the number of compounds and y the experimental value

The output of DT represents the assignment of y value of each leaf for the test set compounds. However, this procedure provides a greedy approach because at each step a split point is defined, which might be good for that specific step but not for the overall of the DT. This limitation of the DTs can be overcome with the use of ensembles DTs like Random Forest (RF).

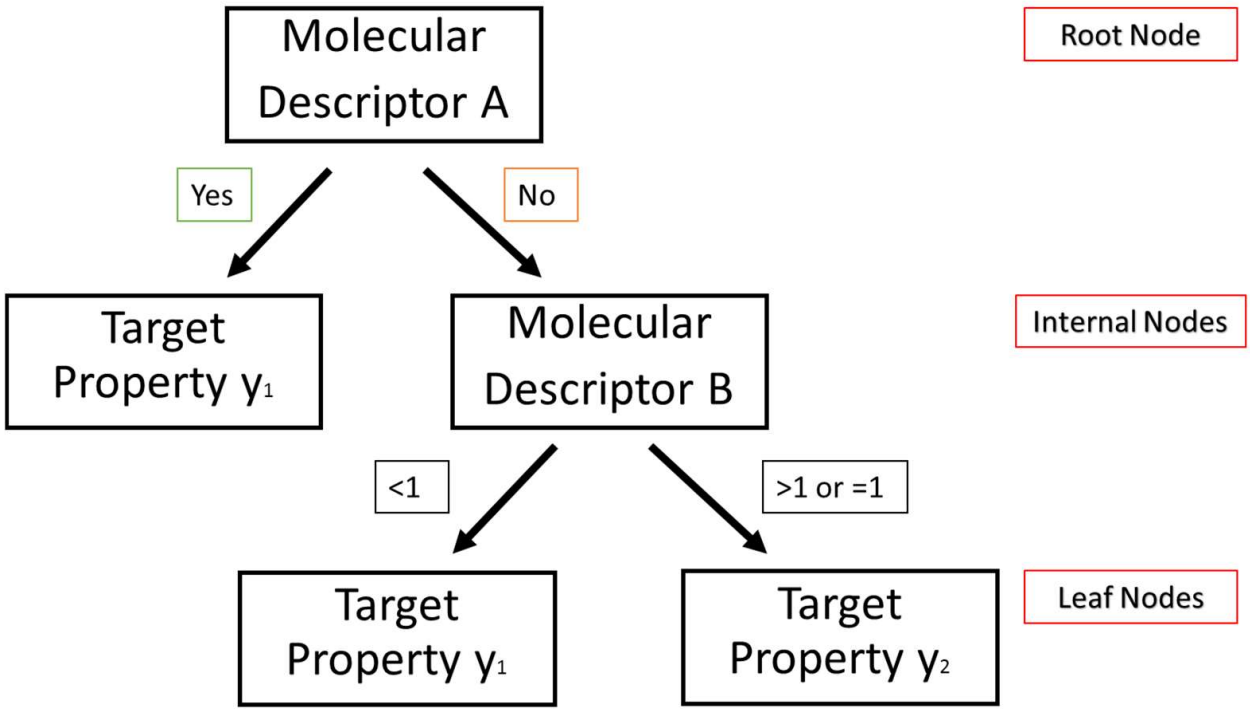


Figure 1.8: Schematic representation of a Decision Tree in a classification task (adapted from Dehmer et al.[145]).

1.4.2.2. Random Forest

RF is based on an ensemble of DTs, which are built by training data on multiple features[146,147]. Ensemble is the process that merges the predictions from multiple predictive algorithms to make a more accurate prediction compared to each individual prediction, as it benefits from the "wisdom of crowds" effect. RF is an improvement of the DTs because the learning algorithm is limited to a random sample compared to DTs, which are searching all the data to identify the ideal split point based on the minimisation of the sum of the squared residuals. The data are partitioned into progressively increasing homogeneous group through the tree. To do this, the algorithm is searching for the best split, which is a combination of a feature and a threshold that will result in the "optimal" separation between the two classes of compounds. To define this optimal split the Gini impurity (G) is used that provides an indication of how "pure" the leaf nodes are (i.e., how mixed are the training data assigned to each node) and is calculated with the following equation:

$$G = 1 - \left(P_{n,A}^2 + P_{n,B}^2\right) \qquad \text{(Equation 1.5)},$$

where $P_{n,A}$ and $P_{n,B}$ are the fractions of instances of classes A and B within the group in the node n (i.e. class probabilities).

The algorithm iterates over all the available features and the possible thresholds and for each threshold the training data is divided into right and left groups. RF searches for the splitting threshold that results in the minimal combined impurity of the two groups as shown in the Equation 1.6.

$$G_{right}\ x\ f_{right} + G_{left}\ x\ f_{left} \quad \text{(Equation 1.6)},$$

where G_{right} and G_{left} are the Gini impurities of the two groups and f_{right} and f_{left} are the fraction of objects in each group so that the sum of the fractions is equal to one. This process is performed for each node of each tree and across all trees. As a result, each terminal node of the DTs is comprised by molecules, which exhibit a similar value of the molecular property evaluated[146]. The randomness of RF is based on two properties; 1) that it trains the different DTs on randomly selected subsets of the full training data, and 2) that is using random subsets of the features in each node of each DT. In this way, the correlation between the different DTs is reduced thus resulting in trees with different conditions in their nodes and different overall structures. Since the prediction of a query compound is performed through a majority vote across the trees, the fraction of the trees that voted for the predicted class serve as a measure of certainty of the resulting prediction. RF has also been used as the basis of a novel algorithm called Probabilistic Random Forest, which considers uncertainty in the data when training a model and it is reviewed, evaluated and compared to classic RF in chapter 2 of this thesis.

1.4.2.3. Support Vector Classifier

Another widely used algorithm in bioinformatics is the Support Vector Machines (SVM), which is an algorithm developed by Vapnik and co-workers[148]. It can be used for both classification and regression problems and when it is used for regression and classification models is referred as Support Vector Regression (SVR) and Support Vector Classifier (SVC) respectively. It is an algorithm extensively used to predict properties like hERG blockade[149], toxicity related properties[146], protein inhibition[150] etc. For example, for a two-class classifier in a 2D space, where the data are linearly separable, the SVM algorithm aims to find the maximum margin hyperplane that divides the data in a way that all the data the belong to a class e.g. 1 lie on the opposite site from those that belong to class e.g. 0[151]. This hyperplane is also referred as

separate hyperplane and the margin is the distance between the separating hyperplane and data samples that are closest to that hyperplane and are called support vectors (Figure 1.9). Therefore, the SVM for a classification problem aims to identify the optimal hyperplane for which the margin of separation between the chemical compounds is maximised[152]. If w is a normal vector to the hyperplane, then the hyperplane equation can be written as:

$$\vec{w}\vec{x} - b = 0 \qquad \text{(Equation 1.7)}$$

and the equations of the two parallel hyperplanes can be written as:

$$\vec{w}\overrightarrow{x_+} - b = 1 \qquad \text{(Equation 1.8)}$$

$$\vec{w}\overrightarrow{x_-} - b = -1 \qquad \text{(Equation 1.9)}$$

As the w vector is perpendicular to the hyperplane it is also perpendicular to the parallel hyperplanes and therefore the vector from the x(-) to x(+) is scalar multiple (r) of the vector w and the following equation can be written:

$$\overrightarrow{x_+} = \overrightarrow{x_-} + r\vec{w} \qquad \text{(Equation 1.10)}$$

By using Equation 1.8 and substitute Equation 1.10 to the x(+), the Equation 1.11 is obtained:

$$(Eq.1.8) \xrightarrow{(Eq.\ 1.10)} \vec{w}(\overrightarrow{x_-} + r\vec{w}) - b = 1 \Rightarrow$$

$$\Rightarrow \vec{w}\overrightarrow{x_-} + r\|w\|^2 - b = 1 \Rightarrow$$

$$\Rightarrow \vec{w}\overrightarrow{x_-} - b + r\|w\|^2 = 1 \Rightarrow$$

$$\Rightarrow -1 + r\|w\|^2 = 1$$

$$\Rightarrow \|w\|^2 = 2 \Rightarrow$$

$$\Rightarrow r = \frac{2}{\|w\|^2} \qquad \text{(Equation 1.11)}$$

The Margin (M) is the half of the distance between x(-) and x(+). Therefore:

$$2M = \|x_+ - x_+\| = \|rw\| \xrightarrow{(Eq.\ 7)}$$

$$\Rightarrow \|rw\| = \frac{2}{\|w\|^2}\|w\| \Rightarrow$$

$$\Rightarrow 2M = \frac{2}{\|w\|} \qquad \text{(Equation 1.12)}$$

Hence, this margin distance should be maximised to identify the optimal hyperplane.

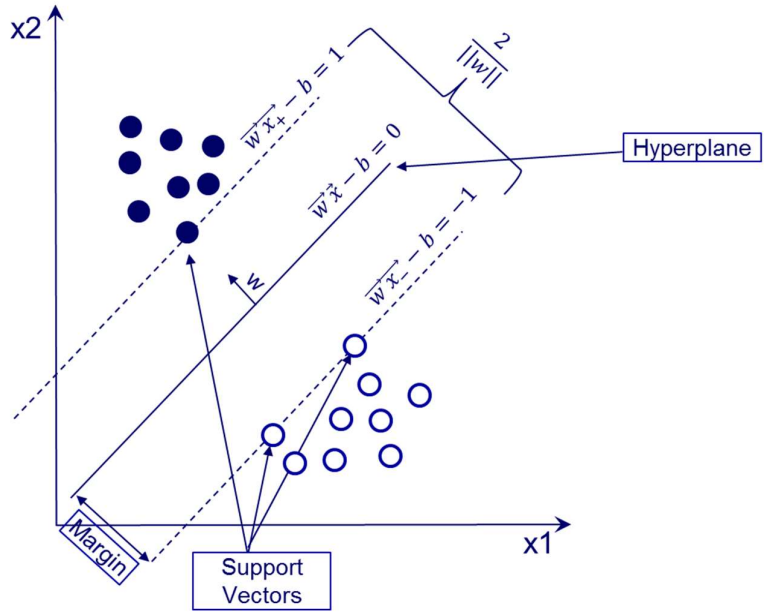


Figure 1.9: Schematic representation of two data classes in a 2D space by the SVM algorithm.

The case outlined above is the simplest case, where the data are linearly separable in a 2D space and can be easily schematically represented. In more complicated cases, where the data 1) are not linearly separable, 2) exist in a higher dimensional space and 3) the goal is the development of a regression model, there are additional strategies to follow. In the non-linearly separable cases, the data are projected in a higher dimension space with the aim to be able to linearly separate them. The kernel trick is used to map the training set data into a higher dimensional space with a mapping function (Φ)[152]. There are various kernels that could be used and one of the most widely used is the radial basis function (rbf) kernel $\left(K(x, x_i)\right)$for two samples/vectors x, x_i of the input space. The rbf kernel can be expressed as the inner product of the projected x, x_i and uses the following equation to map the data in a higher dimension:

$$\left(K(x, x_i)\right) = e^{-\gamma x \sum((x - x_i)^2)} \qquad \text{(Equation 1.13),}$$

where x, x_i are two vectors of the input space and γ is a hyperparameter.

To train the data with the SVM algorithm and the rbf as a kernel, three hyperparameters (ε, γ and C) should be optimised. The ε parameter is affecting the number of support vectors and it can have a value in the range of 0-1. The larger the

ε value is, the lower is the number of support vector[152]. The γ parameter is also taking values in the range of 0-1 and the usual default value is 0.1. If the γ increases, the influence of each data sample is also increased. The C parameter is one of the most important parameters because it can affect both the trained and predicted data[153]. The C value represents a balance between the margin maximisation and the training error minimisation[152]. If the C is too large then the SVM algorithm will produce an overfitted model and if it is too small, insufficient stress is introduced on fitting the training data[152,153].

1.4.2.4. Extreme Gradient Boosting

Another algorithm that has recently gained a lot of attention is the eXtreme Gradient BOOSTing (xgboost), which was designed by Chen Tianqi[154]. As the RF, xgboost is an ensemble DT-based algorithm that uses the gradient boosting framework. The idea behind gradient boosting is to "boost" a single weak model by combining it with several weak models to generate an overall strong model. Therefore, xboost adds a CART steadily and split the features to grow a tree and then one tree is added at a time which is learning a new function[155]. Therefore, xgboost combines the first with the second derivative and the objective function is defined as:

$$Obj = \sum_{i=1}^{m} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \quad \text{(Equation 1.14)}$$

where m is the number of features, K is the number of DTs, k is the model's kth DT, $l(y_i, \hat{y}_i)$ is the training error of sample x_i and $\Omega(f_k)$ is the regular term of the kth classification tree. Following several iterations during the training process, the objective function can be rewritten as:

$$Obj^{(t)} = \sum_{i=1}^{m} l\left[y_i^{(t)}, \widehat{y_i^{(t-1)}} + f_i(x_i)\right] + \sum_{k=1}^{t-1} \Omega(f_k) + \Omega(f_t) \quad \text{(Equation 1.15)}$$

Where $f_i(x_i)$ represents the generated tth classification tree and $\sum_{k=1}^{t-1} \Omega(f_k)$ is the sum of complexity of the first (t-1) classification trees. Then a second order Taylor approximation expansion is used to approximate the original objective function to a function in the Euclidean domain in order to be able to use traditional optimisation techniques and Equation 1.15 is reformed as follows:

$$Obj^{(t)} \approx \sum_{i=1}^{m} l\left[y_i^{(t)}, \widehat{y_i^{(t-1)}} + f_i(x_i) g_i + \frac{1}{2} f_t^2(x_i) h_i\right] + \Omega(f_t) + C \quad \text{(Equation 1.16)}$$

Where g_i and h_i are the first and second derivatives with respect to $y_i^{(t-1)}$ of the loss function $l\left(y_i^{(t)}, \widehat{y_i^{(t-1)}}\right)$, $\Omega(f_t)$ is the penalty term of regularisation (or classification tree complexity) and C is a constant. The classification DT complexity can be written as:

$$\Omega(f_t) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \qquad \text{(Equation 1.17)}$$

Where γ and λ are penalty coefficients, w_j is the weight of the j[th] DT leaf node and T is the number of DT leaf nodes. In the process of segmentation, the weight of each leaf node can be expressed as $w(G_j, H_j)$, G_j and H_j are defined as follows:

$$G_j = \sum_{i \in I_j} g_i \qquad \text{(Equation 1.18)}$$

$$H_j = \sum_{i \in I_j} h_i \qquad \text{(Equation 1.19)}$$

By using Equation 1.18 and 1.19 and substitute equation 1.16, the Equation 1.20 is obtained:

$$Obj^{(t)} \approx \sum_{j=1}^{T} l\left[w_j G_j + \frac{1}{2} w_j^2 (H_j + \lambda)\right] + \gamma T \qquad \text{(Equation 1.20)}$$

Then, by taking the partial derivative of the objective function ($Obj^{(t)}$) with respect to w_j and setting the partial derivative equal to zero to get the optimal weight w_i^*:

$$w_i^* = -\frac{G_j}{H_j + \lambda} \qquad \text{(Equation 1.21)}$$

By using Equation 1.21 and substitute equation 1.20, the Equation 1.22 is obtained:

$$Obj^{(t)} = -\frac{1}{2}\sum_{j=1}^{T} \frac{G_j}{H_j + \lambda} + \gamma T \qquad \text{(Equation 1.22)}$$

Finally, xgboost uses random subspace when selects the optimal split point, which is chosen in order to maximise the gain, which is defined as follows:

$$Gain = \frac{1}{2}\left[\frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{(G_L + G_R)^2}{H_L + H_R + \lambda}\right] - \gamma \qquad \text{(Equation 1.23)}$$

Where the $G_L$and H_L are the gradient values of the subtree on the left of the split point and the $G_R$and $H_R$for the right split point. Therefore, the optimal structure and split are based on the calculation above (Equation 1.23) and the prediction accuracy of the model should increase by adding a new tree[156].

1.4.2.5. Other algorithms

The supervised ML algorithms described above are usually used in a single task setting and i.e., one model is trained to predict one property. However, target prediction models can learn from each other to improve classification accuracy and an example of such an algorithm is the Bayesian Matrix Factorisation (BMF). This algorithm learns multiple tasks (such as predicting multiple drug targets) simultaneously, and the learning tasks can then benefit from each other. The approach works by factorising a sparse matrix Y (N compounds times M targets) containing compound bioactivities to a lower-dimensional representation in latent matrices u and v, for compounds and targets respectively[157]. With BMF it is possible to integrate multiple data types by incorporating side information such as transcriptomic or cell image features (this method is used and described in more detail in chapter 3 of this thesis).

More recently, deep learning (DL) methodologies have attracted more attention for their ability to learn representations of data with multiple levels of abstraction and also their good performance[158]. DL methods are a type of Artificial Neural Network (ANN) with multiple hidden layers in combination with more sophisticated training parameters, which aim to emulate the complex neuronal system (and the process of learning) in the human brain. Specifically, Deep Neural Networks (DNNs) refer to ANNs with many hidden layers, and Convolutional Neural Networks (CNNs) are ANNs which have a convolution layer and a pooling layer (and have shown to be beneficial for processing image data). CNNs in particular can also be used to automatically extract features from cell morphology data[50] for use in further modelling or unsupervised ML approaches such as clustering.

1.4.2.6. Applications of Supervised Machine Learning approaches

The choice of which method to use for bioactivity prediction is not entirely clear and is hence still an area of active research. Different methods have been compared for their ability to predict compound targets, in particular the performance of approaches such as RF and SVM have been compared to NNs. Mayr et al.[159] published a benchmarking study using bioactivity data from ChEMBL which found that deep learning methods outperformed other methods (RF and SVM, as well as k-Nearest Neighbour and Naive

Bayes predictors), and are close to the accuracy of *in vitro* wet lab experiments, based on the AUROC (area under receiver-operating curve, true positive rate/false positive rate) metric[159]. In response to this publication, Robinson et al.[160] performed the study again, this time questioning the usefulness of the AUROC metric for bioactivity prediction and thus also assessing the area under precision-recall curves (AUPRC), which is useful when using imbalanced datasets (i.e., many inactive compounds to a handful of actives, commonly seen in bioactivity data). This study concluded that SVM in fact performs comparably with deep learning methods, in terms of the AUPRC[160]. This highlights the fact that model evaluation is often difficult and has been reviewed previously, with the conclusion that evaluating a model is virtually practically impossible and thus comparing models is a trivial task[161]. How well a method appears to perform depends heavily on the endpoint being modelled, the data going into the model and the evaluation metric being considered. In fact, other important model characteristics such as the applicability domain (where the model works with high reliability and where it doesn't, for example in terms of areas of new chemical space, e.g., Reliability Density Neighbourhoods[162]) and prediction uncertainty (Venn-Abers[163], conformal prediction[164]) should also be considered, as well as performance-based measures such as accuracy, AUROC and AUPRC, but are often neglected in bioactivity model evaluations despite providing a measure of how confident one can be in *new* predictions (which is the ultimate goal of target prediction, and any supervised ML model).

There are numerous examples of target prediction tools developed with supervised ML algorithms but one that offers a broad selection of models and includes all the available activity data from ChEMBL and also inactivity from PubChem is a command line tool called PIDGIN (Prediction IncluDing Inactivity)[62]. The first version of PIDGIN was published in 2015 and proved that the merging of both active and inactive data extracted form ChEMBL, and PubChem respectively costs the performance of target prediction models compared to previous approaches that used only active data. The models in PIDGIN v1 were trained with the Bernoulli Naïve Bayes algorithm and evaluated by using 5-fold cross validation. The model achieved a mean recall and precision equal to 67.7% and 63.8% for the active compounds and 99.6% and 99.7% for the inactive compounds respectively. The model was also externally validated by

using data from WOMBAT database and achieved an average precision-recall AUC (PR-AUC) and BEDROC scores of 0.56 and 0.85 respectively. The second version of PIDGIN was further improved because it was trained using the Random Forest (RF) algorithm with Platt scaling and incorporating orthologue data in the models[74]. The addition of orthologous bioactivity data in target prediction models proved to provide access to molecules in "new areas" of chemical space. However, the influence of orthologue data on models' predictive ability varies between organism and protein types. Therefore, the decision to use orthologue predictions should be considered on a per target basis (e.g., considering chemical diversity of human and orthologue data) and finally based on the biology or disease interested. The third version of PIDGIN further improved by being able to consider the models' applicability domain when making prediction. The methodology used for the applicability domain estimation is called "Reliability Density Neighbourhood"[162], which takes into account how good the models is in predicting its Nearest Neighbour compounds. Therefore, this ensures that a reliability estimate on the prediction can be assigned and this is particularly useful when shortlisting compounds for experimental validation.

However, as mentioned previously a lot of target prediction tools rely on the 'Molecular Similarity Principle', which is not always valid, such as in the case of Activity Cliffs[165]. These are cases when only small transformations to similar structure compounds result in a large difference in potency and bioactivity profiles[82]. Indeed, it has been shown that only 30% of compounds with high similarity (Daylight Fingerprints Tanimoto coefficient > 0.85) to an active compound are themselves active at the same target[83]. As a concrete example, the two very similar antidiabetic drugs, rosiglitazone and troglitazone, share very similar chemical structure but exhibit very different side effect profiles[31]. Hence, it is important to consider different types of compound representations to characterise chemical structure, where even if one representation fails, hopefully, another will pick up the relevant similarity of two structures for the purpose of predicting and understanding its MoA.

Therefore, the use of novel compound information, beyond chemical structure information (such as image-based data), have recently caught the interest of researchers. For example, multitask bioactivity prediction by using image-based data

has been performed by Hofmarcher et al.[166]. The authors used the largest publicly available image dataset consisting of 30,616 small-molecule treatments from the Cell Painting Assay tested on Human Bone Osteosarcoma Epithelial Cells (U2OS)[104]. Hofmarcher et al.,[166] used this dataset and Convolutional Neural Networks (CNNs) with image-based features and found that 32% of the 209 biological assays were predicted with high performance with an Area Under Curve (AUC) larger than 0.9. This result indicated that cell morphological changes upon compound administration can be an informative source to predict compounds bioactivities better. However, only CNNs trained directly on HTI and CNNs trained on image features were compared, without contrasting methods to results obtained from only chemical structure information, that is widely used in target prediction. Moreover, in another study[167] a small subset (1,484 compounds) of the same image dataset was used to predict MoA labels and protein targets, and the model was compared to those based on gene expression profiles and chemical information. Random Forest-based results indicated that the accuracy of the models varied depending on the structural, gene or image descriptor types used and the predicted MoA label/target. More specifically, the best model was obtained using structural descriptors like in the case of GPCR classes, but the gene expression and Cell Painting data outperformed the structural descriptors for other protein families, such as protein kinases. Therefore, this observation highlights those different types of information can be complementary in MoA characterisation and prediction.

A larger scale study using proprietary image-based data (comprising 500,000 compound treatments) has been performed by Simm et al.[105] in two Jansen drug discovery projects[105], to test i) whether image data could overcome limitations employed by chemical descriptors and ii) if image data can be complementary to the chemistry-based models for the sparse and poorly annotated chemical space. Two multitask prediction methods were used, namely Bayesian Matrix Factorization (BMF) Macau and Deep Neural Networks (DNNs). Both methods proved to successfully predict bioactivity using image-based data, performing with an overall AUC-ROC of 0.65 and 0.67 across 535 assays for BMF Macau and DNN respectively. Image-based features were next applied to two discovery projects during virtual screening, increasing the base hit-rate from 50- to 250-fold over that of the chemical structure-

based models. Therefore, image-based data proved to be a rich source of information that can be used to predict the result of biological assays, and hence also for MoA elucidation.

As mentioned previously, an interesting approach is not only to use one type of compound information but also to try to combine then. For example, transcriptomics data can be effectively integrated with cell image data. Changes in cell morphology and gene expression both reflect changes in activity in effector proteins following a perturbation in signalling, where it is not known in detail how these processes interact. Nassiri and McCall developed a pipeline for linking the two types of data together and integrating them for MoA understanding[52]. They utilised the LINCS gene expression dataset as well as the Broad cell morphological image collection to extract a set of 9,515 drugs and small compounds with data on both levels, their 'reference database'. They used the reference database to identify compounds with similar gene expression changes, followed by 'cell morphology enrichment analysis', which involves the identification of significant associations between alterations in cell morphology and gene expression. Least absolute shrinkage and selection operator (LASSO) was used to assess the association between each image-based feature and the landmark genes (i.e., each image feature is modelled as a sparse function of the 978 landmark genes). The enrichment and modelling methods produced a set of genes (with similar expression patterns) associated with each image-based feature. They applied this pipeline to three compounds; Nomilin, Zardaverine and Hydrocotarnine, and were able to better understand the regulatory mechanisms linking the changes on the gene expression and cell morphology levels induced by the compound by performing pathway enrichment with the query-specific cell morphological gene sets. This study revealed a novel interdependence between gene expression and cell morphology and proposed a method to interpret this in terms of compound mechanism of action through the integration of data and methods.

1.4.3. Pathway Enrichment

Pathway enrichment methods require -omics data e.g., transcriptomics, phosphoproteomics or proteomics, and pathway annotations, resulting in a list of significance scores representing the association of the expression data with each

pathway interrogated. In this way, significantly enriched pathways can be related to a compound's mechanism of action in terms of the biological processes and cascades the compound is hypothesised to perturb. The most valuable aspect of pathway enrichment analysis is that they allow large lists of genes or proteins with no biological context (e.g. from transcriptomic, proteomic or phosphoproteomics experiments) to be reduced down to a smaller number of processes, which are inherently more interpretable than gene lists[168], and this biological understanding can help to rationalise the phenotypic finding in question.

The hypergeometric test is considered to be the simplest approach to perform pathway analysis and it works by quantifying the overlap between a set of differentially expressed genes (or other features) detected in the high-throughput data and a background set of genes - also termed ORA or overrepresentation analysis[169]. The background genes are usually the full set of measured genes or the whole human genome. The null hypothesis of this test is that the genes of a pathway are not enriched in the differentially expressed genes. This method provides the advantage of being simple and computationally inexpensive, but it can be biased from the arbitrary cut-off used to define the differentially expressed genes[170], usually a p-value cut-off of 0.05 and absolute log2(fold-change) of between 1-2.

GSEA (Gene Set Enrichment Analysis) on the other hand is a functional class scoring (FCS) method with the underlying hypothesis that the genes that are involved in a similar biological process or pathway (grouped into gene sets) are co-ordinately regulated. Previous benchmarking of FCS methods found that GSEA is a powerful method which is able to detect relevant signalling pathways with a high positive rate[7]. Unlike ORA, this method does not require a defined set of differentially expressed genes, on the contrary it uses some comparison metric for all measured genes. Genes are ranked according to a metric (e.g., differential gene expression significance), and then GSEA aims to identify whether the genes from a set/pathway occur in the top or bottom of the ranked gene list. The null hypothesis of GSEA is that no genes in the expression profile are associated with an observation and occur randomly. A Kolmogorov–Smirnov test is then applied to evaluate the statistical significance of the enrichment. The advantage of GSEA is that it does not require an arbitrary cut-off to

define differentially expressed genes and it provides a more in-depth characterization of pathways representative in the data compared with the hyper-geometric test[171]. However, GSEA and ORA are not able to take into account the topology of the underlying pathways (i.e., the interconnections of genes or other biomolecules within the pathways). Therefore topology-based pathway enrichment analysis methods were developed as the latest generation of pathway enrichment methods[53]. Topology-based methods are similar to FCS methods except they incorporate pathway topology metrics such as number of reactions and position of gene and compute a "pathway impact factor"[172]. A limitation of this approach is that true pathway topology is dependent on cellular context and organism, and such differences are usually not represented in pathway databases.

Pathway analysis has been used with *in-silico* target predictions to elucidate the MoA of compounds and provide further biological insights. In one study, predicted bioactivity profiles (obtained with target prediction) were annotated with pathways and a calculation of enrichment factors revealed targets and pathways that are more likely to be implicated with the studied phenotype, which was the pigmentation phenotype of Xenopus laevis tadpoles based on a genetic screen performed on Xenopus laevis embryos readouts[173]. In more detail, a set of 1,364 compounds was extracted from the National Cancer Institute (NCI) diversity set and a target prediction was applied on these compounds with a Laplacian-modified Naïve Bayes classifier. Moreover, all predicted targets were annotated with pathways extracted from the KEGG database and an enrichment calculation against a background distribution was performed to identify enriched targets and pathways. From the total of the 1,364 compounds, 45 compounds were causing the pigmentation phenotype and these compounds were associated with 236 predicted targets. Out of the 236 predicted targets, 33 were found to be enriched. The top 10 enriched targets implicated in pigmentation and were confirmed by literature analysis. For example, the top enriched target was the Platelet-Derived Growth Factor Receptor alpha (PDGFRα), which according to literature findings is important during the Xenopus laevis developmental effects such as alterations in the pigmentation.

A similar methodology was applied in a different concept and was applied to cellular cytotoxic readouts[8]. In more detail, the pathway annotations improve the MoA information gained from an *in-silico* target prediction by providing a better biological interpretation of the results and additionally providing a better mapping of targets onto pathways. In this study two different datasets were used; a cytotoxicity dataset of 1094 compounds and a smaller apoptotic dataset of 10 compounds and protein targets were also predicted for these compounds. The cytotoxic dataset was extracted from PubChem by selecting 186 bioassays describing molecules that had proved cytotoxic to HeLa cells in cell-based assays and the apoptotic dataset was extracted from the Prestwick Chemical Library based on their activity in killing embryotic mouse stem cells detected by a calorimetric assay based on a metabolic activity performed with the Cell Proliferation Kit II. Target predictions for both datasets (cytotoxic and apoptotic) were annotated with pathways, which were extracted from KEGG, GO biological processes, and GO Slim biological processes. The annotated pathways were further subjected to enrichment calculation. Results of the pathway enrichment for the cytotoxic compounds identified pathways important in cancer development and or immune response and pathways related to DNA and cell cycle. For the smaller apoptotic dataset, pathway enrichment revealed only a small number of enriched pathways. A major disadvantage of this methodology and the apoptotic dataset is that on small datasets this methodology needs to be performed differently and an analysis of absolute targets and pathways seems to be more appropriate. Overall, both studies discussed above showed that target prediction with pathway annotations and enrichment calculation can add meaningful biological insights in MoA and target prediction understanding.

1.5. New Data Modalities and novel MoA

Pharmaceutical companies and academia are focusing on expanding their classical toolbox for drug discovery beyond traditional small molecule therapeutics[174]. New chemical modalities have emerged such as RNA therapeutics, protein degraders, antibody drug conjugates, and gene therapy, which are continuously maturing and demonstrated clinical success and are now considered early in drug discovery process. One promising example of such a new data modality is the protein degraders called PROTelolysis Targeting Chimeras (PROTACs). Despite all the promising

aspects that PTOTACs introduce in the drug discovery process, it is also important to early consider any safety risks related to PROTACs MoA and which will be reviewed in this part of the thesis.

1.5.1. Overview of PROTACs MoA

It is important to emphasise that MoA and understanding of biological effects has extensively been investigated for small-molecule compounds, which are the vast majority of marketed drugs and are based on the concept of occupancy-driven pharmacology as the MoA, in which the protein function is modulated via temporary inhibition[175]. However, there are unavoidable problems associated with small-molecule drug discovery such as drug-resistance and the limited number of drug targets, which consist of 20-25% of all the protein targets that are being studied[176]. Therefore, new data modalities, which operate using alternative MoAs are needed in order to expand to the 'undruggable proteome', which is biologically attractive and include various types of proteins e.g. those without enzymatic function such as transcription factors, proteins functioning via Protein-Protein Interactions (PPIs), scaffolding proteins and others[175,177]. A promising example of such a new data modality is this of the PROteolysis TArgeting Chimeras (PROTACs), which modulate a protein of interest by degrading it[178,179]. In more detail, PROTACs are heterobifunctional molecules which connect a Protein of Interest (POI) ligand to an E3 ubiquitin ligase (E3) recruiting ligand with an optimal linker[180]. PROTACs technology works through active recruitment of an E3 ligase to tag proteins for degradation as shown in Figure 1.10. Bifunctional PROTACs molecules bind to the POI with one end, while the other end binds to an E3 ligase to form a ternary complex. The recruited E3 ligase then mediates the transfer of ubiquitin from an E2 enzyme to the POI. The ternary complex dissociates, the ubiquitylated POI is removed by the proteasome[180].

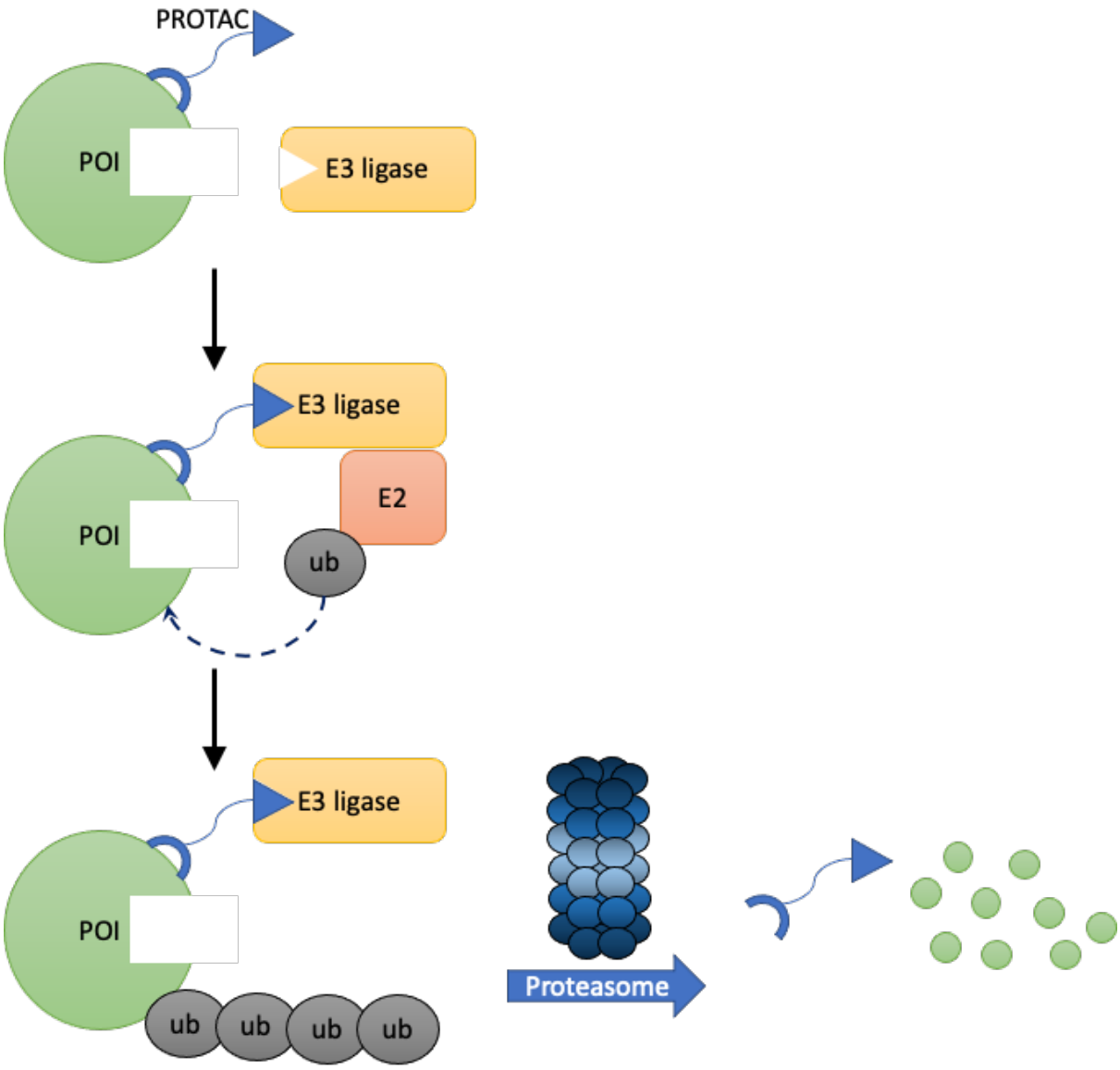


Figure 1.10: Mechanism of PROTACs' induced protein degradation technology. (adapted from Sun et al.[176])

PROTACs are being explored as a therapeutic modality for the past 20 years with the first PROTAC first described by the Crews and Deshaies laboratories in 2001, when the SCFb-TRCP E3 was recruited to induce the degradation of methionine aminopeptidase-2 (MetAp-2)[181]. Since then, significant progress has been made and some key milestones are shown in Figure 1.11. This first milestone was followed by the finding that PROTACs can induce degradation of Androgen (AR) and Estrogen receptor (ER) in 2003 and microinjection of AR- and ER- targeting PROTACs showed that they can function in an intact cell[182]. The next step in 2004 was the design of a PROTAC, which used the HIF-1α peptide-based PROTAC and did not require the use

of a microinjections. This was made possible because the peptide was a cell-penetrating peptide that was able to recruit the Von Hippel-Lindau disease tumor-supressor protein (VHL) E3 in intact cells[183]. Efforts have been made to design optimal peptides and in 2007 a shorter peptide fragment of HIF-1a was incorporated in a PROTAC targeting aryl hydrocarbon receptor nuclear translocator (ARNT)[184]. However, the PROTACs that have been described until then and are usually described as '1st Generation PROTACs' were limited by the fact that they were active only in the low-micromolecular range, showed poor permeability and low cellular activity. These issues limited PROTAC technology for further development of novel therapeutics[175].

The issues described above have been limited by the significant development of PROTACs and the design of the first all-small-molecule-based PROTACs in 2008[185]. In addition, later in 2012 high affinity peptidomimetic ligands were designed for the VHL E3[186–188], which is one of the two most popular E3 ligases being recruited by PROTACs to induce ubiquitination and subsequent proteasomal degradation of a target protein[189]. In 2012 further attempts were made to optimise E3 ligands and improved VHL ligands with improved physical-chemical properties but similar affinities towards VHL were reported by the Ciulli laboratory[190–193]. During the same time period, the E3 cereblon (CRBN) was identified as a molecular target of the immunomodulatory drugs (IMiDs) namely; thalidomide, pomalidomide and lenalidomide[194–200], and which is the other one of the two most popular E3 ligases.

The next milestone was observed in 2013 because of the first evidence that PROTACs function *in-vivo* and can promote inhibition of tumour growth in murine models. This was followed by another important step in PROTACs development in 2015, when the Crews laboratory in collaboration with GSK developed a Receptor-Interacting serine/threonine-protein kinase 2 (RIPK2)-targeting PROTAC, which was able to selectively induce RIPK2 degradation with low nanomolar cellular potency. They have also performed further mechanistic interpretation using an *in-vitro* ubiquitination assay that demonstrated its catalytic nature. In more detail, a negative control PROTAC which incorporated a stereoisomer of the VHL ligand that was unable to recruit VHL, was subsequently unable to reduce RIPK2 levels thus demonstrating the E3 ligase-dependent mechanism for their PROTAC[201].

More recently, in 2016 variations of PROTACs have been emerged. For example, In-cell CLIck- formed Proteolysis TArgeting Chimeras (CLIPTACs) developed by Astex Pharmaceuticals are PROTACs formed intracellularly by biocompatible reactions such as an inverse electron demand Diels–Alder reaction[202]. The treatment of cells sequentially with cell permeable tetrazine substituted thalidomide derivatives and trans-cyclo octene substituted POI ligand results in the formation of active PROTACs that successfully induced POI degradation (BRD4 and the extracellular signal–regulated kinases ERK1/2). Finally, the most recent and important announcements was made by Arvinas Therapeutics with the first-in-man trial of an anticancer PROTAC, which shows how in ~20 years the PROTACs managed to get into trials and actually make an impact in patients[203,204].

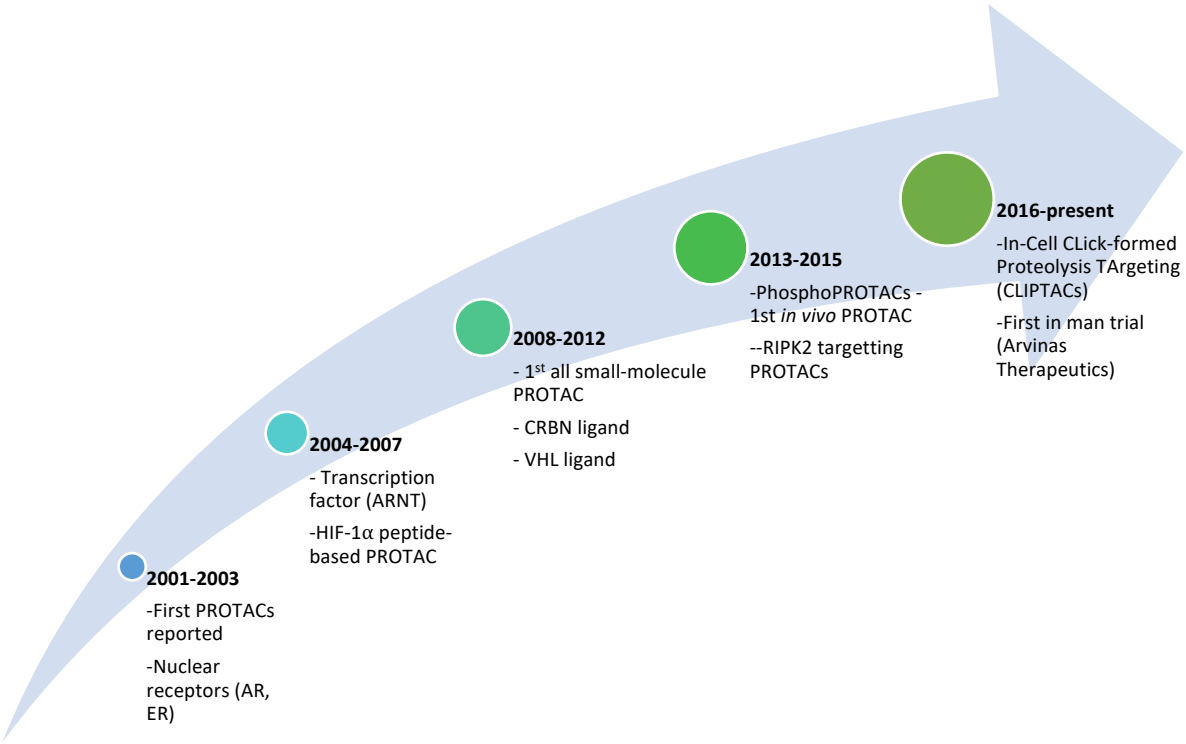


Figure 1.11: Schematic representation of the PROTACs timeline, which describes the evolution of PROTACs since 2001 until present (adapted from Pettersson et al.[175])

1.5.2. Safety concerns related to PROTACs' MoA

The initiation of Phase I clinical trials in 2019 for two PROTACs, one targeting the AR in patients with metastatic castration resistant prostate cancer and the other targeting the ER for locally advanced or metastatic ER positive/HER2 negative breast cancer,

is a significant basis for further PROTACs development. The authorisation of these trials by the US FDA highlights a non-clinical safety profile of these two PROTACs that is aligned with the safety guidelines and standards which are required to enable dosing in cancer patients[205]. However, there are several safety concerns regarding this modality that are already known and are based on the particular way of their MoA. These are: a) prolonged target protein degradation, b) off-target protein degradation, c) disruption of cellular proteostasis and d) implications of the "hook effect"[205].

1.5.2.1. Prolonged target protein degradation

The prolonged target protein degradation is a phenomenon that could happen depending on the pharmacokinetic (PK) profile of the PROTAC and the resynthesis rate of the target protein within the tissue. In general, PROTACs tend to display greater efficacy via the pronounced and extended target degradation compared to reversible small molecule inhibitors. Therefore, their MoA resembles that of the irreversible inhibitors, which have been successfully resulted in approved drugs in different therapeutic areas and prolonged target inhibition can be tolerated, although, this is target and context dependent. Therefore, it is important to optimise PROTACs' dose and schedule to achieve an efficacious level of target degradation in the disease tissue in order to avoid on-target toxicity related to the dose[205].

1.5.2.2. Off-target protein degradation

There are reports that PROTACs are not entirely selective and consequently are able to degrade proteins other than their primary target, which can happen via different ways. The first way is the degradation of a protein that is not directly bound by a PROTAC and can happen because of an effect called "bystander degradation". An example of such degradation is this described by Hsu et al.[206], where an EED-targeting-PROTACs degraded EED and its associated proteins EZH2 and SUZ12 in the PRC2 complex[206]. In addition, off-target effects can also arise from the binary engagement of the target binding ligand of the PROTAC to other proteins in the same way as the POI. The second way is the result of neo-morphic interactions with neo-substrates which become ubiquitinated by the E3 ligase and consequently degraded as shown in Figure 1.12a. Moreover, the third way is related to off-target activities with potential safety issues for PROTACs that contain an immunomodulatory drug (IMiDs)

as the ligand for the CRBN ligase. The IMiDs that are usually used are the thalidomide, pomalidomide and lenalidomide and recent studies have shown that they are recruiting neo-substrates to CRBN and thus these neo-substrates get degraded (Figure 1.12b). These neo-substrates are usually transcription factors such as SALL4[205,207,208] and some Ikaros proteins such as IKZF1 and IKZF3, which regulate important cell-fate decisions during haematopoiesis (which might explain the haematotoxicity seen with IMiDs)[209]. Several PROTACs have been developed by incorporating an IMiD or IMiD derivative as a CRBN ligand and therefore to assess any safety issues and this is of high importance because of, published examples that indicated reasons for concern. For example, a BTK-targeting PROTAC that included pomalidomide as the CRBN warhead resulted in the degradation of the IMiD substrates IKZF1, IKZF3, ZNF827, and ZFP91[210]. Similarly in another study, both IKZF1 and IKZF3 were degraded by a CDK6-targeting PROTAC that used lenalidomide as the E3 ligase ligand as well as a HDAC6-targeting PROTAC that incorporated a pomalidomide ligand[211,212]. There were not reported effects of those PROTACs on haematopoiesis, but it would be important to assess this safety risk using in vitro assays and/or incorporating relevant endpoints in a preclinical, investigative toxicology study and potentially find appropriate descriptors to computationally predict them. Similar effects have also been observed with the sulphonamides (anti-cancer agents) such as indisulam[213,214], which recruit neo-substrates to the DCAF15 E3 ligase for degradation in a similar way as the IMiDs. The neo-substrate degraded by sulphonamides is the mRNA spicing factor RBM39 that is the main target related to the antiproliferative effect of the sulphonamides. Novartis has identified the mechanism by which indisulam mediates interaction between RBM39 and DCAF15[213]. In more detail, indisulam forms a tripartite complex with DCAF15 and RBM39, where DCAF15 and RBM39 bind together (mainly through non-polar interactions). The observation that sulphonamides contact both proteins simultaneously, highlights the challenge to use DCAF15 as a novel E3 ligase and potential safety concerns[215].

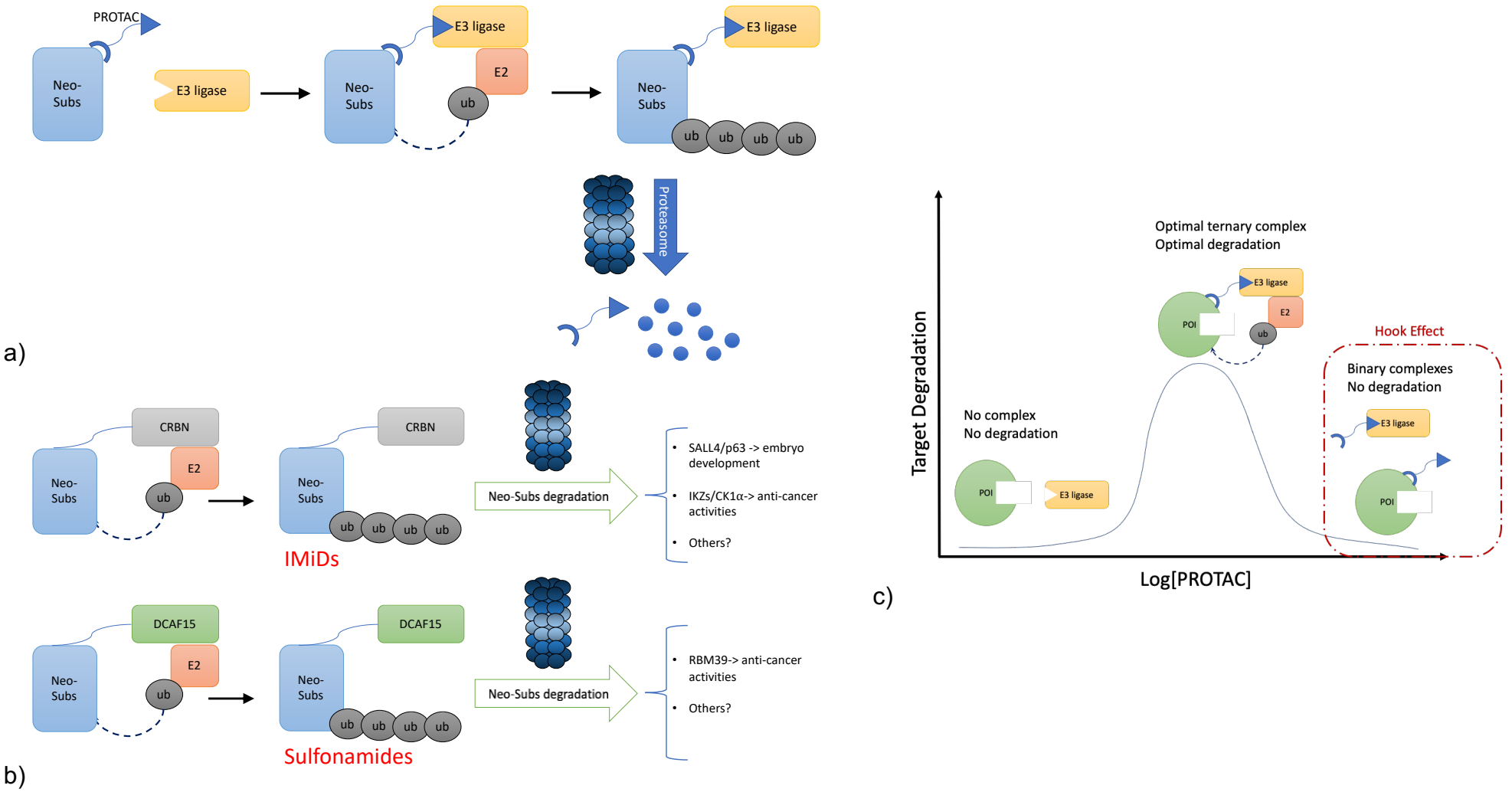


Figure 1.12: Safety challenges inherent to the PROTAC modality. a) PROTAC can degrade other proteins than the primary target via the recruitment of neo-substrates (Neo-subs) to the E3 ligase. b) IMiDs such as thalidomide, pomalidomide and lenalidomide degrade neo- substrates via their recruitment to CRBN. The IMiD substrates are the SALL4 and P63 and Ikaros proteins (IKZs) and CK1α. Anti-cancer sulphonamides degrade neo- substrates via their recruitment to DCAF15. The sulphonamide substrate is RBM39. c)

Hook effect with PROTAC. At high intracellular PROTAC concentration, binary complexes are favoured over ternary complexes, resulting in reduced target degradation. Binary complexes could cause an increase in degradation of neo-substrates as well as the PROTAC acting as a traditional small molecule inhibitor/activator, reflecting the pharmacophore of the POI ligand (adapted from Moreau et al.[205]).

1.5.2.3. Disruption of Cellular Proteostasis

A PROTAC-POI complex is possible to compete with natural substrates for binding to an E3 ligase for ubiquitination and thus for further degradation. This case could result in accumulation of those substrates, which potentially could perturb specific cellular pathways and raise safety concerns. In addition, there is also the risk that PROTACs could increase the cellular protein concentration of ubiquitinated proteins, which could result in changes on cellular homeostasis. This could lead to serious safety concerns because proteasome controls the cellular content of key proteins that regulate aspects such as cell cycle, cell growth, immune homeostasis and metabolic activity[216]. For example proteasome regulates the presence of MHC class I antigens, which play a critical role in the immune response and also dysregulation of proteasome has been associated with neurogenerative and auto-immune diseases[217,218].

1.5.2.4. Implications of the "hook effect"

Another safety concern related to PROTACs' MoA is that PROTACs can saturate binding to the target and to the E3 ligase resulting in the formation of binary complexes instead of the productive ternary complex (usually at high concentrations). This consequently prevents target ubiquitination and degradation, and this situation is referred to as the "hook effect"[219] (Figure 1.12c). This effect could result in two types of adverse consequences, which are: a) increase in off-target degradation activity via the binding of the E3 ligase -PROTAC binary complex to lower affinity targets (other than the POI) and/or b) a pharmacological response driven by the POI-PROTAC complex interaction that is different from the POI degradation. For example, the latter is a concern for PROTACs developed to degrade the AR, which contains mutations within the ligand binding domain[220].

All those safety concerns discussed above explain why PROTACs as a new therapeutic modality is raising multiple concerns on various aspects such as safety, ADME properties, toxicity, and others[205] . There is a need to better understand the way that these molecules interact with the organism. The prediction and/or better understanding of PROTACs' safety profiles related to the issues described above is additionally limited by the lack of descriptors and methodologies for robust safety

profiling. PROTACs belong to a category of compounds which is also referred to as beyond the Rule-of-5 (bRo5) that does not comply with the Lipinski's Rule-of-5 (Ro5). Hence, there is a need for descriptors tailored or 'compatible' with the bRo5 new data modalities[221,222]. This situation results in an increased interest in the development of high throughput data, which could serve as descriptors to better understand PROTACs MoA and safety, and one option could be the use of High Throughput Imaging Assays (HTI). The profiling of PROTACs with a HTI assay for the prediction of a safety issue (mitochondrial toxicity) is discussed in detail in chapter 4 of this thesis.

1.6. Aims of the thesis

The aim of this work is to better understand and predict biological effects of compounds by using and comparing different types of compound data and novel methodologies.

Although it is widely known that bioactivity databases contain an unavoidable error related to the experimental error, this is something usually neglected when *in-silico* target predictions models are trained. Hence, chapter 2 focuses on using and evaluating a novel algorithm called Probabilistic Random Forest, which is a variation of the traditional RF, and which can consider the experimental uncertainty of the bioactivity annotations. This algorithm was developed initially for dealing with noisy astronomical data and the aim was to understand if its application is beneficial and for which cases in bioactivity prediction. Thus, chapter 2 gives insights on the cases where such an algorithm can be an asset when dealing with noisy data.

Most previous studies that are focusing on the prediction of compound-target binding have mainly used chemical structure information, and more recently a few studies are using compounds' high throughout imaging profiles such as those derived from the Cell Painting assay. However, one key question is when to use high throughput imaging profiles, which require more time and cost to be generated compared to computationally calculated chemical structure-based information. Therefore, chapter 3 focuses on the comparison of chemical structure-based information with Cell Painting profiles for compounds' bioactivity predictions across 224 targets. Thus,

chapter 3 provides indications for which targets Cell Painting profiles could be an asset for their bioactivity prediction.

Given the fact that new data modalities such as PROTACs are promising for the future of drug discovery, it is important to find assays to profile them and further use this information to train *in-silico* prediction models to early identify any safety aspects. Therefore, chapter 4 aims to understand whether Cell Painting assay can be used to profile RPOTACs and whether these profiles can be used as a descriptor for training models for mitochondrial toxicity prediction. Hence, chapter 4 outlines the first ML model to predict PROTACs' mitochondrial toxicity using Cell Painting profiles and expands our knowledge for PROTACs' safety profile prediction.

2. Chapter 2: Probabilistic Random Forest improves bioactivity predictions close to the classification threshold by taking into account experimental uncertainty

2.1. Introduction

The majority of research toward small molecule property prediction has predominantly focused on improving the reported accuracy of base algorithms, rather than factoring the experimental error into predictions[223]. Currently, uncertainty estimation as a field is gaining traction due to the application of predictive models toward autonomous decision making within the design-make-test-analyse (DMTA) cycle[224,225]. Various methodologies have been developed and applied in molecular property prediction models to account for the uncertainty in prediction and/or reliability of prediction[226]. Such approaches are the conformal prediction, calibration and Bayesian procedures (shown in appendix Table 7.1**)** and have historically focused on the behavioural characteristics of the base estimator itself (or variants thereof) after initial data processing. Hence, they provide limited consideration toward the true uncertainty in the underlying biological data used to train the algorithm.

Since experimental error influences models' performance, it is important to investigate methods capable of accommodating experimental variability during training the models. This is particularly important for binary classification tasks due to imposing arbitrary cut-off(s) to the activity scale. Such architectures are frequently applied toward biological tasks with poor regression predictivity, as is the case for *in silico* target prediction approaches, where binding probabilities for orphan compounds are calculated at one or more activity thresholds[159,227,228]. Structure activity relationship (SAR) landscapes are highly discontinuous (e.g., presence of activity cliffs) and $IC_{50}/EC_{50}/K_i/K_d$ activities are often heteroscedastic (i.e., the measurement error is unequally distributed across the range of activity values) so regression is not favourable for *in silico* target prediction. The main caveat of binary classification approaches is that they weight minority cases close to the threshold boundary equivalently in distinguishing between activity classes. For example, pXC_{50} activity values of 5.1 or 4.9 are treated equally important in contributing to the opposing activity (e.g., classification threshold of 5), even though experimental error may not afford such

discriminatory accuracy. This is detrimental in practice and therefore it is equally important to evaluate the presence of experimental error in databases and apply methodologies to account for variability in experiments. One potential option to deal with the uncertainty near the classification threshold is the removal of edge cases (i.e., classification marginals), for compounds with activity or property values close to the cut-off value used for classification. This however results in the removal of valuable minority class instances (compounds belonging to the active label) and is likely to hinder the predictivity or applicability of models. For this reason, the removal of "edge cases" of highly imbalanced datasets is not common practice within the field[229] and is considered outside the scope of this work.

Given the studies described in the introduction of this thesis (chapter 1, section 1.3.1.1) regarding the experimental uncertainty of bioactivity data, one can expect a large variation in the range of observed standard deviations between experiments, which should be considered when assimilating a training set dependent on the measurement units and method of aggregation across heterogenous assays. However, there are relatively few previous studies that have framed experimental uncertainty as the natural upper limit of the predictive performance possible, closely monitoring when the maximal performance of a model has been reached[230,231]. For example, an analysis by Brown, Muchmore and Hajduk[230] explored the influence of assay and prediction errors in predictive modelling for drug discovery. The authors calculated the upper performance limit of a model (i.e., correlation between experimental and predicted value), which is likely to be ~80%, given a standard deviation of ~0.3 and the dataset comprised a potency range of only 2 log units. The authors suggested levels of toleration based on the requirements of a particular model application. For example, an upper limit of five standard deviations in prediction errors was suggested for prioritising compounds for HTS, versus an upper limit of one standard deviation for lead optimisation models to ensure a degree of "discovery productivity". Another study took into account the uncertainty in bioactivity data in a systematic analysis of the effect of random experimental errors in the predictive ability of QSAR models. The analysis aimed to evaluate the influence of experimental variability in target prediction models by simulating experimental error on 12 Machine Learning algorithms in bioactivity modelling using 12 diverse data sets (15,840 models

in total) from ChEMBL (version 19)[232]. Noise was artificially defined (which may not reflect real-world situations, where systematic differences between labs etc. exist) by sampling a Gaussian distribution with zero mean and a variance value (defined as a function of the range of bioactivities considered in each data set). Model performance on the test set was used as a proxy to monitor the relative noise sensitivity of these algorithms as function of the level of noise added to the bioactivities from the training set. Overall, Gradient Boosting Machines (GBMs) showed a low tolerance to noisy bioactivities although its performance was comparable to RF, Support Vector Machines (SVM) and Gaussian Process (GP) for low noise levels. The other algorithms showed comparable noise tolerance and a linear decrease of model performance by increasing the level of noise. Therefore, the presence of error in the training data affected the performance of all the algorithms tested and hence should be taken into consideration.

A different approach to account for experimental uncertainty is to explore methodologies that are able to deal with experimental variability. One such method is the Bayesian developed "sum-of-trees" model (BART)[233], where each tree is constrained by a regularization prior to be a weak learner, and fitting and inference are accomplished via an iterative Bayesian backfitting MCMC algorithm that generates samples from a posterior. Effectively, BART is a nonparametric Bayesian regression approach using dimensionally adaptive random basis elements. Motivated by ensemble methods and boosting algorithms in particular, BART is defined by a statistical model using a prior and a likelihood. This approach enables posterior inference including point or interval estimates of the unknown regression function as well as the marginal effects of potential predictors. Although this algorithm presents an interesting comparison, MCMC is slow to perform on larger datasets (as in the case of the many millions of inactive bioactivities held in repositories such as PubChem[67]). Another, more computationally efficient option is the probabilistic random forest (PRF)[234], which is a modification of the long-established Random Forest (RF) algorithm, which can take into account uncertainties in the measurements (i.e., features) as well as in the assigned classes (i.e., activity labels). It is an algorithm recently released for dealing with noisy astronomical data and the scope of this chapter is to use this novel methodology for target prediction[234,235].

In this chapter, the standard deviation of experimental measurements of bioactivity data from the ChEMBL and PubChem repositories is considered in training models by using the Probabilistic Random Forest algorithm. The workflow employed in this chapter can be divided into three main steps (see Figure 2.1). Step 1 is the extraction of bioactivity data from ChEMBL and PubChem databases. Step 2 is the training of models with two different types of algorithms. The first is the classic Random Forest (RF) and the second a modified version of the original RF, namely the Probabilistic Random Forest (PRF), which can take into account uncertainties in assigned classes (i.e., y-labels). The main difference between the two algorithms is that RF uses discrete variables for the activity label (y-label), which is defined by applying a bioactivity threshold in the bioactivity data for each target modelled. On the other hand, PRF algorithm treats the labels as probability distribution functions, rather than deterministic quantities (and are referred to as “ideal y-label”). Multiple PRF models were trained, and different values of noise were injected into bioactivity data. Finally, Step 3 of this work included the comparison between the probabilities returned from RF and PRF algorithms. With this approach, it is presented for the first time an application of probabilistic modelling of activity data for target prediction using a novel algorithm, which is a modification of the well-established RF algorithm.

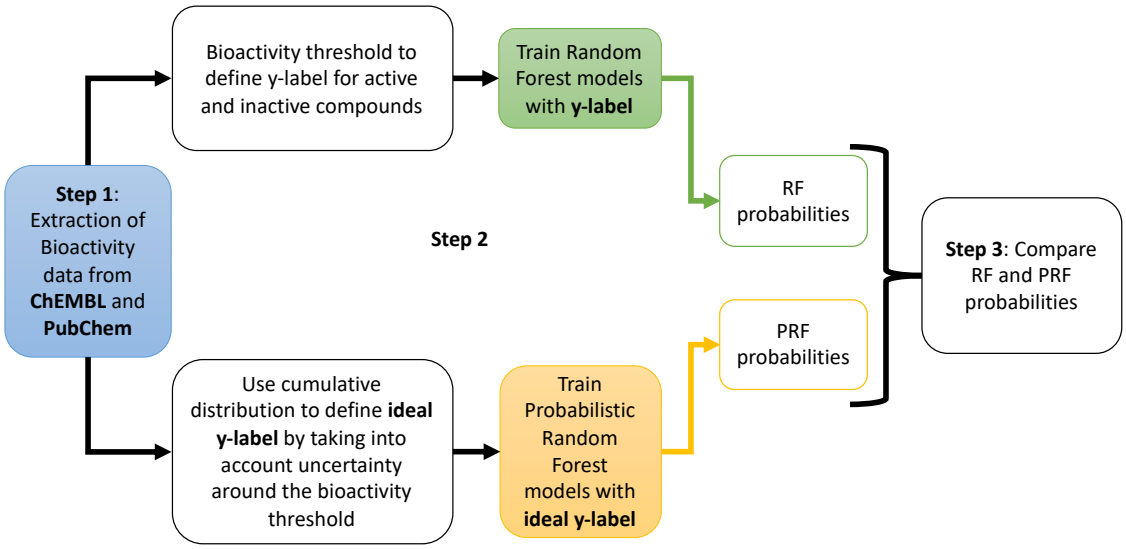


Figure 2.1: Summary of the analysis performed in this work. Random Forest and Probabilistic Random Forest are used to train models and their output probabilities are compared.

2.2. Methods

2.2.1. Bioactivity data set

The ChEMBL (version 27) database[236] was filtered for compounds with a reported pChEMBL (normalized -log[10]) activity value from 'binding' (IC_{50}/EC_{50}/K_i/K_d) human protein assays. Confidence scores of 5 and 8 were employed for the reproducibility comparison when activity values were aggregated at protein complexes or for specific individual proteins, respectively. Compounds were subsequently filtered for a confidence score of 8 for modelling purposes. Targets were also subsequently filtered for greater or equal to 50 active compounds across the activity thresholds for the pChEMBL activity bins 5, 6 and 7 (corresponding to activity values 10, 1 and 0.1 μM) to ensure that only proteins encompassing sufficient chemical space across the activity thresholds were retained for the training set. Models were trained for 557 targets and Figure 7.1 in appendix summarizes the number of active and inactive data points for each model and for which a large variance between the amount of bioactivity data available per target was observed. For example, there was a median of 389, 375, and 386 active compounds per-target for the pChEMBL classification thresholds of 5, 6 and 7, respectively. A median of 1,000 inactive compound datapoints was calculated across targets and thresholds, with a median ratio of 0.4 active compounds to inactive compounds (see appendix Figure 7.1 for details). The dataset for putative inactives per target is available for download as zip files here: https://pidginv3.readthedocs.io/en/latest/install.html.

2.2.2. Compound Standardisation and ECFP calculation

Compound structures were standardized using the IMI eTox molecular structure standardizer (https://github.com/flatkinson/standardiser), with settings to remove salts, waters, solvents, normalize charges, tautomerize (to the most favourable form) and to remove duplicates. RDKit[237] (Version 2019.03.4) was employed to remove structures without carbon, and to retain only compounds with atomic numbers between 21–32, 36–52, and greater than 53, and with a molecular weight between 100 and 1000 Da, to retain small organic molecule chemical space.

RDkit[237] was used to generate ECFP circular Morgan fingerprints (radius=3, 2048 bit length). These fingerprints represent molecular structures using circular atom

neighbourhoods and provide information on the presence or absence of chemical features from fragments within molecules. The ECFP generation process is summarized in the steps described by Rogers et al.[81] The first step in the method initializes the initial integer identifier of each atom (Figure 2.2a). Next, each identifier is updated with the identifier values from the environment of the immediately neighbouring atoms. Once the initial value of each node is specified, the algorithm iterates over neighbouring atom identifiers until a specified diameter is reached (Figure 2.2b). Each iteration captures larger neighbouring atoms and at the end of all steps, duplicate identifies are removed. Finally, the fingerprints are encoded into single integer values using a hashing method. Fingerprints can also be converted to a length of binary values. These fingerprints are often encoded as fixed-sized bit strings, where the presence of each structural feature is captured by the fingerprint and results in a bit becoming set to "1". These groups of binary descriptors are able to represent the presence or absence of features from fragments within molecules using a series of binary digits.

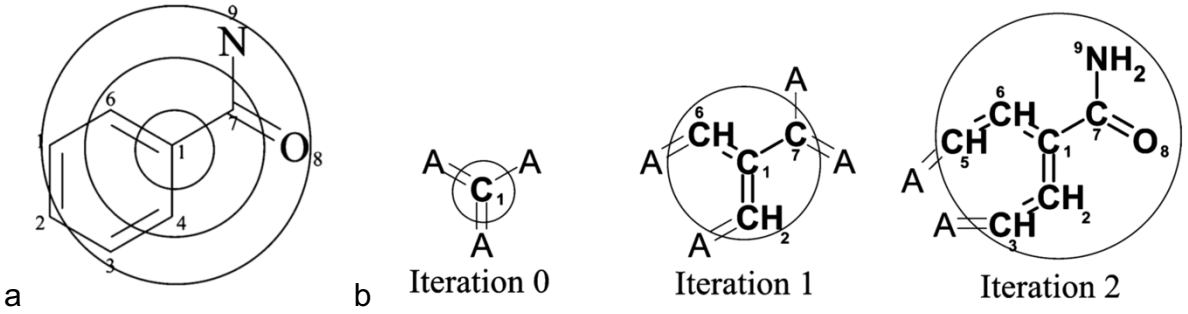


Figure 2.2: An illustrative example of circular atom environments within Benzoic acid amide. a) Atom numbering. The central carbon atom is described by the atom environments at the second and third level, illustrated here via the two larger circles b) Iterative updating on the information represented by an atom identifier. "A" represents any atom other than hydrogen. Reprinted with permission from Rogers et al.[81]. Copyright 2010 American Chemical Society.

2.2.3. Calculating uncertainty values for ChEMBL activity labels

Prior to the application of the PRF algorithm, the calculation of uncertainty in bioactivity labels was required. Since uncertainty originates from the hypothesis that bioactivity data extracted from public bioactivity databases have a degree of uncertainty, an

uncertainty was introduced into the labels. Thereby, labels were treated as probability distribution functions, rather than deterministic values by "injecting noise" in the following way. Bioactivity training data were converted into an uncertainty-based scale on a per-threshold basis ($pActivity^T$), across a range of arbitrary standard deviation (σ) thresholds ranging between 0.0 and 0.6, at increments of 0.2. By varying the standard deviation, σ, the model behaviour over a range of uncertainties was evaluated.

For each bioactivity value ($pActivity$), the cumulative distribution function (cdf) of a normal distribution (Equation 2.1) with a mean equal to the bioactivity threshold for each $pActivity^T$ was used. More concretely, assuming only the mean and variance of activity values is known, the maximum entropy distribution to represent these values is a normal distribution. One can set the mean and variance parameters of this distribution to a threshold value (e.g., 10 μM), and experimental error (e.g., σ of 0.3) and compute the probability of activity values with the cdf. Each $pActivity$ value was converted to a y-label probability (Δy), a value representing the uncertainty in the measurement which was used for PRF training. This value is described as the 'ideal y-label' or simply 'y-ideal' because it represents the ideal case, where experimental error is taken into account when training a target prediction model. For the calculation of Δy, the stats.norm.cdf() function was used from scipy[238] library in python as in Equation 2.1:

$$\Delta y(\vec{c}) = \frac{1}{2}\left[1 + \mathrm{erf}\left(\frac{\overrightarrow{p_{Activity}} - \overrightarrow{p_{Threshold}}}{\sigma\sqrt{2}}\right)\right] \qquad \text{(Equation 2.1),}$$

where Δy were the y-label probabilities, $\vec{c} = (C_1, \ldots, C_n)$ represented the compounds in the training set, $\overrightarrow{p_{Threshold}}$ described the pre-defined binding affinity thresholds for $\overrightarrow{p_{Activity}}$ (-log[10]) values, and σ was the standard deviation defined in this work using arbitrary defined cut-offs (which could also be set as required to the deviation across replicates within or between experiments, screening platforms or activity unit aggregation methods).

Values of Δy hence captured the likelihood that a given compound C_n had binding affinity that falls within the boundary of the active class at the $p_{Threshold}$ given $p_{Activity}$ and given the assumption that most bioactivity data is homoscedastic (which is not always true in practice). Hence, a compound with a pChEMBL value of e.g., 5.1 (8

μM) was assigned a new Δy of ~0.63 for a pChEMBL activity threshold of 5.0 (10 μM) and a user-defined standard deviation σ of 0.3 (Figure 2.3), i.e., there is a 63% chance for that compound to belong to the active class given those parameters compared to traditional RF classifier which assumes that it is 100% active. This enabled representing the activity in a framework in-between the classification and regression architecture, with philosophical differences from either approach. Compared to classification, this approach enables better representation of factors increasing/decreasing inactivity. Conversely, one can utilize all data (even delimited/operand/censored data far from a cut-off) at the same time as taking into account the granularity around the cut-off, compared to a classical regression framework. Thereby, PRF combines characteristics from both classification and regression settings.

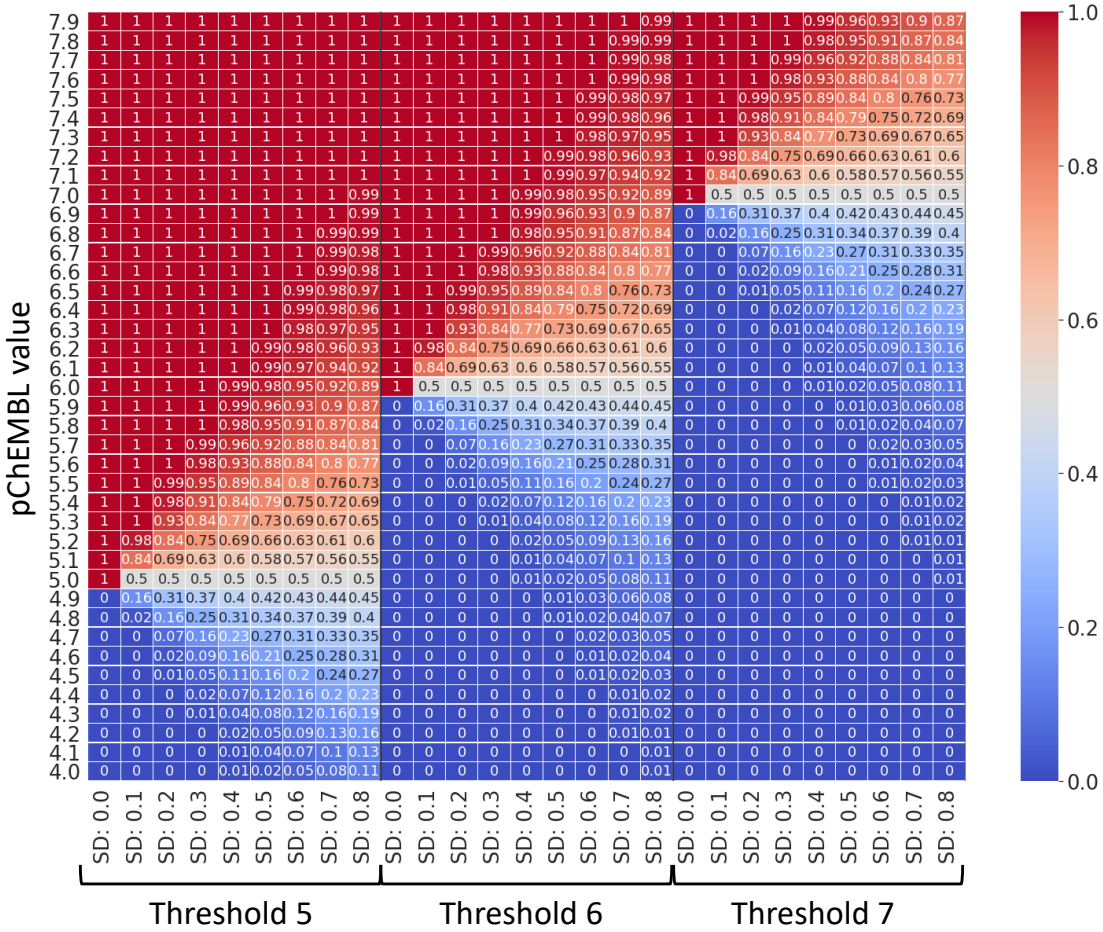


Figure 2.3: Schematic representation of how pChEMBL value is converted into the ideal y-label probability using cdf with different bioactivity thresholds and standard deviation (SD) values. The case when SD is 0 corresponds to traditional RF.

2.2.4. Supplemental inactive data

In order to ensure sufficient chemical space of compounds not binding to targets (hence assigned a constant [$p_{Activity}$= 0] across all test-train standard deviations) an inactive dataset of compounds from PubChem was used as published in Mervin et al.[163] and available at https://github.com/lhm30/PIDGINv3. These supplemental inactive compounds were randomly sampled from PubChem with a Tanimoto coefficient fingerprint similarity to actives lower than 0.4 to obtain the desired number of compounds, which could reasonably be assumed to be inactive against a given target. The dataset included 38,902,310 inactive labelled compound annotations across the full complement of targets. For these inactive datapoints, Δy remained constant across test-train σ thresholds (i.e., only bioactivity data points from ChEMBL were assigned Δy probabilities greater than zero). In more detail, out of a total of 557 models trained (e.g., with a pXC_{50} threshold equal to 5), 310 models (~56%) included at least 1 SE datapoint in the inactive set of compounds and the percentage of SE data included in the inactive data of the 310 models is shown in Figure 2.4. In more detail, 183 models (33% of total models) were trained with a small number of SE data of about 20% of the total inactive compounds and 116 models (21% of total models) were trained with a high number of SE data points (more than 80% SE data in the inactive compound set).

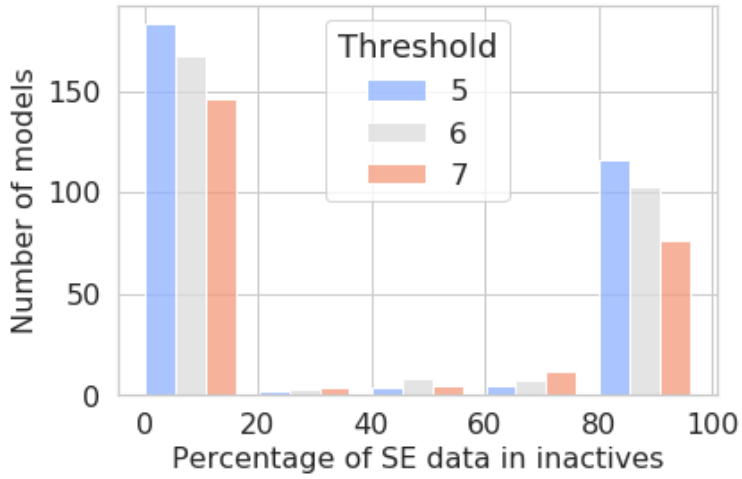


Figure 2.4**:** Percentage of sphere excluded inactive molecules included in the inactive molecule datasets of models across the three different bioactivity thresholds. Statistics show that the putative inactive compounds (calculated with sphere exclusion) account for up to 20% of the total inactive compounds for the majority of the targets that contain putative inactive compounds.

2.2.5. Machine learning modelling and benchmarking

The Probabilistic Random Forest (PRF) is a modification to the original RF algorithm. RF was described in the introduction of this thesis, and it is an ensemble method using a number of decision trees during training. Each decision tree is described via a tree-like graph relating the relationships between (chemical) features and target (activity) variables in a series of conjoined conditions arranged in a top-to-bottom "tree-like" structure. RFs receive a sample of observed random pairs of random variables, $(x_1, y_1), \ldots, (x_n, y_n)$ describing the relation: $h: X \rightarrow Y$ used to predict y for a given value of x. On the other hand, the PRF receives $(x_1, y_1, \Delta x_1, \Delta y_1), \ldots, (x_n, y_n, \Delta x_n, \Delta y_n)$, where Δx and Δy represent uncertainty in features and labels, respectively. Naturally, the focus of this work is concerned with (activity) label uncertainties, and (chemical) feature uncertainties are not specified.

To account for uncertainty, the PRF treats labels as normal distributions, rather than deterministic values. Labels become probability mass functions (PMFs) where each object has a label assigned to it with some probability and the relationship between RF and PRF follow naturally from this concept, since the PRF converges toward a RF when there are low or no (zero) uncertainties in Δy (See Figure 2.3). Another difference between the two algorithms is that randomness of a RF is induced epistemically (i.e., from the model itself) by training different decision trees on randomly selected subgroups of the data and by using random subsets of features in each node of each decision tree. On the other hand, PRF introduces randomness allosterically; since it is not drawn from a defined distribution, but rather the underlying uncertainty (experimental deviation) relevant for classification. Label uncertainties propagate through the splitting criterion during the construction of the tree. Similar to a standard tree, nodes are split left and right, such that resulting subsets are more homogeneous than the set in the parent node. A cost function for minimization is used for this purpose since the transition from y to Δy means that labels now become random variables. Instead of calculating the fraction of objects in node, n, the expectancy value ($\pi_i(n)$) is calculated:

$$P_{n,A} \rightarrow \bar{P}_{n,A} = \frac{\sum_{i \in n} \pi_i(\eta)\ x\ p_{i,A}}{\sum_{i \in n} \pi_i(\eta)} \qquad \text{(Equation 2.2)}$$

$$P_{n,B} \rightarrow \bar{P}_{n,B} = \frac{\sum_{i \in n} \pi_i(\eta)\ x\ p_{i,B}}{\sum_{i \in n} \pi_i(\eta)} \qquad \text{(Equation 2.3)}$$

Hence, Gini impurity is transformed to:

$$G_n \rightarrow \bar{G}_n = 1 - \left(\bar{P}_{n,A}^2 + \bar{P}_{n,B}^2\right) \quad \text{(Equation 2.4)}$$

The cost function (weighted average of the modified impurities of the two nodes) is then:

$$\bar{G}_{(n,r)}\, x\, \frac{\sum_{i\in(n,r)} \pi_i(\eta,r)}{\sum_{i\in n} \pi_i(\eta)} + \bar{G}_{(n,l)}\, x\, \frac{\sum_{i\in(n,l)} \pi_i(\eta,l)}{\sum_{i\in n} \pi_i(\eta)} \quad \text{(Equation 2.5),}$$

The modified propagation scheme and cost functions are the two major conceptual changes separating PRFs and RFs. After training, the PRF classifies new objects which is identical for both training and prediction. Once an object reaches a terminal node the class probability can be used to provide the prediction as in the classical RF, since each object reaches all the terminal nodes a probability. Hence, all the predictions given by all the terminal nodes should be taken into account to obtain the prediction of the tree, which is given by the following equations:

$$Pr_A \rightarrow \sum_{terminal\ nodes} \pi(n)\, x\, \bar{P}_{n,A} \quad \text{(Equation 2.6)}$$

$$Pr_B \rightarrow \sum_{terminal\ nodes} \pi(n)\, x\, \bar{P}_{n,B} \quad \text{(Equation 2.7)}$$

2.2.6. Computational Details

The PRF implementation in Reis, Baron and Shahaf[234] was employed for this work as provided via https://github.com/ireis/PRF. The algorithm was fit with the RDKit fingerprints and the corresponding Δy labels on a per standard deviation (σ) basis, with a lower propagation probability limit ("keep_proba") of 0.05, to ensure that a given object did not propagate to branches with a low probability (reducing runtime without impairing performance). The output of the PRF was recorded as the number of probabilistic decision trees in the forest predicting the label. The RF was implemented using the RandomForestClassifer function from scikit-learn.

Two different metrics were used to compare the PRF and RF prediction probabilities. The first metric is the error margin as described in Equation 2.8:

$$Error\ margin = \begin{bmatrix} (ideal\ ylabel\ - \boldsymbol{RF}\ probabilities) - \\ (ideal\ ylabel\ - \boldsymbol{PRF}\ probabilities) \end{bmatrix} \quad \text{(Equation 2.8),}$$

where ideal ylabel is the activity label, which takes into account experimental uncertainty and RF and PRF probabilities are the probabilities returned from the RF and PRF classifiers respectively.

In addition to the error margin, when two scores are compared (y-probability from 1. RF and 2. PRF) rather than comparing only the absolute values, it is also possible to compare the scores relative to each other. This is achieved by calculating the relative increase toward the potential optimum (i.e., the ideal y-label) as shown in Equation 2.9:

$$Relative\ score = \frac{||error\ margin\ RF| - |error\ margin\ PRF||}{error\ margin\ (worst\ performing\ classifier)}\ x\ 100 \quad \text{(Equation 2.9)}$$

The rationale behind this calculation is that for a metric with an ideal y-label e.g., equal to 0.65 a difference between RF and PRF y-probabilities from 0.75 to 0.70 is more meaningful than a difference from 0.85 to 0.80. In terms of relative score, the latter and the former difference in y-probabilities correspond to 50% and 25% change respectively.

2.2.7. Evaluation of Sphere Exclusion effect on the fraction of improved models by PRF

The effect of including sphere excluded putative inactives on the error margins by Probabilistic RF was evaluated. In this comparison, a) targets that did not contain any putative inactives (models without SE data) and b) targets that 80% of their inactive datapoints were putative inactives (models with SE data) across the three different bioactivity thresholds and different emulated test-train standard deviations were selected. The error margin between the two algorithms was calculated (as described in the section above) separately for models without SE data and models with SE data across different standard deviations. As a result, two error margin distributions were derived and used to compare their means to understand if there is a statistically significant difference. Firstly, a Kolmogorov Smirnov (KS) test in scipy (scipy.stats.kstest) was applied to confirm if the data in error margin distributions are normally distributed. Next, an unpaired t-test (scipy.stats.ttest_ind, with ‘equal_var’ parameter equal to False) was applied to statistically compare the distributions.

2.3. Results & Discussion

2.3.1. ChEMBL experimental variability

The standard deviation across various aggregation schemes for the bioactivity data in ChEMBL 27 was first and evaluated and is outlined in appendix (Table 7.2), to better understand the influence of different approaches toward aggregation as a product of the observed standard deviation between replicate measurements for the same compound-protein target pair. Results from this analysis are presented in Figure 2.5. It can be seen that there is a standard deviation between 0.22–0.41 depending on method of bioactivity data aggregation between the different grouping schemes. As expected, the smallest median deviation in experimental values of 0.04 was observed within the same experiment (replicate) for "intra-assay" aggregation, when compound-target pair replicates were compared within the same experiment. On the other hand, a high standard deviation (0.41) in experimental values was observed across different assay ids and the main reason is that different assay protocols were being used. Though there is an effort to better document and report experimental details regarding assays[239], significant variability was observed between measurements taken in different labs even when assay conditions appear to be the same. In addition, as previously outlined in the work of Kalliokoski, Kramer, Vulpetti and Gedeck[61], aggregation across IC_{50} values was problematic and produced one of the highest median standard deviations of 0.37 for the "Intra IC_{50} type" bin. This is because IC_{50} values are assay-specific and comparable only under certain conditions, which also illustrates the danger of pooling IC_{50} values from different experiments, as is frequently done in the literature (mostly due to lack of alternatives).

Hence, decisions should be taken when aggregating data from databases because of trade-off between increasing data set size versus increasing the discrepancies between the assay technologies and reported activity types (K_i vs IC_{50}). Therefore, one needs to vary the standard deviation depending on the data that is being modelled and how stringent the aggregation function that has been employed.

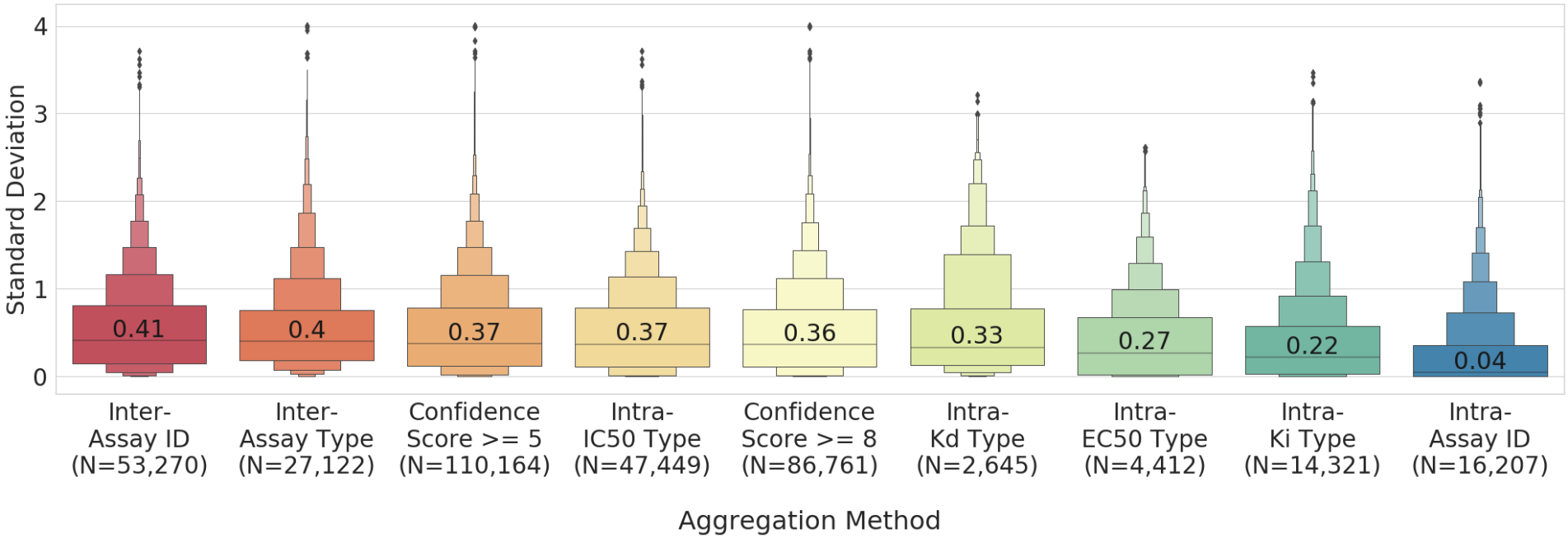


Figure 2.5: Standard deviation of replicate affinity measurements (IC_{50}/EC_{50}/K_i/K_d) across different aggregation types. Standard deviations range between a median of ~0.04 to 0.41 depending on the method of aggregation used for cross-comparison. The median values are shown in each box.

2.3.2. Probabilistic random forest (PRF) performance

In a first step toward benchmarking the PRF, the performance of RF and PRF was compared by taking into account uncertainty around the bioactivity threshold. The difference of performance between PRF and RF was defined as the difference between RF error margin and PRF error margin. Error margin was the difference of each classifier's predicted probability to the 'ideal' y-label probability calculated with the cumulative distribution function (which takes into account both bioactivity threshold and a range of pre-defined values of σ for both test and train sets). Results of this analysis for a pChEMBL cut-off of 5 (0.1 μM) are outlined in Figure 2.6 (complete analysis of pChEMBL of 5, 6 and 7 with different combination of SD in train and test set are included in appendix; Figure 7.2, Figure 7.3 and Figure 7.4, respectively).

Figure 2.6 shows that PRF outperformed RF when there was a degree of uncertainty in the data (i.e., a σ greater or equal to 0.2). For example, when the $\sigma = 0$, the median error margin between the two algorithms was close to 0 (-0.010 to 0.005) across all y-ideal probabilities. However, as the standard deviation in the data increased, the absolute error margin between the two algorithms was increasing too. When e.g., $\sigma = 0.4$ and $\sigma = 0.6$ the median error margin ranged from -0.029 to 0.005 and from -0.039 to 0.004 respectively. Therefore, these results indicated that when σ of training data is 0, there were no substantial differences in the predictions between algorithms and this was not true as the standard deviation increased. This observation is in agreement with previous benchmarking of PRF in a different type of noisy data (astronomical data)[234] and the difference in classification accuracy between the two algorithms (RF and PRF) increased with increasing noise level and complexity.

Moreover, Figure 2.6 highlights areas in the y-ideal probability ranges, where PRF outperformed RF. For example, when there was an uncertainty in the data and σ was equal to 0.2, 0.4 and 0.6, PRF outperformed RF with an average absolute error margin equal to 0.011, 0.024 and 0.037 for y-ideal probability ranges of 0.4 -0.6. However, when y-ideal probability ranged from 0.7 to 1, the absolute error margin between the two algorithms was smaller and equal to 0.005, 0.009 and 0.011 for σ equal to 0.2, 0.4 and 0.6 respectively. A similar trend was observed when the y-ideal probability ranged from 0.0 to 0.3 and the absolute error margin was equal to 0.012, -.015, 0.015.

Therefore, PRF showed a highest absolute error margin and thus was outperforming RF for the y-ideal probabilities closer to midpoint. Therefore, the PRF exhibited the largest benefit over the RF (defined as the lowest delta between PRF error and Scikit-Learn RF error) toward the midpoint of the probability scale, for marginal cases on the binary threshold boundary. This is because the original RF weights the marginal cases as equivalent in distinguishing between activity classes. In this case the PRF classifier was able to better model the granularity around the activity threshold cut-off, as in a regression.

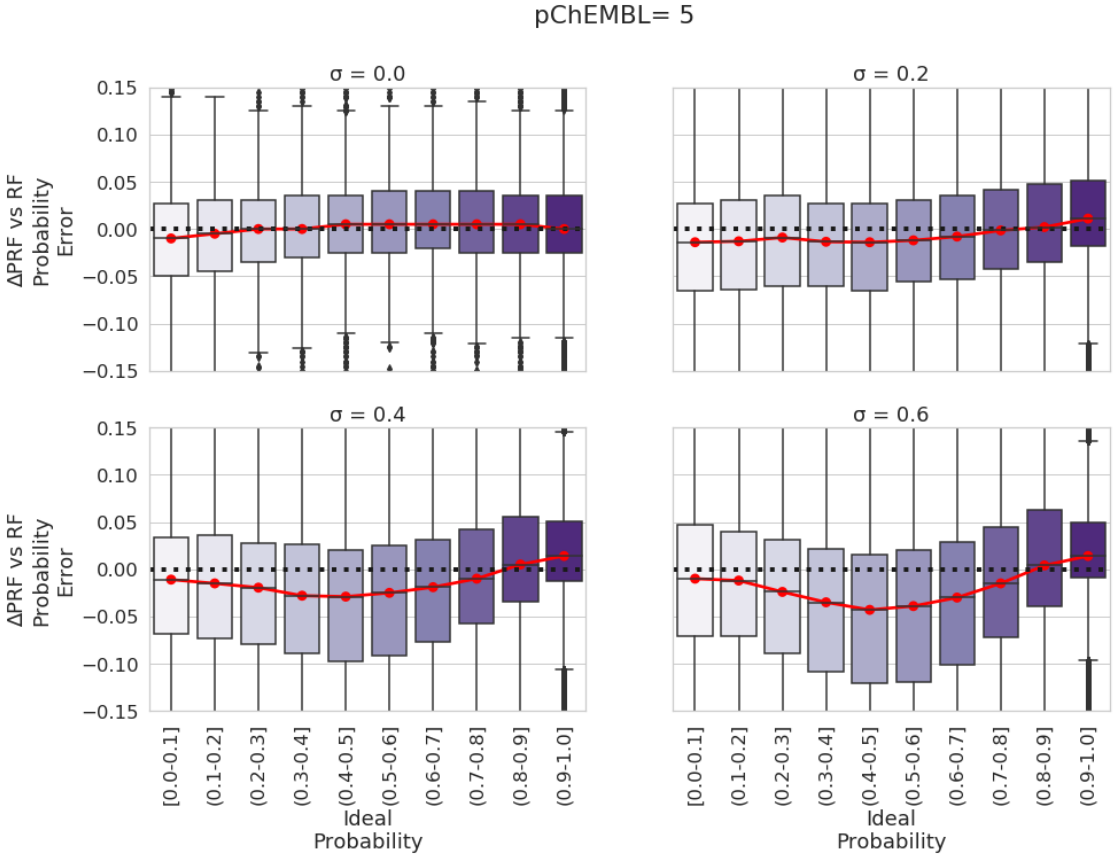


Figure 2.6: Ideal probabilities as a function of the delta of PRF versus RF error margins across emulated train-test standard deviations. Overall, results shown here for a threshold of pChEMBL value of 5 (0.1μM) highlight the most optimal PRF probability estimates were observed in cases when standard deviation in the test set most closely resembled that in the training set. It can also be seen that the largest benefit in terms of error margin for the PRF (lower values on the y-axis) are observed toward the midpoint of the ideal Δy scale, particularly for higher training set standard deviations. This is when the original RF weights the marginal cases equivalent in distinguishing between activity classes. The same observation was observed for pChEMBL thresholds of 6 and 7, as shown in Figure A.2.3 and Figure A.2.4, respectively.

The findings reported above are specific to an analysis using the Scikit-Learn implementation RF. In order to check that the above findings are robust and not due to differences between packages, a similar analysis was conducted emulating a classical RF (i.e., when the binary labels are supplied rather than the probabilities) via the PRF package, as described in the methods. A high overall R^2 correlation between Scikit-Learn RF and the PRF ($\sigma = 0$) ranging between ~0.97-0.98 across the standard deviation test sets was observed (as presented in appendix Figure 7.5), hence the returned predictions from both RF approaches were overall comparable and the findings presented in this study are robust between packages.

The significance of differences between the RF and PRF modelling uncertainties were evaluated, based on the differences between output and expected values (y-ideal probability). To evaluate this, the relative score calculation was applied as described in the methods, to identify the percentage improvement for each algorithm across different emulated train-test standard deviations and different ranges of ideal y-label. As shown in Table 2.1, PRF showed the greatest percentage improvement (~17%) when SD of train and test set ranged from 0.4 to 0.6 and when the ideal y-label ranged from 0.4 to 0.6 and thus the data were close to the bioactivity threshold. Thus, the improvement of correct class assignments showed that PRF has an advantage compared to RF when there was a degree of uncertainty in the data and additionally PRF performed better for values toward the midpoint of the probability scale as also shown across algorithm error margins in Figure 2.6.

Table 2.1: Average percentage improvement between RF and PRF probabilities in relation to ideal y-label values across different emulated train-test standard deviations (SDs) when pChEMBL threshold equals 5.

Standard Deviation in train and test set	y-ideal range (N)	Better- performing Algorithm	% Improvement
SD-train: 0.0-0.4 & SD-test: 0.0-0.4	0.0-0.2 (183,255)	PRF	4.79
	0.2-0.4 (79,890)	PRF	3.83
	0.4-0.6 (124,505)	PRF	10.8
	0.6-0.8 (166,210)	PRF	5.76
	0.8-1.0 (1,007,685)	RF	6.57
SD-train: 0.4-0.8 & SD-test:0.4-0.8	0.0-0.2 (152,835)	PRF	0.27
	0.2-0.4 (194,300)	PRF	9.27
	0.4-0.6 (339,495)	PRF	16.89
	0.6-0.8 (592,575)	PRF	11.04
	0.8-1.0 (5,624,495)	RF	9.59

Overall, results showed that PRFs were able to capture the experimental/aggregational variability in ChEMBL. The maximum achievable accuracy of PRF models was more closely related to the true reproducibility across the experimental data (in this case when aggregated across experiments and measurement data types). In comparison, the baseline RF (when $\sigma = 0$) yielded a reported performance smaller than the experimental uncertainty, which indicated cases of overfitting and/or over-confidence. Therefore, PRF is an algorithm that should

be considered as an alternative to RF when we have *a priori* knowledge that our training data are noisy.

2.3.3. Effect of Sphere Exclusion, dataset imbalance and model set size

Previous studies link Sphere Exclusion (SE) with inflated model performance and poor model calibration (due to the artificial requirement for putative non-binding molecules to be dissimilar to their active counterparts)[240,241]. Conversely, experimentally confirmed inactive compounds are likely to be more skeletally similar to actives and this trend blurs the algorithm's decision boundary between the active and inactive classes. Hence, the next parameter which was evaluated was whether the presence of SE inactives influenced PRF performance by comparing the fraction of targets improved by PRF with the classical RF, for models with/without putative inactives. The error margin was explored between PRF and RF for target protein models that included a high number of putative inactives in Figure 2.7a (detailed comparison shown in appendix Figure 7.6) and for targets that did not include any putative inactives (appendix Figure 7.4). Overall, results showed that the PRF exhibited the largest benefit over the RF toward the midpoint of the probability scale, for marginal cases on the binary threshold boundary and when there was a degree of uncertainty in train and test set (otherwise for low SD PRF converged to classic RF). These observations are in agreement with the previous observations in Figure 2.6**,** where the error margin for all the models was evaluated and thus the addition of putative inactive compounds did not affect the performance of PRF compared to RF.

In addition, the effect of including sphere excluded putative inactive compounds on the error margins between the two algorithms separately for models without SE data and models with SE data across different standard deviations was explored. By applying a Kolmogorov Smirnov (KS) test, the data in error margin distributions were normally distributed and therefore an unpaired t-test was applied to compare them. The error margin distributions and the result of unpaired t-test are shown in Figure 2.7b. Overall, results showed that there was no statistically significant difference between models with and without SE data when SD was equal to 0. However, as the SD increased (0.2, 0.4 and 0.6), there was a statistically significant difference between the error margins of the models with and without SE data with p-values less than 0.05.

The addition of SE data reduced the difference between PRF, and RF compared to models without SE data. The rationale behind this observation could be that for the putative inactives, a pXC50 value cannot be assigned and thus it is not possible to evaluate their uncertainty. Hence, they are considered as inactives with a low uncertainty and i.e., far from the bioactivity threshold. Therefore, a large number of putative inactives could be problematic when combined with PRF but on the other hand their inclusion can enlarge the models' applicability domain.

In addition, it was next investigated how significant are the differences between RF and PRF in terms of how close they are to the real value (y-ideal probability). To this end, the relative score calculation was applied (Equation 2.9) to identify the percentage improvement for each algorithm across different training conditions (standard deviation in train and test set) and different ranges of y-ideal range for the targets that included at least 1 SE datapoint in the inactive dataset as shown in Table 2.2. The main observation is that PRF showed the greatest percentage improvement 11.58% and 14.68% when SD of train and test set ranged from 0.0 to 0.4 and 0.4 to 0.6 respectively and when the ideal y-label ranged from 0.4 to 0.6 and thus the datapoints were located close to the bioactivity threshold. On the other hand, RF showed a ~12% improvement when ideal y-label ranged from 0.8 to 1.0 and therefore RF worked better for datapoints that were assigned as actives with a high confidence. Therefore, the inclusion of SE data did not affect the percentage of improvement in different SDs and y-ideal probability changes.

In a final analysis, an evaluation was performed to explore the influence of dataset size on the performance difference of PRF versus traditional RF models. A correlation analysis, (presented in appendix Figure 7.8) showed no discernible correlation between PRF and RF difference and training size, since no significant Pearson correlation exists across the four arbitrary standard deviations (σ) evaluated (Pearson *r* values ranged between -0.22 to -0.03). Hence, the conclusion was that PRF can be used regardless of the dataset size for cases when experimental uncertainty is large, and where values are distributed around the classification threshold.

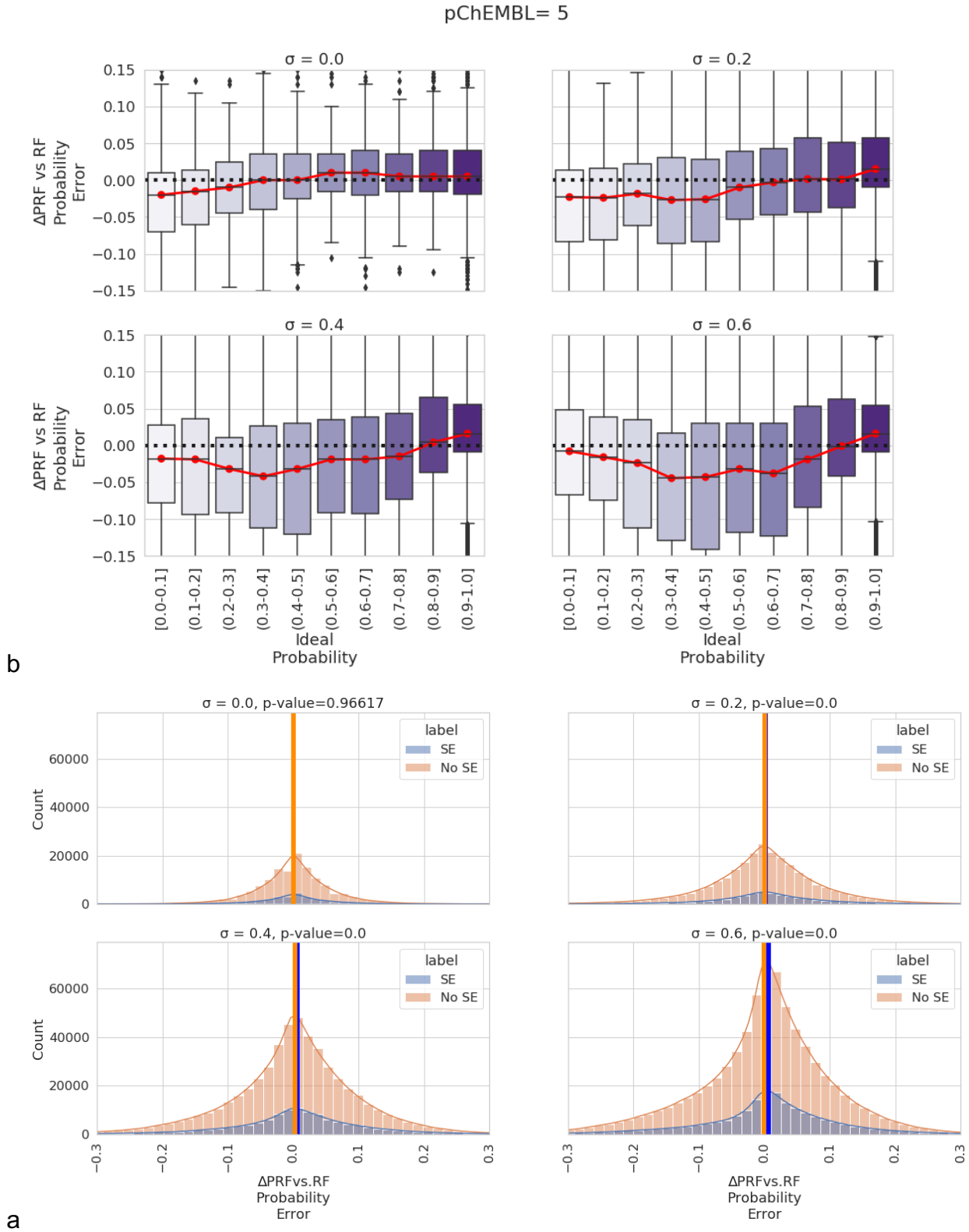


Figure 2.7: a) Ideal y-probabilities as a function of the delta of PRF versus RF error margins across emulated train-test standard deviations. Overall, results shown here for a threshold of pChEMBL value of 5 (10 μM) highlight the most optimal PRF probability estimates were observed in cases when standard deviation in the test set most closely resembled that in the training set. It can also be seen that the largest

benefit in terms of error margin for the PRF (lower values on the y-axis) are observed toward the midpoint of the ideal ∆y scale, particularly for higher training set standard deviations. This is when the original RF weights the marginal cases equivalent in distinguishing between activity classes. b) Effect of Sphere Exclusion (SE) on the error margin between models with and without SE data across different emulated test-train standard deviations. Overall results show that there is no clear advantage of including or excluding SE data when there is no SD in the data. When SD is greater or equal to 0.2, there is a statistically significant difference and hence the inclusion of SE data reduces performance of PRF.

Table 2.2**:** Average percentage improvement between RF and PRF probabilities in relation to ideal y-label values across different emulated train-test standard deviations (SDs) when pChEMBL threshold equals 5.

Standard Deviation in train and test set	y-ideal range (N)	Better - performing Algorithm	% improvement	Average Percentage of SE data
SD-train: 0.0-0.4 & SD-test: 0.0-0.4	0.0-0.2 (104,345)	PRF	6.63	38.66
	0.2-0.4 (42,075)	PRF	5.19	34.03
	0.4-0.6 (63,520)	PRF	6.42	36.74
	0.6-0.8 (86,27)	PRF	3.19	36.01
	0.8-1.0 (530,080)	RF	6.96	31.57
SD-train: 0.4-0.6 & SD-test:0.4-0.6	0.0-0.2 (92,720)	PRF	0.23	42.68
	0.2-0.4 (106,755)	PRF	11.65	35.82
	0.4-0.6 (173,070)	PRF	16.76	36.08
	0.6-0.8 (314,270)	PRF	11.60	33.52
	0.8-1.0 (3,022,800)	RF	9.48	29.99

2.3.4. Case Study: PRF improves PDK1 model performance

After taking into account the learnings from the previous analyses, the conclusion was that PRF exhibited the largest benefit over the RF toward the midpoint of the probability scale, i.e. for marginal cases on the binary threshold boundary. Therefore, one example target was selected to showcase how PRF can be useful to predict compounds near the bioactivity threshold with higher confidence compared to classic RF.

The protein target selected for this analysis was Pyruvate dehydrogenase kinase isozyme 1, encoded by the PDK1 gene, which has been investigated as a potential drug target for breast cancer, due to its essential role in regulating cell migration[242]. This particular target was chosen due to the large proportion of reported activity data measured close to the bioactivity threshold, (i.e., ~60% of the training labels for PRF ranged between 0.3-0.6), as shown in Figure 2.8a. This behaviour can be contrasted to the distribution of binary labels for the classical RF, where the majority of labels (1000 compounds) were assigned (0) for the “non-binding” class. A replicate analysis (analogous to the one presented in section above; 2.3.1 ChEMBL experimental variability) was conducted. One replicate from the same assay showed a low standard deviation of 0.1 whilst the majority of other replicates (across assays and measurement types) showed higher deviations around ~0.3, as outlined in Figure 2.8b. Hence, replicate aggregation is shown for this target to introduce uncertainty into the bioactivity labels in accordance with the global analysis previously outlined. Finally, using different thresholds on the raw probabilities (0.5, 0.6, 0.7) returned by PRF and RF, PRF outperformed the traditional RF and maintained a higher performance compared to RF even when a higher threshold on probabilities was used (Figure 2.8c-e). This illustrates the benefit of taking experimental uncertainty into account using the PRF classifier, as opposed to a RF classifier, for protein targets where much of the data is located around the decision boundary on a concrete dataset.

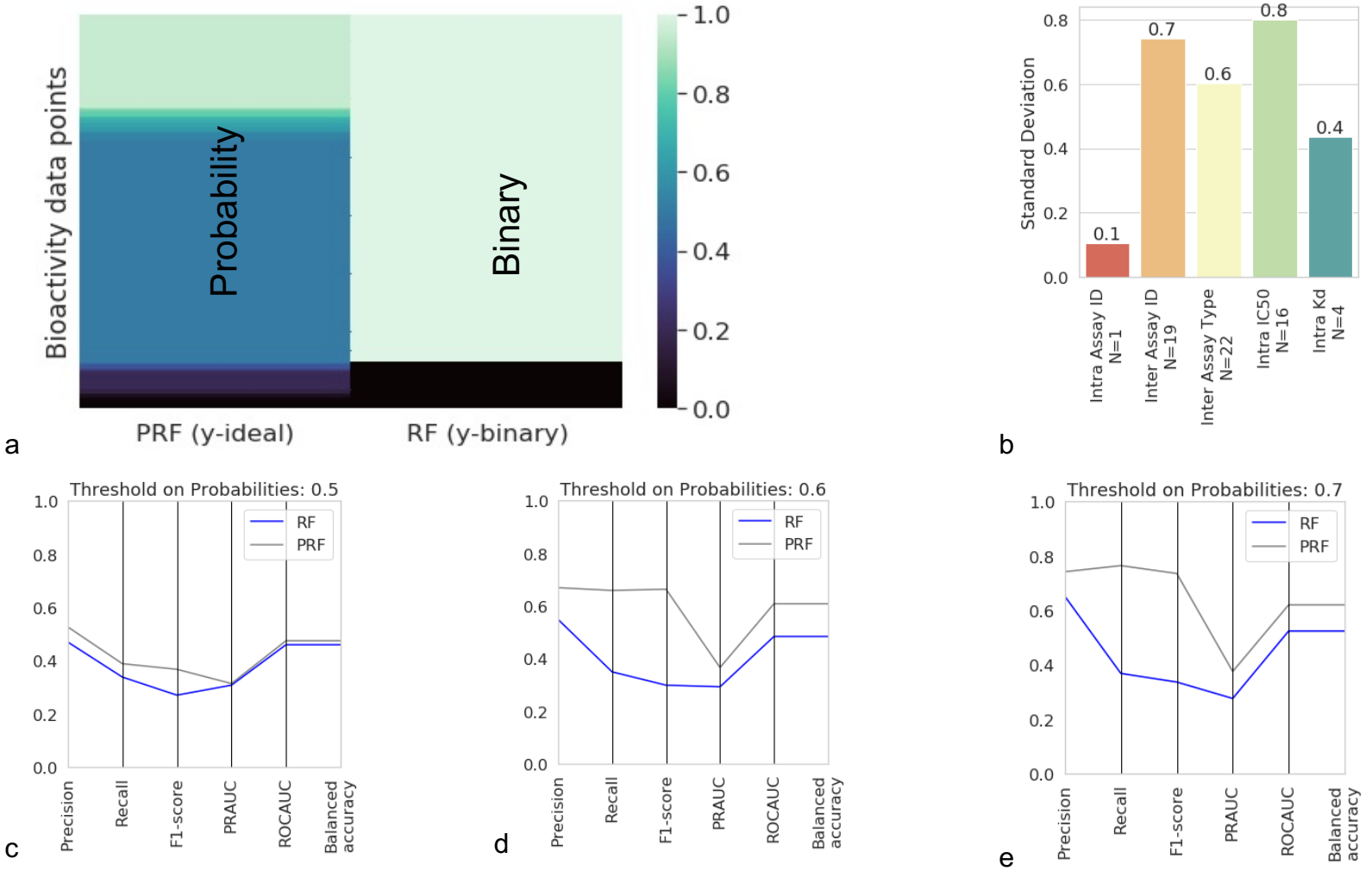


Figure 2.8: a) Distribution of the y-ideal label versus binary y-labels for values close to bioactivity threshold. b) Experimental error in ChEMBL for [Pyruvate dehydrogenase (acetyl-transferring)] kinase isozyme. The error is high when data are derived from different assay IDs and IC_{50} measurements. c-e) Performance of the PRF versus RF classifier using different evaluation metrics and different thresholds on algorithms probabilities and y-ideal labels.

2.4. Conclusion

In conclusion, the aim of this chapter was to investigate the performance of Probabilistic Random Forest (PRF) as a method able to take into account experimental errors, which are usually a neglected aspect of model generation. By evaluating the current experimental error in ChEMBL v27, it was identified that it is very similar to those reported in previous versions of ChEMBL v14. The highest standard deviation in values for the same ligand-target interaction pairs observed for values derived from different assay types and the smallest deviation in experimental values is observed within the same assay id. By applying PRF in target prediction and comparing it to RF there were cases where PRF outperforms RF and *vice versa*. Therefore, the choice should be based on a) training data quality and b) the area of data distribution (i.e., whether they are close to the classification threshold). Firstly, regarding the training data quality, PRF showed a lower error compared to RF when there is a degree of uncertainty in training set (i.e., SD>=0.2). For lower SD in the data (when the uncertainties are set to or close to zero), the PRF converges to the original RF algorithm. When the standard deviation of training set is 0, there are no substantial differences in the prediction of the test set regardless of the standard deviation assigned in the test data. Secondly, PRF exhibits the largest benefit over the RF toward the midpoint of the probability scale, i.e. for marginal cases on the binary threshold boundary. In addition, the addition of sphere excluded inactives affected PRF performance compared to RF and SE data did not affect the observations obtained from the comparison of RF vs PRF. Therefore, PRF can be useful for target prediction and is not affected by the presence of SE data. Based on this chapter observations, PRF can be used for bioactivity prediction in classification setting and in cases where experimental uncertainty is large, and where values are distributed around the classification threshold. Such methods like the PRF, that take into account experimental uncertainty and can result in more accurate predictions are an asset in improving the bioactivity predictions in drug discovery process. Hence, the understanding of compounds' MoA on a target level is performed with less bias from the experimental uncertainty.

3. Chapter 3: Comparison of Structural Chemical and Cell Morphology Information for Multitask Bioactivity Predictions

3.1. Introduction

As discussed in chapter 1 of this thesis, Cell Morphology information is an evolving type of information in the field of MoA understanding. The evolution of this type of information has been significantly influenced by the development of novel high throughput imaging assays such as the Cell Painting. In addition, there is a need for novel compound information which could be complementary to more traditional compound information such as the chemical structure based, gene expression etc.

The use of cell morphology information derived from cell image data to better understand compounds' properties is not new and one of the first approaches to use image-based data to aid MoA understanding was the combination of cytological markers of the cell cycle with chemical knowledge (profiles of chemical similarity and predicted targets). The goal here was to identify compounds that affect cell proliferation and to understand related underlying causal MoAs better[243]. A common factor model was used to reduce the 36 cytological features from images to 6 significant factors, which were further clustered into seven phenotypic categories. Firstly, the comparison between phenotypic and chemical similarity profiles of 211 compounds showed only a moderate but significant positive correlation (Spearman Correlation=0.0746, $p<0.001$), and structurally similar compounds were found in both similar and dissimilar endpoint clusters, underlining the limitations of the 'Molecular Similarity Principle' described above. In particular, similar phenotypes could be caused also by rather distinct molecular structures. For predicting targets, an increased positive correlation (0.136 and $p<0.001$) between phenotypes and targets was observed, showing that the utilization of existing bioactivity data can contribute to meaningful mappings of chemical structure into MoA space. Therefore, this study highlighted those phenotypic profiles are an additional type of information to structural information alone, which can contribute to understanding compound MoA.

In addition, other studies, which are outlined in the introduction have showed that cell morphology information and more specifically profiles derived from the Cell Painting

assay have been useful in target prediction, biological assay prediction and clustering of compounds with the same MoA label. However, in these studies one of the limitations is that Cell Painting profiles are not compared with chemical structure-based information. This is an important consideration because the profiling of compounds with Cell Painting assay is more expensive and time consuming compared to calculating traditional chemical-structure-based descriptors. Therefore, it is important to know when to use Cell Painting profiles for the development of target prediction models. Hence, this chapter aims to predict compound MoA on a target level by comparing cell morphology-based information and chemical structure information. The hypothesis is that there are protein targets that are better predicted by chemical structure information and some others better predicted by image data due to the different activity landscapes across targets and biases in the chemical space of training data. A comparison of the two types of information can, therefore, inform of cases when one particular type of data is more informative compared to the other. To test this hypothesis, a multitask algorithm called Bayesian Matrix Factorization Macau[105,157] and Random Forest (RF) in a single task setting (training and assessing models for each target individually) were applied, for the prediction of the bioactivity of compounds in active and inactive class over a range of targets and using either type of information as side information for each method.

3.2. Methods

3.2.1. Workflow Summary

The workflow employed in this work can be divided into three main steps (see Figure 3.1). The first step is the extraction of side information and in this case, the extraction of image-based features and the calculation of Extended Connectivity Fingerprints (ECFP). The second step is the extraction of the bioactivity data for the main matrix from ExCAPE database[64], which is a merge of ChEMBL[66] and PubChem[67] data. The third step is the application of Bayesian Matrix Factorization Macau and RF to train models and then models are evaluated.

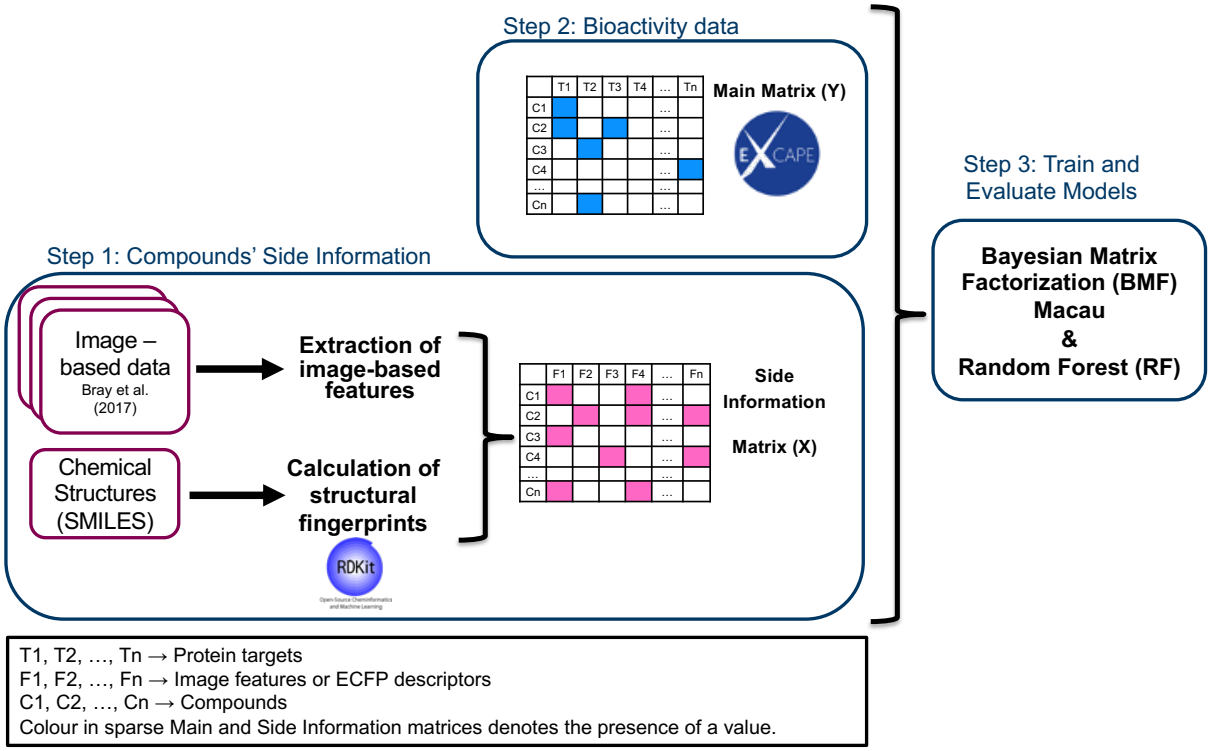


Figure 3.1: Summary of the analysis performed in this work. Image-based data and chemical descriptors (ECFP) are used as features/side information for the training of two models. BMF Macau and RF were used to train models.

3.2.2. Compound information: Image-based features and chemical information

3.2.2.1. Image features extraction and curation

The Cell Painting Assay dataset released by Bray et al.[104] was used to construct the image-based side information matrix. The dataset consists of a variety of perturbations including 10,080 compounds from the Molecular Libraries Small Molecule Repository,

2,260 compounds from the Broad Institute bioactive compound collection and 18,051 novel compounds from diversity-oriented synthesis approaches. The assay stained eight cell components, namely the nucleus, endoplasmic reticulum, F-actin cytoskeleton, Golgi apparatus, plasma membrane, mitochondria, cytoplasmic RNA and the nucleoli. CellProfiler was used to calculate 1,783 morphological features from the cell images. Image-based data (i.e. morphological features) and compound annotations/metadata for the image dataset (e.g. SMILES) were obtained from the GiGaScience Database[244]. Bray et al.[104] performed the assay on 384 well-plates, and there is a minimum of 4 replicates per compound. In each plate, there are 64 wells with DMSO control, which were averaged for each readout dimension. Then the average DMSO vector of each plate was subtracted from the average individual compound's readout features. Finally, four morphological features with zero variance were removed and for the remaining features, Z-score were calculated which were used for further analysis.

3.2.2.2. Compound Standardisation and ECFP calculation

The SMILES of the compounds present in the image dataset were standardised with the ambitcli version 3.1.0[245–247]. The chemical structure processing options follow the same criteria as the compounds in the ExCAPE database[64], namely fragment splitting, isotope removal, implicit hydrogen conversion, tautomer generation and neutralisation. SMILES were also converted to InChIKeys with the ambitcli to query the ExCAPE dataset later.

The non-hashed, count version of ECFP fingerprints was selected as it has previously been shown to successfully capture molecular information in multitask target prediction with BMF Macau[157]. RDkit[237] was used to generate ECFP circular Morgan fingerprints (radius=3, count values, unfolded)[81]. In total, 103,350 distinct features were calculated and features with variance less than 0.005 were removed, resulting in a sparse fingerprint containing the presence or absence of 2,989 features for each compound.

3.2.3. Retrieval of bioactivity data from the ExCAPE database

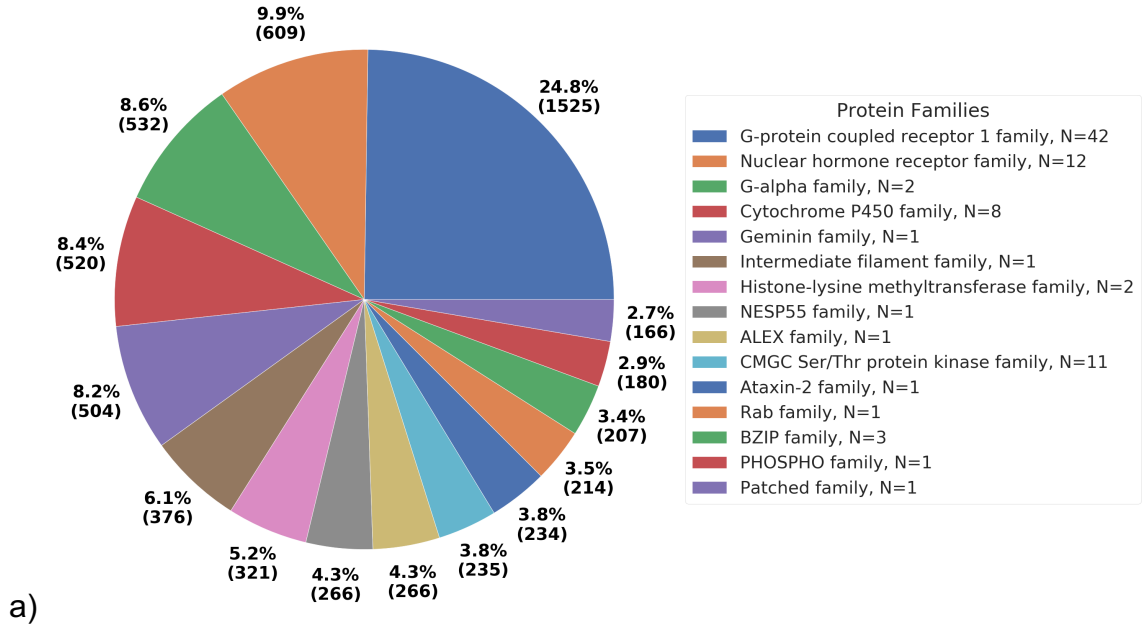
3.2.3.1. Extraction and curation of bioactivity data from the ExCAPE database

The ExCAPE database was queried for the 30,616 InchiKeys of the compounds in the image dataset. This retrieved, 4,148,013 bioactivity datapoints overlapping with the image dataset, and subsequently a curation and cleaning process was applied as follows. The pXC^{50} data was used as provided in the dataset, and there were bioactivity points, where the exact pXC^{50} value was not provided. In this case, only an indication of being active or inactive ('A' or 'I') was given by the dataset. For the compounds-targets combinations that had both an indication of active/ inactive (from a PubChem screen) and an exact pXC^{50} value, only the pXC^{50} value was kept (which was binarized based on a pXC^{50} threshold), thereby resulting in 2,979,179 bioactivity datapoints. The qualitative bioactivity annotations were kept as provided by the database. For the compound- target combinations where bioactivity values existed for different organisms only those from human were kept, resulting in 2,956,174 bioactivity datapoints. Moreover, for compound target combinations where different pXC^{50} values existed, values more than one standard deviation away from the mean (which were assumed to be outliers) were removed and averaged the remaining values. The resulting bioactivity matrix resulted in a sparse matrix of 28,099 compounds, 1,288 targets and 2,907,512 bioactivity datapoints (8% matrix completeness) with very different numbers of data points per class (appendix, Table 8.1). The minimum number of datapoints per target was set to be equal to 5 actives and 5 inactives with a pXC^{50} threshold greater than or equal to 6. Hence filtering the targets with inadequate number of actives and inactives, a 16% complete matrix of 27,753 compounds, 224 targets and 999,602 bioactivities was obtained.

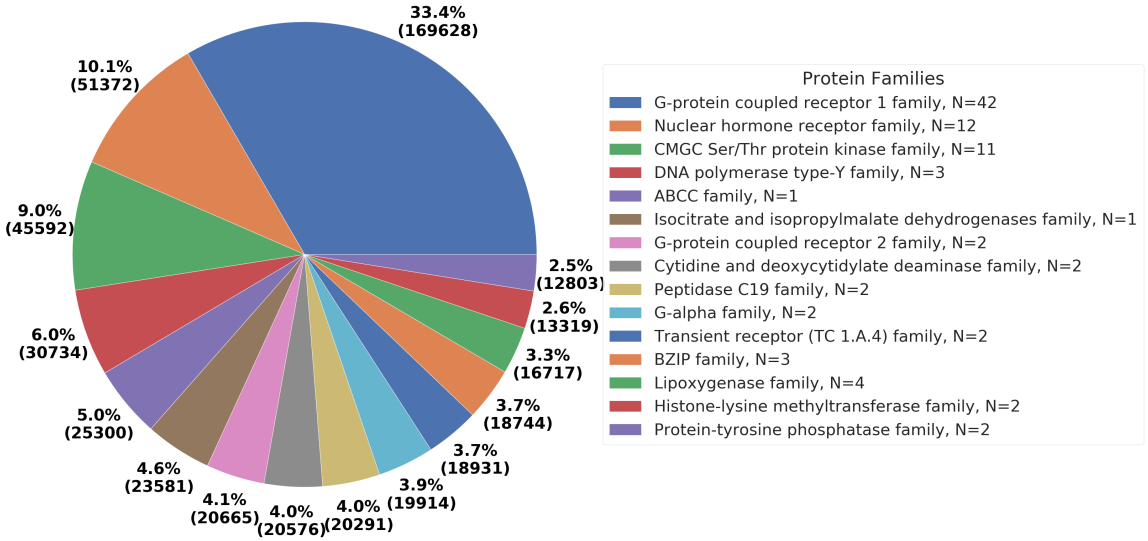
3.2.3.2. Construction of Main Matrices with different ratios of active to inactive datapoints

The 224 targets were categorised into protein families based on information retrieved from UniProt[248] and from the "Family and Domains" column in order to annotate the proteins included in the analysis. The 15 most populated protein families based on the number of actives and inactives are shown Figure 3.2a and b, respectively. The number of bioactivity datapoints is unequally distributed across protein families, where the G-protein Coupled receptor (GPCR) 1 family is the most populated family with

~25% of the total datapoints across 42 targets. The number of active compounds (minority activity class) are many times smaller than the number of inactive compounds (majority activity class), with an overall active to inactive ratio of 1:100 across the 224 targets.



a)



b)

Figure 3.2: Most highly populated protein families with respect to a) active compounds and b) inactive compounds in the ExCAPE database for the compounds in the Cell Painting dataset (protein families were obtained from UniProt[248]). The number of proteins per protein family (N) is shown in the legend and the percentage and actual number of the (a) active and (b) inactive compounds are shown in the main part of the figure.

3.2.3.3. Balancing Dataset for Machine Learning Models

The threshold that was used to separate the actives from the inactives was pXC^{50}=6. There were 224 targets that had a minimum of five active and inactive datapoints, while for 41 out of 224 targets the number of actives (A) is larger than the number of inactives (I). In those cases, all the available bioactivity datapoints were kept in the main bioactivity matrix for all the models trained. For the rest of the targets, the main bioactivity matrix was prepared with a maximum A:I ratio equal to 1:5 and 1:10 (class balanced models) and with all the available bioactivity data from ExCAPE (class imbalanced models). When the A:I ratio was equal to a maximum of 1:5, there were 153/224 targets that had enough inactive compounds to select and 30/224 targets that did not have enough inactives and thus for these targets the A:I ratio is ranging from 0.21 to 0.88. For the targets for which the number of inactives per target was larger than the number of actives and in order to keep the A:I ratio equal to 1:5 and 1:10, the inactive compound space was reduced with the MaxMin algorithm from Rdkit[249]. This algorithm ensures that a small subset of diverse molecules is selected from a larger set and requires a function to calculate distances between compounds and for that reason the Tanimoto similarity was used. The bioactivity data in ExCAPE database either displayed a pXC^{50} value or they have a 'Yes' or 'No' flag if a compound is active or inactive towards a target. Therefore, the compounds with known pXC^{50} values were first included. If the inactives with known pXC^{50} values were still not enough then inactives with a 'No' flag were selected with the MaxMin algorithm from Rdkit[249].

In summary, there are three datasets, one with A:I ratio equal to 1:5, another with A:I equal to 1:10 and one that includes all the available bioactivity datapoints. The models with an A:I ratio of 1:5 consisted of 224 targets, 10,841 compounds and 54,902 bioactivity datapoints with 10,395 out of those being active compounds. When the A:I ratio was equal to a maximum of 1:10, there were 119 out of 224 targets that comprised a sufficient number of inactive compounds, 67 out of 224 targets where this was not the case, and where the A:I ratio was ranging from 0.11 to 0.88 and 35 out of 224 targets that had more actives than inactives. Therefore the models with A:I equal to 1:10 consisted of 224 targets, 11,512 compounds and 93,373 bioactivity datapoints with 10,395 of those being active datapoints. Finally, the imbalanced models that contained all the available bioactivity datapoints from the ExCAPE database consisted

of 224 targets, 27,753 compounds and 999,602 bioactivity datapoints with 10,395 of those being active compounds.

3.2.3.4. Preparation of Train and Test set and Model Evaluation Metrics

Two different methods were used to split the initial main matrix into train and test set five-fold, namely the StratifiedShuffleSplit and GroupShuffleSplit python functions from Scikit-Learn[250]. The Stratified Shuffle Split (SSS) splits a dataset into a training and test set by preserving the same percentage of data for each target class as in the initial dataset. The Group Shuffle Split (GSS) with Murcko scaffolds generates a randomised partition in which a subset of groups (here Murcko scaffolds) is held out and are included in the test set. This methodology ensures that compounds with the same Murcko scaffold will either be in the training or test set, and which is hence more stringent than the SSS. In both cases, the compounds are split for each target into an 80% training set and 20% test set (Figure 3.3). Models were trained and evaluated using five-fold cross-validation.

Figure 3.3: The values in the initial dataset/matrix of compounds n (C1, C2, …, Cn) and targets m (T1, T2, … , Tm) are split by applying SSS per column/target. The 80% of the values are assigned in the training set and the 20% in the test set.

3.2.4. Model Training

BMF Macau is a ML algorithm that is based on Bayesian Probabilistic Matrix Factorization (BPMF), which is usually used in recommender systems[157,251]. This algorithm learns multiple tasks (predicting multiple drug targets for each compound) simultaneously and the learning tasks can hence in principle benefit from each other. The approach works by factorising a sparse matrix Y (N compounds times M targets) containing compound bioactivities to a lower-dimensional representation in latent matrices **u** (Equation 3.2) and **v** (Equation 3.3) as shown in Figure 3.4, for compounds and targets respectively. It is possible to incorporate side information in a matrix Factorization model by taking advantage of a Bayesian framework. In particular,

Macau sets Gaussian distribution priors over latent matrices and uses the side information to form an informative prior. The side information updates the mean parameter of the Gaussian distribution over latent matrix **u**, as it can be seen in Eq. 2 where the coefficient vector β_u refers to importance of each feature in the side information vector X_i for the i_{th} compound.

The prediction for a compound i on target j ($Y_{i,j}$) can then be given as the inner product of $u_i^T\ v_j$ (Equation 3.1). Thus, Macau extends the BPMF methodology by integrating side information for the compounds from a side information matrix. For the image-based model, the side information in the current work was the matrix of image features, while for the fingerprint-based model, the ECFP fingerprints. This information is then combined into the mean of the Gaussian priors, (Equation 3.2), to be used during the model training. BMF Macau models were trained in python with the Macau package[105,157]. The models were trained for 2,000 iterations, discarding the first 400 as burn-in, and 32 latent vectors were used. Different values for the number of iterations, latent vectors and burn-in were tried (appendix; Figure 8.1 and Figure 8.2).

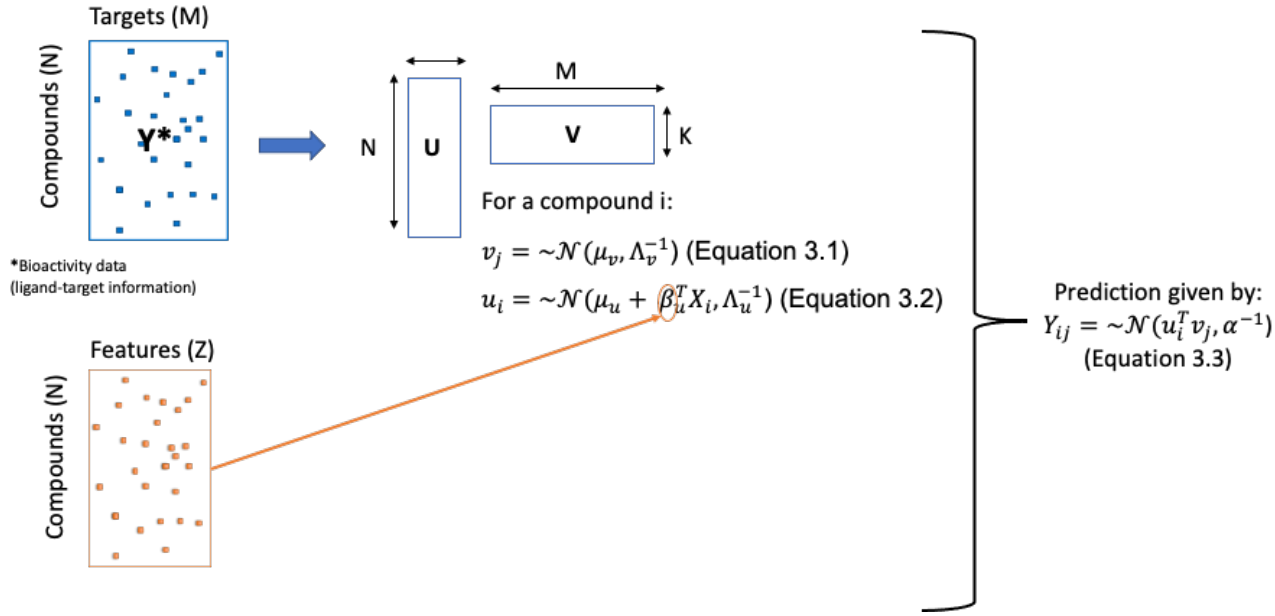


Figure 3.4: Schematic representation of BMF Macau with main equations.

A Random Forest (RF) algorithm in a single task setting (training and assessing models for each target individually) was also selected as a baseline and established method to compare with BMF Macau. Two different types of models were trained, one with the chemical descriptors and one with the image-based features. RF models were

trained with the RandomForestClassifier function from Scikit-Learn[250]. To train the RF models the RandomSearchCV function from scikit-learn was used to select training parameters. Different values were used for the number of trees/estimators (100, 300, 500, 800, 1000, 1200, 1500), the maximum depth (5, 8, 15, 25, 30, 50, 75, 100), the minimum samples split (2, 5, 10, 15, 25,50,75,100) and the minimum samples leaf (1, 2, 5, 10, 15, 20).

3.2.5. Model evaluation

Different model evaluation metrics were used from Scikit-Learn, and the performance of the models was analysed using 5-fold cross-validation, where the different evaluation metrics for each of the folds were averaged to give the overall performance. The Receiver Operating Characteristic - Area Under Curve (AUC), F_1-score and Boltzman Enhanced Discrimination of Receiver Operating Characteristics (BEDROC) score[252] were calculated, with an alpha of 0.2. F_1-score is a metric that is used to evaluate unbalanced classification tasks (although it does not consider true negatives as part of its input)[166], and BEDROC can address the early recognition problem of the actives that the AUC cannot capture[253]. Moreover, the precision and recall were calculated for both classes (actives and inactives) and the balanced accuracy between the two classes.

$$\text{Precision} = \frac{TP}{TP+FP} \quad \text{(Equation 3.4)}$$

$$\text{Recall} = \frac{TP}{TP+FN} \quad \text{(Equation 3.5)}$$

$$\text{F}_1 - \text{score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision+Recall}} \quad \text{(Equation 3. 6)}$$

$$Balanced\ Accuracy = \frac{\left(\frac{TP}{TP+FN}+\frac{TN}{FP+TN}\right)}{2} \quad \text{(Equation 3.7)}$$

where TP denotes true-positives, FP denotes false- positives, TN denotes true-negatives and FN denotes false- negatives.

Finally, y-scrambling[254] was performed in order to evaluate whether the trained models perform better than the y-scrambled models. Y-scrambling was applied by randomly reorganising the bioactivity labels associated with each compound-target pair in the main matrix for each protein for 5 times. Models were rebuilt with the same parameters as the unscrambled models.

3.2.4. Statistical Comparison of Image features and ECFP fingerprints as side information

The Mann Whitney U test, namely the mannwhitneyu python function from scipy package[238] was used to compare the evaluation results from the models trained with image features as side information and the models trained with ECFP fingerprints as side information.

3.3. Results and Discussion

3.3.1 Selection of A:I ratio and side Information compared to no side information models

The effect of A:I ratio in the performance of multitask models was first investigated by training two models with A:I equal to 1:5 and 1:10 (balanced models) and compared with a model that was trained with all the available bioactivity data from ExCAPE (imbalanced model). Overall performance results are shown in Figure 3.5. The balanced models (A:I=1:5, 1:10) yielded overall slightly lower AUC (average across all targets) compared to the unbalanced models regardless of the side information used. For example, the model with ECFP as side information trained with SSS scored an AUC of 0.74, 0.76 and 0.78 when the A:I ratio was equal to 1:5, 1:10 and when all data was used respectively. The reason that the imbalanced models display higher AUC is that they are trained with many more inactive datapoints compared to actives datapoints, and thus the model predicts most of the test points as inactives. Hence, the use of AUC is not ideal for model evaluation because it is a measure which doesn't consider the impact of false positives sufficiently in case of a negative majority class (and hence it is usually not suitable for imbalanced datasets). Still, it was calculated because it has been used in previous studies and to enable comparison of results obtained. That AUC is not a suitable performance measure in the situation encountered here is also supported by the fact that the other metrics (precision and recall for *actives*, balanced accuracy, F_1-score and BEDROC score) indicate a drop in the performance for the unbalanced models compared to the balanced models. For example, the model with ECFP as side information trained with SSS, scored a BEDROC of 0.61, 0.58 and 0.54, when AI ratio was equal to 1:5, 1:10 and when all data were used respectively. Hence, the increasing size of the inactive training data (with a constant number of actives) resulted in a decrease in the recall for actives and, hence the ability of the model to predict the minority (active) class, which is usually however the class of most interest in practice (Figure 3.5d,e). The A:I ratio should be taken into account during the preparation of the main matrix for BMF with Macau as it influences the performance of the classifier and especially influences the ability of the classifier to identify active compounds. In this case, a balanced model has the advantage of performing better in particular for the prediction of the active compounds, but at the cost of a narrower applicability domain (chemical space coverage). As one

attempts to quantify the chemical information contained in the models, the balanced models with A:I ratio equal to 1:5 and 1:10 were trained with 10,302 and 10,824 distinct Murcko scaffolds, whereas the imbalanced model was trained with 13,011 of them. The issue of the A:I ratio has also been investigated in the literature for single-task prediction models, and a ratio of 1:9 – 1:10 was in previous work found to be a suitable ratio for an optimal model performance (though of course generally a trade-off between dataset balance, chemical space coverage, and quantitative performance obtained in a validation setting exists) [241,255].

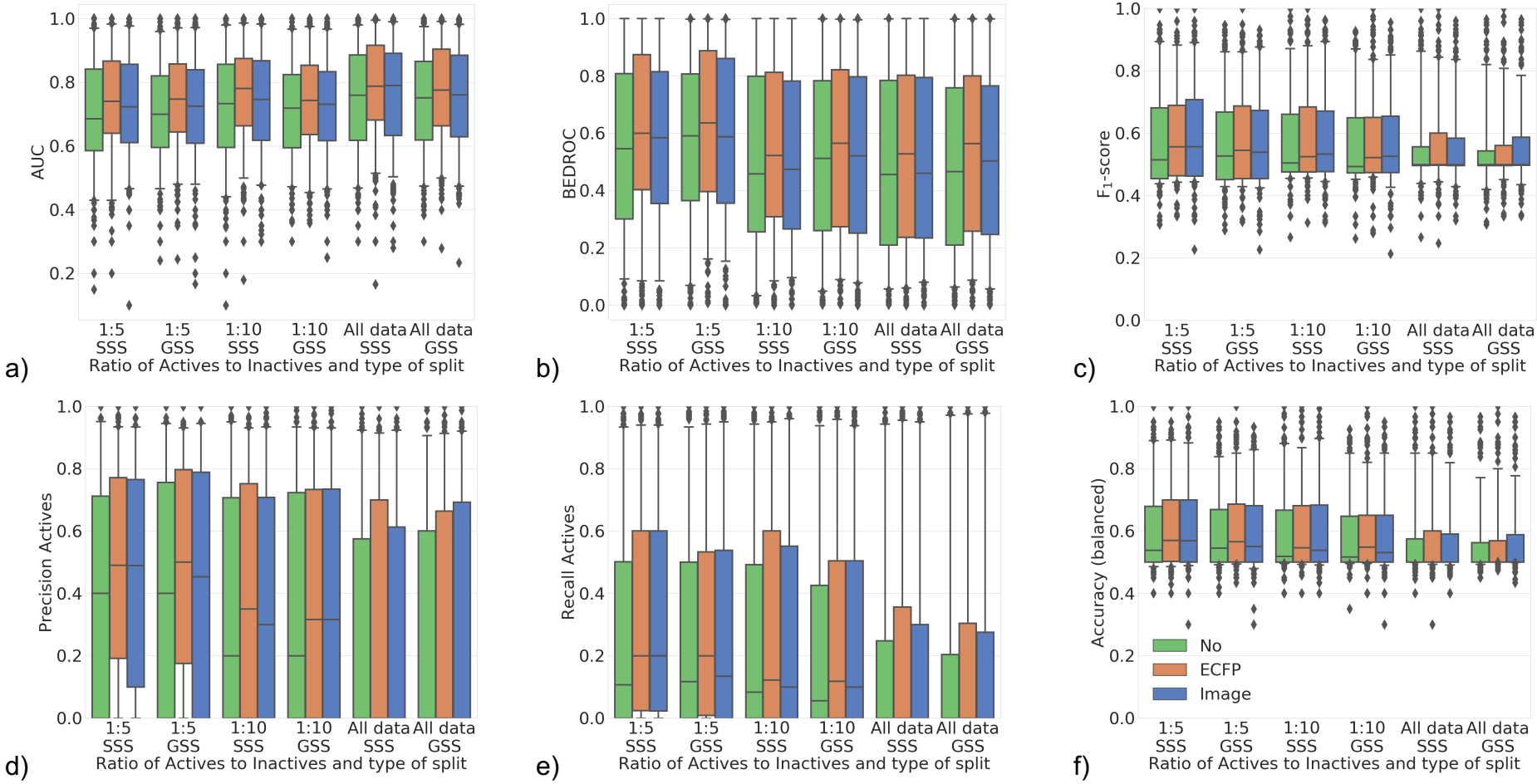


Figure 3.5: Performance of models with no side information (green), ECFP fingerprints (orange) and image- based data (blue) as side information, which were trained with three different ratios of active to inactive datapoints (1:5, 1:10 and all available data) and two different cross validation strategies (SSS and GSS, see methods for details). Each boxplot extends from the 25th (quartile 1) to 75th (quartile 3) percentile, and the median is shown. The outliers outside the 5-95 percentile range are also shown. Side information improves model performance generally, and balanced models (when A:I ratio is equal to 1:5 or 1:10) are better for the prediction of the actives class. Each panel (a-f) corresponds to a different evaluation metric.

To investigate the impact of different types of side information on model performance, a comparison was performed between the models with no side information and the models trained with ECFP and image data as side information. All performance metrics improved when either side information is used to train the models, compared to no-side information models (Figure 3.5a-f). The positive effect of using ECFP side information has been observed in the literature,[157] and in this work, the image-based data can improve the performance of a molecular classification model. For example, the models with A:I ratio equal to 1:10 with SSS scored a BEDROC of 0.51, 0.58 and 0.53 when no side information, ECFP as side information, and image as side information were used for this purpose, respectively. Moreover, the Mann-Whitney U test was used to identify whether there is a statistically significant difference between the average AUC across all targets between no side information and ECFP and Image data as a side information. When ECFP fingerprints were used as side information, they always showed a statistically significant difference in the AUC with a Mann-Whitney U p-values less than 0.05 across all the parameters tested (ranging from 0.002 to 0.047). On the other hand, that was not always the case with the image data with a Mann-Whitney U p–values ranging from 0.036 to 0.217. Although a statistically significant difference was not observed between the model performance when using no side information and image data for all the trained models (different A:I ratios and splits), the inclusion of image data always increased the precision and recall for actives. For example, only 33.5% and 32.14% targets had a precision and recall equal to zero (A:I ratio = 1:5 and stratified split), compared to 39.29% targets when no side information was used. Hence, the addition of side information enhanced the prediction of actives in particular, which is important since many of the models before suffered from rather low recall (see first section of results). Therefore, models with side information outperformed the models without side information generally, and in particular with respect to the retrieval of active compounds.

Finally in this section, it was evaluated whether the models perform better than random models by applying y-scrambling 5 times. The un-scrambled models perform better than the y-scrambled models, which scored mean AUC across all targets in the range of 0.49-0.51 (appendix, Table 8.2). Hence, the models are not obtained by chance.

3.3.2. Comparison of BMF Macau with Random Forest models

Furthermore, the performance of BMF Macau was compared to a widely used training algorithm, namely Random Forests (RF), by using ECFP and image-based data for compounds representation. The results of this analysis are shown in Figure 3.6a and b respectively for the models with A:I ratio equal to 1:10 and for the Group-Shuffle-Split (GSS). For RF, models were trained separately for each target (224 targets in total) and then results were averaged (by calculating the mean) across all targets in order to compare with the BMF Macau model.

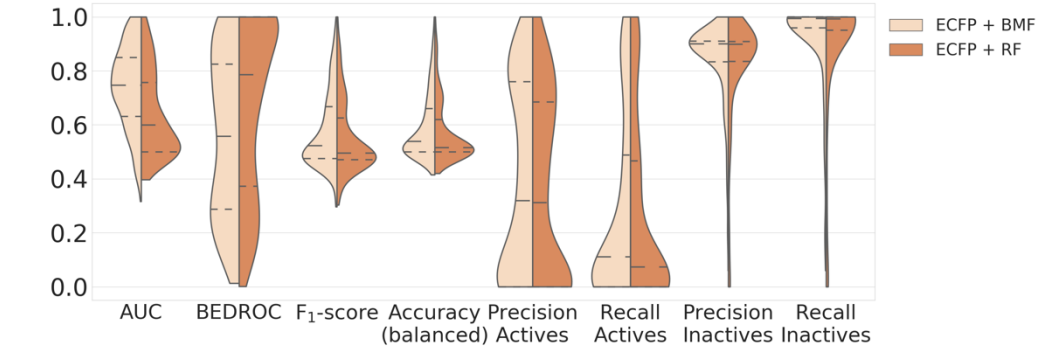
RF and BMF Macau yielded mean AUCs of 0.64 and 0.74 and F_1-scores of 0.55 and 0.58 respectively when ECFP fingerprints were used as a compound representation. In addition, 56.25% of the targets were better predicted by BMF Macau with 20.10% of the targets with an F_1-score difference greater or equal to 0.1, whereas 38.39% of the targets were better predicted by RF with 9.82% of the targets with an F_1-score difference greater or equal to 0.1. Moreover, when the image data were used as compound descriptors, the RF and BMF Macau yielded mean AUCs of 0.56 and 0.73 and F_1-scores of 0.48 and 0.57 respectively. In addition, 71.00% of the targets were better predicted by BMF Macau with 34.38% of the targets with an F_1-score difference greater or equal to 0.1, whereas 23.66% of the targets were better predicted by RF with 1.79% of the targets with an F_1-score difference greater or equal to 0.1 when image data was used as compounds information. Therefore, by taking into account AUC and F1-score, BMF Macau and RF performed overall similar when ECFP fingerprints were used as compound descriptors, and that BMF Macau outperformed RF by a rather large margin when image-based data is used to represent structures.

In addition, Figure 3.6 shows other metrics such as BEDROC score, accuracy and precision and recall reported separately for each class. All the models have high precision and recall for the inactives (mean precision and recall of 0.80 ± 0.04 and 0.88 ± 0.03 across both training methods and compounds representations) and a very low precision and recall for the actives (mean precision and recall of 0.33 ± 0.08 and 0.25 ± 0.03 across both training methods and descriptors used). When ECFP were used as descriptors, the mean precision for actives was equal to 0.36 (increased 3%)

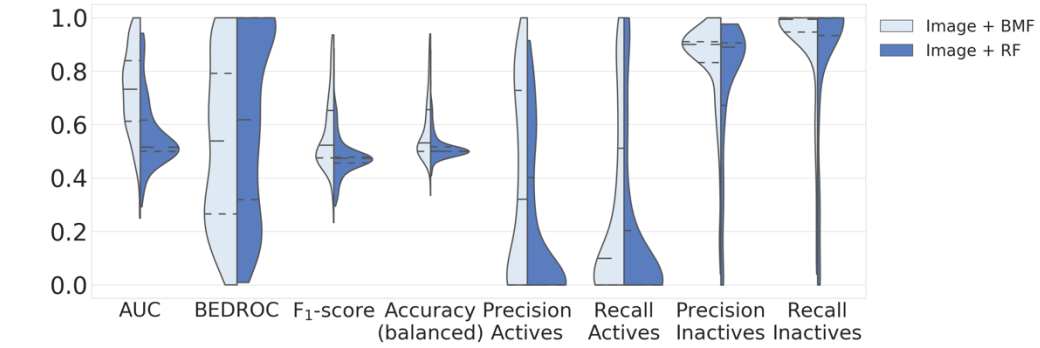
and 0.39 (increased 6%) for RF and BMF Macau respectively, and the recall for actives was equal to a mean of 0.27 (increased 2%) for both algorithms used (Figure 3.6a). When image-data was used as a descriptor, the precision was equal to 0.20 (decreased 13%) and 0.39 (increased 6%), and the recall for actives was equal to a mean of 0.19 (decreased 6%) and 0.27 (increased 2%) for RF and BMF Macau respectively (Figure 3.6b). Those results indicate that both algorithms are similarly challenged to identify compounds from the active minority class when using ECFP fingerprints; however, when image data were used as side information the BMF algorithm performs better in the identification of actives compounds. RF is outperforming BMF Macau only in terms of BEDROC score with a mean of 0.54 and 0.61 respectively across both compounds' representations. Therefore, RF models can rank earlier in the predicted list the active compounds compared to BMF Macau but there is no improvement in terms of how the compounds are classified into the two classes, which is reflected by the other metrics.

To summarise, when ECFP fingerprints are used as compounds' information, the AUC score is higher for BMF Macau compared to RF, BEDROC is higher for RF compared to BMF Macau, while accuracy, F_1-score, precision and recall are similar for both algorithms and thus they perform similarly. On the other hand, when the image-based data is used as compound information, BMF Macau outperforms RF with respect to all the metrics used (except BEDROC). In a comparison of BMF Macau with RF in the literature with ECFP descriptors, a similar performance between the two algorithms was observed in the prediction of bioactivity with data extracted from ChEMBL[251].

In this, section the results for the models trained with A:I ratio equal to 1: 10 and using GSS were discussed. Similar results were also observed with the models trained using other setups when i) A:I ratio was equal to 1:5 with SSS and GSS ii) all the available bioactivity data used with SSS and GSS and iii) A:I ratio was equal to 1:10 with SSS, details of which can be found in appendix in Figure 8.3 - Figure 8.7.



a)



b)

Figure 3.6: Performance of BMF Macau and RF models trained with a) ECFP and b) image-based data as side information with GSS and A:I ratio equal to 1:10 across 224 targets. The dashed lines represent the 25th (quartile 1) to 75th (quartile 3) percentile, and the median of the results distribution and it can be seen that when ECFP were used as compounds' descriptors, the AUC score is higher for BMF Macau compared to RF and the BEDROC score is higher for RF compared to BMF Macau. Still, accuracy, F_1-score, precision and recall were similar for both algorithms. On the other hand, when the image-based data was used as compounds' side information, BMF Macau outperformed RF and that is shown with the all the evaluation metrics used.

3.3.3. Impact on model performance when ECFP descriptors and Image data are used as side information and comparison with the literature

Next it was investigated which type of side information leads to better performance in compound classification. To this end, the results obtained with the models with A:I ratio equal to 1:10 were analysed (see appendix, Table 8.3 and Figure 8.8), which were performing better in identifying actives compared to the unbalanced models (according to increased recall), while at the same time possessing larger chemical space coverage compared to the models with A:I ratio equal to 1:5. The two types of side information compared led to mean AUCs of 0.76 and 0.74 using ECFP

fingerprints and image data as side information, respectively, when the SSS was applied, and mean AUCs of 0.74 and 0.73 when the GSS was used. Both types of side information hence performed similarly well overall, and more than 50% of the targets were predicted with an AUC greater than 0.7 when the image or ECFP data was used. In addition to AUC, other metrics are reported in Figure 8.8, and as in the case of AUC the performance based on BEDROC, accuracy, F_1-score, precision and recall were similar for both types of side information. The difference between the two types of side information was tested by a Mann Whitney test across these 224 AUC values, where the null hypothesis was that the two methods have identical performance. The result was that no statistically significant difference between the two different types of side information was found (Mann-Whitney U p-value > 0.05 for both SSS and GSS). Hence, both types of side information overall contribute similar amounts of information to used can be an asset in bioactivity prediction models.

Moreover, the performance of the models trained with image data as side information was compared with models in the literature. The models trained in this work with image data performed equally well as models described in the literature such as models trained on public bioactivity data from ChEMBL with the same image-based data as descriptors and with ResNet (Residual neural Network) as the training algorithm[166]. These models scored an AUC of 0.73, whereas our models scored AUCs ranging from 0.73 to 0.76, when the image data were used.

3.3.4. Understanding biological differences between targets better predicted by image-based data and ECFP descriptors as side information, compared to each other and a no side information baseline

Moreover, the difference in the predictive performance of image-based data and ECFP descriptors as side information to each other was evaluated, as well as to the no side information (baseline) model, the results of which are shown in Figure 3.7a and b. When SSS was applied (Figure 3.7a), and ECFP fingerprints were used as side information, this improved the performance of 126 targets (56.25%) (based on F_1-score) compared to the baseline model with 14 targets exhibiting a difference in F_1-score greater than or equal to 0.1. When image-based data was used as side

information, this increased the performance of 101 targets (45.01 %) compared to the baseline model and 8 targets showed a difference in F_1-score greater than or equal to 0.1. As a result, there were only 35 targets (15.63 %) whose performance was not improved by the addition of either type of side information. Similar results were obtained when the GSS was applied (Figure 3.7b). As a result, for the large majority of targets (about 85%) the addition of side information improved model performance. However, only 31.25% of targets with improved performance were in common between both types of descriptors, indicating that chemical structure and image-based readouts confer to a good extent complementary information.

In addition, it was evaluated whether the difference in the predictive performance of models with different side information can be related to the modelled target. When SSS was applied, 70 targets were better predicted with image data as side information was used, out of which 13 targets showed a difference in F_1-score greater than or equal to 0.1 (Figure 3.7a). Moreover, 103 targets were better predicted when ECFP descriptors were used, and 18 targets showed a difference in F_1-score greater than or equal to 0.1. Similar results were achieved with the group shuffle split strategy (Figure 3.6b). Therefore, there were targets that were better predicted by image-based data as side information and others by ECFP fingerprints as side information. Detailed tables of pairwise comparisons of ECFP and image data as side information and baseline models performance is provided in appendix Tables 8.4-8.6.

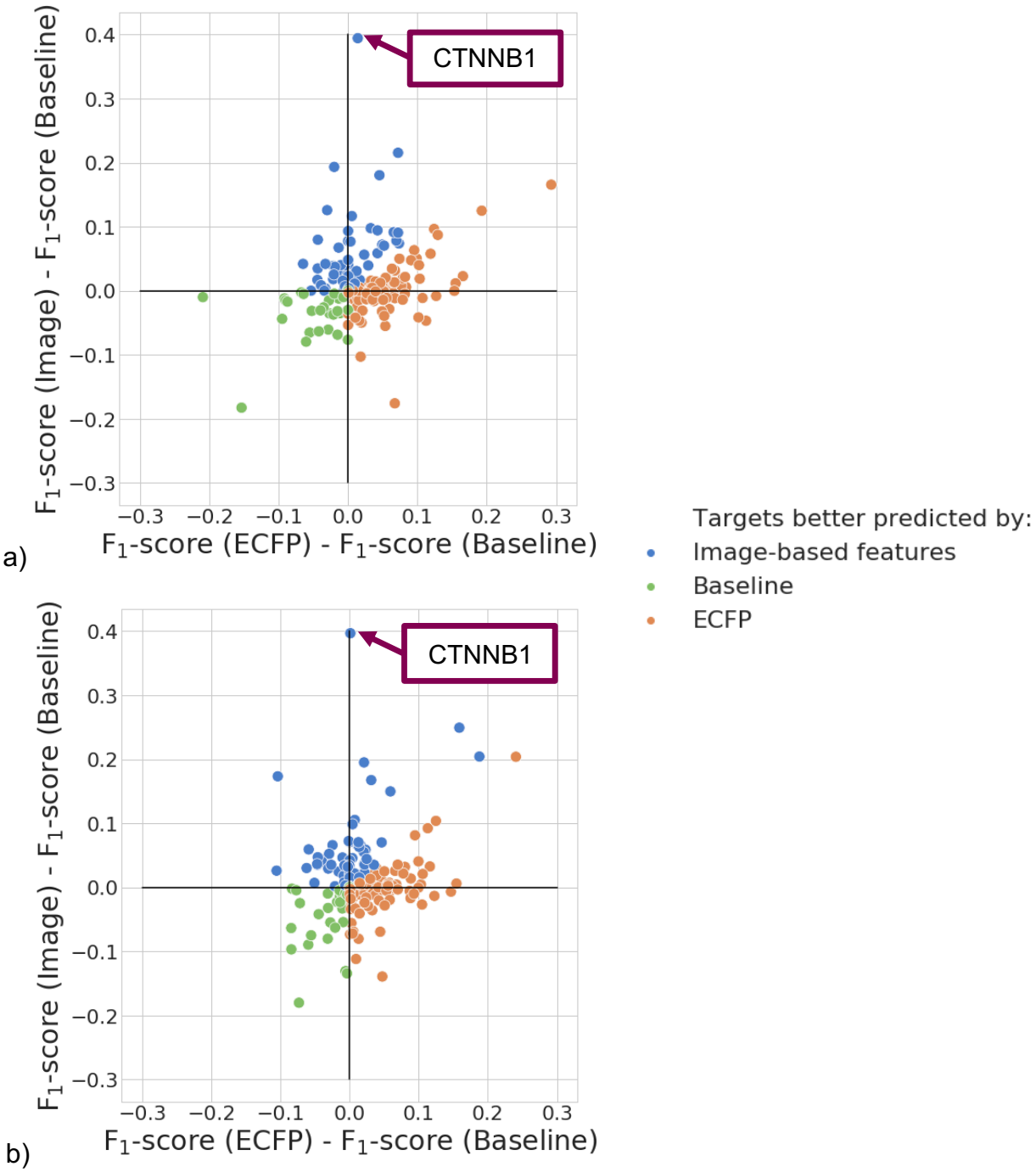


Figure 3.7: Comparison of model performance using ECFP as side information, using Image data as side information and the model using no side information (baseline) overall tasks for the (a) SSS, b) GSS. It can be seen that the majority of targets were either better predicted using either ECFP or image data as side information or both. Image-based data as side information outperformed both ECFP as side information and baseline model (no side information) for the bioactivity prediction of β-catenin (CTNNB1).

There are five targets that were better predicted by image-based features as side information and ouperformed both ECFP as side information and the baseline (no side information) for both splits (SSS and GSS) with a difference in F_1-score greater than or equal to 0.1. These targets are the CTNNB1 (β-catenin), APOBEC3G (a cytidine deaminase), NOD1 (Nucleotide-binding oligomerisation domain-containing protein 1) and two serine/threonine protein kinases: RIPK2 (Receptor-interacting protein kinase 2) and CSNK1D (Casein kinase 1 isoform delta). The β-catenin protein was most significantly better predicted by image data as side information, outperforming both ECFP descriptors as side information and the baseline by a margin of 0.40 F_1-score, as indicated in Figure 3.7. β-catenin is encoded by the CTNNB1 gene, with key roles in the transduction of Wnt signalling pathway, with pivotal roles in carcinogenesis and tumour progression in colon, pancreatic, lung and ovarian cancers[256] and therefore an interesting target for drug discovery[257]. Hence, a question was formed about why the cell morphology data were better predicting β-catenin bioactivity, to better understand the associated changes in model performance. To this end, a Principal Component Analysis (PCA) was performed on image-based feature space which revealed a separation of active compounds from the inactive compounds in the first two principal components, as shown in Figure 3.8a. It can be seen that image data are able to distinguish between active and inactive compounds based on the cell morphology of compounds active on the β-catenin target. To quantify this observation, the intra-class (active vs inactive compounds) and the inter-class (active vs inactive compounds) Tanimoto similarity and Pearson correlation in the image (Figure 3.8b) and ECFP chemical fingerprint (Figure 3.8c) space respectively for β-catenin ligands were evaluated. A high intra-class correlation of the active compounds was observed in the image space with a median of 0.63, compared to a low intra-class Tanimoto similarity in the ECFP space with a median of 0.14, meaning active compounds were significantly more similar to each other in image space, compared to fingerprint space. Furthermore, a high difference was observed in the interclass correlations (between actives and inactives), with medians of -0.09 for image space, and 0.14 for ECFP fingerprint space, respectively. Overall, this means that active compounds on this target are clearly different from inactive compounds in image space (mean similarities of 0.63 vs -0.09), while being virtually indistinguishable in ECFP fingerprint space

(mean similarities of 0.14 vs 0.14). Taken together, this similarity analysis explains why adding image-based features as side information increases model performance. Upon further analysis of β-catenin training data, the large majority (about 99%) of training datapoints are derived from the PubChem BioAssay 1665[258], a high throughput imaging assay identifying the translocation of β-catenin *via* immunostaining of endogenous β-catenin. While hence the type of readout used in both cases is from cellular imaging, this particular assay endpoint uses a *specific* stain for beta-catenin, which the (more generic) Cell Painting does not consider. Assuming that this generally holds, this would mean that generic morphological changes are able to predict more specific morphological changes, and that hence in particular cellular endpoints which involve changes in cell morphology would benefit from using Cell Painting features as side information in mode of action analysis settings.

Furthermore, the intra-class and inter-class similarity was also performed for the other four targets, which are mentioned above and are better predicted by image-based features as side information and a similar trend was observed (Figure 8.10). Upon further analysis of the training data that were used for four out of five targets we observed that the times that image features improve predictions are when the majority of the training data are derived from assays that used fluorescent detection techniques. For one of the targets, the assay information could not be traced (see appendix, Table 8.7).

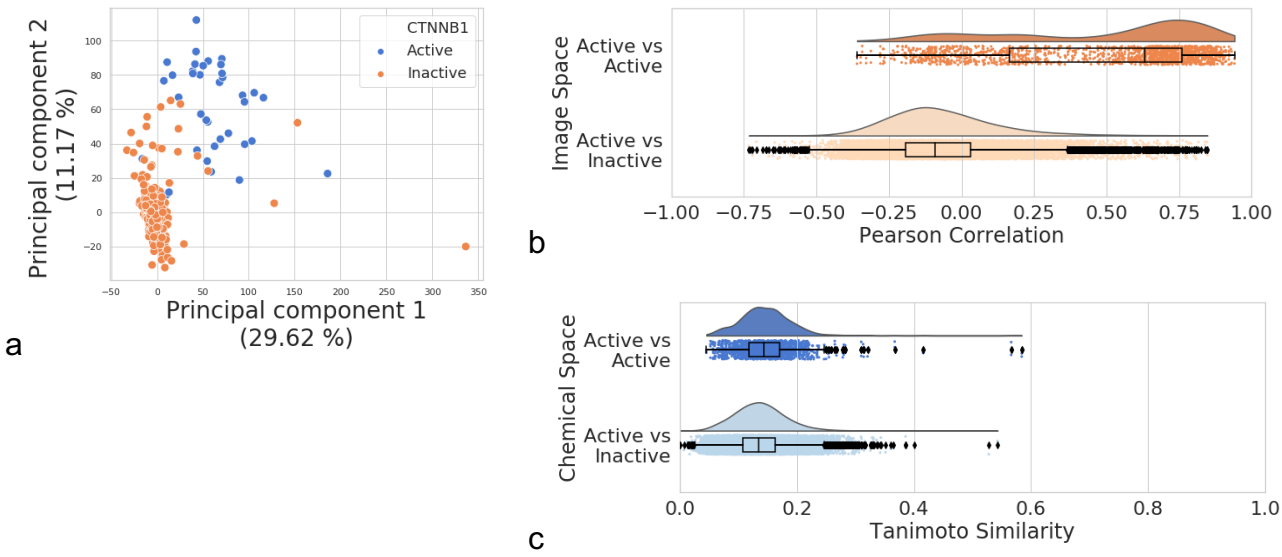


Figure 3.8: a) Principal component (PC) plot of PC 1 vs PC 2 for active (blue) and inactive (orange) compounds for β-catenin in image-based feature space. b) Pearson correlation of the image-based features and c) Tanimoto coefficient similarity in the chemical space for β-catenin ligands. It can be seen in (b) that active compounds for the β-catenin target are more similar in image space to each other (and dissimilar compared to inactives), neither of which is the case in fingerprint space in (c), where both distributions overlap.

The performance of image-based data and ECFP fingerprints as side information for predicting ligands of the GPCR1 family was further evaluated and the results are shown in appendix (Figure 8.9). It can be seen that ligands for proteins of the GPCR1 family tend to be better predicted by ECFP data as side information. Moreover, the performance of the ECFP descriptors as side information drops when the GSS split is used and also the difference in the ECFP and image F_1- score is also smaller (Figure 8.9b). This shows that when chemical scaffolds are used to separate the compounds in train and test set, it is more challenging for the ECFP model to predict accurately. This intuitively makes sense and underlines the general finding that image-based features as side information increase performance in cases where ligands are chemically rather diverse. Example of targets in the GPCR1 family that are better predicted by ECFP as side information, compared to image-based data as side information, are the β-adrenergic receptors (ADRB) 1, 2 and 3. The intra-class

similarity of active compounds was higher than their inter-class similarity in the chemical space, compared to the intra and inter class similarity in the image space (see appendix, Figure 8.11).

3.4 Conclusion

In this work, a target prediction with Bayesian Matrix Factorization Macau was performed and different training parameters were investigated in order to compare the performance of image-based and chemical features as side information over a range of 224 targets. The conclusions are based on the current cell morphology information available in the public domain and their overlap with bioactivity databases and it should be acknowledged that it is a relative comparison of information content and performance. Comparison of different A:I ratios in the preparation of the dataset showed that the balanced models performed better than the unbalanced and especially for the prediction of active compounds. Moreover, the BMF Macau and RF performed very similarly when ECFP fingerprints were used as compound descriptors, but BMF Macau outperformed RF when image-based data was used as compounds' information. In addition, overall (across all the targets) ECFP performed significantly better than the no side information model and that image data was sometimes better than the no side information model. Across all targets, both image and chemical data performed similar in predicting the bioactivity of compounds when used as side information in a BMF model. However, the two types of information performed rather differently on an individual-target level, and thus there are targets that are better predicted by ECFP and targets better predicted by image data. Therefore, the main conclusion of the current study is that image-based data can be a useful source of information for bioactivity prediction which is in many cases complementary to ligand structural information.

4. Chapter 4: Mitochondrial Toxicity Prediction Using Cell Painting Assay on a PROTACs Dataset

4.1. Introduction

Mitochondria play an essential role in the regulation of cellular functions including Adenosime TriPhosphate (ATP) generation, metabolic control, signal transduction, immune response, and apoptosis[259]. Drug-induced mitochondrial toxicity (mitotoxicity) affects many organs such as the liver, heart, kidney, skeletal muscle, and brain[260,261]. Therefore, drug-induced mitotoxicity is a concern during drug discovery process, as first highlighted by Dykens and Will in 2007[262] as a mode of toxicity to late-stage market withdrawals (38 withdrawals between 1994-2006 related to hepatotoxicity and cardiotoxicity) or Black Box warnings. Mitotoxicity has been related to attrition of drugs such as cerivastatin, phenformin, troglitazone and nefazodone and is related to diverse drug classes such as antidiabetics, antilipidemics, antivirals, antibiotics and antidepressants[259,261].

Drug-mediated mitotoxicity is a consequence of different direct or indirect mechanisms as shown in Figure 4.1. A common cause of mitotoxicity is the uncoupling of electron transport chain from ATP generation[263,264]. This in turn increases concentration of reactive oxygen species (ROS), which induce oxidative damage to mitochondrial DNA (mtDNA) and proteins. For example, troglitazone, an antidiabetic drug belonging in the thiazolidinedione class, was withdrawn in 2000 due to lethal hepatotoxicity caused by an off-target effect on the mitochondrial electron transport chain[259,265]. Other indirect mechanisms of mitochondrial toxicity include ROS generation via calcium signalling, wherein calcium ions are passed from the endoplasmic reticulum (ER) to mitochondria “quasi-synaptically”[264]. Calcium promotes ATP synthesis by stimulating ATP synthase and Krebs cycle enzymes[266], and may thus stimulate increased mitochondrial metabolic rate, oxygen consumption, and mitochondrial ROS production. Other indirect mechanisms include irreversible opening of the mitochondrial permeability transition pore, inhibition of fatty acid oxidation and impairment of mtDNA replication or mtDNA-encoded protein synthesis[261,262,264]. For example, the antiviral drug zidovudine disrupts mtDNA replication by inhibiting mDNA polymerase-γ[259]. Evidently,

mitotoxicity is a safety concern that must be addressed early in order to increase the success rate of safe candidates reaching the clinic.

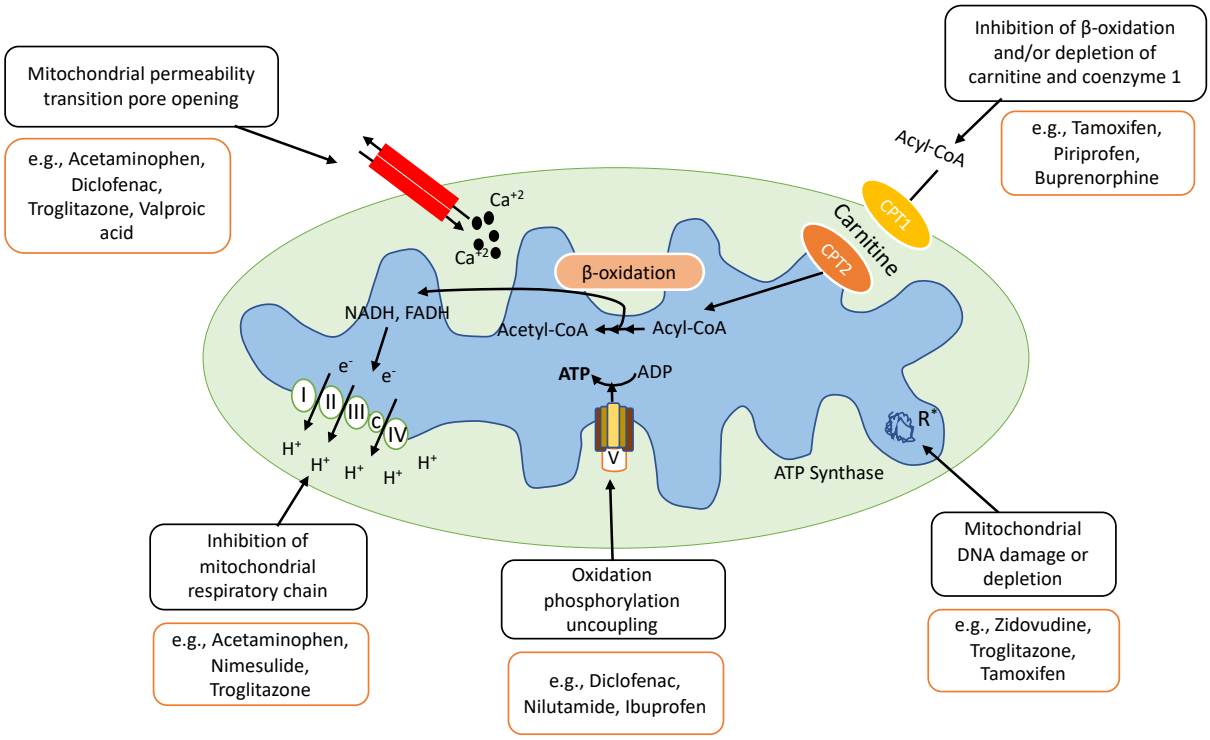


Figure 4.1: Schematic illustration of major mechanisms of Mitochondrial toxicity. Drugs can cause mitochondrial toxicity via multiple mechanisms such as: 1) inhibition of mitochondrial respiratory chain, 2) alteration in mitochondrial permeability transition pore, 3) oxidation phosphorylation uncoupling, 4) mitochondrial DNA damage or depletion and 5) inhibition of β-oxidation and/or depletion of carnitine and coenzyme 1. Examples of drugs that cause mitochondrial toxicity via these mechanisms are provided.

Numerous efforts exist to evaluate or predict small molecule's mitotoxicity and different assays have been developed capturing various mechanisms of drug-induced mitotoxicity[267] (as shown in Table 4.1). An assay that is routinely used to evaluate mitotoxicity is the Glu/Gal assay that uses cells grown in two different media: a high glucose- and galactose- media. Cells grown in high glucose-containing medium use glycolysis for ATP generation and are resistant to mitochondrial insult. On the other hand, cells grown in galactose-containing medium rely almost exclusively on mitochondria for their ATP production and, hence, are very sensitive to mitochondrial insult[268]. However, Hynes et al.[269] showed that the Glu/Gal assay only detects about 2 - 5% of all mitotoxicants, which further highlights the reality that most compounds that cause organ toxicity do so via multiple off-target mechanisms. Another assay, that

is investigating the mitotoxicity differently and not on the ATP production is the Respiratory Screening Technology (RST), which measures mitochondrial respiration in the form of oxygen consumption in isolated rat liver mitochondrial using a time-resolved fluorescent sensitive pore. However, oxygen consumption is only one indication of mitotoxicity, and other supplementary assays could be used. For example, other assays are the mitochondrial mass assay and the mitochondrial membrane potential (MMP) assays. The former, measures the mitochondrial mass[270] or the number of mitochondria in the cell because an increased mitochondrial mass can occur as a result of an adaptive response by the cell to increase energy production. The latter is based on the mitochondrial electron transport chain, which creates an electrochemical gradient through a series of redox reactions. This gradient results in the synthesis of ATP and generates the MMP, which is a key parameter for evaluating mitochondrial function[271].

Table 4.1: Summary of mitochondrial toxicity assays with their description and comments related to their strengths and limitations.

Assay	Description	Comments
Glucose/Galactose (Glu/Gal)	• Utilises HepG2 cells, which are grown in high glucose or galactose media[267] • Usually preferred for early assessment of hundreds or even thousands of compounds[267]	• A limitation is that Glu/Gal assay only detects about 2-5% of all mitotoxicants, which highlights that many compounds that cause organ toxicity do so *via* multiple off-target mechanisms[269] • Not every cell line is amendable to culture in galactose so there is a need for additional assays[267]
Respiratory Screening Technology (RST)	• Measures the effects of compounds on mitochondrial oxygen consumption using soluble oxygen sensors and time-resolved fluorescence[268]	• Distinguishes between inhibition and uncoupling of oxidative phosphorylation[267]
Mitochondrial Membrane Potential	• Fluorescent dyes are used with the majority being cations that distribute to the mitochondrial matrix[271]	• Does not distinguish compounds which inhibit mitochondrial respiration from those than uncouple electron transport from ATP synthesis[271]
Mitochondrial Mass	• Appropriate dyes are used to preferentially target mitochondria regardless of mitochondrial membrane potential this making them suitable tools for determining mitochondrial mass[270]	• Mitochondrial stress, including oxidative stress, can initiate mitochondrial mass upregulation[272]

Therefore, annotations from the above assays are usually used as endpoints to computationally predict them with *in-silico* predictive models in order to minimise the number of compounds that will be tested with *in-vitro*. There have been machine learning approaches for the prediction of mitotoxicity by using physiochemical descriptors, structural alerts and hight throughput imaging data for small molecules[273–275]. However, computational prediction for new data modalities is less investigated. One modality that is currently under evolution are the PROteolysis TArgeting Chimeras (PROTACs). As a new therapeutic modality, they are raising multiple concerns on various aspects such as safety, ADME properties, toxicity, etc[205]. The work described in this chapter explored mitochondrial toxicity prediction, which is one of many possible avenues for safety evaluation of novel compounds. The prediction and/or better understanding of these aspects is additionally limited by the lack of descriptors and methodologies for robust safety profiling. PROTACs belong to a category of compounds also referred to as beyond the Rule-of-5 (bRo5) as they do not comply with the Lipinski's Rule-of-5 (Ro5). Hence, there is a need for descriptors tailored or 'compatible' with the bRo5 new data modalities[221,222].

A potential approach to profile PROTACs and consequently to better understand their safety aspects could be the use of high throughout imaging assays, which have become easier to run over the recent years. HTI assays have not only shown utility in gaining a better understanding of compounds' MoA[105,106,166,276,277] but have also been used to predict a wide range of efficacy and safety endpoints[51,103,278]. For example, Persson et al.[279] used a multi-parametric imaging of cell health to predict human hepatotoxicity and elucidate toxicity mechanisms. They used imaging of bile salt transport inhibition in hepatocytes to predict cholestasis inducing compounds and imaging of micronuclei to detect genotoxicity. Both adverse effects are crucial to detect but unfortunately cannot be readily detected in animal models during preclinical testing, whereas HTI assays were able to. In addition, a high content assay was used by Grimm et al.[280] to evaluate compounds' cardiotoxicity and hepatotoxicity. This was conducted by performing high content imaging on a) cardiomyocytes, which determined cell viability, mitochondrial integrity and ROS formation and on b) hepatocytes, which determined cytotoxicity, cytoskeletal integrity, mitochondrial

integrity and lipid accumulation (as a marker for hepatic steatosis). Moreover, cell morphology and cell proliferation markers have been proved useful in the identification of cytotoxic compounds[278]. A staining and image analysis protocol was applied to a novel chemical library consisting of 329 chemically diverse compounds. Using the changes observed in nuclear morphology, cell shape, and proliferation, they were able to identify compounds with adverse cellular effects such as cell loss or sublethal alteration to cell morphology or cell proliferation with hit rates of 10.0% and 3.6%, respectively. Results further suggested that cell morphology and, in particular, nuclear morphology can be used to identify adverse cellular effects. Considering the studies above, HTI profiling can be an asset in small molecule safety evaluation and therefore could also be useful for PROTACs.

One of the assays that is currently used by academic groups and pharmaceutical companies is the Cell Painting assay[50,104]. Phenotypes from this assay are target agnostic and unbiased and can therefore be considered as image-based fingerprints of a compound covering a wide range of information. Hence, such data can reveal any cell morphological characteristics upon perturbation and therefore readouts have a general nature, being particularly popular in compound MoA and toxicology research[103,277]. Seal et al.[103] compared cell morphology descriptors from the Cell Painting assay and molecular fingerprints (Morgan and ErG), separately and in combination, for the prediction of cytotoxicity and proliferation related in vitro assay endpoints. Cell Painting descriptors proved to be sufficient to train models with good predictive performance and combined with the molecular fingerprints resulted in better models compared to those using only molecular fingerprints. Feature importance analysis showed that CellProfiler features related to nuclei texture, granularity of cells, and cytoplasm as well as cell neighbours and radial distributions were identified to be most contributing, which is plausible given the endpoint considered. Finally in another approach, image-based data from a live-cell imaging assay were used in order to catalogue the morphological phenotypes of 1,008 reference compounds, which were well annotated and their MoA was known [281]. The image data were profiled at four concentrations and in 15 reporter cell lines. Features were calculated with Columbus image analysis software from 12 different fluorescently labelled proteins from their

endogenous chromosomal loci, which enables the monitoring of cell organelle morphology and the activity of various signalling pathways. Image features were projected onto a 2-Dimensional t-distributed Stochastic Neighbour Embedding (t-SNE) plot, which showed clusters of compounds which cause mitochondrial toxicity. Therefore, the studies above highlight that image-based data can be used in both supervised and unsupervised ML approaches and provide information for the safety assessment of compounds such as mitochondrial toxicity.

Considering, the evidence in the literature that HTI assays and in particular Cell Painting assay can be useful in better understanding and predicting safety endpoints, the hypothesis of this work is that the assay could be applied to PROTACs. Therefore, the first novelty of this work was that Cell Painting assay was applied to a compound dataset which included PROTAC compounds, and it was further evaluated whether the profiling of PROTACs with the Cell Painting assay was successful. The second novelty was the use of Cell Painting readouts, which were chosen due to their current interest as outlined above, for the computational prediction of mitochondrial toxicity for PROTACs.

4.2. Methods

4.2.1. Workflow Summary

The study workflow can be divided into four main steps (see Figure 4.2). PROTAC and non-PROTAC compounds were profiled with the Cell Painting assay in U2OS cells (step 1). A total of 341 and 149 PROTAC and non-PROTAC compounds respectively were profiled at three concentrations (0.1, 1 and 10 μM). The non-PROTAC compounds included small-molecule compounds, which were inhibitors of the targets that PROTACs are degrading, E3 ligase ligands and reference compounds that have been tested in Glu/Gal mitochondrial toxicity assay in AstraZeneca. Following the compounds' profiling with the Cell Painting assay, morphological features were calculated with CellProfiler[107] (step 2). The profiling of compounds with the Cell Painting assay and the calculation of morphological features with the CellProfiler was performed by members from the High Throughput Screening team in AstraZeneca. Morphological features were normalised, and feature selection process was applied (step 3). In the final step (4) activity of PROTACs on Cell Painting assay was evaluated and CellProfiler features were used as descriptors for training of mitochondrial toxicity models.

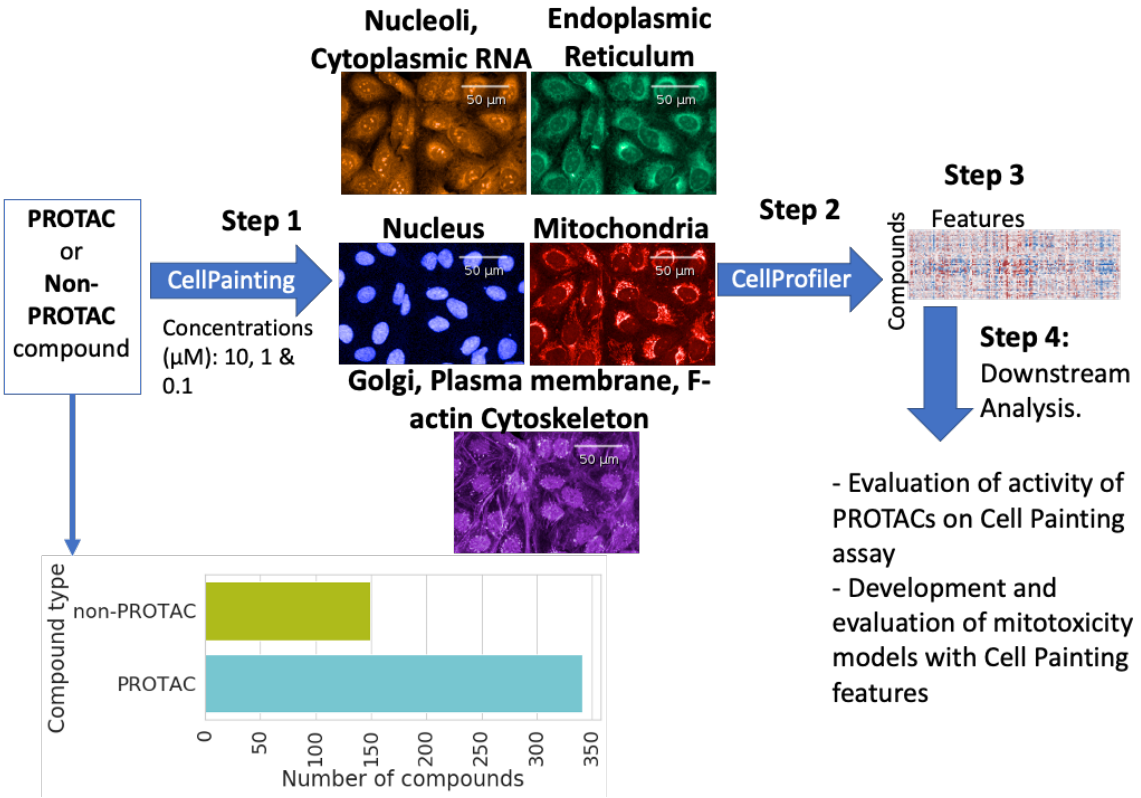


Figure 4.2: Summary of the analysis performed in this work. PROTACs and non-PROTACs compounds were profiled with the Cell Painting assay followed by data normalisation and a downstream analysis.

4.2.2. Data Curation and Normalisation

A normalisation process was applied as described by Way et al.[282] and the reason is that HTS experiments can be affected by systematic row, column and edge effects[283] and thus there is a need for data normalisation in order to reduce false positives in the HTS experiments[284,285]. Firstly, single cell data per well were merged by calculating their median value. Next, data were normalised using the median and the median absolute deviation (MAD) of feature values from empty wells (DMSO) as the centre and scale parameters respectively. We normalised all perturbation profiles by subtracting the centre (median) and dividing by the scale (MAD) and did for each plate individually in Equation 4.1.

$$Z_i = \frac{x_i - \overline{x_{DMSO}}}{x_{MAD}} \quad \text{(Equation 4.1)},$$

where Z_i is the normalised value for feature i, x_i is the raw feature value for feature i, $\overline{x_{DMSO}}$ is the median value of this feature across DMSO wells and x_{MAD} is the mean absolute deviation value of this feature across DMSO wells.

4.2.3. Feature Selection

A Feature Selection was performed to remove features based on a set of criteria. The initial number of features was equal to 3,575. The first criterion was the variance of the features across profiles and hence two features with variance less than one were removed. In addition, features with a high standard deviation were filtered out (75 features) and we used a standard deviation threshold equal to 15. According to Way et al.[282], features with a high standard deviation after normalisation are considered as feature outliers and should be removed. In addition, features with missing values in any profile were filtered out (210 features). Moreover, pairwise correlations were calculated for all the features and randomly removed 1 feature from each pair with a Pearson correlation greater than or equal to 0.9 and 2,564 features were filtered out. A list of 55 ‘block listed’ features were removed because they have been described as problematic by Way et al.[282]. Finally, the data consisted of 669 features.

4.2.4. Evaluation of PROTACs activity on Cell Painting assay

Two different methodologies were used to evaluate whether compounds (PROTACs and non-PROTACs) were active on the Cell Painting assay screen. The first one was a Euclidean distance-based approach and the second is the calculation of grit score. The first approach was described by Cox et al.[51], and it was used it in order to calculate which compounds were "active" on the assay using a 95th percentile cut-off on the null distribution of Euclidean distances between individual DMSO control profiles and the mean DMSO control profile.

In addition, the grit score[286,287] was used, which captures the phenotypic strength of a perturbation in a profiling experiment and combines two concepts. The first is the replicate reproducibility and the second is the difference from the DMSO control. Firstly, for each target profile (i.e., compound) pairwise Pearson correlations were calculated for both compound replicates and control replicates. Hence, the pairwise correlations form two distinct distributions (replicate and control) were obtained. Then using the control profiles only, a z-score transform was obtained, which was then used to transform the compounds' replicates. The mean of compounds' replicates z-scores were calculated, and this was the final score termed grit score. Since grit is based on z-scores, the magnitude can be easily compared between perturbations and is a directly interpretable value. For example, a grit score of 3 for a PROTACs X compared to a neutral control means that on average PROTACs X is 3 standard deviations more similar to replicates than to DMSO controls. Therefore, it is considered as the PROTACs' average reproducibility with respect to the neutral control similarity. Grit score was calculated with cytominer-eval Python package[287], developed by Broad Institute.

4.2.5. Glu/Gal Assay for mitochondrial toxicity assessment

Results from the Glu/Gal assay were used to label the compounds in the dataset. This assay assesses potential test substances that can trigger mitochondrial dysfunction. HepG2 cells are cultured in a) glucose containing and b) galactose containing media and are exposed for 24 h to the test compounds. Following treatment, the IC_{50} (μM)

galactose is measured, and it corresponds to the average galactose signal value which is halfway between the baseline and the average maximal signal for the substance tested. If IC_{50} (μM) galactose is more than 10 then the substance is considered inactive (i.e., does not cause mitochondrial toxicity) and if less than or equal to 10, then it is active and causes mitochondrial toxicity. This mitochondrial toxicity annotation was used to train predictive models for compounds' mitochondrial toxicity prediction. In total 221 compounds (PROTAC and non-PROTAC) were used to train the models with 96 active (mitotoxic) compounds and 125 inactive (not mitotoxic) compounds. Out of the total of 221 compounds, the 149 were PROTACs and in more detail, the 90 PROTACs were mitotoxic with the rest of PROTACs being not mitotoxic.

4.2.6. Mitochondrial toxicity in-silico model training and evaluation

Three times nested five-fold cross-validation was performed with the StratifiedShuffleSplit Python function from Scikit-Learn[250]. The Stratified Shuffle Split (SSS) splits a dataset into a training and test set by preserving the same percentage of data for each class (active and inactive) as in the initial dataset. Schematic representation of the model training process is shown below in Figure 4.3.

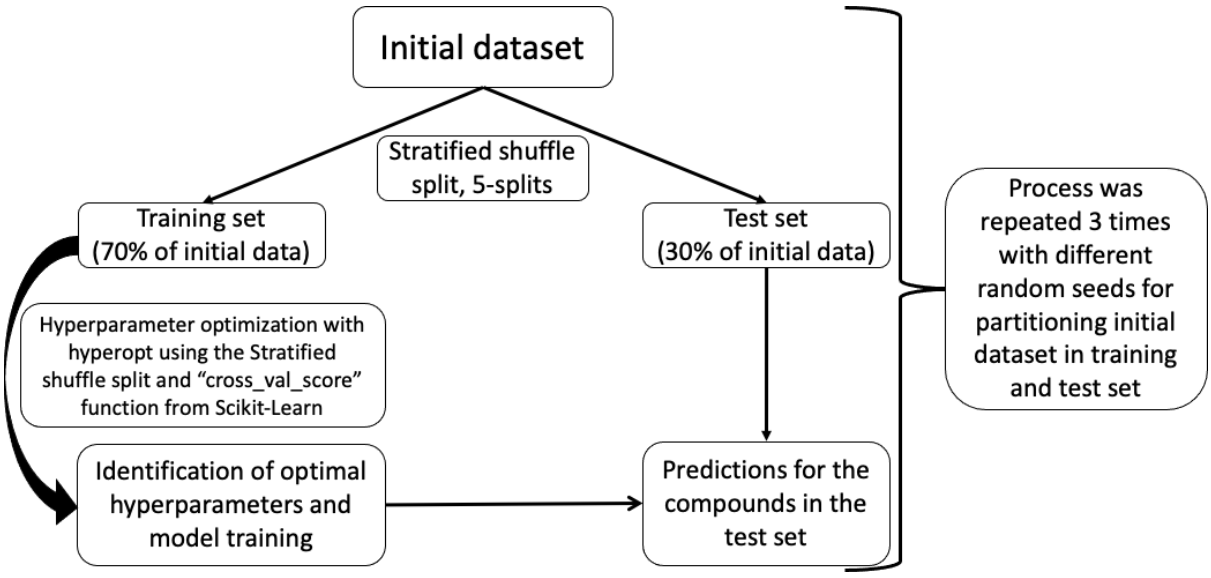


Figure 4.3**:** Schematic representation of model training process. Initial dataset was partitioned in 70% training and 30% test set respectively, 5 times using the stratified shuffle split function from Scikit-Learn. The training set was furhter partitioned 5 times using the stratified shuffle split function from Scikit-Learn to identify the optimal hyperparameters using hyperopt and cross validation score function from Scikit-Learn. When hyperparameters were selected, the models were trained and the compounds

in the test set were predicted. This process was repeated with 3 different random seeds when the initial data were partitioned.

Machine Learning models to predict PROTACs' mitochondrial toxicity were trained (as shown in Figure 4.3) with three different algorithms: a) Random Forest (RF), Support Vector Classifier (SVC) and c) eXtreme Gradient Boosting (XGB). RF and SVC were implemented with the RandomForestClassifier and SupportVectorClassifier functions respectively from Scikit-Learn[250] and XGB with the XGBClassifier from xgboost python package. Hyperparameter selection for each of the algorithms was performed by using hyperopt python package[288,289]. Hyperopt can automate the search for optimal hyperparameter configuration based on a Bayesian optimisation framework and supported by the Sequential Model -Based global Optimisation (SMBO) methodology. Hyperopt ensures that past evaluation results of hyperparameters are tracked and are used to form a probabilistic model mapping the hyperparameters to a probability of a score on the objective function. This probabilistic model is called "surrogate" model for the objective function. In summary, hyperopt trains a surrogate model for the objective function and finds the hyperparameters that perform the best on the surrogate, which is continuously updated by incorporating new results (every time a hyperparameter set is evaluated) and this process was repeated 200 times. The main idea behind this process relies on Bayesian reasoning that the selection of hyperparameters becomes more optimal with more data for the "surrogate" model. Therefore, the Bayesian optimisation with SMBO refers to the sequential running of trials and each time a better set of hyperparameters is tested and the surrogate model is updated. The parameters and the range of values (configuration space), which were explored for each algorithm are included in appendix (Table 9.1). Cell Painting Features were used as descriptors for the models. Different model evaluation metrics were used from Scikit-Learn[250], and they were averaged to give the overall performance across the different folds of cross-validation for Receiver Operating Characteristic – Area Under Curve (ROC-AUC), Precision (Equation 4.2), Recall (Equation 4.3), F_1-score (Equation 4.4), Balanced accuracy (Equation 4.5), brier score (Equation 4.6) and Mathews Correlation Coefficient (MCC, Equation 4.7).

$$\text{Precision} = \frac{TP}{TP+FP} \qquad \text{(Equation 4.2)}$$

$$\text{Recall} = \frac{TP}{TP+FN} \qquad \text{(Equation 4.3)}$$

$$\mathrm{F}_1 - \mathrm{score} = 2 \times \frac{\mathrm{Precision} \times \mathrm{Recall}}{\mathrm{Precision+Recall}} \quad \text{(Equation 4.4)}$$

$$Balanced\ Accuracy = \frac{\left(\frac{TP}{TP+FN}+\frac{TN}{FP+TN}\right)}{2} \quad \text{(Equation 4.5)}$$

$$Brier\ Score = \frac{1}{N}\sum_{true\ value=1}^{N}(predicted\ probability - true\ value)^2 \quad \text{(Equation 4.6)}$$

$$MCC = \frac{TP\ x\ TN - FP\ x\ FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \quad \text{(Equation 4.7)}$$

where TP denotes true-positives, FP denotes false- positives, TN denotes true-negatives and FN denotes false- negatives.

Finally, y-scrambling[254] was performed in order to evaluate whether the trained models performed better than the y-scrambled models. Y-scrambling was applied by randomly reorganising the mitochondrial toxicity labels. Models were rebuilt and evaluated with the same process as the unscrambled (actual) models.

4.2.7. Prospective Model Validation

PROTACs that have been tested on the mitochondrial toxicity assay after the compounds, which have been included in the benchmarking of models were extracted and used as external validation set. This set included five PROTACs that caused mitochondrial toxicity and 34, which did not.

4.3. Results and Discussion

4.3.1. PROTACs can change cellular morphology in Cell Painting Assay

PROTAC profiles together with non-PROTAC molecules were used to understand whether they show systematically different Cell Profiling readouts compared to neutral controls, based on two metrics: a) Euclidean Distance-based and b) grit score activity metric. The results from the Euclidean distance-based method showed that out of the 1,000 (three replicates per PROTAC) profiles obtained from testing PROTACs at concentrations 0.1, 1, and 10 μM, ~17%, ~61% and ~80% profiles, respectively displayed cellular morphology different from the neutral controls (Figure 4.4a-c). In line, higher grit scores were observed with increasing concentration in Figure 4.4g with a mean ± standard deviation of 0.65±0.72, 1.32±1.07 and 2.56±1.49 for concentrations of 0.1, 1 and 10μM respectively. For non-PROTAC compounds, the results (shown in Figure 4.4d-f) from the Euclidean distance-based method showed that out of ~450 profiles (three replicates per compound) tested on concentrations 0.1, 1, and 10μM, the ~22%, ~46% and ~60% profiles, respectively displayed cellular morphology different from the neutral controls. Similarly, higher grit scores were observed with increasing concentration in Figure 4.4g with a mean ± standard deviation of 0.65±1.20, 1.04±1.30 and 1.80±1.60 for concentrations of 0.1, 1 and 10μM, respectively. Next, a dimensionality reduction of the PROTACs-CP profiles was performed with Uniform manifold approximation (UMAP[290]) to understand better which phenotypic responses are clustered together using Cell Profiling readouts using this method. The results of this analysis are shown in appendix in Figure 9.7, which demonstrates a range of cellular phenotypes.

As indicated by this study, not all the compounds (PROTAC and non-PROTAC) tested can change cell morphology, but this is not something unusual for Cell Painting assay. Generally, given an *a priori* knowledge of the activity of genetic and small-molecule perturbations on Cell Painting assay or similar assays, it is not expected all perturbations to be active on the assay. For example, in a Cell Painting CRISPR assay only 50% of the 220 genes yielded detectable morphological profiles[291]. In another genome-wide gene KO screen (21,000 genes), only 1,249 genes were deviating significantly from the controls based on a 200-feature fingerprint calculated from cell

nucleus[292]. Moreover, in the first launch of small-molecule Cell Painting dataset, only 13% (203/1,600) of the compounds were active on the assay[106]. The lack of changes in cellular morphology does not necessarily mean that the PROTACs does not produce a change in cellular morphology, but this lack of changes is a multiparametric condition related to the experimental parameters considered. These parameters include the cell-line used (in this study U2OS cells were used), perturbation time and concentration, and the organelles and cellular sub-compartments that are fluorescently labelled. For example, Cox et al.[51] identified 53% active out of 1,177 compounds tested in a HTI assay in at least one cell line (three in total) and there were compounds annotated with the same MoA label (Glucorticoid receptor agonists), which were active on A549 cell line and not the other two cell lines.

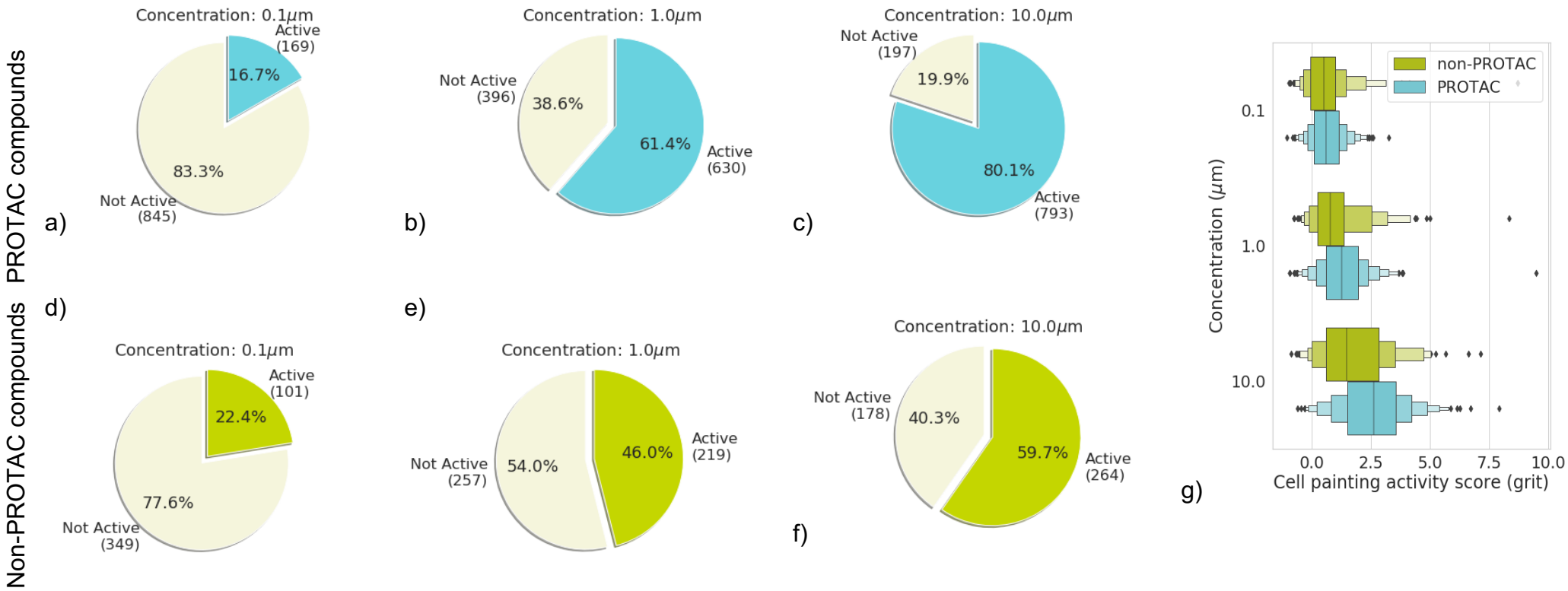


Figure 4.4: Percentage of a-c) PROTACs and d-f) non-PROTACs compounds identified as active on the Cell Painting assay with the Euclidean-based method (i.e. compounds that are able to change the cellular morphology) at concentrations of 0.1, 1 and 10. g) Cell Painting activity score in the form of grit measure across all concentration. Both the grit score and the Euclidean distance-based method showed that the number of active compounds increases as the concentration increases.

4.3.2. PROTACs with mitotoxic annotations are active in the Cell Painting assay

Considering that compounds and in particular PROTACs when combined with Cell Painting assay can alter the cellular morphology, the next question was how to use the image-based profiles. One option was to investigate whether they could be used as a PROTACs' information to evaluate their safety. In more detail, mitochondrial toxicity annotations (from AstraZeneca) for the compounds were extracted. The annotations were further categorised in highly mitotoxic (IC_{50} <1μM; 51 compounds), moderately mitotoxic (IC_{50} between 1μM and 10μM; 44 compounds) and not mitotoxic (IC_{50} >10μM; 126 compounds). The grit score for each category of mitotoxicity and perturbation type (PROTAC or non-PROTAC compound) is shown in Figure 4.5a.

At a concentration of 10μM, the mean grit score was 3.01±1.31, 3.09±1.20, and 1.98±1.59 for highly, moderately, and not mitotoxic PROTACs respectively (Figure 4.5). At concentration 1μM, the mean grit score was 1.75±0.97, 1.24±0.91 and 1.14±1.28 for highly, moderately, and not mitotoxic PROTACS respectively. However, the same trend was not observed at concentration 0.1μM, where the mean grit score was 0.64±0.75, 0.73±0.81, and 0.63±0.56 for highly, moderately, and not mitotoxic PROTACs respectively. Hence, PROTACs that are mitotoxic (highly and moderately) display a higher grit score compared to not mitotoxic especially for concentration 1 and 10μM. Hence, the morphological difference between mitotoxic and not mitotoxic PROTACs indicated by higher grit scores, is more pronounced at concentrations of 1 and 10 μM. Similar trends were observed for the small molecule compounds (non-PROTAC). For example, at concentration 1μM, the mean grit score was 2.36±0.88, 1.36±1.34, and 1.04±1.34 for highly, moderately, and not mitotoxic non-PROTAC compounds respectively. Furthermore, a UMAP dimensionality reduction was performed on the morphological feature space which revealed a separation between mitotoxic and not mitotoxic compounds. Again, this was more evident for the concentrations of 10 and 1μM (Figure 4.5c,d). In summary, results indicated that mitotoxic compounds induced distinct phenotypic changes which are picked by the Cell Painting assay, and which might be used to differentiate between mitotoxic and not mitotoxic compounds. Therefore, it was further investigated whether these profiles

can be used as a descriptor for *in-silico* Machine Learning models for mitochondrial toxicity prediction.

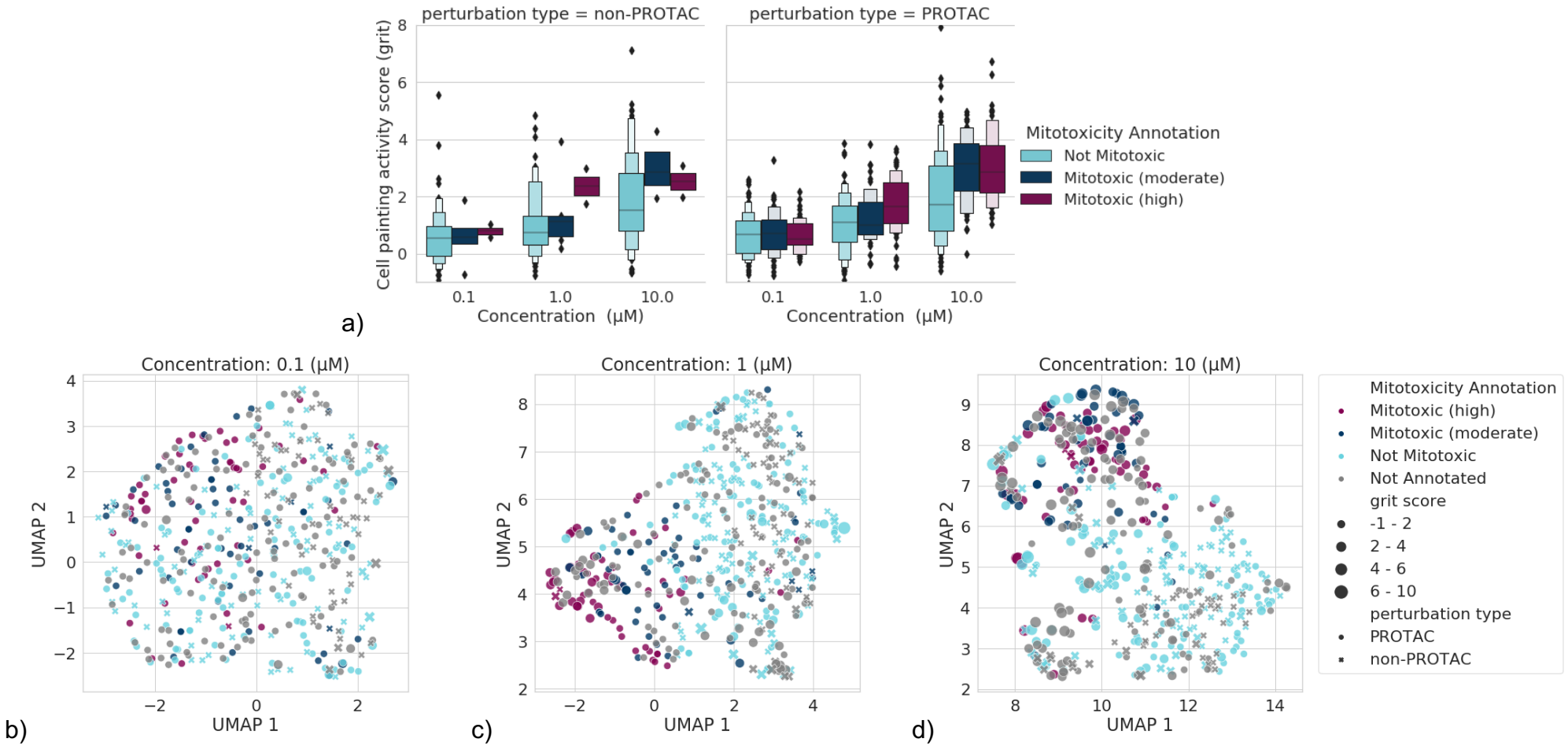


Figure 4.5: a) Cell Painting activity score in the form of grit score across concentrations equal to 0.1, 1.0 and 10.0 μM for each perturbation type. Uniform manifold approximation (UMAP) coordinates of all perturbations labelled with mitotoxicity annotations at concentrations b) 0.1, c) 1 and d) 10 μM.

4.3.3. Evaluation of mitotoxicity prediction models

Next, the Cell Painting profiles were used to train models three different algorithms namely, Random Forest (RF), Support Vector Classifier (SVC) and eXtreme Gradient Boosting (XGB). Model evaluation results are shown in Figure 4.6a. Models were trained for each concentration and scored a ROC-AUC equal to 0.93, 0.93, and 0.80 and F_1-score equal to 0.85, 0.87 and 0.74 for concentrations 10, 1 and 0.1μM respectively. To further validate that the performance was not random, it was evaluated whether the models performed better than random models by applying y-scrambling. The y-scrambled models scored a mean ROC-AUC across all algorithms equal to 0.50, 0.51. and 0.49 for concentrations 0.1, 1 and 10μM respectively (i.e., close to the expected value of 0.5) as shown in Figure 4.6b. Hence, the models performed better than the y-scrambled models and thus they are unlikely to have been obtained by chance.

Using the SVC algorithm, the balanced accuracy was equal to 0.76, 0.88 and 0.85 when the models were trained with profiles from concentrations 0.1, 1 and 10 μM respectively (Figure 4.6a). Hence, models trained with Cell Painting profiles from the two higher concentrations of 1 and 10μM outperformed the models trained on profiles from the concentration of 0.1μM. Similarly, concentrations of 1 and 10μM outperformed concentration 0.1μM regardless of the algorithm used as shown in Figure 4.6a. This is in agreement with the finding described above that grit scores were larger for mitotoxic compounds at the two higher concentrations than at the lower concentration tested. Furthermore, this can be explained by the fact that, a high intra-class correlation was observed between the mitotoxic compounds in the Cell Painting features at a concentration of 10 and 1μM with a median of 0.48 and 0.32 respectively, compared to a lower intra-class Pearson correlation at concentration of 0.1μM with a median of 0.16 (Figure 4.6c-e). Hence, PROTAC and non-PROTAC compounds that cause mitochondrial toxicity were significantly more similar to each other at concentrations 1 and 10μM (Figure 4.6c,d), compared to features derived at 0.1μM (Figure 4.6c). Furthermore, a high difference in the intraclass and interclass correlations (between mitotoxic and not mitotoxic) was observed and were equal to 0.28, 0.21 and 0.07 for concentration 10, 1 and 0.1μM respectively. Overall, this

means that active compounds at concentrations of 10 and 1μM are clearly different from inactive compounds (median similarities of 0.48 vs 0.20 and 0.32 vs 0.11 respectively), while being less indistinguishable at concentration 0.1μM (median similarities of 0.16 vs 0.09). Taken together, this similarity analysis additionally explained why using concentrations 1 and 10μM outperforms concentration 0.1μM model performance.

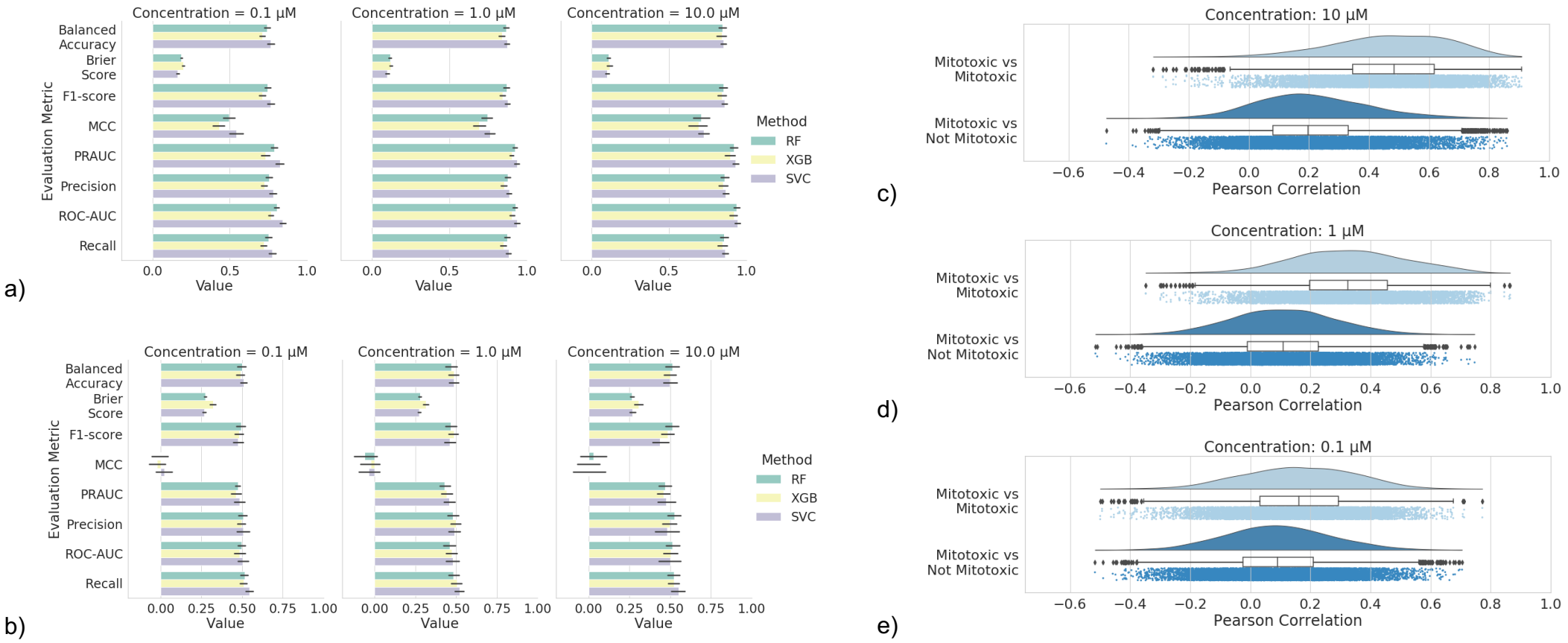


Figure 4.6: a) Model performance and b) y-scrambled model performance, using the Cell Painting features and three different algorithms; RF, XGB and SVC at concentrations a) 10, b) 1 and c) 0.1 µM. The error bars correspond to the confidence interval

across all splits and random states used for cross validation. c-e) Intra-class (Mitotoxic vs Not Mitotoxic) vs Inter-class Pearson correlation of the image-based features at concentrations 0.1, 1 and 10 μM.

4.3.4. Prospective experimental model validation

To further validate the findings, an external validation was performed for the mitochondrial toxicity models. Out of the total compounds tested in the Glu/Gal assay, there were 39 PROTACs that were tested later out of which five were mitotoxic and 34 were not mitotoxic, which were used as a prospective test set. A similarity analysis (by calculating Pearson correlation) was initially performed between the 39 query PROTACs to the compounds which cause mitochondrial toxicity and those which do not (i.e., the compounds in the models) as shown in appendix Figure 9.8**.** For concentrations 1 and 10μM, the mitotoxic query PROTACs showed a higher correlation with the mitotoxic compared to the correlation with the not mitotoxic. In addition, the not mitotoxic query PROTACs did not show a high correlation with the mitotoxic PROTACs in the models. This created the hypothesis that the models have a high chance to be able to also classify the prospective test set correctly.

The mitochondrial toxicity of the 39 PROTACs was hence predicted by all the models and the external validation results are summarised with confusion matrices and a range of evaluation metrics in Figure 4.7a and b respectively. The models trained with data at concentration 1 and 10 μM performed well and outperformed the models trained with data at a concentration of 0.1 μM. For example, the balanced accuracy was equal to 0.68, 0.96 and 0.89 when the models were trained with profiles from concentrations 0.1, 1 and 10μM respectively (Figure 4.7b). Moreover, the models trained with the data at a concentration of 0.1 μM showed relatively high retrieval for mitotoxic PROTACs (more than 60% of mitotoxic PROTACs were correctly classified, but on the other hand showed high false-positive rates (Figure 4.7a). The models trained with the data from concentration 1 and 10 μM were consistently able to predict the majority of the mitotoxic PROTACs (Figure 4.7a), with the models using data from the concentration of 1 μM being able to predict 100% of the mitotoxic PROTACs, regardless of the algorithm used. Models trained with the data from the highest concentration of 10μM were able to correctly detect 60%, 80% and 80% of the mitotoxic PROTACs using the RF, SVC and XGB algorithms respectively (Figure 4.7a). On the other hand, the models trained with data from concentration 10μM have a lower number of false positives and thus a higher number of true negatives

compared to models trained with data from concentration 1μM (Figure 4.7a). Finally, 97% and 91-97% of the not mitotoxic PROTACs are correctly classified using the models trained with data from concentration 10 and 1μM respectively (Figure 4.7a).

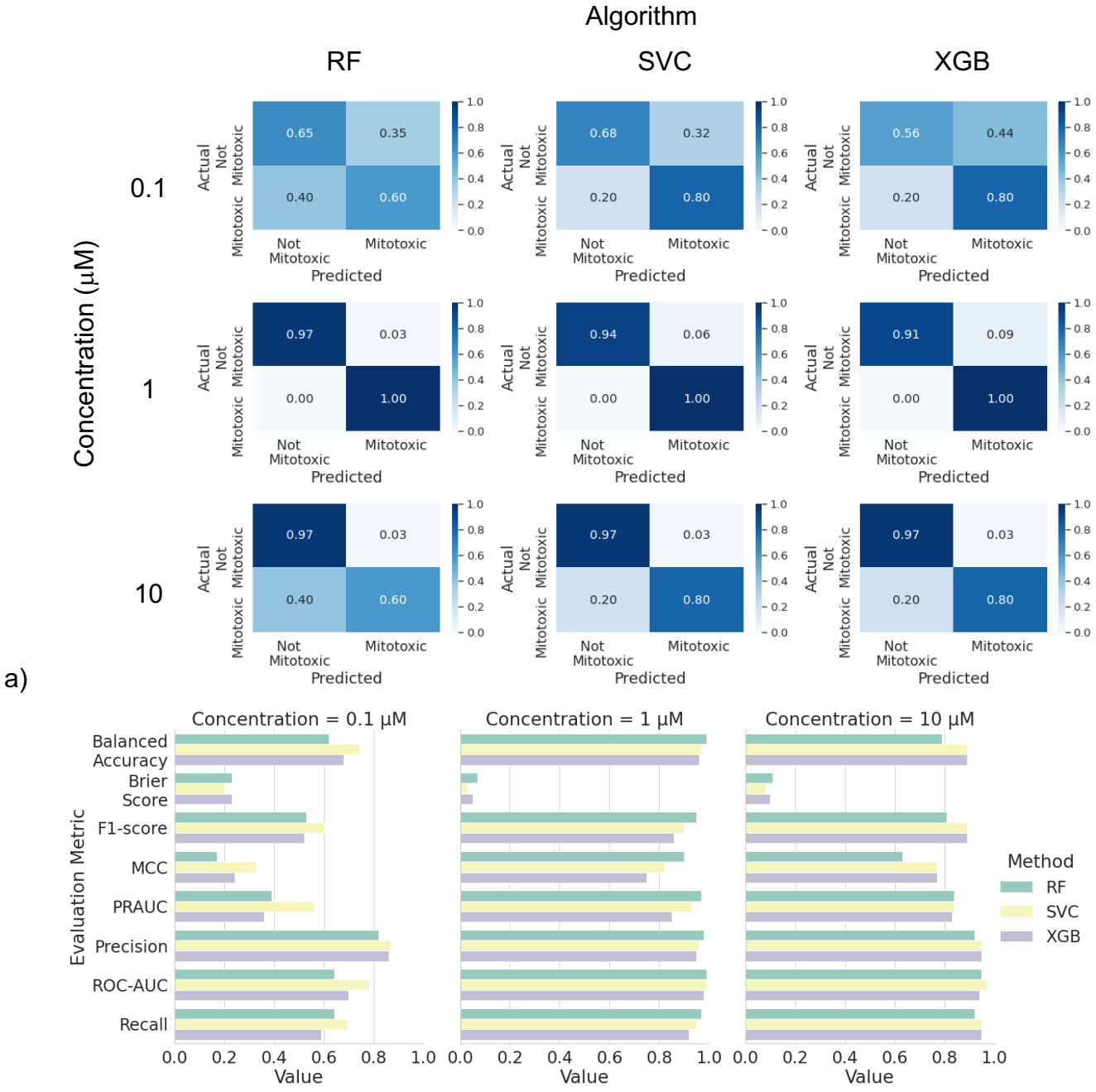


Figure 4.7: a) Prospective experimental model validation results obtained with the models trained with RF, SVC and XGB algorithms and with data from concentration 0.1,1 and 10μM, visualised with confusion matrices. b) Model performance on the prospective experimental model validation using the Cell Painting features and three different algorithms; RF, XGB and SVC at concentrations a) 10, b) 1 and c) 0.1 μM.

The error bars correspond to the confidence interval across all splits and random states used for cross validation.

4.4. Conclusion

In this work, it was evaluated whether PROTACs can be profiled with the Cell Painting assay. In addition, it was evaluated whether the cell morphological profiles derived from the Cell Painting assay could be used as a compounds' descriptor to predict mitochondrial toxicity. Results showed that PROTACs can induce cell morphological changes, and this was proved by using two different metrics: Euclidean distance-based and grit score. In addition, the Cell Painting profiles were used as descriptors in mitochondrial toxicity prediction models and resulted in models with high performance especially when the models were trained with the profiles from concentrations 1 and 10. Finally, the models showed a good performance in predicting an external set of PROTACs. According to our knowledge, this is the first time to show that PROTACs can change the cellular morphology in the Cell Painting assay and this finding create new hypothesis on how the readouts from this assay can be used to better understand this new modality.

5. Chapter 5: Conclusions

5.1. Summary of findings

The aim of this thesis was to better understand compounds' biological effects by using different types of methods and data. In more detail, this thesis provides answers to two questions about which data and methods to use for compounds' target prediction for MoA understanding and how to explore the safety profile of new data modalities such as PROTACs using high throughput imaging data.

The first part of this thesis (chapter 2) focused on applying PRF (which originated in the field of astronomy) in compounds' target prediction because of its ability to take into account the experimental uncertainty close to the classification threshold. This novel method was further compared to the long-established RF and the key conclusion was that PRF exhibited the largest benefit over the RF toward the midpoint of the probability scale, i.e. for marginal cases on the binary threshold boundary. It was also shown that the addition of putative (sphere excluded) inactive compounds affected PRF performance compared to RF but did not affect the observations obtained from the comparison of RF vs PRF. These results are highlighting the cases, where PRF can be a good choice for training models and also how to construct the training space for models. PRF could be applied in cases where experimental uncertainty is large, and where values are distributed around the classification threshold. Finally, the inclusion of putative inactive compounds should be carefully considered by taking into account the diversity of applicability domain this data might introduce in the training data.

The second part of this thesis (chapter 3) was also in the area of target prediction (as chapter 2) but this time focusing on using different types of compound information. Therefore, one commonly used type of chemical structure-based information was compared with cell morphology-based information for compound bioactivity prediction across 224 targets. Results of this analysis showed that cell morphology information was able to predict the bioactivity of compounds for a range of targets, which is in accordance with literature findings. However, the comparison with the chemical structure information provided novel information in terms of which information is

superior for which target. For example, the β-catenin protein was significantly better predicted by cell morphology side information, whereas compounds' bioactivity for proteins belonging in the GPCR1 family were better predicted by the chemical structure-based information. These results provide a rational why for some targets (e.g., β-catenin), it might be beneficial to profile compounds with the Cell Painting assay to derive cell morphology information for target prediction, which requires more time and cost compared to using computationally calculated descriptors (such as ECFP).

The final chapter is again focusing on predicting biological effects of compounds and more specifically mitochondrial toxicity but this time focusing on a new data modality; the PROteolysis TArgeting Chimeras. The aim of this chapter was two-fold; firstly, to identify whether PROTACs can be profiled with the Cell Painting assay and thus alter the cell morphology and secondly whether compounds' Cell Painting profiles could be used for the prediction of mitochondrial toxicity, which is one important safety aspect. Results indicated that PROTACs can alter the cellular morphology in the Cell Painting assay and the derived profiles were used for training models to predict mitochondrial toxicity, which showed good performance. The novelty of these results is that Cell Painting assay could be used in the future for PROTACs profiling and paves the way to investigate what other PROTACs' properties could be better understood or predicted using the Cell Painting assay. In addition, to the best of our knowledge, the *in*-silico mitochondrial toxicity model is the first one reported in the literature for PROTACs.

5.2. Limitations and Future Work

Chapter 2 focused on using and evaluating Probabilistic Random Forest as an algorithm to take into account experimental uncertainty during model training. This means that PRF considered only the uncertainty of the activity labels extracted from bioactivity databases. However, PRF can also consider the uncertainty of the compounds' descriptors (which was not applicable for the chemical structure descriptors used in Chapter 2). Therefore, an interesting future application of PRF is to use it with compound descriptors, which have an experimental uncertainty. For

example, such descriptors could be compound profiles derived from gene expression experiments or profiles from high throughput imaging assays such as the Cell Painting assay.

The main limitation in chapter 3 is the amount of data used to train the models. In more detail, the bioactivity matrix consisted of ~10,000 compounds vs 224 targets with a sparseness 96.4%. However, it was not possible to include more data because the largest publicly available Cell Painting dataset was used. Therefore, results are limited to this amount of data and when more datasets become available in the public domain, they should be used in the future. More Cell Painting data will be available in the near future from a joint effort from Imaging Platform at the Broad Institute of MIT and Harvard with 12 industry and non-profit partners, which has been recently announced. They are going to release a large reference collection of image data with 1 billion cells responding to over 140,000 small molecules and genetic perturbations. In addition, another limitation of chapter 3 is that the two descriptors were not merged in one model because the focus was mainly on the comparison in order to identify the targets for which might be beneficial to screen compounds with the Cell Painting assay. Therefore, an interesting future application would be to identify methods to merge these two types of descriptors.

Finally, the limitation of last chapter is the low number of PROTACs profiled. Since the work in this thesis provides evidence that PROTACs can be successfully profiled with the Cell Painting assay, more PROTACs could be profiled in the future. In addition, it is important to investigate more experimental conditions for the profiling of PROTACs with the Cell Painting. These include different concentration values and perturbation times and also different cell lines. The selection of cell line is important because some proteins might not be expressed in a particular cell line. Therefore, if one wants to profile PROTACs and investigate effects related to their MoA, a cell line should be used in which the protein of interest is expressed. In addition, in the final chapter only the mitochondrial toxicity of PROTACs was predicted using the PROTACs Cell Painting profiles. Therefore, it is important to investigate other safety-related assays for which Cell Painting profiles could be used as a descriptor to computationally predict

them. Finally, PROTACs Cell Painting profiles could be compared with other “PROTACs’ -omics spaces” such as proteomics in order to identify if they provide similar information or complementary.

6. Chapter 6: Bibliography

(1) Vamathevan, J.; Clark, D.; Czodrowski, P.; Dunham, I.; Ferran, E.; Lee, G.; Li, B.; Madabhushi, A.; Shah, P.; Spitzer, M.; Zhao, S. Applications of Machine Learning in Drug Discovery and Development. *Nat. Rev. Drug Discov. 2019 186* **2019**, *18* (6), 463–477. https://doi.org/10.1038/s41573-019-0024-5.

(2) Arrowsmith, J. Phase II Failures: 2008–2010. *Nat. Rev. Drug Discov. 2011 105* **2011**, *23* (5), 87. https://doi.org/10.1038/nrd3439.

(3) Arrowsmith, J. Phase III and Submission Failures: 2007–2010. *Nat. Rev. Drug Discov. 2011 102* **2011**, *10* (2), 87. https://doi.org/10.1038/nrd3375.

(4) Arrowsmith, J.; Miller, P. Trial Watch: Phase II and Phase III Attrition Rates 2011-2012. *Nat. Rev. Drug Discov.* **2013**, *12* (8), 569. https://doi.org/10.1038/NRD4090.

(5) Harrison, R. K. Phase II and Phase III Failures: 2013-2015. *Nat. Rev. Drug Discov.* **2016**, *15* (12), 817–818. https://doi.org/10.1038/NRD.2016.184.

(6) Jenkinson, S.; Schmidt, F.; Rosenbrier Ribeiro, L.; Delaunois, A.; Valentin, J. P. A Practical Guide to Secondary Pharmacology in Drug Discovery. *J. Pharmacol. Toxicol. Methods* **2020**, *105*. https://doi.org/10.1016/J.VASCN.2020.106869.

(7) Mathur, R.; Rotroff, D.; Ma, J.; Shojaie, A.; Motsinger-Reif, A. Gene Set Analysis Methods: A Systematic Comparison. *BioData Min. 2018 111* **2018**, *11* (1), 1–19. https://doi.org/10.1186/S13040-018-0166-8.

(8) Liggi, S.; Drakakis, G.; Koutsoukas, A.; Cortes–Ciriano, I.; Martínez–Alonso, P.; Malliavin, T. E.; Velazquez-Campoy, A.; Brewerton, S. C.; Bodkin, M. J.; Evans, D. A.; Glen, R. C.; Carrodeguas, J. A.; Bender, A. Extending in Silico Mechanism-of-Action Analysis by Annotating Targets with Pathways: Application to Cellular Cytotoxicity Readouts. *Futur. Sci.* **2014**, *6* (18), 2029–2056.

(9) Maddison, J.; Page, S.; Church, D. B. Small Animal Clinical Pharmacology. *Small Anim. Clin. Pharmacol.* **2008**.

(10) Trusheim, M. R.; Berndt, E. R.; Douglas, F. L. Stratified Medicine: Strategic and Economic Implications of Combining Drugs and Clinical Biomarkers. *Nat. Rev. Drug Discov. 2007 64* **2007**, *6* (4), 287–293. https://doi.org/10.1038/nrd2251.

(11) Mechanism Matters. *Nat. Med. 2010 164* **2010**, *16* (4), 347–347.

https://doi.org/10.1038/nm0410-347.

(12) Rovin; Lisa. 22 CASE STUDIES WHERE PHASE 2 AND PHASE 3 TRIALS HAD DIVERGENT RESULTS. **2017**.

(13) Moffat, J. G.; Vincent, F.; Lee, J. A.; Eder, J.; Prunotto, M. Opportunities and Challenges in Phenotypic Drug Discovery: An Industry Perspective. *Nat. Rev. Drug Discov. 2017 168* **2017**, *16* (8), 531–543. https://doi.org/10.1038/nrd.2017.111.

(14) Vane, J. .; Botting, R. . The Mechanism of Action of Aspirin. *Thromb. Res.* **2003**, *110* (5), 255–258. https://doi.org/10.1016/S0049-3848(03)00379-7.

(15) Davis, R. L. Mechanism of Action and Target Identification: A Matter of Timing in Drug Discovery. *iScience* **2020**, *23* (9). https://doi.org/10.1016/J.ISCI.2020.101487.

(16) Bailey, C. J. Metformin: Historical Overview. *Diabetol. 2017 609* **2017**, *60* (9), 1566–1576. https://doi.org/10.1007/S00125-017-4318-Z.

(17) Zhou, G.; Myers, R.; Li, Y.; Chen, Y.; Shen, X.; Fenyk-Melody, J.; Wu, M.; Ventre, J.; Doebber, T.; Fujii, N.; Musi, N.; Hirshman, M. F.; Goodyear, L. J.; Moller, D. E. Role of AMP-Activated Protein Kinase in Mechanism of Metformin Action. *J. Clin. Invest.* **2001**, *108* (8), 1167–1174. https://doi.org/10.1172/JCI13505.

(18) Bachurin, S.; Bukatina, E.; Lermontova, N.; Tkachenko, S.; Afanasiev, A.; Grigoriev, V.; Grigorieva, I.; Ivanov, Y.; Sablin, S.; Zefirov, N. Antihistamine Agent Dimebon as a Novel Neuroprotector and a Cognition Enhancer. *Ann. N. Y. Acad. Sci.* **2001**, *939*, 425–435. https://doi.org/10.1111/J.1749-6632.2001.TB03654.X.

(19) Bachurin, S. O.; Shevtsova, E. P.; Kireeva, E. G.; Oxenkrug, G. F.; Sablin, S. O. Mitochondria as a Target for Neurotoxins and Neuroprotective Agents. *Ann. N.Y. Acad. Sci* **2003**, *993*, 334–344.

(20) Steele, J. W.; Kim, S. H.; Cirrito, J. R.; Verges, D. K.; Restivo, J. L.; Westaway, D.; Fraser, P.; Hyslop, P. S. G.; Sano, M.; Bezprozvanny, I.; Ehrlich, M. E.; Holtzman, D. M.; Gandy, S. Acute Dosing of Latrepirdine (Dimebon™), a Possible Alzheimer Therapeutic, Elevates Extracellular Amyloid-β Levels in Vitro and in Vivo. *Mol. Neurodegener. 2009 41* **2009**, *4* (1), 1–11.

https://doi.org/10.1186/1750-1326-4-51.

(21) Wu, J.; Li, Q.; Bezprozvanny, I. Evaluation of Dimebon in Cellular Model of Huntington's Disease. *Mol. Neurodegener. 2008 31* **2008**, *3* (1), 1–11. https://doi.org/10.1186/1750-1326-3-15.

(22) Downward, J. The Ins and Outs of Signalling. *Nat. 2001 4116839* **2001**, *411* (6839), 759–762. https://doi.org/10.1038/35081138.

(23) Ardito, F.; Giuliani, M.; Perrone, D.; Troiano, G.; Lo Muzio, L. The Crucial Role of Protein Phosphorylation in Cell Signaling and Its Use as Targeted Therapy (Review). *Int. J. Mol. Med.* **2017**, *40* (2), 271–280. https://doi.org/10.3892/IJMM.2017.3036.

(24) Lodish, H.; Berk, A.; Zipursky, S. L.; Matsudaira, P.; Baltimore, D.; Darnell, J. Molecular Cell Biology; 2000.

(25) Ammeux, N.; Housden, B. E.; Georgiadis, A.; Hu, Y.; Perrimon, N. Mapping Signaling Pathway Cross-Talk in Drosophila Cells. *Proc. Natl. Acad. Sci.* **2016**, *113* (35), 9940–9945. https://doi.org/10.1073/PNAS.1610432113.

(26) Dumont, J. E.; Dremier, S.; Pirson, I.; Maenhaut, C. Cross Signaling, Cell Specificity, and Physiology. *https://doi.org/10.1152/ajpcell.00581.2001* **2002**, *283* (1 52-1). https://doi.org/10.1152/AJPCELL.00581.2001.

(27) Trapotsi, M.-A.; Hosseini-Gerami, L.; Bender, A. Computational Analyses of Mechanism of Action (MoA): Data, Methods and Integration. *RSC Chem. Biol.* **2022**. https://doi.org/10.1039/D1CB00069A.

(28) Lee, J.; Bogyo, M. Target Deconvolution Techniques in Modern Phenotypic Profiling. *Curr. Opin. Chem. Biol.* **2013**, *17* (1), 118–126. https://doi.org/10.1016/J.CBPA.2012.12.022.

(29) Sleno, L.; Emili, A. Proteomic Methods for Drug Target Discovery. *Curr. Opin. Chem. Biol.* **2008**, *12* (1), 46–54. https://doi.org/10.1016/J.CBPA.2008.01.022.

(30) Maggiora, G. M.; Freitas, A. A.; Bender, A.; Ghafourian, T.; Breneman, C. M.; Palm, J. On Outliers and Activity CliffsWhy QSAR Often Disappoints. *J. Chem. Inf. Model.* **2006**, *46* (4), 1535. https://doi.org/10.1021/ci060117s.

(31) Chen, B.; Greenside, P.; Paik, H.; Sirota, M.; Hadley, D.; Butte, A. J. Relating Chemical Structure to Cellular Response: An Integrative Analysis of Gene Expression, Bioactivity, and Structural Data Across 11,000 Compounds. *CPT*

Pharmacometrics Syst. Pharmacol. **2015**, *4* (10), 576–584. https://doi.org/10.1002/psp4.12009.

(32) Camp, H. S.; Li, O.; Wise, S. C.; Hong, Y. H.; Frankowski, C. L.; Shen, X.; Vanbogelen, R.; Leff, T. Differential Activation of Peroxisome Proliferator-Activated Receptor-Gamma by Troglitazone and Rosiglitazone. *Diabetes* **2000**, *49* (4), 539–547. https://doi.org/10.2337/DIABETES.49.4.539.

(33) Kores, K.; Konc, J.; Bren, U. Mechanistic Insights into Side Effects of Troglitazone and Rosiglitazone Using a Novel Inverse Molecular Docking Protocol. *Pharmaceutics* **2021**, *13* (3), 1–19. https://doi.org/10.3390/pharmaceutics13030315.

(34) Wetmore, B. A.; Wambaugh, J. F.; Ferguson, S. S.; Sochaski, M. A.; Rotroff, D. M.; Freeman, K.; Clewell, H. J.; Dix, D. J.; Andersen, M. E.; Houck, K. A.; Allen, B.; Judson, R. S.; Singh, R.; Kavlock, R. J.; Richard, A. M.; Thomas, R. S. Integration of Dosimetry, Exposure, and High-Throughput Screening Data in Chemical Toxicity Assessment. *Toxicol. Sci.* **2012**, *125* (1), 157–174. https://doi.org/10.1093/TOXSCI/KFR254.

(35) Subramanian, A.; Narayan, R.; Corsello, S. M.; Peck, D. D.; Natoli, T. E.; Lu, X.; Gould, J.; Davis, J. F.; Tubelli, A. A.; Asiedu, J. K.; Lahr, D. L.; Hirschman, J. E.; Liu, Z.; Donahue, M.; Julian, B.; Khan, M.; Wadden, D.; Smith, I. C.; Lam, D.; Liberzon, A.; Toder, C.; Bagul, M.; Orzechowski, M.; Enache, O. M.; Piccioni, F.; Johnson, S. A.; Lyons, N. J.; Berger, A. H.; Shamji, A. F.; Brooks, A. N.; Vrcic, A.; Flynn, C.; Rosains, J.; Takeda, D. Y.; Hu, R.; Davison, D.; Lamb, J.; Ardlie, K.; Hogstrom, L.; Greenside, P.; Gray, N. S.; Clemons, P. A.; Silver, S.; Wu, X.; Zhao, W.-N.; Read-Button, W.; Wu, X.; Haggarty, S. J.; Ronco, L. V; Boehm, J. S.; Schreiber, S. L.; Doench, J. G.; Bittker, J. A.; Root, D. E.; Wong, B.; Golub, T. R. A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles. *Cell* **2017**, *171* (6), 1437--1452.e17. https://doi.org/10.1016/J.CELL.2017.10.049.

(36) Raser, J. M.; O'Shea, E. K. Molecular Biology - Noise in Gene Expression: Origins, Consequences, and Control. *Science (80-.).* **2005**, *309* (5743), 2010–2013. https://doi.org/10.1126/SCIENCE.1105891.

(37) Kalaitzis, A. A.; Lawrence, N. D. A Simple Approach to Ranking Differentially

Expressed Gene Expression Time Courses through Gaussian Process Regression. *BMC Bioinforma. 2011 121* **2011**, *12* (1), 1–13. https://doi.org/10.1186/1471-2105-12-180.

(38) Nusinow, D. P.; Szpyt, J.; Ghandi, M.; Rose, C. M.; McDonald, E. R.; Kalocsay, M.; Jané-Valbuena, J.; Gelfand, E.; Schweppe, D. K.; Jedrychowski, M.; Golji, J.; Porter, D. A.; Rejtar, T.; Wang, Y. K.; Kryukov, G. V.; Stegmeier, F.; Erickson, B. K.; Garraway, L. A.; Sellers, W. R.; Gygi, S. P. Quantitative Proteomics of the Cancer Cell Line Encyclopedia. *Cell* **2020**, *180* (2), 387-402.e16. https://doi.org/10.1016/J.CELL.2019.12.023.

(39) Piehowski, P. D.; Petyuk, V. A.; Orton, D. J.; Xie, F.; Moore, R. J.; Ramirez-Restrepo, M.; Engel, A.; Lieberman, A. P.; Albin, R. L.; Camp, D. G.; Smith, R. D.; Myers, A. J. Sources of Technical Variability in Quantitative LC–MS Proteomics: Human Brain Tissue Sample Analysis. *J. Proteome Res.* **2013**, *12* (5), 2128–2137. https://doi.org/10.1021/PR301146M.

(40) Medo, M.; Aebersold, D. M.; Medová, M. ProtRank: Bypassing the Imputation of Missing Values in Differential Expression Analysis of Proteomic Data. *BMC Bioinforma. 2019 201* **2019**, *20* (1), 1–12. https://doi.org/10.1186/S12859-019-3144-3.

(41) Johnson, C. H.; Gonzalez, F. J. Challenges and Opportunities of Metabolomics. *J. Cell. Physiol.* **2012**, *227* (8), 2975–2981. https://doi.org/10.1002/JCP.24002.

(42) Ramirez, T.; Daneshian, M.; Kamp, H.; Bois, F. Y.; Clench, M. R.; Coen, M.; Donley, B.; Fischer, S. M.; Ekman, D. R.; Fabian, E.; Guillou, C.; Heuer, J.; Hogberg, H. T.; Jungnickel, H.; Keun, H. C.; Krennrich, G.; Krupp, E.; Luch, A.; Noor, F.; Peter, E.; Riefke, B.; Seymour, M.; Skinner, N.; Smirnova, L.; Verheij, E.; Wagner, S.; Hartung, T.; Ravenzwaay, B. van; Leist, M. Metabolomics in Toxicology and Preclinical Research. *ALTEX* **2013**, *30* (2), 209. https://doi.org/10.14573/ALTEX.2013.2.209.

(43) Livera, A. M. De; Sysi-Aho, M.; Jacob, L.; Gagnon-Bartsch, J. A.; Castillo, S.; Simpson, J. A.; Speed, T. P. Statistical Methods for Handling Unwanted Variation in Metabolomics Data. *Anal. Chem.* **2015**, *87* (7), 3606–3615. https://doi.org/10.1021/AC502439Y.

(44) Chaleckis, R.; Meister, I.; Zhang, P.; Wheelock, C. E. Challenges, Progress and

Promises of Metabolite Annotation for LC–MS-Based Metabolomics. *Curr. Opin. Biotechnol.* **2019**, *55*, 44–50. https://doi.org/10.1016/J.COPBIO.2018.07.010.

(45) Abelin, J. G.; Patel, J.; Lu, X.; Feeney, C. M.; Fagbami, L.; Creech, A. L.; Hu, R.; Lam, D.; Davison, D.; Pino, L.; Qiao, J. W.; Kuhn, E.; Officer, A.; Li, J.; Abbatiello, S.; Subramanian, A.; Sidman, R.; Snyder, E.; Carr, S. A.; Jaffe, J. D. Reduced-Representation Phosphosignatures Measured by Quantitative Targeted MS Capture Cellular States and Enable Large-Scale Comparison of Drug-Induced Phenotypes *. *Mol. Cell. Proteomics* **2016**, *15* (5), 1622–1641. https://doi.org/10.1074/MCP.M116.058354.

(46) Chen, Y. A.; Eschrich, S. A. Computational Methods and Opportunities for Phosphorylation Network Medicine. *Transl. Cancer Res.* **2014**, *3* (3), 266.

(47) Litichevskiy, L.; Peckner, R.; Abelin, J. G.; Asiedu, J. K.; Creech, A. L.; Davis, J. F.; Davison, D.; Dunning, C. M.; Egertson, J. D.; Egri, S.; Gould, J.; Ko, T.; Johnson, S. A.; Lahr, D. L.; Lam, D.; Liu, Z.; Lyons, N. J.; Lu, X.; MacLean, B. X.; Mungenast, A. E.; Officer, A.; Natoli, T. E.; Papanastasiou, M.; Patel, J.; Sharma, V.; Toder, C.; Tubelli, A. A.; Young, J. Z.; Carr, S. A.; Golub, T. R.; Subramanian, A.; MacCoss, M. J.; Tsai, L. H.; Jaffe, J. D. A Library of Phosphoproteomic and Chromatin Signatures for Characterizing Cellular Responses to Drug Perturbations. *Cell Syst.* **2018**, *6* (4), 424-443.e7. https://doi.org/10.1016/J.CELS.2018.03.012.

(48) Bray, M.-A.; Singh, S.; Han, H.; Davis, C. T.; Borgeson, B.; Hartland, C.; Kost-Alimova, M.; Gustafsdottir, S. M.; Gibson, C. C.; Carpenter, A. E. Cell Painting, a High-Content Image-Based Assay for Morphological Profiling Using Multiplexed Fluorescent Dyes. *Nat. Protoc. 2016 119* **2016**, *11* (9), 1757–1774. https://doi.org/10.1038/nprot.2016.105.

(49) Lu, A. X.; Kraus, O. Z.; Cooper, S.; Moses, A. M. Learning Unsupervised Feature Representations for Single Cell Microscopy Images with Paired Cell Inpainting. *PLOS Comput. Biol.* **2019**, *15* (9), e1007348. https://doi.org/10.1371/JOURNAL.PCBI.1007348.

(50) Chandrasekaran, S. N.; Ceulemans, H.; Boyd, J. D.; Carpenter, A. E. Image-Based Profiling for Drug Discovery: Due for a Machine-Learning Upgrade? *Nat. Rev. Drug Discov. 2020 202* **2020**, *20* (2), 145–159.

https://doi.org/10.1038/s41573-020-00117-w.

(51) Cox, M. J.; Jaensch, S.; Van de Waeter, J.; Cougnaud, L.; Seynaeve, D.; Benalla, S.; Koo, S. J.; Van Den Wyngaert, I.; Neefs, J. M.; Malkov, D.; Bittremieux, M.; Steemans, M.; Peeters, P. J.; Wegner, J. K.; Ceulemans, H.; Gustin, E.; Chong, Y. T.; Göhlmann, H. W. H. Tales of 1,008 Small Molecules: Phenomic Profiling through Live-Cell Imaging in a Panel of Reporter Cell Lines. *Sci. Rep.* **2020**, *10* (1). https://doi.org/10.1038/s41598-020-69354-8.

(52) Nassiri, I.; McCall, M. N. Systematic Exploration of Cell Morphological Phenotypes Associated with a Transcriptomic Query. *Nucleic Acids Res.* **2018**, *46* (19), e116–e116. https://doi.org/10.1093/nar/gky626.

(53) Ma, J.; Shojaie, A.; Michailidis, G. A Comparative Study of Topology-Based Pathway Enrichment Analysis Methods. *BMC Bioinforma. 2019 201* **2019**, *20* (1), 1–14. https://doi.org/10.1186/S12859-019-3146-1.

(54) Chowdhury, S.; Sarkar, R. R. Comparison of Human Cell Signaling Pathway Databases—Evolution, Drawbacks and Challenges. *Database* **2015**, *2015*. https://doi.org/10.1093/database/bau126.

(55) Vert, G.; Chory, J. Crosstalk in Cellular Signaling: Background Noise or the Real Thing? *Dev. Cell* **2011**, *21* (6), 985–991. https://doi.org/10.1016/J.DEVCEL.2011.11.006.

(56) Haynes, W. A.; Tomczak, A.; Khatri, P. Gene Annotation Bias Impedes Biomedical Research. *Sci. Reports 2018 81* **2018**, *8* (1), 1–7. https://doi.org/10.1038/s41598-018-19333-x.

(57) Koutsoukas, A.; Lowe, R.; KalantarMotamedi, Y.; Mussa, H. Y.; Klaffke, W.; Mitchell, J. B. O.; Glen, R. C.; Bender, A. In Silico Target Predictions: Defining a Benchmarking Data Set and Comparison of Performance of the Multiclass Naïve Bayes and Parzen-Rosenblatt Window. *J. Chem. Inf. Model.* **2013**, *53* (8), 1957–1966. https://doi.org/10.1021/ci300435j.

(58) Koutsoukas, A.; Simms, B.; Kirchmair, J.; Bond, P. J.; Whitmore, A. V; Zimmer, S.; Young, M. P.; Jenkins, J. L.; Glick, M.; Glen, R. C.; Bender, A. From in Silico Target Prediction to Multi-Target Drug Design: Current Databases, Methods and Applications. *J. Proteomics* **2011**, *74* (12), 2554–2574. https://doi.org/10.1016/j.jprot.2011.05.011.

(59) Hulme, E. C.; Trevethick, M. A. Ligand Binding Assays at Equilibrium: Validation and Interpretation. *Br. J. Pharmacol.* **2010**, *161* (6), 1219–1237. https://doi.org/10.1111/J.1476-5381.2009.00604.X.

(60) MarÉchal, E. Measuring Bioactivity: KI, IC50 and EC50. *Chemogenomics Chem. Genet.* **2011**, 55–65. https://doi.org/10.1007/978-3-642-19615-7_5.

(61) Kalliokoski, T.; Kramer, C.; Vulpetti, A.; Gedeck, P. Comparability of Mixed IC50 Data – A Statistical Analysis. *PLoS One* **2013**, *8* (4), e61007. https://doi.org/10.1371/JOURNAL.PONE.0061007.

(62) Mervin, L. H.; Afzal, A. M.; Drakakis, G.; Lewis, R.; Engkvist, O.; Bender, A. Target Prediction Utilising Negative Bioactivity Data Covering Large Chemical Space. *J. Cheminform.* **2015**, *7* (1), 51. https://doi.org/10.1186/s13321-015-0098-y.

(63) Tanoli, Z.; Seemab, U.; Scherer, A.; Wennerberg, K.; Tang, J.; Vähä-Koskela, M. Exploration of Databases and Methods Supporting Drug Repurposing: A Comprehensive Survey. *Brief. Bioinform.* **2021**, *22* (2), 1656–1678. https://doi.org/10.1093/BIB/BBAA003.

(64) Sun, J.; Jeliazkova, N.; Chupakin, V.; Golib-Dzib, J. F.; Engkvist, O.; Carlsson, L.; Wegner, J.; Ceulemans, H.; Georgiev, I.; Jeliazkov, V.; Kochev, N.; Ashby, T. J.; Chen, H. ExCAPE-DB: An Integrated Large Scale Dataset Facilitating Big Data Analysis in Chemogenomics. *J. Cheminform.* **2017**, *9* (1), 1–9. https://doi.org/10.1186/s13321-017-0203-5.

(65) Papadatos, G.; Overington, J. P. The ChEMBL Database: A Taster for Medicinal Chemists. *Future Med. Chem.* **2014**, *6* (4), 361–364. https://doi.org/10.4155/fmc.14.8.

(66) Mendez, D.; Gaulton, A.; Bento, A. P.; Chambers, J.; De Veij, M.; Félix, E.; Magariños, M. P.; Mosquera, J. F.; Mutowo, P.; Nowotka, M.; Gordillo-Marañón, M.; Hunter, F.; Junco, L.; Mugumbate, G.; Rodriguez-Lopez, M.; Atkinson, F.; Bosc, N.; Radoux, C. J.; Segura-Cabrera, A.; Hersey, A.; Leach, A. R. ChEMBL: Towards Direct Deposition of Bioassay Data. *Nucleic Acids Res.* **2019**, *47* (D1), D930–D940. https://doi.org/10.1093/nar/gky1075.

(67) Kim, S.; Chen, J.; Cheng, T.; Gindulyte, A.; He, J.; He, S.; Li, Q.; Shoemaker, B. A.; Thiessen, P. A.; Yu, B.; Zaslavsky, L.; Zhang, J.; Bolton, E. E. PubChem

2019 Update: Improved Access to Chemical Data. *Nucleic Acids Res.* **2019**, *47* (D1), D1102–D1109. https://doi.org/10.1093/nar/gky1033.

(68) Smit, I. A.; Afzal, A. M.; Allen, C. H. G.; Svensson, F.; Hanser, T.; Bender, A. Systematic Analysis of Protein Targets Associated with Adverse Events of Drugs from Clinical Trials and Postmarketing Reports. *Chem. Res. Toxicol.* **2020**, *34* (2), 365–384. https://doi.org/10.1021/ACS.CHEMRESTOX.0C00294.

(69) Gilson, M. K.; Liu, T.; Baitaluk, M.; Nicola, G.; Hwang, L.; Chong, J. BindingDB in 2015: A Public Database for Medicinal Chemistry, Computational Chemistry and Systems Pharmacology. *Nucleic Acids Res.* **2016**, *44* (D1), D1045–D1053. https://doi.org/10.1093/NAR/GKV1072.

(70) Lin, A.; Horvath, D.; Marcou, G.; Afonina, V.; Reymond, J.-L.; Varnek, A. Quantitative Structure-Property Relationships Methods View Project Various Principles of Machine Learning in Chemoinformatics View Project Mapping of the Available Chemical Space versus the Chemical Universe of Lead-Like Compounds. **2018**. https://doi.org/10.1002/cmdc.201700561.

(71) Schaduangrat, N.; Lampa, S.; Simeon, S.; Gleeson, M. P.; Spjuth, O.; Nantasenamat, C. Towards Reproducible Computational Drug Discovery. *J. Cheminformatics 2020 121* **2020**, *12* (1), 1–30. https://doi.org/10.1186/S13321-020-0408-X.

(72) Kramer, C.; Kalliokoski, T.; Gedeck, P.; Vulpetti, A. The Experimental Uncertainty of Heterogeneous Public Ki Data. *J. Med. Chem.* **2012**, *55* (11), 5165–5173. https://doi.org/10.1021/JM300131X.

(73) Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P. ChEMBL: A Large-Scale Bioactivity Database for Drug Discovery. *Nucleic Acids Res.* **2012**, *40* (D1), D1100–D1107. https://doi.org/10.1093/NAR/GKR777.

(74) Mervin, L. H.; Bulusu, K. C.; Kalash, L.; Afzal, A. M.; Svensson, F.; Firth, M. A.; Barrett, I.; Engkvist, O.; Bender, A. Orthologue Chemical Space and Its Influence on Target Prediction. *Bioinformatics* **2017**, *34* (1), 72–79. https://doi.org/10.1093/bioinformatics/btx525.

(75) Dimova, D.; Stumpfe, D.; Bajorath, J. Identification of Orthologous Target Pairs with Shared Active Compounds and Comparison of Organism-Specific Activity

Patterns. *Chem. Biol. Drug Des.* **2015**, *86* (5), 1105–1114. https://doi.org/10.1111/CBDD.12578.

(76) Gfeller, D.; Zoete, V. Protein Homology Reveals New Targets for Bioactive Small Molecules. *Bioinformatics* **2015**, *31* (16), 2721–2727. https://doi.org/10.1093/BIOINFORMATICS/BTV214.

(77) Kramer, C.; Dahl, G.; Tyrchan, C.; Ulander, J. A Comprehensive Company Database Analysis of Biological Assay Variability. *Drug Discov. Today* **2016**, *21* (8), 1213–1221. https://doi.org/10.1016/J.DRUDIS.2016.03.015.

(78) Fourches, D.; Muratov, E.; Tropsha, A. Trust, But Verify: On the Importance of Chemical Structure Curation in Cheminformatics and QSAR Modeling Research. *J. Chem. Inf. Model.* **2010**, *50* (7), 1189–1204. https://doi.org/10.1021/ci100176x.

(79) Hu, Y.; Bajorath, J. Influence of Search Parameters and Criteria on Compound Selection, Promiscuity, and Pan Assay Interference Characteristics. *J. Chem. Inf. Model.* **2014**, *54* (11), 3056–3066. https://doi.org/10.1021/CI5005509.

(80) Tiikkainen, P.; Bellis, L.; Light, Y.; Franke, L. Estimating Error Rates in Bioactivity Databases. *J. Chem. Inf. Model.* **2013**, *53* (10), 2499–2505. https://doi.org/10.1021/CI400099Q.

(81) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50* (5), 742–754. https://doi.org/10.1021/ci100050t.

(82) Bajorath, J. Representation and Identification of Activity Cliffs. *Expert Opin. Drug Discov.* **2017**, *12* (9), 879–883. https://doi.org/10.1080/17460441.2017.1353494.

(83) Martin, Y. C.; Kofron, J. L.; Traphagen, L. M. Do Structurally Similar Molecules Have Similar Biological Activity? *J. Med. Chem.* **2002**, *45* (19), 4350–4358. https://doi.org/10.1021/jm020155c.

(84) Gao, S.; Han, L.; Luo, D.; Liu, G.; Xiao, Z.; Shan, G.; Zhang, Y.; Zhou, W. Modeling Drug Mechanism of Action with Large Scale Gene-Expression Profiles Using GPAR, an Artificial Intelligence Platform. *BMC Bioinforma. 2020 221* **2021**, *22* (1), 1–13. https://doi.org/10.1186/S12859-020-03915-6.

(85) Scheeder, C.; Heigwer, F.; Boutros, M. Machine Learning and Image-Based Profiling in Drug Discovery. *Curr. Opin. Syst. Biol.* **2018**.

https://doi.org/10.1016/j.coisb.2018.05.004.

(86) Berg, E. L. Systems Biology in Drug Discovery and Development. *Drug Discov. Today* **2014**, *19* (2), 113–125. https://doi.org/10.1016/J.DRUDIS.2013.10.003.

(87) Pabon, N. A.; Xia, Y.; Estabrooks, S. K.; Ye, Z.; Herbrand, A. K.; Süß, E.; Biondi, R. M.; Assimon, V. A.; Gestwicki, J. E.; Brodsky, J. L.; Camacho, C. J.; Bar-Joseph, Z. Predicting Protein Targets for Drug-like Compounds Using Transcriptomics. *PLOS Comput. Biol.* **2018**, *14* (12), e1006651. https://doi.org/10.1371/journal.pcbi.1006651.

(88) Iwata, M.; Sawada, R.; Iwata, H.; Kotera, M.; Reports, Y. Y.-S.; 2017, U. Elucidating the Modes of Action for Bioactive Compounds in a Cell-Specific Manner by Large-Scale Chemically-Induced Transcriptomics. *Sci. Rep.* **2017**, No. 7, 40164.

(89) Kibble, M.; Khan, S. A.; Saarinen, N.; Iorio, F.; Saez-Rodriguez, J.; Mäkelä, S.; Aittokallio, T. Transcriptional Response Networks for Elucidating Mechanisms of Action of Multitargeted Agents. *Drug Discov. Today* **2016**, *21* (7), 1063–1075. https://doi.org/10.1016/J.DRUDIS.2016.03.001.

(90) Iorio, F.; Rittman, T.; Ge, H.; Menden, M.; Saez-Rodriguez, J. Transcriptional Data: A New Gateway to Drug Repositioning? *Drug Discov. Today* **2013**, *18* (7–8), 350–357. https://doi.org/10.1016/J.DRUDIS.2012.07.014.

(91) Lamb, J.; Crawford, E. D.; Peck, D.; Modell, J. W.; Blat, I. C.; Wrobel, M. J.; Lerner, J.; Brunet, J.-P.; Subramanian, A.; Ross, K. N.; Reich, M.; Hieronymus, H.; Wei, G.; Armstrong, S. A.; Haggarty, S. J.; Clemons, P. A.; Wei, R.; Carr, S. A.; Lander, E. S.; Golub, T. R. The Connectivity Map: Using Gene-Expression Signatures to Connect Small Molecules, Genes, and Disease. *Science* **2006**, *313* (5795), 1929–1935. https://doi.org/10.1126/science.1132939.

(92) Saei, A. A.; Beusch, C. M.; Chernobrovkin, A.; Sabatier, P.; Zhang, B.; Tokat, Ü. G.; Stergiou, E.; Gaetani, M.; Végvári, Á.; Zubarev, R. A. ProTargetMiner as a Proteome Signature Library of Anticancer Molecules for Functional Discovery. *Nat. Commun. 2019 101* **2019**, *10* (1), 1–13. https://doi.org/10.1038/s41467-019-13582-8.

(93) Zapalska-Sozoniuk, M.; Chrobak, L.; Kowalczyk, K.; Kankofer, M. Is It Useful to Use Several “Omics” for Obtaining Valuable Results? *Mol. Biol. Reports 2019*

463 **2019**, *46* (3), 3597–3606. https://doi.org/10.1007/S11033-019-04793-9.

(94) Saei, A. A.; Gullberg, H.; Sabatier, P.; Beusch, C. M.; Johansson, K.; Lundgren, B.; Arvidsson, P. I.; Arnér, E. S. J.; Zubarev, R. A. Comprehensive Chemical Proteomics for Target Deconvolution of the Redox Active Drug Auranofin. *Redox Biol.* **2020**, *32*, 101491. https://doi.org/10.1016/J.REDOX.2020.101491.

(95) Hornbeck, P. V.; Kornhauser, J. M.; Tkachev, S.; Zhang, B.; Skrzypek, E.; Murray, B.; Latham, V.; Sullivan, M. PhosphoSitePlus: A Comprehensive Resource for Investigating the Structure and Function of Experimentally Determined Post-Translational Modifications in Man and Mouse. *Nucleic Acids Res.* **2012**, *40* (D1), D261–D270. https://doi.org/10.1093/NAR/GKR1122.

(96) Hollywood, K.; Brison, D. R.; Goodacre, R. Metabolomics: Current Technologies and Future Trends. *Proteomics* **2006**, *6* (17), 4716–4723. https://doi.org/10.1002/PMIC.200600106.

(97) Zhang, L.; Ma, C.; Chao, H.; Long, Y.; Wu, J.; Li, Z.; Ge, X.; Xia, H.; Yin, Y.; Batley, J.; Li, M. Integration of Metabolome and Transcriptome Reveals Flavonoid Accumulation in the Intergeneric Hybrid between Brassica Rapa and Raphanus Sativus. *Sci. Reports 2019 91* **2019**, *9* (1), 1–8. https://doi.org/10.1038/s41598-019-54889-2.

(98) Cavill, R.; Jennen, D.; Kleinjans, J.; Briedé, J. J. Transcriptomic and Metabolomic Data Integration. *Brief. Bioinform.* **2016**, *17* (5), 891–901. https://doi.org/10.1093/BIB/BBV090.

(99) Lu, W.; Su, X.; Klein, M. S.; Lewis, I. A.; Fiehn, O.; Rabinowitz, J. D. Metabolite Measurement: Pitfalls to Avoid and Practices to Follow. *https://doi.org/10.1146/annurev-biochem-061516-044952* **2017**, *86*, 277–304. https://doi.org/10.1146/ANNUREV-BIOCHEM-061516-044952.

(100) Wright Muelas, M.; Roberts, I.; Mughal, F.; O'Hagan, S.; Day, P. J.; Kell, D. B. An Untargeted Metabolomics Strategy to Measure Differences in Metabolite Uptake and Excretion by Mammalian Cell Lines. *Metabolomics 2020 1610* **2020**, *16* (10), 1–12. https://doi.org/10.1007/S11306-020-01725-8.

(101) Lin, Y.; Caldwell, G. W.; Li, Y.; Lang, W.; Masucci, J. Inter-Laboratory Reproducibility of an Untargeted Metabolomics GC–MS Assay for Analysis of Human Plasma. *Sci. Reports 2020 101* **2020**, *10* (1), 1–11.

https://doi.org/10.1038/s41598-020-67939-x.

(102) Bickle, M. The Beautiful Cell: High-Content Screening in Drug Discovery. *Anal. Bioanal. Chem. 2010 3981* **2010**, *398* (1), 219–226. https://doi.org/10.1007/S00216-010-3788-3.

(103) Seal, S.; Yang, H.; Vollmers, L.; Bender, A. Comparison of Cellular Morphological Descriptors and Molecular Fingerprints for the Prediction of Cytotoxicity- And Proliferation-Related Assays. *Chem. Res. Toxicol.* **2021**, *34* (2), 422–437. https://doi.org/10.1021/acs.chemrestox.0c00303.

(104) Bray, M.-A.; Gustafsdottir, S. M.; Rohban, M. H.; Singh, S.; Ljosa, V.; Sokolnicki, K. L.; Bittker, J. A.; Bodycombe, N. E.; Dančík, V.; Hasaka, T. P.; Hon, C. S.; Kemp, M. M.; Li, K.; Walpita, D.; Wawer, M. J.; Golub, T. R.; Schreiber, S. L.; Clemons, P. A.; Shamji, A. F.; Carpenter, A. E. A Dataset of Images and Morphological Profiles of 30 000 Small-Molecule Treatments Using the Cell Painting Assay. *Gigascience* **2017**, *6* (12), 1–5. https://doi.org/10.1093/gigascience/giw014.

(105) Simm, J.; Klambauer, G.; Arany, A.; Steijaert, M.; Wegner, J. K.; Gustin, E.; Chupakhin, V.; Chong, Y. T.; Vialard, J.; Buijnsters, P.; Velter, I.; Vapirev, A.; Singh, S.; Carpenter, A. E.; Wuyts, R.; Hochreiter, S.; Moreau, Y.; Ceulemans, H. Repurposing High-Throughput Image Assays Enables Biological Activity Prediction for Drug Discovery. *Cell Chem. Biol.* **2018**, *25* (5), 611--618.e3. https://doi.org/10.1016/J.CHEMBIOL.2018.01.015.

(106) Gustafsdottir, S. M.; Ljosa, V.; Sokolnicki, K. L.; Anthony Wilson, J.; Walpita, D.; Kemp, M. M.; Petri Seiler, K.; Carrel, H. A.; Golub, T. R.; Schreiber, S. L.; Clemons, P. A.; Carpenter, A. E.; Shamji, A. F. Multiplex Cytological Profiling Assay to Measure Diverse Cellular States. *PLoS One* **2013**, *8* (12), e80999. https://doi.org/10.1371/journal.pone.0080999.

(107) McQuin, C.; Goodman, A.; Chernyshev, V.; Kamentsky, L.; Cimini, B. A.; Karhohs, K. W.; Doan, M.; Ding, L.; Rafelski, S. M.; Thirstrup, D.; Wiegraebe, W.; Singh, S.; Becker, T.; Caicedo, J. C.; Carpenter, A. E. CellProfiler 3.0: Next-Generation Image Processing for Biology. *PLOS Biol.* **2018**, *16* (7), e2005970. https://doi.org/10.1371/journal.pbio.2005970.

(108) Ljosa, V.; Sokolnicki, K. L.; Carpenter, A. E. Annotated High-Throughput

Microscopy Image Sets for Validation. *Nat. Methods 2012 97* **2012**, *9* (7), 637–637. https://doi.org/10.1038/nmeth.2083.

(109) Williams, E.; Moore, J.; Li, S. W.; Rustici, G.; Tarkowska, A.; Chessel, A.; Leo, S.; Antal, B.; Ferguson, R. K.; Sarkans, U.; Brazma, A.; Carazo Salas, R. E.; Swedlow, J. R. Image Data Resource: A Bioimage Data Integration and Publication Platform. *Nat. Methods* **2017**, *14* (8), 775–781. https://doi.org/10.1038/nmeth.4326.

(110) Broad Institute launches academic-industry cell imaging consortium to speed drug discovery and development https://www.broadinstitute.org/news/broad-institute-launches-academic-industry-cell-imaging-consortium-speed-drug-discovery-and (accessed Aug 27, 2021).

(111) Mullard, A. Machine Learning Brings Cell Imaging Promises into Focus. *Nat. Rev. Drug Discov.* **2019**, *18* (9), 653–655. https://doi.org/10.1038/D41573-019-00144-2.

(112) Caicedo, J. C.; Cooper, S.; Heigwer, F.; Warchal, S.; Qiu, P.; Molnar, C.; Vasilevich, A. S.; Barry, J. D.; Bansal, H. S.; Kraus, O.; Wawer, M.; Paavolainen, L.; Herrmann, M. D.; Rohban, M.; Hung, J.; Hennig, H.; Concannon, J.; Smith, I.; Clemons, P. A.; Singh, S.; Rees, P.; Horvath, P.; Linington, R. G.; Carpenter, A. E. Data-Analysis Strategies for Image-Based Cell Profiling. *Nat. Methods* **2017**, *14* (9), 849–863. https://doi.org/10.1038/nmeth.4397.

(113) Caicedo, J. C.; Singh, S.; Carpenter, A. E. Applications in Image-Based Profiling of Perturbations. *Curr. Opin. Biotechnol.* **2016**, *39*, 134–142. https://doi.org/10.1016/J.COPBIO.2016.04.003.

(114) O'Shea, J. J.; Schwartz, D. M.; Villarino, A. V.; Gadina, M.; McInnes, I. B.; Laurence, A. The JAK-STAT Pathway: Impact on Human Disease and Therapeutic Intervention*. *https://doi.org/10.1146/annurev-med-051113-024537* **2015**, *66*, 311–328. https://doi.org/10.1146/ANNUREV-MED-051113-024537.

(115) Sam, S. A.; Teel, J.; Tegge, A. N.; Bharadwaj, A.; Murali, T. M. XTalkDB: A Database of Signaling Pathway Crosstalk. *Nucleic Acids Res.* **2017**, *45* (D1), D432–D439. https://doi.org/10.1093/NAR/GKW1037.

(116) Jassal, B.; Matthews, L.; Viteri, G.; Gong, C.; Lorente, P.; Fabregat, A.;

Sidiropoulos, K.; Cook, J.; Gillespie, M.; Haw, R.; Loney, F.; May, B.; Milacic, M.; Rothfels, K.; Sevilla, C.; Shamovsky, V.; Shorser, S.; Varusai, T.; Weiser, J.; Wu, G.; Stein, L.; Hermjakob, H.; D'Eustachio, P. The Reactome Pathway Knowledgebase. *Nucleic Acids Res.* **2020**, *48* (D1), D498–D503. https://doi.org/10.1093/NAR/GKZ1031.

(117) Kutmon, M.; Riutta, A.; Nunes, N.; Hanspers, K.; Willighagen, E. L.; Bohler, A.; Mélius, J.; Waagmeester, A.; Sinha, S. R.; Miller, R.; Coort, S. L.; Cirillo, E.; Smeets, B.; Evelo, C. T.; Pico, A. R. WikiPathways: Capturing the Full Diversity of Pathway Knowledge. *Nucleic Acids Res.* **2016**, *44* (D1), D488--D494. https://doi.org/10.1093/nar/gkv1024.

(118) Slenter, D. N.; Kutmon, M.; Hanspers, K.; Riutta, A.; Windsor, J.; Nunes, N.; Mélius, J.; Cirillo, E.; Coort, S. L.; Digles, D.; Ehrhart, F.; Giesbertz, P.; Kalafati, M.; Martens, M.; Miller, R.; Nishida, K.; Rieswijk, L.; Waagmeester, A.; Eijssen, L. M. T.; Evelo, C. T.; Pico, A. R.; Willighagen, E. L. WikiPathways: A Multifaceted Pathway Database Bridging Metabolomics to Other Omics Research. *Nucleic Acids Res.* **2018**, *46* (D1), D661--D667. https://doi.org/10.1093/nar/gkx1064.

(119) Kanehisa, M.; Furumichi, M.; Tanabe, M.; Sato, Y.; Morishima, K. KEGG: New Perspectives on Genomes, Pathways, Diseases and Drugs. *Nucleic Acids Res.* **2016**, *45* (D1), D353--D361. https://doi.org/10.1093/nar/gkw1092.

(120) Kanehisa, M.; Sato, Y.; Kawashima, M.; Furumichi, M.; Tanabe, M. KEGG as a Reference Resource for Gene and Protein Annotation. *Nucleic Acids Res.* **2015**, *44* (D1), D457--D462. https://doi.org/10.1093/nar/gkv1070.

(121) Trupp, M.; Altman, T.; Fulcher, C. A.; Caspi, R.; Krummenacker, M.; Paley, S.; Karp, P. D. Beyond the Genome (BTG) Is a (PGDB) Pathway Genome Database: HumanCyc. *Genome Biol. 2010 111* **2010**, *11* (1), 1–1. https://doi.org/10.1186/GB-2010-11-S1-O12.

(122) Rodchenkov, I.; Babur, O.; Luna, A.; Aksoy, B. A.; Wong, J. V; Fong, D.; Franz, M.; Siper, M. C.; Cheung, M.; Wrana, M.; Mistry, H.; Mosier, L.; Dlin, J.; Wen, Q.; O'Callaghan, C.; Li, W.; Elder, G.; Smith, P. T.; Dallago, C.; Cerami, E.; Gross, B.; Dogrusoz, U.; Demir, E.; Bader, G. D.; Sander, C. Pathway Commons 2019 Update: Integration, Analysis and Exploration of Pathway Data. *Nucleic*

Acids Res. **2020**, *48* (D1), D489–D497. https://doi.org/10.1093/NAR/GKZ946.

(123) Geer, L. Y.; Marchler-Bauer, A.; Geer, R. C.; Han, L.; He, J.; He, S.; Liu, C.; Shi, W.; Bryant, S. H. The NCBI BioSystems Database. *Nucleic Acids Res.* **2010**, *38* (suppl_1), D492--D496. https://doi.org/10.1093/nar/gkp858.

(124) Carbon, S.; Douglass, E.; Dunn, N.; Good, B.; Harris, N. L.; Lewis, S. E.; Mungall, C. J.; Basu, S.; Chisholm, R. L.; Dodson, R. J.; Hartline, E.; Fey, P.; Thomas, P. D.; Albou, L. P.; Ebert, D.; Kesling, M. J.; Mi, H.; Muruganujan, A.; Huang, X.; Poudel, S.; Mushayahama, T.; Hu, J. C.; LaBonte, S. A.; Siegele, D. A.; Antonazzo, G.; Attrill, H.; Brown, N. H.; Fexova, S.; Garapati, P.; Jones, T. E. M.; Marygold, S. J.; Millburn, G. H.; Rey, A. J.; Trovisco, V.; Dos Santos, G.; Emmert, D. B.; Falls, K.; Zhou, P.; Goodman, J. L.; Strelets, V. B.; Thurmond, J.; Courtot, M.; Osumi, D. S.; Parkinson, H.; Roncaglia, P.; Acencio, M. L.; Kuiper, M.; Lreid, A.; Logie, C.; Lovering, R. C.; Huntley, R. P.; Denny, P.; Campbell, N. H.; Kramarz, B.; Acquaah, V.; Ahmad, S. H.; Chen, H.; Rawson, J. H.; Chibucos, M. C.; Giglio, M.; Nadendla, S.; Tauber, R.; Duesbury, M. J.; Del, N. T.; Meldal, B. H. M.; Perfetto, L.; Porras, P.; Orchard, S.; Shrivastava, A.; Xie, Z.; Chang, H. Y.; Finn, R. D.; Mitchell, A. L.; Rawlings, N. D.; Richardson, L.; Sangrador-Vegas, A.; Blake, J. A.; Christie, K. R.; Dolan, M. E.; Drabkin, H. J.; Hill, D. P.; Ni, L.; Sitnikov, D.; Harris, M. A.; Oliver, S. G.; Rutherford, K.; Wood, V.; Hayles, J.; Bahler, J.; Lock, A.; Bolton, E. R.; De Pons, J.; Dwinell, M.; Hayman, G. T.; Laulederkind, S. J. F.; Shimoyama, M.; Tutaj, M.; Wang, S. J.; D'Eustachio, P.; Matthews, L.; Balhoff, J. P.; Aleksander, S. A.; Binkley, G.; Dunn, B. L.; Cherry, J. M.; Engel, S. R.; Gondwe, F.; Karra, K.; MacPherson, K. A.; Miyasato, S. R.; Nash, R. S.; Ng, P. C.; Sheppard, T. K.; Shrivatsav Vp, A.; Simison, M.; Skrzypek, M. S.; Weng, S.; Wong, E. D.; Feuermann, M.; Gaudet, P.; Bakker, E.; Berardini, T. Z.; Reiser, L.; Subramaniam, S.; Huala, E.; Arighi, C.; Auchincloss, A.; Axelsen, K.; Argoud, G. P.; Bateman, A.; Bely, B.; Blatter, M. C.; Boutet, E.; Breuza, L.; Bridge, A.; Britto, R.; Bye-A-Jee, H.; Casals-Casas, C.; Coudert, E.; Estreicher, A.; Famiglietti, L.; Garmiri, P.; Georghiou, G.; Gos, A.; Gruaz-Gumowski, N.; Hatton-Ellis, E.; Hinz, U.; Hulo, C.; Ignatchenko, A.; Jungo, F.; Keller, G.; Laiho, K.; Lemercier, P.; Lieberherr, D.; Lussi, Y.; Mac-Dougall, A.; Magrane, M.; Martin, M. J.; Masson, P.; Natale, D. A.; Hyka, N. N.;

Pedruzzi, I.; Pichler, K.; Poux, S.; Rivoire, C.; Rodriguez-Lopez, M.; Sawford, T.; Speretta, E.; Shypitsyna, A.; Stutz, A.; Sundaram, S.; Tognolli, M.; Tyagi, N.; Warner, K.; Zaru, R.; Wu, C.; Chan, J.; Cho, J.; Gao, S.; Grove, C.; Harrison, M. C.; Howe, K.; Lee, R.; Mendel, J.; Muller, H. M.; Raciti, D.; Van Auken, K.; Berriman, M.; Stein, L.; Sternberg, P. W.; Howe, D.; Toro, S.; Westerfield, M. The Gene Ontology Resource: 20 Years and Still GOing Strong. *Nucleic Acids Res.* **2019**, *47* (D1), D330–D338. https://doi.org/10.1093/nar/gky1055.

(125) Jantzen, S. G.; Sutherland, B. J.; Minkley, D. R.; Koop, B. F. GO Trimming: Systematically Reducing Redundancy in Large Gene Ontology Datasets. *BMC Res. Notes 2011 41* **2011**, *4* (1), 1–9. https://doi.org/10.1186/1756-0500-4-267.

(126) Supek, F.; Bošnjak, M.; Škunca, N.; Šmuc, T. REVIGO Summarizes and Visualizes Long Lists of Gene Ontology Terms. *PLoS One* **2011**, *6* (7), e21800. https://doi.org/10.1371/journal.pone.0021800.

(127) Klopfenstein, D. V.; Zhang, L.; Pedersen, B. S.; Ramírez, F.; Warwick Vesztrocy, A.; Naldi, A.; Mungall, C. J.; Yunes, J. M.; Botvinnik, O.; Weigel, M.; Dampier, W.; Dessimoz, C.; Flick, P.; Tang, H. GOATOOLS: A Python Library for Gene Ontology Analyses. *Sci. Reports 2018 81* **2018**, *8* (1), 1–17. https://doi.org/10.1038/s41598-018-28948-z.

(128) Domingo-Fernández, D.; Mubeen, S.; Marín-Llaó, J.; Hoyt, C. T.; Hofmann-Apitius, M. PathMe: Merging and Exploring Mechanistic Pathway Knowledge. *BMC Bioinforma. 2019 201* **2019**, *20* (1), 1–12. https://doi.org/10.1186/S12859-019-2863-9.

(129) Wiwie, C.; Baumbach, J.; Röttger, R. Comparing the Performance of Biomedical Clustering Methods. *Nat. Methods 2015 1211* **2015**, *12* (11), 1033–1038. https://doi.org/10.1038/nmeth.3583.

(130) Karim, M. R.; Beyan, O.; Zappa, A.; Costa, I. G.; Rebholz-Schuhmann, D.; Cochez, M.; Decker, S. Deep Learning-Based Clustering Approaches for Bioinformatics. *Brief. Bioinform.* **2021**, *22* (1), 393–415. https://doi.org/10.1093/BIB/BBZ170.

(131) Kassambara, A. *Practical Guide to Cluster Analysis in R: Unsupervised Machine Learning*, 1st ed.; Sthda, 2017.

(132) Xu, D.; Tian, Y. A Comprehensive Survey of Clustering Algorithms. *Ann. Data*

Sci. 2015 22 **2015**, *2* (2), 165–193. https://doi.org/10.1007/S40745-015-0040-1.

(133) Way, G. P.; Natoli, T.; Adeboye, A.; Litichevskiy, L.; Yang, A.; Lu, X.; Caicedo, J. C.; Cimini, B. A.; Karhohs, K.; Logan, D. J.; Rohban, M.; Kost-Alimova, M.; Hartland, K.; Bornholdt, M.; Chandrasekaran, N.; Haghighi, M.; Singh, S.; Subramanian, A.; Carpenter, A. E. Morphology and Gene Expression Profiling Provide Complementary Information for Mapping Cell State. *bioRxiv* **2021**, 2021.10.21.465335. https://doi.org/10.1101/2021.10.21.465335.

(134) Patel-Murray, N. L.; Adam, M.; Huynh, N.; Wassie, B. T.; Milani, P.; Fraenkel, E. A Multi-Omics Interpretable Machine Learning Model Reveals Modes of Action of Small Molecules. *Sci. Reports 2020 101* **2020**, *10* (1), 1–14. https://doi.org/10.1038/s41598-020-57691-7.

(135) Argelaguet, R.; Velten, B.; Arnol, D.; Dietrich, S.; Zenz, T.; Marioni, J. C.; Buettner, F.; Huber, W.; Stegle, O. Multi-Omics Factor Analysis—a Framework for Unsupervised Integration of Multi-Omics Data Sets. *Mol. Syst. Biol.* **2018**, *14* (6), e8124. https://doi.org/10.15252/MSB.20178124.

(136) Klami, A.; Virtanen, S.; Leppaaho, E.; Kaski, S. Group Factor Analysis. *IEEE Trans. Neural Networks Learn. Syst.* **2015**, *26* (9), 2136–2147. https://doi.org/10.1109/TNNLS.2014.2376974.

(137) Khan, S. A.; Virtanen, S.; Kallioniemi, O. P.; Wennerberg, K.; Poso, A.; Kaski, S. Identification of Structural Features in Chemicals Associated with Cancer Drug Response: A Systematic Data-Driven Analysis. *Bioinformatics* **2014**, *30* (17), i497--i504. https://doi.org/10.1093/bioinformatics/btu456.

(138) Chen, T.; Tyagi, S. Integrative Computational Epigenomics to Build Data-Driven Gene Regulation Hypotheses. *Gigascience* **2020**, *9* (6), 1–13. https://doi.org/10.1093/GIGASCIENCE/GIAA064.

(139) Khan, S. A.; Virtanen, S.; Kallioniemi, O. P.; Wennerberg, K.; Poso, A.; Kaski, S. Identification of Structural Features in Chemicals Associated with Cancer Drug Response: A Systematic Data-Driven Analysis. *Bioinformatics* **2014**, *30* (17), i497–i504. https://doi.org/10.1093/BIOINFORMATICS/BTU456.

(140) Rivello, F.; van Buijtenen, E.; Matuła, K.; van Buggenum, J. A. G. L.; Vink, P.; van Eenennaam, H.; Mulder, K. W.; Huck, W. T. S. Single-Cell Intracellular Epitope and Transcript Detection Reveals Signal Transduction Dynamics. *Cell*

Reports Methods **2021**, *1* (5), 100070. https://doi.org/10.1016/J.CRMETH.2021.100070.

(141) Schenone, M.; Dančík, V.; Wagner, B. K.; Clemons, P. A. Target Identification and Mechanism of Action in Chemical Biology and Drug Discovery. *Nat. Chem. Biol.* **2013**, *9* (4), 232–240. https://doi.org/10.1038/nchembio.1199.

(142) Sheridan, R. P. Time-Split Cross-Validation as a Method for Estimating the Goodness of Prospective Prediction. *J. Chem. Inf. Model.* **2013**, *53* (4), 783–790. https://doi.org/10.1021/CI400084K.

(143) Brownlee, J. *Master Machine Learning Algorithms: Discover How They Work and Implement Them from Scratch*, 1.9.; 2017.

(144) Tsaioun, K.; Kates, S. A. ADMET for Medicinal Chemists: A Practical Guide. **2011**.

(145) Dehmer, M.; Varmuza, K.; Bonchev, D. *Statistical Modelling of Molecular Descriptors in QSAR/QSPR*; Wiley-Blackwell, 2012.

(146) Mitchell B.O., J. B. O. Machine Learning Methods in Chemoinformatics. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2014**, *4* (5), 468–481. https://doi.org/10.1002/WCMS.1183.

(147) Lo, Y. C.; Rensi, S. E.; Torng, W.; Altman, R. B. Machine Learning in Chemoinformatics and Drug Discovery. *Drug Discov. Today* **2018**, *23* (8), 1538–1546. https://doi.org/10.1016/J.DRUDIS.2018.05.010.

(148) Boser, B.; Guyon, I.; ACM, V. V.-P. of the 5th A. A Training Algorithm for Optimal Margin Classifiers. *gautampendse.com*.

(149) Doddareddy, M. R.; Klaasse, E. C.; IJzerman, A. P. Prospective Validation of a Comprehensive in Silico HERG Model and Its Applications to Commercial Compound and Drug Databases. **2010**.

(150) Dong, X.; Jiang, C.; Hu, H.; Yan, J.; Chen, J.; Hu, Y. QSAR Study of Akt/Protein Kinase B (PKB) Inhibitors Using Support Vector Machine. *Eur. J.* **2009**.

(151) Basak, D.; Pal, S.; Patranabis, D. C. Support Vector Regression. *Neural Inf. Process.* **2007**.

(152) Khan, S. A.; Faisal, A.; Mpindi, J.; Parkkinen, J. A.; Kalliokoski, T.; Poso, A.; Kallioniemi, O. P.; Wennerberg, K.; Kaski, S. Comprehensive Data-Driven Analysis of the Impact of Chemoinformatic Structure on the Genome-Wide

Biological Response Profiles of Cancer Cells to 1159 Drugs. *BMC Bioinformatics* **2012**, *13* (1), 112. https://doi.org/10.1186/1471-2105-13-112.

(153) Wang, J.; Hou, T. Advances in Computationally Modeling Human Oral Bioavailability. *Adv. Drug Deliv. Rev.* **2015**, *86*, 11–16. https://doi.org/10.1016/J.ADDR.2015.01.001.

(154) Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.* https://doi.org/10.1145/2939672.

(155) Wang, S.; Liu, S.; Zhang, J.; Che, X.; Yuan, Y.; Wang, Z.; Kong, D. A New Method of Diesel Fuel Brands Identification: SMOTE Oversampling Combined with XGBoost Ensemble Learning. *Fuel* **2020**, *282*, 118848. https://doi.org/10.1016/J.FUEL.2020.118848.

(156) He, J.; Hao, Y.; Wang, X. An Interpretable Aid Decision-Making Model for Flag State Control Ship Detention Based on SMOTE and XGBoost. *J. Mar. Sci. Eng. 2021, Vol. 9, Page 156* **2021**, *9* (2), 156. https://doi.org/10.3390/JMSE9020156.

(157) Simm, J.; Arany, A.; Zakeri, P.; Haber, T.; Wegner, J. K.; Chupakhin, V.; Ceulemans, H.; Moreau, Y. Macau: Scalable Bayesian Multi-Relational Factorization with Side Information Using MCMC. **2015**, 1–13.

(158) Wen, M.; Zhang, Z.; Niu, S.; Sha, H.; Yang, R.; Yun, Y.; Lu, H. Deep-Learning-Based Drug-Target Interaction Prediction. *J. Proteome Res.* **2017**, *16* (4), 1401–1409. https://doi.org/10.1021/acs.jproteome.6b00618.

(159) Mayr, A.; Klambauer, G.; Unterthiner, T.; Steijaert, M.; Wegner, J. K.; Ceulemans, H.; Clevert, D. A.; Hochreiter, S. Large-Scale Comparison of Machine Learning Methods for Drug Target Prediction on ChEMBL. *Chem. Sci.* **2018**, *9* (24), 5441–5451. https://doi.org/10.1039/c8sc00148k.

(160) Robinson, M. C.; Glen, R. C.; Lee, A. A. Validating the Validation: Reanalyzing a Large-Scale Comparison of Deep Learning and Machine Learning Models for Bioactivity Prediction. *J. Comput. Aided. Mol. Des.* **2020**, *34*, 717–730. https://doi.org/10.1007/s10822-019-00274-0.

(161) Bender, A.; Cortés-Ciriano, I. Artificial Intelligence in Drug Discovery: What Is Realistic, What Are Illusions? Part 1: Ways to Make an Impact, and Why We Are Not There Yet. *Drug Discov. Today* **2021**, *26* (2), 511–524.

https://doi.org/10.1016/J.DRUDIS.2020.12.009.

(162) Aniceto, N.; Freitas, A. A.; Bender, A.; Ghafourian, T. A Novel Applicability Domain Technique for Mapping Predictive Reliability across the Chemical Space of a QSAR: Reliability-Density Neighbourhood. *J.* **2016**.

(163) Mervin, L.; Afzal, A. M.; Engkvist, O.; Bender, A. A Comparison of Scaling Methods to Obtain Calibrated Probabilities of Activity for Ligand-Target Predictions. **2020**, No. 1. https://doi.org/10.26434/chemrxiv.11526132.v1.

(164) Bosc, N.; Atkinson, F.; Felix, E.; Gaulton, A.; Hersey, A.; Leach, A. R. Large Scale Comparison of QSAR and Conformal Prediction Methods and Their Applications in Drug Discovery. *J. Cheminformatics 2019 111* **2019**, *11* (1), 1–16. https://doi.org/10.1186/S13321-018-0325-4.

(165) Bajorath, J.; Peltason, L.; Wawer, M.; Guha, R.; Lajiness, M. S.; Van Drie, J. H. Navigating Structure-Activity Landscapes. *Drug Discov. Today* **2009**, *14* (13–14), 698–705. https://doi.org/10.1016/j.drudis.2009.04.003.

(166) Hofmarcher, M.; Rumetshofer, E.; Clevert, D. A.; Hochreiter, S.; Klambauer, G. Accurate Prediction of Biological Assays with High-Throughput Microscopy Images and Convolutional Networks. *J. Chem. Inf. Model.* **2019**, *59* (3), 1163–1171. https://doi.org/10.1021/acs.jcim.8b00670.

(167) Lapins, M.; Spjuth, O. Evaluation of Gene Expression and Phenotypic Profiling Data as Quantitative Descriptors for Predicting Drug Targets and Mechanisms of Action. *bioRxiv* **2019**, 580654. https://doi.org/10.1101/580654.

(168) García-Campos, M. A.; Espinal-Enríquez, J.; Hernández-Lemus, E. Pathway Analysis: State of the Art. *Front. Physiol.* **2015**, *6*, 383. https://doi.org/10.3389/fphys.2015.00383.

(169) Yuryev, A. Pathway Analysis for Drug Discovery : Computational Infrastructure and Applications. **2008**, 303.

(170) Wang, B.; Li, R.; Perrizo, W. (William); IGI Global. *Big Data Analytics in Bioinformatics and Healthcare*; 2015. https://doi.org/10.4018/978-1-4666-6611-5.

(171) Wang, B.; Li, R.; Perrizo, W. Big Data Analytics in Bioinformatics and Healthcare. *Big Data Anal. Bioinforma. Healthc.* **2014**, 1–528. https://doi.org/10.4018/978-1-4666-6611-5.

(172) Khatri, P.; Sirota, M.; Butte, A. J. Ten Years of Pathway Analysis: Current Approaches and Outstanding Challenges. *PLoS Comput. Biol.* **2012**, *8* (2), e1002375. https://doi.org/10.1371/journal.pcbi.1002375.

(173) Liggi, S.; Drakakis, G.; Hendry, A. E.; Hanson, K. M.; Brewerton, S. C.; Wheeler, G. N.; Bodkin, M. J.; Evans, D. A.; Bender, A. Extensions to In Silico Bioactivity Predictions Using Pathway Annotations and Differential Pharmacology Analysis: Application to *Xenopus Laevis*Phenotypic Readouts. *Mol. Inform.* **2013**, *32* (11–12), 1009–1024. https://doi.org/10.1002/minf.201300102.

(174) Blanco, M. J.; Gardinier, K. M. New Chemical Modalities and Strategic Thinking in Early Drug Discovery. *ACS Med. Chem. Lett.* **2020**, *11* (3), 228–231. https://doi.org/10.1021/ACSMEDCHEMLETT.9B00582.

(175) Pettersson, M.; Crews, C. M. PROteolysis TArgeting Chimeras (PROTACs) — Past, Present and Future. *Drug Discov. Today Technol.* **2019**, *31*, 15–27. https://doi.org/10.1016/j.ddtec.2019.01.002.

(176) Sun, X.; Gao, H.; Yang, Y.; He, M.; Wu, Y.; Song, Y.; Tong, Y.; Rao, Y. Protacs: Great Opportunities for Academia and Industry. *Signal Transduct. Target. Ther.* **2019**, *4* (1). https://doi.org/10.1038/s41392-019-0101-6.

(177) Bond, M. J.; Crews, C. M. Proteolysis Targeting Chimeras (PROTACs) Come of Age: Entering the Third Decade of Targeted Protein Degradation. *RSC Chem. Biol.* **2021**, *2* (3), 725–742. https://doi.org/10.1039/d1cb00011j.

(178) Zou, Y.; Ma, D.; Wang, Y. The PROTAC Technology in Drug Development. *Cell Biochem. Funct.* **2019**, *37* (1), 21–30. https://doi.org/10.1002/cbf.3369.

(179) Scheepstra, M.; Hekking, K. F. W.; van Hijfte, L.; Folmer, R. H. A. Bivalent Ligands for Protein Degradation in Drug Discovery. *Comput. Struct. Biotechnol. J.* **2019**, *17*, 160–176. https://doi.org/10.1016/j.csbj.2019.01.006.

(180) Lai, A. C.; Crews, C. M. Induced Protein Degradation: An Emerging Drug Discovery Paradigm. *Nat. Rev. Drug Discov.* **2017**, *16* (2), 101–114. https://doi.org/10.1038/nrd.2016.211.

(181) Sakamoto, K. M.; Kim, K. B.; Kumagai, A.; Mercurio, F.; Crews, C. M.; Deshaies, R. J. Protacs: Chimeric Molecules That Target Proteins to the Skp1–Cullin–F Box Complex for Ubiquitination and Degradation. *Proc. Natl. Acad. Sci.* **2001**, *98* (15), 8554–8559. https://doi.org/10.1073/PNAS.141230798.

(182) Sakamoto, K. M.; Kim, K. B.; Verma, R.; Ransick, A.; Stein, B.; Crews, C. M.; Deshaies, R. J. Development of Protacs to Target Cancer-Promoting Proteins for Ubiquitination and Degradation *. *Mol. Cell. Proteomics* **2003**, *2* (12), 1350–1358. https://doi.org/10.1074/MCP.T300009-MCP200.

(183) Schneekloth, J. S.; Fonseca, F. N.; Koldobskiy, M.; Mandal, A.; Deshaies, R.; Sakamoto, K.; Crews, C. M. Chemical Genetic Control of Protein Levels: Selective in Vivo Targeted Degradation. *J. Am. Chem. Soc.* **2004**, *126* (12), 3748–3754. https://doi.org/10.1021/JA039025Z/SUPPL_FILE/JA039025ZSI20040128_034802.PDF.

(184) Lee, H.; Puppala, D.; Choi, E. Y.; Swanson, H.; Kim, K. B. Targeted Degradation of the Aryl Hydrocarbon Receptor by the PROTAC Approach: A Useful Chemical Genetic Tool. *ChemBioChem* **2007**, *8* (17), 2058–2062. https://doi.org/10.1002/CBIC.200700438.

(185) Schneekloth, A. R.; Pucheault, M.; Tae, H. S.; Crews, C. M. Targeted Intracellular Protein Degradation Induced by a Small Molecule: En Route to Chemical Proteomics. *Bioorg. Med. Chem. Lett.* **2008**, *18* (22), 5904–5908. https://doi.org/10.1016/J.BMCL.2008.07.114.

(186) Buckley, D. L.; Gustafson, J. L.; Van-Molle, I.; Roth, A. G.; Tae, H. S.; Gareiss, P. C.; Jorgensen, W. L.; Ciulli, A.; Crews, C. M. Small-Molecule Inhibitors of the Interaction between the E3 Ligase VHL and HIF1α. *Angew. Chemie Int. Ed.* **2012**, *51* (46), 11463–11467. https://doi.org/10.1002/ANIE.201206231.

(187) Buckley, D. L.; Van Molle, I.; Gareiss, P. C.; Tae, H. S.; Michel, J.; Noblin, D. J.; Jorgensen, W. L.; Ciulli, A.; Crews, C. M. Targeting the von Hippel-Lindau E3 Ubiquitin Ligase Using Small Molecules to Disrupt the VHL/HIF-1α Interaction. *J. Am. Chem. Soc.* **2012**, *134* (10), 4465–4468. https://doi.org/10.1021/JA209924V/SUPPL_FILE/JA209924V_SI_001.PDF.

(188) Van Molle, I.; Thomann, A.; Buckley, D. L.; So, E. C.; Lang, S.; Crews, C. M.; Ciulli, A. Dissecting Fragment-Based Lead Discovery at the von Hippel-Lindau Protein:Hypoxia Inducible Factor 1α Protein-Protein Interface. *Chem. Biol.* **2012**, *19* (10), 1300–1312. https://doi.org/10.1016/J.CHEMBIOL.2012.08.015.

(189) Girardini, M.; Maniaci, C.; Hughes, S. J.; Testa, A.; Ciulli, A. Cereblon versus

VHL: Hijacking E3 Ligases against Each Other Using PROTACs. *Bioorg. Med. Chem.* **2019**, *27* (12), 2466–2479. https://doi.org/10.1016/J.BMC.2019.02.048.

(190) Testa, A.; Lucas, X.; Castro, G. V.; Chan, K. H.; Wright, J. E.; Runcie, A. C.; Gadd, M. S.; Harrison, W. T. A.; Ko, E. J.; Fletcher, D.; Ciulli, A. 3-Fluoro-4-Hydroxyprolines: Synthesis, Conformational Analysis, and Stereoselective Recognition by the VHL E3 Ubiquitin Ligase for Targeted Protein Degradation. *J. Am. Chem. Soc.* **2018**, *140* (29), 9299–9313. https://doi.org/10.1021/JACS.8B05807/SUPPL_FILE/JA8B05807_SI_005.CIF.

(191) Soares, P.; Lucas, X.; Ciulli, A. Thioamide Substitution to Probe the Hydroxyproline Recognition of VHL Ligands. *Bioorg. Med. Chem.* **2018**, *26* (11), 2992–2995. https://doi.org/10.1016/J.BMC.2018.03.034.

(192) Soares, P.; Gadd, M. S.; Frost, J.; Galdeano, C.; Ellis, L.; Epemolu, O.; Rocha, S.; Read, K. D.; Ciulli, A. Group-Based Optimization of Potent and Cell-Active Inhibitors of the von Hippel-Lindau (VHL) E3 Ubiquitin Ligase: Structure-Activity Relationships Leading to the Chemical Probe (2S,4R)-1-((S)-2-(1-Cyanocyclopropanecarboxamido)-3,3-Dimethylbutanoyl)-4-Hydroxy-N-(4-(4-Methylthiazol-5-Yl)Benzyl)Pyrrolidine-2-Carboxamide (VH298). *J. Med. Chem.* **2018**, *61* (2), 599–618. https://doi.org/10.1021/ACS.JMEDCHEM.7B00675/SUPPL_FILE/JM7B00675_SI_001.PDF.

(193) Frost, J.; Galdeano, C.; Soares, P.; Gadd, M. S.; Grzes, K. M.; Ellis, L.; Epemolu, O.; Shimamura, S.; Bantscheff, M.; Grandi, P.; Read, K. D.; Cantrell, D. A.; Rocha, S.; Ciulli, A. Potent and Selective Chemical Probe of Hypoxic Signalling Downstream of HIF-α Hydroxylation via VHL Inhibition. *Nat. Commun. 2016 71* **2016**, *7* (1), 1–12. https://doi.org/10.1038/ncomms13312.

(194) Ito, T.; Ando, H.; Suzuki, T.; Ogura, T.; Hotta, K.; Imamura, Y.; Yamaguchi, Y.; Handa, H. Identification of a Primary Target of Thalidomide Teratogenicity. *Science (80-.).* **2010**, *327* (5971), 1345–1350. https://doi.org/10.1126/SCIENCE.1177319/SUPPL_FILE/ITO.SOM.REVISION1.PDF.

(195) Lopez-Girona, A.; Mendy, D.; Ito, T.; Miller, K.; Gandhi, A. K.; Kang, J.; Karasawa, S.; Carmel, G.; Jackson, P.; Abbasian, M.; Mahmoudi, A.; Cathers,

B.; Rychak, E.; Gaidarova, S.; Chen, R.; Schafer, P. H.; Handa, H.; Daniel, T. O.; Evans, J. F.; Chopra, R. Cereblon Is a Direct Protein Target for Immunomodulatory and Antiproliferative Activities of Lenalidomide and Pomalidomide. *Leuk. 2012 2611* **2012**, *26* (11), 2326–2335. https://doi.org/10.1038/leu.2012.119.

(196) Fischer, E. S.; Böhm, K.; Lydeard, J. R.; Yang, H.; Stadler, M. B.; Cavadini, S.; Nagel, J.; Serluca, F.; Acker, V.; Lingaraju, G. M.; Tichkule, R. B.; Schebesta, M.; Forrester, W. C.; Schirle, M.; Hassiepen, U.; Ottl, J.; Hild, M.; Beckwith, R. E. J.; Harper, J. W.; Jenkins, J. L.; Thomä, N. H. Structure of the DDB1–CRBN E3 Ubiquitin Ligase in Complex with Thalidomide. *Nat. 2014 5127512* **2014**, *512* (7512), 49–53. https://doi.org/10.1038/nature13527.

(197) Krönke, J.; Udeshi, N. D.; Narla, A.; Grauman, P.; Hurst, S. N.; McConkey, M.; Svinkina, T.; Heckl, D.; Comer, E.; Li, X.; Ciarlo, C.; Hartman, E.; Munshi, N.; Schenone, M.; Schreiber, S. L.; Carr, S. A.; Ebert, B. L. Lenalidomide Causes Selective Degradation of IKZF1 and IKZF3 in Multiple Myeloma Cells. *Science (80-.).* **2014**, *343* (6168), 301–305. https://doi.org/10.1126/SCIENCE.1244851/SUPPL_FILE/PAP.PDF.

(198) Petzold, G.; Fischer, E. S.; Thomä, N. H. Structural Basis of Lenalidomide-Induced CK1α Degradation by the CRL4CRBN Ubiquitin Ligase. *Nat. 2016 5327597* **2016**, *532* (7597), 127–130. https://doi.org/10.1038/nature16979.

(199) Krönke, J.; Fink, E. C.; Hollenbach, P. W.; MacBeth, K. J.; Hurst, S. N.; Udeshi, N. D.; Chamberlain, P. P.; Mani, D. R.; Man, H. W.; Gandhi, A. K.; Svinkina, T.; Schneider, R. K.; McConkey, M.; Järås, M.; Griffiths, E.; Wetzler, M.; Bullinger, L.; Cathers, B. E.; Carr, S. A.; Chopra, R.; Ebert, B. L. Lenalidomide Induces Ubiquitination and Degradation of CK1α in Del(5q) MDS. *Nat. 2015 5237559* **2015**, *523* (7559), 183–188. https://doi.org/10.1038/nature14610.

(200) Gandhi, A. K.; Kang, J.; Havens, C. G.; Conklin, T.; Ning, Y.; Wu, L.; Ito, T.; Ando, H.; Waldman, M. F.; Thakurta, A.; Klippel, A.; Handa, H.; Daniel, T. O.; Schafer, P. H.; Chopra, R. Immunomodulatory Agents Lenalidomide and Pomalidomide Co-Stimulate T Cells by Inducing Degradation of T Cell Repressors Ikaros and Aiolos via Modulation of the E3 Ubiquitin Ligase Complex CRL4CRBN. *Br. J. Haematol.* **2014**, *164* (6), 811–821.

https://doi.org/10.1111/BJH.12708.

(201) Bondeson, D. P.; Mares, A.; Smith, I. E. D.; Ko, E.; Campos, S.; Miah, A. H.; Mulholland, K. E.; Routly, N.; Buckley, D. L.; Gustafson, J. L.; Zinn, N.; Grandi, P.; Shimamura, S.; Bergamini, G.; Faelth-Savitski, M.; Bantscheff, M.; Cox, C.; Gordon, D. A.; Willard, R. R.; Flanagan, J. J.; Casillas, L. N.; Votta, B. J.; Den Besten, W.; Famm, K.; Kruidenier, L.; Carter, P. S.; Harling, J. D.; Churcher, I.; Crews, C. M. Catalytic in Vivo Protein Knockdown by Small-Molecule PROTACs. *Nat. Chem. Biol. 2015 118* **2015**, *11* (8), 611–617. https://doi.org/10.1038/nchembio.1858.

(202) Lebraud, H.; Wright, D. J.; Johnson, C. N.; Heightman, T. D. Protein Degradation by In-Cell Self-Assembly of Proteolysis Targeting Chimeras. *ACS Cent. Sci.* **2016**, *2* (12), 927–934. https://doi.org/10.1021/ACSCENTSCI.6B00280/SUPPL_FILE/OC6B00280_SI_001.PDF.

(203) Mullard, A. Targeted Protein Degraders Crowd into the Clinic. *Nat. Rev. Drug Discov.* **2021**, *20* (4), 247–250. https://doi.org/10.1038/D41573-021-00052-4.

(204) Mullard, A. First Targeted Protein Degrader Hits the Clinic. *Nat. Rev. Drug Discov.* **2019**. https://doi.org/10.1038/D41573-019-00043-6.

(205) Moreau, K.; Coen, M.; Zhang, A. X.; Pachl, F.; Castaldi, M. P.; Dahl, G.; Boyd, H.; Scott, C.; Newham, P. Proteolysis-Targeting Chimeras in Drug Development: A Safety Perspective. *Br. J. Pharmacol.* **2020**, *177* (8), 1709–1718. https://doi.org/10.1111/bph.15014.

(206) Hsu, J. H. R.; Rasmusson, T.; Robinson, J.; Pachl, F.; Read, J.; Kawatkar, S.; O' Donovan, D. H.; Bagal, S.; Code, E.; Rawlins, P.; Argyrou, A.; Tomlinson, R.; Gao, N.; Zhu, X.; Chiarparin, E.; Jacques, K.; Shen, M.; Woods, H.; Bednarski, E.; Wilson, D. M.; Drew, L.; Castaldi, M. P.; Fawell, S.; Bloecher, A. EED-Targeted PROTACs Degrade EED, EZH2, and SUZ12 in the PRC2 Complex. *Cell Chem. Biol.* **2020**, *27* (1), 41-46.e17. https://doi.org/10.1016/J.CHEMBIOL.2019.11.004.

(207) Donovan, K. A.; An, J.; Nowak, R. P.; Yuan, J. C.; Fink, E. C.; Berry, B. C.; Ebert, B. L.; Fischer, E. S. Thalidomide Promotes Degradation of SALL4, a Transcription Factor Implicated in Duane Radial Ray Syndrome. *Elife* **2018**, *7*.

https://doi.org/10.7554/ELIFE.38430.

(208) Matyskiela, M. E.; Couto, S.; Zheng, X.; Lu, G.; Hui, J.; Stamp, K.; Drew, C.; Ren, Y.; Wang, M.; Carpenter, A.; Lee, C. W.; Clayton, T.; Fang, W.; Lu, C. C.; Riley, M.; Abdubek, P.; Blease, K.; Hartke, J.; Kumar, G.; Vessey, R.; Rolfe, M.; Hamann, L. G.; Chamberlain, P. P. SALL4 Mediates Teratogenicity as a Thalidomide-Dependent Cereblon Substrate. *Nat. Chem. Biol. 2018 1410* **2018**, *14* (10), 981–987. https://doi.org/10.1038/s41589-018-0129-x.

(209) John, L. B.; Ward, A. C. The Ikaros Gene Family: Transcriptional Regulators of Hematopoiesis and Immunity. *Mol. Immunol.* **2011**, *48* (9–10), 1272–1278. https://doi.org/10.1016/J.MOLIMM.2011.03.006.

(210) Zorba, A.; Nguyen, C.; Xu, Y.; Starr, J.; Borzilleri, K.; Smith, J.; Zhu, H.; Farley, K. A.; Ding, W. D.; Schiemer, J.; Feng, X.; Chang, J. S.; Uccello, D. P.; Young, J. A.; Garcia-Irrizary, C. N.; Czabaniuk, L.; Schuff, B.; Oliver, R.; Montgomery, J.; Hayward, M. M.; Coe, J.; Chen, J.; Niosi, M.; Luthra, S.; Shah, J. C.; El-Kattan, A.; Qiu, X.; West, G. M.; Noe, M. C.; Shanmugasundaram, V.; Gilbert, A. M.; Brown, M. F.; Calabrese, M. F. Delineating the Role of Cooperativity in the Design of Potent PROTACs for BTK. *Proc. Natl. Acad. Sci. U. S. A.* **2018**, *115* (31), E7285–E7292. https://doi.org/10.1073/PNAS.1803662115/-/DCSUPPLEMENTAL.

(211) Wu, H.; Yang, K.; Zhang, Z.; Leisten, E. D.; Li, Z.; Xie, H.; Liu, J.; Smith, K. A.; Novakova, Z.; Barinka, C.; Tang, W. Development of Multifunctional Histone Deacetylase 6 Degraders with Potent Antimyeloma Activity. *J. Med. Chem.* **2019**, *62* (15), 7042–7057. https://doi.org/10.1021/ACS.JMEDCHEM.9B00516/SUPPL_FILE/JM9B00516_SI_002.CSV.

(212) Brand, M.; Jiang, B.; Bauer, S.; Donovan, K. A.; Liang, Y.; Wang, E. S.; Nowak, R. P.; Yuan, J. C.; Zhang, T.; Kwiatkowski, N.; Müller, A. C.; Fischer, E. S.; Gray, N. S.; Winter, G. E. Homolog-Selective Degradation as a Strategy to Probe the Function of CDK6 in AML. *Cell Chem. Biol.* **2019**, *26* (2), 300-306.e9. https://doi.org/10.1016/J.CHEMBIOL.2018.11.006.

(213) Han, T.; Goralski, M.; Gaskill, N.; Capota, E.; Kim, J.; Ting, T. C.; Xie, Y.; Williams, N. S.; Nijhawan, D. Anticancer Sulfonamides Target Splicing by

Inducing RBM39 Degradation via Recruitment to DCAF15. *Science (80-.).* **2017**, *356* (6336). https://doi.org/10.1126/SCIENCE.AAL3755/SUPPL_FILE/PAP.PDF.

(214) Uehara, T.; Minoshima, Y.; Sagane, K.; Sugi, N. H.; Mitsuhashi, K. O.; Yamamoto, N.; Kamiyama, H.; Takahashi, K.; Kotake, Y.; Uesugi, M.; Yokoi, A.; Inoue, A.; Yoshida, T.; Mabuchi, M.; Tanaka, A.; Owa, T. Selective Degradation of Splicing Factor CAPERα by Anticancer Sulfonamides. *Nat. Chem. Biol. 2017 136* **2017**, *13* (6), 675–680. https://doi.org/10.1038/nchembio.2363.

(215) Bussiere, D. E.; Xie, L.; Srinivas, H.; Shu, W.; Burke, A.; Be, C.; Zhao, J.; Godbole, A.; King, D.; Karki, R. G.; Hornak, V.; Xu, F.; Cobb, J.; Carte, N.; Frank, A. O.; Frommlet, A.; Graff, P.; Knapp, M.; Fazal, A.; Okram, B.; Jiang, S.; Michellys, P. Y.; Beckwith, R.; Voshol, H.; Wiesmann, C.; Solomon, J. M.; Paulk, J. Structural Basis of Indisulam-Mediated RBM39 Recruitment to DCAF15 E3 Ligase Complex. *Nat. Chem. Biol. 2019 161* **2019**, *16* (1), 15–23. https://doi.org/10.1038/s41589-019-0411-6.

(216) Thibaudeau, T. A.; Smith, D. M. A Practical Review of Proteasome Pharmacology. *Pharmacol. Rev.* **2019**, *71* (2), 170–197. https://doi.org/10.1124/PR.117.015370.

(217) Santambrogio, L.; Berendam, S. J.; Engelhard, V. H. The Antigen Processing and Presentation Machinery in Lymphatic Endothelial Cells. *Front. Immunol.* **2019**, *10* (MAY), 1033. https://doi.org/10.3389/FIMMU.2019.01033/BIBTEX.

(218) Rousseau, A.; Bertolotti, A. Regulation of Proteasome Assembly and Activity in Health and Disease. *Nat. Rev. Mol. Cell Biol. 2018 1911* **2018**, *19* (11), 697–712. https://doi.org/10.1038/s41580-018-0040-z.

(219) Paiva, S. L.; Crews, C. M. Targeted Protein Degradation: Elements of PROTAC Design. *Curr. Opin. Chem. Biol.* **2019**, *50*, 111–119. https://doi.org/10.1016/J.CBPA.2019.02.022.

(220) Prekovic, S.; Van Royen, M. E.; Voet, A. R. D.; Geverts, B.; Houtman, R.; Melchers, D.; Zhang, K. Y. J.; Van Den Broeck, T.; Smeets, E.; Spans, L.; Houtsmuller, A. B.; Joniau, S.; Claessens, F.; Helsen, C. The Effect of F877L and T878A Mutations on Androgen Receptor Response to Enzalutamide. *Mol. Cancer Ther.* **2016**, *15* (7), 1702–1712. https://doi.org/10.1158/1535-

7163.MCT-15-0892.

(221) Ermondi, G.; Garcia-Jimenez, D.; Caron, G. Protacs and Building Blocks: The 2d Chemical Space in Very Early Drug Discovery. *Molecules* **2021**, *26* (3). https://doi.org/10.3390/molecules26030672.

(222) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and Computational Approaches to Estimate Solubility and Permeability in Drug Discovery and Development Settings. *Adv. Drug Deliv. Rev.* **1997**, *23* (1–3), 3–25.

(223) Begoli, E.; Bhattacharya, T.; Kusnezov, D. The Need for Uncertainty Quantification in Machine-Assisted Medical Decision Making. *Nat. Mach. Intell. 2019 11* **2019**, *1* (1), 20–23. https://doi.org/10.1038/s42256-018-0004-1.

(224) Segler, M. H. S.; Preuss, M.; Waller, M. P. Planning Chemical Syntheses with Deep Neural Networks and Symbolic AI. *Nat. 2018 5557698* **2018**, *555* (7698), 604–610. https://doi.org/10.1038/nature25978.

(225) Dimitrov, T.; Kreisbeck, C.; Becker, J. S.; Aspuru-Guzik, A.; Saikin, S. K. Autonomous Molecular Design: Then and Now. *ACS Appl. Mater. Interfaces* **2019**, *11* (28), 24825–24836. https://doi.org/10.1021/ACSAMI.9B01226.

(226) Mervin, L. H.; Johansson, S.; Semenova, E.; Giblin, K. A.; Engkvist, O. Uncertainty Quantification in Drug Design. *Drug Discov. Today* **2021**, *26* (2), 474–489. https://doi.org/10.1016/J.DRUDIS.2020.11.027.

(227) Nidhi, †; Meir Glick, ‡; John W. Davies, ‡ and; Jeremy L. Jenkins*, ‡. Prediction of Biological Targets for Compounds Using Multiple-Category Bayesian Models Trained on Chemogenomics Databases. *J. Chem. Inf. Model.* **2006**, *46* (3), 1124–1133. https://doi.org/10.1021/CI060003G.

(228) Lenselink, E. B.; ten Dijke, N.; Bongers, B.; Papadatos, G.; van Vlijmen, H. W. T.; Kowalczyk, W.; IJzerman, A. P.; van Westen, G. J. P. Beyond the Hype: Deep Neural Networks Outperform Established Methods Using a ChEMBL Bioactivity Benchmark Set. *J. Cheminformatics 2017 91* **2017**, *9* (1), 1–14. https://doi.org/10.1186/S13321-017-0232-0.

(229) Idakwo, G.; Thangapandian, S.; Luttrell, J.; Li, Y.; Wang, N.; Zhou, Z.; Hong, H.; Yang, B.; Zhang, C.; Gong, P. Structure–Activity Relationship-Based Chemical Classification of Highly Imbalanced Tox21 Datasets. *J. Cheminformatics 2020*

121 **2020**, *12* (1), 1–19. https://doi.org/10.1186/S13321-020-00468-X.

(230) Brown, S. P.; Muchmore, S. W.; Hajduk, P. J. Healthy Skepticism: Assessing Realistic Model Performance. *Drug Discov. Today* **2009**, *14* (7–8), 420–427. https://doi.org/10.1016/J.DRUDIS.2009.01.012.

(231) Li, G.; Zrimec, J.; Ji, B.; Geng, J.; Larsbrink, J.; Zelezniak, A.; Nielsen, J.; Engqvist, M. K. Performance of Regression Models as a Function of Experiment Noise: *https://doi.org/10.1177/11779322211020315* **2021**, *15*, 1–10. https://doi.org/10.1177/11779322211020315.

(232) Cortes-Ciriano, I.; Bender, A.; Malliavin, T. E. Comparing the Influence of Simulated Experimental Errors on 12 Machine Learning Algorithms in Bioactivity Modeling Using 12 Diverse Data Sets. *J. Chem. Inf. Model.* **2015**, *55* (7), 1413–1425. https://doi.org/10.1021/ACS.JCIM.5B00101.

(233) Chipman, H. A.; George, E. I.; McCulloch, R. E. BART: Bayesian Additive Regression Trees. *https://doi.org/10.1214/09-AOAS285* **2010**, *4* (1), 266–298. https://doi.org/10.1214/09-AOAS285.

(234) Reis, I.; Baron, D.; Shahaf, S. Probabilistic Random Forest: A Machine Learning Algorithm for Noisy Data Sets. *Astron. J.* **2018**, *157* (1), 16. https://doi.org/10.3847/1538-3881/AAF101.

(235) Boutsia, K.; Grazian, A.; Calderone, G.; Cristiani, S.; Cupani, G.; Guarneri, F.; Fontanot, F.; Amorin, R.; D'Odorico, V.; Giallongo, E.; Salvato, M.; Omizzolo, A.; Romano, M.; Menci, N. The Spectroscopic Follow-up of the QUBRICS Bright Quasar Survey. *Astrophys. J. Suppl. Ser.* **2020**, *250* (2), 26. https://doi.org/10.3847/1538-4365/ABAFC1.

(236) Bento, A. P.; Gaulton, A.; Hersey, A.; Bellis, L. J.; Chambers, J.; Davies, M.; Krüger, F. A.; Light, Y.; Mak, L.; McGlinchey, S.; Nowotka, M.; Papadatos, G.; Santos, R.; Overington, J. P. The ChEMBL Bioactivity Database: An Update. *Nucleic Acids Res.* **2014**, *42* (D1), D1083–D1090. https://doi.org/10.1093/NAR/GKT1031.

(237) RDKit: Cheminformatics and Machine Learning Software.

(238) Virtanen, P.; Gommers, R.; Oliphant, T. E.; Haberland, M.; Reddy, T.; Cournapeau, D.; Burovski, E.; Peterson, P.; Weckesser, W.; Bright, J.; van der Walt, S. J.; Brett, M.; Wilson, J.; Millman, K. J.; Mayorov, N.; Nelson, A. R. J.;

Jones, E.; Kern, R.; Larson, E.; Carey, C. J.; Polat, İ.; Feng, Y.; Moore, E. W.; VanderPlas, J.; Laxalde, D.; Perktold, J.; Cimrman, R.; Henriksen, I.; Quintero, E. A.; Harris, C. R.; Archibald, A. M.; Ribeiro, A. H.; Pedregosa, F.; van Mulbregt, P.; Vijaykumar, A.; Bardelli, A. Pietro; Rothberg, A.; Hilboll, A.; Kloeckner, A.; Scopatz, A.; Lee, A.; Rokem, A.; Woods, C. N.; Fulton, C.; Masson, C.; Häggström, C.; Fitzgerald, C.; Nicholson, D. A.; Hagen, D. R.; Pasechnik, D. V.; Olivetti, E.; Martin, E.; Wieser, E.; Silva, F.; Lenders, F.; Wilhelm, F.; Young, G.; Price, G. A.; Ingold, G. L.; Allen, G. E.; Lee, G. R.; Audren, H.; Probst, I.; Dietrich, J. P.; Silterra, J.; Webber, J. T.; Slavič, J.; Nothman, J.; Buchner, J.; Kulick, J.; Schönberger, J. L.; de Miranda Cardoso, J. V.; Reimer, J.; Harrington, J.; Rodríguez, J. L. C.; Nunez-Iglesias, J.; Kuczynski, J.; Tritz, K.; Thoma, M.; Newville, M.; Kümmerer, M.; Bolingbroke, M.; Tartre, M.; Pak, M.; Smith, N. J.; Nowaczyk, N.; Shebanov, N.; Pavlyk, O.; Brodtkorb, P. A.; Lee, P.; McGibbon, R. T.; Feldbauer, R.; Lewis, S.; Tygier, S.; Sievert, S.; Vigna, S.; Peterson, S.; More, S.; Pudlik, T.; Oshima, T.; Pingel, T. J.; Robitaille, T. P.; Spura, T.; Jones, T. R.; Cera, T.; Leslie, T.; Zito, T.; Krauss, T.; Upadhyay, U.; Halchenko, Y. O.; Vázquez-Baeza, Y. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nat. Methods* **2020**, *17* (3), 261–272. https://doi.org/10.1038/s41592-019-0686-2.

(239) Papadatos, G.; Gaulton, A.; Hersey, A.; Overington, J. P. Activity, Assay and Target Data Curation and Quality in the ChEMBL Database. *J. Comput. Mol. Des. 2015 299* **2015**, *29* (9), 885–896. https://doi.org/10.1007/S10822-015-9860-5.

(240) Vaicenavicius, J.; Widmann, D.; Andersson, C.; Lindsten, F.; Roll, J.; Schön, T. Evaluating Model Calibration in Classification. PMLR April 11, 2019, pp 3459–3467.

(241) Kurczab, R.; Smusz, S.; Bojarski, A. J. The Influence of Negative Training Set Size on Machine Learning-Based Virtual Screening. *J. Cheminform.* **2014**, *6* (1), 1–9. https://doi.org/10.1186/1758-2946-6-32.

(242) Raimondi, C.; Falasca, M. Targeting PDK1 in Cancer. *Curr. Med. Chem.* **2011**, *18* (18), 2763–2769. https://doi.org/10.2174/092986711796011238.

(243) Young, D. W.; Bender, A.; Hoyt, J.; McWhinnie, E.; Chirn, G. W.; Tao, C. Y.;

Tallarico, J. A.; Labow, M.; Jenkins, J. L.; Mitchison, T. J.; Feng, Y. Integrating High-Content Screening and Ligand-Target Prediction to Identify Mechanism of Action. *Nat. Chem. Biol.* **2008**, *4* (1), 59–68. https://doi.org/10.1038/nchembio.2007.53.

(244) Sneddon, T. P.; Li, P.; Edmunds, S. C. GigaDB: Announcing the GigaScience Database. *Gigascience* **2012**, *1* (1), 1–2. https://doi.org/10.1186/2047-217X-1-11.

(245) Jeliazkova, N.; Jeliazkov, V. AMBIT RESTful Web Services: An Implementation of the OpenTox Application Programming Interface. *J. Cheminform.* **2011**, *3* (1), 1–18. https://doi.org/10.1186/1758-2946-3-18.

(246) Kochev, N. T.; Paskaleva, V. H.; Jeliazkova, N. Ambit-Tautomer: An Open Source Tool for Tautomer Generation. *Mol. Inform.* **2013**, *32* (5–6), 481–504. https://doi.org/10.1002/minf.201200133.

(247) Jeliazkova, N.; Kochev, N. AMBIT-SMARTS: Efficient Searching of Chemical Structures and Fragments. *Mol. Inform.* **2011**, *30* (8), 707–720. https://doi.org/10.1002/minf.201100028.

(248) Bateman, A. UniProt: A Worldwide Hub of Protein Knowledge. *Nucleic Acids Res.* **2019**, *47* (D1), D506–D515. https://doi.org/10.1093/nar/gky1049.

(249) Ashton, M.; Barnard, J.; Casset, F.; Charlton, M.; Downs, G.; Gorse, D.; Holliday, J.; Lahana, R.; Willett, P. Identification of Diverse Database Subsets Using Property-Based and Fragment-Based Molecular Descriptions. *Quant. Struct. Relationships* **2002**, *21* (6), 598–604. https://doi.org/10.1002/qsar.200290002.

(250) F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, M. G. O.; Blondel, P. Prettenhofer, R. Weiss, V. D. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.

(251) de la Vega de León, A.; Chen, B.; Gillet, V. J. Effect of Missing Data on Multitask Prediction Methods. *J. Cheminform.* **2018**, *10* (1), 1–12. https://doi.org/10.1186/s13321-018-0281-z.

(252) Zhao, W.; Hevener, K. E.; White, S. W.; Lee, R. E.; Boyett, J. M. A Statistical Framework to Evaluate Virtual Screening. *BMC Bioinformatics* **2009**, *10*, 1–13. https://doi.org/10.1186/1471-2105-10-225.

(253) Truchon, J. F.; Bayly, C. I. Evaluating Virtual Screening Methods: Good and Bad Metrics for the “Early Recognition” Problem. *J. Chem. Inf. Model.* **2007**, *47* (2), 488–508. https://doi.org/10.1021/ci600426e.

(254) Lipiński, P. F. J.; Szurmak, P. SCRAMBLE’N’GAMBLE: A Tool for Fast and Facile Generation of Random Data for Statistical Evaluation of QSAR Models. *Chem. Pap.* **2017**, *71* (11), 2217–2232. https://doi.org/10.1007/s11696-017-0215-7.

(255) Kurczab, R.; Bojarski, A. J. The Influence of the Negative-Positive Ratio and Screening Database Size on the Performance of Machine Learning-Based Virtual Screening. *PLoS One* **2017**, *12* (4), 1–12. https://doi.org/10.1371/journal.pone.0175410.

(256) Shang, S.; Hua, F.; Hu, Z. W. The Regulation of β-Catenin Activity and Function in Cancer: Therapeutic Opportunities. *Oncotarget* **2017**, *8* (20), 33972–33989. https://doi.org/10.18632/oncotarget.15687.

(257) Cui, C.; Zhou, X.; Zhang, W.; Qu, Y.; Ke, X. Is β-Catenin a Druggable Target for Cancer Therapy? *Trends Biochem. Sci.* **2018**, *43* (8), 623–634. https://doi.org/10.1016/j.tibs.2018.06.003.

(258) National Center for Biotechnology Information. PubChem Database. Source: Burnham Center for Chemical Genomics, AID=1665.

(259) Lin, Y.-T.; Lin, K.-H.; Huang, C.-J.; Wei, A.-C. MitoTox: A Comprehensive Mitochondrial Toxicity Database. *BMC Bioinforma. 2021 2210* **2021**, *22* (10), 1–14. https://doi.org/10.1186/S12859-021-04285-3.

(260) Chan, K.; Truong, D.; Shangari, N.; O’Brien, P. J. Drug-Induced Mitochondrial Toxicity. *http://dx.doi.org/10.1517/17425255.1.4.655* **2005**, *1* (4), 655–669. https://doi.org/10.1517/17425255.1.4.655.

(261) Vuda, M.; Kamath, A. Drug Induced Mitochondrial Dysfunction: Mechanisms and Adverse Clinical Consequences. *Mitochondrion* **2016**, *31*, 63–74. https://doi.org/10.1016/J.MITO.2016.10.005.

(262) Dykens, J. A.; Will, Y. The Significance of Mitochondrial Toxicity Testing in Drug Development. *Drug Discov. Today* **2007**, *12* (17–18), 777–785. https://doi.org/10.1016/J.DRUDIS.2007.07.013.

(263) Nadanaciva, S.; Will, Y. Current Concepts in Drug-Induced Mitochondrial

Toxicity. *Curr. Protoc. Toxicol.* **2009**, No. SUPPL. 40. https://doi.org/10.1002/0471140856.TX0215S40.

(264) Meyer, J. N.; Hartman, J. H.; Mello, D. F. Mitochondrial Toxicity. *Toxicol. Sci.* **2018**, *162* (1), 15–23. https://doi.org/10.1093/TOXSCI/KFY008.

(265) Julie, N. L.; Julie, I. M.; Kende, A. I.; Wilson, G. L. Mitochondrial Dysfunction and Delayed Hepatotoxicity: Another Lesson from Troglitazone. *Diabetol. 2008 5111* **2008**, *51* (11), 2108–2116. https://doi.org/10.1007/S00125-008-1133-6.

(266) Rizzuto, R.; Bernardi, P.; Pozzan, T. Mitochondria as All-Round Players of the Calcium Game. *J. Physiol.* **2000**, *529* (1), 37–47. https://doi.org/10.1111/J.1469-7793.2000.00037.X.

(267) Will, Y.; Dykens, J. Mitochondrial Toxicity Assessment in Industry – a Decade of Technology Development and Insight. *https://doi.org/10.1517/17425255.2014.939628* **2014**, *10* (8), 1061–1067. https://doi.org/10.1517/17425255.2014.939628.

(268) Rana, P.; Aleo, M. D.; Gosink, M.; Will, Y. Evaluation of in Vitro Mitochondrial Toxicity Assays and Physicochemical Properties for Prediction of Organ Toxicity Using 228 Pharmaceutical Drugs. *Chem. Res. Toxicol.* **2018**, *32* (1), 156–167. https://doi.org/10.1021/ACS.CHEMRESTOX.8B00246.

(269) Hynes, J.; Nadanaciva, S.; Swiss, R.; Carey, C.; Kirwan, S.; Will, Y. A High-Throughput Dual Parameter Assay for Assessing Drug-Induced Mitochondrial Dysfunction Provides Additional Predictivity over Two Established Mitochondrial Toxicity Assays. *Toxicol. Vitr.* **2013**, *27* (2), 560–569. https://doi.org/10.1016/J.TIV.2012.11.002.

(270) Mitochondrial Toxicity https://www.cyprotex.com/toxicology/mitochondrial-toxicity (accessed Oct 25, 2021).

(271) Sakamuru, S.; Attene-Ramos, M. S.; Xia, M. Mitochondrial Membrane Potential Assay. *Methods Mol. Biol.* **2016**, *1473*, 17–22. https://doi.org/10.1007/978-1-4939-6346-1_2.

(272) Wilkins, H. M.; Weidling, I.; Koppel, S.; Wang, X.; von Schulze, A.; Swerdlow, R. H. Mitochondrial Function and Neurodegenerative Diseases. *Mol. Cell. Basis Neurodegener. Dis. Underlying Mech.* **2018**, 369–414. https://doi.org/10.1016/B978-0-12-811304-2.00013-4.

(273) Zhao, P.; Peng, Y.; Xu, X.; Wang, Z.; Wu, Z.; Li, W.; Tang, Y.; Liu, G. In Silico Prediction of Mitochondrial Toxicity of Chemicals Using Machine Learning Methods. *J. Appl. Toxicol.* **2021**, *41* (10), 1518–1526. https://doi.org/10.1002/JAT.4141.

(274) Hemmerich, J.; Troger, F.; Füzi, B.; F.Ecker, G. Using Machine Learning Methods and Structural Alerts for Prediction of Mitochondrial Toxicity. *Mol. Inform.* **2020**, *39* (5), 2000005. https://doi.org/10.1002/MINF.202000005.

(275) Zhang, H.; Chen, Q. Y.; Xiang, M. L.; Ma, C. Y.; Huang, Q.; Yang, S. Y. In Silico Prediction of Mitochondrial Toxicity by Using GA-CG-SVM Approach. *Toxicol. Vitr.* **2009**, *23* (1), 134–140. https://doi.org/10.1016/J.TIV.2008.09.017.

(276) Trapotsi, M. A.; Mervin, L. H.; Afzal, A. M.; Sturm, N.; Engkvist, O.; Barrett, I. P.; Bender, A. Comparison of Chemical Structure and Cell Morphology Information for Multitask Bioactivity Predictions. *J. Chem. Inf. Model.* **2021**, *61* (3), 1444–1456. https://doi.org/10.1021/acs.jcim.0c00864.

(277) Trapotsi, M.; Barrett, I.; Engkvist, O.; Bender, A. Bioinformatic Approaches in the Understanding of Mechanism of Action (MoA). In Target Discovery and Validation; Plowright, A. T., Ed.; 2020. https://doi.org/10.1002/9783527818242.ch11.

(278) Martin, H. L.; Adams, M.; Higgins, J.; Bond, J.; Morrison, E. E.; Bell, S. M.; Warriner, S.; Nelson, A.; Tomlinson, D. C. High-Content, High-Throughput Screening for the Identification of Cytotoxic Compounds Based on Cell Morphology and Cell Proliferation Markers. *PLoS One* **2014**, *9* (2), e88338. https://doi.org/10.1371/JOURNAL.PONE.0088338.

(279) Persson, M.; Løye, A. F.; Jacquet, M.; Mow, N. S.; Thougaard, A. V.; Mow, T.; Hornberg, J. J. High-Content Analysis/Screening for Predictive Toxicology: Application to Hepatotoxicity and Genotoxicity. *Basic Clin. Pharmacol. Toxicol.* **2014**, *115* (1), 18–23. https://doi.org/10.1111/BCPT.12200.

(280) Alexander, G.; IwataYasuhiro; SirenkoOksana; BittnerMichael; RusynIvan. High-Content Assay Multiplexing for Toxicity Screening in Induced Pluripotent Stem Cell-Derived Cardiomyocytes and Hepatocytes. *https://home.liebertpub.com/adt* **2015**, *13* (9), 529–546. https://doi.org/10.1089/ADT.2015.659.

(281) Cox, M. J.; Jaensch, S.; Waeter, J. Van de; Cougnaud, L.; Seynaeve, D.; Benalla, S.; Koo, S. J.; Wyngaert, I. Van Den; Neefs, J.-M.; Malkov, D.; Bittremieux, M.; Steemans, M.; Peeters, P. J.; Wegner, J. K.; Ceulemans, H.; Gustin, E.; Chong, Y. T.; Göhlmann, H. W. H. Tales of 1,008 Small Molecules: Phenomic Profiling through Live-Cell Imaging in a Panel of Reporter Cell Lines. *bioRxiv* **2020**, 2020.03.13.990093. https://doi.org/10.1101/2020.03.13.990093.

(282) Way, G. P.; Kost-Alimova, M.; Shibue, T.; Harrington, W. F.; Gill, S.; Piccioni, F.; Becker, T.; Hahn, W. C.; Carpenter, A. E.; Vazquez, F.; Singh, S. Predicting Cell Health Phenotypes Using Image-Based Morphology Profiling. *bioRxiv* **2020**, 1–27. https://doi.org/10.1101/2020.07.08.193938.

(283) Makarenkov, V.; Zentilli, P.; Kevorkov, D.; Gagarin, A.; Malo, N.; Nadon, R. An Efficient Method for the Detection and Elimination of Systematic Error in High-Throughput Screening. *Bioinformatics* **2007**, *23* (13), 1648–1657. https://doi.org/10.1093/bioinformatics/btm145.

(284) Dragiev, P.; Nadon, R.; Makarenkov, V. Two Effective Methods for Correcting Experimental High-Throughput Screening Data. *Bioinformatics* **2012**, *28* (13), 1775–1782. https://doi.org/10.1093/bioinformatics/bts262.

(285) Dragiev, P.; Nadon, R.; Makarenkov, V. Systematic Error Detection in Experimental High-Throughput Screening. *BMC Bioinformatics* **2011**, *12*. https://doi.org/10.1186/1471-2105-12-25.

(286) GitHub - broadinstitute/grit-benchmark: Benchmarking a metric used to evaluate a perturbation strength https://github.com/broadinstitute/grit-benchmark (accessed Oct 25, 2021).

(287) GitHub - cytomining/cytominer-eval: Common Evaluation Metrics for DataFrames https://github.com/cytomining/cytominer-eval (accessed Oct 25, 2021).

(288) Bergstra, J.; Yamins, D.; Cox, D. Making a Science of Model Search: Hyperparameter Optimization in Hundreds of Dimensions for Vision Architectures. PMLR February 13, 2013, pp 115–123.

(289) Bergstra, J.; Komer, B.; Eliasmith, C.; Yamins, D.; Cox, D. D. Hyperopt: A Python Library for Model Selection and Hyperparameter Optimization. *Comput. Sci. Discov.* **2015**, *8* (1), 014008. https://doi.org/10.1088/1749-

4699/8/1/014008.

(290) McInnes, L.; Healy, J.; Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. **2018**.

(291) Rohban, M. H.; Singh, S.; Wu, X.; Berthet, J. B.; Bray, M. A.; Shrestha, Y.; Varelas, X.; Boehm, J. S.; Carpenter, A. E. Systematic Morphological Profiling of Human Gene and Allele Function via Cell Painting. *Elife* **2017**, *6*, 1–23. https://doi.org/10.7554/eLife.24060.

(292) Neumann, B.; Walter, T.; Hériché, J. K.; Bulkescher, J.; Erfle, H.; Conrad, C.; Rogers, P.; Poser, I.; Held, M.; Liebel, U.; Cetin, C.; Sieckmann, F.; Pau, G.; Kabbe, R.; Wünsche, A.; Satagopam, V.; Schmitz, M. H. A.; Chapuis, C.; Gerlich, D. W.; Schneider, R.; Eils, R.; Huber, W.; Peters, J. M.; Hyman, A. A.; Durbin, R.; Pepperkok, R.; Ellenberg, J. Phenotypic Profiling of the Human Genome by Time-Lapse Microscopy Reveals Cell Division Genes. *Nature* **2010**, *464* (7289), 721–727. https://doi.org/10.1038/nature08869.

(293) Weaver, S.; Gleeson, M. P. The Importance of the Domain of Applicability in QSAR Modeling. *J. Mol. Graph. Model.* **2008**, *26* (8), 1315–1326. https://doi.org/10.1016/J.JMGM.2008.01.002.

(294) Janet, J. P.; Duan, C.; Yang, T.; Nandy, A.; Kulik, H. Uncertain Times Call for Quantitative Uncertainty Metrics: Controlling Error in Neural Network Predictions for Chemical Discovery. **2019**. https://doi.org/10.26434/CHEMRXIV.7900277.V1.

(295) Hanser, T.; Barber, C.; Guesné, S.; Marchaland, J. F.; Werner, S. Applicability Domain: Towards a More Formal Framework to Express the Applicability of a Model and the Confidence in Individual Predictions. *Challenges Adv. Comput. Chem. Phys.* **2019**, *30*, 215–232. https://doi.org/10.1007/978-3-030-16443-0_11.

(296) Sahigara, F.; Mansouri, K.; Ballabio, D.; Mauri, A.; Consonni, V.; Todeschini, R. Comparison of Different Approaches to Define the Applicability Domain of QSAR Models. *Mol. 2012, Vol. 17, Pages 4791-4810* **2012**, *17* (5), 4791–4810. https://doi.org/10.3390/MOLECULES17054791.

(297) Pereira, T.; Cardoso, S.; Guerreiro, M.; Mendonça, A.; Madeira, S. C. Targeting the Uncertainty of Predictions at Patient-Level Using an Ensemble of Classifiers

Coupled with Calibration Methods, Venn-ABERS, and Conformal Predictors: A Case Study in AD. *J. Biomed. Inform.* **2020**, *101*, 103350. https://doi.org/10.1016/J.JBI.2019.103350.

(298) Norinder, U.; Spjuth, O.; Svensson, F. Using Predicted Bioactivity Profiles to Improve Predictive Modeling. *J. Chem. Inf. Model.* **2020**, *60* (6), 2830–2837. https://doi.org/10.1021/ACS.JCIM.0C00250.

(299) Cortés-Ciriano, I.; Bender, A. Deep Confidence: A Computationally Efficient Framework for Calculating Reliable Prediction Errors for Deep Neural Networks. *J. Chem. Inf. Model.* **2018**, *59* (3), 1269–1281. https://doi.org/10.1021/ACS.JCIM.8B00542.

7. Appendix – Chapter 2

Table 7.1: Description of Methodologies, which are used to take into account uncertainty in predictions, and their advantages and disadvantages

Method	Description	Advantage	Disadvantage
Applicability Domain (AD) estimation	Provides an estimate of whether the assumptions of a model are fulfilled for a given input, e.g., distance to model AD provides a reliability based on whether a query compound is close to model training data[293–296].	Provides estimates in uncertainty when making predictions for new compounds.	Do not commonly take into account the uncertainty related to the underlying data.
Conformal Prediction	Produces error bands around the predictions, with the underlying assumption that inputs less similar to model training data should lead to less certain estimates. This is captured using a nonconformity measure, i.e., the nonconformity score for a new query compound is calculated[297–299].	Provides estimates in uncertainty when making predictions for new compounds.	Do not commonly take into account the uncertainty related to the underlying data.
Probability Calibration	Addresses the question of obtaining accurate likelihoods of predictions based on the	There are advantages related to specific	Performance depends on the reference observations used.

	distributions of reference observations for a given dataset[163].	calibration methodologies. e.g., Isotonic regression methodology makes no assumptions on the curve form. Inductive methods must split data in order to create 'proper' calibration splits.	Limitations related to specific calibration methodologies: e.g., Isotonic regression methodology requires a large number of calibration points and has a tendency to overfit.
Gaussian Processes (GP, Bayesian Methodology)	Probability distributions over possible functions are used to evaluate confidence intervals and decide based on those if one should refit the prediction in some region of interest[226].	Allow the incorporation of data prior knowledge. The uncertainty of a fitted GP increases away from the training data.	Gaussian processes can be computationally expensive (because of their non-parametric nature and they need to take into account all the training data each time they make a prediction)

Table 7.2: Standard deviation of replicate affinity measurements (IC_{50}/EC_{50}/K_i/K_d) across different aggregation methods. The mean, median and variance of standard deviation values across replicates are presented.

Aggregation Method	Number of replicate measurements	Mean	Median	Variance
Inter-Assay ID	53,270	0.55	0.41	0.26
Inter Assay Type	27,122	0.55	0.40	0.27
Confidence score >= 5	110,164	0.53	0.37	0.27
Confidence score >= 8	86,761	0.51	0.37	0.26
Intra-IC_{50} Type	47,449	0.51	0.36	0.25
Intra-K_d Type	2,645	0.56	0.33	0.37
Intra- EC_{50} Type	4,412	0.42	0.27	0.23
Intra-K_i Type	14,321	0.40	0.22	0.25
Intra-Assay ID	16,207	0.26	0.04	0.18

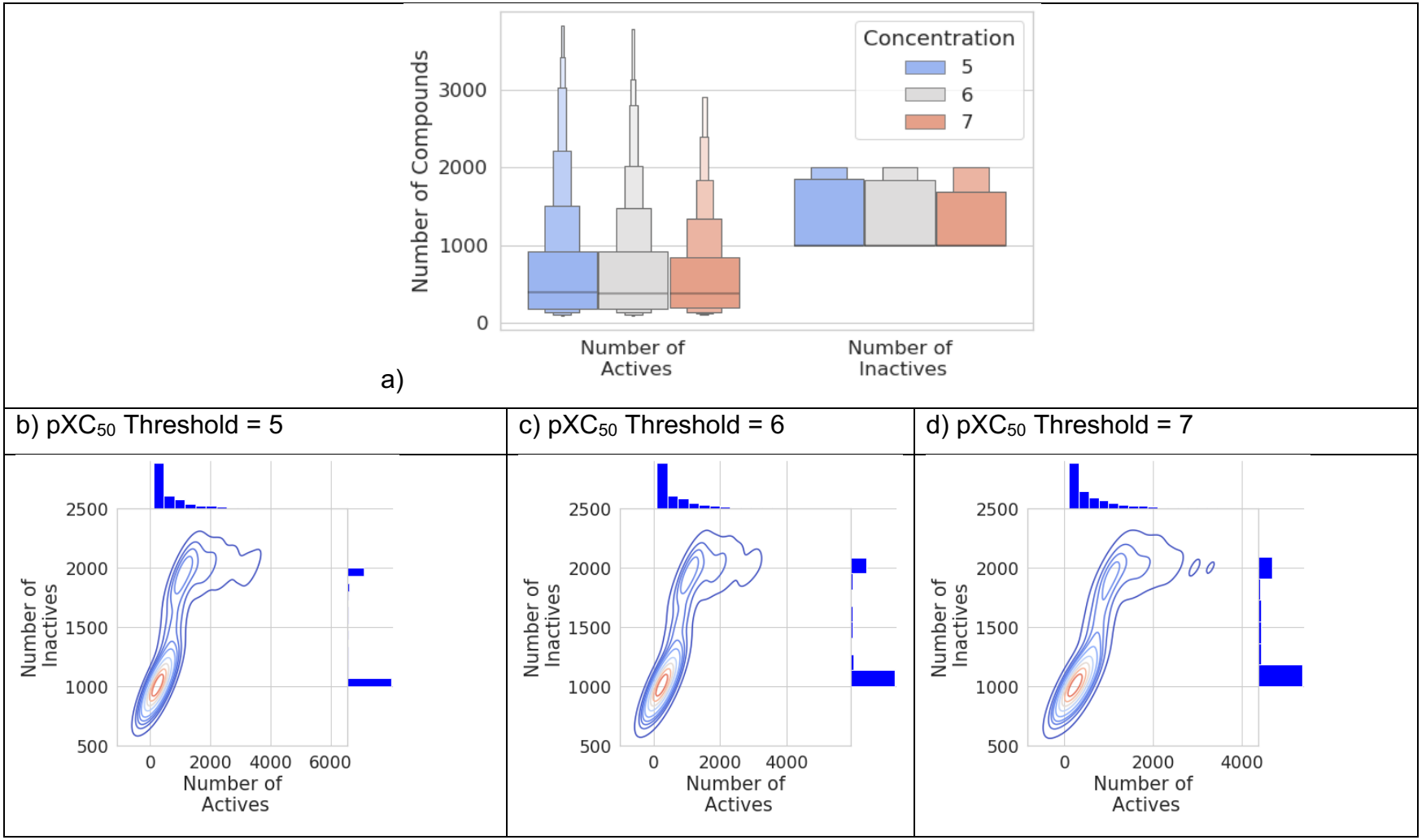


Figure 7.1: a) Number of Active and Inactive compounds across the 557 models and across the three different pXC_{50} Thresholds (5, 6, and 7). Number of Actives vs the Number of Inactives per model across the three different pXC_{50} Thresholds: b) 6, c) 6 and d) 7.

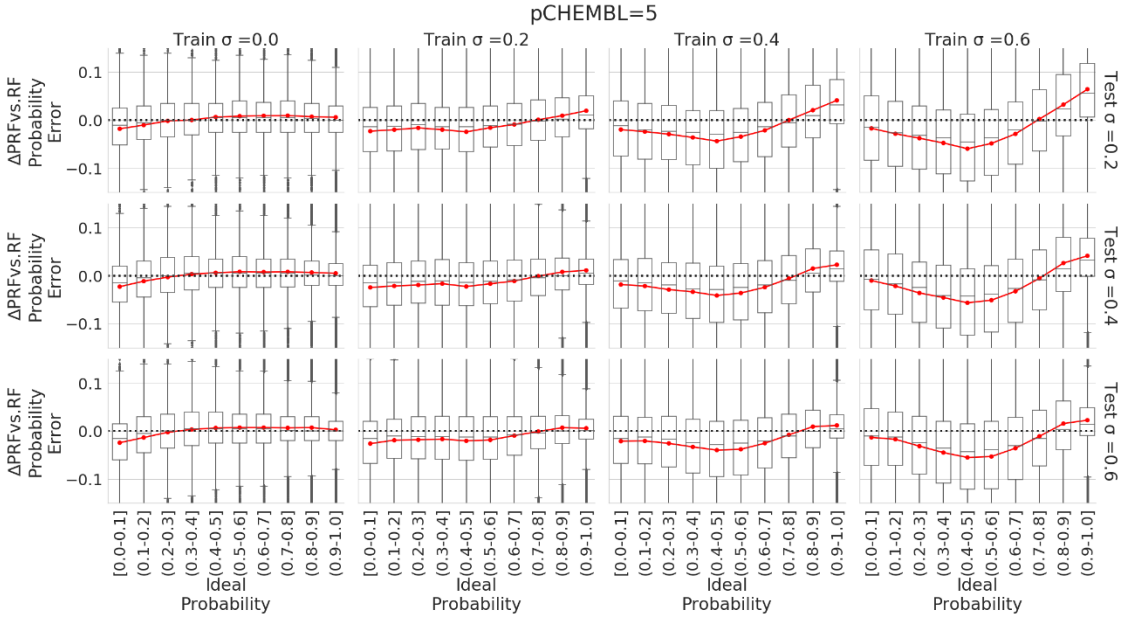


Figure 7.2: Ideal probabilities as a function of the delta of PRF versus RF error margins across emulated train-test standard deviations. Overall, results shown here for a threshold of pChEMBL value of 5 highlight the most optimal PRF probability estimates were observed in cases when standard deviation in the test set most closely resembled that in the training set. It can also be seen that the largest benefit in terms of error margin for the PRF (lower values on the y-axis) are observed toward the midpoint of the ideal Δy scale, particularly for higher training set standard deviations. This is when the original RF weights the marginal cases equivalent in distinguishing between activity classes.

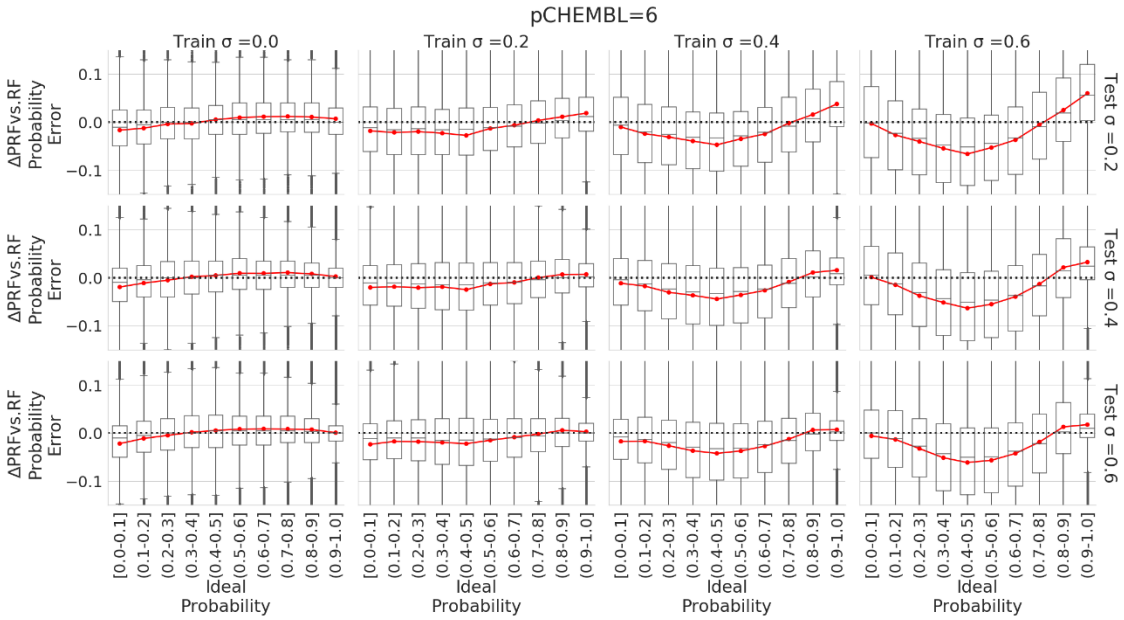


Figure 7.3: Ideal probabilities as a function of the delta of PRF versus RF error margins across emulated train-test standard deviations. Overall, results shown here for a threshold of pChEMBL value of 6 highlight the most optimal PRF probability estimates were observed in cases when standard deviation in the test set most closely resembled that in the training set. It can also be seen that the largest benefit in terms of error margin for the PRF (lower values on the y-axis) are observed toward the midpoint of the ideal Δy scale, particularly for higher training set standard deviations. This is when the original RF weights the marginal cases equivalent in distinguishing between activity classes.

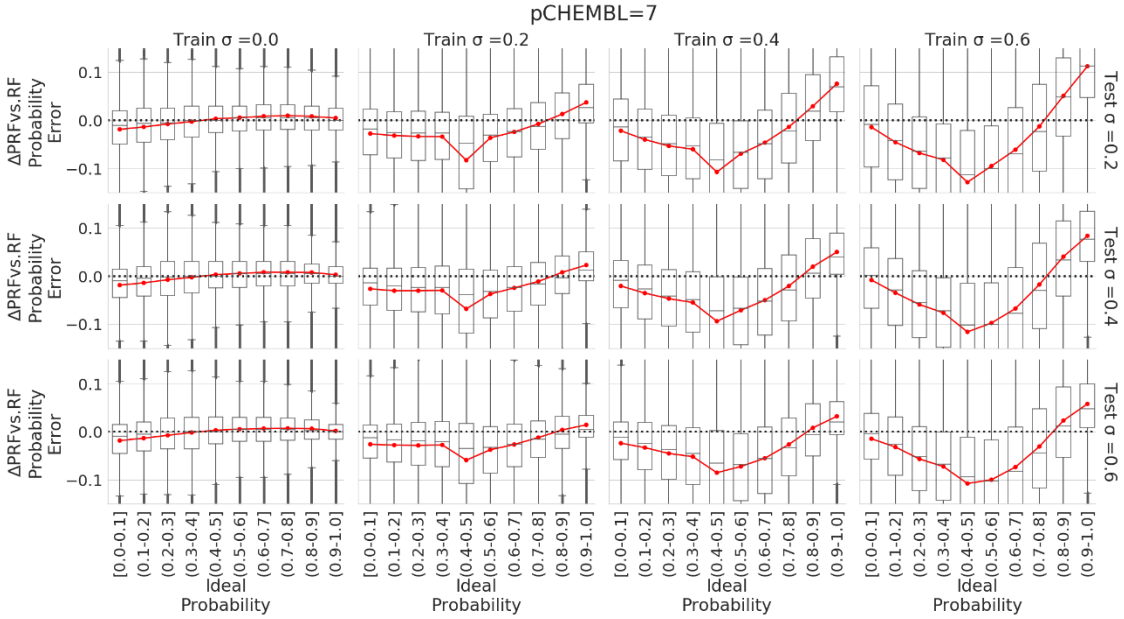


Figure 7.4: Ideal probabilities as a function of the delta of PRF versus RF error margins across emulated train-test standard deviations. Overall, results shown here for a threshold of pChEMBL value of 7 highlight the most optimal PRF probability estimates were observed in cases when standard deviation in the test set most closely resembled that in the training set. It can also be seen that the largest benefit in terms of error margin for the PRF (lower values on the y-axis) are observed toward the midpoint of the ideal Δy scale, particularly for higher training set standard deviations. This is when the original RF weights the marginal cases equivalent in distinguishing between activity classes.

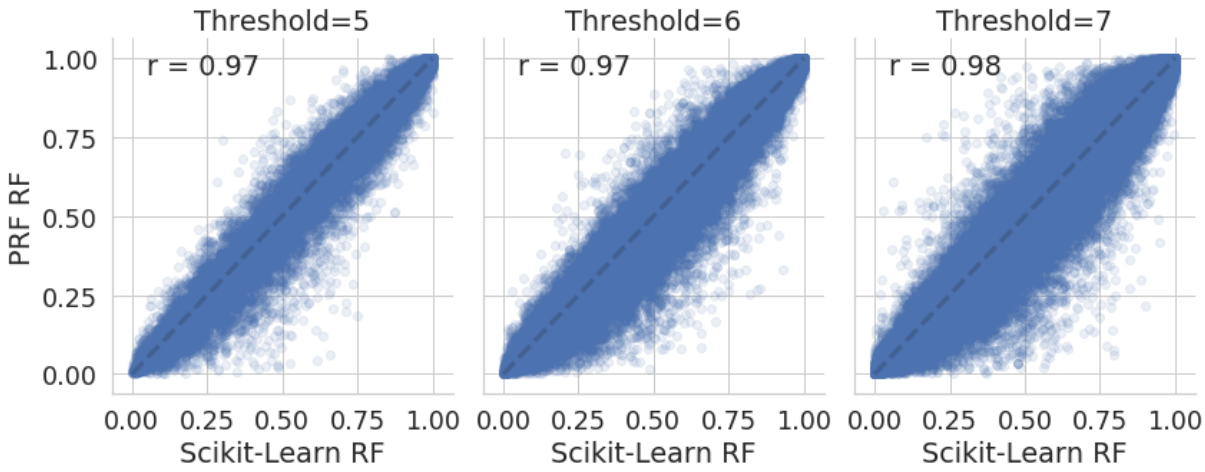


Figure 7.5: Comparison between RF scikit-learn implementation and PRF (when σ = 0). There is a high overall R^2 correlation between Scikit-Learn RF and the PRF (σ = 0) ranging between ~0.97-0.98 across the standard deviation test sets, hence the returned predictions from both RF approaches are overall comparable.

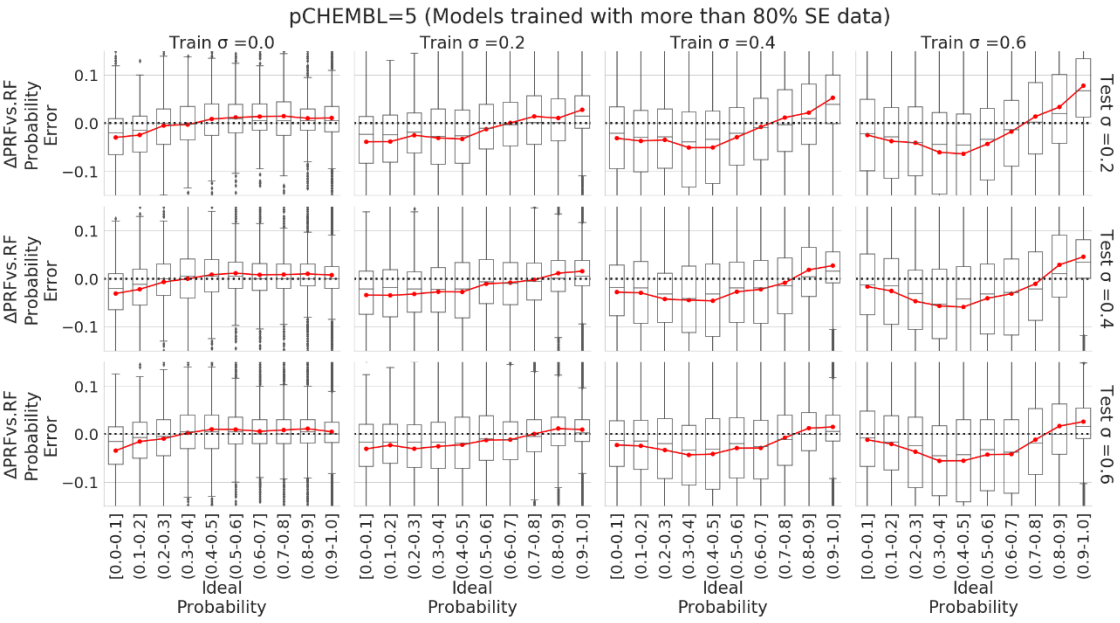


Figure 7.6: Ideal probabilities as a function of the delta of PRF versus RF error margins across emulated train-test standard deviations for models trained with a min of 80% putative inactives. Overall, results shown here for a threshold of pChEMBL value of 5 (10 μM) highlight the most optimal PRF probability estimates were observed in cases when standard deviation in the test set most closely resembled that in the training set. It can also be seen that the largest benefit in terms of error margin for the PRF (lower values on the y-axis) are observed toward the midpoint of the ideal Δy scale, particularly for higher training set standard deviations. This is when the original RF weights the marginal cases equivalent in distinguishing between activity classes.

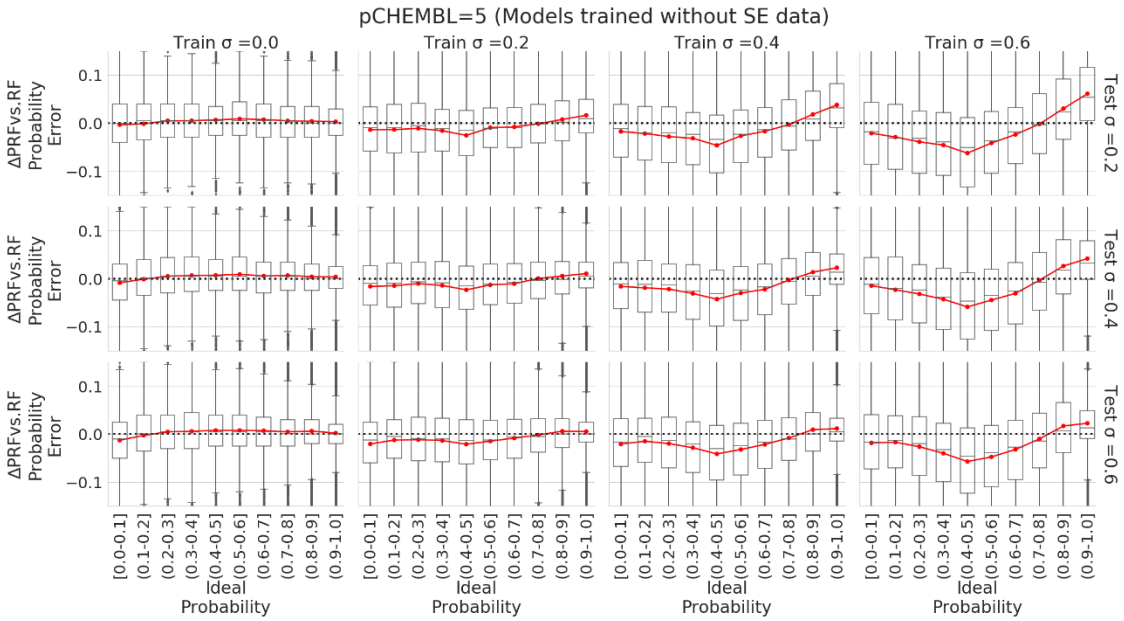


Figure 7.7: Ideal probabilities as a function of the delta of PRF versus RF error margins across emulated train-test standard deviations for models trained without putative inactives. Overall, results shown here for a threshold of pChEMBL value of 5 (10 μM) highlight the most optimal PRF probability estimates were observed in cases when standard deviation in the test set most closely resembled that in the training set. It can also be seen that the largest benefit in terms of error margin for the PRF (lower values on the y-axis) are observed toward the midpoint of the ideal ∆y scale, particularly for higher training set standard deviations. This is when the original RF weights the marginal cases equivalent in distinguishing between activity classes.

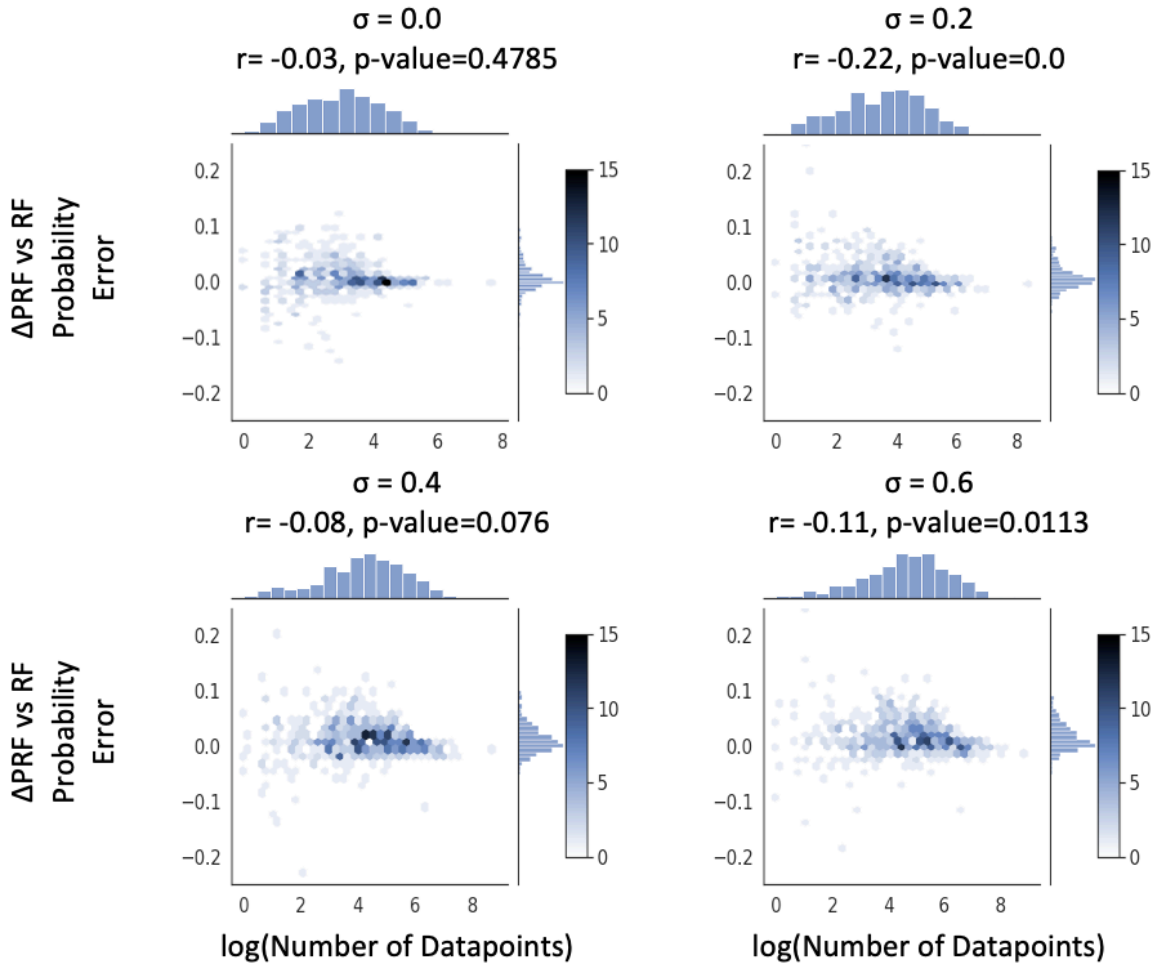


Figure 7.8: Correlation analysis of model sizes (when pChEMBL threshold is 5 [10 μM]) as a function of PRF improvement. Overall, model dataset size is shown to have no discernible effect on the improvement of PRF versus the vanilla RF, since no significant Pearson correlation exists across the four arbitrary standard deviations (σ) evaluated. Increased density is represented via the blue hex marker intensity.

8. Appendix – Chapter 3

Table 8.1: Targets in models and number of actives and inactives per target and A:I ratio per-target.

Protein Uniprot Name	Number of Active Compounds	Number of Inactive Compounds	Ratio AI number
ABL1	5	10234	0.0005
ACHE	11	32	0.3438
ADORA1	15	27	0.5556
ADORA2A	12	24	0.5000
ADORA2B	7	8	0.8750
ADORA3	28	34	0.8235
ADRA1A	31	5	6.2000
ADRA1B	42	12	3.5000
ADRA1D	40	21	1.9048
ADRA2A	43	34	1.2647
ADRA2B	40	27	1.4815
ADRA2C	42	11	3.8182
ADRB1	24	16	1.5000
ADRB2	33	9935	0.0033
ADRB3	12	19	0.6316
AHR	13	9492	0.0014
AKR1B1	8	28	0.2857
AKR1C3	5	9	0.5556
ALDH1A1	130	9148	0.0142
ALOX12	12	7296	0.0016
ALOX15	32	4008	0.0080
ALOX15B	18	5387	0.0033
ALOX5	6	26	0.2308
ALPL	9	7997	0.0011
APEX1	10	70	0.1429
APOBEC3F	13	10032	0.0013

APOBEC3G	20	10544	0.0019
AR	53	778	0.0681
ARSA	34	324	0.1049
ATAD5	123	10029	0.0123
ATM	9	9259	0.0010
ATXN2	235	10102	0.0233
AURKB	5	53	0.0943
BAZ2B	141	9459	0.0149
BLM	133	10227	0.0130
BRCA1	46	10256	0.0045
CA1	11	30	0.3667
CA12	18	18	1.0000
CA13	7	7	1.0000
CA14	11	15	0.7333
CA2	18	21	0.8571
CA4	8	17	0.4706
CA5A	12	13	0.9231
CA5B	12	11	1.0909
CA6	11	15	0.7333
CA7	13	18	0.7222
CA9	20	16	1.2500
CAR13	11	8	1.3750
CAR15	10	8	1.2500
CBX1	44	9944	0.0044
CDK1	7	13	0.5385
CDK2	6	25	0.2400
CDK5	7	9631	0.0007
CFTR	7	25301	0.0003
CGA	61	9329	0.0065
CHRM1	91	10246	0.0089
CHRM2	38	13	2.9231

CHRM3	33	16	2.0625
CHRM4	31	10241	0.0030
CHRM5	28	10243	0.0027
CHRNA7	7	9	0.7778
CLK4	9	10	0.9000
CNR1	13	36	0.3611
CNR2	12	31	0.3871
CSF1R	7	52	0.1346
CSNK1A1	6	12	0.5000
CSNK1D	10	55	0.1818
CSNK1G1	35	23	1.5217
CSNK1G3	27	32	0.8438
CTNNB1	42	7575	0.0055
CXCL8	5	71	0.0704
CYP19A1	43	756	0.0569
CYP1A1	5	5	1.0000
CYP1A2	126	614	0.2052
CYP2C19	95	3055	0.0311
CYP2C9	85	4864	0.0175
CYP2D6	93	761	0.1222
CYP3A4	64	1323	0.0484
DDIT3	7	8295	0.0008
DRD1	85	9619	0.0088
DRD2	90	9060	0.0099
DRD3	73	10309	0.0071
DRD4	27	15	1.8000
DUSP3	37	9358	0.0040
DYRK1A	23	9154	0.0025
EEF2K	52	8	6.5000
EGFR	13	89	0.1461
EHMT2	307	3540	0.0867

EIF4H	152	9085	0.0167
ERG	12	9267	0.0013
ESR1	50	4469	0.0112
ESR2	18	3932	0.0046
F2	5	8547	0.0006
FEN1	16	9904	0.0016
FLT1	5	52	0.0962
FLT3	8	55	0.1455
FYN	9	73	0.1233
GAA	31	10308	0.0030
GBA	23	10395	0.0022
GFER	26	8099	0.0032
GLA	10	9821	0.0010
GLP1R	117	10527	0.0111
GLS	16	10406	0.0015
GMNN	505	9614	0.0525
GNAS	266	9957	0.0267
GPR35	6	9362	0.0006
GPR55	13	9370	0.0014
GSK3A	5	9362	0.0005
GSK3B	9	9406	0.0010
HIF1A	38	715	0.0531
HIPK1	45	10	4.5000
HPGD	14	7180	0.0019
HRH1	35	11	3.1818
HRH2	8	22	0.3636
HSD17B10	33	1341	0.0246
HSF1	12	9157	0.0013
HSP90AA1	13	9420	0.0014
HSPA5	6	207	0.0290
HTR1A	75	3253	0.0231

HTR1B	33	25	1.3200
HTR2A	85	10178	0.0084
HTR2B	50	34	1.4706
HTR2C	74	32	2.3125
HTR3A	9	7	1.2857
HTR5A	7	9583	0.0007
HTR6	34	28	1.2143
HTR7	30	7	4.2857
HTT	94	8895	0.0106
IDH1	48	23582	0.0020
IL2	22	10268	0.0021
IMPA1	118	8860	0.0133
KAT2A	83	10335	0.0080
KCNH2	47	10344	0.0045
KDM4A	17	2171	0.0078
KDM4E	25	2925	0.0085
KDR	14	59	0.2373
KIT	6	58	0.1034
KMT2A	14	9779	0.0014
L3MBTL1	74	7438	0.0099
LCK	9	75	0.1200
LMNA	376	7155	0.0526
MAP2K1	45	22	2.0455
MAP4K4	6	9	0.6667
MAPK1	73	7902	0.0092
MAPK14	10	66	0.1515
MAPT	115	9750	0.0118
MBNL1	9	7726	0.0012
MCL1	13	8929	0.0015
MPHOSPH8	40	469	0.0853
MTOR	9	560	0.0161

NEK2	47	19	2.4737
NEK3	48	6	8.0000
NFE2L2	194	10385	0.0187
NFKB1	30	10086	0.0030
NOD1	148	9104	0.0163
NOD2	7	10389	0.0007
NPC1	166	9656	0.0172
NPSR1	49	8973	0.0055
NR1H4	12	699	0.0172
NR3C1	49	747	0.0656
OPRD1	10	9588	0.0010
OPRK1	11	9379	0.0012
OPRM1	14	9594	0.0015
PAX8	11	19381	0.0006
PDGFRA	6	56	0.1071
PHOSPHO1	181	9174	0.0197
PIM1	7	66	0.1061
PIN1	11	9856	0.0011
PIP4K2A	161	8349	0.0193
PKM	36	9524	0.0038
PLK1	9	10015	0.0009
PLK2	50	5	10.0000
PLK4	6	7	0.8571
PMP22	41	765	0.0536
POLB	62	9542	0.0065
POLH	12	10291	0.0012
POLI	46	9977	0.0046
POLK	11	10467	0.0011
PPARD	10	682	0.0147
PPARG	9	8316	0.0011
PTGS1	17	33	0.5152

PTGS2	15	26	0.5769
PTH1R	25	10138	0.0025
PTPN7	11	3445	0.0032
RAB9A	214	9606	0.0223
RARA	12	709	0.0169
RECQL	29	264	0.1098
RET	5	58	0.0862
RGS4	99	10168	0.0097
RIPK2	38	10248	0.0037
ROCK2	8	3055	0.0026
RORC	293	10150	0.0289
RXRA	6	553	0.0108
SIGMAR1	64	25	2.5600
SLC47A1	5	19	0.2632
SLC5A7	6	9656	0.0006
SLC6A2	26	50	0.5200
SLC6A3	26	4988	0.0052
SLC6A4	42	33	1.2727
SLCO1B1	6	74	0.0811
SMAD3	160	10207	0.0157
SMN1	258	1908	0.1352
SMN2	164	7499	0.0219
SNCA	9	10347	0.0009
SRPK3	42	13	3.2308
TAAR1	6	10190	0.0006
TARDBP	74	10226	0.0072
TBXAS1	9	7	1.2857
TDP1	147	9072	0.0162
TGFBR1	45	10	4.5000
THPO	19	310	0.0613
THRB	76	9858	0.0077

TNF	8	142	0.0563
TP53	42	11938	0.0035
TRPC4	13	9366	0.0014
TRPV1	14	9566	0.0015
TSHR	96	9997	0.0096
TXNRD1	158	9813	0.0161
USP1	6	10200	0.0006
USP2	5	10092	0.0005
VDR	21	10479	0.0020
VRK2	47	7	6.7143
WNK2	45	9	5.0000
XBP1	7	64	0.1094

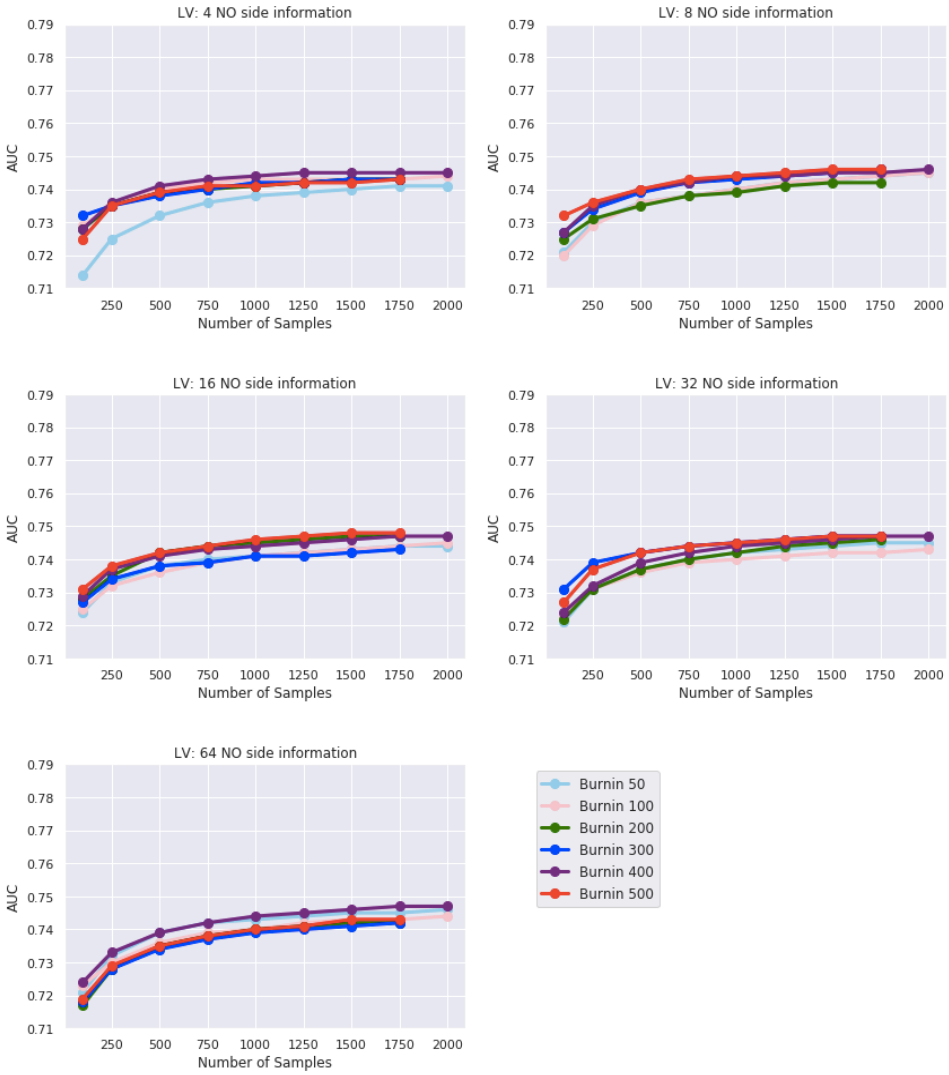


Figure 8.1: Performance of BMF Macau with no side information by using different combinations of latent vectors (LV), burn-in and samples.

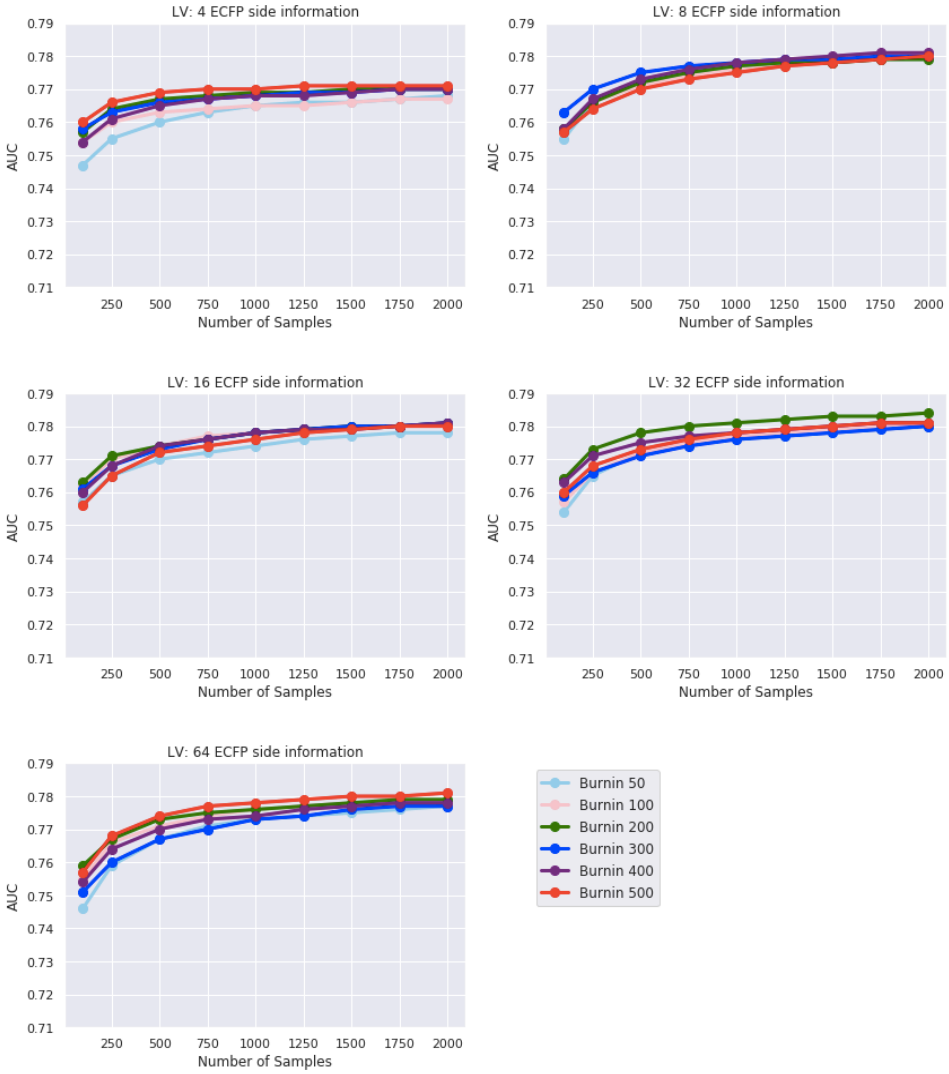


Figure 8.2: Performance of BMF Macau with ECFP as side information by using different combinations of latent vectors, burn-in and samples. Three parameters need to be defined during the training of the models with BMF Macau and these are the number of: Latent Vectors (LV), Number of burn-in and the Number of samples. For that reason, we tried a different combination of these parameters and as we can see there is no change in the performance of the models after e.g. 1000 samples (of which 400 are used for the burn-in) and 8 latent vectors.

Table 8.2: Model Performance of actual and random (y-scrambled) models.

Split	Side Information	Active:Inactive Compound Ratio	AUC (y-scrambled models)
Stratified Shuffle Split	No	1:5	0.51
		1:10	0.50
		all	0.51
	ECFP	1:5	0.50
		1:10	0.51
		all	0.51
	Image	1:5	0.51
		1:10	0.50
		all	0.51
Group Shuffle Split	No	1:5	0.49
		1:10	0.50
		all	0.50
	ECFP	1:5	0.49
		1:10	0.50
		all	0.50
	Image	1:5	0.49
		1:10	0.50
		all	0.50

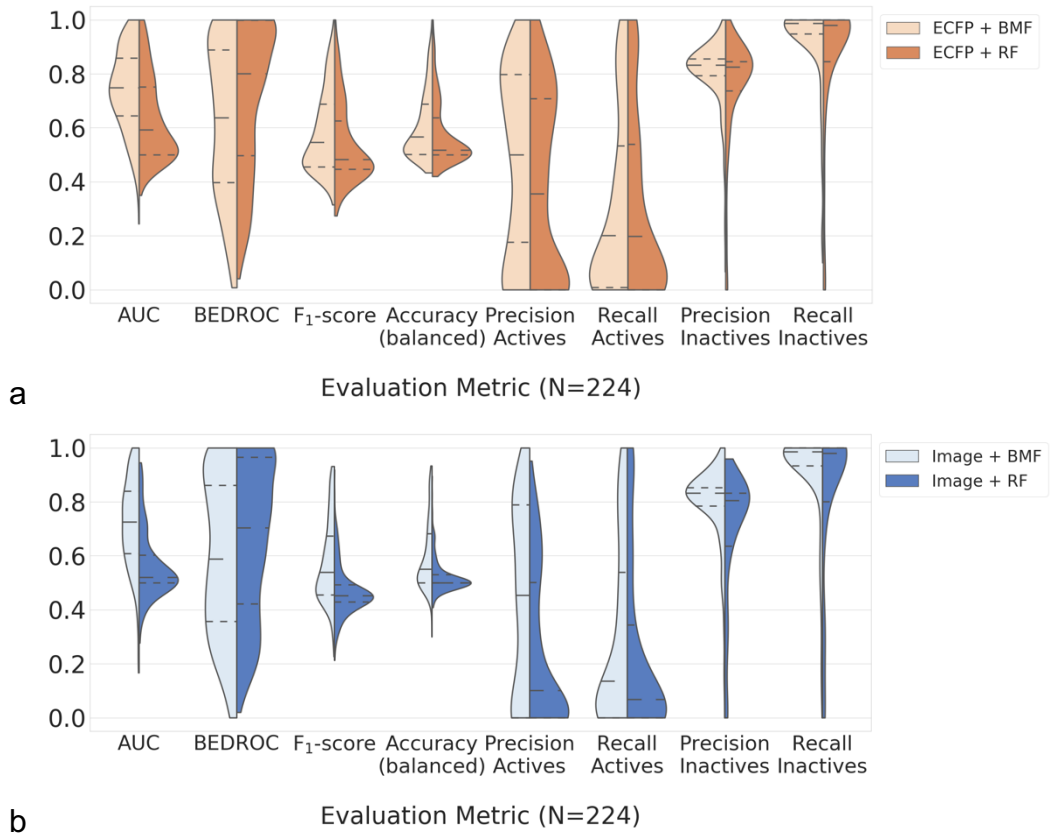
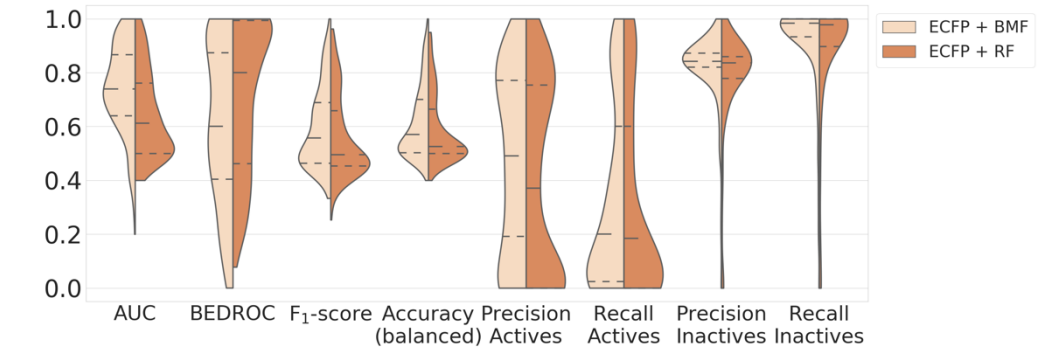
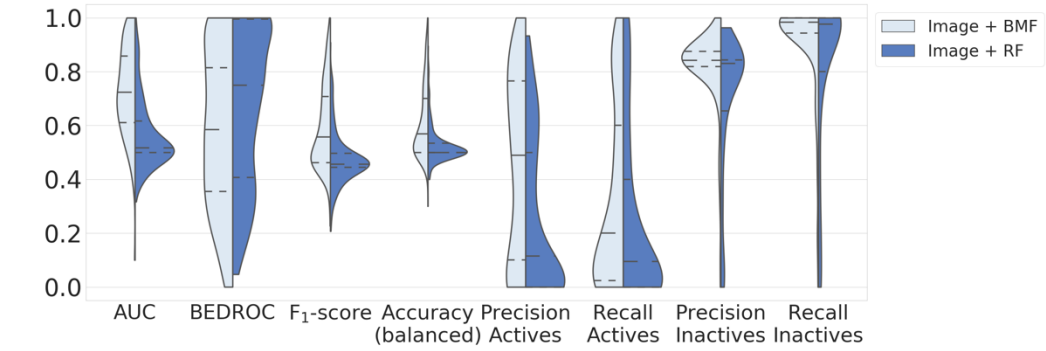


Figure 8.3: Performance of BMF Macau and RF models trained with a) ECFP and b) image-based data as side information with GSS and A:I ratio equal to 1:5 across 224 targets. The dashed lines represent the 25th (quartile 1) to 75th (quartile 3) percentile, and the median of the results distribution and it can be seen that when ECFP are used as compounds' descriptors, the AUC score is higher for BMF Macau compared to RF and the BEDROC score is higher for RF compared to BMF Macau. Still, accuracy, F_1-score, precision and recall are similar for both algorithms. On the other hand, when the image-based data is used as compounds' side information, BMF Macau outperforms RF and that is shown with the all the evaluation metrics used.

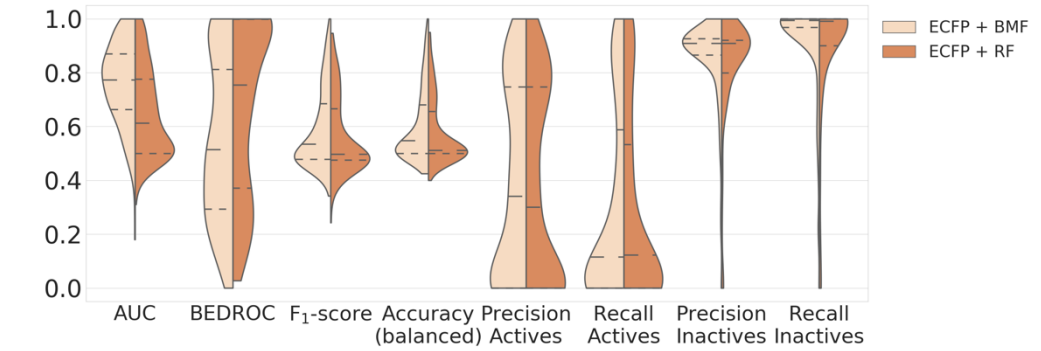


a

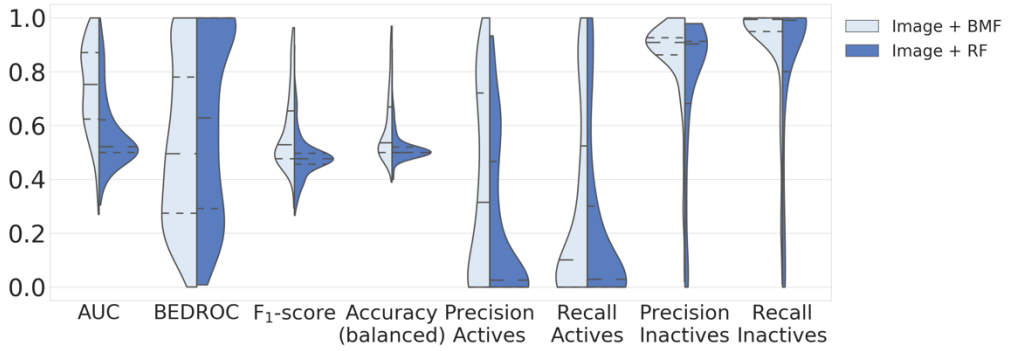


b

Figure 8.4: Performance of BMF Macau and RF models trained with a) ECFP and b) image-based data as side information with SSS and A:I ratio equal to 1:5 across 224 targets. The dashed lines represent the 25th (quartile 1) to 75th (quartile 3) percentile, and the median of the results distribution and it can be seen that when ECFP are used as compounds' descriptors, the AUC score is higher for BMF Macau compared to RF and the BEDROC score is higher for RF compared to BMF Macau. Still, accuracy, F_1-score, precision and recall are similar for both algorithms. On the other hand, when the image-based data is used as compounds' side information, BMF Macau outperforms RF and that is shown with the all the evaluation metrics used.

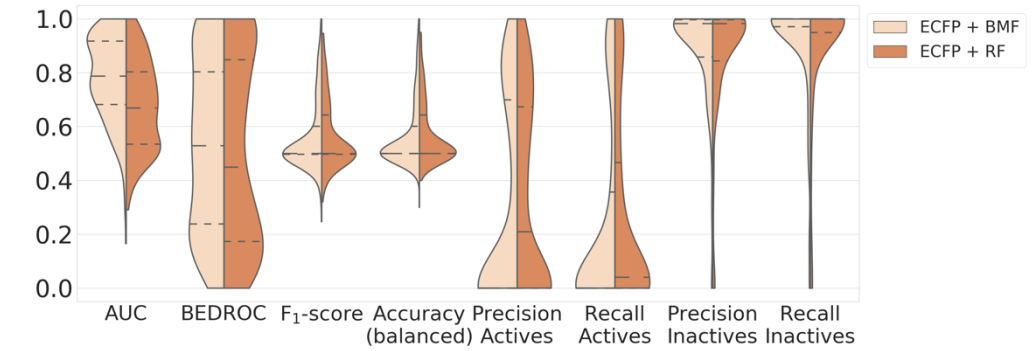


a

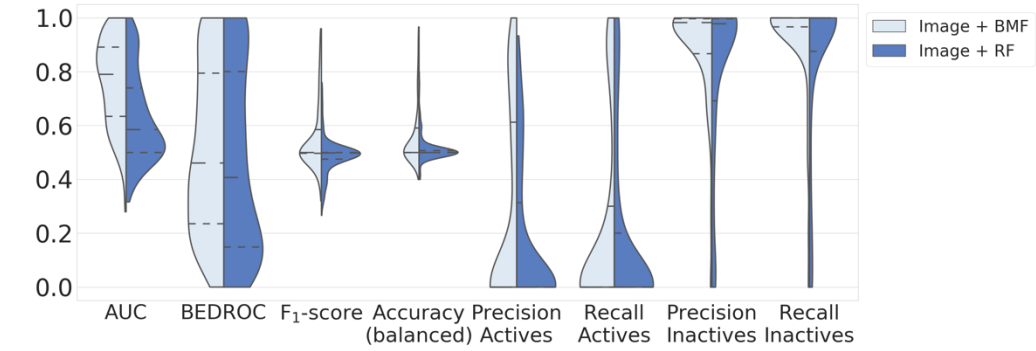


b

Figure 8.5: Performance of BMF Macau and RF models trained with a) ECFP and b) image-based data as side information with SSS and A:I ratio equal to 1:10 across 224 targets. The dashed lines represent the 25th (quartile 1) to 75th (quartile 3) percentile, and the median of the results distribution and it can be seen that when ECFP are used as compounds' descriptors, the AUC score is higher for BMF Macau compared to RF and the BEDROC score is higher for RF compared to BMF Macau. Still, accuracy, F_1-score, precision and recall are similar for both algorithms. On the other hand, when the image-based data is used as compounds' side information, BMF Macau outperforms RF and that is shown with the all the evaluation metrics used.

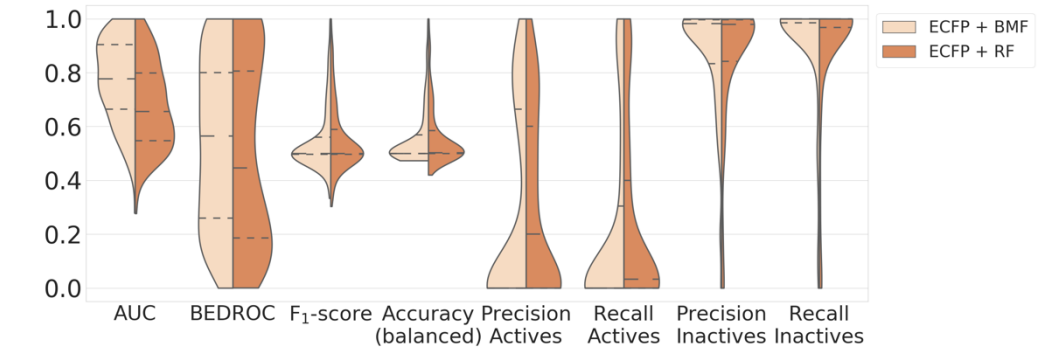


a

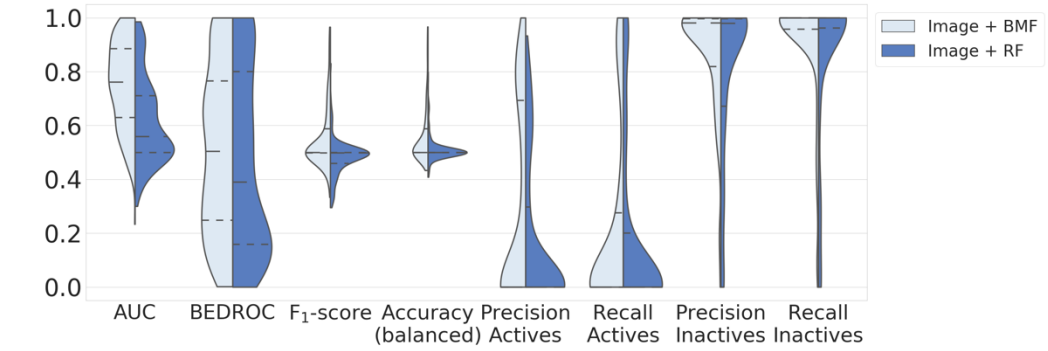


b

Figure 8.6: Performance of BMF Macau and RF models trained with a) ECFP and b) image-based data as side information with SSS and all available bioactivity data across 224 targets. The dashed lines represent the 25th (quartile 1) to 75th (quartile 3) percentile, and the median of the results distribution and it can be seen that when ECFP are used as compounds' descriptors, the AUC and BEDROC score are higher for BMF Macau compared to RF. Still, accuracy, F_1-score, precision and recall are similar for both algorithms. On the other hand, when the image-based data is used as compounds' side information, BMF Macau outperforms RF and that is shown with the all the evaluation metrics used.



a

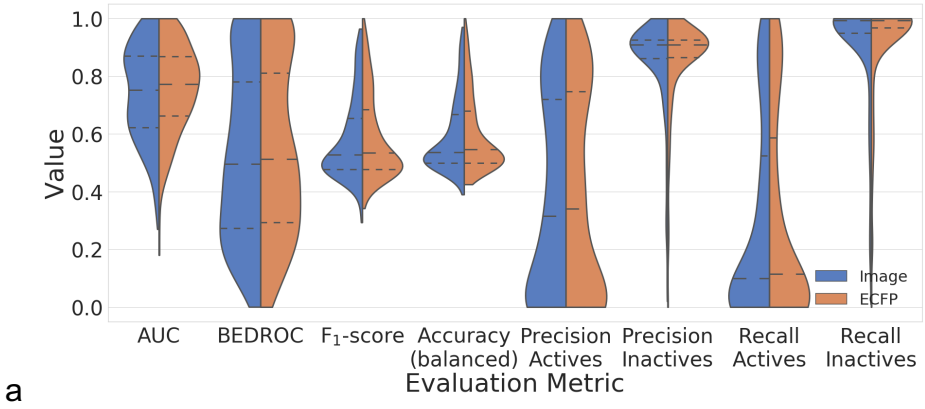


b

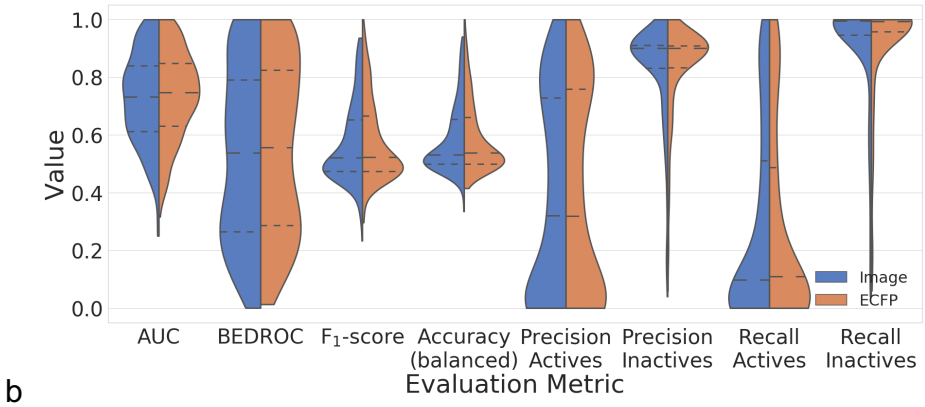
Figure 8.7: Performance of BMF Macau and RF models trained with a) ECFP and b) image-based data as side information with GSS and all available bioactivity data across 224 targets. The dashed lines represent the 25th (quartile 1) to 75th (quartile 3) percentile, and the median of the results distribution and it can be seen that when ECFP are used as compounds' descriptors, the AUC and BEDROC score are higher for BMF Macau compared to RF. Still, accuracy, F_1-score, precision and recall are similar for both algorithms. On the other hand, when the image-based data is used as compounds' side information, BMF Macau outperforms RF and that is shown with the all the evaluation metrics used.

Table 8.3: Model Performance and the number of biological assays that can be predicted with an AUC Better than 0.9, 0.8, and 0.7.

Split	Side Information	AUC	BEDROC	F_1-score	AUC > 0.9	AUC > 0.8	AUC > 0.7	A:I Ratio
Stratified	ECFP	0.74 ± 0.16	0.61 ± 0.30	0.60 ± 0.15	48	87	144	1:5
	Image	0.73 ± 0.17	0.58 ± 0.30	0.59 ± 0.15	41	81	126	
Scaffold - based	ECFP	0.74 ± 0.15	0.62 ± 0.28	0.59 ± 0.15	42	82	141	
	Image	0.72 ± 0.17	0.59 ± 0.29	0.58 ± 0.15	39	78	121	
Stratified	ECFP	0.76 ± 0.16	0.55 ± 0.30	0.59± 0.14	46	101	156	1:10
	Image	0.74± 0.17	0.53 ± 0.30	0.59± 0.14	46	97	131	
Scaffold - based	ECFP	0.74± 0.15	0.56 ± 0.30	0.58± 0.14	39	86	142	
	Image	0.73± 0.16	0.53 ± 0.30	0.57± 0.14	40	84	128	
Stratified	ECFP	0.78 ± 0.16	0.54 ± 0.31	0.56± 0.13	71	109	158	All available data
	Image	0.76 ± 0.16	0.51 ± 0.30	0.55± 0.12	53	107	143	
Scaffold - based	ECFP	0.77 ± 0.15	0.54 ± 0.30	0.55± 0.12	60	105	153	
	Image	0.75 ± 0.16	0.51 ± 0.31	0.55± 0.12	48	95	146	



a



b

Figure 8.8: Performance of models using image data (green) and ECFP data (orange) as side information and two different cross validation strategies (a) Stratified Shuffle Split and b) Group Shuffle Split with A:I ratio equal to 1:10. It can be seen that both types of side information perform well and with no statistically significant difference.

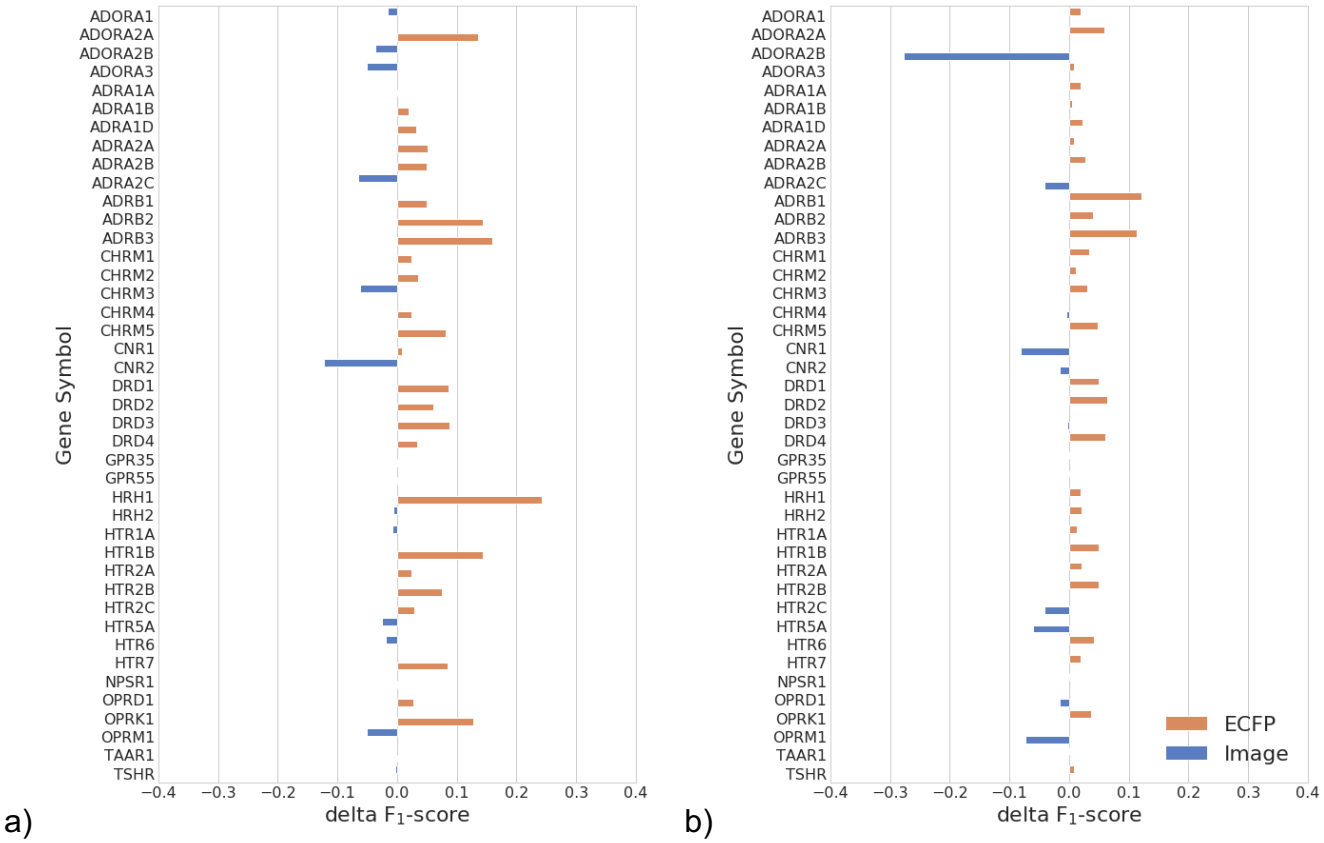


Figure 8.9: Difference between image model F_1-score and the ECFP model F_1-score (delta F_1-score) per Protein in the G-Protein Coupled Receptor 1 family. The bars show the performance of each protein when 1:10 actives to inactives ratio and a) SSS and b) GSS were applied.

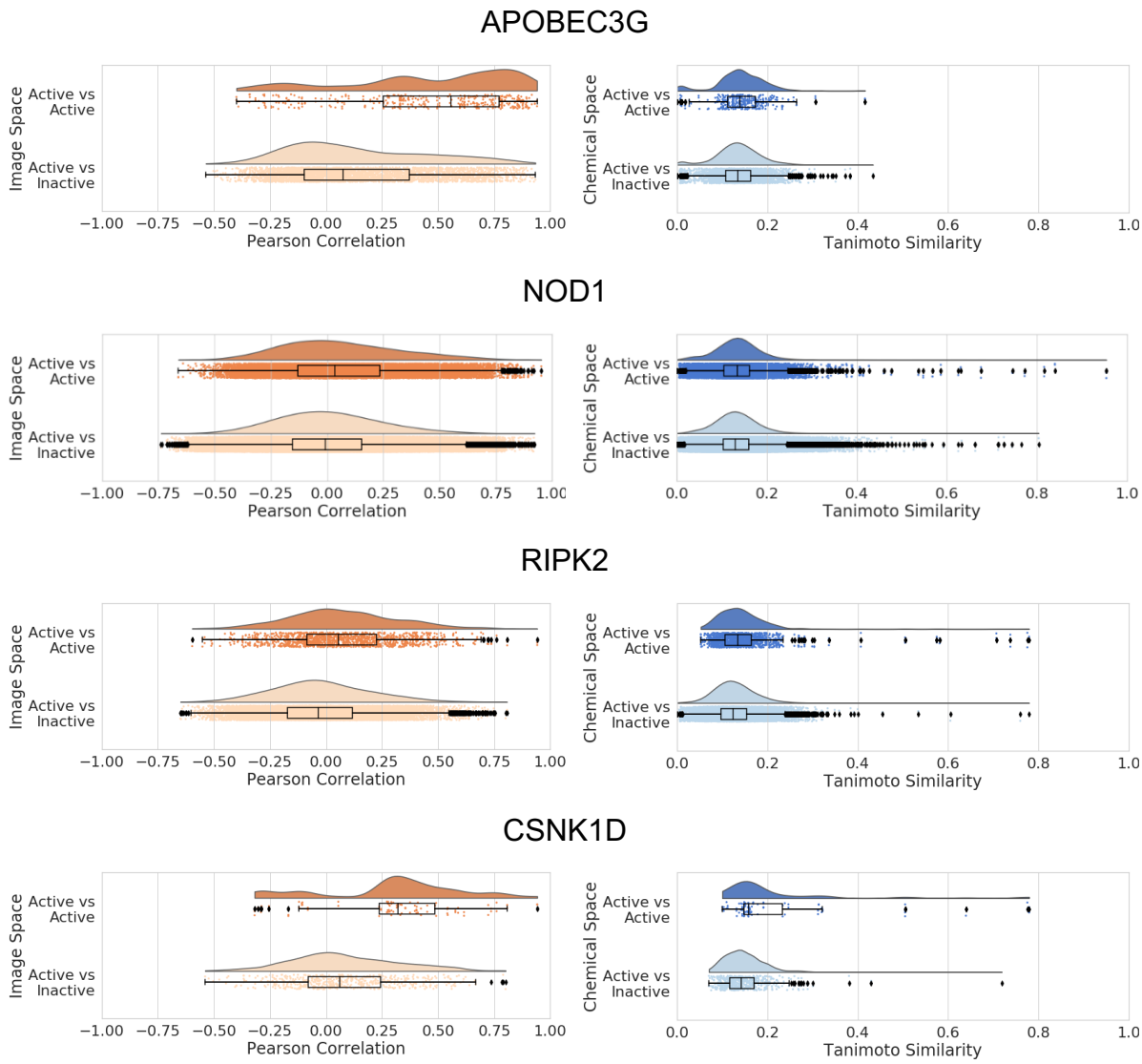


Figure 8.10: Pearson correlation in the image space and Tanimoto coefficient similarity in the chemical space for APOBEC3G, PAX8, RIPK2 and CSNK1D. There is a higher intra-class similarity of chemical features compared to inter-class, something that cannot be observed with the intra- and inter- class similarity in the image-based feature space.

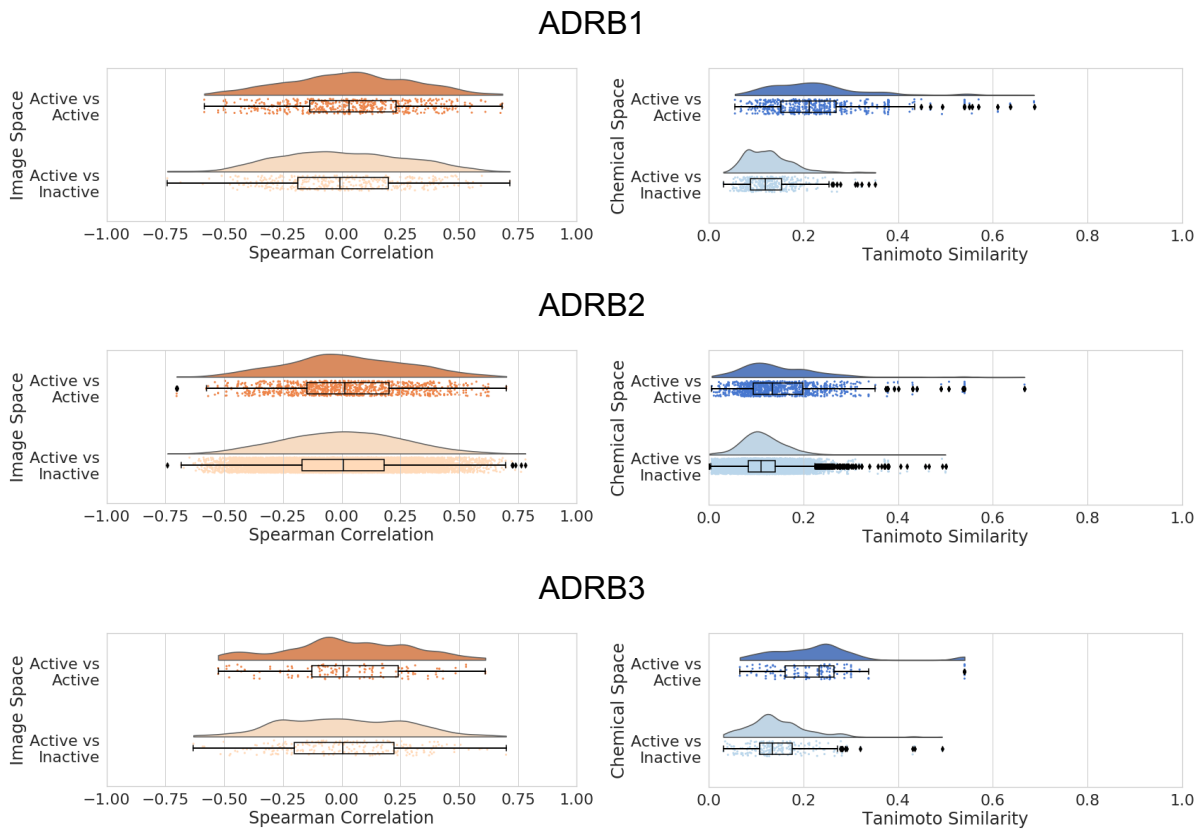


Figure 8.11: Pearson correlation in the image space and Tanimoto coefficient similarity in the chemical space for ADRB1, ADRB2 and ADRB3. There is a higher intra-class similarity of chemical features compared to inter-class, something that cannot be observed with the intra- and inter- class similarity in the image-based feature space.

Table 8.4: Model Performance in terms of F_1-score for the models trained with ECFP as side information and no side information (baseline). The column “Fingerpirnt Type” column indicates, which model performed better (the one trained with ECFP as side information or the other trained with no side information). The colour gradiend indicates the magnitude of the performance difference between the two types of side information.

Target	F_1-score (no side information - Baseline)	F_1-score (ECFP as side information)	Fingerprint type	Δ (F_1-score (Baseline) - F_1-score (ECFP))
OPRK1	0.4894	0.7300	ECFP	-0.2406
OPRD1	0.4969	0.6847	ECFP	-0.1878
PAX8	0.5450	0.7038	ECFP	-0.1587
PIP4K2A	0.4758	0.6304	ECFP	-0.1546
CYP1A1	0.3400	0.4867	ECFP	-0.1467
SLC6A3	0.5631	0.6880	ECFP	-0.1249
TBXAS1	0.4160	0.5387	ECFP	-0.1227
ESR2	0.4781	0.5943	ECFP	-0.1161
PDGFRA	0.5003	0.6133	ECFP	-0.1130
SLC6A4	0.5707	0.6765	ECFP	-0.1058
CA1	0.5961	0.7009	ECFP	-0.1048
ALOX5	0.4571	0.5603	ECFP	-0.1032
ADORA2A	0.5784	0.6780	ECFP	-0.0996
KCNH2	0.4730	0.5726	ECFP	-0.0995
HTR1A	0.6712	0.7658	ECFP	-0.0946
POLK	0.4780	0.5710	ECFP	-0.0930
BLM	0.5807	0.6693	ECFP	-0.0886
CA6	0.8060	0.8940	ECFP	-0.0881
AR	0.7363	0.8231	ECFP	-0.0868
CAR15	0.7600	0.8400	ECFP	-0.0800
CFTR	0.4659	0.5435	ECFP	-0.0776
ROCK2	0.4709	0.5406	ECFP	-0.0697
CHRM1	0.5280	0.5975	ECFP	-0.0695
DRD2	0.8465	0.9142	ECFP	-0.0677

CDK1	0.4574	0.5231	ECFP	-0.0657
DRD4	0.7255	0.7880	ECFP	-0.0625
CA14	0.5560	0.6174	ECFP	-0.0614
RIPK2	0.4884	0.5473	ECFP	-0.0590
CA13	0.9413	1.0000	ECFP	-0.0587
SIGMAR1	0.4481	0.5061	ECFP	-0.0580
DRD1	0.5953	0.6526	ECFP	-0.0573
TXNRD1	0.4869	0.5425	ECFP	-0.0556
NFE2L2	0.5325	0.5878	ECFP	-0.0553
HTR1B	0.5582	0.6124	ECFP	-0.0542
NR3C1	0.7733	0.8271	ECFP	-0.0538
PIM1	0.4623	0.5155	ECFP	-0.0533
CHRM5	0.8828	0.9360	ECFP	-0.0532
SNCA	0.4991	0.5498	ECFP	-0.0507
PTGS2	0.4359	0.4861	ECFP	-0.0502
MAPK14	0.5062	0.5544	ECFP	-0.0483
CSF1R	0.6616	0.7089	ECFP	-0.0473
HIF1A	0.6099	0.6567	ECFP	-0.0468
NR1H4	0.5625	0.6088	ECFP	-0.0464
ADRB3	0.7196	0.7639	ECFP	-0.0443
KDR	0.5409	0.5833	ECFP	-0.0424
MBNL1	0.4737	0.5152	ECFP	-0.0415
IDH1	0.7333	0.7725	ECFP	-0.0391
ADRB2	0.5724	0.6112	ECFP	-0.0387
ADRA2A	0.6427	0.6814	ECFP	-0.0387
HTR2B	0.6639	0.7009	ECFP	-0.0370
RXRA	0.7945	0.8297	ECFP	-0.0352
GLP1R	0.4924	0.5271	ECFP	-0.0346
FLT3	0.5402	0.5733	ECFP	-0.0331
CA2	0.8989	0.9316	ECFP	-0.0327
PIN1	0.4889	0.5201	ECFP	-0.0312

CYP2C9	0.4792	0.5099	ECFP	-0.0307
THPO	0.4647	0.4946	ECFP	-0.0299
NEK2	0.7008	0.7274	ECFP	-0.0266
TP53	0.5659	0.5905	ECFP	-0.0246
ESR1	0.4900	0.5144	ECFP	-0.0244
HTR7	0.7414	0.7657	ECFP	-0.0243
EIF4H	0.4757	0.4999	ECFP	-0.0242
NFKB1	0.5393	0.5633	ECFP	-0.0240
PMP22	0.4813	0.5044	ECFP	-0.0230
BAZ2B	0.4773	0.4999	ECFP	-0.0226
RARA	0.5905	0.6121	ECFP	-0.0216
RGS4	0.4859	0.5075	ECFP	-0.0216
GMNN	0.4962	0.5178	ECFP	-0.0215
SLC5A7	0.4763	0.4971	ECFP	-0.0208
HSF1	0.5648	0.5856	ECFP	-0.0208
PPARG	0.4738	0.4943	ECFP	-0.0205
HRH2	0.3965	0.4165	ECFP	-0.0199
CA9	0.7476	0.7656	ECFP	-0.0180
SLC47A1	0.4389	0.4556	ECFP	-0.0167
HTR2A	0.8177	0.8343	ECFP	-0.0166
MAP4K4	0.2803	0.2963	ECFP	-0.0160
HSD17B10	0.4889	0.5046	ECFP	-0.0157
CHRM4	0.8134	0.8284	ECFP	-0.0150
CYP2D6	0.4711	0.4859	ECFP	-0.0148
TSHR	0.4769	0.4914	ECFP	-0.0145
CHRM2	0.6464	0.6609	ECFP	-0.0145
PPARD	0.6462	0.6606	ECFP	-0.0144
THRB	0.5383	0.5522	ECFP	-0.0139
PTGS1	0.3899	0.4037	ECFP	-0.0138
ADORA1	0.6648	0.6780	ECFP	-0.0132
HSP90AA1	0.4727	0.4857	ECFP	-0.0129

ACHE	0.5078	0.5207	ECFP	-0.0129
TDP1	0.5756	0.5883	ECFP	-0.0127
CA7	0.8042	0.8158	ECFP	-0.0115
BRCA1	0.4902	0.5004	ECFP	-0.0102
ADORA3	0.6665	0.6766	ECFP	-0.0101
CXCL8	0.6648	0.6741	ECFP	-0.0094
ADRB1	0.7339	0.7432	ECFP	-0.0093
CYP2C19	0.4721	0.4812	ECFP	-0.0091
CBX1	0.5164	0.5252	ECFP	-0.0088
L3MBTL1	0.4744	0.4830	ECFP	-0.0086
HTR6	0.6490	0.6571	ECFP	-0.0081
CYP1A2	0.4539	0.4617	ECFP	-0.0078
CSNK1D	0.5296	0.5370	ECFP	-0.0074
LMNA	0.5028	0.5096	ECFP	-0.0068
POLI	0.5322	0.5382	ECFP	-0.0060
MTOR	0.6924	0.6979	ECFP	-0.0055
PHOSPHO1	0.4757	0.4802	ECFP	-0.0045
SRPK3	0.5263	0.5307	ECFP	-0.0044
CYP3A4	0.4756	0.4798	ECFP	-0.0042
NOD1	0.4761	0.4803	ECFP	-0.0042
DYRK1A	0.4739	0.4778	ECFP	-0.0039
ADRA1D	0.7437	0.7475	ECFP	-0.0038
MPHOSPH8	0.4641	0.4672	ECFP	-0.0031
GPR35	0.4596	0.4626	ECFP	-0.0030
CHRM3	0.8908	0.8935	ECFP	-0.0026
CA12	0.9721	0.9746	ECFP	-0.0025
CSNK1A1	0.4452	0.4470	ECFP	-0.0018
EGFR	0.4564	0.4581	ECFP	-0.0017
WNK2	0.5138	0.5154	ECFP	-0.0016
CTNNB1	0.4770	0.4781	ECFP	-0.0011
HTT	0.4759	0.4769	ECFP	-0.0010

SMN1	0.8208	0.8218	ECFP	-0.0010
SMAD3	0.5008	0.5017	ECFP	-0.0008
VDR	0.5649	0.5657	ECFP	-0.0008
CHRNA7	0.3067	0.3073	ECFP	-0.0007
RORC	0.4782	0.4788	ECFP	-0.0007
MAPT	0.4757	0.4760	ECFP	-0.0003
KAT2A	0.4743	0.4744	ECFP	-0.0001
ALOX15	0.4728	0.4728	ECFP	-0.0001
MAPK1	0.4757	0.4758	ECFP	0.0000
IMPA1	0.4742	0.4742	No	0.0000
ALDH1A1	0.4762	0.4761	No	0.0001
TRPC4	0.4727	0.4725	No	0.0002
PKM	0.4735	0.4733	No	0.0003
TRPV1	0.4747	0.4744	No	0.0003
MAP2K1	0.6133	0.6129	No	0.0004
SLCO1B1	0.4615	0.4611	No	0.0004
ALOX15B	0.4723	0.4717	No	0.0006
IL2	0.4731	0.4725	No	0.0006
FEN1	0.4681	0.4674	No	0.0006
KDM4E	0.4762	0.4755	No	0.0007
MCL1	0.4727	0.4719	No	0.0008
NPSR1	0.4755	0.4747	No	0.0008
GLA	0.4762	0.4752	No	0.0010
KDM4A	0.4722	0.4709	No	0.0013
OPRM1	0.4747	0.4733	No	0.0014
PLK4	0.3333	0.3307	No	0.0027
GLS	0.4738	0.4706	No	0.0032
SMN2	0.9274	0.9240	No	0.0034
EEF2K	0.8832	0.8794	No	0.0038
EHMT2	0.4852	0.4804	No	0.0048
CYP19A1	0.7561	0.7506	No	0.0056

ALPL	0.4737	0.4680	No	0.0057
RECQL	0.4822	0.4763	No	0.0059
CLK4	0.4545	0.4485	No	0.0060
TARDBP	0.4805	0.4744	No	0.0060
APEX1	0.4667	0.4603	No	0.0063
HTR2C	0.6674	0.6598	No	0.0076
CNR2	0.6029	0.5950	No	0.0079
ATXN2	0.4827	0.4741	No	0.0085
HIPK1	0.5628	0.5534	No	0.0093
APOBEC3F	0.4831	0.4727	No	0.0104
ADRA1A	0.4985	0.4880	No	0.0106
CDK2	0.4389	0.4283	No	0.0106
DRD3	0.8527	0.8410	No	0.0117
PLK1	0.4945	0.4806	No	0.0138
DDIT3	0.4806	0.4667	No	0.0140
NPC1	0.6989	0.6842	No	0.0146
ADRA2C	0.7881	0.7722	No	0.0160
ADRA1B	0.6746	0.6566	No	0.0181
PLK2	0.6174	0.5965	No	0.0210
CA4	0.9093	0.8881	No	0.0212
APOBEC3G	0.5466	0.5217	No	0.0248
NOD2	0.6975	0.6708	No	0.0267
ADRA2B	0.8325	0.8045	No	0.0280
FLT1	0.6027	0.5730	No	0.0296
HTR5A	0.6970	0.6660	No	0.0310
CSNK1G1	0.8594	0.8282	No	0.0312
ATAD5	0.6552	0.6237	No	0.0315
SLC6A2	0.7473	0.7156	No	0.0317
ABL1	0.5850	0.5471	No	0.0379
FYN	0.6774	0.6374	No	0.0399
RAB9A	0.7514	0.7065	No	0.0448

POLH	0.5204	0.4743	No	0.0461
AHR	0.5195	0.4729	No	0.0466
TNF	0.8752	0.8282	No	0.0470
KIT	0.5752	0.5243	No	0.0510
HRH1	0.7937	0.7380	No	0.0557
CA5A	0.8722	0.8129	No	0.0593
AKR1C3	0.4455	0.3857	No	0.0598
LCK	0.6622	0.5999	No	0.0623
NEK3	0.5972	0.5253	No	0.0719
HTR3A	0.5147	0.4413	No	0.0733
CSNK1G3	0.6813	0.6042	No	0.0771
CNR1	0.5563	0.4729	No	0.0835
TGFBR1	0.6792	0.5949	No	0.0844
XBP1	0.5939	0.5094	No	0.0845
ADORA2B	0.4507	0.3467	No	0.1040
CA5B	0.8985	0.7928	No	0.1057

Table 8.5: Model Performance in terms of F_1-score for the models trained with image-base data as side information and no side information (baseline). The column "Fingerpirnt Type" column indicates, which model performed better (the one trained with image-based data as side information or the other trained with no side information). The colour gradiend indicates the magnitude of the performance difference between the two types of side information.

Target	F_1-score (no side information - Baseline)	F_1-score (image-based data as side information)	Fingerprint type	Δ (F_1-score (Baseline) - F_1-score (Image))
CTNNB1	0.4770	0.8737	Image	-0.3966
PAX8	0.5450	0.7942	Image	-0.2492
OPRD1	0.4969	0.7010	Image	-0.2042
OPRK1	0.4894	0.6933	Image	-0.2038
SLC5A7	0.4763	0.6713	Image	-0.1950
ADORA2B	0.4507	0.6240	Image	-0.1733
PIN1	0.4889	0.6563	Image	-0.1674
RIPK2	0.4884	0.6381	Image	-0.1498
CSNK1D	0.5296	0.6349	Image	-0.1053
SLC6A3	0.5631	0.6668	Image	-0.1038
NOD1	0.4761	0.5747	Image	-0.0986
PDGFRA	0.5003	0.5926	Image	-0.0923
HTR1A	0.6712	0.7524	Image	-0.0812
OPRM1	0.4747	0.5470	Image	-0.0723
GSK3B	0.4737	0.5444	Image	-0.0708
TDP1	0.5756	0.6461	Image	-0.0705
NR1H4	0.5625	0.6326	Image	-0.0702
CA7	0.8042	0.8703	Image	-0.0661
APOBEC3G	0.5466	0.6125	Image	-0.0659
MAP4K4	0.2803	0.3457	Image	-0.0653
PTGS1	0.3899	0.4531	Image	-0.0632
CA5A	0.8722	0.9314	Image	-0.0593
PMP22	0.4813	0.5401	Image	-0.0588
PPARG	0.4738	0.5289	Image	-0.0551

FLT1	0.6027	0.6550	Image	-0.0524
POLH	0.5204	0.5674	Image	-0.0470
APOBEC3F	0.4831	0.5297	Image	-0.0466
DYRK1A	0.4739	0.5196	Image	-0.0457
TP53	0.5659	0.6102	Image	-0.0443
ABL1	0.5850	0.6286	Image	-0.0436
HSPA5	0.4615	0.5038	Image	-0.0422
ADORA2A	0.5784	0.6190	Image	-0.0406
FYN	0.6774	0.7174	Image	-0.0400
TNF	0.8752	0.9116	Image	-0.0364
HSF1	0.5648	0.6011	Image	-0.0363
RGS4	0.4859	0.5215	Image	-0.0355
CHRM1	0.5280	0.5635	Image	-0.0355
NOD2	0.6975	0.7328	Image	-0.0353
AHR	0.5195	0.5547	Image	-0.0352
RXRA	0.7945	0.8297	Image	-0.0352
CGA	0.4748	0.5095	Image	-0.0346
IL2	0.4731	0.5075	Image	-0.0344
HTR2C	0.6674	0.7016	Image	-0.0342
ESR2	0.4781	0.5108	Image	-0.0327
CAR15	0.7600	0.7920	Image	-0.0320
PLK4	0.3333	0.3653	Image	-0.0320
LCK	0.6622	0.6923	Image	-0.0302
ADRA2A	0.6427	0.6728	Image	-0.0301
HTR5A	0.6970	0.7263	Image	-0.0293
CA5B	0.8985	0.9246	Image	-0.0261
CDK1	0.4574	0.4831	Image	-0.0258
EGFR	0.4564	0.4817	Image	-0.0253
PTGS2	0.4359	0.4610	Image	-0.0251
ADRA2C	0.7881	0.8128	Image	-0.0247
GAA	0.4733	0.4979	Image	-0.0247

GMNN	0.4962	0.5193	Image	-0.0231
LMNA	0.5028	0.5247	Image	-0.0219
CFTR	0.4659	0.4876	Image	-0.0217
CXCL8	0.6648	0.6857	Image	-0.0210
SLC6A4	0.5707	0.5917	Image	-0.0209
GLP1R	0.4924	0.5130	Image	-0.0206
DUSP3	0.4738	0.4942	Image	-0.0205
CHRM4	0.8134	0.8328	Image	-0.0194
IDH1	0.7333	0.7523	Image	-0.0190
ATXN2	0.4827	0.5009	Image	-0.0183
ALDH1A1	0.4762	0.4932	Image	-0.0170
PTPN7	0.4681	0.4837	Image	-0.0156
THRB	0.5383	0.5537	Image	-0.0154
ESR1	0.4900	0.5053	Image	-0.0153
BLM	0.5807	0.5945	Image	-0.0138
THPO	0.4647	0.4781	Image	-0.0134
HIF1A	0.6099	0.6197	Image	-0.0098
CYP2C9	0.4792	0.4890	Image	-0.0098
CBX1	0.5164	0.5256	Image	-0.0092
CAR13	0.8367	0.8453	Image	-0.0087
CNR2	0.6029	0.6107	Image	-0.0078
DRD1	0.5953	0.6025	Image	-0.0072
KIT	0.5752	0.5823	Image	-0.0070
TSHR	0.4769	0.4836	Image	-0.0066
FEN1	0.4681	0.4744	Image	-0.0064
PIP4K2A	0.4758	0.4818	Image	-0.0060
CHRM5	0.8828	0.8884	Image	-0.0056
NFE2L2	0.5325	0.5375	Image	-0.0050
HTR7	0.7414	0.7463	Image	-0.0049
HTR1B	0.5582	0.5630	Image	-0.0048
KMT2A	0.4728	0.4775	Image	-0.0047

ALOX5	0.4571	0.4615	Image	-0.0045
DRD2	0.8465	0.8508	Image	-0.0043
SLC47A1	0.4389	0.4431	Image	-0.0042
CHRM2	0.6464	0.6502	Image	-0.0038
TARDBP	0.4805	0.4836	Image	-0.0031
GPR35	0.4596	0.4626	Image	-0.0030
TXNRD1	0.4869	0.4893	Image	-0.0024
DRD4	0.7255	0.7279	Image	-0.0024
ADORA3	0.6665	0.6688	Image	-0.0023
NR3C1	0.7733	0.7752	Image	-0.0019
CYP1A2	0.4539	0.4557	Image	-0.0018
CA4	0.9093	0.9111	Image	-0.0018
EIF4H	0.4757	0.4774	Image	-0.0017
MPHOSPH8	0.4641	0.4656	Image	-0.0015
SMN1	0.8208	0.8223	Image	-0.0015
BRCA1	0.4902	0.4915	Image	-0.0013
CYP2C19	0.4721	0.4730	Image	-0.0009
ALOX15	0.4728	0.4728	Image	-0.0001
PKM	0.4735	0.4736	Image	-0.0001
L3MBTL1	0.4744	0.4744	Image	0.0000
KAT2A	0.4743	0.4743	Image	0.0000
MAPT	0.4757	0.4757	Image	0.0000
PHOSPHO1	0.4757	0.4757	No	0.0000
ARSA	0.4744	0.4744	No	0.0001
MAPK1	0.4757	0.4756	No	0.0001
GNAS	0.4749	0.4748	No	0.0001
POLB	0.4751	0.4749	No	0.0002
KDM4E	0.4762	0.4759	No	0.0003
KCNH2	0.4730	0.4726	No	0.0004
CYP2D6	0.4711	0.4706	No	0.0005
GSK3A	0.4762	0.4757	No	0.0005

ALOX15B	0.4723	0.4718	No	0.0005
ADRB2	0.5724	0.5719	No	0.0006
KDM4A	0.4722	0.4716	No	0.0006
HRH2	0.3965	0.3959	No	0.0006
ERG	0.4706	0.4695	No	0.0011
CA14	0.5560	0.5549	No	0.0011
NPSR1	0.4755	0.4743	No	0.0011
HSP90AA1	0.4727	0.4716	No	0.0012
RORC	0.4782	0.4767	No	0.0014
MAP2K1	0.6133	0.6115	No	0.0018
CNR1	0.5563	0.5543	No	0.0020
SMAD3	0.5008	0.4985	No	0.0024
BAZ2B	0.4773	0.4749	No	0.0024
VDR	0.5649	0.5623	No	0.0026
GLS	0.4738	0.4706	No	0.0032
ROCK2	0.4709	0.4677	No	0.0032
RET	0.4770	0.4734	No	0.0037
HTR2A	0.8177	0.8139	No	0.0038
MAPK14	0.5062	0.5018	No	0.0044
CSNK1G3	0.6813	0.6768	No	0.0045
NPC1	0.6989	0.6942	No	0.0047
PIM1	0.4623	0.4570	No	0.0053
CA13	0.9413	0.9360	No	0.0053
AR	0.7363	0.7308	No	0.0055
FLT3	0.5402	0.5343	No	0.0059
ADORA1	0.6648	0.6587	No	0.0061
CYP1A1	0.3400	0.3333	No	0.0067
ALPL	0.4737	0.4664	No	0.0073
EHMT2	0.4852	0.4776	No	0.0076
CA9	0.7476	0.7398	No	0.0079
RECQL	0.4822	0.4734	No	0.0088

DRD3	0.8527	0.8437	No	0.0090
ATAD5	0.6552	0.6456	No	0.0096
POLI	0.5322	0.5224	No	0.0098
POLK	0.4780	0.4681	No	0.0099
VRK2	0.5933	0.5824	No	0.0109
SMN2	0.9274	0.9160	No	0.0114
HSD17B10	0.4889	0.4775	No	0.0115
HTR2B	0.6639	0.6514	No	0.0124
TBXAS1	0.4160	0.4027	No	0.0133
DDIT3	0.4806	0.4667	No	0.0140
CYP19A1	0.7561	0.7411	No	0.0150
NFKB1	0.5393	0.5234	No	0.0159
CA6	0.8060	0.7891	No	0.0168
WNK2	0.5138	0.4960	No	0.0178
ADRA1D	0.7437	0.7251	No	0.0187
SIGMAR1	0.4481	0.4286	No	0.0195
KDR	0.5409	0.5207	No	0.0203
ADRA1B	0.6746	0.6520	No	0.0226
PLK1	0.4945	0.4714	No	0.0230
RARA	0.5905	0.5665	No	0.0239
NEK3	0.5972	0.5726	No	0.0246
CA1	0.5961	0.5693	No	0.0268
CHRM3	0.8908	0.8639	No	0.0270
SNCA	0.4991	0.4708	No	0.0283
NEK2	0.7008	0.6717	No	0.0291
ADRA1A	0.4985	0.4691	No	0.0294
CSNK1G1	0.8594	0.8273	No	0.0321
CDK2	0.4389	0.4061	No	0.0329
HTR6	0.6490	0.6160	No	0.0330
CSNK1A1	0.4452	0.4113	No	0.0339
CA2	0.8989	0.8632	No	0.0357

PPARD	0.6462	0.6054	No	0.0408
RAB9A	0.7514	0.7095	No	0.0418
HIPK1	0.5628	0.5085	No	0.0542
ADRA2B	0.8325	0.7778	No	0.0547
CA12	0.9721	0.9163	No	0.0557
PLK2	0.6174	0.5546	No	0.0629
XBP1	0.5939	0.5304	No	0.0635
MTOR	0.6924	0.6232	No	0.0692
ADRB3	0.7196	0.6504	No	0.0692
SRPK3	0.5263	0.4543	No	0.0720
CHRNA7	0.3067	0.2333	No	0.0733
HRH1	0.7937	0.7189	No	0.0748
SLC6A2	0.7473	0.6672	No	0.0801
ACHE	0.5078	0.4276	No	0.0802
AKR1C3	0.4455	0.3560	No	0.0894
TGFBR1	0.6792	0.5825	No	0.0967
ADRB1	0.7339	0.6221	No	0.1118
CLK4	0.4545	0.3238	No	0.1307
EEF2K	0.8832	0.7489	No	0.1342
CSF1R	0.6616	0.5225	No	0.1391
HTR3A	0.5147	0.3347	No	0.1800

Table 8.6: Model Performance in terms of F_1-score for the models trained with ECFP and the models trained with image-based data as side information. The column "Fingerpirnt Type" column indicates, which model performed better (the one trained with image-based data as side information or the other trained with ECFP as side information). The colour gradiend indicates the magnitude of the performance difference between the two types of side information.

Target	F_1-score (image-based data as side information)	F_1-score (ECFP as side information)	Fingerprint type	Δ (F1-score (Image) - F1-score (ECFP))
CSF1R	0.5225	0.7089	ECFP	-0.1864
CYP1A1	0.3333	0.4867	ECFP	-0.1533
PIP4K2A	0.4818	0.6304	ECFP	-0.1486
TBXAS1	0.4027	0.5387	ECFP	-0.1360
CA1	0.5693	0.7009	ECFP	-0.1316
EEF2K	0.7489	0.8794	ECFP	-0.1305
CLK4	0.3238	0.4485	ECFP	-0.1247
ADRB1	0.6221	0.7432	ECFP	-0.1210
ADRB3	0.6504	0.7639	ECFP	-0.1135
HTR3A	0.3347	0.4413	ECFP	-0.1067
CA6	0.7891	0.8940	ECFP	-0.1049
POLK	0.4681	0.5710	ECFP	-0.1029
KCNH2	0.4726	0.5726	ECFP	-0.1000
ALOX5	0.4615	0.5603	ECFP	-0.0987
ACHE	0.4276	0.5207	ECFP	-0.0932
AR	0.7308	0.8231	ECFP	-0.0923
SLC6A4	0.5917	0.6765	ECFP	-0.0848
ESR2	0.5108	0.5943	ECFP	-0.0835
SNCA	0.4708	0.5498	ECFP	-0.0790
SIGMAR1	0.4286	0.5061	ECFP	-0.0775
SRPK3	0.4543	0.5307	ECFP	-0.0763
BLM	0.5945	0.6693	ECFP	-0.0748
MTOR	0.6232	0.6979	ECFP	-0.0747
CHRNA7	0.2333	0.3073	ECFP	-0.0740

ROCK2	0.4677	0.5406	ECFP	-0.0729
CA2	0.8632	0.9316	ECFP	-0.0684
CA13	0.9360	1.0000	ECFP	-0.0640
DRD2	0.8508	0.9142	ECFP	-0.0634
KDR	0.5207	0.5833	ECFP	-0.0626
CA14	0.5549	0.6174	ECFP	-0.0625
DRD4	0.7279	0.7880	ECFP	-0.0601
ADORA2A	0.6190	0.6780	ECFP	-0.0590
PIM1	0.4570	0.5155	ECFP	-0.0585
CA12	0.9163	0.9746	ECFP	-0.0583
CFTR	0.4876	0.5435	ECFP	-0.0559
NEK2	0.6717	0.7274	ECFP	-0.0556
PPARD	0.6054	0.6606	ECFP	-0.0552
TXNRD1	0.4893	0.5425	ECFP	-0.0532
MAPK14	0.5018	0.5544	ECFP	-0.0526
NR3C1	0.7752	0.8271	ECFP	-0.0519
NFE2L2	0.5375	0.5878	ECFP	-0.0503
DRD1	0.6025	0.6526	ECFP	-0.0501
HTR1B	0.5630	0.6124	ECFP	-0.0494
HTR2B	0.6514	0.7009	ECFP	-0.0494
SLC6A2	0.6672	0.7156	ECFP	-0.0484
CAR15	0.7920	0.8400	ECFP	-0.0480
CHRM5	0.8884	0.9360	ECFP	-0.0477
RARA	0.5665	0.6121	ECFP	-0.0456
HIPK1	0.5085	0.5534	ECFP	-0.0449
PLK2	0.5546	0.5965	ECFP	-0.0419
MBNL1	0.4737	0.5152	ECFP	-0.0415
HTR6	0.6160	0.6571	ECFP	-0.0411
CDK1	0.4831	0.5231	ECFP	-0.0400
NFKB1	0.5234	0.5633	ECFP	-0.0399
ADRB2	0.5719	0.6112	ECFP	-0.0393

FLT3	0.5343	0.5733	ECFP	-0.0390
HIF1A	0.6197	0.6567	ECFP	-0.0370
OPRK1	0.6933	0.7300	ECFP	-0.0368
CSNK1A1	0.4113	0.4470	ECFP	-0.0357
CHRM1	0.5635	0.5975	ECFP	-0.0341
AKR1C3	0.3560	0.3857	ECFP	-0.0296
CHRM3	0.8639	0.8935	ECFP	-0.0296
HSD17B10	0.4775	0.5046	ECFP	-0.0272
ADRA2B	0.7778	0.8045	ECFP	-0.0267
CA9	0.7398	0.7656	ECFP	-0.0259
PTGS2	0.4610	0.4861	ECFP	-0.0252
BAZ2B	0.4749	0.4999	ECFP	-0.0249
EIF4H	0.4774	0.4999	ECFP	-0.0226
ADRA1D	0.7251	0.7475	ECFP	-0.0224
CDK2	0.4061	0.4283	ECFP	-0.0223
SLC6A3	0.6668	0.6880	ECFP	-0.0211
CYP2C9	0.4890	0.5099	ECFP	-0.0209
PDGFRA	0.5926	0.6133	ECFP	-0.0207
HRH2	0.3959	0.4165	ECFP	-0.0206
HTR2A	0.8139	0.8343	ECFP	-0.0204
IDH1	0.7523	0.7725	ECFP	-0.0201
HTR7	0.7463	0.7657	ECFP	-0.0194
WNK2	0.4960	0.5154	ECFP	-0.0193
ADORA1	0.6587	0.6780	ECFP	-0.0193
HRH1	0.7189	0.7380	ECFP	-0.0191
ADRA1A	0.4691	0.4880	ECFP	-0.0189
THPO	0.4781	0.4946	ECFP	-0.0165
POLI	0.5224	0.5382	ECFP	-0.0157
CYP2D6	0.4706	0.4859	ECFP	-0.0153
HSP90AA1	0.4716	0.4857	ECFP	-0.0141
GLP1R	0.5130	0.5271	ECFP	-0.0141

HTR1A	0.7524	0.7658	ECFP	-0.0134
SLC47A1	0.4431	0.4556	ECFP	-0.0125
TGFBR1	0.5825	0.5949	ECFP	-0.0124
VRK2	0.5824	0.5933	ECFP	-0.0109
CHRM2	0.6502	0.6609	ECFP	-0.0106
CYP19A1	0.7411	0.7506	ECFP	-0.0094
PLK1	0.4714	0.4806	ECFP	-0.0092
ESR1	0.5053	0.5144	ECFP	-0.0091
BRCA1	0.4915	0.5004	ECFP	-0.0089
ADRA2A	0.6728	0.6814	ECFP	-0.0086
L3MBTL1	0.4744	0.4830	ECFP	-0.0086
CYP2C19	0.4730	0.4812	ECFP	-0.0083
SMN2	0.9160	0.9240	ECFP	-0.0080
TSHR	0.4836	0.4914	ECFP	-0.0078
ADORA3	0.6688	0.6766	ECFP	-0.0078
CYP1A2	0.4557	0.4617	ECFP	-0.0060
ADRA1B	0.6520	0.6566	ECFP	-0.0046
PHOSPHO1	0.4757	0.4802	ECFP	-0.0045
CYP3A4	0.4756	0.4798	ECFP	-0.0042
RET	0.4734	0.4770	ECFP	-0.0037
VDR	0.5623	0.5657	ECFP	-0.0034
SMAD3	0.4985	0.5017	ECFP	-0.0032
RECQL	0.4734	0.4763	ECFP	-0.0029
EHMT2	0.4776	0.4804	ECFP	-0.0028
RORC	0.4767	0.4788	ECFP	-0.0021
MPHOSPH8	0.4656	0.4672	ECFP	-0.0016
ALPL	0.4664	0.4680	ECFP	-0.0016
MAP2K1	0.6115	0.6129	ECFP	-0.0013
ERG	0.4695	0.4706	ECFP	-0.0011
HTT	0.4759	0.4769	ECFP	-0.0010
CSNK1G1	0.8273	0.8282	ECFP	-0.0009

GSK3A	0.4757	0.4762	ECFP	-0.0005
NPSR1	0.4743	0.4747	ECFP	-0.0003
MAPT	0.4757	0.4760	ECFP	-0.0003
POLB	0.4749	0.4751	ECFP	-0.0002
MAPK1	0.4756	0.4758	ECFP	-0.0001
GNAS	0.4748	0.4749	ECFP	-0.0001
KAT2A	0.4743	0.4744	ECFP	-0.0001
ARSA	0.4744	0.4744	ECFP	-0.0001
ALOX15B	0.4718	0.4717	Image	0.0000
IMPA1	0.4742	0.4742	Image	0.0000
TRPC4	0.4727	0.4725	Image	0.0002
PKM	0.4736	0.4733	Image	0.0003
TRPV1	0.4747	0.4744	Image	0.0003
KDM4E	0.4759	0.4755	Image	0.0004
CBX1	0.5256	0.5252	Image	0.0004
SLCO1B1	0.4615	0.4611	Image	0.0004
SMN1	0.8223	0.8218	Image	0.0005
KDM4A	0.4716	0.4709	Image	0.0007
MCL1	0.4727	0.4719	Image	0.0008
GLA	0.4762	0.4752	Image	0.0010
THRB	0.5537	0.5522	Image	0.0015
GMNN	0.5193	0.5178	Image	0.0015
DRD3	0.8437	0.8410	Image	0.0027
RAB9A	0.7095	0.7065	Image	0.0030
CHRM4	0.8328	0.8284	Image	0.0044
KMT2A	0.4775	0.4728	Image	0.0047
APEX1	0.4667	0.4603	Image	0.0063
FEN1	0.4744	0.4674	Image	0.0070
CAR13	0.8453	0.8367	Image	0.0087
TARDBP	0.4836	0.4744	Image	0.0091
NPC1	0.6942	0.6842	Image	0.0100

CXCL8	0.6857	0.6741	Image	0.0116
RGS4	0.5215	0.5075	Image	0.0140
LMNA	0.5247	0.5096	Image	0.0151
HSF1	0.6011	0.5856	Image	0.0156
PTPN7	0.4837	0.4681	Image	0.0156
CNR2	0.6107	0.5950	Image	0.0157
OPRD1	0.7010	0.6847	Image	0.0164
ALDH1A1	0.4932	0.4761	Image	0.0172
TP53	0.6102	0.5905	Image	0.0197
DUSP3	0.4942	0.4738	Image	0.0205
XBP1	0.5304	0.5094	Image	0.0210
ATAD5	0.6456	0.6237	Image	0.0219
CA4	0.9111	0.8881	Image	0.0230
EGFR	0.4817	0.4581	Image	0.0236
NR1H4	0.6326	0.6088	Image	0.0238
GAA	0.4979	0.4733	Image	0.0247
ATXN2	0.5009	0.4741	Image	0.0268
PPARG	0.5289	0.4943	Image	0.0346
CGA	0.5095	0.4748	Image	0.0346
PLK4	0.3653	0.3307	Image	0.0347
IL2	0.5075	0.4725	Image	0.0349
PMP22	0.5401	0.5044	Image	0.0357
ADRA2C	0.8128	0.7722	Image	0.0407
DYRK1A	0.5196	0.4778	Image	0.0417
HTR2C	0.7016	0.6598	Image	0.0417
HSPA5	0.5038	0.4615	Image	0.0422
NEK3	0.5726	0.5253	Image	0.0473
MAP4K4	0.3457	0.2963	Image	0.0493
PTGS1	0.4531	0.4037	Image	0.0494
CA7	0.8703	0.8158	Image	0.0545
APOBEC3F	0.5297	0.4727	Image	0.0569

TDP1	0.6461	0.5883	Image	0.0578
KIT	0.5823	0.5243	Image	0.0580
HTR5A	0.7263	0.6660	Image	0.0603
NOD2	0.7328	0.6708	Image	0.0619
GSK3B	0.5444	0.4737	Image	0.0708
CSNK1G3	0.6768	0.6042	Image	0.0726
OPRM1	0.5470	0.4733	Image	0.0737
FYN	0.7174	0.6374	Image	0.0800
CNR1	0.5543	0.4729	Image	0.0815
ABL1	0.6286	0.5471	Image	0.0815
AHR	0.5547	0.4729	Image	0.0818
FLT1	0.6550	0.5730	Image	0.0820
TNF	0.9116	0.8282	Image	0.0834
PAX8	0.7942	0.7038	Image	0.0905
APOBEC3G	0.6125	0.5217	Image	0.0907
RIPK2	0.6381	0.5473	Image	0.0908
LCK	0.6923	0.5999	Image	0.0925
POLH	0.5674	0.4743	Image	0.0931
NOD1	0.5747	0.4803	Image	0.0944
CSNK1D	0.6349	0.5370	Image	0.0979
CA5A	0.9314	0.8129	Image	0.1185
CA5B	0.9246	0.7928	Image	0.1318
PIN1	0.6563	0.5201	Image	0.1362
SLC5A7	0.6713	0.4971	Image	0.1742
ADORA2B	0.6240	0.3467	Image	0.2773
CTNNB1	0.8737	0.4781	Image	0.3956

Table 8.7: Background information about the source of bioactivity datapoints and the assay ids and detection methods used to produce the bioactivity datapoints. These bioactivity datapoints when used with image features as side information showed an f-score difference of 0.5 compared to when used with ECFP fingerprints as side information or no side inforation for both Stratified Shuffle Split and Group Shuffle Split.

Target (gene symbol)	Source	Assay id (percentage of datapoints in each assay)	Detection method
CTNNB1	PubChem	1665 (99%)	Fluorescence Intensity
NOD1	PubChem	1578 (93%)	
RIPK2	PubChem	624267 (75%)	
APOBEC3G	PubChem	493012 (29%) 602310 (8%) 651812 (7%)	
	ChEMBL20	809311 (30%)	Assay no longer available
CSNK1D	ChEMBL20	887829 (37%) 887830 (34%)	

9. Appendix – Chapter 4

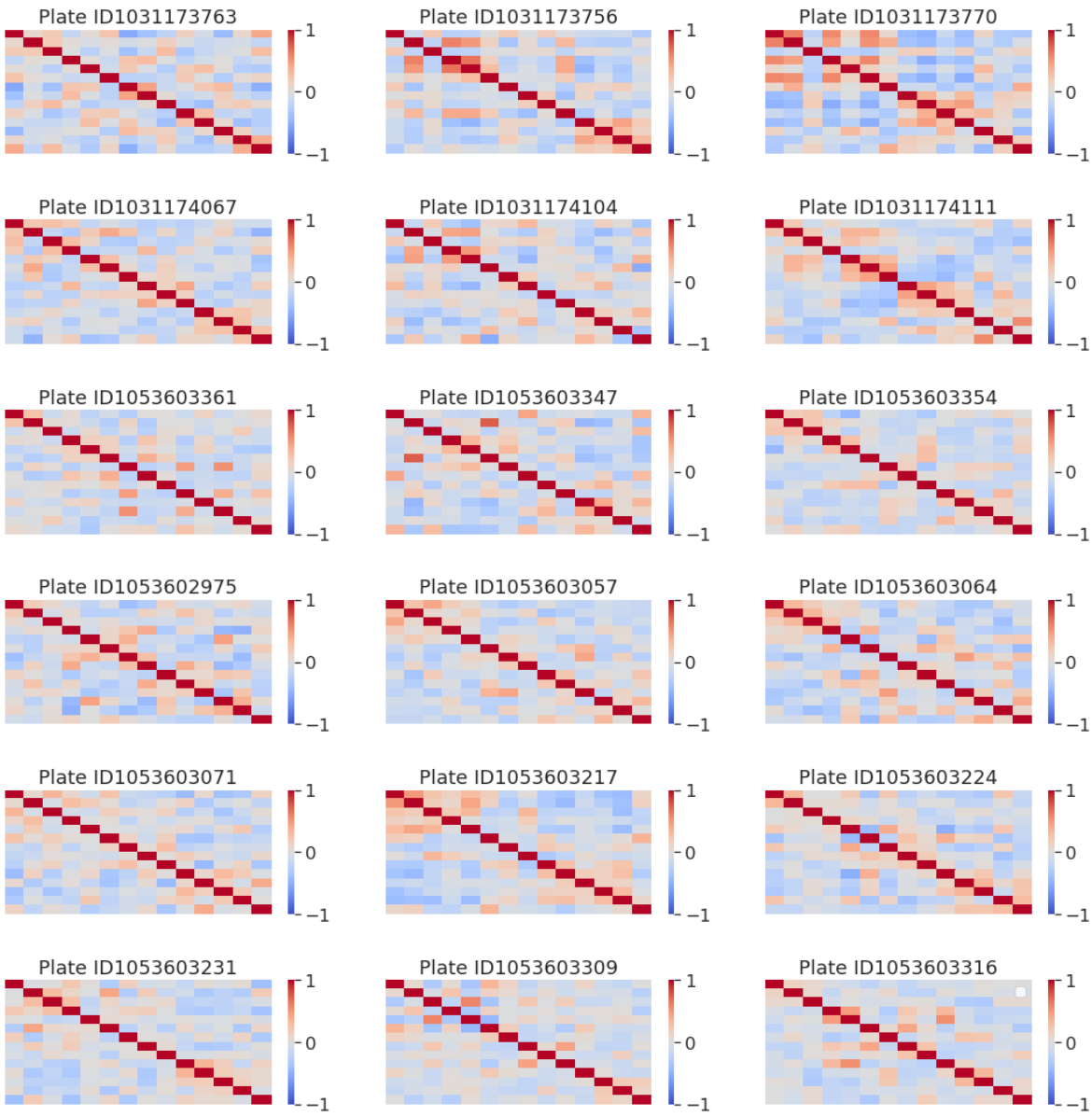


Figure 9.1: Intra-plate Pearson correlation across plates for the neutral control (DMSO) wells.

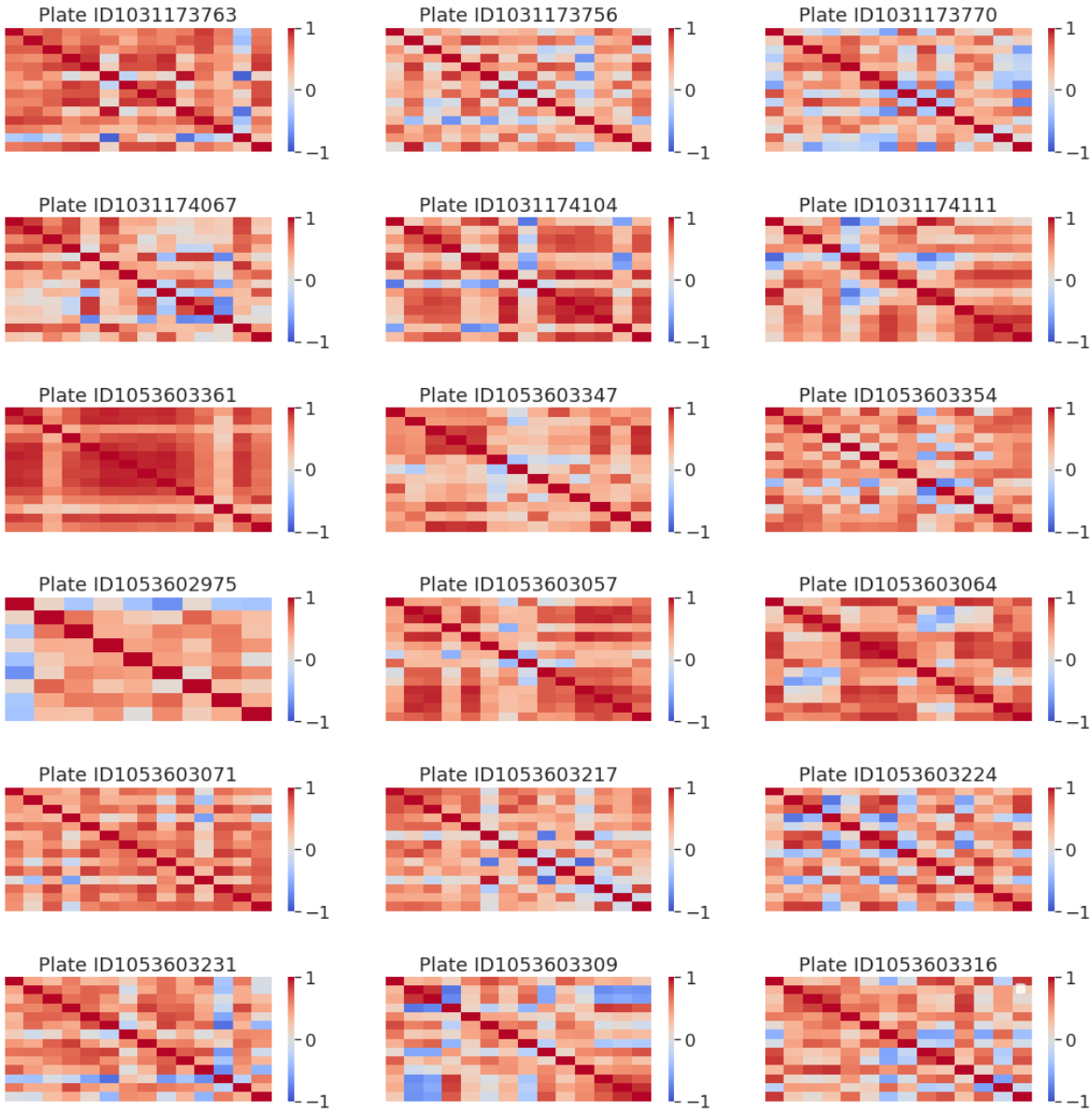


Figure 9.2: Intra-plate Pearson correlation across plates for the positive control (mitoxantrone) wells.

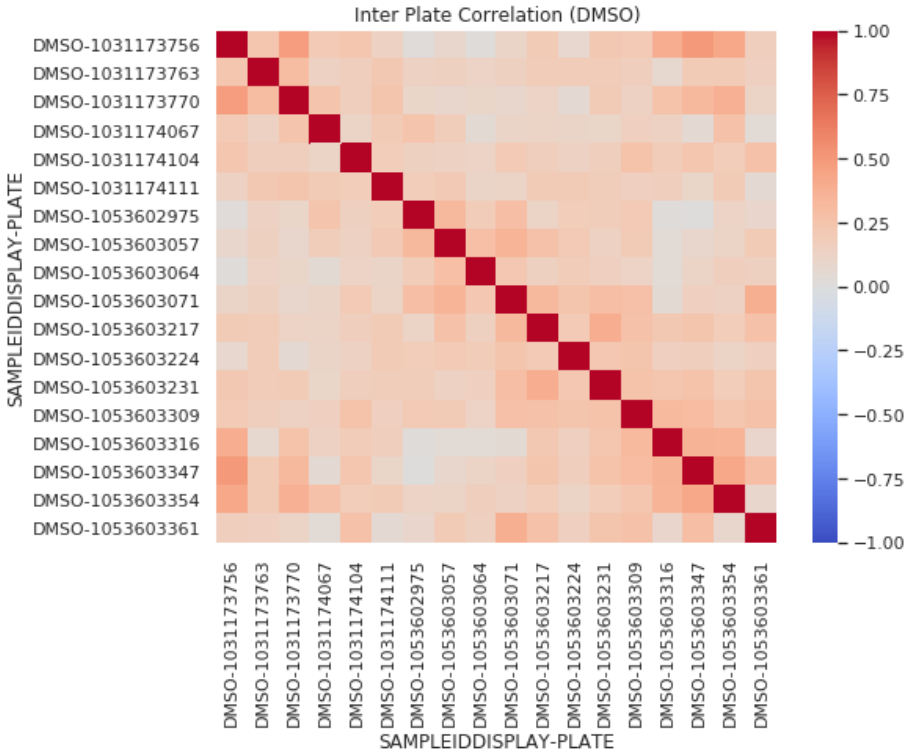


Figure 9.3: Inter-plate Pearson correlation across plates for the neutral control (DMSO) wells.

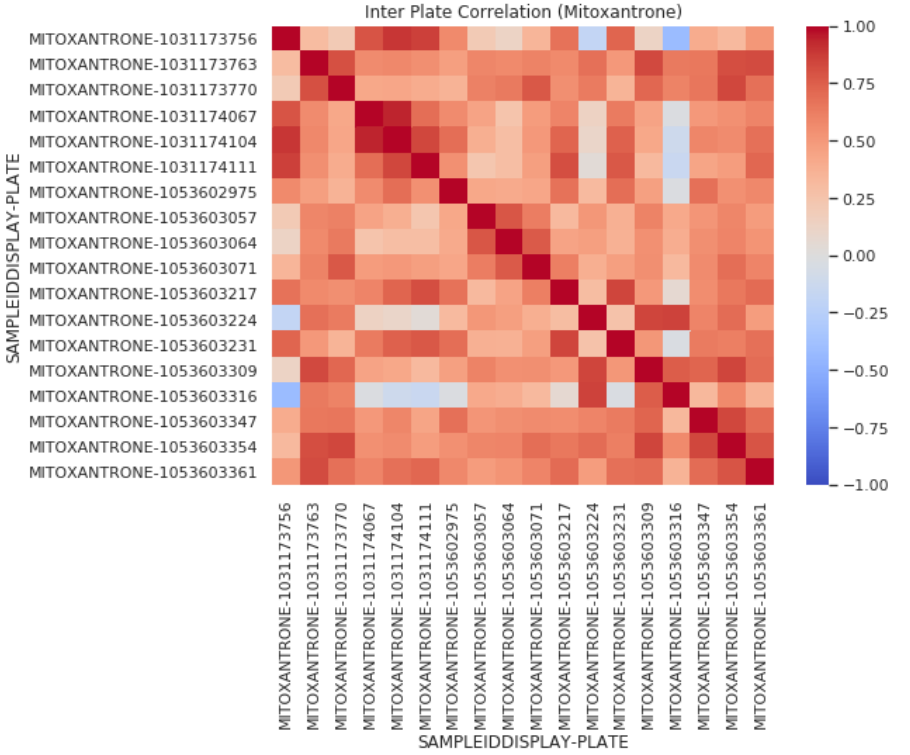


Figure 9.4: Inter-plate Pearson correlation across plates for the neutral control (DMSO) wells.

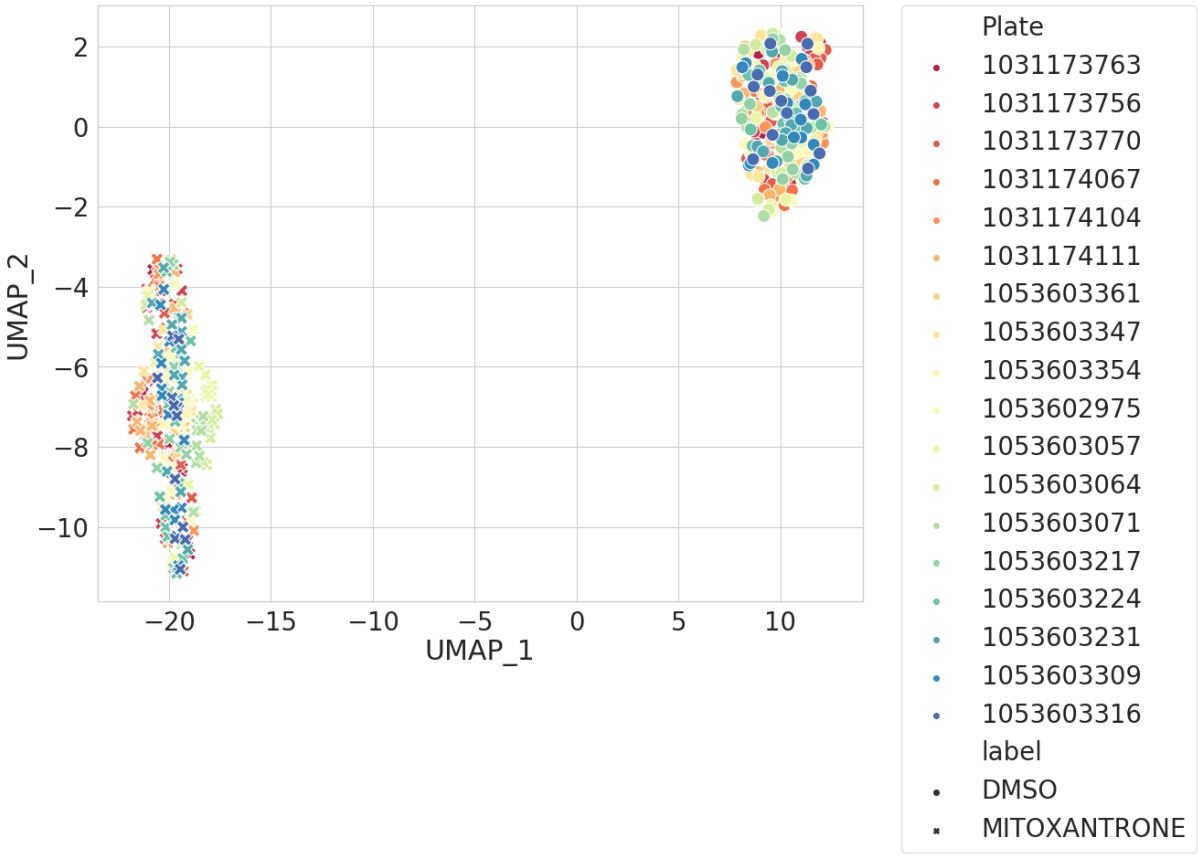


Figure 9.5: UMAP dimensionality reduction on positive control (mitoxantrone) and neutral control (DMSO) wells.

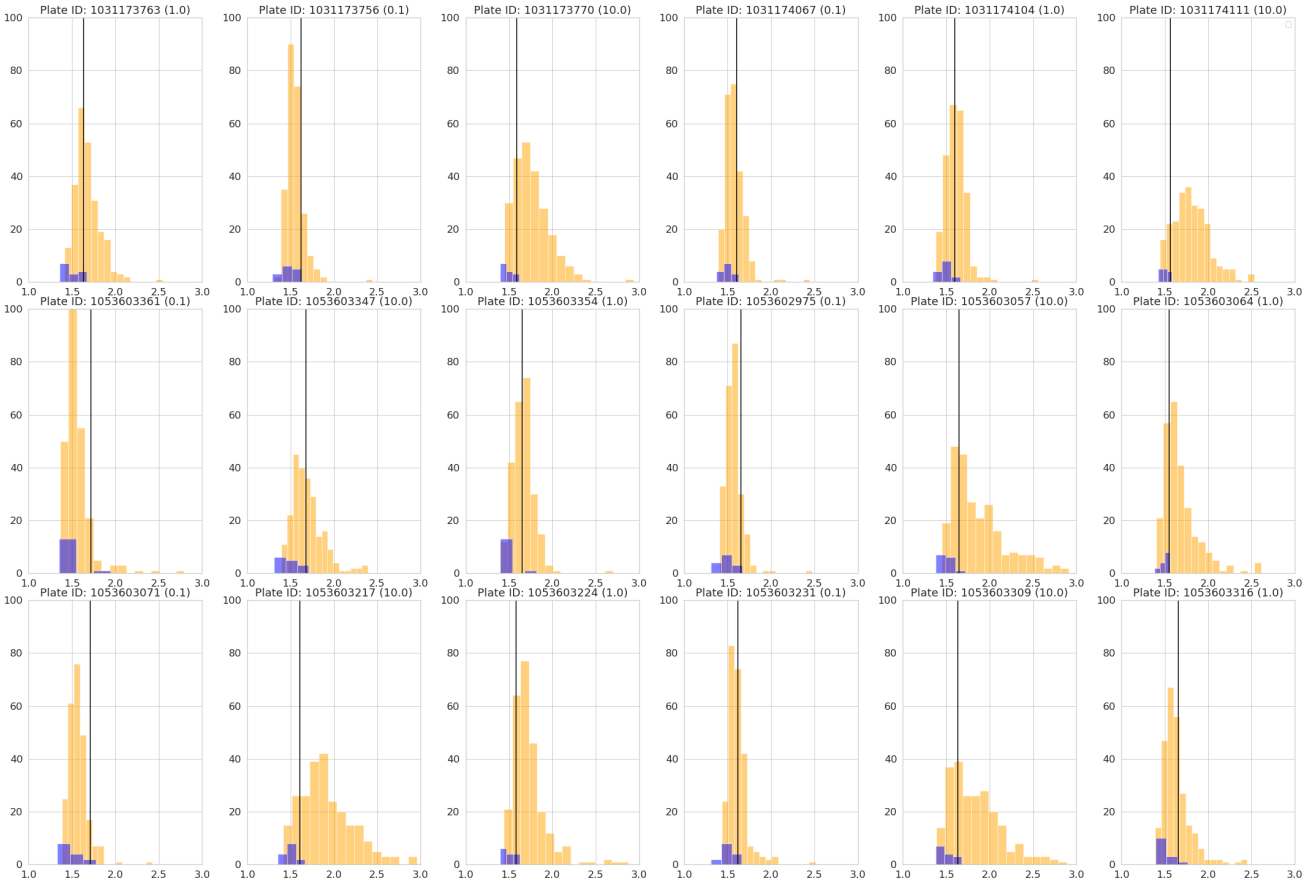


Figure 9.6: Euclidean distance between compounds and the mean of the neutral control (orange distribution) and between the controls and the mean of the controls (blue distribution). The vertical black line corresponds to the 95th percentile of the blue distribution and corresponds to the threshold to separate compound profiles between active (above the threshold) and inactives (below the threshold) on Cell Painting assay. The term “active” refers to whether compounds are able to change the cell morphology in the Cell Painting assay.

Table 9.1: Considered machine learning hyperparameters. Hyperparameters were systematically evaluated using hyperopt python package.

Algorithm	Hyperparameter	Values
Random Forest	max_depth	3-18 with increments of 1
	n_estimators	100-1000 with increments of 100
	min_samples_split	2-50
	min_samples_leaf	1-15 with increments of 1
Support Vector Classifier (with rbf kernel)	gamma	10^{-10}-1
	C	10^{-4}-1000
eXtreme Gradient Boosting (XGB)	max_depth	3-18 with increments of 1
	n_estimators	100-1000 with increments of 100
	gamma	0-9
	reg_alpha	0.1-100 with increments of 1
	colsample_bytree	0-1
	min_child_weight	0-10 with increments of 1

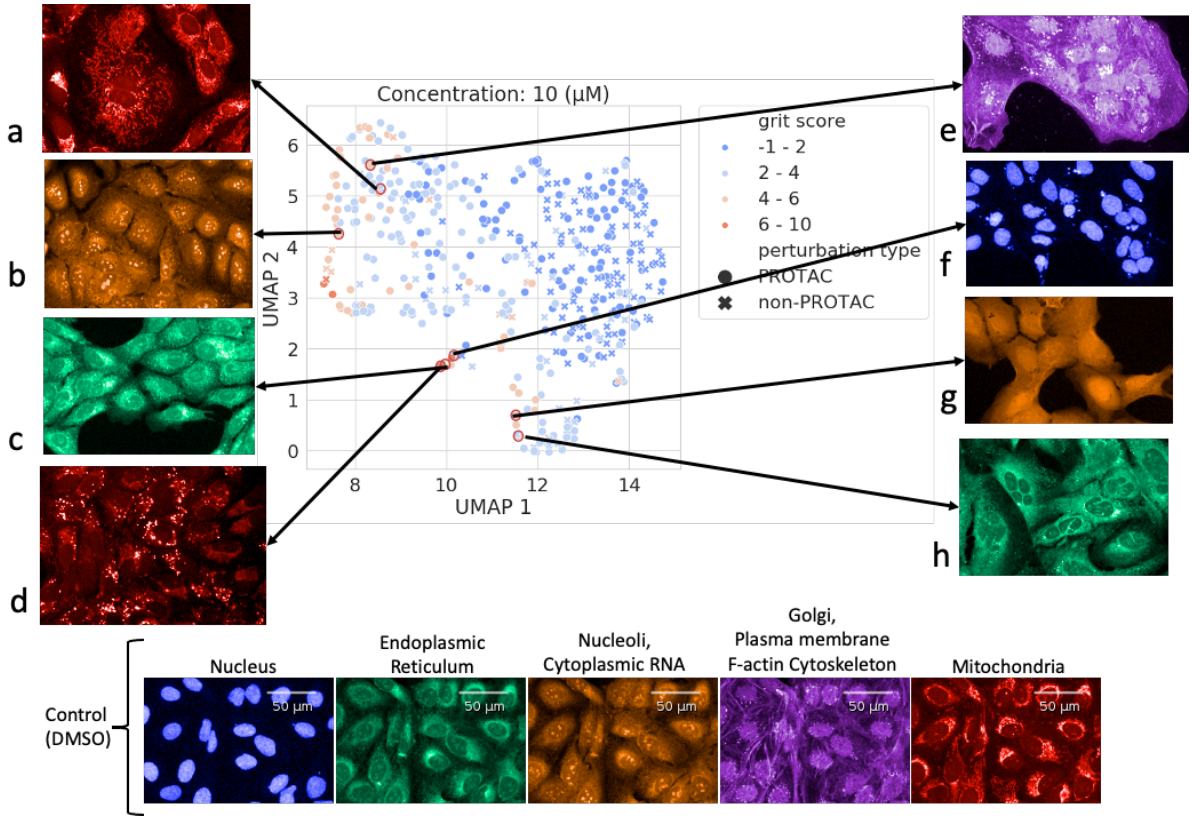
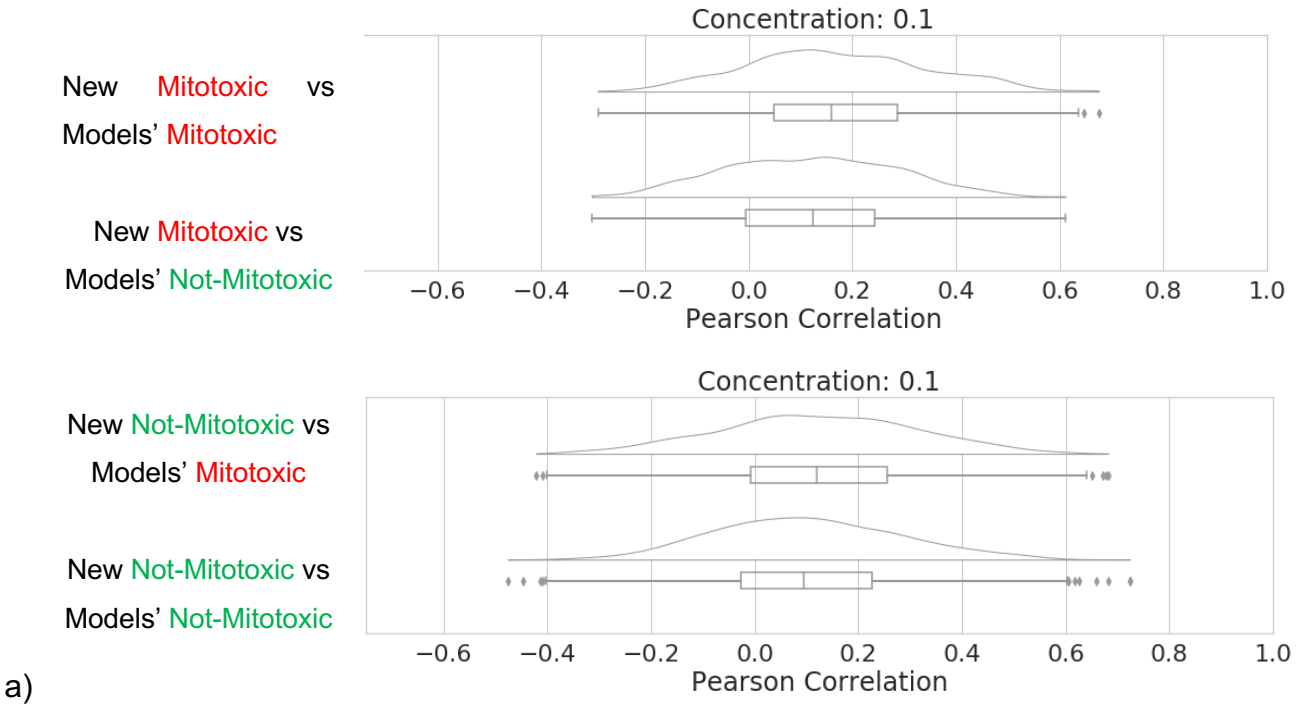
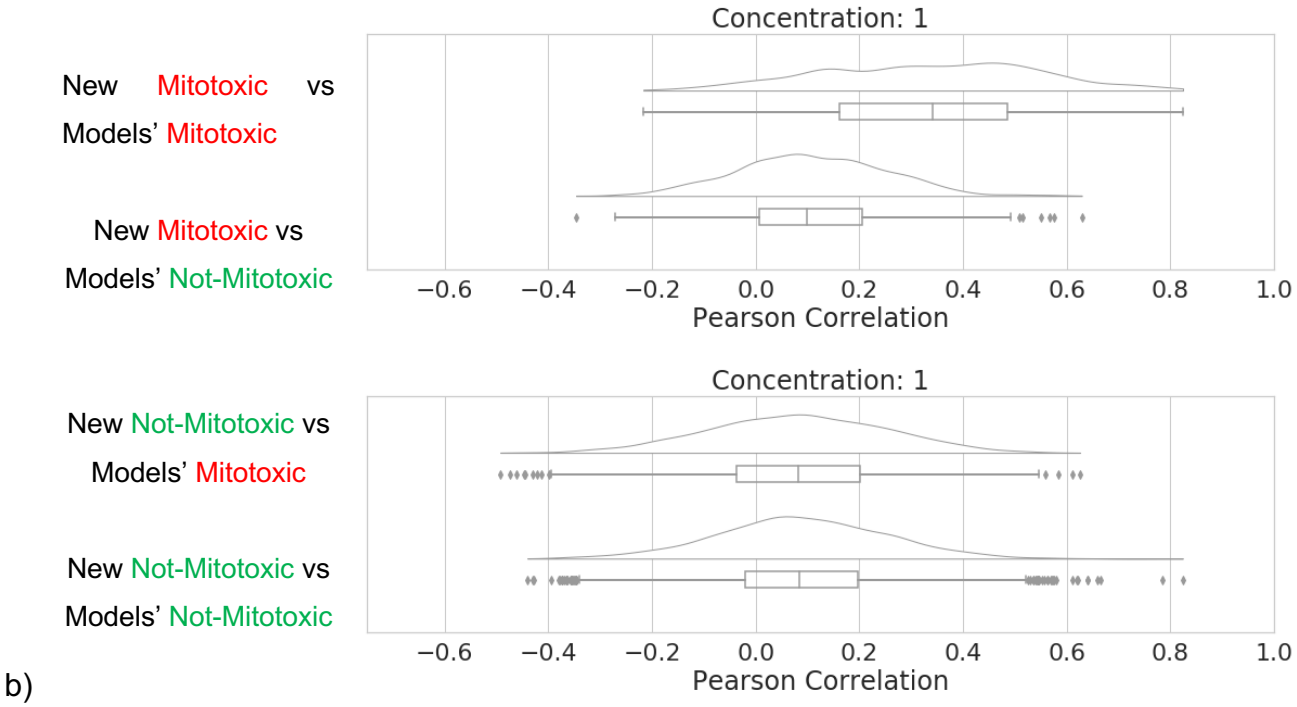


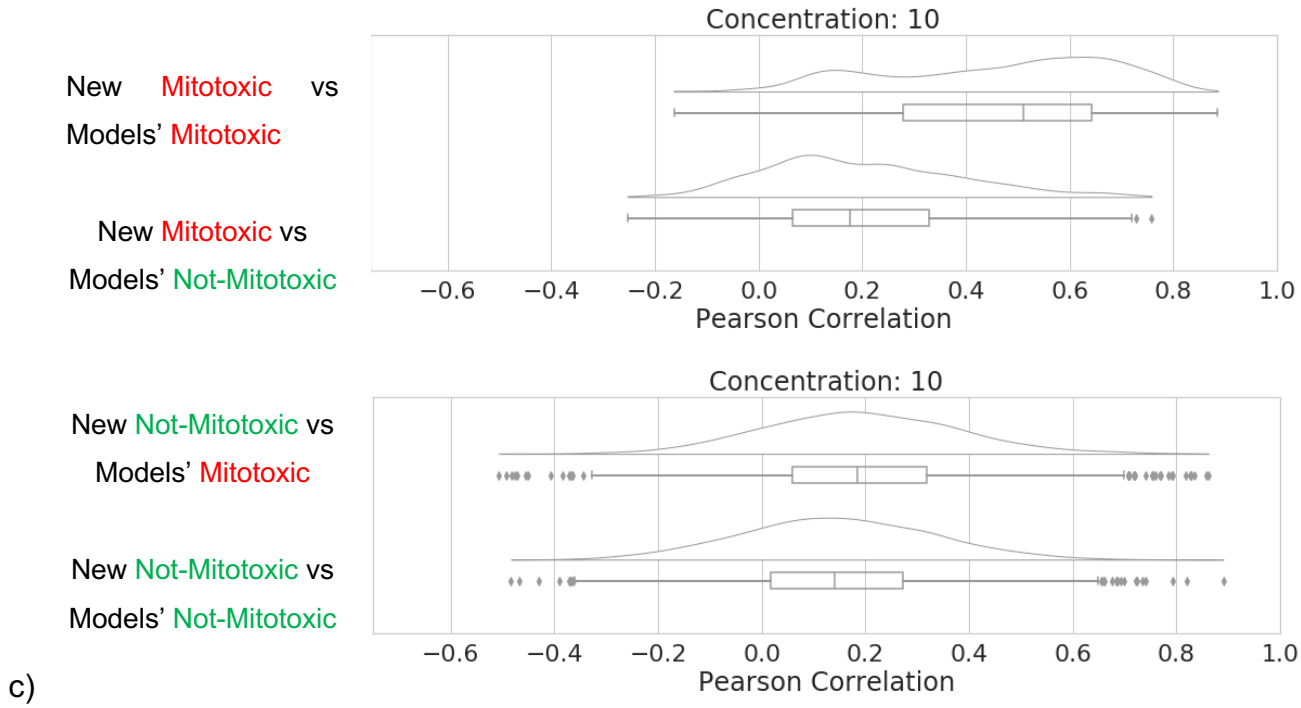
Figure 9.7: Uniform manifold approximation (UMAP) coordinates of all perturbations at concentration 10 µM. Phenotypic examples induced by PROTACs on Cell Painting assay are shown with the raw images. These include: a) networked mitochondria, b) small micronuclei, c) abundant ER staining, d) redistribution of mitochondria, e) abundant Golgi staining, f) presence of micronuclei (sign of genotoxicity), g) no nucleoli and h) clustered nuclei. Examples of neutral control (DMSO) are also included, and the scale bar is set to 50 µm.


New Mitotoxic vs Models' Mitotoxic
New Mitotoxic vs Models' Not-Mitotoxic
Concentration: 0.1
−0.6 −0.4 −0.2 0.0 0.2 0.4 0.6 0.8 1.0
Pearson Correlation
New Not-Mitotoxic vs Models' Mitotoxic
New Not-Mitotoxic vs Models' Not-Mitotoxic
Concentration: 0.1
−0.6 −0.4 −0.2 0.0 0.2 0.4 0.6 0.8 1.0
Pearson Correlation

a)


New Mitotoxic vs Models' Mitotoxic
New Mitotoxic vs Models' Not-Mitotoxic
Concentration: 1
−0.6 −0.4 −0.2 0.0 0.2 0.4 0.6 0.8 1.0
Pearson Correlation
New Not-Mitotoxic vs Models' Mitotoxic
New Not-Mitotoxic vs Models' Not-Mitotoxic
Concentration: 1
−0.6 −0.4 −0.2 0.0 0.2 0.4 0.6 0.8 1.0
Pearson Correlation

b)



c)

Figure 9.8:Pairwise Pearson correlation in the Cell Painting features space between the PROTACs in the external validation set and the compounds (PROTACs and non-PROTACs) in the mitochondrial toxicity models. The four following comparisons are performed. "New Mitotoxic vs Models' Mitotoxic" corresponds to the pairwise Pearson correlation calculation between the mitotoxic PROTACs in the external validation set and the mitotoxic compounds in the model. "New Mitotoxic vs Models' Not-Mitotoxic" corresponds to the pairwise Pearson correlation calculation between the mitotoxic PROTACs in the external validation set and the not-mitotoxic compounds in the model. "New Not-Mitotoxic vs Models' Mitotoxic" corresponds to the pairwise Pearson correlation calculation between the not mitotoxic PROTACs in the external validation set and the mitotoxic compounds in the model. "New Not-Mitotoxic vs Models' Not-Mitotoxic" corresponds to the pairwise Pearson correlation calculation between the not-mitotoxic PROTACs in the external validation set and the not-mitotoxic compounds in the model. These calculations are performed for concentration a)0.1, b)1 and c)10 µM.