Computational analyses of small molecules activity from phenotypic screens

Azedine Zoufir

Hughes Hall

This dissertation is submitted for the degree of Doctor of Philosophy

July 2018

Declaration

This thesis is submitted as the result of my own work and includes nothing which is the outcome of work done in collaboration except where specifically indicated in the text. It is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or 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 or similar institution except as declared in the Preface and specified in the text.

This dissertation does not exceed the word limit of 60,000 words.

Azedine Zoufir

July 2018

Summary

Title: Computational analyses of small molecules activity from phenotypic screens

Author: Azedine Zoufir

Drug discovery is no longer relying on the one gene-one disease paradigm nor on target-based screening alone to discover new drugs. Phenotypic-based screening is regaining momentum to discover new compounds since those assays provide an environment closer to the physiological state of the disease and allow to better anticipate off-target effects and other factors that can limit the efficacy of the drugs. However, uncovering the mechanism of action of the compounds active in those assays relies on *in vitro* techniques that are expensive and time-consuming. *In silico* approaches are therefore beneficial to prioritise mechanism of action hypotheses to be tested in such systems.

In this thesis, the use of machine learning algorithms for *in silico* ligand-target prediction for target deconvolution in phenotypic screening datasets was investigated. A computational workflow is presented in Chapter 2, that allows to improve the coverage of mechanism of action hypotheses obtained by combining two conceptually different target prediction algorithms.

These models rely on the principle that two structurally similar compounds are likely to have the same target. In Chapter 3 of this thesis, it was shown that structural similarity and the similarity in phenotypic activity are correlated, and the fraction of phenotypically similar compounds that can be expected for an increase in structural similarity was subsequently

quantified. Morgan fingerprints were also found to be less sensitive to the dataset employed in these analyses than two other commonly used molecular descriptors.

In Chapter 4, the mechanism of action hypotheses obtained through target prediction was compared to those obtained by extracting experimental bioactivity data of compounds active in phenotypic assays. It was then showed that the mechanism of action hypotheses generated from these two types of approach agreed where a large number of compounds were active in the phenotypic assay. When there were fewer compounds active in the phenotypic assay, target prediction complemented the use of experimental bioactivity data and allowed to uncover alternative mechanisms of action for compounds active in these assays.

Finally, the *in silico* target prediction workflow described in Chapter 2 was applied in Chapter 5 to deconvolute the activity of compounds in a kidney cyst growth reduction assay, aimed at discovering novel therapeutic opportunities for polycystic kidney disease. A metric was developed to rank predicted targets according to the activity of the compounds driving their prediction. Gene expression data and occurrences in the literature were combined with the target predictions to further narrow down the most probable mechanisms of action of cyst growth reducing compounds in the screen. Two target predictions were proposed as a potential mechanism for the reduction of kidney cyst growth, one of which agreed with docking studies.

Acknowledgements

I would like to thank my supervisor Dr Andreas Bender for allowing me to be a part of his lab. I thank him for his continuous guidance and patience through all the revisions of my work. I also thank my collaborators from the University of Leiden, Dr Tijmen Booij and Dr Leo Price for providing the kidney cyst screening data, and Dr Dorien Peters and Tareq Malas for providing their gene expression dataset. I am grateful to Dr Xitong Li and Dr Ellen Berg for providing the BioMAP dataset. I thank the European Research Council for funding my research.

Next, I thank Dr Fredrik Svensson, Dr Krishna Bulusu, Dr Avid Afzal and Dr Deszo Modos, for providing excellent scientific advice and very constructive feedback on my work. I am grateful to the whole of the Bender group for being supportive colleagues and always friendly. Rich and Lewis are thanked for their technical help in using our computing server. Also, my time in this group would not have been the same without Sefer, Siti, Fatima, Nitin, Avid, Leen, Nadia, Fredrik, Deszo and Krishna, who all have been very supportive and helpful with me, and I am very grateful to all of them. I really enjoyed working among such a diverse and talented group of people. I also thank Susan Begg without whom the lab would not be running so smoothly.

Last but not least, I thank my family and particularly my parents for their encouragements throughout my studies. My deepest gratitude goes to my friends Ain, Charles, Ben and Cristian for being there in those times when friends are needed, and for keeping me away of my thesis when I would become too preoccupied about it.

Table of Contents

List of figures

List of tables

Abbreviations

A

ACE Angiotensin-converting enzyme

ADPKD Autosomal Dominant Polycystic Kidney Disease

ADSC Adipose-Derived Stem Cells

ALDH2 Aldehyde Dehydrogenase

ALP Alkaline Phosphatase

ARTS Assay Related Target Similarity

B

BioMAP Biologically Multiplexed Activity Profiling

BMU Best Matching Unit

C

CA14 Carbonic Anhydrase XIV

CAD Cyst Area Deviation

CDK Cyclin-Dependent Kinase

CFTR Cystic fibrosis transmembrane conductance regulator

CGR Cyst growth reducing/ Cyst growth reduction

COX-2 Cyclooxygenase 2

CTD Comparative Toxicogenomics Database

CYP Cytochrome P

CYP2D6 Cytochrome P enzyme isoform 2D6

E

EBP Emopamil binding protein

ECFC Endothelial Colony Forming Cells

ECFP Extended connectivity fingerprint

ELISA Enzyme-linked immunosorbent assay

ELPD Expected log pointwise predictive density

G

GFP Green Fluorescent Protein

GLP-1 Glucagon-like peptide-1

GO Gene Ontology

GPCR G-protein-coupled receptor

H

HCS High-content screening

HDAC Histone Deacetylase

HERG Human Ether-A-Go-Go-Related Gene

HGF Hepatocyte Growth Factor

HIF-1α Hypoxia-inducible factor-1 alpha

HSD11B2 11 beta-hydroxysteroid dehydrogenase type II

HTS High-throughput screening

I

IC Information Content

IDH1 Isocitrate Dehydrogenase 1

K

KLK1 Kallikrein 1

kNN k-Nearest Neighbour

L

LINCS Library of Integrated Network-based Cellular Signatures

M

MAPK Mitogen-Activated Protein Kinase

MC3R Melanocortin-3 Receptor

MC4R Melanocortin-4 Receptor

MGEA5 Meningioma Expressed Antigen 5

MNB Multinomial Naïve Bayes

MoA(s) Mechanism(s) of action

mTOR Mammalian Target Of Rapamycin

N

NB Naïve Bayes

NCATS National Centre for Advancing Translational Sciences

NCOR2 Nuclear Receptor Corepressor 2

nIC Normalised Information Content

NIH National Institutes of Health

NOS3 Nitric oxide synthase 3

NPC National Centre for Advancing Translational Sciences Pharmaceutical Collection

O

OIDD Open Innovation Drug Discovery

P

PA Podophyllin Acetate

PAR-1 Proteinase-Activated Receptor 1

PASS Prediction of Activity Spectra For Substances

PBMC Peripheral Blood Mononuclear Cells

PC-1 Polycystin-1

PC-2 Polycystin-2

PDB Protein Data Bank

PDGFC Platelet-Derived Growth Factor C

PKC Protein Kinase C

PKD Polycystic Kidney Disease

PPA Picropodophyllin Acetate

PPAR-γ Peroxisome-Proliferator-Activated Receptor Gamma

PPB Polypharmacology Browser

PPP Potential Pharmacophoric Points

PXR Pregnane X Receptor

Q

QSAR Quantitative structure-activity relationship

R

RF Random Forest

S

SALI Structure-Activity Landscape Index

SARI Structure-Activity Relationship Index

SEA Similarity Ensemble Approach

SLC5A1 Solute Carrier Family 5 Member 1

SOM Self Organising Maps

SRB Sulforhodamine B

sSOM Supervised Self-Organising Maps

SVM Support Vector Machine

T

Tc(s) Tanimoto coefficient(s)

TCM Traditional Chinese Medicine

TNF-α Tumor Necrosis Factor Alpha

TRAIL TNF-Related Apoptosis-Inducing ligand

V

VEGF Vascular Endothelial Growth Factor

VEGFR Vascular Endothelial Growth Factor Receptor

Chapter 1 Introduction

A recent report by the World Health Organization (WHO) reported that 54% of the mortality observed in 2016 was due to 10 causes, 9 of which being diseases. These included various heart and respiratory diseases, Alzheimer's disease, diabetes and tuberculosis.[1] Furthermore, there are about 7000 orphan diseases for which only about 100 drugs exist, but affect more than 30 million people in Europe and 25 million in North America.[2–4] Therefore, there is no question regarding the societal impact and the strong need for the development of drugs that can either prevent or stop those disorders.

Drug discovery involves interdisciplinary research aimed at discovering novel therapies. Chemistry combined with progress made in enzymology, biochemistry and pharmacology enabled the discovery and validation of protein targets related to diseases.[5] These led to the development of the nowadays called target-based screening assays.

1.1 From target-based to phenotypic-based drug discovery

1.1.1 Target-based screening and limitations

The aim of target-based drug discovery is to either block a protein's signalling activity that is responsible for the disease state or on the contrary to re-establish normal signalling activity in the cell.[6] To discover compounds that can modulate these proteins, target-based screening assays were developed to measure the binding of compounds to a defined protein that has been obtained through recombinant technology and genetics.[7] The pharmaceutical industry has been relying on these assays over the last 30 years to discover new drugs, and the majority of first-in-class drugs have been discovered through target-based assays.[8]

However, the number of drugs reaching the market has progressively decreased over the last two decades. While pre-clinical toxicology and solubility are the main reason of attrition rates at pre-clinical and phase I trials, efficacy remains the main reason for drugs failing in phase II.[9] Indeed, the percentage of drugs that failed in phase II increased from 43% to 66% between 1990 and 2010, and those that failed in phase III increased from 20% to 30% over the same period of time.[10] The lack of efficacy of the drug was the reason for failure in phase II in 51% of the trials between 2008 and 2010, and 59% of the trials between 2011 and 2012.[11–13] In phase III trials, efficacy was the reason for failure in 66% of the trials between 2007 and 2010 and reduced to 52% between 2011 and 2012.[11–13]

This lack of efficacy of drugs was attributed to downsides of target-based approaches.[14] In a study of the reproducibility of published data in the drug target literature, only about 20% of the scientific literature was in line with in-house findings.[15] Another article claimed that the

number of unreproducible findings in the literature is about 50%.[16] Another reason for the lack of efficacy is the difficulties and the lack of resources in identifying the precise molecular binding mode of the drug to the target, a pre-requisite of the usage of confirmatory target-based assays.[17] Therapeutic opportunities identified in target-based approaches may translate poorly to the desired effect *in vivo*, as they fail to capture complex disease biology *in vivo* and/or cannot account for polypharmacology (the desired effect may be exerted through binding of several targets).[7]

1.1.2 Phenotypic-based screening compensates for the limitations of target-based screening

In those cases where target-based approaches are not sufficient on their own, phenotypic-based screening assays have been developed that overcome the challenges of target-based assays. In these assays, rather than assessing whether a compound can modulate a specific protein target, a certain feature of a disease is exploited *e.g.* selective eliminating of a specific cell population or modulation of a specific pathway within a cell are instead measured.[18] Measuring such readouts paints a more comprehensive picture of a compound's effect on a native cellular environment or tissue since they use living cells in which a compound will modulate the enzymatic or signalling activity of several targets, change signalling cascades and affect various cellular processes.[7,18] Under the current polypharmacology paradigm, drugs are more likely to exhibit the desired effect by affecting several targets at a time rather than modulating single targets, and the modulation of multiple targets is more likely to be assessed in phenotypic screens than in target-based assays.[19]

Phenotypic screens have had varying degrees of success and usage over the years. Between 1999 and 2008, 37% of the first-in-class drugs were discovered using phenotypic-screening while target-based approaches led to the discovery of 27% of the first-in-class drugs over the same period.[20] When increasing the timeframe of the analysis and including biological therapeutics (*e.g.* antibodies) in addition to small molecules, only 7% of the first-in-class drugs were discovered through phenotypic screens between 1999 and 2013, while 41% of the first-in-class drugs were discovered through target-based assays.[8] Rather than a way to discover new chemical entities that can serve as drugs, phenotypic screening is seen as an approach that can complement target-based approaches. Indeed, phenotypic screens can be used to discover new indications for an already marketed drug and combine them with target-based approaches to identify its mechanism of action (MoA).[21] These assays are described in the next section.

1.1.3 Assays used in phenotypic-based screening

Mainly three types of assays can be found in phenotypic screening: cell viability assays, cell signalling pathway assays and disease-related assays.[21] There are three types of cell viability and proliferation assays in phenotypic compound activity databases. One of the most common types of assays in this category is colourimetric assays measuring tetrazolium dyes[22–24] in which living cells are detected by how they metabolise these substances using mitochondrial enzymes. Another frequently encountered cell viability assay is using Sulforhodamine B (SRB), an aminoxanthene dye which binds intracellular proteins, indicating the protein content present in a cell culture, which in turn is correlated to cell mass.[25] One of the most popular cell viability assay, the Alamar Blue assay, employs a resazurin dye changing colour with the oxidation/reduction potential observed in the cell media as cells proliferate.[26]

Cell signalling pathway assays detect the modulation of a pathway by a chemical. For example, certain assays label a protein effector and member of the target pathway with Green Fluorescent Protein (GFP) and measuring the nuclear translocation of this protein to the nucleus upon a compound's binding to the target receptor of this pathway.[21] Another example includes assays where compounds are screened for their effects modulating the Wnt pathway by measuring alkaline phosphatase activity which is expressed downstream of this pathway.[27,28]

The last type of phenotypic assay measures disease-specific endpoints. Typical assays from this category include those found in the Biologically Multiplexed Activity Profiling (BioMAP) systems which use human primary cells to mimic a broad range of physiological responses to compounds such as inflammation, angiogenesis or microtubule function among others by measuring biomarkers (protein readouts) from 8 different cell systems[29,30]. High-content screening (HCS) is also very popular in the phenotypic screening literature, where cells are grown to mimic their 3D arrangement, similarly to their arrangement in biological tissues. This would, therefore, maintain their physiological properties as opposed to what would be observed within Petri dish cultures,[31] and would allow developing assays which are more relevant to the *in vivo* disease state.

1.1.4 *In vitro* deconvolution in phenotypic screens and limitations

Once a compound has been successfully screened for its modulation of the phenotype of interest, several methods are available to deconvolute the activity of the compound in the phenotypic assay (**Table 1**), *i.e.* to identify the MoA of the compound responsible for the phenotypic readout in the assay.[32,33] Affinity chromatography is one of the oldest

deconvolution methods and probes an ensemble of proteins using beads followed by identification of the bound targets with Western blots or mass spectrometry.[34] However, this method is time-consuming and expensive, and therefore impractical if a relatively large number of compound signals need to be deconvoluted (**Table 1**).

Other techniques utilise gene expression for target identification. The three-hybrid system involves the isolation of cells associated with the expression of a reporter gene activated by the interaction of the compound with the target expressed in those cells (**Table 1**).[35] A similar concept uses the amplification of phage colonies specifically displaying the protein targets that interact with the compound.[36] The challenges with this type of assays are that post-translational modifications are often lost and that it is not possible to screen for protein complexes (**Table 1**).[32]

Protein microarrays allow probing for binding to nearly all targets in the proteome of a species.[37] Each protein is immobilised in each well of a plate, and therefore each protein has a known position on the plate. The compound can then be engineered to react to a fluorescent or radioactive conjugate enabling the detection of the compound-protein complexes upon binding. Even though this approach seems more efficient than the other approaches, the issue of detecting compound-target interactions as observed *in vivo* remains problematic, since post-translational modifications, location and complexes may not be amenable to this type of assay (**Table 1**).[32] All these techniques are time-consuming and costly. It is, therefore, necessary to shortlist and prioritize compounds that are to be used in these *in vitro* deconvolution experiments. *In silico* deconvolution can help in making such decisions. These approaches rely on the molecular and biological similarity concepts which are introduced in the following section.

1.2 Molecular and biological similarity

This section focuses on the processing of chemical information and how it relates to the bioactivity data observed in target-based and phenotypic assays. This starts by finding an appropriate representation for the compounds in the dataset. This representation is then used to calculate the structural similarity of a query compound to other compounds of known properties.

1.2.1 Representation of chemicals

The starting point of computational approaches aimed at understanding the behaviour of compounds in *in vitro* assays is to find an effective representation of the chemicals. In its most simple form, a molecule may be represented as a graph, that is to say, a set of atoms (nodes) linked by bonds (edges) with labels (simple, double, etc.) from which a more complex encoding of the compound can be derived.[38]

Method	Description	Advantages	Limitations	Reference
Affinity chromatography	Attaching beads to compounds; purify bound protein targets using microbeads; characterise protein	Applicable to any small-molecule; targets maintain conformation and post-translational modification	Long; impractical for high-throughput screening; require high binding affinities	27
Three hybrid system	Activation expression of reporter gene when compound binds target; Identification of cell expressing the reporter gene; purification and characterization of target	Association of compound to target occurs in living cell and hence information about subcellular localisation and stability can be obtained	False positives due to activation of reporter other than binding of the compound to the target	28
Phage display	Selective amplification of phage displaying protein-target which binds the compound of target	Identification of proteins with low abundance possible through amplification	Post-translational modifications and subcellular locations lost and may affect binding	29
Protein microarray	One protein per well; bound protein fluoresces; fluorescence detected by imaging	Detection against nearly all proteome of an organism; all proteins are exposed equally and allows identification of low abundance proteins	Post-translational modifications and subcellular locations lost and may affect binding	30

Table 1. In vitro deconvolution methods discussed in this chapter.

One possible encoding may be numeric through the molecular properties of the compound such as solubility, partition coefficient, melting point, molecular weight and electronic properties.[38–40] Other types of molecular encodings are based on torsions and angles between the atoms of a molecule, which is found to perform consistently well, according to several performance metrics, for virtual screening approaches in which query compounds are compared to reference compounds with desirable properties or activity.[41] Some representations are based on pairs or triplets of atoms and represent yet another possibility to represent compounds but are not often used in cheminformatics applications.[38–40]

Alternative and more frequently encountered chemical representations employ a 2D molecular encoding called a fingerprint. Fingerprints are binary vectors indicating either the presence or absence of a certain structural motif in a chemical or count vectors of the molecule's substructures [38–40]. Two types of 2D fingerprints are found in the literature: molecule-based fingerprints where each bit is computed from the structure of the compound through hashing algorithms, and the dictionary-based type where each element of the binary vector represents a pre-defined chemical substructure.[42]

In the first category, the Morgan fingerprints or Extended-Connectivity Fingerprints (ECFP) can be found,[43] which are by far the most used 2D fingerprint in the cheminformatics literature. Each bit in the vector represents hashed identifiers extracted from individual atom properties of the compound in the first iteration of the algorithm. The bit vector is then supplemented by combinations of the previous identifiers, which include up to 2 neighbouring (non-hydrogen) atoms (ECFP4) or up to 3 neighbours (ECFP6). The ECFPs outperformed most 2D fingerprints for virtual screening tasks[41], which explains their popularity. The Daylight fingerprint is another type of fingerprint which is computed from linear substructures of the query compound. Each substructure is computed with a maximum pre-defined length, and identifiers

describing atom and bond properties are hashed to produce a binary fingerprint.[44] MOLPRINT2D is another type of molecule-based fingerprint in which strings are employed to describe atomic environments for each atom,[45] instead of identifiers as in the ECFP algorithm. The rest of the MOLPRINT2D fingerprint generation is similar to that of the ECFP algorithm.

In the second category of 2D fingerprints, the dictionary-based fingerprints, the MACCS MDL keys are found, which encode an ensemble of 166 keys corresponding to atom and bond types,[46] while PubChem fingerprints encode 881 and have additional atom counts, atomic neighbourhoods and substructures.[47] The Unity fingerprint is a 988-bit vector which is a hybrid between dictionary-based and molecule-based representation since pre-defined generic atom and bond types are encoded, but paths of specified lengths are generated in a similar fashion to the Daylight algorithm.[48]

3D descriptors such as pharmacophores are derived from steric and electronic properties of a chemical, which are important for describing the interactions with the binding pocket of a specific target.[38–40] Finally, another encoding uses projections of 3D structures into 2D circular planes, which are then used to derive potential pharmacophoric points (PPP), generating a 1D descriptor for which it becomes possible to apply bioinformatics algorithms designed to work with sequences of letters as input such as base-pair alignment.[49]

1.2.2 Molecular similarity principle in virtual screening and neighbourhood property

As mentioned previously, one of the main application of the fingerprints is virtual screening.[50,51] In this approach, compounds which have desired properties or activity in a

target-based or in a phenotypic-based screening assay are employed as a reference. After converting compounds to a suitable representation, query compounds are "screened" against the reference by measuring the similarity of the representation of the query compounds to that of the reference. This leads to filtering of the query compounds to those with the desired activity. This important property is based on the **molecular similarity principle** which states that structurally similar compounds should have similar properties and yield similar readouts in target-based and phenotypic-based assays.[52,53] Even though virtual screening is discussed in this introduction, it is also acknowledged that the molecular similarity principle is useful for combinatorial chemists when designing new libraries of compounds based on existing libraries.[54]

The Tanimoto coefficient (Tc) is a measure that assesses the number of bits in common between two binary vectors and is therefore usually well-suited to measure the structural similarity between two compound's fingerprints.[55] Otherwise, the Pearson correlation coefficient is used instead to measure the similarity of two non-binary fingerprints or bioactivity profiles.

Several studies investigated the molecular similarity principle by comparing the chemical and biological similarity of pairs of compounds in a dataset. A study of the correlation between chemical similarity and activity in a monoamine oxidase inhibition assay in which Daylight fingerprints were employed. The authors found that a structural Tc similarity of 0.85 for two compounds corresponds to a 30% probability that two compounds are active in the monoamine oxidase inhibition assay used in the study.[56] This observation was later repeated with 23 assays measuring the inhibition of various protein targets, mostly kinases and aminergic receptors.[57] With the use of both ECFP6 fingerprints and MACCS keys, it was shown that two chemical similarity definitions of ECFP6 Tc similarity >0.4 and MACCS key Tc similarity > 0.8 correctly identified the target of one compound based on structural similarity to another

compound with 90% and 87% accuracy, respectively, although this was attributed to the composition of the WOMBAT dataset (congeneric chemical series) rather than the performance of the descriptors themselves.[58]

It is evident from the above studies that different descriptors capture the molecular similarity principle differently. Aiming to quantitively compare these differences, Patterson et al. found that descriptors commonly used in virtual screening applications, such as 2D fingerprints, exhibit a "neighbourhood property", meaning that they can be used to search for compounds that fall within activity regions of interest in those descriptor spaces (**Figure 1**).[59] Fingerprints for which two compounds are highly similar should yield similar biological activity and have such neighbourhood property (**Figure 1**). In contrast, descriptors which yield structurally similar compound pairs with large differences in biological activity do not display such property and are undesirable for cheminformatics applications (**Figure 1**).

However, the relationship between molecular similarity and activity is often more complex in reality, since small changes in chemical structure can lead to important modification of the activity against targets between two highly structurally similar analogues,[53,60,61] and the "neighbourhood property" concept also allows for the evaluation of descriptors in terms of their sensitivity to such activity cliffs.[59] Other studies attempted to capture and rationalize activity-cliffs in target-based bioactivity datasets by developing metrics such as the Structure-Activity Relationship Index (SARI)[62] and the Structure-Activity Landscape Index (SALI)[63] which have been shown to retrieve known activity-cliffs for classes of targets where these are predominant.

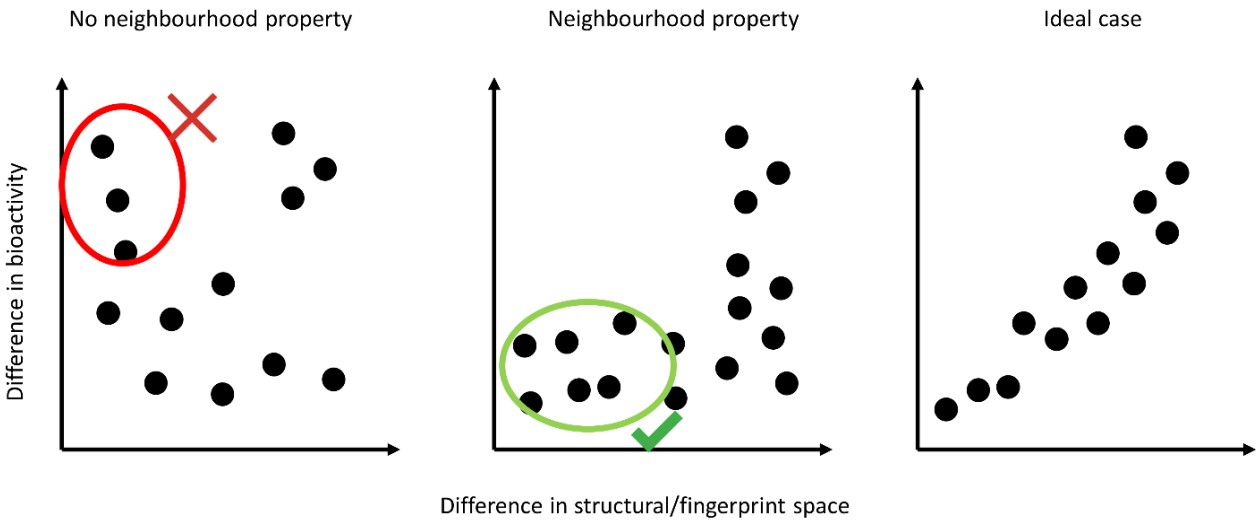

Figure 1. Illustration of the neighbourhood property principle. A descriptor for which compound pairs display no or very little differences in bioactivity when compounds are structurally similar *i.e.* compound pairs with low structural similarity. On the contrary, if compounds show high differences in bioactivity despite their structural similarity, then the descriptor doesn't have the neighbourhood property. In the ideal scenario, differences in structural similarity and bioactivity are correlated.

1.3 *In silico* deconvolution methods of compound activity in phenotypic screens

1.3.1 Data-driven deconvolution

The development of various omics techniques and the public availability of historical bioactivity datasets gave rise to *in silico* deconvolution techniques exploiting these data.[33] Gene expression databases such as the Library of Integrated Network-based Cellular Signatures (LINCS) can be used to query compounds active in phenotypic screens for their gene expression profiles, already narrowing down the number of candidate protein targets to explore using *in vitro* deconvolution methods.[64] However, due to the higher cost and lower throughput of gene expression technologies, bioactivity databases usually contain more data.

Hence, the wealth of bioactivity data can be exploited in cheminformatics analyses aimed at deconvoluting the signal of phenotypic screening campaigns. For example, bioactivity data from ChEMBL was extracted and corresponding targets were grouped by pathway annotations from the Gene Ontology (GO) framework. This grouping was used to deconvolute the activity of compounds in a screen measuring the inhibition of tumour necrosis factor alpha (TNF-α) production in leukemic cells.[65] The authors found that the targets they identified for their compounds from ChEMBL were consistent with the literature on TNF-α production. In another study, enrichment analyses using Sanofi historical high-throughput screening allowed to explain the MoA of compounds active in DNA fragmentation and TNF-related apoptosis-inducing ligand (TRAIL) assays: a large number of targets were identified for the TRAIL assay

including a range of cyclin-dependent kinases (CDKs), vascular endothelial growth factor receptors (VEGFRs) and other kinases; whereas fewer but very relevant targets were identified for the DNA fragmentation assay, including GSK3b, Tubulin, Aurora 2 and Eg5.[66]

In addition to relying on the structural similarity of the compounds, the correlation of biological spectra can be used for the deconvolution. For example, the activity of a pyrimidine compound in an assay aimed at measuring the induction of senescence was attributed to targeting tubulin.[67] This was performed by computing the correlation of the biological profile of this compound to the profiles of other compounds in a proprietary bioactivity database as well as employing structural similarity.

It is also possible to facilitate the deconvolution step by pre-selecting compounds with known bioactivity prior to their use in phenotypic screens. Indeed, a recent study employed a biologically annotated library of compounds, which was collected from publicly available databases and used in a phenotypic screening of compound combinations measuring the inhibition of serine palmitoyltransferase, a model for the necrosis of lung cancer. Since the compounds were already annotated from the public databases, this facilitated the identification of the putative MoA of cyclooxygenase 2 (COX-2) modulation behind the activities observed in the combination screen.[68]

1.3.2 Deconvolution methods based on *in silico* ligand-target predictions

While current *in silico* deconvolution techniques employ pre-existing bioactivity data, other methods rely on the molecular similarity principle in bioactivity datasets to predict putative targets from the fingerprint of compounds (**Figure 2**). These methods are the emphasis of this section.

1.3.2.1 Bioactivity datasets and limitations relevant to target prediction

Target prediction methods were developed thanks to the increasing availability of large-scale chemical information publicly available.[69] While many bioactivity databases exist, only the datasets which are mentioned or used in this thesis are discussed here (**Table 2**) and readers are referred to Gaulton et al.[69] for a more comprehensive overview of such datasets.

Database	Number of compounds	Number of Targets	References
Pubchem	2,570,179	10,857	[70,71]
ChEMBL	1,735,442	11,538	[72,73]
Drugmatrix	1,291	132	[75,76]
WOMBAT	136,091	1,320	[77,78]

Table 2. Bioactivity databases discussed in this chapter and statistics (as of 5/4/18)

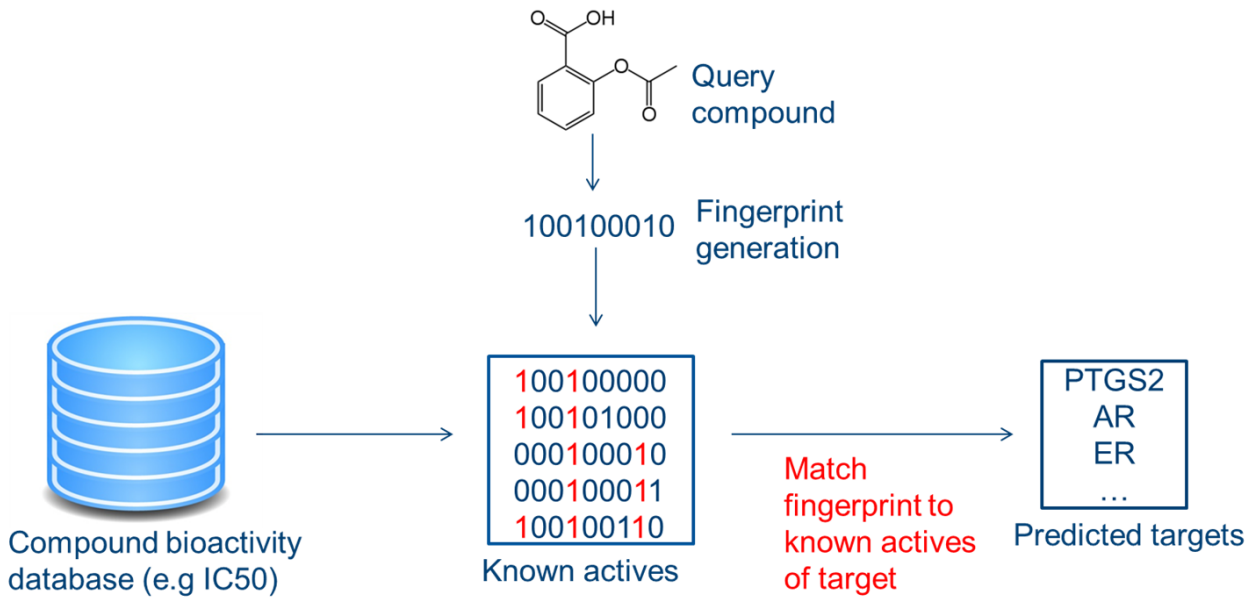

Figure 2. Target prediction approach. A compound bioactivity database is employed to extract known active ligands for a large set of targets. These are converted into fingerprints which machine learning algorithms identify and "learn". When the fingerprint of a query compound is presented to the algorithm, it is matched to those of known actives. If the matching is successful, then the target is predicted for the compound. Repeating this process against all the known active ligand datasets of other targets will generate a predicted target profile for the query compound from which novel MoA hypotheses can be drawn.

The Molecular Libraries and Imaging program launched by the National Institutes of Health (NIH) gave rise to Pubchem[47,70,71,74], one of the largest repository of high-throughput screening (HTS) data of both target-based screens and phenotypic screens. It comprises three related sub-databases: Substance, Compound and BioAssay. The latter contains bioactivity data (mostly IC50, EC50, Ki, Kd) for small molecules and RNAi, along with a suite of tools to query,

analyse and summarise the data. [70,71] In the latest versions, PubChem BioAssay contained more than 700,000 bioassays for more than 2,000,000 tested small molecules contributed by 50 organisations (**Table 2**).[71]

PubChem is the biggest repository and shares data with many other databases including ChEMBL, another comprehensive database developed by the EMBL-EBI containing bioactivity data extracted from more than 50,000 publications, and calculated molecular properties of chemicals, which yield data points for a total number of more than 1.3 million bioactive compounds in relation to more than 2,800 human targets (**Table 2**).[72,73] Recent developments included information about the development stage of a chemical, new target annotation (*e.g.*, binding site information), and the possibility to filter results based on the quality of the data via several additional fields identifying duplicates or data validity.[73] While ChEMBL overlaps with PubChem, ChEMBL active compounds to inactive compounds ratio is very high, meaning that ChEMBL contains mainly potent compounds.

ChEMBL is also large enough to overlap with other databases. In particular, the Drugmatrix is a comprehensive pharmacogenomics database. It comprises gene expression profiles of chemicals, on-target binding affinities and ADME assays results, along with pathology data such as haematology, histopathology and clinical readouts in relation to these chemicals. [75,76] It has recently been in integrated with ChEMBL. It contains a complete bioactivity matrix for 1,291 compounds and 130 *in vitro* assays, even though the inactivity data points are not quantified in the matrix (**Table 2**).

PubChem also overlaps with WOMBAT, another bioactivity database which resulted from a collaboration of Astra Zeneca with the Romanian Academy Institute in Timisoara and contains 307,700 activity points on 1320 targets, and covers 136,091 unique compounds (**Table 2**).[77,78]

However, these datasets are not without limitations. The sparse nature of this data is the main limitation and HTS is indeed still far from giving the full picture of drug-target interactions. In addition, certain compounds do not exhibit any activities in HTS (also called "dark matter compounds"), despite them being active in other types of screening such as gene expression experiments and antifungal assays.[79]

While integrating data from different vendors helped in generating more complete datasets, inconsistencies were found between the different datasets.[80] These inconsistencies were later attributed to errors in manual extraction and curation of bioactivity values from the literature.[81] It is therefore important to recognise those limitations when using ligand-target prediction models and to keep in mind that these could potentially affect the validity of the MoA uncovered with these models. It is noted however that the likelihood of invalid results is proportional to the errors rate encountered in the databases, which in databases such as ChEMBL or PubChem, is constantly reduced thanks to the ongoing curation efforts.[71,73]

1.3.2.2 Current target prediction methods

Many algorithms are currently employed for ligand-target predictions. Predicting targets from fingerprints using machine learning is one of the most used approaches in the literature (**Table 3**). Nidhi et al. developed a multi-category Naïve Bayes (NB) model, a model which combines the probabilities of all the targets in WOMBAT.[82] The model retrieved the correct first target for 82% of the compounds and the correct first target class for 89% of the compounds in the test set. They employed this model for chemicals which were only associated with a therapeutic class, in order to generate a putative novel MoA for these chemicals. This model has been extended and successfully used in other MoA studies of potential therapeutics for tuberculosis[83]

or to improve hit list triaging in a Luciferase gene reporter assay.[84] Target prediction models similar to the one developed by Nidhi et al. are so far the most frequently used and have been integrated with other types of data such as high-content screening,[85] gene expression profiles,[86] or proteochemometric features[87] in order to draw more interpretable MoA hypotheses.

Even though Naïve Bayes models such as the one by Nidhi et al. are popular for target prediction, other machine learning approaches also exist (**Table 3**). Support Vector Machine (SVM) models were developed on the same database (WOMBAT) to generate more accurate target prediction (mean balanced accuracy of 0.912 ± 0.093), and which were used to profile drugs and hypothesise the MoA leading to liver-related adverse events.[88] Self-organising maps (SOMs) were also successfully employed to predict the selectivity of glutamate receptor antagonists,[89] or to predict human targets for *de novo* synthesised compounds.[90] A more recent approach called DeepDTI employed deep neural networks to predict 10 novel drug-target interactions from drug target annotations in Drugbank which were in agreement with the literature, and the model outperformed some of the most frequently used machine learning algorithms used in target prediction, such as NB and Random Forest (RF).[91]

However, machine learning is not the only mean for target prediction and approaches based on similarity are also frequently employed (**Table 3**). The earliest attempt to employ similarity-based target prediction approach, called "prediction of activity spectra for substances" or PASS, employs a scoring function that relates the number of compounds with a certain chemical descriptor to the number of compounds active against a target.[92] The popular Similarity Ensemble Approach (SEA) approach employs the Tc to compute the chemical similarity to known ligands of targets to generate predictions and evaluates the statistical significance of the scores by employing an expected value similar to the e-value employed in NCBI's BLAST for gene alignment, which measures the likelihood of a result being random.[93]

Nickel et al. developed a web-server for target prediction called SuperPred, which is based on normalised similarity to target known actives, which takes into account differences in ligand number for the targets, and which achieved 94% accuracy (when predictions are filtered by quality), but performed poorly for targets bound by structurally diverse compounds.[94] The authors behind the Polypharmacology browser (PPB), another similarity-based target prediction tool, recognised that the choice of the fingerprint impacts the predictions, and therefore decided to use a consensus prediction based on 10 descriptors (6 fingerprints and 4 combinations thereof).[95] The choice of the Manhattan distance (called "city block" in the manuscript) used in this algorithm was based on computational speed but is questionable, as it was established that Tc similarity and similar methods outperform Manhattan distances among others for molecular similarity application.[55]

Other similarity studies using similarity were based on other descriptors than 2D fingerprints (**Table 3**). Nigsch et al. investigated the use of gene expression profiles for target prediction based on profile correlation.[96] They found that a minimal number of 128 genes achieved the highest accuracy of 0.3. The iRaise target prediction algorithm is based on structural information where triangle pharmacophores are computed from the Protein Data Bank (PDB) structures and which describe hydrogen bond acceptors, donors or hydrophobic interactions.[97] A similar approach is employed by the PharmMapper web server which employs pharmacophores that take into account additional properties such as positive and negative charges.[98] The LT-scanner algorithm takes as input a ligand-protein complex and uses a scoring function to identify whether similar interactions can be found in other proteins across the genome.[99] However, such structural information is not always available and can limit the number of targets this type of approach can find. Moreover, availability of gene expression profiles can be an issue for certain compounds, and protein structure database may not entirely cover certain protein families such as transmembrane proteins. Structure-based target

prediction is also more time-consuming and resource-intensive compared to approaches employing 2D fingerprints.

A more recent approach combined the similarity methods with machine learning algorithms, and used the actual potency of compounds in the training set to weight the Tc similarity to known ligands of targets when making predictions (**Table 3**).[100] Downsides of this approach is the reliance on older versions of ChEMBL (versions 19 and 20) for the training set, as well as the imbalanced training sets which were biased towards potent compounds, and contained fewer inactive compounds as a result.

Network-based analyses are also employed to predict putative ligand-target pairs (**Table 3**). Yaminishi et al. extracted drug-target interaction network from the KEGG database and employed those in an attempt to predict novel drug-target interactions based on the proximity of both novel compounds and novel targets to already known drug-target interactions.[101] The proximity of compounds was computed as the size of the common substructure set, while the proximity in target space was computed using amino acid sequence similarity. They then integrated these similarities using bi-partite graphs to predict new compound-target pairs, which is a form of network representation where edges can only be found between two sets of vertices but not within the sets themselves. He et al. have also employed KEGG networks and used instead functional group composition and amino acid composition as feature vectors for compounds and targets, respectively.[102] They combined a k-Nearest Neighbour (kNN) algorithm with two feature selection methods in order to achieve accuracies around 80% depending on the target family. A more recent approach called nAnnoLyze was developed based on the construction of several sub-networks linking query compounds to ligands co-crystallised with protein targets, themselves linked to human protein targets 3D models through both sequence and structural similarity.[103] This last step enhanced the number of predictable

targets. Additionally, compounds in a benchmark dataset compiled from Drugbank were associated with these 3D target models through Djikstra-like shortest path analyses from which scores were computed and combined in random forest classifiers. The prediction on the benchmark dataset yielded a precision of 73% and a recall of 66%. Finally, signalling and metabolic networks were also utilized for target prediction[104]. However, these two network approaches are at an early stage and lack evaluation on benchmarking datasets.

Type	Input	Model name	Description	References
Machine learning	ECFP6 fingerprints		Multiple-category Laplacian-modified naïve Bayesian model	82–85
	Tc similarity to reference compounds		Support Vector Machine	88
	Pharmacophores		Self-organising maps	89
	Pharmacophores and physicochemical properties	SPIDER	Self-organising maps	90
	ECFP2, ECFP4, ECFP6 for compounds and amino acid, dipeptide and tripeptide composition	DeepDTI	Deep neural networks	91
Similarity & Scoring	Multilevel neighbourhood of atoms (second level)	Prediction of activity spectra for substances (PASS)	Scoring function taking into account the number of compounds active against the target and the number of compounds	92
	Daylight fingerprints	Similarity Ensemble Approach (SEA)	Tc similarity and conversion to Z-Score	93
	FP24, MDL MACCS keys, ECFP4 fingerprints	SuperPred	Tc similarity and conversion to Z-Score	94
	Apfp, Xfp, MQN, SMIfp, Sfp, ECFP4 fingerprints and combinations	Polypharmacology browser (PPB)	Manhattan similarity	95
	Gene expression profiles		Nearest Neighbours based on Pearson product-moment correlation coefficient	96
	Pharmacophores	iRaise	Cascade of binding mode scoring based on spatial alignment to reference ligand and pocket coverage	97
	Pharmacophores	PharmMapper	Pharmacophoric fit score and conversion to Z-Score	98
	Ligand-protein complex	LT-scanner	Scoring function that identifies proteins with similar binding site interactions than the one observed in ligand-target complex	99

Machine learning & Similarity	ECFP4, FP2 fingerprints	MOST	Tc similarity to reference compound followed by p-value computation by machine learning	[100]
Network-based prediction	Functional groups for compounds and biochemical/physicochemical properties for targets		Nearest Neighbours based on normalised dot products of drug and target with feature vectors	[102]
	Maximum common substructure scores for compound similarities; Smith-Waterman score for target sequence similarity; Known drug-target interactions		Bipartite graph learning and connecting new compound-target pairs via pre-existing compound-target interactions	[101]
	Structural and physicochemical similarity for ligand similarities; Structural alignments for protein similarities; Existing binding interaction in PDB for ligand-target edges	nAnnoLyze	Bipartite graph construction and Djikstra shortest path weighted sum of edges converted to Z-Score and random forest classifier	[103]

Table 3. Target prediction algorithms discussed in this chapter.

Since the computation of fingerprints is possible for most compounds, they provide a more readily available feature space for compounds than gene expression profiles or 3D structures which are not always available. Machine learning models may be preferred over similarity-based methods for target prediction since they achieve better accuracies on average, which may be due to the former better accounting for non-linearities in the bioactivity training data than the latter approach. Moreover, even though some of these similarity approaches were successful and some of their prediction validated,[93] their predictive power is expected to be limited when it comes to novel pharmacological actions since this type of approach is based on achieving high similarity to already known ligands. Machine learning algorithms are more flexible, hence less affected by such biases. Indeed, applicability domain analyses generally circumvent those issues[105] or a scoring mechanism that allows obtaining the confidence of such predictions through *e.g.* conformal predictions.[106] Finally, target predictions based on machine learning models have been evaluated more thoroughly over the past decade than the more recent approaches based on networks, are more intuitive and are simpler to implement through the use of open-source and maintained programming libraries.

1.3.2.3 Applications to deconvolution of compounds active in phenotypic screens

Several publications illustrate the use of ligand-target prediction for the deconvolution of the activity of compounds in phenotypic screens. Such target prediction models were employed to rationalise the effects of compounds in high-content cell screens aimed at identifying compounds that modulate the cell cycle.[107] They found that while most clusters putatively targeted tubulin, the group of kinase inhibitors was associated with CDK1 and CDK2 predictions, which indeed play a role in cell cycle regulation. Similar approaches were used

to understand the difference of MoA between cytostatic and cytotoxic compounds, where cytostatic compounds were linked to DNA damage reversal, metabolism and processes regulating the cytoskeleton.[108] The same type of approach was used in rationalising the MoA for compounds present in plant extracts used in Traditional Chinese Medicine (TCM), where the phenotypic effect of the active ingredients is usually known, but not the molecular targets.[109] Compounds with the cold nature affected targets involved in detoxification and sedation processes, while compounds with hot nature were linked to targets that affect fertility and cardioprotection. These target prediction methods can be combined with decision trees, such as in Liggi et al., to identify the MoA of compounds inducing different phenotypes in *Xenopus Laevis*.[110] Compounds affecting pigmentation were predicted to bind Carbonic anhydrase II which is in accordance with the literature on pigmentation biology. They also found that compounds affecting melanophore function targeted alpha-2a adrenergic receptor, delta-type opioid receptor metabotropic glutamate receptor 1 and tyrosine-protein kinase Fyn. A similar approach was used to rationalise the MoA of compounds that have sedative effects and were subsequently used to predict sedative and/or hypnotic function in Drugbank compounds.[111] The predicted target profile implicated in the sedative-hypnotic effect was comprised of a variety of aminergic G-protein-coupled receptors (GPCRs) such as Dopamine receptors D1B, D2, D4, muscarinic receptors M1 and M4, histamine receptors H1 and adrenergic receptor alpha among others.

Methods involving similarity networks of compounds can also be used as powerful deconvoluting tools. Indeed, using a large-scale network of compound-compound similarities (named CSNAP) compounds active in a microtubule polymerization assay were grouped into five distinct target profiles, which were subsequently validated *in vitro*.[112]

1.4 Conclusions and aims of the thesis

Many studies have used the molecular similarity principle to generate target predictions to identify the MoA of compounds active in phenotypic screens. While the molecular similarity principle has been well-characterised in bioactivity datasets generated by target-based screening, it is not entirely clear how well the molecular similarity principle holds in datasets generated from phenotypic-based screening. In this thesis, the molecular representations described in Chapter 2 were employed to evaluate and quantify the molecular similarity principle in two datasets generated by phenotypic screening assays in Chapter 3.

The molecular similarity principle is implicitly applied in a target prediction workflow described in Chapter 2. This workflow is based on two conceptually different target prediction algorithms and it is shown how this workflow allowed to augment the coverage of MoA hypotheses.

Furthermore, while target prediction algorithms relying on this molecular similarity principle have been successfully employed in the literature to deconvolute signals in phenotypic screens, there was not any study, which assessed how the MoAs uncovered by these algorithms compare to the MoAs obtained using publicly available experimental bioactivity data of the compounds active in phenotypic screens. This is the focus of Chapter 4 where the target prediction workflow described in Chapter 2 was used to predict targets for compounds active in a phenotypic screening dataset. The MoA generated through this target prediction workflow were then compared to MoA hypotheses obtained through experimental bioactivity data of compounds active in similar phenotypic endpoints. This comparison will allow generating more insights into why these *in silico* methods are appropriate for the

deconvolution of compounds in phenotypic screens and evaluating their contribution in generating novel MoA hypotheses.

Finally, based on the insights gained from Chapter 4, the target prediction workflow was applied to generate MoA hypotheses for compounds active in a kidney cyst screening dataset in Chapter 5. It will be shown how these predictions can be improved by additional metrics to rank the predicted targets according to their relevance to the phenotypic endpoint measured in the assay. These MoA hypotheses will be further prioritised through their integration with gene expression profiles and occurrence counts in the literature. Structural bioinformatics studies of the shortlisted MoA hypotheses were also performed to strengthen their confidence. Therefore, this chapter shows how target prediction can be integrated with additional data domains to narrow down relevant MoA hypotheses for compounds active in a phenotypic screening assay.

Chapter 2 Computational methods

This chapter introduces the target prediction workflow employed in this thesis. This chapter introduces molecular descriptors and similarity scoring metrics used mainly in Chapter 3. One of the descriptors, namely ECFP4, is then used as input for the target prediction workflow employed in Chapter 4 and 5, and details about these algorithms are also given in this chapter.

2.1 Workflow overview

The target prediction workflow is depicted on **Figure 3**. The input is a query compound structure in SMILES format. The structure is first standardized as recommended by Fourches *et al.*[113] The ChemAxon standardizer[114] (version 15.1.19.0) was used with the options "Remove Fragment" (keep largest), "Neutralize", "RemoveExplicitH", "Clean2d", "Mesomerize", and "Tautomerize". Then a fingerprint is generated from the compound structure. For target prediction, compounds are converted to ECFP4 fingerprints and this process is described in the next section (**Figure 3**).

Generated ECFP4 fingerprints are then used as input for two machine learning models that will output binding probabilities for a large number of protein targets. These will be combined and Z-scaled to filter predictions that may be obtained randomly and/or predictions that are outside the applicability domain represented by molecules active against the corresponding targets (**Figure 3**). More details about all of these steps are given in this chapter.

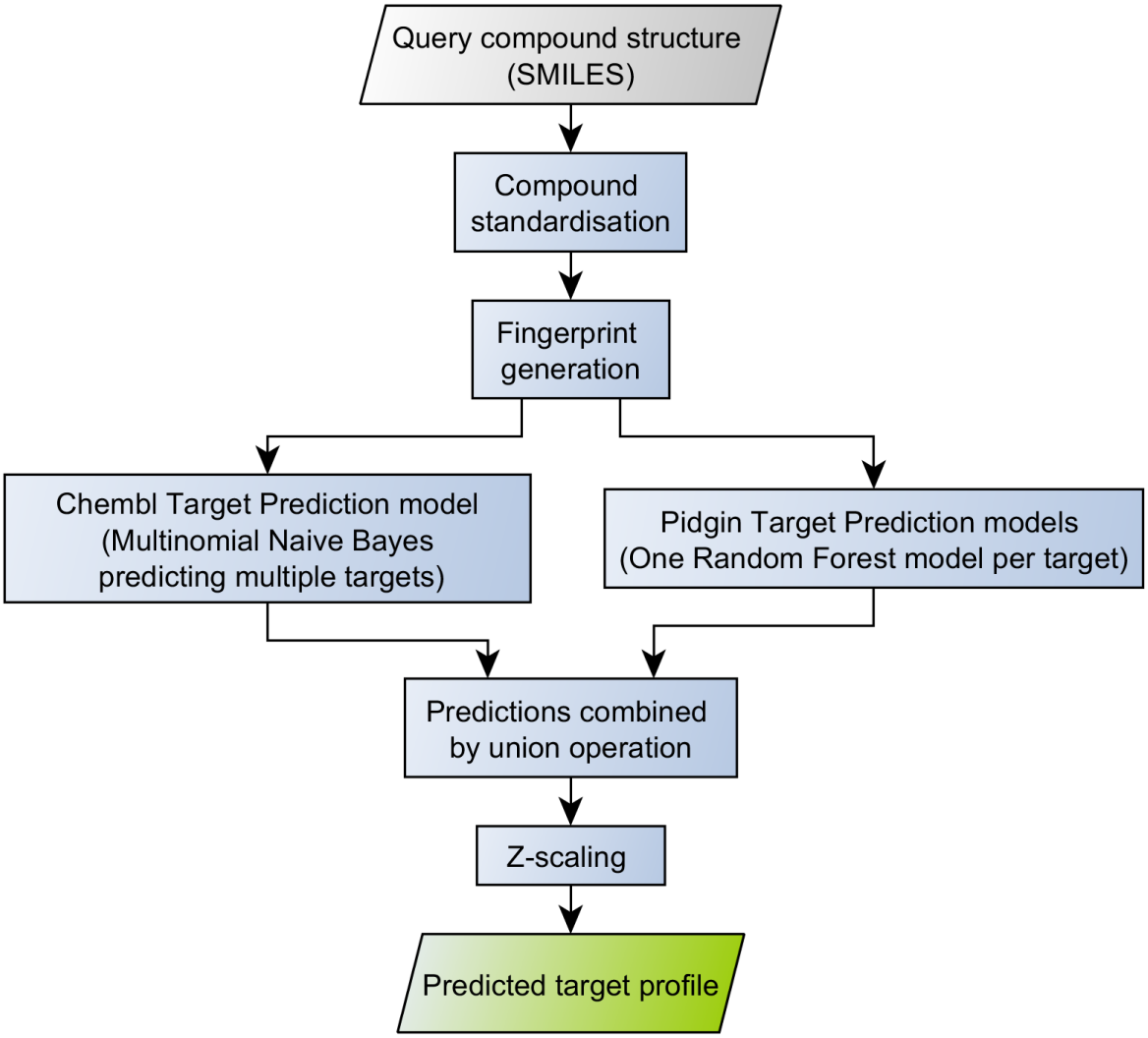

Figure 3. Target prediction workflow employed in this thesis.

2.2 Molecular fingerprints

2.2.1 ECFP4 fingerprints

The generation of ECFP4 fingerprints was performed using the cheminformatics programming libraries rdkit and scikit-chem.[115,116] ECFP4 fingerprints correspond to a binary vector of various bit length, in general, either 1024 or 2048 bits. Each bit represents an atom, substructure or any atomic environment of a certain radius *i.e.* a certain number of bonds away from a given atom of the query molecule. The radius has a maximum of 2 bonds for ECFP4 fingerprints. The algorithm implemented in the cheminformatics libraries follows four steps (**Figure 4**).[43]

In the first step, each atom (except for hydrogen atoms) are associated with a unique integer identifier (**Figure 4**). The generation of the identifier is based on seven atomic properties. These are the number of neighbouring atoms which are not hydrogen atoms, the valence minus the number of hydrogen atoms, the atomic number, the atomic mass, the atomic charge, the number of neighbouring hydrogen atoms, and whether the atom is contained in a ring or not. These values are then hashed together into a single integer which constitutes the initial identifier for this atom.

Once all heavy atoms in the molecule have an identifier, an iterative process starts in which each atom identifier is updated based on its identifier and its neighbours' identifiers (**Figure 4**). A hash function maps the identifier and neighbours' identifier into a new single identifier. Once all atoms have been updated the next iteration starts and the same procedure is repeated

with an increased radius. Hence, at each iteration, more neighbouring atoms are considered than in the previous iteration (**Figure 4**). Note that the previous identifiers, including the initial identifiers, are kept for the next phase of the algorithm.

The third step involves the removal of duplicated identifiers and structures. The final stage involves the use of a hash function to map these identifiers into the final binary vector, with a length defined by the user in general either 1024 or 2048 bits (**Figure 4**). While it has been shown that a small amount of information is lost during this step, two different identifiers can be mapped to the same bit in the final vector and can render the interpretation of the corresponding bit somewhat difficult.[43] Nonetheless, ECFP4 fingerprints, and more generally circular fingerprints, are among the highest performing fingerprints for virtual screening applications such as scaffold retrieval and similarity searching,[117,118] and this motivated their use in this thesis and for the machine learning algorithms described in the next sections.

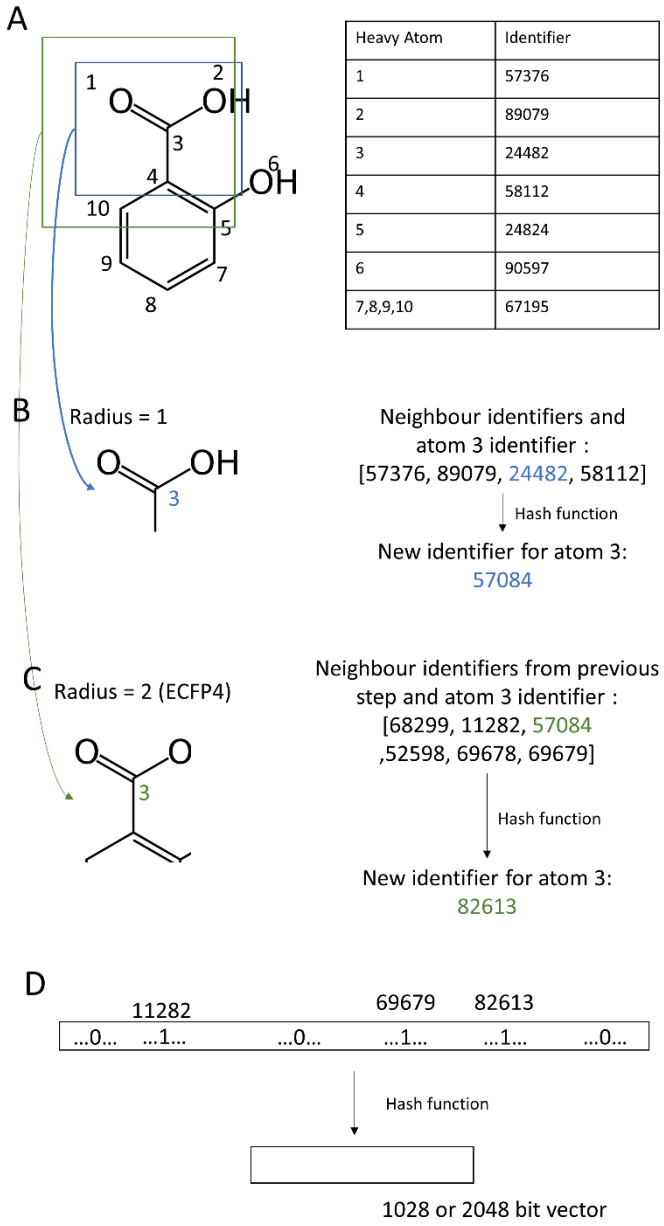

Figure 4. ECFP4 generation algorithm. The first step consists of generating an identifier for each non-hydrogen atom (A). Then for every single atom, the identifiers in the neighbourhood of the atom and the identifier of the atom are hashed into one single identifier. First, only immediate neighbours are considered (B) but then eventually the neighbourhood radius increases (C). In the end, a binary vector is created out of all identifiers from all steps, and this vector is generally hashed into a smaller bit vector of either 1024 or 2048 bits.

2.2.2 MACCS keys and PubChem fingerprints

MACCS keys fingerprints comprise a set of 166 keys representing various structural aspects of the molecules.[46] These encode single atom-based properties and bond types (**Appendix A**). The resulting fingerprint is then a binary vector encoding the presence or absence of each of the 166 keys in the query molecule.

Similarly, to the MACCS keys, the PubChem fingerprint is based on a set of 881 structural keys (**Appendix B**) and the resulting fingerprint is a binary vector representing the presence or absence of each key in the query molecule. The keys encode atomic counts, ring types, atom pairs, atom environments and specific SMARTS patterns.

2.3 Similarity scoring

2.3.1 Structural similarity scoring

Once a representation of compounds has been selected, the assessment of similarity requires computation of similarity metrics. The Tc is usually employed for the similarity of compounds using binary fingerprints and is computed as:

$$Tc = \frac{c}{a + b - c} \quad (1)$$

where c is the number of structural features in common between the two fingerprints, and a and b are the number of features in the first and second molecules respectively. It has been compared to various comparable similarity scoring and performs equally well despite its simplicity.[55,119]

2.3.2 Biological similarity scoring

The similarity of compounds can also be assessed via the biological activity profile of the compounds. In this case, since the profiles are usually made of continuous data (e.g. pIC50s, GI50s etc.), the Pearson correlation coefficient is used.[120] However, the Pearson correlation coefficient tends to be affected by the amount of missing data in the dataset and it is better suited to estimate linear correlations, which is often not the case with two bioactivity profiles. The Spearman correlation coefficient is an alternative that does not assume a linear relationship between the two profiles which is what motivated its use in Chapter 3.

Another biological similarity metric is the Assay Related Target Similarity (ARTS)[121] in which two compounds are similar if both compounds have a similar bioactivity profile *and* if their bioactivities are higher than that of other compound pairs. In other words, similar activities are weighted by their observed activity and ARTS tends to rank lower similar compound pairs with low bioactivities:

$$ARTS_{xy} = \frac{Sim_{xy}}{\sqrt{Sim_{xx}Sim_{yy}}} = \frac{\sum_{i=1}^{n_{xy}}(k + \frac{x_i + y_i}{2})^2 e^{-(x_i - y_i)^2}}{\sqrt{\sum_{i=1}^{n_x}(k + x_i)^2 \sum_{i=1}^{n_y}(k + y_i)^2}} \quad (2)$$

where Sim_{xy} represents the calculated similarity between compounds x and y, Sim_{xx} and Sim_{yy} represent the calculated similarity of compound x with itself and y with itself respectively, n_x, n_y and n_{xy} represent the number of elements for profiles x, y and both respectively, and k is a constant set to the minimal bioactivity observed in the dataset. ARTS has also been showed to display less variability with the amount of missing data compared to Pearson correlation.[121]

The Tc is used throughout this thesis and the Spearman correlation coefficient and ARTS were used in Chapter 3.

2.4 Ligand-target prediction models

The target prediction approach used in this thesis combined two different target prediction algorithms. The first algorithm initially developed by Nidhi et al.[82] was implemented as part of the ChEMBL database and is based on a multinomial NB (MNB) model.[83] The second target prediction model, named PIDGIN, was developed by Mervin et al., which uses a separate RF model per target.[105,122,123] This section is divided into two subsections, one for each model. In each subsection, the machine learning algorithm employed in the target prediction workflow are described, as well as technical details about the algorithm itself. Then additional output processing steps performed in this thesis are described. Finally, the rationale behind combining these two target prediction models is explained.

2.4.1 ChEMBL target prediction model

2.4.1.1 Laplacian-corrected multinomial Naïve Bayes machine learning model

The probability $p(A \mid Fi)$ of a compound to be active against a target A (event A) given the presence of a certain chemical feature Fi (event Fi) in the ECFP fingerprint of the query compound is estimated by the proportion of active compounds with feature Fi in the dataset:

$$p(A \mid Fi) = \frac{A_{Fi}}{N_{Fi}} \tag{3}$$

where A_{Fi} is the number of times Fi is found in actives and N_{Fi} is the number of times Fi is found in all the compounds in the training set. However, since the presence of a certain feature in the ECFP fingerprint can be quite rare, this probability tends to be overestimated, which is why the Laplacian correction is used.

The Laplacian correction is based on the observation that the overestimation bias can be overcome by sampling this feature K times which modifies the probability accordingly:

$$\frac{A_{Fi} + fK}{N_{Fi} + K} \tag{4}$$

where f is the proportion of active compounds in the training set, which is also an estimate for $p(A)$ in this model. The Laplacian correction substitutes K by $1/P(A) = 1/f$ in equation (4) to yield:

$$\frac{A_{Fi} + 1}{N_{Fi} + 1/f} = \frac{(A_{Fi} + 1)f}{N_{Fi}f + 1} \tag{5}$$

The relative estimate of the activity probability given feature Fi is:

$$p_{rel}(A\,|Fi) = \frac{p(A\,|Fi)}{p(A)} = \frac{(A_{Fi} + 1)}{N_{Fi}f + 1} \tag{6}$$

Similarly, relative estimates of the inactivity probability (event $\bar{A}$) can be computed as:

$$p_{rel}(\bar{A}|Fi) = \frac{(\bar{A}_{Fi} + 1)}{N_{Fi}(1 - f) + 1} \tag{7}$$

where $\bar{A}_{Fi}$ is the number of times feature Fi is found in inactive compounds in the training set for target A.

The weights Wi that will be used in the prediction score for query compounds, are computed as the ratio of the relative estimate of activity and inactivity given feature Fi.

$$Wi = \frac{p_{rel}(A\,|Fi)}{p_{rel}(\bar{A}|Fi)} \tag{8}$$

Log-probabilities were used instead in order to avoid numerical issues with small weights. The final prediction score for activity against target A of a query compound is derived by

summing the products of the individual bits from the ECFP4 fingerprint of the query compound ($F_{q,i}$) and the log weights estimated from the training set:

$$Score_{A,q} = \sum_i \log (Wi) F_{q,i}$$

$$= \sum_i \log(\frac{p_{rel}(A\,|Fi)}{p_{rel}(\bar{A}|Fi)}) F_{q,i} \quad (9)$$

$$= \sum_i (\log p_{rel}(A\,|Fi) - \log p_{rel}(\bar{A}|Fi))\, F_{q,i}$$

The same procedure is then repeated for all targets covered by the model to yield a vector of prediction scores for the query compounds. The most likely targets for the query compounds can be found by ranking the targets by the corresponding prediction score. Alternatively, those probabilities can be processed further to ascertain the statistical significance of the prediction. This is described in the next section.

2.4.1.2 Multinomial Naïve Bayes target prediction model and training data

The target prediction model provided in ChEMBL used the MNB approach described above. The model was downloaded from the ChEMBL ftp services (ftp://ftp.ebi.ac.uk/pub/databases/chembl/target_predictions/). Compounds in the training set were converted to ECFP4 fingerprints with 2048 bits in the ChEMBL model. The 10 μM version of the model was used. In this version, active compounds (actives) in the training set

(for each target) were all ChEMBL compounds with a potency of 10 μM (or lower), and the inactive compounds were the remainder.

In total the classifier has 1,290 classes *i.e.* 1,290 targets. The model corresponding to ChEMBL release 22 was used and was implemented using the Python library scikit-learn (version 0.18)[124] using the OneVsRestClassifier() function with parameter estimator=MultinomialNB() and n_jobs=1. The parameter used in the MNB model were alpha=1, class_prior=None and fit_prior=True.

2.4.2 PIDGIN target prediction models

2.4.2.1 Random Forest machine learning model

The RF model developed by Leo Breiman builds on the idea of exploiting the predictive power of many decision trees.[125] Decision trees are partitions of the instances of a dataset based on one variable at a time. The higher the number of such partitions is performed, the deeper the tree is. The resulting groups of instances after all the successive partitions are the most homogenous sets of instances that is possible to obtain. In RF, each individual tree is grown according to the CART methodology,[125,126] in which at each iteration, all candidate variables are evaluated in terms of the purity of their partition, *i.e.* to determine which variable

yields the partition with the most homogenous groups of instances. This is done by calculating the Gini impurity metric for each sub-nodes or resulting groups of instances after the partition using that variable:

$$Gini = \sum_{i}^{2} \frac{n_i}{N}(1 - (p_i^2 + (1 - p_i)^2)) \quad (10)$$

where n_i represents the number of instances in partition i, N is the number of instances that are currently partitioned, and p_i represents the fraction of instances that belong to the class of interest in the resulting partition i. The variable with the highest Gini value is selected for the partition. Then the same procedure is applied to the remaining partitions with the other variables until no variable remains or until all instances belong to homogeneous partitions.

The RF uses this principle to train multiple decision trees based on various random samples of the instances in the training set (sampling made with replacement). In addition, when the decision trees are trained, a small number of randomly selected variables are considered at each split, and those variables will be different in each decision tree.

Predictions are obtained by performing a "vote by the majority" in the forest. Hence, in the context of the PIDGIN model, a compound will be classified as active if most of the individual trees classified this compound as active. The probability score is, therefore, the proportion of decision trees in the forest that classified the compound as active.

2.4.2.2 Target prediction model based on Random Forest and Training data

Unlike the ChEMBL target prediction MNB models described above, PIDGIN is an ensemble of single RF machine learning models, each developed to predict a specific target. The models corresponding to the second version of PIDGIN were downloaded from https://github.com/lhm30/PIDGINv2/. These were implemented with scikit-learn (version 0.17) with number of trees equal to 100, class weight set to 'balanced', and sample weight to the ratio of inactive to active. In the PIDGIN models, the training set data consisted of a combination of 2,089,404 ChEMBL (release 21) active compounds (*i.e.* compounds with PCHEMBL = 5) and 11,829,475 PubChem[70,71] inactive compounds (mined on 21/06/16). The compounds in the training set were represented as ECFP4 fingerprints with 2048 bits. 3,394 RF models are available (one per target).

The output RF probabilities were already calibrated using Platt scaling.[127] This is performed because the probabilities obtained as output from the RF model do not reflect the confidence that the model gives in this prediction. Indeed, it has been previously reported that RF models rarely assign probabilities close to 0 and 1, and this is due to the unlikelihood to have all of the individual trees classifying the query compound has inactive or as active respectively (some noise is always involved in the data which will prevent such 'perfect predictions').[128] The Platt scaling algorithm employs a logistic transformation of the scores from a classifier, which will convert these scores into probabilities. Even though the scores in RF are already probabilities, this transformation effectively corrects the aforementioned behaviour. RF probabilities corrected in this way, therefore, represent the true fraction of actives in the training set. In PIDGIN, this was performed with scikit-learn using the

calibrate_classification_cv with a number of folds set to 3, number of calibration also set to 3, and using the 'sigmoid' method.

2.5 Additional output processing performed in this thesis

Output probabilities from each model were converted to two Z-Scores: one z-score to assess the randomness of the prediction, and the other to assess the similarity to known actives against a target. Hence, output probabilities of the query compounds were compared to the overall probability distribution in the query dataset for a certain target. In addition, known actives of all human targets in ChEMBL22 were extracted for PCHEMBL_VALUE >= 6 with STANDARD_TYPE being one of 'EC50', 'IC50' or 'Ki', and CONFIDENCE_SCORE > 5. The target prediction models were then applied to these known actives to obtain probability distributions for the known actives of each target. Then the output probabilities of the query compounds were also compared to the probability distribution obtained for known actives. Z-Scores corresponding to these two comparisons were computed as:

$$ZScore_{ij} = \frac{p_{ij} - \mu_j}{\sigma_j} \qquad ZScore_{ACTIVESij} = \frac{p_{ij} - \mu_{ACTIVESj}}{\sigma_{ACTIVESj}} \qquad (11)$$

where p_{ij} represents the likelihood of compound i to be active against target j, μ and σ represent the average and standard deviation of the probability distribution obtained for target j for all compounds in the query library, $\mu_{ACTIVESj}$ and $\sigma_{ACTIVESj}$ represent the average and

standard deviation of the probability distribution obtained for known actives in the training set of target j. Therefore $ZScore_{ij}$ measures how far from random the score computed for compound i on target j is, while $ZScore_{ACTIVESij}$ quantifies how the score computed for compound i on target j is far from the average score computed for known actives on the same target.

$ZScore_{ij}$ is largely dataset-dependant and can be adjusted depending on the query dataset and aims of the analysis, whereas the $ZScore_{ACTIVESij}$ depends on the training set and is not adjusted depending on the analysis or query dataset.

A correct prediction should have a score that is as far away from the mean of the prediction score as the average score obtained for the query compound library ($ZScore_{ij}$ >1 or 2), but needs to fall within the distribution obtained for the actives, and preferably as close to the mean of the known actives distribution as possible (-1 < $ZScore_{ACTIVESij}$ < 1). The choice of the cutoff value for $ZScore_{ij}$ will be discussed specifically in each chapter in which the target prediction workflow is used (Chapter 4 and 5).

2.6 Combination of the predictions from both algorithms

The predictions from these two models were combined by taking the **union of all the target binary predictions**. Combining these algorithms also allowed to predict more targets than the use of either of the algorithms alone. Indeed, in the NCATS dataset (which is introduced in detail in Chapter 4), although both algorithms predicted 494 targets in common, they also predict targets which are not found by the other algorithm *i.e.* out of the 761 targets predicted

by Pidgin, 267 are only predicted by Pidgin, and out of the 510 targets predicted by the ChEMBL model, 16 are predicted only by this model and not by Pidgin (**Figure 5**). In addition, the frequency to which they predict certain targets is different by the two algorithms (**Figure 5**). Indeed, 25 targets were predicted for more than 100 compounds by the ChEMBL target prediction models and less than 20 times for Pidgin. Conversely, Pidgin predicted 54 targets more than 100 times, whereas the ChEMBL target prediction models predicted those less than 20 times. In this way, predictions which are not accessible by the use of one algorithm is complemented by the use of the other. This can be used to probe the space of possible MoAs more effectively than with the use of one algorithm alone.

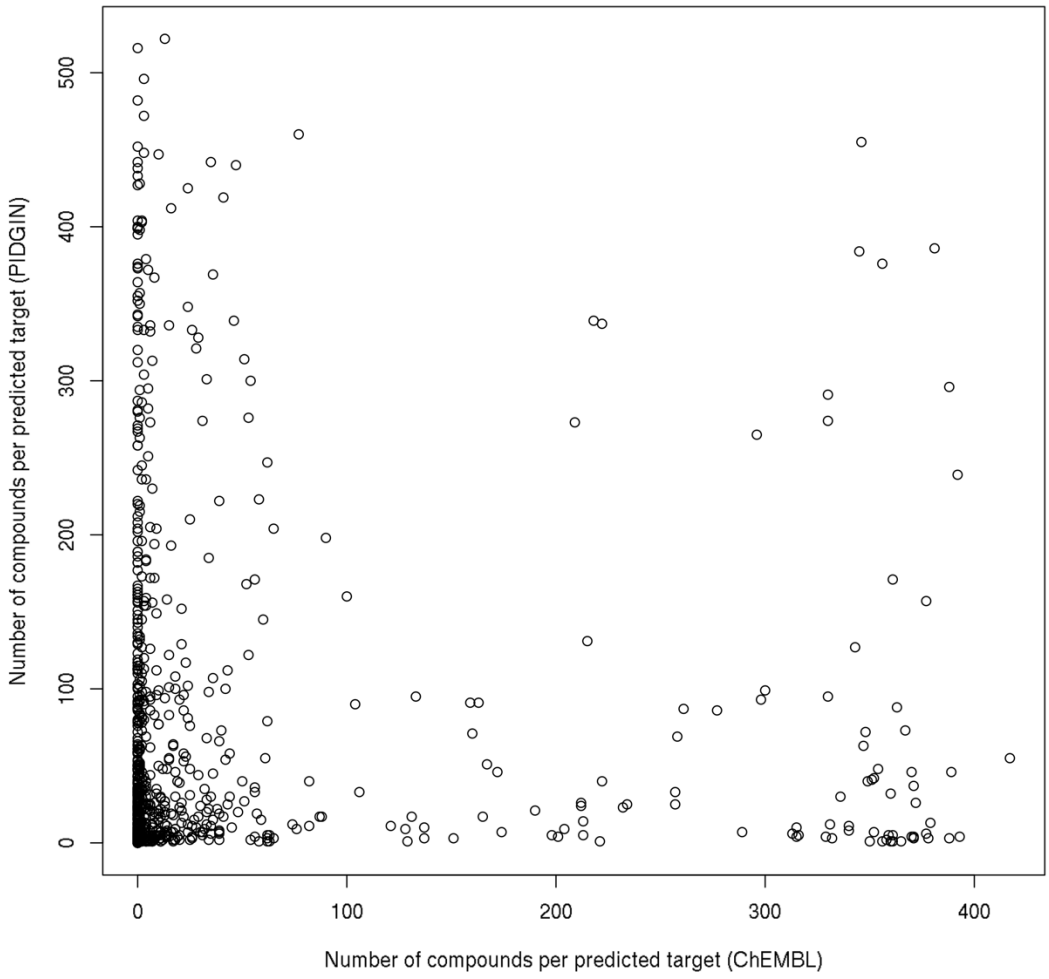

Figure 5. Comparison of compound counts in the NCATS library per predicted target for PIDGIN (y-axis) and CHEMBL (x-axis). Each point represents a target and the coordinates represent the number of times the target is predicted in the corresponding models (x-axis: ChEMBL models, y-axis: PIDGIN). It is clear from this figure that the models prioritise different target predictions.

Chapter 3 Quantifying the molecular similarity principle in phenotypic screening datasets

3.1 Introduction

The molecular similarity principle, which states that two structurally similar compounds may have the same activity in an assay (*i.e* they both bind or modulate the target in the same way), is central to all cheminformatics applications for drug discovery including target prediction. Further, target prediction algorithms assume that query compounds being similar to the known actives of a target is enough to hypothesise the activity of the query compound against this target as well. While the molecular similarity principle is well characterised in target-based screening datasets, there is still a lack of clear evidence that this molecular similarity principle is valid in phenotypic screening datasets, where the cytotoxic activity of compounds against cell-lines or modulation of pathways of interest is studied. Moreover, the molecular similarity principle is more difficult to study in phenotypic screens as a change in the structure of a compound can affect the binding of the compound to different targets and hence multiple

cellular processes can change. This structural change may hence result in a bigger change of activity in the phenotypic assay compared to the change of activity it might yield in a target-binding assay. Because the molecular similarity principle is an implicit assumption of ligand-target prediction models used to deconvolute signals in phenotypic screens, it is therefore important that the molecular similarity principle is studied and quantified with phenotypic screening datasets as well.

Previous attempts to study the molecular similarity principle in phenotypic screening datasets used integrative analyses of structural data and gene expression of cancer cell lines. Chen et al. for example, found that there is a 20% chance for highly structural compounds with Tc > 0.85 to share the same gene expression profile.[129] Another study looked at the chemical/phenotypical similarity relationship while considering cytological features of HeLa cells as their phenotypic profiles, and the authors defined ways to measure and represent correlations between chemical similarity and phenotypic similarity, which could then be used to predict and validate compounds that triggered specific cellular events such as microtubule depolymerization and mitotic arrest.[85] It was found that 96% compound pairs have a concordant structure-activity relationship meaning that they have a Tc > 0.3 and have very similar effects on the cytology of the cells. Finally, Tiikkainen et al. also studied the correlation between chemical similarity and cytotoxic activity against a panel of 60 cancer cell lines and found that the fraction of chemically similar compounds increases for larger values of cytotoxicity profile similarity for Pearson correlation of 0.5 or more.[130] These studies were based on cancer cell lines but few studies considered the relationship between chemical similarity and activity similarity in a phenotypic assay using primary cell lines instead of cancer cell-lines which are not reflecting the biology of cells in normal tissues. This limitation was addressed in Shah et al, where biological similarity of compounds was based on biological profiles that combined both target-based and phenotypic-based assays from the

ToxCast dataset. They found that compound pairs that shared both a chemical similarity measured by Tc > 0.6 and a biological profile correlation > 0.6 had similar known on-target effects (e.g caffeine and theophylline were among such pairs). Unfortunately, only a low number of chemical pairs with such properties were found, which might be attributed to the very sparse nature of the ToxCast bioactivity matrix, as well as the poor correlation between pharmaceuticals and pesticides.[131]

This chapter is concerned with studying the relationship between chemical and phenotypic similarity while addressing the issues encountered in the previous publications. It is important to study and quantify the molecular similarity principle across cells from different tissues and organs, to show that this is applicable for as many phenotypic assays as possible. Otherwise, the molecular similarity principle is only applicable to those cells that are currently studied, which are mostly cancer cells. In addition, this work also quantifies the molecular similarity principle, and this is the novel aspect of this work compared to previous research on the subject. It is investigated how phenotypic similarity develops as a function of Tc chemical similarity based on three widely used fingerprints in cheminformatics namely, ECFP4, MACCS keys, and PubChem fingerprints, which were discussed in Chapter 2. Employing Bayesian regression models, the increase in phenotypically similar compounds with increasing chemical similarity bins is estimated. Eventually, the neighbourhood behaviour of the fingerprints is analysed to explain the differences in trends observed between the fingerprints, and it is showed that ECFP4 is less sensitive to the choice of the dataset used in the analysis, compared to the other two fingerprints.

3.2 Materials and methods

3.2.1 Phenotypic profile data collection and preparation

3.2.1.1 BioMAP dataset

Dataset background and extraction

The BioMAP dataset is a collection of phenotypic screening assays in which compounds are assessed in term of their effects on various human cells by measuring protein biomarkers characterising the state of the cells. The activity readouts in this dataset are therefore protein expression levels of the markers of interests. More specifically, BioMAP assays report the logarithm of the expression level of the protein biomarker measured after treatment, divided by the expression levels of the biomarker in the control samples (DMSO). This dataset is used in this analysis as they are representative of the disease-based category of phenotypic assays but also because they represent non-cancerous human cell lines. Regarding details on the generation of BioMAP profiles, readers are referred to previous publications for more information on the experimental details on reagents used, cell culture details and enzyme-linked immunosorbent assay (ELISA) measurements.[29,132] In total, 8 cell lines were used in this analysis. The ELISA readouts in each cell line consisted of protein expression levels which were converted into a log ratio of expression of the protein in the treated sample, as

compared to the protein level in control samples (DMSO). 1,120 compounds along with their structural information were obtained from the ToxCast website.[133]

Compound filtering

Overtly cytotoxic compounds were first identified by cytotoxicity against Peripheral blood mononuclear cells (PBMC) and also in an SRB assay (the amount of protein-bound dye in the assay approximates cell mass) of various cell lines leading to the removal of 284 compounds out of 1,120. Assays and corresponding cut-offs employed are summarised in **Table 4**.

CELL SYSTEM	ASSAY NAME	ACTIVITY CUTOFF
3C	SRB	-.24
4H	SRB	-.16
SAg	SRB	-.14
SAg	PBMC Cytotoxicity	-.14
LPS	SRB	-.14
BE3C	SRB	-.08
KF3CT	SRB	-.08
HDF3CGF	SRB	-.2
CASM3C	SRB	-.2

Table 4. Assays and cut-offs employed to filter out cytotoxic compounds in the BioMAP dataset.

For a given compound in the BioMAP dataset, the ensemble of all 84 protein level readouts constituted the phenotypic profile of this compound. 95% significance envelopes of activity values created from historical controls[29] were used in this study to discard false positive

readouts from a compound profile (activity readout falling inside the 95% envelope) *i.e.* these activity points were set to 0.

Creation of the BioMAP profiles

The ensemble of all the assays for a compound constituted the profile for a compound. Moreover, positive readouts were separated from negative readouts in order to obtain profiles with one unique direction. For example, if a compound was measured with a protein expression of -0.5 for "AID = 114", this activity was transformed as 0.5 under a new readout "AID = 114_dn" in this compound profile. Similarly, a protein expression of 0.5 for the same variable "AID = 114", would be transformed as "AID = 114_up". This provides the same effect as taking the absolute value without losing the directionality of the activity.

Clustering and reduction of the missing data

The resulting matrix was clustered to obtain a dataset that is as complete as possible. Hence both dimensions of the matrix were sorted by the number of activity points yielding a dataset of 365 compounds that consisted of protein levels in 168 assay descriptors (84 readouts with bidirectional profiles) which is 59% complete.

3.2.1.2 ChEMBL compound dataset

Dataset background and extraction

To study the molecular similarity principle for a type of phenotypic readout different than the biomarker-based BioMAP dataset, an additional dataset was extracted from ChEMBL21[72,73] This dataset is comprised of cell viability against cancerous cell lines and this makes the analysis conducted in this chapter more comprehensive than the previous research: combined with the BioMAP dataset, both primary human cells and cancer cells are used.

An initial bioactivity matrix was created by extracting all the bioactivity data *i.e.* pCHEMBL values (-log10 IC50/EC50) of compounds against any human cell line assay reported in ChEMBL21, yielding a dataset of 14,743 bioactivity data points. 21 data points corresponding to assays in which the protective effect of compounds was sought (rather than the inhibition/cytotoxicity) were removed from the dataset for consistency.

Clustering and reduction of the missing data

Next, this matrix was clustered to get a dataset that is as complete as possible, while also retaining a number of compounds and assays comparable to that of the BioMAP dataset. The top 400 compounds were kept and 87 cell lines resulting in a cytotoxicity matrix which is 21% complete. The structures corresponding to the 400 compounds in SMILES format were also extracted from ChEMBL21. This dataset contained mostly cancer cell lines. In a similar

way to the BioMAP dataset, the phenotypic profiles of the compounds were created by creating a vector of all pChEMBL values available for that compound. Missing data are kept as is, as they will not be used in the latter part of the analysis (as explained in the following).

3.2.2 Standardisation, fingerprint generation and chemical similarity

Compounds were standardized as recommended by Fourches *et al.*[113] The ChemAxon standardizer[114] (version 15.1.19.0) was used with the options "Remove Fragment" (keep largest), "Neutralize", "RemoveExplicitH", "Clean2d", "Mesomerize", and "Tautomerize".

ECFP4 fingerprints (2,048 bits),[43] MACCS keys (166 keys)[46] were generated using scikit-chem.[116] Additionally, PubChem fingerprints (881 bits) were extracted from the PubChem database[70,74] using the python API pubchempy.[134] Tcs were then computed for all pairs of compounds for all three fingerprints, hence yielding three chemical similarity matrices. Chemical and structural similarity are used interchangeably in the following.

3.2.3 Phenotypic similarity coefficients

Computations in this section were performed using R (version 3.2.4).[135,136] Spearman correlation coefficients were computed as a metric to assess the phenotypic similarity of the compounds in both datasets. Because missing data were present in the compound profiles, the correlation was calculated only on the readouts that are in common between the two compound profiles.

The phenotypic similarity of compounds was also assessed based on the ARTS metric. Since the datasets used in this study are sparse and quantifying similarity requires robust metrics, the ARTS metric was selected since it has previously been shown to assign meaningful biological similarity values between assay readouts also in case of a large fraction of missing data.[121]

Because the ARTS metric was originally developed to compare pIC50 profiles, the BioMAP profiles were first converted to pIC50 range prior to the use of ARTS using the following transformation:

$$y' = \log_{10}(y) + \mathrm{c} \tag{12}$$

where y' is the transformed BioMAP readout, y is the original BioMAP readout and c=$(4 - \min(\log_{10}(y)))$. ARTS was applied in a similar way to the Spearman coefficient, *i.e.* only on readouts that are in common between the two profiles.

3.2.4 Modelling of the relationship between the fraction of phenotypically similar compounds with increasing chemical similarity using Bayesian regression models

To quantify how phenotypic similarity varies on average with increasing values of chemical similarity, compounds were binned into 20 chemical similarity bins, ranging from Tc = 0 to Tc = 1 and each bin spanning Tc = 0.05. Then for each combination of the three fingerprints and the two biological similarity measures described above, 9 thresholds of phenotypic similarity were explored *i.e.* every 0.1 unit in Spearman correlation and ARTS metric. For each phenotypic similarity threshold in combination with one of the two phenotypic similarity measures and one of the three fingerprints, the fraction of compounds was computed in each chemical similarity bin that was above this threshold. Therefore, there was one such curve per combination of threshold, phenotypic similarity metric and fingerprint.

Curves for which the threshold yielded less than 1% of the compound pairs below or above the threshold were removed since this yielded constant lines around either 0% or 100% phenotypically similar pairs respectively. These included thresholds between 0.1 and 0.6 ARTS or above 0.9 on the BioMAP dataset, and above 0.9 Spearman correlation also on the BioMAP dataset. No such curves were found and removed on the ChEMBL dataset.

The remaining curves corresponding to all the thresholds were averaged, and the curves were grouped by phenotypic similarity metric, fingerprint and dataset. There were, therefore, twelve averaged curves (six per dataset).

Bayesian regression models were employed to these averaged curves using the rstanarm R package (v. 2.17.2)[137] to model and quantify the increase in phenotypically similar pairs with increasing chemical similarity. The reason for employing the Bayesian framework instead of

the traditional least squares regression is that the model assumptions about normality and linearity of the residuals were not always obtained for some combinations of fingerprints, phenotypic similarity metric and thresholds.

Using Bayesian regression models allowed for the estimation of the slopes of the above curves and fit different models to determine the trend of the relationship between chemical similarity and phenotypic similarity. When the model is a simple linear regression model, the slope represents the increase in the fraction of phenotypically similar compounds per 0.05 or 5% increase in Tc chemical similarity.

The regression models were obtained through the *stan_glm* function with parameters family=gaussian(link='identity') for linear regression, and chains=3 for faster convergence. The initial values of the Markov chains were constrained to be sampled from [-0.5 and 0.5] through the use of the init_r parameter since this improved the convergence of the models. The weakly informative default priors were used for the estimation of the intercept, slope and all the other coefficients (discussed in the following), which in the case of the regression are modelled by Gaussian distribution with $\mu=0$ and $\sigma=5$ for the intercept, or $\mu=0$ and $\sigma=10$ for the other coefficients of the regression. For the auxiliary *i.e.* the error or residuals of the model (ε_i), the default prior was used for the error standard deviation of the response variable (fraction of phenotypically similar compounds) which was modelled by an exponential model with rate=1.

Four models were evaluated per averaged curves. The first one modelled a constant linear relationship between the fraction of phenotypically similar compounds and chemical similarity:

$$f_{i,FP,BM} = \alpha + \varepsilon_i \quad (13)$$

where $f_{i,FP,BM}$ is the fraction of compounds with phenotypic similarity metric BM (BM = Spearman correlation or ARTS) in chemical similarity bin i, and α is the intercept of the line. When the relationship between the fraction of phenotypically similar compounds with increasing chemical similarity bin was not constant, the following model was applied:

$$f_{i,FP,BM} = \alpha + \beta x_{i,FP} + \varepsilon_i \quad (14)$$

where $x_{i,FP}$ is the ith chemical similarity bin using fingerprint FP, while α, β and ε_i are the intercept, slope and residuals of the model respectively. Sometimes, the trend was not linear but had a parabolic shape, in which case the following model provided a better fit:

$$f_{i,FP,BM} = \alpha + \beta_1 x_{i,FP} + \beta_2 x_{i,FP}^2 + \varepsilon_i \quad (15)$$

where β_1 is the slope for the first chemical bin (0<Tc<0.5), and β_2 indicates the direction of the curvature and represents the variation of the slope from one chemical similarity bin to the next. In the last case, the trend was a combination of two regression lines which was modelled according to the following piecewise linear model:

$$f_{i,FP,BM} = \alpha + \beta'_1 x_{i,FP} + \beta'_2 [x_{i,FP} - b_{BM,FP}]_+ + \varepsilon_i \quad (16)$$

where β'_1 is the slope of the first regression line and β'_2 is the difference in slopes between the first and the second regression line. b is the breakpoint *i.e.* the chemical similarity bin defining the end of the first regression line, and the start of the second line. b was estimated independently to this model using the r package segmented[138] (v.0.5.3.0) for each combination

of phenotypic similarity metric BM, phenotypic similarity threshold t and fingerprint FP. $[x_{i,FP} - b_{BM,FP}]_+$ represents a hinge function, for which only the positive values of the expression within the brackets are retained, so that the second slope is estimated for chemical similarity bin b and above.

3.2.5 Model selection and estimation of slopes and breakpoint

After fitting these four models to the averaged curves, the R package loo[139] (v. 1.1.0) was utilised to compute the expected log pointwise predictive density (elpd) for each model. This measure estimated the quality of the fit. The higher the elpd, the better the model fits the data. Hence, for each averaged curve corresponding to one of the 12 combinations of biosimilarity metric, fingerprint and dataset, the model with the highest elpd was selected.

Since the aim of this paper is the quantification of how phenotypic similarity varies with increasing values of chemical similarity, the coefficients representing the slope (or equivalent depending on the models) were estimated by taking the median of the posterior probability distribution. A 95% confidence interval was also directly calculated by obtaining the quantiles 2.5% and 97.5% of the posterior distribution.

Since some of the trends were parabolic or bi-linear, estimates for the breakpoint were calculated *i.e.* an estimate of the chemical similarity bins at which the second slope or increasing trend starts. For the quadratic models, these can be obtained by computing the x-axis of the point corresponding to the vertex of the parabola ($= -\beta_1/2\beta_2$). For the piecewise linear models, the estimates of the breakpoint were computed using the R package segmented as mentioned in the previous section.

3.2.6 Neighbourhood Enhancement

Since the variation of phenotypic similarity with increasing chemical similarity behaved differently depending on the fingerprints, it was hypothesized that fingerprints had different neighbourhood enhancements also in the context of phenotypic screening. This concept has previously been established by Patterson et al.[59] by plotting the distance in biological readout space over the distance in chemical descriptor space. An 'ideal' descriptor would lead to high biological readout similarity for the large majority of chemically similar compounds, with only a minority of compounds showing different biological response for similar compounds, what is today called 'activity cliffs'.[61,140]

The enhancement ratio[59] is the score Patterson et al. developed to quantify how many such activity-cliffs points are present on any plot comparing chemical and biological similarity. In this analysis, enhancement ratios were calculated per fingerprint, per assay, and per dataset. The maximal enhancement ratio that can be computed is 2. Hence the closer to 2 the enhancement ratio gets, the better the descriptor was to predict neighbourhood regions and hence yield less activity-cliffs. In addition, as recommended by Paterson et al., a Chi-squared test ($p < 0.01$, one degree of freedom) was utilised to assess whether the point density in neighbourhood regions was higher than expected under a uniform distribution.

3.3 Results and discussion

3.3.1 Phenotypic similarity increases with structural similarity

Using activities from phenotypic datasets that are different in nature, *i.e.* a protein biomarker expression dataset (BioMAP) and a cytotoxicity dataset (ChEMBL), the relationship between phenotypic similarity and chemical similarity was evaluated.

To this aim, phenotypic profiles for each compound were created, and the Spearman correlation and ARTS correlation quantified the pairwise phenotypic similarity of those phenotypic profiles. Similarly, chemical representations for each compound were obtained using ECFP4, MACCS and PubChem fingerprints, and the pairwise chemical similarity was obtained using Tc similarity. Compounds were binned into several chemical similarity intervals, and the fraction of phenotypically similar compounds was calculated for various thresholds of phenotypic similarity.

In the BioMAP dataset, the fraction of phenotypically similar compounds increased linearly only for certain thresholds of ARTS and Spearman correlation coefficients (**Figure 6**). The fraction of phenotypically similar compounds remained constant for thresholds of ARTS between 0.1 and 0.7 and an increasing trend was observed only for thresholds of ARTS between 0.8 and 0.9. This was due to the fact that less than 1% of compounds had an ARTS score below 0.7. For the Spearman correlation coefficient, the fraction of phenotypically similar compounds increased with chemical similarity for Spearman correlation thresholds between 0.4 and 0.7. Those values were in accordance with the study by Shah et al. who used BioMAP readouts (among others in the ToxCast dataset) where compounds that had a

biological similarity correlation of >0.6 coincided with compounds that were structurally similar.[131] Hence the molecular similarity principle is observed in the BioMAP dataset regardless of the fingerprint used, which may suggest that the molecular similarity principle is also applicable for biomarker-based phenotypic screening assays similar to the BioMAP assays.

In the ChEMBL dataset, the fraction of phenotypically similar compounds increased with chemical similarity for all thresholds (*i.e.* all subpanels of phenotypic similarity thresholds) of ARTS in a piecewise linear manner *i.e.*, where the fraction of phenotypically similar compounds increased only after a certain chemical similarity has been reached (**Figure 7**). This is in accordance with the study by Young et al. which showed that the molecular similarity principle holds in so-called "structure-phenotype concordance regions" defined by Tc > 0.3.[85]

When the Spearman correlation coefficient was used, the fraction of phenotypically similar compounds increased mainly for ECFP4 and only up to a Spearman correlation threshold of 0.5, after which the fraction of phenotypically similar compounds remained constant. For the other two fingerprints, the fraction of phenotypically similar pairs decreased until reaching a certain chemical similarity and increased again slightly for most thresholds. These results suggested that the Spearman correlation metric is not appropriate for molecular similarity principle studies on cell viability readouts comparable to those of the ChEMBL dataset.

Nonetheless, the fraction of phenotypically similar compounds increased with chemical similarity for most thresholds of Spearman and ARTS metrics in both datasets. These results also agreed with the current literature since the molecular similarity principle was observed only for a certain range of Tc values. In the following, the molecular similarity principle is quantified along with the Tc values after which the increase in phenotypic similarity with increasing chemical similarity is observed.

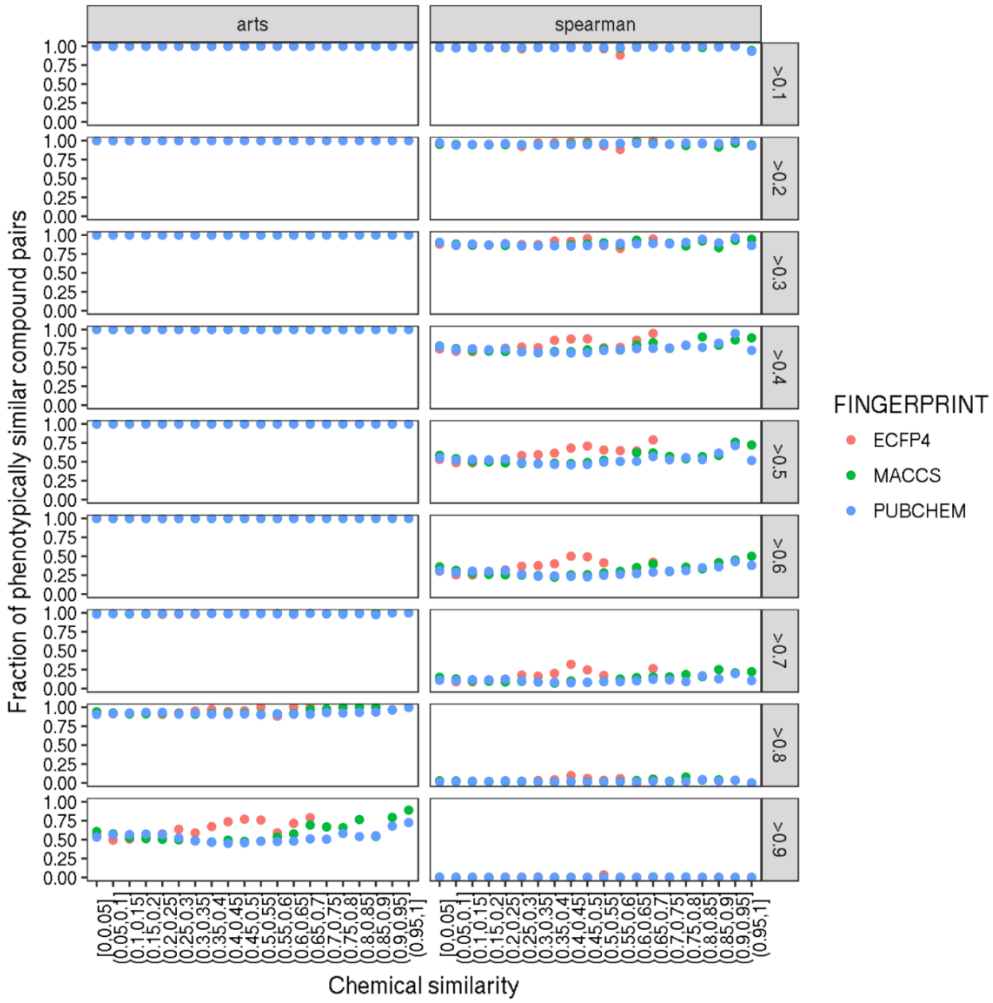

Figure 6. Evaluation of the relationship between the fraction of phenotypically similar compound pairs and chemical similarity in the BioMAP dataset for various thresholds of Spearman or ARTS. Each panel represents a combination of phenotypic similarity measure (column) and the corresponding threshold (row). Then in each panel, the fraction of compound pairs with phenotypic similarity above the threshold is depicted as function of increasing chemical similarity divided into 20 bins. For ARTS, the fraction of phenotypically similar compounds increased with chemical similarity only for ARTS thresholds of 0.8 or higher. For the Spearman correlation coefficient, the fraction of phenotypically similar compounds increased with chemical similarity for Spearman correlation thresholds between 0.4 and 0.7. For other thresholds, the fraction of phenotypically similar compounds did not seem to vary with increasing chemical similarity. ECFP4 fingerprints started to increase for lower chemical similarity values compared to the other two fingerprints.

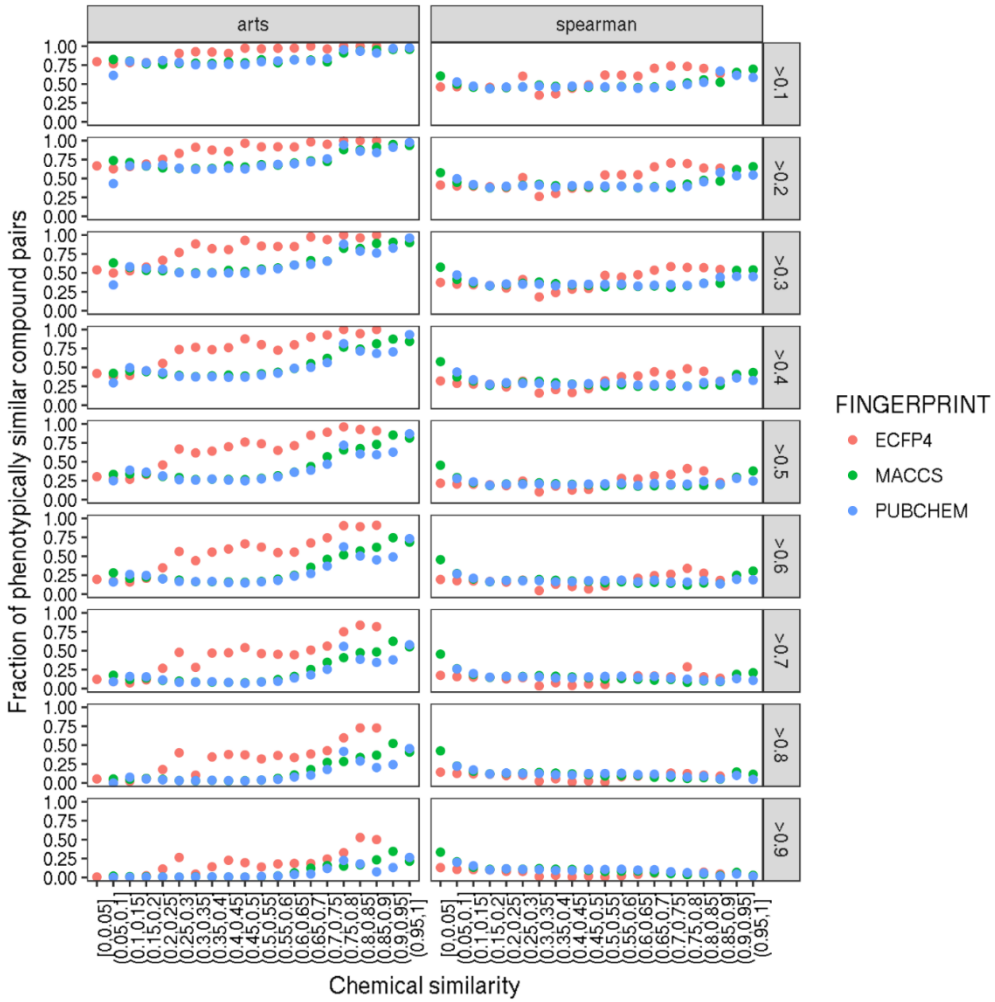

Figure 7. Evaluation of the relationship between the fraction of phenotypically similar compound pairs and chemical similarity in the ChEMBL dataset for various thresholds of Spearman correlation or ARTS. Each panel represents a combination of phenotypic similarity measure (column) and the corresponding threshold (row). Then in each panel, the fraction of compound pairs with phenotypic similarity above the threshold is depicted as a function of increasing chemical similarity divided into 20 bins. With ARTS, the fraction of phenotypically similar compounds increased with chemical similarity. With Spearman the fraction of phenotypically similar pairs decreased up to a certain Tc after which it increases slightly again. Again, ECFP4 fingerprints started to increase for lower Tc values compared to the other two fingerprints.

3.3.2 Quantification of the molecular similarity principle through Bayesian regression models

To quantify the average increase in the fraction of phenotypically similar compounds per a small increase in chemical similarity, curves corresponding to all combinations of phenotypic similarity metrics and fingerprints were averaged. Then four Bayesian regression models were applied to these twelve averaged curves (six per dataset) and the model which best fitted the curves were selected. One model, the constant linear model, was employed to test for a flat relationship of the phenotypic similarity with chemical similarity. In other words, if this constant model better fitted those curves compared to the other three models, the slope would be null and the molecular principle would therefore not be observed.

In the BioMAP dataset, the averaged variation of the fraction of phenotypically similar pairs increased moderately with increasing chemical similarity (**Figure 8**). Even though this increase was slight, the fact that the constant model did not have the best fit compared to the other three models, in all cases, indicated that the observed increase in phenotypic similarity with increasing chemical similarity was statistically above 0. When ECFP4 was used, the linear regression model fitted best, and the slope was estimated as a 0.9% increase in the fraction of phenotypically similar compound pairs per 5% increase in Tc chemical similarity (**Table 5**). When the MACCS was utilised, either the quadratic model or the piecewise linear model fitted best depending on the phenotypic similarity metric employed (**Figure 8**). For the PubChem fingerprint, the piecewise linear model had the best fits compared to the other models for both phenotypic similarity metrics (**Figure 8**).

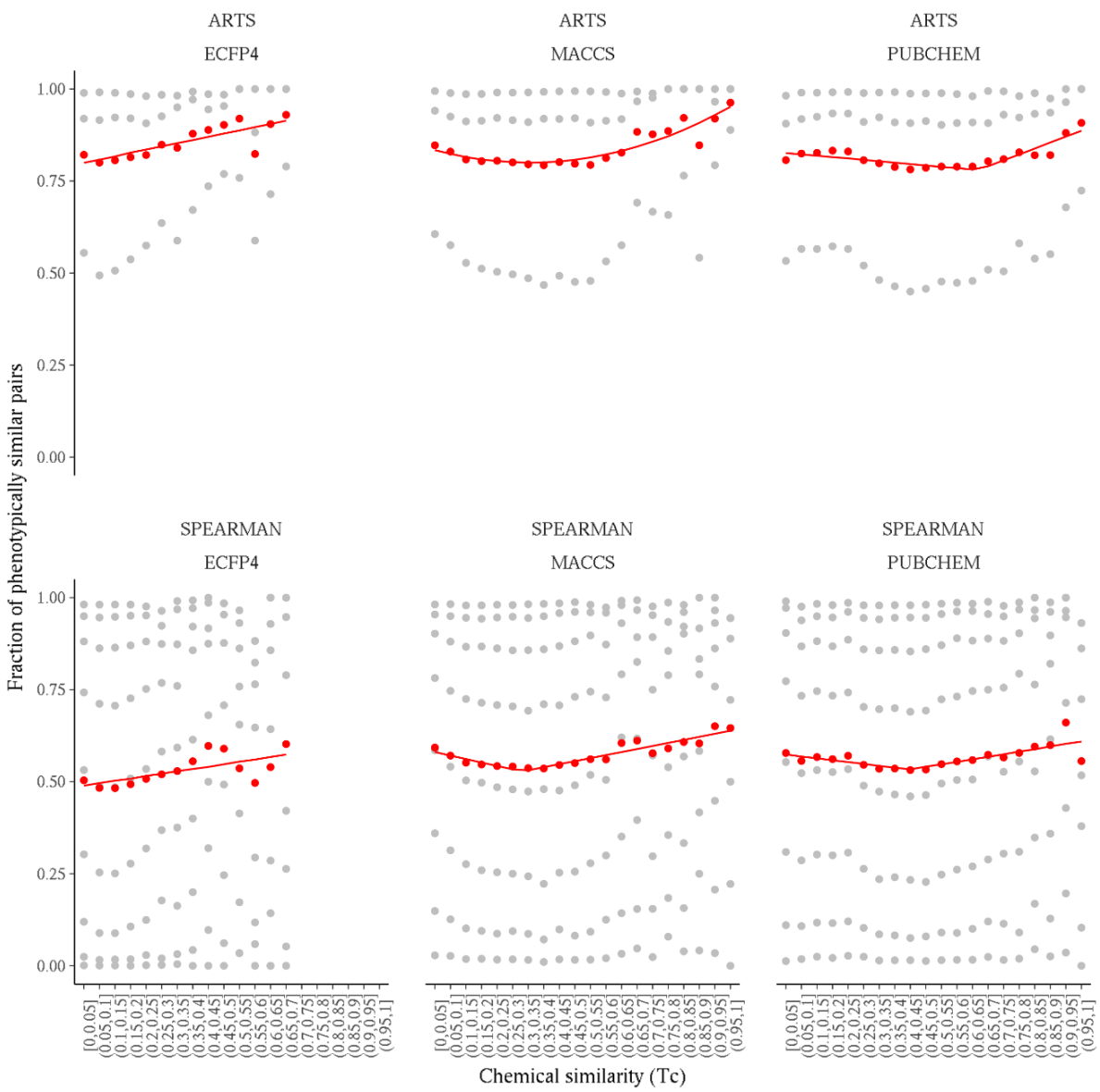

Figure 8. Curve averaging and Bayesian regression modelling for the BioMAP dataset. Each panel represents a combination of fingerprint and phenotypic similarity metric. In each panel, the grey curves represent the fraction of phenotypically similar pairs with similarity above a specific threshold (**Figure 6**) for increasing chemical similarity. The red curve is the point-by-point average of those curves in each panel. The Bayesian model with the best fit to the average curve is also overlaid.

The molecular similarity principle is observed for the whole chemical similarity scale for ECFP4 whereas, for the other two fingerprints, the molecular similarity principle was only observed after a certain Tc threshold. This Tc threshold will be referred to as **breakpoint** in the following. This observation of a breakpoint is in agreement with most of the studies which also examined the relationship between phenotypic similarity and chemical similarity in a similar way.[129,130]

For the MACCS and PubChem fingerprint, since the molecular similarity principle was only observed after the breakpoint, an estimate for this value must first be calculated. Since the trend was parabolic for MACCS with the ARTS metric, the breakpoint corresponded to the x-coordinate of the vertex of the parabolic trend, which was estimated as Tc = (0.3,0.35] (**Table 5**). Using MACCS in combination with the Spearman correlation metric, a piecewise linear trend was obtained which meant that the breakpoint was estimated as the Tc which marks the beginning of the second slope. Remarkably, this was also estimated as Tc = (0.3,0.35] (**Table 5**).

For PubChem, the trend was piecewise linear for both ARTS and Spearman correlation metrics but the breakpoint was estimated as Tc = (0.6;0.65] for the former and Tc = (0.4,0.45] for the latter (**Table 5**). MACCS displayed earlier breakpoint than PubChem, meaning that the molecular similarity principle was observed for a higher range of chemical similarities for MACCS compared to PubChem.

Once the breakpoints were estimated, this value was used to limit the portion of the curves for the estimation of the increase in phenotypic similarity for increasing chemical similarity *i.e.* provide an estimate of the molecular similarity principle.

For MACCS, when using ARTS to quantify the phenotypic similarity, since the curve was parabolic, this estimate corresponded to the average slope observed from the breakpoint to the maximal Tc bin ([0.95, 1]) which was estimated as 1.2% increase in the fraction of

phenotypically similar compounds per 5% increase in Tc chemical similarity. For MACCS using the Spearman correlation to quantify the phenotypic similarity, the second slope provided an estimate of the molecular similarity principle and this was estimated as a 0.9% increase in the fraction of phenotypically similar compounds per 5% increase in chemical similarity (**Table 5**).

Since the PubChem trend was piecewise linear for both phenotypic similarity metrics, the second slope provided the molecular similarity principle in both cases. For ARTS, this was estimated as 1.6% increase in the fraction of phenotypically similar compounds pairs per 5% increase in chemical similarity (**Table 5**). In the case of the Spearman correlation, this was estimated as a 0.7% increase in the fraction of phenotypically similar compounds pairs per 5% increase in chemical similarity (**Table 5**).

For the ChEMBL dataset, the constant model did not win over the other three models for any combination of fingerprint and phenotypic similarity metric. This once again showed that the observed increase in phenotypic similarity with increasing chemical similarity was significantly different than 0. In this dataset, the piecewise linear model predominated in most combinations of fingerprints and phenotypic similarity metrics (**Figure 9**).

When ECFP4 and ARTS were used, the fit consisted of two slightly differing increasing slopes. For the remaining of the combinations where the piecewise linear model was the best fit, the first slope was negative or almost null and only the second slope was positive (**Figure 9**). The latter behaviour reinforces the current knowledge that biological similarity increases only after a Tc breakpoint,[129,130] suggesting that phenotypic similarity behaved in the same manner as well. When the Spearman coefficient was used, the quadratic model had a better fit over the other models for MACCS and PubChem fingerprint but not for ECFP4 (**Figure 9**).

	ARTS			SPEARMAN		
	selected model	estimate	95% confidence interval	selected model	estimate	95% confidence interval
ECFP4	linear	$\beta = 0.009$	[0.005, 0.013]	linear	$\beta = 0.007$	[0.002, 0.011]
MACCS	quadratic	vertex = (0.3,0.35]		piecewise linear	b = (0.3,0.35]	(0.2,0.4]
		$\beta_1 = -0.013$	[-0.02, -0.005]		$\beta'_1 = -0.009$	[-0.014, -0.005]
		$\beta_2 = 0.0009$	[0.0005, 0.0012]		$\beta'_2 = 0.018$	[0.012,0.023]
					$\beta'_1 + \beta'_2 = 0.009$	
PUBCHEM	piecewise linear	b = (0.6;0.65]	(0.55;0.75]	piecewise linear	b = (0.4,0.45]	(0.25,0.6]
		$\beta'_1 = -0.004$	[-0.006, -0.002]		$\beta'_1 = -0.005$	[-0.009, -0.001]
		$\beta'_2 = 0.02$	[0.014,0.026]		$\beta'_2 = 0.012$	[0.005,0.019]
		$\beta'_1 + \beta'_2 = 0.016$			$\beta'_1 + \beta'_2 = 0.007$	

Table 5. Key estimates and associated 95% confidence interval for the models fitted on the averaged curves for the BioMAP dataset (displayed in **Figure 8**). Only estimates related to the increase of phenotypic similarity with increasing chemical similarity are reported. If the selected model (*i.e.* the model with the best fit to the average curve) is the linear model, then the slope (β) is reported. If the selected model is quadratic, then the initial slope (β_1) is reported as well as the increase in slope for each 5% increase in chemical similarity (β_2). The x-coordinate of the vertex (*i.e.* the breakpoint where the curve changes direction) is also reported. Finally, if the model is piecewise linear, then the estimate of the breakpoint (b), the estimate of the first slope (β'_1), the estimate of the difference in two slopes (β'_2) and the estimate of the second slope ($\beta'_1 + \beta'_2$) are reported.

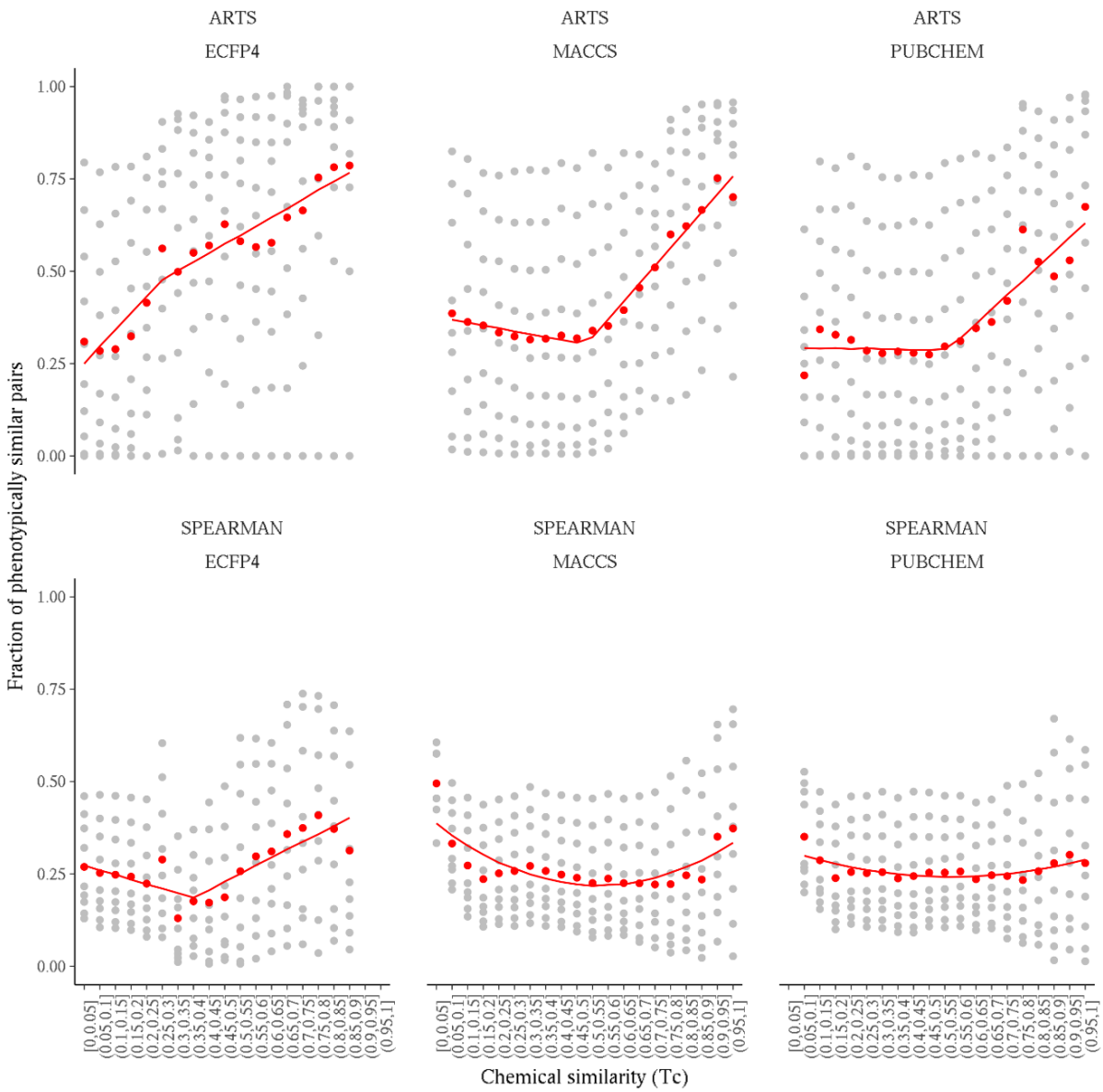

Figure 9. Curve averaging and Bayesian modelling for the ChEMBL dataset. Each panel represents a combination of fingerprint and phenotypic similarity metric. In each panel, the grey curves represent the fraction of phenotypically similar pairs with similarity above a specific threshold (**Figure 7**) for increasing chemical similarity. The red curve is the point-by-point average of those curves in each panel. The Bayesian model with the best fit to the average curve is also overlaid.

When the quadratic model was selected, very slight increases of phenotypic similarity with chemical similarity can be observed after the breakpoint is reached, especially in the case of PubChem (**Figure 9**). This suggested that the Pearson correlation coefficient is an inadequate metric to model phenotypic similarity for cytotoxicity readouts such as those present in the ChEMBL dataset.

When estimating the breakpoints, ECFP4 had lower Tc values than the other fingerprints. As a matter of fact, the breakpoint was observed for Tc = (0.25,0.3] with the ARTS metric, and Tc = (0.35,0.4] with the Spearman correlation coefficient (**Table 6**). Conversely, the breakpoints estimated with the MACCS fingerprint was Tc = (0.5,0.55] for both the ARTS and Spearman correlation coefficient (**Table 6**). For PubChem, the breakpoint was Tc = (0.5,0.55] for ARTS and Tc = (0.6,0.65] for the Spearman correlation coefficient (**Table 6**). This indicated that ECFP4 allowed to observe the molecular similarity principle for a higher range of chemical similarity than the other two fingerprints employed in the analysis. However, the fact that the breakpoint was observed for a higher Tc, around 0.5, agreed not only with Tiikkainen et al. who employed cytotoxicity readouts,[130] but also with Shah et al. who used readouts similar to those used in this study.[131] Hence ECFP4 displayed an atypical behaviour compared to the literature in which phenotypic screening datasets were used.

The estimate for the molecular similarity principle, when the phenotypic similarity was measured with the ARTS metric was always modelled by a piecewise linear model for all three fingerprints and is hence estimated using the second slope of those fits. For ECFP4, this was estimated as a 2.5% increase of the fraction of phenotypically similar compounds per 5% increase in chemical similarity (**Table 6**). For MACCS keys and PubChem, this was estimated as 4.8% and 3.9% respectively per 5% increase in chemical similarity (**Table 6**).

If the Spearman correlation coefficient was used, different trends were obtained. For ECFP4, the piecewise linear model had the best fit, and the molecular similarity principle was 2.1%

increase in the fraction of phenotypically similar compounds per 5% increase in chemical similarity (**Table 6**). For MACCS and PubChem, since the model was quadratic, the molecular similarity principle was estimated as the average slope from the breakpoint to the maximal chemical similarity bin. These estimates were respectively 1.6% and 0.6% increase in the fraction of phenotypically similar compounds per 5% increase in chemical similarity.

Overall, the fraction of phenotypically similar compounds increased by approximately 1% per 5% increase in chemical similarity in the BioMAP dataset, whereas it increased by approximately 3% per 5% increase in chemical similarity in the ChEMBL dataset. Despite these small values, the fact that both estimates are positive showed that the molecular similarity principle is valid in phenotypic screening datasets since this means that more phenotypically similar compounds can be expected with higher structural similarity. This also shows how much phenotypically similar pairs can be expected for a given Tc value depending on the fingerprint. This is of importance for analyses such as virtual screening and quantitative structure-activity relationship (QSAR) studies, which rely on thresholds of structural similarities to obtain sets of compounds with desired properties.

In addition, this study showed that mostly linear or piecewise linear trends were obtained when ECFP4 was used (**Figure 8** and **Figure 9**). Combined with the lower estimated breakpoints when ECFP4 was used, suggested that this fingerprint was more appropriate for molecular similarity principle studies in phenotypic screens compared to the other two fingerprints, since the molecular similarity principle can then be observed for a larger chemical similarity range.

	ARTS			SPEARMAN		
	selected model	estimate	95% confidence interval	selected model	estimate	95% confidence interval
ECFP4	piecewise linear	b = (0.25,0.3]	(0,0.55]	piecewise linear	b = (0.35,0.4]	(0.2,0.55]
		β'_1 = 0.046	[0.025,0.064]		β_1 = -0.013	[-0.023, -0.001]
		β'_2 = -0.021	[-0.044,0.004]		β'_2 = 0.034	[0.017,0.05]
		$\beta'_1+\beta'_2$ = 0.025			$\beta'_1+\beta'_2$ = 0.021	
MACCS	piecewise linear	b = (0.5,0.55]	(0.45,0.55]	quadratic	Vertex = (0.5,0.55]	
		β_1 = -0.008	[-0.013, -0.003]		β_1 = -0.036	[-0.048, -0.022]
		β'_2 = 0.056	[0.048,0.065]		β_2 = 0.0016	[0.0009,0.0021]
		$\beta'_1+\beta'_2$ = 0.048				
PUBCHEM	piecewise linear	b = (0.5,0.55]	(0.4,0.7]	quadratic	Vertex = (0.6,0.65]	
		β'_1 = -0.001	[-0.01,0.009]		β_1 = -0.015	[-0.022, -0.007]
		β'_2 = 0.04	[0.022, 0.058]		β_2 = 0.0006	[0.0003, 0.001]
		$\beta'_1+\beta'_2$ = 0.039				

Table 6. Key estimates and associated 95% confidence interval for the models fitted on the averaged curves for the ChEMBL dataset (displayed in **Figure 9**). Only estimates related to the increase of phenotypic similarity with increasing chemical similarity are reported. If the selected model (*i.e.* the model with the best fit to the average curve) is the linear model, then the slope (β) is reported. If the selected model is quadratic, then the initial slope (β_1) is reported as well as the increase in slope for each 5% increase in chemical similarity (β_2). The x-coordinate of the vertex (*i.e.* the breakpoint where the curve changes direction) is also reported. Finally, if the model is piecewise linear, then the estimate of the breakpoint (b), the estimate of the first slope (β'_1), the estimate of the difference in two slopes (β'_2) and the estimate of the second slope ($\beta'_1 + \beta'_2$) are reported.

3.3.3 Neighbourhood enhancement indicated that ECFP4 are better descriptors for similarity analyses involving phenotypic screening data

In order to explain these differences observed with ECFP4 compared to the other two descriptors, the concept of neighbourhood behaviour defined in Patterson *et al*.[59] was employed. This concept implies that molecular descriptors differ in how small structural changes impact the change in biological activity, and this is quantified with the enhancement ratio. The enhancement ratio quantified the number of compound pairs which have a correlated biological and chemical similarity, compared to the number of pairs with anticorrelated values. Therefore, ideal descriptors would have a high enhancement ratio close to 2, which would then translate into a predictable structure-activity relationship. Enhancement ratio were calculated per assay and per fingerprint. They were then averaged per fingerprint.

In the BioMAP dataset, the neighbourhood enhancements were overall very high with an averaged enhancement ratio of 1.97 for ECFP4, 1.94 for MACCS keys and 1.91 for PubChem fingerprints (**Figure 10**). This showed that all three descriptors were appropriate to capture structural similarity to phenotype similarity relationships in the context of the BioMAP dataset. With the ChEMBL dataset, however, only ECFP4 achieved a high averaged neighbourhood enhancement of 1.93, while MACCS keys and PubChem fingerprints had averaged neighbourhood enhancement values of 1.64 and 1.66, respectively (**Figure 10**).

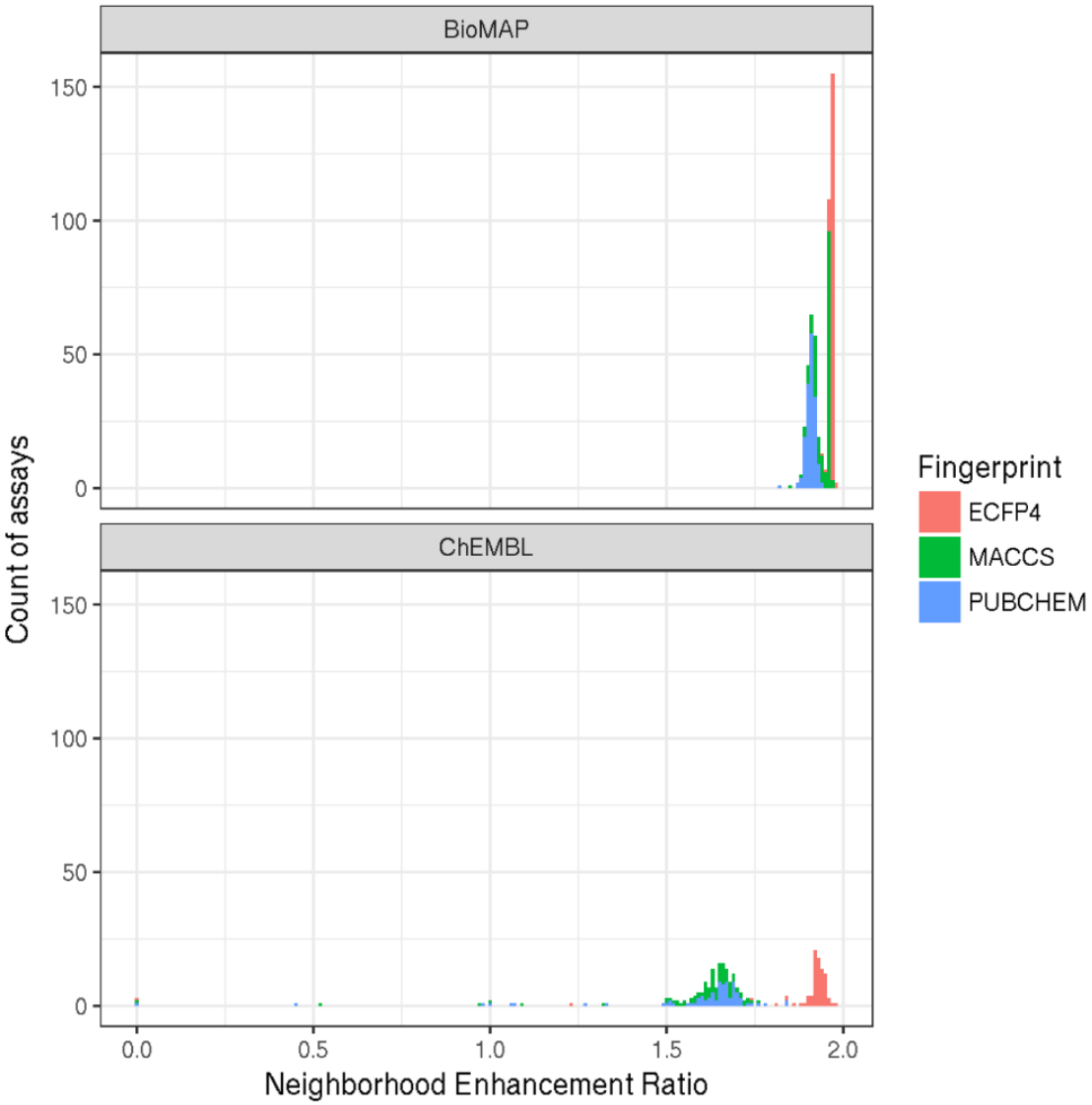

Figure 10. Neighbourhood enhancement ratio distributions for each combination of fingerprint and dataset. The distributions were derived from neighbourhood enhancement ratios calculated over all assays and grouped by fingerprint and dataset. Neighbourhood enhancement ratios calculated with ECFP4 are always higher on average than for the other two fingerprints. Also, neighbourhood enhancement ratios were generally lower in the case of the ChEMBL dataset.

The difference in ratios between the descriptors implied not only that the neighbourhood behaviour observed with ECFP4 fingerprints was consistent across datasets, but also that

ECFP4 is able to capture bioactivity differences better than the other two descriptors in chemical similarity space since the neighbourhood enhancement ratio was higher in both datasets. Compared to the neighbourhood enhancement values in Patterson *et al.*,[59] the values obtained in this analysis were much higher. This implied that the fingerprints used in this thesis are now more advanced in terms of their neighbourhood behaviour compared to the ones they used. It may also imply that it is easier to obtain neighbourhood behaviour with the current phenotypic datasets compared to the datasets used in their study.

In the ChEMBL dataset, even though the neighbourhood enhancement ratios of the fingerprints were high for most cell lines, the fingerprints did not exhibit the neighbourhood behaviour for certain cell lines in the ChEMBL dataset. In the following, it is explained why the neighbourhood behaviour cannot be obtained for these specific cell lines. The cell lines CEM-SS, HaCaT, and SK-N-SH in the CHEMBL datasets, displayed neighbourhood enhancements that were not statistically better than random for all three descriptors, and those had very few data points compared to the other cell lines (3, 9 and 63 compound pairs respectively) and this impacted the statistical significance of the enhancement ratio computations. On the other hand, the enhancement ratio was not significantly higher than random for the cell line Bel-7402 only in the case of the PubChem dataset. This prompted the examination of scatterplots comparing pairwise phenotypic and chemical similarities for this cell line in order to gain insights into how the neighbourhood enhancement ratio varied with the choice of descriptor.

Hence, a comparison of the scatter plots for Bel-7402 revealed further differences between the three descriptors, regarding compound pairs with higher structural similarity (**Figure 11**). Indeed, with ECFP4, very few compound pairs had a high structural similarity, and those stretched between $0.3 < Tc < 1$ and generally had low pairwise phenotypic differences in the Bel-7402 assay, whereas in the case of MACCS and PubChem, a second group of highly structurally similar compound pairs was formed, some with relatively high pairwise phenotypic

differences. In the case of PubChem, there was an even higher density of such high structurally similar pairs, which reduced the density of points in the lower left part of the plot. Because the density became lower in this area of the plot, the Chi-squared test employed to assess the statistical plausibility of the enhancement ratio failed. For five cell-lines, the neighbourhood enhancement was statistically significant for ECFP4 alone: BGC-823, HEp-2, KB3-1, MT4 and MX1. Comparably to Bel-7402, much higher densities in highly structurally similar compound pairs were observed in the case of MACCS and PubChem than for ECFP4 with these cell lines (data not shown).

Hence the neighbourhood enhancement ratios were useful at quantifying how well fingerprints capture the relationship between phenotypic similarity and chemical similarity. ECFP4 had a consistently higher neighbourhood enhancement ratio compared to MACCS keys and PubChem fingerprints across all readouts/cell lines in both datasets. The reason behind this observation is that they yielded less stringent Tc similarity scores compared to ECFP4 descriptors resulting in higher densities of points in the upper right end of the chemical similarity/phenotypic similarity plot. This is perhaps due to the higher resolution of ECFP4 fingerprints, which yield longer binary vectors than the other two fingerprints.

Hence ECFP4 fingerprints was a better descriptor for cheminformatics analyses involving chemical similarity in the context of phenotypic screens such as virtual screening. This is consistent with previous studies in which ECFP4 outperformed MACCS keys in virtual screening applications[41,117] and in drug-target interaction prediction studies.[141]

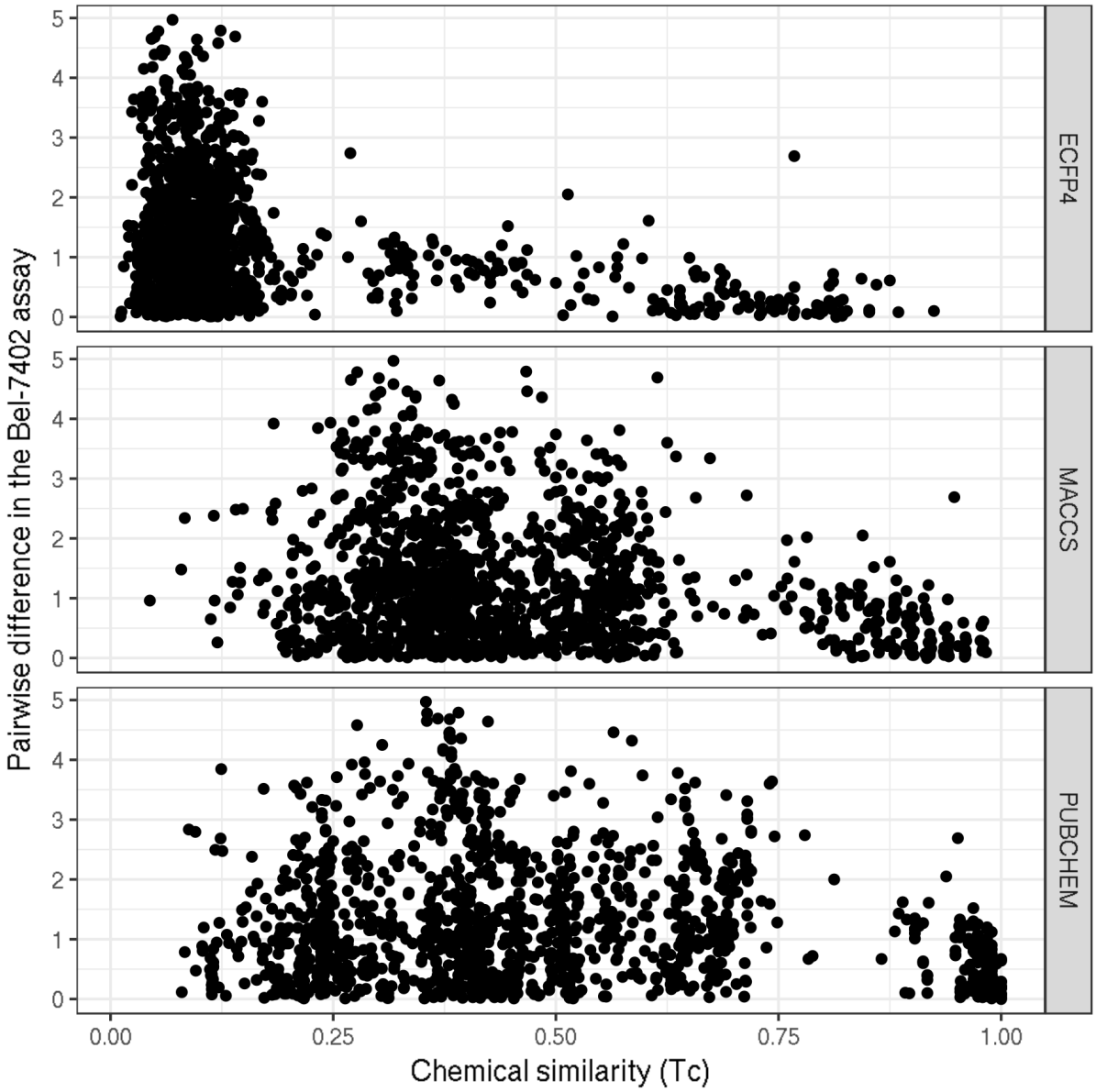

Figure 11. Pairwise differences between compound activities in the Bel-7402 assay compared to chemical similarity with all three fingerprints. A higher density of points was observed for low chemical similarity in ECFP4 whereas the density is more spread out for the other fingerprints. This explained why ECFP4 exhibited the neighbourhood property compared to the other two fingerprints.

3.4 Conclusion

This study investigated the correlation between chemical similarity and similarity of phenotypic profiles comprising about 80 diverse cytotoxicity and biomarker readouts. The molecular similarity principle was found to be valid when using phenotypic readouts and this study quantified the correlation between chemical similarity and phenotypic similarity. Indeed, the fraction of compounds that are phenotypically similar increased by an average of 1% and 3% per 5% increase in Tc chemical similarity in the BioMAP and ChEMBL datasets respectively. In the BioMAP dataset, the fraction of phenotypically similar compounds increased linearly with chemical similarity when ECFP4 was employed, whereas the trend was parabolic or piecewise linear for MACCS and PubChem fingerprints. In the ChEMBL dataset, the piecewise linear model predominated. When the trend was quadratic or piecewise linear, the breakpoint was estimated to happen for a Tc between 0.4 and 0.6 on average, but the breakpoint was on average lower for ECFP4, as this happened for a Tc around 0.3.

This led to comparing fingerprints as to their effects on differentiating the phenotypic profiles of compounds based on structural similarity, and it was found that ECFP4 consistently yielded higher averaged neighbourhood enhancement ratios across readouts and datasets (1.97 and 1.93 for the BioMAP and ChEMBL dataset respectively). MACCS and PubChem produced relatively high enhancement ratios in the BioMAP dataset (1.94 and 1.91 respectively) but yieleded lower enhancement ratios in the cytotoxicity-based ChEMBL dataset (1.64 and 1.66 respectively). This showed that ECFP4 performed better in tasks involving chemical similarity with application to phenotypic screens. In conclusion, the findings of these studies showed that the molecular similarity principle used in target-based screens is also relevant to phenotypic screening datasets.

Chapter 4 Comparative study of the mechanism of action hypotheses obtained in the NCATS dataset using experimental bioactivity versus *in silico* bioactivity

4.1 Introduction

Phenotypic assays are coming back as a screening method to discover new chemical entities. However, the MoA of compounds active in this type of assays, *i.e.* the actual protein target(s) by which the compounds elicit their activity in the screening assay remain to be uncovered. Deconvolution methods involve assays aimed at identifying candidate protein targets for these active compounds. But difficulties in deconvoluting the MoA of active compounds experimentally hinder the discovery of new compounds through phenotypic screening since *in vitro* deconvolution methods are time-consuming and expensive. Therefore, MoA hypotheses

that need to be confirmed experimentally must be prioritised with the help of computational methods.

As a result, computational analyses using the wealth of experimental data in gene expression databases[64] or bioactivity databases[33,65,112] were employed to prioritise such hypotheses to be tested experimentally. Other methods employ machine learning models applied to a 2D representation of the compounds to predict the MoA based on large bioactivity training sets.[83,108,110] So far, *in silico* deconvolution studies either used experimental target activity datasets or *in silico* predictions thereof to deconvolute the effects of compounds in phenotypic screens. To our knowledge, no studies compared the effect of using one or the other data type on target deconvolution. Consequently, it is presently not known whether the MoA hypotheses generated from predicted bioactivity can replace or even complement experimental bioactivities when deconvoluting compound activity in phenotypic screening datasets.

The aim of the present chapter is therefore to compare MoA hypotheses obtained from experimental bioactivity datasets against MoA hypotheses obtained from purely *in silico* target predictions (**Figure 12**). Compounds from the National Centre for Advancing Translational Sciences (NCATS) pharmaceutical collection (NPC)[28] were employed, and experimental data for those compounds were extracted from Drugmatrix. In parallel, targets were predicted via the target prediction workflow described in Chapter 2 (**Figure 12**). MoA hypotheses obtained from experimental bioactivities were then compared to MoA hypotheses obtained from predictions. Additional steps were taken to ensure that the comparison between those two datasets was unbiased *i.e.* the compounds in the target prediction dataset were not part of the experimental/Drugmatrix dataset or in ChEMBL which is a larger repository of drug-like compounds.

Supervised SOMs (sSOM) were employed to narrow the number of MoA hypotheses generated by the two approaches described above. Even though SOMs were successfully employed to generate target predictions,[89,142] in this chapter, the use of sSOM was motivated by two different reasons: 1. Since the bioactivity datasets are sparse, sSOMs have been shown to yield good clustering performance where training sets are sparse such as the Drugmatrix and target prediction dataset;[143] 2. sSOM also allowed to remove noisy target-phenotypic associations and to reduce the number of MoA hypotheses being made per phenotypic activity, by analysing the weights of each target associated to certain clusters of compounds with a specific phenotypic activity cluster (**Figure 12**). Indeed, it is often the case that compounds are active in different phenotypic assays, and the relationship between these assays is complex. sSOM allows modelling these relationships and to isolate phenotypic activities based on the most relevant targets. Targets in the compound clusters obtained from the experimental bioactivity dataset were then compared to those associated with the target prediction clusters both quantitatively using the GO framework[144], and qualitatively with literature analysis of the individually selected targets (**Figure 12**).

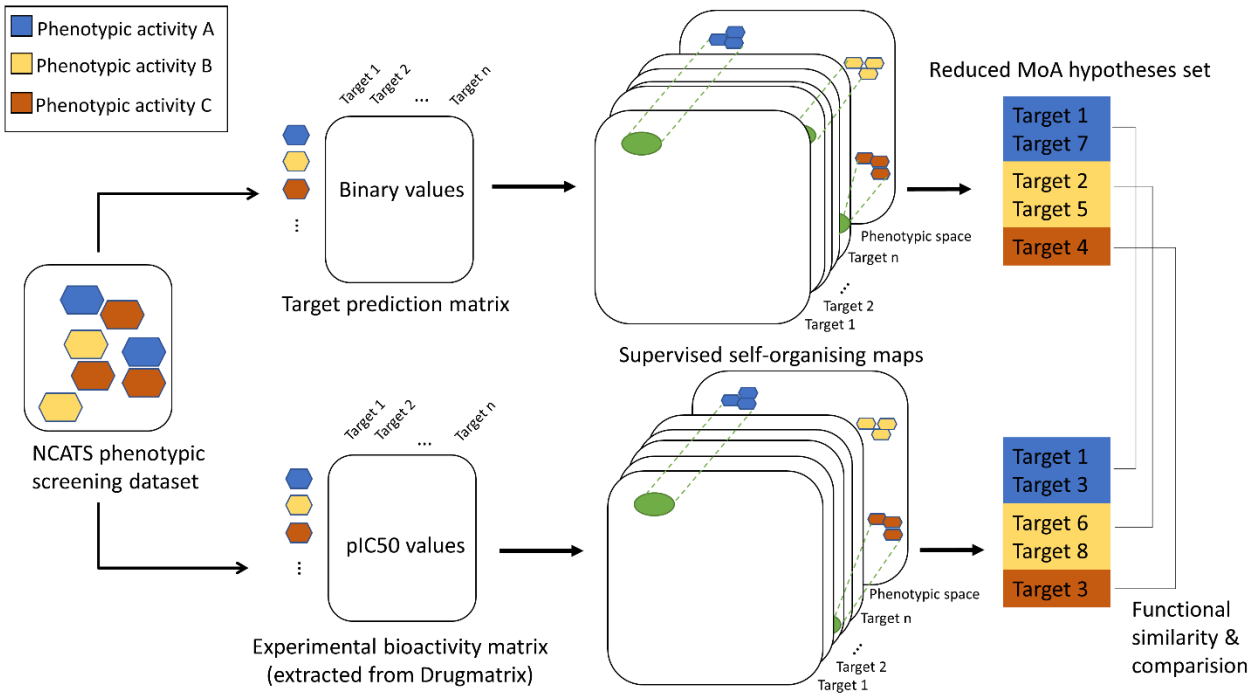

Figure 12. Overview of the workflow employed in this chapter to separate phenotypic activity clusters based on relevant targets. This effectively reduces the target space to relevant targets for each phenotypic space. Ultimately selected targets obtained from target prediction and the use of experimental bioactivity data alone are compared with functional similarity and also with literature analysis.

4.2 Materials and methods

4.2.1 NCATS phenotypic compound (NPC) library and generation of the phenotypic outcomes matrix

The dataset originated from a collaboration of the Open Innovation Drug Discovery (OIDD) between the NCATS and Eli Lilly, which was aimed at exploring the effects of a large collection of drugs in cell-based and various *in vitro* disease models. The compound and assay metadata CSV files, along with the assay experiment results CSV file, were downloaded from https://ncats.nih.gov/expertise/preclinical/pd2. The library (as of 21/11/2017) comprised 2,511 drugs which were measured at 4 doses (0.2, 2, 10 and 20 μM depending on the type of assay *i.e.* screening, preliminary or confirmatory) in 47 assays spanning across the 5 phenotypic annotations described in **Table 7**.[28] Compounds were also screened in an additional 5 assays for their effect on the G2/M cell-cycle phases. Compounds exhibiting an effect on both cell-cycle and on one of the endpoints measured by the OIDD (e.g. anti-angiogenesis) may only have a secondary effect on this endpoint. Since the aim was to generate a phenotypic readout matrix, 5 assays measuring on-target activities as measured by the kinase panels were not considered in the analysis.

Phenotypic annotation	Model	Description
Anti-angiogenesis	Inhibition of angiogenesis in oncology	Assays measuring endothelial tube formation (presence of CD31/PECAM-1 marker) and cell nuclei number by imaging of co-cultures of human clonal endothelial colony forming cells (ECFCs) and adipose-derived stem cells (ADSCs) treated with vascular endothelial growth factor (VEGF). Compounds are labelled as active if the decrease in tube area is > 40% at either 2 μM or 10 μM.
Insulin secretion	Diabetes	Immunoassay to quantify insulin secretion from the INS-1E cell line using AlphaLISA kit[145] (immunoassay similar to ELISA). Compounds yielding a 20% increase in insulin secretion at either 2 μM or 10 μM were considered active.
GLP-Secretion	Diabetes	Quantification of Glucagon-Like Peptide 1 (GLP-1) in supernatants using AlphaLISA kit[145] in media containing NCI-H716 cells and STC-1 cells. Compounds yielding a 20% increase in insulin secretion at either 2 μM or 10 μM were considered active.
Kras/Wnt SL	Selective cytotoxicity of colorectal cancer	Selective cytotoxicity to 7 colon cancer cell lines bearing combinations of mutations in KRAS APC, PI3K and BRAF genes against wild-type under conditions that mimic tumour metastasis using non-adherent and non-proliferating cells; one assay tests for the modulation of the Wnt pathway. Compounds were screened in the cytotoxicity assays at 0.2, 2 and 20 μM yielding dose-response curves. Compounds were considered active with an IC50 < 2 μM.
Wnt pathway	Osteoporosis	β-catenin translocation and alkaline phosphatase (ALP) activity using fluorescence in multilineage potential C2C12 cell line. Compounds are labelled as active if they yield an increased nuclear β-catenin > 40% at either 2 μM or 10 μM
Cell cycle G2/M	Cell-cycle arrest	Fluorescence assays aimed at determining DNA content and condensation as well as levels of Cyclin B within HeLa cells; this assay is used to discriminate compounds which arrest cell-cycle e.g. compounds active in both this assay and in an anti-angiogenesis assay is a weak potentiator of anti-angiogenesis since its effect may be due to cytostaticity or cytotoxicity. Compounds with IC50 < 20 μM in this assay was deemed cytotoxic.

Table 7. Phenotypic annotations in the NCATS dataset[28], their target biological endpoint, a description of experimental measurements and the number of assays measuring them.

A binary readout matrix was created based on the <OUTCOME> variable of the "assay experiment results" CSV file. More specifically, a compound was considered active in an assay if it was active at any of the concentrations measured for this assay. For each assay, **Table 7** specifies what constitutes activity and the threshold at which the <OUTCOME> variable is annotated with “active” (“inactive” by default). This yielded a matrix of 2,511 compounds by 27 assays. Assays were then grouped by their respective project which meant that if a compound is active in at least one of the assay, then it will be active for one of the 6 readouts in the dataset (**Table 7**). Hence the final matrix comprised 2,511 compounds by 6 readouts.

4.2.2 Drugmatrix and generation of the on-target activity matrix

The pharmacology assay dataset can be downloaded from https://www.niehs.nih.gov/results/dbsearch/index.html. This dataset comprises 1,291 compounds with IC50 measured across 132 targets. Even though only 69% of the assays in this dataset are based on the expression of human targets (the remainder originated from rat, mouse, guinea-pig, rabbit, bovine, and bacteria), this dataset was selected because it is complete and such a characteristic is rare for pharmaceutical datasets.[146] Inactive values (denoted by NA in the dataset) were replaced by the highest IC50 observed *i.e.* the lowest potency (140 μM in this case).

In order to match the compounds in the Drugmatrix dataset to those in the NCATS dataset, the InChIKeys of the compounds in both datasets were extracted through the PubChem idexchange service (https://pubchem.ncbi.nlm.nih.gov/idexchange/). For the Drugmatrix compounds, this

was performed using the chemical names as input and extracting the InChIKeys with Operator Type set to "Same CID". For the NCATS compounds, the PubChem CIDs were already provided and were directly used as input to extract the InChIKeys via the same setting. Then the compounds in the Drugmatrix dataset were matched to the compounds in the NCATS phenotypic outcome matrix through inchikey matching. In total 363 Drugmatrix compounds were found in the NCATS dataset.

4.2.3 Target prediction matrix

The target prediction workflow described in Chapter 2 was used to predict the targets for the NCATS compounds using $Zscore > 1$, since the aim of this analysis was exploratory and therefore stringent predictions were not required. This generated a matrix of 2,511 compounds by 777 putative human targets.

Since the aim of this chapter was to compare target profiles from clustering results based on experimental against predicted activity values, steps were taken to ensure that the comparison was not biased. Therefore, the target prediction compound dataset was created with the condition that none of the compounds in this dataset was in the Drugmatrix dataset. Moreover, the target prediction algorithms described in Chapter 2 are based on data from the ChEMBL database. The Drugmatrix compounds are comprised of drugs mainly and this dataset is now a subset of ChEMBL. As a matter of fact, an overlap of 875 compounds was calculated between

the ChEMBL and Drugmatrix *i.e.* 68% of the Drugmatrix compounds are found in ChEMBL22.

Therefore, a subset of NCATS compounds that were not in ChEMBL was used for the comparison of clustering with Drugmatrix and found 157 such compounds. Hence in this chapter, a subset of the target prediction matrix was created and which contained 157 compounds by 777 targets.

4.2.4 Supervised clustering with supervised self-organising maps

A supervised clustering approach was used to reduce the target space leading to the generation of a fewer, more specific MoA hypotheses behind the phenotypic annotations of the compounds in the NCATS dataset. sSOMs[147–149] were implemented with the R package Kohonen (v. 3.0.4).[150] sSOMs are a type of neural networks, and as such, are very suited to the clustering of sparse data.[143] However, the subset of the target prediction matrix comprised of the 157 NCATS compounds which did not overlap with ChEMBL was used for reasons mentioned in the previous section.

In the sSOM algorithm, individual instances in the data (compounds in this case) are assigned to a node on a map of a pre-defined size, and each node can be regarded as a cluster of instances. sSOMs rely on the principle that nodes that are neighbours are likely to be close in input space as well (bioactivity in this case). The algorithm to build the sSOMs starts by initialising the nodes with random weights. Then, at each iteration, a random data point is considered and the

closest node in Euclidean distance in the sSOM is chosen to be the best matching unit (BMU). The BMU and all the nodes within a radius σ of the BMU are then updated with the difference between the current weights of the nodes and the variable values of the data point. In other words, the weight w_i of a node i is updated by:

$$w_i = w_i + h(BMU, i)\left(\alpha(x - w_i) + (1 - \alpha)(y - w_i)\right) \quad (17)$$

where x and y represent the independent and dependant set of variables of the data point respectively (*i.e.* the values of that point on the target and phenotypic space respectively), α weights the influence x has on the update of w_i compared to y, and $h(BMU, i)$ is the Gaussian neighbourhood kernel that defines the distance between node i and the BMU and is defined by:

$$h(BMU, i) = \exp\left(-\frac{\| l_{BMU} - l_i \|^2}{2\sigma}\right) \quad (18)$$

where l_{BMU} and l_i are the location on the sSOM for the BMU and node i respectively.

Both α and the radius of the neighbourhood σ are linearly decreased at each iteration. The effect of decreasing α reduces the influence of the target space on the update of the nodes, and at the last iterations, both the target and phenotypic space have the same importance on the update of the nodes. On the other hand, the effect of reducing σ will lead to nodes that are more and more specialised at the later iterations of the algorithm, since less nodes will be updated around the BMU after each iteration. The linear decay of those two parameters ensure the convergence of the algorithm.

sSOMs were constructed in this work using the *xyf()* function of the Kohonen package to cluster the bioactivity space (x) in relation to the phenotypic annotations of the compounds (y). The input parameters were selected based on a qualitative assessment of the phenotypic annotation map, *i.e.* values which yielded the most specialised clusters possible (containing one or at most two phenotypic annotations). The resulting parameters were: 1. The biggest map was used for each dataset (*i.e.* 19x19 nodes for the Drugmatrix dataset and 11x11 nodes for the target prediction dataset), 2. rlen=10,000 iterations and 3. alpha=c(1 to 0.01). The alpha parameter represents a vector of initial α and ending α, which the sSOM algorithm will explore and which is decreased at each iteration as mentioned above. The bioactivity data was scaled prior to applying the sSOM algorithm to give equal weights to all targets.

The output of the algorithm consisted in a set of different weight mappings of the nodes in the sSOM, each representing either a target (or predicted target in the case of the target prediction dataset) or a phenotypic annotation. Each target mapping can be represented using a gradient of colour on top of the nodes of the sSOM. Because the individual node intensities of certain maps were much higher than those of the other maps, each map was normalised separately to have intensities between 0 and 1. Since there are only 6 phenotypic annotations, these were represented as individual pie charts for each node, showing the relative weights of each phenotypic endpoint for each node.

Group of nodes were selected on the phenotypic annotation map, for which the majority of nodes within these groups corresponded to one phenotypic annotation. For each of these groups, targets were deemed associated with the phenotypic annotation of the group if at least one node in the group had a value higher than the quantile 95% of the distribution of the intensities measured on all maps.

4.2.5 Quantitative comparison of targets associated with experimental clusters vs in silico clusters using Gene Ontology-based functional similarity

To quantitatively assess the difference of the clustering obtained by the experimental and predicted bioactivity matrices, comparable identifiers were obtained for the protein targets identified in the two datasets. Drugmatrix target names were matched to their corresponding UniProt accession Identifiers. For the protein complexes in the Drugmatrix dataset, the protein complexes were matched to the UniProt accession of all their subunits instead. UniProt Identifiers of the predicted targets were already available. Then, the UniProt Identifiers of both datasets were mapped to their corresponding Entrez gene identifiers.

The R package GOSim (v. 1.16.0)[151] was utilised to compare the gene identifiers associated with the experimental bioactivity sSOM clusters, with the gene identifiers associated with the predicted target sSOM clusters via the functional annotation of the genes. More specifically, the *getGeneSim()* function was employed with parameters similarity='max' and similarityTerm='Resnik'. The first parameter defines how the functional similarity between any two genes is computed. Indeed, since some of the pairwise comparisons were large, the maximum Go Term similarity observed between two genes (instead of more computationally expensive techniques) was used to quantify their functional similarity.

The other parameter relates to the way the similarity between the GO Term themselves is computed. In this case, the Resnik distance was used, which was also the most straightforward

to interpret and is based on the information content of all the lowest common ancestors common to the two GO Terms in the GO hierarchy:

$$Sim(GO1, GO2) = max_{GO\ \epsilon\ Lca}(nIC_{GO}) \tag{19}$$

where $GO1$ and $GO2$ are the two Go Terms for which the similarity is computed, Lca is the set of all lowest common ancestors of $GO1$ and $GO2$, nIC_{GO} is the normalised information content of GO Term GO. The lowest common ancestors for two GO-Terms $GO1$ and $GO2$, are the lowest GO-Terms in the tree that has both $GO1$ and $GO2$ as descendants.

The normalised information content (nIC) is defined as:

$$nIC_{GO} = \frac{IC_{GO}}{max_{GT\ \epsilon\ Pall}(IC_{GT})} = \frac{-\log\,(p_{GO})}{max_{GT\ \epsilon\ Pall}(-\log\,(p_{GT}))} \tag{20}$$

where IC_{GO} is the information content of GO Term GO *i.e.* the negative logarithm of the probability p_{GO} of observing functional annotation GO across the human genome, and $Pall$ is the set of all GO Terms in the GO hierarchy. All the information content values were already pre-computed in the GOSim package.

At this stage, a pairwise similarity matrix for the genes in the two lists was generated, with the Drugmatrix cluster genes in the rows and the genes from the target prediction clusters in columns and each cell representing a nIC. To assess the strength of the functional similarity of gene pairs based on the nIC value, the nIC was represented as a function of the probability of occurrence of any GO Term p . It was found that the nIC decreased with the frequency of the

GO-Term for which it is computed (**Figure 13**) and hence, the higher the nIC the rarest the GO-Term. In fact, a value of $nIC \geq 0.4$ corresponded with frequencies close to 0 and therefore represented rare GO Terms with high information content.

Since the functional similarity of two genes is computed as the maximal nIC observed for the GO Term of the lowest common ancestors of the GO Terms of these genes, gene pairs with $nIC \geq 0.4$ share rare functional annotations with high information content, reflecting specific functions. Therefore, gene pairs with a similarity of ≥ 0.4 were deemed very similar in this analysis.

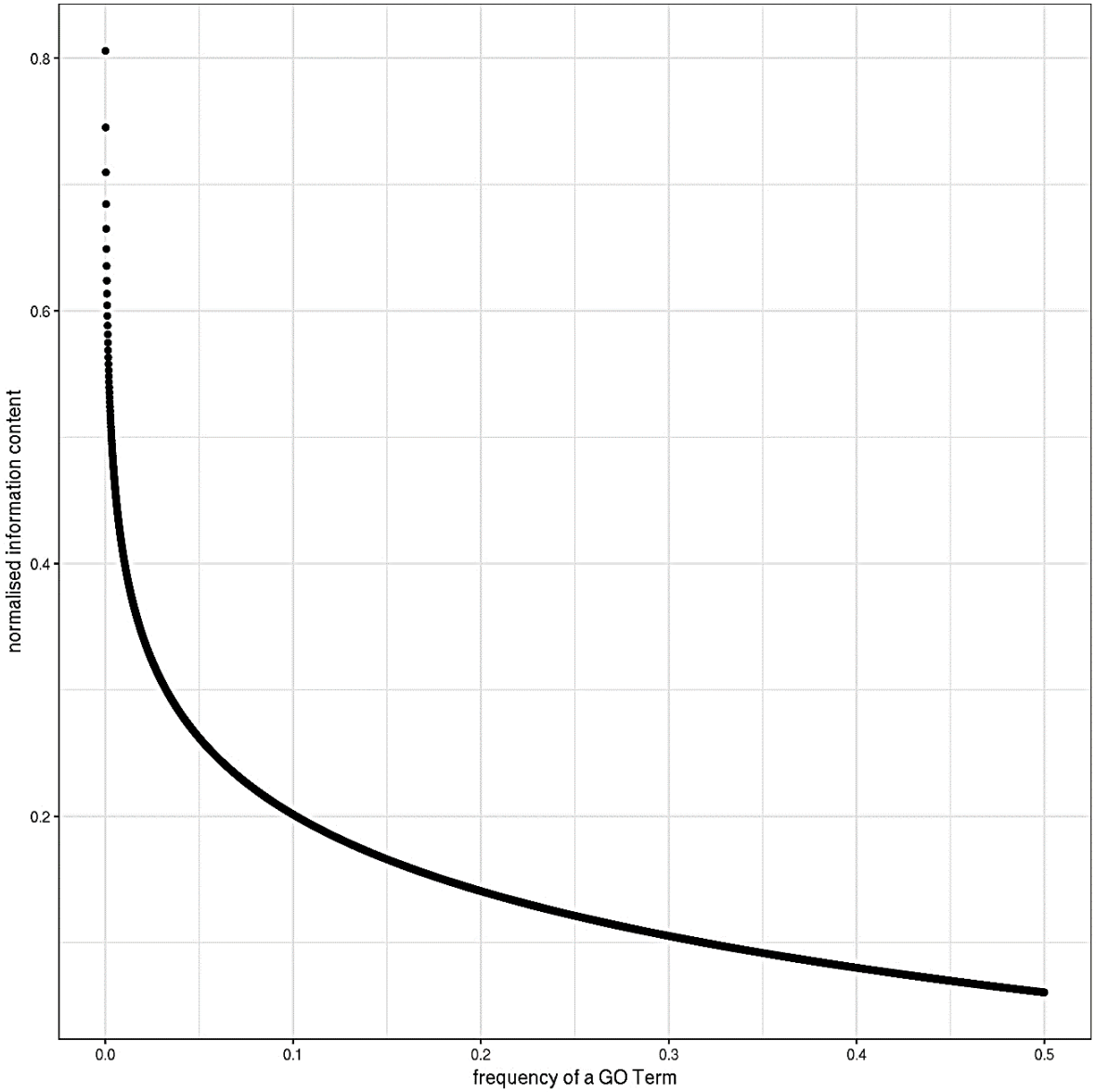

Figure 13. Normalised information content (nIC) in function of the frequency of a GO Term (p). Values of $nIC \geq 0.4$ corresponded with frequencies close to 0.

4.3 Results and discussion

4.3.1 Analysis of the relationship between the phenotypic annotations and selection of phenotypic neighbourhoods

A supervised clustering approach with sSOMs was employed to eliminate noisy target-phenotypic endpoints associations and find groups of compounds based on relevant MoA for the phenotypic endpoints in the NCATS dataset. Two sSOMs were employed: one based on experimental bioactivity data from the Drugmatrix dataset (**Figure 14**) and one based on *in silico* generated bioactivity data (**Figure 15**).

Interestingly, sSOMs allowed to analyse the relationship between the various phenotypic annotations of the compounds in the NCATS library. In the sSOM of the Drugmatrix dataset (**Figure 14**), nodes with anti-angiogenesis clustered close to the nodes annotated with Glucagon-like peptide 1 (GLP-1) secretion, which agreed with the literature since GLP-1 promotes angiogenesis *in vitro*,[152] and therefore compounds modulating the secretion of GLP-1 might also promote anti-angiogenesis. On the right side of the map (and to a lesser extent; the bottom left side as well), nodes with dual phenotypic annotations, namely compounds active on both the Kras/Wnt module and the anti-angiogenesis module. These may be comprised of anti-carcinogenic compounds that specifically target anti-angiogenesis in cancer, a major area of research in oncology.[153,154] Compounds in this group included Floxuridine, an anti-carcinogenic agent used in the treatment of colorectal cancer,[155] and angiogenic

modulators Enalapril[156] and alpha-lipoic acids.[157] This suggested that the modulation of angiogenesis may be a plausible MoA by which Floxuridine exerted its effect on colon cancer cells. Three nodes were located at the bottom-left corner of the map and were annotated with the combination of anti-angiogenesis, GLP-1 secretion, and Kras/Wnt modulation, which reflect anti-carcinogenic compounds, that induce anti-angiogenic effects by modulation of GLP-1 secretion. Compounds in this cluster comprised aminergic GPCRs inhibitors Levosulpiride and Domperidone. Aminergic GPCRs are not only associated with colorectal cancer but are also known to regulate pancreatic islet function and insulin secretion.[158] Dopaminergic receptors, in particular, subtype D2 which is a target of both Levosulpiride and Domperidone, are linked to anti-angiogenic effects in tumour mice models.[159]

On the other hand, nodes annotated with insulin-secretion were located closer to compounds annotated with Kras/Wnt synthetic lethality (**Figure 14**). This is also in agreement with the literature, where it was shown that Wnt signalling potentially influences insulin sensitivity and mediate glucose homeostasis,[160] and therefore compounds mediating Wnt signalling may also modulate insulin secretion.

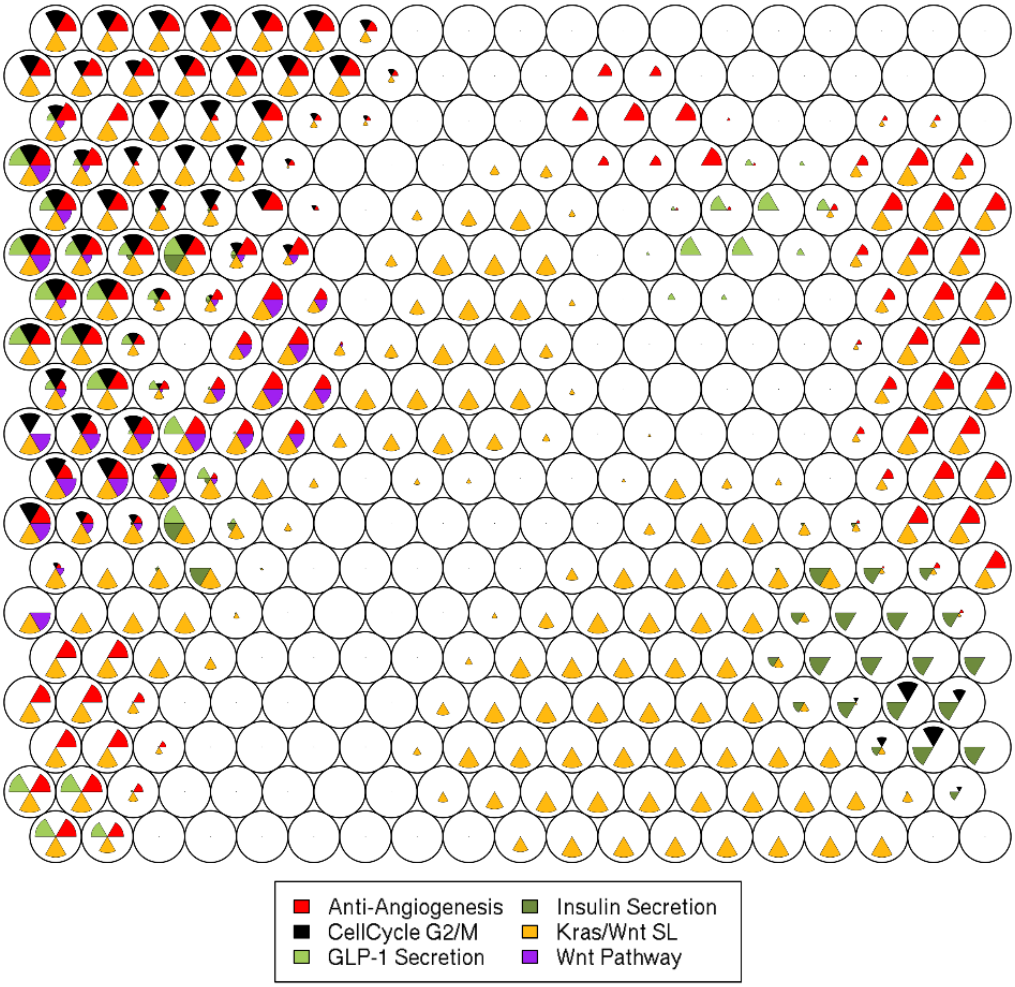

Figure 14. Supervised self-organising maps for the Drugmatrix dataset. Five groups with single phenotypic annotations (one colour predominates in a node) could be detected: 2 for the Kras/Wnt module (yellow nodes middle and middle-bottom of the map), 1 for the angiogenesis module (red nodes at the top of the map in the middle), 1 for the GLP-1 secretion (light green nodes at the top of the map) module and 1 for the insulin secretion module (dark green nodes at bottom-right of the map). Nodes annotated with Wnt pathway (purple nodes) and nodes affecting the cell-cycle (black nodes at the upper-left corner of the map) are annotated with many different phenotypic annotations.

Interestingly, the top-left corner of the map is mostly comprised of nodes with multiple phenotypic annotations. These included nodes with compounds affecting the cell-cycle or the Wnt pathway. These nodes might, therefore, be comprised of compounds whose modulation of those endpoints is mainly through cell-cycle arrest, which suggested that anti-angiogenic effects or the modulation of the GLP-1 secretion were a side-effect rather than the primary activity for the compounds in these clusters. This shows the practical utility of using sSOMs, as this allowed to study the MoA of compounds with one or multiple phenotypic annotations, separately from those compounds which effect can be attributed to unspecific action via cell-cycle arrest.

The sSOM on the target prediction dataset was smaller and had overall fewer nodes with single annotations (**Figure 15**). This is a result of the size of the target prediction dataset which was built so that there is no overlap in terms of compounds with either ChEMBL or Drugmatrix. This stringent condition led to a target prediction dataset of only 157 compounds. Despite the smaller size of the sSOM, many associations observed in the Durgmatrix sSOM were also retrieved in this sSOM as well. There was a large group of nodes with the single Kras/Wnt annotation, and many of the nodes displayed the dual annotation anti-angiogenesis and Kras/Wnt which was previously discussed (top-right corner of the sSOM, **Figure 15**).

As observed on the Drugmatrix sSOM, nodes annotated with the cell-cycle were also associated with other annotations but never on their own on the target prediction sSOM (nodes with black colours, **Figure 15**), which reflected modulators of the cell cycle which modulate other endpoints as a side effect. However, there was no node annotated with the Wnt pathway, since this annotation was not present in the target prediction dataset.

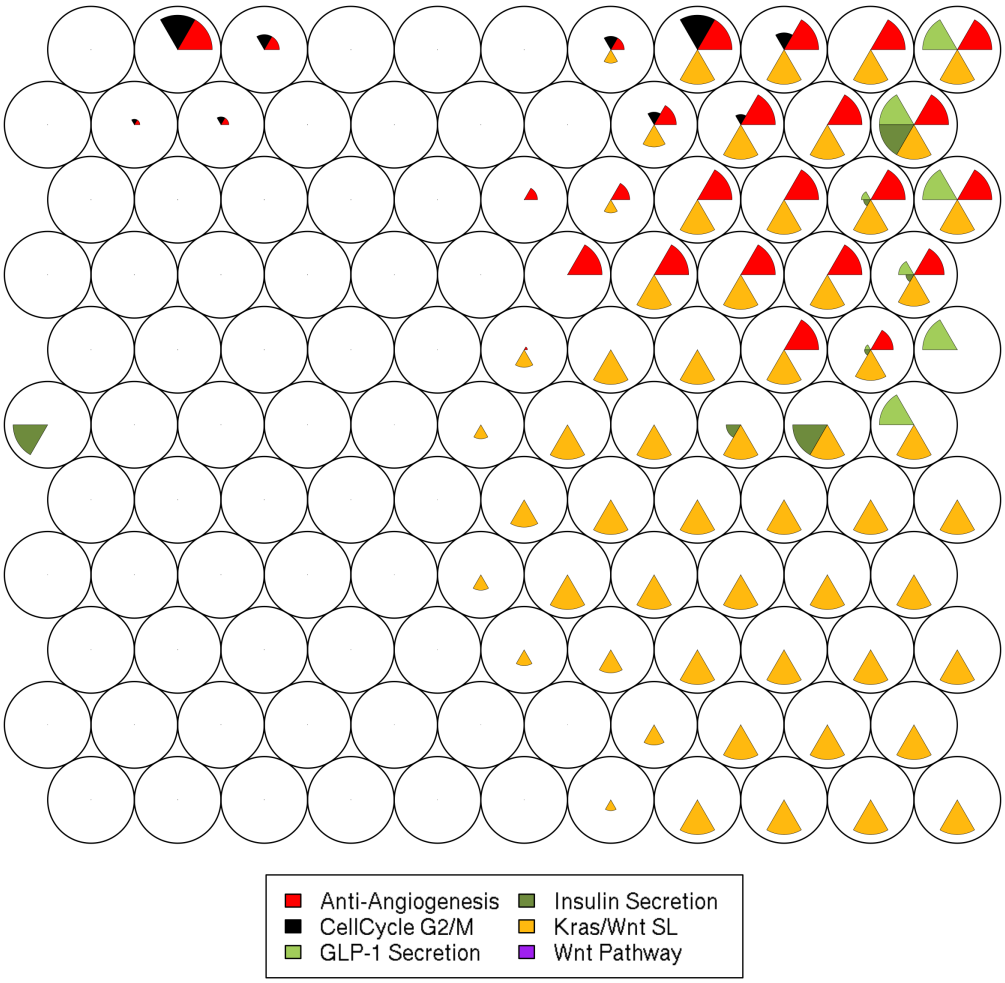

Figure 15. Supervised self-organising map for the target prediction dataset. Groups of nodes with dual annotations were also found on this map as well, including the group of anti-angiogenesis and Kras/Wnt nodes (red and yellow nodes at the top-right corner of the map), and the group of GLP-1 secretion and Kras/Wnt nodes (light green and yellow nodes on the right side of the map). Again, nodes affecting the cell-cycle (black nodes at the top of the map) are also associated with other phenotypic annotations.

4.3.2 Analysis of the functional similarity of targets associated with phenotypic neighbourhoods in both sSOMs

Next, node neighbourhoods where only one phenotypic annotation dominated were identified and target weights associated with these neighbourhoods were calculated to yield MoA hypotheses for both sSOMs/datasets. This step effectively filters the most relevant MoA hypotheses for each phenotypic endpoint (see material and methods for more details) and for each sSOM/dataset.

The functional similarity of the genes encoding the targets was used to compare targets obtained by the two sSOMs, for each phenotypic endpoint. This was performed using the maximal normalised information content (nIC) value which quantifies the rarity of the GO-Term annotations of all the lowest common ancestors of two GO Terms (see Materials and methods section and **Figure 13**). The higher the nIC, the rarer and more specific the GO-Term annotation, and therefore the highest the similarity between the genes encoding the targets of interest.

The functional similarity distribution of nIC values were calculated for Anti-angiogenesis, Insulin and GLP-1 secretion forming the Diabetes module, and the Kras/Wnt synthetic lethal phenotypic annotations. The target prediction dataset did not have compounds annotated with the Wnt pathway, and the Cell-cycle G2/M annotation was always associated with several other phenotypic annotations on both sSOM and hence a target deconvolution of this effect would be difficult. Therefore, the focus of the chapter will be on Anti-angiogenesis, Diabetes

(comprising both GLP-1 and Insulin secretion nodes), and Kras/Wnt synthetic lethal phenotypic annotations.

The nIC distribution (or GO term similarity) for the Anti-angiogenesis, Diabetes and Kras/Wnt annotations were distributed around an average of 0.37, 0.41 and 0.47 respectively (**Figure 16**). This suggested that targets identified in the Drugmatrix sSOM shared high functional similarities with the proteins from the target prediction sSOM, since a $nIC \geq 0.4$ indicated common functional annotations with high information content, as previously mentioned (see Methods section and **Figure 13**).

The distributions for the Anti-angiogenesis and Diabetes modules were approximately normally distributed with a small skew towards higher nIC values (**Figure 16**). This indicated that the targets associated with these phenotypic annotations in the Drugmatrix sSOM shared very high functional similarities with targets associated with the corresponding phenotypic annotations in the target prediction sSOM. On the other hand, the distribution for the Kras/Wnt synthetic lethal module displayed many target pairs with no functional similarity (**Figure 16**). This is a result of the large number of nodes associated with Kras/Wnt compared to other annotations. The higher the number of nodes, the higher the number of compounds and therefore the higher the number of targets associated with the annotation. Furthermore, Kras/Wnt assays are less specialised than the other endpoints, since it measures selective cytotoxicity of colorectal cancer for which multiple MoAs exist. Nonetheless, the Kras/Wnt also displayed the largest nIC average of 0.47, which implied that targets associated with the Kras/Wnt neighbourhoods in the Drugmatrix sSOM shared high functional similarities to the corresponding targets in the Kras/Wnt.

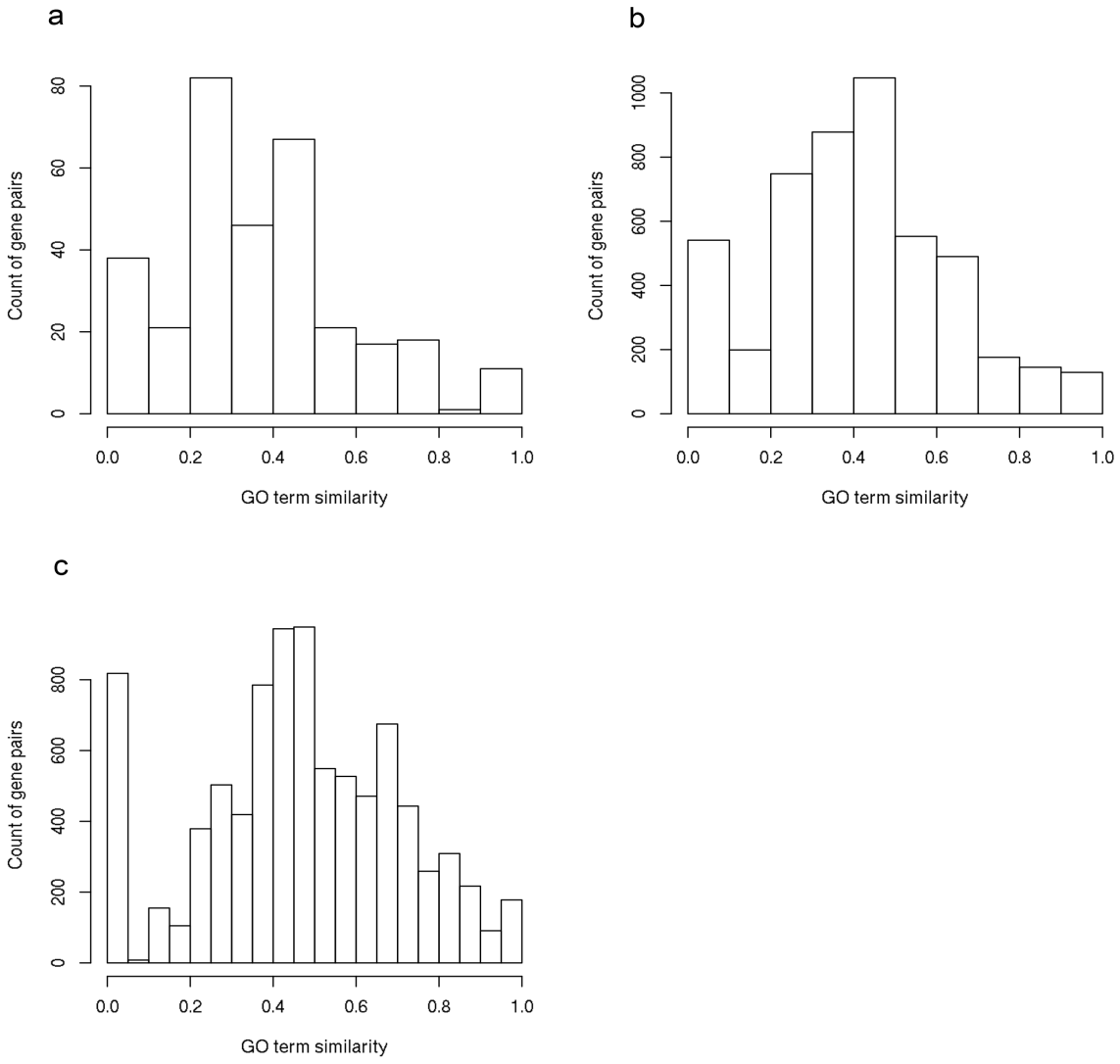

Figure 16. Distribution of pairwise gene functional similarity (nIC) between the genes of the Drugmatrix-based sSOM and the genes of the target prediction sSOM for a) Anti-angiogenesis node neighbourhoods b) Diabetes which combine both GLP-1 secretion and insulin secretion neighbourhoods c) Kras/Wnt neighbourhoods. The functional similarity distribution being clustered around nIC averages of 0.4, it can be concluded that targets in both datasets shared high functional similarities for all three endpoints.

4.3.3 Comparison of targets associated with Anti-angiogenesis neighbourhoods

The target weights behind the six nodes which only had the Anti-angiogenesis annotation in the Drugmatrix sSOM were extracted. Only two targets were associated with this cluster of nodes (**Table 8**). These two targets were strongly linked to angiogenesis and regulation of vascular networks, namely the Acetylcholinesterase and the GABAA complex. Indeed, Acetylcholinesterase inhibitors were shown to reduce angiogenesis in mice[161] and GABAA signalling activates cell-proliferation and angiogenesis in mice models as well.[162]

The 19 proteins identified by the target prediction sSOM for Anti-angiogenesis were different than the ones identified in the Drugmatrix sSOM. However, most of these targets belonged to three protein families that are strongly linked to angiogenesis in the literature namely adenosine receptors, histone deacetylases (HDACs) and the protein kinase C (PKC) family (**Table 8**). Adenosine intervenes through adenosine receptors under hypoxic conditions to form new blood vessels and therefore these receptors are studied to develop anti-angiogenic inhibitors.[163] HDACs have been shown to be overexpressed in hypoxic conditions and their inhibition have been shown to reduce anti-angiogenesis *in vivo*.[164] Inhibition of members of the PKC family, in particular, PKC-α, promoted angiogenesis by re-establishing platelet-derived growth factor C (PDGFC) signalling.[165]

The next five predicted targets could not be grouped in any protein family but were also strongly linked to angiogenesis (**Table 8**). Fatty acid synthase inhibitor Orlistat has been shown to have antiangiogenic activity in an ex-vivo model of human angiogenesis.[166] A systems

biology analysis of angiogenesis uncovered six proteins central to this process, including the meningioma expressed antigen 5 (MGEA5).[167] As for the nuclear receptor corepressor 2 (NCOR2), this protein has been shown to interact with HDAC1 to repress apoptosis of endothelial cells and its displacement from Peroxisome proliferator-activated receptor gamma (PPAR-γ) have been shown to promote apoptosis of endothelial cells and subsequent inhibition of angiogenesis.[168] The activation of pregnane X receptor (PXR) by Rifaximin led to inhibition of pro-angiogenic factors such as Hypoxia-inducible factor-1 alpha (HIF-1α) in an intestinal cell-line *in vitro*.[169] Finally, P-selectin overexpression correlated with ischemia-induced angiogenesis and this suggested that this target plays an important role in inflammatory processes that precede ischemia-induced angiogenesis.[170]

The remaining four targets had a weaker link to anti-angiogenesis in the literature (**Table 8**). The Emopamil binding protein (EBP) has been shown to be up-regulated in endothelial cells treated with a combination of vascular endothelial growth factor (VEGF) and Hepatocyte growth factor (HGF),[171] suggesting a connection between VEGF angiogenic activity and EBP. Glucosylceramides have been shown to inhibit angiogenesis through reduction of VEGF and HIF-1α in a mouse xenograft model of cancer cells,[172] which might suggest an involvement of glucosylceramidase beta in the observed anti-angiogenic effect. The melanocortin-3 receptor (MC3R) was not directly implicated with angiogenesis, but another member of this protein family melanocortin-4 receptor (MC4R) has been linked to angiogenic balance in rat models of obesity.[173] The solute carrier family 5 member 1 (A1) was down-regulated in Vhl tumour suppressor gene-depleted mice, where this deletion caused angiogenesis.[174]

In summary, even though both sSOMs identified different targets, these targets were equally relevant to angiogenesis and this shows the complementarity target prediction offers to experimental bioactivity datasets in deconvoluting phenotypic screening signals.

sSOM	Target group	Gene Symbol	Target	Reference
Drugmatrix		*ACHE*	acetylcholinesterase	[161]
		---	GABAA	[162]
Target prediction	Adenosine receptors	*ADORA1*	adenosine A1 receptor	[163]
		ADORA2A	adenosine A2a receptor	[163]
		ADORA3	adenosine A3 receptor	[163]
	Histone deacetylases	*HDAC10*	histone deacetylase 10	[164]
		HDAC4	histone deacetylase 4	[164]
		HDAC5	histone deacetylase 5	[164]
		HDAC7	histone deacetylase 7	[164]
		HDAC9	histone deacetylase 9	[164]
	Protein kinase C	*PRKCB*	protein kinase C beta	[165]
		PRKCG	protein kinase C gamma	[165]
		FASN	fatty acid synthase	[166]
		MGEA5	meningioma expressed antigen 5	[167]
		NCOR2	nuclear receptor corepressor 2	[168]
		NR1I2	pregnane X receptor	[169]
		SELP	selectin P	[170]
		EBP	emopamil binding protein	[171]
		GBA	glucosylceramidase beta	[172]
		MC3R	melanocortin 3 receptor	[173]
		SLC5A1	solute carrier family 5 member 1	[174]

Table 8. Targets associated with anti-angiogenesis neighbourhoods in each sSOM along with the number of nodes for which the target is associated. The reference for each target refers to any link this target has with angiogenesis or anti-angiogenesis in the literature.

4.3.4 Comparison of targets associated with Diabetes neighbourhoods using GLP-1 secretion and Insulin secretion nodes

The nodes annotated with GLP-1 or insulin secretion in the Drugmatrix sSOM were associated with 19 targets, mainly aminergic GPCRs and the transporters of their ligands (**Table 9**). Their role in insulin secretion was already discussed above (see section Comparison of the phenotypic annotations and selection of phenotypic neighbourhoods).[158] The calcium channel type L, potassium and sodium channels were also associated with the insulin secretion nodes (**Table 9**), which is in agreement with the literature. Voltage-gated channels, principally calcium and sodium channels are required for exocytosis and glucose-induced insulin secretion.[175] Two targets were not related to insulin secretion or diabetes, namely CYP450 subtype 2D6 (CYP2D6) and the Sigma-1 receptor.

In the target prediction sSOM, 58 targets were associated with either GLP-1 and/or Insulin secretion. Among these targets, 9 targets belonged to the renin-angiotensin pathway[176] which if inhibited, had been shown to reduce the incidence of type 2 diabetes (**Table 10**).[177,178] Another 24 targets were relevant to diabetes and/or glucose homeostasis (**Table 10**). The remaining targets did not show a clear link to diabetes.

Target family	Gene Symbol	Target	Reference
Amingergic GPCRs & transporters	*ADRA2A*	adrenoreceptor alpha 2A	[158]
	ADRB1	adrenoreceptor beta 1	[158]
	ADRB2	adrenoreceptor beta 2	[158]
	ADRB3	adrenoreceptor beta3	[158]
	SLC6A2	adrenergic norepinephrine transporter	[158]
	DRD2	dopamine receptor D2L	[158, 159]
	DRD3	dopamine receptor D3	[158]
	SLC6A3	dopamine transporter	[158]
	HRH1	histamine receptor H1	[158]
	HTR1A	serotonin receptor 1A	[158]
	HTR2A	serotonin receptor 2A	[158]
	HTR2B	serotonin receptor 2B	[158]
	HTR2C	serotonin receptor 2C	[175]
Ion channels	(complex)	calcium channel type L (benzothiazepine site)	[175]
	(complex)	calcium channel type L (phenylalkylamine site)	[175]
	(complex)	ATP-sensitive potassium channel	[175]
	(complex)	sodium channel (site 2)	[175]

Table 9. Targets associated with Insulin secretion and/or GLP-1 secretion neighbourhoods in the Drugmatrix sSOM. The reference for each target refers to any link this target has with diabetes and its complications in the literature. CYP2D6 and the Sigma-1 receptor were also found by this analysis but not reported in this table.

The target prediction dataset found very relevant targets for the compounds annotated with the Diabetes endpoints in the NCATS dataset, which were not redundant and instead complemented those obtained with the Drugmatrix dataset. Thus, this shows once again, how target prediction complements experimental bioactivities in the deconvolution of the signals in the phenotypic assays testing for modulation of factors relevant to diabetes.

Target group	Gene Symbol	Target	Reference
Renin-angiotensin pathway	*ACE*	angiotensin I converting enzyme	176–178
	ACE2	angiotensin I converting enzyme 2	176–178
	ANPEP	alanyl aminopeptidase, membrane	176–178
	CMA1	chymase 1	176–178
	CTSG	cathepsin G	176–178
	*CTSB**	cathepsin B*	176–178
	KLK1	kallikrein 1	176–178
	*KLKB1**	kallikrein B1*	176–178
	MME	membrane metalloendopeptidase	176–178
Targets involved in other aspects of glucose homeostasis or diabetes	*ACVRL1*	activin A receptor-like type 1	179
	CALCA	calcitonin related polypeptide alpha	180
	CAMK1D	calcium/calmodulin dependent protein kinase ID	181
	CAMK2A	calcium/calmodulin dependent protein kinase IIA	181
	CCKBR	cholecystokinin B receptor	182
	ECE1	endothelin converting enzyme 1	183
	F11	coagulation factor XI	184
	GRK5	G protein-coupled receptor kinase 5	185
	GRM1	glutamate metabotropic receptor 1	186
	HSP90AA1	heat shock protein 90 alpha family class A member 1	187
	ITGA4	integrin subunit alpha 4	188
	ITGA5	integrin subunit alpha 5	188
	ITGB1	integrin subunit beta 1	188
	ITGB5	integrin subunit beta 5	188
	ITGB7	integrin subunit beta 7	188
	MC1R	melanocortin 1 receptor	189
	MC3R	melanocortin 3 receptor	189
	MC4R	melanocortin 4 receptor	189
	MC5R	melanocortin 5 receptor	189, 190
	PIN1	Rotamase Pin1	191
	PRKCB	protein kinase C beta	192
	PTGER2	prostaglandin E receptor 2	193
	PTGES	prostaglandin E synthase	194
	XIAP	X-linked inhibitor of apoptosis	195

Table 10. Targets associated with Insulin secretion and/or GLP-1 secretion neighbourhoods in the target prediction sSOM. Here only 33 out of the 58 targets associated with these nodes were reported in this table and relevant to diabetes according to the literature. Gene symbols marked with a star do not belong to the pathway but are closely associated to one of the members of the pathway.

4.3.5 Comparison of targets associated with Kras/Wnt synthetic lethal neighbourhoods

The Kras/Wnt module of the NCATS dataset measures the selective cytotoxicity against colon cancer cell lines bearing different mutations of Kras and other genes. Mutations in Kras and these other genes activate the Mitogen-Activated Protein Kinase (MAPK) pathway.[196,197] Since the Kras/Wnt phenotypic activity was predominant in the NCATS dataset, many targets were associated with this phenotypic annotation in both the Drugmatrix and target prediction sSOM (**Table 11**). In total 118 targets were associated with the Kras/Wnt nodes in the target prediction sSOM, and 56 such targets were extracted from the Drugmatrix sSOM. Although only 8 targets were in common between the two sSOMs, many target families in the Drugmatrix sSOM overlapped with those of the target prediction sSOM.

The majority of these targets were the aminergic GPCRs and the transporters for their ligand which have long been investigated in colon cancer.[198–211] Despite the fact that 6 aminergic GPCR receptors (including all members of the muscarinic receptor sub-family) were found by both the Drugmatrix and the target prediction sSOM, each sSOM uncovered additional members of the various GPCR receptors subfamilies (**Table 11**).

Moreover, many members from the MAPK pathway were found from the target prediction sSOM whereas the Drugmatrix sSOM selected two central members of this pathway namely, MAPK1 and EGFR (**Table 11**). As mentioned previously, MAPK is activated as a result of mutated KRAS in colon cancer cells,[196,197] and therefore targeting members of this pathway constitutes a plausible MoA for the compounds active in the Kras/Wnt module. The fact that target prediction uncovered different but equally important members of the MAPK pathway,

further stresses the complementarity that target prediction hypotheses offer to hypotheses generated from experimental data from Drugmatrix, in order to deconvolute compound activity in the NCATS phenotypic assays. Among the other targets which were found in both the Drugmatrix and target prediction, ion channels correlated with Kras/Wnt annotated nodes (**Table 11**). Both the calcium and human Ether-à-go-go-Related Gene (hERG) channels were directly linked to colon cancer.[212–214] However, the sodium channel encoded by the *SCN9A* gene only shared an indirect link to colorectal cancer, since another member of the sodium channel SCN5A was found to regulate the expression of genes involved in the invasiveness of colorectal cancer.[215] In addition, members of the Cytochrome P (CYP) family were found for which polymorphisms[216,217] and differential expression was observed in colorectal cancer patients.[218]

In essence, the intersection of the sSOMs of both Drugmatrix and target prediction of the Kras/Wnt nodes in the sSOM was mainly comprised of aminergic GPCR receptors and members of the MAPK signalling pathway. Both datasets agreed as to the importance of these two families for colorectal cancer cytotoxicity, as well as targeting CYPs and ion channels (**Table 11**) Since nodes with compounds active in the cell cycle assay were excluded from the analysis, the MoA hypotheses generated for the Kras/Wnt module are representative of cytotoxicity against colorectal cancer cells that do not rely on cell-cycle arrest. However, since the aforementioned target families are part of major signalling pathways, a careful selection of these MoA hypotheses is required to avoid adverse events.

Target group	sSOM	Gene Symbol	Target	References
Aminergic GPCRs and transporters	Both	*ADRA2C*	adrenoceptor alpha 2C	198
		CHRM1	cholinergic receptor muscarinic 1	199,204
		CHRM2	cholinergic receptor muscarinic 2	199,204
		CHRM3	cholinergic receptor muscarinic 3	199,204,205
		CHRM4	cholinergic receptor muscarinic 4	199,204
		CHRM5	cholinergic receptor muscarinic 5	199,204
	Drugmatrix	*ADRA1B*	adrenoceptor alpha 1B	206
		ADRA2A	adrenoceptor alpha 2A	207
		ADRA2B	adrenoceptor alpha 2B	208
		ADRB2	adrenoreceptor beta 2	209
		ADRB3	adrenoreceptor beta 3	210
		DRD3	dopamine receptor D3	
		HTR2A	serotonin receptor 2A	211
		HTR2B	serotonin receptor 2B	200
		HTR2C	serotonin receptor 2C	
		HTR6	serotonin receptor 6	201
		HRH2	histamine receptor H2	202
		OPRD1	opioid receptor delta 1	
		OPRK1	opioid receptor kappa 1	
		OPRM1	opioid receptor mu 1	203
		SLC6A2	adrenergic norepinephrine transporter	
		SLC6A3	dopamine transporter	
		SLC6A4	serotonin transporter	
	Target prediction	*HTR1B*	serotonin receptor 1B	219
		HTR1D	serotonin receptor 1D	220
		HTR3A	serotonin receptor 3A	
		HTR5A	serotonin receptor 5A	
		CHRNA3	cholinergic receptor nicotinic	221

		CHRNA7	alpha 3 subunit cholinergic receptor nicotinic alpha 7 subunit	221
		SSTR3	somatostatin receptor 3	
		SSTR4	somatostatin receptor 4	222
MAPK signaling	Drugmatrix	*EGFR*	EGF receptor	223
		MAPK1	mitogen-activated protein kinase 1	224,225
	Target prediction	*FER*	FER tyrosine kinase	226
		PTPN1	protein tyrosine phosphatase, non-receptor type 1	227
		ROS1	ROS proto-oncogene 1, receptor tyrosine kinase	228
		MAP2K1	mitogen-activated protein kinase kinase 1	196,197
		MAP2K2	mitogen-activated protein kinase kinase 2	196,197
		MAP3K5	mitogen-activated protein kinase kinase kinase 5	196,197
		MAPK12	mitogen-activated protein kinase 12	196,197
		MAPK3	mitogen-activated protein kinase 3	196,197
		MAPKAPK5	mitogen-activated protein kinase-activated protein kinase 5	196,197
		CAMK1D	calcium/calmodulin-dependent protein kinase ID	
		CAMK2A	calcium/calmodulin dependent protein kinase II alpha	229
		CAMK2G	calcium/calmodulin dependent protein kinase II gamma	229
		FYN	FYN proto-oncogene, Src family tyrosine kinase	226,230,231
		FGFR1	fibroblast growth factor receptor 1	
		FGFR2	fibroblast growth factor receptor 2	232

		FGFR3	fibroblast growth factor receptor 3	[233]
		GRIA3	glutamate ionotropic receptor AMPA type subunit 3	
		GRIN1	glutamate ionotropic receptor NMDA type subunit 1	
		GRIN2A	glutamate ionotropic receptor NMDA type subunit 2A	[234]
		GRIN2B	glutamate ionotropic receptor NMDA type subunit 2B	[234]
		MARK2	microtubule affinity-regulating kinase 2	[235]
		MARK3	microtubule affinity-regulating kinase 3	
		RET	RET proto-oncogene	[236]
Ion channels	Both	*SCN9A*	sodium voltage-gated channel alpha subunit 9	[215]
	Drugmatrix	*(complex)*	calcium channel type 1 (benzothiazepine site)	[212,213]
		(complex)	calcium channel type 1 (dihydropyridine site)	[212,213]
		(complex)	calcium channel type 1 (phenylalkylamine site)	[212,213]
	Target Prediction	*KCNH2*	potassium voltage-gated channel subfamily H member 2	[214]
Cytochrome P450	Both	*CYP1A2*	cytochrome P450 1A2	[218]
	Drugmatrix	*CYP2C9*	cytochrome P450 2C9	[217]
		CYP2D6	cytochrome P450 family 2 subfamily D member 6	[216]
		CYP2C19	cytochrome P450 2C19	

Table 11. Target families associated with the Kras/Wnt module in both sSOM. The reference for each target refers to any link this target shares with colorectal cancer in the literature.

4.4 Conclusion

In this work, sSOMs were used on experimental and predicted bioactivity datasets, for the deconvolution of the MoA of compounds with a certain phenotypic annotation, based on supervised clustering analyses on the bioactivity space of these compounds. This was performed to narrow down MoA hypotheses to the most relevant targets associated with the phenotypic activities in the NCATS dataset.

This not only allowed to retrieve compound clusters for 4 out of the 5 phenotypic endpoints in the NCATS dataset but also allowed the analysis of the relationship between phenotypic endpoints themselves where it was confirmed that Anti-angiogenesis, GLP-1 secretion and Kras/Wnt synthetic lethal annotations are closely related, in agreement with the literature.

The functional annotations of the targets obtained with experimental data were compared to the annotations of the targets obtained through prediction. Functional similarity values between the two sSOMs reflected high functional similarities with averages of 0.37, 0.41 and 0.47 for Anti-angiogenesis, Diabetes and Kras/Wnt modules respectively. For Anti-angiogenesis and the phenotypic annotations associated with Diabetes (insulin and GLP-1 secretion), it was found that even though the targets obtained from the two sSOMs did not overlap individually, the targets were still very relevant to these endpoints and complemented each other well. For the Kras/Wnt endpoint, targets belonging to similar protein families (namely aminergic GPCRs, Cytochrome P450s, and ion channels) or to the MAPK pathway overlapped between the two sSOMs. The target prediction retrieved different members of the MAPK pathway than

the ones identified with the Drugmatrix sSOMs confirming that *in silico* predictions can complement the MoA obtained with the Drugmatrix experimental bioactivity dataset.

The analysis presented in this chapter, therefore, demonstrated that target prediction can uncover alternative MoA hypotheses that were not detected using experimental data as was the case with the comparison performed with the Anti-angiogenesis and Diabetes modules. In conclusion, both approaches allowed to retrieve different but relevant MoA hypotheses according to the extensive literature search conducted in this chapter. Hence, it is here suggested that *in silico* bioactivity predictions should be used to complement historical bioactivity datasets for target deconvolution tasks in phenotypic screens.

Chapter 5 Computational studies of the mechanism-of-action of kidney cyst growth reducing compounds

5.1 Introduction

PKD is a hereditary disease in which kidneys are enlarged due to the presence of many small fluid-filled sacs called cysts, leading in some cases to hypertension and kidney failure.[237,238] In fact, PKD is responsible for approximately 10% of the patients with kidney failure, ranking PKD as the fourth most common disease leading to such stage.[239] Autosomal dominant PKD (ADPKD) is caused by mutations in *PKD1* (in 85% of the cases) and *PKD2* genes (the remaining 15%).[237,240] Polycystin-1 (PC1) and polycystin-2 (PC2), the resulting proteins of these genes in the healthy phenotype, may interact to form a protein complex which is not only able to control intracellular calcium concentration, but also to regulate tubular cell proliferation and apoptosis via the interaction with key proteins in several pathways including the Wnt and the mammalian target of Rapamycin (mTOR) signaling pathways.[237,240] In the disease state, this complex is not formed and as a result, this disturbs the signalling of many genes. These molecular changes translate to the cellular level as a perturbation of the arrangement of the

tubular cells constituting the nephrons, and a transformation of the cell function from ion-absorbing to ion-excreting cells. This ultimately leads to the perturbation of intracellular ion concentration and water osmosis towards the forming cyst, eventually resulting in an increase in cyst fluid volume.[240] As with many diseases, PKD is accompanied not only by local ischemia[241–243] but also inflammation and fibrosis.[244]

A number of therapeutic options to reduce cyst growth have been investigated. Since PKD is frequently associated with hypertension,[238] enabling blood regulation control seemed like an interesting MoA to investigate. Unfortunately, angiotensin-converting enzyme (ACE) inhibitors such as Enalapril did not show any improvement in renal function.[245] Instead, many compounds targeting pathways involved in the control of the cell cycle and cellular proliferation have been explored. Glycosphingolipids are molecules which are involved in cell proliferation control, and blocking their intracellular accumulation through the inhibition of the glucosylceramide synthase led to inhibition of cystogenesis in mouse models.[246] PPAR-γ agonists such as pioglitazone and rosiglitazone also showed significant renal cyst inhibition in mice models.[247,248] The MoA of PPAR-γ inhibitors may also inhibit the cystic fibrosis transmembrane conductance regulator (CFTR) expression,[247,249] which has been linked to reducing kidney cyst progression *in vivo*.[250] Concomitant inhibition of the phosphorylation of two important targets in cell proliferation, mTOR and Akt, have been successful in reducing cyst growth in rodent models.[251] Cyclin-dependant kinase (CDK) inhibitor, Roscovitine, was also found to be effective in inhibiting disease progression in mice.[252] The regulation of cell proliferation through the retrieval of intracellular Ca^{2+} levels control is also a MoA commonly investigated in the area. For example, it has been shown that triptolide was able to release Ca^{2+} levels conditionally to the expression of PC2 and limit the proliferation of the disease in *Pkd1*$^{-/-}$ mice.[253] To date, only Tolvaptan (also called Jinarc), a vasopressin receptor antagonist,[254,255]

has been authorized for the treatment of ADPKD in the EU and the UK markets.[256,257] Despite slowing down the progression of the disease, this drug does not completely prevent cyst growth and is associated with strong diuresis and thirst.[255,258]

Since most of the currently MoA hypotheses are associated with potential adverse events or have not shown efficacy in clinical studies yet, more suitable drugs need to be discovered through the generation of novel therapeutic opportunities in PKD drug discovery. This study is concerned with using target prediction models to find under-studied MoA of compounds able to modulate cyst growth. A kidney cyst screening generated by collaborators from the University of Leiden was employed as a starting point of the target prediction models. While previous analyses performed by this group tried to identify blockers of cyst growth using a library of kinase inhibitors,[259] the analysis described in this chapter employs a more comprehensive dataset containing a collection of diverse compounds through the use of the SPECTRUM screening library.[260] Target predictions were integrated with gene expression data as well as literature occurrence to shortlist two MoAs. Furthermore, docking algorithms were employed to further assess the plausibility of two shortlisted target predictions. To the best of our knowledge, this constitutes the first study in which target prediction algorithms are combined with structural bioinformatics approaches to hypothesize the MoA of cyst growth reducing (CGR) compounds in an *in vitro* kidney cyst screening platform.

5.2 Material and Methods

5.2.1 Spectrum library and screening for kidney cyst growth reduction

The experimental procedures described in this section were all performed by Tijmen Booij from the University of Leiden. 2,320 compounds from the Spectrum collection[260] were screened for kidney cyst growth reduction following the procedure described in Booij et al.[259] Mouse inner medullary connecting duct cells (mIMCD3, ATCCCRL-2123) with down-regulated *Pkd1* expression were generated by shRNA mediated knockdown. This was achieved by Lentivirus transfection expressing shRNA which specifically target the *Pkd1* gene in those cells. Reduced *Pkd1* expression was confirmed by quantitative polymerase chain reaction. The cells were then cryopreserved in a 2D culture for 72 hours to allow their recovery.

Prior to transfer in 3D culture and screening of the compounds, the mIMCD3 cells were defrosted in a 37°C water bath and cultured in a medium with 5% CO_2 for another 72 hours. Cells were next washed with PBS and trypsinised. These were then grown on 384 well plates filled with a hydrogel. The gel-cell mix had a final density of 2,175 cells per well. After gel polymerisation and the addition of a cell medium, combinations of Forskolin and the compounds of the Spectrum library were then added at 1 and 10 μM for 3 days. The Forskolin allowed the formation of multicellular cysts, and the compounds of the library were tested for their ability to reduce the growth of these cysts. Cultures were then fixated with formaldehyde and stained with Hoechst 33,258 and Rhodamine-Phalloidin (actin).

The cyst area in each well was measured using a BD Pathway 855 imager. These cyst area measurements were normalised to control:

$$Normalised\ cyst\ area_i = \frac{100\ \times (\ median(cyst\ area\ conrol_+) - cyst\ area_i\ )}{median(cyst\ area\ conrol_+) - \ median(cyst\ area\ conrol_-)} \quad (21)$$

where $cyst\ area\ conrol_+$ represents all replicates of cyst area measurements obtained for the positive control $i.e.$ wells treated with DMSO and normalise to 0% stimulation, and $cyst\ area\ conrol_-$ correspond to all the replicates of cyst area measurements obtained for the negative control $i.e.$ Forskolin-exposed wells without library compounds, which induce large cysts and would hence normalize to 100% stimulation. $cyst\ area_i$ is the raw cyst area measured for compound i.

Hence, for a particular concentration (either 1 or 10 μM), the closer a normalised cyst area measurement gets to 0%, the smaller the cysts get, and the more effective a compound becomes at inhibiting cyst growth (although this can also be due to toxicity, which will be addressed in the next section). Compounds reaching approximately 0% or lower for any concentration were therefore labelled as having a CGR effect. Out of the 2,320 compounds in the Spectrum library, 81 such compounds had a notable CGR effect.

5.2.2 Compound dataset pre-processing and filtering

Compounds containing metal groups and with a molecular weight of more than 900 g/mol were removed using Schrodinger canvas.[261] Those filters combined with duplicate compounds removal yielded a final dataset of 2,279 compounds with 50 compounds having a CGR effect. Salts and fragments were removed from the structures.

Additionally, compounds which CGR effect might potentially be attributed to general cytotoxicity rather than by the result of a specific on-target effect were separated from the remainder of the CGR compounds. For this purpose, conformal machine learning models were used to predict cytotoxicity of kidney cancer cells.[106] More specifically, the HEK293 model (embryonic kidney cells) was used and predictions were made with a confidence of 94%. This threshold was selected as a compromise for keeping the largest amount of CGR compounds while removing most of the known antineoplastic compounds from the dataset. Out of the 16 models available, this model was the only one coming from kidney tissues, and this motivated its use in this analysis.[106] Any compounds with single label predictions for cytotoxicity were considered cytotoxic. All the other cases were considered non-cytotoxic predictions (there was no single label prediction for non-cytotoxicity). In total, 1,240 compounds were not labelled as cytotoxic by the model, and among the 50 CGR compounds, 17 were not labelled as cytotoxic.

5.2.3 Target prediction and statistical association with effect on cyst growth

The workflow previously described in Chapter 2 was employed to predict the targets of the 50 standardised CGR compounds, as well as the 17 CGR compounds which were not labelled as cytotoxic per the cytotoxicity model. Since the purpose of this work is to narrow down a small number of MoA hypotheses to be experimentally tested, the threshold $Zscore > 2$ was implemented to binarize the compound-target predictions.

A metric called Cyst Area Deviation (CAD) was developed in this thesis to rank targets based on the CGR effect of their compounds. This metric prioritized predicted targets which are associated with CGR compounds having the biggest effect on cyst growth compared to the other compounds. It is therefore based on the cyst area variations induced by the compounds driving the prediction.

To compute the CAD, an offset equal to the minimal induced cyst area observed in the dataset was applied to shift the normalised cyst area distribution to positive values, so that the smallest area measurement is at 0. Then, for each predicted target, the normalised cyst area distribution obtained for CGR compounds driving the prediction was compared to the distribution of compounds with no effect on cyst growth and that are driving the prediction of the same target. The difference in medians between the two cyst area distributions was subtracted to yield a CAD for each target:

$$CAD_i = Median_{NO\ EFFECT}(Cyst\ Area)_i - Median_{CGR}(Cyst\ Area)_i \quad (22)$$

where i is the i^{th} target, $Median_{NO\ EFFECT}(Cyst\ Area)_i$ the median of the cyst area distribution for the compounds driving the prediction for the target and had no effect on cyst growth in the screen, and $Median_{CGR}(Cyst\ Area)_i$ is the median of the cyst area distribution for the CGR compounds driving the prediction for the i^{th} target.

To assess the validity of this metrics, the CAD was compared to the "Literature score" from Open Targets[262] where targets corresponding to "polycystic kidney diseases" were extracted, and 26 overlapped with the predicted targets obtained at the previous step of this analysis

5.2.4 Target shortlisting based on CAD values, literature occurrence, and gene expression studies

CAD scores were converted to Z-Score scale and targets with CAD Z-Score above 1 were retained in the analysis. Because the focus of this study was to select less researched MoAs, targets that had at most 2 PubMed referenced associations to PKD were retained in the analysis. This filter was performed using the Comparative Toxicogenomics Database (CTD)[263,264] in which gene-disease associations are ranked by the number of PubMed references in which both the gene and disease co-localised in the abstract, or if a compound that is known to modulate the disease co-localised with the target, or if a compound known to modulate the target co-localised with the disease. These associations were manually collected from PubMed articles by expert curators. Finally, targets were selected if their corresponding genes were present in a list of differentially expressed genes in PKD mice models curated and contributed by Malas et al. (**Appendix C**).[265]

5.2.5 Docking

Docking algorithms were employed to evaluate the shortlisted predicted associations. Hence crystal structures of the targets shortlisted from the previous step were extracted from the Protein Data Bank (PDB) with resolution <2.2Å. For Proteinase-Activated Receptor 1 (PAR-1) and Kallikrein 1 (KLK1), only one such structure could be found (PDB ids: 3VW7 and 1SPJ respectively).

Protein structures were imported into Maestro (release 2015-3) [266] where the protein preparation wizard was used to remove all water molecules, correct the orientation of the amino-acid side-chains, optimise protonation states of residues for PH=7 (using the PROPKA option) and run a restrained minimization with the "convergence of heavy atoms" parameter set to a RMSD of 0.3Å (this allows for the relaxation of the hydrogen bonding network within the protein and with the co-crystallized ligand). Since KLK1 did not have a co-crystallised ligand, the binding sites were inferred via Sitemap 3.6 [267] and the paper accompanying the structure confirmed the approximate location of the active site that was used for docking.

Ligands accompanying the structure (Vorapaxar for 3VW7) and the compounds driving the prediction of the protein to be docked (Podophyllin acetate, Picropodophyllin acetate for PAR-1 and Anthotecol for KLK1) were imported into Maestro's Ligprep 3.5 wizard for which all states corresponding to PH = 7 ± 2 were generated using Epik 3.3 [268] with the default OPLS_2005 force field. Specified chiralities were retained when available where other chiral centres were left free to vary. Only one low energy ring conformation was retained per ligand.

The grid generation and docking of the ligands and compounds of interest were performed with Glide 6.8[269,270] with extra-precision scoring[271] and expanded pose sampling was enabled.

Additionally, docking poses were superimposed to the co-crystallised inhibitor in order to check whether the compound poses occupied the same area as the co-crystallised inhibitor.

5.3 Results and discussion

5.3.1 Cyst Area Deviation ranked targets that are known to be involved in PKD higher than other targets

Targets were predicted for the compounds of the Spectrum dataset using the target prediction workflow described in Chapter 2. For the 33 CGR compounds that were labelled as potential cytotoxic compounds, predictions were obtained for 277 targets, resulting in 642 compound-predicted target pairs. For the reduced set of 17 CGR compounds which were not labelled as cytotoxic, 178 unique targets were predicted with 287 CGR compound-target predicted pairs.

To assess the relevance of the target predictions to PKD, a metric was developed in this thesis to prioritise predicted targets that were associated with compounds having the biggest CGR effect (Cyst Area Deviation, or CAD). No differences were found when comparing the distributions of CAD between the two target prediction models (**Figure 17, left)** which indicated that none of the models predicted targets that were more relevant to PKD than the other.

To assess if the CAD ranked targets according to their relevance to the disease, these rankings were compared to associations scores extracted from Open Targets[262]. For a specific disease of

interest, this database ranks genes according to their association to this disease using several criteria such as the presence of mutations in the gene, known drugs for the targets of those genes and the maximum phase they reached in their development, and the strength of the evidence from animal models. The literature score, which ranks genes according to the number of evidence found in the literature that links them to PKD, was selected because it provided the highest coverage with the predicted targets obtained with this analysis.

In total, 26 predicted targets matched one of the genes in the "Autosomal dominant polycystic kidney disease" gene set from Open Targets literature score. The CAD metric was compared with the literature score from Open Targets (**Figure 17, right**). A correlation of 0.31 was obtained between the CAD and the literature score from Open Targets, meaning that a higher ranking per this metric also provided a higher literature score. There was one clear target outlier which had a high literature score in Open Targets but a low rank per the CAD metric employed in this analysis (**Figure 17**, **right**) and which corresponded to the Nitric oxide synthase 3 (NOS3). The NOS3 was ranked very poorly with the CAD but is currently being investigated in PKD since certain polymorphisms of this gene coding for this target was linked to chronic kidney disease progression in PKD[272], and also for its effect on hypertension aspects of the disease[273]. Since the CAD is based on the compounds that were tested in the screen, it is therefore limited towards ranking proteins which are likely to be targeted by these compounds.

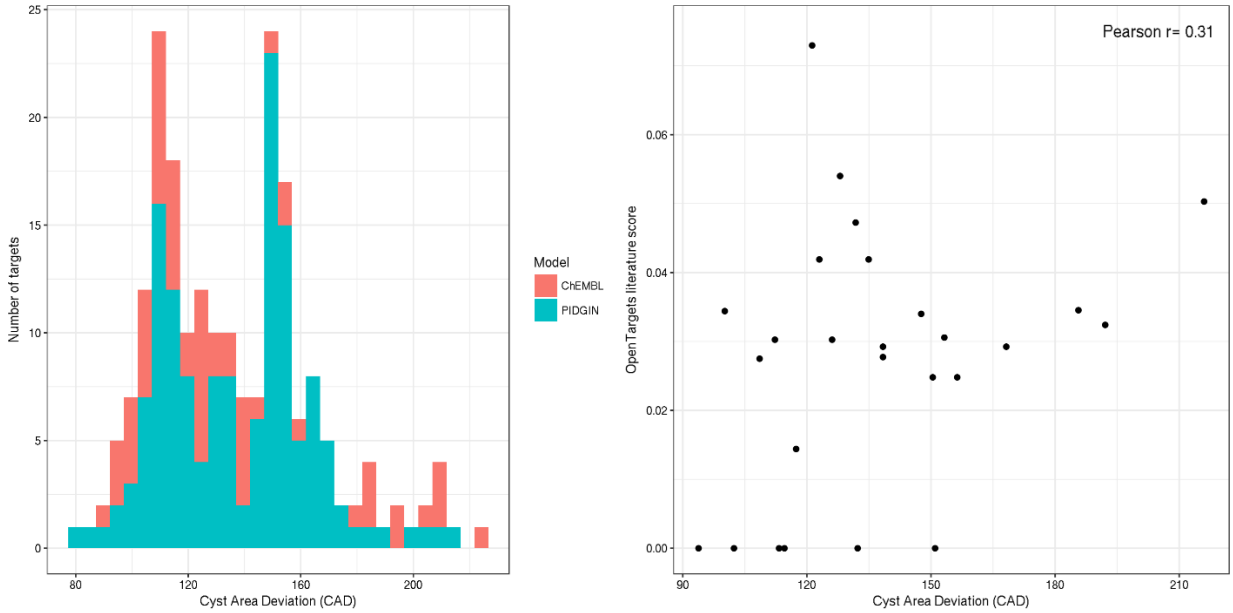

Figure 17. Characteristics of the Cyst Area Deviation (CAD) distribution measuring the strength of association of a target to cyst growth reduction and validation against Open Targets metrics. (Left) The distribution of targets predicted with the ChEMBL target prediction model is in red, whilst the distribution for those predicted with PIDGIN is depicted in blue. The CAD distribution is comparable for the two models. (Right) Each point represents a target, and the plot shows how the CAD correlated with the literature score. For most targets, an increase in CAD reflected a higher association with PKD, as reported by the higher literature score in Open Targets.

Despite this limitation, this analysis demonstrated that with a higher CAD, the rankings from Open Targets also increased, meaning that the CAD ranked targets in accordance with the current knowledge in the field of PKD drug discovery. It is hence proposed to be employed to identify targets which are less researched in the context of PKD drug discovery. However, none of the targets had both a high score per this metric and a low Open Targets Literature score

(**Figure 17, right**). Instead, gene-disease literature rankings from the CTD were used as a proxy for less researched targets in PKD.

5.3.2 Integration of gene expression studies and target occurrences in literature with the list of targets scoring high for CAD

Gene-disease associations from the CTD database were employed in order to get a sense of which targets are currently researched in PKD and, focus the analysis towards those that are comparatively less researched. 249 predicted targets had such an information in the CTD database (**Figure 18a**), 147 of which are predicted for non-cytotoxic CGR compounds (**Figure 18b**). This set of targets was then intersected with targets that had a higher CAD metric which narrowed the number of targets to 64 for all predicted targets in the CGR compound group (**Figure 18a**) and 26 for the set of non-cytotoxic CGR compounds (**Figure 18b**).

Additionally, Malas et al. compiled gene expression studies using PKD samples to extract differentially expressed genes in the disease, and also conducted their own gene expression analysis (**Appendix C**).[265] It was recently suggested that expression of the relevant targets in the disease-tissue of interest is of value when generating novel MoA analysis,[274] and for this reason, the 64 targets discussed above were intersected with the list of dysregulated genes in PKD which ultimately narrowed the focus of the study to 10 targets (**Figure 18a** and **Table 12**).

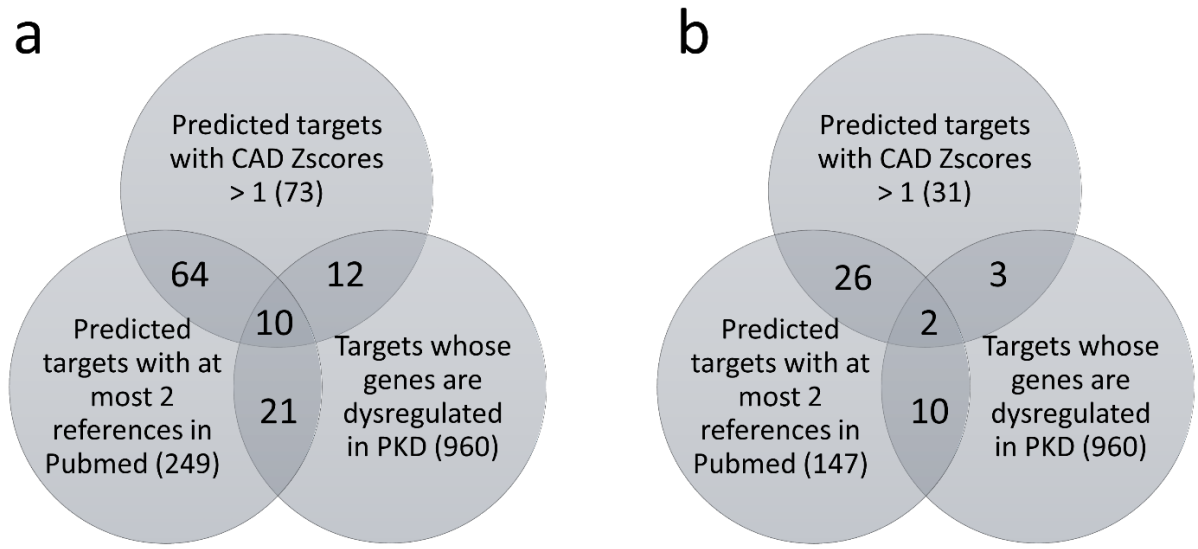

Figure 18. Venn diagram representing how targets were shortlisted for novelty, their effect in the kidney cyst screen and/or their enrichment in the CGR set of compounds, and their agreement with gene expression studies. (a) Counts for all the targets predicted by CGR compounds. (b) Counts for all the targets predicted for non-cytotoxic CGR compounds for which three targets passed all the filters and were shortlisted for structure-based studies.

Three targets were associated with renal processes observed in PKD (**Table 12**). Casein kinase II is investigated for its inhibition of survival and DNA repair in cancer cells through inhibition of Caspase.[275] This is also of interest in PKD since caspase-mediated apoptosis has been observed in PKD rat models.[276] Proteinase-activated receptor 1 (PAR-1) is downregulated in PKD and is involved in signalling that leads to fibrosis in the kidneys[277,278]. It is hence very relevant to PKD since interstitial fibrosis is associated with cystogenesis.[244,279] Serotonin receptor HT-2b is expressed in the kidneys[280] and serotonin itself may play a role in renal fibrogenesis through serotonin receptors.[281]

The next three targets were instead associated with other renal disorders (**Table 12**). Markedly, CYP2D6 was down-regulated in end-stage renal failure patients.[282] Mutations in 11 beta-hydroxysteroid dehydrogenase type II (HSD11B2) was also found with patients in end-stage renal failure.[283] Single nucleotide polymorphisms were observed in the gene encoding for mitochondrial aldehyde dehydrogenase (ALDH2) in a cohort comprising patients with chronic kidney disease.[284]

Even though Carbonic Anhydrase XIV (CA14) is not directly involved in PKD or any other renal disease (**Table 12**), it is expressed in the nephrons and may play a role in the acidification of urine taking place in the kidneys.[285] The remainder of the targets namely, the 2-acylglycerol O-acyltransferase 2, the microtubule-associated protein tau and the mitochondrial aldehyde dehydrogenase were not linked to PKD in the literature (**Table 12**).

If cytotoxic compounds predicted per the cytotoxicity model were removed, the list of shortlisted targets decreases to two targets (**Figure 18b** and **Table 13**). These two targets are discussed in the next section.

Gene ID	Target	Gene Symbol	Gene Regulation	CTD PubMed associations
1457	Casein kinase II subunit alpha	*CSNK2A1*	DOWN	0
2149	Proteinase-activated receptor 1	*F2R*	UP	0
3357	Serotonin HT-2b receptor	*HTR2B*	UP	0
3248	Prostaglandin dehydrogenase 1	*HPGD*	DOWN	2
1565	Cytochrome P450 2D6	*CYP2D6*	DOWN	1
3291	11-beta-hydroxysteroid dehydrogenase 2	*HSD11B2*	DOWN	2
23632	Carbonic anhydrase XIV	*CA14*	DOWN	0
80168	2-acylglycerol O-acyltransferase 2	*MOGAT2*	DOWN	0
4137	Microtubule-associated protein tau	*MAPT*	DOWN	1
217	Aldehyde dehydrogenase, mitochondrial	*ALDH2*	DOWN	1

Table 12. Targets selected as a result of the intersection between the three filters namely, differential expression filter, association to PKD filter and occurrence in the literature filter. The direction of gene expression dysregulation is also given for each target, along with the number of association to PKD in CTD.

Predicted Target	Gene	Protein family	Compound(s) driving the prediction	PDB Structure used for docking	Docking
Kallikrein 1	*KLK1*	Protease	Anthothecol	1SPJ	Inconclusive
Protease-activated receptor 1	*F2R*	GPCR	Podophyllin acetate; Picropodophyllin acetate	3VW7	Yes

Table 13. Shortlisted predicted targets when cytotoxicity filter was included and outcome of structure-based studies.

5.3.3 Docking analyses agreed with 1 out of 2 shortlisted target predictions

In order to give more weight to the shortlisted predictions reported in **Table 13**, docking studies were performed. This involved docking the compound driving the prediction to its corresponding target (see Material and Methods). The first shortlisted compound-target prediction was comprised of the CGR compound Anthotecol and the Kallikrein 1 (KLK1) protein (**Table 13**). KLK1 has been linked to recessive polycystic kidneys in *cpk* mice models[286] and also to end-stage renal disease (to which PKD develops into).[287] Unfortunately, the only PDB structure available for KLK1 did not have a co-crystallised inhibitor or ligand in the binding site (PDB ID: 1SPJ). Hence, the site reported in the article accompanying the structure, along with predictions of the binding site were performed (see Material and Methods). A docking score of -4.417 was obtained with Glide. Since acceptable Glide docking

scores fall below -6, it cannot be considered that this docking was indicative of a potential binding of Anthotecol to KLK1.

The second shortlisted target was PAR-1 (**Table 13**), which is involved in signalling that leads to renal fibrosis [277,278]. This prediction was driven by two CGR stereoisomers, podophyllin acetate (PA) and picropodophyllin acetate (PPA). The Glide docking score obtained for both compounds was optimal as it fell below -8 for both compounds, even though Vorapaxar, the co-crystallized inhibitor of the PAR-1 structure, had a much lower docking score, and hence docked better than PA and PPA (**Table 14**).

Compound	Docking score
Podophyllin acetate	-8.549
Picropodophyllin acetate	-8.132
Vorapaxar (co-crystallised inhibitor, PDB: 3VW7)	-15.144

Table 14. Glide docking scores of podophyllin acetate, picropodophyllin acetate and the co-crystallized inhibitor Vorapaxar for the PAR-1 structure (PDB: 3VW7). Acceptable Glide docking scores fall below -6 and optimal Glide docking scores fall below -8.

A superimposition of the docking pose obtained for PA to that of Vorapaxar showed that PA occupied a very similar conformation in the binding site (**Figure 19**, top). PPA also occupied the same area with some minor variations compared to PA (**Figure 19**, bottom). The interaction diagram of Vorapaxar showed that it made hydrogen bounds with TYR337 and LEU258 (**Figure 20**), and these interactions were also found in the diagrams of PA and PPA (**Figure 21**). Those results showed that PA and PPA may bind to the PAR-1 receptor. This suggested

that PA and PPA exerted their effect on cyst growth through the binding of PAR-1, potentially by reducing fibrosis in the polycystic kidney tissues.

5.4 Conclusion

The target prediction workflow developed in this thesis was applied to rationalise the MoA of compounds active in a phenotypic assay aimed at identifying modulators of Forskolin-induced kidney cyst growth. The CAD was developed to rank predicted targets according to their relevance to PKD. It was found that the CAD agreed with the literature association score from Open Targets. As a result, 2 targets were identified for which little research was done in the context of PKD drug discovery and which were differentially expressed in mice models of PKD and had a high CAD: these were KLK1 and PAR-1. For the latter target, it was shown that the compounds driving the prediction occupied the same area of the co-crystallised inhibitors in the binding site, and shared the same interactions to those seen for the inhibitor as well. This chapter presented a workflow in which cheminformatics analyses of phenotypic screens were integrated with gene expression studies and structure-based approaches to generate MoA hypotheses for the identification of therapeutic opportunities in PKD research.

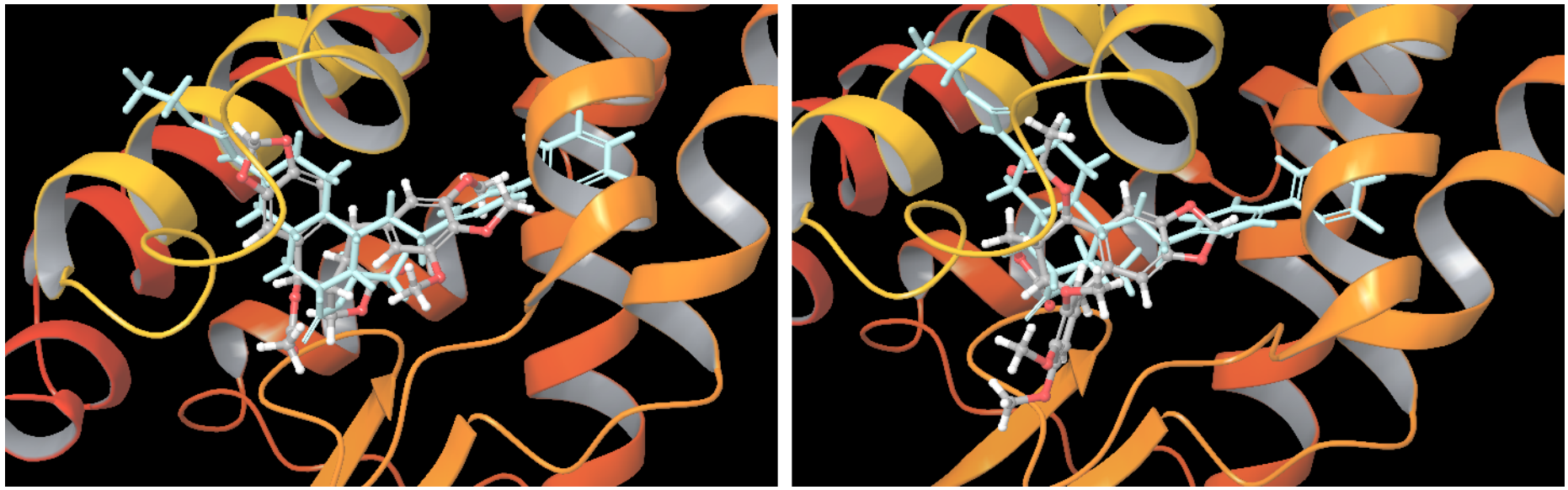

Figure 19. Superimposition of Podophyllin (left) and Picropodophyllin (right) with Vorapaxar (co-crystallised ligand, coloured in cyan) in the binding site of protease-activated receptor 1 (PAR-1). Podophyllin adopted a very similar binding pose to the one of Vorapaxar in the binding site. The pose of Picropodophyllin aligned with that of Vorapaxar with slightly more variations.

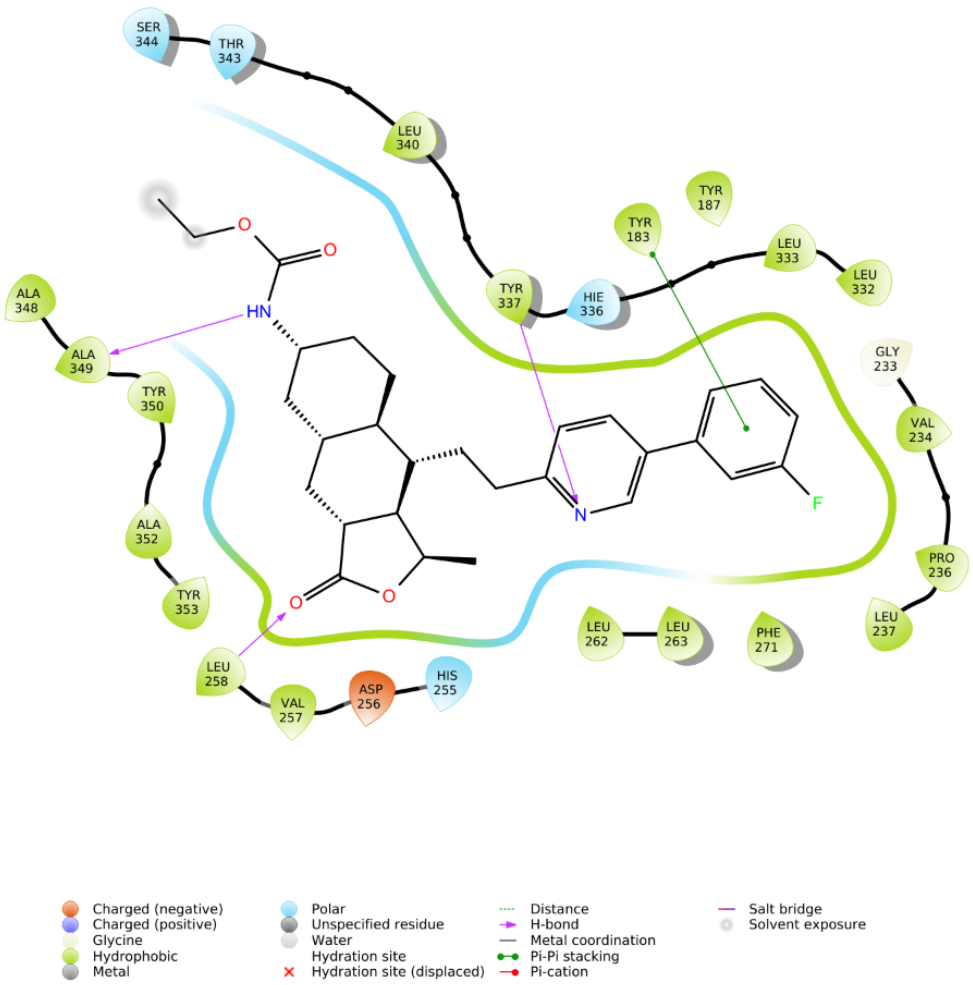

Figure 20. Interaction diagram for Vorapaxar (co-crystallised) inhibitor in the binding site of PAR-1. Notable interactions such as the hydrogen bonds with TYR337 and LEU258 were also found for Podphylin acetate and Picropodophyllin acetate.

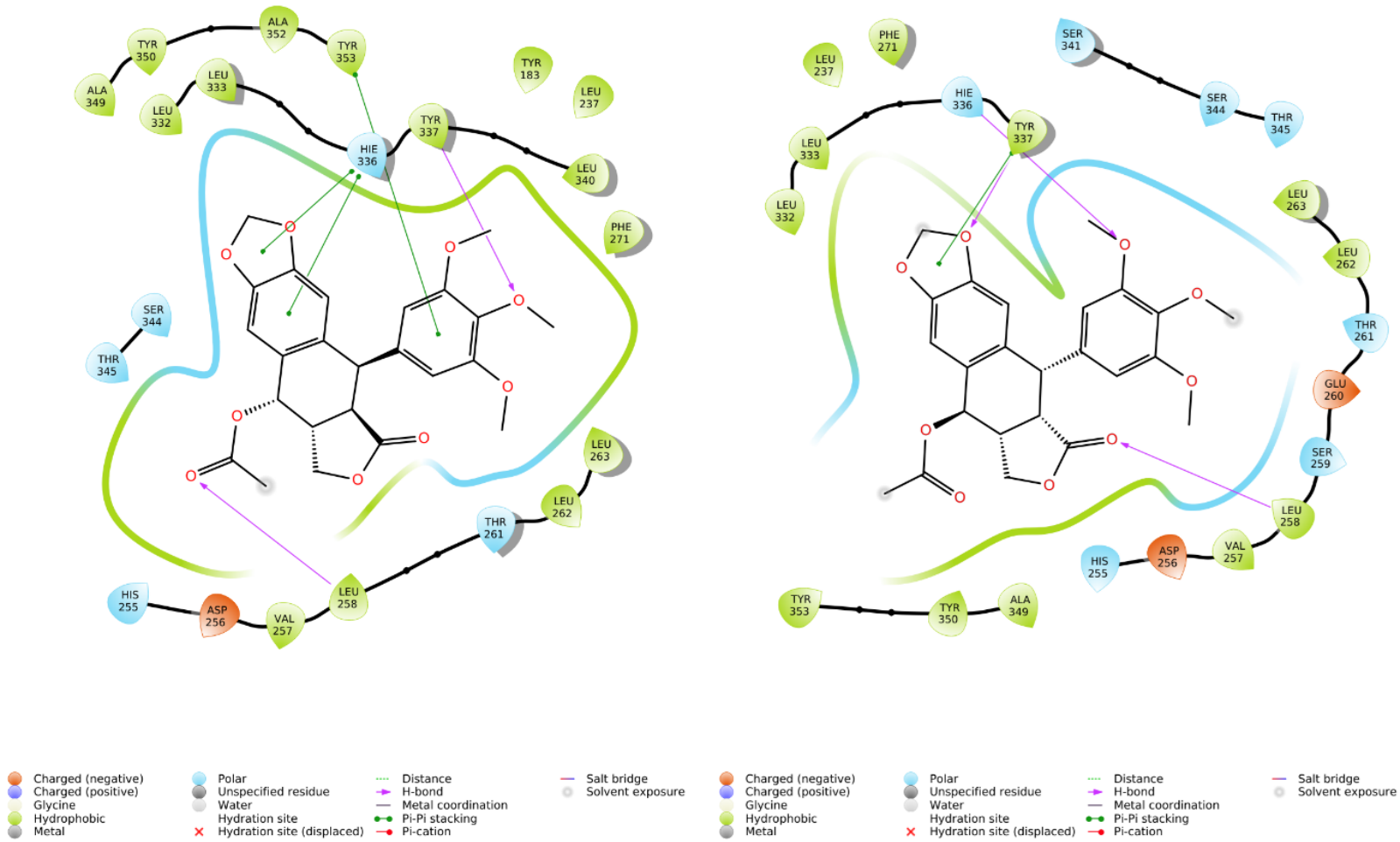

Figure 21. Interaction diagram of Podophyllin acetate (left) and Picropodophyllin acetate (right) with the amino acids in the binding site of PAR-1. Comparable interactions to the one Vorapaxar make in the binding site can be found, namely, the hydrogen bonds to TYR337 and LEU258

Conclusion

Drug discovery is currently hampered by major drug attrition rates. This is due to the focus of drug development efforts on target-based screening approaches since those do not provide the full picture behind the MoA of the lead compounds with regard to efficacy and safety. Phenotypic-based screening approaches, in combination with target-based approaches, are regarded by researchers in the field, as a way to solve these problems, but the identification of the MoA responsible for the activity of the compounds observed in those screens is a strenuous task, as it involves the use of expensive and time-consuming biochemical assays. *In silico* approaches that make use of databases of gene expression profiles of compounds or experimental bioactivity can identify potentially efficacious candidate compounds and pinpoint unsafe ones. Another *in silico* approach uses the molecular similarity principle to make compound-target predictions and both these concepts have been the focus of this thesis. The aim of the thesis has been to 1. assess and quantify whether the molecular similarity principle also applies to phenotypic screening datasets, 2. investigate how similar or different the MoA predicted by those algorithms are compared to those that can be generated from experimental bioactivity, and 3. combine the output of these models with other information to propose disease-relevant MoA for compounds active in phenotypic screens.

While the molecular similarity principle has been extensively studied for target-based screening data, the literature analysis conducted in this thesis highlighted that this was less the case for phenotypic-based screening datasets. In addition, the molecular similarity principle

has so far, not been quantified for phenotypic-based screening datasets such as those employed in this thesis. In Chapter 3, the Bayesian framework was employed to estimate the variation of phenotypically similar pairs that can be expected for an increasing chemical similarity in the BioMAP and ChEMBL phenotypic datasets, using three of the most employed fingerprints in the cheminformatics literature. The fraction of phenotypically similar compounds was estimated to increase by 1% in the ELISA based readout dataset and 3% in the cell-viability dataset, for an increase in 5% Tc similarity. In this way, it was shown that the molecular similarity principle is valid in phenotypic screens. Despite that a limited number of datasets was used, a framework was provided to be applied to other types of phenotypic screening data, in order to compute estimates for the molecular similarity principle in phenotypic screens. In addition, the variation of the fraction of phenotypically similar pairs was not always linear and followed different trends for each of the three fingerprints. When the trend was piecewise-linear or quadratic, it was shown that ECFP4 detected the molecular similarity principle for lower values of Tc compared to the other two fingerprints used in chapter 3. The neighbourhood enhancement ratio was employed to demonstrate that ECFP4 were less sensitive to activity-cliffs in the ChEMBL dataset compared to MACCS keys and PubChem fingerprints. This showed that ECFP4 is a reasonable fingerprint choice for any cheminformatics applications such as target prediction analyses or QSAR analyses involving phenotypic screening datasets such as those employed in this thesis.

Since the molecular similarity principle was found to be valid in phenotypic screening datasets and knowing that ECFP4 is a recommended choice for these datasets, the target prediction workflow described in Chapter 2 was applied to the compounds in the NCATS phenotypic dataset. The aim of that work, which was described in Chapter 4, was to assess how the MoA hypotheses obtained through the target prediction workflow described in Chapter 2 compared to the MoA hypotheses obtained through the use of experimental data in the Drugmatrix

dataset. Indeed, this type of comparison has not been performed and highlighted the practicality of target prediction algorithms for the deconvolution of activity signals in phenotypic screens. Indeed, there was a high agreement between the MoA obtained through prediction and experimental data. While the individual target mostly differed between the predicted MoA and the experimental MoA for all the phenotypic assays in the NCATS dataset, the predicted and experimental target genes were highly functionally similar. This highlighted that while some functional agreement was seen between prediction and experimental data, the target prediction workflow developed in this thesis obtained targets that were still different from those which were observed experimentally but still very relevant to the phenotypic assay. This suggested that target prediction algorithms can complement the use of experimental bioactivity data available publicly to generate relevant MoA hypotheses for the deconvolution of compound activity in phenotypic screens.

In Chapter 5, the target prediction workflow was applied to deconvolute the activity of compounds in a kidney cyst growth assay. A metric was developed to rank targets by their relevance to PKD. Predicted targets with a high ranking per the above metric were intersected with differentially expressed genes whilst ensuring that the targets were not highly investigated in drug development for PKD. This analysis shortlisted two candidate targets. Where structural information was available, it was shown that the plausible docking of candidate compounds active in the kidney cyst screening assay to their predicted targets. This application chapter showed how target prediction can be integrated with various information and other approaches to improve target deconvolution.

In conclusion, this work reinforces the recommendation of *in silico* target prediction algorithms in the deconvolution of phenotypic screens. They indeed rely on the molecular similarity principle which was not only valid but also quantifiable in phenotypic screening datasets and

used in combination with experimental data, can lead to alternative MoA. This work showed that predictions can easily be integrated to other types of “evidence” such as gene expression profiles of compounds and structural bioinformatics and this will contribute to not only alleviate some of the shortcomings associated with the training set of these models regarding data quality and quantity but also to build confidence in these predictions.

References

(1) World Health Organization. The top 10 causes of death http://www.who.int/en/news-room/fact-sheets/detail/the-top-10-causes-of-death (accessed May 27, 2018).

(2) Haffner, M. E.; Whitley, J.; Moses, M. *Nat. Rev. Drug Discov.* **2002**, *1* (10), 821–825.

(3) Schieppati, A.; Henter, J.-I.; Daina, E.; Aperia, A. *Lancet* **2008**, *371* (9629), 2039–2041.

(4) Luzzatto, L.; Hollak, C. E. M.; Cox, T. M.; Schieppati, A.; Licht, C.; Kääriäinen, H.; Merlini, G.; Schaefer, F.; Simoens, S.; Pani, L.; Garattini, S.; Remuzzi, G. *Lancet* **2015**, *385* (9970), 750–752.

(5) Drews, J. *Science* **2000**, *287* (5460), 1960–1964.

(6) Drews, J.; Ryser, S. *Drug Inf. J.* **1996**, *30* (1), 97–108.

(7) Croston, G. E. *Expert Opin. Drug Discov.* **2017**, *12* (5), 427–429.

(8) Eder, J.; Sedrani, R.; Wiesmann, C. *Nat. Rev. Drug Discov.* **2014**, *13* (8), 577–587.

(9) Waring, M. J.; Arrowsmith, J.; Leach, A. R.; Leeson, P. D.; Mandrell, S.; Owen, R. M.; Pairaudeau, G.; Pennie, W. D.; Pickett, S. D.; Wang, J.; Wallace, O.; Weir, A. *Nat. Rev. Drug Discov.* **2015**, *14* (7), 475–486.

(10) Mignani, S.; Huber, S.; Tomás, H.; Rodrigues, J.; Majoral, J.-P. *Drug Discov. Today* **2016**, *21* (2), 239–249.

(11) Arrowsmith, J. *Nat. Rev. Drug Discov.* **2011**, *10* (5), 328–329.

(12) Arrowsmith, J. *Nat. Rev. Drug Discov.* **2011**, *10* (2), 87–87.

(13) Arrowsmith, J.; Miller, P. *Nat. Rev. Drug Discov.* **2013**, *12* (8), 569–569.

(14) Sams-Dodd, F. *Drug Discov. Today* **2005**, *10* (2), 139–147.

(15) Prinz, F.; Schlange, T.; Asadullah, K. *Nat. Rev. Drug Discov.* **2011**, *10* (9), 712–712.

(16) Osherovich, L. *SciBX* **2011**, *4* (15).

(17) Swinney, D. C.; Anthony, J. *Nat. Rev. Drug Discov.* **2011**, *10* (7), 507–519.

(18) Zheng, W.; Thorne, N.; McKew, J. C. *Drug Discov. Today* **2013**, *18* (21), 1067–1073.

(19) Anighoro, A.; Bajorath, J.; Rastelli, G. *J. Med. Chem.* **2014**, *57* (19), 7874–7887.

(20) Swinney, D. C.; Anthony, J. *Nat. Rev. Drug Discov.* **2011**, *10* (7), 507–519.

(21) Zheng, W.; Thorne, N.; McKew, J. C. *Drug Discov. Today* **2013**, *18* (21–22), 1067–1073.

(22) Mosmann, T. *J. Immunol. Methods* **1983**, *65* (1–2), 55–63.

(23) Buttke, T. M.; McCubrey, J. A.; Owen, T. C. *J. Immunol. Methods* **1993**, *157* (1–2), 233–240.

(24) Roehm, N. W.; Rodgers, G. H.; Hatfield, S. M.; Glasebrook, A. L. *J. Immunol. Methods* **1991**, *142* (2), 257–265.

(25) Vichai, V.; Kirtikara, K. *Nat. Protoc.* **2006**, *1* (3), 1112–1116.

(26) Ansar Ahmed, S.; Gogal, R. M.; Walsh, J. E. *J. Immunol. Methods* **1994**, *170* (2), 211–

224.

(27) Lee, J. A.; Chu, S.; Willard, F. S.; Cox, K. L.; Sells Galvin, R. J.; Peery, R. B.; Oliver, S. E.; Oler, J.; Meredith, T. D.; Heidler, S. A.; Gough, W. H.; Husain, S.; Palkowitz, A. D.; Moxham, C. M. *J. Biomol. Screen.* **2011**, *16* (6), 588–602.

(28) Lee, J. A.; Shinn, P.; Jaken, S.; Oliver, S.; Willard, F. S.; Heidler, S.; Peery, R. B.; Oler, J.; Chu, S.; Southall, N.; Dexheimer, T. S.; Smallwood, J.; Huang, R.; Guha, R.; Jadhav, A.; Cox, K.; Austin, C. P.; Simeonov, A.; Sittampalam, G. S.; Husain, S.; Franklin, N.; Wild, D. J.; Yang, J. J.; Sutherland, J. J.; Thomas, C. J. *PLoS One* **2015**, *10* (7), e0130796.

(29) Plavec, I.; Sirenko, O.; Privat, S.; Wang, Y.; Dajee, M.; Melrose, J.; Nakao, B.; Hytopoulos, E.; Berg, E. L.; Butcher, E. C. *Proc. Natl. Acad. Sci. U. S. A.* **2004**, *101* (5), 1223–1228.

(30) Kunkel, E. J.; Dea, M.; Ebens, A.; Hytopoulos, E.; Melrose, J.; Nguyen, D.; Ota, K. S.; Plavec, I.; Wang, Y.; Watson, S. R.; Butcher, E. C.; Berg, E. L. *FASEB J.* **2004**, *18*, 1279–1281.

(31) Pampaloni, F.; Reynaud, E. G.; Stelzer, E. H. K. *Nat. Rev. Mol. Cell Biol.* **2007**, *8* (10), 839–845.

(32) Terstappen, G. C.; Schlüpen, C.; Raggiaschi, R.; Gaviraghi, G. *Nat. Rev. Drug Discov.* **2007**, *6* (11), 891–903.

(33) Lee, J.; Bogyo, M. *Curr. Opin. Chem. Biol.* **2013**, *17* (1), 118–126.

(34) Guiffant, D.; Tribouillard, D.; Gug, F.; Galons, H.; Meijer, L.; Blondel, M.; Bach, S. *Biotechnol. J.* **2007**, *2* (1), 68–75.

(35) Licitra, E. J.; Liu, J. O. *Proc. Natl. Acad. Sci. U. S. A.* **1996**, *93* (23), 12817–12821.

(36) Rossenu, S.; Dewitte, D.; Vandekerckhove, J.; Ampe, C. *J. Protein Chem.* **1997**, *16* (5), 499–503.

(37) Zhu, H.; Snyder, M. *Curr. Opin. Chem. Biol.* **2003**, *7* (1), 55–63.

(38) Bajorath, J. *Chemoinformatics: Concepts, Methods, and Tools for Drug Discovery*; Methods in Molecular Biology; Humana Press, 2004; Vol. 275.

(39) Guha, R.; Bender, A. *Computational Approaches in Cheminformatics and Bioinformatics*; John Wiley and Sons, 2011.

(40) Baskin, I.; Varnek, A. In *Chemoinformatics Approaches to Virtual Screening*; Varnek, A., Tropsha, A., Eds.; 2008; pp 1–44.

(41) Riniker, S.; Landrum, G. *J. Cheminform.* **2013**, *5* (1).

(42) Gardiner, E. J.; Gillet, V. J.; Haranczyk, M.; Hert, J.; Holliday, J. D.; Malim, N.; Patel, Y.; Willett, P. *Stat. Anal. Data Min.* **2009**, *2* (2), 103–114.

(43) Rogers, D.; Hahn, M. *J. Chem. Inf. Model.* **2010**, *50* (5), 742–754.

(44) DAYLIGHT Fingerprint, DAYLIGHT Inc., Mission Viejo, CA, USA.

(45) Bender, A.; Mussa, H. Y.; Glen, R. C.; Reiling, S. *J. Chem. Inf. Comput. Sci.* **2004**, *44* (5), 1708–1718.

(46) MDL Maccs Keys, 166 bit keyset, MDL Information Systems, San Leandro, CA 94577, USA.

(47) Kim, S.; Thiessen, P. A.; Bolton, E. E.; Chen, J.; Fu, G.; Gindulyte, A.; Han, L.; He, J.; He, S.; Shoemaker, B. A.; Wang, J.; Yu, B.; Zhang, J.; Bryant, S. H. *Nucleic Acids*

Res. **2016**, *44* (D1), D1202–D1213.

(48) UNITY, Tripos Inc., 2006., 1699 S. Hanley Road, St. Louis, MO 63144, USA.

(49) Sebastian, A.; Bender, A.; Ramakrishnan, V. *Mol. Inform.* **2010**, *29* (11), 773–779.

(50) Yu, X.; Geer, L. Y.; Han, L.; Bryant, S. H. *J. Cheminform.* **2015**, *7* (1), 1–12.

(51) Paricharak, S.; IJzerman, A. P.; Bender, A.; Nigsch, F. *Submitt. to ACS Chem. Biol.* **2016**.

(52) Johnson, M. A.; Maggiora, G. M. *Concepts and Applications of Molecular Similarity*; John Wiley & Sons: New York, 1990.

(53) Maggiora, G.; Vogt, M.; Stumpfe, D.; Bajorath, J. *J. Med. Chem.* **2013**, *57* (8), 3186–3204.

(54) Leach, A. R.; Hann, M. M. *Drug Discov. Today* **2000**, *5* (8), 326–336.

(55) Bajusz, D.; Rácz, A.; Héberger, K. *J. Cheminform.* **2015**, *7* (1), 20.

(56) Martin, Y. C.; Kofron, J. L.; Traphagen, L. M. *J. Med. Chem.* **2002**, *45* (19), 4350–4358.

(57) Muchmore, S. W.; Debe, D. A.; Metz, J. T.; Brown, S. P.; Martin, Y. C.; Hajduk, P. J. *J. Chem. Inf. Model.* **2008**, *48* (5), 941–948.

(58) Nettles, J. H.; Jenkins, J. L.; Bender, A.; Deng, Z.; Davies, J. W.; Glick, M. *J. Med. Chem.* **2006**, *49* (23), 6802–6810.

(59) Patterson, D. E.; Cramer, R. D.; Ferguson, A. M.; Clark, R. D.; Weinberger, L. E. *J. Med. Chem.* **1996**, *39* (16), 3049–3059.

(60) Kubinyi, H. *J. Braz. Chem. Soc.* **2002**, *13* (6), 717–726.

(61) Stumpfe, D.; Bajorath, J. *J. Med. Chem.* **2012**, *55* (7), 2932–2942.

(62) Peltason, L.; Bajorath, J. *J. Med. Chem.* **2007**, *50* (23), 5571–5578.

(63) Guha, R.; Van Drie, J. H. *J. Chem. Inf. Model.* **2008**, *48* (3), 646–658.

(64) Liu, C.; Su, J.; Yang, F.; Wei, K.; Ma, J.; Zhou, X. *Mol. Biosyst.* **2015**, *11* (3), 714–722.

(65) Bornot, A.; Blackett, C.; Engkvist, O.; Murray, C.; Bendtsen, C. *J. Biomol. Screen.* **2014**, *19* (5), 696–706.

(66) Polyakov, V. R.; Moorcroft, N. D.; Drawid, A. *J. Chem. Inf. Model.* **2014**, *54* (2), 377–386.

(67) Fukuda, Y.; Sano, O.; Kazetani, K.; Yamamoto, K.; Iwata, H.; Matsui, J. *BMC Biochem.* **2016**, *17* (1), 9.

(68) Sano, O.; Kazetani, K.; Adachi, R.; Kurasawa, O.; Kawamoto, T.; Iwata, H. *FEBS Open Bio* **2017**, *7* (4), 495–503.

(69) Gaulton, A.; Overington, J. P. *Future Med. Chem.* **2010**, *2* (6), 903–907.

(70) Wang, Y.; Xiao, J.; Suzek, T. O.; Zhang, J.; Wang, J.; Bryant, S. H. *Nucleic Acids Res.* **2009**, *37* (SUPPL. 2), 623–633.

(71) Wang, Y.; Suzek, T.; Zhang, J.; Wang, J.; He, S.; Cheng, T.; Shoemaker, B. a.; Gindulyte, A.; Bryant, S. H. *Nucleic Acids Res.* **2014**, *42* (D1), 1–8.

(72) 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. *Nucleic Acids*

Res. **2012**, *40*, 1100–1107.

(73) 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. *Nucleic Acids Res.* **2014**, *42*, 1083–1090.

(74) Li, Q.; Cheng, T.; Wang, Y.; Bryant, S. H. *Drug Discov. Today* **2010**, *15* (23–24), 1052–1057.

(75) Ganter, B.; Tugendreich, S.; Pearson, C. I.; Ayanoglu, E.; Baumhueter, S.; Bostian, K. A.; Brady, L.; Browne, L. J.; Calvin, J. T.; Day, G.-J.; Breckenridge, N.; Dunlea, S.; Eynon, B. P.; Furness, L. M.; Ferng, J.; Fielden, M. R.; Fujimoto, S. Y.; Gong, L.; Hu, C.; Idury, R.; Judo, M. S. B.; Kolaja, K. L.; Lee, M. D.; McSorley, C.; Minor, J. M.; Nair, R. V.; Natsoulis, G.; Nguyen, P.; Nicholson, S. M.; Pham, H.; Roter, A. H.; Sun, D.; Tan, S.; Thode, S.; Tolley, A. M.; Vladimirova, A.; Yang, J.; Zhou, Z.; Jarnagin, K. *J. Biotechnol.* **2005**, *119* (3), 219–244.

(76) Ganter, B.; Snyder, R. D.; Halbert, D. N.; Lee, M. D. *Pharmacogenomics* **2006**, *7* (7), 1025–1044.

(77) Olah, M.; Mracec, M.; Ostopovici, L.; Rad, R.; Bora, A.; Hadaruga, N.; Olah, I.; Banda, M.; Simon, Z.; Mracec, M.; Oprea, T. I. In *Chemoinformatics in Drug Discovery*; Wiley-Blackwell, 2005; Vol. 23, pp 221–239.

(78) Olah, M.; Rad, R.; Ostopovici, L.; Bora, A.; Hadaruga, N.; Hadaruga, D.; Moldovan, R.; Fulias, A.; Mractc, M.; Oprea, T. I. *Chem. Biol. From Small Mol. to Syst. Biol. Drug Des. Vol. 1-3* **2008**, *2*, 760–786.

(79) Wassermann, A. M.; Lounkine, E.; Hoepfner, D.; Le Goff, G.; King, F. J.; Studer, C.;

Peltier, J. M.; Grippo, M. L.; Prindle, V.; Tao, J.; Schuffenhauer, A.; Wallace, I. M.; Chen, S.; Krastel, P.; Cobos-Correa, A.; Parker, C. N.; Davies, J. W.; Glick, M. *Nat. Chem. Biol.* **2015**, *11* (12), 958–966.

(80) Tiikkainen, P.; Franke, L. *J. Chem. Inf. Model.* **2012**, *52* (2), 319–326.

(81) Tiikkainen, P.; Bellis, L.; Light, Y.; Franke, L. *J. Chem. Inf. Model.* **2013**, *53* (10), 2499–2505.

(82) Nidhi; Glick, M.; Davies, J. W.; Jenkins, J. L. *J. Chem. Inf. Model.* **2006**, *46* (3), 1124–1133.

(83) Martínez-Jiménez, F.; Papadatos, G.; Yang, L.; Wallace, I. M.; Kumar, V.; Pieper, U.; Sali, A.; Brown, J. R.; Overington, J. P.; Marti-Renom, M. a. *PLoS Comput. Biol.* **2013**, *9* (10).

(84) Crisman, T. J.; Parker, C. N.; Jenkins, J. L.; Scheiber, J.; Thoma, M.; Kang, Z. Bin; Kim, R.; Bender, A.; Nettles, J. H.; Davies, J. W.; Glick, M. *J. Chem. Inf. Model.* **2007**, *47* (4), 1319–1327.

(85) 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. *Nat. Chem. Biol.* **2008**, *4* (1), 59–68.

(86) Ravindranath, A. C.; Perualila-Tan, N.; Kasim, A.; Drakakis, G.; Liggi, S.; Brewerton, S. C.; Mason, D.; Bodkin, M. J.; Evans, D. A.; Bhagwat, A.; Talloen, W.; Göhlmann, H. W. H.; QSTAR Consortium, Q. C.; Shkedy, Z.; Bender, A. *Mol. BioSyst.* **2015**, *11* (1), 86–96.

(87) Paricharak, S.; Cortes-Ciriano, I.; IJzerman, A. P.; Malliavin, T. E.; Bender, A. *J.*

Cheminform. **2015**, 1–11.

(88) Sato, T.; Matsuo, Y.; Honma, T.; Yokoyama, S. *J. Med. Chem.* **2008**, *51* (24), 7705–7716.

(89) Noeske, T.; Sasse, B. C.; Stark, H.; Parsons, C. G.; Weil, T.; Schneider, G. *ChemMedChem* **2006**, *1* (10), 1066–1068.

(90) Reker, D.; Rodrigues, T.; Schneider, P.; Schneider, G. *Proc. Natl. Acad. Sci. U. S. A.* **2014**, *111* (11), 4067–4072.

(91) Wen, M.; Zhang, Z.; Niu, S.; Sha, H.; Yang, R.; Yun, Y.; Lu, H. *J. Proteome Res.* **2017**, *16* (4), 1401–1409.

(92) Lagunin, A.; Stepanchikova, A.; Filimonov, D.; Poroikov, V. *Bioinformatics* **2000**, *16* (8), 747–748.

(93) Keiser, M. J.; Roth, B. L.; Armbruster, B. N.; Ernsberger, P.; Irwin, J. J.; Shoichet, B. K. *Nat. Biotechnol.* **2007**, *25* (2), 197–206.

(94) Nickel, J.; Gohlke, B. O.; Erehman, J.; Banerjee, P.; Rong, W. W.; Goede, A.; Dunkel, M.; Preissner, R. *Nucleic Acids Res.* **2014**, *42* (W1), 26–31.

(95) Awale, M.; Reymond, J.-L. *J. Cheminform.* **2017**, *9* (1), 11.

(96) Nigsch, F.; Hutz, J.; Cornett, B.; Selinger, D. W.; McAllister, G.; Bandyopadhyay, S.; Loureiro, J.; Jenkins, J. L. *EURASIP J. Bioinforma. Syst. Biol.* **2012**, *2012* (1), 2.

(97) Schomburg, K. T.; Bietz, S.; Briem, H.; Henzler, A. M.; Urbaczek, S.; Rarey, M. *J. Chem. Inf. Model.* **2014**, *54* (6), 1676–1686.

(98) Wang, X.; Shen, Y.; Wang, S.; Li, S.; Zhang, W.; Liu, X.; Lai, L.; Pei, J.; Li, H.

Nucleic Acids Res. **2017**, *45* (W1), W356–W360.

(99) Hwang, H.; Dey, F.; Petrey, D.; Honig, B. *Proc. Natl. Acad. Sci. U. S. A.* **2017**, *114* (52), 13685–13690.

(100) Huang, T.; Mi, H.; Lin, C.; Zhao, L.; Zhong, L. L. D.; Liu, F.; Zhang, G.; Lu, A.; Bian, Z. *BMC Bioinformatics* **2017**, *18* (1), 165.

(101) Yamanishi, Y.; Araki, M.; Gutteridge, A.; Honda, W.; Kanehisa, M. *Bioinformatics* **2008**, *24* (13), 232–240.

(102) He, Z.; Zhang, J.; Shi, X.-H.; Hu, L.-L.; Kong, X.; Cai, Y.-D.; Chou, K.-C. *PLoS One* **2010**, *5* (3), e9603.

(103) Martínez-Jiménez, F.; Marti-Renom, M. a. *PLOS Comput. Biol.* **2015**, *11* (3), e1004157.

(104) Klipp, E.; Wade, R. C.; Kummer, U. *Curr. Opin. Biotechnol.* **2010**, *21* (4), 511–516.

(105) Mervin, L. H.; Afzal, A. M.; Drakakis, G.; Lewis, R.; Engkvist, O.; Bender, A. *J. Cheminform.* **2015**, *7* (1), 51.

(106) Svensson, F.; Norinder, U.; Bender, A. *Toxicol. Res.* **2017**, *6* (1), 73–80.

(107) 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. *Nat. Chem. Biol.* **2008**, *4* (1), 59–68.

(108) Mervin, L. H.; Cao, Q.; Barrett, I. P.; Firth, M. A.; Murray, D.; McWilliams, L.; Haddrick, M.; Wigglesworth, M.; Engkvist, O.; Bender, A. *ACS Chem. Biol.* **2016**, *11* (11), 3007–3023.

(109) Fu, X.; Mervin, L. H.; Li, X.; Yu, H.; Li, J.; Mohamad Zobir, S. Z.; Zoufir, A.; Zhou, Y.; Song, Y.; Wang, Z.; Bender, A. *J. Chem. Inf. Model.* **2017**, *57* (3).

(110) Liggi, S.; Drakakis, G.; Hendry, A. E.; Hanson, K. M.; Brewerton, S. C.; Wheeler, G. N.; Bodkin, M. J.; Evans, D. a.; Bender, A. *Mol. Inform.* **2013**, *32* (11–12), 1009–1024.

(111) Drakakis, G.; Wafford, K. A.; Brewerton, S. C.; Bodkin, M. J.; Evans, D. A.; Bender, A. *ACS Chem. Biol.* **2017**, *12* (6), 1593–1602.

(112) Lo, Y.-C.; Senese, S.; Li, C.-M.; Hu, Q.; Huang, Y.; Damoiseaux, R.; Torres, J. Z. *PLOS Comput. Biol.* **2015**, *11* (3), e1004153.

(113) Fourches, D.; Muratov, E.; Tropsha, A. *J. Chem. Inf. Model.* **2010**, *50* (7), 1189–1204.

(114) Standardizer, version 15.1.19.0, ChemAxon, 2015.

(115) RDKit: Open-source cheminformatics, version 2013.09.1, Greg Landlum, 2013.

(116) Scikit-chem, pre-alpha version 0.0.6, Richard P. I. Lewis, 2016.

(117) Bender, A.; Jenkins, J. L.; Scheiber, J.; Sukuru, S. C. K.; Glick, M.; Davies, J. W. *J. Chem. Inf. Model.* **2009**, *49*, 108–119.

(118) Alvarsson, J.; Eklund, M.; Andersson, C.; Carlsson, L.; Spjuth, O.; Wikberg, J. E. S. *J. Chem. Inf. Model.* **2014**.

(119) Willett, P. *Drug Discov. Today* **2006**, *11* (23–24), 1046–1053.

(120) Jansen, G.; Lee, A. Y.; Epp, E.; Fredette, A.; Surprenant, J.; Harcus, D.; Scott, M.; Tan, E.; Nishimura, T.; Whiteway, M.; Hallett, M.; Thomas, D. Y. *Mol. Syst. Biol.* **2009**, *5*, 338.

(121) Bieler, M.; Heilker, R.; Köppen, H.; Schneider, G. *J. Chem. Inf. Model.* **2011**, *51* (8), 1897–1905.

(122) Mervin, L. H. PIDGIN, version 2, Platt Scaled Random Forest Protein Target Prediction Tool trained on SARs from PubChem (Mined 21/06/16) and ChEMBL21, 2016, https://github.com/lhm30/PIDGINv2.

(123) Mervin, L. H.; Bulusu, K. C.; Kalash, L.; Afzal, A.; Svensson, F.; Firth, M. A.; Barrett, I.; Engkvist, O.; Bender, A. *Bioinformatics* **2017**.

(124) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, É. *J. Mach. Learn. Res.* **2012**, *12*, 2825–2830.

(125) Breiman, L. *Mach. Learn.* **2001**, *45* (1), 5–32.

(126) Breiman, L.; Friedman, J. H.; Olshen, R. A.; Stone, C. J. *Classification and Regression Trees*; 1984; Vol. 19.

(127) Platt, J. C.; Platt, J. C. *Adv. LARGE MARGIN Classif.* **1999**, 61--74.

(128) Niculescu-Mizil, A.; Caruana, R. In *Proceedings of the 22nd international conference on Machine learning - ICML '05*; ACM Press: New York, New York, USA, 2005; pp 625–632.

(129) Chen, B.; Greenside, P.; Paik, H.; Sirota, M.; Hadley, D.; Butte, A. *CPT Pharmacometrics Syst. Pharmacol.* **2015**, *4* (10), 576–584.

(130) Tiikkainen, P.; Poso, A.; Kallioniemi, O. *J. Comput. Aided. Mol. Des.* **2009**, *23* (4), 227–239.

(131) Shah, F.; Greene, N. *Chem. Res. Toxicol.* **2014**, *27* (1), 86–98.

(132) Berg, E. L.; Kunkel, E. J.; Hytopoulos, E.; Plavec, I. *J. Pharmacol. Toxicol. Methods* **2006**, *53* (1), 67–74.

(133) Kavlock, R.; Chandler, K.; Houck, K.; Hunter, S.; Judson, R.; Kleinstreuer, N.; Knudsen, T.; Martin, M.; Padilla, S.; Reif, D.; Richard, A.; Rotroff, D.; Sipes, N.; Dix, D. *Chem. Res. Toxicol.* **2012**, *25* (7), 1287–1302.

(134) PubChemPy, version 1.0.3, Matt Swain, Eka A. Kurniawan, 2015.

(135) R Core Team. R: A Language and Environment for Statistical Computing, 2016.

(136) Dessau, R. B.; Pipper, C. B. *Ugeskr. Laeger* **2008**, *170* (5), 328–330.

(137) rstanarm: Bayesian applied regression modeling via Stan, version 2.17.2, Stan Development Team, 2016.

(138) Muggeo, V. M. R. *Stat. Med.* **2003**, *22* (19), 3055–3071.

(139) Vehtari, A.; Gelman, A.; Gabry, J. *Stat. Comput.* **2017**, *27* (5), 1413–1432.

(140) Cruz-Monteagudo, M.; Medina-Franco, J. L.; Pérez-Castillo, Y.; Nicolotti, O.; Cordeiro, M. N. D. S.; Borges, F. *Drug Discov. Today* **2014**, *19* (8), 1069–1080.

(141) Sawada, R.; Kotera, M.; Yamanishi, Y. *Mol. Inform.* **2014**, *33* (11-12), 719–731.

(142) Reker, D.; Rodrigues, T.; Schneider, P.; Schneider, G. *Proc. Natl. Acad. Sci. U. S. A.* **2014**, *111* (11), 4067–4072.

(143) Abdel, B.; Mercier, G. In *Self-Organizing Maps*; InTech, 2010.

(144) Gene Ontology Consortium. *Nucleic Acids Res.* **2004**, *32* (90001), 258D–261.

(145) Bielefeld-Sevigny, M. *Assay Drug Dev. Technol.* **2009**, *7* (1), 90–92.

(146) Cherkasov, A.; Muratov, E. N.; Fourches, D.; Varnek, A.; Baskin, I. I.; Cronin, M.; Dearden, J.; Gramatica, P.; Martin, Y. C.; Todeschini, R.; Consonni, V.; Kuz'min, V. E.; Cramer, R.; Benigni, R.; Yang, C.; Rathman, J.; Terfloth, L.; Gasteiger, J.; Richard, A.; Tropsha, A. *J. Med. Chem.* **2014**, *57* (12), 4977–5010.

(147) Kohonen, T. *Neurocomputing* **1998**, *21* (1–3), 1–6.

(148) Xiao, Y. De; Clauset, A.; Harris, R.; Bayram, E.; Santago, P.; Schmitt, J. D. *J. Chem. Inf. Model.* **2005**, *45* (6), 1749–1758.

(149) Melssen, W.; Wehrens, R.; Buydens, L. *Chemom. Intell. Lab. Syst.* **2006**, *83* (2), 99–113.

(150) Wehrens, R.; Buydens, L. M. C. *J. Stat. Softw.* **2007**, *21* (5), 1–19.

(151) Fröhlich, H.; Speer, N.; Poustka, A.; Beißbarth, T. *BMC Bioinformatics* **2007**, *8* (1), 166.

(152) Aronis, K. N.; Chamberland, J. P.; Mantzoros, C. S. *Metabolism* **2013**, *62* (9), 1279–1286.

(153) Jain, R. K.; Duda, D. G.; Clark, J. W.; Loeffler, J. S. *Nat. Clin. Pract. Oncol.* **2006**, *3* (1), 24–40.

(154) Shojaei, F. *Cancer Lett.* **2012**, *320* (2), 130–137.

(155) Allen-Mersh, T. .; Earlam, S.; Fordy, C.; Abrams, K.; Houghton, J. *Lancet* **1994**, *344* (8932), 1255–1260.

(156) Sharma, A.; Bettis, D. I.; Cowden, J. W.; Mohan, R. R. *Mol. Vis.* **2010**, *16*, 720–728.

(157) Dworacka, M.; Iskakova, S.; Krzyżagórska, E.; Wesołowska, A.; Kurmambayev, Y.; Dworacki, G. *Diabetes Res. Clin. Pract.* **2015**, *107* (2), 273–279.

(158) Layden Brian T; Durai Vivek; L, L. W. J. *Nat. Educ.* **2010**, *3* (9), 13.

(159) Basu, S.; Nagy, J. A.; Pal, S.; Vasile, E.; Eckelhoefer, I. A.; Susan Bliss, V.; Manseau, E. J.; Dasgupta, P. S.; Dvorak, H. F.; Mukhopadhyay, D. *Nat. Med.* **2001**, *7* (5), 569–574.

(160) Yoon, J. C.; Ng, A.; Kim, B. H.; Bianco, A.; Xavier, R. J.; Elledge, S. J. *Genes Dev.* **2010**, *24* (14), 1507–1518.

(161) Miyazaki, R.; Ichiki, T.; Hashimoto, T.; Ikeda, J.; Kamiharaguchi, A.; Narabayashi, E.; Matsuura, H.; Takeda, K.; Sunagawa, K. *Clin. Sci.* **2012**, *123* (4), 241–249.

(162) Li, S.; Kumar T, P.; Joshee, S.; Kirschstein, T.; Subburaju, S.; Khalili, J. S.; Kloepper, J.; Du, C.; Elkhal, A.; Szabó, G.; Jain, R. K.; Köhling, R.; Vasudevan, A. *Cell Res.* **2017**.

(163) Auchampach, J. A. *Circ. Res.* **2007**, *101* (11), 1075–1077.

(164) Ellis, L.; Hammers, H.; Pili, R. *Cancer Lett.* **2009**, *280* (2), 145–153.

(165) Moriya, J.; Ferrara, N. *Cardiovasc. Diabetol.* **2015**, *14*, 19.

(166) Browne, C. D.; Hindmarsh, E. J.; Smith, J. W. *FASEB J.* **2006**, *20* (12), 2027–2035.

(167) Rivera, C. G.; Bader, J. S.; Popel, A. S. *Ann. Biomed. Eng.* **2011**, *39* (8), 2213–2222.

(168) Biyashev, D.; Veliceasa, D.; Kwiatek, A.; Sutanto, M. M.; Cohen, R. N.; Volpert, O. V. *J. Biol. Chem.* **2010**, *285* (18), 13517–13524.

(169) Esposito, G.; Gigli, S.; Seguella, L.; Nobile, N.; D'Alessandro, A.; Pesce, M.;

Capoccia, E.; Steardo, L.; Cirillo, C.; Cuomo, R.; Sarnelli, G. *Int. J. Oncol.* **2016**, *49* (2), 639–645.

(170) Egami, K.; Murohara, T.; Aoki, M.; Matsuishi, T. *J. Leukoc. Biol.* **2006**, *79* (5), 971–976.

(171) Gerritsen, M. E.; Tomlinson, J. E.; Zlot, C.; Ziman, M.; Hwang, S. *Br. J. Pharmacol.* **2003**, *140* (4), 595–610.

(172) Yazama, H.; Kitatani, K.; Fujiwara, K.; Kato, M.; Hashimoto-Nishimura, M.; Kawamoto, K.; Hasegawa, K.; Kitano, H.; Bielawska, A.; Bielawski, J.; Okazaki, T. *Int. J. Clin. Oncol.* **2015**, *20* (3), 438–446.

(173) Spradley, F. T.; Palei, A. C.; Granger, J. P. *Physiol. Rep.* **2013**, *1* (4), e00081.

(174) Schönenberger, D.; Rajski, M.; Harlander, S.; Frew, I. J. *Oncotarget* **2016**, *7* (38), 60971–60985.

(175) Braun, M.; Ramracheya, R.; Bengtsson, M.; Zhang, Q.; Karanauskaite, J.; Partridge, C.; Johnson, P. R.; Rorsman, P. *Diabetes* **2008**, *57* (6), 1618–1628.

(176) Vaswani, K.; Chan, H.-W.; Verma, P.; Dekker Nitert, M.; Peiris, H. N.; Wood-Bradley, R. J.; Armitage, J. A.; Rice, G. E.; Mitchell, M. D. *Reprod. Biol. Endocrinol.* **2015**, *13* (1), 89.

(177) Jandeleit-Dahm, K. A.; Tikellis, C.; Reid, C. M.; Johnston, C. I.; Cooper, M. E. *J. Hypertens.* **2005**, *23* (3), 463–473.

(178) Scheen, A. J. *Diabetes Metab.* **2004**, *30* (6), 487–496.

(179) Wang, S.; Hirschberg, R. *Microvasc. Res.* **2009**, *78* (2), 174–179.

(180) Jiang, Y.; Nyengaard, J. R.; Zhang, J. S.; Jakobsen, J. *Diabetes* **2004**, *53* (10), 2669–2675.

(181) Yu, X.; Murao, K.; Sayo, Y.; Imachi, H.; Cao, W. M.; Ohtsuka, S.; Niimi, M.; Tokumitsu, H.; Inuzuka, H.; Wong, N. C. W.; Kobayashi, R.; Ishida, T. *Diabetes* **2004**, *53* (6), 1475–1481.

(182) Julien, S.; Laine, J.; Morisset, J. *Diabetes* **2004**, *53* (6), 1526–1534.

(183) Anstadt, M. P.; Hutchinson, J.; Portik-Dobos, V.; Jafri, F.; Bannan, M.; Mawulawde, K.; Ergul, A. *Ethn. Dis.* **2002**, *12* (4), S3-5–9.

(184) Dayer, M. R.; Mard-Soltani, M.; Dayer, M. S.; Alavi, S. M. R. *Med. J. Islam. Repub. Iran* **2014**, *28*, 59.

(185) Jorgensen, R.; Martini, L.; Schwartz, T. W.; Elling, C. E. *Mol. Endocrinol.* **2005**, *19* (3), 812–823.

(186) Anjaneyulu, M.; Berent-Spillson, A.; Russell, J. W. *Curr. Drug Targets* **2008**, *9* (1), 85–93.

(187) Bellini, S.; Barutta, F.; Mastrocola, R.; Imperatore, L.; Bruno, G.; Gruden, G. *Int. J. Mol. Sci.* **2017**, *18* (12).

(188) Ning, A.; Cui, J.; Maberley, D.; Ma, P.; Matsubara, J. *Can. J. Ophthalmol.* **2008**, *43* (6), 683–688.

(189) Obici, S.; Feng, Z.; Tan, J.; Liu, L.; Karkanias, G.; Rossetti, L. *J. Clin. Invest.* **2001**, *108* (7), 1079–1085.

(190) Zhang, L.; Dong, L.; Liu, X.; Jiang, Y.; Zhang, L.; Zhang, X.; Li, X.; Zhang, Y. *PLoS*

One **2014**, *9* (4), e93433.

(191) Lv, L.; Zhang, J.; Zhang, L.; Xue, G.; Wang, P.; Meng, Q.; Liang, W. *J. Cell. Mol. Med.* **2013**, *17* (8), 989–1005.

(192) Koya, D.; King, G. L. *Diabetes* **1998**, *47* (6), 859–866.

(193) Vennemann, A.; Gerstner, A.; Kern, N.; Ferreiros Bouzas, N.; Narumiya, S.; Maruyama, T.; Nüsing, R. M. *Diabetes* **2012**, *61* (7), 1879–1887.

(194) Kimple, M. E.; Keller, M. P.; Rabaglia, M. R.; Pasker, R. L.; Neuman, J. C.; Truchan, N. A.; Brar, H. K.; Attie, A. D. *Diabetes* **2013**, *62* (6), 1904–1912.

(195) Wu, H.; Panakanti, R.; Li, F.; Mahato, R. I. *Mol. Pharm.* **2010**, *7* (5), 1655–1666.

(196) Fang, J. Y.; Richardson, B. C. *Lancet Oncol.* **2005**, *6* (5), 322–327.

(197) Urosevic, J.; Nebreda, A. R.; Gomis, R. R. *Cell Cycle* **2014**, *13* (17), 2641–2642.

(198) Lee, H.; Flaherty, P.; Ji, H. P. *BMC Med. Genomics* **2013**, *6* (1).

(199) Von Rosenvinge, E. C.; Raufman, J.-P. *Cancers (Basel).* **2011**, *3* (1), 971–981.

(200) De Sousa E Melo, F.; Wang, X.; Jansen, M.; Fessler, E.; Trinh, A.; De Rooij, L. P. M. H.; De Jong, J. H.; De Boer, O. J.; Van Leersum, R.; Bijlsma, M. F.; Rodermond, H.; Van Der Heijden, M.; Van Noesel, C. J. M.; Tuynman, J. B.; Dekker, E.; Markowetz, F.; Medema, J. P.; Vermeulen, L. *Nat. Med.* **2013**, *19* (5), 614–618.

(201) Ferracin, M.; Gafà, R.; Miotto, E.; Veronese, A.; Pultrone, C.; Sabbioni, S.; Lanza, G.; Negrini, M. *J. Pathol.* **2008**, *214* (5), 594–602.

(202) Tutton, P. J.; Steel, G. G. *Br. J. Cancer* **1979**, *40* (5), 743–749.

(203) Nylund, G.; Pettersson, A.; Bengtsson, C.; Khorram-Manesh, A.; Nordgren, S.; Delbro, D. S. *Dig. Dis. Sci.* **2008**, *53* (2), 461–466.

(204) Xie, G.; Raufman, J.-P. *J. Cancer Metastasis Treat.* **2016**, *2* (6), 195.

(205) Felton, J.; Hu, S.; Raufman, J.-P. *Curr. Mol. Pharmacol.* **2018**, *11*.

(206) Liang, B.; Li, C.; Zhao, J. *Med. Oncol.* **2016**, *33* (10), 111.

(207) Ieta, K.; Tanaka, F.; Haraguchi, N.; Kita, Y.; Sakashita, H.; Mimori, K.; Matsumoto, T.; Inoue, H.; Kuwano, H.; Mori, M. *Ann. Surg. Oncol.* **2008**, *15* (2), 638–648.

(208) Joyce, T.; Oikonomou, E.; Kosmidou, V.; Makrodouli, E.; Bantounas, I.; Avlonitis, S.; Zografos, G.; Pintzas, a. *Curr. Cancer Drug Targets* **2012**, *12* (7), 873–898.

(209) İşeri, Ö. D.; Sahin, F. I.; Terzi, Y. K.; Yurtcu, E.; Erdem, S. R.; Sarialioglu, F. *Pharm. Biol.* **2014**, *52* (11), 1374–1381.

(210) Takezaki, T.; Hamajima, N.; Matsuo, K.; Tanaka, R.; Hirai, T.; Kato, T.; Ohashi, K.; Tajima, K. *Int. J. Clin. Oncol.* **2001**, *6* (3), 117–122.

(211) Ahmadi, A. A.; Shadifar, M.; Ataee, R.; Vaillancourt, C.; Ataee, A.; Oufkir, T.; Jafari-Sabet, M. *Int. Biol. Biomed. J.* **2015**, *1* (2), 56–65.

(212) Wang, X. T.; Nagaba, Y.; Cross, H. S.; Wrba, F.; Zhang, L.; Guggino, S. E. *Am. J. Pathol.* **2000**, *157* (5), 1549–1562.

(213) Zawadzki, A.; Liu, Q.; Wang, Y.; Melander, A.; Jeppsson, B.; Thorlacius, H. *Dis. Colon Rectum* **2008**, *51* (11), 1696–1702.

(214) Lastraioli, E.; Guasti, L.; Crociani, O.; Polvani, S.; Hofmann, G.; Witchel, H.; Bencini, L.; Calistri, M.; Messerini, L.; Scatizzi, M.; Moretti, R.; Wanke, E.; Olivotto, M.;

Mugnai, G.; Arcangeli, A. *Cancer Res.* **2004**, *64* (2), 606–611.

(215) House, C. D.; Vaske, C. J.; Schwartz, A. M.; Obias, V.; Frank, B.; Luu, T.; Sarvazyan, N.; Irby, R.; Strausberg, R. L.; Hales, T. G.; Stuart, J. M.; Lee, N. H. *Cancer Res.* **2010**, *70* (17), 6957–6967.

(216) Goetz, M. P.; Suman, V. J.; Hoskin, T. L.; Gnant, M.; Filipits, M.; Safgren, S. L.; Kuffel, M.; Jakesz, R.; Rudas, M.; Greil, R.; Dietze, O.; Lang, A.; Offner, F.; Reynolds, C. A.; Weinshilboum, R. M.; Ames, M. M.; Ingle, J. N. *Clin. Cancer Res.* **2013**, *19* (2), 500–507.

(217) Bigler, J.; Whitton, J.; Lampe, J. W.; Fosdick, L.; Bostick, R. M.; Potter, J. D. *Cancer Res.* **2001**, *61* (9), 3566–3569.

(218) Sachse, C.; Bhambra, U.; Smith, G.; Lightfoot, T. J.; Barrett, J. H.; Scollay, J.; Garner, R. C.; Boobis, A. R.; Wolf, C. R.; Gooderham, N. J. *Br. J. Clin. Pharmacol.* **2003**, *55* (1), 68–76.

(219) Ataee, R.; Ajdary, S.; Zarrindast, M.; Rezayat, M.; Hayatbakhsh, M. R. *J. Cancer Res. Clin. Oncol.* **2010**, *136* (10), 1461–1469.

(220) Sui, H.; Xu, H.; Ji, Q.; Liu, X.; Zhou, L.; Song, H.; Zhou, X.; Xu, Y.; Chen, Z.-S.; Ji, G.; Li, Q. *Oncotarget* **2015**, *6* (28), 25975–25987.

(221) Xiang, T.; Fei, R.; Wang, Z.; Shen, Z.; Qian, J.; Chen, W. *Oncol. Rep.* **2016**, *35* (1), 205–210.

(222) Laws, S.; Gough, A.; Evans, A.; Bains, M.; Primrose, J. *Br. J. Cancer* **1997**, *75* (3), 360–366.

(223) Spano, J.-P.; Lagorce, C.; Atlan, D.; Milano, G.; Domont, J.; Benamouzig, R.; Attar,

A.; Benichou, J.; Martin, A.; Morere, J.-F.; Raphael, M.; Penault-Llorca, F.; Breau, J.-L.; Fagard, R.; Khayat, D.; Wind, P. *Ann. Oncol.* **2005**, *16* (1), 102–108.

(224) Kumar, R.; Srinivasan, S.; Pahari, P.; Rohr, J.; Damodaran, C. *Mol. Cancer Ther.* **2010**, *9* (9), 2488–2496.

(225) Najar, A. G.; Pashaei-Asl, R.; Omidi, Y.; Farajnia, S.; Nourazarian, A. R. *Asian Pacific J. Cancer Prev.* **2013**, *14* (1), 495–498.

(226) Piedra, J.; Miravet, S.; Castaño, J.; Pálmer, H. G.; Heisterkamp, N.; García de Herreros, A.; Duñach, M. *Mol. Cell. Biol.* **2003**, *23* (7), 2287–2297.

(227) Radhakrishnan, V. M.; Kojs, P.; Young, G.; Ramalingam, R.; Jagadish, B.; Mash, E. A.; Martinez, J. D.; Ghishan, F. K.; Kiela, P. R. *PLoS One* **2014**, *9* (1), e85796.

(228) Aisner, D. L.; Nguyen, T. T.; Paskulin, D. D.; Le, A. T.; Haney, J.; Schulte, N.; Chionh, F.; Hardingham, J.; Mariadason, J.; Tebbutt, N.; Doebele, R. C.; Weickhardt, A. J.; Varella-Garcia, M. *Mol. Cancer Res.* **2014**, *12* (1), 111–118.

(229) Li, N.; Wang, C.; Wu, Y.; Liu, X.; Cao, X. *J. Biol. Chem.* **2009**, *284* (5), 3021–3027.

(230) Summy, J. M.; Gallick, G. E. *Cancer Metastasis Rev.* **2003**, *22* (4), 337–358.

(231) Kim, L. C.; Song, L.; Haura, E. B. *Nat. Rev. Clin. Oncol.* **2009**, *6* (10), 587–595.

(232) Carter, J. H.; Cottrell, C. E.; McNulty, S. N.; Vigh-Conrad, K. A.; Lamp, S.; Heusel, J. W.; Duncavage, E. J. *Mol. Case Stud.* **2017**, *3* (6), a001495.

(233) Wei, C.-H.; Wu, G.; Cai, Q.; Gao, X.-C.; Tong, F.; Zhou, R.; Zhang, R.-G.; Dong, J.-H.; Hu, Y.; Dong, X.-R. *J. Hematol. Oncol.* **2017**, *10* (1), 125.

(234) Palaniappan, A.; Ramar, K.; Ramalingam, S. *PLoS One* **2016**, *11* (5), e0156665.

(235) Espinosa, L.; Navarro, E. *Cytogenet. Genome Res.* **1998**, *81* (3–4), 278–282.

(236) Luo, Y.; Tsuchiya, K. D.; Il Park, D.; Fausel, R.; Kanngurn, S.; Welcsh, P.; Dzieciatkowski, S.; Wang, J.; Grady, W. M. *Oncogene* **2013**, *32* (16), 2037–2047.

(237) Tan, Y.-C.; Blumenfeld, J.; Rennert, H. *Biochim. Biophys. Acta* **2011**, *1812* (10), 1202–1212.

(238) Yoder, B. K.; Mulroy, S.; Eustace, H.; Boucher, C.; Sandford, R. *Expert Rev. Mol. Med.* **2006**, *8* (2), 1–22.

(239) Spithoven, E. M.; Kramer, A.; Meijer, E.; Orskov, B.; Wanner, C.; Abad, J. M.; Areste, N.; Alonso de la Torre, R.; Caskey, F.; Couchoud, C.; Finne, P.; Heaf, J.; Hoitsma, A.; de Meester, J.; Pascual, J.; Postorino, M.; Ravani, P.; Zurriaga, O.; Jager, K. J.; Gansevoort, R. T.; de los Angeles Garcia Bazaga, M.; Metcalfe, W.; Rodrigo, E.; Quiros, J. R.; the EuroCYST Consortium, K.; Budde, K.; Devuyst, O.; Ecder, T.; Eckardt, K. U.; Gansevoort, R. T.; Kottgen, A.; Ong, A. C.; Petzold, K.; Pirson, Y.; Remuzzi, G.; Torra, R.; Sandford, R. N.; Serra, A. L.; Tesar, V.; Walz, G.; the WGIKD, C.; Wuthrich, R. P.; Antignac, C.; Bindels, R.; Chauveau, D.; Devuyst, O.; Emma, F.; Gansevoort, R. T.; Maxwell, P. H.; Ong, A. C.; Remuzzi, G.; Ronco, P.; Schaefer, F. *Nephrol. Dial. Transplant.* **2014**, *29* (suppl 4), iv15-iv25.

(240) Chapin, H. C.; Caplan, M. J. *J. Cell Biol.* **2010**, *191* (4), 701–710.

(241) Graham, P. C.; Lindop, G. B. M. *Kidney Int.* **1988**, *33*, 1084–1090.

(242) Chapman, A. B.; Johnson, A.; Gabow, P. A.; Schrier, R. W. *N. Engl. J. Med.* **1990**, *323* (16), 1091–1096.

(243) Chapman, A. B.; Stepniakowski, K.; Rahbari-Oskoui, F. *Adv. Chronic Kidney Dis.*

2010, *17* (2), 153–163.

(244) Song, C. J.; Zimmerman, K. A.; Henke, S. J.; Yoder, B. K. In *Results and problems in cell differentiation*; 2017; Vol. 60, pp 323–344.

(245) van Dijk, M. A.; Breuning, M. H.; Duiser, R.; van Es, L. A.; Westendorp, R. G. J. *Nephrol. Dial. Transplant* **2003**, *18* (11), 2314–2320.

(246) Natoli, T. A.; Smith, L. A.; Rogers, K. A.; Wang, B.; Komarnitsky, S.; Budman, Y.; Belenky, A.; Bukanov, N. O.; Dackowski, W. R.; Husson, H.; Russo, R. J.; Shayman, J. A.; Ledbetter, S. R.; Leonard, J. P.; Ibraghimov-Beskrovnaya, O. *Nat. Med.* **2010**, *16* (7), 788–792.

(247) Blazer-Yost, B. L.; Haydon, J.; Eggleston-Gulyas, T.; Chen, J.-H.; Wang, X.; Gattone, V.; Torres, V. E. *PPAR Res.* **2010**, *2010*, 274376.

(248) Dai, B.; Liu, Y.; Mei, C.; Fu, L.; Xiong, X.; Zhang, Y.; Shen, X.; Hua, Z. *Clin. Sci.* **2010**, *119* (8).

(249) Nofziger, C.; Brown, K. K.; Smith, C. D.; Harrington, W.; Murray, D.; Bisi, J.; Ashton, T. T.; Maurio, F. P.; Kalsi, K.; West, T. A.; Baines, D.; Blazer-Yost, B. L. *Am. J. Physiol. - Ren. Physiol.* **2009**, *297* (1).

(250) Li, H.; Sheppard, D. N. *BioDrugs* **2009**, *23* (4), 203–216.

(251) Liu, C.; Li, H.; Gao, X.; Yang, M.; Yuan, L.; Fu, L.; Wang, X.; Mei, C. *Am. J. Physiol. - Ren. Physiol.* **2016**.

(252) Bukanov, N. O.; Smith, L. A.; Klinger, K. W.; Ledbetter, S. R.; Ibraghimov-Beskrovnaya, O. *Nature* **2006**, *444* (7121), 949–952.

(253) Leuenroth, S. J.; Okuhara, D.; Shotwell, J. D.; Markowitz, G. S.; Yu, Z.; Somlo, S.; Crews, C. M. *Proc. Natl. Acad. Sci. U. S. A.* **2007**, *104* (11), 4389–4394.

(254) Irazabal, M. V.; Torres, V. E.; Hogan, M. C.; Glockner, J.; King, B. F.; Ofstie, T. G.; Krasa, H. B.; Ouyang, J.; Czerwiec, F. S. *Kidney Int.* **2011**, *80* (3), 295–301.

(255) Meijer, E.; de Jong, P. E.; Peters, D. J.; Gansevoort, R. T. *J. Nephrol.* **2008**, *21* (2), 133–138.

(256) European Medicines Agency. Public assessment report - Jinarc http://www.ema.europa.eu/docs/en_GB/document_library/EPAR_-_Public_assessment_report/human/002788/WC500187923.pdf (accessed Oct 15, 2017).

(257) National institute for health and care excellence. Final appraisal determination - Tolvaptan for treating autosomal dominant polycystic kidney disease https://www.nice.org.uk/guidance/TA358/documents/kidney-disease-autosomal-dominant-polycystic-tolvaptan-id652-final-appraisal-determination-document2 (accessed Nov 26, 2017).

(258) Ohnishi, A.; Orita, Y.; Takagi, N.; Fujita, T.; Toyoki, T.; Ihara, Y.; Yamamura, Y.; Inoue, T.; Tanaka, T. *J. Pharmacol. Exp. Ther.* **1995**, *272* (2).

(259) Booij, T. H.; Bange, H.; Leonhard, W. N.; Yan, K.; Fokkelman, M.; Kunnen, S. J.; Dauwerse, J. G.; Qin, Y.; van de Water, B.; van Westen, G. J. P.; Peters, D. J. M.; Price, L. S. *SLAS Discov. Adv. Life Sci. R&D* **2017**, *22* (8), 974–984.

(260) Spectrum collection, MicroSource Discovery System Inc., Gaylordsville, CT 06755 USA, http://www.msdiscovery.com/spectrum.html (accessed Jan 10, 2015).

(261) SCHRÖDINGER Canvas, version 2.5.015, Schrödinger, LLC, New York, NY, USA, 2015.

(262) Koscielny, G.; An, P.; Carvalho-Silva, D.; Cham, J. A.; Fumis, L.; Gasparyan, R.; Hasan, S.; Karamanis, N.; Maguire, M.; Papa, E.; Pierleoni, A.; Pignatelli, M.; Platt, T.; Rowland, F.; Wankar, P.; Bento, A. P.; Burdett, T.; Fabregat, A.; Forbes, S.; Gaulton, A.; Gonzalez, C. Y.; Hermjakob, H.; Hersey, A.; Jupe, S.; Kafkas, Ş.; Keays, M.; Leroy, C.; Lopez, F.-J.; Magarinos, M. P.; Malone, J.; McEntyre, J.; Munoz-Pomer Fuentes, A.; O'Donovan, C.; Papatheodorou, I.; Parkinson, H.; Palka, B.; Paschall, J.; Petryszak, R.; Pratanwanich, N.; Sarntivijal, S.; Saunders, G.; Sidiropoulos, K.; Smith, T.; Sondka, Z.; Stegle, O.; Tang, Y. A.; Turner, E.; Vaughan, B.; Vrousgou, O.; Watkins, X.; Martin, M.-J.; Sanseau, P.; Vamathevan, J.; Birney, E.; Barrett, J.; Dunham, I. *Nucleic Acids Res.* **2017**, *45* (D1), D985–D994.

(263) King, B. L.; Davis, A. P.; Rosenstein, M. C.; Wiegers, T. C.; Mattingly, C. J. *PLoS One* **2012**, *7* (11).

(264) Davis, a. P.; Grondin, C. J.; Lennon-Hopkins, K.; Saraceni-Richards, C.; Sciaky, D.; King, B. L.; Wiegers, T. C.; Mattingly, C. J. *Nucleic Acids Res.* **2014**, *43* (D1), D914–D920.

(265) Malas, T. B.; Formica, C.; Leonhard, W. N.; Rao, P.; Granchi, Z.; Roos, M.; Peters, D. J. M.; 't Hoen, P. A. C. *Am. J. Physiol. - Ren. Physiol.* **2017**, *312* (4), F806–F817.

(266) SCHRÖDINGER Maestro version 10.3.015, Schrödinger, LLC, New York, NY, USA, 2015.

(267) Halgren, T. A. *J. Chem. Inf. Model.* **2009**, *49* (2), 377–389.

(268) Shelley, J. C.; Cholleti, A.; Frye, L. L.; Greenwood, J. R.; Timlin, M. R.; Uchimaya, M. *J. Comput. Aided. Mol. Des.* **2007**, *21* (12), 681–691.

(269) Halgren, T. A.; Murphy, R. B.; Friesner, R. A.; Beard, H. S.; Frye, L. L.; Pollard, W. T.; Banks, J. L. *J. Med. Chem.* **2004**, *47* (7), 1750–1759.

(270) Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shelley, M.; Perry, J. K.; Shaw, D. E.; Francis, P.; Shenkin, P. S. *J. Med. Chem.* **2004**, *47* (7), 1739–1749.

(271) Friesner, R. A.; Murphy, R. B.; Repasky, M. P.; Frye, L. L.; Greenwood, J. R.; Halgren, T. A.; Sanschagrin, P. C.; Mainz, D. T. *J. Med. Chem.* **2006**, *49* (21), 6177–6196.

(272) Elumalai, R.; Periasamy, S.; Ramanathan, G.; Lakkakula, B. V. *J. Ren. Inj. Prev.* **2014**, *3* (3), 69–73.

(273) Wiwanitkit, V. *J. nephropharmacology* **2015**, *4* (2), 85–87.

(274) Kumar, V.; Sanseau, P.; Simola, D. F.; Hurle, M. R.; Agarwal, P. *Sci. Rep.* **2016**, *6* (1), 36205.

(275) Rabalski, A. J.; Gyenis, L.; Litchfield, D. W. *Clin. Cancer Res.* **2016**, *22* (12), 2840–2847.

(276) Tao, Y.; Kim, J.; Stanley, M.; He, Z.; Faubel, S.; Schrier, R. W.; Edelstein, C. L. *Kidney Int.* **2005**, *67* (3), 909–919.

(277) Zhang, G.; Kernan, K. A.; Collins, S. J.; Cai, X.; Lopez-Guisa, J. M.; Degen, J. L.; Shvil, Y.; Eddy, A. A. *J. Am. Soc. Nephrol.* **2007**, *18* (3), 846–859.

(278) Fragiadaki, M.; Mason, R. M. *Int. J. Exp. Pathol.* **2011**, *92* (3), 143–150.

(279) Norman, J. *Biochim. Biophys. Acta - Mol. Basis Dis.* **2011**, *1812* (10), 1327–1336.

(280) Hannon, J.; Hoyer, D. *Behav. Brain Res.* **2008**, *195* (1), 198–213.

(281) Erikci, A.; Ucar, G.; Yabanoglu-Ciftci, S. *Ren. Fail.* **2016**, *38* (7), 1141–1150.

(282) Yan, F.-X.; Langub, M. C.; Ihnen, M. A.; Hornung, C.; Juronen, E.; Rayens, M. K.; Cai, W.-M.; Wedlund, P. J.; Fanti, P. *Br. J. Clin. Pharmacol.* **2003**, *56* (1), 68–77.

(283) Smolenicka, Z.; Bach, E.; Schaer, A.; Liechti-Gallati, S.; Frey, B. M.; Frey, F. J.; Ferrari, P. *J. Clin. Endocrinol. Metab.* **1998**, *83* (5), 1814–1814.

(284) Lee, J.; Lee, Y.; Park, B.; Won, S.; Han, J. S.; Heo, N. J. *PLoS One* **2018**, *13* (3), e0194044.

(285) Kaunisto, K.; Parkkila, S.; Rajaniemi, H.; Waheed, A.; Grubb, J.; Sly, W. S. *Kidney Int.* **2002**, *61* (6), 2111–2118.

(286) Mrug, M.; Zhou, J.; Woo, Y.; Cui, X.; Szalai, A. J.; Novak, J.; Churchill, G. A.; Guay-Woodford, L. M. *Kidney Int.* **2008**, *73* (1), 63–76.

(287) Yu, H.; Song, Q.; Freedman, B. I.; Chao, J.; Chao, L.; Rich, S. S.; Bowden, D. W. *Kidney Int.* **2002**, *61* (3), 1030–1039.

Appendices

Appendix A. Structural features corresponding to the 166 keys of the MACCS MDL keys.[46] Atom symbols are the following: A: Any valid periodic table element symbol; Q: Heteroatoms; any non-C or non-H atom; X: Halogens; F, Cl, Br, I; Z: Others; other than H, C, N, O, Si, P, S, F, Cl, Br, I

KEY	STRUCTURAL FEATURE
1	ISOTOPE
2	103 < ATOMIC NO. < 256
3	GROUP IVA, VA, VIA PERIODS 4-6 (Ge...)
4	ACTINIDE
5	GROUP IIIB, IVB (Sc...)
6	LANTHANIDE
7	GROUP VB, VIB, VIIB (V...)
8	QAAA@1
9	GROUP VIII (Fe...)
10	GROUP IIA (ALKALINE EARTH)
11	4M RING
12	GROUP IB, IIB (Cu...)
13	ON(C)C
14	S-S
15	OC(O)O
16	QAA@1
17	CTC
18	GROUP IIIA (B...)
19	7M RING
20	SI
21	C=C(Q)Q
22	3M RING
23	NC(O)O
24	N-O
25	NC(N)N
26	C$=C($A)$A
27	I
28	QCH2Q
29	P
30	CQ(C)(C)A
31	QX
32	CSN
33	NS
34	CH2=A
35	GROUP IA (ALKALI METAL)

36	S HETEROCYCLE
37	NC(O)N
38	NC(C)N
39	OS(O)O
40	S-O
41	CTN
42	F
43	QHAQH
44	OTHER
45	C=CN
46	BR
47	SAN
48	OQ(O)O
49	CHARGE
50	C=C(C)C
51	CSO
52	NN
53	QHAAAQH
54	QHAAQH
55	OSO
56	ON(O)C
57	O HETEROCYCLE
58	QSQ
59	Snot%A%A
60	S=O
61	AS(A)A
62	A$A!A$A
63	N=O
64	A$A!S
65	C%N

66	CC(C)(C)A
67	QS
68	QHQH (&...)
69	QQH
70	QNQ
71	NO
72	OAAO
73	S=A
74	CH3ACH3
75	A!N$A
76	C=C(A)A
77	NAN
78	C=N
79	NAAN
80	NAAAN
81	SA(A)A
82	ACH2QH
83	QAAAA@1
84	NH2
85	CN(C)C
86	CH2QCH2
87	X!A$A
88	S
89	OAAAO
90	QHAACH2A
91	QHAAACH2A
92	OC(N)C
93	QCH3
94	QN
95	NAAO

96	5M RING
97	NAAAO
98	QAAAAA@1
99	C=C
100	ACH2N
101	8M RING
102	QO
103	CL
104	QHACH2A
105	A$A($A)$A
106	QA(Q)Q
107	XA(A)A
108	CH3AAACH2A
109	ACH2O
110	NCO
111	NACH2A
112	AA(A)(A)A
113	Onot%A%A
114	CH3CH2A
115	CH3ACH2A
116	CH3AACH2A
117	NAO
118	ACH2CH2A > 1
119	N=A
120	HETEROCYCLIC ATOM > 1 (&...)
121	N HETEROCYCLE
122	AN(A)A
123	OCO
124	QQ
125	AROMATIC RING > 1

126	A!O!A
127	A$A!O > 1 (&...)
128	ACH2AAACH2A
129	ACH2AACH2A
130	QQ > 1 (&...)
131	QH > 1
132	OACH2A
133	A$A!N
134	X (HALOGEN)
135	Nnot%A%A
136	O=A > 1
137	HETEROCYCLE
138	QCH2A > 1 (&...)
139	OH
140	O > 3 (&...)
141	CH3 > 2 (&...)
142	N > 1
143	A$A!O
144	Anot%A%Anot%A
145	6M RING > 1
146	O > 2
147	ACH2CH2A
148	AQ(A)A
149	CH3 > 1
150	A!A$A!A
151	NH
152	OC(C)C
153	QCH2A
154	C=O
155	A!CH2!A

156	NA(A)A
157	C-O
158	C-N
159	O > 1
160	CH3
161	N

162	AROMATIC
163	6M RING
164	O
165	RING
166	FRAGMENTS

Appendix B. Structural features corresponding to the 881 keys of the PubChem fingerprints.[74]

KEY	FEATURE
0	>= 4 H
1	>= 8 H
2	>= 16 H
3	>= 32 H
4	>= 1 Li
5	>= 2 Li
6	>= 1 B
7	>= 2 B
8	>= 4 B
9	>= 2 C
10	>= 4 C
11	>= 8 C
12	>= 16 C
13	>= 32 C
14	>= 1 N
15	>= 2 N
16	>= 4 N
17	>= 8 N
18	>= 1 O
19	>= 2 O
20	>= 4 O
21	>= 8 O
22	>= 16 O
23	>= 1 F
24	>= 2 F
25	>= 4 F
26	>= 1 Na
27	>= 2 Na
28	>= 1 Si
29	>= 2 Si
30	>= 1 P
31	>= 2 P
32	>= 4 P
33	>= 1 S
34	>= 2 S
35	>= 4 S
36	>= 8 S
37	>= 1 Cl
38	>= 2 Cl
39	>= 4 Cl
40	>= 8 Cl
41	>= 1 K
42	>= 2 K
43	>= 1 Br
44	>= 2 Br
45	>= 4 Br
46	>= 1 I
47	>= 2 I
48	>= 4 I
49	>= 1 Be
50	>= 1 Mg
51	>= 1 Al
52	>= 1 Ca

53	>= 1 Sc
54	>= 1 Ti
55	>= 1 V
56	>= 1 Cr
57	>= 1 Mn
58	>= 1 Fe
59	>= 1 Co
60	>= 1 Ni
61	>= 1 Cu
62	>= 1 Zn
63	>= 1 Ga
64	>= 1 Ge
65	>= 1 As
66	>= 1 Se
67	>= 1 Kr
68	>= 1 Rb
69	>= 1 Sr
70	>= 1 Y
71	>= 1 Zr
72	>= 1 Nb
73	>= 1 Mo
74	>= 1 Ru
75	>= 1 Rh
76	>= 1 Pd
77	>= 1 Ag
78	>= 1 Cd
79	>= 1 In
80	>= 1 Sn
81	>= 1 Sb
82	>= 1 Te

83	>= 1 Xe
84	>= 1 Cs
85	>= 1 Ba
86	>= 1 Lu
87	>= 1 Hf
88	>= 1 Ta
89	>= 1 W
90	>= 1 Re
91	>= 1 Os
92	>= 1 Ir
93	>= 1 Pt
94	>= 1 Au
95	>= 1 Hg
96	>= 1 Tl
97	>= 1 Pb
98	>= 1 Bi
99	>= 1 La
100	>= 1 Ce
101	>= 1 Pr
102	>= 1 Nd
103	>= 1 Pm
104	>= 1 Sm
105	>= 1 Eu
106	>= 1 Gd
107	>= 1 Tb
108	>= 1 Dy
109	>= 1 Ho
110	>= 1 Er
111	>= 1 Tm
112	>= 1 Yb

113	>= 1 Tc
114	>= 1 U
115	>= 1 any ring size 3
116	>= 1 saturated or aromatic carbon-only ring size 3
117	>= 1 saturated or aromatic nitrogen-containing ring size 3
118	>= 1 saturated or aromatic heteroatom-containing ring size 3
119	>= 1 unsaturated non-aromatic carbon-only ring size 3
120	>= 1 unsaturated non-aromatic nitrogen-containing ring size 3
121	>= 1 unsaturated non-aromatic heteroatom-containing ring size 3
122	>= 2 any ring size 3
123	>= 2 saturated or aromatic carbon-only ring size 3
124	>= 2 saturated or aromatic nitrogen-containing ring size 3
125	>= 2 saturated or aromatic heteroatom-containing ring size 3
126	>= 2 unsaturated non-aromatic carbon-only ring size 3
127	>= 2 unsaturated non-aromatic nitrogen-containing ring size 3
128	>= 2 unsaturated non-aromatic heteroatom-containing ring size 3
129	>= 1 any ring size 4
130	>= 1 saturated or aromatic carbon-only ring size 4
131	>= 1 saturated or aromatic nitrogen-containing ring size 4
132	>= 1 saturated or aromatic heteroatom-containing ring size 4
133	>= 1 unsaturated non-aromatic carbon-only ring size 4
134	>= 1 unsaturated non-aromatic nitrogen-containing ring size 4
135	>= 1 unsaturated non-aromatic heteroatom-containing ring size 4
136	>= 2 any ring size 4
137	>= 2 saturated or aromatic carbon-only ring size 4
138	>= 2 saturated or aromatic nitrogen-containing ring size 4
139	>= 2 saturated or aromatic heteroatom-containing ring size 4
140	>= 2 unsaturated non-aromatic carbon-only ring size 4
141	>= 2 unsaturated non-aromatic nitrogen-containing ring size 4
142	>= 2 unsaturated non-aromatic heteroatom-containing ring size 4
143	>= 1 any ring size 5
144	>= 1 saturated or aromatic carbon-only ring size 5
145	>= 1 saturated or aromatic nitrogen-containing ring size 5
146	>= 1 saturated or aromatic heteroatom-containing ring size 5
147	>= 1 unsaturated non-aromatic carbon-only ring size 5
148	>= 1 unsaturated non-aromatic nitrogen-containing ring size 5

149	>= 1 unsaturated non-aromatic heteroatom-containing ring size 5
150	>= 2 any ring size 5
151	>= 2 saturated or aromatic carbon-only ring size 5
152	>= 2 saturated or aromatic nitrogen-containing ring size 5
153	>= 2 saturated or aromatic heteroatom-containing ring size 5
154	>= 2 unsaturated non-aromatic carbon-only ring size 5
155	>= 2 unsaturated non-aromatic nitrogen-containing ring size 5
156	>= 2 unsaturated non-aromatic heteroatom-containing ring size 5
157	>= 3 any ring size 5
158	>= 3 saturated or aromatic carbon-only ring size 5
159	>= 3 saturated or aromatic nitrogen-containing ring size 5
160	>= 3 saturated or aromatic heteroatom-containing ring size 5
161	>= 3 unsaturated non-aromatic carbon-only ring size 5
162	>= 3 unsaturated non-aromatic nitrogen-containing ring size 5
163	>= 3 unsaturated non-aromatic heteroatom-containing ring size 5
164	>= 4 any ring size 5
165	>= 4 saturated or aromatic carbon-only ring size 5
166	>= 4 saturated or aromatic nitrogen-containing ring size 5
167	>= 4 saturated or aromatic heteroatom-containing ring size 5
168	>= 4 unsaturated non-aromatic carbon-only ring size 5
169	>= 4 unsaturated non-aromatic nitrogen-containing ring size 5
170	>= 4 unsaturated non-aromatic heteroatom-containing ring size 5
171	>= 5 any ring size 5
172	>= 5 saturated or aromatic carbon-only ring size 5
173	>= 5 saturated or aromatic nitrogen-containing ring size 5
174	>= 5 saturated or aromatic heteroatom-containing ring size 5
175	>= 5 unsaturated non-aromatic carbon-only ring size 5
176	>= 5 unsaturated non-aromatic nitrogen-containing ring size 5
177	>= 5 unsaturated non-aromatic heteroatom-containing ring size 5
178	>= 1 any ring size 6
179	>= 1 saturated or aromatic carbon-only ring size 6
180	>= 1 saturated or aromatic nitrogen-containing ring size 6
181	>= 1 saturated or aromatic heteroatom-containing ring size 6
182	>= 1 unsaturated non-aromatic carbon-only ring size 6
183	>= 1 unsaturated non-aromatic nitrogen-containing ring size 6

184	>= 1 unsaturated non-aromatic heteroatom-containing ring size 6
185	>= 2 any ring size 6
186	>= 2 saturated or aromatic carbon-only ring size 6
187	>= 2 saturated or aromatic nitrogen-containing ring size 6
188	>= 2 saturated or aromatic heteroatom-containing ring size 6
189	>= 2 unsaturated non-aromatic carbon-only ring size 6
190	>= 2 unsaturated non-aromatic nitrogen-containing ring size 6
191	>= 2 unsaturated non-aromatic heteroatom-containing ring size 6
192	>= 3 any ring size 6
193	>= 3 saturated or aromatic carbon-only ring size 6
194	>= 3 saturated or aromatic nitrogen-containing ring size 6
195	>= 3 saturated or aromatic heteroatom-containing ring size 6
196	>= 3 unsaturated non-aromatic carbon-only ring size 6
197	>= 3 unsaturated non-aromatic nitrogen-containing ring size 6
198	>= 3 unsaturated non-aromatic heteroatom-containing ring size 6
199	>= 4 any ring size 6
200	>= 4 saturated or aromatic carbon-only ring size 6
201	>= 4 saturated or aromatic nitrogen-containing ring size 6
202	>= 4 saturated or aromatic heteroatom-containing ring size 6
203	>= 4 unsaturated non-aromatic carbon-only ring size 6
204	>= 4 unsaturated non-aromatic nitrogen-containing ring size 6
205	>= 4 unsaturated non-aromatic heteroatom-containing ring size 6
206	>= 5 any ring size 6
207	>= 5 saturated or aromatic carbon-only ring size 6
208	>= 5 saturated or aromatic nitrogen-containing ring size 6
209	>= 5 saturated or aromatic heteroatom-containing ring size 6
210	>= 5 unsaturated non-aromatic carbon-only ring size 6
211	>= 5 unsaturated non-aromatic nitrogen-containing ring size 6
212	>= 5 unsaturated non-aromatic heteroatom-containing ring size 6
213	>= 1 any ring size 7
214	>= 1 saturated or aromatic carbon-only ring size 7
215	>= 1 saturated or aromatic nitrogen-containing ring size 7
216	>= 1 saturated or aromatic heteroatom-containing ring size 7
217	>= 1 unsaturated non-aromatic carbon-only ring size 7
218	>= 1 unsaturated non-aromatic nitrogen-containing ring size 7

219	>= 1 unsaturated non-aromatic heteroatom-containing ring size 7
220	>= 2 any ring size 7
221	>= 2 saturated or aromatic carbon-only ring size 7
222	>= 2 saturated or aromatic nitrogen-containing ring size 7
223	>= 2 saturated or aromatic heteroatom-containing ring size 7
224	>= 2 unsaturated non-aromatic carbon-only ring size 7
225	>= 2 unsaturated non-aromatic nitrogen-containing ring size 7
226	>= 2 unsaturated non-aromatic heteroatom-containing ring size 7
227	>= 1 any ring size 8
228	>= 1 saturated or aromatic carbon-only ring size 8
229	>= 1 saturated or aromatic nitrogen-containing ring size 8
230	>= 1 saturated or aromatic heteroatom-containing ring size 8
231	>= 1 unsaturated non-aromatic carbon-only ring size 8
232	>= 1 unsaturated non-aromatic nitrogen-containing ring size 8
233	>= 1 unsaturated non-aromatic heteroatom-containing ring size 8
234	>= 2 any ring size 8
235	>= 2 saturated or aromatic carbon-only ring size 8
236	>= 2 saturated or aromatic nitrogen-containing ring size 8
237	>= 2 saturated or aromatic heteroatom-containing ring size 8
238	>= 2 unsaturated non-aromatic carbon-only ring size 8
239	>= 2 unsaturated non-aromatic nitrogen-containing ring size 8
240	>= 2 unsaturated non-aromatic heteroatom-containing ring size 8
241	>= 1 any ring size 9
242	>= 1 saturated or aromatic carbon-only ring size 9
243	>= 1 saturated or aromatic nitrogen-containing ring size 9
244	>= 1 saturated or aromatic heteroatom-containing ring size 9
245	>= 1 unsaturated non-aromatic carbon-only ring size 9
246	>= 1 unsaturated non-aromatic nitrogen-containing ring size 9
247	>= 1 unsaturated non-aromatic heteroatom-containing ring size 9
248	>= 1 any ring size 10
249	>= 1 saturated or aromatic carbon-only ring size 10
250	>= 1 saturated or aromatic nitrogen-containing ring size 10
251	>= 1 saturated or aromatic heteroatom-containing ring size 10
252	>= 1 unsaturated non-aromatic carbon-only ring size 10
253	>= 1 unsaturated non-aromatic nitrogen-containing ring size 10

254	>= 1 unsaturated non-aromatic heteroatom-containing ring size 10
255	>= 1 aromatic ring
256	>= 1 hetero-aromatic ring
257	>= 2 aromatic rings
258	>= 2 hetero-aromatic rings
259	>= 3 aromatic rings
260	>= 3 hetero-aromatic rings
261	>= 4 aromatic rings
262	>= 4 hetero-aromatic rings
263	Li-H
264	Li-Li
265	Li-B
266	Li-C
267	Li-O
268	Li-F
269	Li-P
270	Li-S
271	Li-Cl
272	B-H
273	B-B
274	B-C
275	B-N
276	B-O
277	B-F
278	B-Si
279	B-P
280	B-S
281	B-Cl
282	B-Br
283	C-H
284	C-C
285	C-N
286	C-O
287	C-F
288	C-Na
289	C-Mg
290	C-Al
291	C-Si
292	C-P
293	C-S
294	C-Cl
295	C-As
296	C-Se
297	C-Br
298	C-I
299	N-H
300	N-N
301	N-O
302	N-F
303	N-Si
304	N-P
305	N-S
306	N-Cl
307	N-Br
308	O-H
309	O-O
310	O-Mg
311	O-Na
312	O-Al

313	O-Si
314	O-P
315	O-K
316	F-P
317	F-S
318	Al-H
319	Al-Cl
320	Si-H
321	Si-Si
322	Si-Cl
323	P-H
324	P-P
325	As-H
326	As-As
327	C(~Br)(~C)
328	C(~Br)(~C)(~C)
329	C(~Br)(~H)
330	C(~Br)(:C)
331	C(~Br)(:N)
332	C(~C)(~C)
333	C(~C)(~C)(~C)
334	C(~C)(~C)(~C)(~C)
335	C(~C)(~C)(~C)(~H)
336	C(~C)(~C)(~C)(~N)
337	C(~C)(~C)(~C)(~O)
338	C(~C)(~C)(~H)(~N)
339	C(~C)(~C)(~H)(~O)
340	C(~C)(~C)(~N)
341	C(~C)(~C)(~O)
342	C(~C)(~Cl)
343	C(~C)(~Cl)(~H)
344	C(~C)(~H)
345	C(~C)(~H)(~N)
346	C(~C)(~H)(~O)
347	C(~C)(~H)(~O)(~O)
348	C(~C)(~H)(~P)
349	C(~C)(~H)(~S)
350	C(~C)(~I)
351	C(~C)(~N)
352	C(~C)(~O)
353	C(~C)(~S)
354	C(~C)(~Si)
355	C(~C)(:C)
356	C(~C)(:C)(:C)
357	C(~C)(:C)(:N)
358	C(~C)(:N)
359	C(~C)(:N)(:N)
360	C(~Cl)(~Cl)
361	C(~Cl)(~H)
362	C(~Cl)(:C)
363	C(~F)(~F)
364	C(~F)(:C)
365	C(~H)(~N)
366	C(~H)(~O)
367	C(~H)(~O)(~O)
368	C(~H)(~S)
369	C(~H)(~Si)
370	C(~H)(:C)
371	C(~H)(:C)(:C)
372	C(~H)(:C)(:N)

373	C(~H)(:N)
374	C(~H)(~H)(~H)
375	C(~N)(~N)
376	C(~N)(:C)
377	C(~N)(:C)(:C)
378	C(~N)(:C)(:N)
379	C(~N)(:N)
380	C(~O)(~O)
381	C(~O)(:C)
382	C(~O)(:C)(:C)
383	C(~S)(:C)
384	C(:C)(:C)
385	C(:C)(:C)(:C)
386	C(:C)(:C)(:N)
387	C(:C)(:N)
388	C(:C)(:N)(:N)
389	C(:N)(:N)
390	N(~C)(~C)
391	N(~C)(~C)(~C)
392	N(~C)(~C)(~H)
393	N(~C)(~H)
394	N(~C)(~H)(~N)
395	N(~C)(~O)
396	N(~C)(:C)
397	N(~C)(:C)(:C)
398	N(~H)(~N)
399	N(~H)(:C)
400	N(~H)(:C)(:C)
401	N(~O)(~O)
402	N(~O)(:O)
403	N(:C)(:C)
404	N(:C)(:C)(:C)
405	O(~C)(~C)
406	O(~C)(~H)
407	O(~C)(~P)
408	O(~H)(~S)
409	O(:C)(:C)
410	P(~C)(~C)
411	P(~O)(~O)
412	S(~C)(~C)
413	S(~C)(~H)
414	S(~C)(~O)
415	Si(~C)(~C)
416	C=C
417	C#C
418	C=N
419	C#N
420	C=O
421	C=S
422	N=N
423	N=O
424	N=P
425	P=O
426	P=P
427	C(#C)(-C)
428	C(#C)(-H)
429	C(#N)(-C)
430	C(-C)(-C)(=C)
431	C(-C)(-C)(=N)
432	C(-C)(-C)(=O)

433	C(-C)(-Cl)(=O)
434	C(-C)(-H)(=C)
435	C(-C)(-H)(=N)
436	C(-C)(-H)(=O)
437	C(-C)(-N)(=C)
438	C(-C)(-N)(=N)
439	C(-C)(-N)(=O)
440	C(-C)(-O)(=O)
441	C(-C)(=C)
442	C(-C)(=N)
443	C(-C)(=O)
444	C(-Cl)(=O)
445	C(-H)(-N)(=C)
446	C(-H)(=C)
447	C(-H)(=N)
448	C(-H)(=O)
449	C(-N)(=C)
450	C(-N)(=N)
451	C(-N)(=O)
452	C(-O)(=O)
453	N(-C)(=C)
454	N(-C)(=O)
455	N(-O)(=O)
456	P(-O)(=O)
457	S(-C)(=O)
458	S(-O)(=O)
459	S(=O)(=O)
460	C-C-C#C
461	O-C-C=N
462	O-C-C=O
463	N:C-S-[#1]
464	N-C-C=C
465	O=S-C-C
466	N#C-C=C
467	C=N-N-C
468	O=S-C-N
469	S-S-C:C
470	C:C-C=C
471	S:C:C:C
472	C:N:C-C
473	S-C:N:C
474	S:C:C:N
475	S-C=N-C
476	C-O-C=C
477	N-N-C:C
478	S-C=N-[#1]
479	S-C-S-C
480	C:S:C-C
481	O-S-C:C
482	C:N-C:C
483	N-S-C:C
484	N-C:N:C
485	N:C:C:N
486	N-C:N:N
487	N-C=N-C
488	N-C=N-[#1]
489	N-C-S-C
490	C-C-C=C
491	C-N:C-[#1]
492	N-C:O:C

493	O=C-C:C
494	O=C-C:N
495	C-N-C:C
496	N:N-C-[#1]
497	O-C:C:N
498	O-C=C-C
499	N-C:C:N
500	C-S-C:C
501	Cl-C:C-C
502	N-C=C-[#1]
503	Cl-C:C-[#1]
504	N:C:N-C
505	Cl-C:C-O
506	C-C:N:C
507	C-C-S-C
508	S=C-N-C
509	Br-C:C-C
510	[#1]-N-N-[#1]
511	S=C-N-[#1]
512	C-[As]-O-[#1]
513	S:C:C-[#1]
514	O-N-C-C
515	N-N-C-C
516	[#1]-C=C-[#1]
517	N-N-C-N
518	O=C-N-N
519	N=C-N-C
520	C=C-C:C
521	C:N-C-[#1]
522	C-N-N-[#1]

523	N:C:C-C
524	C-C=C-C
525	[As]-C:C-[#1]
526	Cl-C:C-Cl
527	C:C:N-[#1]
528	[#1]-N-C-[#1]
529	Cl-C-C-Cl
530	N:C-C:C
531	S-C:C-C
532	S-C:C-[#1]
533	S-C:C-N
534	S-C:C-O
535	O=C-C-C
536	O=C-C-N
537	O=C-C-O
538	N=C-C-C
539	N=C-C-[#1]
540	C-N-C-[#1]
541	O-C:C-C
542	O-C:C-[#1]
543	O-C:C-N
544	O-C:C-O
545	N-C:C-C
546	N-C:C-[#1]
547	N-C:C-N
548	O-C-C:C
549	N-C-C:C
550	Cl-C-C-C
551	Cl-C-C-O
552	C:C-C:C

553	O=C-C=C
554	Br-C-C-C
555	N=C-C=C
556	C=C-C-C
557	N:C-O-[#1]
558	O=N-C:C
559	O-C-N-[#1]
560	N-C-N-C
561	Cl-C-C=O
562	Br-C-C=O
563	O-C-O-C
564	C=C-C=C
565	C:C-O-C
566	O-C-C-N
567	O-C-C-O
568	N#C-C-C
569	N-C-C-N
570	C:C-C-C
571	[#1]-C-O-[#1]
572	N:C:N:C
573	O-C-C=C
574	O-C-C:C-C
575	O-C-C:C-O
576	N=C-C:C-[#1]
577	C:C-N-C:C
578	C-C:C-C:C
579	O=C-C-C-C
580	O=C-C-C-N
581	O=C-C-C-O
582	C-C-C-C-C

583	Cl-C:C-O-C
584	C:C-C=C-C
585	C-C:C-N-C
586	C-S-C-C-C
587	N-C:C-O-[#1]
588	O=C-C-C=O
589	C-C:C-O-C
590	C-C:C-O-[#1]
591	Cl-C-C-C-C
592	N-C-C-C-C
593	N-C-C-C-N
594	C-O-C-C=C
595	C:C-C-C-C
596	N=C-N-C-C
597	O=C-C-C:C
598	Cl-C:C:C-C
599	[#1]-C-C=C-[#1]
600	N-C:C:C-C
601	N-C:C:C-N
602	O=C-C-N-C
603	C-C:C:C-C
604	C-O-C-C:C
605	O=C-C-O-C
606	O-C:C-C-C
607	N-C-C-C:C
608	C-C-C-C:C
609	Cl-C-C-N-C
610	C-O-C-O-C
611	N-C-C-N-C
612	N-C-O-C-C

613	C-N-C-C-C
614	C-C-O-C-C
615	N-C-C-O-C
616	C:C:N:N:C
617	C-C-C-O-[#1]
618	C:C-C-C:C
619	O-C-C=C-C
620	C:C-O-C-C
621	N-C:C:C:N
622	O=C-O-C:C
623	O=C-C:C-C
624	O=C-C:C-N
625	O=C-C:C-O
626	C-O-C:C-C
627	O=[As]-C:C:C
628	C-N-C-C:C
629	S-C:C:C-N
630	O-C:C-O-C
631	O-C:C-O-[#1]
632	C-C-O-C:C
633	N-C-C:C-C
634	C-C-C:C-C
635	N-N-C-N-[#1]
636	C-N-C-N-C
637	O-C-C-C-C
638	O-C-C-C-N
639	O-C-C-C-O
640	C=C-C-C-C
641	O-C-C-C=C
642	O-C-C-C=O

643	[#1]-C-C-N-[#1]
644	C-C=N-N-C
645	O=C-N-C-C
646	O=C-N-C-[#1]
647	O=C-N-C-N
648	O=N-C:C-N
649	O=N-C:C-O
650	O=C-N-C=O
651	O-C:C:C-C
652	O-C:C:C-N
653	O-C:C:C-O
654	N-C-N-C-C
655	O-C-C-C:C
656	C-C-N-C-C
657	C-N-C:C-C
658	C-C-S-C-C
659	O-C-C-N-C
660	C-C=C-C-C
661	O-C-O-C-C
662	O-C-C-O-C
663	O-C-C-O-[#1]
664	C-C=C-C=C
665	N-C:C-C-C
666	C=C-C-O-C
667	C=C-C-O-[#1]
668	C-C:C-C-C
669	Cl-C:C-C=O
670	Br-C:C:C-C
671	O=C-C=C-C
672	O=C-C=C-[#1]

673	O=C-C=C-N
674	N-C-N-C:C
675	Br-C-C-C:C
676	N#C-C-C-C
677	C-C=C-C:C
678	C-C-C=C-C
679	C-C-C-C-C-C
680	O-C-C-C-C-C
681	O-C-C-C-C-O
682	O-C-C-C-C-N
683	N-C-C-C-C-C
684	O=C-C-C-C-C
685	O=C-C-C-C-N
686	O=C-C-C-C-O
687	O=C-C-C-C=O
688	C-C-C-C-C-C-C
689	O-C-C-C-C-C-C
690	O-C-C-C-C-C-O
691	O-C-C-C-C-C-N
692	O=C-C-C-C-C-C
693	O=C-C-C-C-C-O
694	O=C-C-C-C-C=O
695	O=C-C-C-C-C-N
696	C-C-C-C-C-C-C-C
697	C-C-C-C-C-C(C)-C
698	O-C-C-C-C-C-C-C
699	O-C-C-C-C-C(C)-C
700	O-C-C-C-C-C-O-C
701	O-C-C-C-C-C(O)-C
702	O-C-C-C-C-C-N-C

703	O-C-C-C-C-C(N)-C
704	O=C-C-C-C-C-C-C
705	O=C-C-C-C-C(O)-C
706	O=C-C-C-C-C(=O)-C
707	O=C-C-C-C-C(N)-C
708	C-C(C)-C-C
709	C-C(C)-C-C-C
710	C-C-C(C)-C-C
711	C-C(C)(C)-C-C
712	C-C(C)-C(C)-C
713	Cc1ccc(C)cc1
714	Cc1ccc(O)cc1
715	Cc1ccc(S)cc1
716	Cc1ccc(N)cc1
717	Cc1ccc(Cl)cc1
718	Cc1ccc(Br)cc1
719	Oc1ccc(O)cc1
720	Oc1ccc(S)cc1
721	Oc1ccc(N)cc1
722	Oc1ccc(Cl)cc1
723	Oc1ccc(Br)cc1
724	Sc1ccc(S)cc1
725	Sc1ccc(N)cc1
726	Sc1ccc(Cl)cc1
727	Sc1ccc(Br)cc1
728	Nc1ccc(N)cc1
729	Nc1ccc(Cl)cc1
730	Nc1ccc(Br)cc1
731	Clc1ccc(Cl)cc1
732	Clc1ccc(Br)cc1

733	Brc1ccc(Br)cc1
734	Cc1cc(C)ccc1
735	Cc1cc(O)ccc1
736	Cc1cc(S)ccc1
737	Cc1cc(N)ccc1
738	Cc1cc(Cl)ccc1
739	Cc1cc(Br)ccc1
740	Oc1cc(O)ccc1
741	Oc1cc(S)ccc1
742	Oc1cc(N)ccc1
743	Oc1cc(Cl)ccc1
744	Oc1cc(Br)ccc1
745	Sc1cc(S)ccc1
746	Sc1cc(N)ccc1
747	Sc1cc(Cl)ccc1
748	Sc1cc(Br)ccc1
749	Nc1cc(N)ccc1
750	Nc1cc(Cl)ccc1
751	Nc1cc(Br)ccc1
752	Clc1cc(Cl)ccc1
753	Clc1cc(Br)ccc1
754	Brc1cc(Br)ccc1
755	Cc1c(C)cccc1
756	Cc1c(O)cccc1
757	Cc1c(S)cccc1
758	Cc1c(N)cccc1
759	Cc1c(Cl)cccc1
760	Cc1c(Br)cccc1
761	Oc1c(O)cccc1
762	Oc1c(S)cccc1
763	Oc1c(N)cccc1
764	Oc1c(Cl)cccc1
765	Oc1c(Br)cccc1
766	Sc1c(S)cccc1
767	Sc1c(N)cccc1
768	Sc1c(Cl)cccc1
769	Sc1c(Br)cccc1
770	Nc1c(N)cccc1
771	Nc1c(Cl)cccc1
772	Nc1c(Br)cccc1
773	Clc1c(Cl)cccc1
774	Clc1c(Br)cccc1
775	Brc1c(Br)cccc1
776	CC1CCC(C)CC1
777	CC1CCC(O)CC1
778	CC1CCC(S)CC1
779	CC1CCC(N)CC1
780	CC1CCC(Cl)CC1
781	CC1CCC(Br)CC1
782	OC1CCC(O)CC1
783	OC1CCC(S)CC1
784	OC1CCC(N)CC1
785	OC1CCC(Cl)CC1
786	OC1CCC(Br)CC1
787	SC1CCC(S)CC1
788	SC1CCC(N)CC1
789	SC1CCC(Cl)CC1
790	SC1CCC(Br)CC1
791	NC1CCC(N)CC1
792	NC1CCC(Cl)CC1

793	NC1CCC(Br)CC1
794	ClC1CCC(Cl)CC1
795	ClC1CCC(Br)CC1
796	BrC1CCC(Br)CC1
797	CC1CC(C)CCC1
798	CC1CC(O)CCC1
799	CC1CC(S)CCC1
800	CC1CC(N)CCC1
801	CC1CC(Cl)CCC1
802	CC1CC(Br)CCC1
803	OC1CC(O)CCC1
804	OC1CC(S)CCC1
805	OC1CC(N)CCC1
806	OC1CC(Cl)CCC1
807	OC1CC(Br)CCC1
808	SC1CC(S)CCC1
809	SC1CC(N)CCC1
810	SC1CC(Cl)CCC1
811	SC1CC(Br)CCC1
812	NC1CC(N)CCC1
813	NC1CC(Cl)CCC1
814	NC1CC(Br)CCC1
815	ClC1CC(Cl)CCC1
816	ClC1CC(Br)CCC1
817	BrC1CC(Br)CCC1
818	CC1C(C)CCCC1
819	CC1C(O)CCCC1
820	CC1C(S)CCCC1
821	CC1C(N)CCCC1
822	CC1C(Cl)CCCC1
823	CC1C(Br)CCCC1
824	OC1C(O)CCCC1
825	OC1C(S)CCCC1
826	OC1C(N)CCCC1
827	OC1C(Cl)CCCC1
828	OC1C(Br)CCCC1
829	SC1C(S)CCCC1
830	SC1C(N)CCCC1
831	SC1C(Cl)CCCC1
832	SC1C(Br)CCCC1
833	NC1C(N)CCCC1
834	NC1C(Cl)CCCC1
835	NC1C(Br)CCCC1
836	ClC1C(Cl)CCCC1
837	ClC1C(Br)CCCC1
838	BrC1C(Br)CCCC1
839	CC1CC(C)CC1
840	CC1CC(O)CC1
841	CC1CC(S)CC1
842	CC1CC(N)CC1
843	CC1CC(Cl)CC1
844	CC1CC(Br)CC1
845	OC1CC(O)CC1
846	OC1CC(S)CC1
847	OC1CC(N)CC1
848	OC1CC(Cl)CC1
849	OC1CC(Br)CC1
850	SC1CC(S)CC1
851	SC1CC(N)CC1
852	SC1CC(Cl)CC1

853	SC1CC(Br)CC1
854	NC1CC(N)CC1
855	NC1CC(Cl)CC1
856	NC1CC(Br)CC1
857	ClC1CC(Cl)CC1
858	ClC1CC(Br)CC1
859	BrC1CC(Br)CC1
860	CC1C(C)CCC1
861	CC1C(O)CCC1
862	CC1C(S)CCC1
863	CC1C(N)CCC1
864	CC1C(Cl)CCC1
865	CC1C(Br)CCC1
866	OC1C(O)CCC1

867	OC1C(S)CCC1
868	OC1C(N)CCC1
869	OC1C(Cl)CCC1
870	OC1C(Br)CCC1
871	SC1C(S)CCC1
872	SC1C(N)CCC1
873	SC1C(Cl)CCC1
874	SC1C(Br)CCC1
875	NC1C(N)CCC1
876	NC1C(Cl)CC1
877	NC1C(Br)CCC1
878	ClC1C(Cl)CCC1
879	ClC1C(Br)CCC1
880	BrC1C(Br)CCC1

Appendix C. Up- and down-regulated genes in PKD used in Chapter 5 and compiled by Malas et al.[265] These genes were intersected with the predicted targets of the compounds in the SPECTRUM library, and the compounds with a low number of references associating them with PKD.

Gene ID	Gene Symbol	Gene name	Number of studies	Direction of Regulation
9563	H6PD	hexose-6-phosphate dehydrogenase (glucose 1-dehydrogenase)	4	UP
57016	AKR1B10	aldo-keto reductase family 1, member B10 (aldose reductase)	3	UP
11167	FSTL1	follistatin-like 1	3	UP
23612	PHLDA3	pleckstrin homology-like domain, family A, member 3	3	UP
3835	KIF22	kinesin family member 22	3	UP
6281	S100A10	S100 calcium binding protein A10	3	UP
59	ACTA2	actin, alpha 2, smooth muscle, aorta	3	UP
27122	DKK3	dickkopf WNT signaling pathway inhibitor 3	3	UP
79039	DDX54	DEAD (Asp-Glu-Ala-Asp) box polypeptide 54	3	UP
1956	EGFR	epidermal growth factor receptor	3	UP
5493	PPL	periplakin	3	UP
8507	ENC1	ectodermal-neural cortex 1 (with BTB domain)	3	UP
12	SERPINA3	serpin peptidase inhibitor, clade A (alpha-1 antiproteinase, antitrypsin), member 3	3	UP

688	KLF5	Kruppel-like factor 5 (intestinal)	3	UP
1026	CDKN1A	cyclin-dependent kinase inhibitor 1A (p21, Cip1)	3	UP
3106	HLA-B	major histocompatibility complex, class I, B	3	UP
130814	PQLC3	PQ loop repeat containing 3	3	UP
23589	CARHSP1	calcium regulated heat stable protein 1, 24kDa	3	UP
6277	S100A6	S100 calcium binding protein A6	3	UP
1843	DUSP1	dual specificity phosphatase 1	3	UP
710	SERPING1	serpin peptidase inhibitor, clade G (C1 inhibitor), member 1	3	UP
3119	HLA-DQB1	major histocompatibility complex, class II, DQ beta 1	3	UP
1465	CSRP1	cysteine and glycine-rich protein 1	3	UP
4239	MFAP4	microfibrillar-associated protein 4	3	UP
80301	PLEKHO2	pleckstrin homology domain containing, family O member 2	3	UP
7739	ZNF185	zinc finger protein 185 (LIM domain)	3	UP
308	ANXA5	annexin A5	3	UP
3958	LGALS3	lectin, galactoside-binding, soluble, 3	3	UP
51421	AMOTL2	angiomotin like 2	3	UP
7414	VCL	vinculin	3	UP
8714	ABCC3	ATP-binding cassette, sub-family C (CFTR/MRP), member 3	3	UP
64393	ZMAT3	zinc finger, matrin-type 3	3	UP
2487	FRZB	frizzled-related protein	3	UP

64094	SMOC2	SPARC related modular calcium binding 2	3	UP
5099	PCDH7	protocadherin 7	3	UP
5328	PLAU	plasminogen activator, urokinase	3	UP
26585	GREM1	gremlin 1, DAN family BMP antagonist	3	UP
4016	LOXL1	lysyl oxidase-like 1	3	UP
84662	GLIS2	GLIS family zinc finger 2	3	UP
1634	DCN	decorin	3	UP
55343	SLC35C1	solute carrier family 35 (GDP-fucose transporter), member C1	2	UP
115908	CTHRC1	collagen triple helix repeat containing 1	2	UP
27286	SRPX2	sushi-repeat containing protein, X-linked 2	2	UP
3992	FADS1	fatty acid desaturase 1	2	UP
10964	IFI44L	interferon-induced protein 44-like	2	UP
2316	FLNA	filamin A, alpha	2	UP
586	BCAT1	branched chain amino-acid transaminase 1, cytosolic	2	UP
824	CAPN2	calpain 2, (m/II) large subunit	2	UP
161291	TMEM30B	transmembrane protein 30B	2	UP
81620	CDT1	chromatin licensing and DNA replication factor 1	2	UP
2266	FGG	fibrinogen gamma chain	2	UP
387923	SERP2	stress-associated endoplasmic reticulum protein family member 2	2	UP
2145	EZH1	enhancer of zeste 1 polycomb repressive complex 2 subunit	2	UP

3691	ITGB4	integrin, beta 4	2	UP
26022	TMEM98	transmembrane protein 98	2	UP
26020	LRP10	low density lipoprotein receptor-related protein 10	2	UP
10417	SPON2	spondin 2, extracellular matrix protein	2	UP
3693	ITGB5	integrin, beta 5	2	UP
5292	PIM1	Pim-1 proto-oncogene, serine/threonine kinase	2	UP
10568	SLC34A2	solute carrier family 34 (type II sodium/phosphate cotransporter), member 2	2	UP
1999	ELF3	E74-like factor 3 (ets domain transcription factor, epithelial-specific)	2	UP
29968	PSAT1	phosphoserine aminotransferase 1	2	UP
7076	TIMP1	TIMP metallopeptidase inhibitor 1	2	UP
1281	COL3A1	collagen, type III, alpha 1	2	UP
57045	TWSG1	twisted gastrulation BMP signaling modulator 1	2	UP
6574	SLC20A1	solute carrier family 20 (phosphate transporter), member 1	2	UP
55118	CRTAC1	cartilage acidic protein 1	2	UP
406	ARNTL	aryl hydrocarbon receptor nuclear translocator-like	2	UP
2033	EP300	E1A binding protein p300	2	UP
261734	NPHP4	nephronophthisis 4	2	UP
51006	SLC35C2	solute carrier family 35 (GDP-fucose transporter), member C2	2	UP
5990	RFX2	regulatory factor X, 2 (influences HLA class II expression)	2	UP
253430	IPMK	inositol polyphosphate multikinase	2	UP

2319	FLOT2	flotillin 2	2	UP
55023	PHIP	pleckstrin homology domain interacting protein	2	UP
2305	FOXM1	forkhead box M1	2	UP
170506	DHX36	DEAH (Asp-Glu-Ala-His) box polypeptide 36	2	UP
4628	MYH10	myosin, heavy chain 10, non-muscle	2	UP
81615	TMEM163	transmembrane protein 163	2	UP
259266	ASPM	asp (abnormal spindle) homolog, microcephaly associated (Drosophila)	2	UP
51100	SH3GLB1	SH3-domain GRB2-like endophilin B1	2	UP
9246	UBE2L6	ubiquitin-conjugating enzyme E2L 6	2	UP
3694	ITGB6	integrin, beta 6	2	UP
7980	TFPI2	tissue factor pathway inhibitor 2	2	UP
9332	CD163	CD163 molecule	2	UP
22801	ITGA11	integrin, alpha 11	2	UP
63874	ABHD4	abhydrolase domain containing 4	2	UP
2697	GJA1	gap junction protein, alpha 1, 43kDa	2	UP
6819	SULT1C2	sulfotransferase family, cytosolic, 1C, member 2	2	UP
10572	SIVA1	SIVA1, apoptosis-inducing factor	2	UP
5196	PF4	platelet factor 4	2	UP
83461	CDCA3	cell division cycle associated 3	2	UP
3357	HTR2B	5-hydroxytryptamine (serotonin) receptor 2B, G protein-coupled	2	UP

1545	CYP1B1	cytochrome P450, family 1, subfamily B, polypeptide 1	2	UP
694	BTG1	B-cell translocation gene 1, anti-proliferative	2	UP
79709	COLGALT1	collagen beta(1-O)galactosyltransferase 1	2	UP
79705	LRRK1	leucine-rich repeat kinase 1	2	UP
79701	OGFOD3	2-oxoglutarate and iron-dependent oxygenase domain containing 3	2	UP
23603	CORO1C	coronin, actin binding protein, 1C	2	UP
6566	SLC16A1	solute carrier family 16 (monocarboxylate transporter), member 1	2	UP
6560	SLC12A4	solute carrier family 12 (potassium/chloride transporter), member 4	2	UP
55329	MNS1	meiosis-specific nuclear structural 1	2	UP
55320	MIS18BP1	MIS18 binding protein 1	2	UP
4281	MID1	midline 1	2	UP
3588	IL10RB	interleukin 10 receptor, beta	2	UP
55742	PARVA	parvin, alpha	2	UP
347733	TUBB2B	tubulin, beta 2B class IIb	2	UP
7058	THBS2	thrombospondin 2	2	UP
64857	PLEKHG2	pleckstrin homology domain containing, family G (with RhoGef domain) member 2	2	UP
60485	SAV1	salvador family WW domain containing protein 1	2	UP
2192	FBLN1	fibulin 1	2	UP
2200	FBN1	fibrillin 1	2	UP

54751	FBLIM1	filamin binding LIM protein 1	2	UP
9507	ADAMTS4	ADAM metallopeptidase with thrombospondin type 1 motif, 4	2	UP
165	AEBP1	AE binding protein 1	2	UP
3689	ITGB2	integrin, beta 2 (complement component 3 receptor 3 and 4 subunit)	2	UP
3688	ITGB1	integrin, beta 1 (fibronectin receptor, beta polypeptide, antigen CD29 includes MDF2, MSK12)	2	UP
23344	ESYT1	extended synaptotagmin-like protein 1	2	UP
4837	NNMT	nicotinamide N-methyltransferase	2	UP
9619	ABCG1	ATP-binding cassette, sub-family G (WHITE), member 1	2	UP
9411	ARHGAP29	Rho GTPase activating protein 29	2	UP
10769	PLK2	polo-like kinase 2	2	UP
80003	PCNXL2	pecanex-like 2 (Drosophila)	2	UP
54407	SLC38A2	solute carrier family 38, member 2	2	UP
84294	UTP23	UTP23, small subunit (SSU) processome component, homolog (yeast)	2	UP
285313	IGSF10	immunoglobulin superfamily, member 10	2	UP
1600	DAB1	Dab, reelin signal transducer, homolog 1 (Drosophila)	2	UP
56548	CHST7	carbohydrate (N-acetylglucosamine 6-O) sulfotransferase 7	2	UP
1604	CD55	CD55 molecule, decay accelerating factor for complement (Cromer blood group)	2	UP
83452	RAB33B	RAB33B, member RAS oncogene family	2	UP
79026	AHNAK	AHNAK nucleoprotein	2	UP

7050	TGIF1	TGFB-induced factor homeobox 1	2	UP
114990	VASN	vasorin	2	UP
5515	PPP2CA	protein phosphatase 2, catalytic subunit, alpha isozyme	2	UP
2995	GYPC	glycophorin C (Gerbich blood group)	2	UP
3290	HSD11B1	hydroxysteroid (11-beta) dehydrogenase 1	2	UP
286133	SCARA5	scavenger receptor class A, member 5	2	UP
114898	C1QTNF2	C1q and tumor necrosis factor related protein 2	2	UP
23254	KAZN	kazrin, periplakin interacting protein	2	UP
50509	COL5A3	collagen, type V, alpha 3	2	UP
81566	CSRNP2	cysteine-serine-rich nuclear protein 2	2	UP
10992	SF3B2	splicing factor 3b, subunit 2, 145kDa	2	UP
55315	SLC29A3	solute carrier family 29 (equilibrative nucleoside transporter), member 3	2	UP
4609	MYC	v-myc avian myelocytomatosis viral oncogene homolog	2	UP
220	ALDH1A3	aldehyde dehydrogenase 1 family, member A3	2	UP
57333	RCN3	reticulocalbin 3, EF-hand calcium binding domain	2	UP
22822	PHLDA1	pleckstrin homology-like domain, family A, member 1	2	UP
10370	CITED2	Cbp/p300-interacting transactivator, with Glu/Asp-rich carboxy-terminal domain, 2	2	UP
25960	GPR124	G protein-coupled receptor 124	2	UP
1809	DPYSL3	dihydropyrimidinase-like 3	2	UP

6237	RRAS	related RAS viral (r-ras) oncogene homolog	2	UP
83442	SH3BGRL3	SH3 domain binding glutamate-rich protein like 3	2	UP
9021	SOCS3	suppressor of cytokine signaling 3	2	UP
255738	PCSK9	proprotein convertase subtilisin/kexin type 9	2	UP
1397	CRIP2	cysteine-rich protein 2	2	UP
7043	TGFB3	transforming growth factor, beta 3	2	UP
10193	RNF41	ring finger protein 41, E3 ubiquitin protein ligase	2	UP
54033	RBM11	RNA binding motif protein 11	2	UP
1014	CDH16	cadherin 16, KSP-cadherin	2	UP
83716	CRISPLD2	cysteine-rich secretory protein LCCL domain containing 2	2	UP
1012	CDH13	cadherin 13	2	UP
8510	MMP23B	matrix metallopeptidase 23B	2	UP
388135	C15orf59	chromosome 15 open reading frame 59	2	UP
4192	MDK	midkine (neurite growth-promoting factor 2)	2	UP
6507	SLC1A3	solute carrier family 1 (glial high affinity glutamate transporter), member 3	2	UP
4905	NSF	N-ethylmaleimide-sensitive factor	2	UP
23480	SEC61G	Sec61 gamma subunit	2	UP
11340	EXOSC8	exosome component 8	2	UP
55014	STX17	syntaxin 17	2	UP
55789	DEPDC1B	DEP domain containing 1B	2	UP

4174	MCM5	minichromosome maintenance complex component 5	2	UP
4175	MCM6	minichromosome maintenance complex component 6	2	UP
4176	MCM7	minichromosome maintenance complex component 7	2	UP
4170	MCL1	myeloid cell leukemia 1	2	UP
4171	MCM2	minichromosome maintenance complex component 2	2	UP
9948	WDR1	WD repeat domain 1	2	UP
56477	CCL28	chemokine (C-C motif) ligand 28	2	UP
51330	TNFRSF12A	tumor necrosis factor receptor superfamily, member 12A	2	UP
56475	RPRM	reprimo, TP53 dependent G2 arrest mediator candidate	2	UP
219902	TMEM136	transmembrane protein 136	2	UP
348932	SLC6A18	solute carrier family 6 (neutral amino acid transporter), member 18	2	UP
4507	MTAP	methylthioadenosine phosphorylase	2	UP
85477	SCIN	scinderin	2	UP
64943	NT5DC2	5'-nucleotidase domain containing 2	2	UP
57326	PBXIP1	pre-B-cell leukemia homeobox interacting protein 1	2	UP
8985	PLOD3	procollagen-lysine, 2-oxoglutarate 5-dioxygenase 3	2	UP
10347	ABCA7	ATP-binding cassette, sub-family A (ABC1), member 7	2	UP
1958	EGR1	early growth response 1	2	UP
827	CAPN6	calpain 6	2	UP
9019	MPZL1	myelin protein zero-like 1	2	UP

6259	RYK	receptor-like tyrosine kinase	2	UP
83892	KCTD10	potassium channel tetramerization domain containing 10	2	UP
2149	F2R	coagulation factor II (thrombin) receptor	2	UP
6764	ST5	suppression of tumorigenicity 5	2	UP
6539	SLC6A12	solute carrier family 6 (neurotransmitter transporter), member 12	2	UP
5784	PTPN14	protein tyrosine phosphatase, non-receptor type 14	2	UP
25983	NGDN	neuroguidin, EIF4E binding protein	2	UP
11168	PSIP1	PC4 and SFRS1 interacting protein 1	2	UP
23345	SYNE1	spectrin repeat containing, nuclear envelope 1	2	UP
929	CD14	CD14 molecule	2	UP
11245	GPR176	G protein-coupled receptor 176	2	UP
51148	CERCAM	cerebral endothelial cell adhesion molecule	2	UP
7187	TRAF3	TNF receptor-associated factor 3	2	UP
3491	CYR61	cysteine-rich, angiogenic inducer, 61	2	UP
85480	TSLP	thymic stromal lymphopoietin	2	UP
4886	NPY1R	neuropeptide Y receptor Y1	2	UP
2147	F2	coagulation factor II (thrombin)	2	UP
387758	FIBIN	fin bud initiation factor homolog (zebrafish)	2	UP
10537	UBD	ubiquitin D	2	UP
3275	PRMT2	protein arginine methyltransferase 2	2	UP

83959	SLC4A11	solute carrier family 4, sodium borate transporter, member 11	2	UP
837	CASP4	caspase 4, apoptosis-related cysteine peptidase	2	UP
2934	GSN	gelsolin	2	UP
9839	ZEB2	zinc finger E-box binding homeobox 2	2	UP
3316	HSPB2	heat shock 27kDa protein 2	2	UP
7791	ZYX	zyxin	2	UP
5589	PRKCSH	protein kinase C substrate 80K-H	2	UP
79188	TMEM43	transmembrane protein 43	2	UP
1030	CDKN2B	cyclin-dependent kinase inhibitor 2B (p15, inhibits CDK4)	2	UP
1031	CDKN2C	cyclin-dependent kinase inhibitor 2C (p18, inhibits CDK4)	2	UP
1036	CDO1	cysteine dioxygenase type 1	2	UP
114899	C1QTNF3	C1q and tumor necrosis factor related protein 3	2	UP
4245	MGAT1	mannosyl (alpha-1,3-)-glycoprotein beta-1,2-N-acetylglucosaminyltransferase	2	UP
23645	PPP1R15A	protein phosphatase 1, regulatory subunit 15A	2	UP
5654	HTRA1	HtrA serine peptidase 1	2	UP
11065	UBE2C	ubiquitin-conjugating enzyme E2C	2	UP
11067	C10orf10	chromosome 10 open reading frame 10	2	UP
11325	DDX42	DEAD (Asp-Glu-Ala-Asp) box helicase 42	2	UP
3913	LAMB2	laminin, beta 2 (laminin S)	2	UP
3918	LAMC2	laminin, gamma 2	2	UP

8313	AXIN2	axin 2	2	UP
4794	NFKBIE	nuclear factor of kappa light polypeptide gene enhancer in B-cells inhibitor, epsilon	2	UP
683	BST1	bone marrow stromal cell antigen 1	2	UP
81552	VOPP1	vesicular, overexpressed in cancer, prosurvival protein 1	2	UP
55573	CDV3	CDV3 homolog (mouse)	2	UP
216	ALDH1A1	aldehyde dehydrogenase 1 family, member A1	2	UP
51170	HSD17B11	hydroxysteroid (17-beta) dehydrogenase 11	2	UP
5900	RALGDS	ral guanine nucleotide dissociation stimulator	2	UP
3488	IGFBP5	insulin-like growth factor binding protein 5	2	UP
60	ACTB	actin, beta	2	UP
6397	SEC14L1	SEC14-like 1 (S. cerevisiae)	2	UP
116496	FAM129A	family with sequence similarity 129, member A	2	UP
64145	RBSN	rabenosyn, RAB effector	2	UP
4899	NRF1	nuclear respiratory factor 1	2	UP
7975	MAFK	v-maf avian musculoaponeurotic fibrosarcoma oncogene homolog K	2	UP
5167	ENPP1	ectonucleotide pyrophosphatase/phosphodiesterase 1	2	UP
10468	FST	follistatin	2	UP
55089	SLC38A4	solute carrier family 38, member 4	2	UP
5125	PCSK5	proprotein convertase subtilisin/kexin type 5	2	UP

22980	TCF25	transcription factor 25 (basic helix-loop-helix)	2	UP
116039	OSR2	odd-skipped related transciption factor 2	2	UP
54825	CDHR2	cadherin-related family member 2	2	UP
54820	NDE1	nudE neurodevelopment protein 1	2	UP
80863	PRRT1	proline-rich transmembrane protein 1	2	UP
4087	SMAD2	SMAD family member 2	2	UP
768211	RELL1	RELT-like 1	2	UP
2956	MSH6	mutS homolog 6	2	UP
2952	GSTT1	glutathione S-transferase theta 1	2	UP
7162	TPBG	trophoblast glycoprotein	2	UP
4360	MRC1	mannose receptor, C type 1	2	UP
8076	MFAP5	microfibrillar associated protein 5	2	UP
6632	SNRPD1	small nuclear ribonucleoprotein D1 polypeptide 16kDa	2	UP
8676	STX11	syntaxin 11	2	UP
126374	WTIP	Wilms tumor 1 interacting protein	2	UP
9805	SCRN1	secernin 1	2	UP
26471	NUPR1	nuclear protein, transcriptional regulator, 1	2	UP
4783	NFIL3	nuclear factor, interleukin 3 regulated	2	UP
8099	CDK2AP1	cyclin-dependent kinase 2 associated protein 1	2	UP
5918	RARRES1	retinoic acid receptor responder (tazarotene induced) 1	2	UP

5914	RARA	retinoic acid receptor, alpha	2	UP
3664	IRF6	interferon regulatory factor 6	2	UP
196463	PLBD2	phospholipase B domain containing 2	2	UP
3675	ITGA3	integrin, alpha 3 (antigen CD49C, alpha 3 subunit of VLA-3 receptor)	2	UP
254228	FAM26E	family with sequence similarity 26, member E	2	UP
2762	GMDS	GDP-mannose 4,6-dehydratase	2	UP
2760	GM2A	GM2 ganglioside activator	2	UP
1848	DUSP6	dual specificity phosphatase 6	2	UP
4958	OMD	osteomodulin	2	UP
9962	SLC23A2	solute carrier family 23 (ascorbic acid transporter), member 2	2	UP
9735	KNTC1	kinetochore associated 1	2	UP
1902	LPAR1	lysophosphatidic acid receptor 1	2	UP
26872	STEAP1	six transmembrane epithelial antigen of the prostate 1	2	UP
64332	NFKBIZ	nuclear factor of kappa light polypeptide gene enhancer in B-cells inhibitor, zeta	2	UP
22881	ANKRD6	ankyrin repeat domain 6	2	UP
5886	RAD23A	RAD23 homolog A (S. cerevisiae)	2	UP
79098	C1orf116	chromosome 1 open reading frame 116	2	UP
132884	EVC2	Ellis van Creveld syndrome 2	2	UP
11183	MAP4K5	mitogen-activated protein kinase kinase kinase kinase 5	2	UP
7175	TPR	translocated promoter region, nuclear basket protein	2	UP

6892	TAPBP	TAP binding protein (tapasin)	2	UP
5355	PLP2	proteolipid protein 2 (colonic epithelium-enriched)	2	UP
1058	CENPA	centromere protein A	2	UP
716	C1S	complement component 1, s subcomponent	2	UP
8796	SCEL	sciellin	2	UP
7045	TGFBI	transforming growth factor, beta-induced, 68kDa	2	UP
6415	SEPW1	selenoprotein W, 1	2	UP
1476	CSTB	cystatin B (stefin B)	2	UP
7041	TGFB1I1	transforming growth factor beta 1 induced transcript 1	2	UP
66004	LYNX1	Ly6/neurotoxin 1	2	UP
4224	MEP1A	meprin A, alpha (PABA peptide hydrolase)	2	UP
3597	IL13RA1	interleukin 13 receptor, alpha 1	2	UP
25820	ARIH1	ariadne RBR E3 ubiquitin protein ligase 1	2	UP
3934	LCN2	lipocalin 2	2	UP
9532	BAG2	BCL2-associated athanogene 2	2	UP
11270	NRM	nurim (nuclear envelope membrane protein)	2	UP
146760	RTN4RL1	reticulon 4 receptor-like 1	2	UP
22795	NID2	nidogen 2 (osteonidogen)	2	UP
131177	FAM3D	family with sequence similarity 3, member D	2	UP
4542	MYO1F	myosin IF	2	UP

6731	SRP72	signal recognition particle 72kDa	2	UP
9124	PDLIM1	PDZ and LIM domain 1	2	UP
375790	AGRN	agrin	2	UP
6282	S100A11	S100 calcium binding protein A11	2	UP
153830	RNF145	ring finger protein 145	2	UP
182	JAG1	jagged 1	2	UP
26136	TES	testin LIM domain protein	2	UP
54836	BSPRY	B-box and SPRY domain containing	2	UP
861	RUNX1	runt-related transcription factor 1	2	UP
84186	ZCCHC7	zinc finger, CCHC domain containing 7	2	UP
9055	PRC1	protein regulator of cytokinesis 1	2	UP
9057	SLC7A6	solute carrier family 7 (amino acid transporter light chain, y+L system), member 6	2	UP
72	ACTG2	actin, gamma 2, smooth muscle, enteric	2	UP
57819	LSM2	LSM2 homolog, U6 small nuclear RNA associated (S. cerevisiae)	2	UP
666	BOK	BCL2-related ovarian killer	2	UP
663	BNIP2	BCL2/adenovirus E1B 19kDa interacting protein 2	2	UP
1265	CNN2	calponin 2	2	UP
22809	ATF5	activating transcription factor 5	2	UP
3123	HLA-DRB1	major histocompatibility complex, class II, DR beta 1	2	UP

1462	VCAN	versican	2	UP
1316	KLF6	Kruppel-like factor 6	2	UP
23764	MAFF	v-maf avian musculoaponeurotic fibrosarcoma oncogene homolog F	2	UP
23564	DDAH2	dimethylarginine dimethylaminohydrolase 2	2	UP
23166	STAB1	stabilin 1	2	UP
6709	SPTAN1	spectrin, alpha, non-erythrocytic 1	2	UP
8321	FZD1	frizzled class receptor 1	2	UP
9912	ARHGAP44	Rho GTPase activating protein 44	2	UP
51365	PLA1A	phospholipase A1 member A	2	UP
7431	VIM	vimentin	2	UP
6515	SLC2A3	solute carrier family 2 (facilitated glucose transporter), member 3	2	UP
51186	WBP5	WW domain binding protein 5	2	UP
4059	BCAM	basal cell adhesion molecule (Lutheran blood group)	2	UP
50618	ITSN2	intersectin 2	2	UP
90417	KNSTRN	kinetochore-localized astrin/SPAG5 binding protein	2	UP
57007	ACKR3	atypical chemokine receptor 3	2	UP
3725	JUN	jun proto-oncogene	2	UP
3726	JUNB	jun B proto-oncogene	2	UP
3727	JUND	jun D proto-oncogene	2	UP
960	CD44	CD44 molecule (Indian blood group)	2	UP

57486	NLN	neurolysin (metallopeptidase M3 family)	2	UP
5118	PCOLCE	procollagen C-endopeptidase enhancer	2	UP
124540	MSI2	musashi RNA-binding protein 2	2	UP
9590	AKAP12	A kinase (PRKA) anchor protein 12	2	UP
203068	TUBB	tubulin, beta class I	2	UP
10267	RAMP1	receptor (G protein-coupled) activity modifying protein 1	2	UP
1783	DYNC1LI2	dynein, cytoplasmic 1, light intermediate chain 2	2	UP
1786	DNMT1	DNA (cytosine-5-)-methyltransferase 1	2	UP
6876	TAGLN	transgelin	2	UP
84552	PARD6G	par-6 family cell polarity regulator gamma	2	UP
3134	HLA-F	major histocompatibility complex, class I, F	2	UP
3133	HLA-E	major histocompatibility complex, class I, E	2	UP
1522	CTSZ	cathepsin Z	2	UP
50807	ASAP1	ArfGAP with SH3 domain, ankyrin repeat and PH domain 1	2	UP
84886	C1orf198	chromosome 1 open reading frame 198	2	UP
301	ANXA1	annexin A1	2	UP
302	ANXA2	annexin A2	2	UP
55723	ASF1B	anti-silencing function 1B histone chaperone	2	UP
11332	ACOT7	acyl-CoA thioesterase 7	2	UP
25903	OLFML2B	olfactomedin-like 2B	2	UP

6118	RPA2	replication protein A2, 32kDa	2	UP
6117	RPA1	replication protein A1, 70kDa	2	UP
4692	NDN	necdin, melanoma antigen (MAGE) family member	2	UP
2131	EXT1	exostosin glycosyltransferase 1	2	UP
4928	NUP98	nucleoporin 98kDa	2	UP
93986	FOXP2	forkhead box P2	2	UP
9214	FAIM3	Fas apoptotic inhibitory molecule 3	2	UP
4854	NOTCH3	notch 3	2	UP
9076	CLDN1	claudin 1	2	UP
388341	LRRC75A	leucine rich repeat containing 75A	2	UP
10457	GPNMB	glycoprotein (transmembrane) nmb	2	UP
10581	IFITM2	interferon induced transmembrane protein 2	2	UP
714	C1QC	complement component 1, q subcomponent, C chain	2	UP
713	C1QB	complement component 1, q subcomponent, B chain	2	UP
283149	BCL9L	B-cell CLL/lymphoma 9-like	2	UP
1066	CES1	carboxylesterase 1	2	UP
10628	TXNIP	thioredoxin interacting protein	2	UP
5579	PRKCB	protein kinase C, beta	2	UP
6422	SFRP1	secreted frizzled-related protein 1	2	UP
3267	AGFG1	ArfGAP with FG repeats 1	2	UP

330	BIRC3	baculoviral IAP repeat containing 3	2	UP
23328	SASH1	SAM and SH3 domain containing 1	2	UP
51435	SCARA3	scavenger receptor class A, member 3	2	UP
25907	TMEM158	transmembrane protein 158 (gene/pseudogene)	2	UP
23433	RHOQ	ras homolog family member Q	2	UP
51230	PHF20	PHD finger protein 20	2	UP
4071	TM4SF1	transmembrane 4 L six family member 1	2	UP
4070	TACSTD2	tumor-associated calcium signal transducer 2	2	UP
5176	SERPINF1	serpin peptidase inhibitor, clade F (alpha-2 antiplasmin, pigment epithelium derived factor), member 1	2	UP
9208	LRRFIP1	leucine rich repeat (in FLII) interacting protein 1	2	UP
10493	VAT1	vesicle amine transport 1	2	UP
4862	NPAS2	neuronal PAS domain protein 2	2	UP
2017	CTTN	cortactin	2	UP
6850	SYK	spleen tyrosine kinase	2	UP
706	TSPO	translocator protein (18kDa)	2	UP
7132	TNFRSF1A	tumor necrosis factor receptor superfamily, member 1A	2	UP
9111	NMI	N-myc (and STAT) interactor	2	UP
30846	EHD2	EH-domain containing 2	2	UP
11082	ESM1	endothelial cell-specific molecule 1	2	UP

1435	CSF1	colony stimulating factor 1 (macrophage)	2	UP
79050	NOC4L	nucleolar complex associated 4 homolog (S. cerevisiae)	2	UP
6923	TCEB2	transcription elongation factor B (SIII), polypeptide 2 (18kDa, elongin B)	2	UP
79586	CHPF	chondroitin polymerizing factor	2	UP
5549	PRELP	proline/arginine-rich end leucine-rich repeat protein	2	UP
23158	TBC1D9	TBC1 domain family, member 9 (with GRAM domain)	2	UP
1992	SERPINB1	serpin peptidase inhibitor, clade B (ovalbumin), member 1	2	UP
3880	KRT19	keratin 19, type I	2	UP
7405	UVRAG	UV radiation resistance associated	2	UP
4650	MYO9B	myosin IXB	2	UP
257364	SNX33	sorting nexin 33	2	UP
6627	SNRPA1	small nuclear ribonucleoprotein polypeptide A'	2	UP
84935	MEDAG	mesenteric estrogen-dependent adipogenesis	2	UP
55700	MAP7D1	MAP7 domain containing 1	2	UP
55701	ARHGEF40	Rho guanine nucleotide exchange factor (GEF) 40	2	UP
3625	INHBB	inhibin, beta B	2	UP
60370	AVPI1	arginine vasopressin-induced 1	2	UP
2335	FN1	fibronectin 1	2	UP
3071	NCKAP1L	NCK-associated protein 1-like	2	UP
6004	RGS16	regulator of G-protein signaling 16	2	UP

200185	KRTCAP2	keratinocyte associated protein 2	2	UP
4673	NAP1L1	nucleosome assembly protein 1-like 1	2	UP
4670	HNRNPM	heterogeneous nuclear ribonucleoprotein M	2	UP
7169	TPM2	tropomyosin 2 (beta)	2	UP
10699	CORIN	corin, serine peptidase	2	UP
80139	ZNF703	zinc finger protein 703	2	UP
64388	GREM2	gremlin 2, DAN family BMP antagonist	2	UP
9659	PDE4DIP	phosphodiesterase 4D interacting protein	2	UP
10435	CDC42EP2	CDC42 effector protein (Rho GTPase binding) 2	2	UP
3572	IL6ST	interleukin 6 signal transducer	2	UP
9785	DHX38	DEAH (Asp-Glu-Ala-His) box polypeptide 38	2	UP
9787	DLGAP5	discs, large (Drosophila) homolog-associated protein 5	2	UP
5321	PLA2G4A	phospholipase A2, group IVA (cytosolic, calcium-dependent)	2	UP
10659	CELF2	CUGBP, Elav-like family member 2	2	UP
1356	CP	ceruloplasmin (ferroxidase)	2	UP
283209	PGM2L1	phosphoglucomutase 2-like 1	2	UP
6774	STAT3	signal transducer and activator of transcription 3 (acute-phase response factor)	2	UP
131583	FAM43A	family with sequence similarity 43, member A	2	UP
8560	DEGS1	delta(4)-desaturase, sphingolipid 1	2	UP
5888	RAD51	RAD51 recombinase	2	UP

55603	FAM46A	family with sequence similarity 46, member A	2	UP
6633	SNRPD2	small nuclear ribonucleoprotein D2 polypeptide 16.5kDa	2	UP
4015	LOX	lysyl oxidase	2	UP
64343	AZI2	5-azacytidine induced 2	2	UP
983	CDK1	cyclin-dependent kinase 1	2	UP
79801	SHCBP1	SHC SH2-domain binding protein 1	2	UP
10130	PDIA6	protein disulfide isomerase family A, member 6	2	UP
8971	H1FX	H1 histone family, member X	2	UP
24137	KIF4A	kinesin family member 4A	2	UP
10556	RPP30	ribonuclease P/MRP 30kDa subunit	2	UP
1633	DCK	deoxycytidine kinase	2	UP
2817	GPC1	glypican 1	2	UP
7089	TLE2	transducin-like enhancer of split 2	2	UP
5426	POLE	polymerase (DNA directed), epsilon, catalytic subunit	2	UP
3371	TNC	tenascin C	2	UP
726	CAPN5	calpain 5	2	UP
727	C5	complement component 5	2	UP
55040	EPN3	epsin 3	2	UP
8000	PSCA	prostate stem cell antigen	2	UP
23621	BACE1	beta-site APP-cleaving enzyme 1	2	UP

6541	SLC7A1	solute carrier family 7 (cationic amino acid transporter, y+ system), member 1	2	UP
790	CAD	carbamoyl-phosphate synthetase 2, aspartate transcarbamylase, and dihydroorotase	2	UP
1364	CLDN4	claudin 4	2	UP
55586	MIOX	myo-inositol oxygenase	3	DN
275	AMT	aminomethyltransferase	3	DN
64849	SLC13A3	solute carrier family 13 (sodium-dependent dicarboxylate transporter), member 3	3	DN
2053	EPHX2	epoxide hydrolase 2, cytoplasmic	3	DN
54988	ACSM5	acyl-CoA synthetase medium-chain family member 5	3	DN
5502	PPP1R1A	protein phosphatase 1, regulatory (inhibitor) subunit 1A	3	DN
5618	PRLR	prolactin receptor	3	DN
64902	AGXT2	alanine--glyoxylate aminotransferase 2	3	DN
10166	SLC25A15	solute carrier family 25 (mitochondrial carrier; ornithine transporter) member 15	3	DN
80168	MOGAT2	monoacylglycerol O-acyltransferase 2	3	DN
1800	DPEP1	dipeptidase 1 (renal)	3	DN
6505	SLC1A1	solute carrier family 1 (neuronal/epithelial high affinity glutamate transporter, system Xag), member 1	3	DN
9365	KL	klotho	3	DN
56898	BDH2	3-hydroxybutyrate dehydrogenase, type 2	3	DN
1950	EGF	epidermal growth factor	3	DN

18	ABAT	4-aminobutyrate aminotransferase	3	DN
7021	TFAP2B	transcription factor AP-2 beta (activating enhancer binding protein 2 beta)	3	DN
6521	SLC4A1	solute carrier family 4 (anion exchanger), member 1 (Diego blood group)	3	DN
7809	BSND	barttin CLCNK-type chloride channel accessory beta subunit	3	DN
3483	IGFALS	insulin-like growth factor binding protein, acid labile subunit	3	DN
1852	DUSP9	dual specificity phosphatase 9	3	DN
152404	IGSF11	immunoglobulin superfamily, member 11	3	DN
6887	TAL2	T-cell acute lymphocytic leukemia 2	3	DN
79774	GRTP1	growth hormone regulated TBC protein 1	3	DN
3291	HSD11B2	hydroxysteroid (11-beta) dehydrogenase 2	3	DN
5037	PEBP1	phosphatidylethanolamine binding protein 1	3	DN
83697	SLC4A9	solute carrier family 4, sodium bicarbonate cotransporter, member 9	3	DN
9056	SLC7A7	solute carrier family 7 (amino acid transporter light chain, y+L system), member 7	3	DN
123264	SLC51B	solute carrier family 51, beta subunit	3	DN
670	BPHL	biphenyl hydrolase-like (serine hydrolase)	3	DN
127124	ATP6V1G3	ATPase, H+ transporting, lysosomal 13kDa, V1 subunit G3	3	DN
94081	SFXN1	sideroflexin 1	3	DN
7284	TUFM	Tu translation elongation factor, mitochondrial	3	DN
171586	ABHD3	abhydrolase domain containing 3	3	DN

36	ACADSB	acyl-CoA dehydrogenase, short/branched chain	3	DN
6611	SMS	spermine synthase	3	DN
57715	SEMA4G	sema domain, immunoglobulin domain (Ig), transmembrane domain (TM) and short cytoplasmic domain, (semaphorin) 4G	3	DN
90507	SCRN2	secernin 2	3	DN
64081	PBLD	phenazine biosynthesis-like protein domain containing	3	DN
5172	SLC26A4	solute carrier family 26 (anion exchanger), member 4	3	DN
3712	IVD	isovaleryl-CoA dehydrogenase	3	DN
22928	SEPHS2	selenophosphate synthetase 2	3	DN
143941	TTC36	tetratricopeptide repeat domain 36	3	DN
1429	CRYZ	crystallin, zeta (quinone reductase)	3	DN
6557	SLC12A1	solute carrier family 12 (sodium/potassium/chloride transporter), member 1	3	DN
6559	SLC12A3	solute carrier family 12 (sodium/chloride transporter), member 3	3	DN
58510	PRODH2	proline dehydrogenase (oxidase) 2	3	DN
92840	REEP6	receptor accessory protein 6	3	DN
55825	PECR	peroxisomal trans-2-enoyl-CoA reductase	3	DN
793	CALB1	calbindin 1, 28kDa	3	DN
84912	SLC35B4	solute carrier family 35 (UDP-xylose/UDP-N-acetylglucosamine transporter), member B4	2	DN
23443	SLC35A3	solute carrier family 35 (UDP-N-acetylglucosamine (UDP-GlcNAc) transporter), member A3	2	DN

55296	TBC1D19	TBC1 domain family, member 19	2	DN
115817	DHRS1	dehydrogenase/reductase (SDR family) member 1	2	DN
7512	XPNPEP2	X-prolyl aminopeptidase (aminopeptidase P) 2, membrane-bound	2	DN
119467	CLRN3	clarin 3	2	DN
55902	ACSS2	acyl-CoA synthetase short-chain family member 2	2	DN
10966	RAB40B	RAB40B, member RAS oncogene family	2	DN
292	SLC25A5	solute carrier family 25 (mitochondrial carrier; adenine nucleotide translocator), member 5	2	DN
100506658	OCLN	occludin	2	DN
51115	RMDN1	regulator of microtubule dynamics 1	2	DN
51110	LACTB2	lactamase, beta 2	2	DN
2261	FGFR3	fibroblast growth factor receptor 3	2	DN
51300	TIMMDC1	translocase of inner mitochondrial membrane domain containing 1	2	DN
2103	ESRRB	estrogen-related receptor beta	2	DN
22981	NINL	ninein-like	2	DN
116238	TLCD1	TLC domain containing 1	2	DN
51268	PIPOX	pipecolic acid oxidase	2	DN
254295	PHYHD1	phytanoyl-CoA dioxygenase domain containing 1	2	DN
10380	BPNT1	3'(2'), 5'-bisphosphate nucleotidase 1	2	DN
9325	TRIP4	thyroid hormone receptor interactor 4	2	DN

8942	KYNU	kynureninase	2	DN
7263	TST	thiosulfate sulfurtransferase (rhodanese)	2	DN
552	AVPR1A	arginine vasopressin receptor 1A	2	DN
5052	PRDX1	peroxiredoxin 1	2	DN
133522	PPARGC1B	peroxisome proliferator-activated receptor gamma, coactivator 1 beta	2	DN
23504	RIMBP2	RIMS binding protein 2	2	DN
5789	PTPRD	protein tyrosine phosphatase, receptor type, D	2	DN
11162	NUDT6	nudix (nucleoside diphosphate linked moiety X)-type motif 6	2	DN
8763	CD164	CD164 molecule, sialomucin	2	DN
29100	TMEM208	transmembrane protein 208	2	DN
686	BTD	biotinidase	2	DN
29104	N6AMT1	N-6 adenine-specific DNA methyltransferase 1 (putative)	2	DN
51471	NAT8B	N-acetyltransferase 8B (GCN5-related, putative, gene/pseudogene)	2	DN
258010	SVIP	small VCP/p97-interacting protein	2	DN
5629	PROX1	prospero homeobox 1	2	DN
5624	PROC	protein C (inactivator of coagulation factors Va and VIIIa)	2	DN
80727	TTYH3	tweety family member 3	2	DN
23475	QPRT	quinolinate phosphoribosyltransferase	2	DN
131474	CHCHD4	coiled-coil-helix-coiled-coil-helix domain containing 4	2	DN
51074	APIP	APAF1 interacting protein	2	DN

65266	WNK4	WNK lysine deficient protein kinase 4	2	DN
8110	DPF3	D4, zinc and double PHD fingers, family 3	2	DN
6780	STAU1	staufen double-stranded RNA binding protein 1	2	DN
4720	NDUFS2	NADH dehydrogenase (ubiquinone) Fe-S protein 2, 49kDa (NADH-coenzyme Q reductase)	2	DN
117247	SLC16A10	solute carrier family 16 (aromatic amino acid transporter), member 10	2	DN
249	ALPL	alkaline phosphatase, liver/bone/kidney	2	DN
2271	FH	fumarate hydratase	2	DN
122970	ACOT4	acyl-CoA thioesterase 4	2	DN
57406	ABHD6	abhydrolase domain containing 6	2	DN
79828	METTL8	methyltransferase like 8	2	DN
3698	ITIH2	inter-alpha-trypsin inhibitor heavy chain 2	2	DN
54566	EPB41L4B	erythrocyte membrane protein band 4.1 like 4B	2	DN
57085	AGTRAP	angiotensin II receptor-associated protein	2	DN
83594	NUDT12	nudix (nucleoside diphosphate linked moiety X)-type motif 12	2	DN
24146	CLDN15	claudin 15	2	DN
7326	UBE2G1	ubiquitin-conjugating enzyme E2G 1	2	DN
3234	HOXD8	homeobox D8	2	DN
1962	EHHADH	enoyl-CoA, hydratase/3-hydroxyacyl CoA dehydrogenase	2	DN
10083	USH1C	Usher syndrome 1C (autosomal recessive, severe)	2	DN

79017	GGCT	gamma-glutamylcyclotransferase	2	DN
7069	THRSP	thyroid hormone responsive	2	DN
5447	POR	P450 (cytochrome) oxidoreductase	2	DN
284541	CYP4A22	cytochrome P450, family 4, subfamily A, polypeptide 22	2	DN
2965	GTF2H1	general transcription factor IIH, polypeptide 1, 62kDa	2	DN
11112	HIBADH	3-hydroxyisobutyrate dehydrogenase	2	DN
29118	DDX25	DEAD (Asp-Glu-Ala-Asp) box helicase 25	2	DN
119032	C10orf32	chromosome 10 open reading frame 32	2	DN
3827	KNG1	kininogen 1	2	DN
23464	GCAT	glycine C-acetyltransferase	2	DN
4285	MIPEP	mitochondrial intermediate peptidase	2	DN
55745	AP5M1	adaptor-related protein complex 5, mu 1 subunit	2	DN
360	AQP3	aquaporin 3 (Gill blood group)	2	DN
3816	KLK1	kallikrein 1	2	DN
55258	THNSL2	threonine synthase-like 2 (S. cerevisiae)	2	DN
10171	RCL1	RNA terminal phosphate cyclase-like 1	2	DN
3758	KCNJ1	potassium channel, inwardly rectifying subfamily J, member 1	2	DN
60488	MRPS35	mitochondrial ribosomal protein S35	2	DN
124935	SLC43A2	solute carrier family 43 (amino acid system L transporter), member 2	2	DN
159371	SLC35G1	solute carrier family 35, member G1	2	DN

51138	COPS4	COP9 signalosome subunit 4	2	DN
63917	GALNT11	polypeptide N-acetylgalactosaminyltransferase 11	2	DN
363	AQP6	aquaporin 6, kidney specific	2	DN
6745	SSR1	signal sequence receptor, alpha	2	DN
2731	GLDC	glycine dehydrogenase (decarboxylating)	2	DN
134147	CMBL	carboxymethylenebutenolidase homolog (Pseudomonas)	2	DN
10165	SLC25A13	solute carrier family 25 (aspartate/glutamate carrier), member 13	2	DN
1810	DR1	down-regulator of transcription 1, TBP-binding (negative cofactor 2)	2	DN
84650	EBPL	emopamil binding protein-like	2	DN
9761	MLEC	malectin	2	DN
132321	C4orf33	chromosome 4 open reading frame 33	2	DN
27109	ATP5S	ATP synthase, H+ transporting, mitochondrial Fo complex, subunit s (factor B)	2	DN
10213	PSMD14	proteasome (prosome, macropain) 26S subunit, non-ATPase, 14	2	DN
3329	HSPD1	heat shock 60kDa protein 1 (chaperonin)	2	DN
112840	WDR89	WD repeat domain 89	2	DN
1773	DNASE1	deoxyribonuclease I	2	DN
157724	SLC7A13	solute carrier family 7 (anionic amino acid transporter), member 13	2	DN
23078	VWA8	von Willebrand factor A domain containing 8	2	DN
51	ACOX1	acyl-CoA oxidase 1, palmitoyl	2	DN
3033	HADH	hydroxyacyl-CoA dehydrogenase	2	DN

1559	CYP2C9	cytochrome P450, family 2, subfamily C, polypeptide 9	2	DN
51458	RHCG	Rh family, C glycoprotein	2	DN
9990	SLC12A6	solute carrier family 12 (potassium/chloride transporter), member 6	2	DN
6518	SLC2A5	solute carrier family 2 (facilitated glucose/fructose transporter), member 5	2	DN
59307	SIGIRR	single immunoglobulin and toll-interleukin 1 receptor (TIR) domain	2	DN
23498	HAAO	3-hydroxyanthranilate 3,4-dioxygenase	2	DN
23382	AHCYL2	adenosylhomocysteinase-like 2	2	DN
50507	NOX4	NADPH oxidase 4	2	DN
55268	ECHDC2	enoyl CoA hydratase domain containing 2	2	DN
55640	FLVCR2	feline leukemia virus subgroup C cellular receptor family, member 2	2	DN
2181	ACSL3	acyl-CoA synthetase long-chain family member 3	2	DN
2180	ACSL1	acyl-CoA synthetase long-chain family member 1	2	DN
55862	ECHDC1	ethylmalonyl-CoA decarboxylase 1	2	DN
51126	NAA20	N(alpha)-acetyltransferase 20, NatB catalytic subunit	2	DN
6341	SCO1	SCO1 cytochrome c oxidase assembly protein	2	DN
6342	SCP2	sterol carrier protein 2	2	DN
5608	MAP2K6	mitogen-activated protein kinase kinase 6	2	DN
81706	PPP1R14C	protein phosphatase 1, regulatory (inhibitor) subunit 14C	2	DN
9351	SLC9A3R2	solute carrier family 9, subfamily A (NHE3, cation proton antiporter 3), member 3 regulator 2	2	DN

5284	PIGR	polymeric immunoglobulin receptor	2	DN
1807	DPYS	dihydropyrimidinase	2	DN
8864	PER2	period circadian clock 2	2	DN
8863	PER3	period circadian clock 3	2	DN
10247	HRSP12	heat-responsive protein 12	2	DN
54502	RBM47	RNA binding motif protein 47	2	DN
388595	TMEM82	transmembrane protein 82	2	DN
5265	SERPINA1	serpin peptidase inhibitor, clade A (alpha-1 antiproteinase, antitrypsin), member 1	2	DN
149483	CCDC17	coiled-coil domain containing 17	2	DN
10099	TSPAN3	tetraspanin 3	2	DN
1628	DBP	D site of albumin promoter (albumin D-box) binding protein	2	DN
11136	SLC7A9	solute carrier family 7 (amino acid transporter light chain, bo,+ system), member 9	2	DN
171425	CLYBL	citrate lyase beta like	2	DN
2981	GUCA2B	guanylate cyclase activator 2B (uroguanylin)	2	DN
525	ATP6V1B1	ATPase, H+ transporting, lysosomal 56/58kDa, V1 subunit B1	2	DN
526	ATP6V1B2	ATPase, H+ transporting, lysosomal 56/58kDa, V1 subunit B2	2	DN
29911	HOOK2	hook microtubule-tethering protein 2	2	DN
523	ATP6V1A	ATPase, H+ transporting, lysosomal 70kDa, V1 subunit A	2	DN

85301	COL27A1	collagen, type XXVII, alpha 1	2	DN
117177	RAB3IP	RAB3A interacting protein	2	DN
4190	MDH1	malate dehydrogenase 1, NAD (soluble)	2	DN
131669	UROC1	urocanate hydratase 1	2	DN
11212	PROSC	proline synthetase co-transcribed homolog (bacterial)	2	DN
11001	SLC27A2	solute carrier family 27 (fatty acid transporter), member 2	2	DN
435	ASL	argininosuccinate lyase	2	DN
23704	KCNE4	potassium channel, voltage gated subfamily E regulatory beta subunit 4	2	DN
145482	PTGR2	prostaglandin reductase 2	2	DN
23710	GABARAPL1	GABA(A) receptor-associated protein like 1	2	DN
8195	MKKS	McKusick-Kaufman syndrome	2	DN
57224	NHSL1	NHS-like 1	2	DN
55277	FGGY	FGGY carbohydrate kinase domain containing	2	DN
93100	NAPRT	nicotinate phosphoribosyltransferase	2	DN
9942	XYLB	xylulokinase homolog (H. influenzae)	2	DN
5816	PVALB	parvalbumin	2	DN
130752	MDH1B	malate dehydrogenase 1B, NAD (soluble)	2	DN
9375	TM9SF2	transmembrane 9 superfamily member 2	2	DN
132	ADK	adenosine kinase	2	DN
7827	NPHS2	nephrosis 2, idiopathic, steroid-resistant (podocin)	2	DN

286676	ILDR1	immunoglobulin-like domain containing receptor 1	2	DN
80157	CWH43	cell wall biogenesis 43 C-terminal homolog (S. cerevisiae)	2	DN
26268	FBXO9	F-box protein 9	2	DN
56894	AGPAT3	1-acylglycerol-3-phosphate O-acyltransferase 3	2	DN
3795	KHK	ketohexokinase (fructokinase)	2	DN
9562	MINPP1	multiple inositol-polyphosphate phosphatase 1	2	DN
134548	SOWAHA	sosondowah ankyrin repeat domain family member A	2	DN
192668	CYS1	cystin 1	2	DN
3249	HPN	hepsin	2	DN
3248	HPGD	hydroxyprostaglandin dehydrogenase 15-(NAD)	2	DN
1486	CTBS	chitobiase, di-N-acetyl-	2	DN
5257	PHKB	phosphorylase kinase, beta	2	DN
1719	DHFR	dihydrofolate reductase	2	DN
10618	TGOLN2	trans-golgi network protein 2	2	DN
7780	SLC30A2	solute carrier family 30 (zinc transporter), member 2	2	DN
114880	OSBPL6	oxysterol binding protein-like 6	2	DN
8504	PEX3	peroxisomal biogenesis factor 3	2	DN
6584	SLC22A5	solute carrier family 22 (organic cation/carnitine transporter), member 5	2	DN
27034	ACAD8	acyl-CoA dehydrogenase family, member 8	2	DN
205	AK4	adenylate kinase 4	2	DN

204	AK2	adenylate kinase 2	2	DN
55840	EAF2	ELL associated factor 2	2	DN
440503	PLIN5	perilipin 5	2	DN
5091	PC	pyruvate carboxylase	2	DN
57132	CHMP1B	charged multivesicular body protein 1B	2	DN
26273	FBXO3	F-box protein 3	2	DN
9376	SLC22A8	solute carrier family 22 (organic anion transporter), member 8	2	DN
10682	EBP	emopamil binding protein (sterol isomerase)	2	DN
56889	TM9SF3	transmembrane 9 superfamily member 3	2	DN
9570	GOSR2	golgi SNAP receptor complex member 2	2	DN
9719	ADAMTSL2	ADAMTS-like 2	2	DN
2936	GSR	glutathione reductase	2	DN
84735	CNDP1	carnosine dipeptidase 1 (metallopeptidase M20 family)	2	DN
3313	HSPA9	heat shock 70kDa protein 9 (mortalin)	2	DN
785	CACNB4	calcium channel, voltage-dependent, beta 4 subunit	2	DN
9391	CIAO1	cytosolic iron-sulfur assembly component 1	2	DN
9390	SLC22A13	solute carrier family 22 (organic anion/urate transporter), member 13	2	DN
79746	ECHDC3	enoyl CoA hydratase domain containing 3	2	DN
509	ATP5C1	ATP synthase, H+ transporting, mitochondrial F1 complex, gamma polypeptide 1	2	DN

638	BIK	BCL2-interacting killer (apoptosis-inducing)	2	DN
6520	SLC3A2	solute carrier family 3 (amino acid transporter heavy chain), member 2	2	DN
6524	SLC5A2	solute carrier family 5 (sodium/glucose cotransporter), member 2	2	DN
151531	UPP2	uridine phosphorylase 2	2	DN
5723	PSPH	phosphoserine phosphatase	2	DN
5833	PCYT2	phosphate cytidylyltransferase 2, ethanolamine	2	DN
217	ALDH2	aldehyde dehydrogenase 2 family (mitochondrial)	2	DN
213	ALB	albumin	2	DN
51179	HAO2	hydroxyacid oxidase 2 (long chain)	2	DN
5909	RAP1GAP	RAP1 GTPase activating protein	2	DN
56521	DNAJC12	DnaJ (Hsp40) homolog, subfamily C, member 12	2	DN
57127	RHBG	Rh family, B glycoprotein (gene/pseudogene)	2	DN
6392	SDHD	succinate dehydrogenase complex, subunit D, integral membrane protein	2	DN
26249	KLHL3	kelch-like family member 3	2	DN
81029	WNT5B	wingless-type MMTV integration site family, member 5B	2	DN
9547	CXCL14	chemokine (C-X-C motif) ligand 14	2	DN
3161	HMMR	hyaluronan-mediated motility receptor (RHAMM)	2	DN
5283	PIGH	phosphatidylinositol glycan anchor biosynthesis, class H	2	DN
51705	EMCN	endomucin	2	DN

84725	PLEKHA8	pleckstrin homology domain containing, family A (phosphoinositide binding specific) member 8	2	DN
1427	CRYGS	crystallin, gamma S	2	DN
53841	CDHR5	cadherin-related family member 5	2	DN
9031	BAZ1B	bromodomain adjacent to zinc finger domain, 1B	2	DN
196740	VSTM4	V-set and transmembrane domain containing 4	2	DN
1738	DLD	dihydrolipoamide dehydrogenase	2	DN
5230	PGK1	phosphoglycerate kinase 1	2	DN
22948	CCT5	chaperonin containing TCP1, subunit 5 (epsilon)	2	DN
1733	DIO1	deiodinase, iodothyronine, type I	2	DN
1737	DLAT	dihydrolipoamide S-acetyltransferase	2	DN
1595	CYP51A1	cytochrome P450, family 51, subfamily A, polypeptide 1	2	DN
5345	SERPINF2	serpin peptidase inhibitor, clade F (alpha-2 antiplasmin, pigment epithelium derived factor), member 2	2	DN
1020	CDK5	cyclin-dependent kinase 5	2	DN
23216	TBC1D1	TBC1 (tre-2/USP6, BUB2, cdc16) domain family, member 1	2	DN
23743	BHMT2	betaine--homocysteine S-methyltransferase 2	2	DN
5733	PTGER3	prostaglandin E receptor 3 (subtype EP3)	2	DN
11318	GPR182	G protein-coupled receptor 182	2	DN
26503	SLC17A5	solute carrier family 17 (acidic sugar transporter), member 5	2	DN

482	ATP1B2	ATPase, Na+/K+ transporting, beta 2 polypeptide	2	DN
51540	SCLY	selenocysteine lyase	2	DN
8301	PICALM	phosphatidylinositol binding clathrin assembly protein	2	DN
130589	GALM	galactose mutarotase (aldose 1-epimerase)	2	DN
200931	SLC51A	solute carrier family 51, alpha subunit	2	DN
27141	CIDEB	cell death-inducing DFFA-like effector b	2	DN
11264	PXMP4	peroxisomal membrane protein 4, 24kDa	2	DN
51205	ACP6	acid phosphatase 6, lysophosphatidic	2	DN
6097	RORC	RAR-related orphan receptor C	2	DN
6652	SORD	sorbitol dehydrogenase	2	DN
199	AIF1	allograft inflammatory factor 1	2	DN
191	AHCY	adenosylhomocysteinase	2	DN
2542	SLC37A4	solute carrier family 37 (glucose-6-phosphate transporter), member 4	2	DN
1457	CSNK2A1	casein kinase 2, alpha 1 polypeptide	2	DN
10330	CNPY2	canopy FGF signaling regulator 2	2	DN
908	CCT6A	chaperonin containing TCP1, subunit 6A (zeta 1)	2	DN
54499	TMCO1	transmembrane and coiled-coil domains 1	2	DN
22977	AKR7A3	aldo-keto reductase family 7, member A3 (aflatoxin aldehyde reductase)	2	DN
54810	GIPC2	GIPC PDZ domain containing family, member 2	2	DN
7009	TMBIM6	transmembrane BAX inhibitor motif containing 6	2	DN

7008	TEF	thyrotrophic embryonic factor	2	DN
2272	FHIT	fragile histidine triad	2	DN
90161	HS6ST2	heparan sulfate 6-O-sulfotransferase 2	2	DN
79090	TRAPPC6A	trafficking protein particle complex 6A	2	DN
2940	GSTA3	glutathione S-transferase alpha 3	2	DN
23632	CA14	carbonic anhydrase XIV	2	DN
79783	SUGCT	succinyl-CoA:glutarate-CoA transferase	2	DN
4318	MMP9	matrix metallopeptidase 9	2	DN
222389	BEND7	BEN domain containing 7	2	DN
495	ATP4A	ATPase, H+/K+ exchanging, alpha polypeptide	2	DN
92815	HIST3H2A	histone cluster 3, H2a	2	DN
4137	MAPT	microtubule-associated protein tau	2	DN
4047	LSS	lanosterol synthase (2,3-oxidosqualene-lanosterol cyclase)	2	DN
55233	MOB1A	MOB kinase activator 1A	2	DN
26228	STAP1	signal transducing adaptor family member 1	2	DN
10901	DHRS4	dehydrogenase/reductase (SDR family) member 4	2	DN
2593	GAMT	guanidinoacetate N-methyltransferase	2	DN
390916	NUDT19	nudix (nucleoside diphosphate linked moiety X)-type motif 19	2	DN
183	AGT	angiotensinogen (serpin peptidase inhibitor, clade A, member 8)	2	DN
2110	ETFDH	electron-transferring-flavoprotein dehydrogenase	2	DN

6048	RNF5	ring finger protein 5, E3 ubiquitin protein ligase	2	DN
2625	GATA3	GATA binding protein 3	2	DN
133121	ENPP6	ectonucleotide pyrophosphatase/phosphodiesterase 6	2	DN
54662	TBC1D13	TBC1 domain family, member 13	2	DN
10471	PFDN6	prefoldin subunit 6	2	DN
9054	NFS1	NFS1 cysteine desulfurase	2	DN
7108	TM7SF2	transmembrane 7 superfamily member 2	2	DN
79152	FA2H	fatty acid 2-hydroxylase	2	DN
79154	DHRS11	dehydrogenase/reductase (SDR family) member 11	2	DN
669	BPGM	2,3-bisphosphoglycerate mutase	2	DN
92558	CCDC64	coiled-coil domain containing 64	2	DN
9829	DNAJC6	DnaJ (Hsp40) homolog, subfamily C, member 6	2	DN
643236	TMEM72	transmembrane protein 72	2	DN
81693	AMN	amnion associated transmembrane protein	2	DN
51185	CRBN	cereblon	2	DN
23788	MTCH2	mitochondrial carrier 2	2	DN
27095	TRAPPC3	trafficking protein particle complex 3	2	DN
2299	FOXI1	forkhead box I1	2	DN
8125	ANP32A	acidic (leucine-rich) nuclear phosphoprotein 32 family, member A	2	DN
140735	DYNLL2	dynein, light chain, LC8-type 2	2	DN

5028	P2RY1	purinergic receptor P2Y, G-protein coupled, 1	2	DN
5025	P2RX4	purinergic receptor P2X, ligand gated ion channel, 4	2	DN
2108	ETFA	electron-transfer-flavoprotein, alpha polypeptide	2	DN
10005	ACOT8	acyl-CoA thioesterase 8	2	DN
2639	GCDH	glutaryl-CoA dehydrogenase	2	DN
2638	GC	group-specific component (vitamin D binding protein)	2	DN
27072	VPS41	vacuolar protein sorting 41 homolog (S. cerevisiae)	2	DN
170961	ANKRD24	ankyrin repeat domain 24	2	DN
84925	DIRC2	disrupted in renal carcinoma 2	2	DN
762	CA4	carbonic anhydrase IV	2	DN
767	CA8	carbonic anhydrase VIII	2	DN
148808	MFSD4	major facilitator superfamily domain containing 4	2	DN
144110	TMEM86A	transmembrane protein 86A	2	DN
27069	GHITM	growth hormone inducible transmembrane protein	2	DN
6456	SH3GL2	SH3-domain GRB2-like 2	2	DN
643008	SMIM5	small integral membrane protein 5	2	DN
91614	DEPDC7	DEP domain containing 7	2	DN
93611	FBXO44	F-box protein 44	2	DN
7429	VIL1	villin 1	2	DN
81689	ISCA1	iron-sulfur cluster assembly 1	2	DN

27242	TNFRSF21	tumor necrosis factor receptor superfamily, member 21	2	DN
150209	AIFM3	apoptosis-inducing factor, mitochondrion-associated, 3	2	DN
245973	ATP6V1C2	ATPase, H+ transporting, lysosomal 42kDa, V1 subunit C2	2	DN
57017	COQ9	coenzyme Q9	2	DN
9104	RGN	regucalcin	2	DN
57498	KIDINS220	kinase D-interacting substrate, 220kDa	2	DN
957	ENTPD5	ectonucleoside triphosphate diphosphohydrolase 5	2	DN
3029	HAGH	hydroxyacylglutathione hydrolase	2	DN
5164	PDK2	pyruvate dehydrogenase kinase, isozyme 2	2	DN
22797	TFEC	transcription factor EC	2	DN
2642	GCGR	glucagon receptor	2	DN
2643	GCH1	GTP cyclohydrolase 1	2	DN
3420	IDH3B	isocitrate dehydrogenase 3 (NAD+) beta	2	DN
9073	CLDN8	claudin 8	2	DN
9071	CLDN10	claudin 10	2	DN
54602	NDFIP2	Nedd4 family interacting protein 2	2	DN
10103	TSPAN1	tetraspanin 1	2	DN
11181	TREH	trehalase (brush-border membrane glycoprotein)	2	DN
125206	SLC5A10	solute carrier family 5 (sodium/sugar cotransporter), member 10	2	DN
79135	APOO	apolipoprotein O	2	DN

4329	ALDH6A1	aldehyde dehydrogenase 6 family, member A1	2	DN
593	BCKDHA	branched chain keto acid dehydrogenase E1, alpha polypeptide	2	DN
594	BCKDHB	branched chain keto acid dehydrogenase E1, beta polypeptide	2	DN
493856	CISD2	CDGSH iron sulfur domain 2	2	DN
445	ASS1	argininosuccinate synthase 1	2	DN
8697	CDC23	cell division cycle 23	2	DN
38	ACAT1	acetyl-CoA acetyltransferase 1	2	DN
35	ACADS	acyl-CoA dehydrogenase, C-2 to C-3 short chain	2	DN
338094	FAM151A	family with sequence similarity 151, member A	2	DN
3898	LAD1	ladinin 1	2	DN
7416	VDAC1	voltage-dependent anion channel 1	2	DN
55157	DARS2	aspartyl-tRNA synthetase 2, mitochondrial	2	DN
1468	SLC25A10	solute carrier family 25 (mitochondrial carrier; dicarboxylate transporter), member 10	2	DN
846	CASR	calcium-sensing receptor	2	DN
223082	ZNRF2	zinc and ring finger 2, E3 ubiquitin protein ligase	2	DN
127845	GOLT1A	golgi transport 1A	2	DN
1317	SLC31A1	solute carrier family 31 (copper transporter), member 1	2	DN
2348	FOLR1	folate receptor 1 (adult)	2	DN
27343	POLL	polymerase (DNA directed), lambda	2	DN

57447	NDRG2	NDRG family member 2	2	DN
9200	HACD1	3-hydroxyacyl-CoA dehydratase 1	2	DN
134288	TMEM174	transmembrane protein 174	2	DN
2651	GCNT2	glucosaminyl (N-acetyl) transferase 2, I-branching enzyme (I blood group)	2	DN
3417	IDH1	isocitrate dehydrogenase 1 (NADP+), soluble	2	DN
116085	SLC22A12	solute carrier family 22 (organic anion/urate transporter), member 12	2	DN
7385	UQCRC2	ubiquinol-cytochrome c reductase core protein II	2	DN
56954	NIT2	nitrilase family, member 2	2	DN
9685	CLINT1	clathrin interactor 1	2	DN
84331	FAM195A	family with sequence similarity 195, member A	2	DN
4482	MSRA	methionine sulfoxide reductase A	2	DN
51156	SERPINA10	serpin peptidase inhibitor, clade A (alpha-1 antiproteinase, antitrypsin), member 10	2	DN
8802	SUCLG1	succinate-CoA ligase, alpha subunit	2	DN
283130	SLC25A45	solute carrier family 25, member 45	2	DN
91749	KIAA1919	KIAA1919	2	DN
83641	FAM107B	family with sequence similarity 107, member B	2	DN
2122	MECOM	MDS1 and EVI1 complex locus	2	DN
51409	HEMK1	HemK methyltransferase family member 1	2	DN
55144	LRRC8D	leucine rich repeat containing 8 family, member D	2	DN

23428	SLC7A8	solute carrier family 7 (amino acid transporter light chain, L system), member 8	2	DN
51084	CRYL1	crystallin, lambda 1	2	DN
327	APEH	acylaminoacyl-peptide hydrolase	2	DN
65125	WNK1	WNK lysine deficient protein kinase 1	2	DN
4121	MAN1A1	mannosidase, alpha, class 1A, member 1	2	DN
9476	NAPSA	napsin A aspartic peptidase	2	DN
124976	SPNS2	spinster homolog 2 (Drosophila)	2	DN
134526	ACOT12	acyl-CoA thioesterase 12	2	DN
3925	STMN1	stathmin 1	2	DN
22921	MSRB2	methionine sulfoxide reductase B2	2	DN
81889	FAHD1	fumarylacetoacetate hydrolase domain containing 1	2	DN
60592	SCOC	short coiled-coil protein	2	DN
56922	MCCC1	methylcrotonoyl-CoA carboxylase 1 (alpha)	2	DN
10615	SPAG5	sperm associated antigen 5	2	DN
9099	USP2	ubiquitin specific peptidase 2	2	DN
10434	LYPLA1	lysophospholipase I	2	DN
10125	RASGRP1	RAS guanyl releasing protein 1 (calcium and DAG-regulated)	2	DN
5095	PCCA	propionyl CoA carboxylase, alpha polypeptide	2	DN
5096	PCCB	propionyl CoA carboxylase, beta polypeptide	2	DN
2665	GDI2	GDP dissociation inhibitor 2	2	DN

6821	SUOX	sulfite oxidase	2	DN
10542	LAMTOR5	late endosomal/lysosomal adaptor, MAPK and MTOR activator 5	2	DN
1644	DDC	dopa decarboxylase (aromatic L-amino acid decarboxylase)	2	DN
84216	TMEM117	transmembrane protein 117	2	DN
7010	TEK	TEK tyrosine kinase, endothelial	2	DN
342527	SMTNL2	smoothelin-like 2	2	DN
2805	GOT1	glutamic-oxaloacetic transaminase 1, soluble	2	DN
732	C8B	complement component 8, beta polypeptide	2	DN
9027	NAT8	N-acetyltransferase 8 (GCN5-related, putative)	2	DN
131920	TMEM207	transmembrane protein 207	2	DN
8564	KMO	kynurenine 3-monooxygenase (kynurenine 3-hydroxylase)	2	DN
23185	LARP4B	La ribonucleoprotein domain family, member 4B	2	DN
55353	LAPTM4B	lysosomal protein transmembrane 4 beta	2	DN
112817	HOGA1	4-hydroxy-2-oxoglutarate aldolase 1	2	DN
51382	ATP6V1D	ATPase, H+ transporting, lysosomal 34kDa, V1 subunit D	2	DN
27293	SMPDL3B	sphingomyelin phosphodiesterase, acid-like 3B	2	DN
51090	PLLP	plasmolipin	2	DN
354	KLK3	kallikrein-related peptidase 3	2	DN
359	AQP2	aquaporin 2 (collecting duct)	2	DN
2326	FMO1	flavin containing monooxygenase 1	2	DN

2329	FMO4	flavin containing monooxygenase 4	2	DN
4708	NDUFB2	NADH dehydrogenase (ubiquinone) 1 beta subcomplex, 2, 8kDa	2	DN
288	ANK3	ankyrin 3, node of Ranvier (ankyrin G)	2	DN
287	ANK2	ankyrin 2, neuronal	2	DN
3766	KCNJ10	potassium channel, inwardly rectifying subfamily J, member 10	2	DN
166815	TIGD2	tigger transposable element derived 2	2	DN
22849	CPEB3	cytoplasmic polyadenylation element binding protein 3	2	DN
387338	NSUN4	NOP2/Sun domain family, member 4	2	DN
1565	CYP2D6	cytochrome P450, family 2, subfamily D, polypeptide 6	2	DN
5151	PDE8A	phosphodiesterase 8A	2	DN
10605	PAIP1	poly(A) binding protein interacting protein 1	2	DN
4969	OGN	osteoglycin	2	DN
9317	PTER	phosphotriesterase related	2	DN
120939	TMEM52B	transmembrane protein 52B	2	DN
8417	STX7	syntaxin 7	2	DN
202151	RANBP3L	RAN binding protein 3-like	2	DN
29958	DMGDH	dimethylglycine dehydrogenase	2	DN
84888	SPPL2A	signal peptide peptidase like 2A	2	DN
1188	CLCNKB	chloride channel, voltage-sensitive Kb	2	DN
1183	CLCN4	chloride channel, voltage-sensitive 4	2	DN

51015	ISOC1	isochorismatase domain containing 1	2	DN
48	ACO1	aconitase 1, soluble	2	DN
11179	ZNF277	zinc finger protein 277	2	DN
56606	SLC2A9	solute carrier family 2 (facilitated glucose transporter), member 9	2	DN
795	S100G	S100 calcium binding protein G	2	DN
27199	OXGR1	oxoglutarate (alpha-ketoglutarate) receptor 1	2	DN
6546	SLC8A1	solute carrier family 8 (sodium/calcium exchanger), member 1	2	DN
28976	ACAD9	acyl-CoA dehydrogenase family, member 9	2	DN
479	ATP12A	ATPase, H+/K+ transporting, nongastric, alpha polypeptide	2	DN