A systematic pathway-based network approach for *in silico* drug repositioning

Katjuša Koler

Department of Computer Science
Faculty of Engineering
The University of Sheffield

A thesis submitted in partial fulfilment of the requirements for the degree of
Doctor of Philosophy

March 2020

Acknowledgements

A big thank you to the Department of Computer Science and my main supervisor Winston Hide for supporting me financially and to the Sheffield Institute of Translational Neuroscience for giving me a place where I could do all the thinking, tea-drinking, and researching.

Thank you once again to Winston Hide and Wenbin Wei for supervising me and my science and to Dennis Wang for taking me in during the final stages of my PhD. Thank you to Gabriel Altschuler and Yered Pita-Juárez for giving me a great starting point for the development of my project. To Lester Kobzik, thank you for continuous enthusiasm and interest in my progress, as well as valuable advice and support on the Juvenile Arthritis case study. Thank you to Oliver Bandmann and Doo Yeon Kim for trusting me with their data and believing in my project. Thank you to Sokratis Kariotis for providing me with coding support for the computational improvements of the prototype methods and to Pourya Naderi for continuing my work and reviewing the arguably driest of my thesis chapters. Thank you to Sarah Morgan for invaluable feedback on several chapters.

Special thanks to my friend DRJ for making me a better programmer and for being a friend throughout. From teaching me so many programming concepts, regular expressions, and command-line magic to continuing being a great lunch buddy. Thank you for your significant input in the development of the KATdb idea and for opening my eyes to the world of kerning.

I would also like to thank my dear friends from different walks of PhD and Sheffield life. Mollie and Chris, thanks for fancy beer-drinking breaks and Peak District walks. Claire, Mogs, Ellen, Monika, and Hannah, thank you for keeping me sane over teas, lunches, cake clubs, and many drinks.

Hvala moji družini in prijateljem za podporo, razumevanje in razvajanje ob vsakem obisku doma. Hvala za interes v moje delo, čeprav mi je velikokrat zmanjkalo slovenskih

besed in vam nisem znala opisati kaj sem sploh počela. Predvsem pa hvala za vedno optimistične spodbudne besede in za sproščujoče počitnice.

My biggest thanks to my *najboljšejši* partner David for getting me out of bed, making me breakfast (and dinner), and most importantly for loving and supporting me every day. I am grateful (and also really impressed) for the levels of patience and understanding you have shown me during the hardest of times. I will love and support you, but also feed and cuddle you during your journey. I wouldn't have come so far without you. I am forever grateful to have you in my life.

Abstract

Drug repositioning, the method of finding new uses for existing drugs, holds the potential to reduce the cost and time of drug development. Successful drug repositioning strategies depend heavily on the availability and aggregation of different drug and disease databases. Moreover, to yield greater understanding of drug prioritisation approaches, it is necessary to objectively assess (benchmark) and compare different methods.

Data aggregation requires extensive curation of non-standardised drug nomenclature. To overcome this, we used a graph-theoretic approach to construct a drug synonym resource that collected drug identifiers from a range of publicly available sources, establishing missing links between databases. Thus, we could systematically assess the performance of available *in silico* drug repositioning methodologies with increased power for scoring true positive drug-disease pairs.

We developed a novel pathway-based drug repositioning pipeline, based on a bipartite network of pathway- and drug-gene set correlations that captured functional relationships. To prioritise drugs, we used our bipartite network and the differentially expressed pathways in a given disease that formed a disease signature. We then took the cumulative network correlation between disease pathway and drug signatures to generate a drug prioritisation score. We prioritised drugs for three case studies: juvenile idiopathic arthritis, Alzheimer's and Parkinson's disease. We explored the use of different true positive lists in the evaluation of drug repositioning performance, providing insight into the most appropriate benchmark designs.

We have identified several promising drug candidates and showed that our method successfully prioritises disease-modifying treatments over drugs offering symptomatic relief. We have compared the pipeline's performance to an alternative well-established method and showed that our method has increased sensitivity to current treatment trends. The successful translation of drug candidates identified in this thesis has the potential to speed up the drug-discovery pipeline and thus more rapidly and efficiently deliver disease-modifying treatments to patients.

Table of contents

List of figures

List of tables

Nomenclature

Gene and Protein Acronyms

α-Syn	α-synuclein
3’UTR	3 prime untranslated region
6-OHDA	6-hydroxydopamine
Aβ	amyloid-β
AChE	acetylcholinesterase
AHSP	alpha-haemoglobin stabilising protein
ALDH	aldehyde dehydrogenase
AP1	activator protein 1
APOBEC3G	apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like 3G
APP	β-amyloid precursor protein
ASH2L	absent, small, or homeotic-like
Bcl2	B-cell lymphoma 2
BCLAF1	BCL-associated factor 1
C2	complement component 2
C4	complement component 4
CBX4	chromobox 4
CD40	cluster of differentiation 40
CDC42	cell division control protein 42 homolog
CD	cluster of differentiation
CDK	cyclin-dependent kinase
CII	type II collagen
COX	cyclooxygenases
CYP	cytochrome P450
EGFR	epidermal growth factor receptor
EphB1	ephrin
EPO	erythropoietin
ERK	extracellular signal-regulated kinase
FLT3	FMS-like receptor tyrosine kinase 3
GABAB	gamma-aminobutyric acid B
HDAC	histone deacetylase
HIF	hypoxia inducible factor

HLA	human leukocyte antigen
Hsp90	heat shock protein 90
IgM	immunoglobulin M
I-κB	inhibitor of NF-κB
IL-1	interleukin-1
IL-6	interleukin-6
iNOS	inducible nitric oxide synthase
JAK2	janus kinase 2
JNK	c-Jun N-terminal kinase
KO	knockout
MAPK	mitogen-activated protein kinases
MHC	major histocompatibility complex
MPTP	1-methyl-4-phenyl-1,2,3,6-tetrahydropyridine
mTOR	mammalian target of rapamycin
NF-κB	nuclear factor κ-light-chain-enhancer of activated B cells
NFT	neurofibrillary tangle
NK	natural killer
NLRP3	NOD-like receptor family, pyrin domain containing 3
NMDA	N-methyl-D-aspartate
PAR4	protease-activated receptor 4
PARP	poly(ADP-Ribose) polymerase
PDE3A	phosphodiesterase 3A
PI3K	phosphoinositide 3-kinase
PKC	protein kinase C
PKL1	polo-like kinase 1
PSEN1	presenilin 1
Pyk2	protein-tyrosine kinase 2
RAC1	Ras-related C3 botulinum toxin substrate 1
RAP1	repressor activator protein 1
RNA pol III	RNA polymerase III
ROS	reactive oxygen species
SAHA	suberoylanilide hydroxamic acid
SEPT2	septin 2
SFPQ	splicing factor proline- and glutamine-rich
SLC	solute carrier
SNCA	gene encoding α-synuclein
STAT3	signal transducer and activator of transcription 3
SUMO	small ubiquitin-like modifier
Tcf3	T-cell factor 3
TCRA	T-cell receptor activation
TCR	T-cell receptor
TGFβ	transforming growth factor beta

Th1	T helper 1 cells
Th2	T helper 2 cells
TIA1	T-cell intracellular antigen-1
TMD	transmembrane domain
TNFα	tumour necrosis factor alpha
TNF	tumour necrosis factor
UPS	ubiquitin proteasome system
VEGFR	vascular endothelial growth factor receptor 2
VEGF	vascular endothelial growth factor
ZAP70	zeta-associated protein of 70kD

Other Symbols

$\leftrightarrow$	edge between two nodes
E	edge between two vertices
G	graph in graph theory sense, consisting of edges *E*, and vertices *V*
$\hat{r}$	correlation estimate
t	t-test statistic
V	vertex or node

Glossary

Aβ42/40 ratio	plays a role in Alzheimer's disease pathogenesis
A5	Alzheimer's disease 3D cell culture model with low Aβ42/40 ratio
accuracy	$\frac{\text{TP+TN}}{\text{total}} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}}$
authority	regulatory body responsible for maintaining the identifier, type of name, e.g. synonym or DrugBank
betweenness centrality	measures the extent to which a vertex lies on the shortest path between other vertices
Biocarta	pathway database
bipartite	or two-mode network, is a network where nodes belong to two distinct sets and there are only edges between two nodes from different sets and no edges between nodes in the same set
BRD ID	Broad ID, the primary drug identifier in LINCS and LINCS-derived drug signature datasets
CAS number	unique numerical identifier assigned by CAS to every chemical substance described in open scientific literature, also known as CAS Registry Number, CASRN
ChEBI	Chemical entities of biological interest database developed by EBI
ChEMBL	chemical database developed by EMBL
chirality	the geometric property of a rigid object or spatial arrangement of points or atoms
CMap	Connectivity Map, predecessor to LINCS, database of drug perturbation signatures
connected component	is a subnetwork in which any two nodes are connected to each other but are not connected to any nodes outside that subnetwork
edge betweenness centrality	measures the extent to which the shortest path between other vertices passes through an edge

enantiomer	one of two molecules which are mirror images of each other, also known as optical isomers
FN	false negative, the number of incorrect classifications of the positive case
FP	false positive, the number of incorrect classifications of the negative case
FPR	false positive rate, $\frac{\text{FP}}{\text{actual false}} = \frac{\text{FP}}{\text{FP+TN}}$
H10	Alzheimer's disease 3D cell culture model with high Aβ42/40 ratio
I45F	Alzheimer's disease 3D cell culture model with high Aβ42/40 ratio
I47F	Alzheimer's disease 3D cell culture model with low Aβ42/40 ratio
InChI	the IUPAC International Chemical Identifier, textual identifier for chemical substances
InChIKey	condensed, 27-character version of hashed InChI (using the SHA-256 algorithm)
INN	international non-proprietary name, generic name
KEGG	pathway database
L1000CDS	drug signature database consisting of LINCS Level 3 data processed with a characteristic direction measure
L1000	gene expression profiling method for cost-effective, high-throughput screening, profiling 978 landmark transcripts
LINCS	database of drug perturbation signatures
MOA	mechanism of action refers to the specific biochemical interaction through which a drug substance produces its pharmacological effect
MoA	mode of action describes a functional or anatomical change, resulting from the exposure of a living organism to a substance
modularity	is a measure of the strength of division of a network into clusters or modules measuring the density of edges inside communities to edges outside communities
MSigDB	pathway database
PID	pathway database
precision	$\frac{\text{TP}}{\text{predicted true}} = \frac{\text{TP}}{\text{TP+FP}}$
PubChem CID	PubChem Compound ID, unique chemical structures in PubChem
PubChem SID	PubChem Substance ID, depositor-provided substance descriptions
PubChem	database maintained by the US NIH
PubMed	search engine accessing database of references and abstracts on life sciences and biomedical topics
Reactome	pathway database
recall	$\frac{\text{TP}}{\text{actual true}} = \frac{\text{TP}}{\text{TP+FN}}$, also known as true positive rate or sensitivity
ROC curve	receiver operating characteristics curve, showing false positive rate on x axis and sensitivity on y
RxNorm CUI	identifier for the normalised clinical drug dictionary of the Unified Medical Language System
sensitivity	$\frac{\text{TP}}{\text{actual true}} = \frac{\text{TP}}{\text{TP+FN}}$, also known as true positive rate or recall
specificity	1 - false positive rate

stereoisomerism	or spatial isomerism is when molecules have the same molecular formula and the same sequence of bonded atoms, but differ in 3D orientation of their atoms in space
TNR	true negative rate, $\frac{\text{TN}}{\text{actual false}} = \frac{\text{TN}}{\text{TN}+\text{FP}}$, also known as specificity, 1 − FPR
TN	true negative, the number of correct classifications of the negative case
TPR	true positive rate, $\frac{\text{TP}}{\text{actual true}} = \frac{\text{TP}}{\text{TP}+\text{FN}}$, also known as sensitivity or recall
TP	true positive, the number of correct classifications of the positive case
unipartite	or one-mode network, is a network where all nodes belong to one set

Acronyms and Abbreviations

3DDS	3D drugs screen
AACT	Aggregate analysis of clinicaltrials.gov database
AD	Alzheimer's disease
ADCL	Alzheimer's disease curated list
ADHD	attention deficit hyperactivity disorder
ADMET	Absorption, Distribution, Metabolism, Excretion, and Toxicity
AIDS	acquired immune deficiency syndrome
ALS	amyotrophic lateral sclerosis
ASSESS	Analysis of sample set enrichment scores
ATC	Anatomical Therapeutic Chemical
AUC	area under the receiver operating characteristics (ROC) curve
AUPRC	Area under Precision-recall curve
AUROC	Area under the receiver operating characteristics curve
BBB	blood-brain barrier
BRD	Broad ID
CAS	Chemical Abstract Service
CasRN	CAS Registry Number, also known as CASRN or CAS Number
ChEBI	Chemical Entities of Biological Interest
ChEMBL	Chemicals EMBL
CIA	collagen-induced arthritis
CID	Compound ID
CMap	Connectivity Map
conc.	concentration
COVID-19	coronavirus disease 2019
COVID-19	coronavirus disease 19
cpm	Counts Per Million
csv	comma separated values
CTD	Comparative Toxicogenomics Database
CUI	Concept Unique Identifier
DB	database
db	database
DE	differential expression
DEG	differentially expressed gene

DEP	differentially expressed pathway
DMARD	disease modifying antirheumatic drug
DNI	Drugs of New Indications (Liu et al., 2013), also referred to as Liu2013
DO	Disease ontology
DPD	Drugs Product Database
EBI	European Bioinformatics Institute
ECHA	European Chemicals Agency
ECM	extracellular matrix
EHR	electronic health records
EINECS	European Inventory of Existing Commercial Chemical Substances
EMA	European Medicines Agency
EMBL	European Molecular Biology Laboratory
EPAR	European public assessment report
Eq.	equation
ER	endoplasmic reticulum
EU	European Union
fAD	familial Alzheimer's disease
FAERS	FDA adverse event reporting system
FC	Fold Change
FDA	the US Food and Drug Administration
FDR	False Discovery Rate
Fig.	figure
FLS	fibroblasts-like synoviocytes
FN	false negative
FP	false positive
FPR	false positive rate
fRMA	frozen robust multi-array analysis
FTP	file transfer protocol
GB	gigabyte
GCH1	GTP cyclohydrolase 1
GEO	Gene Expression Omnibus
GNUSE	global normalized unscaled standard error
GO	gene ontology
GPL11154	Illumina HiSeq 2000
GPL570	Affymetrix Human Genome U133 Plus 2.0 Array
GPL96	Affymetrix Human Genome U133A Array
GPL97	Affymetrix Human Genome U133B Array
GSEA	gene set enrichment analysis
GSRS	Global Substance Registration System
GSVA	Gene set enrichment analysis
GTP	guanosine-5'-triphosphate

GWAS	genome wide association studies
h	hour
HIV	human immunodeficiency virus
hNPCs	human neuronal progenitor cells
HOM	homozygous
HPV	human papillomavirus
ID	identifier
ILAR	International League of Associations for Rheumatology
InChI	IUPAC International Chemical Identifier
INN	international non-proprietary name
IUPAC	International Union of Pure and Applied Chemistry
JIA	juvenile idiopathic arthritis
polyJIA	polyarticular juvenile idiopathic arthritis
sJIA	systemic juvenile idiopathic arthritis
KEGG	Kyoto Encyclopedia of Genes and Genomes
L1000CDS^2	LINCS L1000 characteristic direction signature search engine
L1000CDS	L1000 Characteristic Direction Signature
L1000FWD	L1000 fireworks display
limma	Linear models for microarray data
LINCS	Library of Integrated Network-based Cellular Signatures
logFC	$\log_2$ fold change
LPAwb+	Label Propagation Algorithm weighted bipartite plus
LSM	LINCS standardized unique small molecule
MESH	Medical Subject Headings
MOA	mechanism of action
MoA	mode of action
MSigDB	Molecular Signatures Database
NaB	sodium butyrate
n/a	not applicable
NDF-RT	National Drug File - Reference Terminology
NIH	National Institutes of Health (United States)
N	negative
NSAID	non-steroidal anti-inflammatory drug
OMIM	Online Mendelian Inheritance in Man
PBMC	peripheral blood mononuclear cells
PCA	principal component analysis
PCxN	Pathway Coexpression Network
PDN	Pathway Drug Network
PD	Parkinson’s disease
PDxN	Pathway Drug Coexpression Network
pert id	perturbagen ID
pert time	perturbagen time

PharmGKB	Pharmacogenomics Knowledge Base
PheWAS	phenome wide association studies
PID	Pathway Interaction Database
P	positive
PRC	Precision-Recall curve
PREDICT	Prediction of drug indications
pval	*p*-value
QC	quality control
qval	*q*-value, adjusted *p*-value
RA	rheumatoid arthritis
RF-	rheumatoid factor negative
RIN	RNA integrity number
ROC	receiver operating characteristics
SD	standard deviation
SID	Substance ID
SITraN	Sheffield Institute for Translational Neuroscience
SMILES	simplified molecular-input line-entry system
sPD	sporadic Parkinson's disease
SPL	Structured Product Labels
SSRI	selective serotonin reuptake inhibitor
TMM	trimmed mean of M values
TNR	true negative rate
TN	true negative
TPR	true positive rate
TP	true positive
TTD	Therapeutic Target Database
UK	The United Kingdom of Great Britain and Northern Ireland
μM	micro (μ) molar, concentration unit
UMLS	Unified Medical Language System
UNII	Unique Ingredient Identifier
URL	uniform resource locator
USD	United States dollar
US	The United States of America
VHA	Veterans Health Administration
vmem	virtual memory
WHOCC	World Health Organisation Collaborating Centre
WHO	World Health Organisation
WT	wild type
yrs	years

Chapter 1

Introduction to the Thesis

1.1 Motivation

Drug repositioning, also referred to as drug repurposing, is the process of finding new uses for existing drugs. It holds the potential to reduce the price of developing new drug candidates and fast-track drug discoveries. It has significant translational potential to identify novel drug candidates for disease with no approved or disease-modifying interventions. Successful drug repositioning strategies depend heavily on the availability and aggregation of different information resources. With the increased availability of publicly available data, we can now develop powerful new methods that can provide new insights into drug discovery and disease progression.

However, with large amounts of data, come organisational and aggregation challenges because of non-standardised naming conventions. In addition, to successfully translate proposed novel indications from studies to clinical treatments, it is necessary to objectively assess and compare currently available drug repositioning methods. It has become challenging to perform this comparison, because poorly established links between databases lead to incorrect scoring of true positives as false negatives and thus affect the performance score, clouding insight into best drug repositioning approaches.

To overcome the multiple chemical naming conventions a drug synonym database was constructed to aid in benchmarking of drug repositioning methods. Applying this system together with a novel well-characterised disease-agnostic signature-based drug repositioning method, we have leveraged the increasing amounts of publicly available data in order to expand the landscape of the current drug repositioning methods and to deliver a new pathway-based approach to gain insight into the mechanism of disease.

1.2 Contributions

There are four main contributions to knowledge from work described in this thesis:

(i) A novel well-characterised pathway-based coexpression network drug repositioning pipeline,
(ii) Development of a drug synonym database aiding in drug repositioning pipeline performance assessments,
(iii) Characterisation and performance evaluation for *in silico* drug repositioning results,
(iv) Computational performance improvement of a published method increasing availability and further development.

In addition, this work has already contributed to securing financial support for development of this project beyond the completion of the current project. The work described in this thesis has been a significant contribution to NIH R01 research grant proposal entitled "The Alzheimer's Disease Resiliome: Pathway Analysis and Drug Discovery"*(http://grantome.com/grant/NIH/R01-AG062547-01)* awarded $668,819 per year for 5 years (subject to annual review, $3,344,097 total).

1.3 Publications

Y. Pita-Juárez, G. Altschuler, S. Kariotis, W. Wei, K. Koler, C. Green, R. Tanzi, and W. Hide. The pathway coexpression network: Revealing pathway relationships. *PLoS Comput. Biol.*, 14(3):e1006042, 2018.

K. Koler, S. L. Morgan, D. R. Jones, D. Wang, and W. A. Hide. KATdb: a graph theoretic approach to unification of drug names. Manuscript in Preparation, 2020.

H. Larbalestier, M. Keatinge, L. Trollope, E. White, S. Gowda, W. Wei, K. Koler, S. Semenova, N. Rimmer, S. Sweeney, J. Mazzolini, D. Sieger, W. A. Hide, R. Macdonald, J. McDearmid, P. Panula, and O. Bandmann. Tyrosine hydroxylase depletion and inflammatory dysregulation in a zebrafish gch1$^{-/-}$ Parkinson's disease model. Manuscript in Preparation, 2020.

S. L. Morgan, P. Naderi, K. Koler, Y. Pita-Juarez, I. Vlachos, and W. A. Hide. Are all pathways related to Alzheimer's disease? Manuscript in Preparation, 2020.

D. von Maydell, K. Koler, M. Jorfi, E. Brand, J. Aronson, K. J. Washicosky, S. S. Kwak, M. Cetinbas, R. Sadreyev, J. Park, S. L. Wagner, W. Hide, R. E. Tanzi, and D. Y. Kim. Identifying shared enriched pathways driven by pathogenic Aβ accumulation in 3D cellular models and human Alzheimer's brain. Manuscript in Preparation, 2020.

Acknowledged in M. P. Menden, D. Wang, M. J. Mason, B. Szalai, K. C. Bulusu, Y. Guan, T. Yu, J. Kang, M. Jeon, R. Wolfinger, T. Nguyen, M. Zaslavskiy, AstraZeneca-Sanger Drug Combination DREAM Consortium, I. S. Jang, Z. Ghazoui, M. E. Ahsen, R. Vogel, E. C. Neto, T. Norman, E. K. Y. Tang, M. J. Garnett, G. Y. D. Veroli, S. Fawell, G. Stolovitzky, J. Guinney, J. R. Dry, and J. Saez-Rodriguez. Community assessment to advance computational prediction of cancer drug combinations in a pharmacogenomic screen. *Nat. Commun.*, 10(1):2674, 2019. ISSN 2041-1723. doi: 10.1038/s41467-019-09799-2.

1.4 Organisation of the Thesis

The thesis is divided into the following chapters:

- Chapter 1: Introduction to the Thesis
- Chapter 2: Background
- Chapter 3: Materials and Methods
- Chapter 4: KATdb, the Drug Synonym Database
- Chapter 5: The Drug Repositioning Pipeline
- Chapter 6: Evaluation of the System: Application to juvenile idiopathic arthritis (JIA)
- Chapter 7: Case Studies: Neurodegenerative Diseases
- Chapter 8: Conclusions

Chapter 1 summarises the key contributions and publications (including manuscripts in preparation) that have resulted from the work related to or the work described in this thesis. Chapter 2 offers context and background information to the drug repositioning field and its relevance to the work presented in this thesis. Chapter 3 is a description of the materials and methods used in later chapters. Chapter 4 is a stand-alone project describing the development of a drug synonym database that has been utilised in benchmarking the drug repositioning pipeline described in the following chapters. Chapter 5 describes and characterises the drug repositioning pipeline presented as the key contribution of this thesis. It includes a description of development and characterisation of each individual step of the pipeline that was then applied to three different case studies in Chapters 6 and 7. Chapters 6 and 7 describe in detail the application of the drug repositioning pipeline to juvenile rheumatoid arthritis (Chapter 6), and two neurodegenerative diseases: Alzheimer's and Parkinson's disease (Chapter 7). Chapter 8 offers a summary discussion of the work presented and outlines points for future development.

1.5 Others' Contributions to the Thesis

Chapter 4: KATdb, the Drug Synonym Database: David R. Jones (Sheffield Institute for Translational Neuroscience (SITraN), University of Sheffield) has contributed to the concept of using graph theory to resolve heterogeneous drug nomenclature.

Chapter 5: The Drug Repositioning Pipeline: Gabriel Altschuler (SITraN, University of Sheffield) was instrumental in the design of the initial PhD project proposal and has demonstrated the adopted overarching concept for a drug repurposing method in Joachim et al. (2018). Yered Pita-Juárez's PhD project (Department of Biostatistics, Harvard T.H. Chan School of Public Health) served as the basis for the underlying method used in the coexpression network construction. This work was described in Pita-Juárez et al. (2018). Sokratis Kariotis (SITraN, University of Sheffield) provided coding support in the computational improvement of the Pita-Juárez et al. (2018) method.

Chapter 6: Evaluation of the System: Application to juvenile idiopathic arthritis (JIA): Professor Lester Kobzik (Department of Environmental Health, Harvard T.H. Chan School of Public Health) assisted in the curation of the publicly available juvenile idiopathic arthritis studies and advised on the interpretation of the results in relation to the standard treatment practice.

Chapter 7: Case Studies: Neurodegenerative Diseases: Alzheimer's disease: Assistant Professor Doo Yeon Kim (Department of Neurology, Massachusetts General Hospital, Harvard Medical School) and Professor Rudolph E. Tanzi (Department of Neurology, Massachusetts General Hospital, Harvard Medical School) shared preprocessed 3D cell model RNA-Seq data described in Kwak et al. (2020), and the positive drug hits from the 3D drug screen. The RNA-Seq data was preprocessed by the Harvard Bioinformatics Core. Assistant Professor Doo Yeon Kim has also kindly provided context for their 3D models and their characterisation. Sarah Morgan (SITraN, University of Sheffield; Department of Pathology, Beth Israel Deaconess Medical Center, Harvard Medical School) advised on the Mayo Alzheimer's disease dataset sample selection. **Parkinson's disease:** Professor Oliver Bandmann (SITraN, University of Sheffield) shared the human sporadic Parkinson's disease RNA-Seq data described in Carling et al. (2020) and the zebrafish GCH1 mutant data described in Larbalestier et al. (2020). Professor Oliver Bandmann has also kindly curated a list of neuroprotective drugs. The Carling et al. (2020) RNA-Seq data was preprocessed by Claire Green (SITraN, University of Sheffield) and the Larbalestier et al. (2020) zebrafish data by Wenbin Wei (SITraN, University of Sheffield).

All draft chapters were reviewed by my main supervisor Winston A. Hide. In addition, Sarah Morgan provided valuable feedback on Chapters 1, 2, 6 and 7, David R. Jones for Chapter 4, Pourya Naderi for Chapters 3 and 5, Lester Kobzik for Chapter 6, and David Rapley reviewed Chapter 8.

Chapter 2

Background

2.1 The Value of Drug Repositioning

With high rates of clinical trial failures and stricter rules on drug safety, *de novo* drug discovery is both expensive, time consuming, and has a high failure rate. Recent reports are putting the cost of drug development at between \$0.7-2.6 billion (in 2013 USD) with a yearly increase of 8.5% over general price inflation (DiMasi et al., 2016; Prasad and Mailankody, 2017). The number of drugs that are approved has not shown a concomitant increase, despite major advances in science and technology, as well as increased investments in *de novo* drug discovery (Oprea and Overington, 2015). With decreasing productivity, finding new uses for already existing drugs has become an appealing option.

Drug repositioning, or drug repurposing, is the development of new drugs by using already approved or investigational compounds (Shameer et al., 2017). It offers a cheaper and lower-risk alternative to *de novo* drug discovery. Importantly, reuse of existing approved drugs has the added benefit of rapid use for patient treatment—drugs have known toxicity assessments and known side effects as well as history of patient use. There are numerous successful drug repositioning examples and even more in development (Table 2.1). In particular, the past decade has seen an increasing number of drug repositioning studies (Brum et al., 2015; Cheng et al., 2014; Dudley et al., 2011b; Ferrero and Agarwal, 2018; Sirota et al., 2011; Zhang et al., 2016) (Fig. 2.1). However, the majority of research and development spending is still allocated for new drug discoveries (Booth and Zemmel, 2004).

Table (2.1) Examples of approved repositioned drugs and those in clinical trials. Drugs and their indications were extracted from Gns et al. (2019); Park (2019); Pushpakom et al. (2019); Rosa and Santos (2020) (referenced as † § ‡ þ, respectively). Year of approval for the new indication is listed for successfully repositioned drugs and the phase number and its associated clinical trial IDs are listed for drugs in clinical trials. AD — Alzheimer's disease; ADHD — attention deficit hyperactivity disorder; AIDS — acquired immune deficiency syndrome; ALS — amyotrophic lateral sclerosis; COVID-19 — coronavirus disease 2019; HIV — human immunodeficiency virus; JIA — juvenile idiopathic arthritis; PD — Parkinson's disease; RA — rheumatoid arthritis; ‡ — Pushpakom et al. (2019); § — Park (2019); þ — Rosa and Santos (2020); † — Gns et al. (2019); ∗ — curated by Katjuša Koler.

Drug	First indication	New indications	Year of approval / Clinical trial phase	Source
atomoxetine	PD	ADHD	2002	‡
aspirin	inflammation, pain	antiplatelet, colorectal cancer	1998, 2015	‡§
raloxifene	osteoporosis	breast cancer	2007	‡§
ketoconazole	fungal infections	cushing syndrome	2014	‡
sildenafil	angina	erectile dysfunction	1998	‡§
celecoxib	inflammation, pain	familial adenomatous polyps	2000	‡
allopurinol	cancer	gout	1966	§
minoxidil	hypertension	hair loss	1988	‡
finasteride	benign prostatic hyperplasia	hair loss	1997	§
zidovudine	cancer	HIV/AIDS	1987	‡§
propranolol	hypertension	migraine headache	1974	§
gabapentin	epilepsy	neuropathic pain	2002	§
gemcitabine	antiviral	non-small cell lung cancer, metastatic breast cancer	1998, 2004	§
topiramate	epilepsy	obesity	2012	‡
dapoxetine	analgesia and depression	premature ejaculation	2012	‡
methotrexate	cancer	RA	1988	§
rituximab	cancer	RA	2006	‡
bupropion	depression	smoking cessation	1997	§
duloxetine	depression	stress urinary incontinence	2004	‡§
thalidomide	morning sickness	leprosy, multiple myeloma, COVID-19	1998,2006, 2 (NCT04273529, NCT04273581)	‡§þ
fingolimod	transplant rejection	multiple sclerosis, COVID-19	2010, 2 (NCT04280588)	‡þ
bromocriptine	PD	type 2 diabetes, AD	2009, 1-2 (NCT04413344)	§∗
levetiracetam	epilepsy	AD	2 (NCT04004702, NCT02002819, NCT03875638, NCT03489044)	∗

continues on the next page

Table 2.1 continued

Drug	**First indication**	**New indications**	**Year of approval / Clinical trial phase**	**Source**
vitamin C	vitamin C deficiency, wounds	COVID-19	2 (NCT04264533, NCT04363216, NCT04401150, NCT04344184)	þ
darunavir-cobicistat	HIV/AIDS	COVID-19	3 (NCT04252274, NCT04425382)	þ
dexamethasone	RA	COVID-19	4 (NCT04325061), 3 (NCT04395105, NCT04327401)	*
sarilumab	RA	JIA	2 (NCT02991469, NCT02776735)	*
baricitinib	RA	JIA	3 (NCT03773978, NCT03773965)	*
nelfinavir	HIV/AIDS	Kaposi sarcoma, other cancers	2 (NCT03077451)	†‡
saracatinib	experimental anticancer drug	lymphangioleiomyomatosis	2 (NCT02737202)	‡
paracetamol	analgesic, antipyretic	malaria	3 (NCT03056391)	†
riluzole	ALS	mild AD	2 (NCT01703117)	†
propranolol	cardiovascular diseases	osteoporosis	1 (NCT02467400)	†
nilotinib	chronic myeloid leukemia	PD	2 (NCT03205488)	†
isradipine	hypertension	PD	3 (NCT02168842)	†
exenatide	type 2 diabetes	PD	3 (NCT04232969), 2 (NCT04305002)	*
droxidopa	neurogenic orthostatic hypotension	PD	4 (NCT03229174), 2 (NCT03446807, NCT03567447)	*
ibuprofen	JIA, osteoarthritis	tuberculosis	2 (NCT02781909)	†

Repurposed drug candidates have a higher probability of success, as well as a faster development pipeline (Fig. 2.2). The development time for *de novo* drug discovery ranges from 10–17 years (Fig. 2.2A), while drug repositioning studies can produce results in 3–12 years (Fig. 2.2B) (Ashburn and Thor, 2004). Since these drugs have already been through several stages of development and the toxicology, safety and pharmacokinetic profiles have been conducted, the risks of repositioning these drugs are lower. While overall *de novo* drug discovery has a reported less than 10% success rate, with Phase II clinical trials having the highest attrition rate (Mullard, 2016).

Lower costs and lower risks in drug repositioning also mean that rare diseases, mostly overlooked by pharma, can be considered without the need for large scale preclinical investment (Hodos et al., 2016). There are about 7000 rare diseases that together affect approximately 10% of the world's population, but only a handful have known treatments.

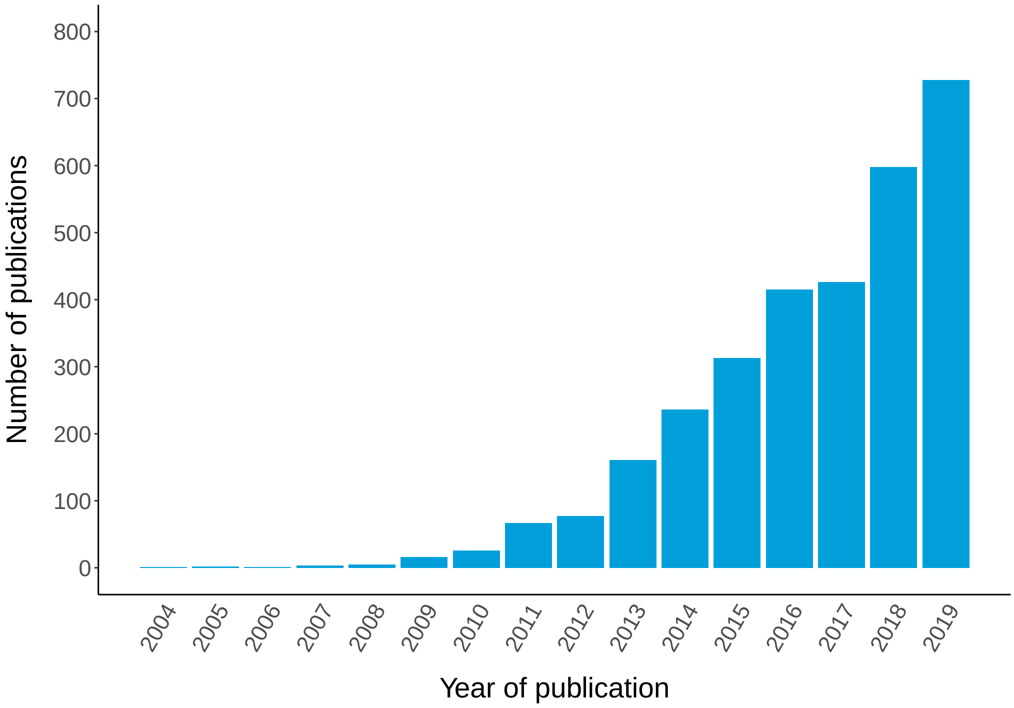


Fig. (2.1) The number of publications in drug repositioning is growing every year. The number of publications reflects publications on PubMed with keywords "drug repositioning" or "drug repositioning". Every year includes publications from 1st January to 31st December during a given year. The number of publications was extracted from PubMed in March 2020.

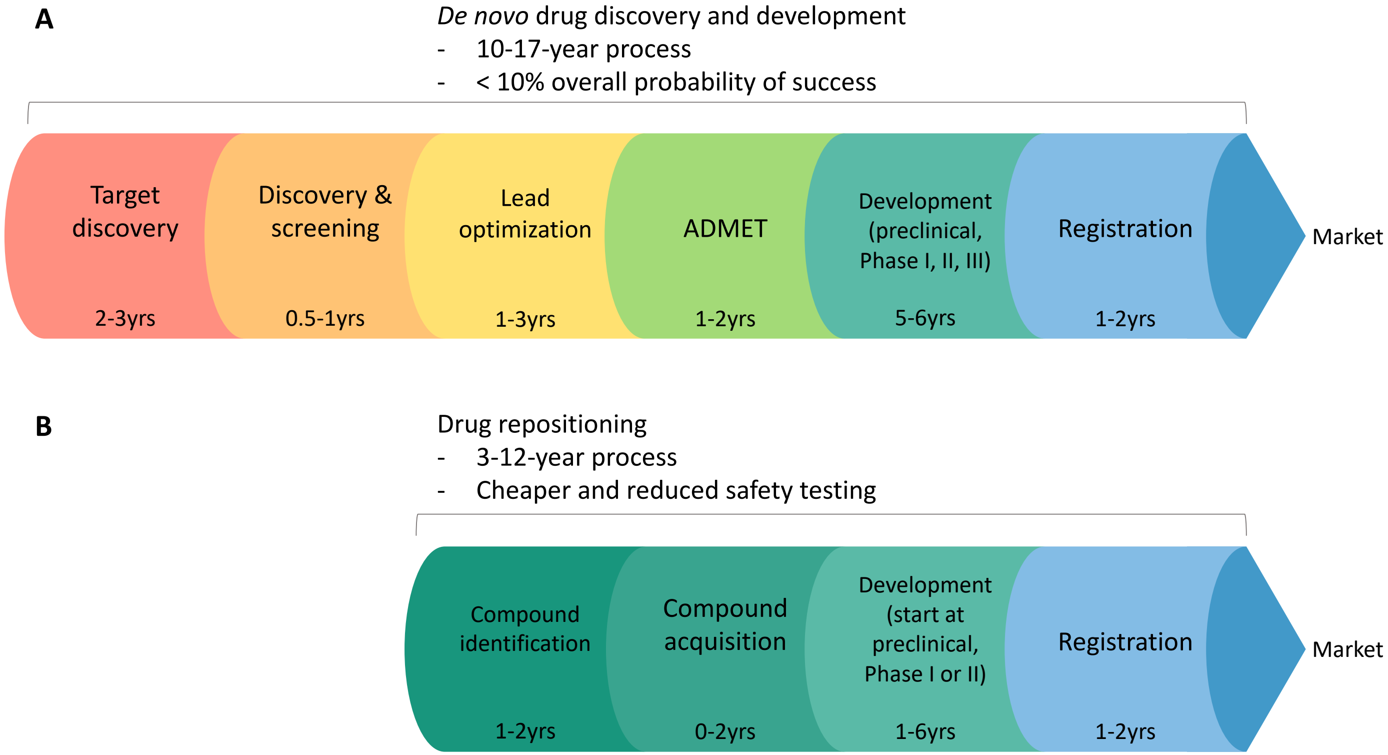


Fig. (2.2) *De novo* drug discovery and drug repositioning pipelines. (A) *De novo* discovery takes about 10–17 years and its reported success rate is under 10%, while in (B) drug repositioning getting a drug with new indication to market takes only 3–12 years. The shorter development time is mostly due to reduced safety testing, typically more toxicology, safety and pharmacokinetic information is known for drug repositioning candidates. This enables the development to start either at the preclinical stage or further down the line at Phase I or II, while in *de novo* drug discovery it always starts with preclinical testing. ADMET — absorption, distribution, metabolism, excretion, and toxicity; yrs — years. Modified from Ashburn and Thor (2004).

There has been an increasing rate of orphan drugs, drugs that treat rare diseases, on the market, with about one in five resulting from repositioning studies (Davies et al., 2017). In addition, rare disease treatments, excluding cancer therapies, have also shown a higher than average overall success rate of 25% in clinical trials (Mullard, 2016).

There may be limited interest in drug repositioning from big pharmaceutical companies as a drug with a new indication usually does not produce a big financial return due to the complex patenting rules. Focusing on the United States (US), first the drug needs to be shown to constitute patentable subject matter. The second step is an application of the new use patent which assesses the novelty of a drug itself and its use. It requires no previous patent, description in publication or access to the drug in the public domain before the patent application is filed (Conour, 2018). A common strategy to meet the drug novelty requirements is patenting a novel drug combination or a novel dosage form or route of administration (Shaughnessy, 2011; Smith, 2011). In addition to the novelty of the drug, a novel method of use strengthens the new use claim. The main obstacle in the new use claim is to overcome the obviousness criteria i.e. showing an average clinician or researcher would have not expected the drug to be useful for the new indication. The originality criteria can be met by showing unexpected results, e.g. drug combinations showing synergy, working at surprisingly low dosage or showing a drug acts via a novel mechanism of action (Conour, 2018). Another obstacle is collecting sufficient amounts of data to demonstrate the new use, in which time the obviousness criteria might be harder to overcome as more research is done and published by others (Breckenridge and Jacob, 2019). To complicate the patenting landscape further, the patenting rules vary between jurisdictions e.g. similar new use patents are available in the US, Australia, European Union (EU) and China, but not in India and Brazil (Conour, 2018). To encourage drug repositioning the EU has extended the new use patent to 8 years of data protection and 2 years of market exclusivity, while in the US there is an initial 5 year period with a 3 year extension. Nevertheless if the generic version of that drug with the same active ingredient is available, it is often prescribed/purchased over the more expensive branded new drug protected by the new use patent, e.g. in Germany it is obligatory to prescribe the generic, cheaper medicine if available, decreasing profits for the repositioned drug (Breckenridge and Jacob, 2019)

In spite of the patent regulations, in recent years, there has been a shift from *de novo* drug development supported by pharma to drug repositioning backed by governments, non-profit organisations and academia. The government-started initiatives, the National Center for Advancing Translational Sciences (US) and Medical Research Council (UK) have started large scale funding programs in partnership with industry to find new indications for significantly researched drugs developed in industry (Mullard, 2012). The US Food and Drug Administration (FDA) has also been creating online resources designed for computational drug repositioning. With more government funds and public resources available, academia has an opportunity to further push drug development forward.

Repositioning study approaches range across *in vitro*, *ex vivo* and *in silico* screenings. Initially, repositioning success stories have happened serendipitously or observationally (Table 2.1). With the availability of high-performance computing, high throughput- and high content cell-based screening methods, and large amounts of omics data, drug repositioning methods have become more systematic and knowledge-driven (Sleigh and Barton, 2012). The increasing size of large-scale publicly available genomic and phenotypic databases makes scalable computational methods development of particular interest as they offer a much cheaper alternative to wet lab hypothesis-based approaches. In addition, combining genomic, phenotypic and other clinical data can help elucidate the drug's mode of action. More importantly, it provides an opportunity to gain insight into the mechanism of a disease.

Early studies (Campillos et al., 2008; Keiser et al., 2009; Kinnings et al., 2009) mostly focused on drug similarities, such as chemical structure or side effect, while newer studies (Glicksberg et al., 2015; Xu and Wang, 2016; Zhang et al., 2016, 2013) have focused on integrating heterogeneous data from multiple databases. Most studies promise superior methods for prioritising drug candidates, but how do we know if their claims are correct? We need a gold-standard data set. Without standardised and centralised benchmarks to allow an unbiased evaluation of performance and comparison, there are no means to define the relative performance of approaches. Currently, studies rely on different benchmarking approaches using a varying selection of reference databases or skip to validation with a case study. Demonstrating the rapid evolution of the field; a recent standardised repositioning database RepurposeDB[1]; has been launched to provide a platform where all successful attempts have been logged and new attempts can be submitted in a repositioning investigations-specific format (Shameer et al., 2017). As a publicly available resource, RepurposeDB would allow the generation of a benchmarking dataset that could be systematically applied and used in drug repositioning studies.

[1] RepurposeDB (Shameer et al., 2017) is no longer available online (27th May 2020)

2.2 Drug Repositioning Data Types

Computational strategies can be roughly classified based on the type of data used. Approaches predominantly relying on drug databases can be classified as drug-based and those depending on disease data are disease-based. In this classification, the basis of discovery is an important distinction. For example, in drug-based strategies, the discovery comes from the perspective of the chemical and its properties, whereas in disease-based approaches, the discovery comes from pathology or clinical characteristics of the disease (Dudley et al., 2011a).

The main difference is that they are powered by distinct hypotheses. In a drug-based approach, the assumption is that drugs with similar properties, such as structure or bioactivity, to an already approved drug will have similar therapeutic effects and thus be as effective. A disease-based approach hypothesises that diseases with similar properties, such as symptoms or transcriptional signature, require the same therapies and can therefore be treated with the same drugs (Liu et al., 2013). Both assumptions rely on an assessment of similarity between drugs or diseases, which in turn is the source of key differences in repositioning approaches. Similarity criteria, for example, can range from the drugs' chemical structures in molecular docking approaches to gene expression under drug and disease conditions. The increasing diversity in data sources available for supporting these computational approaches contributes to the variety of different repositioning methods. The databases available further split the methods into: genome-wide, phenome-focused, and drug-orientated (Li et al., 2016) (Fig. 2.3).

2.2.1 Genome-wide approaches

Advances in genomics have significantly increased the amount and availability of genetic and transcriptomic data in a large set of different drug and disease settings in addition to controls in cell lines, tissue samples and animal models. In particular, gene expression data has been the most widely used for systematic repositioning methods, with Connectivity Map (CMap) (Lamb et al., 2006), later extended into the Library of Integrated Network-based Cellular Signatures (LINCS) (Subramanian et al., 2017), as one of the key resources. These databases provide a rich resource of gene expression profiles using a selection of mostly cancer cell lines treated with different perturbations under a selection of different conditions. Lamb et al. (2006) constructed a detailed map designed for exploration of functional connections among diseases, genes and drug perturbations. They proposed a

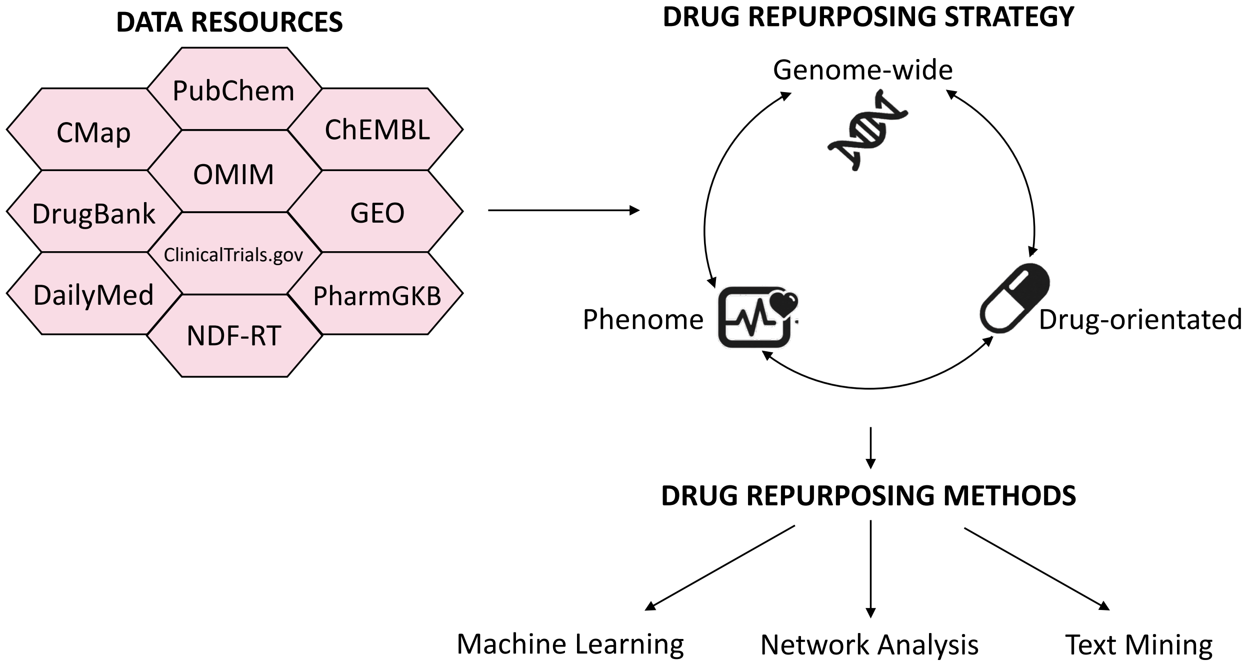


Fig. (2.3) Overview of the classification of computational drug repositioning data types and methods. The use of data resources determines the data types taken that can be developed with the use of three main repositioning methods: machine learning, network analysis, and text mining. CMap — Connectivity Map; ChEMBL — Chemical database developed by the European Molecular Biology Laboratory (EMBL); GEO — Gene Expression Omnibus; NDF-RT — National Drug File Reference Terminology; OMIM — Online Mendelian Inheritance in Man; PharmGKB — Pharmacogenomics Knowledge Base.

gene signature-based approach (Fig. 2.4), where, by systematically extracting differentially expressed genes under drug perturbation, they obtain a unique drug signature that they later can match to a user input disease signature. They prioritise drug candidates that have high negative connectivity to the input disease signature. The concept is formulated so that the aberrant disease signature is counterbalanced by the drug's opposing signature and reverts the cell to normal state (Lamb et al., 2006).

The power of this approach has been demonstrated in several successful studies (Chang et al., 2010; Chen et al., 2011; Dudley et al., 2011b; Johnstone et al., 2012; Kunkel et al., 2011), and adopted in this thesis (Chapters 5, 6 and 7). Very few systematic evaluations of each method's performance have been completed (Cheng et al., 2014). This can be partially explained by the lack of a gold-standard truth table consisting of true drug-disease relationships (addressed in Chapters 6 and 7), as well as the effort to generate the disease signatures required for querying CMap (Cheng et al., 2014). NCBI's Gene Expression Omnibus (GEO) (Barrett et al., 2005), the largest public data repository for transcriptomic data, is one such resource that is widely used for extracting disease signatures, allowing for a more systematic use of CMap. An important consideration for the use of GEO is the bias from overrepresentation of cancer-related samples.

Hu and Agarwal (2009) constructed drug↔disease networks by using the CMap drug signature system, employing part of its methodology and adding GEO disease and drug data. They constructed drug and disease genomic signatures and used two approaches for network construction. The first approach used correlation and the second was based on enrichment (Hu and Agarwal, 2009). Similarly, Jadamba and Shin (2016) constructed an enrichment based pathway↔drug network. They constructed the network by establishing pathway↔drug associations with Gene set enrichment analysis (GSEA) (Subramanian et al., 2005) on CMap drug signatures. They then generated disease pathways by obtaining disease gene expression from GEO and performed GSEA enrichment on the top dysregulated genes. Semi-supervised learning approach based on known drugs for the disease was then used to prioritise novel drug candidates. All these approaches are largely limited by the availability and quality of input data, so careful assessment of this data is required before integrating it into the drug repositioning system. Another consideration is that many diseases usually affect different tissues in the system, so a representative transcriptional profile of disease can be hard to obtain and model (Dudley et al., 2011a). This thesis introduces a novel network-based method (Chapter 5) that addresses the disease tissue specificity by prioritising drug candidates based on a disease case study specific signature.

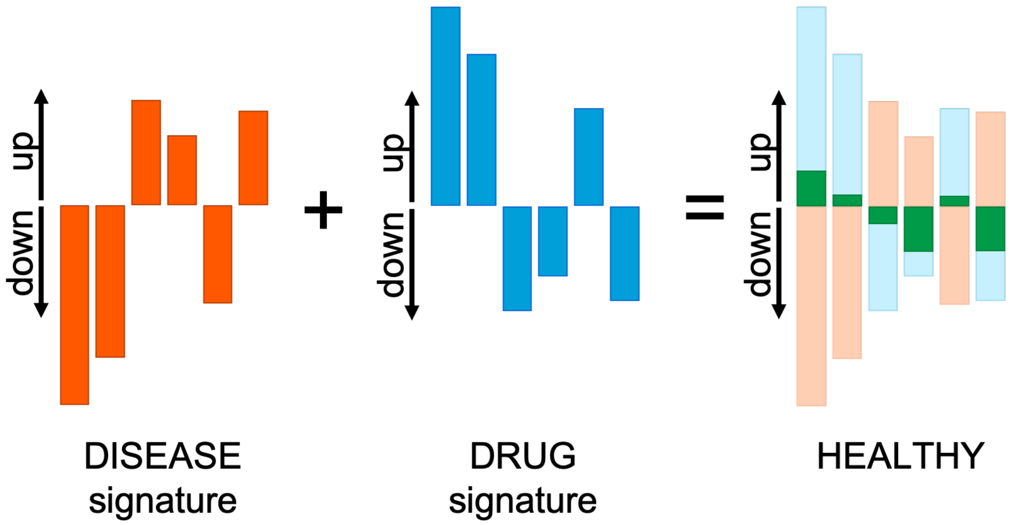


Fig. (2.4) Hypothesis under-pinning signature-based drug repositioning methods. The signature concept is formulated so that opposite drug and disease signatures can counterbalance the effects and revert the cell to normal state (Lamb et al., 2006). Signature-based methods systematically extract differentially expressed genes (up- or down-regulated) under drug perturbation, then obtain a unique drug signature that can be matched to a user input disease signature. The most highly anti-correlated signatures to the input disease signature are the prioritised drug candidates.

Guilt-by-association approaches

A drug-centric approach can also be taken with CMap data. It finds perturbations with similar signatures, and then extrapolates one drug's use to the other drug with a matching signature. A drug-centric approach is also known as the guilt-by-association approach (Chiang and Butte, 2009). Applying the drug-based strategy hypothesis, if one drug causes a particular transcriptional response and another drug causes a similar response, they are likely to work through a similar mode of action (MoA) and thus might be beneficial for similar diseases. In a case where one drug's MoA is known, and the other's is not, this approach can provide insight into the second's MoA. MoA transcriptional similarity approaches have also been successfully used by combining FDA approved indications with their off-label use, generating approximately 57,000 new drug indications that were enriched in clinical trials (Chiang and Butte, 2009).

Genetic association drug study approaches

Drugs with known protein targets that have been genetically associated with a disease have also been included in drug repositioning studies. Genome-wide association studies (GWAS) show association between genetic variants and complex diseases. Sanseau et al. (2012) have shown that 15.6% of GWAS genes are associated with a drug, compared to only 5.7% of all human genes, making GWAS genes promising drug repositioning targets. From GWAS-associated drugs more than half had the marketed indication different from that of GWAS, which identifies potential candidates for repositioning (Sanseau et al., 2012). An example of this approach was performed by Okada et al. (2014) who conducted a GWAS meta-analysis looking at rheumatoid arthritis and identified promising drug repositioning candidates (Okada et al., 2014). There are, however, some challenges that this approach carries; the biggest being the problem of directionality of the therapeutic effect. For example, it is not evident from GWAS information alone whether an inhibitor or an activator is required. In GWAS approaches, rare diseases are overlooked as high participant numbers are necessary for reliable GWAS.

2.2.2 Phenome studies

In addition to GWAS-driven repositioning studies that examine the relationship between the phenotype to a genetic variant, a phenome-wide association studies (PheWAS) approach has been used. PheWAS make use of electronic medical record systems to make connections from genotype to phenotype. A powerful advantage of PheWAS is that they can measure genetic associations with many diseases simultaneously (Hebbring, 2014). Like GWAS, PheWAS can be used for hypothesis generation, rather than full systematic repositioning studies. PheWAS has the ability to identify novel genetic associations with human diseases that can be later integrated with other drug and disease resources in repositioning approaches.

Clinical observations of side effects are another source for repositioning hypotheses. Perhaps the most famous example in repositioning comes from the unexpected side effect of sildenafil citrate, for which the primary indication was angina. As a result of the commonly reported side effect of penile erection during clinical trials, this drug was later approved for erectile dysfunction (Ghofrani et al., 2006). More recently, side effect data from drug labels and clinical trials have been used to predict new drug targets as well as provide insight on MoA of drugs with similar uncommon side effects. Side effects allow for the transitive linking of drugs to diseases as they present the drugs' physiological effect. With about 70 side effects reported per drug on average (Duke et al., 2011), studies constructing and matching drugs' side effect profiles have emerged, hypothesising that drugs with similar profiles will share similar therapeutic properties through related MoA (Ye et al., 2014). An advantage in this type of approach is that the side effects reported have been observed on human subjects rather than animal models, which avoids translational issues that commonly occur in transition from preclinical studies to clinical trials. However, it does require having well-defined side effect profiles, which are harder to obtain for newer drugs as clinical testing takes years. Even though repositioning with the use of side effect data seems promising, it does not provide a deeper understanding of the underlying mechanisms required for drug development studies. However, PheWAS and side effect data could be exploited by integrating them with other resources to improve a method's performance and increase the potential of identified repositioning candidates.

2.2.3 Drug-orientated approaches

One of the key aspects driving drug repositioning is the fact that drugs show promiscuous targeting. Drugs often interact with more than one target in the biological system; a phenomenon that is widely explored in chemical similarity drug repositioning efforts. Cheminformatics explores and predicts new targets for existing drugs as well as similar drugs for known targets by looking at structural and chemical properties. Chemicals with related structures often have related biological properties and affect biological systems in similar ways (Swamidass, 2011). Current methods can integrate different sets of resources and predict similar drugs working on the same target (Chiang and Butte, 2009; Keiser et al., 2009; Li and Lu, 2012a). These leverage the availability of high-throughput screening data, literature-mined biochemical data and databases, such as PubChem (Bolton et al., 2008) and ChEMBL (Gaulton et al., 2017), which are populated with chemical structures and other properties. Chemical-similarity approaches are usually based on extracting a set of chemical properties for a selection of drugs and then constructing networks where the more similar chemicals would cluster together. Although the predictive activity from structural similarity has been well established, some pairs of chemicals that are structurally similar can have a very different activity (Guha and van Drie, 2008). Another problem with chemical-similarity approaches is the quality and availability of chemical data. Many structures and other chemical properties contain errors or the information on them is withheld by pharma (Warren et al., 2012). Because drugs get metabolised and distributed in the biological system differently, considering only structural properties is not always predictive of a drug's physiological effects. Integrative methods combining different kinds of data address these issues by using different data types to improve performance of these approaches.

Molecular docking is another drug-based strategy relying on structural and chemical information. It is a modelling approach that is aimed at predicting physical interactions between existing drugs and finding new therapeutic targets associated with disease. Molecular docking uses 3D modelling and simulation to find a fit for a drug into a protein-binding site and then calculate the binding affinity (Meng et al., 2011). As more protein 3D structures become resolved, this has become a popular repositioning strategy. These approaches are computationally demanding so they can be used either on a target-by-target basis exploring repositioning opportunities of well-defined targets or can be more widely applied by proposing drug↔target interaction networks (Dudley et al., 2011a). An advantage of this strategy is its ability to predict new targets for drugs and, more importantly, to predict side effects by identifying off-target interactions. However, an even greater concern than with the chemical similarity approach is the correctness of structural information, as the core of these knowledge-driven predictions is based on the chemical 3D structure. In particular, stereoisomers and protonation states need to be carefully considered, which makes this type of approach very challenging (Oprea and Overington, 2015). For now, molecular docking strategies are known to have high false positive rates, but as more curated 3D structures become available, they are likely to increase in prediction accuracy (Dudley et al., 2011a).

2.3 Current Methods

Independent of data types defined by the type of resources used, different methods can be applied in the drug repositioning pipeline. The methods can be generally split into machine learning algorithms, network-based and text mining methods (Fig. 2.3).

2.3.1 Network methods

With the intention of capturing complex interdependent relationships, any relationship between drugs, diseases and targets can be modelled using networks. Network-based methods aim to organise these relationships and provide further insight into biological processes. The ability to connect molecular signatures to phenotypes and investigate the effect and mechanisms of different perturbations on biological systems within the network framework is of particular significance to drug repositioning (Wu et al., 2013). With the combination of different data resources, studies have proposed various methods for network construction: drug↔drug (Zhou et al., 2015), drug↔target (Wang et al., 2013), drug↔disease (Paik et al., 2015), drug↔target↔disease (Li and Lu, 2012b), disease↔side effect (Yang and Agarwal, 2011) and transcriptional networks (Hu and Agarwal, 2009) have been used, with some studies integrating one or more different types of network together. Similarly, in this thesis we introduce a pathway-drug coexpression network that is based on transcriptomic data, as well as curated and experimental gene sets (Chapter 5).

One of the earliest network integration studies was performed by Nacher and Schwartz (2008), who constructed a drug↔therapy bipartite network alongside its drug and therapy network projections. Drugs sharing at least one therapy were connected in the drug network whilst therapies with shared drugs were connected in the therapy network. They found that drugs involved in many treatments had a higher betweenness centrality value in the drug-therapy network and thus were hypothesized to interact with multiple targets. This study provided a global map of known drugs and therapies, serving as a knowledge base for repositioning hypothesis generation (Nacher and Schwartz, 2008). While this method used known relationships, many use networks to predict new relationships. One such method, Prediction of drug indications (PREDICT), leverages several drug↔drug and disease↔disease similarity measures, one of which, a protein↔protein interaction network-based similarity, is used to predict novel drug indications with a machine learning algorithm (Gottlieb et al., 2011). When making predictions an important consideration is the data quality used in the model. The predictions are only as good as the information they are based on. This is why most studies opt for manually curated data and make an effort to integrate multiple sources to account for data acquisition bias. The downside of any prediction method is that in addition to a need for successful benchmarking and evidence of high precision, the predictions still need to be experimentally validated.

2.3.2 Machine learning approaches

Machine learning methods can leverage the wide range of drug repositioning resources to study the underlying systems and predict novel associations between drugs and diseases. Machine learning methods utilise similarity measures to construct classification features and then a learning classification rule that separates between a true and a false node association. Machine learning methods take several different approaches to drug repositioning using various sets of data and the particular design of the method relies on the data resources used (Vanhaelen et al., 2017). Gottlieb et al. (2011) used a machine learning algorithm to predict novel drug indications from drug and disease similarity measures. They used a set of known drug–disease associations, constructed from DrugBank (Law et al., 2014) and Online Mendelian Inheritance in Man (OMIM) (Hamosh et al., 2002), as a training set, then used the similarity measures to construct classification features and predict novel drug indications with a logistic regression classifier. An advantage of this method is that it can be applied to new drugs with no previous indication information. The authors performed integration of additional similarity measures such as disease-specific gene expression signatures, which could potentially lead to personalised medicine (Gottlieb et al., 2011). Many different machine learning algorithms are being developed, each with their own strengths and weaknesses. In general, methods improve their performance and prediction power by integrating different methodological approaches. A big advantage of machine learning methods is that they can be benchmarked with cross-validation approaches, but still require a well-defined positive and negative benchmarking data set.

2.3.3 Text-mining

With the exponential growth of biomedical literature, it has become more challenging to extract knowledge, especially since most of the biomedical knowledge is recorded in free-text format (Hunter and Cohen, 2006). Text mining methods have been developed to automatically extract desirable information and assist researches in new discoveries. When applied to drug repositioning, information on drugs and diseases is being mined from biomedical literature as well as electronic health records (EHR), drug labels, clinical trials, disease and drug databases. EHR in particular are an information rich source containing longitudinal data on millions of patients, with access to lab results, standardised diagnosis codes, treatment plans and physician notes. However, for now EHR still have some legal, ethical and financial concerns, as well as errors in diagnosis codes, irregularity and differences in reporting between institutions (Yao et al., 2011).

Text-based methods aim to extract terms and their inter-relations, which can be organised into ontologies with controlled vocabularies and provide a framework for mapping associations between concepts. It is now possible to detect novel drug indications by extracting relevant knowledge and inferring relationships between drugs, targets, diseases and side effects even if they were not mentioned in the same abstract (Andronis et al., 2011). Even though the methods in text mining are developing fast, they are still mainly used in addition to either machine learning or network-based methods. For example, Sun et al. (2016) constructed a tripartite network connecting associated genes, drugs and diseases. They then integrated a text mining method to evaluate the findings and score the confidence level of predicted associations (Sun et al., 2016). There is great potential in integrating text mining methods as they can efficiently extract the knowledge about reagents in drug repositioning approaches. Furthermore, novel associations and predictions can gain power if integrated with a wide range of *a priori* knowledge. Currently many repositioning data resources are being constructed with text mining and curation methods. The hidden knowledge from text has the potential to serve as a source for a standardised benchmarking dataset needed to help evaluate current drug repositioning studies.

2.4 Benchmarking

Considering the heterogeneity of approaches to drug repositioning, it is not surprising that there are an increased number of publications offering improved and superior methodology (Fig. 2.1). Claims are usually based on comparison between study results and existing biomedical knowledge. However, no standardised way of comparing and evaluating the power of different methodologies has been adopted. Studies adopt two main approaches to assess individual success: benchmarking and validation. There is an important distinction between benchmarking and validation, which is often overlooked but will be considered in this thesis. Benchmarking is based on method evaluation with a structured gold-standard data set. Validation is the assessment of the validity of the method's results. For example, benchmarking would consider how many of the already known drug↔disease relationships the method can find, thus, assessing the general performance and reliability. Validation, on the other hand, would show that the method has the power to identify novel indications that can be translated into practice. Benchmarking deals with computationally assessing the method's ability to give reliable predictions and is currently based on looking at the recall rate of approved drug indications obtained with text-mining or curation methods. Validation of biological relevance is still examined by wet lab experiments or during preclinical animal studies (Vanhaelen et al., 2017).

2.4.1 Benchmarking methods

Many repositioning studies claim that they have analytically benchmarked their methods in some way. However, no agreed best practice of benchmarking computational predictions exists, which makes it hard to assess how useful a particular approach is in producing reliable repositioning hypotheses. There are three main techniques to assess the accuracy of results:

(i) overlap between predictions and a set of known drug indications,

(ii) sensitivity- and specificity-based methods,

(iii) cross-validation in machine learning methods (Brown and Patel, 2016).

Overlap methods analytically measure the extent to which an approach correctly identifies known indications. Overlap methods can include currently approved and/or investigational drug uses. They can assess the general ability of the method to make valid claims by measuring the sensitivity, also called recall, of the approach, which is the rate of correctly identified true positives. The advantage of this approach is that it only needs a list of true indications to compare with their predictions (Brown and Patel, 2016), while sensitivity and specificity methods also require a list of negative interactions, which are harder to compile. A disadvantage of the overlap method is that it cannot be used for machine learning approaches. Meanwhile, cross-validation is a well-established method of benchmarking machine learning algorithms. With cross-validation, the algorithm is already optimised on part of the data called the training set and tested on another previously unseen set. Testing the algorithm can highlight overfitting of the algorithm to the training data set and more importantly, its results are more representative of its future performance (Lever et al., 2016). Cross-validation allows for the calculation of both sensitivity and specificity.

Commonly, a combination of both sensitivity and specificity will be used in benchmarking, which describe the recall or true positive rate (TPR) and the true negative rate

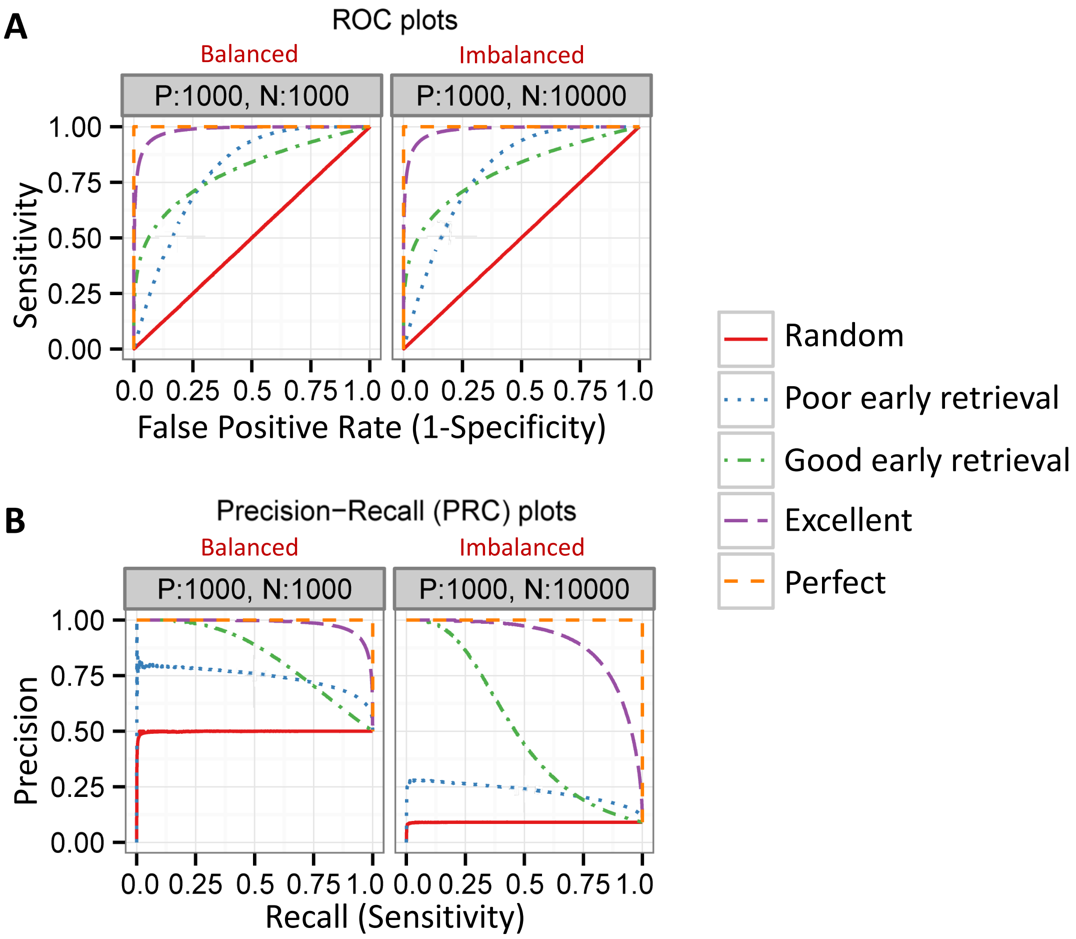


Fig. (2.5) Comparison between receiver operating characteristic curve (ROC, A) and precision-recall curve (PRC, B) in class balanced and imbalanced algorithms. PRC changes with imbalanced classes, while there is no change on ROC plot. Each panel contains two plots with balanced (left) and imbalanced (right) for (A) ROC and (B) PRC. Five curves represent different performance levels: random (red), poor early retrieval (blue), good early retrieval (green), excellent (purple), and perfect (orange). N — negative; P — positive. Modified from Saito and Rehmsmeier (2015).

(TNR), respectively. The benchmarking measures are calculated using a confusion matrix (Table 2.2). The false positive rate (FPR) and TPR can be plotted into a receiver operating characteristics curve (ROC, Fig. 2.5A) and a commonly reported value for benchmarking is the area under the ROC curve (AUROC) (Cheng et al., 2014; Guney, 2016). The ROC curve is created by plotting the TPR (sensitivity) against the FPR (1 - specificity, Fig. 2.5A). Results are summarised by calculating the AUROC of the predictions.

Table (2.2) Confusion matrix describing performance of a classification model with true known values. FN — False Negative; FP — False Positive; TN — True Negative; TP — True Positive.

		Predicted	
		True	False
Actual	True	TP	FN
	False	FP	TN

This method relies on two benchmarking datasets: a gold standard for true positives (TPs) and one for false positives (FPs) or true negatives (TNs). However, many only define a TP set of known drug indications and assume the rest of their predictions as FPs (Gottlieb et al., 2011; Takarabe et al., 2012; Yang et al., 2013a). The assumption that all predictions apart from those in a TP set are false, is counterintuitive, as the purpose of the algorithm is to establish new drug indications, whereas this benchmarking approach assumes all novel predictions as FPs. Furthermore, the definition of different gold-standard data sets impacts the AUROC estimates and makes the reported sensitivity and specificity non-comparable between studies if different resources are used for the benchmarking data set. We explore the strengths and weaknesses of using different types of "true positive" drug sets in three case studies in Chapters 6 and 7. These benchmarking shortcomings could be avoided with the universal use of a well-defined and standardised gold-standard set that includes both successful drug indications as well as negative drug–indication pairs from failed clinical trials. Finally, this method creates a significant imbalance in the number of true and FPs, which has been shown to reduce the accuracy of AUROC and other sensitivity and specificity estimates (Davis and Goadrich, 2006; Fawcett, 2006) (Fig. 2.5). Precision-recall curve (PRC, Fig. 2.5B) and the area under PRC (AUPRC) have been suggested as alternative measures, because they are sensitive to the class imbalance (Fig. 2.5B). In contrast to the FPR, used in AUROC, that measures the fraction of negative examples that are mislabelled as positive, precision (positive predictive value) measures the fraction of examples classified as positive that are truly positive (Davis and Goadrich, 2006).

AUROC and AUPRC offer the most rigorous analytical assessment of repositioning methods. They both come with drawbacks that have currently not been overcome. The lack of a well-defined positive and negative gold-standard set compromises the reliability and reproducibility of both benchmarking methods. An additional factor compromising both benchmarking methods is the data used in the repositioning method. If the data is mainly cancer-based it is likely to perform better in cancer predictions than, for example, neurological conditions. Thus, if a generalised benchmarking set including all diseases is applied it could penalise a good performance in cancer, thus a disease area specific benchmarking might be more appropriate when repositioning methods are tailored to one disease area. For the moment, overlap benchmarking methods offer a more accurate assessment as more time and effort has been invested into formulating a data set of approved drug indications including both first time and second use drugs. However, even for overlap methods, a gold-standard data set of approved indications should be used.

2.4.2 Benchmarking data sets

The field has recognised the lack and importance of a structured benchmarking data set that includes both true and failed indications. Several recent studies have attempted to tackle this problem. Some identified the lack of structured data sets (Brown and Patel, 2016; Li et al., 2016), some constructed their own (Cheng et al., 2014; Li and Lu, 2012a; Yang et al., 2013a; Zhang et al., 2013), and others proposed a gold-standard data set (Brown and Patel, 2017; Khare et al., 2013; Kissa and Tsatsaronis, 2015; Liu et al., 2013; Shameer et al., 2017). However, the sets from different studies are usually compiled for a variety of drug uses. For example, Cheng et al. (2014) used FDA-approved drug indications, in Kissa and Tsatsaronis (2015), and Liu et al. (2013) they included successfully repurposed drugs, while Chiang and Butte (2009) include off-label use. Brown and Patel (2017) took a step further and included drugs from failed clinical trials. Even though methods included the same type of drugs, for example FDA-approved drugs, they usually used a different combination of data resources to construct their gold-standard data set. Most are compiled from publicly available databases, such as DrugBank (Law et al., 2014), PharmGKB (Gong et al., 2002), National Drug File - Reference Terminology (NDF-RT) (Bodenreider, 2004) and DailyMed (NIH U.S. National Library of Medicine, 2016). For example, Khare et al. (2013) used DailyMed for FDA-approved drug labels, while Cheng et al. (2014) used Pharmaprojects (Pharma Intelligence, 2017) and FDA Adverse Event Reporting System

(FAERS). Each data resource contains different information, therefore even the methods that have been benchmarked cannot be compared, further highlighting the necessity of a standardised and widely adopted benchmark data set.

A potential explanation to why there is no widely accepted gold-standard data set might be because systematic computational drug repositioning is a relatively new and fast developing field. In the past few years, the lack of a benchmark has been noted, with no agreement on what type of drug indications it should consist of, nor what resources those should come from. One of the problems is that indications are still mostly described in free-text which makes them harder to extract. Text mining efforts are close to overcoming this, but manual curation is still needed in most cases to ensure only true drug↔indication pairs are included (Khare et al., 2013). However, there is still a lack of a wider agreement on what vocabulary and coding system to use. Benchmarking is also confounded by drug terminology. There is considerable heterogeneity in the vocabulary used in different databases. Drugs often have many synonyms. Different standardised databases exist employing controlled vocabularies for drugs and disease, but are lacking interoperability to allow integration across a wider range of resources. For example, the Unified Medical Language System (UMLS) (Bodenreider, 2004) consists of a set of programs that work towards unifying the vocabulary and codes from other systems. This problem is further addressed in Chapter 4 where a drug synonym database was constructed to resolve heterogeneity in drug nomenclature.

Ideally, a list of approved and failed indications across a wide range of diseases is needed. It should not be restricted only to FDA-approved drugs, but also include drugs approved by other corresponding authorities with well-defined approval protocols. Care should be taken when identifying true indications as some might be false positives (i.e. drugs that only provide symptomatic relief rather than directly affecting the disease) (Cheng et al., 2014). In addition, when defining negative indications, only drugs failing clinical trials because of toxicity or lack of efficacy should be included. These measures might highly reduce the numbers of negative indications, as many clinical trial databases such as Aggregate analysis of clinicaltrials.gov (AACT) database (Tasneem et al., 2012) often lack detailed explanations on why a trial has been terminated. An accurate, complete and up-to-date resource should be used for extraction of this information. Finally, drugs and indications should be in one of the widely used controlled vocabularies such as UMLS (Bodenreider, 2004) to allow interoperability. Further steps would be to include not only indication information but also the dosage and form (Khare et al., 2013) and grouping indications into disease areas would be beneficial when assessing methods tailored to a specific disease.

A publicly available database, RepoDB, provides a systematic extraction of approved and failed drug–indication pairs that has a high potential to become a widely used benchmarking data set. It is closest to fulfilling the requirements of a good benchmark, because it includes positive and negative indications. It consists of 6677 approved and 4123 failed drug indications across 1571 unique drugs and 2051 indications. For approved indications: the drug list has been extracted from DrugBank (Law et al., 2014); all drug synonyms, their DrugBank ID and UMLS indication were extracted from DrugCentral (Ursu et al., 2017); while failed clinical trials information was extracted from AACT that is based on ClinicalTrials.gov. Data from AACT included details on failed trials, UMLS indications, and MeSH interventions that were mapped to the DrugCentral synonyms to compile the final database (Brown and Patel, 2016). However, some care should be taken using failed drug–indication pairs, as many trials ended due to the lack of funding or, more commonly, with no information on the reason for termination.

Establishing a widely accepted benchmarking set has the potential to improve consistency and reproducibility in the field, as well as to allow objective comparison between methods and increase the accuracy of reported benchmarking values. Consequently, positive validation results will become more likely and, from that, improved translation into clinical treatments.

2.4.3 Validation

Studies usually perform validation of a few selected candidates to show the applied potential of their method and to satisfy the ultimate goal, which is to find a new drug indication that can be translated into clinical applications. If the novel predictions are to be pursued, experimental validations consisting of *in vitro* and *in vivo* preclinical drug evaluations become necessary. For example, once potential candidates have been identified, their biological significance can be explored in the literature. Drug properties such as side effects, cost, availability, intake form and distribution, should also be considered when narrowing down the search space for candidates. Brown and Patel (2016) observed that some studies completed this in an addition to previous benchmarking whereas others omitted assessing the general performance of their method, skipping straight to validation of carefully selected predictions. An example of a benchmark omission by Sirota et al. (2011) involved identifying one novel drug indication out of over 16,000 drug-indication pairs tested, from which there were 2664 statistically significant associations with more than half of therapeutic interest. They tested a drug with a moderate score for the association to lung adenocarcinoma, on the basis that it is an inexpensive off-patent drug with a favourable side effect profile. The authors then demonstrated its success with a series of *in vitro* and *in vivo* experiments as well as evidence from the literature (Sirota et al., 2011). Their promising result identified in one case study shows the potential of their method, however, it cannot be extrapolated to all predictions made by the method. In repositioning studies focused only on one disease or one drug, this approach might be more acceptable, as their method is closely suited to the specific conditions of that disease and drug, and their search space will be smaller and better defined. However, it is best practice to generally evaluate the method with a well-defined benchmark and proceed with validation of carefully selected candidates if the method shows high sensitivity and specificity. This way the method's power is systematically assessed, and its predictions will be more likely to succeed further down the drug development pipeline.

2.5 Summary of the Current State

Drug repositioning carries the potential of large economic and public health benefits for the public, academia and government. Repositioning can accelerate drug discovery and systematically investigate many previously overlooked diseases, such as rare diseases as well as disease areas with poor *de novo* drug development success, for example, Alzheimer's disease. Drug repositioning is highly dependent on the availability and quality of data resources, such as CMap, GEO and ClinicalTrials.gov. As more large-scale assays are generated, a more detailed molecular understanding of drug and disease mechanisms is becoming possible. In particular, this can lead to improving disease classification, such as cancer subtypes, which can yield better results in drug repositioning studies as well as improve personalised medicine, where therapies are targeted at specific patient sub-populations.

More repositioning methods can be developed as more data from different techniques, formats and domains become available. The richness of resources allows the use of various different combinations and consequently the design of many approaches that can be used to identify candidates for repositioning. Several different approaches have been considered in this chapter, each with their own advantages and disadvantages. Integration of these methods often results in higher sensitivity and specificity, which indicate improved success in identifying novel indications. With improved benchmarking practices, the repositioning methods could be compared more vigorously in order to prioritise the most effective one. In addition, identification of best methods can accelerate prioritisation of novel drug indications as well as further development of repositioning approaches.

At the moment, most publicly available repositioning studies have not yet been translated into clinical trials. Even though repositioning provides the opportunity to shorten drug development time, the requirement for preclinical testing and clinical trials often remains. A number of repositioning candidates have, however, already undergone phase I, which assesses safety. Phase II, which addresses efficacy, requires more resources for further drug validation. A combination of different funding sources, such as pharma, governments and charities, could help overcome the costs of accelerated drug repositioning trials. As more recognition for repositioning studies occurs, more funding opportunities are becoming available.

In summary, recent methodological advances in the rapidly emerging field of drug repositioning provide a new perspective to revolutionise drug development, offering a potentially faster and cheaper alternative to current *de novo* drug discovery practices. We identify five key areas for improvement:

(i) systematic database development for increasing amounts of data,

(ii) increasing public access to studies and resources,

(iii) developing a standardised protocol for recording clinical trials with well-annotated outcomes,

(iv) development of a widely applied gold-standard benchmark,

(v) increasing funding for translation of candidates from computational studies into clinical trials.

With such advances, we predict drug repositioning will become an invaluable resource to lead the way in drug development.

2.6 Positioning of the Current Study

In this chapter, we considered strategies for computational repositioning to provide context for our drug repositioning system. We assessed different types of computational methods and the type of databases used in each (Fig. 2.3). Highlighting that to our knowledge there are no pathway$\leftrightarrow$gene set correlation approaches. Furthermore, we highlighted the shortcomings of benchmarking and validation attempts adopted by studies in the field of computational drug repositioning. We have identified a set of requirements for a standardised benchmark that we explored further in the benchmarking methodology developed in this project.

In this thesis, we have developed a disease-based drug repositioning method based on the signature reversal hypothesis (Fig. 2.4). In Chapter 5, we present a novel pathway-based correlation network approach where we estimated correlation coefficients between pathways and LINCS drug signatures on a background of a large curated collection of GEO gene expression data. We have identified several related methods in this chapter, such as Hu and Agarwal (2009), and Jadamba and Shin (2016), however, we took the unique approach of constructing a bipartite-network based on the pathway- and drug-gene set correlation capturing functional relationships.

To overcome some of the benchmarking challenges discussed, we have built a drug synonym resource that can overcome the use of heterogeneous drug nomenclature (Chapter 4). We explored the use of different true positive lists in the evaluation of drug repositioning performance (Chapters 6 and 7), providing insight into most appropriate benchmark designs.

Chapter 3

Materials and Methods

Others' contributions to this chapter

Network construction method. Yered Pita-Juárez's PhD project (Department of Biostatistics, Harvard T.H. Chan School of Public Health) served as the basis for the underlying method used in the pathway-drug coexpression network construction (Pita-Juárez et al., 2018). Wenbin Wei (Sheffield Institute of Translational Neuroscience (SITraN), University of Sheffield) and Sokratis Kariotis (SITraN, University of Sheffield) were both involved in the code review to identify points for improvement. Sokratis Kariotis provided coding support in the computational improvement of the Pita-Juárez et al. (2018) method.

Juvenile idiopathic arthritis study curation. Professor Lester Kobzik (Department of Environmental Health, Harvard T.H. Chan School of Public Health) assisted in the curation of the publicly available juvenile idiopathic arthritis studies.

Alzheimer's disease data and preprocessing. Assistant Professor Doo Yeon Kim (Department of Neurology, Massachusetts General Hospital, Harvard Medical School) and Professor Rudolph E. Tanzi (Department of Neurology, Massachusetts General Hospital, Harvard Medical School) shared preprocessed 3D cell model RNA-Seq data described in Kwak et al. (2020), and the positive drug hits from the 3D drug screen. The RNA-Seq data was preprocessed by the Harvard Bioinformatics Core. Sarah Morgan (SITraN, University of Sheffield; Department of Pathology, Beth Israel Deaconess Medical Center, Harvard Medical School) advised on the Mayo Alzheimer's disease dataset sample selection.

Parkinson's disease data and preprocessing. Professor Oliver Bandmann (SITraN, University of Sheffield) shared the human sporadic Parkinson's disease RNA-Seq data

(Carling et al., 2020) and the zebrafish GCH1 mutant data (Larbalestier et al., 2020). Professor Oliver Bandmann has also curated a list of neuroprotective drugs. The Carling et al. (2020) RNA-Seq data was preprocessed by Claire Green (SITraN, University of Sheffield) and the Larbalestier et al. (2020) zebrafish data by Wenbin Wei (SITraN, University of Sheffield).

3.1 KATdb, the Drug Synonym Database

This section describes the resources and methods used in the construction of KATdb, the drug synonym database. KATdb is a resource containing drug and chemical identifiers, names and synonyms from several databases. The databases described in Table 3.1. were used as the source of synonym information for KATdb.

3.1.1 Synonym extraction

From each database listed in Table 3.1, we extracted the database-specific unique identifier, synonyms, external identifiers, and systematic identifiers (see Supplementary Table A.1). We formatted the names, so they retain information of the type of name they include, termed: the authority, followed by the identifier provided in the database.

From each of the resources (Table 3.1) the dataset was reshaped into a database of edges. Each edge $(V_a, V_b) \in E_{ab}$ was defined by one unique relationship listed in a source database between a vertex, V_a, represented by the unique identifier authority:value pair and a vertex, V_b, for synonym or external identifier:value pair for a given dataset. An edge between two vertices was:

$$V_a \leftrightarrow V_b \tag{3.1}$$

then each relationship from each database is reshaped into:

$$\text{database authority:unique identifier} \leftrightarrow \text{synonym authority:value} \tag{3.2}$$

For example, one relationship from ChEMBL database would be represented as:

$$\text{ChEMBL:CHEMBL25} \leftrightarrow \text{Wikipedia:Aspirin} \tag{3.3}$$

The synonyms extracted from each database are listed in Supplementary Table A.1.

3.1.2 Drug synonym source databases

We utilised the databases in Table 3.1 for their drug synonym information. The main applications of the databases are briefly described in Table 3.1. The details of data obtained from each database are described below.

Table (3.1) Description of individual source databases used in KATdb. ATC — Anatomical Therapeutic Chemical; ChEMBL — Chemicals database by European Molecular Biology Laboratory; CMap — Connectivity Map; CTD — Comparative Toxicogenomics Database; DB — database; DNI — Drugs of New Indications (Liu et al., 2013); EMA — European Medicines Agency; EU — the European Union; KEGG — Kyoto Encyclopedia of Genes and Genomes; LINCS — Library of Integrated Network-based Cellular Signatures; PharmGKB — Pharmacogenomics Knowledge Base; TTD — Therapeutic Target Database.

Database	Website	Description
ATC	https://www.whocc.no/atc/	A hierarchical therapeutic chemical classification system.
BindingDB	https://www.bindingdb.org/	Database of measured binding affinities, focusing on interactions of drug-targets to drugs.
ChEMBL	https://www.ebi.ac.uk/chembl/	Manually curated database of bioactive molecules with drug-like properties.
CMap	https://portals.broadinstitute.org/cmap/	Cellular signatures from genetic and pharmacologic perturbagens.
CMaptoATC	http://www.gepedia.org/instances.html	Dataset of CMap instances annotated with names and ATC codes.
CTD	http://ctdbase.org/	Curated database describing relationships between drugs, genes, diseases, pathways and other annotations.
DNI	http://dx.doi.org/10.1016/j.drudis.2012.08.005	Database of drugs with novel, repurposed indications.
DrugBank	https://www.drugbank.ca/	Resource that combines detailed drug data with drug target information.
DrugCentral	http://drugcentral.org/download	Online drug information resource including now drug approval data.
EMA	https://www.ema.europa.eu/en	The EU agency responsible for the scientific evaluation, supervision and safety monitoring of medicines in the EU.
KEGG	https://www.genome.jp/kegg/	Collection of databases for understanding high-level functions of the biological system.

Table 3.1 continued

Database	**Website**	**Description**
LINCS	http://lincsportal.ccs.miami.edu/	Reference library of cell-based perturbation-response signatures.
PharmGKB	https://www.pharmgkb.org/	Curated database of clinical information for drugs and diseases, including gene-drug and genotype-phenotype relationships.
RepoDB	http://apps.chiragjpgroup.org/repoDB/	Database of approved and failed drug indications.
RepurposeDB	http://repurposedb.dudleylab.org	A collection of repurposed drugs, drug targets and diseases, which was assembled, indexed and annotated from public data.
TTD	http://bidd.nus.edu.sg/group/cjttd/	Database of known and explored therapeutic protein and nucleic acid targets, the targeted disease, pathway information and the corresponding drugs.
Wikipedia	https://en.wikipedia.org/	Free online encyclopaedia, created and edited by volunteers, hosted by the Wikimedia Foundation.

Anatomical Therapeutic Chemical (ATC) Classification System (WHOCC, 2018). The 2018AB version released on 5th November 2018 and uploaded on 29th April 2019 was downloaded from *https://bioportal.bioontology.org/ontologies/ATC* on 6th August 2019. The ATC code was used as the unique identifier and it was extracted with the World Health Organisation (WHO) name, RxNorm Concept Unique Identifier (CUI) and other synonyms.

BindingDB (Gilson et al., 2016). The lists and identifier mappings from BindingDB were retrieved from *https://www.bindingdb.org/bind/chemsearch/marvin/SDFdownload.jsp?all_download=yes* on 6th August 2019. The BindingDB ID was used as the unique identifier, while we extracted PubChem Concept Unique Identifier (CID), ChEMBL and Chemical Entities of Biological Interest (ChEBI) database identifiers using Python 3.6 (van Rossum and Drake Jr, 1995).

ChEMBL (Gaulton et al., 2017). The ChEMBL version 25 SQLite database was downloaded from *http://ftp.ebi.ac.uk/pub/databases/chembl/ChEMBLdb/releases/chembl_25/*, DOI:*10.6019/CHEMBL.database.25* on 26th July 2019. ChEMBL ID, Compound name, ATC classification and WHO name were extracted from "compound records", "molecule atc classification", "atc classification" and "molecule dictionary" tables using

SQLite 3. The output was processed with Python 3.6 (van Rossum and Drake Jr, 1995) to arrange the synonym relationships as a list of edges for further processing.

Connectivity Map (CMap) (Lamb et al., 2006). The instance inventory from Affymetrix-based CMap "build 02" dataset was downloaded on 2nd August 2019 from *https://portals.broadinstitute.org/cmap/* accessible upon creating a free account. CMap Name, CMap Instance ID and Catalog Name were extracted and processed in Python 3.6 (van Rossum and Drake Jr, 1995).

CMaptoATC (Gepedia, 2010). The CMap to ATC relationships were retrieved from *http://www.gepedia.org/instances.html* on 5th August 2019. CMap Name, CMap Instance ID and ATC classification were extracted using Python 3.6 (van Rossum and Drake Jr, 1995).

Comparative Toxicogenomics Database (CTD) (Davis et al., 2019). The curated Chemical vocabulary (26th June 2019 release) was retrieved from CTD, MDI Biological Laboratory, Salisbury Cove, Maine, and NC State University, Raleigh, North Carolina, accessed at *http://ctdbase.org/downloads/* on 5th August 2019. CTD Chemical ID, CTD Name, Chemical Abstract Service (CAS) number, synonyms, DrugBank ID, PubChem CID and Medical Subject Headings (MESH) terms were extracted using Python 3.6 (van Rossum and Drake Jr, 1995). CTD Chemical ID was used as the unique identifier.

Drugs of New Indications (Liu et al., 2013), also referred to as Liu2013. Drugs of New Indications (DNI) database described in Liu et al. (2013) was downloaded from *Supplementary Table S2* at *http://dx.doi.org/10.1016/j.drudis.2012.08.005* on 8th November 2017. The drug name and the CAS number were extracted from the database using Python 3.6 (van Rossum and Drake Jr, 1995).

DrugBank (Law et al., 2014). The DrugBank version 5.1.4 (released 2nd July 2019) was retrieved from *https://www.drugbank.ca/releases* on 25th July 2019 accessible upon creating a free account. The database was processed in Python 3.6 (van Rossum and Drake Jr, 1995) using BeautifulSoup. DrugBank ID, DrugBank Name, CAS number, ATC classification, ChEMBL, Drugs Product Database (DPD), Kyoto Encyclopedia of Genes and Genomes (KEGG), Pharmacogenomics Knowledge Base (PharmGKB), ChEBI, ChemSpider, PubChem CID, Wikipedia and other external identifiers and associations were extracted into a list of edges.

DrugCentral (Ursu et al., 2017). The SMILES and InChI file was retrieved from *http://drugcentral.org/download* on 5th August 2019. DrugCentral ID, Simplified Molecular-

Input Line-Entry System (SMILES), the IUPAC International Chemical Identifier (InChI), InChI Key, International Non-proprietary Name (INN) and CAS number were extracted using Python 3.6 (van Rossum and Drake Jr, 1995).

European Medicines Agency (EMA) (European Medicines Agency, 2019). The European public assessment reports (EPAR) for all human and veterinary medicines were retrieved from *https://www.ema.europa.eu/en/medicines/download-medicine-data* on 6th August 2019. The EPARs are full scientific assessment reports of medicines authorised at EU level. They include information on medicines that have been refused a marketing authorisation or that have been suspended or withdrawn after being approved. EMA Id, Active substance, EMA Name and synonyms were extracted using Python 3.6 (van Rossum and Drake Jr, 1995).

Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa and Goto, 2000). All drug entries from KEGG were retrieved using their GenomeNet FTP download of KEGG MEDICUS from *https://www.kegg.jp/kegg/download/* on 5th August 2019. KEGG Drug ID as the unique identifier and CAS number, ChEBI, ChEMBL, DrugBank and other external identifiers or synonyms were extracted using Python 3.6 (van Rossum and Drake Jr, 1995).

Library of Integrated Network-based Cellular Signatures (LINCS) (Koleti et al., 2018). The LINCS Standardized Unique Small Molecule (LSM) release number 28 (Release Date: 20th February 2018) was retrieved on 10th July 2018 from *http://lincsportal.ccs.miami.edu/dcic-portal/#/list/Small_Molecules*. LINCS ID, LINCS Center ID, LINCS Name, Alternative Name, PubChem CID, SMILES, InChI, InChI Key, ChEMBL, ChEBI, DrugCentral and other synonyms were extracted and processed in Python 3.6 (van Rossum and Drake Jr, 1995). In addition, the LINCS Data Portal Small Molecule Catalog was retrieved from *http://lincsportal.ccs.miami.edu/SmallMolecules/catalog* on 6th August 2019. LINCS ID, LINCS Name, Alternative Name, PubChem CID, SMILES, InChI, and ChEBI ID were extracted and processed in Python 3.6 (van Rossum and Drake Jr, 1995).

Pharmacogenomics Knowledge Base (PharmGKB) (Whirl-Carrillo et al., 2012). The summaries of chemical information (Primary data — Drugs/Chemicals) annotated by PharmGKB were retrieved from *https://www.pharmgkb.org/downloads* on 5th August 2019. PharmGKB ID was used as the unique identifier with synonym relationships to PharmGKB name, Generic name, Trade name, SMILES, InChI and other external identifiers.

RepoDB (Brown and Patel, 2017). The RepoDB dataset version 1.2 was retrieved from *http://apps.chiragjpgroup.org/repoDB/* on 3rd November 2017. RepoDB was built using the 25th October 2016 build of DrugCentral (Ursu et al., 2017), the 27th March 2016 build of the Aggregate analysis of clinicaltrials.gov database (AACT) database by Clinical Trials Transformation Initiative *http://www.ctti-clinicaltrials.org*, and the 2016AB release of the Unified Medical Language System (UMLS) (Bodenreider, 2004). Drug name and DrugBank ID have been extracted using Python 3.6 (van Rossum and Drake Jr, 1995).

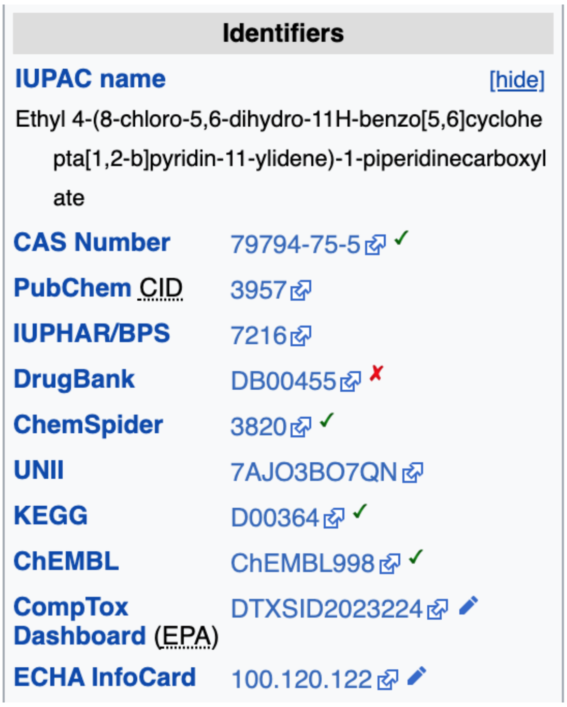
RepurposeDB (Shameer et al., 2017). The Chemo-Genomic Enrichment Analysis (CGEA) subset of Version 1.0 was retrieved from *http://repurposedb.dudleylab.org/data* on 12th October 2017. RepurposeDB identifier together with CAS number, ChEBI, ChEMBL, DrugBank, InChI, KEGG, MESH, PubChem CID, SMILES and other synonyms were extracted using Python 3.6 (van Rossum and Drake Jr, 1995).

Therapeutic Target Database (TTD) (Li et al., 2018). Synonyms of drugs and small molecules and cross-matching ID between TTD drugs and public databases were retrieved from the September 2017 release available at *https://db.idrblab.org/ttd/full-data-download* on 1st November 2017. TTD identifier, CAS number, ChEBI, PubChem CID and other synonyms were extracted using Python 3.6 (van Rossum and Drake Jr, 1995).

Wikipedia (Wikipedia contributors, 2018). The drug info boxes as seen on Fig. 3.1 were scraped from all Wikipedia pages in: "Infobox drug tracking categories" and "Chembox tracking categories" on 2nd August 2019 using `pywikibot` in Python 3.6 (van Rossum and Drake Jr, 1995). Three different info boxes were considered from the scraped templates: "Drugbox", "Infobox drug" and "Chembox". The Wikipedia name for the page entry was extracted as the unique identifier. The synonyms were extracted from the identifiers section in drug info boxes (Fig. 3.1). Identifiers that have not been successfully verified (red cross ✘) have been excluded. The following identifiers were extracted: CAS Number, ChEBI, ChEMBL, ChemSpider, DrugBank, InChI, InChI Key, KEGG, Unique Ingredient Identifier (UNII), European Inventory of Existing Commercial Chemical Substances (EINECS), IUPAC, KEGG, MESH, PubChem CID, SMILES, synonym (see Supplementary Table A.1 for a list of types of synonyms extracted)

3.1.3 Data cleaning

KATdb edge entities for the described databases were extracted, cleaned, and processed in R (R Core Team, 2019) using `tidyr` (Wickham and Henry, 2019), `dplyr` (Wickham et al., 2019) and `stringr` (Wickham, 2019). Extracted identifiers relating to protein structure, genes or targets and associated disease indications were removed. The URLs



Identifiers

IUPAC name [hide]

Ethyl 4-(8-chloro-5,6-dihydro-11H-benzo[5,6]cyclohepta[1,2-b]pyridin-11-ylidene)-1-piperidinecarboxylate

CAS Number	79794-75-5 ✓
PubChem CID	3957
IUPHAR/BPS	7216
DrugBank	DB00455 ✗
ChemSpider	3820 ✓
UNII	7AJO3BO7QN
KEGG	D00364 ✓
ChEMBL	ChEMBL998 ✓
CompTox Dashboard (EPA)	DTXSID2023224 ✎
ECHA InfoCard	100.120.122 ✎

Fig. (3.1) Loratadine example of Wikipedia Identifiers section of the info box. The info box includes key compound information including identifiers from different authorities. The green tick ✓ (yes) or red cross ✘ (no) indicate if the identifier has been automatically authenticated on the external authority's site. Screen captured 25th July 2019 at *https://en.wikipedia.org/wiki/Loratadine*.

from Wikipedia and other web addresses, as well as names and identifiers including URL encoding were URL decoded. Different spellings and abbreviations of authority names were unified to one representative authority name. Where there was one primary human-readable name provided, the name was assigned to *Database-name Name* name type, e.g. DrugBank Name. When the value in an *authority:value* pair did not match the authority standard and followed another standard, the value was reassigned to the predicted correct authority. Curated relationships are annotated with "katkoler" next to the original source. Typographical errors from source databases were removed.

3.1.4 Database construction

From the cleaned edge database, we constructed a graph in R (R Core Team, 2019) using `igraph` (Csardi and Nepusz, 2006). We constructed graph $G_c = (V, E)$, where G_c is a graph on drug (compound) synonym information. V, the set of vertices, represent a database *authority:value pair* (Eq. (3.2)) and E, the set of edges, represents each synonym relationship from a source database.

Vertices were unified where they shared a name and authority, hence each source database provided a unique source of edges, but a shared source of vertices. Edges extracted from each source database correspond to a subset of all edges. An edge only exists if the relationship is present in one of the source databases.

We identified connected components of graph G_c, under expectation that each connected component represented a set of drug synonyms for one drug. Each connected component was assigned a KATdb unique identifier.

A MySQL database was constructed and implemented into the KATdb shiny app.

3.1.5 Estimation of correctness

The 10 largest components were selected for partial manual curation. The edge betweenness algorithm, as described by Brandes (2001) and implemented in `igraph` (Csardi and Nepusz, 2006), was calculated for all edges in the 10 largest components and 2% of edges with the highest edge betweenness score were manually assessed for correctness. Each node of the manually checked edges was searched on Google (*https://www.google.com/*) to manually establish the connection to the other node in at least two databases (excluding the source edge database). Each edge was assigned one of the following terms: *correct*, *related structure*, *related* or *incorrect*, signifying the level of correctness. An edge was assigned *correct* if the relationship could be established manually between the two vertices from two additional databases (not including the source database). Spatial isomers and small structural changes were marked as *related structures*. Edges connecting individual ingredients of the same mixture were marked as *related*. Edges connecting non-related drugs were marked as *incorrect*.

A source database identified with high levels of incorrect edges was removed from KATdb. The synonym database was reconstructed, and the estimation of correctness was repeated. We again identified 10 largest connected components and manually validated the 2% of edges with the highest betweenness score.

3.1.6 Robustness

We tested KATdb robustness by removing a random subset of nodes, 5% at a time for 10 iterations. We measured the size of the largest component (Lordan and Albareda-Sambola, 2019), the mean component size and the number of components. These robustness measures were averaged across 10 iterations. We performed the test on the whole network and also on network variations with one source database removed at a time.

3.1.7 KATdb shiny app

A visual interface for KATdb was developed using `shiny` (Chang et al., 2019) to allow exploration of KATdb's components and the relationships connected into one synonym entity. It provides an overview of KATdb and provides a mapping functionality, where drug names and identifiers can be translated from one synonym authority to another. Additionally, it facilitates exploration of synonym relationships in table format and as plots by exploring information encompassed in each connected component.

3.1.8 Translation success and redundancy

We defined the *input name* as the name that was used as the query name, the number of *found names* as the number of input names that were found in KATdb, the *goal name* as the name to which we were aiming to map the input name, and *mapped terms* as all found names that had at least one goal name.

We defined:

$$\text{translation success} = \frac{|\text{mapped names}|}{|\text{found names}|} \tag{3.4}$$

measuring what proportion of input names present in KATdb that were mapped to at least one goal synonym, and:

$$\text{translation redundancy} = \frac{|\text{goal names}|}{|\text{mapped names}|} \tag{3.5}$$

measuring how many successfully mapped names, mapped to more than one goal name.

Translation test cases (A-K) and their details are listed in Supplementary Table A.2 and summarised in Chapter 4 Table 4.2. For KATdb cases we used the translation feature of KATdb visual interface to map drug names into the goal name type. For manual test case A (BRD ID to name) we used L1000CDS^2 metadata and mapped *pert_id* to *pert_desc*. For manual case E (RepoDB name to LINCS name) we mapped names by matching lower-case names from input type to goal name type. For manual case F (RepoDB name to BRD ID) we used mapping results from E and then matched them to mapping results from A, so that the mapping path was as follows: RepoDB name → LINCS name → BRD ID.

All translations ignored the letter case during mapping, summarisation, and translation success and redundancy calculation.

3.2 Pathway-Drug Coexpression Network (PDxN)

The method for the Pathway-Drug Coexpression Network (PDxN) was adapted from Pita-Juárez et al. (2018). The published method designed for pathway gene sets was applied here to an updated and heavily extended gene set collection including pathway gene sets and drug signatures.

3.2.1 Gene expression background data

We used 134 experiments with 3207 Affymetrix Human Genome U133 Plus 2.0 microarrays (GPL570) from 72 normal human tissues manually curated in Barcode 3.0 (McCall et al., 2014). The curated microarrays in Barcode 3.0 have been filtered to exclude poor quality samples using the global normalized unscaled standard error (GNUSE) method

(McCall et al., 2011a) which was a single-array version of a multi-array quality metrics (McCall et al., 2011b, 2014). We used the R package `GEOquery` (Davis and Meltzer, 2007) to retrieve raw files from the Gene Expression Omnibus (GEO) (Barrett et al., 2007) and we processed the raw data with frozen robust multi-array analysis (fRMA) (McCall et al., 2010). Redundancies in probe annotations were solved by mapping multiple probes to unique Entrez Gene IDs by their mean expression level. To reduce batch effect, the genes in each experiment were ranked, from 1 (low expression) to K (high expression), giving the same dynamic range to each experiment.

3.2.2 Gene sets construction

Two independent sets of gene sets were constructed for PDxN; set P, the set of pathways and set D, the set of chemical or drug gene sets. Set P consists of gene sets represented by canonical pathways from MSigDB (Subramanian et al., 2005) and static modules from `pathprint` (Altschuler et al., 2013; Wu et al., 2010b). Set D includes two gene sets for each drug, an up- and down-regulated, genes separately for a given drug from L1000 characteristic direction signature search engine (L1000CDS^2) (Duan et al., 2016).

MSigDB (Subramanian et al., 2005). C2: Canonical Pathways collection from MSigDB (Subramanian et al., 2005) (v6.2 updated July 2018) was retrieved from *http://software.broadinstitute.org/gsea/downloads.jsp* on 26th July 2018. The collection is a curated selection of 1329 pathway annotations from other databases: Reactome (Croft et al., 2011), KEGG (Kanehisa et al., 2014), the Pathway Interaction Database (PID) (Schaefer et al., 2009), Biocarta (Nishimura, 2001), the Matrisome Project (Naba et al., 2012), Signal Transduction Knowledge Environment (Gough, 2002), SigmaAldrich and Signalling Gateway (Saunders et al., 2008).

Pathprint (Altschuler et al., 2013). `pathprint` (Altschuler et al., 2013) is a Bioconductor (Huber et al., 2015) package that includes 633 gene sets. The gene sets represent pathways derived from a range of pathway databases (Reactome (Croft et al., 2011), KEGG (Kanehisa et al., 2014), Wikipathways (Kelder et al., 2012), Netpath (Kandasamy et al., 2010)) in addition to static modules derived from a functional gene interaction network (Wu et al., 2010b). Only the static modules from the `pathprint` package were used.

L1000CDS (Duan et al., 2016). We used drug signatures from the L1000 characteristic direction signature (L1000CDS) dataset available at L1000CDS search engine (L1000CDS^2) (Duan et al., 2016). The L1000CDS' underlying dataset is the drug expression data from LINCS Level 3: Normalised gene expression profiles of landmark genes and imputed transcripts (Subramanian et al., 2017). Level 3 data has been normalised using invariant set scaling followed by quantile normalisation first within-plate and then across replicate plates. LINCS Level 3 has been processed into drug signatures with a characteristic direction (CD) method. Each drug signature represents a set of replicates under specific conditions, a combination of drug, cell line, exposure time, concentration and batch. Cell lines include primary cell lines, cancer cell lines, stem cell lines, and differentiated cell lines from different tissue types (Supplementary Table D.9).

CD measures (Clark et al., 2014) is a multivariate method that first identifies the linear hyperplane that best separates the control samples from the case samples using linear discriminant analysis, and then uses the normal to the hyperplane to define the direction of change in expression space for each gene.

L1000CDS has used LINCS L1000 Level 3 normalised data to calculate a CD unit vector for each experiment replicate in comparison with all the control replicates on the same plate. A CD signature has been computed for each experimental condition by averaging the CDs across replicates. The mean of the pairwise cosine distance between the CDs across replicates has been used as a test statistic to assess the significance of each signature. The mean has been compared with a null distribution constructed from random sampling of irrelevant CD replicates to compute a p-value. The differentially expressed genes have been calculated using the random product algorithm.

The CD signatures and associated metadata are stored in a MongoDB database and are available for download from L1000CDS[2]. L1000CD signatures were downloaded from *http://amp.pharm.mssm.edu/public/L1000CDS_download/* on 13th July 2018. Drug signatures were extracted using MongoDB and Python.

On 13th July 2018, the L1000CDS database included 119,156 drug signatures of which 26,124 met the significance threshold of p-value < 0.05 and had more than 5 genes in each direction.

Each drug signature was split into two individual and non-overlapping gene sets: up-regulated, and down-regulated gene set according to member gene CD value, CD > 0 for up- and CD < 0 for down-regulated. Thus, each drug node in PDxN represents a particular drug-direction gene set.

3.2.3 Computational improvements to Pathway Coexpression Network (PCxN) method

The code review and rewriting R functions into C++ were done in collaboration with thesis supervisor Wenbin Wei and junior programmer Sokratis Kariotis. The performance testing was done by the author of this thesis.

We accessed the Pathway Coexpression Network (PCxN) method (Pita-Juárez et al., 2018) code base at *https://github.com/yeredh/pcxn_plos* on 19th March 2018. We reviewed the 4-part method and concluded that part 0 and 3 had negligible resource requirements. In part 1, we identified a repeated calculation when summarising the pathway expression for each pathway pair. We resolved it by pre-calculating a pathway summary matrix. We increased the parallelisation of part 1 by calculating one experiment per task rather than one tissue per task. We increased from 72 to 134 tasks for part 1. In part 2, we identified that the part was split into $\frac{\text{number of pairs}}{1000}$ tasks where the most resource consuming step was repeated by each task. Each task was required to read into memory all study-level correlation estimates, which was identified as the bottleneck. We decreased the number of tasks and thus the redundantly repeated reading into memory by increasing the number of pairs per task from 1000 to 100,000.

We added two additional features that improve performance at a large number of gene sets. We implemented an option to limit the calculation to relationships between different types of nodes or only within one type of node. Additionally, we added an optional feature that joins different completed versions of the network. These two features were not compared to the original method.

We measured the performance improvement by running a set of test runs with the original and improved PCxN method. We varied the number of the gene sets and the number of relationships calculated. We used 1473 pathway gene sets from the PDxN pathway set. In each test we calculated all possible relationships, i.e. all pathway↔pathway pairs, resulting in $\frac{n(n-1)}{2}$ number of pairs, where n is the number of gene sets.

In order to calculate the whole network from the original method, we were required to increase the number of tasks part 2 calculated and part 3 joined from 2 to $\frac{\text{number of pairs}}{1000}$, because the code available on GitHub (*https://github.com/yeredh/pcxn_plos*, 19th March 2018) only calculates 2000 pairs. In addition, we used an updated pathway set compared to the PCxN pathway set described in Pita-Juárez et al. (2018). The updated pathway set consisted of 1473 pathways compared to 1330 in the method paper.

We tested both methods at 10, 20, 50, 100, 200, 400, 800, 1000, 1200 and 1473 random gene sets from the PDxN pathway set, where the 1473-version represented the whole PDxN pathway node set. We ran each part with 8 cores with 4GB virtual memory (vmem) each. We measured maximum vmem and wall clock time. We ran each of the tests twice and averaged the max vmem and wall clock time. We reported a total maximum vmem and wall clock time per part and for all parts together with an increasing number of pairs.

3.2.4 Network construction

We adapted the network construction method described in Pita-Juárez et al. (2018) to an extended collection including drug and pathway gene sets and limiting the relationships to pathway↔drug rather than all possible relationships.

A bipartite weighted network was constructed with two independent sets: P, the set of pathway nodes and D, the set of chemical nodes including up- and down-regulated genes as separate gene sets for each drug. The Pearson correlation estimates (R Core Team, 2019), $\hat{r}$, were calculated for every node pair between P and D. An edge between nodes exists if the correlation estimate between those nodes is below the adjusted p-value (q-value) threshold < 0.05. The edges were weighted by the absolute value of the correlation estimate, $|\hat{r}|$.

The network was constructed based on the expression correlation between a pair of gene sets, from set P to set D. The gene set correlation coefficients and their corresponding p-values were first estimated for each experiment, then the experiment-level estimates were combined into global estimates.

The genes in each experiment were first ranked, from 1 (low expression) to K (high expression), giving the same dynamic range to each experiment. As each gene set is represented by a set of genes, the gene set expression E is a gene set summary statistic based on the expression ranks of the gene set genes. The gene set expression E is the mean of the expression ranks of the gene set genes. We used the shrinkage estimator from R package `corpcor` (Schafer et al., 2017) to compute the experiment-level gene set correlation coefficients conditioning on gene overlap. The overall correlation from experiment-level estimates between two gene sets is calculated with Hunter-Schmidt weighted average estimator. The overall p-value was calculated with Liptak p-value aggregation. The combined p-values were then adjusted with Benjamini-Hochberg FDR method.

The overlap coefficient was calculated by the size of the intersection divided by the size of the smaller of the two sets

$$o_{GH} = \frac{|G \cap H|}{min\{|G|,|H|\}} \tag{3.6}$$

where o is the overlap between gene set G and H. Two disjointed (i.e. non-overlapping) gene sets have the overlap coefficient = 0. If a gene set is fully contained within the other, then the overlap coefficient = 1.

Pathway Coexpression Network (PCxN) construction

A new version of PCxN was constructed with the improved method to enable direct comparisons with PDxN. We constructed PCxN using the MSigDB v6.2 C2 Canonical Pathways (Subramanian et al., 2005) and Pathprint Static Modules (Altschuler et al., 2013; Wu et al., 2010b) as in PDxN pathway set P. Correlation estimates were calculated between each pair of pathway gene sets as described in 3.2.4.

3.3 Disease Signature

3.3.1 Disease gene expression data

Gene expression studies used in case studies are summarised in Table 3.2.

Juvenile idiopathic arthritis

We queried the GEO datasets search engine for juvenile idiopathic arthritis (JIA) in blood or peripheral blood mononuclear cells (PBMC) studies analysed on either microarray or high throughput sequencing. We used the query:

"(Juvenile Idiopathic Arthritis
OR juvenile idiopathic arthritis)
AND (blood OR PBMC)
AND "gse"[Filter]
AND ("Expression profiling by array"[Filter]
OR "Expression profiling by high throughput sequencing"[Filter])"

at *https://www.ncbi.nlm.nih.gov/gds* on 9th January 2019. The search yielded 30 studies (Supplementary Table A.3). An additional 20 studies were curated by rheumatoid arthritis expert Professor Lester Kobzik (Department of Environmental Health, Harvard T.H. Chan School of Public Health) (personal communication). 18 studies with less than 20 samples were not considered due to low sample numbers. Six studies were excluded because of no control samples. Two studies were excluded because they did not include any JIA samples. Three studies were excluded because they only included treated JIA samples. Five were analysed on not commonly used arrays and thus excluded due to their platform. One was excluded because of the study design, it investigated monozygotic twins. One study was a duplicate of another study and one was removed because it represented a super series, from which the individual series were already included in the search results. The remaining 16 studies were then curated by Lester Kobzik. Ten studies were selected based on having untreated systemic or polyarticular JIA with control samples in either PBMC or whole blood (Table 3.2). In addition, GSE79970 was removed, because the data was identified to be unsuitable to study JIA by the authors (Wong et al., 2016).

Alzheimer's disease (AD)

Preprocessed RNA-Seq data on G2B2, H10, A5, I45F, and I47F Alzheimer's 3D cell models (Kwak et al., 2020) was shared by collaborators in the Tanzi and Kim labs (Doo Yeon Kim, Assistant Professor of Neurology, Harvard Medical School, Building 114, Charlestown Navy Yard, 16th Street, Charlestown, Massachusetts 02129, USA). The 3D cell model RNA-Seq data was preprocessed by the Harvard Bioinformatics Core.

Preprocessed RNA-Seq from the deceased human brain was obtained from the Mayo dataset (doi:10.7303/syn2580853) (Allen et al., 2016), accessible at *https://www.synapse.org/#!Synapse:syn3163039*. Sarah Morgan (SITraN, University of Sheffield; Department of Pathology, Beth Israel Deaconess Medical Center, Harvard Medical School) advised on the Mayo Alzheimer's disease dataset sample selection. Control and Alzheimer's disease samples from temporal cortex were used with the age at death >= 75 and RNA integrity number (RIN) >= 7.5 (Gallego Romero et al., 2014).

Parkinson's disease (PD)

RNA-Seq data from sporadic Parkinson's patients' fibroblasts with lysosomal or mitochondrial dysfunction (Carling et al., 2020), and zebrafish (*D. rerio*) homozygous GCH1

Table (3.2) Summary of gene expression datasets used for disease signature generation in case studies as well as assessing the method. Number of control and disease samples post-quality control is listed in brackets in their respective columns. AD — Alzheimer's disease; GCH1 — GTP cyclohydrolase 1; GPL570 — U133 Plus 2.0 Array; GPL96 — U133A Array; GPL97 — U133B Array; GPL11154 — Illumina HiSeq 2000; JIA — juvenile idiopathic arthritis; n/a — not applicable; PD — Parkinson's disease; PBMC — peripheral blood mononuclear cells; polyJIA — polyarticular JIA; sJIA — systemic JIA.

Dataset	Platform	Sample Tissue	Control Samples	Disease Samples	Disease	Organism
GSE15645	GPL570	PBMC	13 (12)	14 (13)	RF-polyJIA	human
GSE26554	GPL570	PBMC	23 (22)	38 (35)	RF-polyJIA	human
GSE20307	GPL570	PBMC	56 (52)	20 (20)	sJIA	human
GSE21521	GPL570	PBMC	29 (26)	18 (18)	sJIA	human
GSE7753	GPL570	PBMC	30 (27)	17 (17)	sJIA	human
GSE80060	GPL570	whole blood	22 (22)	22 (21)	sJIA	human
GSE8650	GPL96	PBMC	21 (21)	16 (14)	sJIA	human
GSE8650	GPL97	PBMC	21 (21)	12 (12)	sJIA	human
GSE112057	RNA-Seq	whole blood	12 (12)	26 (26)	sJIA	human
H10, A5, G2B2	RNA-Seq	neurons	3 (3)	3, 3 (3, 3)	AD	3D cell line
I45F, I47F, G2B2	RNA-Seq	neurons	3 (3)	3, 3 (3, 3)	AD	3D cell line
Mayo	RNA-Seq	temporal cortex	37 (37)	69 (69)	AD	human
Lysosomal dysfunction	RNA-Seq	fibroblast	4 (4)	5 (4)	PD	human
Mitochondrial dysfunction	RNA-Seq	fibroblast	5 (4)	5 (4)	PD	human
GCH1 mutant	RNA-Seq	neurons	4 (3)	4 (4)	PD	zebrafish
GSE133815	GPL570	liver	12 (11)	11 (11)	n/a	human

mutant neurons (Larbalestier et al., 2020) were shared by collaborators in the Bandmann lab (Oliver Bandmann, Professor of Movement Disorders Neurology, Department of Neuroscience, Sheffield Institute for Translational Neuroscience, University of Sheffield, 385a Glossop Road, Sheffield, S10 2HQ).

Young and Old Liver

GSE133815 microarray human dataset not related to JIA, AD or PD was selected as control. GSE133815, accessible from GEO, includes liver samples of young (21–45 years) and old (69+ years) men and women.

3.3.2 Disease signature generation

Preprocessing

Microarray. The raw microarray data was downloaded from GEO with series matrix files including metadata. Quality control was carried out to identify and remove any outliers. In particular, R package `arrayQualityMetrics` was used to identify outliers in microarray studies (Kauffmann et al., 2009). `arrayQualityMetrics` assesses: (i) between array comparison by calculating distances between arrays and doing principle component analysis, (ii) homogeneity between arrays with boxplots of $\log_2$ intensities and density estimate plots, (iii) variance mean dependence with standard deviation versus rank of the mean, and (iv) relative log expression (Brettschneider et al., 2007). Given a gene expression dataset with n samples, appropriate samples were chosen for comparison. Raw microarray data was processed with fRMA available in R package `frma` (McCall et al., 2010). For array number GPL96 and GPL97, from the GSE8650 data set, RMA from `affy` (Gautier et al., 2004) was used.

RNA-Seq. All but the PD datasets were obtained as count-level matrices. The Carling et al. (2020) Lysosomal and mitochondrial dysfunction RNA-Seq data was preprocessed by Claire Green (SITraN, University of Sheffield) and the Larbalestier et al. (2020) zebrafish data by Wenbin Wei (SITraN, University of Sheffield). The PD RNA-Seq samples were preprocessed with RNA-Seq pipeline bcbio (*https://github.com/bcbio/bcbio-nextgen*) using Salmon quantification (Patro et al., 2017). In all RNA-Seq studies, outliers were identified with principal component analysis (PCA). Transcripts with low expression (less than 10 counts in more samples than the size of the smallest group) were excluded from further analysis. The data was normalised by using trimmed mean of M values (TMM) normalisation implemented in `edgeR` (Robinson et al., 2010).

The remaining genes in the RNA-Seq and microarray datasets were mapped to Entrez Gene IDs with Biomart (Durinck et al., 2009). The homolog mapping from zebrafish to human Entrez Gene IDs was performed with Biomart for GCH1 mutant dataset. To resolve redundancies, multiple probes were mapped to unique Entrez Gene IDs by their mean expression level. The pathway data using Entrez Gene IDs was downloaded from MSigDB C2 collection (August 2018) (Subramanian et al., 2005). The gene expression dataset was then further filtered to *m* pathway member genes that were represented in the data.

The RNA-Seq normalised counts were then log-transformed using voom from `limma` (Ritchie et al., 2015). The RNA-Seq data was transformed to an array-comparable scale so that the same statistical tests could be applied to both types of data in the downstream analysis.

Gene set summary statistic — Mean top 50%

The mean top 50% gene set summary statistic was adapted from Hwang (2012) (Fig. 5.13). The gene expression profiles were *z*-scaled for each gene across samples. Then the summary statistic was calculated across samples, *n*, for each gene set or pathway, *m*, whose member genes were represented in the data.

The resulting $m \times n$ matrix X is then a *z*-scaled gene expression profile of the pathway's member genes across samples and each element x_{ij} is a *z*-scaled expression level of a member gene i in sample j.

The member genes' expression profile was subject to Welch's t-test. Then, the member genes were sorted by $|t|$ in descending order, or equivalently, by *p*-value in ascending order. The top 50% of the member genes were selected. If there was an odd number of member genes, then the gene member with median $|t|$ was also selected. Selected genes' gene expression profile was averaged:

$$a_j = \frac{1}{m} \sum_{i=l}^{m} x_{ij} \tag{3.7}$$

This pathway-level aggregation method derived a pathway expression profile, *a*, which was a vector with *n* elements.

Due to different rates of mapping from a particular probe set m, the number of genes in a pathway varied between different studies analysed with this method.

Differential pathway expression

The pathway summary matrix of case and control samples was then used to generate differential pathway expression profile using `limma`. The pathways with q-value < 0.5 were considered differentially expressed. The top up-regulated pathways were defined as pathways with the highest positive logFC score and the top down-regulated pathways were defined as pathways with the lowest negative logFC score. The top up- and down-regulated pathways together were considered as the disease pathway signature.

3.4 Drug Prioritisation

3.4.1 Disease cluster definition

The pathways from the disease signature (Section 3.3) consisting of n number of top up-regulated pathways and n number of top down-regulated pathways (q-value < 0.05) were separated into clusters. We ran drug prioritisation for $n = 5, 10, 15, 20$.

3.4.2 Disease sub-network generation

PDxN was filtered so that only correlation edges with q-value < 0.05 were considered. A disease sub-network was generated consisting of pathway nodes representing pathways from the disease clusters and all connected drug nodes.

3.4.3 Disease cluster score

Each disease cluster was considered separately. First the average correlation was calculated per drug-direction node (one for up and one for down per drug) for each cluster, forming a correlation cluster score (Eq. (3.8)). The edge p-values were combined with Fisher's method (Eq. (3.10)) (Mosteller and Fisher, 1948).

The pathway cluster↔drug-direction summary:

$$\hat{r}_{Pd} = \sum_{p \in P}^{n} (r_{pd}) \times \frac{1}{|P|} \tag{3.8}$$

where $\hat{r}_{Pd}$ is the cluster P to drug-direction d summary correlation estimate, n is the number of pathway to drug-direction edges (q-value < 0.05), $\hat{r}_{pd}$ is the correlation edge between pathway node p and drug-direction node d.

The difference between the cluster correlations was then calculated between the up and the down part of the drug node (Eq. (3.9)) and the p-values were again combined with Fisher's method (Eq. (3.10)).

The pathway cluster↔drug summary:

$$\hat{r}_{PD} = \hat{r}_{Pd_{\text{up}}} - \hat{r}_{Pd_{\text{dn}}} \tag{3.9}$$

where $\hat{r}_{Pd_{\text{up}}}$ is the cluster P to up-regulated drug d_{up} summary correlation estimate, $\hat{r}_{Pd_{\text{dn}}}$ is the cluster P to down-regulated drug d_{dn} summary correlation estimate, $\hat{r}_{PD}$ is the correlation edge between pathway cluster P and drug D, previously represented as two drug-direction nodes d_{up} and d_{dn}.

Fisher's method:

$$-2\sum log(p_{pd}) \tag{3.10}$$

Where p_{pd} is the q-value for the edge between pathway p and drug-direction node d.

3.5 Benchmarking

3.5.1 Disease signature

Correlation between differentially expressed pathways. Spearman's rank correlation was calculated between differential pathway expression $\log_2$ fold change (logFC) values for all pathways (no p-value cut off) between every pair of studies. Spearman's correlation was chosen due to its increased robustness to outliers compared to Pearson's. A heatmap was plotted with `ComplexHeatmap` (Gu et al., 2016).

Overlap between differentially expressed pathways. Pairwise overlap of differentially expressed pathways (q-value < 0.05) was calculated between all JIA studies as well as the liver study (GSE133815). The overlap coefficient was calculated by the size of the intersection divided by the size of the smaller of the two sets of pathways

$$o_{PQ} = \frac{|P \cap Q|}{min\{|P|, |Q|\}} \tag{3.11}$$

where o was the overlap coefficient between differentially expressed pathway signatures (q-value < 0.05) from studies P and Q. Two disjoint sets of differentially expressed pathways (DEP) have the overlap coefficient of 0 and if a DEP set is fully contained within the other, then the overlap coefficient = 1.

Overlap sJIA signature. Overlapping sJIA differentially expressed pathways were generated from differentially expressed pathway profiles (q-value < 0.05) from PBMC systemic JIA (sJIA) studies (GSE7753, GSE20307, GSE21521, GSE8650_GPL96, and GSE8650_GPL96). The significance of the size of the overlap was tested with a permutation test (number of permutations =10,000) (Phipson and Smyth, 2010).

Gene-level disease signature. Differential gene expression was performed for two sJIA studies: GSE7753 and GSE112057. We quality controlled and preprocessed the studies as described in Section 3.3.2. The transcripts were mapped to Entrez Gene IDs with Biomart (Durinck et al., 2009). To resolve redundancies, multiple probes were mapped to unique Entrez Gene IDs by their mean expression level. The RNA-Seq normalised counts were then log-scaled using voom from `limma` (Ritchie et al., 2015). We generated differentially expressed pathways with `limma` (Ritchie et al., 2015). If not stated otherwise we defined differentially expressed genes as all genes that meet the significance threshold of q-value < 0.05 and absolute log fold change |logFC| > 1.

Enrichment. Disease ontology (Schriml et al., 2019) enrichment was performed with `clusterProfiler` (Yu et al., 2012), `DOSE` (Yu et al., 2015) and `enrichplot` (Yu, 2019) on the top differentially expressed genes and pathways. We define the top differentially expressed genes as the 1000 genes with the highest absolute logFC (q-value < 0.05), and top differentially expressed pathways as the top n pathways with the highest absolute logFC (q-value < 0.05) whose genes members added up to 1000 unique genes (Table 3.3).

Table (3.3) Number of genes and pathways used in the Disease ontology enrichment query. The numeric columns reflect the number of genes used for gene-level, and the total number of pathways with number of genes in brackets for pathway-level enrichment.

Study	Level	Total	Up-regulated	Down-regulated
GSE7753	gene	1000	580	420
GSE112057	gene	1000	350	650
GSE7753	pathway	45 (999)	40 (553)	5 (448)
GSE112057	pathway	75 (991)	52 (654)	23 (369)
sJIA PBMC overlap	pathway	23 (980)	6 (210)	17 (770)

3.5.2 Pathway Drug Coexpression Network (PDxN)

Pathway and drug annotation enrichment

Network projections. A projection of Pathway Drug Coexpression Network (PDxN) (q-value < 0.05) was made for each of the node sets, P and D using `igraph` (Csardi and Nepusz, 2006), making a pathway projected graph and a drug projected graph.

PCxN (q-value < 0.05), and PDxN pathway projection were clustered with Louvain, also known as Multi-level, community-finding method with `igraph` R package (Blondel et al., 2008). PDxN (q-value < 0.05) was clustered using Label Propagation Algorithm weighted bipartite plus (LPAwb+) (Beckett, 2016) from `bipartite` (Dormann et al., 2009) R package. Absolute edge correlation estimates were used as edge weights in weighted clustering for PDxN and PCxN. PDxN drug projection was first split into drug-up and drug-down subgraphs and then clustered with Louvain method.

KEGG and Reactome enrichment. Pathways in PDxN, PCxN and PDxN pathway projection were annotated with KEGG B-level terms and Reactome pathway group terms. KEGG hierarchical structure of pathways was downloaded on 20th September 2019 from *https://www.kegg.jp/kegg-bin/get_htext?hsa00001* and Reactome Pathways hierarchy relationship file was downloaded on 20th September 2019 from *https://reactome.org/download-data*. The Reactome pathway network was constructed with `igraph` and clustered with Louvain method (Blondel et al., 2008). The clusters were annotated with Reactome group pathway terms.

Every pathway in PDxN or PCxN that was present in the KEGG hierarchical structure of pathways or Reactome pathway network was annotated with the higher-level pathway terms (KEGG B-level and Reactome group term). Making an $i \times j$ matrix x where i represents the number of pathway annotation classes and j represents the number of PDxN or PCxN clusters. Over- or under-representation of annotated pathways in a particular network cluster was assessed with an enrichment score.

Let x be an $i \times j$ matrix. The enrichment score for x_{ij} is then:

$$enr(x_{ij}) = \frac{observed(x_{ij})}{expected(x_{ij})} \tag{3.12}$$

where:

$$observed(x_{ij}) = x_{ij} \tag{3.13}$$

$$expected(x_{ij}) = \frac{(\sum_{l=1}^{n_i} x_{il})(\sum_{l=1}^{n_j} x_{lj})}{\sum_{l=1,k=1}^{n_i,n_j} x_{lk}} \tag{3.14}$$

with n_i and n_j as the number of rows and columns in x, respectively. Then:

$$enr(x_{ij}) = \frac{x_{ij}(\sum_{l=1,k=1}^{n_i,n_j} x_{lk})}{(\sum_{l=1}^{n_i} x_{il})(\sum_{l=1}^{n_j} x_{lj})} \tag{3.15}$$

enrichment scores below 1 indicating under-representation or depletion were transformed:

$$enr(x_{ij}) < 1 = -\frac{1}{enr(x_{ij})} \tag{3.16}$$

so that $enr(x_{ij}) > 1$ represents enrichment and $enr(x_{ij}) < -1$ represents depletion. For example, -2 would indicate twice less than expected and $+2$ indicates twice as many as expected. An enrichment score of $|enr(x_{ij})| = 1$ indicates that the observed number of terms is the same as the number of terms expected by chance.

We used a contingency table based on shared terms between the network cluster and annotation group to calculate the enrichment score p-value using a two-sided minimum likelihood hypergeometric test (Table 3.4).

ATC class enrichment. Using KATdb, drug names from PDxN were mapped to Anatomical Therapeutic Chemical (ATC) classification system codes. Drug nodes in PDxN and the subgraphs of the PDxN drug projection were then annotated with ATC classification system level 3 codes. ATC codes were accessed as described in Section 3.1.2. Redundancies in ATC annotations were resolved by counting the drug matching to multiple

Table (3.4) Contingency table for shared terms between network clusters and annotation groups. The contingency table splits the pathway terms in a network cluster and an annotation group in 4 disjoint sets. x_{ij} — terms in the cluster and in the annotation group; $y = \sum_{l=1}^{n_i} x_{il}$ — terms in the annotation group; $z = \sum_{l=1,k=1}^{n_i,n_j} x_{lk} - y$ — terms not in the annotation group; $m = \sum_{l=1}^{n_j} x_{lj}$ — terms in the network cluster.

		Terms in network cluster		
		IN	NOT IN	Total
Terms in annotation group	IN	x_{ij}	$y - x_{ij}$	y
	NOT IN	$m - x_{ij}$	$z - (m - x_{ij})$	z
	Total	m	$(y + z - m)$	$y + z$

ATC classes in each class. Enrichment scores were calculated from a $i \times j$ matrix x where i represents the number of ATC annotation classes and j represents the number of PDxN or projection clusters. The enrichment was calculated as described above in Eqs. (3.15), (3.16). The enrichment p-value was calculated using the hypergeometric test from contingency table (Table 3.4) adapted for ATC class terms.

3.5.3 Drug prioritisation

Prioritised drug lists were scored for approved or experimental drugs for a given disease. We assessed the performance by generating receiver operating characteristic (ROC) curves and calculating the area under the ROC curve (AUC). Different definitions of "true positive" drug lists were explored in order to establish strengths and weaknesses of each approach. Current gold-standard lists of approved drugs were used in each case study. ATC classes were used to highlight the weaknesses of approved lists in JIA case study. In AD, an experimental *in vitro* drug screen true positive list was used and in PD we used a list of expert-curated experimental neuroprotective drugs.

True positive drug lists

Approved drug-disease pairs. RepoDB (Brown and Patel, 2017) and European Medicines Agency (EMA) (European Medicines Agency, 2019) were used to extract approved drugs for:

(i) juvenile idiopathic arthritis with query "rheumatoid OR juvenile arthritis",

(ii) juvenile idiopathic arthritis with query "juvenile arthritis",

(iii) Alzheimer's disease with query "alzheimer",

(iv) Parkinson's disease with query "parkinson",

RepoDB (Brown and Patel, 2017) a resource that when downloaded included 1571 drugs and 2051 diseases. It included 6677 approved and 4123 failed FDA drug-disease pairs. When downloaded European Medicines Agency (EMA) (European Medicines Agency, 2019) included 1111 authorised, 46 refused, and 226 withdrawn drug-disease pairs. It included 1038 different drugs and 1547 free-text descriptions of indications. The extracted approved lists were mapped to BRD IDs with KATdb (Supplementary Tables D.5, D.6, E.4, and E.6).

ATC class lists. Two ATC drug classes have been used in benchmarking sJIA prioritised drug lists: M01 – Anti-inflammatory and antirheumatic products, and L04A – Immunosuppressants. All BRD IDs annotated with M01 or L04A were extracted and marked as true positives in their respective tests (Supplementary Tables D.7 and D.8).

AD *in vitro* drug screen list. A confidential true positive (TP) drug list was shared with us by our collaborator Tanzi and Kim group. The list consists of drugs which have shown reduction in Aβ or reduction in Aβ and tau as part of the high-throughput 3D drugs screen (3DDS) using 3D human neural culture systems related to the A5 3D cell culture. Out of approximately 1200 FDA and other biologically active drugs screened, 38 ameliorated AD-related pathology.

Curated neuroprotective drugs. A list of 6 neuroprotective drugs was curated by Professor Oliver Bandmann (SITraN, University of Sheffield) as potential true positives for Parkinson's disease. The neuroprotective drugs were mapped to BRD IDs with KATdb (Supplementary Table E.7).

Evaluating drug signature feature ranking

The prioritised drug lists were ranked from 1 (best) to *n* (worst), where *n* is the number of drug signatures associated with a prioritisation score in a given list. The ranks were scaled to a range of 0–1, where 0 is the best and 1 represents the worst rank. The list was then subsetted to include only drug signatures in a TP list. The TP drug signature ranks were averaged between lists from 5 and 10 pathway clusters, and 15 and 20 pathway clusters, yielding two instead of four ranked drug lists. Scaled ranks were then averaged across each drug signature feature (batch, drug id, concentration, cell line, perturbation time) per each 5–10 or 15–20 pathway prioritised drug list. The mean scaled rank was calculated representing the average feature rank across all prioritised drug lists. Features that were on average in the bottom 40% were marked as low-performing and removed from the prioritised drug lists.

ROC curve generation

The ROC curves and AUC scores were generated using R package `precrec`. The prioritised drug list was assessed for its specificity and sensitivity by scoring it with a selected TP drug list. Specificity and sensitivity were generated based on the confusion matrix (Table 3.5).

Table (3.5) Confusion matrix describing performance of generated prioritised drug lists scored with approved or experimental true drugs. True drugs are drugs from a true positive list, which can be a list of approved, experimental or predicted drugs. FN — False Negative; FP — False Positive; TN — True Negative; TP — True Positive.

		Predicted drug	
		above threshold	below threshold
Approved drug	True	TP	FN
	False	FP	TN

The threshold values for generation of the ROC curve equate to the increasing length of the ranked list, from 0 to n, where n is the number of drug signatures with an associated score. The AUC scores were summarised in a heatmap plotted with `ComplexHeatmap` (Gu et al., 2016).

Comparison with LINCS on clue.io

Drug prioritisation from official LINCS platform available at *https://clue.io/* was performed on 30th January 2020 to compare PDxN pipeline to an alternative method. The gene-level GSE7753 disease signature consisting of the most up- and down-regulated genes was used for the query. The query was run at 175, 100, 50 and 20 up- and down-regulated genes, as recommended by the platform. The results were downloaded and `.gctx` files analysed in R with `cmapR` (CMap Group at The Broad Institute, 2018). The summary across cell lines score was used for assessing the method's performance. The resulting prioritised drug lists were benchmarked with the approved JIA and RA list, and ATC M01 and L04A lists.

Chapter 4

KATdb, the Drug Synonym Database

Others' contributions to this chapter. David R. Jones (Sheffield Institute for Translational Neuroscience (SITraN), University of Sheffield) has contributed to the concept of using graph theory to resolve heterogeneous drug nomenclature described in this chapter.

4.1 Abstract

Computational drug repositioning depends heavily on the availability and aggregation of different information resources. Successful translation of novel findings to clinical treatments requires objective assessment and benchmarking of drug repositioning methods. Proposed benchmarking methods include scoring indications against a well-defined gold-standard truth table. It has become challenging to perform this comparison, because poorly established links between databases lead to incorrect scoring of true positives as false negatives, affecting the performance score.

Drug repositioning methods often combine information from various drug and disease databases, each consisting of disparate collections of entries and entry-specific properties. The increased availability of publicly available data and its organisation come with aggregation challenges for repositioning because naming conventions are non-standardised. Each drug and chemical database offers a variety of drug identifiers with heterogeneous associated names. The associated names include external identifier links to other similar databases. When we aggregate drug data from different sources in a unified system, the data contains duplicate entries under different identifiers, solely due to poorly established links between databases caused by the lack of interoperability between databases and use

of inconsistent naming conventions. This harms our ability to build systems that use and develop drug knowledge.

To overcome some of the complexity, we have constructed a semantic translation resource, **KATdb**, which addresses the need for cross-database assessment of drug names, by collecting drug identifiers and synonyms from diverse publicly available sources. We have used graph theoretic approaches to connect identifiers and synonyms from different sources, establishing missing links between databases. **KATdb** is based on many databases with the aim to unify and connect the synonym information, and make a step toward a complete resource of drug-synonym information, currently connecting drug names and identifiers from 17 drug databases. 45 different types of drug synonyms have been extracted including standardised chemical descriptors, names and database identifiers. **KATdb** helps leverage the increased amounts of data and can overcome some of the aggregation challenges. **KATdb**, the drug synonym database, increases the mapping and the usability of benchmarking datasets by translating drug names and identifiers into a shared language.

To improve the objective comparison of repositioning methods, we have used **KATdb** in benchmarking. It has improved translation between a drug repositioning system built on the Library of Integrated Network-based Cellular Signatures (LINCS) (Koleti et al., 2018; Subramanian et al., 2017) and two well-defined drug-disease relationship truth tables, RepoDB (Brown and Patel, 2017) and European Medicines Agency (EMA) (European Medicines Agency, 2019). We have shown that **KATdb** increased the rate of translation 1.9-fold between approved drug names to LINCS identifiers. Thus, we were able to systematically assess performance of currently available *in silico* drug repositioning methodologies with increased power for scoring true positive indications enhanced with KATdb.

4.2 Introduction

There is an increasing amount of medical data becoming available every day and with it, new methods using the data in complex ways (Dinov, 2016). In drug repositioning, methods often combine knowledge from many different sources, introducing data aggregation challenges. Data aggregation challenges arise when different sources use non-uniform metadata language and nomenclature. For example, most drug repositioning methods rely on compiling different types of drug data and information from different sources. It is expected that the nomenclature used is not uniform. Disparate drug naming can lead to methods where one drug might be presented as multiple different drug entities under different names because the information comes from different sources. The heterogeneous nature of the nomenclature introduces duplicate information in systems and consequently, affects the power of benchmarking. In drug repositioning, methods are usually benchmarked by scoring true drug-disease relationships prioritised by the method. Typically, when the benchmarking method is scoring the retrieval of known relationships, the novel proposed relationships are only marked as true if drug names match. Therefore, the repositioning method will have inferior performance, if non-matching nomenclature is used between drug names utilised in the method and the true drug names.

Several studies and reviews have identified the problem of inconsistent and non-standardised nomenclature (Akhondi et al., 2015; Chambers et al., 2013; Dashti et al., 2019; Drug and Therapeutics Bulletin, 2018; Wohlgemuth et al., 2010). Numerous drug and chemical databases each introduce a unique identifier to identify a particular drug or chemical entry in their database. While the unique identifiers make their database more searchable, the poorly established links to other databases make aggregation more challenging. With each database establishing their own identifiers, the drugs can be referred to with at least as many names as there are databases. There are several well-defined naming conventions for chemicals, such as simplified molecular-input line-entry system (SMILES) or the International Union of Pure and Applied Chemistry (IUPAC) International Chemical Identifier (InChI), and several human-readable names such as International non-proprietary names (INNs) or trade names. There has been a move towards providing cross-referencing to other databases (Ursu et al., 2017; Wishart et al., 2018), however, these sections are limited to only a few links, are not systematic, and are sometimes erroneous.

4.2.1 Drug nomenclature

In this chapter, we distinguished between drugs and chemicals. We defined a *chemical* as one chemical compound, while a *drug* was defined as one or more chemical compounds. Each drug has several names and identifiers. We defined *drug names* as human-readable names, including generic and trade names. While we mostly considered *identifiers* as alphanumeric terms relating to drug identity. We used the term *synonym* to indicate any combination of names and identifiers predicted to refer to one drug identity. A database or synonym *authority* was defined as the synonym or name type, often matching the database name or the nomenclature standard. We used name type and synonym type interchangeably.

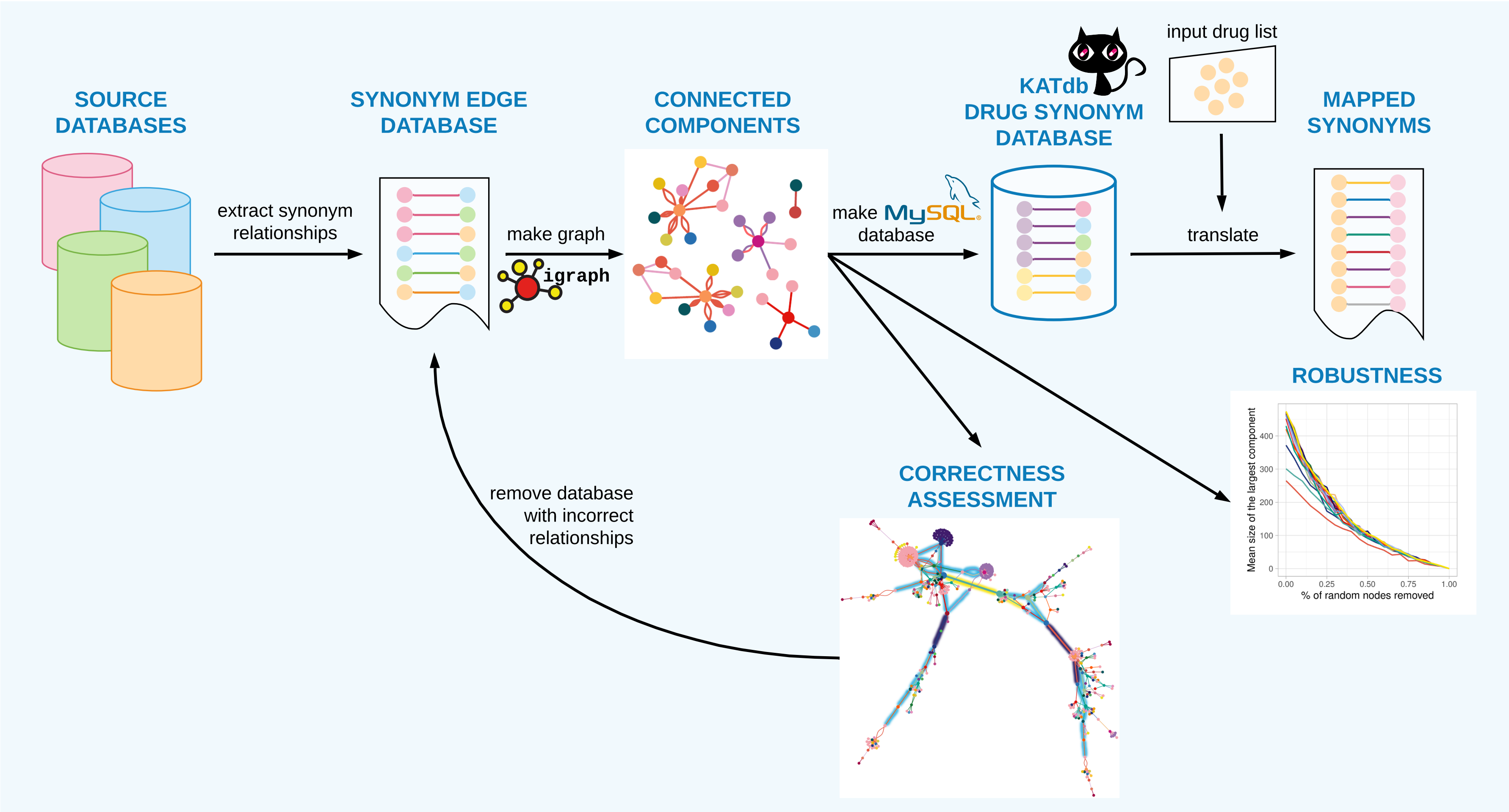


Fig. (4.1) KATdb, the drug synonym database chapter overview. In this chapter we describe the approach taken to construct KATdb using igraph (Csardi and Nepusz, 2006) and MySQL. We assess its correctness and revise the database by removing the source database with the most incorrect relationships. We look at KATdb robustness by removing random nodes from the network and apply it to 3 translation test cases.

There are well-defined chemical nomenclature guidelines and standards, but because one or more chemicals can define a drug, the drug nomenclature becomes less clear. In addition to the *many chemicals in one drug* challenge, there are several localised authorities governing drug nomenclature (Drug and Therapeutics Bulletin, 2018).

Each database uses one primary, unique identifier which is governed by a defined organisation or authority. The database-specific identifiers are traditionally regulated by the database creators, while the chemical nomenclature, drug classification systems, and generic names have their own authorities, e.g. the World Health Organisation (WHO) oversees the INN and a sub-organisation WHO Collaborating Centre for Drug Statistics Methodology (WHOCC) regulates the Anatomical Therapeutic Chemical (ATC) classification. There are some types of drug names and identifiers that are based on nomenclature principles released by an individual or organisation and then used and created by users, e.g. SMILES. These are not overseen by one single authority.

To add to the drug nomenclature complexity, approved and marketed drugs are often known under different names depending on the country (e.g. paracetamol (UK), acetaminophen (USA)) and the pharmaceutical company brand. They sometimes include different ingredients depending on the country and their names are regulated by a local governing body, e.g. European Medicines Agency (EMA, EU), the US Food and Drug Administration (FDA, US) (Drug and Therapeutics Bulletin, 2018). In Supplementary Table B.1, we show an example of each name type collected in KATdb for aspirin. For each name type we found an identifier linked to the synonym "aspirin" in either the KATdb or by manually searching for it online. We collected at least one synonym for 45 different name types out of which most have been found in KATdb.

We developed a semantic translation network, named KATdb (Fig. 4.1), based on several drug and chemical databases with the aim to unify and connect the synonym information. We aggregated synonym information from several databases with only a subset of these developed with the main aim to provide a comprehensive collection of knowledge about chemicals and drugs (Table 3.1). However, each of the source databases used in KATdb has, in addition to the primary identifier, also provided at least one other synonym. We connected 45 different types of drug names and identifiers from 17 drug databases. The identifiers included standardised chemical descriptors, names and database IDs. KATdb increased the semantic translation and the usability of benchmarking datasets by translating drug names and identifiers into a shared language.

4.3 Approach

We used a graph theoretic approach to connect all synonyms into a network (Steyvers and Tenenbaum, 2005), where synonyms representing one drug connected into one connected component. A graph theoretic approach facilitated the construction of a semantic network where its topology was exploited to provide novel insights. The approach is described in Fig. 4.2.

We selected a set of drug and chemical databases based on their availability, popularity and comprehensiveness (Fig. 4.2a). We additionally selected a set of published drug–disease relationship datasets with two or more synonyms as our aim was to use KATdb in mapping drug names from truth-table name to drug repositioning method name. From each database listed in Materials and Methods Table 3.1 we extracted the database specific unique identifier, name, external identifiers, and other associated synonyms. Each extracted relationship between a unique identifier and a synonym was added to a database of edges (Fig. 4.2b). The edges were first curated and cleaned to only include synonyms related to drugs and chemicals. Where there was one primary human-readable name provided, the name was assigned to *Database-name Name* name type, e.g. DrugBank Name. Table A.1 lists all extracted and used types of synonyms.

We constructed a graph from the cleaned edge database, where each node represents an *authority:value* pair and each edge is a synonym relationship between two nodes (Fig. 4.2c). The synonym network was decomposed into connected components. A connected component, or simply component, is a subnetwork in which any two nodes are connected to each other but are not connected to any nodes outside the subnetwork. Under assumption that all extracted relationships were correct, we hypothesised that each connected component represented a set of synonyms for one drug. Each connected component included relationships between synonyms from one or more databases. We assigned a KATdb unique identifier to each component (Fig. 4.2d). A MySQL database was constructed and a KATdb visual interface was created with `shiny` (Chang et al., 2019).

4.4 Initial Database Overview

In this section we provide a brief overview of the initial KATdb, followed by a section exploring the database correctness. After correction assessment, we reconstructed KATdb

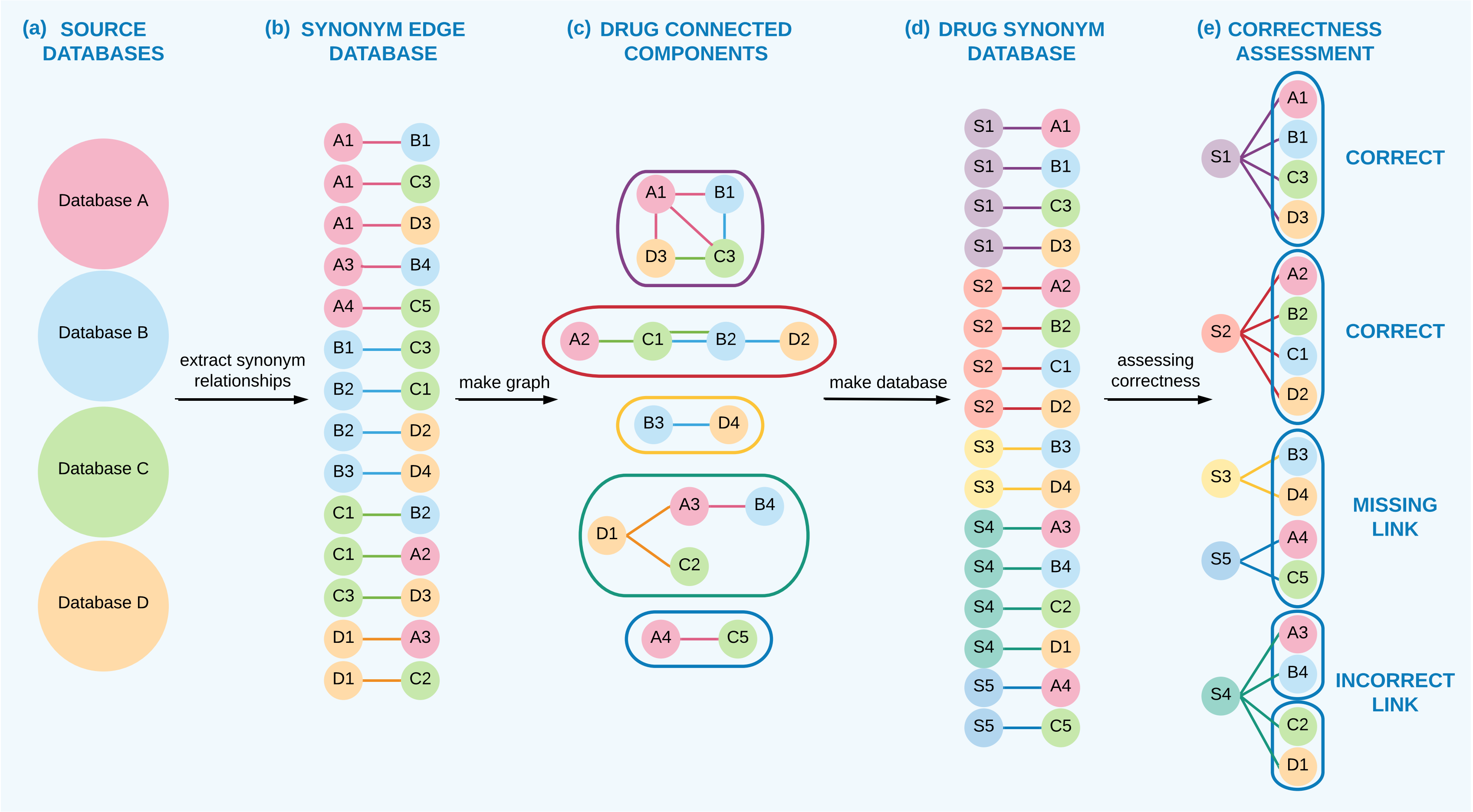


Fig. (4.2) KATdb, the drug synonym database method overview. (a) Relationships between drug synonyms are extracted from source databases. (b) The relationships are then used as edges and synonyms as nodes when constructing a network. (c) The network is decomposed into connected components which are then processed into (d) a drug synonym database where each connected component represents a group of synonyms connected to represent one drug. (e) We predict 3 major correctness evaluation outcomes.

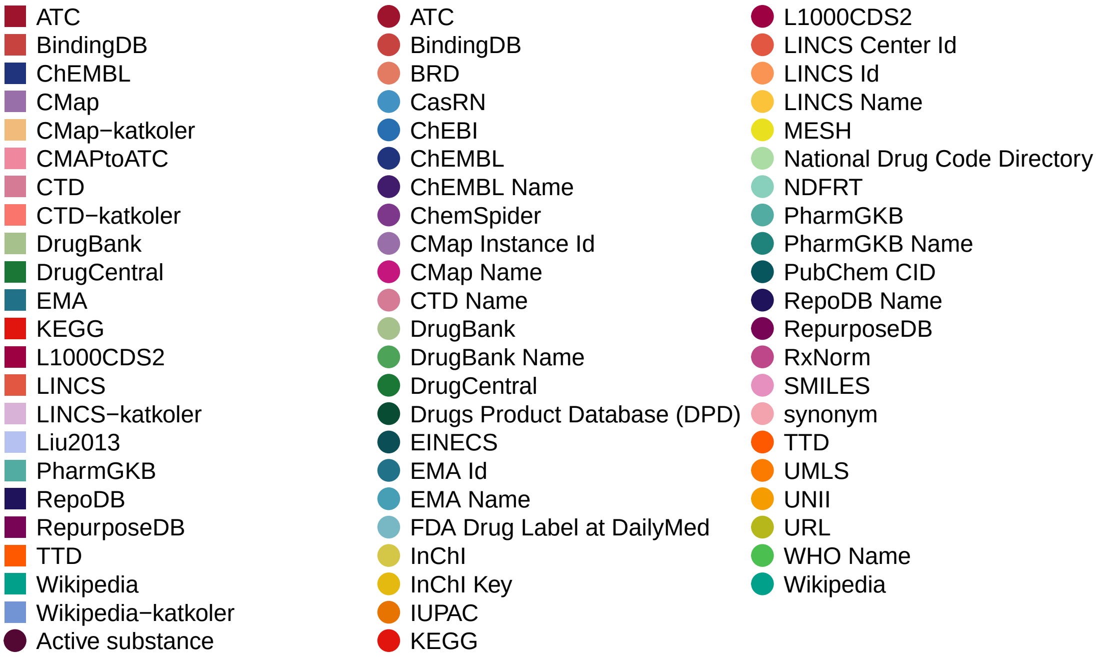


Fig. (4.3) KATdb list of databases and name types (colour legend). Colours in squares annotate source databases and colours in circles annotate name types. Manually curated relationships are annotated with "katkoler" next to the original source. Please refer to these colour annotations throughout this chapter. On Figs. 4.5, 4.7, 4.8, and 4.9 the squares annotate edges and circles annotate nodes. On Fig. 4.4 squares annotate the columns and circles the rows. On Fig. 4.10 the squares annotate the lines which represent the removed database. ATC — Anatomical Therapeutic Chemical; BRD — Broad ID; CasRN — CAS Registry Number; ChEBI — Chemical Entities of Biological Interest; ChEMBL — Chemicals database by European Molecular Biology Laboratory; CID — Compound ID; CMap — Connectivity Map; CTD — Comparative Toxicogenomics Database; DB — database; DPD — Drugs Product Database; EINECS — European Inventory of Existing Commercial Chemical Substances; EMA — European Medicines Agency; FDA — the US Food and Drug Administration; InChI — the IUPAC International Chemical Identifier; IUPAC — International Union of Pure and Applied Chemistry; KEGG — Kyoto Encyclopedia of Genes and Genomes; L1000CDS^2 — LINCS L1000 characteristic direction signature search engine; LINCS — Library of Integrated Network-based Cellular Signatures; MESH — Medical Subject Headings; NDFRT — National Drug File - Reference Terminology; PharmGKB — Pharmacogenomics Knowledge Base; SMILES — Simplified molecular-input line-entry system; TTD — Therapeutic Target Database; UMLS — Unified Medical Language System; UNII — Unique Ingredient Identifier; URL — Uniform Resource Locator; WHO — World Health Organisation.

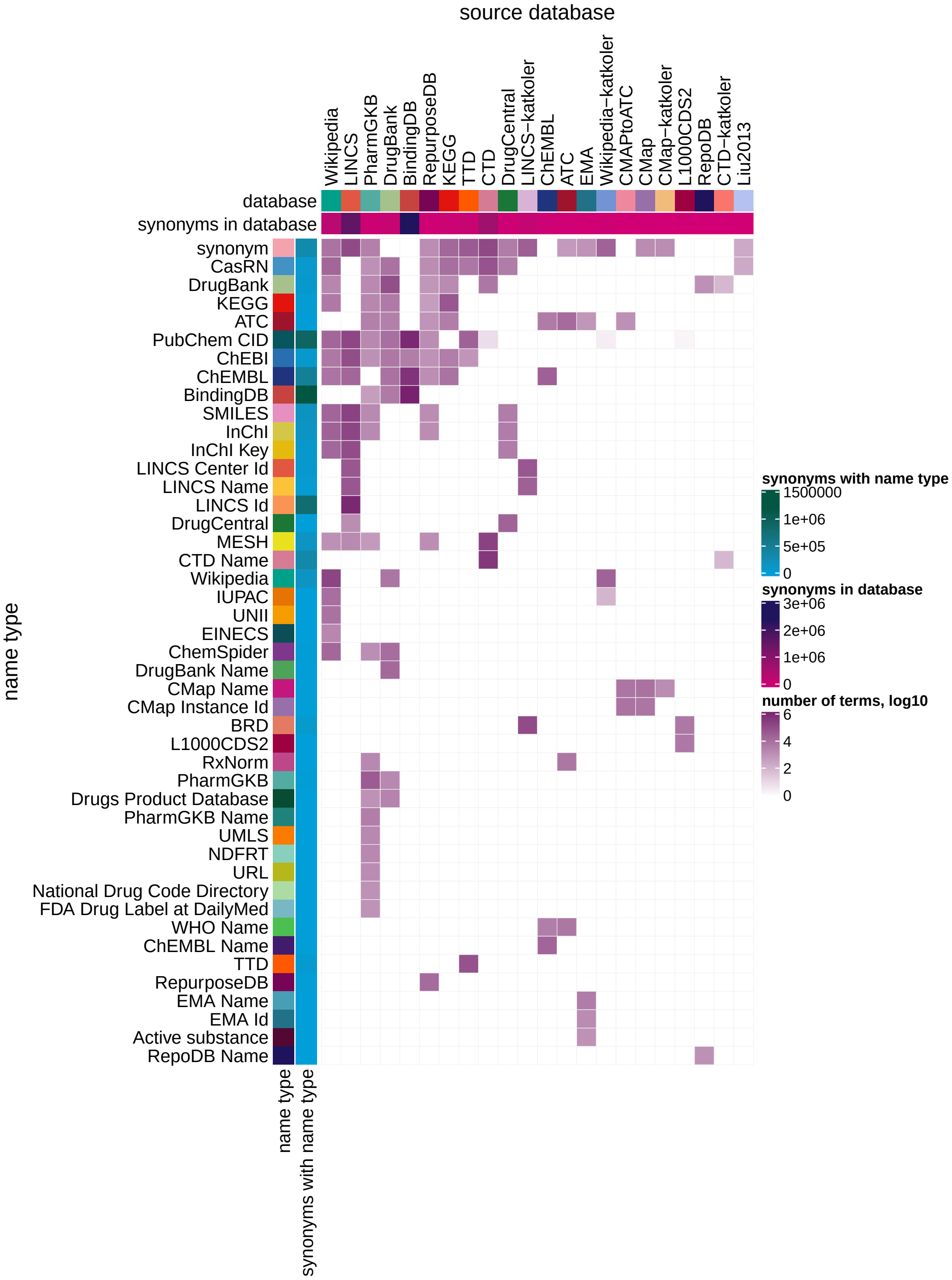


Fig. (4.4) Number of terms per name type contributed by each source database. Each database (columns) contributes a set of synonyms connecting different name types (rows). Some databases contributed many more synonyms than others. There are name types that appear in more databases. Manually curated relationships are annotated with "katkoler" next to the original source. Please refer to Fig. 4.3 (page 68) for the colour legend and acronyms for source database and name type annotations. The number of terms is $\log_{10}$ scaled. Each column is summarised in the top margin (number of synonyms in database), and each row is summarised in the left margin (number of synonyms with name type). The node and edge contributions to KATdb per database are listed in Supplementary Table B.2.

removing the most erroneous database. A more detailed overview of KATdb is discussed after the correctness section.

We successfully extracted over 2.8 million synonym relationships between 3.3 million different drug names and identifiers. KATdb initially consisted of synonyms extracted from 18 different databases representing 45 different synonym types. The source databases and synonym types are listed in Fig. 4.3.

Name types present in more than one database had the potential to establish cross-database connections, which increased the translation power of KATdb. The numbers of synonyms of a particular name type present in source databases are summarised in Fig. 4.4. Most databases have provided poorly-characterised synonyms, where the database did not include the information on what type of synonym the value was. These were commonly found in *synonym* sections, and predominantly included human-readable names. The second most abundant name type across databases was the PubChem Compound ID (CID), which is present in 10 different databases.

The most popular synonym type by number of terms was the BindingDB ID, present in only three databases: BindingDB, PharmGKB, and DrugBank. Another highly used name type was PubChem Compound ID (CID), present in 10 databases. The PubChem CID was one of the most frequent name types in KATdb, although we had not included PubChem as a source database, suggesting the community-recognised importance of the database. PubChem was the world's largest freely accessible chemistry database (Kim et al., 2016). However, due to computational limits we were unable to integrate the large PubChem Compound database into KATdb.

PharmGKB provided the highest variety of name types, connecting 19 different name types, followed by Wikipedia with 16 different name types. While PharmGKB provided the highest number of different name types, only 13 types were present in at least one other database. 13 out of 16 name types used in Wikipedia were used in at least 2 other databases.

Overall, BindingDB contributed the most relationships, followed by the Library of Integrated Network-based Cellular Signatures (LINCS) (Supplementary Table B.2). BindingDB included relationships from 4 different name types, the BindingDB identifier to PubChem CID, Chemical Entities of Biological Interest (ChEBI), and ChEMBL. LINCS contributed to connections between 12 different name types.

Liu et al. (2013) and RepoDB (Brown and Patel, 2017) contributed only relationships between *synonym↔CasRN*, and *RepoDB Name↔DrugBank*, respectively. Even though both added a small amount of connections to the overall KATdb, they could both be used as gold-standard true positive benchmarking datasets. Thus, it was important to be able to map the drug names from truth-tables to identifiers used in drug repositioning methods.

4.5 Correctness

To assess the reliability of our approach, we assessed how faithfully each component represented one drug. We predicted that there would be the following correctness outcomes (Fig. 4.2e):

1. one component representing one drug, synonyms connected by correct links,
2. one component representing many drugs, synonyms connected by incorrect links. The links could be incorrect to a different degree:
    (a) one component representing structurally similar drugs,
    (b) one component representing related drugs, e.g. a mixture and a single ingredient of a drug,
    (c) one component representing non-related drugs,
3. two or more components representing one drug as a result of missing links.

Each of the correctness outcomes had a different set of impacts on the overall correctness of the synonym graph. The correct links, unifying synonyms belonging to one drug, had no negative impact. The most severe repercussions happened when the incorrect links connected non-related drugs. Any mapping or translation from one synonym type to another would therefore be partially incorrect. Incorrect links connecting related drugs were less severe where one drug ingredient existing on its own also connected to other ingredients in a particular drug mixture. The mixture compounds likely complemented each other, but translation of single compound identifiers would also erroneously map to other mixture components.

The links connecting structurally similar drugs could be considered ambiguous. We classed spatial isomers and oxidised or reduced versions of one compound as structurally similar. Active and inactive forms of chemicals might be structurally distinct, but they commonly appear together. Spatial isomerism or stereoisomerism is when molecules have the same molecular formula and the same sequence of bonded atoms, but differ in 3D orientation of their atoms in space (McNaught et al., 1997). An enantiomer is one of two molecules which are mirror images of each other (McNaught et al., 1997), also known as optical isomers. Although stereoisomers in mixtures often differ in properties, they appear together, e.g. thalidomide. Approximately half of marketed drugs are chiral, and of these, approximately half are mixtures of enantiomers rather than single enantiomers (Hutt, 2002; McConathy and Owens, 2003). Thus, we acknowledge that although structurally similar drugs that are connected into one component would be technically incorrect, we considered that mapping synonyms to a stereoisomer would be less detrimental than mapping to an unrelated drug.

The last predicted correctness outcome was the missing links. While they were not desirable, as they represent the lack of information, they did not carry many negative effects. Mapping would still be correct, even though a particular search for synonyms for one drug might reference two or more connected components.

4.5.1 Correctness estimation

To evaluate the correctness of KATdb, we manually investigated a set of connections. We investigated edges that we predicted as topologically important and thus more likely to influence the correctness of downstream reasoning.

We chose to investigate the edges in the largest components as these components included more synonyms than we expected. We estimated that there are $\sim$100 synonyms for one drug, with at least one for each name type (45) and an additional 45 to roughly account for non-unique identifiers. Many name types mapped to more than one synonym per drug. Considering our estimate, it was unlikely that one drug could be identified by more than 600 synonyms.

The second parameter for prioritisation of edges related to the topological role of a given edge. *Edge betweenness centrality* is defined as the number of shortest paths that go through an edge in a graph (Girvan and Newman, 2002). An edge with a high edge betweenness centrality score is represented by a bridge-like connector between two parts of a graph. These edges are important links that connect two well-connected parts of the connected component. The removal of such edge would affect the communication between many pairs of nodes in the two parts of a graph (Lu and Zhang, 2013).

The largest component (Fig. 4.5) displayed both of these concepts. It consisted of 609 nodes connected with 1043 edges. The edges with high edge betweenness centrality often provided the only bridge between one part of the component to another. Several edges

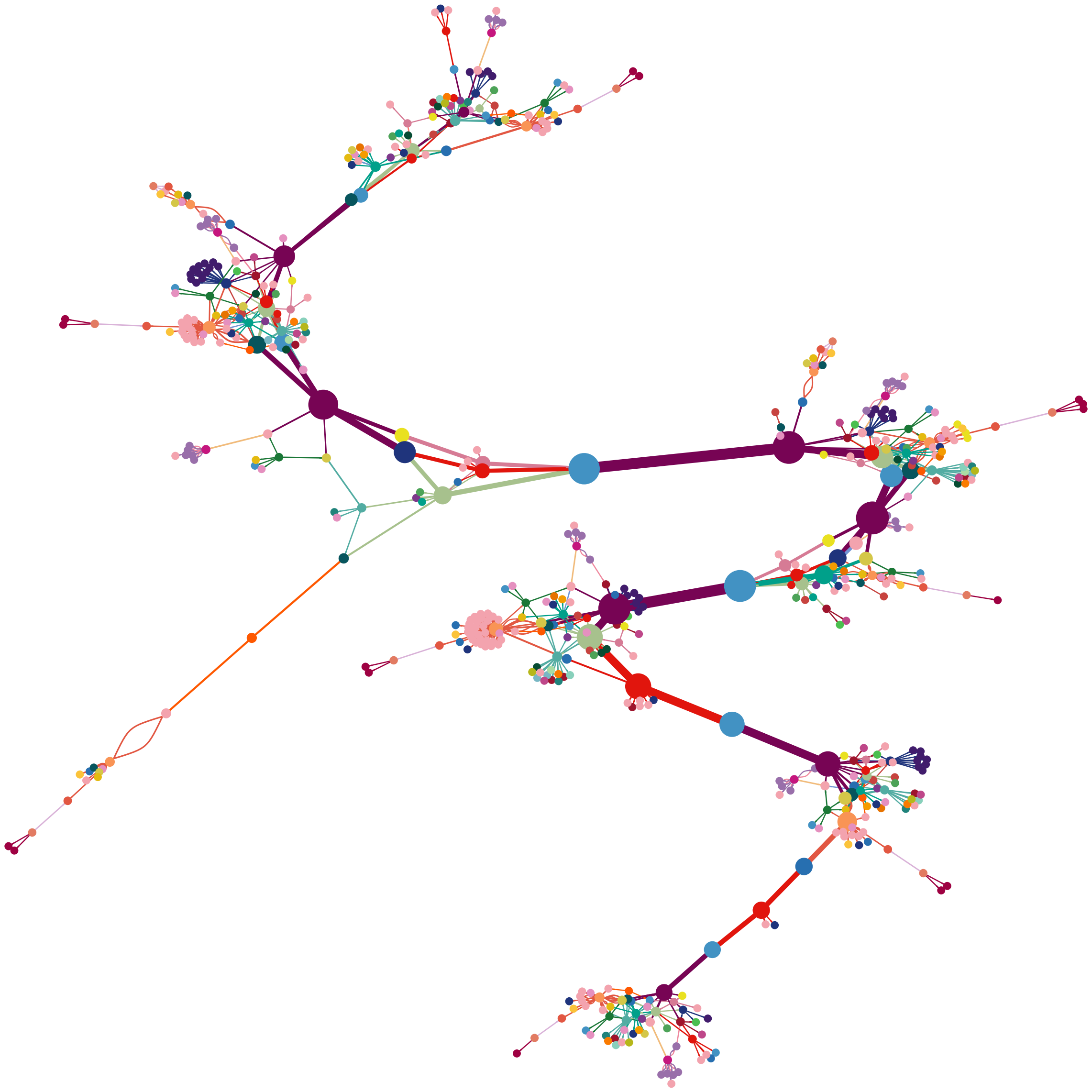


Fig. (4.5) The largest connected component prior to assessing correctness. There are 609 nodes and 1043 edges in the component. The node size and edge width are proportional to the node and edge betweenness centrality. The key connector edges between different well-connected parts of the component are thicker, indicating the topological importance of that edge as well as the increased likelihood of that edge to be incorrect. A node represents one drug name or identifier and an edge represents the direct synonym relationship between two nodes extracted from a source database. The node colour signifies the name type and the edge colour matches the source database. Please refer to Fig. 4.3 (page 68) for the colour legend for edges (square — source database) and nodes (circle — name type).

showed disproportionately high edge betweenness centrality (thicker edges on Fig. 4.5). For instance, there were 7 edges from RepurposeDB with high edge betweenness centrality (thick purple edges on Fig. 4.5) that were key for the overall connectivity of this component. It was important to the overall component structure that these edges were correct. However, we predicted that there were several incorrect edges in this component, particularly the edges with a high edge betweenness centrality score.

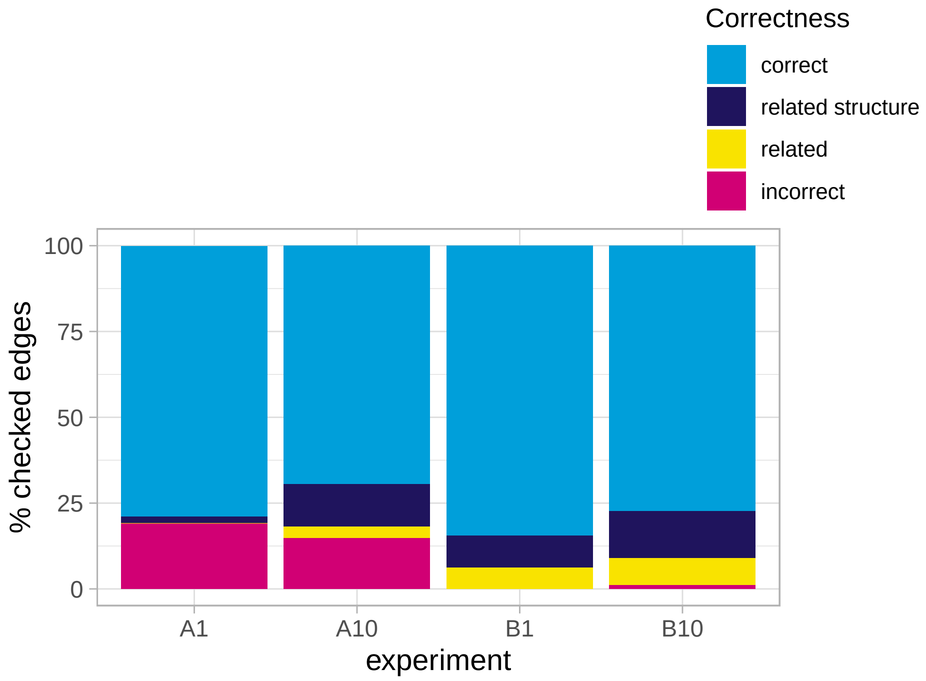
To systematically assess the correctness of the graph, we incorporated both component size and edge betweenness centrality into our test. We first selected edges in the top 10 largest connected components and then prioritised the edges with the highest edge betweenness. The top 2% of edges with the highest edge betweenness centrality in the top 10 largest connected components were manually checked. While a source database can be a key source of providing missing links, it can at the same time be the most detrimental to component correctness if those relationships are in fact incorrect.

Four different iterations were considered in estimating correctness. First, we assessed the largest and the 10 largest components from the initial KATdb, referred to as A1 and A10, respectively. Followed by the second assessment of the largest and the 10 largest components after removing the source database with the highest proportion of incorrect edges, referred to as B1 and B10, respectively.

Each node of the manually checked edges was searched on Google (*https://www.google.com/*) to manually establish the connection to the other node in at least two databases (excluding the source edge database). We assigned one of the following values to the level of correctness for each checked edge:

(i) correct
(ii) related structure
(iii) related
(iv) incorrect

An edge was marked as *correct* if the relationship between synonyms was established in two other source databases. Spatial isomers and small structural changes were marked as *related structures*. Edges connecting individual ingredients of the same mixture were marked as *related*. Edges connecting non-related drugs were marked as *incorrect*.



	A1	A10	B1	B10
DB removed	n/a	n/a	RepurposeDB	RepurposeDB
components	1	10	1	10
nodes	609	3787	474	2900
edges	1043	6001	722	4391
checked	52 (4.99%)	121 (2.02%)	32 (4.43%)	88 (2%)

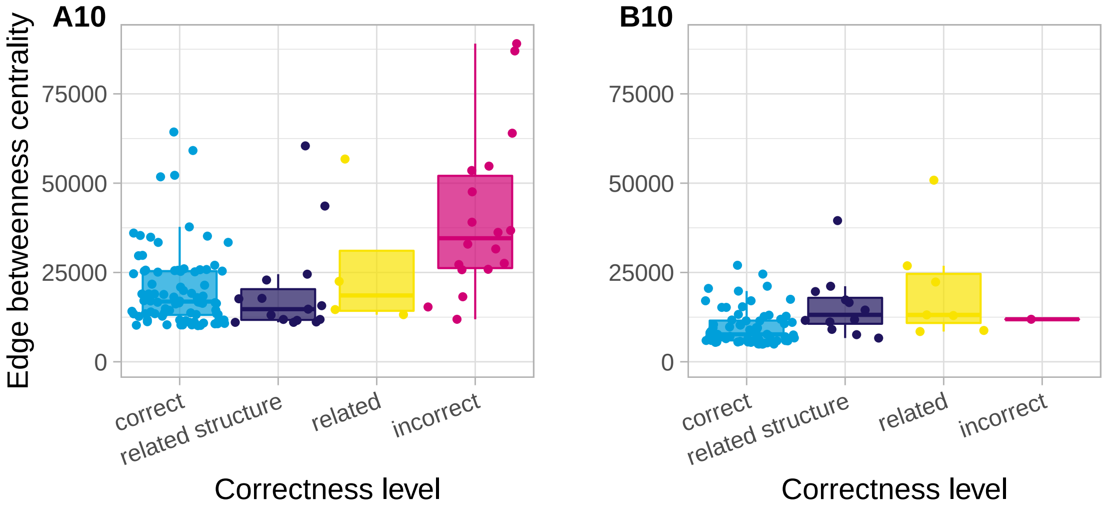


Fig. (4.6) KATdb correctness assessment. Four different iterations were considered in estimating correctness. A1 and A10 assess the largest and the 10 largest components from the initial KATdb, respectively. B1 and B10 assess the largest and the 10 largest components, respectively, after removing RepurposeDB as a source database. We assigned one of 4 different levels of correctness to each manually checked synonym relationship. The top plot summaries the proportion of each correctness level in an experiment. The table (middle) summarises the experiment properties. The two plots (bottom row) show the edge betweenness values for a manually checked edge with assigned correctness level.

The initial assessment

The top 10 largest components ranged from 240 (10^{th} largest) to 609 (the largest) nodes, with a total of 6001 edges. We manually investigated 121 (2%) edges with the highest betweenness centrality (Fig. 4.6 – A10). We identified 84 correct, 15 structurally related, 4 related, and 18 incorrect edges. Estimating $> 69.4\%$ of correct edges at first iteration of KATdb and $< 14.9\%$ of incorrect edges. The incorrectly linked components were driven by errors in the source databases, which were propagated in the KATdb.

Out of the 18 incorrect edges, 10 were found in the largest component and all 10 were from one database, RepurposeDB[1] (Supplementary Fig. B.1). The largest component highlighted a systematic fault in 11 consecutive rows of RepurposeDB (rows 520 (betazole) – 531 (naproxen)), where the CasRN identifiers were the correct identifiers for the drug listed in the row above. An additional small erroneous section of the RepurposeDB table was identified (rows 117 (indometacin) – 120 (prednisolone)), where the KEGG IDs were correct for the drug name in the row above. RepurposeDB was removed as a consequence of contributing a high proportion of incorrect edges, however it is possible that the systematic faults were limited to the identified sections and that the remaining relationships were correct. This highlights the strength of using edge betweenness centrality as a measure of correctness for individual edges, however a more systematic curation would be required to extrapolate the limited findings from faulty RepurposeDB relationships to the remainder of the database. We re-evaluated the largest component by removing all RepurposeDB edges (Fig. 4.7). The component separated into 13 smaller components, from which, 5 were smaller than the rest. The size of the resulting largest component was 107 nodes, with no edges that had disproportionately high edge betweenness centrality scores. While we have reduced the proportion of incorrect edges to $\sim 0\%$, we have also removed 14 at least correct links. The relatively small components were likely connected to the larger group of synonyms through one of the correct RepurposeDB edges, that are now missing links.

In the top 10 largest components of the initial KATdb, 16 incorrect edges were extracted from RepurposeDB, 1 from TTD and 1 from LINCS. We removed RepurposeDB as a source database in the next iteration of the database, because it contributed 89% of the detected incorrect edges. We repeated the test after removing all edges from RepurposeDB (Fig. 4.6 – B1, B10).

[1] RepurposeDB (Shameer et al., 2017) is no longer available online (27^{th} May 2020)

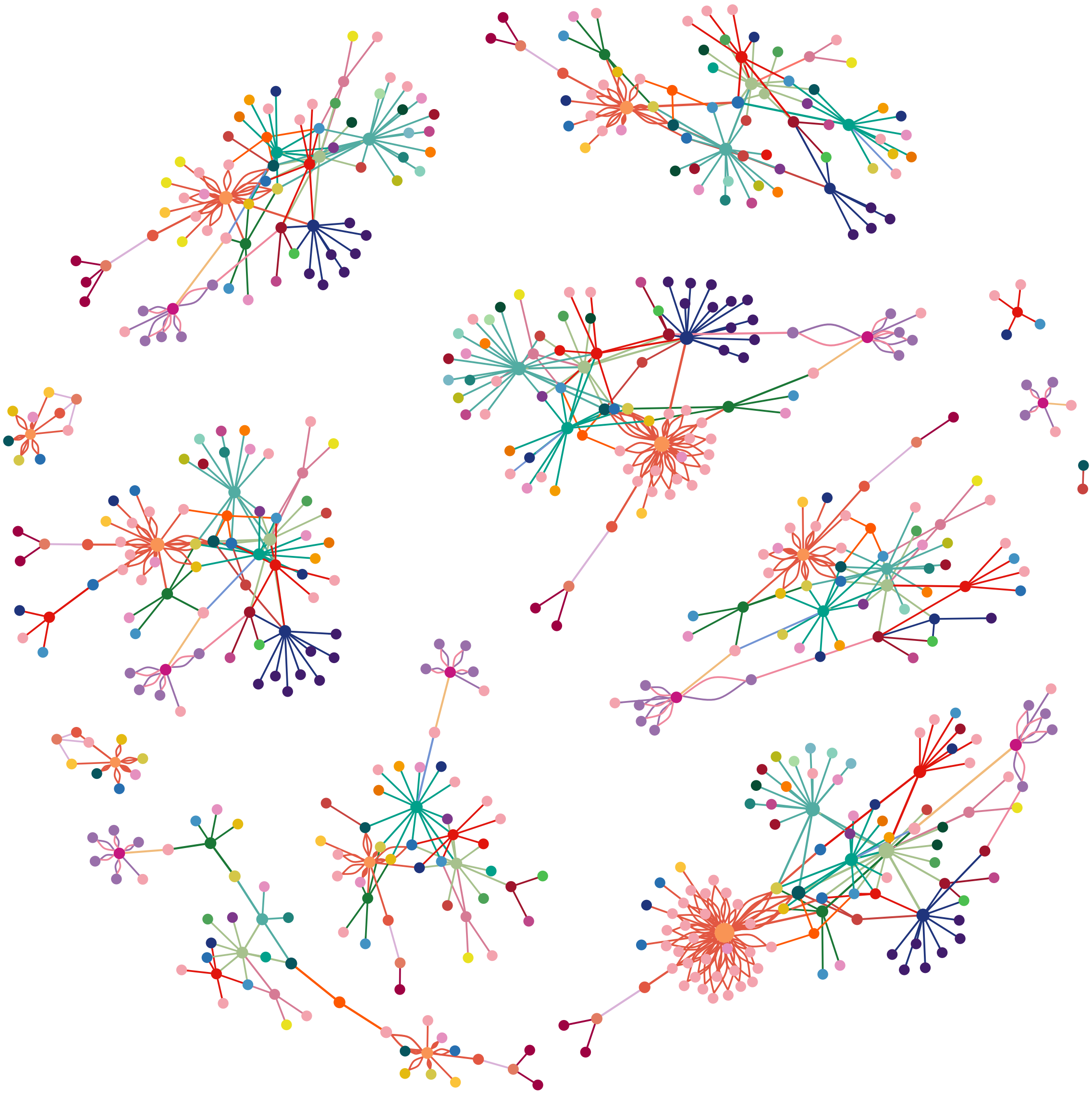


Fig. (4.7) The largest component after removing RepurposeDB. The component (Fig. 4.5) split into 13 separate components. 5 of those were relatively small compared to the remaining 8. After removing all RepurposeDB edges, there were no assessed incorrect edges remaining. A node represents one drug name or identifier and an edge represents the direct synonym relationship between two nodes extracted from a source database. The node and edge width is proportional to node and edge betweenness. Please refer to Fig. 4.3 (page 68) for the colour legend for edges (square - source database) and nodes (circle - name type).

The second assessment

After RepurposeDB removal, we identified 68 correct, 12 structurally related, 7 related, and 1 incorrect edge in the top 2% of edges with the highest edge betweenness centrality score from the newly defined 10 largest components. Estimating $> 77.3\%$ of correct and $< 1.1\%$ of incorrect edges overall. The new largest component consisted of 474 nodes and the 10th largest consisted of 219 nodes.

The top 10 components were still larger than expected. While we have reduced the number of *incorrect* edges, we still retained $\sim$8.0% of edges connecting related drugs and $\sim$13.6% of edges linking structurally related drugs. These edges have been extracted from several different source databases: *related* from 4 and *structurally related* from 7. If we removed all of the source databases with errors, we would significantly decrease the size and connectivity of KATdb. Instead we indicated the edge correctness with the edge betweenness centrality score.

Edge betweenness centrality as correctness predictor

After manually checking $\sim$200 edges, we assessed our hypothesis that the high edge betweenness centrality score would indicate a higher likelihood that the edge was incorrect (Fig. 4.6 bottom row). We assessed each version of KATdb separately as the edge betweenness centrality score was influenced by the topological features of the network. We showed that edges connecting random drug synonyms had a higher edge betweenness scores than those connecting same, structurally related or broadly related drugs (t-test p-value$= 0.00166$, Fig. 4.6 bottom row – A10). In addition, the links connecting loosely related drugs had a higher median compared to those that connected structurally related synonyms in both iterations, but their scores were not significantly different (A10 – p-value=0.564, B10 – p-value$= 0.450$). After removing RepurposeDB edges (B10), the correct relationships had a significantly lower edge betweenness score compared to any other assessed edge (t-test p-value$= 0.00846$).

By removing RepurposeDB we have increased the estimated correctness from 69.4% to 77.3% and more importantly, we reduced the estimated proportion of incorrect edges from 14.9% to 1.1%. The largest components were still larger than expected, thus it was important to use translated terms while acknowledging their limitations by investigating their network properties, such as edge betweenness centrality. We have demonstrated that edges with a high betweenness centrality score were more likely to be incorrect. Therefore, we could hypothesise that our correctness estimate was in reality even higher, $> 77.3\%$, and that there were $< 1.1\%$ incorrect, $< 8.0\%$ related, and $< 13.6\%$ structurally related edges.

4.6 Database Overview

After assessing correctness on the initial KATdb, we have found a disproportionate amount of incorrect edges extracted from RepurposeDB. To be able to use KATdb with more confidence, we have removed all edges from RepurposeDB, resulting in KATdb consisting of 2.85 million edges from 17 different databases between 3.31 million nodes (Table 4.1). The database was separated into 983,560 connected components, with each connected component hypothesised to represent synonyms for one drug.

The node with the highest degree was CMap Name:trichostatin A, connected to another 370 synonyms (Fig. 4.8, central node in ball-like component). 364 of those were CMap Instance IDs and the remaining 6 were synonyms. Searching "trichostatin A" in the database, resulted in 3 components: 2 large, of those one was CMap-themed, one from mixed sources, and 1 component of only 2 nodes from PharmGKB (Fig. 4.8).

Overall, 67.5% of the nodes were connected to only one other node, contributing non-essential relationships for overall connectivity. 39.1% of all components represented the smallest by definition. They included only two synonym nodes. 93.3% of all KATdb components had the diameter ≤ 2. However, only a few specific topological shapes have the diameter ≤ 2:

- a connection between two nodes: V1$\leftrightarrow$V2, Fig. 4.8 – small two-node component,
- linear connection between three nodes: V1$\leftrightarrow$V2$\leftrightarrow$V3,
- connected triangle between three nodes: same as above with V1$\leftrightarrow$V3, $\triangle$,
- ring or more connected square between 4 nodes: $\square$, $\boxtimes$,
- ring or more connected pentagon between 5 nodes: ⬠,
- star-like components with one central node: $+$, $\star$, $\ast$, Fig. 4.8 – bottom, ball-like component.
- star-like components with one central node and one or more extra edges between star rays: $\bowtie$, $\ltimes$, $\rtimes$, $\ast\!|$.

These components did not add much information as they connected two synonyms with no or one intermediate synonym. They represented poorly connected and/or seldom used

Table (4.1) KATdb statistical overview. Statistics including general graph, node and connected component properties for the drug synonym graph.

Property Type	Property	Value
graph	Total number of edges	2851870
graph	Total number of sources	17
graph	Total number of nodes	3305952
graph	Total number of name types	44
graph	Total number of connected components	983560
node	Highest degree node	370 (CMap Name:trichostatin A)
node	Median degree of a node	1
node	% degree == 1	67.5%
component	Largest component size	474
component	Median size	3
component	Mean size	3.36
component	% size == 2	39.1%
component	Max diameter	24
component	% diameter <= 2	93.3%
component	Mean shortest path	3.01

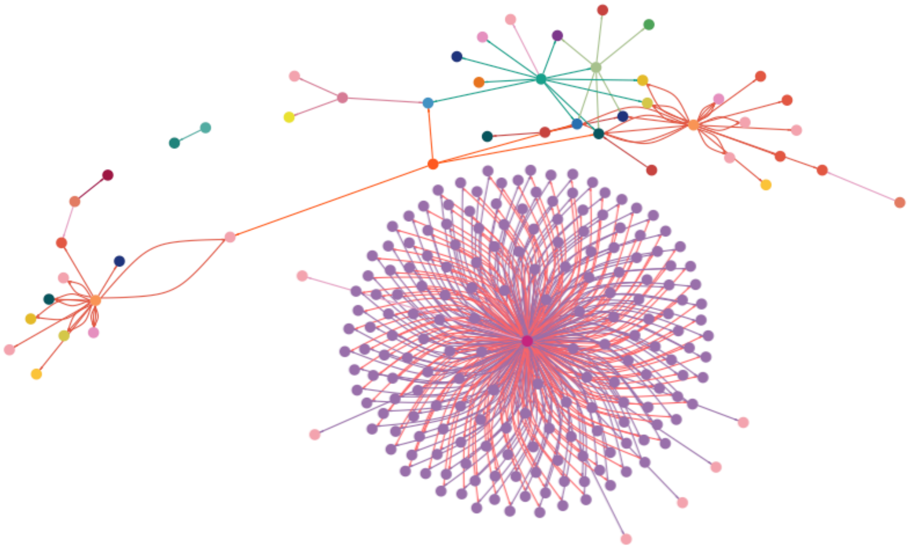


Fig. (4.8) Trichostatin A in KATdb. Three connected components represented trichostatin A in KATdb. The CMap Name node in the centre of the ball-like component was the node with the highest degree in the whole database. There were 3, rather than only 1, separate components because of missing links failing to connect all the synonyms. A node represents one drug name or identifier and an edge represents the direct synonym relationship between two nodes extracted from a source database. Please refer to Fig. 4.3 (page 68) for the colour legend for edges (square - source database) and nodes (circle - name type).

synonyms. However, if additional databases were added, they would have the potential to connect to bigger components. A major reason for the high proportion of components with small diameter was that we do not connect names based on their human-readable names, unless those names were well-defined e.g. INN or Wikipedia names. Therefore, most connections between different source databases were established through alphanumeric identifiers.

The remaining 6.7% of components with diameter > 2 connected at least two source databases. The widest component with diameter $= 24$ was also the largest component with 474 nodes (Fig. 4.9). During correctness estimation, we checked 32 of the 722 edges in the widest component with the highest betweenness score (highlighted in Fig. 4.9). It could be seen that the largest component represented at least 2 related drugs, as the removal of the *related* (Fig. 4.9 yellow highlight) edge would have split the component in two. The component represented a drug called *conjugated estrogens* that:

“contains a mix of estrogen *from which about 50% is represented by* estrone sulfate *followed by 25% of* equilin sulfate*, 15% of* 17-alpha-dehydroequilenin sulfate*, 3% of* equilenin sulfate*, 5% of* 17-alpha *and* 17-beta-dihydroequilenin sulfate*, 2% of* 17-alpha-estradiolsulfate *and 3% of* 17-beta-estradiolsulfate*. It also presents a large number of* unidentified molecules *with weak estrogenic activity as well as* non-human molecules *when it is obtained from pregnant mares’ urine.”* (DrugBank, 2005; Lauritzen and Studd, 2005)

Due to the compounds appearing in a complex oestrogen mixture, the largest component represented a set of chemically related compounds, rather than one single one. All checked edges link to a human-readable name that includes "estro/a" as the root of the name. It represented one drug, however, that drug was a mixture of many individual but related compounds. We have not investigated if each of the chemicals was used as an individual compound. However, analysing the component, there were 160 synonyms, 48 ChEMBL Names, 41 CMap Instance IDs, 16 SMILES, 11 PubChem CIDs, 8 Wikipedia pages, 7 InChI Keys, 5 DrugBank IDs, etc. To estimate how many *different* compounds were in the component, we could consider the frequency of the name types. The median frequency in this component was 5.5, however, there were many non-unique identifiers, such as synonyms or CMap Instance IDs, that might have skewed this estimate.

To resolve the largest component into single compounds would require extensive curation of all edges and detailed knowledge of the chemical structures. With decreasing component size the probability of erroneous edges decreased. Considering that only 235 (0.024%) out of $\sim$1 million components consisted of more than 100 nodes, we

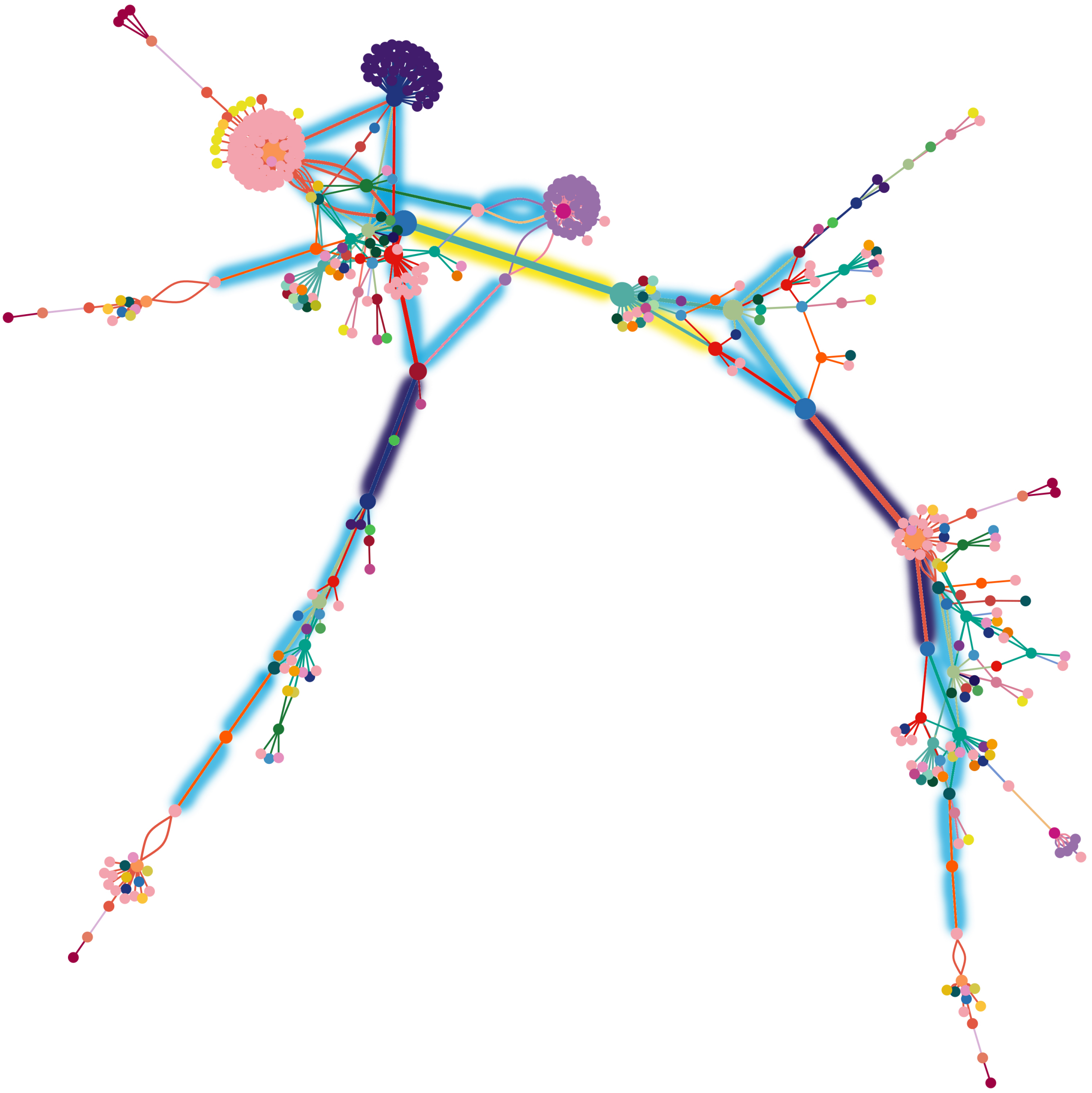


Fig. (4.9) KATdb widest and largest component - conjugated estrogens. It consisted of 474 synonyms and had the diameter of 24. It included at least 2 edges connecting related (yellow) and 3 connecting structurally related synonyms (dark blue). A node represents one drug name or identifier and an edge represents the direct synonym relationship between two nodes extracted from a source database. Please refer to Fig. 4.3 (page 68) for the colour legend for edges (square - source database) and nodes (circle - name type) and to Fig. 4.6 (page 75) for the colour legend on edge correctness level highlights (blue — correct; dark blue — structurally related; yellow — related; none — not manually assessed for correctness).

proceeded with the investigation of KATdb properties, acknowledging the currently defined limitations.

4.7 Robustness

The KATdb network was tested for robustness to assess its ability to cope with errors and perturbations. Robustness assesses the ability to maintain connectivity after deletion of nodes. The connectivity of the resulting network was measured by the size of the largest connected component (Lordan and Albareda-Sambola, 2019). In addition to the size of the largest component, we measured the mean component size and number of components. We tested KATdb by removing a random subset of nodes 5% at a time for 10 iterations. We performed the test on the whole network and also on the network with one source database removed at the time.

The largest component from the whole KATdb network drastically decreased in size when we started removing random nodes (Fig. 4.10A). It reduced to less than half the size with 25% of the nodes removed. It was approximately one fifth of the size when removing 50% of the nodes and one tenth with 75% of nodes removed. Fig. 4.10A shows that removing the majority of source databases had a small effect on the overall structure of the network, suggesting that the edges extracted from those databases were supported by other edges. There were three databases that changed the resilience of the network: LINCS, PharmGKB and ChEMBL. After removing edges from LINCS, the largest component reduced from 474 to 265 nodes. This indicated that LINCS provided edges that connect several smaller components to a bigger component. It suggested that LINCS is a key contributor to overall connectivity of the synonym components. It also indicated that there may be several incorrect edges contributed by LINCS. The other two databases that had a large effect on the size of the largest component were PharmGKB with 301 nodes in the largest component and ChEMBL with 372 nodes. Similar assumptions could be made for these as for LINCS contributions. The correctness evaluation (Supplementary Fig. B.1 – B10) supported the idea that these databases may have had a higher proportion of incorrect, related and structurally related edges. When estimating correctness for the top 10 largest components by manually checking top 2% of edges with the highest edge betweenness centrality, all three were found with some related and/or structurally related edges. PharmGKB and LINCS had a high, but similar proportion of related and structurally related edges (43% and 31%, respectively), and the only two ChEMBL edges checked were both connecting structurally related synonyms.

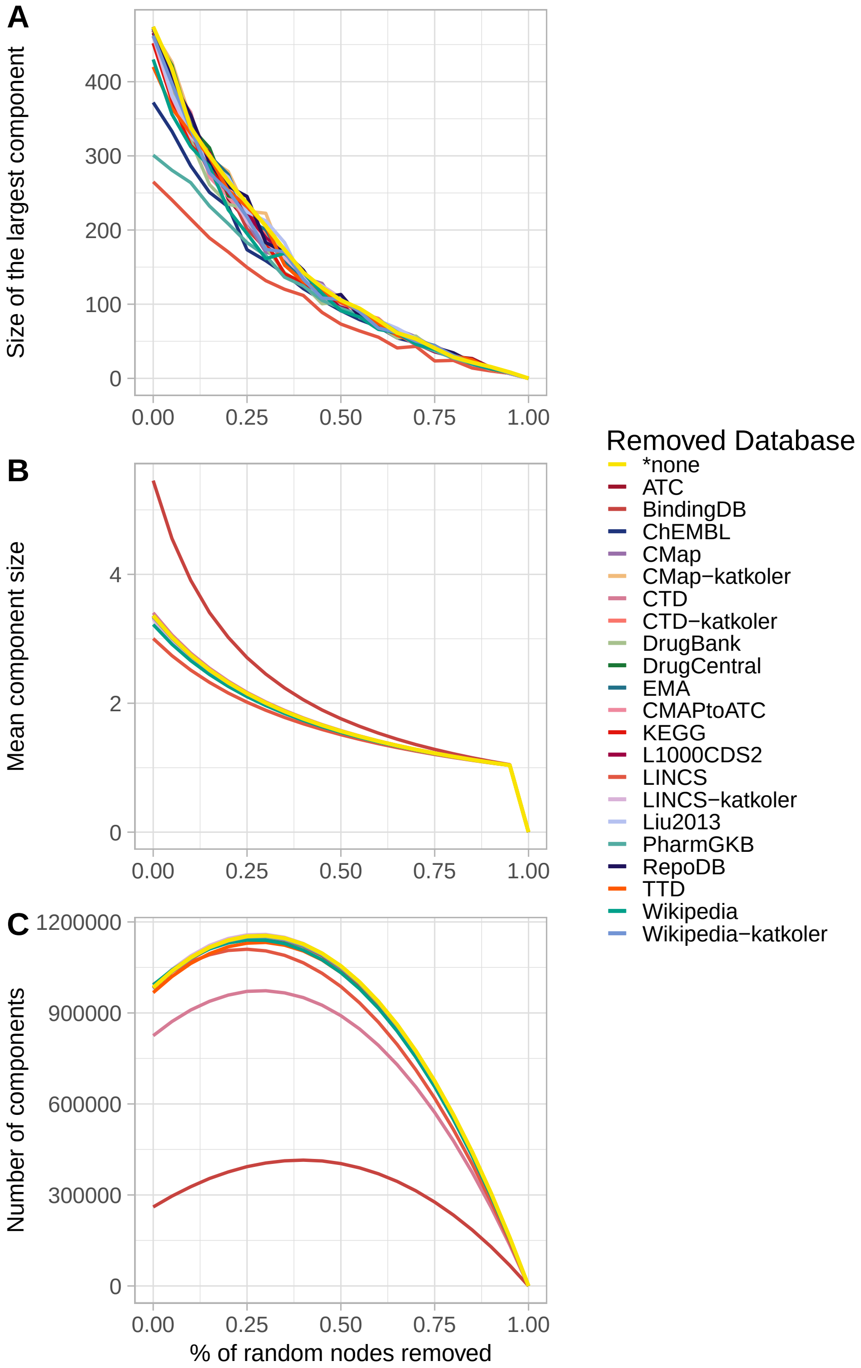


Fig. (4.10) KATdb robustness. The KATdb network robustness was tested by removing a random subset of nodes, 5% at the time. We explored the influence of each source database on the KATdb synonym network by removing one source database at a time, followed by the removal of a random subset of nodes. Random nodes were removed to assess the resulting graphs' robustness. Robustness was assessed by (A) the mean size of the largest component, (B) the mean of the mean component size, (C) the mean number of components. *none — no database removed apart from RepurposeDB that was not included in the final version of KATdb. ATC — Anatomical Therapeutic Chemical; ChEMBL — Chemicals database by European Molecular Biology Laboratory; CMap — Connectivity Map; CTD — Comparative Toxicogenomics Database; EMA — European Medicines Agency; KEGG — Kyoto Encyclopedia of Genes and Genomes; L1000CDS^2 — L1000 characteristic direction signature search engine; LINCS — Library of Integrated Network-based Cellular Signatures; PharmGKB — Pharmacogenomics Knowledge Base; TTD — Therapeutic Target Database.

We investigated the mean component size and the number of components in our robustness test, because a large proportion of KATdb components were relatively small. Removal of most databases did not alter mean component size decrease compared to the whole network (Fig. 4.10B). Similar to changes observed in the largest component size, removing LINCS resulted in a smaller mean component size of 3.00, again suggesting that LINCS provided key connections linking smaller components. However, the largest difference was observed when removing BindingDB. Removing BindingDB increased the mean component size from 3.36 to 5.45, suggesting that BindingDB contributed many small components that were not connecting to other databases. BindingDB contributed 41% of all edges and included 58% of all nodes from KATdb, covering names from only 4 different name types. Therefore, it was expected that many of those edges were not connecting to the rest of the network. This was supported by Fig. 4.10C, where we observed that removing BindingDB reduced the number of components from 983,560 to 260,360 (26.5%). The drastic decrease in number of components, together with increased average when removing BindingDB further supported that BindingDB primarily added small components. The second largest decrease in component numbers was observed by removing CTD, where the number of components decreased to 825,061 (83.9%). CTD contributed 12% of all edges, thus it was expected that removing so many edges would have reduced the number of overall components. Removing ∼30-35% of nodes increased the number of components for each test run. After removing more than ∼35% the number of components then gradually decreased to 0. This behaviour suggested that initially larger components were split into smaller ones, however, by removing more than 35% of nodes, we started removing whole components and thus the number of components decreased.

The robustness analysis has indicated different types of contributions from some databases. BindingDB has been shown to contribute mostly small connected components, while LINCS, PharmGKB and ChEMBL contributed important, but possibly also incorrect, links to the largest components. Overall, the removal of individual databases mimicked the robustness of the whole network. BindingDB which contributed approximately half of KATdb terms, could be considered an exception, as it mimicked the whole network when investigating the size of the largest component, but its removal increased the mean component size and decreased the number of components.

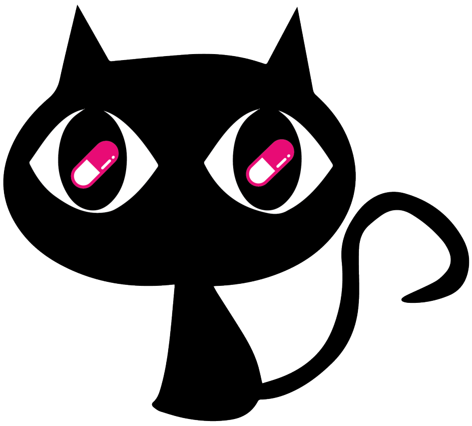


Fig. (4.11) The KATdb logo.

4.8 KATdb Visual Interface

A prototype visual interface was developed with `shiny` (Chang et al., 2019) to allow exploration of KATdb's components and the relationships connected into one synonym entity as well as mapping from one synonym authority to another. As part of the brand identity development, KATdb was associated with a cat-themed logo (Fig. 4.11). The minimal viable product consists of:

(i) welcome page with summary statistic of KATdb (Supplementary Fig. B.2),

(ii) translate your drug list page (Supplementary Fig. B.3),

(iii) explore components page (Supplementary Fig. B.4),

(iv) table of all nodes,

(v) table of all edges.

The welcome page (Supplementary Fig. B.2) offers an overview of the current KATdb version. It briefly introduces the aims of KATdb and lists the key summary statistics, such as: the number of nodes, edges and connected components as well as the number of different sources and name types.

Translate your list page (Supplementary Fig. B.3) can be used to retrieve drug synonyms from a user-defined list of drug names or identifiers. It allows translating from one or a mix of many authorities to all or a selected list of target authorities. An example translation from WHO or ChEMBL names for *aspirin* and *paracetamol* translated to

DrugBank identifier is pre-prepared to demonstrate one possible search. The user has the option of exploring the translated list within the app or downloading it as a CSV file (Shafranovich, 2005).

Explore components page (Supplementary Fig. B.4) plots a selected number of components. The components are ordered by the decreasing number of nodes. The user can choose to plot a window of up to 100 components decreasing in size at one time or provide a list of KATdb identifiers from tables on *translate your list page* or *tables of nodes and edges*. The component structure can be further explored by interactively removing edges and nodes from a particular source database or name type, respectively. The edge and node size on a zoomed-in plot correspond to edge and node betweenness centrality, allowing the user to estimate overall correctness of the component.

Tables of nodes and edges pages provide complete lists of nodes and edges currently present in KATdb. Table of all nodes lists the component KATdb ID that the node is part of, name type, name and degree for each node. Table of all edges lists the KATdb ID that the edge is present in, the two nodes the edge is connecting and the source database from which the edge has been extracted.

The KATdb shiny app has been used throughout this thesis. It has been used in the following sections: 4.9, 5.3.5, 6.4.1, 7.2.4, 7.3.4.

4.9 Application Test Cases

We have applied KATdb to three different test cases, related to work described in this thesis. In the following chapters, we used the LINCS L1000 characteristic direction signatures search engine (L1000CDS^2) (Duan et al., 2016) as the source database for drug perturbation signatures. We thus focused on unifying the nomenclature used in L1000CDS^2, the BRD IDs, in three different use cases:

(i) BRD ID to any name,

(ii) BRD ID to ATC code,

(iii) approved drug names to BRD ID.

The LINCS nomenclature. BRD IDs are the primary identifiers in LINCS (Subramanian et al., 2017) and L1000CDS^2 (Duan et al., 2016). BRD ID, or Broad ID, is an identifier used to uniquely address a particular small molecule. It was developed by the

Broad Institute and consists of 13 characters, "BRD-" followed by 9 characters, to uniquely identify the physical batch of the chemical, e.g. BRD-K11433652 for aspirin. In the L1000CDS^2 metadata the BRD IDs are listed as perturbagen ID (*pert_id*) and associated with perturbagen description (*pert_desc*) which is either "-666" for missing/not known names or a more common, often human-readable name. In this section, we have referred to it as the LINCS name.

Translation measures. In each test case, we used the translation feature of KATdb visual interface, for a comparison we performed manual mapping in 3 test cases. We used the following nomenclature: *Input name* is the name that was used as query name, the number of *found names* is the number of input names that were found in KATdb, the *goal name* is the name to which we were aiming to map the input name. We classed all found names that have at least one goal name as *mapped terms*.

To determine translation success and redundancy, we counted the number of unique input, found, and goal names, different connected components, as well as the number of mapped and not mapped terms. We defined *translation success* as:

$$\text{translation success} = \frac{\text{number of mapped terms}}{\text{number of found names}} \tag{4.1}$$

measuring what proportion of input names present in KATdb were mapped to at least one synonym, and *translation redundancy* as:

$$\text{translation redundancy} = \frac{\text{number of unique goal names}}{\text{number of mapped terms}} \tag{4.2}$$

measuring how many successfully mapped names, mapped to more than one goal name.

Translation success score ranges from 0–1. Translation success = 1 indicated that all input names present in KATdb have been mapped to at least one goal synonym and < 1 meant that not all input names found in KATdb have been mapped to the goal name type.

Translation redundancy = 1 indicates 1:1 mapping, > 1 indicated that more than one goal name per mapped name and < 1 suggested that more than one mapped name matches the same goal name. Some name types were by definition non-unique so redundancy score > 1 was expected. In addition, if our translation was aiming to map to more than one goal name, a 1:1 mapping would be indicated by 1:number of goal name types.

Table (4.2) KATdb and manual translation test cases. Success improvement is calculated as $\frac{\text{Translation success B}}{\text{Translation success A}}$ for B and similarly $\frac{\text{Translation success F}}{\text{Translation success G}}$ for F. The test case details can be found in Supplementary Table A.2. * — unique perturbagen IDs from drug signatures below significance threshold p-value < 0.05 in L1000CDS^2 signature database. ATC — Anatomical Therapeutic Chemical; BRD ID — Broad ID; EMA — European Medicines Agency; LINCS — Library of Integrated Network-based Cellular Signatures.

	Method	Input name type	Goal name type	Unique input names	Unique found names	Unique goal names	Different components	Mapped terms	Not mapped terms	Translation success	Translation redundancy	Success improvement
A	manual	BRD ID	name	5913	5913	5183	n/a	4521	1392	0.765	1.146	n/a
B	KATdb	BRD ID	name	5913	5910	36999	5646	5907	3	0.999	6.264	1.307
C	KATdb	BRD ID	ATC	5913	5910	3174	5646	1294	4616	0.219	2.453	n/a
D	KATdb	BRD ID*	ATC	3123	3121	1676	2994	641	2480	0.205	2.615	n/a
E	manual	RepoDB name	LINCS name	1572	1572	548	n/a	548	1024	0.349	1.000	n/a
F	manual	RepoDB name	BRD ID	1572	1572	684	n/a	548	1024	0.349	1.248	n/a
G	KATdb	RepoDB name	BRD ID	1572	1564	1410	1779	1052	512	0.673	1.340	1.930
H	KATdb	RepoDB DB	BRD ID	1572	1562	1279	1512	1036	526	0.663	1.235	n/a
I	KATdb	EMA name	BRD ID	1615	1381	400	1427	421	960	0.305	0.950	n/a
J	KATdb	EMA INN	BRD ID	1040	883	442	1060	350	533	0.396	1.263	n/a
K	KATdb	RepoDB + EMA name	BRD ID	5802	4968	1459	2862	2608	2360	0.525	0.559	n/a

We summarised each test case (A–K) and its outcomes in Table 4.2, the test case details can be found in Supplementary Table A.2.

4.9.1 LINCS drug IDs to drug names (A–B)

In order to explore how many BRD IDs from L1000CDS^2 we could map to names (case A–B, Table 4.2), we first (A) manually mapped all BRD IDs (*perturbagen ID*) to LINCS names (*perturbagen description*) using the L1000CDS^2 metadata. Initially, no name was assigned to 2878 BRD IDs, however, 1486 BRD IDs that appeared in more than one experimental condition were assigned a human-readable name in one experiment, but not in another. In A, we gave preference to human-readable names when one BRD ID mapped to "-666" and a name. In B, we utilised KATdb to perform the same test. The translation success was higher in B (0.99) compared to A (0.76). KATdb was unable to map only 3 out of 5913 BRD IDs, compared to 1392, if done manually. KATdb improved the translation success 1.3-fold. B had a much higher translation redundancy, where it retrieved 6.3 names for each BRD ID, compared to 1.1 in A. In B, we queried KATdb to return 13 different name types that consist of human-readable names (Supplementary Table A.2), thus a higher than 1:1 mapping was expected.

4.9.2 Annotating drug names with ATC codes (C-D)

We explored the therapeutic classification of L1000CDS^2 drug signatures by mapping all unique L1000CDS^2 identifiers to Anatomical Therapeutic Chemical (ATC) classification codes. We performed two tests C and D (Table 4.2). In both cases we used KATdb to map BRD IDs to ATC codes. In C, we used all unique BRD IDs available in L1000CDS^2 and in D we used BRD IDs from signatures that met a significance threshold (p-value < 0.05). In C we started with 5913 unique BRD IDs and in D with 3123. In both cases, the translation success was similar (C:0.22 and D:0.21), which was to be expected as D input was a subset of C. The translation redundancy was also similar with 1:2.5 ACT codes for C and 1:2.6 for D.

The translation success in C–D was much lower than in A–B. A–B mapped to names which many chemicals and drugs get associated with, however, not every drug has an ATC code. ATC codes are assigned by the WHO Collaborating Centre for Drug Statistics Methodology (WHOCC) if requested. They are linked to specific indications so investigational drugs are not included in the ATC classification. LINCS L1000 data consists of signatures from ~20,000 small molecules, of which only ~3000 are well-characterised and ~16,000 unannotated small molecules. We thus estimated that only ~3000 out of ~20,000 drugs (~15%) would have an associated ATC code, thus we hypothesised that the translation rate reflected the underlying LINCS database structure. More information on L1000CDS^2 data and ATC is discussed in Chapter 5: The Drug Repositioning Pipeline sections 5.3.1 and 5.3.5, respectively.

4.9.3 Translating approved drug names to LINCS drug IDs (E–K)

As a final step of the project described in this thesis, we wanted to assess the performance of the novel drug repositioning method (Chapters 6 and 7. We scored the performance based on the methods' ability to retrieve already approved drugs for a particular indication. We used subsets of two lists of approved drugs: RepoDB and EMA. RepoDB and EMA include information on the FDA and EMA approved, withdrawn, and failed drugs. To score a prioritised list of BRD ID drugs with a list of approved drugs, we were required to map the approved names into BRD IDs. To explore different mapping characteristics of KATdb, we prepared 7 different test cases (E-K, Table 4.2), but only used K in benchmarking the drug repositioning method described as the core contribution of this thesis.

Translating RepoDB (E–H)

We manually mapped RepoDB names to LINCS names in E and to BRD IDs in F. From 1572 RepoDB names we successfully mapped 548 to LINCS names, with translation success of 0.35. In F we first mapped RepoDB names to LINCS names (as in E) and then to BRD IDs. The translation success was 0.35, limited by the first mapping step.

We next repeated the mapping from F with KATdb (G). The translation success score was improved 1.9-fold, to 0.67, compared to manual mapping with the intermediate step. With KATdb the mapping from one name type to another was not limited to one intermediate step, but could simultaneously use all synonym information. In F, the mapping was done in two steps: RepoDB name → LINCS name → BRD ID, while KATdb automated multiple steps with various intermediates. This could be achieved with manual mapping as well, but it would require extensive work, including database search, data cleaning, and data aggregation. For each additional database that could yield additional synonyms these steps would have to be repeated and more complex translation maps of possible synonym information flow would have to be constructed. The network nature of KATdb efficiently connected all extracted relationships.

In H, we explored the translation success of translating from one identifier to another. In addition to names, RepoDB includes a DrugBank ID for every drug name. Instead of the RepoDB name, we used DrugBank ID to map to BRD IDs. The success rate was similar to G, 0.674 compared to 0.665 in H. We hypothesised that using unique identifiers from a curated database would yield higher quality results compared to using drug names. However, the translation success scores were comparable in G-H, so we predicted comparable confidence in mapping from name to BRD ID. The similar translation success was probably due to similar mapping paths from either RepoDB name or RepoDB DrugBank ID to BRD ID, because we included RepoDB in KATdb. Thus, the RepoDB name and RepoDB DrugBank IDs were directly connected in KATdb and any mapping path from either of them would have likely passed through the other.

Translating EMA (I–J)

EMA dataset included two human-readable names, the EMA name, which is a trade name and an INN. In I and J we translated from EMA name and INN to BRD ID, respectively. There were 1615 unique EMA names and 1040 INNs present in EMA. Every EMA name was assigned an INN. 1385 EMA names and 886 INNs were found in KATdb, that mapped to 955 and 1022 BRD IDs, respectively. Although there were $\sim$1.6-times more EMA names than INNs, they mapped to approximately the same number of BRD IDs. EMA names were mapped with translation success of 0.3 and INNs with 0.4. The translation redundancy was also higher in J; these findings suggested that INNs are more frequently used nomenclature than EMA names.

Translating RepoDB and EMA (K)

In the final test case, K, we joined all approved drug identifiers: RepoDB and EMA names, DrugBank IDs, and INNs to map them to BRD IDs. From $\sim$5000 input names found in KATdb, we successfully mapped approximately 50%. Translation redundancy score indicated that on average there were 1:0.6 BRD IDs for each search name. This indicated that approximately 2 input names matched to 1 BRD ID. As we initiated the search with two identifiers per drug, we would expect that two input names matched one goal name.

4.10 Discussion

In this chapter, we developed and characterised a drug synonym resource, KATdb, that overcomes challenges in unifying drug nomenclature. KATdb collected drug identifiers and synonyms from 17 different databases. It improved the translation rate by exploiting graph theoretic principles and is able to estimate the correctness of source database information.

Several studies and reviews have identified the problem of inconsistent and non-standardised nomenclature (Akhondi et al., 2015; Chambers et al., 2013; Dashti et al., 2019; Drug and Therapeutics Bulletin, 2018; Wohlgemuth et al., 2010). Most chemical and drug databases have increased the number of synonyms and external references to other databases (Ursu et al., 2017; Wishart et al., 2018). Wohlgemuth et al. (2010) have developed a service offering single and batch conversion of chemical names to improve the nomenclature used in metabolomic reports. Their service allows one name type per query and the user needs to specify one out of over 200 name types. Although the service is more comprehensive than KATdb, it is chemical focused, therefore there are key drug and drug repositioning resources missing that have contributed to KATdb. In addition, their service relies on established 1:1 links, while KATdb is able to translate with several intermediate steps, linking synonyms that are not directly linked in any of the source databases.

We have shown that KATdb increased the translation success rate compared to manual mapping. When published, the use of the KATdb visual interface has been created to represent an intuitive and fast way to translate drug lists from and to one or many name types. Each group of synonyms will be possible to investigate by examining the component structure with edge betweenness indicating correctness confidence.

The creation of KATdb has relied upon, and has been limited by, source database correctness, where some databases are less vigilant on the correctness of external synonyms. Consequently, KATdb has required more data curation. Thus, we propose that drug databases develop a simple user reporting system where reporting incorrect information is possible, so that each user of that database can benefit from increased quality. Wikipedia is one resource where it is user input reliant and allows real-time corrections and updates. During this investigation we were able to correct several pieces of information on Wikipedia, and thus improve the quality of consequent versions of any databases gaining information from Wikipedia. We acknowledge that the errors were introduced by users.

Due to poorly established nomenclature practices for drug mixtures, we have not resolved them in KATdb. Many structurally similar drugs also appeared as one marketed drug. In larger KATdb connected components, synonyms from related and structurally similar drugs grouped together as one drug. In some cases, this offered additional insight into therapeutic value. For any particular investigation a set of well-defined criteria for drug mixtures and further curation may be required.

A graph theoretic approach to resolving drug mixtures could be implemented when more relationships are added. One of the most widely used external identifiers, PubChem CID, is the primary unique identifier for the comprehensive public PubChem Compound database, that has not yet been included in KATdb. In addition to the Compound database, PubChem also provides a Substance database, which is based on non-unique entries of chemical information by the public. The confusion between the two types of identifiers, CID and SID, for Compound and Substance database, respectively, presents a challenge as many databases using the PubChem IDs as external identifiers do not specify whether they are using CID or SID. Due to their similarity and unclear distinctions between CID and SID in public databases, we omitted using SIDs as PubChem curates and assigns them to a more-frequently used and unique CID. Implementing PubChem into KATdb could reinforce already established relationships between synonyms and consequently, tighten clusters within connected components. This would allow the identification of "bridge" edges that we showed are more likely to convey false synonym relationships.

An important further step is to increase the availability of KATdb by publishing the visual interface and hosting the database online. Additional features, such as user curation and edge betweenness filtering, would enrich the user experience and improve the quality of the resource. We predict that the development and publication of a KATdb R package would increase its implementation into benchmarking pipelines.

KATdb was developed with the aim to increase translation between different types of drug names used in truth-tables and drug repositioning methods. Mapping approved drugs to other drug identifiers presents a benchmarking challenge that KATdb can overcome. Upon publication, we encourage users to use KATdb in benchmarking drug repositioning methods and increase the ability to objectively compare different repositioning methods.

Chapter 5

The Drug Repositioning Pipeline

This chapter describes the development of the drug repositioning pipeline that is the key contribution of this thesis. The work has been developed as the result of two prototype concepts: the Pathway Drug Network (PDN) (Joachim et al., 2018) and the Pathway Coexpression Network (PCxN) (Pita-Juárez et al., 2018). After a brief overview of the pipeline, we describe the core component: the Pathway Drug Coexpression Network (PDxN). Development and characterisation of PDxN is the main focus of this chapter. The remaining pipeline components are briefly described and discussed in relation to PDxN, but further explored in later chapters when applied to 3 case studies.

The Alzheimer's disease (AD) case study of PCxN and Complement to GSEA section published in Pita-Juárez et al. (2018) were conducted and written in collaboration with the first author as initial analysis of work contributing to this thesis.

Others' contributions to this chapter. Gabriel Altschuler (Sheffield Institute for Translational Neuroscience (SITraN), University of Sheffield) designed the overarching concept for a drug repositioning method (Joachim et al., 2018) adopted in this chapter. Yered Pita-Juárez's PhD project (Department of Biostatistics, Harvard T.H. Chan School of Public Health) served as the basis for the underlying method (Pita-Juárez et al., 2018) used for the network construction. Wenbin Wei (SITraN, University of Sheffield) and Sokratis Kariotis (SITraN, University of Sheffield) were both involved in the code review to identify points for improvement. Sokratis Kariotis provided coding support in the computational improvement of the Pita-Juárez et al. (2018) method.

5.1 Pipeline Overview

This chapter presents the Pathway-Drug Coexpression Network (PDxN): a signature-based drug repositioning pipeline. It is a novel pathway-based network method, relying upon gene set correlations that capture the inherent relationships between pathways and drug response signatures across a background of gene expression data. It is unique in that it utilises gene set correlation to create global relationships, giving insight into the degree to which they show similar and/or opposite functionality. The PDxN drug repositioning pipeline is a disease-agnostic approach; each disease pathway signature is used to interrogate PDxN separately.

The PDxN drug repositioning pipeline (Fig. 5.1) has 5 main components that are used to process input data:

(i) the PDxN base system generation,

(ii) disease signature generation,

(iii) signature processing and drug prioritisation,

(iv) benchmarking,

(v) *in vitro* drug testing.

The first component is used as the base system and the other four are dynamically applied to new data sets or case studies.

System generation. We generated PDxN by applying an adaptation of the method described in Pita-Juárez et al. (2018). It is a correlation-based method that summarises the expression of gene set members across a curated background of gene expression microarrays from the Barcode project (McCall et al., 2014). We applied it to two types of gene sets: a pathway set from the Molecular Signatures Database (MSigDB) (Subramanian et al., 2005) and Pathprint (Altschuler et al., 2013) (details in Section 5.3.1 Data resources), and a drug set of signatures from L1000 characteristic direction signature search engine (L1000CDS^2) (Duan et al., 2016) that were split into an up- and down-regulated gene set for each drug signature. The genes in the up- or down-regulated drug gene sets were defined as genes that were either up- or down-regulated upon drug perturbation compared to control. We calculated the correlation between each pathway$\leftrightarrow$drug-direction pair. The correlation estimates represented the edges in the resulting bipartite network with 1473 pathway nodes, 26,124 up-regulated drug nodes, 26,124 down-regulated drug nodes and

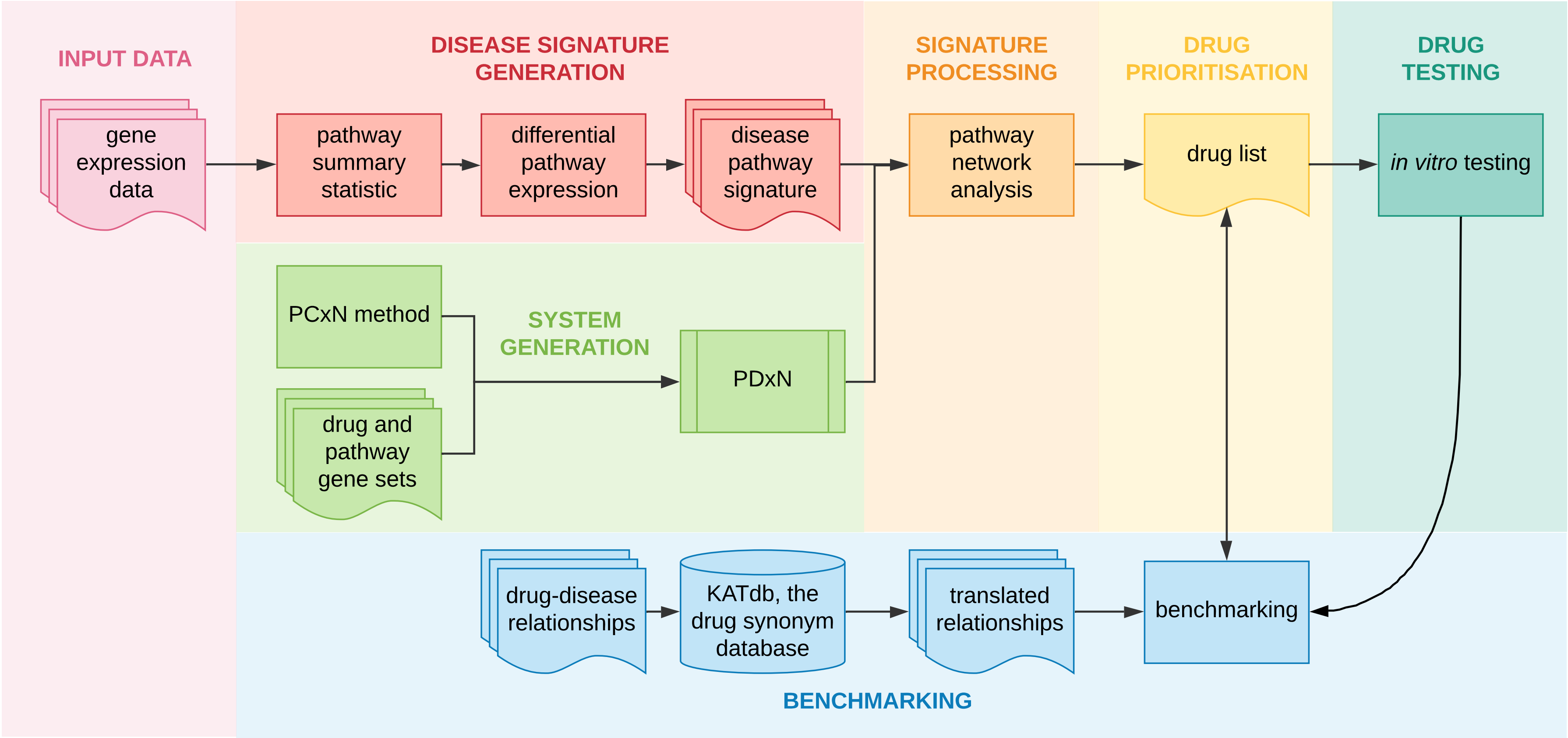


Fig. (5.1) Drug repositioning pipeline overview. The pipeline consists of 5 main components that are used to process input data: (i) the Pathway-Drug Coexpression Network (PDxN) generation (green, centre), (ii) Disease signature generation (red, top), (iii) Signature processing and drug prioritisation (orange and yellow, middle-right), (iv) Benchmarking (blue, bottom) and (v) drug testing (teal, rightmost). The pipeline can be applied to any disease with available gene expression data to generate a disease pathway signature. PDxN relationships between pathways in disease signature and drug signatures are summarised to generate a prioritised drug list. The top drug candidates can then be tested *in vitro* for further investigation.

76,961,304 edges between each pathway↔drug-direction pair. We filtered the network to only include correlation estimates meeting a significance threshold (q-value < 0.05).

Disease signature generation. We generated a disease pathway signature from either microarray or RNA-Seq data, comparing disease samples against controls. We summarised z-scaled gene expression for each pathway by calculating the mean expression of the top 50% pathway member genes with highest $|t|$ score (Section 5.4). We then used `limma` (Ritchie et al., 2015) to generate differentially expressed pathways. We extracted the 5, 10, 15 and 20 most up-regulated ($\log_2$ fold change (logFC) > 0, adjusted p-value (q-value) < 0.05) and the 5, 10, 15 and 20 most down-regulated (logFC < 0, q-value < 0.05) pathways to construct the disease signature.

Signature processing and drug prioritisation. The pathways from the disease signature were split into up- and down-regulated pathway clusters depending on their fold change. We summarised PDxN correlation estimates between each pathway cluster to each drug-direction node. We then summarised the up- and down-direction node of each drug to each pathway cluster, so that we derived one summarised edge score representing the relationship between each *pathway cluster↔drug* pair. The drug list was then prioritised with a decreasing correlation summary score for down- and increasing for up-regulated cluster.

Benchmarking. The prioritised drug list was benchmarked with approved drugs for the disease of interest. If there were no approved drugs for a particular disease, drugs predicted to have a beneficial effect were used instead. A receiver operating characteristic (ROC) curve was calculated for each cluster, assessing the overall performance of PDxN, drug signature generation and signature processing steps. An area under the ROC curve (AUC) was calculated and the AUC score was used to identify the most successful pathway cluster.

Drug testing. A prioritised drug list for the most successful pathway cluster was further investigated. The top drugs were manually curated based on their availability, toxicity and known beneficial effects. The curated drug candidates were then provided to our collaborators to be validated in a wet lab setting e.g. *in vitro* or *in vivo*.

Data resources. The drug repositioning method focused on using publicly available data with a range of different types of databases used at different stages of the pipeline. There were three main data types used in the pipeline:

(i) gene expression datasets for the disease signature generation and background construction in the base network,

(ii) gene set databases for pathway and drug signatures in the network,

(iii) benchmarking datasets to evaluate the performance of the method.

The data provided from collaborators were used when applying the pipeline to new case studies for disease signature generation. When available, we used existing drug screen results from disease models for benchmarking.

5.2 Prototype Network Methods

Two prototype methods have been used as the basis for the pathway-based network repositioning system described in this thesis with the intent to combine the strengths from both while improving on the weaknesses. The first, the Pathway Drug Network (PDN) (Joachim et al., 2018), was developed as a drug repositioning system featuring `pathprint` (Altschuler et al., 2013) for pathway signature generation, and drug and disease gene sets extracted from Connectivity Map (CMap), PharmGKB and CTD. In this prototype, the correlation estimates were calculated across 58,475 microarrays. Specific user-input pathway clusters were then considered for correlation with drug-based signatures. The second, the Pathway Coexpression Network (PCxN) (Pita-Juárez et al., 2018), aims to interpret functional interaction between pathways by systematically quantifying coexpression between canonical pathways from the MSigDB (Subramanian et al., 2005). The correlation was estimated on a curated collection of 3207 microarrays from 72 normal human tissues. The PCxN method accounts for shared genes between annotations to establish significant correlations between pathways with related functions rather than with similar annotations.

We incorporated the network construction concept from PCxN, and from PDN we took the concept of combining pathway and drug signatures, while considering disease pathway clusters for drug prioritisation.

5.2.1 Pathway Drug Network (PDN) — Sepsis case study (Joachim et al., 2018)

The first prototype system, Pathway Drug Network (PDN), described in Joachim et al. (2018), tests whether an experimental gene signature is positively or negatively correlated to a gene signature associated with drug perturbation response from a collection of cell lines. The base network was constructed by calculating the expression correlation between 16,150 drug, disease and pathway gene sets averaged across 58,475 microarrays. By measuring correlation between gene sets across thousands of experiments, they hypothesised that the action that regulates, or is regulated by two gene sets may be linked and/or have similar or opposite functional roles. The drug candidates might therefore promote beneficial pathways or inhibit harmful ones.

The PDN was applied to a sepsis case study (Fig. 5.2). The network considered specific pathway clusters for correlation with drug-based signatures. The query pathway clusters were defined after identifying differences in pathway activity using Pathprint (Altschuler et al., 2013) on publicly available sepsis microarrays. Separate pathway clusters were considered based on combinations of up- and down-regulated pathways between adult and children samples. The network neighbourhood of the sepsis cluster pathways was used to identify drugs that were most positively or most negatively linked to the pathway clusters. For each cluster, a sub-network was constructed that contained nodes representing each of the member pathways of that cluster, together with all the base network nodes with connecting edges to the cluster members. The significance of the base nodes was ranked using the edge p-values aggregated by Fisher's method (Mosteller and Fisher, 1948). The p-values were then further aggregated across drug nodes and clusters accounting for direction of the correlation, resulting in one p-value for each *drug-correlation direction-cluster* combination. The final *drug↔cluster* score was calculated by combining the p-value ranks of correlation direction for each *drug↔cluster* pair.

The top drug candidates were identified by extensive curation of the top scoring drugs for each of the clusters. The curation considered published data collected in preclinical animal models of sepsis on top scoring drug candidates and drugs similar to those top scoring candidates. Evaluation and validation of the resulting drug list by both literature curation and direct experimentation showed substantial enrichment for promising drug candidates. Joachim et al. (2018) have demonstrated that their methodology was more effective at generating positive drug leads than a gene-level method Library of Integrated Network-based Cellular Signatures (LINCS) (Wang et al. (2016b), Appendix Tables S1-S5 in Joachim et al. (2018)).

In summary, PDN offers an alternative to traditional signature-based methods. It tests whether an experimental gene signature is correlated or anti-correlated to the gene signature associated with drug treatment. It quantifies the relationship between two pathway signatures across many experiments rather than accessing their similarity in a

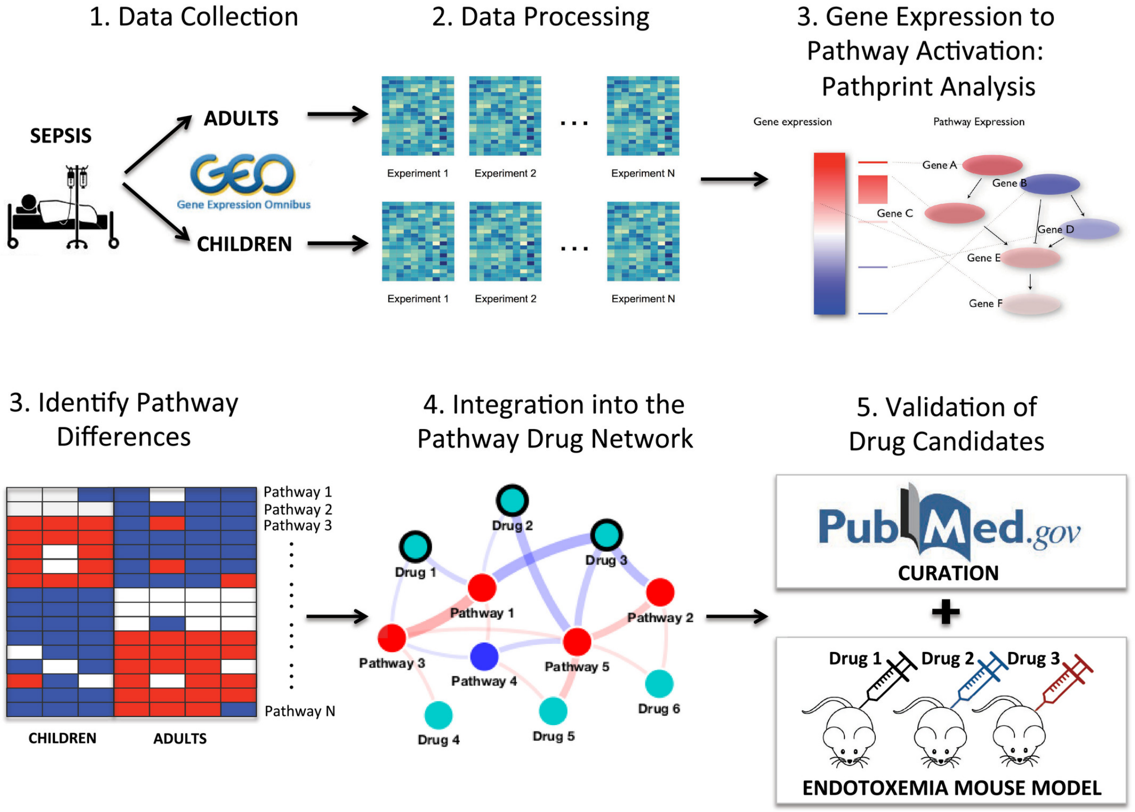


Fig. (5.2) Pathway Drug Network (PDN) overview. (1) Publicly available datasets from transcriptome profiling experiments are identified that include blood leukocyte samples from adult and child sepsis patients. (2) After data processing, (3) Pathprint is used to translate the gene expression patterns at the pathway activity level. After identifying age-associated differences in pathway activity, the pathways are used to facilitate drug discovery by constructing targeted pathway drug networks (PDNs). (4) The method works by incorporating target pathways into a base network built upon the correlation in the expression of $>$ 16,000 disease, pathway, and drug gene signatures across $>$ 50,000 individual microarrays. The resultant network neighbourhood was used to identify drugs with positive or negative association with high-survival (child) or high-mortality (adult) pathways, respectively. (5) The top drug leads were validated by curating and analysing prior data collected in preclinical models of sepsis and also by directly testing their ability to improve survival in a mouse model of fatal endotoxemia. Figure taken, and legend modified from Joachim et al. (2018).

single test. Using human transcriptomics data in both network construction and sepsis signature development increased potential value of subsequent analyses. It has the potential to link any pair of gene signatures in terms of their transcriptional regulation, irrespective of their source, thus providing the possibility to consider use-case specific curated gene signatures. Having a base correlation network meant that case specific clusters of pathways could be used to detect any associated drugs.

Limitations

The main limitations of this approach are related to the meta-analysis of microarray data. The disease signature generation was limited to platforms available in Pathprint, hence overlooking the vast amount of RNA-Seq datasets. The reliance on a relatively small collection of CMap drug perturbation signatures that are quantified on cancer cell lines may have also skewed results. Updating the system to include a much larger successor of CMap, LINCS as well as a newer version of pathway gene sets might have also provided additional power. In addition, final drug candidate score based on aggregation of the correlation estimates, instead of *p*-values, might have revealed more biologically meaningful drug candidates for a given pathway cluster. Despite these limitations, the method served as proof of concept for pathway-based signature driven network methods.

5.2.2 Pathway Coexpression Network (PCxN) — Alzheimer's disease case study (Pita-Juárez et al., 2018)

The Pathway Coexpression Network (PCxN), described in Pita-Juárez et al. (2018), is based on a coexpression method that describes global relationships between pathways. It provides an interpretation of functional interactions between pathways by quantifying coexpression between 1330 canonical pathways using a curated collection of 3207 microarrays in 134 experiments from 72 normal human tissues from GEO curated in Barcode 3.0 (McCall et al., 2014). PCxN is a weighted undirected network where the nodes represent pathway gene sets, and the edges are based on the correlation between the expression of the pathways. We integrated a wide range of experiments by estimating the correlation between summaries of the pathway expression, testing their significance in every experiment, and then aggregating the experiment-level estimates into global estimates. In Pita-Juárez et al. (2018), we demonstrated that PCxN provides novel insight into mechanisms of complex diseases using an Alzheimer's disease (AD) case study.

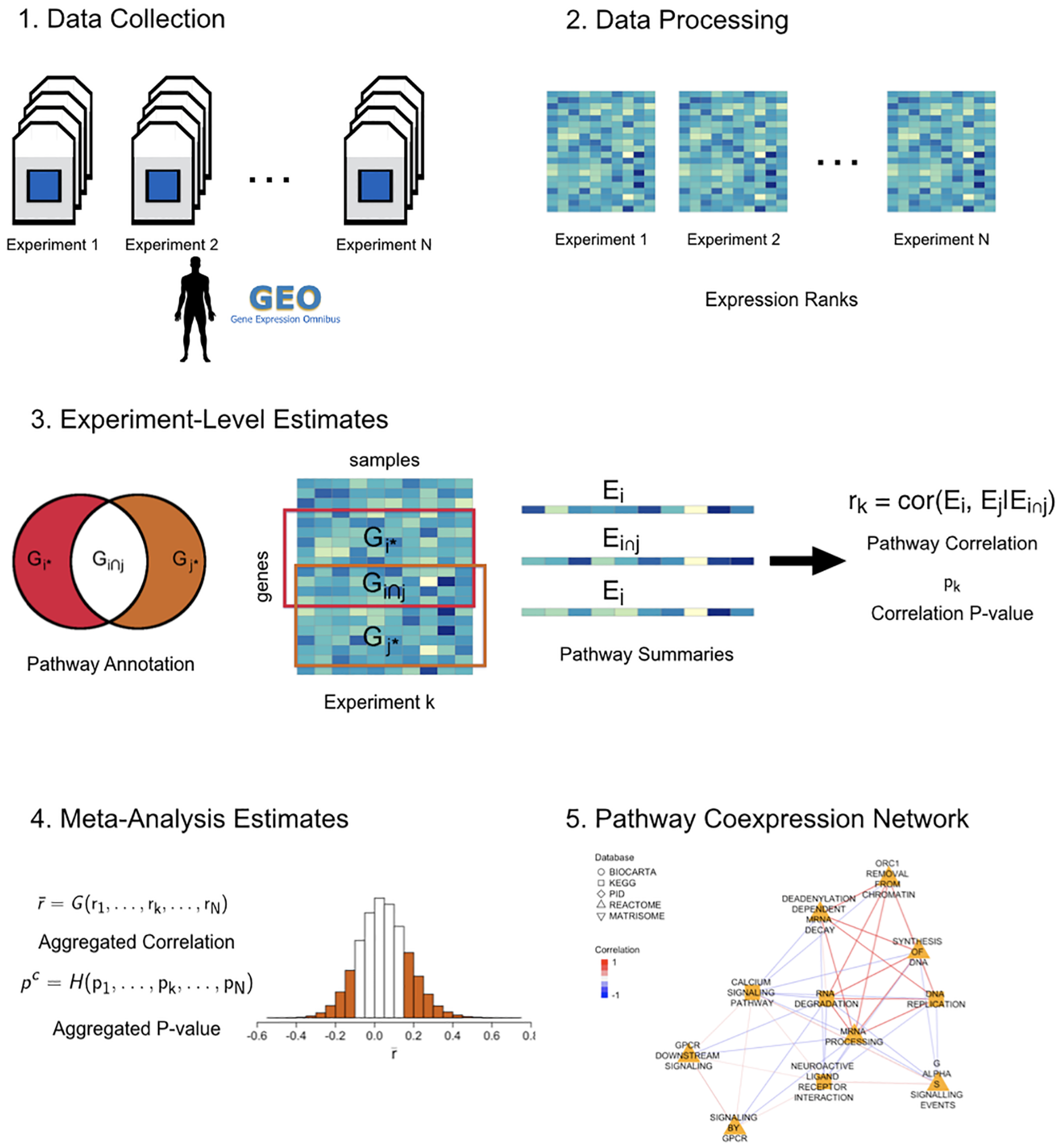


Fig. (5.3) Pathway Coexpression Network (PCxN) overview. (1) Human gene expression arrays for normal human tissues curated from GEO in Barcode 3.0. (2) The gene expression levels were replaced by their ranks so all arrays share a common scale. (3) For each microarray experiment, we first estimated the pathway expression-based on the mean of the expression ranks, then the pathway correlation adjusted for shared genes and tested the significance of the correlation. (4) We aggregated the experiment-level estimates to get the global pathway correlation and its corresponding significance. (5) We built a pathway coexpression network based on the significant pathway correlations. Figure taken and legend modified from Pita-Juárez et al. (2018).

The network was created (Fig. 5.3) by first ranking normalised gene expression levels to provide uniform scale for all samples. Ranks provided a robust summary statistic to calculate expression scores that were independent of the dynamic range of an array. Pathways were assigned an expression summary in each sample based on the mean rank of its member genes. Since the gene expression background was composed of several experiments representing different tissues, the estimated correlation between each pair of canonical pathways summarised expression and tested for significance in every experiment. We then combined the experiment-level estimates into global estimates. Two pathways were connected in the resulting coexpression network if the correlation coefficient between them was significant after adjusting for multiple comparisons.

PCxN adjusts the correlation between pathways by conditioning on the shared genes (Fig. 5.3 part 3) to overcome the pathway annotation redundancies. Pathway databases often include pathways that share genes to varying degrees and shared genes between pathways can either be a consequence of closely related functions or redundant annotation from different sources. Therefore, not accounting for such redundancies during pathway analysis could lead to identifying pathways relationships due to high content-similarity, rather than truly related biological mechanisms. The advantage of PCxN approach is that, by accounting for shared genes between pathways, the relationship between canonical pathways were established when their functions were related, rather than when their annotations had similar content.

PCxN provides a powerful means to generate models for complex diseases by providing pathways significantly correlated with an assay-independent disease gene signature. We applied PCxN to identify key processes related to AD, interpreting a mixed genetic association and experimental derived set of disease genes in the context of gene co-expression. PCxN retrieved pathways significantly correlated with an expert curated AD gene list.

The PCxN case study: Alzheimer's disease

In the PCxN paper (Pita-Juárez et al., 2018), we used genes within an AD curated list (ADCL) as the disease gene signature. The ADCL is a set of association-derived and experimentally-derived genes related to AD that were curated by an AD expert (Professor Rudolph Tanzi, Harvard Medical School) to represent the current understanding of AD (list published in Pita-Juárez et al. (2018), Supplementary Table C.1). We integrated the ADCL to PCxN as an additional gene set, following the same method as for the other pathways. PCxN allowed us to identify canonical pathways significantly correlated with the curated AD gene list.

Since PCxN does not rely on shared genes, PCxN uncovered relationships that would have been missed by methods that rely only on gene overlap to describe the relationships between pathways. All of the top ten correlated pathways had no genes in common with the ADCL. These pathways have known relationships with AD, amyloid pathology or immune system. Furthermore, the correlated pathways were significantly enriched for genes associated with AD independently derived from genome wide association studies. These results showed the value of PCxN in finding biological processes associated with complex diseases using gene signatures.

PCxN provided a powerful contribution to the interpretation of the gene set enrichment methods by describing the relationships between enriched pathways independent of gene overlap. We used PCxN to describe the relationships between pathways identified as enriched by gene set enrichment analysis (GSEA) in a published microarray gene expression experiment profiling the effect of AD in the superior frontal gyrus. We expanded the scope of gene set enrichment results by retrieving pathways correlated with the enriched pathways. The PCxN revealed that correlated pathways from an AD expression profiling study include functional clusters involved in cell adhesion and oxidative stress. It provided a powerful new framework for interrogation of global pathway relationships.

Limitations

PCxN relied on the completeness and correctness of pathway annotations to relate biological processes. A limitation is that PCxN only considers a pathway as a gene list, omitting any knowledge of the interaction between its members. Compared to pathways as a gene list, pathway topology-based methods have been shown to perform better (Nguyen et al., 2019). It is also limited by the gene expression data used to estimate the correlations. The current implementation only used one microarray platform and a curated expression background. This implementation of PCxN did not take advantage of the growing number of publicly available RNA-Seq datasets. The method could be expanded to include a wider range of pathway annotations and to use gene expression data from other platforms such as RNA-Seq.

PCxN established the utility of describing relationships between pathways in a broad context. By using a diverse set of gene expression experiments, it leveraged correlation estimates across various human tissues effectively capturing relationships regardless of shared genes.

5.3 Pathway-Drug Coexpression Network (PDxN)

The Pathway-Drug Coexpression Network (PDxN) is a weighted bipartite network connecting pathway and drug gene sets. The PDxN is, as the name suggests, a pathway-based network estimating the correlation between pathway and drug nodes. It is the core of the drug repositioning pipeline described in this thesis. It serves as the base network that is then interrogated with user-specific disease pathway signatures. While most other pipeline components were generated for each individual application, PDxN was pre-calculated and remained relatively static, with the dynamic possibility to be expanded to include other gene sets.

The relationships between gene sets were based on the methodology of PCxN (Pita-Juárez et al., 2018) (Section 5.2.2). The PDxN relationships estimated correlation by summarising the expression of gene set members across a curated normal background of gene expression microarrays from Barcode 3.0 (McCall et al., 2014) and then calculating the correlation between each pair of pathway↔drug gene sets. The background gene expression data consisted of 134 experiments with 3207 Affymetrix Human Genome U133 Plus 2.0 microarrays (GPL570) from 72 normal human tissues.

The correlation in PDxN was calculated between two types of gene sets, a pathway set of 1329 pathways from MSigDB (Subramanian et al., 2005) and a subset of 144 static modules from Pathprint (Altschuler et al., 2013; Wu et al., 2010b), and a drug set of 26,124 drug signatures from L1000CDS^2 (Duan et al., 2016) that were split into an up- and down-regulated gene set for each drug signature. We calculated the correlation between each pathway↔drug-direction pair, resulting in a bipartite network with 1473 pathway nodes, 26,124 up-regulated drug nodes, 26,124 down-regulated drug nodes and 76,961,304 edges between each pathway↔drug-direction pair.

We adapted the idea of using many types of gene sets from PDN, which included pathway, drug and disease gene sets. However, a part of drug- and all disease-gene sets in PDN were limited to genes researched and annotated to be associated with the drug/disease, skewed by a reporting and curation bias towards well-studied genes and processes. Curated drug and disease gene sets also lack directionality that can be preserved with signatures generated from gene expression data. Therefore, we have only included experimentally-derived drug signatures in PDxN. The gene expression-based drug signatures used in PDN were updated from the pilot CMap (Lamb et al., 2006) dataset to a larger successor database LINCS (Keenan et al., 2018; Subramanian et al., 2017), processed by using the characteristic direction as part of the L1000CDS database (Duan et al., 2016). While the pathway gene sets were in their definition limited by their annotations, we updated and expanded the pathway gene sets to MSigDB v6.2 (Subramanian et al., 2005) and added data-driven static modules from `pathprint` (Altschuler et al., 2013; Wu et al., 2010b).

To be able to apply it to an expanded set of gene sets, we have improved the computational efficiency of the PCxN methodology (Section 5.3.2). We applied it to two types of gene sets, pathway and drug, and limited the relationships calculated to only pathway↔drug, rather than all possible relationships, to further decrease the computational requirements. We were able to apply the PCxN method to a ∼36-times bigger node set, calculating ∼71-times more edges.

5.3.1 Data resources

The PDxN was built from 4 different resources: MSigDB (Subramanian et al., 2005), Pathprint (Altschuler et al., 2013) and L1000CDS (Duan et al., 2016) for gene sets and Barcode 3.0 (McCall et al., 2014) for background gene expression data.

MSigDB (Subramanian et al., 2005)

The network nodes were represented by 1473 pathway nodes of which 1329 are MSigDB v6.2 C2 canonical pathways. The Molecular Signatures Database (MSigDB) is a collection of annotated gene sets designed for use with Gene Set Enrichment Analysis (GSEA) software. It is divided into 8 main collections:

(i) H: hallmark gene sets,

(ii) C1: positional gene sets,

(iii) C2: curated gene sets,

(iv) C3: motif gene sets,

(v) C4: computational gene sets,

(vi) C5: gene ontology (GO) gene sets,

(vii) C6: oncogenic gene sets,

(viii) C7: immunologic gene sets.

We considered using H and C2, but due to low numbers (50) of gene sets present in H, we proceeded with only C2. C2 consisted of 4276 gene sets that were further split into CGP: chemical and genetic perturbations (3433), and CP: Canonical pathways (1329). As we utilised a more expansive set of chemical perturbations, we only used CP Canonical Pathways subset of C2 curated gene sets in our pathway set of gene sets.

The C2 Canonical Pathways include canonical representations of biological processes compiled by domain experts. They were aggregated from the following pathway databases:

(i) Reactome (Matthews et al., 2009),

(ii) KEGG (Kanehisa et al., 2014),

(iii) the Pathway Interaction Database (PID) (Schaefer et al., 2009),

(iv) Biocarta (Nishimura, 2001),

(v) the Matrisome Project (Naba et al., 2012),

(vi) Signal Transduction Knowledge Environment (Gough, 2002),

(vii) SigmaAldrich (SA) and Signaling Gateway (Saunders et al., 2008).

Pathprint static modules (Altschuler et al., 2013; Wu et al., 2010b)

In addition to MSigDB C2 Canonical Pathways we included 144 static modules from `pathprint` R package (Altschuler et al., 2013). Static modules are a set of data-derived functional-interaction gene clusters (Wu et al., 2010b). Pathprint included 633 pathways of which there were 489 canonical pathways and 144 static modules. The canonical pathways were constructed from KEGG, Reactome and Wikipathways. We incorporated only static modules, because the canonical pathways include pathway annotations from outdated versions of databases used in MSigDB.

We used static modules (Wu et al., 2010b) to counterbalance inherent curation bias towards well-studied genes introduced in MSigDB gene sets. Static modules are non-curated gene sets of highly connected genes from a functional-interaction network. They enabled us to examine the activity of less studied or annotated biological processes.

As described in Wu et al. (2010b), the functional-interaction network was constructed by extending curated pathways with non-curated sources of information, including protein-protein interactions, gene co-expression, protein domain interaction, GO annotations and text-mined protein interactions. The functional-interaction network consisted of 181,706 interactions between 9452 genes. A Markov cluster algorithm was used to decompose the network, yielding 144 highly connected functional-interaction clusters, termed static modules, ranging in size from 10 to 743 nodes or member genes. Each cluster was named according to the member gene with the highest degree, i.e. the hub gene. The modules cover 6458 genes, 1551 of which are not represented in MSigDB v6.2 C2 Canonical pathways.

Top GO terms associated with all the static modules can be found in Altschuler et al. (2013) "Additional file 2" to provide additional biological context for the static modules.

LINCS L1000 (Subramanian et al., 2017) and L1000CDS (Duan et al., 2016)

The Library of Integrated Network-based Cellular Signatures (LINCS) L1000 small molecule expression profiles is the underlying dataset for the L1000 characteristic direction signature (L1000CDS) database used as the source of drug signatures in PDxN.

The LINCS L1000 dataset utilised a new gene expression profiling method, L1000, that drastically lowered cost and therefore enabled cost-effective, high-throughput screenings currently totalling 1.30 million L1000 profiles available in LINCS. The L1000 profiling method measures 978 landmark transcripts and then imputes the rest of the transcripts. They showed that the ∼1000 data-driven landmark transcripts were sufficient to recover 82% of the information in the full transcriptome (Subramanian et al., 2017).

The LINCS L1000 project has collected gene expression profiles for over 25,000 genetic and small molecule perturbagens at a variety of time points, concentrations, and cell lines, generating ∼470,000 signatures (consolidating replicates) from 1.3 million profiles. There were ∼20,000 small molecules of which a small subset of ∼3000 represented the annotated small molecules that were systematically profiled in 9 core cell lines (A375, A549, HA1E, HCC515, HEPG2, HT29, MCF7, PC3, and VCAP) — the touchstone subset, and the rest (∼16,000) were unannotated small molecules that were tested in variable cell lines — the discover subset.

The LINCS L1000 data was separated into five data levels at different points in the analysis pipeline:

(i) Level 1: Raw unprocessed data,

(ii) Level 2: Gene expression values per 1000 genes,

(iii) Level 3: Normalised gene expression profiles of landmark genes and imputed transcripts,

(iv) Level 4: Gene signatures computed using z-scores relative to the plate population as control or relative to the plate vehicle control

(v) Level 5: Differential gene expression signatures computed using the moderate z-score method.

The L1000CDS dataset used characteristic direction (CD) method to calculate the differentially expressed genes of the profiles in LINCS L1000. The CD (Clark et al., 2014) is a multivariate method that first identifies the linear hyperplane that best separates the control from the case samples using linear discriminant analysis. It then uses the normal to the hyperplane to define the direction of change in expression space for each gene. It gives less weight to individual genes that display a large change in magnitude when comparing two conditions. Some genes that change in magnitude substantially may be given a lower score, or a p-value, compared with other methods such as the fold-change method. The CD method gives more weight to genes that move together in the same direction across repeats. Therefore, a gene that changed less but moved together with a large group of genes in other repeats may have been scored higher than a gene that changed more in overall magnitude (Duan et al., 2016).

The L1000CDS used LINCS L1000 Level 3 normalised data to calculate a CD unit vector for each experimental replicate compared to all the control replicates on the same plate. The CDs across replicates were then averaged. The mean of the pairwise cosine distance between the CDs across replicates was used as a test statistic to generate a p-value. Duan et al. (2016) showed that processing the L1000 data with CD method significantly improved signal to noise compared with the moderate z-score method currently used by original LINCS to compute L1000 signatures.

To remain directionality of drug signatures, we split each signature into two non-overlapping gene sets: up-regulated, and down-regulated gene set according to the CD value of each member gene. CD > 0 for up- and CD < 0 for down-regulated. So that each drug node in PDxN represented a particular direction of drug expression in specific conditions.

The L1000CDS included 119,156 drug signatures of which 26,124 met a significance threshold (p-value < 0.05) and included more than 5 genes in each direction. Resulting in 26,124 up- and 26,124 down-regulated drug nodes, contributing to 52,248 drug nodes.

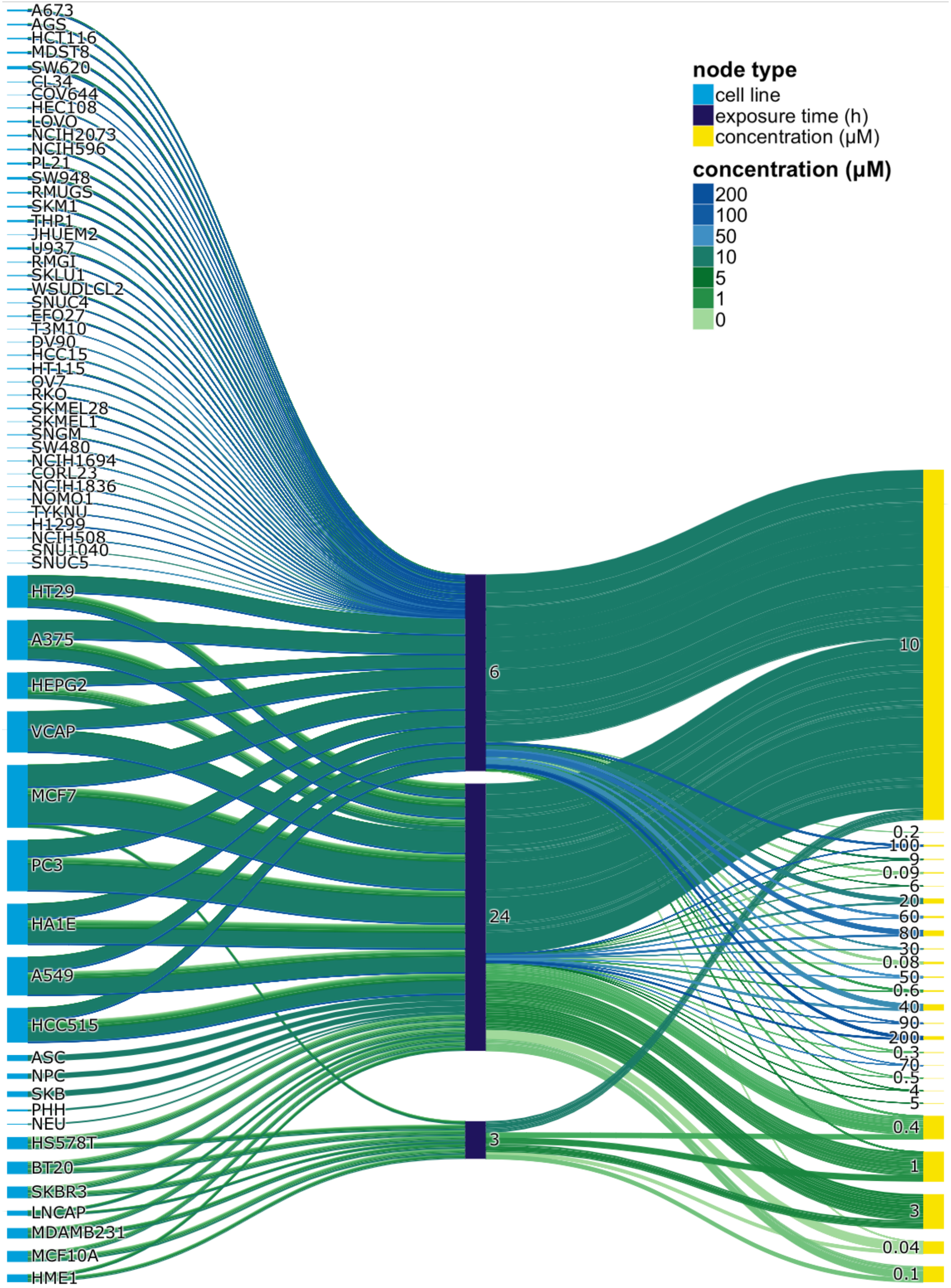


Fig. (5.4) The L1000CDS drug signatures in PDxN. Distribution of 26,124 drug signatures (p-value < 0.05) representing 3105 different drugs across cell types (light blue), exposure times (h, dark blue) and concentrations (μM, yellow). Showing the 9 most predominant cell lines, at mostly 24h exposure time at $10\mu M$ concentration. Each drug signature is combined across several replicates. Node colour indicates metadata property type (node type). Edge colour represents the concentration value in μM. Concentration values are rounded to 1 significant digit. h — hours; L1000CDS — L1000 Characteristic Direction Signature; PDxN — Pathway Drug Coexpression Network.

Each drug signature represented a specific set of conditions, varying drug, exposure time, cell type and concentration (Fig. 5.4). Cell types included primary cell lines, cancer cell lines, stem cell lines, and differentiated cell lines from different tissue types. It can be seen from Fig. 5.4 that the distribution of 26,124 drug signatures from L1000CDS still mimics the structure of LINCS L1000 data where there were 9 predominant cell lines (A375, A549, HA1E, HCC515, HEPG2, HT29, MCF7, PC3, and VCAP) systematically tested at two exposure times (24h and 6h) and several concentrations, but predominantly at 10μM. There were fewer signatures tested in other cell lines (not the main 9) and it can be seen from Fig. 5.4 that these were tested in a less systematic way.

Barcode 3.0

We calculated the correlation estimates between pathway and drug gene sets in a curated normal background of gene expression microarrays from Barcode 3.0 (McCall et al., 2014). We summarised the expression of each gene set in each sample. We then calculated the correlation between each pair of pathway↔drug gene set across samples and combined it across experiments. The background gene expression data consisted of 134 experiments with 3207 Affymetrix Human Genome U133 Plus 2.0 microarrays (GPL570) from 72 normal human tissues. The curated microarrays in Barcode 3.0 were filtered to exclude poor quality samples (McCall et al., 2011b, 2014).

We based our pipeline on a normal background of gene expression data, so that we could follow the signature hypothesis and match opposite drug and disease signatures to try drive the system back to normal, healthy state (Fig. 2.4). The disease signature was defined in disease gene expression data and drug signatures were defined from drug perturbations in a set of conditions on a variety of cell lines. We brought them together and investigated their correlation relationship on the curated background of normal gene expression in a collection of tissues. We captured the functional relationships between pathways and drug signatures under normal conditions. We could identify pathway↔drug pairs that were closely related across many tissues. We quantified relationships between pathways that were dysregulated in disease conditions and drug genes that were dysregulated upon drug treatment.

5.3.2 Reducing computational resource requirements

The PCxN method used to construct PDxN, was initially designed to compute relationships between 1330 pathways. In order to be able to apply this method to a much larger pathway and drug set, we were required to perform significant computational improvements. We focused on reducing computational requirements and improving the computational speed. The original method had been split into four parts:

(i) Part 0: Filter gene set annotations to keep only genes present in the gene expression background

(ii) Part 1: Get experiment-level estimate; estimate all pairwise pathway correlation coefficients and corresponding p-values

(iii) Part 2: Aggregate the experiment-level correlation estimates and p-values

(iv) Part 3: Aggregate results into a single data structure

Each part needed to be completed in full before the next could start.

In collaboration with thesis supervisor Wenbin Wei and junior programmer Sokratis Kariotis we started with a desk review of the code. This led us to identifying a set of possible slow points. We tested the key functions to identify the bottlenecks. The largest bottlenecks were partially rewritten in C++ by Sokratis Kariotis, but because of the large amount of R dependencies it did not provide any speedup. Part 0 and 3 required a negligible amount of time and memory compared to part 1 and 2. Therefore, we focused our improvements only on part 1 and 2.

In part 1, we identified that there was a redundantly repeated calculation when summarising the pathway expression for each pathway pair. Instead of summarising the pathway once for every experiment, the pathway summary was calculated whenever the pathway was in a pathway pair. To resolve this, we pre-calculated the pathway summary matrix and then called it when calculating the pathway pair correlation.

In the original method part 1 was split in 72 tasks, one task per tissue, with tissues represented by 1–11 experiments. The tissues tasks with more experiments were therefore queuing the experiments and were predicted to take longer. To allow further parallelisation, we split part 1 into even more tasks. We calculated estimates for one experiment per task, splitting experiment-level estimates in 134 smaller tasks.

In part 2 where we combined the experiment-level estimates, we increased the number of pairs calculated per task from 1000 to 100,000. While it was faster to combine 1000 pairs per task, the number of parts increased linearly with an increasing size of relationships, e.g. 5 tasks for 100 gene sets with 4950 pairs to 500 tasks for 1000 gene sets with 499,500 pairs. In each task all the experiment-level estimates needed to be read into memory and then subsetted to the task pairs before the experiment-level estimates were combined for each task pair.

The original PCxN was thus split into 1 task for part 0, 72 for part 1, $\frac{\text{number of pairs}}{1000}$ tasks for part 2 and 1 for part 3. We restructured it so that it ran in 1 task for part 0, 134 for part 1, $\frac{\text{number of pairs}}{100\,000}$ tasks for part 2 and 1 for part 3. We predicted that increasing the number of tasks in part 1 and decreasing in part 2 will improve computational efficiency of the method.

We added two new features that made it possible to apply it to a much larger number of gene sets. We first limited the pairs calculated to those of interest rather than all possible pairs. For example, in PDxN we were only interested in pathway↔drug pairs, while PCxN calculates all possible pairs. Additionally, due to memory constraints when considering larger gene sets, we added an extra optional joining step that can join multiple subsets of the network together. For example, with the current version of PDxN, we split the drug gene sets into 6 sets and calculated 6 sub-versions of PDxN, one for each drug subset, that were then joined and p-values corrected for multiple testing.

We measured the performance improvement by running a set of test runs with the original and improved PCxN method (Fig. 5.5). We varied the number of the gene sets and consequently, also the number of relationships calculated. We used the PDxN pathway set in our tests where we calculated all possible relationships, i.e. all pathway↔pathway pairs. We tested both methods at 10, 20, 50, 100, 200, 400, 800, 1000, 1200 and 1473 gene sets, with 1473 representing the updated size of PCxN. We ran each part with 8 cores with 4GB virtual memory (vmem) each. We repeated each test two times. We measured maximum vmem (max vmem) and wall clock time. Wall clock time is the actual amount of time taken to perform a task. We reported max vmem and wall clock time per part and total for all parts together with an increasing number of pairs.

Pre-calculating pathway summary scores in part 1 and keeping them in memory decreased the computational time but increased the initial memory burden. The memory requirement stabilised with an increasing number of gene sets. As the number of pairs increased, the total max vmem of the original version surpassed the memory requirements of the improved version. The wall clock time of the original part 1 increased at ∼2-times the rate of the improved version. The most significant improvement was increasing the number of pairs calculated in part 2 from 1000 per task to 100,000. In each task, all the

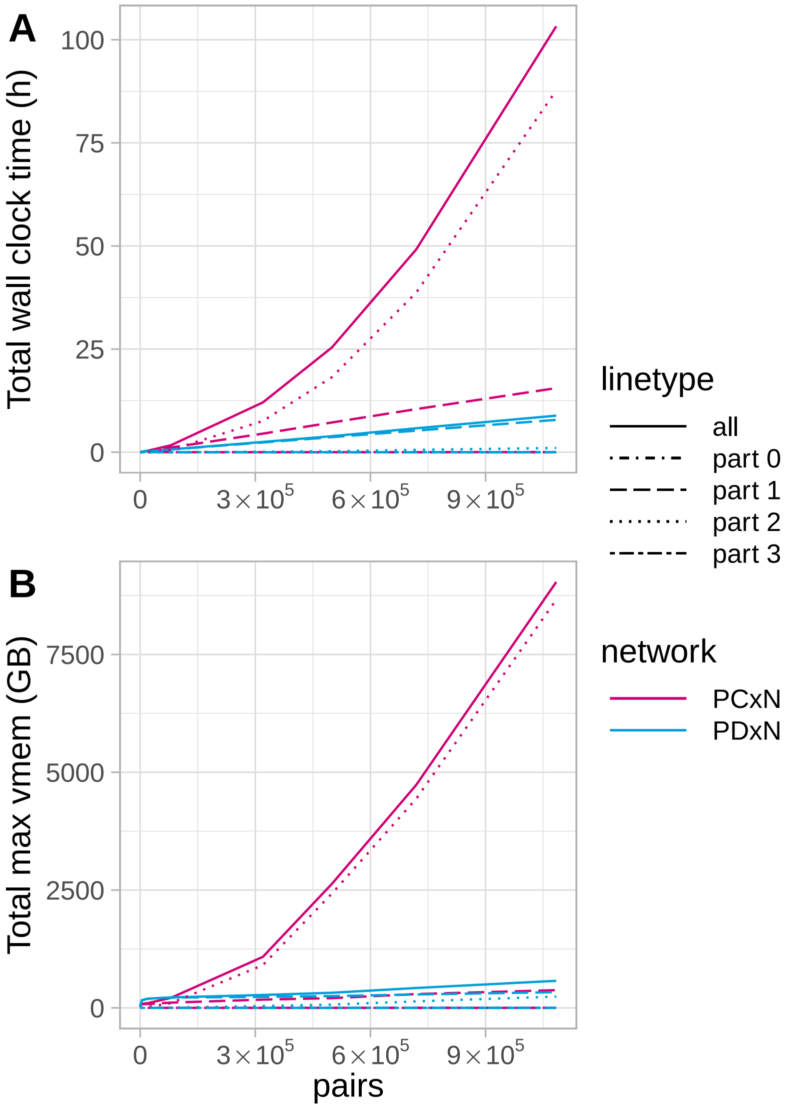


Fig. (5.5) Computational improvement of PCxN method. We improved the original PCxN method to be able to apply it to a larger number of gene sets. The PCxN method is split into 4 parts. Part 0 and 3 represent a negligible part of total time and memory. We thus focused on improving part 1 and 2. We tested the original and the improved method with an increasing number of gene sets and thus pairs. We measured (A) the total wall clock time (in h) and (B) the total maximum virtual memory (max vmem, in GB). There is ∼ 10-times improvement in total wall clock time and ∼ 15-times improvement in total max vmem. GB — gigabyte; h — hours; max vmem — maximum virtual memory; PCxN — Pathway Coexpression Network; PDxN — Pathway Drug Coexpression Network.

experiment-level estimates were read into the environment and then subsetted to the task pairs. The more tasks part was split into, the more times this was repeated, introducing large memory and time burdens. It is possible that a further increase in the number of pairs per task would further decrease the total memory and time. In addition, the new feature of joining the sub-networks made it possible to dynamically add extra gene sets, without the need to re-calculate the existing network.

Table (5.1) Summary of the computational improvements of the PCxN method (Pita-Juárez et al., 2018). Improvement is calculated as the original method's parameter divided by the improved. An improvement score > 1 indicates an improvement and < 1 indicates a decrease in performance. max vmem — maximum virtual memory; PCxN — Pathway Coexpression Network.

Number of gene sets	Number of pairs	Total max vmem improvement	Total wall clock time improvement
10	45	0.633	0.450
20	190	0.767	0.435
50	1225	0.704	0.614
100	4950	0.506	0.921
200	19900	0.517	1.520
400	79800	0.954	2.500
800	319600	3.950	4.850
1000	499500	8.210	6.540
1200	719400	11.200	8.480
1473	1084128	15.700	11.700

The final results showed that changes caused a decrease in performance when considering a small number of pairs ($\lesssim 5000$) but an increasing improvement when testing a larger number of pairs. We measured $\sim$12-times speedup at 1473 gene sets and 1,084,128 pairs, and $\sim$16-times lower total max vmem. Testing showed no change in results.

The improved performance of the method made it more accessible and available for reuse and further development. For example, with increasing availability of data and continuous evolving curated databases, the improved performance makes it possible to update the system with new information available.

5.3.3 Network topology

We constructed PDxN with computationally improved PCxN method (Pita-Juárez et al., 2018), making use of additionally implemented features that made it possible to apply the method to an extended number of gene sets. We described the resulting network in more detail by considering its topological features.

Network topology describes the topological structure of a network in particular the arrangement of the network elements. We modelled gene sets as nodes and the correlation estimates between nodes as weighted connections termed edges. Using graph theory, network topology gives insight in the underlying structure and properties of a particular network. We investigated the PDxN topology and compared it to the PCxN to evaluate their structure at varying significance thresholds and to investigate whether functional relationships were also captured in PDxN even though we were applying the PCxN method to two types of gene sets. An updated version of PCxN was generated using the same pathway set as used in PDxN.

A distinctive feature of PDxN is that it is a bipartite network, where nodes are separated in two independent sets, drugs and pathways, and connections exist only between two nodes from different sets (Fig. 5.6). In comparison, PCxN is a fully connected network where all nodes are connected to all other nodes (at significance threshold p-value ≤ 1). It has only one set of nodes, pathway nodes. In both networks the connections between nodes represent the correlation estimates calculated for the pair of gene sets that the connected nodes are representing. The correlation estimate is a summarised correlation score for given nodes across experiments on a background of gene expression. The edges can be positive or negative depending on the correlation estimate, where two gene sets are either positively or negatively correlated. Each correlation estimate has an associated q-value, which is used to filter the edges in a fully connected PCxN and a complete bipartite PDxN. Absolute correlation values were used as edge weights in clustering of both graphs and generating PDxN projections.

Theoretically, PDxN could be a unipartite network where we would be investigating all possible relationships between gene sets, ignoring the underlying information the gene sets represent. However, we have only considered relationships between pathway↔drug pairs rather than all possible relationships, i.e. pathway↔drug, pathway↔pathway (PCxN), drug↔drug, due to computational limitations as well as the aims of the proposed pipeline. The omitted relationships could be calculated to further investigate pathway↔pathway and drug↔drug relationships. By limiting our relationships to only pathway↔drug we were required to calculate only $n \times m = 76{,}961{,}304$ rather than $\frac{(n+m)((n+m)-1)}{2} = 1{,}442{,}946{,}060$ of all possible edges, where $|P| = n$ is the number of nodes in pathway set P and $|C| = m$ the number of nodes in drug set C. Thus, PDxN (at significance threshold p-value ≤ 1) represented 5.33% of all possible relationships, significantly reducing the computational burden.

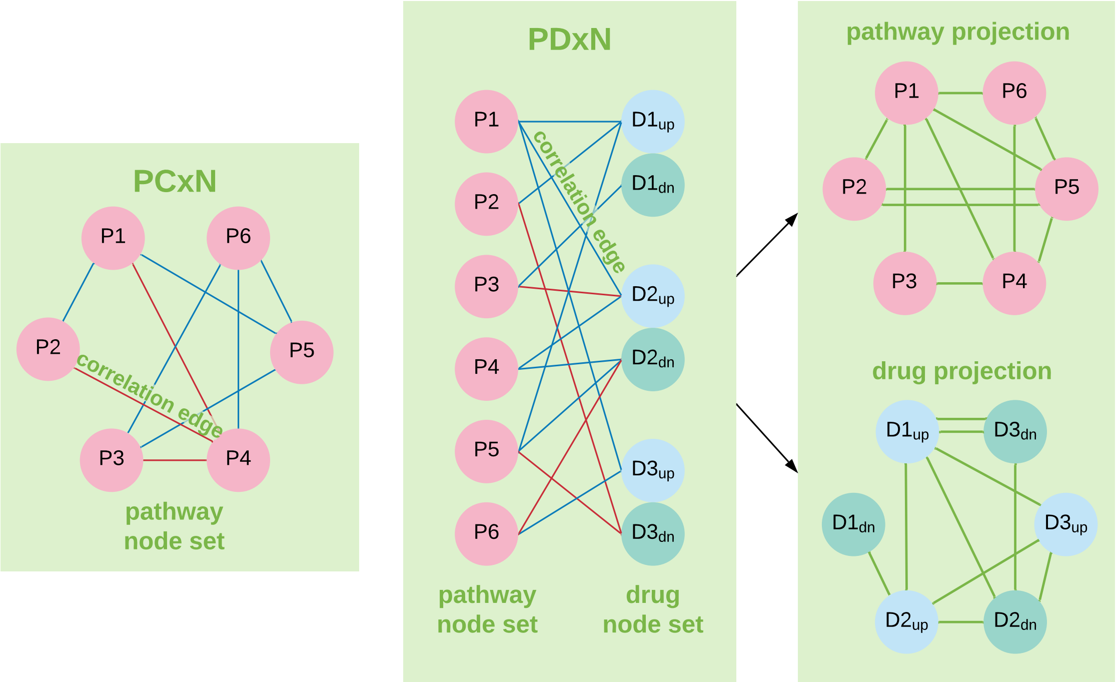


Fig. (5.6) PCxN, PDxN and PDxN projections. PCxN consists of one set of nodes: a pathway node-set, P; while PDxN consists of two sets of nodes: a pathway set, P, and a drug node-set, D. The pathway node set (pink) consists of gene sets annotated with a particular function i.e. pathway and static modules representing functionally connected gene communities. The drug node sets consist of two nodes for each drug, one representing up- (blue, D$_{up}$) and one down-regulated (teal, D$_{dn}$) genes from drug perturbation experiments. The connections between nodes in PDxN and PCxN (dark blue, red) represent correlation estimates of gene sets represented by the two nodes on a background of gene expression data. Connections exist only between two nodes from different sets in PDxN. In PCxN, connections exist between nodes in the pathway node set. The edges can be positive (red) or negative (blue) depending on the correlation estimate, where two gene sets are positively or negatively correlated. PDxN is projected into drug and pathway network projections, where the nodes in each projection are connected if they are connected to the same node from the other node set in the original bipartite network. PCxN — Pathway Coexpression Network; PDxN — Pathway Drug Coexpression Network.

To further characterise PDxN and PCxN we investigated their topological features. We first looked at how the networks change at q-value thresholds. Tables 5.2 (PCxN) and 5.3 (PDxN) show that over 90% of network edges in both networks have an q-value > 0.99. This was expected as gene sets were likely to be differentially regulated in different tissues and hence were not likely to have a consistent relationship across experiments. Table 5.2 shows that on average, pathways only have significant relationships with less than 10% of other pathways (Table 5.2). PDxN has even fewer connections which could be explained by a higher number of gene sets. With higher number of gene sets (1473 for PCxN and 53,721 for PDxN), the system was trying to summarise a higher number of individual member genes (10,450 unique genes for PCxN and 14,781 for PDxN) across experiments and tissues, thus we could expect fewer consistently expressed relationships.

We generated sub-networks for PDxN and PCxN to further investigate how mean degree and the total number of nodes, and the total number of edges changed with the q-value threshold for the edge correlation values (Fig. 5.7A-B). As seen in tables 5.2 and 5.3, there was a drastic decrease at q-value $\cong 1$ in edge properties (number of edges (green), mean degree of pathway nodes (dark blue) and mean degree of drug nodes (blue), right side for both, PCxN (5.7A) and PDxN (5.7B)). The networks were relatively unaffected at $0.99 > q\text{-value} > 0.05$ (Fig. 5.7A-B right, dashed line — $q\text{-value} = 0.05$) with all the metrics decreasing slowly with lower p-value thresholds. The networks gradually reduced in size at q-values < 0.05 (Fig. 5.7A–B left) in a linear manner.

At q-value ≤ 0.05 there were 4.32% of edges remaining connecting 95.2% of pathway and 89.5% of drug nodes, PCxN, in contrast, is higher with 7.26% of edges remaining intra-connecting 97.6% of pathway nodes.

In the remainder of this thesis we used PDxN and PCxN with a q-value threshold < 0.05. At that threshold there were 1402 pathways, 21,227 drug signatures with both direction nodes, 2442 with only up-regulated signatures, and 1841 with only down-regulated signatures. On average each pathway was connected to ~ 2370 drug nodes of which 1345.6 were up-regulated and 1024.2 were down-regulated. Each drug node was connected to ~ 71 pathways (up-drug nodes to 79.7 pathway nodes and down-drug nodes to 62.2 pathways). There were 1437 pathway gene sets remaining in PCxN connected to ~ 110 other pathways.

We compared the PDxN correlation value and q-value distributions to PCxN. We observed similar q-value and correlation distributions (Fig.5.7D, E, respectively) in PDxN and PCxN, despite the fact that PDxN has over 40-times more edges than PCxN (at q-value < 0.05) and there were no overlapping edges between the networks. It could be argued

Table (5.2) Summary of PCxN properties at different significance thresholds. Each q-value threshold represents values $\leq$ than the value stated. The value in brackets states the % of the total number of nodes or edges. PCxN — Pathway Coexpression Network.

q-value	Number of pathway nodes	Mean degree of pathway nodes	Total number of edges
0.001	1417 (96.2%)	92.2	65341 (6.03%)
0.005	1424 (96.7%)	98.5	70122 (6.47%)
0.01	1428 (96.9%)	101.6	72541 (6.69%)
0.05	1437 (97.6%)	109.5	78685 (7.26%)
0.99	1448 (98.3%)	132.9	96222 (8.88%)
1	1473 (100%)	1472.0	1084128 (100%)

Table (5.3) Summary of PDxN properties at different significance thresholds. Each q-value threshold represents values $\leq$ than the value stated. The value in brackets states the % of the total number of a given property. PDxN — Pathway Drug Coexpression Network.

q-value	Number of pathway nodes	Mean degree of pathway nodes	Number of drug nodes	Mean degree of drug nodes	Total number of edges
0.001	1335 (90.6%)	1969.3	44316 (84.8%)	59.3	2629043 (3.42%)
0.005	1361 (92.4%)	2112.0	45240 (86.6%)	63.5	2874376 (3.73%)
0.01	1382 (93.8%)	2167.1	45668 (87.4%)	65.6	2994883 (3.89%)
0.05	1402 (95.2%)	2369.7	46737 (89.5%)	71.1	3322333 (4.32%)
0.99	1429 (97%)	2983.9	48923 (93.6%)	87.2	4264050 (5.54%)
1	1473 (100%)	52248.0	52248 (100%)	1473.0	76961304 (100%)

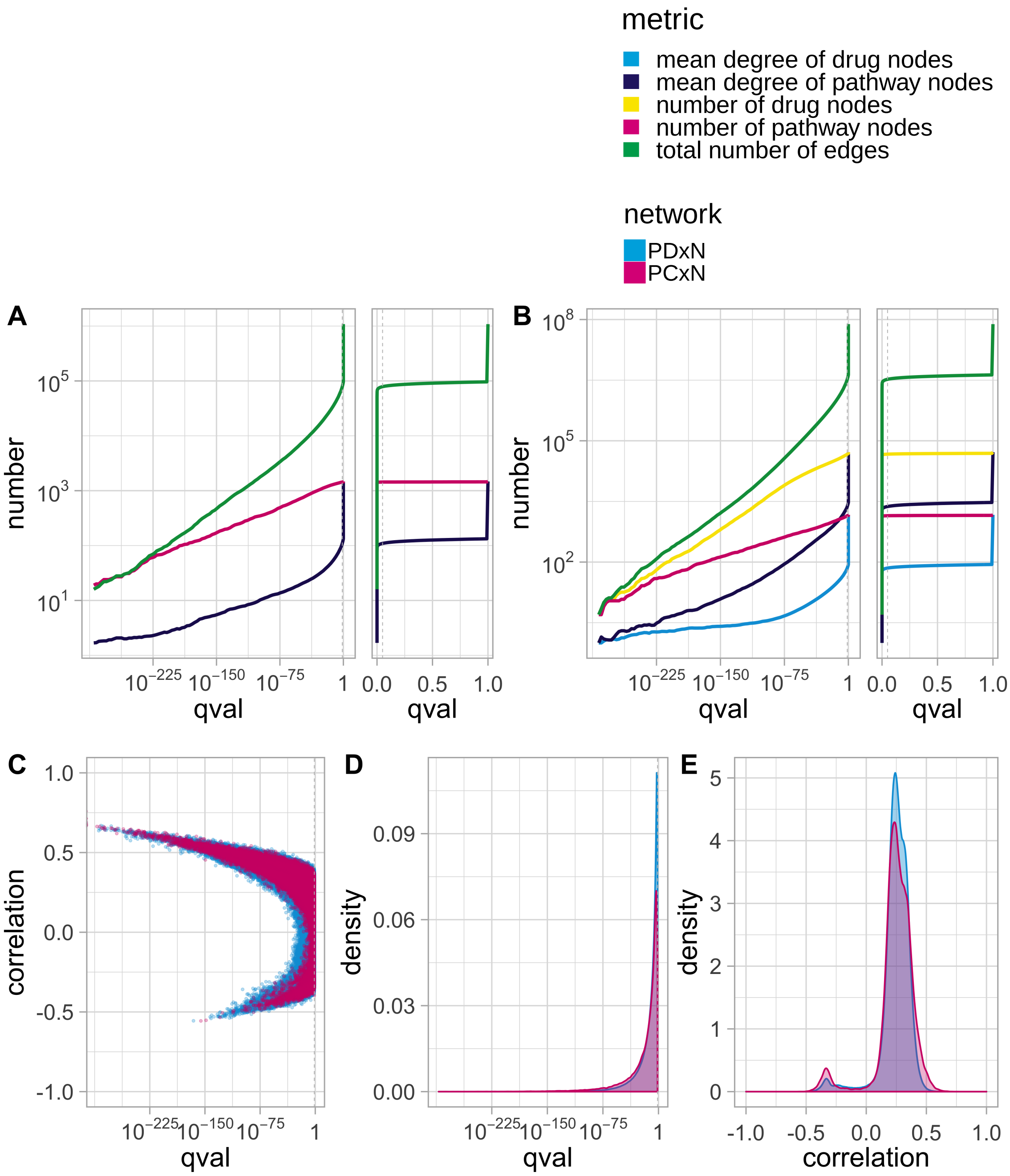


Fig. (5.7) PCxN and PDxN network properties. Mean degree and number of drug and pathway nodes and total number edges for (A) PCxN and (B) PDxN are shown at q-value cut-offs for the edge correlation. The networks are relatively unaffected at q-values > 0.05. The drastic fall at q-value $\cong 1$ in edge properties on the right reflects the variation in gene expression across experiments and tissues. Note that the x-axis on the A–B left is $log10$-scaled. Relationship between edge correlation and q-value (C), q-value (D) and correlation (E) distributions for PDxN (blue) and PCxN (pink) at q-value < 0.05. The networks show similar q-value and correlation distributions even though PDxN is much larger and has no-overlapping edges with PCxN. The dashed line represents q-value = 0.05. correlation — edge correlation value, correlation between two gene sets on a background of gene expression data; PCxN — Pathway Coexpression Network; PDxN — Pathway Drug Coexpression Network; qval — q-value.

that a much smaller PCxN still modelled the discussed topology of a larger PDxN with limited calculated relationships. Both PDxN and PCxN had a much higher proportion of positively correlated gene sets (Fig.5.7C, E), with positive edges reaching higher absolute correlation values and higher levels of significance. The imbalance between the number of positive and negative edges was more disproportionate in PDxN with 96.1% of edges with correlation > 0 in PDxN, compared to 94.8% in PCxN. This imbalance was important for the consideration for the downstream drug repositioning pipeline.

5.3.4 Network clustering

To investigate the functional context of PDxN and PCxN topology we clustered each network (q-value < 0.05) and then annotated pathway nodes with KEGG and Reactome pathway groups, and drug nodes with the Anatomical Therapeutic Chemical (ATC) classification classes. In this section, we provide a summary of clustering algorithms used and introduce two different approaches to analysing a bipartite network. We then describe the PDxN community structure and discuss the functional annotation patterns detected by clustering PCxN and PDxN.

Network clustering methods

We clustered each network with a modularity-based community detection method. We used the Louvain, also known as multi-level, method (Blondel et al., 2008) for PCxN and the PDxN network projections, and the Label Propagation Algorithm weighted bipartite plus (LPAwb+) (Beckett, 2016) method for PDxN. We chose two different methods, because Louvain is designed for unipartite graphs and LPAwb+ for weighted bipartite networks. Both methods are based on optimising modularity and are implemented in R (R Core Team, 2019) packages, `igraph` (Csardi and Nepusz, 2006) and `bipartite` (Dormann et al., 2009), respectively.

Network modularity is a measure of the strength of division of a network into clusters or modules. It compares the density of edges inside communities to edges outside communities. Networks with high modularity have dense connections between the nodes in the same cluster but sparse connections between nodes in different clusters. Modularity values lie in the range between -1 and 1. Modularity is the fraction of the edges that fall within the given groups minus the expected fraction if edges were distributed at random (Li and Schuurmans, 2011). It is positive if the number of edges within groups exceeds the number expected by chance. There are several methods for network community structure detection that use modularity optimisation.

The Louvain method (Blondel et al., 2008), developed at University of Louvain, extracts communities or clusters from large networks by optimising modularity as the algorithm progresses. Optimising the modularity value theoretically results in the best possible grouping of the nodes of a given network, however, going through all possible iterations of the nodes into groups is unfeasible. In the Louvain method, first small communities are found by optimising modularity locally on all nodes, then each small community is grouped into one node and the first step is repeated, limiting the number of iterations required and therefore applicable to large networks. The Louvain method was chosen for PCxN and PDxN projections as it was shown to be an appropriate method for networks with less than 6000 nodes, but also for large networks with not well-defined communities (Yang et al., 2016). We applied weighted Louvain clustering to PCxN, the edges were weighted by the absolute values of correlation estimates.

In contrast to simply applying a clustering algorithm to a bipartite network, we could approach the characterisation of bipartite network communities in two common ways: we could create two unipartite projections from a bipartite network (Fig. 5.6) or investigate the original bipartite network accounting for its bipartite nature.

A network projection can be used to reduce dimensionality in a bipartite network, by summarising it into two, unipartite graphs, one for each set of nodes. Two nodes within the same set are connected in the projection if they are connected to the same (share a common neighbouring) node from the second set in the original bipartite network (Fig. 5.6). This approach is particularly common as it allows studying a bipartite network as two unipartite networks and can therefore use the powerful tools provided for classical, unipartite networks. However, there are potential drawbacks to projections. By reducing the dimensionality and complexity of the original network, there is some information in the bipartite structure that may disappear after projection. For example, two drug nodes that had low correlation with a pathway node are now equivalent to two drug nodes that were connected with a large correlation value. On the other hand, if two drug nodes have a lot of common neighbours, that will be reflected in the projection as increased edge weight. When projecting, we also introduce an inflation of the number of edges, which introduces large demands for computational resources and thus makes further processing more challenging. Lastly, some properties of the projection may be due to projection itself rather than the underlying network (Latapy et al., 2008). There are a few ways to compensate for limitations introduced by projections. To reduce the size of the projection and retain the most meaningful information, a filtering approach is usually applied to the network before the projection. In addition, a weighting projection approach is commonly used to decrease the loss of information (Pavlopoulos et al., 2018).

An alternative to using network projections is to use the original network. This approach considers and accounts for the bipartite nature of networks. There are fewer tools available for topological analysis of bipartite networks than for unipartite, but there is a growing number of proposed approaches applicable to bipartite networks, some of which are extensions of the traditional unipartite measures. An additional difficulty is that the available approaches are usually developed for particular applications and are thus very case specific and lack generality (Latapy et al., 2008).

We used three main criteria for selecting an appropriate clustering approach for a bipartite network: finding a clustering approach designed for weighted bipartite networks that outperformed other methods, an ability to analyse large networks, and its implementation within an R package. We identified the LPAwb+ (Beckett, 2016) implementation in `bipartite` (Dormann et al., 2009) as the most appropriate choice given our criteria.

The LPAwb+ (Beckett, 2016) is based on LPAb+ (Liu et al., 2010) which uses label propagation and multi-step agglomeration to maximise modularity identifying joint communities composed of both types of nodes in non-weighted bipartite networks. Similarly to LPAb+, but developed to account for the edge weights, LPAwb+ consists of two stages: step 1 that maximises modularity on a node-by-node basis using label propagation, and step 2 that joins modules together when it results in increased network modularity. In more detail, first a unique label is given to each of the nodes in the smallest of the two sets. Then the label propagation stage is initiated where the algorithm asynchronous updates the labels of each node set to locally maximise modularity. The nodes in one set can only use information about the nodes in the other set to update their labels. The labels are updated in turns until modularity can no longer be increased. The second agglomeration step seeks the global maximum by merging groups of communities together as the local maxima identified in step 1 may not be the global maximum. Each identified community module is composed of nodes from both sets that share the same label. Merging of two different communities can only occur if that results in an increase in network modularity and there is no third community whose merger with either of the two communities would result in a larger increase in modularity. Step 1 and then step 2 are repeated until it is no longer possible to increase network modularity by merging any of the communities together (Beckett, 2016). LPAwb+ is only the second weighted bipartite algorithm proposed. It robustly identifies partitions with high modularity scores, it is appropriate for large networks and outperforms the first proposed bipartite weighted algorithm QuanBiMo (Dormann and Strauss, 2014).

Weighted bipartite clustering of PDxN

We first clustered PDxN with LPAwb+ (Beckett, 2016) and identified 12 clusters with both pathway and drug members (Fig. 5.8). The clusters varied in size, with 6 large ($>$ 1% nodes in cluster 1, 2, 3, 5, 6, 9) and 6 very small ($<$ 1% nodes in cluster 4, 7, 8, 10, 11, 12) clusters (Table 5.4). The largest cluster, according to the number of nodes, was cluster 6 with 29% of all nodes and largest by the number of edges was cluster 2 with 32% of all edges from pathway nodes and also 32% of all edges from drug nodes. Cluster 2 predominantly included edges from pathway↔up-regulated drug nodes and cluster 9 included pathway↔down-regulated drug node relationships. The separation and clustering of up- and down-regulated signatures from one another was expected.

Table (5.4) Summary of node membership and internal edges in PDxN clusters. % for total nodes is the proportion of nodes in that cluster compared to all, % for internal edges is the proportion of internal edges compared to all edges from pathways and drugs in that cluster, respectively.

Cluster	Pathway nodes	Up-regulated drug nodes	Down-regulated drug nodes	Total nodes	Internal edges
1	137	2713	2362	5212 (11%)	121288 (34%, 34%)
2	348	7162	2559	10069 (21%)	576343 (54%, 54%)
3	216	3348	3299	6863 (14%)	109761 (29%, 29%)
4	9	40	3	52 (0.11%)	135 (28%, 45%)
5	11	835	1017	1863 (3.9%)	11913 (58%, 60%)
6	321	7431	6056	13808 (29%)	403005 (48%, 54%)
7	5	218	12	235 (0.49%)	674 (29%, 34%)
8	1	0	1	2 (0.0042%)	1 (100%, 25%)
9	351	1921	7757	10029 (21%)	524025 (79%, 70%)
10	1	0	1	2 (0.0042%)	1 (25%, 25%)
11	1	1	0	2 (0.0042%)	1 (100%, 100%)
12	1	0	1	2 (0.0042%)	1 (50%, 20%)

Although we have defined PDxN clusters and assigned cluster membership to each node, we could see high levels of inter-connectivity between clusters. This is particularly noticeable on Fig. 5.8 with smaller clusters e.g. 4, 5, 7, and 10, where we can observe many connections from both pathway and drug nodes in a particular cluster that lead do other clusters. Cluster 11 was the only cluster that has only connections within the cluster and

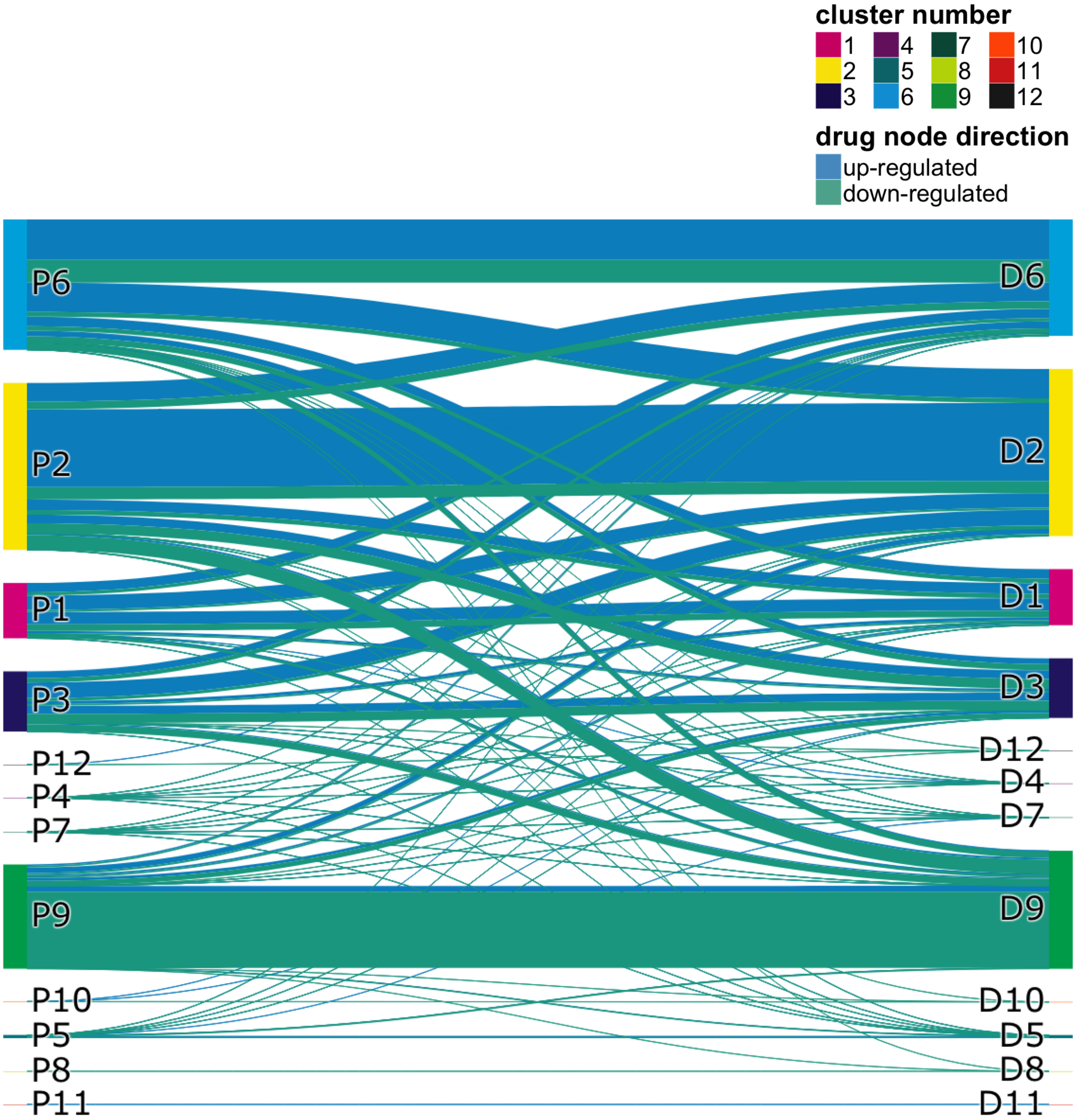


Fig. (5.8) PDxN inter- and intra-cluster interactions. PDxN was clustered using LPAwb+ (Beckett, 2016) identifying 12 clusters varying in size. The clusters are interconnected. Cluster member nodes interact with nodes from the same cluster e.g. P1↔D1 and nodes from other clusters e.g. P1↔D2. Each cluster is made from nodes in pathway (P, left) and drug (D, right) node sets. The number after P or D indicates the cluster the P or D partition belongs to. The size of the node is proportional to the number of edges from pathway nodes (P) to drug nodes (D) in a given cluster. The node colour represents the cluster membership. The edge colour represents the colour of the drug node direction (in D, right) a particular pathway is connected to. PDxN — Pathway Drug Coexpression Network.

none outside, however, cluster 11 only includes two nodes (pathway↔up-regulated drug) and one edge, so by definition the edge could only be connecting the two member nodes and thus, was only connected within the cluster. Considering only large clusters, cluster 9 was the most intra-connected (79% internal pathway and 70% internal drug edges) and cluster 3 was the most inter-connected (71% external pathway and drug edges).

An advantage of using a bipartite weighted approach was that we have identified clusters that consist of both pathway and drug nodes. We were therefore able to explore relationships between pathways and drugs within the same cluster as well as pathways and drugs responsible for connections with other clusters.

5.3.5 Functional annotation of network clusters

Below we explore the functional landscape of PDxN and compare it with PCxN. We describe PCxN and PDxN clusters annotated with pathway groups from KEGG (Kanehisa and Goto, 2000) and Reactome (Matthews et al., 2009). In addition, we also annotated the PDxN clusters with the ATC drug classes.

Pathway annotation of PCxN and PDxN clusters

To explore if PDxN not only mimics PCxN topology but also retains the functional relationships, we annotated each network with KEGG (Kanehisa and Goto, 2000) and Reactome (Matthews et al., 2009) pathway group terms (Supplementary Table C.2). Both KEGG and Reactome offer hierarchical relationships between pathway terms. As both resources were present in PCxN and PDxN, we leveraged their hierarchical nature to explore the functional landscape of both networks and look for enrichment of higher-level pathway terms in network clusters.

The KEGG hierarchical pathway structure is split into 4 levels: A–D. Where A is the most general and D the most detailed pathway annotation. Level A consists of 6 groups (*Metabolism, Genetic Information Processing, Environmental Information Processing, Cellular Processes, Organismal Systems, Human Diseases*) and level B of 46 group terms. Level C pathway annotations are included in MSigDB C2 Canonical Pathway set. The Reactome pathway annotations are linked into a network. We clustered the Reactome pathway annotations network into 27 pathway groups. Pathway names from pathway node set were name-matched to KEGG level C and Reactome pathway annotations. We matched

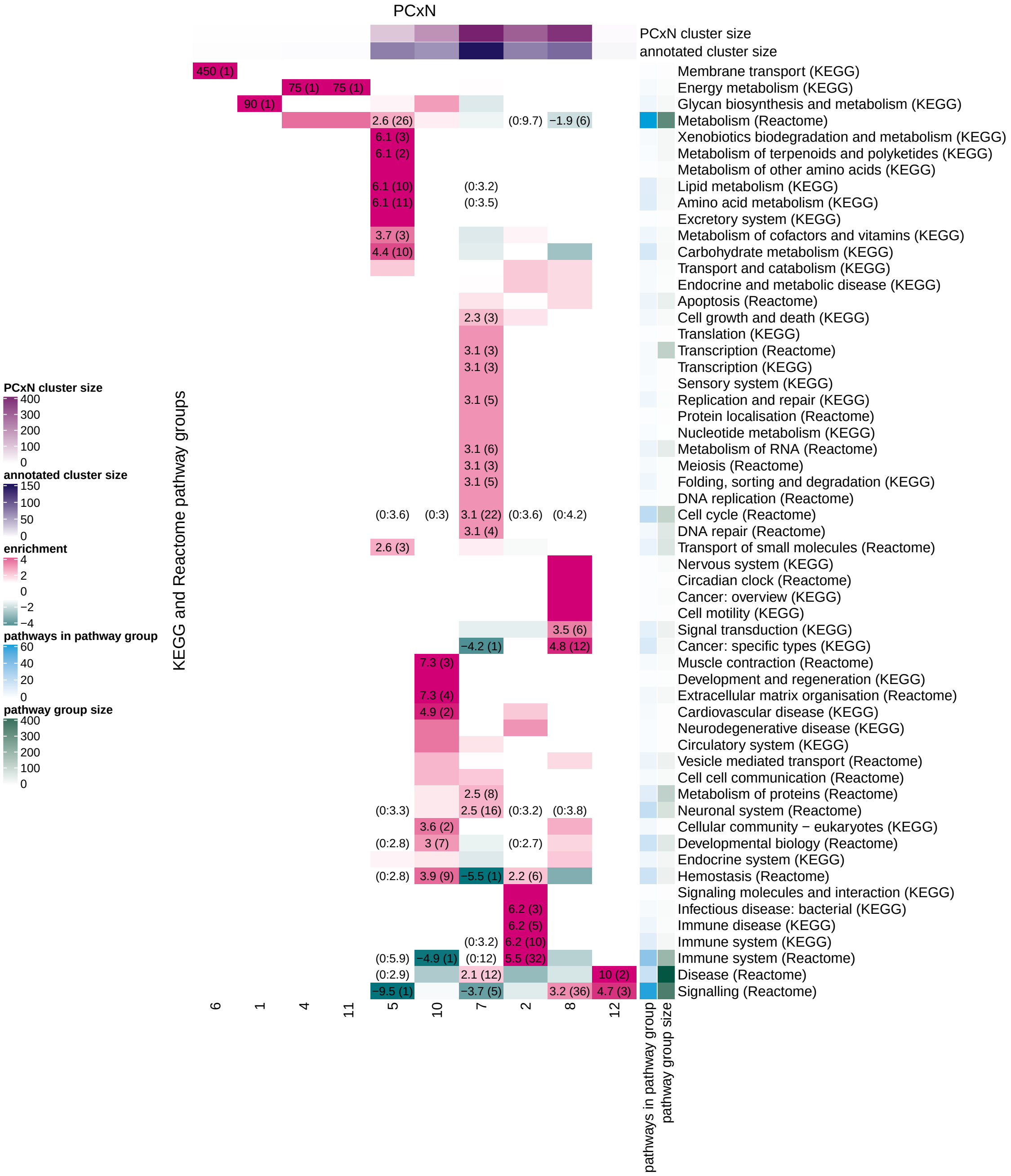


Fig. (5.9) KEGG and Reactome pathway annotation enrichment of PCxN clusters. The heatmap summarises enrichment (pink) or depletion (teal) score. Absolute enrichment values of 0 or above 1.5 and below significance threshold p-value < 0.05 are displayed. The first number displayed is the enrichment or depletion score, while the number in () is the observed number. When the observed number is 0, the depletion score cannot be calculated, thus we displayed number-of-observed terms : the-number-of-expected terms in (). Column names correspond to arbitrary PCxN cluster numbers. Cluster numbers are not related to PDxN clusters, but some clusters show similar enrichment patterns (PCxN-cluster↔PDxN-clusters: 2↔1, (1,4–6,11)↔(4,5,7,10,11), 7↔9, 8↔3, 10↔6). Row names are KEGG and Reactome pathway group names. Only clusters with at least one pathway group annotation are shown. KEGG — Kyoto Encyclopedia of Genes and Genomes; PCxN — Pathway Coexpression Network; PDxN — Pathway Drug Coexpression Network.

312 different Reactome pathways from both PCxN and PDxN to 23 group terms and 146 of KEGG pathways to 36 B level terms (Supplementary Table C.2). We matched 296/660 Reactome pathways in PCxN, 307/636 in PDxN, and 144/183 KEGG pathways in PCxN and 148/176 in PDxN.

We calculated enrichment or depletion for a given pathway group for each cluster. We summarised enrichment values in Fig. 5.9 for PCxN and Fig. 5.10 for PDxN.

The PCxN consisted of 12 clusters, of which 10 (cluster numbers 1, 2, 4–8 and 10–12) had at least one pathway group annotation. Cluster 1, 4, 6 and 11 were the smallest clusters with 1–2 pathway annotations. Enriched biological pathway sets existed for nearly every cluster. Cluster 1, 4 and 11 include metabolism pathways and cluster 6 included the only membrane transport pathway. Cluster 5 was mostly enriched in metabolism pathways, it was enriched in *Transport of small molecules*, general *Metabolism* from Reactome and several metabolism groups from KEGG: *Xenobiotics biodegradation and metabolism, Metabolism of terpenoids and polyketides, Lipid metabolism, Amino acid metabolism, Metabolism of cofactors and vitamins,* and *Carbohydrate metabolism.* Cluster 10 enriched in *Cardiovascular disease, Hemostasis, Developmental biology, Cellular community, Extracellular matrix organisation* and *Muscle contraction.* Cluster 7 enriched in core cell processes (genetic information processing), it enriched in *Metabolism of proteins, DNA repair, Cell cycle, Meiosis, Metabolism of RNA, Transcription, Replication and repair, Folding sorting and degradation*, and *Cell growth and Death.* Both transcription groups from Reactome and KEGG enrich in this cluster. It was also enriched in *Disease* pathways and *Neuronal system.* Cluster 2 enriched in immune system related groups: *Immune system* (KEGG and Reactome), *Infectious disease: Bacterial, Immune disease, Hemostasis*. Cluster 8 enriched in signalling groups: *Signalling and Signal transduction*, and most specific cancer type pathways. A small cluster 12 consisting of 5 annotated pathways enriched in *Signalling* and *Disease*.

In summary, there were a few clusters that enrich for mainly one type of pathway annotation. Cluster 5 mostly included metabolism pathways, cluster 10 included developmental pathways and cell-cell communications, cluster 7 included core cell processes like transcription, and cluster 2 enriched in immune pathways. Considering that the PCxN method accounted for shared genes and assigned significant correlation coefficients between pathways representing related functions and non-significant correlation coefficients for pathways with redundant annotations representing the same function, we have shown that PCxN can be clustered in several functionally enriched clusters. PCxN explored functional relationships between pathways and provided a framework for interrogation of global pathway relationships (Pita-Juárez et al., 2018).

We repeated the functional pathway annotation with bipartite clustered PDxN discussed above. As with PCxN, PDxN also consisted of 12 clusters, of which 10 (cluster numbers 1–7 and 9–11) had at least one pathway group annotation. There were several small clusters, cluster 10 and 11 have only one pathway annotation. They both were related to metabolism. Cluster 7 with 3 annotated pathways enriched in *Lipid metabolism* and cluster 5 with 4 pathways in *Metabolism of proteins* and *Translation*. Cluster 4 with 8 annotated pathways enriched in *Xenobiotics biodegradation and metabolism* and *Carbohydrate metabolism*. We observed that the small clusters were mostly related to metabolism. Cluster 9 mostly involved the core cell processes (genetic information processing): *DNA repair, Cell cycle, Folding, sorting and degradation, Meiosis, Metabolism of RNA, Transcription*, and *Replication and repair*, but also, *Disease pathways, Amino acid* and *Carbohydrate metabolism*. Cluster 6 enriched in *Extracellular matrix organisation, Muscle contraction, Cardiovascular disease, Endocrine system, Developmental biology*, and *Hemostasis*. Cluster 2 enriched in the *Neuronal* and *Sensory system.* Cluster 3 in Signalling and specific cancer types, and cluster 1 in the *Immune system* (KEGG and Reactome) and *Immune disease*.

As in PCxN, there were a few clusters that enriched in mainly one type of pathway annotations. The small clusters mostly included metabolism pathways, cluster 6 included developmental pathways and cell-cell communications, cluster 9 included core cell processes and cluster 1 enriched in immune pathways.

PCxN estimated the correlation between pathway pairs and PDxN included only pathway↔drug pairs. Every PDxN pathway pair put in the same cluster depended on a similar relationship to a set of drug nodes. As in PCxN, the correlation coefficients in PDxN also accounted for shared genes and thus gave significant relationships to pathways representing related functions rather than redundant annotation. Considering that we clustered PDxN by using information from the other set of nodes, we still retrieved significantly enriched functional clusters that modelled the PCxN clusters. Even though we used different clustering approaches and the networks were different in their topology, we see consistent patterns of enrichment between PCxN and PDxN. The similar PCxN-cluster↔PDxN-cluster pairs were: 2↔1, (1,4–6,11)↔(4,5,7,10,11), 7↔9, 8↔3, 10↔6.

It was encouraging to see such functional overlap in PCxN and PDxN clusters. It showed that pathway↔pathway relationships were transmitted across pathway↔drug↔pathway edges. Although we accounted for gene set overlap, we still clustered similarly annotated pathway groups together. For example, both Reactome and KEGG *Immune system* groups

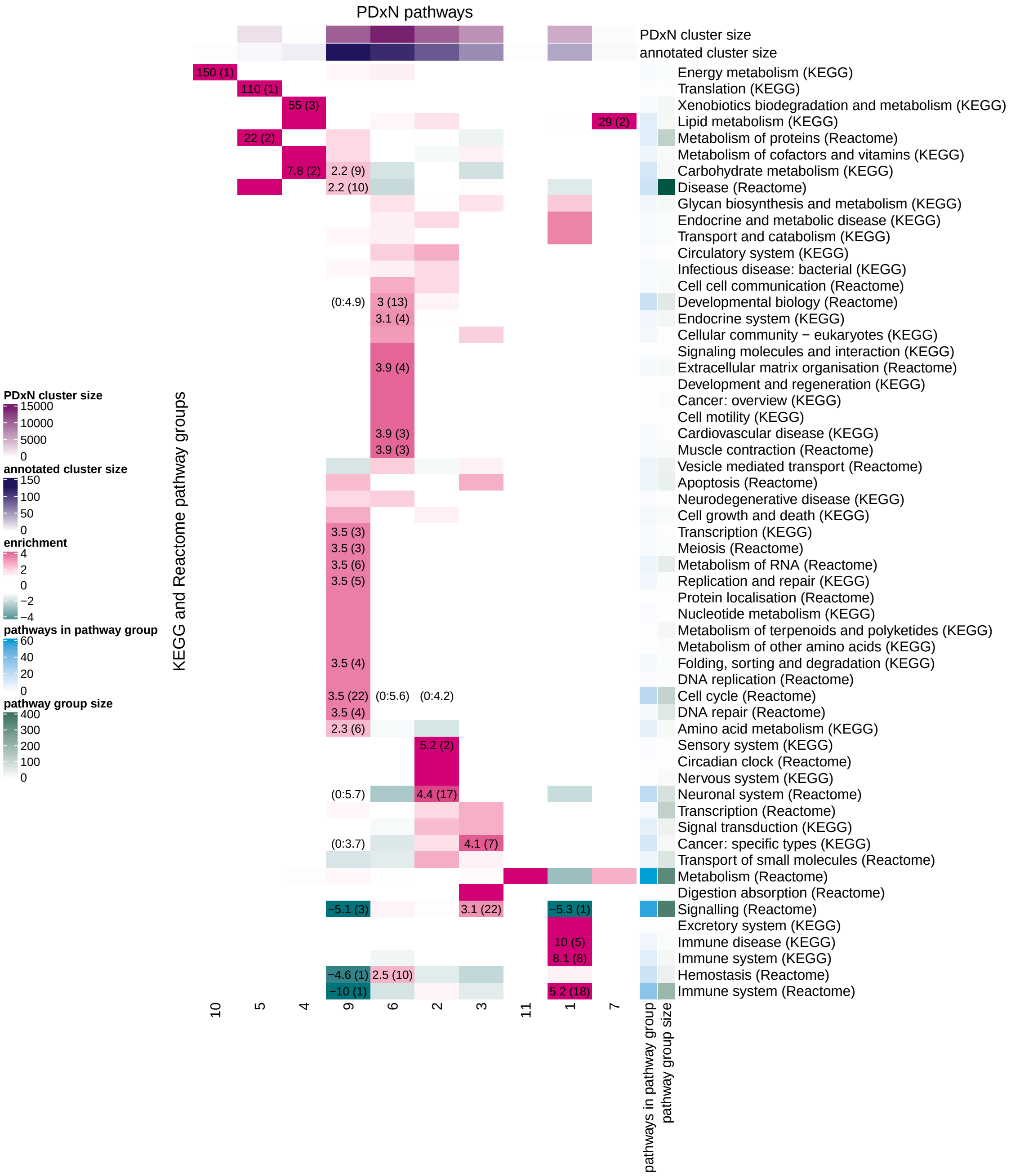


Fig. (5.10) KEGG and Reactome pathway annotation enrichment of PDxN bipartite clusters. The heatmap summarises enrichment (pink) or depletion (teal) scores. Absolute enrichment values of 0 or above 1.5 and below significance threshold p-value < 0.05 are displayed. The first number displayed is the enrichment or depletion score, while the number in () is the observed number. When the observed number is 0, the depletion score cannot be calculated, thus we displayed number-of-observed terms : the-number-of-expected terms in (). Cluster numbers are not corresponding to PCxN clusters. Column names are PDxN clusters as seen in Figs.5.8, 5.11 and Table 5.4. Some clusters show similar enrichment patterns to PCxN (PCxN-cluster↔PDxN-clusters: 2↔1, (1,4–6,11)↔(4,5,7,10,11), 7↔9, 8↔3, 10↔6). Row names are KEGG and Reactome pathway group names. Only clusters with at least one pathway group annotation are shown. KEGG — Kyoto Encyclopedia of Genes and Genomes; PCxN — Pathway Coexpression Network; PDxN — Pathway Drug Coexpression Network.

clustered away from other pathway groups in both PCxN (cluster 2) and PDxN (cluster 1) with relatively high enrichment scores (PCxN: KEGG 6.2, Reactome 5.5; PDxN: KEGG 8.1, Reactome 5.2). In addition to the *Immune system* group, there are also other immune related groups: *Infectious disease: bacterial* (KEGG: PCxN 6.2) and *Immune disease* (KEGG: PCxN 6.2, PDxN 10) that were highly enriched in either or both networks.

In addition, PCxN cluster 7 and PDxN cluster 9 overlapped in most enriched pathway groups involving genetic information processing: *DNA repair, Cell cycle, Meiosis, Metabolism of RNA, Transcription, Replication and repair,* and *Folding, sorting and degradation.* Each of these groups was enriched 3.1-times in PCxN and 3.5-times in PDxN. Both of these clusters were also enriched in the Reactome *Disease* group (PCxN 2.1; PDxN 2.2) which consisted of several viral life cycle pathways (Supplementary Table C.2).

Drug class annotation of PDxN clusters

With the possibility that there may be shared functional relationships between drugs, we annotated drug nodes with Anatomical Therapeutic Chemical (ATC) classification.

The World Health Organisation (WHO) Collaborating Centre for Drug Statistics Methodology (WHOCC)-controlled ATC classification system is a drug system that classifies the active ingredients of drugs according to the organ or system on which they act and their therapeutic, pharmacological and chemical properties. It is a hierarchical classification consisting of 5 levels, each level describing a particular property and encoded by a predefined set of digits or letters:

(i) ATC level 1: Main anatomical (or pharmacological) group, there are 14 groups, consists of one letter which is the first letter of the code.

(ii) ATC level 2: Pharmacological or therapeutic subgroup, consists of two digits.

(iii) ATC level 3 and 4: Chemical, Pharmacological or Therapeutic subgroup, consists of one letter each.

(iv) ATC level 5: Chemical substance, consists of two digits.

The predefined letters and digits from each level form the ATC code, with first letter indicating the level 1 group and last two digits the level 5 group, e.g. aspirin: B - *Blood and blood forming organs,* B01 - *antithrombotic agents,* B01A - *antithrombotic agents,* B01AC - *Platelet aggregation inhibitors excl. heparin,* B01AC06 - *acetylsalicylic acid.*

With the aim for one ATC code for each medical product, medicinal substances are classified according to their main therapeutic use. However, they sometimes have multiple ATC codes for different strengths or routes of administration with distinct indications. Combinations of two or more active ingredients in a medicinal substance also have their own ATC code, different to when they are prescribed individually. The coverage of the system is not comprehensive. An application has to be sent to the WHO to create a new ATC code. Obsolete drugs or drugs withdrawn from the market are kept in the system. A substance is not included if no request for inclusion has been received (WHOCC, 2018).

We have taken advantage of the hierarchical nature of the ATC classification system. We used KATdb to annotate drug IDs from the PDxN drug set with ATC codes. We matched 1294 BRD IDs to 2022 ATC codes. 634 annotated drugs were still in PDxN at q-value < 0.05 with 378 mapped to one ATC code, 105 to two, 77 to three and 74 to four or more. The drugs that mapped to the most ATC codes were hydrocortisone, dexamethasone and betamethasone. They matched to 17, 13 and 12 ATC codes, respectively with 21, 17 and 18 listed on the official ATC index website (*https://www.whocc.no/atc_ddd_index/*).

We annotated PDxN drugs with level 1 (Fig. 5.11) and level 2 (Supplementary Fig. C.1) ATC codes. Hierarchical relationship between level 1 and level 2 codes can be found in Supplementary Table C.3. Up- and down-regulated drug signatures were considered separately as we would not expect the up- and down-regulated signature from one or similar drugs to cluster together. Where there were multiple ATC codes for one drug, we considered the drug in all matched ATC categories. Each annotated drug could be present in multiple conditions: variations of cell-line, exposure time, concentration, and batch.

From 12 PDxN clusters, 7 (clusters 1–3,5–7,9) had at least one drug annotation in up- and down-regulated drug subset. When compared to pathway-level annotation, there were fewer overall enriched functional terms. Assessing level 1 enrichment (Fig. 5.11), two clusters had enriched terms in both up- and down-regulated subset. The smallest cluster 7 enriched in *Nervous system* and *Cardiovascular system drugs* when considering the up-regulated subset, and enriched in *Respiratory system drugs* in the down-regulated subset. Cluster 9 enriched in *Antineoplastic and immunomodulating agents* in the up-regulated subset, and in *Various drug classes* in the down-regulated subset. The level 2 ATC groups enriched in *Various drug classes* are *Diagnostic agents* (21/25 drug signatures), and *All other therapeutic products* (22/25 drug signatures) (Supplementary Fig. C.1, Supplementary Table C.3).

The rest of the clusters were either enriched in one subset (cluster up:1–2 and down:5–6), mostly depleted in one (up:9) or both (3) subsets or not showing any significant

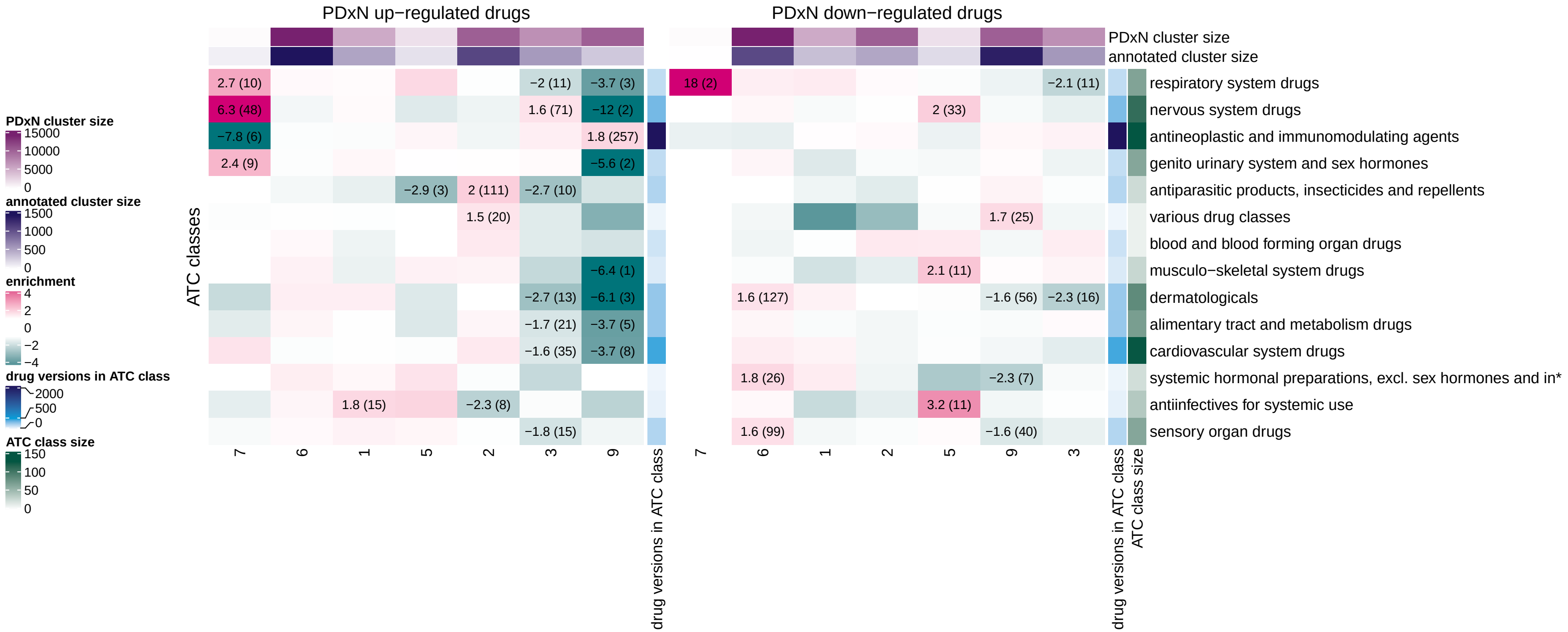


Fig. (5.11)

ATC level 1 drug annotation enrichment of PDxN bipartite clusters. The heatmap summarises enrichment (pink) or depletion (teal) scores for drug nodes consisting of either up- or down-regulated genes. Absolute enrichment values of 0 or above 1.5 and below significance threshold p-value < 0.05 are displayed. The first number displayed is the enrichment or depletion score, while the number in () is the observed number. When the observed number is 0, the depletion score cannot be calculated, thus we displayed number-of-observed terms : the-number-of-expected terms in (). Column names are PDxN clusters as seen in Figs.5.8, 5.11 and Table 5.4. Row names are ATC level 1 class names. Only clusters with at least one ATC class annotation are shown. * — ATC term was shortened, full terms are listed in Supplementary Table C.3; ATC — Anatomical Therapeutic Chemical; PDxN — Pathway Drug Coexpression Network.

enrichment/depletion patterns (up:6, down:1–2). In the up-regulated set, cluster 2 enriched in *Antiparasitic products, insecticides and repellents, Blood and blood forming organ drugs,* and *Various drug classes* (*Diagnostic agents* and *Contrast media*). Cluster 1 enriched in *Anti-infectives for systemic use*. In the down-regulated set, cluster 6 enriched in *Dermatologicals, Alimentary tract and metabolism,* and *Sensory organ drugs*. Cluster 5 enriched in *Anti-infectives for systemic use, Nervous system* and *Muscolo-skeletal system drugs*.

There were a few ATC level 1 classes that enriched in different clusters depending on the direction of the signature: *Various drugs* enriched in up:2 *(drug signature direction:cluster number)* and down:9, *Nervous system drugs* enriched in up:7, up:3 and down:5, and *Anti-infectives for systemic use* were up in 1 and down in 5. These relationships were interesting as the drugs belonging to that class consistently cluster together in both up- and down-regulated signatures. In both cases, the *Various drug class* enriched in the cluster where there was a clear imbalance between direction of signatures. In cluster 2, there were mostly up-regulated signatures and in cluster 9 there were mostly down-regulated signatures (Fig. 5.8).

We investigated how enrichment patterns within clusters relate pathway and drug annotations, suggesting pathway-level interaction with specific drugs or classes of drugs. For example, cluster 1 enriched in *Anti-infectives for systemic use* (ATC level 1 code: J) in the up-regulated drugs and immune response-related pathway groups. Another example was cluster 9 that enriched *Antineoplastic and immunomodulating agents* (ATC level 1 code: L) in up-regulated drugs as well as in genetic information processing pathway groups. The antineoplastic ATC class included drugs used for treatment of malignant neoplastic diseases, with drugs that prevent, inhibit or halt the development of tumours. Errors in genetic information processing pathways like *Transcription, Replication and Repair, DNA repair,* and *Cell cycle* have been a well-known mode of action for tumour development and cancer (Hartwell and Kastan, 1994). We observed that drugs that likely promote high fidelity of core cell processes clustered together with pathways involving those genes.

PDxN pathway and drug projection annotation

In addition to annotating bipartite clustered PDxN, we investigated the relationships between the same type of nodes by calculating PDxN projections for each set of nodes. Two nodes within the same node set were connected in the projection only if they had been connected to the same node in the opposite set in the original network. If a node pair within the same set of nodes connected to the more than one node in the opposite set, the multiple connections in the projection were represented as increased weight of the projected edge. We used the q-value < 0.05 threshold on edges to project only the most relevant parts of the network. Consequently, the projections included the information that was used downstream the repositioning pipeline. We summarised 4.3% of all the possible edges between pathway and drug node sets in PDxN by applying the significance threshold. We thus hypothesised that the projection would still provide insight in key relationships in the network.

We clustered each projection with the Louvain method (Blondel et al., 2008) for community detection and annotated each cluster with pathway and drug annotations. We clustered the up- and down-regulated signatures in the drug projection separately. The projections clustered in 3 clusters each. The pathway projection split into two similarly sized clusters of $\sim$700 nodes and one with one node. There were about 300 annotated pathways in the large pathway projection clusters. The up-regulated signatures in drug projection clustered into 3 clusters with $\sim$1500, $\sim$1000 and $\sim$500 annotated nodes. The down-regulated signatures clustered into 3 clusters with $\sim$2000, $\sim$1000 and $\sim$50 annotated nodes. Both projections were less granular likely due to losing weight information from the original bipartite network. The annotations in projection clusters (Supplementary Fig. C.2 for annotation of pathway projection clusters, and Supplementary Fig. C.3 for drug projection annotation at ATC level 1) showed similar trends to that of bipartite clustering.

Most pathway projection enriched terms overlapped with PDxN. *The Extracellular matrix organisation, Immune disease, Immune system* (Reactome, KEGG), *Hemostasis, Cancer:specific types* enriched in projection cluster 1 also enriched in clusters 1, 3 and 6 in PDxN. *Neuronal system, Disease, Carbohydrate metabolism, Cell cycle, Metabolism of RNA* are enriched in pathway projection cluster 2. *Neuronal system* also enriched in PDxN cluster 2 and the rest enriched in PDxN cluster 9. Only *Cellular community - eukaryotes* and *Apoptosis* pathway groups were enriched in pathway projection, but not in PDxN. *Cellular community* group clustered with *Extracellular matrix organisation* in projection cluster 1 and *Apoptosis* clustered with *Cell cycle* in projection cluster 2. Although the pathway annotations did not form as many well-defined clusters as in PDxN, they still showed separation into functional clusters with one or two main functions like genetic information processing or immune related pathway groups.

The drug projection showed similar patterns of mimicking PDxN clustering as the pathway projection. All drug projection enriched terms were also enriched in PDxN. The up-regulated drug projection sub-network showed enrichments only in cluster 3. The enriched ATC classes were: *Antiparasitic products, insecticides and repellents, Blood and blood forming organ drugs,* and *Cardiovascular system drugs*. The first two also enriched in up-regulated subset in PDxN cluster 2 and the last enriched in cluster 7. The down-regulated drug projection sub-network enriched in: *Anti-infectives for systemic use, Nervous system drugs* and *Muscolo-skeletal system drugs* in projection sub-network cluster 1 and also in PDxN cluster 5.

Although information was lost due to dimensionality reduction, we still observed similar functionally annotated clustering patterns between PDxN and its projections.

5.3.6 Annotation method limitations

We demonstrated functional clusters in both pathway and drug subset, nonetheless, we considered the limitations of each annotation method.

We used KEGG and Reactome pathway groups to annotate PCxN, PDxN and its pathway projection. Both these resources are manually curated. Curation of pathways is an ongoing process and, with time, we will be able to annotate pathways with more accuracy by re-defining established pathways and splitting generic pathways into more detailed processes. An example of pathway curation is the C2 Canonical pathways from MSigDB version 6.2 (July 2018) used in this thesis, which has been updated to include 2199 (increase of 738 pathways) in version 7.0 (August 2019). To avoid misannotation we only annotated pathways if the KEGG/Reactome name matched perfectly to MSigDB version 6.2. We were only able to annotate 56% of PDxN pathways (52% in PCxN). In addition, not all pathways had clear membership to a particular larger pathway group as the cell processes are interlinked, thus not all pathways can be annotated with a single larger functional group.

The drug annotation comes with its own disadvantages. The main drawback of using ATC classification is that the system first splits drugs by main anatomical group rather than the mode of action. For example, anti-infectives can be found not only in ATC level 1 class J: Anti-infectives for systemic use, but also in 14 other ATC classes across 6 different level 1 groups. The mode of action is considered on lower levels of the classification. ATC represents an established drug classification system (World Health Organization et al., 2003). Despite these disadvantages, it is still a well-defined drug classification system that was used to gain insight into functional grouping of drugs.

Whilst there are limitations in the annotation approach taken, we were able to find functional consistency between PCxN, PDxN and its projections. We argue that PDxN represents functionally meaningful relationships within and between pathways and drugs.

5.4 Disease Signature Generation

In order to utilise PDxN in drug repositioning, we developed a set of accompanying steps. PDxN provides a way to investigate the general functional landscape and relationships between pathways and drug gene sets. We developed a disease signature generation step that generates a pathway signature that is used to query PDxN and provide a closeup on the disease signature related processes. The PDxN drug repositioning pipeline is disease agnostic in its nature, meaning that we can apply it to any disease as long as we can generate a disease signature. Here we propose a pathway disease signature generation method that consists of three main parts: a pathway summary step, differential pathway expression and from that, identification of the most up- and down-regulated pathways (Fig. 5.12).

A pathway-centred approach was taken, as transforming gene expression data to a pathway space is expected to yield a more robust representation of the data in which technological and biological variance across samples is reduced (Guo et al., 2005). As outlined in Subramanian et al. (2005) gene-level differential expression (DE) approaches have a few major limitations:

(i) after multiple hypotheses testing, no individual gene may meet the significance threshold,

(ii) alternatively, a long list of genes within a significance threshold could be yielded with no underlying biological theme, followed by ad hoc interpretation depending on a biologist's area of expertise,

(iii) single gene analysis may miss important effects on pathways,

(iv) study of the same biological system with different data sets might result in low overlap of genes below the significance threshold.

A pathway-centred approach can overcome some of these limitations, by looking at the consensus expression of gene members. While expression of individual genes in a pathway may vary considerably across samples with similar phenotypic characteristics, expression of the pathway as a whole may become consistent across the samples, thus

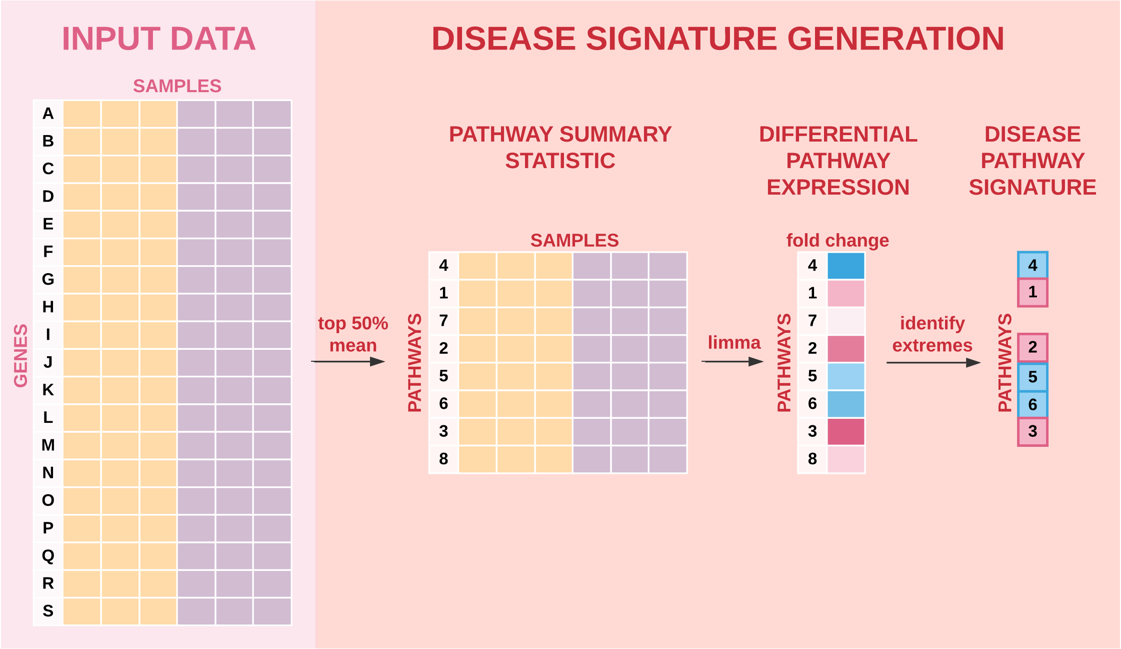


Fig. (5.12) Disease signature generation overview. Disease input gene expression data is processed to generate a disease signature. First genes are summarised into pathway-level expression with the top 50% mean method. The pathway summary score is then analysed with `limma` (Ritchie et al., 2015) to yield a list of differentially expressed pathways. The extremes, the most up- and down-regulated pathways, are identified and used as a disease pathway signature used to interrogate PDxN. Two sample groups, case and control (yellow and purple), are required for the disease signature generation step. Pink fold change indicates up- and blue down-regulated pathways. limma — Linear models for microarray data.

giving more importance to multiple genes that display a similar change in expression rather than a single gene that changes expression independently.

To generate a disease pathway signature, we first require disease and control gene expression data. Both microarray and RNA-Seq expression data can be analysed with the current pipeline. First, the gene expression data was quality-controlled, so that outliers or poor-quality samples were removed. To make RNA-Seq comparable to microarray data, we log-transform normalised and quality-controlled RNA-Seq with `voom` from linear models for microarray data (`limma`) (Ritchie et al., 2015). The gene expression data was then *z*-scaled for each gene across samples and then summarised to a pathway-level with the top 50% mean summary statistic (Hwang, 2012). The summarised pathway expression scores were then analysed with `limma` package (Ritchie et al., 2015) to identify differentially expressed pathways. The top 5, 10, 15, and 20 most up- and down-regulated pathways according to the fold change were defined as the disease signatures. The disease pathway signatures were then used to explore the disease PDxN sub-network and identify potential drug candidates.

5.4.1 Pathway summary statistic

The key step of the disease signature generation was the pathway summary statistic. There are several ways gene expression data is used in pathway analysis, most focusing on gene set enrichment, e.g. Gene Set Enrichment Analysis (GSEA) (Subramanian et al., 2005), gene set variation analysis (GSVA) (Hänzelmann et al., 2013) or analysis of sample set enrichment scores (ASSESS) (Edelman et al., 2006). An alternative approach is pathway aggregation methods, that transform gene expression data from gene to pathway level. The pathway-level scores can then be used in downstream analysis applying analysis pipelines normally applied to gene-level data. Hwang (2012) systematically compared the 6 most prominent aggregation methods. They assessed performance of 5 existing methods and proposed a 6th, the mean top 50% method. The mean top 50% method is a variant of a simple all mean approach, where the mean of all member genes is used as a summary statistic. They show that Mean top 50% and ASSESS (Edelman et al., 2006) are the two best performing methods at classification accuracy and correlative extent of pathway signatures between the dataset pairs. ASSESS is considered to be a sample-level extension of GSEA (Subramanian et al., 2005) as they both calculate enrichment scores for each pathway. We chose the mean top 50% as our pathway aggregation method as it achieved the highest accuracy rank in both their external and internal validation.

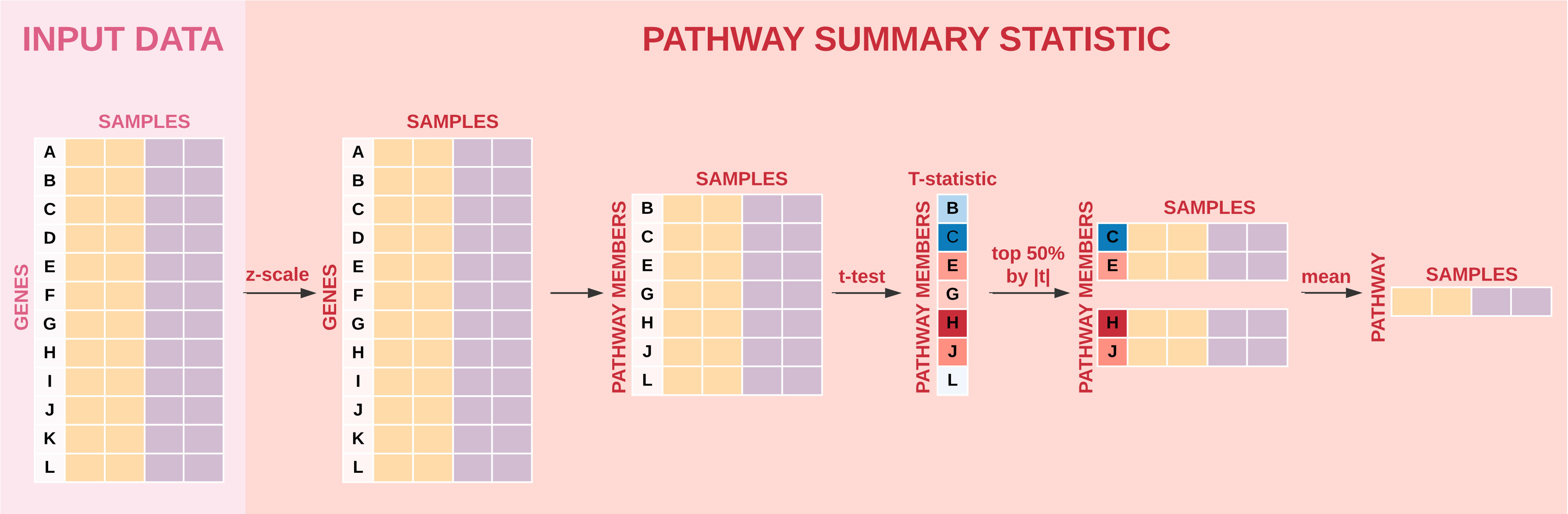


Fig. (5.13) Pathway summary statistic. In the first step of disease signature generation, the input gene expression data is summarised into pathway-level expression. The top 50% mean method modified from Hwang (2012) is used as a pathway summary statistic. First, genes are z-scaled across samples so that each gene has mean expression $= 0$ and standard deviation $= 1$. Then each pathway is considered individually. Welch's t-test is performed on pathway member genes. The genes are then ordered by absolute value of t-statistic, $|t|$. The z-scaled expression of the top 50% of genes is averaged for every sample. The pathway profile is thus a pathway summary score for each sample. Two sample groups, case and control (yellow and purple), are required for the pathway summary statistic method. Red indicates a positive and blue a negative t-statistic. Diagram modified from Hwang (2012).

The mean top 50% (Fig. 5.13) method takes z-scaled gene expression data. The data is scaled for each gene across all samples, so that the gene expression has mean of 0 and standard deviation $= 1$. For each pathway in a pathway set, the scaled gene expression matrix is subsetted to pathway gene members. Welch's t-test (Welch, 1947) is performed on the pathway gene member sub-matrix, identifying the genes that are most different between two conditions. The genes are ranked by absolute t-statistic, $|t|$. The expression of the top 50% of genes with highest $|t|$ is then averaged, yielding a pathway profile, consisting of a pathway score for each sample. The summary statistic is repeated for every pathway in a pathway set resulting in a $p \times n$ matrix, where p is the number of pathways and n is the number of samples.

5.4.2 Differential pathway expression and disease pathway signature

The pathway-level summarised expression data mimicking the traditional gene expression data could be applied to well-established tools for downstream processing of gene-level data. We used limma (Ritchie et al., 2015) for one main reason: the R package provides integrated comparable analysis pipelines for microarray and RNA-Seq gene expression data.

Limma is as the name suggests, based on linear modelling. It requires a matrix of expression values, where each row represents a genomic feature relevant to the current study, and each column corresponds to a sample. It fits a linear model to each row of data. The empirical Bayes framework borrows information across genes in a dynamic way to smooth out variances, allowing for different levels of variability between genes and between samples. The method uses posterior variances in a t-test setting, making statistical conclusions more reliable when the number of samples is small. The Benjamini-Hochberg correction is applied to estimate the false discovery rate (FDR). The `voom` function applied to RNA-Seq data, $\log_2$-transforms the normalised counts and estimates the mean-variance relationship for the transformed data to assign weights for each observation. The limma-voom approach makes RNA-Seq count data comparable to microarray datasets. It allowed the application of the same methods to both types of data in the downstream analysis. In addition, using limma for both, microarray and RNA-Seq, meant that the same statistical tests were applied to both data types.

We analysed our pathway-summarised matrix with `limma` and identified differentially expressed pathways that meet a significance threshold FDR < 0.05. We ordered the pathways by decreasing log fold change (logFC), where the top pathways represented the most up-regulated pathways between case and control and the bottom pathways (with the lowest logFC) represented the most down-regulated pathways. We took top n most up- and down-regulated pathways as disease pathway signatures used in downstream analysis for disease-specific drug prioritisation. We generated prioritised drug lists for drug signatures with $n = 5, 10, 15, 20$.

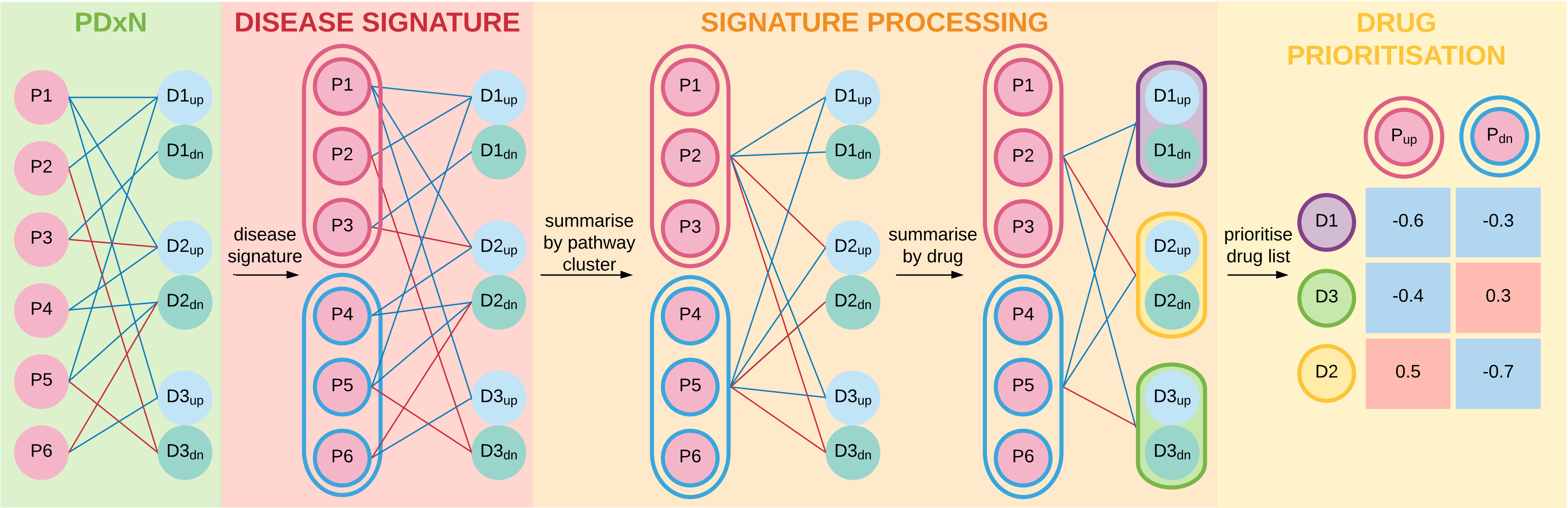
5.5 Signature Processing and Drug Prioritisation

The signature processing step brings together a pre-calculated PDxN and a disease-specific pathway signature. In this step, the edges between pathway and drug nodes were summarised (Fig. 5.14). To preserve directionality of the disease pathway signature, we separated the up- and the down-regulated pathways into two pathway clusters (an up- and a down-regulated cluster). While we took inspiration from the PDN prototype method to consider pathways together (Joachim et al., 2018), we deviated from their approach by focusing on summarising correlation estimates rather than the p-values.

The PDxN sub-networks were constructed for each pathway cluster. Each sub-network consisted of pathway nodes that were in a pathway cluster and all drug nodes that were connected to the cluster pathways. The correlation edges between cluster pathways and drugs were then summarised. First, the edges were summarised by pathway cluster, meaning that we join individual pathways in a cluster, forming a pathway cluster node. The edges, which connect individual pathways in a cluster with drug-direction nodes, were summarised by taking the mean of the edge weights, resulting in one edge between each *pathway cluster↔drug-direction* node. In the next step we summarised the correlation by drug, so that we join the up- and down-regulated drug nodes into one. We took the difference between *pathway cluster↔up-regulated drug* and *pathway cluster↔down-regulated drug* so that we derived one summarised edge for each *pathway cluster↔drug*.

5.5.1 Score interpretation

To interpret the resulting score, we applied the signature-driven hypothesis (Fig. 2.4), where we hypothesised that a treatment with a drug working in the opposite direction than the disease (compared to control) would drive the diseased system back to a healthy balance.



	P_{up}	P_{dn}
D1	-0.6	-0.3
D3	-0.4	0.3
D2	0.5	-0.7

Fig. (5.14) Interrogating PDxN with the disease signature. The pathways in the disease signature are used to summarise the edges between disease pathway (P) and drug (D) nodes. In the signature processing step, the up- and down-regulated differentially expressed disease pathways are used to form two separate clusters. First, the edges are summarised across pathways in a pathway cluster, then the up- and down-regulated drug nodes (D_{up}, D_{dn}, respectively) are summarised. Resulting in one edge between each *pathway cluster↔drug* pair. The summarised edges are then used to generate a prioritised drug list. PDxN — Pathway Drug Coexpression Network.

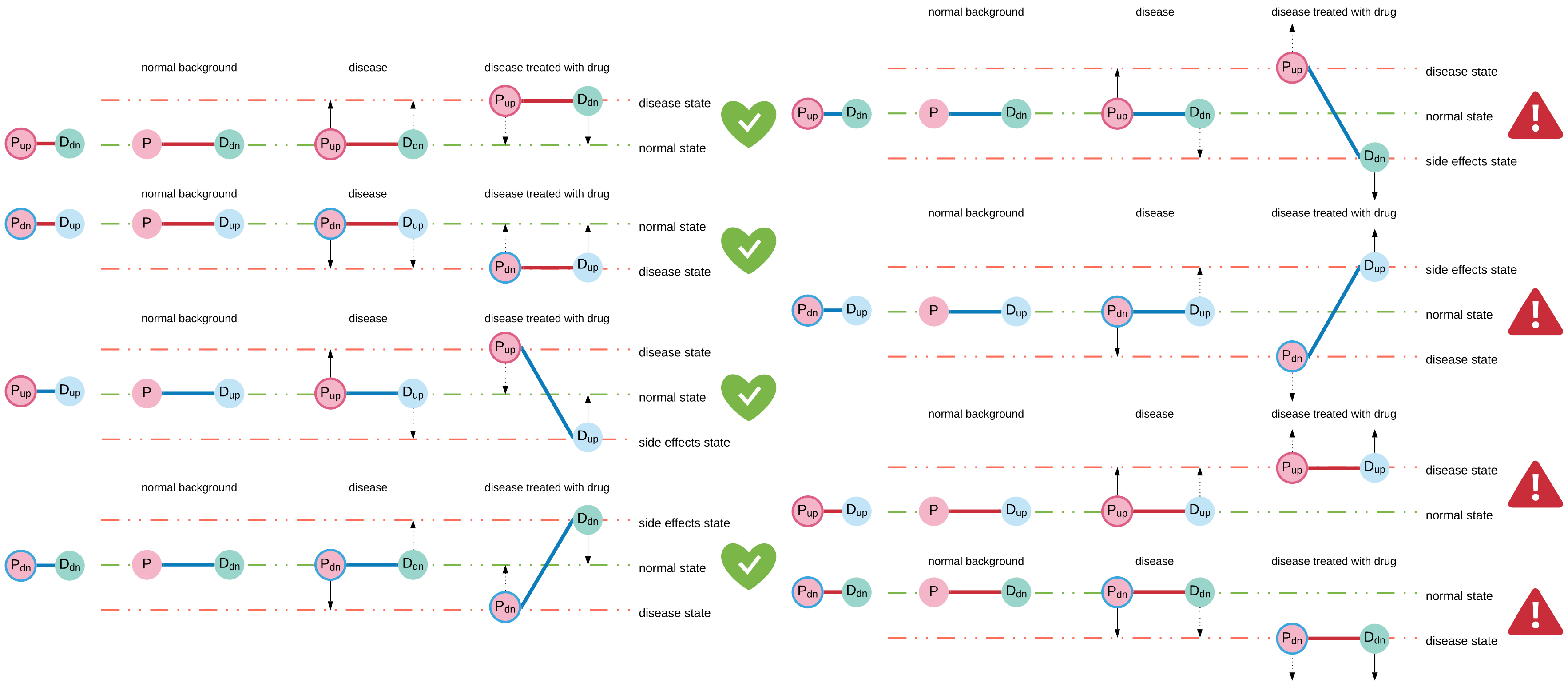


Fig. (5.15) Predicted health outcomes based on pathway and drug directionality in PDxN. Each pathway node (P) in PDxN has the potential to be up- or down-regulated in a particular disease signature. Each pathway node can be positively (red edge) or negatively (blue edge) correlated to an up- or down-regulated drug node (D_{up}, D_{dn}, respectively). Listed are the 8 possible outcomes supported by the signature hypothesis where two opposite dysregulated states can move to the normal state when combined. We assume that positively correlated node pairs will move in the same direction in disease and drug state, where negatively correlated pairs will move into opposite directions. We prioritise the drug scores based on returning to normal or healthy state following this diagram (Supplementary Fig. C.4). Full arrow represents the initiated movement by the nature of the node e.g. up-regulated pathway in disease will move upwards. The dotted arrow represents the followed movement of the other node based on the edge correlation value i.e. the same direction for positively correlated pairs and opposite for negatively correlated pairs. PDxN — Pathway Drug Coexpression Network.

We were interested in identifying a disease signature that was the opposite to the drug signature (Fig. 5.15), therefore we would expect the best success from:

(i) an up-regulated drug signature that was positively correlated with the down-regulated pathway cluster,

(ii) a down-regulated drug signature that was positively correlated with an up-regulated cluster,

(iii) an up-regulated drug signature that was negatively correlated with the up-regulated pathway cluster,

(iv) a down-regulated drug signature that was negatively correlated with the down-regulated pathway cluster.

When we considered the summarised drug score, that we calculated by taking the difference between an *up-* and a *down-regulated drug↔pathway cluster* score, we would expect a final negative score for the up-regulated pathway cluster and a positive score for the down-regulated pathway cluster to yield beneficial outcomes (Fig. 5.15).

5.6 Benchmarking

An important step in the development of any new method is benchmarking. Benchmarking systematically assesses the performance of a given method. In this drug repositioning pipeline we benchmarked the final output, which was the prioritised drug list. The prioritised drug list was assessed on its ability to prioritise already known drugs with beneficial effects on a disease of interest. Where available, we used the approved indications by the US Food and Drug Administration (FDA) and the European Medicines Agency (EMA) as a true positive list. If no drugs had been approved for the disease of interest, we extended the true positive list to drugs predicted by a disease expert to have a beneficial effect on the underlying mechanism of disease.

5.6.1 Score evaluation

A receiver operating characteristic (ROC) curve was generated for each cluster, summarising its performance by plotting the sensitivity (TPR) and specificity (1 - FPR). The ROC curve illustrated the ability of a cluster score to recall the true positive drugs. The area under the ROC curve (AUC) was calculated for each cluster, where AUC = 1 indicated perfect performance, i.e. the true positive drugs had the best score and all other drugs had a lower score, and AUC = 0.5 indicated randomness, i.e. drugs were prioritised at random. Therefore, a cluster with a higher AUC score was more successful at predicting true positive drugs than others.

5.6.2 Assumptions

There were two main assumptions made in this benchmarking approach:

(i) the approved drugs treat the disease and not its symptoms,

(ii) the predicted, prioritised drugs are all true negatives.

Drugs providing symptomatic relief often present the majority of approved drugs for use in a particular disease, thus the first assumption might have lowered the "true" performance score as our data-driven predictions based on the pathway disease signature were aimed at treating the disease and not the symptoms. As we did not design our system to prioritise drugs for symptomatic relief, but rather to reverse the disease state back to a healthy balance, the drugs that are approved for symptomatic relief e.g. pain relief for headache resulting from a brain tumour, were likely to lower the algorithms performance score. A possible solution to account for this limitation is to curate a set of drugs that are targeting the disease mechanism and not only the symptoms. However, not all disease mechanisms or drug modes of actions are well-defined, and not many diseases have approved medications that target only the disease itself.

The second assumption is contradictory to the aim and purpose of this study. The method was designed to predict novel drug candidates for a given disease, while the benchmark assumed that the predicted drugs were true negatives. It is possible that if the method predicted drug candidates that could be validated later, but have not yet been approved, the initial benchmarking performance of the algorithm would be low. It is thus important to re-evaluate the performance of any method with any results from validation.

Although there were two significant assumptions made in this benchmarking approach, it was the best available systematic assessment method to guide us in identifying best-performing methods.

5.7 Testing Drug Candidates in Disease Models

Testing drug candidates in disease models was out of scope for this thesis, however, the results from the work described in Chapters 6 and 7 have been utilised for potential validation by our wet-lab collaborators at the Harvard Medical School and the Sheffield Institute for Translational Neuroscience.

The prioritised drug lists from our pipeline were further investigated to identify top drug candidates. The top scoring drugs could be manually curated based on their availability, toxicity, previous test results and known beneficial effects. The available drug candidates could then be tested by our collaborators in disease models. Depending on funding and facilities available, a selected number of top scoring candidates from each pathway cluster will be tested. A subset of drugs could be further investigated in a combination of different conditions: varying disease model age, concentration and exposure time.

Any drug screen results could be used to additionally benchmark and characterise the drug repositioning pipeline and influence further method development.

5.8 Discussion

In this chapter, we have described the drug repositioning pipeline, that is the core focus of this thesis, we also provided detailed description and reasoning for each of the pipeline components, and characterised the underlying network that powers the identification of novel drug candidates.

5.8.1 PDxN strengths and limitations

There are several limitations and points of improvement to the PDxN method. Although we preserved directionality of signatures where possible, we did not account for internal pathway structure. Whilst, we heavily relied on the correctness of the pathway annotations, we also included data-driven static modules and drug gene sets derived from gene expression experiments. Nevertheless, the system requires continuous version updates with new versions of pathway annotations, as both the size and quality of annotation databases increase with time (Wadi et al., 2016). Although we defined drug gene sets from gene expression experiments, those drugs were mostly tested on cancer cell lines and other immortalised cell lines and were thus unlikely to faithfully represent the reaction of either control or diseased cells. An additional weakness of the drug-perturbation database is that due to logistics, only a subset of the drugs tested have been performed in a systematic set of conditions. This is understandable given the scale of the project and the database ($\sim$1.3 million signatures) as it would be an unrealistic expectation for all the conditions tested within the same laboratory with the same drug batches. Even for those, the signatures may lack consistency, given the high variation between replicates. The most frequently tested conditions in the drug-perturbation datasets pose their own limitations. For example the most commonly tested concentration of 10μM is likely unsuitable for use in humans so any conclusions made from those experiments would require further testing and dose optimisation. Similarly the exposure times currently tested do not consider the absorption rate or bioavailability of a drug in a human body.

Another important limitation is the gene expression background data used in the PDxN methodology. It involved a curated set of microarray data, and so it was limited to a small subset of available gene expression data as it did not yet include any experiments from the growing plethora of RNA-Seq data. In addition, we currently looked for consistent relationships between gene sets across many tissues. While this gave us insight into global relationships, it would be beneficial to look at tissue-specific background gene expression data for drug repositioning using PDxN.

Because we were interested in capturing a global overview of relationships that could be interrogated in our drug repositioning pipeline, we reduced dimensionality in multiple summarising steps. We reduced the information carried onward in the pipeline with each step. Therefore, we were likely losing some meaningful, and keeping some irrelevant or random, information in each step. We acknowledge these limitations in the final result interpretation as noise is likely to accumulate.

In addition to biological implications of the current method, we considered its computational limitations. PDxN has already significantly reduced the computational requirements of the PCxN method, while maintaining the original idea. The PCxN method was designed for 1330 gene sets taking approximately 100h of computational time. We were able to consider over 50,000 gene sets by limiting the relationships calculated, restructuring the code, increased parallelisation, and calculating subsets of the network and combining them at the end. The current version with 53,721 nodes and 76,961,304 edges was approaching the limit with currently available computational resources. It would require further computational improvements if larger datasets were to be implemented. However, the rapid development in computational power over the last decade is promising for the implementation of large datasets.

Despite these limitations, we have shown here, and in Pita-Juárez et al. (2018) that the method captured biologically relevant information. Here we compared PCxN to PDxN and found several topological and functional similarities. While PCxN consisted of pathway$\leftrightarrow$pathway relationships, PDxN only considered relationships between pathway and drug nodes. Whilst, PDxN was topologically similar to the much smaller PCxN, PDxN included $\sim$36-times as many nodes and $\sim$71-times as many edges as PCxN. As in PCxN, the PDxN clusters enriched in similar functionally-related pathway groups. The PDxN pathways formed similar clusters to PCxN, suggesting that the PDxN captured comparable biologically relevant information despite its bipartite nature. Annotation enrichment in drugs showed fewer clear patterns. However, the annotation patterns were present also when investigating the PDxN pathway and drug projections. Through functional annotation of PDxN clusters, we have shown that PDxN captured functionally meaningful relationships and can thus confidently use PDxN in further analysis.

5.8.2 Disease signature strengths and limitations

We have developed a disease signature generation method that was designed to identify key dysregulated pathways in a disease data set. We implemented a pathway-level approach in order to capture the consensus expression of gene members rather than individual gene dysregulation. As in PDxN, the method not only relied on correctness of curated pathway annotations, but also included data-derived static modules, allowing us to counterbalance inherent curation bias. However, the approach did not take into account pathway topology, which has shown to improve performance of pathway-based methods (Nguyen et al., 2019). We preserved directionality, which is lost by most enrichment-based methods, by generating up- and down-regulated sets of pathways that could then be interrogated separately in the downstream PDxN analysis.

By keeping this step of the drug repositioning pipeline separated from the underlying relatively static PDxN, we enabled a drug repositioning approach that was not restricted to a particular disease. Any use-cases could be considered upon disease signature generation from case and control gene expression data. Given the method's reliance on use-case expression data, we provide additional disease signature characteristics and biological interpretation of the results in the following chapters.

5.8.3 Drug prioritisation strengths and limitations

We have implemented a correlation-based drug prioritisation, that leveraged the PDxN topology and use-case defined disease pathway signatures. The approach preserved directionality by considering up- and down-regulated disease pathways separately. It generated PDxN sub-networks between the up- or down-regulated pathways and the drug nodes connected to those. It consisted of several summarisation steps that reduced the PDxN correlation connections into one summary score. While the approach leveraged the PDxN topology, it only considered the disease pathways' first neighbours, overlooking all other topological information.

Our direction-sensitive approach allowed the implementation of the signature reversal hypothesis (Lamb et al., 2006), which is based on prioritising drugs that have an opposite signature to that of the disease, so that together their signatures counterbalance toward a healthy state. Further strengths and limitations are explored in the following chapters. We evaluated this method's performance and identified the top prioritised drug candidates in three different case studies described in the following chapters.

In this chapter, we have identified several strengths and limitations to the current approach. We have characterised the underlying PDxN, however in order to further explore all the remaining components of the drug prioritisation pipeline we have applied it to three case studies: juvenile idiopathic arthritis, Alzheimer's and Parkinson's disease.

Chapter 6

Evaluation of the System: Application to juvenile idiopathic arthritis (JIA)

In this chapter we characterise the drug repositioning pipeline components by applying them to a juvenile idiopathic arthritis (JIA) case study. We applied methods described in Chapters 3 and 5 to publicly available JIA datasets.

In this chapter, we explored the method's strengths and weaknesses by applying it to a large set of studies. In the following chapter, Chapter 7: Case Studies: Neurodegenerative Diseases, we applied the same pipeline to another two case studies: Alzheimer's and Parkinson's disease. These two chapters follow a similar structure with the following chapter focusing on neurodegenerative disease drug repositioning pipeline results.

Others' contributions to this chapter. Professor Lester Kobzik (Department of Environmental Health, Harvard T.H. Chan School of Public Health) assisted in the curation of the publicly available juvenile idiopathic arthritis studies and advised on the interpretation of the results in relation to the standard treatment practice.

6.1 Disease Introduction

Juvenile idiopathic arthritis (JIA) describes a group of clinically heterogeneous chronic arthritic diseases of unknown cause, present for a continuous period of at least 6 weeks in juveniles less than 16 years old (Petty et al., 2004). JIA is the most common chronic rheumatic disease in children and can cause short- and long-term disability. It has prevalence of approximately 1 per 1000 in developed countries (Beukelman et al., 2011). The cause of JIA is poorly understood but it is thought to be related to both genetic and environmental factors, which lead to disease heterogeneity. It has been divided by the International League of Associations for Rheumatology (ILAR) into several subcategories with distinct presentation, clinical manifestations and genetic backgrounds (Giancane et al., 2016). Although ILAR classification is based on the biological basis of the subtypes, there is increasing evidence for heterogeneity within JIA subtypes (Fall et al., 2007).

In this chapter we focused on systemic JIA (sJIA), characterised by systemic inflammation including recurrent fever and rash. It involves arthritis in one or more joints with or preceded by fever (Akioka, 2019). Arthritis is swelling within a joint, or limitation in the range of joint movement with joint pain or tenderness (Petty et al., 2004). This subtype is the most likely to combine with macrophage activation syndrome — a life-threatening complication. sJIA patients show a different inflammatory profile compared to the other subtypes (Akioka, 2019). In particular, the concentrations of interleukin-6 (IL-6) are increased in patients during active systemic disease and correlate with the extent of joint involvement (de Benedetti et al., 1991). The overproduction of IL-6 can also explain many of the extra-articular manifestations of this disease, such as stunted growth (de Benedetti et al., 1997).

In addition to sJIA datasets, we analysed two that present as Immunoglobulin M (IgM) rheumatoid factor negative (RF-) polyarticular JIA (polyJIA). RF- polyJIA is characterised by having 5 or more affected joints within the first 6 months of disease with the absence of the RF (Petty et al., 2004). RF- polyJIA shows higher prevalence in girls with its onset more common in early childhood (Ringold et al., 2014). Patients with sJIA who later develop arthritis in multiple joints can all have polyarticular disease but are excluded from the polyJIA subtype based on the ILAR classification. polyJIA children tend to have a more complex course of treatment compared to children with fewer affected joints. They are at increased risk for joint damage, resulting in poorer functional outcomes and decreased quality of life (Oberle et al., 2014). Several genetic risk loci have been identified indicating increased susceptibility to JIA, many within the human leukocyte antigen (HLA) region (Ombrello et al., 2015).

None of the current available drugs have curative properties, but they greatly improve disease management. Treatment options include non-steroidal anti-inflammatory drugs (NSAIDs), intra-articular glucocorticoid injections, traditional disease modifying antirheumatic drugs (DMARDs), and biologic therapy, including tumour necrosis factor (TNF) inhibitors, interleukin-1 (IL-1), and IL-6 inhibitors (Beukelman et al., 2011).

Treatment is tailored to control the disease and limit disability while balancing these with excessive immunosuppression and side effects from these medications (Ravelli et al., 2018). The more joints that are involved in the disease or the more severe the systemic symptoms, the greater the amounts of immunosuppressant necessary, causing greater side effects (Harris et al., 2013).

Although the functional outcomes have greatly improved over the last few decades, approximately 40% of sJIA and one third of RF- polyJIA patients develop moderate or severe disability. In addition, JIA often extends into adulthood where higher unemployment rates among patients are observed, suggesting difficulty adapting to adult life (Oen et al., 2002). It is thus necessary to establish better treatments that would increase recovery and decrease long-term disability.

6.2 Case Study Design

In this chapter, we focused on generating prioritised drug lists with the Pathway Drug co-expression Network (PDxN) pipeline for treatment of sJIA. We first generated pathway-level disease signatures from a curated set of sJIA and RF- polyJIA studies (Section 5.4). We then applied the sJIA disease signatures to the signature processing method (Section 5.5) and evaluated the resulting prioritised drug lists by comparison with an approved drug list. We explored the sensitivity of the PDxN method by further analysing the prioritised drug lists and comparing our results to an online Library of Integrated Network-based Cellular Signatures (LINCS) query tool (*https://clue.io/*), an alternative gene-based method based on the same drug signatures.

In order to be able to differentiate between studies representing the same disease and tissue, we referred to studies by their Gene Expression Omnibus (GEO) accession number (starting with GSE). We treated GSE88650 as two separate studies, because the samples were profiled on two different platforms (GPL96, GPL97) with discrete probe sets. We provided disease, tissue and array annotations on all figures. We referred to microarray platforms by their GEO accession number (starting with GPL) and referred to GPL11154 (Illumina HiSeq 2000) as RNA-Seq.

We focused on 8 different JIA studies: 6 included sJIA, and 2 RF- polyJIA samples (Table 6.1). Two sJIA studies included samples from the whole blood, the rest were profiled in peripheral blood mononuclear cells (PBMC), a fraction of the whole blood. Most studies were analysed on GPL570. In addition to JIA studies, we included an unrelated microarray dataset consisting of liver samples from old and young individuals (GSE133815, Table 3.2) to serve as the control study. We used it as a control in the disease signature generation step because it represented a study from the same organism, on the most commonly used platform in our curated selection of JIA studies, but of non-inflammatory condition.

Table (6.1) Overview of JIA studies. The control and disease sample numbers in the table reflect the number of samples post quality control (QC). Sample numbers before QC are listed in Table 3.2. * — in GSE26554, the 3 sJIA samples were removed due to the low number, instead the larger group of RF- PolyJIA samples was used. GPL570 — U133 Plus 2.0 Array; GPL96 — U133A Array; GPL97 — U133B Array; GPL11154 — Illumina HiSeq 2000; JIA — juvenile idiopathic arthritis; PBMC — peripheral blood mononuclear cells; RF- polyJIA — rheumatoid factor negative polyarticular JIA; sJIA — systemic JIA.

GEO accession	Platform	Tissue	Control samples	Disease samples	Disease	Reference
GSE15645	GPL570	PBMC	12	13	RF- polyJIA	Knowlton et al. (2009)
GSE26554	GPL570	PBMC	22	3*, 35	sJIA*, RF- polyJIA	Thompson et al. (2012)
GSE20307	GPL570	PBMC	52	20	sJIA	Barnes et al. (2010)
GSE21521	GPL570	PBMC	26	18	sJIA	Hinze et al. (2010)
GSE7753	GPL570	PBMC	27	17	sJIA	Fall et al. (2007)
GSE8650	GPL96	PBMC	21	14	sJIA	Allantaz et al. (2007)
GSE8650	GPL97	PBMC	21	12	sJIA	Allantaz et al. (2007)
GSE112057	GPL11154	whole blood	12	26	sJIA	Mo et al. (2018)
GSE80060	GPL570	whole blood	22	21	sJIA	Brachat et al. (2017)

6.2.1 Representative studies

As there were several studies being analysed separately, we, for clarity, chose only two representative studies, GSE7753 and GSE112057, that we discussed in detail. We interpreted both studies for disease signature characterisation and only GSE7753 for further downstream analysis. Both representative studies included samples from sJIA patients. GSE7753 samples were extracted from PBMC and analysed on Affymetrix Human Genome U133 Plus 2.0 Array (GPL570) which were the most common tissue and array in the curated set of JIA studies. GSE7753 thus represents the most common sJIA subtype, tissue and platform among the curated set of JIA studies considered in this chapter. GSE112057 was selected because it was the only curated RNA-Seq sJIA study.

We generated disease signatures for GSE112057, but the study was excluded from further downstream analysis because we focused our analysis on only sJIA PBMC studies, and GSE112057 consisted of tissue samples from whole blood. While we showed summarised results of the curated studies, we also show a further break-down of the representative study results.

6.3 Disease Signatures

We generated signatures (Sections 3.3 and 5.4) for a set of curated sJIA and polyJIA gene expression studies. An additional non-JIA microarray dataset was analysed to serve as control in characterisation of the disease signature generation. The studies were profiled on 4 different platforms in three different tissues. We generated pathway signatures for all 10 studies and gene-level signatures for 2 representative studies (one microarray and one RNA-Seq JIA).

6.3.1 Pathway signatures

We employed a pathway-level approach, as it was expected to yield a more robust representation of the gene expression data as the technological and biological variance across samples is reduced (Guo et al., 2005). The pathway-level analysis captured the consensus expression of gene members, rather than individual gene variance. The pathway signatures generated were compatible with the PDxN system, thus allowing identification of potential drug candidates.

Pathway signature correlation and overlap

We investigated the sensitivity of the pathway disease signature generation method, before applying the signatures in downstream analysis. To determine whether the signature generation method was sensitive to the underlying biological similarities and differences between the studies, we analysed studies from two JIA subtypes and a non-JIA liver study comparing tissue samples from old and young individuals (GSE133815). Out of 7 sJIA studies, 5 included samples from PBMC and two from whole blood. We wanted to determine whether our differential pathway expression approach could identify the biological differences and similarities between JIA subtypes and tissues, and overcome the

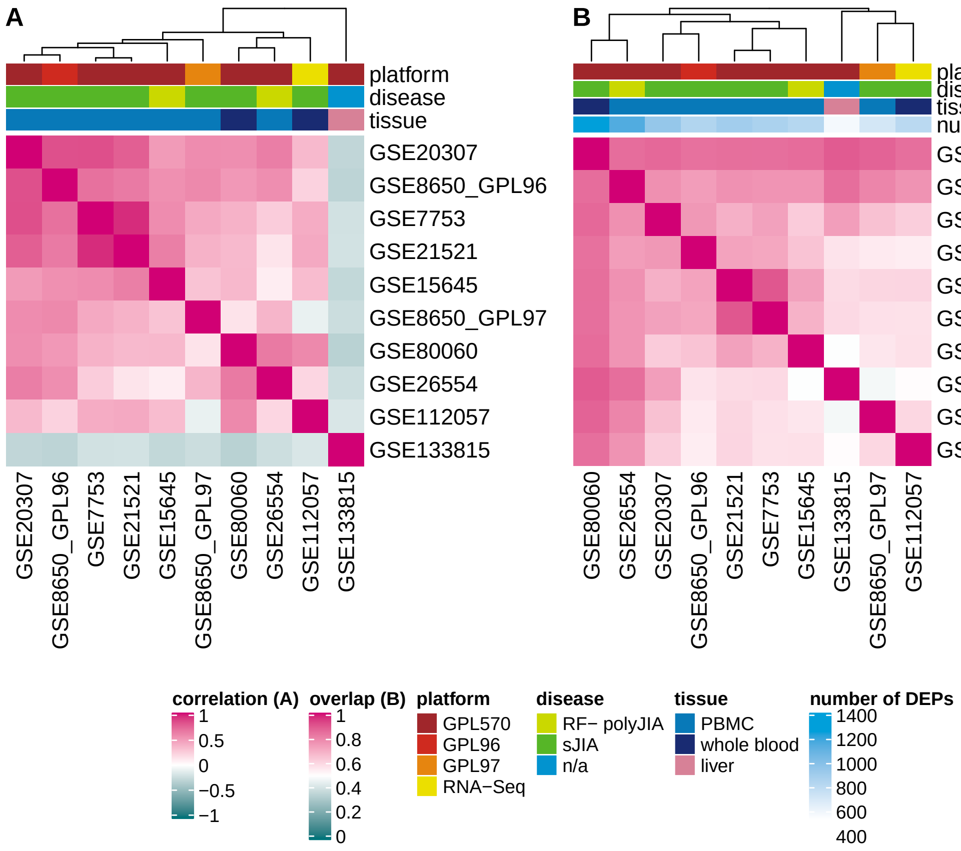


Fig. (6.1) Correlation and overlap between JIA disease pathway signatures. (A) Spearman correlation between logFCs of JIA differentially expressed pathways (DEPs) and DEPs from one non-related study (GSE133815) with no p-value threshold. (B) Overlap between disease pathway signatures (q-value < 0.05). Supplementary Fig. D.1 shows the differential pathway expression profile (q-value < 0.05) for each of the 10 studies. DEP — differentially expressed pathway; JIA — juvenile idiopathic arthritis; logFC — $\log_2$ fold change; PBMC — peripheral blood mononuclear cells; RF- polyJIA — rheumatoid factor negative polyarticular JIA; sJIA — systemic JIA.

platform effect. We hypothesised that sJIA PBMC studies will produce similar pathway profiles, distinct from other JIA studies and sJIA studies profiled in the whole blood. Additionally, we hypothesised that the liver study will produce the most contrasting pathway signature.

In order to assess the similarities of differentially expressed pathway lists, we calculated Spearman's correlation and the overlap between differential pathway expression profiles for each pair of the analysed studies (Fig. 6.1). We calculated the correlation of $\log_2$ fold change (logFC) values for all pathways with no p-value threshold (Fig. 6.1A). All JIA studies formed a positively correlated cluster, while the control liver study was negatively correlated with all JIA studies and consequently clustered separately from them. 5 out of 7 sJIA studies clustered together with one polyJIA study. GSE20307, GSE8650_GPL96, GSE7753 and GSE21521 formed a tight cluster with high positive correlation scores between the study pairs. The other two sJIA studies (GSE80060, GSE112057) clustered with the second polyJIA study away from the other JIA studies. The two sJIA studies, clustering apart from the majority of studies, were from the whole blood rather than the PBMC.

We assessed the overlap size between differently expressed pathways (DEPs) (q-value < 0.05) for each pair of studies, ignoring the DEP direction (Fig. 6.1B). The overlap values were relatively high (> 0.5) due to the high number of DEPs meeting the significance threshold. A possible reason for the high number of significant DEPs is that the pathway summary scores represent the mean expression of only the most different genes between two conditions thus leading to enhanced pathway-level differences between two conditions. The pair of studies with the highest correlation (GSE7753 and GSE21521) also showed a high overlap value. Due to a large number of DEPs (1250/1473) GSE80060 showed high overlap with all other studies, including the control liver study. The liver study showed low overlap with studies with fewer DEPs. As the liver study showed negative correlation with all JIA studies (no q-value threshold), it suggested that the DEPs were expressed in opposite directions despite the high overlap. The differential pathway expression profiles for each of the 10 studies (q-value < 0.05, Supplementary Fig. D.1) showed that the clear outlier by direction of DEPs was the liver study we included as the control. Its profile contradicted all other studies by having mostly up-regulated pathways when the rest have down-regulated and vice versa.

The whole blood sJIA RNA-Seq study (GSE112057) had higher correlation with other JIA studies compared to the negatively correlated control study. It showed high correlation with the other whole blood sJIA microarray study (GSE80060) suggesting that the method could partially overcome the platform effect. No RNA-Seq data from PBMC sJIA samples were available at the time of study selection; therefore, we could not make definitive conclusions about the signature method's inherent ability to overcome microarray and RNA-Seq platform differences.

Clustering of the differential pathway expression analysis results suggested that the pathway-level signature generation method was sensitive to the underlying biology of the samples, while partially overcoming the platform effect. It was able to distinguish between JIA and non-JIA studies, in addition, it showed the potential to differentiate different JIA subtypes or studies from different tissues. The sJIA GPL96 study (GSE8650) consistently clustered with the other sJIA studies analysed on the more common GPL570 platform. Due to discrete probe sets GSE8650 analysed on GPL97 did not show similar patterns of logFC values to the GPL96, despite analysing the same samples (Supplementary Fig. D.1). Both GPL96 and GPL97 probe sets are subsets of GPL570 (Affymetrix, 2003a,b). The differential pathway expression analysis suggested that GPL96 included a more relevant probe set for JIA compared to GPL97. Returning to the hypothesis we showed that the sJIA PBMC studies produced similar pathway profiles, distinct from other JIA studies and sJIA studies profiled in the whole blood. Additionally, we confirmed that the control liver study produced the most contrasting pathway signature.

Interpretation of differentially expressed pathways

While we generated the differential pathway expression profiles for all studies listed in Table 6.1, we conducted a literature review for biological relevance of the 10 most up- and down-regulated pathways from only the two representative studies (GSE7753 and GSE112057). We investigated whether each pathway has been linked to JIA or rheumatoid arthritis pathology. The top 20 most up- and down-regulated pathways for GSE7753 and GSE112057 are listed in Tables 6.2 and 6.3, respectively.

Table (6.2) GSE7753 sJIA disease signature pathways. The top 20 most up- (rank: 1 to 20) and down- (rank: -1 to -20) regulated pathways (q-value < 0.05). Drugs were prioritised for the up- and down-regulated pathway clusters at: the top 5, 10, 15 or 20 pathways by decreasing log fold change (LogFC) for up- and decreasing for down-regulated pathways. The genes in pathway column represents the number of possible genes in that pathway, while genes in data is the number of pathway genes found in data. sJIA — systemic juvenile idiopathic arthritis.

Rank	Pathway	LogFC	q-value	Genes in pathway	Genes in data
1	Biocarta AHSP pathway	1.300	1.22e-06	13	12
2	Reactome G1 S SPECIFIC TRANSCRIPTION	1.140	6.98e-06	19	17
3	SA G2 and M PHASES	1.110	8.55e-06	8	8
4	Reactome POST CHAPERONIN TUBULIN FOLDING pathway	1.060	4.68e-06	19	16
5	Reactome METABOLISM of PORPHYRINS	1.040	4.68e-06	14	13
6	Reactome RNA POL I PROMOTER OPENING	0.982	6.98e-06	62	43
7	Reactome ACTIVATION of the AP1 FAMILY of TRANSCRIPTION FACTORS	0.971	4.00e-06	10	10
8	Biocarta GLYCOLYSIS pathway	0.942	4.62e-04	10	10
9	Reactome CREATION of C4 and C2 ACTIVATORS	0.941	1.09e-04	10	8
10	Reactome UNWINDING of DNA	0.937	7.05e-05	11	11
11	SPTAN1 10 Static Module	0.892	1.81e-05	10	10
12	Reactome AMYLOIDS	0.879	8.23e-06	83	63
13	Biocarta EPONFKB pathway	0.858	6.90e-06	11	11
14	Biocarta SKP2E2F pathway	0.845	6.98e-06	10	10
15	Reactome DEGRADATION of the EXTRACELLULAR MATRIX	0.844	3.90e-06	29	28
16	Biocarta DREAM pathway	0.843	6.64e-06	14	14
17	Reactome PACKAGING of TELOMERE ENDS	0.831	1.16e-04	48	36
18	SPI1 10 Static Module	0.823	2.08e-05	10	9
19	Reactome SYNTHESIS of BILE ACIDS and BILE SALTS via 24 HYDROXYCHOLESTEROL	0.810	3.90e-06	10	10
20	Biocarta GRANULOCYTES pathway	0.809	1.62e-05	14	14
-1	SFPQ 10 Static Module	-0.919	8.06e-05	10	8
-2	BCLAF1 25 Static Module	-0.898	1.52e-04	25	25
-3	ASH2L 391 Static Module	-0.799	4.06e-05	390	335
-4	TIA1 10 Static Module	-0.763	1.90e-03	10	10
-5	Biocarta TCRA pathway	-0.758	1.26e-03	13	11
-6	Reactome GLUCURONIDATION	-0.702	2.70e-05	18	9
-7	KEGG CIRCADIAN RHYTHM MAMMAL	-0.691	6.98e-06	13	13
-8	Reactome PEPTIDE CHAIN ELONGATION	-0.684	1.63e-03	153	83
-9	Reactome 3 UTR MEDIATED TRANSLATIONAL REGULATION	-0.665	1.04e-03	176	103
-10	KEGG RIBOSOME	-0.660	2.73e-03	88	85
-11	Reactome PROCESSING of INTRONLESS PRE MRNAS	-0.659	6.59e-04	14	14
-12	Reactome FORMATION of the TERNARY COMPLEX and SUBSEQUENTLY the 43S COMPLEX	-0.659	1.97e-03	74	47
-13	Reactome TRANSLOCATION of ZAP 70 to IMMUNOLOGICAL SYNAPSE	-0.657	4.99e-03	14	12

continues on the next page

Table 6.2 continued

Rank	Pathway	LogFC	q-value	Genes in pathway	Genes in data
-14	Reactome ACTIVATION of the MRNA UPON BINDING of the CAP BINDING COMPLEX and EIFS and SUBSEQUENT BINDING to 43S	-0.615	1.62e-03	84	55
-15	PID CIRCADIAN pathway	-0.613	4.06e-05	16	16
-16	Reactome INFLUENZA VIRAL RNA TRANSCRIPTION and REPLICATION	-0.607	2.69e-03	169	99
-17	Reactome NONSENSE MEDIATED DECAY ENHANCED by the EXON JUNCTION COMPLEX	-0.600	1.14e-03	176	104
-18	Reactome GENERIC TRANSCRIPTION pathway	-0.596	6.24e-05	352	329
-19	Reactome SRP DEPENDENT COTRANSLATIONAL PROTEIN TARGETING to MEMBRANE	-0.595	3.78e-03	179	106
-20	HLA-A 53 Static Module	-0.594	1.89e-04	51	42

The top 10 up-regulated pathways in GSE7753 were involved in the cell cycle (*Reactome G1/S specific transcription*, *SA G2 and M phases*, *Reactome Unwinding of DNA* (Howard, 1953)), erythropoiesis (*Biocarta AHSP pathway*, *Reactome Metabolism of porphyrins*), transcription (*Reactome RNA Pol I promoter opening*, *Reactome Activation of the Activator protein 1 (AP1) family of transcription factors* (Angel and Karin, 1991)), complement activation (*Reactome Creation of C4 and C2 activators*), glycolysis (*Biocarta Glycolysis pathway*), and tubulin folding (*Reactome Post-chaperonin tubulin folding pathway*) (Table 6.2).

Erythropoiesis. The top up-regulated pathway in GSE7753 was the *Biocarta alpha haemoglobin stabilising protein (AHSP) pathway* involved in erythropoiesis, production of mature red blood cells (Weiss et al., 2005). It is considered to be an indicator of ineffective erythropoiesis as higher proportion of immature PBMC sub-populations were identified with flow cytometry in samples from this study (GSE7753, Fall et al. (2007)) and supported by later studies (GSE21521 (Hinze et al., 2010), and GSE13501 (duplicate sJIA samples from GSE20307), (Barnes et al., 2009)). The 5th most up-regulated pathway (*Reactome Metabolism of porphyrins*) was also related to the production of red blood cells. Porphyrins form a porphyrin complex that is in haem, the pigment in red blood cells (Perutz et al., 1960). **Complement activation.** The complement system is part of the innate immune system, it is also a key mediator of inflammatory injury (Walport, 2001). It was shown that the levels of bound C3 and C4 to the circulating immune complexes were significantly higher in RF- polyJIA (Gilliam et al., 2011), as well as that the levels of plasma complement activation fragments correlate with active JIA (Aggarwal et al., 2000; Jarvis et al., 1993). **Glycolysis.** An increased glycolytic activity has been observed in rheumatoid arthritis (RA) synovial fluid (Ciurtin et al., 2006; Henderson et al., 1979; Naughton et al., 1993). Additionally, it has been shown that the hypoxia-inducible factor 1 α (HIF-1α), whose expression is also increased in synovial fluid (Hollander et al., 2001), can increase the expression of glycolytic enzymes in the inflammatory synovium (Distler et al., 2004).

The top 10 down-regulated pathways in GSE7753 were related to TNFα (*SFPQ Static module*, *TIA1 Static module*), HLA (*Biocarta TCRA pathway*), drug metabolism (*Reactome Glucuronidation*), transcription (*ASH2L Static module*), translation (*Reactome Peptide chain elongation*, *Reactome 3'UTR mediated translation*, *KEGG Ribosome*, *BCLAF1 Static module*), and circadian rhythm (*KEGG Circadian rhythm mammal*) (Table 6.2).

TNFα. The most down-regulated pathway in GSE7753 was the *SFPQ Static module* with splicing factor proline- and glutamine-rich (SFPQ) as its hub gene. SFPQ modulates phosphodiesterase 3A (PDE3A), whose locus has been established as a strong genetic marker of anti-TNF therapy response (Acosta-Colman et al., 2013; Rhee et al., 2017). TNFα is one of the most important cytokines involved in JIA pathogenesis and is thought to account for the articular manifestation of JIA together with other pro-inflammatory cytokines like IL-1α and IL-6 (Kutukculer et al., 1998; Lepore et al., 1994). The 4th most down-regulated pathway (*TIA1 Static module*) is also associated with TNFα. The arthritis suppressor gene, T-cell intracellular antigen-1 (TIA-1), lowers the expression of TNFα (Phillips et al., 2004). BCL-associated factor 1 (BCLAF1) is the hub gene in another down-regulated Static module. BCLAF1 is induced by nuclear factor κ-light-chain-enhancer of activated B cells (NF-κB) transcriptional activity (Shao et al., 2016) and NF-κB is induced by TNFα (Hayden and Ghosh, 2014). **HLA.** The *Biocarta TCRA pathway* involving T-cell receptor (TCR) activation was linked to JIA's best established genetic risk factor, the HLA complex (Ombrello et al., 2015). HLA encodes the major histocompatibility complex (MHC) that stimulates TCRs (Dausset, 1958; Zinkernagel and Doherty, 1979). **Drug metabolism.** The 6th most down-regulated pathway was the *Reactome Glucuronidation*. Glucuronidation is commonly involved in drug metabolism including the metabolism of NSAIDs which are commonly administered to JIA patients (Kuehl et al., 2005; Williams et al., 2004). NSAIDs have also shown inhibitory potential of glucuronidation (Mano et al., 2007). While GSE7753 sJIA samples were considered untreated, the authors only specified that samples were taken prior to the initiation of treatment with DMARDs or steroids (Fall et al., 2007). Therefore, the glucuronidation pathway could be dysregulated due to patients receiving treatment other than DMARDs or steroids.

The top up-regulated pathways in GSE112057 were related to the cytoskeleton (*Biocarta Salmonella pathway*, *Biocarta RAB pathway*, *Biocarta CDC42RAC pathway*, *Biocarta ACTINY pathway*, *SEPT2 Static module*), hypoxia (*Reactome Oxygen dependent proline hydroxylation of hypoxia inducible factor alpha*, *Reactome Regulation of hypoxia inducible factor HIF by oxygen*), NF-κB (*Reactome RAP1 signalling*, *Biocarta EPO/NFKB pathway*), and platelet activation (*PID Thrombin PAR4 pathway*) (Table 6.3).

Table (6.3) GSE112057 sJIA disease signature pathways. The top 20 most up- (rank: 1 to 20) and down- (rank: -1 to -20) regulated pathways (q-value < 0.05). Drugs were prioritised for up- and down-regulated pathway clusters at: the top 5, 10, 15 or 20 pathways by decreasing log fold change (LogFC) for up- and decreasing for down-regulated pathways. The genes in pathway column represents the number of possible genes in that pathway, while genes in data is the number of pathway genes found in data. sJIA — systemic juvenile idiopathic arthritis.

Rank	**Pathway**	**LogFC**	***q*-value**	**Genes in pathway**	**Genes in data**
1	Biocarta SALMONELLA pathway	1.100	0.005650	13	11
2	Biocarta RAB pathway	1.090	0.005800	12	9
3	Biocarta CDC42RAC pathway	1.090	0.005690	16	14
4	Reactome OXYGEN DEPENDENT PROLINE HYDROXYLATION of HYPOXIA INDUCIBLE FACTOR ALPHA	1.070	0.001440	18	11
5	PID THROMBIN PAR4 pathway	1.050	0.006610	15	10
6	Biocarta ACTINY pathway	1.020	0.006930	20	15
7	SEPT2 21 Static Module	1.000	0.004050	20	5
8	Reactome REGULATION of HYPOXIA INDUCIBLE FACTOR HIF by OXYGEN	0.994	0.003480	25	16
9	Reactome RAP1 SIGNALLING	0.988	0.004080	17	12
10	Biocarta EPONFKB pathway	0.986	0.006990	11	9
11	Reactome DSCAM INTERACTIONS	0.976	0.008430	11	6
12	KEGG DORSO VENTRAL AXIS FORMATION	0.976	0.007780	25	14
13	PID IL5 pathway	0.976	0.008080	14	13
14	Biocarta HCMV pathway	0.965	0.005230	17	17
15	PRKCA 14 Static Module	0.965	0.008430	14	5
16	Reactome SIGNALING by NOTCH2	0.965	0.007630	12	10
17	Biocarta MONOCYTE pathway	0.957	0.006250	11	10
18	Biocarta RANKL pathway	0.956	0.007200	14	10
19	Reactome RECEPTOR LIGAND BINDING INITIATES the SECOND PROTEOLYTIC CLEAVAGE of NOTCH RECEPTOR	0.945	0.009520	12	6
20	Reactome IL 6 SIGNALING	0.940	0.007270	11	9
-1	Reactome PROSTANOID LIGAND RECEPTORS	-1.040	0.001440	10	6
-2	Reactome COLLAGEN FORMATION	-0.976	0.003400	58	16
-3	Biocarta NUCLEARRS pathway	-0.962	0.001610	15	7
-4	Reactome VEGF LIGAND RECEPTOR INTERACTIONS	-0.961	0.003800	10	5
-5	Reactome EICOSANOID LIGAND BINDING RECEPTORS	-0.955	0.001740	16	8
-6	Reactome TRANSLOCATION of ZAP 70 to IMMUNOLOGICAL SYNAPSE	-0.954	0.008080	14	7
-7	PID CONE pathway	-0.946	0.002520	23	5
-8	Reactome APOBEC3G MEDIATED RESISTANCE to HIV1 INFECTION	-0.936	0.005690	12	5

continues on the next page

Table 6.3 continued

Rank	Pathway	LogFC	q-value	Genes in pathway	Genes in data
-9	Reactome RNA POL III TRANSCRIPTION INITIATION FROM TYPE 2 PROMOTER	-0.918	0.006830	23	22
-10	Reactome RNA POL III CHAIN ELONGATION	-0.887	0.007570	17	16
-11	Reactome RNA POL III TRANSCRIPTION TERMINATION	-0.874	0.006930	19	17
-12	KEGG VALINE LEUCINE and ISOLEUCINE BIOSYNTHESIS	-0.847	0.011700	11	9
-13	PID CIRCADIAN pathway	-0.830	0.012200	16	11
-14	KEGG RNA POLYMERASE	-0.830	0.007270	29	27
-15	Reactome NCAM1 INTERACTIONS	-0.803	0.000994	39	12
-16	Biocarta NO2IL12 pathway	-0.799	0.001910	17	13
-17	Reactome RECRUITMENT of NUMA to MITOTIC CENTROSOMES	-0.799	0.006990	10	10
-18	Biocarta SET pathway	-0.797	0.001440	11	10
-19	Reactome ABCA TRANSPORTERS in LIPID HOMEOSTASIS	-0.791	0.001960	18	8
-20	Reactome BIOSYNTHESIS of the N GLYCAN PRECURSOR DOLICHOL LIPID LINKED OLIGOSACCHARIDE LLO and TRANSFER to A NASCENT PROTEIN	-0.784	0.005690	29	25

Actin cytoskeleton. The Salmonella pathway includes two Rho GTPases, cell division control protein 42 homolog (CDC42) and Ras-related C3 botulinum toxin substrate 1 (RAC1) (also key proteins in *Biocarta CDC42/RAC pathway*), that regulate the actin cytoskeleton (Chen et al., 1996). Rab GTPase coordinates with Rho GTPases in regulation of cytoskeleton organisation (reviewed in Kjos et al. (2018)), the *ACTINY pathway* includes proteins involved in actin polymerization including Rac1. Septin 2 (SEPT2) is involved in stabilisation of actin fibres (Kinoshita et al., 2002; Schmidt and Nichols, 2004). These pathways suggested actin cytoskeleton dysregulation in JIA. Actin cytoskeleton and its components have previously been identified as dysregulated in JIA plasma (Gibson et al., 2012), polyJIA synovium (Finnegan et al., 2014) and RA synovial fibroblasts (Aidinis et al., 2005; Matsuo et al., 2006; Vasilopoulos et al., 2007). **Hypoxia and NF-κB.** The NF-κB and hypoxia inducible factor (HIF) have been shown to act interdependently in hypoxia and inflammation (Belaiba et al., 2007; Rius et al., 2008; Walmsley et al., 2005). Additionally, synovial hypoxia has been a constant feature of RA (Quiñonez-Flores et al., 2016). HIF-2α, overexpressed in synovial fibroblasts, regulates the expression of receptor activator of NF-κB ligand (Ryu et al., 2014) and is involved in cartilage erosion (Huh et al., 2015). HIF-1α can be activated via NF-κB pathway in the presence of bacterial lipopolysaccharides (Frede et al., 2006). In addition to HIF related pathways, RAP1 signalling and EPO/NFKB pathway were also linked to NF-κB. Repressor activator protein 1 (RAP1) regulates NF-κB-dependent gene expression (Teo et al., 2010). Erythropoietin (Epo) secreted by the kidney stimulates red blood cell production and is also secreted in the brain in response to hypoxia, induced by HIF-1 (Digicaylioglu and Lipton, 2001). **Platelet activation.** Thrombin, the most potent platelet activator, mediates the process through two proteinase-activated receptors (PARs) 1 and 4 (Wu et al., 2010a). Active JIA was associated with increased mean platelet volume, which is an indicator of systemic inflammation (Güneş et al., 2015).

The top down-regulated pathways in GSE112057 were related to drug response (*Reactome Prostanoid ligand receptors*, *Biocarta NUCLEARRS pathway*, *Eicosanoid ligand binding receptors*), NF-κB (*Reactome VEGF ligand receptor interactions*), immune response (*Reactome Translocation of ZAP70 to immunological synapse*, *Reactome APOBEC3G mediated resistance to HIV1 infection*, *Reactome RNA pol III transcription initiation from type 2 promoter*, *Reactome RNA pol III chain elongation*), and collagen formation (*Reactome Collagen formation*) (Table 6.3).

Drug response. Arachidonic acid can be metabolised into different classes of eicosanoids by cyclooxygenases (COX), lipoxygenases, or cytochrome P450 (CYP) enzymes and Prostanoid receptors bind prostanoids, a subclass of eicosanoids, which are COX metabolites. Most NSAIDs are non-selective inhibitors of COX, thus can limit the rate of arachidonic acid metabolism and consequently, prostanoid formation (Allaj et al., 2013). COX-2 expression is induced by various inflammatory stimuli and can be suppressed by glucocorticoids like dexamethasone (Smith et al., 2000). CYPs present in the *Biocarta NUCLEARRS pathway* are responsible for clearing approximately 80% of the top 200 prescribed drugs in the US (Zanger et al., 2008). Only partial treatment information was accessible in Mo et al. (2018) for GSE112057. In their analysis they adjusted for treatment based on 3 non-exclusive categories of medication: known treatment with DMARDs, biologics, and steroids, however, no treatment information was provided in the metadata in the GEO dataset. Thus, it is likely that the patients have received various treatments that we were unable to correct for due to the lack of treatment metadata. The disease pathways were likely a result of drug perturbations, as suggested by dysregulation of drug metabolism related pathways. **NF-κB.** The vascular endothelial growth factor (VEGF) expression is regulated by TNFα and NF-κB (Yoshida et al., 1997), its concentrations have also been positively correlated with the disease activity in JIA (Maeno et al., 1999; Świdrowska et al., 2015). **Immune response.** Zeta-associated protein of 70kD (ZAP70) normally expressed near the surface membrane of T and natural killer (NK) cells, is essential in the T cell activation. Additionally, it has also been found in B cells obtained from the synovial fluid and tissue of RA patients. The percentage of ZAP-70^{+} B cells correlates with levels of IL-6 (Tolusso et al., 2009). The Apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like 3G (APOBEC3G) family provide innate resistance to retroviruses by mutating the viral DNA into non-functional provirus and inducing the innate immune response (Harris and Dudley, 2015). RNA polymerase III has been shown to transcribe pathogenic DNA and trigger the innate immune response (Chiu et al., 2009). **Collagen formation.** Autoantibodies to type II collagen (CII), the predominant structural protein of articular cartilage, have been detected in the serum of sJIA patients (Myers et al., 2001).

In summary, the differential pathway analysis of GSE7753 and GSE112057 has identified several JIA-related dysregulated pathways. In both studies there were several inflammatory and immune-related processes, some associated with well-established JIA proteins like TNFα and NF-κB. In both studies drug metabolism-related pathways were down-regulated, suggesting patients might have received some treatment before sample extraction, leading to dysregulation of those pathways. These results were further supported by Ramanathan et al. (2018). They have identified similar trends in their pathway analysis of public JIA datasets (including GSE21521 and two other studies). Their analysis identified pathways elevated in JIA included cytokine signalling pathways, kinase pathways (NF-κB), pathways relating to cell migration (actin cytoskeleton), pathways relating to production of reactive oxygen species (ROS, hypoxia-related) and regulation of the cell cycle checkpoints (Ramanathan et al., 2018). These groups were also found in GSE7753 and GSE112057 discussed in this section. The pathological relevance of the top up- and down-regulated pathways further demonstrated the sensitivity of the differential pathway expression method.

6.3.2 Overlap pathway signature

In order to assess the consensus sJIA PBMC signature, we identified the overlapping DEPs from all 5 sJIA PBMC studies (GSE20307, GSE21521, GSE7753, GSE8650_GPL96, GSE8650_GPL97). From the 5 studies there were 208 overlapping DEPs (q-value < 0.05 in each study). The expected range of intersect size from a random permutation test was 54–113 pathways (expected mean = 82.8), with the observed overlap size of 208 (p-value = 1e-04, Supplementary Fig. D.2). We ranked the overlap pathways by mean logFC. The 20 most up- and down-regulated pathways are listed in Table 6.4.

Table (6.4) Overlap PBMC sJIA disease signature pathways. The top 20 most up- (rank: 1 to 20) and down- (rank: -1 to -20) regulated overlap pathways from GSE20307, GSE21521, GSE7753, GSE8650_GPL96, GSE8650_GPL96 (q-value < 0.05 in each study). Drugs were prioritized for up- and down-regulated pathway clusters at: the top 5, 10, 15 or 20 pathways by decreasing mean log fold change (mean logFC). The genes in pathway column represents the number of possible genes in that pathway. Genes in data is the number of pathway genes found in data. PBMC — peripheral blood mononuclear cells; sJIA — systemic juvenile idiopathic arthritis.

Rank	Pathway	mean LogFC
1	Reactome DEGRADATION of the EXTRACELLULAR MATRIX	0.889
2	Reactome AMYLOIDS	0.775
3	Reactome UNWINDING of DNA	0.746
4	KEGG COMPLEMENT and COAGULATION CASCADES	0.716
5	Reactome INHIBITION of VOLTAGE GATED CA2 CHANNELS via GBETA GAMMA SUBUNITS	0.694
6	Reactome FORMATION of FIBRIN CLOT CLOTTING CASCADE	0.692
7	Reactome ORGANIC CATION ANION ZWITTERION TRANSPORT	0.686
8	Reactome AMINE COMPOUND SLC TRANSPORTERS	0.679
9	Reactome PLATELET CALCIUM HOMEOSTASIS	0.676
10	PID INTEGRIN A9B1 pathway	0.667
11	NFIC 21 Static Module	0.662
12	Reactome SYNTHESIS of PC	0.654
13	Reactome INWARDLY RECTIFYING K CHANNELS	0.641
14	F2 46 Static Module	0.633
15	Biocarta UCALPAIN pathway	0.623
16	SA G1 and S PHASES	0.608
17	KEGG SYSTEMIC LUPUS ERYTHEMATOSUS	0.594
18	PID P38 GAMMA DELTA pathway	0.590
19	Reactome NITRIC OXIDE STIMULATES GUANYLATE CYCLASE	0.582
20	Reactome GABA B RECEPTOR ACTIVATION	0.575
-1	Reactome PEPTIDE CHAIN ELONGATION	-0.880
-2	Reactome FORMATION of the TERNARY COMPLEX and SUBSEQUENTLY the 43S COMPLEX	-0.870
-3	Reactome 3 UTR MEDIATED TRANSLATIONAL REGULATION	-0.867
-4	BCLAF1 25 Static Module	-0.867
-5	KEGG RIBOSOME	-0.864
-6	Reactome ACTIVATION of the MRNA UPON BINDING of the CAP BINDING COMPLEX and EIFS and SUBSEQUENT BINDING to 43S	-0.837
-7	Reactome NONSENSE MEDIATED DECAY ENHANCED by the EXON JUNCTION COMPLEX	-0.833
-8	Reactome INFLUENZA VIRAL RNA TRANSCRIPTION and REPLICATION	-0.830
-9	Reactome SRP DEPENDENT COTRANSLATIONAL PROTEIN TARGETING to MEMBRANE	-0.826
-10	Reactome TRANSLATION	-0.810
-11	Reactome INFLUENZA LIFE CYCLE	-0.803
-12	RPS27A 138 Static Module	-0.797
-13	ASH2L 391 Static Module	-0.789
-14	Reactome MRNA 3 END PROCESSING	-0.740
-15	Reactome CLEAVAGE of GROWING TRANSCRIPT in the TERMINATION REGION	-0.702
-16	KEGG SPLICEOSOME	-0.678
-17	EPRS 15 Static Module	-0.671
-18	Reactome PROCESSING of CAPPED INTRON CONTAINING PRE MRNA	-0.670
-19	Reactome NEP NS2 INTERACTS with the CELLULAR EXPORT MACHINERY	-0.669
-20	POLR2A 195 Static Module	-0.667

Briefly, the top up-regulated overlap pathways were related to the extracellular matrix (*Reactome Degradation of the extracellular matrix*, *Reactome Amyloids*, *PID Integrin A9B1 pathway*), coagulation (*KEGG Complement and coagulation cascades*, *Reactome Formation of fibrin clot clotting cascade*, *Reactome Platelet calcium homeostasis*), transport (*Reactome Organic cation/anion/zwitterion transport*, *Reactome Amine compound SLC transporters*), DNA unwinding (*Reactome Unwinding of DNA*), and gamma-aminobutyric acid B (GABAB) receptors (*Reactome Inhibition of voltage gated Ca*$^{2+}$ *channels via Gbeta/gamma subunits*).

The most down-regulated overlap pathways were related to transcription (*Reactome Nonsense mediated decay enhanced by the exon junction complex*, *Reactome Influenza viral RNA transcription and replication*), translation (*Reactome Peptide chain elongation*, *Reactome Formation of the ternary complex and subsequently the 43S complex*, *Reactome 3'UTR mediated translational regulation*, *KEGG Ribosome*, *Reactome Activation of the mRNA upon binding of the CAP binding complex and EIFS and subsequent binding to 43S*, *Reactome SRP-dependent cotranslational protein targeting to membrane*, *Reactome Translation*), and NF-κB induced BCLAF1 (*BCLAF1 Static module*).

Although there was a large overlap between the sJIA PBMC overlap pathways, the top pathways were related to generic cell processes rather than processes related to JIA pathology. We hypothesised that this was due to including GSE8050_GPL97 in the overlap, because the GSE8050_GPL97 pathway profile was the least correlated with the lowest overlap to the other 4 PBMC sJIA studies (see Supplementary Fig. D.1 and Fig. 6.1). We tested whether GSE8050_GPL97 reduced the overlap size by generating the sJIA PBMC overlap by leaving out either GSE8050_GPL96 or GSE8050_GPL97. The overlap size leaving out GSE8050_GPL97 was 455 with expected range 144–217 (mean = 178), compared to overlap size of 255 with expected range of 114–183 (mean = 148) when leaving out GSE8050_GPL96. Thus, confirming that GSE8050_GPL97 inclusion restricted the size and likely also the disease-relevant pathways. We continued our analysis with the overlap signature including GSE8050_GPL97 to highlight the limitations of the consensus signature as well as avoid overfitting to a small subset of sJIA studies.

6.3.3 Gene-level signatures

We generated gene signatures from 2 representative studies: GSE7753 (microarray) and GSE112057 (RNA-Seq). We identified 2239 (1145 up- and 1094 down-regulated, q-value < 0.05) differentially expressed genes (DEGs) for GSE7753 and 2496 (1156 up-

and 1340 down-regulated, q-value < 0.05). The top 20 up- and down-regulated genes are listed in Supplementary Tables D.1 and D.2 for GSE7753 and GSE112057, respectively.

Overall, there were 4 solute carrier (SLC) family members dysregulated in the top 20 genes (per direction) in GSE7753 and 3 in GSE112057. This family of transporters is key in two transport associated pathways identified in the overlap pathway signature. 8 out of 20 GSE7753 up-regulated genes had "blood" or "hemoglobin" in the description, likely influencing the red blood cell-associated pathways up-regulated in GSE7753 pathway analysis. In each study there was one C4-related gene, a C4 binding protein in GSE112057 and a C4 receptor in GSE7753. GSE112057 had two down-regulated keratin genes that were likely driving the down-regulation of *Reactome Collagen formation* in GSE112057 pathway analysis (Supplementary Tables D.1 and D.2).

6.3.4 Disease ontology (DO) enrichment of gene and pathway signatures

To investigate the similarities and differences between gene- and pathway-level JIA signatures we performed enrichment analysis of disease ontology (DO) terms. We investigated enrichment of the top 1000 differentially expressed genes by decreasing |logFC| and the top DEPs (decreasing |logFC|) whose member genes add up to 1000. In addition to gene- and pathway-level enrichment of GSE7753 and GSE112057, we also included the sJIA PBMC pathway overlap signature. For clarity, we referred to enrichment tests with test letters A–D:

(A) gene-level GSE112057,

(B) pathway-level GSE112057,

(C) gene-level GSE7753,

(D) pathway-level GSE7753,

(E) sJIA PBMC overlap pathways.

The pathway-level enrichment analysis yielded 7.1- (GSE112057) and 4.5-times (GSE7753) more enriched DO terms (q-value < 0.05) than the gene-level analysis (Table 6.5). This was likely due to pathways consisting of curated sets of genes that are involved in the same process, and at the same time, DO terms encompassing biologically connected genes. The overlap between single-study pathway-level enriched terms was 88.3% (B∩D) compared to 40.2% (A∩C) between gene-level studies, suggesting that the pathway-level analysis has identified highly related processes overcoming the platform effect, while the gene-level analysis provided more disparate sets of differentially expressed genes. The overlap in enriched DO terms from the same study at pathway- and gene-level was similar between both studies (GSE112057 (A∩B): 97.5% and GSE7753 (C∩D): 93.5%), suggesting the biology represented by the top DEGs was also captured by the pathway-level analysis.

Table (6.5) Number of enriched Disease ontology terms from sJIA pathway- and gene-level differential expression analysis. Letters in () refer to enrichment tests in 6.3.4: (A) gene-level GSE112057, (B) pathway-level GSE112057, (C) gene-level GSE7753, (D) pathway-level GSE7753. Overlap % is calculated as the number of overlapping terms divided by the smaller of the two sets of enriched terms (Chapter 3 Eq. (3.11)). sJIA — systemic juvenile idiopathic arthritis.

		study		
		GSE112057	GSE7753	overlap
level	gene	(A) 82	(C) 124	33 (40.2%)
	pathway	(B) 582	(D) 556	491 (88.3%)
	overlap	80 (97.5%)	116 (93.5%)	33

We next investigated the top 10 terms from each enrichment test and the overlap between those (Fig. 6.2). All cases enriched in several blood- and immune system-related diseases, several of which have been previously related or co-studied with JIA or RA (*hepatitis* (Canna et al., 2009), *coagulation abnormalities* (Bloom et al., 1998; Hadchouel et al., 1985), *bacterial infections* (Beukelman et al., 2012), *malignancy/lymphoma/leukaemia* (Demir et al., 2019; Kok et al., 2014; Murray et al., 2004; Wilton and Matteson, 2017), *atherosclerosis* (reviewed in Jednacz and Rutkowska-Sak (2012)), *anaemia* (Koerper et al., 1978; Raj, 2009)).

Five out of 36 listed top DO terms related to female reproductive organ cancer (*female reproductive organ cancer*, *malignant ovarian surface epithelial-stromal neoplasm*, *ovary epithelial cancer*, *cervix carcinoma*, *cervical cancer*). While these could have been driven by female to male imbalance in patients versus controls, persistent human papillomavirus (HPV) infections and increased risk of cervical dysplasia and cervical cancer has been reported in adults with RA (Kim et al., 2015; Rojo-Contreras et al., 2012; Waisberg et al., 2015). In addition, JIA patients with abnormal cervical cytology had a higher frequency of HPV infection with a lower frequency of HPV vaccination compared to non-JIA controls (Ferreira et al., 2019).

From the 36 listed DO terms in Fig. 6.2, 15 were overlapping in test A–D, including *rheumatic disease*. Due to pathways consisting of biologically related genes, the enrichment values from pathway tests had higher gene ratios and lower p-values. The

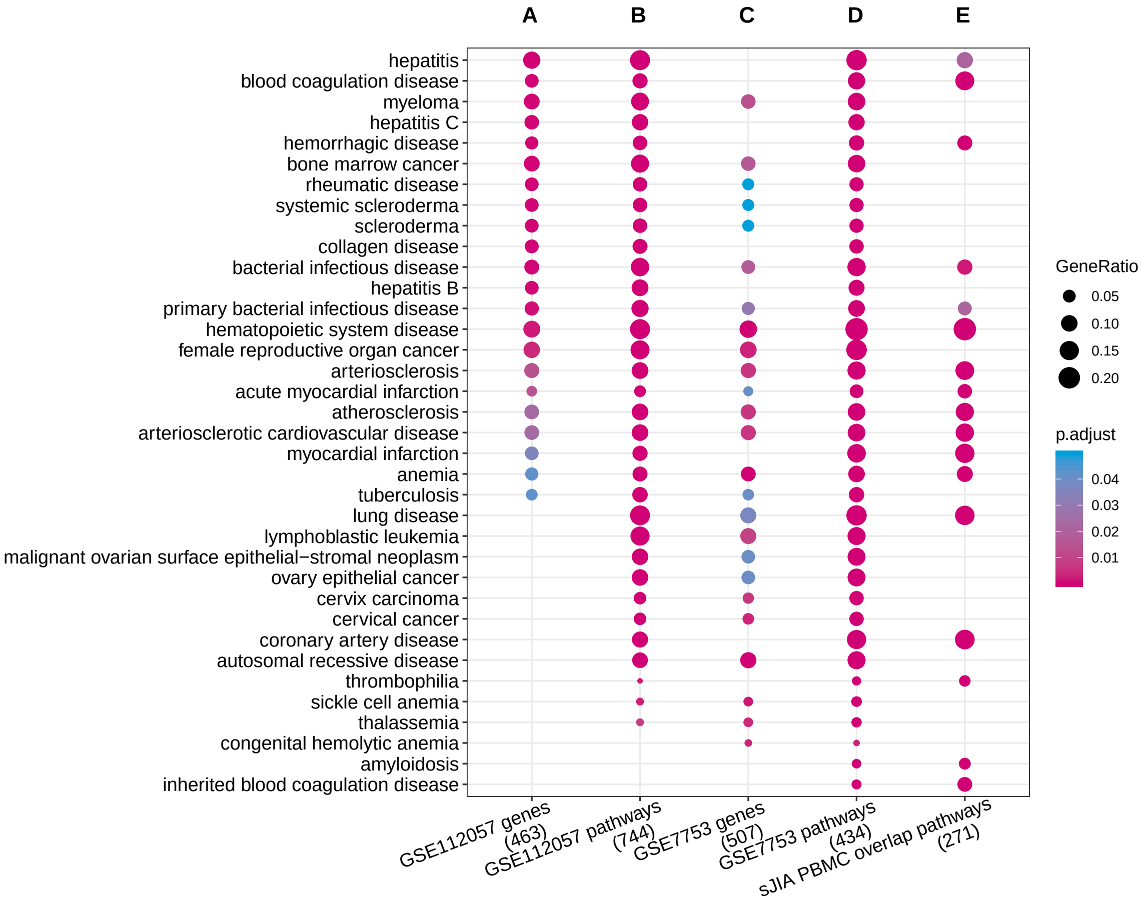


Fig. (6.2) Disease ontology enrichment in sJIA differential gene and pathway expression analysis. Two sJIA studies, GSE112057 (A–B) and GSE7753 (C–D), were analysed on gene- (A, C) and pathway-level (B, D). Enrichment results for the overlapping DEPs from 5 sJIA studies from PBMC are in E (rightmost). Enrichment was calculated for the top 1000 differentially expressed genes and the top ~1000 gene members from the most differentially expressed pathways according to highest |logFC| (p-value < 0.05). The top 10 terms (ranked by p-value) are listed per study. Additional terms are marked in each case if they overlap with the top 10 from another test. DEP — differentially expressed pathway; p.adjust — q-value; PBMC — peripheral blood mononuclear cells; sJIA — systemic juvenile idiopathic arthritis.

pathway-level tests (Fig. 6.2B,D) had a higher overlap between the top DO terms, including comparable gene ratios, e.g. higher gene ratio in *hepatitis* and *hematopoietic system disease*, and lower ratios in *thrombophilia* and *sickle cell anaemia*.

The sJIA PBMC overlap signature enriched in 122 DO terms (A∩E = 25, B∩E = 113, C∩E = 35, D∩E = 115, A∩B∩C∩D∩E = 13). There were fewer enriched terms than in two other pathway-level tests (B, D), suggesting that when identifying the intersecting pathways across multiple studies, we lost some information, but retained the consensus signature. Based on previous analysis of DEP profiles (Supplementary Fig. D.1 and top DEPs (Table 6.4), it was likely that these differences were driven by GSE8050_GPL97. The top enriched terms overlapped with terms found in the other tests (Fig. 6.2E), in particular all top terms from E were also in D.

The enrichment analysis of DO terms suggested that there was concordance between top terms identified by gene- and pathway-level tests. However, the pathway-level analysis enriches in more DO terms and led to fewer disparate terms when comparing two pathway-level analyses from different platforms compared to gene-level analysis comparing the same two studies. The sJIA PBMC overlap pathway signature enriched in fewer DO terms than the other two pathway-level tests (B, D), suggesting reduction of signal and increase in noise when joining 5 studies from the same disease and tissue, but from 3 different platforms.

6.4 Evaluating Drug Prioritisation

After the identification of disease pathway signatures, we continued with the signature processing and drug prioritisation step in the PDxN drug repositioning pipeline (Section 5.5). In brief, the most up- and down-regulated disease signature pathways were separated into up- and down-regulated pathway clusters. We defined the pathway clusters as 5, 10, 15 or 20 up- or down-regulated pathways. The PDxN sub-networks were then constructed for each pathway cluster including all pathway nodes from the cluster and all the drug nodes that were connected to at least one cluster pathway. The PDxN correlation edges were then summarised per pathway cluster followed by summarisation step per drug direction (joining the up- and down-regulated drug signature node). The remaining summary edges between each *pathway cluster*↔*drug* pair were prioritised by increasing summary score for up-regulated pathway cluster and decreasing for the down-regulated pathway cluster (see Fig. 5.15 and Supplementary Fig. C.4).

In this section, we first looked at the top prioritised drug candidates for GSE7753 and then assessed the performance of the pipeline by scoring the resulting prioritised drug list with a gold-standard list of true positive drugs. We assessed the influence of the drug signature features (e.g. cell line, concentration, batch) and the quality of the true positive list on the performance score.

6.4.1 Prioritised drug candidates

We prioritised drug candidates for each of the sJIA PBMC studies as well as the sJIA PBMC overlap signature. For each study we analysed up- and down-regulated pathway clusters with 5, 10, 15 or 20 pathways. For clarity, we referred to these clusters by their abbreviated names consisting of direction and size, e.g. *up10* represents the cluster consisting of the top 10 up-regulated cluster. In addition, clusters: *up10*, *up15* and *up20* were together notated as *up10:20*. A prioritised drug list was generated by interrogating the PDxN with each pathway cluster. The resulting prioritised list returned a cluster score and the score q-value for each drug signature that was connected to the pathway cluster (Section 5.5). A drug signature is a unique combination of drug id, cell type, perturbation time, drug concentration and batch. Prioritised drug lists varied in length depending on how well-connected the disease pathway clusters are to the drug signature nodes (the lengths of each drug list are in Supplementary Table D.3).

The sJIA pathway clusters (including the overlap pathway signature) prioritised 2740–14,542 drug signatures (Section 5.5). In particular, the up-regulated clusters yielded 3467–14,542 drug signatures and the down-regulated yielded 2740–9810. Up-regulated clusters had a higher mean length compared to the down-regulated ones. The list length positively correlated with the mean size of the cluster separated by direction, indicating that in PDxN, larger pathway clusters (i.e. more pathways) were connected to more drug signatures.

The top drug candidates

While we prioritised drug candidates for each of the sJIA PBMC studies, we only conducted a literature review for the top drug candidates identified from GSE7753 disease pathway clusters. We investigated the top 10 drug signatures from the eight pathway clusters generated for GSE7753 (up- and down-regulated at 5, 10, 15, and 20 pathways each, Supplementary Table D.4). Across the 8 lists (*up5:20, down5:20*), 43 unique drugs were prioritised in the top 10 drugs. Of those, 23 drugs only ranked in one of the lists, and 11 of those appeared in clusters consisting of 5 pathways. Suggesting that drug lists from clusters with 5 pathways were less robust than from larger clusters.

Seven drugs ranked in the top 10 of *down10:20* (*diethylstilbestrol, H5902, MK-2206, SPECTRUM_000090, amlodipine base, SPB02303, geldanamycin*), one, wortmannin (BRD-A75409952), ranked in the *up5* and *up15:20* clusters, and 4 drugs ranked in the *up15:20* clusters (*SUGA1_008424, cobaltous chloride, 2541665-P1, FPA1_000240*). From these, MK-2206 (BRD-K68065987) consistently ranked 4th, SPECTRUM_000090 (BRD-A80151636) 5th, and geldanamycin (BRD-A19500257) 8th in the *down10:20*. SUGA1_008424 (BRD-K33164466) ranked 1st in the *up15* and *up-20*.

MK-2206 is an Akt inhibitor that has been shown to enhance the antitumour effect of chemotherapeutic agents. Higher levels of phosphorylated Akt have been found in synovial tissue from RA patients, which could be further increased with TNFα stimulation, suggesting that Akt contributes to stimulatory effects of TNFα (Zhang et al., 2001). Additionally, Pan et al. (2018) suppressed the Akt pathway with Quetiapine, an antipsychotic, which decreased the levels of pro-inflammatory cytokines such as IL-6 and IL-1β in an arthritis mouse model through subsequent NF-κB and CREB signalling pathways. **Geldanamycin** is an antitumour antibiotic that inhibits Hsp90 and downregulates Akt. It induces apoptosis and inhibits inflammation by suppressing NF-κB in RA fibroblasts-like synoviocytes (FLS). Together with infiltrated leukocytes, FLS contribute to RA progression (Ma et al., 2019). Geldanamycin has also been shown to inhibit the production of TNFα, IL-6, and IL-1β in activated macrophages (Wax et al., 2003). Our literature search did not yield any information on biological effects of SPECTRUM_000090 and SUGA1_008424. However, open-label studies of Bromocriptine, a **SPECTRUM_000090** stereoisomer, have resulted in significant improvement of clinical measures of RA (Figueroa et al., 1998, 1997).

Two drugs ranked in the top 10 in both up- and down-regulated clusters. **Gemcitabine** (BRD-K15108141) drug signature from 3 different cell types ranked in the top 10 of three up-regulated clusters and one down-regulated cluster. Gemcitabine is used in several cancers, and it also has antiproliferative effects on lymphocytes which contribute to RA pathogenicity. A study in collagen-induced arthritis (CIA) rats showed decreased TNF-α levels as well as reduced inflammation and cartilage destruction in the Gemcitabine-treated group (Dağli et al., 2017). **ALW-II-38-3** (BRD-K68191783) was present in two up- and one down-regulated clusters. ALW-II-38-3 is an Eph kinase inhibitor (Choi et al., 2009). Ephrin-B1 (EphB1), an ALW-II-38-3 substrate, expression was increased in synovial fibroblast cells of RA patients compared with osteoarthritis patients. An increase was also seen in peripheral blood lymphocytes of RA patients compared with healthy controls. In an RA animal model, activation of the EphB1 receptor resulted in an increase in TNFα and IL-6 production (Kitamura et al., 2008).

Cytarabine (BRD-K33106058) signature from two different cell types ranked 1st and 2nd in the *up10* cluster, 2nd in *up15*, and 6th and 7th in *up5*. It is used for haematological malignancies due to its ability to inhibit the DNA polymerase, which results in decrease of DNA replication and repair (Momparler, 2013). It has no anti-inflammatory activity, but it has shown immunosuppressant activity by inhibiting the onset of adjuvant-induced arthritis in rats as well as inhibiting local inflammation (Glenn, 1968; Glenn et al., 1977). **GDC-0980** (BRD-A18328003) ranked in the top 10 in all 4 down-regulated clusters (at rank number 9, 2, 3, 1, with increasing cluster size). GDC-0980 is a Phosphoinositide 3-kinase (PI3K)/mammalian target of rapamycin (mTOR) kinase inhibitor, that inhibits tumour cell growth and triggers apoptosis. Selective inhibition of PI3Kγ and PI3Kδ suppressed joint inflammation and damage in RA mice models (Camps et al., 2005; Randis et al., 2008).

Out of the 43 drugs, we were able to associate 7 with Anatomical Therapeutic Chemical (ATC) classification codes. ATC classification is a drug system that classifies the active ingredients of drugs according to the organ or system on which they act and their therapeutic, pharmacological and chemical properties (Section 5.3.5). 4 belong to L01B class of Antimetabolites, and one to each of the following classes: CO1BA Antiarrhythmics, class Ia, GO3C/LO2AA Estrogens, L04A Immunosuppressants, and NO6A Antidepressants. **Clofarabine**, which is in the L01B class with Gemcitabine and Cytarabine, is a purine nucleoside analogue that interferes with nucleic acid synthesis, terminating DNA chain elongation and inhibiting DNA repair. It is used for treatment of acute lymphoblastic leukaemia in children and young people. Clofarabine can inhibit methotrexate transport, by competitive binding of ABCG2, a methotrexate transporter (Nagai et al., 2011). No other connection to RA or JIA was found. **Pepstatin A**, in CO1BA ATC class, inhibits aspartyl proteases like HIV proteases and cathepsins D. Pepstatin increases the collagenolytic activity leading to collagen degradation (McAdoo et al., 1973). It does not reduce inflammation or joint destruction in RA rats (Biroc et al., 2001). **Diethylstilbestrol**, also known as stilboestrol, part of G03C and LO2AA, is an oestrogen agonist. It has been linked to cause many cancers after *in utero* exposure (Hatch et al., 1998). However, maintained pregnancy levels of oestrogen have been shown to sustain complete protection from postpartum exacerbation of CIA in mice, suggesting immunosuppressive effects of oestrogen (Mattsson et al., 1991). **2-chloro-2-deoxyadenosine**, also known as cladribine, is an immunosuppressant in L04A class. Schirmer et al. (1997) have shown that it can decrease T and B cell populations in patients with refractory RA. **Sertraline**, from the N06A class, is an antidepressant from selective serotonin reuptake inhibitor (SSRI) family. RA model rats showed significant reduction in arthritis upon treatment with Sertraline, this was accompanied by an increase in IL-10 and decrease in TNFα. The improvement was not significantly different to treatment with methotrexate, which is the most commonly used RA drug (Baharav et al., 2012). Additionally, another two SSRIs showed anti-inflammatory activity in human and mouse models of RA (Sacre et al., 2010).

In summary, we assessed the relevance of 13 out of 43 top drug candidates, based on their consistent presence in the top 10 in more than one prioritised list or their known ATC classification. Six drugs have shown improvement in RA patients, rat or mouse models, an additional three have shown to be involved in RA- and JIA-related processes, one has been shown to not improve the condition in rats and one has no recorded connection to arthritis. We could not find any biological information on two drugs, but an isomer of one was shown to improve RA. Although these results are encouraging, a more systematic approach is necessary for the evaluation of the pipeline's performance.

6.4.2 True positive lists for JIA

To systematically assess the pipeline's performance and the quality of the prioritised lists, we assessed how well our method prioritised approved drugs and disease-relevant ATC drug classes. We used four different lists of true positives (TPs): two consisting of approved drugs and two investigating the performance of two relevant ATC classes. The ATC class-based lists were used to overcome some approved-list limitations. Drugs are often approved for either treatment or symptomatic relief for a particular disease. In addition, they consist of several types of drugs, e.g. in JIA, it includes pain-relief as well as immunosuppressants. To compensate for these limitations, we compared our performance scored by an anti-inflammatory and antirheumatic class as well as an immunosuppressant ATC class of drugs.

Approved drugs for JIA (Supplementary Table D.5). This list would be expected to be the gold standard for testing the pipeline's performance. However, from 37 drug names (including international nonproprietary names (INN) and trade names) only 8 were mapped to 4 Broad (BRD) IDs and only 2 of those BRD IDs were present in PDxN. Both of those represented methotrexate. The 2 BRD IDs were present in 5 different drug signatures. Due to the limited number of TP drugs we decided to extend the list after the initial analysis to include RA-approved drugs.

Approved drugs for JIA and RA (Supplementary Table D.6). We included drugs from the list mentioned above and added approved RA drug names from EMA (European Medicines Agency, 2019) and RepoDB (Brown and Patel, 2016). We retrieved 106 INN and trade names for JIA and RA, 55 mapped to 76 BRD IDs. From those 40 were in PDxN in 183 different signatures. The approved drugs with the most signatures in PDxN are cyclosporine and auranofin with 62 and 31 drug signatures, respectively. Together they represent more than half of the approved signatures. Most benchmarking tests in this section used the approved drugs for JIA and RA as the TP list. We specified when alternative lists erre used.

ATC class M01: Anti-inflammatory and antirheumatic products (Supplementary Table D.7). 47 BRDs from the whole LINCS were mapped to 49 M01 codes. From those, 20 BRDs were present in PDxN in 88 signatures. Again, the majority, 31, of the drug signatures represent auranofin, and 21 represent valdecoxib. 13/20 PDxN BRD IDs in this list were also in the JIA and RA approved list.

ATC class L04A: Immunosuppressants (Supplementary Table D.8). This group included TNFα, IL-1, and IL-6 inhibitors as well as methotrexate. 14 BRDs from the whole LINCS were mapped to 20 L04A codes. From those, 17 BRDs were present in PDxN in 130 signatures. Nearly half (62/130) of the signatures represent ciclosporin (also known as cyclosporine). 9/17 PDxN BRD IDs in this list were also in the JIA and RA approved list. There was no overlap between M01 and L04A lists.

6.4.3 Benchmarking with JIA approved drugs

We first evaluated the performance of the method by scoring the generated prioritised drug lists with a TP list consisting of approved JIA drugs (Fig. 6.3).

We generated ROC curves for sJIA prioritised drug lists scoring their sensitivity and specificity with the approved JIA TP list. ROC curves and areas under the ROC curves (AUCs) are a standard way of reporting the method performance. The ROC curve measures sensitivity (recall) and specificity (true negative rate) of a given method based on a confusion matrix (Table 3.5). In this case, it assumed all predicted drugs that were not on the TP list were false positives or true negatives, depending on the threshold. This assumption was counterintuitive as it considered all novel drugs as incorrect. Even if drugs prioritised by this method were more appropriate for treatment than the approved drugs, our method would generate AUC $<< 1$. AUC $= 1$ indicates perfect performance, AUC $=$

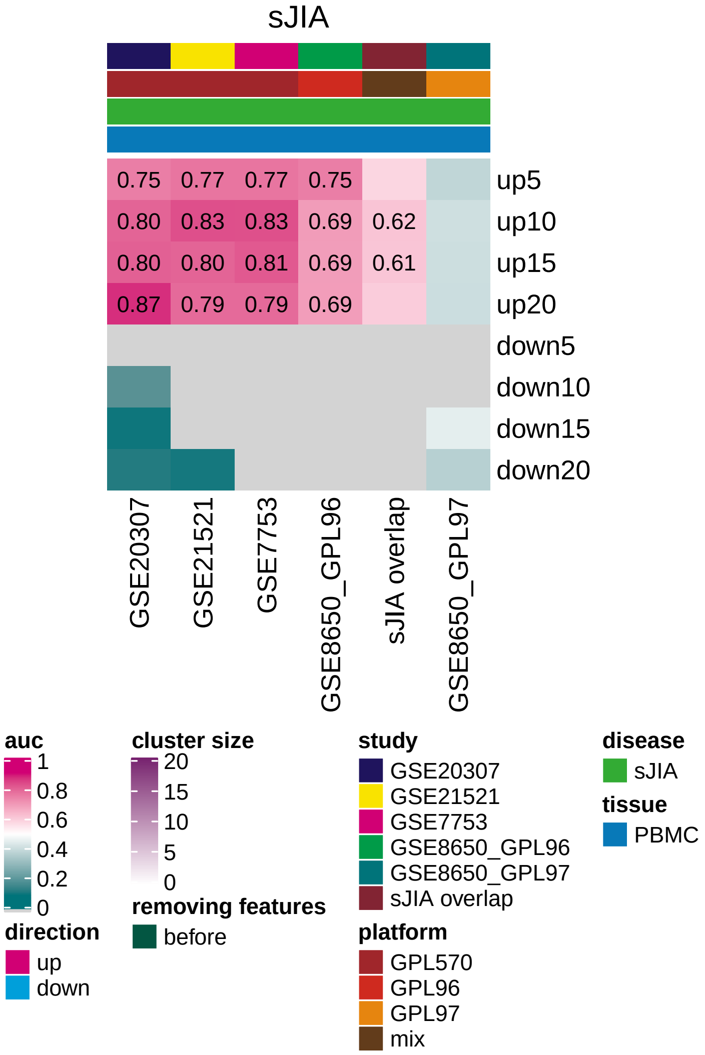


Fig. (6.3) AUC score summary for sJIA benchmarked with approved drugs for each condition. sJIA studies were benchmarked with JIA approved drugs (Supplementary Table D.5), Several prioritised drug lists did not associate a score with any TP drugs (grey square). The AUCs > 0.6 are displayed. AUC — area under the ROC curve; PBMC — peripheral blood mononuclear cell; ROC — receiver operating characteristics; sJIA — systemic juvenile idiopathic arthritis; TP — true positive.

0.5 is random performance, and AUC $<$ 0.5 is worse than random. The AUC scores from different disease clusters are summarised in Fig. 6.3.

The results showed superior performance for the sJIA up-regulated clusters compared to the down-regulated clusters. Due to the low number of TP in the list, there was a high number of prioritised lists without a score, meaning that no TP drugs were associated with a score in that prioritised drug list. All 5 signatures considered as TP represent methotrexate, the most commonly used JIA treatment (Ferrara et al., 2018), suggesting that our method was prioritising the current gold standard disease modifying treatment.

There were several high-scoring lists (AUC $>$ 0.7), but there were also a few poorly scoring lists, such as GSE20307 *down10:20*. These results suggested that prioritising drugs separately for the up- and down-regulated clusters might provide better drug candidates than if the disease signatures from both directions were considered together. Further work considering the effect of combining the prioritisation scores would provide further insight into this observed performance property. While our method showed great sensitivity and selectivity for the JIA-approved drug list, the list was very limited in length. However, it could have indicated that fewer, but better fitting drugs might serve as superior predictors of performance than the TP lists with a broader selection of mixed-effect drugs. To explore this further we benchmarked JIA with two different JIA-relevant therapeutic classes later in the chapter.

Due to the limited number of approved JIA drugs in PDxN, we decided to use an extended list including RA-approved drugs in the remainder of the chapter.

6.4.4 The effect of drug signature features on drug signature rank

When evaluating the performance of our pipeline with a larger TP list including JIA and RA approved drugs, we first assessed the effect of the drug signature experimental features on the ranking of the drug signatures in the prioritised list. Each drug signature had the following features: batch, drug ID, concentration of perturbagen, cell line, and perturbation time. We assessed the rank of the approved JIA and RA drugs across prioritised drug lists from sJIA studies' disease signatures (including sJIA overlap), for *up5:20* and *down5:20* pathway clusters. For each drug signature feature we computed the mean rank of TP signatures with that particular feature in each prioritised drug list (Fig. 6.4). For example, we calculated the mean rank of TP signatures tested in batch number CPC020, and separately we also calculated the mean rank of all signatures in the U937 cell line for each

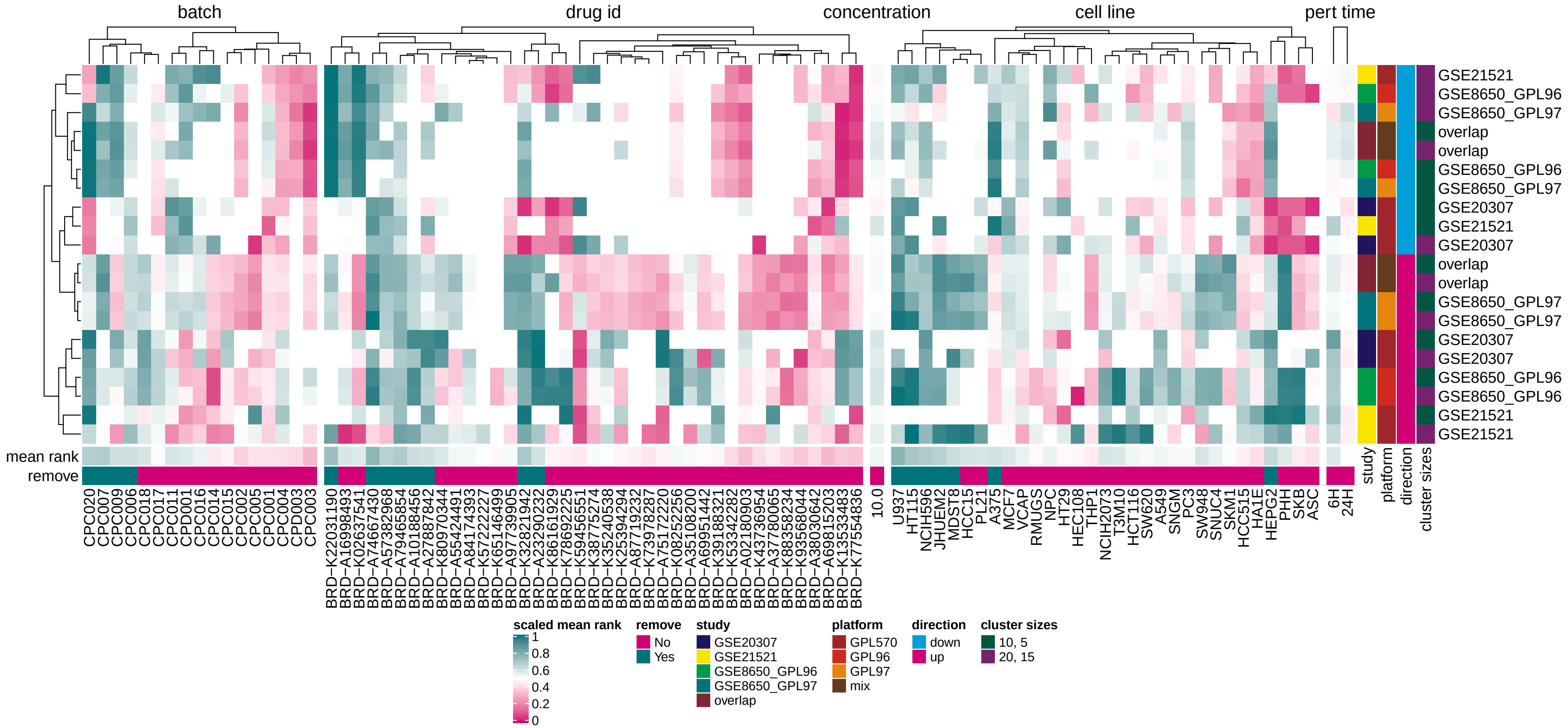


Fig. (6.4) Mean true positive (TP) drug rank in sJIA prioritised drug lists per drug signature feature. Drug lists prioritised by up- and down-regulated pathway signatures with 5, 10, 15 or 20 pathways, were scored with approved drugs for rheumatoid arthritis (RA) and JIA (Supplementary Table D.6). The pathway signatures were defined from sJIA PBMC studies and their overlap. Mean rank of 1 (top prioritised drug) was scaled to 0 and the worst rank to 1. Each row represents a combined prioritised list between 5 and 10 or 15 and 20 pathways up- or down-regulated pathways identified in one of the sJIA studies. Each column represents a drug signature feature (batch, drug id, concentration, cell line or perturbation time). Scaled rank of ∼0 (pink) indicated high ranking features in that prioritised drug list, while rank of 1 (teal) indicates low ranking drug signature features. The bottom margin (mean rank) indicates the average feature performance across all prioritised drug lists from all listed studies and pathway clusters. Drug signature features were selected for removal if the TP drugs with that feature ranked in the bottom 40% on average across different sJIA PBMC dataset signatures and their overlap signatures, leaving out the GSE7753 dataset. Only the TP signature features are used in this assessment. batch — experimental batch; JIA — juvenile idiopathic arthritis; PBMC — peripheral blood mononuclear cells; pert time — perturbation time; sJIA — systemic JIA.

drug list. Due to similarity in rankings in similarly sized pathway clusters, we combined the mean ranks between lists from 5 and 10 pathways, and 15 and 20. The mean ranks were scaled so that the best mean rank (of 1) was scaled to 0 and the worst rank (length of the list) to 1. The scaled rank represented the percentile at which the drug appears, e.g. a drug signature with scaled rank = 0.01 was at the top 1% of the prioritised drug signatures. We excluded GSE7753 from this analysis, so that we could cross-validate by assessing whether the patterns defined in other studies also held for GSE7753.

Similar patterns of rank appeared in the lists from the same direction signatures, suggesting non-random behaviour because disease signatures from different studies similarly ranked the TP drugs. In addition, lists from the same study, but different size clusters group together. The prioritised lists from the sJIA overlap clustered with GSE8650_GPL97 for up- and with GSE8650_GPL96 and GSE8650_GPL97 for down-regulated signatures. The overlap signature generated without GSE8650_GPL97 (data not shown) clustered with GSE8650_GPL96 for up- and with GSE8650_GPL96 and GSE21521 for down-regulated signatures. These results suggested that the overlap signature results in the downstream analysis would be the most similar to GSE8650_GPL97, while if GSE8650_GPL97 was not included in the overlap generation they would be the most similar to GSE8650_GPL96. There were batches, drugs and cell lines which consistently rank at the top and some that ranked at bottom of the lists, independent of pathway cluster direction (Fig. 6.4). We calculated the mean scaled rank per feature in order to identify best and worst performing features. If the feature was not present in one of the prioritised drug lists, we assigned a mean rank of 0.5.

Batch. 4 batches were identified as consistently low-ranking. Some batches included signatures from a limited number of drugs or cell ids, therefore the batch effect was confounded to other variable features. **Drug ID.** Out of 8 drugs identified as low ranking across all lists, 3 were in S02BA Corticosteroids, 3 in M01 Anti-inflammatory and Antirheumatics products, one in L04A Immunosuppressants and one in N02 Analgesics ATC class. Among the M01 drugs was auranofin (BRD-A79465854), which was one of the most frequent drugs in JIA and RA approved drug signatures. **Concentration.** All TP signatures were tested in 10μM. As expected, the mean rank across all drug lists was $\sim$0.5. We expected this as on average, if the drug signature feature displayed no ranking bias, the scaled mean rank would equal 0.5, indicating that the feature ranks in the middle of the prioritised drug list. **Cell type.** From the 7 cell lines identified as consistently ranking in the bottom of the prioritised list, two were from colon, and the rest from blood, liver, lung, skin and uterus. This could suggest that drug signatures tested in these cell lines were not appropriate for JIA drug evaluation. **Perturbation time.** Although signatures perturbed for 6 hours ranked lower on average than 24h signatures, they did not consistently rank in the bottom 40% of the lists.

We assessed the effect of removing the worst performing features by removing them from the drug list when assessing specificity and selectivity of the whole prioritised list. We defined the worst performing features as the features whose mean scaled rank across all drug lists was above 0.6, in other words, that feature was on average in the bottom 40% of the list, independent of direction, study, or size of the pathway cluster.

We generated ROC curves and their AUCs for GSE7753 prioritised drug lists (Fig. 6.5A–B) scoring their sensitivity and specificity with the approved JIA and RA TP list. The performance AUC scores for the sJIA studies including GSE7753 are summarised in Fig. 6.5C. The initial benchmarking with JIA- and RA-approved TP list suggested random or worse than random performance of our method (AUC $<$ 0.5). However, removing consistently low-ranking features, identified from all but GSE7753 sJIA prioritised drug lists, improved the overall performance of GSE7753, as well as other studies' performance. However, those were involved in the identification of low-ranking features, thus directly influencing the ranking (Fig. 6.5C). Therefore, there was a risk that the identified features are overfitted to those studies. Prioritised lists for both directions improved after removing low-ranking features, however AUCs from more down-regulated clusters surpassed AUC $>$ 0.6.

6.4.5 Benchmarking with immunosuppressant, and anti-inflammatory and antirheumatic drugs

To further assess the influence of TP drug lists on the performance score, we investigated the individual contribution of two major drug groups approved for JIA and RA: immunosuppressants, and anti-inflammatory and antirheumatic drugs. When identifying individual effects of drug signature parameters, the consistently low-ranking drugs were predominately in M01 anti-inflammatory and antirheumatic drugs ATC class, suggesting that those drugs were deprioritised by our system. We hypothesised that M01 class drugs will generate lower AUC scores compared to those of L04A Immunosuppressants ATC class based on relative potency and the current treatment trends in JIA. We included all drugs annotated with their respective class, not only those approved for JIA and RA.

While scoring prioritised drug lists with approved JIA and RA drugs suggested random or worse than random performance (Fig. 6.6A), scoring it with M01 indicated much

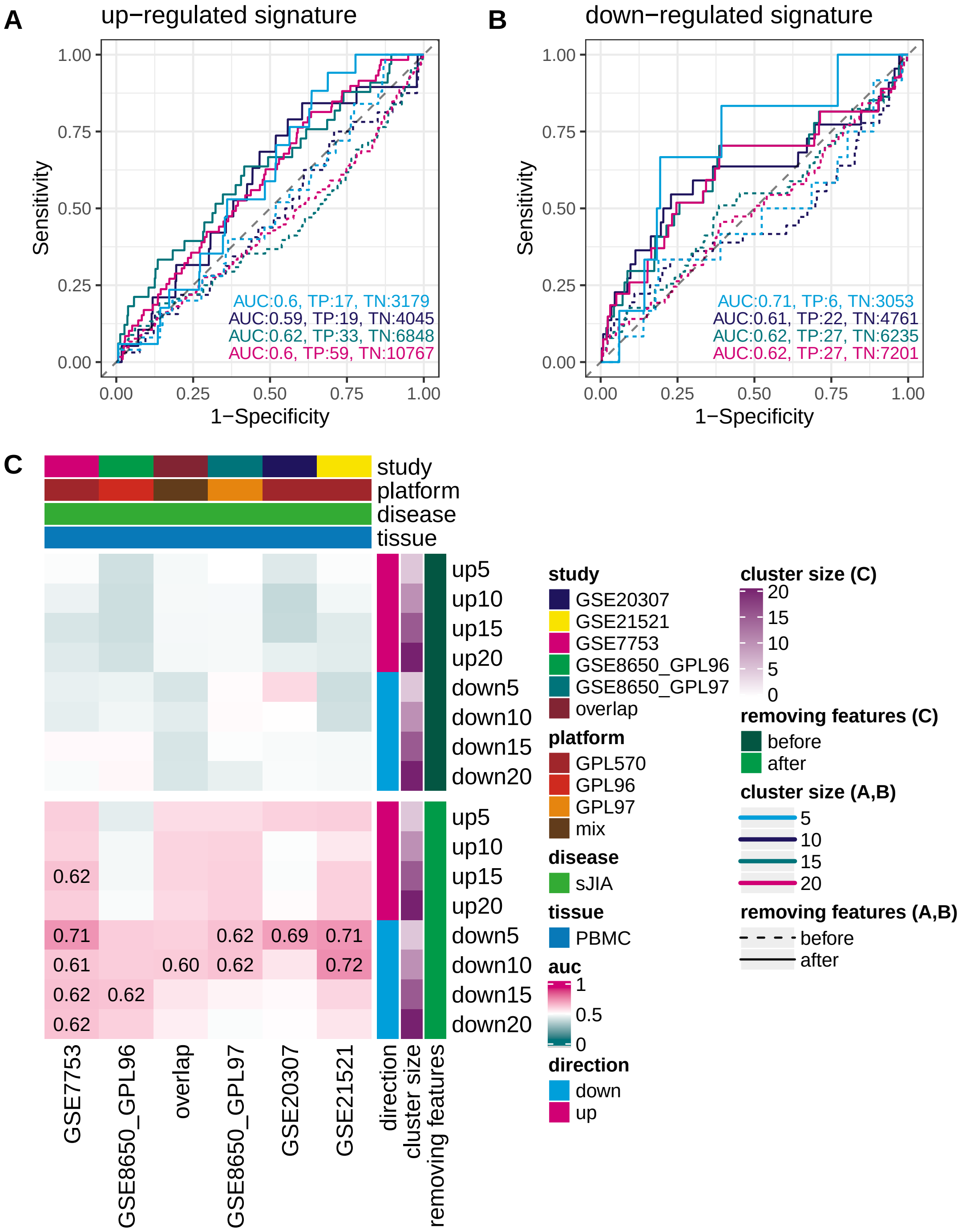


Fig. (6.5) AUC score summary for sJIA studies before and after removing low-ranking drug signature features. (A–B) ROC curves for the PDxN-derived drug lists, prioritised by the disease signatures defined from the GSE7753 study. Lists were scored against an approved drug list for rheumatoid arthritis (RA) and juvenile idiopathic arthritis (JIA) (Supplementary Table D.6). Performance is indicated for (A) up- and (B) down-regulated disease pathway clusters of different sizes before (dotted line) and after (full line) the removal of low-ranking drug signature features (identified in Fig. 6.4). The AUC, TP and TN counts are displayed for the *after* ROC curves. (C) Summary of AUCs per sJIA PBMC study, signature direction and pathway cluster size before (C top) and after (C bottom) the removal of low-ranking features. AUC — area under the ROC curve; PBMC — peripheral blood mononuclear cells; PDxN — Pathway Drug Coexpression Network; ROC — receiver operating characteristics; sJIA — systemic JIA; TN — true negative; TP — true positive.

worse performance (Fig. 6.6B), but scoring with the L04A class indicated better than random in several studies (Fig. 6.6C). The drug lists from down-regulated pathway clusters ranked the M05 drugs predominantly in the bottom half of the lists (AUC $<$ 0.5). The up-regulated clusters ranked the L04A drugs in the top half, generating AUCs $>$ 0.7 in three studies. These results indicated that the drug prioritisation method prioritised several immunosuppressant drugs in most sJIA signatures, while associating low ranks with anti-inflammatory and antirheumatic drugs.

As mentioned above, the many NSAIDs in the anti-inflammatory category, while the most common, are often only used for pain relief, or as initial treatment during JIA diagnosis (Armon, 2018; Ringold et al., 2013). Systemic patients benefit most significantly from immunosuppressant treatment (Ravelli et al., 2018). There has been a recent move towards immediate aggressive treatment, yielding better outcomes for patients (Ravelli et al., 2018). Our prioritisation results reflected the currently emerging trends in treatment of sJIA. In addition, although every effort had been made to curate only studies with previously untreated samples, it is unrealistic to expect that children with sJIA refrained from pain-relief medications included in the M01 group. Thus, the fact that our method deprioritised the M01 group could be due to patients already under treatment from M01 drugs, or perhaps due to M01 not being the most appropriate course of treatment for sJIA patients.

6.4.6 Comparison with LINCS clue.io method

We compared our method to LINCS (Subramanian et al., 2017). LINCS is an alternative, well-established repositioning pipeline available online at *https://clue.io/lincs*. The proof of concept for LINCS, the Connectivity Map (Lamb et al., 2006), has been cited 3334-times and the extended LINCS dataset (Subramanian et al., 2017) has been cited 457 times (both as of 2nd March 2020 on Google Scholar). We chose LINCS as a comparable method to PDxN as both are based on the same hypothesis of signature reversion (Fig. 2.4). In addition, the drug signature data used in PDxN has been developed by the LINCS project, thus the underlying drug data is the same between the two methods. The main difference between the methods is that PDxN is a pathway-based, and LINCS is a gene-based drug prioritisation method.

We queried *https://clue.io/lincs* with differentially expressed genes from sJIA study GSE7753. Because we investigated the top prioritised drugs at 4 different-sized clusters in PDxN, we submitted 4 different-sized sets of genes to LINCS. Per website recommenda-

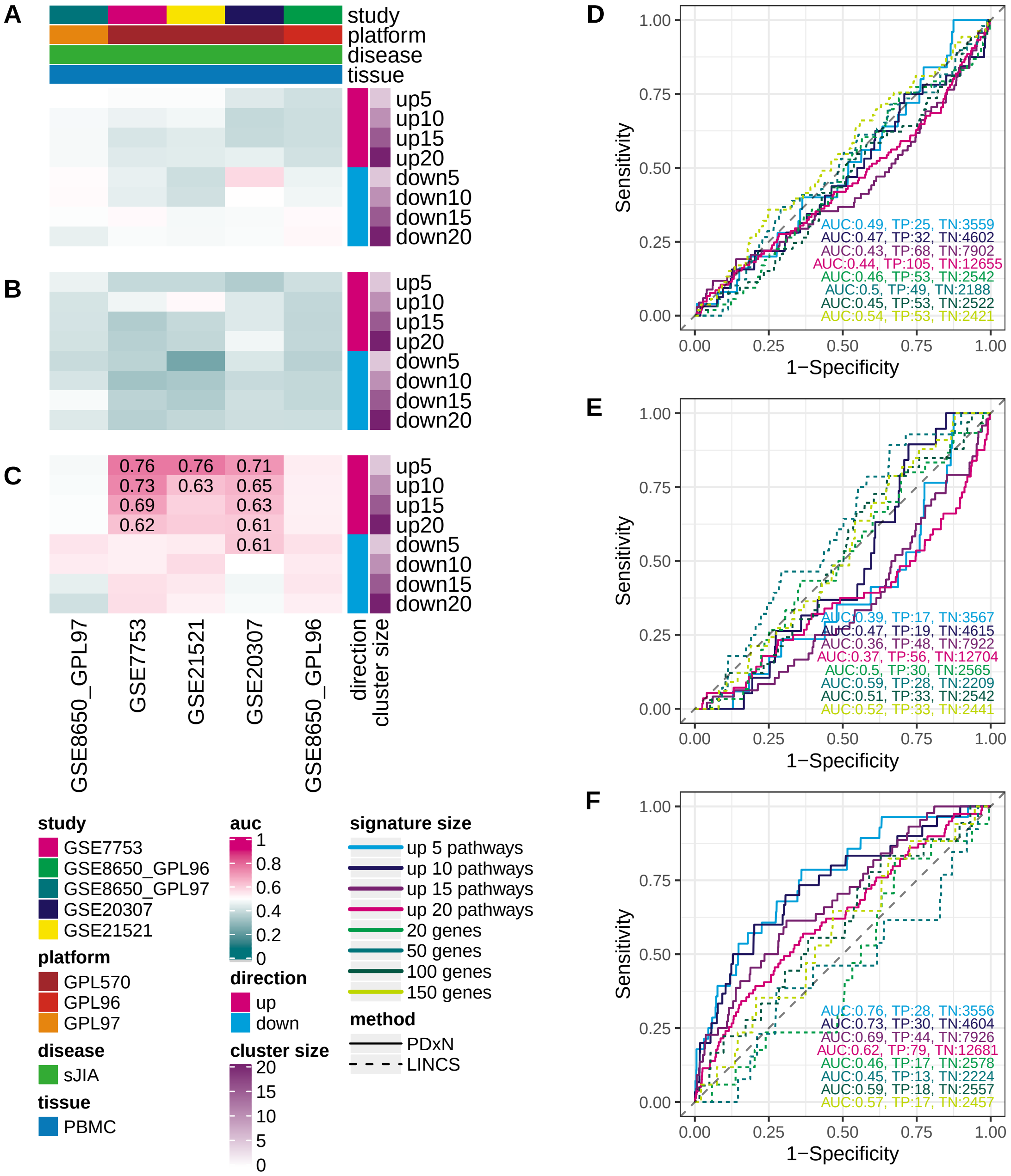


Fig. (6.6) AUC score summary for sJIA benchmarked with 3 different true positive lists. True positive lists used for scoring the sensitivity and specificity of prioritised drug lists: (A,D) EMA and FDA RA and JIA approved drugs, (B,E) Anti-inflammatory and antirheumatic products (ATC M01), (C,F) Immunosuppressants (ATC L04A). (A–C) AUC summary heatmaps for sJIA PBMC studies with different true positive lists. (D–F) ROC curves scored with 3 different TP lists comparing PDxN vs LINCS performance from GSE7753 disease signatures. ATC — Anatomical Therapeutic Chemical; AUC — area under the ROC curve; EMA — European Medicines Agency; FDA — the US Food and Drug Administration; JIA — juvenile idiopathic arthritis; LINCS — Library of Integrated Network-based Cellular Signatures; PBMC —– peripheral blood mononuclear cells; PDxN — Pathway Drug Coexpression Network; RA — rheumatoid arthritis; ROC — receiver operating characteristics; sJIA — systemic JIA; TN — true negative; TP — true positive.

tions, we submitted the top up- and down-regulated gene lists between 10–150 genes (20, 50, 100, and 150 per direction, q-value $<$ 0.05). We used the recommended summarised score across cell lines for comparison with PDxN.

Benchmarking the LINCS-prioritised lists with approved JIA and RA drugs yielded similar AUC scores to PDxN (LINCS: 0.45–0.54 (mean AUC = 0.49), PDxN: 0.43–0.49 (mean AUC = 0.46), Fig. 6.6D). Because we demonstrated that the AUC$\approx$0.5 could be driven by opposite performance from two major drug classes represented in the approved TP list, we benchmarked LINCS with M01 and L04A ATC class drugs (Fig. 6.6E–F). While PDxN-analysed GSE7753 signatures decreased their performance in M01 (mean AUC = 0.40) and increase it in L04A-benchmarked tests (mean AUC = 0.70), LINCS showed little change in performance when benchmarked with different lists (LINCS mean AUCs: M01: 0.53, L04A: 0.52).

The differences in performance between PDxN and LINCS could be due to underlying differences in the two methods. The factors influencing the different performance could be that:

(i) the pathway-level signatures were more representative of underlying biology,

(ii) the correlation network approach was more appropriate for drug repositioning,

(iii) joining opposite-direction signatures decreased sensitivity.

In summary, PDxN and LINCS gave a comparable AUC score when benchmarked with the full panel of JIA and RA approved drugs. PDxN outperformed LINCS when scoring for the immunosuppressant group of drugs, which has been shown to be more relevant in current treatment of JIA. LINCS outperformed PDxN when scoring for M01 drugs, due to the PDxN deprioritisation of anti-inflammatory drugs.

6.5 Discussion

In this chapter we have demonstrated the pipeline's ability to prioritise drugs relevant to JIA. We have shown that our disease signature generation method was robust for its ability to generate a consensus signature. It yielded highly concordant lists of pathways across disparate JIA studies derived from different platforms. We have shown that the JIA disease signatures were contrasting to that of the liver control study (GSE133815). The liver samples from old and young individuals were a good control as they came from the same organism and were profiled on the same platform as the majority of the JIA studies. Additionally, they represented a non-inflammatory condition. A drawback of using GSE133815 as control was that the study included a relatively low number of samples as well as that the samples were not age-matched to those of JIA. We have shown that the top dysregulated pathways we have identified in our representative studies were related to JIA. They contained dysregulated genes and pathways enriched in blood and immune response-related disease ontology terms as well as *rheumatic disease*.

We have identified promising prioritised drug candidates from GSE7753. Out of 13 literature-searched drugs, six have shown improvement in RA patients, rat or mouse models, an additional three have shown to be involved in RA- and JIA-related processes. Two had no associated biological information, but an isomer of one of them has shown an improvement in RA. In order for these top drug candidates to be validated experimentally, the findings from the literature review would be used and the most promising drug candidates and drugs related to those would be further theoretically evaluated for their solubility, bioavailability and toxicity. The top candidates with beneficial solubility and toxicity profiles would then be tested in an *in vitro* disease model to assess its efficacy and to optimise their dose and exposure times, before any further *in vitro* and *in vivo* tests would be carried out. Therefore, each promising drug candidate identified in an *in silico* drug repositioning pipeline undergoes extensive *in vitro* and *in vivo* validation before it would be used in a clinical trial.

While we have shown promising disease signatures and drug candidates we need to consider an important limitation in this case study. We excluded JIA studies focusing on the investigation of particular treatments during our study selection process, several JIA studies used in this chapter included samples from children treated with NSAIDs. In addition, the selected studies varied in their sample sizes. A weighted approach to reflect study sample sizes and confounding use of NSAIDs could be explored in the future development of the consensus disease signature generation. This approach could enhance the value of results from studies with high sample numbers as well as decrease the importance of studies in which the patients have received treatment. In addition, because the GSE7753 included samples treated with NSAIDs, the repositioning candidates from the literature search might be associated with NSAID use and not the underlying sJIA pathology. To correct for this confounding effect drug signatures from NSAIDs or other administered medications could be used. Alternatively, results from studies with samples from untreated patients could be prioritised.

We have demonstrated the influence of drug signature features such as batch, drug ID and cell type, on the drug signature ranking. We explored two definitions of TP lists for benchmarking the method's performance: (i) two lists of approved drugs and (ii) two therapeutically relevant ATC classes. We highlighted the limitation of using approved drugs as a gold-standard TP list for diseases with low number of approved drugs in addition to showing distinct performance of two therapeutic classes of drugs that were combined in the extended approved TP list. Usage of ATC classes as TP drug lists for benchmarking showed that a particular therapeutic class could be prioritised over another.

We have shown that our method prioritises immunosuppressant over anti-inflammatory and antirheumatic drugs, which reflects the current treatment recommendations. By assessing performance with therapeutic classes we have highlighted how PDxN could be used to prioritise not only individual drugs but also classes of drugs. Prioritising the immunosuppressant and disease-modifying group of drugs over the anti-inflammatory group that provides only symptomatic relief pointed to the possibility of *de novo* discovery of best therapeutic classes for diseases with unknown or reduced treatment options.

Chapter 7

Case Studies: Neurodegenerative Diseases

In this chapter, we have applied our method to two additional case studies: Alzheimer's and Parkinson's disease. We have chosen these devastating neurodegenerative diseases due to the current lack of disease-modifying treatments. This work has been performed in collaboration with two world-leading groups in their respective areas: the Tanzi and Kim group from Harvard Medical School and Massachusetts General Hospital for Alzheimer's disease (AD), and Bandmann group from the Sheffield Institute for Translational Neuroscience for Parkinson's disease (PD). Our collaborators have provided relevant RNA-Seq data, analysed in this chapter, together with drug lists used for evaluation of the method. The results from this chapter provided a unique translational opportunity to our collaborators, enabling them to test prioritised drug candidates that have been carefully benchmarked, *in vitro* and *in vivo*.

Each case study section follows the outline of the previous chapter (Chapter 6: Evaluation of the System: Application to juvenile idiopathic arthritis (JIA)) focusing on the drug repositioning pipeline results. We conclude the chapter with parallel evaluation of drug prioritisation results for approved drugs for the two neurodegenerative case studies: AD and PD.

Others' contributions to this chapter

Alzheimer's disease. Assistant Professor Doo Yeon Kim (Department of Neurology, Massachusetts General Hospital, Harvard Medical School) and Professor Rudolph E. Tanzi (Department of Neurology, Massachusetts General Hospital, Harvard Medical School) shared preprocessed 3D cell model RNA-Seq data described in Kwak et al. (2020), and the positive drug hits from the 3D drug screen. The RNA-Seq data was preprocessed by the Harvard Bioinformatics Core. Assistant Professor Doo Yeon Kim has also kindly provided context for their 3D models and their characterisation. Sarah Morgan (Sheffield Institute for Translational Neuroscience (SITraN), University of Sheffield; Department of Pathology, Beth Israel Deaconess Medical Center, Harvard Medical School) advised on the Mayo Alzheimer's disease dataset sample selection.

Parkinson's disease. Professor Oliver Bandmann (SITraN, University of Sheffield) shared the human sporadic Parkinson's disease RNA-Seq data described in Carling et al. (2020) and the zebrafish GCH1 mutant data described in Larbalestier et al. (2020). Professor Oliver Bandmann has also kindly curated a list of neuroprotective drugs. The Carling et al. (2020) RNA-Seq data was preprocessed by Claire Green (SITraN, University of Sheffield) and the Larbalestier et al. (2020) zebrafish data by Wenbin Wei (SITraN, University of Sheffield).

7.1 Case Study Design

The pipeline was first applied to AD, results were interpreted, and then we applied the pipeline to PD. For each case study, as in Chapter 6, we generated pathway-level disease signatures (Section 5.4), which were then applied to the signature processing method (Section 5.5). We evaluated the resulting prioritised drug lists with known or predicted true positive (TP) drugs. Both case studies were designed with the aim of prioritising drugs based on a disease signature defined from human data. The top prioritised drugs have the potential to be validated *in vitro* and *in vivo*. However, to establish the confidence in the method's performance we focused on prioritising drugs from three-dimensional (3D) cell models for AD. We focused our analysis on the PD human fibroblast data to test the limitations of our current approach. The collaborator-provided PD datasets include a low number of case and control samples, thus we aimed to explore how low-sample number affects the pipeline's performance.

Across both case studies, we analysed 8 different datasets (Table 7.1). In AD, our analysis was based on AD organoid models and comparison with human subjects. We analysed 4 different 3D disease model cell cultures (Choi et al., 2014; Kwak et al., 2020) and compared the cell line results to the Mayo dataset (Allen et al., 2016) which included human temporal cortex samples from deceased subjects with AD. In PD, our analysis was based on fibroblasts from sporadic PD (sPD) patients with either characterised lysosomal or mitochondrial dysfunction (Carling et al., 2020). We subsequently compared those results to a GCH1 mutant zebrafish PD model (Larbalestier et al., 2020).

Table (7.1) Overview of neurodegenerative datasets. All datasets include RNA-Seq data. The control and disease samples reflect the number after quality control (QC). Sample numbers before QC are listed in Table 3.2. AD — Alzheimer's disease; Lyso — lysosomal dysfunction; Mito — mitochondrial dysfunction; PD — Parkinson's disease.

Dataset	Tissue	Control samples	Disease samples	Disease	Organism	Feature	Reference
A5	neurons	3	3	AD	3D cell line	low Aβ42/40	Choi et al. (2014)
H10	neurons	3	3	AD	3D cell line	high Aβ42/40	Choi et al. (2014)
I47F	neurons	3	3	AD	3D cell line	low Aβ42/40	Kwak et al. (2020)
I45F	neurons	3	3	AD	3D cell line	high Aβ42/40	Kwak et al. (2020)
Mayo	temporal cortex	37	69	AD	human	Braak ≥ 4	Allen et al. (2016)
Lyso	skin fibroblast	4	4	PD	human	high lysosome counts	Carling et al. (2020)
Mito	skin fibroblast	4	4	PD	human	low ATP levels	Carling et al. (2020)
GCH1	neurons	3	4	PD	zebrafish	low *gch1* expression	Larbalestier et al. (2020)

7.2 Alzheimer's Disease

7.2.1 Disease introduction

Alzheimer's disease (AD) is a progressive, multifarious, neurodegenerative disorder. It is the most common type of dementia and one of the great health-care challenges of the 21st century. In 2017, it was the sixth leading cause of death in the US. The unpaid care provided by family members and caregivers to people with Alzheimer's and other dementias is valued at ~$230 billion (Alzheimer's Association, 2019).

AD-related pathology is thought to begin 20 years or more before AD symptoms arise, whereby an accumulation of brain changes cause memory loss and language problems (Villemagne et al., 2013). As the disease progresses, the accumulation of toxic protein or protein fragment deposits leads to neurodegeneration. Pathologically AD is characterised by intracellular neurofibrillary tangles and extracellular amyloid protein deposits contributing to senile plaques (Braak and Braak, 1991). The pathological changes lead to inflammation and contribute to atrophy. The accumulation of toxic β-amyloid and tau activates microglia, the primary innate immune cells in brains responsible for clearance of toxic fragments and proteins. While the neuropathological features of AD are recognised, little is known about the causes of the disease and no curative treatments are available (Selkoe and Hardy, 2016).

The majority of the approved drugs for AD temporarily improve symptoms by increasing the number of neurotransmitters in the brain. They are mostly acetylcholinesterase (AChE) inhibitors, that improve cognition, but do not slow disease progression. Memantine blocks glutamate receptors in the brain to reduce excess simulation that can damage nerve cells. The effectiveness of these drugs varies from person to person and is limited in duration (Yiannopoulou and Papageorgiou, 2013).

The greatest risk factors for late-onset AD are age, carrying the APOE-ε4 mutation and having family history of the disease. APOE encodes a cholesterol transporter (Fratiglioni et al., 1993; Hebert et al., 2010; Saunders et al., 1993). Regular physical activity, management of cardiovascular risk factors, lifelong learning and cognitive training are associated with reduced risk of cognitive decline (Baumgart et al., 2015).

7.2.2 Alzheimer's disease datasets

In our AD case study, we used RNA-Seq data from 4 different 3D cell culture models of AD developed by Tanzi and Kim lab, and one publicly accessible dataset from deceased AD human brains (Table 7.1).

3D cell cultures (Choi et al., 2014; Kwak et al., 2020)

We used data from 4 different 3D human neural cell culture models of AD (Table 7.1). A5 and H10 are human neuronal progenitor cells (hNPCs) with high levels of toxic amyloid-β (Aβ) species overexpressing human amyloid precursor protein (APP), and APP and

Presenilin 1 (PSEN1), respectively. I47F and I45F are hNPCs with an APP transmembrane domain (TMD) mutation.

A5 and H10 are constructs with a Swedish and London familial AD (fAD) APP mutation. In addition, H10 also has the fAD PSEN1ΔE9 mutation. I47F and I45F include the Swedish fAD APP mutation and have their respective isoleucine (I) to phenylalanine (F) mutations in the APP TMD.

Patients with the Swedish or London mutation display early-onset AD. The Swedish double (KM670/671NL) mutation is immediately adjacent to the β-secretase cleavage site in APP. It increases production of Aβ by competitive β-secretase cleavage of N-terminal APP (Haass et al., 1995). The London (V717I) mutation occurs in the APP TMD and increases the Aβ42/40 ratio by having little effect on Aβ40 (Eckman et al., 1997; Goate et al., 1991; Hardy and Allsop, 1991). The PSEN1ΔE9 (S290C;T291_S319del) mutation removes a 5.9kb sequence from PSEN1 due to an amino acid substitution at the splice junction of exon 8 causing the skipping of exon 9 (Smith et al., 2001). The mutation causes impaired APP processing and increased Aβ42/40 ratio (Borchelt et al., 1996). I47F (I718F) causes Aβ48–42 blockage decreasing the Aβ42/40 ratio and I45F (I716F) causes Aβ49–40 blockage increasing the Aβ42/40 ratio (Kwak et al., 2020; Lichtenthaler et al., 1999).

The A5 and I47F cell cultures develop low Aβ42/40 ratio, while the H10 and I45F develop a high ratio (Table 7.1). Most fAD mutations increase the Aβ42/40 ratio, strongly suggesting that an increased ratio plays an important role in AD pathogenesis (Tanzi and Bertram, 2005). The clonal hNPCs recapitulate amyloid-β and tau pathology. They exhibit phosphorylated tau in the soma and neurites, as well as filamentous tau (Choi et al., 2014; Kwak et al., 2020). Kwak et al. (2020) have shown that the Aβ42/40 ratio drives the tau pathology in 3D cell culture models.

Mayo dataset (Allen et al., 2016)

The Mayo dataset includes deceased human temporal cortex samples from the Mayo Brain Bank and Banner Sun Health cohort. The AD cases are diagnosed based on Braak score $\geq$ 4, while controls include elderly brains (Braak $\leq$ 3) without any diagnosed cognitive decline (Allen et al., 2016). The Braak score reflects the degree of pathology in AD, with a higher score indicating higher degree of accumulation of tau constituting neurofibrillary tangles (Braak and Braak, 1991). We selected AD and control samples over the age of 75.

The representative study

We focused our interpretation on the results from the A5 cell culture. The A5 was chosen because it has been optimised for drug screens. Due to their high Aβ42/40 ratio the H10 and I45F cell lines show increased neuronal death, thus making them less suitable for drug screens. Additionally, the I47F has only recently been developed and has not yet been optimised for drug screens.

We focused on the A5 model system rather than the human dataset, because the model system enabled the use of true positives from a drug screen targeting the molecular pathology of AD whereas the approved drugs for AD in humans offer only symptomatic relief. The Mayo dataset was used to identify promising top drug candidates that scored highly in both the AD model and the human dataset, highlighting the drugs with increased model-to-human translation potential.

7.2.3 Disease pathway signatures

The top 10 up-regulated pathways for A5 were mostly related to the inflammatory response (*Biocarta INFLAM pathway*, *KEGG Asthma*, *KEGG Graft versus host disease*, *KEGG Autoimmune thyroid disease*, *KEGG Allograft rejection*, *Reactome Phosphorylation of CD3 and TCR zeta chains*, *Reactome PD1 signaling*), brain development (*SIX3 Static module*, *CBX4 Static module*), and fatty acid metabolism (*KEGG Linoleic acid metabolism*) (Table 7.2).

Table (7.2) Alzheimer's disease (AD) disease signature pathways (A5 3D cell model). The top 20 most up- (rank: 1 to 20) and down- (rank: -1 to -20) regulated pathways (q-value < 0.05). Drugs were prioritised for up- and down-regulated pathway clusters at: the top 5, 10, 15 or 20 pathways by decreasing log fold change (LogFC) for up- and decreasing for down-regulated pathways. The genes in pathway column represents the number of possible genes in that pathway, while genes in data is the number of pathway genes found in data.

Rank	Pathway	LogFC	q-value	Genes in path-way	Genes in data
1	Biocarta INFLAM pathway	1.74	4.02e-05	29	9
2	KEGG ASTHMA	1.73	2.80e-04	30	6
3	KEGG GRAFT VERSUS HOST DISEASE	1.73	2.80e-04	42	6
4	KEGG AUTOIMMUNE THYROID DISEASE	1.72	2.94e-04	53	7
5	KEGG ALLOGRAFT REJECTION	1.72	2.94e-04	38	7
6	Reactome PHOSPHORYLATION of CD3 and TCR ZETA CHAINS	1.71	3.34e-04	16	5
7	Reactome PD1 SIGNALING	1.71	3.34e-04	18	6

continues on the next page

Table 7.2 continued

Rank	**Pathway**	**LogFC**	***q*-value**	**Genes in pathway**	**Genes in data**
8	SIX3 11 Static Module	1.64	3.06e-04	11	6
9	KEGG LINOLEIC ACID METABOLISM	1.64	1.11e-04	29	6
10	CBX4 10 Static Module	1.62	3.61e-04	10	8
11	MLH1 20 Static Module	1.62	1.98e-04	16	8
12	Reactome ACTIVATION of the AP1 FAMILY of TRANSCRIPTION FACTORS	1.61	1.09e-04	10	10
13	HIST3H3 14 Static Module	1.61	4.07e-04	14	9
14	Reactome PROLACTIN RECEPTOR SIGNALING	1.54	1.57e-03	14	8
15	HTATIP 20 Static Module	1.53	7.92e-05	19	19
16	Reactome POL SWITCHING	1.53	2.41e-04	13	13
17	Reactome TRYPTOPHAN CATABOLISM	1.49	4.22e-04	11	8
18	KEGG PRIMARY BILE ACID BIOSYNTHESIS	1.45	1.64e-04	16	8
19	KEGG SYSTEMIC LUPUS ERYTHEMATOSUS	1.30	1.11e-04	140	38
20	KEGG TYPE I DIABETES MELLITUS	1.23	3.61e-04	44	13
-1	Biocarta SRCRPTP pathway	-1.77	3.47e-05	11	10
-2	Reactome PEPTIDE CHAIN ELONGATION	-1.76	3.61e-05	153	81
-3	KEGG RIBOSOME	-1.76	3.61e-05	88	83
-4	TCF3 20 Static Module	-1.75	4.02e-05	20	12
-5	Reactome 3 UTR MEDIATED TRANSLATIONAL REGULATION	-1.75	3.61e-05	176	101
-6	Biocarta CBL pathway	-1.74	3.47e-05	13	11
-7	Reactome SRP DEPENDENT COTRANSLATIONAL PROTEIN TARGETING to MEMBRANE	-1.74	4.02e-05	179	104
-8	Reactome INFLUENZA VIRAL RNA TRANSCRIPTION and REPLICATION	-1.74	5.52e-05	169	97
-9	Reactome TRANSLATION	-1.72	3.61e-05	222	140
-10	Biocarta GLYCOLYSIS pathway	-1.71	3.47e-05	10	8
-11	Reactome FORMATION of the TERNARY COMPLEX and SUBSEQUENTLY the 43S COMPLEX	-1.71	3.61e-05	74	46
-12	Reactome RETROGRADE NEUROTROPHIN SIGNALLING	-1.70	3.47e-05	13	11
-13	Reactome ACTIVATION of the MRNA UPON BINDING of the CAP BINDING COMPLEX and EIFS and SUBSEQUENT BINDING to 43S	-1.70	4.02e-05	84	54
-14	Reactome FACILITATIVE NA INDEPENDENT GLUCOSE TRANSPORTERS	-1.69	1.00e-04	12	8
-15	Biocarta AMI pathway	-1.69	4.02e-05	20	10
-16	Reactome GLUTATHIONE CONJUGATION	-1.69	3.47e-05	23	16
-17	RPS27A 138 Static Module	-1.68	3.61e-05	138	132
-18	Biocarta KREB pathway	-1.67	8.72e-05	8	8
-19	Biocarta PTC1 pathway	-1.66	9.65e-05	11	9
-20	Reactome ORGANIC CATION ANION ZWITTERION TRANSPORT	-1.66	2.89e-04	13	7

Inflammatory response. All the inflammatory pathways identified in the top 10 upregulated pathway included cytokines such as interleukin-6 (IL-6), interleukin-1 (IL-1), cluster of differentiation 40 (CD40) and tumour necrosis factor (TNF) as well as major histocompatibility complex (MHC). In AD, neuroinflammation is triggered by microglia detecting misfolded and aggregated proteins. However, an inflammatory signal is not expected from the 3D AD model without microglia. Kwak et al. (2020) have shown that their clonal hNPCs differentiate into neurons as well as astrocytes, another primary innate immune cell in brains, in 3D cultures. Reactive astrocytes and microglia have been identified in the vicinity of the Aβ deposits (Verkhratsky et al., 2016). It has been shown that astrocytes are activated by Aβ, leading to increase in inflammatory factors like IL-1β, IL-6 and TNFα, which reduce synaptic and neuronal health in cell models of AD (Garwood et al., 2011; Phillips et al., 2014). **Brain development.** SIX3, the hub gene in *SIX3 Static module*, was identified in 3 different DNA methylation studies conducted in different cohorts and brain areas as differentially methylated in AD (Altuna et al., 2019; De Jager et al., 2014; Qin et al., 2020). SIX3 is involved in the equilibrium control between proliferation and differentiation of neural progenitor cells during mammalian neurogenesis (Appolloni et al., 2008). Its activation leads to eye and forebrain hypoplasia in zebrafish (Kobayashi et al., 2001). Chromobox 4 (CBX4), from the *CBX4 Static module*, is an E3 Small ubiquitin-like modifier (SUMO)-protein ligase that mediates SUMO modification of BMI1, which is required for the accumulation of BMI1 at sites of DNA damage (Ismail et al., 2012). BMI1 deficiency in mice results in growth retardation and neurodegeneration. BMI1 expression was shown to be silenced in AD brains, but not in early-onset fAD, frontotemporal dementia, or Lewy body disease (Flamier et al., 2018). **Fatty acid metabolism.** Linoleic acid is an essential unsaturated fatty acid metabolised by many cytochrome P450s (CYPs). Lower levels of linoleic acid have been found in the middle frontal gyrus in samples from individuals with significant AD neuropathology, but no cognitive decline, and even lower in AD patients compared to controls (Snowden et al., 2017). A study of Saudi elderly women also showed lower levels of linoleic acid in AD patient blood (Alsumari et al., 2019).

The top 10 down-regulated pathways were mostly involved in translation (*Reactome Peptide chain elongation*, *KEGG Ribosome*, *Reactome 3'UTR mediated translational regulation*, *Reactome SRP-dependent cotranslational protein targeting to membrane*, *Reactome Translation*, *Reactome Influenza viral RNA transcription and replication*), glycolysis (*Biocarta Glycolysis pathway*), cell proliferation (*Biocarta SRCRPTP pathway*, *Biocarta CBL pathway*), and Wnt signalling (*TCF3 Static module*) (Table 7.2).

Translation. Decreased levels of several genes encoding ribosomal proteins and reduced protein levels of elongation factors were characterised in advanced stages of AD. Changes are even more marked in rapid course AD (Garcia-Esparcia et al., 2017). **Cell proliferation.** Both Src from *Biocarta SRCRPTP pathway* and its target Cbl (*Biocarta CBL pathways*) are involved in cell proliferation. Both pathways include PRKCA and PRKCB, encoding Protein kinase Cα (PKCα) and Cβ (PKCβ), respectively. PRKCB levels were found to be dysregulated in AD brains (Gerschütz et al., 2014), and three highly penetrant variants were defined in the PRKCA gene in families with late-onset AD (Alfonso et al., 2016). The authors suggest that enhanced PKCα activity may contribute to AD, possibly by mediating the actions of Aβ on synapses. **Wnt signalling.** Aβ is thought to trigger the Wnt signalling pathway dysregulation, which could contribute to synaptic dysfunction and degradation (Palomer et al., 2019). T-cell factor 3 (Tcf3), from *TCF3 Static module*, and several other components of the Wnt pathway signalling were dysregulated in AD brains (Riise et al., 2015). Nuclear Tcf7l1/Tcf3 gene expression was found to be associated with disease stage, neurofibrillary tangles (NFTs) and cognition in the hippocampus of AD samples (Blalock et al., 2004; Gómez Ravetti et al., 2010).

In summary, the top dysregulated pathways identified in the 3D cell culture model reflected well-established AD-processes. The majority of the top 10 up-regulated pathways were related to inflammatory response, and the down-regulated pathways were mostly related to translation.

7.2.4 Evaluating drug prioritisation

The top drug candidates

We assessed the top 10 prioritised drugs for each of the 8 pathway clusters using the same approach as in the JIA representative study (Section 6.4.1). The pathway clusters used for drug prioritisation represented the top 5, 10, 15, and 20 up- and down-regulated pathways. For clarity, we referred to them with an abbreviated name: *up5* or *up10* for the top 5 or 10 up-regulated pathways, respectively. *up5:20* notation includes all 4 up-regulated pathway clusters. We performed a literature search on drugs appearing in the top 10 in more than one cluster. We prioritised drugs that were prioritised by opposite direction clusters, drugs that appeared in the highest number of clusters and drugs that consistently scored at the top. Additionally, we looked for association with AD for two drugs that have ranked in the top 10 in our representative 3D cell culture model and also in top 10 drug lists from Mayo disease signatures. The top 10 drugs in *up10*, *up20*, *down10* and *down20* are listed in Supplementary Table E.2.

Across the eight A5 pathway clusters, 46 unique BRD IDs were identified. 27 of those appear in one cluster, from that 21 were in an up- and six in a down-regulated cluster. **MLS003329221** (BRD-K81814927) has been prioritised in two clusters with opposite directions, in *up15* and *down15*. However, no related biological information was found for this drug.

Six drugs have been prioritised in the top 10 of the A5 down-regulated clusters (*down5:20*): H5902 (BRD-K15402119), diethylstilbestrol (BRD-K45330754), SPECTRUM_000090 (BRD-A80151636), SCHEMBL2560033 (BRD-K17739445), SPB02303 (BRD-K99532291), geldanamycin (BRD-A19500257). **H5902**, also known as huperzine A, ranked first in three, and second in one of the *down5:20*. H5902 acts as a cholinesterase inhibitor, studies suggest it improves memory and protects nerve cells, which could slow the cognitive decline associated with AD (Huang et al., 2014; Qian and Ke, 2014; Wang et al., 2011). It is a natural AChE inhibitor derived from *Huperzia serrata* used in Chinese folk medicine. It is a licensed AD treatment in China (Zangara, 2003). In addition to the symptomatic, cognitive-enhancing effect via the AChE inhibition (Rafii et al., 2011; Yang et al., 2013b), studies have shown the potential for it to serve as a disease-modifying agent for AD (reviewed in Qian and Ke (2014)). It was shown that huperzine A could suppress Aβ accumulation and hyperphosphorylated tau formation when administered at an early stage of AD in AD animal models (Huang et al., 2014; Wang et al., 2011).

Diethylstilbestrol ranked second in 3 and first in 1 of the down-regulated clusters. SPECTRUM_000090 (BRD-A80151636) ranked third in *down5:20*. Geldanamycin (BRD-A19500257) ranked between 6–9th across the *down5:20* clusters. Diethylstilbestrol, SPECTRUM_000090 and geldanamycin have all also appeared in the top drug candidates for JIA (Section 6.4.1), the similarities in the top drugs could be due to artefacts in prioritisation scoring or due to the inflammation-related pathways in both JIA and AD down-regulated clusters. There is no evidence of **diethylstilbestrol**, an oestrogen agonist, having an effect on AD. However, some evidence from animal studies, and from both observational studies and clinical trials suggest that oestrogen is neuroprotective. For example, long-term self-reported postmenopausal hormone therapy was associated with reduced AD risk (Imtiaz et al., 2017), but recent clinical trials and observation studies have not found an association between hormone therapy and AD (Gleason et al., 2015). Additionally, the Women's Health Initiative Study has shown that oestrogen treatment in late post-menopause increases risk (Shumaker et al., 2004). Rocca et al. (2011) have proposed that the conflicting evidence could be explained by the window of opportunity hypothesis that indicates that the results are highly dependent on the cohort age assessed in each trial. As in JIA, no biological information was associated with **SPECTRUM_000090**, but its isomer, bromocriptine, was identified in a drug screen for its ability to lower Aβ in a dose-dependent manner. Together with cromolyn and topiramate, it has been identified as part of the most effective combination that reduced the Aβ content in familial and sporadic AD patient neurons (Kondo et al., 2017). **Geldanamycin**, a heat shock protein 90 (Hsp90) inhibitor, has been shown to potently and preferentially reduce phospho-tau levels (Petrucelli et al., 2004), although with a high degree of toxicity (Ansar et al., 2007). Hsp90 is a chaperone protein that regulates tau metabolism and Aβ processing. Several geldanamycin derivatives have been developed to reduce toxicity and improve potency but have failed in clinical trials due to low solubility and toxicity (Blair et al., 2014). No information could be found for **SCHEMBL2560033** (BRD-K17739445) and **SPB02303** (BRD-K99532291). However, SCHEMBL2560033 was predicted to be a dopamine receptor antagonist (probability = 0.72) and SPB02303 a histone deacetylase (HDAC) inhibitor (probability = 0.76) by L1000 fireworks display (L1000FWD) (Wang et al., 2018). The mechanism of action (MOA) was predicted by clustering drug signatures and assigning MOAs of known, well-characterised drugs to less known preclinical drugs. HDAC inhibitors were one of their well-defined clusters, hence, the high probability value. Other drugs without any associated biological information mentioned in this section did not yield any high-probability MOAs. HDAC3 inhibition has been shown to reverse AD-related pathology *in vitro* and in an AD mouse model (Janczura et al., 2018). HDAC2 inhibitor, vorinostat is currently in clinical trials (ClinicalTrials.gov, 2019). Another HDAC inhibitor sodium butyrate (NaB) has been demonstrated to improve memory performance and rescue neurodegeneration in several AD mouse models (Cao et al., 2018). More HDCA inhibitors and their effects are reviewed in De Simone and Milelli (2019). Post-mortem studies showed loss of dopamine D2 receptors in the temporal lobes in AD and a decrease of D2 receptor availability was shown in AD patient hippocampus, suggesting that dopamine receptors antagonists are unlikely to improve AD, but more likely to worsen the condition (Kemppainen et al., 2003).

TG101348 (BRD-K12502280), also known as fedratinib, is a semi-selective Janus kinase 2 (JAK2) and FMS-like Receptor Tyrosine Kinase 3 (FLT3) inhibitor. It ranked in the top 10 in two signatures in each *up5* and *up10.* It has been shown that Aβ-dependent inactivation of the JAK2/signal transducer and activator of transcription 3 (STAT3) axis in hippocampal neurons causes cholinergic dysfunction, which leads to memory impairment related to AD. In addition, activation of JAK2/STAT3 axis with a humanin derivative restored cognitive function in an AD model (Chiba et al., 2009). It is thus likely that fedratinib would worsen AD due to inactivation of JAK2. However, evidence suggests that inhibition of FLT3 might be beneficial in AD by reducing neuronal oxidative stress. FLT3 inhibitors blocked ferroptotic cell death in neurons by preventing lipid peroxidation that triggers glutamate toxicity (Kang et al., 2014).

Amlodipine base (BRD-A64297288), prioritised by the *down5:10*, and *down20* pathway clusters, is a calcium channel blocker with antihypertensive and antianginal properties. While amlodipine did not show cognitive improvement, likely due to poor brain availability, another related drug, nilvadipine, stabilised cognition in patients with mild cognitive impairment (Hanyu et al., 2007). Although both reduced hypertension (a risk factor in

AD), nilvadipine was shown to increase cerebral blood flow in the hippocampus, a feature that is reduced early in the development of AD (de Jong et al., 2019).

BI 2536 (BRD-K64890080) was the top scoring drug in the *up5:10* clusters. It is a polo-like kinase 1 (PLK1) inhibitor that reduces β-amyloid-induced neuronal cell death in an AD cell model (Song et al., 2011). PLK1 is present in hippocampal and cortical neurons of AD patients, but not controls (Harris et al., 2000). PLK1 is a cell cycle regulator and its expression in AD patients suggests neuronal cell-cycle re-entry triggered by Aβ (Peng et al., 2015; Song et al., 2011). **S-8599** (BRD-K49810818), also known as sorafenib was the top-ranking drug in *up20*, and second in *up15*. It is a multikinase inhibitor of Raf1, BRaf, vascular endothelial growth factor receptor 2 (VEGFR2) and other kinases, including FLT3. Active form of Raf-1, cRaf-1, is up-regulated in post-mortem AD brains and AD mouse models. Sorafenib inhibits cRaf-1 and nuclear factor κ-light-chain-enhancer of activated B cells (NF-κB), it decreases APP, inducible nitric oxide synthase (iNOS) and cyclooxygenase-2 (COX-2) expression and restores working memory in AD mice (Echeverria et al., 2009).

Two drugs, JNK-9L and AG14361, from the top 10 in the A5 clusters were also found in the top 10 in the Mayo clusters (Supplementary Table E.3). JNK-9L ranked in the top 10 in Mayo *up20* and A5 *up15*. AG14361 ranked in A5 *up15* and Mayo *down5:10*. **JNK-9L** (BRD-K19220233) is a c-Jun N-terminal kinase (JNK) inhibitor. Aβ activates JNK and caspase-8 leading to neuronal apoptosis (Wei et al., 2002). In addition, JNK3 enhances Aβ production and plays a role in maturation and development of neurofibrillary tangles. Increased levels of phosphorylated JNK have been shown in human post-mortem brains of AD patients that co-localised with Aβ (Killick et al., 2014; Zhu et al., 2001). It has also been correlated with the rate of cognitive decline (Gourmaud et al., 2015). JNK3 is the major kinase for APP phosphorylation and a depletion of JNK3 in AD mice resulted in a reduction of Aβ42 peptide levels, overall plaque load and an increase of number of neurons and improved cognition (Yoon et al., 2012). It has been shown that JNK3-mediated phosphorylation regulated APP cleavage by inducing the amyloidogenic processing of the protein, while JNK inhibition reduced this processing and increased the non-amyloidogenic route *in vitro* by blocking APP phosphorylation (Colombo et al., 2009; Morishima et al., 2001; Savage et al., 2002). **AG14361** (BRD-K00615600) is a poly (ADP-Ribose) polymerase (PARP) inhibitor. PARP1 is responsible for the maintenance of genome stability, transcriptional regulation, and long-term potentiation in neurons. However, the extensive activation of it under pathological conditions may lead to cell death (Strosznajder et al., 2012). Enhanced PARP activity has been demonstrated in the AD human brain (Love et al., 1999). Aβ and inflammation can lead to activation of

PARP1 and cell death (Abeti et al., 2011; Adamczyk et al., 2005; Strosznajder et al., 2000). A PARP1 gene polymorphism has also been associated with the risk of AD (Liu et al., 2010). PARP1 inhibition and deficiency have been shown to prevent Aβ-triggered toxicity, increase release of neurotrophic factors such as Transforming growth factor β (TGFβ) and vascular endothelial growth factor (VEGF), and preserve the ability of microglia to phagocytose Aβ peptides (Kauppinen et al., 2011).

In summary, even though the AD disease signature was developed on a 3D cell model, the top prioritised drugs showed a remarkably close relevance to human AD physiology and symptoms, such as targeting cognitive impairment. Out of 11 evaluated drugs most showed direct or indirect evidence of improvement in cell and animal models. Several related drugs have been in clinical trials. The top-ranking drug in down-regulated clusters, huperzine A, is approved for treating AD in China and there is supporting evidence for not only cognitive improvement in AD patients, but also disease-modifying beneficial effects (Huang et al. (2014); Wang et al. (2011), reviewed in Qian and Ke (2014)). SPB02303's predicted mechanism of action would probably cause deteriorating effects. The two drugs that have also been identified in the top prioritised drugs for the Mayo dataset show high relevance to AD pathology and present promising translational drug candidates. The current cell model reflects advanced AD pathology, and several top prioritised drugs target those processes. However, if pre-symptomatic AD processes are identified, our system could potentially prioritise drugs for those dysregulated processes providing early therapeutic intervention.

True positive lists for AD

Approved drugs for AD (Supplementary Table E.4). Ideally, this list would be the gold standard for testing the pipeline's performance. However, there are not many available treatments for AD. From 29 brand and international non-proprietary names (INN), only 11 were mapped to 5 Broad (BRD) IDs and only 2 of those BRD IDs were present in PDxN. The 2 BRD IDs (donepezil and memantine) were present in 5 different drug signatures. Due to the limited number of AD approved drugs in PDxN we used a list of TPs from an *in vitro* drug screen.

AD *in vitro* drug screen. A confidential list was shared with us by our collaborator Tanzi and Kim group. Due to its confidential nature, the list cannot be shared in this thesis. They identified drugs that have shown reduction in Aβ or reduction in Aβ and tau as part of the 3D drugs screen (3DDS). The 3DDS was a high-throughput drug screening for AD using 3D human neural culture systems related to A5 3D cell culture. Approximately

1200 of the Food and Drug Administration (FDA) and other biologically active drugs were screened. They have identified 38 drugs that had a beneficial effect in the 3D cell line model. Only 18 of those were present in PDxN. We mapped them to 31 BRD IDs and 30 of those BRD IDs were present in PDxN. The 30 BRD IDs were present in 300 different drug signatures. This was the primary TP list used in benchmarking tests in this section.

Benchmarking

We showed in the previous chapter, Chapter 6, that the drug signature features had an effect on the drug signature rank (Section 6.4.4). We calculated the mean scaled rank of each drug signature feature from the *in vitro* drug screen TP list for each of the cell lines, leaving out A5 for cross-validation (Figure 7.1). Again, we showed that some signature features were consistently associated with low ranks and some with high. The I47F (low Aβ42/40 ratio) lists showed differential ranking to those of H10 and I45F (low Aβ42/40 ratio). In particular, the drug lists from H10 and I45F produced similar rankings of drug signature features depending on the direction of the input pathway clusters.

Two batches (CPC005, CPC015), two drugs, and two cell lines (MDST8, SW620) consistently ranked in the bottom 40%. There were no concentration values or perturbation times that on average rank in the last 40% of the lists. Both cell lines were developed from colon carcinoma (Supplementary Table D.9). MDST8 was also consistently ranked in the bottom 40% across JIA prioritised drug lists.

We assessed the effect of the worst performing features by removing any drug signature that included features that were on average ranked in the bottom 40% of the list. We generated receiver operating characteristic (ROC) curves for A5 prioritised drug lists by scoring the specificity and sensitivity with 3DDS drugs (Fig. 7.2). The area under the ROC curve (AUC) of 1 indicates perfect, AUC $= 0.5$ indicates random, and AUC $<$ 0.5 worse than random performance. The performance before and after removal of poor scoring features was comparable and indicated random performance (AUC ≈ 0.5). The best performing cluster in A5 that yielded the highest AUC score was the *up20* (AUC $= 0.55$). The down-regulated clusters yielded AUC scores in the range of 0.43–0.47. If we took the performance of < 0.5 AUC as a guide, then these results suggested that the developed drug repositioning pipeline was not able to prioritise drugs that ameliorate AD pathology in a 3D model of AD. However, the literature search of the top 10 drugs across each prioritised list suggested otherwise. The poor AUC scores could be explained by the length of the prioritised lists. Each of the prioritised drug lists in A5 has between

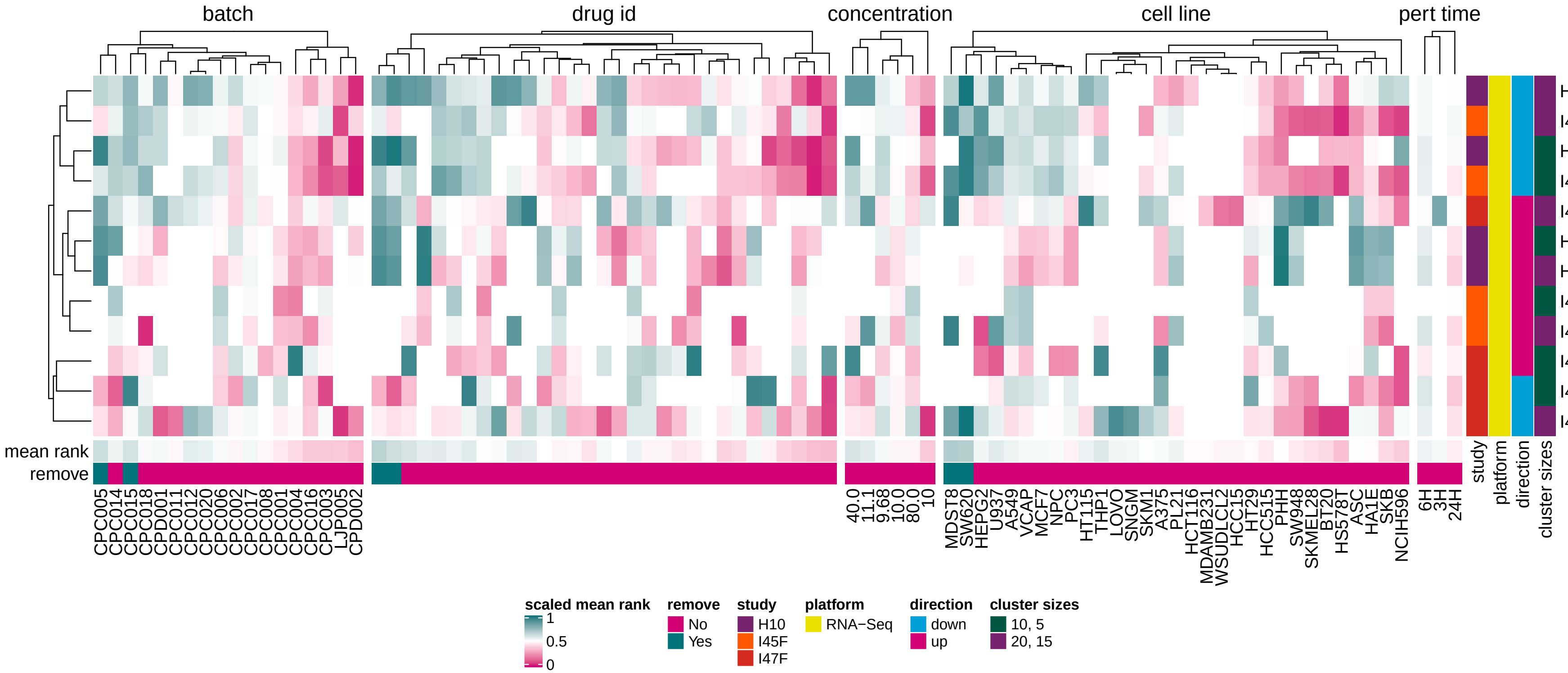


Fig. (7.1) Mean true positive (TP) drug rank in prioritised drug lists per drug signature feature. Drug lists prioritised by up- and down-regulated pathway signatures with 5, 10, 15 or 20 pathways, were scored with the positive hits from *in vitro* drug screen on A5-related cell culture. The pathway signatures were defined from the Alzheimer's disease (AD) 3D cell culture models. Mean rank of 1 (top prioritised drug) was scaled to 0 and the worst rank to 1. Each row represents a combined prioritised list between 5 and 10 or 15 and 20 pathways up- or down-regulated pathways identified in one of the AD 3D cell model studies. Each column represents a drug signature feature (batch, drug id, concentration, cell line or perturbation time). Scaled rank of ∼0 (pink) indicated high ranking features in that prioritised drug list, while rank of 1 (teal) indicates low ranking drug signature features. The bottom margin (mean rank) indicates the average feature performance across all prioritised drug lists from all listed studies and pathway clusters. Drug signature features were selected for removal if the true positive drugs with that feature ranked in the bottom 40% on average across three different 3D cell culture dataset signatures, leaving out the A5 dataset. Only the TP signature features are used in this assessment. batch — experimental batch; pert time — perturbation time.

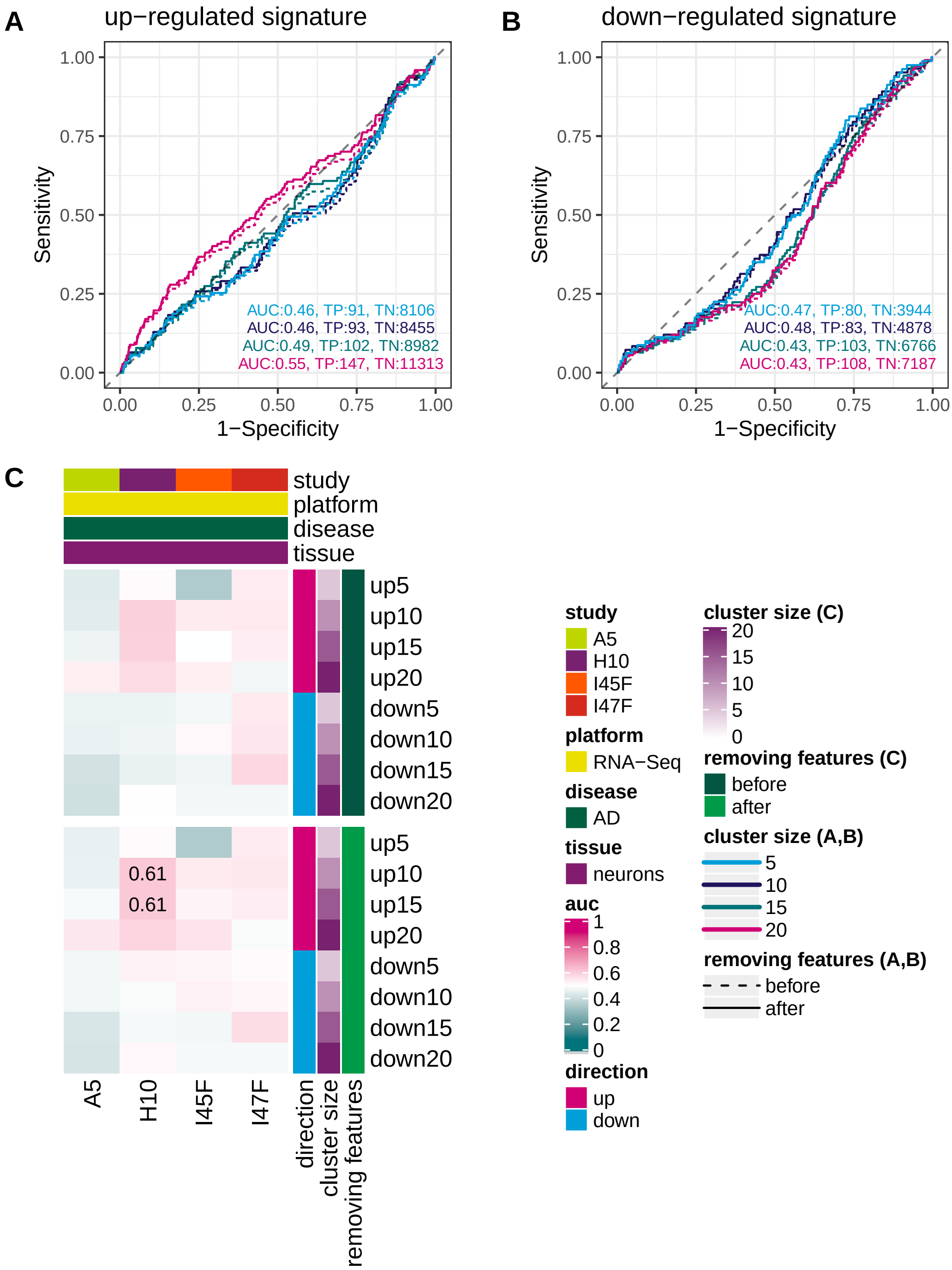


Fig. (7.2) AUC score summary for Alzheimer's disease (AD) studies before and after removing selected drug signature features. (A–B) ROC curves for the PDxN-derived drug lists, prioritised by the disease signatures defined from the A5 3D cell culture model. Lists were scored against positive hits from an *in vitro* drug screen on an A5-related cell culture. Performance is indicated for (A) up- and (B) down-regulated disease pathway clusters of different sizes before (dotted line) and after (full line) the removal of low-ranking drug signature features (identified in Fig. 7.1). The AUC, TP and TN counts are displayed for the *after* ROC curves. Dashed diagonal line signifies random performance (AUC ≤ 0.5). (C) Summary of AUCs per AD study, signature direction and pathway cluster size before (C top) and after (C bottom) removing selected drug signature features. AUC — area under the ROC curve; PDxN — Pathway Drug Coexpression Network; ROC — receiver operating characteristics; TN — true negative; TP — true positive.

3944–11313 true negative (TN) signatures and 80–147 TP signatures. Translationally, only the top 10–100 would be of realistic interest for further validation. Thus, when assessing our method, the steepness of the initial ROC curve slope would be of higher importance, because it represents the performance of the top scoring drugs. The performance of the top 1000 drugs in A5 clusters is represented at *x*-axis values of *1-Specificity* < 0.25 for the shortest list (Fig. 7.2B cluster size 5 - light blue) and < 0.088 for the longest list (Fig. 7.2A cluster size 20 - pink). The slope on all, but in particular the A5 *up20* (Fig. 7.2A cluster size 20 - pink), is much steeper at lower *1-Specificity* values. For example, if we considered the top 1000 drugs for *up20* the slope is steeper than the diagonal indicating random performance ($m_{1000} = 1.85$) and the slope for only the top 100 is $m_{100} = 3.70$, suggesting high sensitivity at the topmost prioritised drugs. Even the worst performing cluster, cluster *down20* ((Fig. 7.2B cluster size 20 - pink, AUC = 0.43), has a steep initial slope $\mathrm{m}_{100} = 3.83$ that then lowers to $m_{1000} = 0.79$.

The literature search and the benchmarking results suggested that our method can effectively recall the 3DDS drugs that have shown an improvement in AD-related pathology in an AD 3D cell culture model. Assessing the performance of all drug signatures assigned a prioritisation score suggested random performance. In this analysis, the benchmarking results were limited by the lack of more appropriate true positive drugs. However, literature searched top drugs and the initial steepness of the ROC curve indicated promising results. Further characterisation of the benchmarking and alternative benchmarking designs are necessary for more conclusive results.

7.3 Parkinson's Disease

7.3.1 Disease introduction

Parkinson's disease (PD) is a progressive, multifactorial, neurodegenerative disease. It is the second most common neurodegenerative disorder, after AD, affecting approximately 1% of people over the age of 60 (Tysnes and Storstein, 2017). As in AD, PD-related pathology is thought to begin many years before the clinical manifestation.

Pathologically, it is characterised by aggregations of α-synuclein (α-Syn) into mostly insoluble Lewy bodies (Braak et al., 2003) and progressive neurodegeneration of dopaminergic neurons in the substantia nigra pars compacta. Severe motor symptoms, such as resting tremor, rigidity, bradykinesia, gait, and balance dysfunction result in severely reduced mobility and strongly impair patients' quality of life (Jankovic, 2008; Thomas and Beal, 2007). Non-motor symptoms have been of increasing interest (Chaudhuri et al., 2006). The cellular mechanisms underlying dopaminergic cell death in PD are still not fully understood, but mitochondrial dysfunction, oxidative stress and inflammation are strongly implicated in the pathogenesis of both familial and sporadic PD cases. Aberrant post-translational modifications and age-dependent insufficient quality control systems lead to cellular overload of dysfunctional proteins (Dauer and Przedborski, 2003).

Approximately 90% of PD cases have a sporadic origin, which may be caused by environmental factors together with genetic susceptibility. The remaining 10% represent familial forms of the disease often associated with an earlier onset (Thomas and Beal, 2007). The main risk factor for developing PD is ageing (Tysnes and Storstein, 2017). The α-synuclein gene (SNCA) is a major risk factor linked to sporadic PD (Simón-Sánchez et al., 2009).

Current treatments offer symptomatic relief, but none stop or decrease the dopaminergic neuron degeneration (Oertel, 2017). The discovery of levodopa, a dopamine precursor, revolutionised the treatment of PD. However, after several years of treatment most patients develop dyskinesias, which are difficult to treat and can develop into a significant source of disability (Poewe et al., 2010). Current research is directed toward modulation and clearance of α-Syn aggregates (Oertel, 2017).

7.3.2 Parkinson's disease datasets

In the PD case study, we used data from 2 different sporadic PD (sPD) subgroups and one from a PD zebrafish model (Table 7.1).

Lysosomal and mitochondrial dysfunction data

We used data from 2 distinct sPD groups collected and characterised in one dataset (Carling et al., 2020). Lysosomal and mitochondrial dysfunction were characterised in skin fibroblasts from sPD patient cohort. Patients with lysosomal counts greater than 3 standard deviations (SD) from the mean of the age-matched controls were characterised as lysosomal dysfunction group and patients with ATP activity lower than 3 SD from the control mean were grouped into mitochondrial dysfunction group.

GHC1 zebrafish

The zebrafish dataset consisted of GCH1 mutant and wild type (WT) control samples. GCH1 encodes the enzyme GTP cyclohydrolase 1, which is essential for dopamine synthesis in nigrostriatal cells (Kurian et al., 2011). Rare GCH1 variants have been associated with increased risk of PD (Mencacci et al., 2014).

The gch1 mutant zebrafish line was generated with CRISPR/Cas9 targeting exon 1 inducing a 94 base pair deletion. It resulted in a frameshift mutation, with a predicted nonsense protein product, removing the GTP cyclohydrolase domain. The homozygous (HOM) mutant showed 21.3% reduction in gch1 mRNA levels compared to WT (Larbalestier et al., 2020).

The dataset includes RNA-Seq data from WT and $gch1^{-/-}$ zebrafish larval brain tissue from pooled brain samples from 4 biological replicates per genotype ($gch1^{-/-}$ and WT) at 8 days post fertilisation (Larbalestier et al., 2020).

The representative study

Due to the low sample number in both the lysosomal and mitochondrial dysfunction group, the lysosomal dysfunction group was chosen for further analysis because it yielded more differentially expressed genes than the mitochondrial dysfunction analysis in Carling et al. (2020), suggesting more measurable differences between case and control subjects.

The zebrafish dataset was used to identify promising top drug candidates that scored highly in both the human lysosomal deficiency and the zebrafish model, highlighting the drugs with increased model-to-human translation potential.

7.3.3 Disease pathway signatures

We performed a literature search to establish whether the top dysregulated pathways in the lysosomal dysfunction dataset were linked to known PD processes. We searched the top 10 most up- or down-regulated pathways. The top 20 up- and down-regulated pathways are listed in Table 7.3.

Table (7.3) Parkinson's disease (PD) lysosomal dysfunction disease signature pathways. The top 20 most up- (rank: 1 to 20) and down- (rank: -1 to -20) regulated pathways (q-value $<$ 0.05). Drugs were prioritised for up- and down-regulated pathway clusters at: the top 5, 10, 15 or 20 pathways by decreasing log fold change (LogFC) for up- and decreasing for down-regulated pathways. The genes in pathway column represents the number of possible genes in that pathway, while genes in data is the number of pathway genes found in data.

Rank	**Pathway**	**LogFC**	***q*-value**	**Genes in pathway**	**Genes in data**
1	Reactome N GLYCAN ANTENNAE ELONGATION	1.41	0.00226	14	10
2	Biocarta UCALPAIN pathway	1.40	0.00234	18	14
3	Biocarta ECM pathway	1.38	0.00226	24	22
4	Biocarta MCALPAIN pathway	1.37	0.00226	25	20
5	SPI1 10 Static Module	1.36	0.00226	10	8
6	Biocarta CBL pathway	1.34	0.00918	13	12
7	Reactome N GLYCAN TRIMMING in the ER and CALNEXIN CALRETICULIN CYCLE	1.34	0.00236	13	13
8	Reactome CALNEXIN CALRETICULIN CYCLE	1.32	0.00267	11	11
9	Reactome REGULATION of INSULIN SECRETION by ACETYLCHOLINE	1.31	0.02520	11	6
10	Reactome PLATELET CALCIUM HOMEOSTASIS	1.29	0.00617	18	9
11	Reactome COPI MEDIATED TRANSPORT	1.29	0.00844	10	10
12	Reactome FATTY ACYL COA BIOSYNTHESIS	1.28	0.00603	18	15
13	Biocarta CFTR pathway	1.28	0.00442	12	10
14	Biocarta KREB pathway	1.27	0.00849	8	8
15	Biocarta CCR3 pathway	1.26	0.00691	23	16
16	Biocarta GLYCOLYSIS pathway	1.26	0.04610	10	8
17	KEGG STEROID BIOSYNTHESIS	1.25	0.00907	17	16
18	Reactome SYNTHESIS of VERY LONG CHAIN FATTY ACYL COAS	1.24	0.00617	14	12
19	Biocarta EPHA4 pathway	1.24	0.00267	10	7
20	Reactome TRANSFERRIN ENDOCYTOSIS and RECYCLING	1.24	0.00472	25	22
-1	Reactome PEPTIDE CHAIN ELONGATION	-1.55	0.00472	153	83
-2	CBX4 10 Static Module	-1.51	0.00234	10	10
-3	Biocarta IL7 pathway	-1.50	0.00234	17	14
-4	SIX3 11 Static Module	-1.49	0.00234	11	8
-5	KEGG RIBOSOME	-1.48	0.00617	88	85
-6	PAX6 19 Static Module	-1.47	0.00579	19	6
-7	Reactome 3 UTR MEDIATED TRANSLATIONAL REGULATION	-1.46	0.00691	176	103
-8	Reactome INFLUENZA VIRAL RNA TRANSCRIPTION and REPLICATION	-1.45	0.00791	169	99
-9	Reactome ETHANOL OXIDATION	-1.45	0.00271	10	6
-10	Biocarta TH1TH2 pathway	-1.45	0.00226	19	7
-11	KEGG CIRCADIAN RHYTHM MAMMAL	-1.43	0.00226	13	13
-12	RARA 17 Static Module	-1.41	0.00226	17	11
-13	SMAD4 27 Static Module	-1.41	0.00234	26	18
-14	Reactome BMAL1 CLOCK NPAS2 ACTIVATES CIRCADIAN EXPRESSION	-1.40	0.00226	36	34
-15	Reactome NONSENSE MEDIATED DECAY ENHANCED by the EXON JUNCTION COMPLEX	-1.40	0.00894	176	104
-16	Reactome FORMATION of the TERNARY COMPLEX and SUBSEQUENTLY the 43S COMPLEX	-1.40	0.00840	74	48
-17	STX5 12 Static Module	-1.38	0.00226	12	11
-18	Biocarta CLASSIC pathway	-1.38	0.00791	14	7
-19	ST IL 13 pathway	-1.38	0.00601	7	6
-20	Biocarta NKT pathway	-1.38	0.00226	29	9

The top 10 up-regulated pathways in the lysosomal dysfunction dataset were related to endoplasmic reticulum (ER) stress and unfolded protein response (*Reactome N-glycan antennae elongation*, *Reactome N-glycan trimming in the ER and calnexin/calreticulin cycle*, *Reactome Calnexin/calreticulin cycle*), α-Syn cleavage (*Biocarta UCALPAIN pathway*, *Biocarta MCALPAIN pathway*), and other miscellaneous processes (*Biocarta ECM pathway*, *SPI1 Static module*, *Biocarta CBL pathway*, *Reactome Regulation of insulin secretion by acetylcholine*, *Reactome Platelet calcium homeostasis*) (Table 7.3).

ER stress and unfolded protein response. Calnexin and calreticulin are ER chaperone proteins that recognise unfolded proteins that have been glycosylated (Williams, 2006). They promote proper folding and protect glycoproteins from aggregation or premature export from the ER. They detect unfolded and terminally misfolded proteins and trigger the ER-associated degradation via the ubiquitin proteasome system (UPS) (Ellgaard and Helenius, 2003). Calnexin has been shown to associate with the α-Syn protein complex in the cytoplasm. Internalisation of aggregated α-Syn has been related to calnexin (Liu et al., 2007). Overexpression of wild-type or mutant α-Syn caused accumulation in the ER and activation of the unfolded protein response in yeast PD models (Cooper et al., 2006). Activation of the unfolded protein response has been characterised in post-mortem PD brain (Hoozemans et al., 2007). **α-Syn cleavage.** Calpains are calcium-activated non-lysosomal intracellular cysteine proteases. Increased M-calpain expression has been detected in midbrain of PD patients (Mouatt-Prigent et al., 1996). α-Syn is a substrate for calpain cleavage (Mishizen-Eberz et al., 2003). Calpain-cleaved α-Syn fragments generate a high-molecular weight species and convert α-Syn from a random coil into a β-sheet structure, which enhances the ability of α-Syn to aggregate. Calpain cleaves α-Syn in human brains of PD and dementia with Lewy bodies patients. The cleaved α-Syn fragments co-localise with activated calpain (Dufty et al., 2007). Inhibition of calpain reduced α-Syn pathology and improved activity performance in α-Syn mice (Hassen et al., 2018). **SPI1**, encoding the PU.1 transcription factor, is expressed in microglia (Walton et al., 2000) and is vital for microglial survival (Smith et al., 2013). Microarray analysis in mixed glial cultures identified several genes altered by PU.1 silencing that are also risk variants in PD (Gosselin et al., 2017; Rustenhoven et al., 2018). **CBL** pathway includes CBL-mediated ligand-induced downregulation of epidermal growth factor receptors (EGFR) through degradation (de Melker et al., 2004). Similarly to CBL, a Parkin KO accelerates EGFR endocytosis and degradation (Fallon et al., 2006). EGFR is degraded by both the proteasome and lysosome (Levkowitz et al., 1999). **Calcium homeostasis** is disrupted by α-Syn aggregation. Its significance in PD is extensively reviewed in Zaichick et al. (2017). **Insulin secretion by acetylcholine** is driven by two Ca^{2+} dependent mechanisms (Gilon and Henquin, 2001). Acetylcholine levels are abnormally low in PD (Rizzi and Tan, 2017).

Extracellular matrix (**ECM**) is disrupted by α-synuclein moving to the extracellular space in PD and related neurodegenerative diseases (Lee et al., 2014a).

The top 10 down-regulated pathways were involved in translation (*Reactome Peptide chain elongation*, *KEGG Ribosome*, *Reactome 3'UTR mediated translational regulation*, *Reactome Influenza viral RNA transcription and replication*), brain development (*CBX4 Static module*, *SIX3 Static module*, *PAX6 Static module*), immune response (*Biocarta IL7 pathway*, *Biocarta TH1TH2 pathway*), and ethanol metabolism (*Reactome Ethanol oxidation*) (Table 7.3).

Translation. Translation is highly dysregulated in PD. Several of the familial PD mutations are linked to deregulation of mRNA translation, suggesting its importance in PD onset (Correddu and Leung, 2019). A US patent for measuring decreased translation in PD demonstrated that sPD patients showed reduced levels of translation in skin fibroblasts (Flinkman et al., 2019). **Brain development.** CBX4 is one of the Polycomb-group proteins that bind and repress genes in embryonic stem cells through lineage commitment to the terminal differentiated state. It has been identified as downregulated in PD blood (Tan et al., 2018). Dopamine signalling has been shown to lead to a loss of Polycomb repression and aberrant gene activation in parkinsonian mice (Södersten et al., 2014). SIX3 is a Wnt1 suppressor (Lagutin et al., 2003), its down-regulation could thus dysregulate the Wnt signalling pathway, which has been linked to neurodegeneration (Berwick and Harvey, 2012). Parkin knockout (KO) in mice resulted in excessive Wnt signalling and was associated with primary dopaminergic neuron death (Rawal et al., 2009). PAX6 is expressed in selective populations of dopaminergic neurons. Post-mortem tissue from PD patients showed decreased levels of substantia nigra PAX6^{+} cells. Over-expression of PAX6 in cells treated with PD neurotoxins resulted in increased cell survival and reduction of apoptosis and oxidative stress markers (Thomas et al., 2016). **Immune response.** IL-7 is a key cytokine essential for normal development of B and T cells. JAK3-phosphorylated IL-7 activates STAT5, Src kinases, phosphoinositide 3-kinase (PI3K), protein-tyrosine kinase 2 (Pyk2) and B-cell lymphoma 2 (Bcl2) proteins (Foxwell et al., 1995). Induction of these correlated with the growth-promoting effects of IL-7. T helper 1 and 2 cell (Th1/Th2) differentiation creates helper T cells that produce and respond to two different sets of cytokines. The Th1 cell cytokines stimulate the phagocytosis and destruction of microbial pathogens while Th2 cytokines like IL-4 generally stimulate the production of antibodies directed toward large extracellular parasites. PD patients have reduced Th2 with naive T cells differentiating towards Th1 (Kustrimovic et al., 2018). **Ethanol metabolism.** Aldehyde dehydrogenases (ALDH) play an important role in ethanol and dopamine metabolism (Yu et al., 2016). Growing evidence suggests that the aldehyde metabolites from dopamine are neurotoxic, and their intraneuronal accumulation has been associated with neuronal cell death leading to neurodegeneration (Li et al., 2001). ALDH2 polymorphism that reduces the enzyme activity has been associated with decreased cognitive function in PD patients (Yu et al., 2016).

In summary, the top dysregulated pathways in the lysosomal dysfunction group were linked to known PD processes. The up-regulated pathways were related to ER stress and unfolded protein response as well as to α-Syn aggregation. As in AD, the down-regulated pathways were mostly linked to translation.

7.3.4 Evaluating drug prioritisation

The top drug candidates

We assessed the top 10 drugs for each of the 8 pathway clusters defined from the sPD lysosomal deficiency disease signature (Supplementary Table E.5). Across the 8 clusters 32 unique drugs were prioritised, with 11 appearing in only one cluster. Seven drugs were prioritised across all lists from the up-regulated signatures (*up5:20*), and 6 were prioritised across drug lists from the down-regulated signatures (*down5:20*).

All 6 drugs prioritised in *down5:20* were also prioritised in *down5:20* for A5 AD cell culture. These similarities were likely due to high overlap in the top 10 down-regulated pathways. **Diethylstilbestrol**, an oestrogen agonist, has not been tested in PD patients or models, however, PD displays greater prevalence and earlier age at onset in men (Elbaz et al., 2002). Treatment of male and female PD mice with a brain-selective oestrogen improved PD symptoms in both. Untreated female mice displayed less severe symptoms at a later age that oestrogen treatment still further improved, while male mice treated with oestrogen displayed reduced α-Syn build-up and improved their motor performance (Rajsombath et al., 2019). **H5902**, or huperzine A, is an AChE inhibitor. It has also been shown to inhibit reactive oxygen species formation, caspase-3 and N-methyl-D-aspartate (NMDA) glutamate receptor activity (Wang and Tang, 2005). Caspase-3 is an important factor in neuronal death. There is an increased amount of active caspase-3-positive dopaminergic neurons in PD patients (Hartmann et al., 2000). 6-hydroxydopamine (6-OHDA), a neurotoxin commonly used to model PD (Simola et al., 2007), also causes activation of caspase-3 (Singh et al., 2010). 6-OHDA-induced dopaminergic neuronal degeneration is attenuated by caspase inhibitors (von Coelln et al., 2001). In addition, NMDA receptor antagonists are effective antiparkinsonian agents and can reduce the complications of chronic dopaminergic therapy in animal models (Nash and Brotchie, 2002; Papa et al., 1995). Dextromethorphan, an NMDA antagonist, reduced dyskinesia in PD patients, but its beneficial effects are limited by adverse effects (Verhagen Metman et al., 1998a,b). Remacemide, another NMDA antagonist has shown no clear effect on parkinsonian symptoms over a 5-week trial (Shoulson et al., 2001). No biological information is associated with SPECTRUM_000090 and SPB02303. However, bromocriptine, a **SPECTRUM_000090** isomer, is a dopamine agonist used in treatment of PD (Lieberman and Goldstein, 1985), and **SPB02303** has been predicted to be an HDAC inhibitor (probability $= 0.76$) by L1000FWD (Wang et al., 2018). Harrison et al. (2018) show that there is a disease-dependant increase in histone acetylation observed in PD humans. Several specific and non-specific HDAC inhibitors have been shown to be neuroprotective in PD models (Chen et al., 2015; Choong et al., 2016; Di Fruscia et al., 2015; Harrison et al., 2016; Jian et al., 2017; Outeiro et al., 2007; Pinho et al., 2016; Suo et al., 2015). **Geldanamycin**, a Hsp90 inhibitor, has been shown to suppress α-Syn neurotoxicity in Drosophila despite the continued presence of Lewy body-like inclusions (Auluck et al., 2005). It also protects against 1-methyl-4-phenyl-1,2,3,6-tetrahydropyridine (MPTP)-induced dopaminergic neurotoxicity in mice (Shen et al., 2005). Due to high toxicity, poor solubility, and poor blood-brain barrier (BBB) permeability geldanamycin derivatives have been explored for translation (Friesen et al., 2017). Higher cumulative use of **amlodipine base**, a calcium channel blocker used in hypertension has been associated with reduced risk of PD even though amlodipine shows low BBB permeability (Lee et al., 2014b; Qiu et al., 2011).

The drugs prioritised by the up-regulated pathway clusters (*up5:20*) were: Hydroquinidine, CGP-60474, HY-11001, 5284616, Menadione, BRD-K60067222 and U18666A in increasing (best to worst) mean rank order. **Hydroquinidine** (BRD-A06390036) scored first in 3 and second in one of the clusters. It is a Class I antiarrhythmic and an inward sodium channel inhibitor. Lower serum sodium was identified with longer duration and higher levodopa dose PD patients with dyskinesia. The serum sodium inversely correlated with the duration of disease (Mao et al., 2017). Excessive membrane depolarisation contributing to PD pathology could be reduced by depressing postsynaptic sodium influx with a sodium channel inhibitor. Lamotrigine and riluzole, sodium channel inhibitors, have shown beneficial effects in PD animal models (Caputi et al., 2003). **CGP-60474** (BRD-K79090631) is a VEGFR-2 and PKC inhibitor that scored with a mean rank of 1.75 in *up5:20*. Increased levels of VEGF and its receptors were characterised in PD (Shim and Madsen, 2018). Other angiogenesis markers are associated with gait difficulties and BBB dysfunction (Janelidze et al., 2015). Deferoxamine-mediated up-regulation of hypoxia-inducible factor 1α (HIF-1α) that also up-regulates VEGFR has shown to prevent dopaminergic neuronal death in PD mice (Guo et al., 2016). Several other studies have demonstrated neuroprotective effects of VEGF (Falk et al., 2009; Poesen et al., 2008; Wada et al., 2006; Yasuhara et al., 2004, 2005). On the other hand, PKC inhibitor rottlerin demonstrated neuroprotective effects in PD cell culture and animal models (Zhang et al., 2007). **HY-11001** (BRD-K64800655), more commonly referred to as PHA-793887, is a pan-cyclin-dependent kinase (CDK) inhibitor. CDK5 has been shown to be a mediator of dopaminergic neuron loss in a PD mouse model, with its inhibition attenuating hypolocomotion (Smith et al., 2003). Elevated levels of CDK5 have been reported in dopamine neurons in human post-mortem PD brains (Nakamura et al., 1997). Increased activity of cyclin D3/CDK6/pRB pathway was shown in lymphoblasts from sPD. This pathway was targeted upstream with two HDAC inhibitors that reduced 6-OHDA-induced cell death (Alquézar et al., 2015). Inhibition of CDK1 provides neuroprotection against ischemic neuronal death (Marlier et al., 2018). **5284616** (BRD-K89626439), better known as sirolimus or rapamycin, is the mammalian target of rapamycin (mTOR) inhibitor. It is an immunosuppressant commonly used in organ transplant patients. Sirolimus has been shown to revert cognitive and affective deficits in PD mouse model (Masini et al., 2018). Eight other studies showing improvement upon sirolimus treatment in animal models are reviewed in Bové et al. (2011). One of them demonstrates that sirolimus protects against dopaminergic neurodegeneration by increasing lysosomal biogenesis, autophagosome-lysosome fusion, and lysosome-mediated clearance of accumulated autophagosomes (Dehay et al., 2010). A combination therapy with sirolimus and trehalose in PD mouse model activated autophagy and reversed both neuronal dopaminergic and behavioural deficits (Pupyshev et al., 2019). Another combination therapy with sirolimus and RTB101 is currently in Phase 1/2 trial for PD (resTORbio, Inc., 2019). **Menadione** (BRD-K78126613), vitamin K3, is a superoxide releasing oxidative stressor. It is a potent activator of mitogen-activated protein kinases (MAPKs) like extracellular signal-regulated kinases 1 and 2 (ERK1, ERK2). Menadione addition and α-Syn expression decrease yeast cell viability (Zampol and Barros, 2018). It has been shown to have anti-fibrillation activity by reducing Aβ42 aggregations and reducing neuronal cytotoxicity in a human neuronal cell line (Alam et al., 2016). Menadione increases neuronal cell death, release of TNF-α, IL-1α, and IL-1β in cultured neurons (Tripathy and Grammas, 2009). Menadione treated neuronal cells over-expressing α-Syn localised α-Syn in the plasma membrane and also developed Lewy body-like inclusions (Wang et al., 2016a). No biological information was associated with **BRD-K60067222**, and only low probability MOAs were predicted by L1000FWD (Wang et al., 2018). **U18666A** (BRD-A81795050) is a cholesterol synthesis and transport inhibitor, it thus increases intracellular cholesterol. It is used to simulate intracellular accumulation of cholesterol leading to neurodegeneration in cells of Niemann-Pick type C patients (Nunes et al., 2013). The lysosomal cholesterol accumulation in neurons resulted in increased lysosomal stability and sensitivity, and reduced oxidative stress-induced apoptosis (Appelqvist et al., 2012). Higher serum cholesterol has been associated with decreased PD risk in men (Rozani et al., 2018). Cathepsin D enzyme activity which reflects lysosomal activity was decreased in lysosomal dysfunction sPD patients (Carling et al., 2020). It was shown that cathepsin D expression and enzyme activity increased upon U18666A-mediated toxicity (Amritraj et al., 2013). A treatment of neuroblastoma cells showed that lysosomal cholesterol accumulation is a stress response protecting lysosomal membrane integrity in response to early apoptotic stress. However, high cholesterol also stimulated α-Syn aggregation (Eriksson et al., 2017).

Six drugs from the top 10 in the lysosomal dysfunction clusters were also found in the top 10 in the mitochondrial dysfunction and/or the zebrafish dataset. There were three drugs (emetine, vorinostat, and sirolimus) that scored in the top 10 in clusters from all three datasets. In addition, geldanamycin, cephaeline, and homoharringtonine scored in the top 10 for the lysosomal dysfunction and GCH1 mutant data. Additional four drugs (alvocidib, cefixime, prucalopride, and quinacrine hydrochloride) overlapped in the clusters from the mitochondrial dysfunction and zebrafish datasets. The PD-relevant **geldanamycin** and **sirolimus** have been reviewed above. **Emetine** (BRD-A25687296) and the structurally similar **cephaeline** (BRD-K80348542) are used in blocking protein synthesis. They can be metabolised from dopamine in a medicinal plant *Psychotria ipecacuanha*, where they were discovered (Nomura and Kutchan, 2010), but have been shown to not affect the central dopaminergic mechanisms in rats (Lal et al., 1972). **Vorinostat** (BRD-K81418486) also known as suberoylanilide hydroxamic acid (SAHA), is a HDAC inhibitor which has shown neuroprotective and survival promoting effects on dopaminergic neurons in neuron-glia cultures (Chen et al., 2012) as well as show early neuroprotection by preventing mitochondrial fragmentation in a PD cell model (Alquézar et al., 2015). **Homoharringtonine** (BRD-K76674262) is an apoptosis inducing compound which was shown to initially decrease mitofilin expression, followed by a rapid increase within 6 hours of treatment (Jin et al., 2004). Mitofilin deficiency is detrimental to cell viability and response to stress, in particular it is essential for maintaining mitochondrial structure (reviewed in van Laar et al. (2018)).

In summary, several drugs showed promising beneficial effects on PD-related pathology, while some suggest further deterioration. Notably, two drugs had potential lysosomal implications. All drugs that were commonly prioritised by the *down5:20* lysosomal dysfunction sPD signatures were also in the top 10 of *down5:20* from the A5 signatures. These also overlapped with drugs prioritised in JIA *down10:20*. This could be due to system artefacts or due to common inflammatory and other processes driving these diseases.

All case studies had several translation-associated down-regulated pathways. We found several plausible connections to commonly prioritised drugs for each of the case studies. The top drug prioritisation results were encouraging, given that the disease signature was generated from a human dataset with a small sample size and consequently the pathway clusters were likely of reduced quality.

True positive lists for PD

Approved drugs for PD (Supplementary Table E.6). We extracted 58 trade and INN drug names from the FDA and European Medicines Agency (EMA) approved treatments for PD. 38 were mapped to 27 BRD IDs and only seven of those BRD IDs were present in PDxN. These appeared in 11 different drug signatures. As with other case studies, this list would be expected to be the gold standard for testing the pipeline's performance in PD. However, PD treatment is focused on symptomatic relief, therefore we initially benchmark PD with an expert-curated list of neuroprotective drugs.

PD neuroprotective drugs (Supplementary Table E.7). Our PD collaborator, Oliver Bandmann, provided us with a list of neuroprotective drugs that are predicted to target plausible causal PD mechanisms. From the six curated drug names, five were mapped to eight BRD IDs. Seven of those BRD IDs were present in PDxN, in 44 different drug signatures.

Benchmarking

As with the two other case studies, we explored the effects of drug signature features on their ranking (Fig. 7.3A). Given a much smaller set of studies, only mitochondrial dysfunction disease signature was used to assess rankings, leaving out lysosomal dysfunction prioritised drug lists for cross-validation. We took an alternative benchmarking approach for PD. Rather than scoring our prioritised drug lists for recovery of approved drugs (i.e. true positives) as means to objectively evaluate the method's performance, we benchmarked it with an expert curated neuroprotective drug list. As discussed above, a big limitation of using approved lists is that we make an assumption that the approved drugs are a treatment for the causal basis of the given disease rather than a treatment of its symptoms. In particular, PD's approved treatments are focused on symptomatic relief. Therefore, a drug repositioning method focused on treating the causal basis of disease

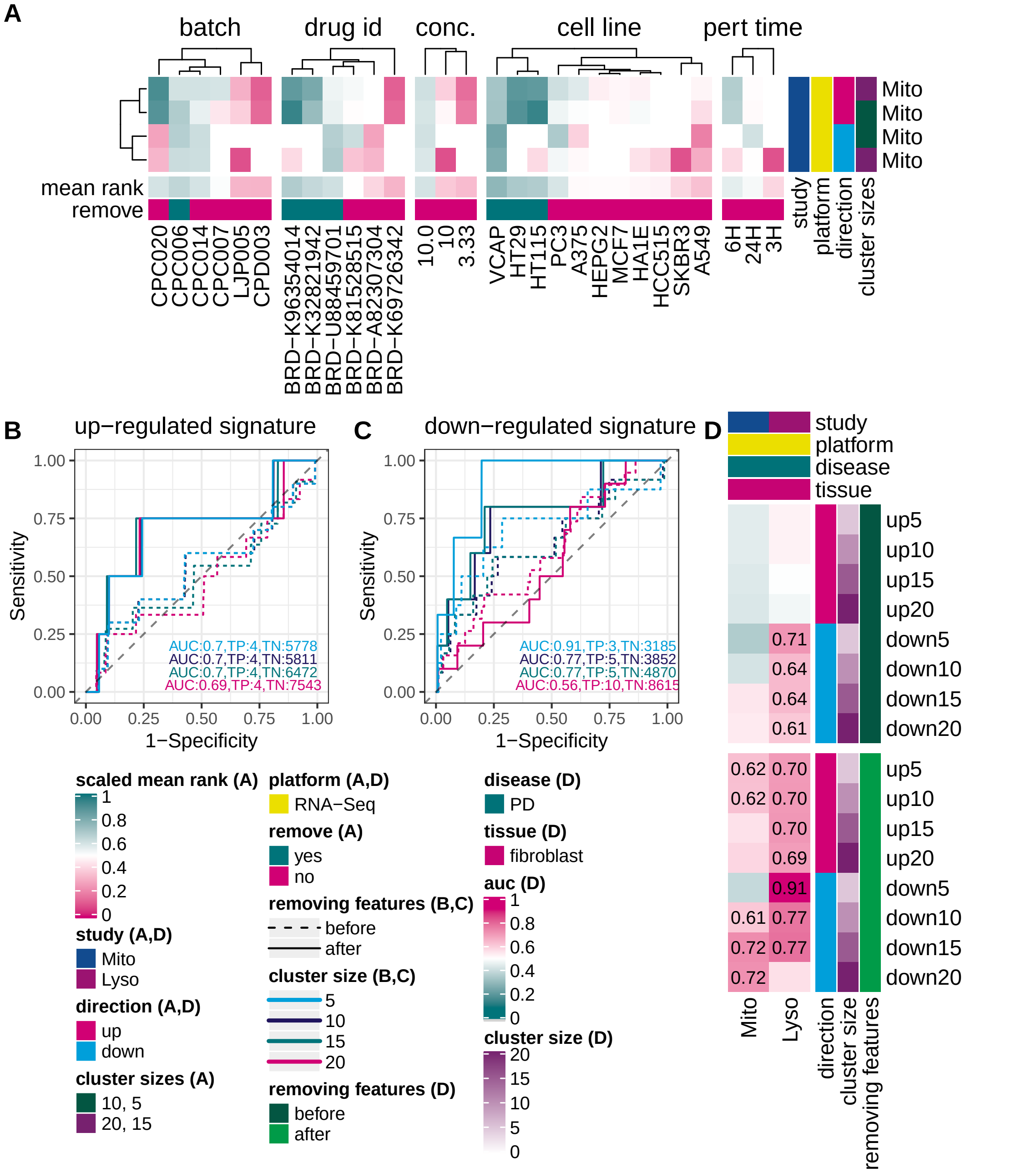


Fig. (7.3) Benchmarking Parkinson's disease (PD) prioritised drug lists with neuroprotective drugs. (A) Scaled mean TP drug ranks in PD drug lists per drug signature feature. Prioritised drug lists from the mitochondrial deficiency (Mito) up- and down-regulated pathway signatures with 5, 10, 15 or 20 pathways were scored with a curated list of neuroprotective drugs (Supplementary Table E.7). Mean rank of 1 (top prioritised drug) was scaled to 0 and the worst rank to 1. Drug signature features were selected for removal if the TP drug signatures with that feature ranked on average in the bottom 40% across mitochondrial dataset signatures, leaving out the lysosomal dataset. (C–D) ROC curves for the PDxN-derived drug lists, prioritised by the lysosomal dysfunction (Lyso) disease signatures including (B) up- and (C) down-regulated disease pathway clusters of different sizes before (dotted line) and after (full line) the removal of low-ranking drug signature features (A). The AUC, TP and TN counts are displayed for the *after* ROC curves. (D) Summary of AUCs per PD study, signature direction and pathway cluster size before (C top) and after (C bottom) removing selected drug signature features. AUC — area under the ROC curve; conc. — concentration; PDxN — Pathway Drug Coexpression Network; pert time — perturbation time; ROC — receiver operating characteristics; TN — true negative; TP — true positive.

rather than its symptoms would be penalised in approved drug-orientated scoring, if the approved drugs only treat symptoms.

Given a small neuroprotective TP list of 8 BRD IDs, from which only 6 ranked in the mitochondrial dysfunction drug lists, the ranking patterns were more extreme. The ranked drug lists clustered according to the direction of the pathway signatures. This was expected as the smaller pathway clusters (5, 10 pathways) were subsets of the larger clusters (15, 20 pathways). One batch (CPC006), three TP drugs (BRD-K96354014 (nifedipine), BRD-K32821942 (azathioprine), BRD-U88459701 (atorvastatin)) and three cell lines (VCAP, HT29, HT115) were identified as consistently low ranking. The remaining three BRD IDs represent nilotinib (BRD-K81528515), and atorvastatin (BRD-K69726342, BRD-A82307304). The two of the three poorly performing cell lines were from colon carcinoma and one from prostate cancer (Supplementary Table D.9). Due to only including mitochondrial dysfunction drug lists in this analysis, the identification of poorly performing features was highly skewed to mitochondrial dysfunction drug lists. Upon removal of these features the performance in both lysosomal and mitochondrial dysfunction improved, with the best lysosomal cluster *(*down5) improving from AUC = 0.71 to AUC = 0.91. The prioritised lists from down-regulated lysosomal dysfunction signatures showed relatively encouraging performance (AUC: 0.61–0.71), even before the removal of the identified poorly-performing features (Fig. 7.3C–D). Even though there was a limited amount of the TP drug signatures (3–10), the initial slope is relatively steep in lists from up- and down-regulated pathway clusters.

These results suggested that carefully defined TP lists might have been of higher value when assessing method's performance than approved treatments. However, a major limitation of this analysis was the low number of lysosomal and mitochondrial samples as well as only one currently included study for deciding poorly performing features. The addition of extra sPD studies would have provided clearer indications of ranking trends in drug signature features.

7.4 Evaluating Drug Prioritisation Results with Approved Drugs

In each of the case studies we evaluated the performance of the method by scoring the generated prioritised drug lists with alternative TP lists. We opted for alternative TP lists for several reasons:

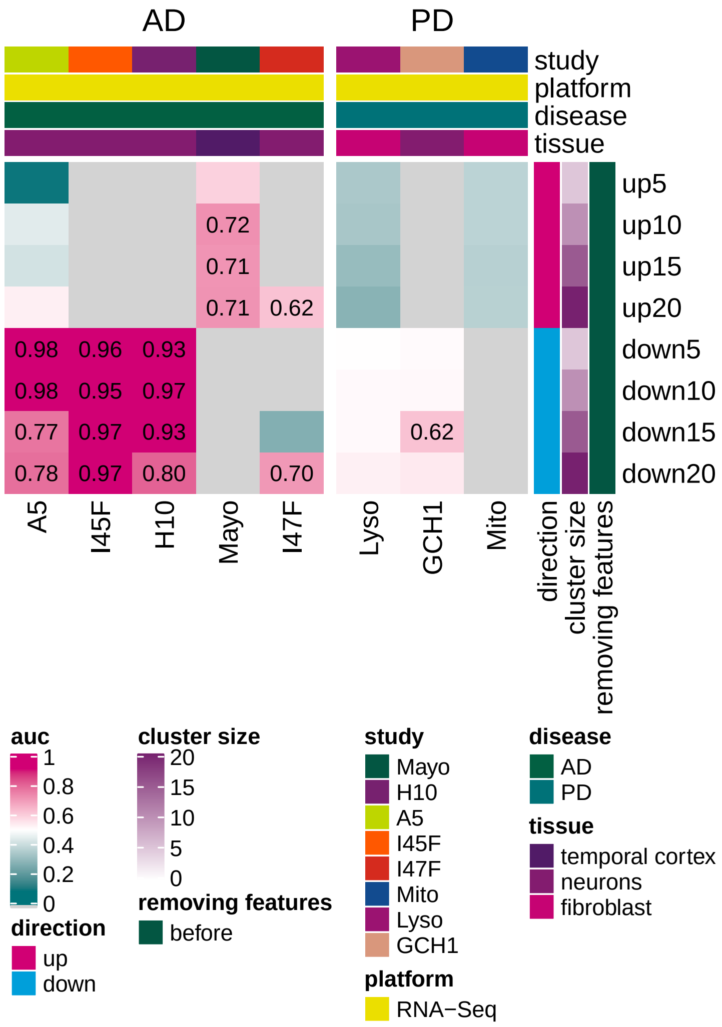


Fig. (7.4) AUC score summary for AD and PD benchmarked with approved drugs for each condition. AD 3D cell culture models and Mayo dataset were scored with AD approved drugs (Supplementary Table E.4), and PD lysosomal and mitochondrial dysfunction (Lyso, Mito, respectively), and zebrafish GCH1 were assessed with PD approved drugs (Supplementary Table E.6). Several prioritised drug lists did not associate a score with any TP drugs (grey square). The AUCs > 0.6 are displayed. AD — Alzheimer's disease; AUC — area under the ROC curve; PD — Parkinson's disease; ROC — receiver operating characteristics; TP — true positive.

(i) there were very few approved drugs for the chosen case study,

(ii) only a limited number of the approved drugs were present in PDxN,

(iii) the approved drugs were predominately meant for symptom management rather than disease-modification.

Although we excluded the approved lists in the main evaluation of the method's performance, we benchmarked the AD cell culture and Mayo dataset with AD-approved drugs, and sPD and the GCH1 zebrafish drug lists with PD-approved drugs (Fig. 7.4). The results showed distinct patterns between drug lists from up- and down-regulated pathway clusters. For example, Mayo up-regulated clusters performed much better than the down-regulated, while the down-regulated AD cell culture signatures performed better than the corresponding up-regulated clusters. Due to the low number of TP in each of the lists, there were several prioritised lists without a score, meaning that no TP drugs were associated with a score in that prioritised drug list. There were many very high-scoring lists (AUC $>$ 0.9), but there were also a few poorly scoring lists, such as A5 *up5* and I47F *down15*. Nevertheless, it is hard to make any conclusive statements as the number of the TP drugs was limited, and several drug lists had no score associated with the TP drug signatures. The PD-approved list scored poorly in the up-regulated clusters, and approximately at random in the down-regulated lysosomal dysfunction and GCH1 lists.

These results suggested that prioritising drugs separately for the up- and down-regulated clusters might have provided better drug candidates than if the disease signatures from both directions were considered together. Further work considering the effect of combining the prioritisation scores would provide further insight into this observed performance property. While our method showed great sensitivity and selectivity for the AD-approved drug list, the list was very limited in length. However, it could indicate that fewer, but better fitting drugs might serve as superior predictors of performance than the TP lists with a broader selection of mixed-effect drugs. For example, we showed that our method is better at recalling neuroprotective than approved PD drugs.

7.5 Discussion

In this chapter, we applied our drug repositioning method to two case studies: Alzheimer's and Parkinson's disease. Both represent common neurodegenerative diseases with dysregulated protein aggregation and clearance mechanisms. Both associate increased risk with increased age. The two diseases were selected because although they have both been heavily researched, neither has available disease-modifying treatments. In addition, we had the impactful opportunity for our collaborators to test our prioritised drug candidates *in vitro* and *in vivo*.

In contrast to Chapter 6, where we used a set of publicly available datasets, we wanted to evaluate a closer to real-life application scenario by analysing collaborator-provided datasets which often come with limited sample sizes. Even though there were low numbers of case and control samples available, our approach identified key pathways in both disease systems. In AD, a set of biologically-relevant dysregulated pathways were determined using a 3D cell culture model displaying low Aβ42/40 ratio. In PD, we identified a set of PD-related pathways in a lysosomal dysfunction dataset. No pathways with known links to lysosomal dysfunction were identified. Intriguingly, both AD and PD dysregulated pathways included a set of down-regulated translation-related pathways, suggesting partially overlapping neurodegenerative disease mechanisms.

The commonalities between down-regulated pathways led to an overlapping set of the top prioritised drugs. We have successfully linked the top prioritised drugs to their respective diseases for both AD and PD signatures. Promisingly, the top prioritised drug for AD, huperzine A, has already been shown to improve cognition in AD and is hypothesised to also be disease-modifying.

In this chapter, we analysed a mix of human and disease model data in order to increase the chance of translational success in the drug development pipeline by prioritising the drug candidates that have been prioritised for human as well as disease model data. Before these are validated experimentally, the findings from the literature review would be used and the most promising drug candidates and drugs related to those would be further theoretically evaluated for their solubility, bioavailability, and toxicity. In particular, an important consideration in neurodegenerative diseases that the PDxN method did not account for is the drug's ability to penetrate the blood-brain barrier. The top candidates with beneficial pharmacokinetic profiles would then be tested in an *in vitro* disease model to assess its efficacy and to optimise their dose and exposure times, before any further *in vitro* and *in vivo* tests would be carried out. Therefore, each promising drug candidate identified in an *in silico* drug repositioning pipeline would undergo extensive *in vitro* and *in vivo* validation before it would be used in a clinical trial.

We benchmarked AD with a set of drugs identified to ameliorate AD-related pathology in a large-scale drug screen. Our method showed high sensitivity and selectivity in the top 100–1000 ranked drugs, but overall showed near-random performance. We benchmarked PD with a curated list of neuroprotective drugs. The down-regulated signatures from lysosomal dysfunction yielded relatively high AUC scores in the range of 0.61–0.71. Prioritised drug lists for the lysosomal dysfunction signatures demonstrated better sensitivity and specificity with neuroprotective drugs compared to PD-approved drugs. The difference in performance could be due to heterogeneity of drug classes or due to the overabundance of symptomatic relief over disease-modifying treatments in the approved list of PD drugs. Both sJIA and AD studies yielded high AUC scores when benchmarked with their respective approved lists. However, these lists included a low number of TP drug signatures, leading to unreliable performance scores. This highlighted another weakness of the current gold-standard approach to benchmarking for diseases with few approved treatments.

The sample numbers were limited in both AD and PD collaborator datasets. However, we estimate that the cell culture and animal model heterogeneity was lower compared to that in human samples. Thus, we expected that our results included lower signal-to-noise ratio in lysosomal and mitochondrial dysfunction results compared to that in 3D culture models and the GCH1 zebrafish dataset. We could implement several publicly available sPD datasets to increase the signal-to-noise ratio in the sPD datasets. We could thus increase the confidence in the quality of lysosomal and mitochondrial dysfunction results. Due to the low number of samples available in both case studies as well as the low number of TP drugs approved for AD and PD case studies we did not compare the PDxN method's performance in AD and PD with an alternative drug repositioning method.

In addition to ambiguity in selecting the most appropriate TP list, there are several limitations to the current benchmarking design. The TP drugs from a large-scale screen, used in AD benchmarking, may not include all the drugs that could have a beneficial effect on the 3D cell model. Although 1200 were tested, the screen was done at one concentration for one exposure time in one cell line, thus potentially missing effective drugs. We did not investigate whether some drug–cell line pairs were consistently performing poorly. However, anecdotal evidence (not shown) suggested that certain drug–cell line combinations performed better than if the drug was tested on a different cell line. We could therefore construct a benchmark that investigates drug signature feature combinations and their accumulative effect. The major drawback of the current benchmarking approach is that all predicted drugs are marked as TN. We could partially overcome this by using the toxicity data from the 3DDS; Drugs with increased toxicity in the 3D model could be used as TNs.

We observed that shorter TP lists indicated better performance. This could be due to those lists including a highly selective range of TP drugs compared to larger lists where drugs with mixed effects and mechanism of action are combined. In the previous chapter, we have demonstrated that the method was sensitive to the differences in anti-inflammatory and immunosuppressive drug classes. Benchmarking with ATC classes or drug groups with similar MOAs would have provided additional insight into the sensitivity of our system as well as into potential key drug and disease mechanisms. We observed that the performance was highly dependent on the direction of the signature used for drug prioritisation. The down-regulated clusters performed better than the up-regulated clusters in AD cell cultures and mitochondrial datasets. Prioritising drugs for directional clusters provides a unique opportunity to consider combination therapy where the top drugs from the up- and down-regulated clusters could be considered in pairs targeting independent mechanisms contributing to disease severity.

In addition to using our method to prioritise drug candidates, it could be used to validate preclinical models. We could investigate the concordance of drugs predicted by an *in vitro* model and drugs predicted in the human data. It would allow us to assess the suitability of the model system as a preclinical model and prioritise them when multiple models are available.

We demonstrated that our drug repositioning pipeline was able to prioritise drug candidates with disease-relevant MOAs. We applied the method to three distinct case studies and characterised its weaknesses and strengths. We are optimistic of the translational potential of this newly developed drug repositioning pipeline as well as its role in increasing insight into the causal basis of disease.

Chapter 8

Conclusions

8.1 Summary

In this thesis, we described the development of a drug synonym database, and a novel drug repositioning pipeline as well as its application to three case studies. We supported the pipeline development results with extensive characterisation of the method features. In addition, we explored the properties of the current gold standard benchmarking approaches.

In **Chapter 2: Background** we introduced the drug repositioning field and evaluated the current computational methods. We identified the opportunity for a novel pathway-based correlation network drug repositioning method. In addition, we evaluated the current benchmarking practices in the field, identifying their lack of consistency.

In **Chapter 3: Materials and Methods** we described concise, detailed and reproducible materials and methods that powered the development of the work described in the subsequent chapters.

In **Chapter 4: KATdb, the Drug Synonym Database** we provided a novel database that can be utilised in drug discovery, development of new drug repositioning methods and benchmarking. We proposed a systematic method to evaluate the correctness of extracted relationships, which carries important implications for data quality and curation in other semantic-databases and knowledge-graphs based on public databases. We developed a user-friendly visual interface that facilitates the translation and exploration of drug synonym relationships. We have demonstrated increased translation rates between different drug name types by using KATdb.

In **Chapter 5: The Drug Repositioning Pipeline** we described the drug repositioning pipeline in detail and characterised the underlying Pathway-Drug coexpression Network (PDxN) that powers the prioritisation of drug candidates for *in vitro* and *in vivo* testing. We have computationally improved the original Pathway Correlation Network (PCxN) method (Pita-Juárez et al., 2018) and thus increased its potential for further applications and development. We evaluated the biological significance of PDxN based on bipartite clustering of the whole network and enrichment with pathway groups and drug classification terms. We demonstrated that PDxN clusters functionally enrich with pathway and drug annotations, pointing to its ability to capture biologically meaningful relationships. In addition, we describe the remaining drug repositioning pipeline components: the disease signature generation, signature processing and drug prioritisation, and benchmarking.

In **Chapter 6: Evaluation of the System: Application to juvenile idiopathic arthritis (JIA)** we applied the pipeline to JIA and explored the strengths and weaknesses of our repositioning method and the current benchmarking approaches. We identified several biologically relevant pathways and, more importantly, several promising drug candidates. We demonstrated that the differential pathway expression could overcome platform effect and that it yielded pathway expression profiles that clustered based on underlying biological differences and similarities. We compared our pathway-level approach to a standard gene-level differential expression and demonstrated that the pathway-level results were more comparable across platforms. We investigated the top prioritised drug candidates and successfully linked the majority to established rheumatoid arthritis and JIA treatments or processes. We benchmarked the method's performance with three different true positive (TP) lists: an approved list, a list of anti-inflammatory and antirheumatic drugs, and a list of immunosuppressant drugs. We have shown that our method prioritises drugs offering disease-modifying treatments over drugs offering symptomatic relief. We have compared the pipeline's performance to an alternative well-established method LINCS (Subramanian et al., 2017) and showed the increased sensitivity of our method to current treatment trends in JIA.

In **Chapter 7: Case Studies: Neurodegenerative Diseases** we applied the pipeline to two additional case studies: Alzheimer's and Parkinson's disease (AD and PD, respectively), where we analysed collaborator-provided RNA-Seq datasets. We identified several biologically relevant pathways for each disease and, as in the previous chapter, we prioritised several promising drug candidates that have the potential to be tested *in vitro* and *in vivo* as part of our established collaborations. We benchmarked AD with results from a large drug screen performed on 3D cell culture models similar to those analysed in the chapter. While the overall performance indicated randomness of the prioritised list, the initial increased steepness of the ROC curve slope suggested that the top prioritised drugs showed greater specificity and selectivity than the remainder of the prioritised drug list. We benchmarked PD with an expert-curated list of 6 neuroprotective drugs and a list of PD-approved drugs. We showed that our method's recall of neuroprotective drugs is better than those of current symptomatic PD-approved drugs.

8.2 Scope and Limitations

The drug repositioning pipeline presented in this thesis was aimed at novel drug candidate identification and hypotheses generation, and thus, comes with some substantial limitations. The current pipeline did not consider dose, toxicity and disease tissue bioavailability of the prioritised drugs. Any further translational steps would therefore require including these into consideration throughout extensive *in vitro* and *in vivo* testing, before this work could begin to make beneficial impacts in general clinical practice.

Before *in vitro* testing, the *in silico* prioritised drug candidates would first be manually assessed for their translation potential using existing literature. The findings from the literature review would be used and the most promising drug candidates and drugs related to those would be further theoretically evaluated for their solubility, bioavailability, and toxicity. In particular, an important consideration in neurodegenerative diseases that the PDxN method did not account for is the drug's ability to penetrate the blood-brain barrier. The top candidates with beneficial pharmacokinetic profiles would then be tested in an *in vitro* disease model to assess its efficacy and to optimise their dose and exposure times, before any further *in vitro* and *in vivo* tests would be carried out. The assessment of pharmacokinetic profiles for top prioritised drugs could be automated in the future by integrating a weighted approach to scoring the final prioritised list that would be promoting drugs with beneficial pharmacokinetic features. Only drugs with promising *in vitro*, followed by promising *in vivo* results have the potential progress into clinical trials, thus it is also important to assess the suitability of preclinical disease models.

There were several overlapping drugs identified in the top prioritised candidates in all three case studies, highlighting a potential method bias towards those drugs that could be further investigated with a negative control using the non-diseased liver study. However, the drug overlap could be explained by the overlap in the pathway signatures. All three case studies had several transcriptional pathways in the top differentially expressed pathways. Therefore the similarity in the disease signatures could have driven the similarities in the top prioritised drugs. As all three case studies: JIA, AD, and PD have an immune response component in the disease mechanism the concordance in top pathways and top drug candidates might be due to disease similarities. Further investigation of non immune response-related diseases would help determine whether the overlap is due to method bias or common disease mechanisms.

An underlying limitation of the JIA case study is that due to the disease presentation and treatment protocol, several samples were from treated patients. These samples have likely biased the resulting disease signatures and consequently also the drug prioritisation. To overcome this limitation, we could use the drug signatures from the confounding medications to correct the patient expression signatures or employ further study curation to include data only from untreated patients. However, the latter could significantly reduce the number of studies and samples and thus reduce the reliability of the results.

In the AD case study, we focused on prioritising drug candidates from preclinical models, because the human data comes with considerable limitations. In AD, the human data represents the disease end stage due to the inaccessibility of the diseased tissue during the disease progression. While in PD, the disease tissue availability is circumvented by analysing a more readily available tissue, fibroblasts, as proxy. However, it is unlikely that fibroblasts encompass all mechanisms involved in PD progression. Another limitation of human data is that it is more heterogeneous and often harder to obtain than cell or animal models, which leads to decreased statistical power unless well-designed large cohort studies are conducted. Nevertheless, human data is more representative of human disease compared to any *in vitro* or *in vivo* disease model, but the models allow more extensive tests e.g. large drug screens as well as detailed characterisation of specific disease mechanisms.

Despite these important considerations, we provided a tool which has the potential to speed-up the novel drug repositioning candidate discovery and guide in delivering disease-modifying treatments to patients.

8.3 Future Work

KATdb. We are aiming to establish KATdb as an online resource available for use and exploration. In addition, further work into increasing correctness would benefit not only our resource, but also other semantic databases and knowledge graphs. Methods and approaches to resolving individual drugs from drug combinations will be explored.

The Drug Repositioning Pipeline. Further interrogation of PDxN correlation relationships has the potential to increase insight into poorly characterised drugs and their mechanisms of action. Additionally, interrogation of network topology with disease signature pathway clusters could yield causal disease mechanisms. Rather than employing a summarisation-based signature processing, we could exploit network topology-based approaches such as seeded random walks. The current network approach was based on a background of mixed tissues. In order to increase the sensitivity for tissue-specific disease dysregulation, tissue-specific networks could be explored. Tissue-specific networks could assess the coexpression of drug signatures from cell lines representing that tissue on a background of tissue-specific gene expression data. The drug signatures were generated predominantly on cancer data. Thus, whilst these signatures may be more suitable for drug repositioning for cancer treatments, they will be less predictive of drug success in other disease areas (Paranjpe et al., 2019).

Case Studies. In addition to using our method to prioritise drug candidates, it could be used to validate preclinical models. We could investigate the concordance of disease signatures and drugs predicted by an *in vitro* or *in vivo* model compared to those defined from the human data. It would allow us to assess the suitability of the model systems as preclinical models and prioritise the most suitable when multiple models are available.

Benchmarking. Current benchmarking trends, which largely focus on using approved drug-disease pairs, are limited in their evaluation of, and application to poorly characterised diseases that are scarce in available treatments. Furthermore, approved treatments often only alleviate symptoms, as opposed to targeting the causal basis of the disease. We have shown that benchmarking with specific Anatomical Therapeutic Chemical (ATC) classification classes might yield differential estimations of the method's performance. Assessing sensitivity and specificity for individual drug classes might even lead to the identification of more successful treatment groups, and hence also improve our understanding of key disease mechanisms.

Validation of identified drug candidates. Top drug candidates prioritised for AD and PD will move to *in vitro* and *in vivo* testing, providing validation of the work described here as well as, more importantly, identifying possible treatments for the diseases in question. In collaboration with the drug testing facilities at the Tanzi and Kim labs, we will explore our system's utilisation of prioritising drug combinations.

8.4 Impact

The importance of rapid drug repositioning is being demonstrated during the current coronavirus disease 2019 (COVID-19) pandemic. The immediate need for accessible treatments has been recognised globally, with cheap approved treatments primarily investigated for their efficacy against COVID-19. The PDxN methodology could be applied to either cell models of COVID-19 or to patient data in order to rapidly prioritise promising treatments. At the same time, KATdb could be used to assess the current COVID-19 clinical trials to provide an overview of most promising treatments and therapeutic classes.

This thesis provides a well-characterised drug repositioning method with potential for high translational impact. We have developed a systems biology approach to drug repositioning, by exploiting functional relationships between pathways and drug signatures. In the process, we have improved the usability of a published method by decreasing its computational requirements, making it suitable for further applications and development. Our method is based on a pathway-drug coexpression network that can be interrogated with disease-specific signatures for prioritising drug candidates, providing insight into mechanisms of disease. We have identified several promising drug candidates for juvenile idiopathic arthritis, Alzheimer's and Parkinson's disease. We showed that our method prioritises the disease-modifying treatments over drugs offering symptomatic relief. We compared the method's performance to an alternative well-established method and show the increased sensitivity of our method to current treatment trends when applied to the JIA case study. The successful translation of drug candidates identified as part of this project can speed up the drug-discovery pipeline and thus more rapidly and efficiently deliver disease-modifying treatments to patients.

References

R. Abeti, A. Y. Abramov, and M. R. Duchen. Beta-amyloid activates PARP causing astrocytic metabolic failure and neuronal death. *Brain*, 134(Pt 6):1658–1672, 2011. ISSN 0006-8950, 1460-2156. doi: 10.1093/brain/awr104.

I. Acosta-Colman, N. Palau, J. Tornero, A. Fernández-Nebro, F. Blanco, I. González-Alvaro, J. D. Cañete, J. Maymó, J. Ballina, B. Fernández-Gutiérrez, A. Olivé, H. Corominas, A. Erra, O. Canela-Xandri, A. Alonso, M. López Lasanta, R. Tortosa, A. Julià, and S. Marsal. GWAS replication study confirms the association of PDE3A-SLCO1C1 with anti-TNF therapy response in rheumatoid arthritis. *Pharmacogenomics*, 14(7):727–734, 2013.

A. Adamczyk, G. A. Czapski, H. Jeśko, and R. P. Strosznajder. Non a beta component of Alzheimer's disease amyloid and amyloid beta peptides evoked poly(ADP-ribose) polymerase-dependent release of apoptosis-inducing factor from rat brain mitochondria. *J. Physiol. Pharmacol.*, 56 Suppl 2:5–13, 2005. ISSN 0867-5910, 1899-1505.

Affymetrix. Data sheet, GeneChip human genome arrays. *RNA ARRAYS AND REAGENTS*, 2003a.

Affymetrix. Technical note, design and performance of the GeneChip® human genome U133 plus 2.0 and human genome U133A 2.0 array. *RNA ARRAYS AND REAGENTS*, 2003b.

A. Aggarwal, A. Bhardwaj, S. Alam, and R. Misra. Evidence for activation of the alternate complement pathway in patients with juvenile rheumatoid arthritis. *Rheumatology*, 39 (2):189–192, 2000. ISSN 1462-0324. doi: 10.1093/rheumatology/39.2.189.

V. Aidinis, P. Carninci, M. Armaka, W. Witke, V. Harokopos, N. Pavelka, D. Koczan, C. Argyropoulos, M.-M. Thwin, S. Möller, K. Waki, P. Gopalakrishnakone, P. Ricciardi-Castagnoli, H.-J. Thiesen, Y. Hayashizaki, and G. Kollias. Cytoskeletal rearrangements in synovial fibroblasts as a novel pathophysiological determinant of modeled rheumatoid arthritis. *PLoS Genet.*, 1(4):e48, 2005.

S. A. Akhondi, S. Muresan, A. J. Williams, and J. A. Kors. Ambiguity of non-systematic chemical identifiers within and between small-molecule databases. *J. Cheminform.*, 7: 54, 2015.

S. Akioka. Interleukin-6 in juvenile idiopathic arthritis. *Mod. Rheumatol.*, 29(2):275–286, 2019.

P. Alam, S. K. Chaturvedi, M. K. Siddiqi, R. K. Rajpoot, M. R. Ajmal, M. Zaman, and R. H. Khan. Vitamin k3 inhibits protein aggregation: Implication in the treatment of amyloid diseases. *Sci. Rep.*, 6:26759, 2016. ISSN 2045-2322. doi: 10.1038/srep26759.

S. I. Alfonso, J. A. Callender, B. Hooli, C. E. Antal, K. Mullin, M. A. Sherman, S. E. Lesné, M. Leitges, A. C. Newton, R. E. Tanzi, and R. Malinow. Gain-of-function mutations in protein kinase Cα (PKCα) may promote synaptic defects in Alzheimer's disease. *Sci. Signal.*, 9(427):ra47, 2016. ISSN 1937-9145, 1945-0877. doi: 10.1126/scisignal.aaf6209.

V. Allaj, C. Guo, and D. Nie. Non-steroid anti-inflammatory drugs, prostaglandins, and cancer. *Cell Biosci.*, 3(1):8, 2013.

F. Allantaz, D. Chaussabel, D. Stichweh, L. Bennett, W. Allman, A. Mejias, M. Ardura, W. Chung, E. Smith, C. Wise, K. Palucka, O. Ramilo, M. Punaro, J. Banchereau, and V. Pascual. Blood leukocyte microarrays to diagnose systemic onset juvenile idiopathic arthritis and follow the response to IL-1 blockade. *J. Exp. Med.*, 204(9):2131–2144, 2007.

M. Allen, M. M. Carrasquillo, C. Funk, B. D. Heavner, F. Zou, C. S. Younkin, J. D. Burgess, H.-S. Chai, J. Crook, J. A. Eddy, H. Li, B. Logsdon, M. A. Peters, K. K. Dang, X. Wang, D. Serie, C. Wang, T. Nguyen, S. Lincoln, K. Malphrus, G. Bisceglio, M. Li, T. E. Golde, L. M. Mangravite, Y. Asmann, N. D. Price, R. C. Petersen, N. R. Graff-Radford, D. W. Dickson, S. G. Younkin, and N. Ertekin-Taner. Human whole genome genotype and transcriptome data for Alzheimer's and other neurodegenerative diseases. *Sci Data*, 3: 160089, 2016.

C. Alquézar, E. Barrio, N. Esteras, A. de la Encarnación, F. Bartolomé, J. A. Molina, and Á. Martín-Requero. Targeting cyclin D3/CDK6 activity for treatment of Parkinson's disease. *J. Neurochem.*, 133(6):886–897, 2015. ISSN 0022-3042, 1471-4159. doi: 10.1111/jnc.13070.

S. R. Alsumari, D. M. AlNouri, M. M. A. El-Sayed, M. F. S. El-Din, and S. Arzoo. The sociodemographic characteristics and dietary and blood plasma fatty acid profiles of elderly saudi women with Alzheimer disease. *Lipids Health Dis.*, 18(1):77, 2019. ISSN 1476-511X. doi: 10.1186/s12944-019-1029-0.

G. M. Altschuler, O. Hofmann, I. Kalatskaya, R. Payne, S. J. Ho Sui, U. Saxena, A. V. Krivtsov, S. A. Armstrong, T. Cai, L. Stein, and W. A. Hide. Pathprinting: An integrative approach to understand the functional basis of disease. *Genome Med.*, 5(7):68, 2013.

M. Altuna, A. Urdánoz-Casado, J. Sánchez-Ruiz de Gordoa, M. V. Zelaya, A. Labarga, J. M. J. Lepesant, M. Roldán, I. Blanco-Luquin, Á. Perdones, R. Larumbe, I. Jericó, C. Echavarri, I. Méndez-López, L. Di Stefano, and M. Mendioroz. DNA methylation signature of human hippocampus in Alzheimer's disease is linked to neurogenesis. *Clin. Epigenetics*, 11(1):91, 2019. ISSN 1868-7075, 1868-7083. doi: 10.1186/s13148-019-0672-7.

Alzheimer's Association. 2019 Alzheimer's disease facts and figures. *Alzheimers. Dement.*, 15(3):321–387, 2019. ISSN 1552-5260. doi: 10.1016/j.jalz.2019.01.010.

A. Amritraj, Y. Wang, T. J. Revett, D. Vergote, D. Westaway, and S. Kar. Role of cathepsin D in U18666A-induced neuronal cell death: potential implication in Niemann-Pick type C disease pathogenesis. *J. Biol. Chem.*, 288(5):3136–3152, 2013. ISSN 0021-9258, 1083-351X. doi: 10.1074/jbc.M112.412460.

C. Andronis, A. Sharma, V. Virvilis, S. Deftereos, and A. Persidis. Literature mining, ontologies and information visualization for drug repurposing. *Brief. Bioinform.*, 12(4): 357–368, 2011.

P. Angel and M. Karin. The role of jun, fos and the AP-1 complex in cell-proliferation and transformation. *Biochim. Biophys. Acta*, 1072(2-3):129–157, 1991.

S. Ansar, J. A. Burlison, M. K. Hadden, X. M. Yu, K. E. Desino, J. Bean, L. Neckers, K. L. Audus, M. L. Michaelis, and B. S. J. Blagg. A non-toxic hsp90 inhibitor protects neurons from abeta-induced toxicity. *Bioorg. Med. Chem. Lett.*, 17(7):1984–1990, 2007. ISSN 0960-894X. doi: 10.1016/j.bmcl.2007.01.017.

H. Appelqvist, L. Sandin, K. Björnström, P. Saftig, B. Garner, K. Ollinger, and K. Kågedal. Sensitivity to lysosome-dependent cell death is directly regulated by lysosomal cholesterol content. *PLoS One*, 7(11):e50262, 2012. ISSN 1932-6203. doi: 10.1371/journal.pone.0050262.

I. Appolloni, F. Calzolari, G. Corte, R. Perris, and P. Malatesta. Six3 controls the neural progenitor status in the murine CNS. *Cereb. Cortex*, 18(3):553–562, 2008. ISSN 1047-3211, 1460-2199. doi: 10.1093/cercor/bhm092.

K. Armon. Outcomes for juvenile idiopathic arthritis. *Paediatr. Child Health*, 28(2):64–72, 2018. ISSN 1751-7222, 1878-206X. doi: 10.1016/j.paed.2017.10.010.

T. T. Ashburn and K. B. Thor. Drug repositioning: identifying and developing new uses for existing drugs. *Nat. Rev. Drug Discov.*, 3(8):673–683, 2004.

P. K. Auluck, M. C. Meulener, and N. M. Bonini. Mechanisms of suppression of {alpha}-synuclein neurotoxicity by geldanamycin in drosophila. *J. Biol. Chem.*, 280(4):2873–2878, 2005. ISSN 0021-9258. doi: 10.1074/jbc.M412106200.

E. Baharav, M. Bar, M. Taler, I. Gil-Ad, L. Karp, A. Weinberger, and A. Weizman. Immunomodulatory effect of sertraline in a rat model of rheumatoid arthritis. *Neuroimmunomodulation*, 19(5):309–318, 2012. ISSN 1021-7401, 1423-0216. doi: 10.1159/000339109.

M. G. Barnes, A. A. Grom, S. D. Thompson, T. A. Griffin, P. Pavlidis, L. Itert, N. Fall, D. P. Sowders, C. H. Hinze, B. J. Aronow, L. K. Luyrink, S. Srivastava, N. T. Ilowite, B. S. Gottlieb, J. C. Olson, D. D. Sherry, D. N. Glass, and R. A. Colbert. Subtype-specific peripheral blood gene expression profiles in recent-onset juvenile idiopathic arthritis. *Arthritis Rheum.*, 60(7):2102–2112, 2009.

M. G. Barnes, A. A. Grom, S. D. Thompson, T. A. Griffin, L. K. Luyrink, R. A. Colbert, and D. N. Glass. Biologic similarities based on age at onset in oligoarticular and polyarticular subtypes of juvenile idiopathic arthritis. *Arthritis Rheum.*, 62(11):3249–3258, 2010.

T. Barrett, T. O. Suzek, D. B. Troup, S. E. Wilhite, and others. NCBI GEO: mining millions of expression profiles—database and tools. *Nucleic acids*, 2005.

T. Barrett, D. B. Troup, S. E. Wilhite, P. Ledoux, D. Rudnev, C. Evangelista, I. F. Kim, A. Soboleva, M. Tomashevsky, and R. Edgar. NCBI GEO: mining tens of millions of expression profiles–database and tools update. *Nucleic Acids Res.*, 35(Database issue): D760–5, 2007.

M. Baumgart, H. M. Snyder, M. C. Carrillo, S. Fazio, H. Kim, and H. Johns. Summary of the evidence on modifiable risk factors for cognitive decline and dementia: A population-based perspective. *Alzheimers. Dement.*, 11(6):718–726, 2015. ISSN 1552-5260, 1552-5279. doi: 10.1016/j.jalz.2015.05.016.

S. J. Beckett. Improved community detection in weighted bipartite networks. *R Soc Open Sci*, 3(1):140536, 2016.

R. S. Belaiba, S. Bonello, C. Zähringer, S. Schmidt, J. Hess, T. Kietzmann, and A. Görlach. Hypoxia up-regulates hypoxia-inducible factor-1alpha transcription by involving phosphatidylinositol 3-kinase and nuclear factor kappab in pulmonary artery smooth muscle cells. *Mol. Biol. Cell*, 18(12):4691–4697, 2007.

D. C. Berwick and K. Harvey. The importance of wnt signalling for neurodegeneration in Parkinson's disease. *Biochem. Soc. Trans.*, 40(5):1123–1128, 2012. ISSN 0300-5127, 1470-8752. doi: 10.1042/BST20120122.

T. Beukelman, N. M. Patkar, K. G. Saag, S. Tolleson-Rinehart, R. Q. Cron, E. M. DeWitt, N. T. Ilowite, Y. Kimura, R. M. Laxer, D. J. Lovell, A. Martini, C. E. Rabinovich, and N. Ruperto. 2011 american college of rheumatology recommendations for the treatment of juvenile idiopathic arthritis: initiation and safety monitoring of therapeutic agents for the treatment of arthritis and systemic features. *Arthritis Care Res.*, 63(4):465–482, 2011.

T. Beukelman, F. Xie, L. Chen, J. W. Baddley, E. Delzell, C. G. Grijalva, J. D. Lewis, R. Ouellet-Hellstrom, N. M. Patkar, K. G. Saag, K. L. Winthrop, J. R. Curtis, and SABER Collaboration. Rates of hospitalized bacterial infection associated with juvenile idiopathic arthritis and its treatment. *Arthritis Rheum.*, 64(8):2773–2780, 2012. ISSN 0004-3591, 1529-0131. doi: 10.1002/art.34458.

S. L. Biroc, S. Gay, K. Hummel, C. Magill, J. T. Palmer, D. R. Spencer, S. Sa, J. L. Klaus, B. A. Michel, D. Rasnick, and R. E. Gay. Cysteine protease activity is up-regulated in inflamed ankle joints of rats with adjuvant-induced arthritis and decreases with in vivo administration of a vinyl sulfone cysteine protease inhibitor. *Arthritis Rheum.*, 44(3):703–711, 2001. ISSN 0004-3591. doi: 10.1002/1529-0131(200103)44:3<703::AID-ANR120>3.0.CO;2-2.

L. J. Blair, J. J. Sabbagh, and C. A. Dickey. Targeting hsp90 and its co-chaperones to treat Alzheimer's disease. *Expert Opin. Ther. Targets*, 18(10):1219–1232, 2014. ISSN 1472-8222, 1744-7631. doi: 10.1517/14728222.2014.943185.

E. M. Blalock, J. W. Geddes, K. C. Chen, N. M. Porter, W. R. Markesbery, and P. W. Landfield. Incipient Alzheimer's disease: microarray correlation analyses reveal major transcriptional and tumor suppressor responses. *Proc. Natl. Acad. Sci. U. S. A.*, 101(7): 2173–2178, 2004. ISSN 0027-8424. doi: 10.1073/pnas.0308512100.

V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *J. Stat. Mech.*, 2008(10):P10008, 2008.

B. J. Bloom, L. B. Tucker, L. C. Miller, and J. G. Schaller. Fibrin d-dimer as a marker of disease activity in systemic onset juvenile rheumatoid arthritis. *J. Rheumatol.*, 25(8): 1620–1625, 1998. ISSN 0315-162X.

O. Bodenreider. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic Acids Res.*, 32(Database issue):D267–70, 2004.

E. E. Bolton, Y. Wang, P. A. Thiessen, and S. H. Bryant. Chapter 12 - PubChem: Integrated platform of small molecules and biological activities. In R. A. Wheeler and avid C. Spellmeyer, editors, *Annual Reports in Computational Chemistry*, volume 4, pages 217–241. Elsevier, 2008.

B. Booth and R. Zemmel. Prospects for productivity. *Nat. Rev. Drug Discov.*, 3(5):451–456, 2004.

D. R. Borchelt, G. Thinakaran, C. B. Eckman, M. K. Lee, F. Davenport, T. Ratovitsky, C. M. Prada, G. Kim, S. Seekins, D. Yager, H. H. Slunt, R. Wang, M. Seeger, A. I. Levey, S. E. Gandy, N. G. Copeland, N. A. Jenkins, D. L. Price, S. G. Younkin, and S. S. Sisodia. Familial Alzheimer's disease-linked presenilin 1 variants elevate abeta1-42/1-40 ratio in vitro and in vivo. *Neuron*, 17(5):1005–1013, 1996. ISSN 0896-6273. doi: 10.1016/s0896-6273(00)80230-5.

J. Bové, M. Martínez-Vicente, and M. Vila. Fighting neurodegeneration with rapamycin: mechanistic insights. *Nat. Rev. Neurosci.*, 12(8):437–452, 2011. ISSN 1471-003X, 1471-0048. doi: 10.1038/nrn3068.

H. Braak and E. Braak. Neuropathological stageing of Alzheimer-related changes. *Acta Neuropathol.*, 82(4):239–259, 1991. ISSN 0001-6322. doi: 10.1007/bf00308809.

H. Braak, K. Del Tredici, U. Rüb, R. A. I. de Vos, E. N. H. Jansen Steur, and E. Braak. Staging of brain pathology related to sporadic Parkinson's disease. *Neurobiol. Aging*, 24(2):197–211, 2003. ISSN 0197-4580. doi: 10.1016/s0197-4580(02)00065-9.

A. H. Brachat, A. A. Grom, N. Wulffraat, H. I. Brunner, P. Quartier, R. Brik, L. McCann, H. Ozdogan, L. Rutkowska-Sak, R. Schneider, V. Gerloni, L. Harel, M. Terreri, K. Houghton, R. Joos, D. Kingsbury, J. M. Lopez-Benitez, S. Bek, M. Schumacher, M.-A. Valentin, H. Gram, K. Abrams, A. Martini, D. J. Lovell, N. R. Nirmala, N. Ruperto, and Pediatric Rheumatology International Trials Organization (PRINTO) and the Pediatric Rheumatology Collaborative Study Group (PRCSG). Early changes in gene expression and inflammatory proteins in systemic juvenile idiopathic arthritis patients on canakinumab therapy. *Arthritis Res. Ther.*, 19(1):13, 2017.

U. Brandes. A faster algorithm for betweenness centrality. *J. Math. Sociol.*, 25(2):163–177, 2001.

A. Breckenridge and R. Jacob. Overcoming the legal and regulatory barriers to drug repurposing. *Nat. Rev. Drug Discov.*, 18(1):1–2, 2019. ISSN 1474-1776, 1474-1784. doi: 10.1038/nrd.2018.92.

J. Brettschneider, F. Collin, B. M. Bolstad, and T. P. Speed. Quality assessment for short oligonucleotide microarray data. *arXiv*, 2007.

A. S. Brown and C. J. Patel. A review of validation strategies for computational drug repositioning. *Brief. Bioinform.*, 2016.

A. S. Brown and C. J. Patel. A standard database for drug repositioning. *Scientific Data*, 4: 170029, 2017.

A. M. Brum, J. van de Peppel, C. S. van der Leije, M. Schreuders-Koedam, M. Eijken, B. C. J. van der Eerden, and J. P. T. M. van Leeuwen. Connectivity map-based discovery of parbendazole reveals targetable human osteogenic pathway. *Proc. Natl. Acad. Sci. U. S. A.*, 112(41):12711–12716, 2015. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1501597112.

M. Campillos, M. Kuhn, A.-C. Gavin, L. J. Jensen, and P. Bork. Drug target identification using side-effect similarity. *Science*, 321(5886):263–266, 2008.

M. Camps, T. Rückle, H. Ji, V. Ardissone, F. Rintelen, J. Shaw, C. Ferrandi, C. Chabert, C. Gillieron, B. Françon, T. Martin, D. Gretener, D. Perrin, D. Leroy, P.-A. Vitte, E. Hirsch, M. P. Wymann, R. Cirillo, M. K. Schwarz, and C. Rommel. Blockade of PI3Kgamma suppresses joint inflammation and damage in mouse models of rheumatoid arthritis. *Nat. Med.*, 11(9):936–943, 2005. ISSN 1078-8956. doi: 10.1038/nm1284.

S. Canna, J. Frankovich, G. Higgins, M. R. Narkewicz, S. R. Nash, J. R. Hollister, J. B. Soep, and L. L. Dragone. Acute hepatitis in three patients with systemic juvenile idiopathic arthritis taking interleukin-1 receptor antagonist. *Pediatr. Rheumatol. Online J.*, 7:21, 2009. ISSN 1546-0096. doi: 10.1186/1546-0096-7-21.

T. Cao, X. Zhou, X. Zheng, Y. Cui, J. Z. Tsien, C. Li, and H. Wang. Histone deacetylase inhibitor alleviates the neurodegenerative phenotypes and histone dysregulation in Presenilins-Deficient mice. *Front. Aging Neurosci.*, 10:137, 2018. ISSN 1663-4365. doi: 10.3389/fnagi.2018.00137.

L. Caputi, A. Hainsworth, E. Guatteo, A. Tozzi, A. Stefani, F. Spadoni, M. Leach, G. Bernardi, and N. B. Mercuri. Actions of the sodium channel inhibitor 202W92 on rat midbrain dopaminergic neurons. *Synapse*, 48(3):123–130, 2003. ISSN 0887-4476. doi: 10.1002/syn.10195.

P. J. Carling, H. Mortiboys, C. Green, S. Mihaylov, C. Sandor, A. Schwartzentruber, R. Taylor, W. Wei, C. Hastings, S. Wong, C. Lo, S. Evetts, H. Clemmens, M. Wyles, S. Willcox, T. Payne, R. Hughes, L. Ferraiuolo, C. Webber, W. Hide, R. Wade-Martins, K. Talbot, M. T. Hu, and O. Bandmann. Deep phenotyping of peripheral tissue facilitates mechanistic disease stratification in sporadic Parkinson's disease. *Prog. Neurobiol.*, page 101772, 2020. ISSN 0301-0082, 1873-5118. doi: 10.1016/j.pneurobio.2020.101772.

J. Chambers, M. Davies, A. Gaulton, A. Hersey, S. Velankar, R. Petryszak, J. Hastings, L. Bellis, S. McGlinchey, and J. P. Overington. UniChem: a unified chemical structure cross-referencing and identifier tracking system. *J. Cheminform.*, 5(1):3, 2013.

M. Chang, S. Smith, A. Thorpe, M. J. Barratt, and F. Karim. Evaluation of phenoxybenzamine in the CFA model of pain following gene expression studies and connectivity mapping. *Mol. Pain*, 6:56, 2010.

W. Chang, J. Cheng, J. Allaire, Y. Xie, and J. McPherson. *shiny: Web Application Framework for R*, 2019. URL https://CRAN.R-project.org/package=shiny. R package version 1.3.2.

K. R. Chaudhuri, D. G. Healy, A. H. V. Schapira, and National Institute for Clinical Excellence. Non-motor symptoms of Parkinson's disease: diagnosis and management. *Lancet Neurol.*, 5(3):235–245, 2006. ISSN 1474-4422. doi: 10.1016/S1474-4422(06)70373-8.

L. M. Chen, S. Hobbie, and J. E. Galán. Requirement of CDC42 for salmonella-induced cytoskeletal and nuclear responses. *Science*, 274(5295):2115–2118, 1996.

M.-H. Chen, W.-L. R. Yang, K.-T. Lin, C.-H. Liu, Y.-W. Liu, K.-W. Huang, P. M.-H. Chang, J.-M. Lai, C.-N. Hsu, K.-M. Chao, C.-Y. Kao, and C.-Y. F. Huang. Gene expression-based chemical genomics identifies potential therapeutic drugs in hepatocellular carcinoma. *PLoS One*, 6(11):e27186, 2011.

S. H. Chen, H. M. Wu, B. Ossola, N. Schendzielorz, B. C. Wilson, C. H. Chu, S. L. Chen, Q. Wang, D. Zhang, L. Qian, X. Li, J. S. Hong, and R. B. Lu. Suberoylanilide hydroxamic acid, a histone deacetylase inhibitor, protects dopaminergic neurons from neurotoxin-induced damage. *Br. J. Pharmacol.*, 165(2):494–505, 2012. ISSN 0007-1188, 1476-5381. doi: 10.1111/j.1476-5381.2011.01575.x.

X. Chen, P. Wales, L. Quinti, F. Zuo, S. Moniot, F. Herisson, N. A. Rauf, H. Wang, R. B. Silverman, C. Ayata, M. M. Maxwell, C. Steegborn, M. A. Schwarzschild, T. F. Outeiro, and A. G. Kazantsev. The sirtuin-2 inhibitor AK7 is neuroprotective in models of Parkinson's disease but not amyotrophic lateral sclerosis and cerebral ischemia. *PLoS One*, 10(1):e0116919, 2015. ISSN 1932-6203. doi: 10.1371/journal.pone.0116919.

J. Cheng, L. Yang, V. Kumar, and P. Agarwal. Systematic evaluation of connectivity map for disease indications. *Genome Med.*, 6(12):540, 2014.

A. P. Chiang and A. J. Butte. Systematic evaluation of drug-disease relationships to identify leads for novel drug uses. *Clin. Pharmacol. Ther.*, 86(5):507–510, 2009.

T. Chiba, M. Yamada, and S. Aiso. Targeting the JAK2/STAT3 axis in Alzheimer's disease. *Expert Opin. Ther. Targets*, 13(10):1155–1167, 2009. ISSN 1472-8222, 1744-7631. doi: 10.1517/14728220903213426.

Y.-H. Chiu, J. B. Macmillan, and Z. J. Chen. RNA polymerase III detects cytosolic DNA and induces type I interferons through the RIG-I pathway. *Cell*, 138(3):576–591, 2009.

S. H. Choi, Y. H. Kim, M. Hebisch, C. Sliwinski, S. Lee, C. D'Avanzo, H. Chen, B. Hooli, C. Asselin, J. Muffat, J. B. Klee, C. Zhang, B. J. Wainger, M. Peitz, D. M. Kovacs, C. J. Woolf, S. L. Wagner, R. E. Tanzi, and D. Y. Kim. A three-dimensional human neural cell culture model of Alzheimer's disease. *Nature*, 515(7526):274–278, 2014. ISSN 0028-0836, 1476-4687. doi: 10.1038/nature13800.

Y. Choi, F. Syeda, J. R. Walker, P. J. Finerty, Jr, D. Cuerrier, A. Wojciechowski, Q. Liu, S. Dhe-Paganon, and N. S. Gray. Discovery and structural analysis of eph receptor tyrosine kinase inhibitors. *Bioorg. Med. Chem. Lett.*, 19(15):4467–4470, 2009. ISSN 0960-894X, 1464-3405. doi: 10.1016/j.bmcl.2009.05.029.

C.-J. Choong, T. Sasaki, H. Hayakawa, T. Yasuda, K. Baba, Y. Hirata, S. Uesato, and H. Mochizuki. A novel histone deacetylase 1 and 2 isoform-specific inhibitor alleviates experimental Parkinson's disease. *Neurobiol. Aging*, 37:103–116, 2016. ISSN 0197-4580, 1558-1497. doi: 10.1016/j.neurobiolaging.2015.10.001.

C. Ciurtin, V. M. Cojocaru, I. M. Miron, F. Preda, M. Milicescu, M. Bojincă, O. Costan, A. Nicolescu, C. Deleanu, E. Kovàcs, and V. Stoica. Correlation between different components of synovial fluid and pathogenesis of rheumatic diseases. *Rom. J. Intern. Med.*, 44(2):171–181, 2006. ISSN 1220-4749.

N. R. Clark, K. S. Hu, A. S. Feldmann, Y. Kou, E. Y. Chen, Q. Duan, and A. Ma'ayan. The characteristic direction: a geometrical approach to identify differentially expressed genes. *BMC Bioinformatics*, 15:79, 2014.

ClinicalTrials.gov. Clinical trial to determine tolerable dosis of vorinostat in patients with mild Alzheimer disease - full text view - ClinicalTrials.gov. https://clinicaltrials.gov/ct2/show/NCT03056495, 2019. URL https://clinicaltrials.gov/ct2/show/NCT03056495. Accessed: 2020-3-13.

CMap Group at The Broad Institute. *cmapR: CMap tools in R*, 2018. R package version 1.0.1.

A. Colombo, A. Bastone, C. Ploia, A. Sclip, M. Salmona, G. Forloni, and T. Borsello. JNK regulates APP cleavage and degradation in a model of Alzheimer's disease. *Neurobiol. Dis.*, 33(3):518–525, 2009. ISSN 0969-9961, 1095-953X. doi: 10.1016/j.nbd.2008.12.014.

J. E. Conour. Patenting repurposed drugs. https://www.patentdocs.org/2018/09/patenting-repurposed-drugs.html, 2018. URL https://www.patentdocs.org/2018/09/patenting-repurposed-drugs.html. Accessed: 2020-7-16.

A. A. Cooper, A. D. Gitler, A. Cashikar, C. M. Haynes, K. J. Hill, B. Bhullar, K. Liu, K. Xu, K. E. Strathearn, F. Liu, S. Cao, K. A. Caldwell, G. A. Caldwell, G. Marsischky, R. D. Kolodner, J. LaBaer, J.-C. Rochet, N. M. Bonini, and S. Lindquist. α-synuclein blocks ER-Golgi traffic and rab1 rescues neuron loss in Parkinson's models. *Science*, 313(5785):324–328, 2006. ISSN 0036-8075. doi: 10.1126/science.1129462.

D. Correddu and I. K. H. Leung. Targeting mRNA translation in Parkinson's disease. *Drug Discov. Today*, 24(6):1295–1303, 2019. ISSN 1359-6446, 1878-5832. doi: 10.1016/j.drudis.2019.04.003.

D. Croft, G. O'Kelly, G. Wu, R. Haw, M. Gillespie, L. Matthews, M. Caudy, P. Garapati, G. Gopinath, B. Jassal, S. Jupe, I. Kalatskaya, S. Mahajan, B. May, N. Ndegwa, E. Schmidt, V. Shamovsky, C. Yung, E. Birney, H. Hermjakob, P. D'Eustachio, and L. Stein. Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Res.*, 39(Database issue):D691–7, 2011.

G. Csardi and T. Nepusz. The igraph software package for complex network research. *InterJournal*, Complex Systems:1695, 2006. URL http://igraph.org.

A. F. Dağli, A. Karataş, C. Orhan, M. Tuzcu, M. Özgen, K. Şahin, and S. S. Koca. Antiinflammatory and antioxidant effects of gemcitabine in collagen-induced arthritis model. *Turk J Med Sci*, 47(3):1037–1044, 2017. ISSN 1300-0144. doi: 10.3906/sag-1606-80.

H. Dashti, J. R. Wedell, W. M. Westler, J. L. Markley, and H. R. Eghbalnia. Automated evaluation of consistency within the PubChem compound database. *Sci Data*, 6:190023, 2019.

W. Dauer and S. Przedborski. Parkinson's disease: mechanisms and models. *Neuron*, 39 (6):889–909, 2003. ISSN 0896-6273. doi: 10.1016/s0896-6273(03)00568-3.

J. Dausset. [iso-leuko-antibodies]. *Acta Haematol.*, 20(1-4):156–166, 1958.

E. H. Davies, E. Fulton, D. Brook, and D. A. Hughes. Affordable orphan drugs: a role for not-for-profit organizations. *Br. J. Clin. Pharmacol.*, 2017.

A. P. Davis, C. J. Grondin, R. J. Johnson, D. Sciaky, R. McMorran, J. Wiegers, T. C. Wiegers, and C. J. Mattingly. The comparative toxicogenomics database: update 2019. *Nucleic Acids Res.*, 47(D1):D948–D954, 2019.

J. Davis and M. Goadrich. The relationship between Precision-Recall and ROC curves. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, pages 233–240, New York, NY, USA, 2006. ACM.

S. Davis and P. Meltzer. GEOquery: a bridge between the gene expression omnibus (GEO) and bioconductor. *Bioinformatics*, 14:1846–1847, 2007.

F. de Benedetti, M. Massa, P. Robbioni, A. Ravelli, G. R. Burgio, and A. Martini. Correlation of serum interleukin-6 levels with joint involvement and thrombocytosis in systemic juvenile rheumatoid arthritis. *Arthritis Rheum.*, 34(9):1158–1163, 1991.

F. de Benedetti, T. Alonzi, A. Moretta, D. Lazzaro, P. Costa, V. Poli, A. Martini, G. Ciliberto, and E. Fattori. Interleukin 6 causes growth impairment in transgenic mice through a decrease in insulin-like growth factor-i. a model for stunted growth in children with chronic inflammation. *J. Clin. Invest.*, 99(4):643–650, 1997.

P. L. De Jager, G. Srivastava, K. Lunnon, J. Burgess, L. C. Schalkwyk, L. Yu, M. L. Eaton, B. T. Keenan, J. Ernst, C. McCabe, A. Tang, T. Raj, J. Replogle, W. Brodeur, S. Gabriel, H. S. Chai, C. Younkin, S. G. Younkin, F. Zou, M. Szyf, C. B. Epstein, J. A. Schneider, B. E. Bernstein, A. Meissner, N. Ertekin-Taner, L. B. Chibnik, M. Kellis, J. Mill, and D. A. Bennett. Alzheimer's disease: Early alterations in brain DNA methylation at ANK1, BIN1, RHBDF2 and other loci. *Nat. Neurosci.*, 17(9):1156–1163, 2014. ISSN 1097-6256. doi: 10.1038/nn.3786.

D. L. K. de Jong, R. A. A. de Heus, A. Rijpma, R. Donders, M. G. M. Olde Rikkert, M. Günther, B. A. Lawlor, M. J. P. van Osch, and J. A. H. R. Claassen. Effects of nilvadipine on cerebral blood flow in patients with Alzheimer disease. *Hypertension*, 74 (2):413–420, 2019. ISSN 0194-911X, 1524-4563. doi: 10.1161/HYPERTENSIONAHA.119.12892.

A. A. de Melker, G. van der Horst, and J. Borst. c-cbl directs EGF receptors into an endocytic pathway that involves the ubiquitin-interacting motif of eps15. *J. Cell Sci.*, 117(Pt 21):5001–5012, 2004. ISSN 0021-9533. doi: 10.1242/jcs.01354.

A. De Simone and A. Milelli. Histone deacetylase inhibitors as multitarget ligands: New players in Alzheimer's disease drug discovery? *ChemMedChem*, 14(11):1067–1073, 2019. ISSN 1860-7179, 1860-7187. doi: 10.1002/cmdc.201900174.

B. Dehay, J. Bové, N. Rodríguez-Muela, C. Perier, A. Recasens, P. Boya, and M. Vila. Pathogenic lysosomal depletion in Parkinson's disease. *J. Neurosci.*, 30(37):12535–12544, 2010. ISSN 0270-6474, 1529-2401. doi: 10.1523/JNEUROSCI.1920-10.2010.

F. Demir, N. Eroglu, A. Bahadir, and M. Kalyoncu. A case of acute lymphoblastic leukemia mimicking juvenile idiopathic arthritis. *North Clin Istanb*, 6(2):184–188, 2019. ISSN 2536-4553. doi: 10.14744/nci.2018.48658.

P. Di Fruscia, E. Zacharioudakis, C. Liu, S. Moniot, S. Laohasinnarong, M. Khongkow, I. F. Harrison, K. Koltsida, C. R. Reynolds, K. Schmidtkunz, M. Jung, K. L. Chapman, C. Steegborn, D. T. Dexter, M. J. E. Sternberg, E. W.-F. Lam, and M. J. Fuchter. The discovery of a highly selective 5,6,7,8-tetrahydrobenzo[4,5]thieno[2,3- d]pyrimidin-4(3 H)-one SIRT2 inhibitor that is neuroprotective in an in vitro Parkinson's disease model. *ChemMedChem*, 10(1):69–82, 2015. ISSN 1860-7179. doi: 10.1002/cmdc.201402431.

M. Digicaylioglu and S. A. Lipton. Erythropoietin-mediated neuroprotection involves cross-talk between jak2 and NF-kappaB signalling cascades. *Nature*, 412(6847):641–647, 2001.

J. A. DiMasi, H. G. Grabowski, and R. W. Hansen. Innovation in the pharmaceutical industry: New estimates of R&D costs. *J. Health Econ.*, 47:20–33, 2016.

I. D. Dinov. Volume and value of big healthcare data. *J Med Stat Inform*, 4, 2016. ISSN 2053-7662. doi: 10.7243/2053-7662-4-3.

J. H. W. Distler, R. H. Wenger, M. Gassmann, M. Kurowska, A. Hirth, S. Gay, and O. Distler. Physiologic responses to hypoxia and implications for hypoxia-inducible factors in the pathogenesis of rheumatoid arthritis. *Arthritis Rheum.*, 50(1):10–23, 2004. ISSN 0004-3591. doi: 10.1002/art.11425.

C. F. Dormann and R. Strauss. A method for detecting modules in quantitative bipartite networks. *Methods Ecol. Evol.*, 5(1):90–98, 2014.

C. F. Dormann, J. Frueund, N. Bluethgen, and B. Gruber. Indices, graphs and null models: analyzing bipartite ecological networks. *The Open Ecology Journal*, 2:7–24, 2009.

Drug and Therapeutics Bulletin. Drugs and their names. *Drug and Therapeutics Bulletin*, 56(3):33–36, 2018.

DrugBank. Conjugated estrogens. https://www.drugbank.ca/drugs/DB00286, 2005. Accessed: 2019-12-12.

Q. Duan, S. P. Reid, N. R. Clark, Z. Wang, N. F. Fernandez, A. D. Rouillard, B. Readhead, S. R. Tritsch, R. Hodos, M. Hafner, M. Niepel, P. K. Sorger, J. T. Dudley, S. Bavari, R. G. Panchal, and A. Ma'ayan. L1000CDS2: LINCS L1000 characteristic direction signatures search engine. *npj Systems Biology and Applications*, 2:16015, 2016.

J. T. Dudley, T. Deshpande, and A. J. Butte. Exploiting drug-disease relationships for computational drug repositioning. *Brief. Bioinform.*, 12(4):303–311, 2011a.

J. T. Dudley, M. Sirota, M. Shenoy, R. K. Pai, S. Roedder, A. P. Chiang, A. A. Morgan, M. M. Sarwal, P. J. Pasricha, and A. J. Butte. Computational repositioning of the anticonvulsant topiramate for inflammatory bowel disease. *Sci. Transl. Med.*, 3(96): 96ra76, 2011b.

B. M. Dufty, L. R. Warner, S. T. Hou, S. X. Jiang, T. Gomez-Isla, K. M. Leenhouts, J. T. Oxford, M. B. Feany, E. Masliah, and T. T. Rohn. Calpain-cleavage of alpha-synuclein: connecting proteolytic processing to disease-linked aggregation. *Am. J. Pathol.*, 170(5): 1725–1738, 2007. ISSN 0002-9440. doi: 10.2353/ajpath.2007.061232.

J. Duke, J. Friedlin, and P. Ryan. A quantitative analysis of adverse events and "overwarning" in drug labeling. *Arch. Intern. Med.*, 171(10):941–954, 2011.

S. Durinck, P. T. Spellman, E. Birney, and W. Huber. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomart. *Nat. Protoc.*, 4(8):1184–1191, 2009.

V. Echeverria, S. Burgess, J. Gamble-George, R. Zeitlin, X. Lin, C. Cao, and G. W. Arendash. Sorafenib inhibits nuclear factor kappa b, decreases inducible nitric oxide synthase and cyclooxygenase-2 expression, and restores working memory in APPswe mice. *Neuroscience*, 162(4):1220–1231, 2009. ISSN 0306-4522, 1873-7544. doi: 10.1016/j.neuroscience.2009.05.019.

C. B. Eckman, N. D. Mehta, R. Crook, J. Perez-tur, G. Prihar, E. Pfeiffer, N. Graff-Radford, P. Hinder, D. Yager, B. Zenk, L. M. Refolo, C. M. Prada, S. G. Younkin, M. Hutton, and J. Hardy. A new pathogenic mutation in the APP gene (I716V) increases the relative proportion of a beta 42(43). *Hum. Mol. Genet.*, 6(12):2087–2089, 1997. ISSN 0964-6906. doi: 10.1093/hmg/6.12.2087.

E. Edelman, A. Porrello, J. Guinney, B. Balakumaran, A. Bild, P. G. Febbo, and S. Mukherjee. Analysis of sample set enrichment scores: assaying the enrichment of sets of genes for individual samples in genome-wide expression profiles. *Bioinformatics*, 22(14): e108–16, 2006.

A. Elbaz, J. H. Bower, D. M. Maraganore, S. K. McDonnell, B. J. Peterson, J. E. Ahlskog, D. J. Schaid, and W. A. Rocca. Risk tables for parkinsonism and Parkinson's disease. *J. Clin. Epidemiol.*, 55(1):25–31, 2002. ISSN 0895-4356. doi: 10.1016/s0895-4356(01) 00425-5.

L. Ellgaard and A. Helenius. Quality control in the endoplasmic reticulum. *Nat. Rev. Mol. Cell Biol.*, 4(3):181–191, 2003. ISSN 1471-0072. doi: 10.1038/nrm1052.

I. Eriksson, S. Nath, P. Bornefall, A. M. V. Giraldo, and K. Öllinger. Impact of high cholesterol in a Parkinson's disease model: Prevention of lysosomal leakage versus stimulation of α-synuclein aggregation. *Eur. J. Cell Biol.*, 96(2):99–109, 2017. ISSN 0171-9335, 1618-1298. doi: 10.1016/j.ejcb.2017.01.002.

European Medicines Agency. European medicines agency. https://www.ema.europa.eu/en, 2019. Accessed: 2019-8-6.

T. Falk, S. Zhang, and S. J. Sherman. Vascular endothelial growth factor B (VEGF-B) is up-regulated and exogenous VEGF-B is neuroprotective in a culture model of Parkinson's disease. *Mol. Neurodegener.*, 4:49, 2009. ISSN 1750-1326. doi: 10.1186/ 1750-1326-4-49.

N. Fall, M. Barnes, S. Thornton, L. Luyrink, J. Olson, N. T. Ilowite, B. S. Gottlieb, T. Griffin, D. D. Sherry, S. Thompson, D. N. Glass, R. A. Colbert, and A. A. Grom. Gene expression profiling of peripheral blood from patients with untreated new-onset systemic juvenile idiopathic arthritis reveals molecular heterogeneity that may predict macrophage activation syndrome. *Arthritis Rheum.*, 56(11):3793–3804, 2007.

L. Fallon, C. M. L. Bélanger, A. T. Corera, M. Kontogiannea, E. Regan-Klapisz, F. Moreau, J. Voortman, M. Haber, G. Rouleau, T. Thorarinsdottir, A. Brice, P. M. P. van Bergen En Henegouwen, and E. A. Fon. A regulated interaction with the UIM protein eps15 implicates parkin in EGF receptor trafficking and PI(3)K-Akt signalling. *Nat. Cell Biol.*, 8(8):834–842, 2006. ISSN 1465-7392. doi: 10.1038/ncb1441.

T. Fawcett. An introduction to ROC analysis. *Pattern Recognit. Lett.*, 27(8):861–874, 2006.

G. Ferrara, G. Mastrangelo, P. Barone, F. La Torre, S. Martino, G. Pappagallo, A. Ravelli, A. Taddio, F. Zulian, R. Cimaz, and Rheumatology Italian Study Group. Methotrexate in juvenile idiopathic arthritis: advice and recommendations from the MARAJIA expert consensus meeting. *Pediatr. Rheumatol. Online J.*, 16(1):46, 2018. ISSN 1546-0096. doi: 10.1186/s12969-018-0255-8.

G. R. V. Ferreira, R. B. Tomioka, L. B. Queiroz, K. Kozu, N. E. Aikawa, A. M. E. Sallum, P. Serafini, M. Tacla, E. C. Baracat, R. M. R. Pereira, E. Bonfá, and C. A. Silva. Lower genital tract infections in young female juvenile idiopathic arthritis patients. *Adv Rheumatol*, 59(1):50, 2019. ISSN 2523-3106. doi: 10.1186/s42358-019-0092-6.

E. Ferrero and P. Agarwal. Connecting genetics and gene expression data for target prioritisation and drug repositioning. *BioData Min.*, 11:7, 2018. ISSN 1756-0381. doi: 10.1186/s13040-018-0171-y.

F. Figueroa, F. Carrión, M. E. Martínez, S. Rivero, I. Mamani, and G. González. Effects of bromocriptine in patients with active rheumatoid arthritis. *Rev. Med. Chil.*, 126(1): 33–41, 1998. ISSN 0034-9887.

F. E. Figueroa, F. Carrión, M. E. Martínez, S. Rivero, and I. Mamani. Bromocriptine induces immunological changes related to disease parameters in rheumatoid arthritis. *Br. J. Rheumatol.*, 36(9):1022–1023, 1997. ISSN 0263-7103. doi: 10.1093/rheumatology/36.9.1022.

S. Finnegan, J. Robson, C. Scaife, C. McAllister, S. R. Pennington, D. S. Gibson, and M. E. Rooney. Synovial membrane protein expression differs between juvenile idiopathic arthritis subtypes in early disease. *Arthritis Res. Ther.*, 16(1):R8, 2014.

A. Flamier, J. El Hajjar, J. Adjaye, K. J. Fernandes, M. Abdouh, and G. Bernier. Modeling Late-Onset sporadic Alzheimer's disease through BMI1 deficiency. *Cell Rep.*, 23(9): 2653–2666, 2018. ISSN 2211-1247. doi: 10.1016/j.celrep.2018.04.097.

D. Flinkman, Y. Hong, P. S. Deshpande, T. L.-P. Laurén, S. Peltonen, V. Kaasinen, P. H. James, and E. T. Coffey. DIAGNOSIS OF PARKINSON'S DISEASE ON THE BASIS OF DECREASED OVERALL TRANSLATION, 2019. URL http://www.freepatentsonline.com/y2019/0390277.html.

B. M. Foxwell, C. Beadling, D. Guschin, I. Kerr, and D. Cantrell. Interleukin-7 can induce the activation of jak 1, jak 3 and STAT 5 proteins in murine T cells. *Eur. J. Immunol.*, 25 (11):3041–3046, 1995. ISSN 0014-2980. doi: 10.1002/eji.1830251109.

L. Fratiglioni, A. Ahlbom, M. Viitanen, and B. Winblad. Risk factors for late-onset Alzheimer's disease: a population-based, case-control study. *Ann. Neurol.*, 33(3): 258–266, 1993. ISSN 0364-5134. doi: 10.1002/ana.410330306.

S. Frede, C. Stockmann, P. Freitag, and J. Fandrey. Bacterial lipopolysaccharide induces HIF-1 activation in human monocytes via p44/42 MAPK and NF-kappaB. *Biochem. J*, 396(3):517–527, 2006.

E. L. Friesen, M. L. de Snoo, L. Rajendran, L. V. Kalia, and S. K. Kalia. Chaperone-Based therapies for disease modification in Parkinson's disease. *Parkinson's Disease*, 2017, 2017. ISSN 2090-8083, 2090-8083. doi: 10.1155/2017/5015307.

I. Gallego Romero, A. A. Pai, J. Tung, and Y. Gilad. RNA-seq: impact of RNA degradation on transcript quantification. *BMC Biol.*, 12:42, 2014.

P. Garcia-Esparcia, G. Sideris-Lampretsas, K. Hernandez-Ortega, O. Grau-Rivera, T. Sklaviadis, E. Gelpi, and I. Ferrer. Altered mechanisms of protein synthesis in frontal cortex in Alzheimer disease and a mouse model. *Am. J. Neurodegener. Dis.*, 6(2): 15–25, 2017. ISSN 2165-591X.

C. J. Garwood, A. M. Pooler, J. Atherton, D. P. Hanger, and W. Noble. Astrocytes are important mediators of Aβ-induced neurotoxicity and tau phosphorylation in primary culture. *Cell Death Dis.*, 2:e167, 2011. ISSN 2041-4889. doi: 10.1038/cddis.2011.50.

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

L. Gautier, L. Cope, B. M. Bolstad, and R. A. Irizarry. affy—analysis of affymetrix genechip data at the probe level. *Bioinformatics*, 20(3):307–315, 2004. ISSN 1367-4803. doi: 10.1093/bioinformatics/btg405.

Gepedia. Instances | GEPedia.org | gepedia. http://www.gepedia.org/instances.html, 2010. URL http://www.gepedia.org/instances.html. Accessed: 2017-10-22.

A. Gerschütz, H. Heinsen, E. Grünblatt, A. K. Wagner, J. Bartl, C. Meissner, A. J. Fallgatter, S. Al-Sarraj, C. Troakes, I. Ferrer, T. Arzberger, J. Deckert, P. Riederer, M. Fischer, T. Tatschner, and C. M. Monoranu. Neuron-specific alterations in signal transduction pathways associated with Alzheimer's disease. *J. Alzheimers. Dis.*, 40(1):135–142, 2014. ISSN 1387-2877, 1875-8908. doi: 10.3233/JAD-131280.

H. A. Ghofrani, I. H. Osterloh, and F. Grimminger. Sildenafil: from angina to erectile dysfunction to pulmonary hypertension and beyond. *Nat. Rev. Drug Discov.*, 5(8): 689–702, 2006.

G. Giancane, A. Consolaro, S. Lanni, S. Davì, B. Schiappapietra, and A. Ravelli. Juvenile idiopathic arthritis: Diagnosis and treatment. *Rheumatol Ther*, 3(2):187–207, 2016.

D. S. Gibson, J. Qiu, E. A. Mendoza, K. Barker, M. E. Rooney, and J. LaBaer. Circulating and synovial antibody profiling of juvenile arthritis patients by nucleic acid programmable protein arrays. *Arthritis Res. Ther.*, 14(2):R77, 2012.

B. E. Gilliam, M. R. Reed, A. K. Chauhan, A. B. Dehlendorf, and T. L. Moore. Significance of complement components c1q and C4 bound to circulating immune complexes in juvenile idiopathic arthritis: support for classical complement pathway activation. *Clin. Exp. Rheumatol.*, 29(6):1049–1056, 2011. ISSN 0392-856X.

P. Gilon and J. C. Henquin. Mechanisms and physiological significance of the cholinergic control of pancreatic beta-cell function. *Endocr. Rev.*, 22(5):565–604, 2001. ISSN 0163-769X. doi: 10.1210/edrv.22.5.0440.

M. K. Gilson, T. Liu, M. Baitaluk, G. Nicola, L. Hwang, and J. Chong. BindingDB in 2015: A public database for medicinal chemistry, computational chemistry and systems pharmacology. *Nucleic Acids Res.*, 44(D1):D1045–53, 2016.

M. Girvan and M. E. J. Newman. Community structure in social and biological networks. *Proc. Natl. Acad. Sci. U. S. A.*, 99(12):7821–7826, 2002.

C. E. Gleason, N. M. Dowling, W. Wharton, J. E. Manson, V. M. Miller, C. S. Atwood, E. A. Brinton, M. I. Cedars, R. A. Lobo, G. R. Merriam, G. Neal-Perry, N. F. Santoro, H. S. Taylor, D. M. Black, M. J. Budoff, H. N. Hodis, F. Naftolin, S. M. Harman, and S. Asthana. Effects of hormone therapy on cognition and mood in recently postmenopausal women: Findings from the randomized, controlled KEEPS-Cognitive and affective study. *PLoS Med.*, 12(6):e1001833; discussion e1001833, 2015. ISSN 1549-1277, 1549-1676. doi: 10.1371/journal.pmed.1001833.

E. M. Glenn. Inhibition of adjuvant-induced polyarthritis with cytarabine. *Proc. Soc. Exp. Biol. Med.*, 129(3):860–865, 1968. ISSN 0037-9727. doi: 10.3181/00379727-129-33443.

E. M. Glenn, B. J. Bowman, N. A. Rohloff, and R. J. Seely. A major contributory cause of arthritis in adjuvant-inoculated rats: granulocytes. *Agents Actions*, 7(2):265–282, 1977. ISSN 0065-4299. doi: 10.1007/bf01969985.

B. S. Glicksberg, L. Li, W.-Y. Cheng, K. Shameer, J. Hakenberg, R. Castellanos, M. Ma, L. Shi, H. Shah, J. T. Dudley, and R. Chen. An integrative pipeline for multi-modal discovery of disease relationships. *Pac. Symp. Biocomput.*, pages 407–418, 2015.

H. S. Gns, S. Gr, M. Murahari, and M. Krishnamurthy. An update on drug repurposing: Re-written saga of the drug's fate. *Biomed. Pharmacother.*, 110:700–716, 2019. ISSN 0753-3322, 1950-6007. doi: 10.1016/j.biopha.2018.11.127.

A. Goate, M. C. Chartier-Harlin, M. Mullan, J. Brown, F. Crawford, L. Fidani, L. Giuffra, A. Haynes, N. Irving, and L. James. Segregation of a missense mutation in the amyloid precursor protein gene with familial Alzheimer's disease. *Nature*, 349(6311):704–706, 1991. ISSN 0028-0836. doi: 10.1038/349704a0.

M. Gómez Ravetti, O. A. Rosso, R. Berretta, and P. Moscato. Uncovering molecular biomarkers that correlate cognitive decline with the changes of hippocampus' gene expression profiles in Alzheimer's disease. *PLoS One*, 5(4):e10153, 2010. ISSN 1932-6203. doi: 10.1371/journal.pone.0010153.

L. Gong, R. P. Owen, W. Gor, R. B. Altman, and T. E. Klein. PharmGKB: An integrated resource of pharmacogenomic data and knowledge. In *Current Protocols in Bioinformatics*. John Wiley & Sons, Inc., 2002.

D. Gosselin, D. Skola, N. G. Coufal, I. R. Holtman, J. C. M. Schlachetzki, E. Sajti, B. N. Jaeger, C. O'Connor, C. Fitzpatrick, M. P. Pasillas, M. Pena, A. Adair, D. D. Gonda, M. L. Levy, R. M. Ransohoff, F. H. Gage, and C. K. Glass. An environment-dependent transcriptional network specifies human microglia identity. *Science*, 356(6344), 2017. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.aal3222.

A. Gottlieb, G. Y. Stein, E. Ruppin, and R. Sharan. PREDICT: a method for inferring novel drug indications with application to personalized medicine. *Mol. Syst. Biol.*, 7(1): 496, 2011.

N. R. Gough. Science's signal transduction knowledge environment: the connections maps database. *Ann. N. Y. Acad. Sci.*, 971:585–587, 2002.

S. Gourmaud, C. Paquet, J. Dumurgier, C. Pace, C. Bouras, F. Gray, J.-L. Laplanche, E. F. Meurs, F. Mouton-Liger, and J. Hugon. Increased levels of cerebrospinal fluid JNK3 associated with amyloid pathology: links to cognitive decline. *J. Psychiatry Neurosci.*, 40(3):151–161, 2015. ISSN 1180-4882, 1488-2434. doi: 10.1503/jpn.140062.

Z. Gu, R. Eils, and M. Schlesner. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics*, 2016.

R. Guha and J. H. van Drie. Structure- activity landscape index: Identifying and quantifying activity cliffs. *J. Chem. Inf. Model.*, 48(3):646–658, 2008.

A. Güneş, A. Ece, V. Şen, Ü. Uluca, F. Aktar, İ. Tan, S. Yel, and İ. Yolbaş. Correlation of mean platelet volume, neutrophil-to-lymphocyte ratio, and disease activity in children with juvenile ıdiopathic arthritis. *Int. J. Clin. Exp. Med.*, 8(7):11337–11341, 2015. ISSN 1940-5901.

E. Guney. Reproducible drug repurposing: When similarity does not suffice. *Pac. Symp. Biocomput.*, 22:132–143, 2016.

C. Guo, L.-J. Hao, Z.-H. Yang, R. Chai, S. Zhang, Y. Gu, H.-L. Gao, M.-L. Zhong, T. Wang, J.-Y. Li, and Z.-Y. Wang. Deferoxamine-mediated up-regulation of HIF-1α prevents dopaminergic neuronal death via the activation of MAPK family proteins in MPTP-treated mice. *Exp. Neurol.*, 280:13–23, 2016. ISSN 0014-4886, 1090-2430. doi: 10.1016/j.expneurol.2016.03.016.

Z. Guo, T. Zhang, X. Li, Q. Wang, J. Xu, H. Yu, J. Zhu, H. Wang, C. Wang, E. J. Topol, Q. Wang, and S. Rao. Towards precise classification of cancers based on robust gene functional expression profiles. *BMC Bioinformatics*, 6:58, 2005.

C. Haass, C. A. Lemere, A. Capell, M. Citron, P. Seubert, D. Schenk, L. Lannfelt, and D. J. Selkoe. The swedish mutation causes early-onset Alzheimer's disease by beta-secretase cleavage within the secretory pathway. *Nat. Med.*, 1(12):1291–1296, 1995. ISSN 1078-8956. doi: 10.1038/nm1295-1291.

M. Hadchouel, A. M. Prieur, and C. Griscelli. Acute hemorrhagic, hepatic, and neurologic manifestations in juvenile rheumatoid arthritis: possible relationship to drugs or infection. *J. Pediatr.*, 106(4):561–566, 1985. ISSN 0022-3476. doi: 10.1016/s0022-3476(85) 80072-x.

A. Hamosh, A. F. Scott, J. Amberger, C. Bocchini, D. Valle, and V. A. McKusick. Online mendelian inheritance in man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.*, 30(1):52–55, 2002.

H. Hanyu, K. Hirao, S. Shimizu, T. Sato, A. Kiuchi, and T. Iwamoto. Nilvadipine prevents cognitive decline of patients with mild cognitive impairment. *Int. J. Geriatr. Psychiatry*, 22(12):1264–1266, 2007. ISSN 0885-6230. doi: 10.1002/gps.1851.

S. Hänzelmann, R. Castelo, and J. Guinney. GSVA: gene set variation analysis for microarray and RNA-seq data. *BMC Bioinformatics*, 14:7, 2013. ISSN 1471-2105. doi: 10.1186/1471-2105-14-7.

J. Hardy and D. Allsop. Amyloid deposition as the central event in the aetiology of Alzheimer's disease. *Trends Pharmacol. Sci.*, 12(10):383–388, 1991. ISSN 0165-6147. doi: 10.1016/0165-6147(91)90609-v.

J. G. Harris, E. A. Kessler, and J. W. Verbsky. Update on the treatment of juvenile idiopathic arthritis. *Curr. Allergy Asthma Rep.*, 13(4):337–346, 2013.

P. L. Harris, X. Zhu, C. Pamies, C. A. Rottkamp, H. A. Ghanbari, A. McShea, Y. Feng, D. K. Ferris, and M. A. Smith. Neuronal polo-like kinase in Alzheimer disease indicates cell cycle changes. *Neurobiol. Aging*, 21(6):837–841, 2000. ISSN 0197-4580. doi: 10.1016/s0197-4580(00)00218-9.

R. S. Harris and J. P. Dudley. APOBECs and virus restriction. *Virology*, 479-480:131–145, 2015.

I. F. Harrison, H. K. Anis, and D. T. Dexter. Associated degeneration of ventral tegmental area dopaminergic neurons in the rat nigrostriatal lactacystin model of parkinsonism and their neuroprotection by valproate. *Neurosci. Lett.*, 614:16–23, 2016. ISSN 0304-3940, 1872-7972. doi: 10.1016/j.neulet.2015.12.052.

I. F. Harrison, A. D. Smith, and D. T. Dexter. Pathological histone acetylation in Parkinson's disease: Neuroprotection and inhibition of microglial activation through SIRT 2 inhibition. *Neurosci. Lett.*, 666:48–57, 2018. ISSN 0304-3940, 1872-7972. doi: 10.1016/j.neulet.2017.12.037.

A. Hartmann, S. Hunot, P. P. Michel, M. P. Muriel, S. Vyas, B. A. Faucheux, A. Mouatt-Prigent, H. Turmel, A. Srinivasan, M. Ruberg, G. I. Evan, Y. Agid, and E. C. Hirsch. Caspase-3: A vulnerability factor and final effector in apoptotic death of dopaminergic neurons in Parkinson's disease. *Proc. Natl. Acad. Sci. U. S. A.*, 97(6):2875–2880, 2000. ISSN 0027-8424. doi: 10.1073/pnas.040556597.

L. H. Hartwell and M. B. Kastan. Cell cycle control and cancer. *Science*, 266(5192): 1821–1828, 1994.

G. W. Hassen, L. Kesner, A. Stracher, A. Shulman, E. Rockenstein, M. Mante, A. Adame, C. Overk, R. A. Rissman, and E. Masliah. Effects of novel calpain inhibitors in transgenic animal model of Parkinson's disease/dementia with lewy bodies. *Sci. Rep.*, 8(1):18083, 2018. ISSN 2045-2322. doi: 10.1038/s41598-018-35729-1.

E. E. Hatch, J. R. Palmer, L. Titus-Ernstoff, K. L. Noller, R. H. Kaufman, R. Mittendorf, S. J. Robboy, M. Hyer, C. M. Cowan, E. Adam, T. Colton, P. Hartge, and R. N. Hoover. Cancer risk in women exposed to diethylstilbestrol in utero. *JAMA*, 280(7):630–634, 1998. ISSN 0098-7484. doi: 10.1001/jama.280.7.630.

M. S. Hayden and S. Ghosh. Regulation of NF-κB by TNF family cytokines. *Semin. Immunol.*, 26(3):253–266, 2014.

S. J. Hebbring. The challenges, advantages and future of phenome-wide association studies. *Immunology*, 141(2):157–165, 2014.

L. E. Hebert, J. L. Bienias, N. T. Aggarwal, R. S. Wilson, D. A. Bennett, R. C. Shah, and D. A. Evans. Change in risk of Alzheimer disease over time. *Neurology*, 75(9):786–791, 2010. ISSN 0028-3878, 1526-632X. doi: 10.1212/WNL.0b013e3181f0754f.

B. Henderson, L. Bitensky, and J. Chayen. Glycolytic activity in human synovial lining cells in rheumatoid arthritis. *Ann. Rheum. Dis.*, 38(1):63–67, 1979. ISSN 0003-4967. doi: 10.1136/ard.38.1.63.

C. H. Hinze, N. Fall, S. Thornton, J. Q. Mo, B. J. Aronow, G. Layh-Schmitt, T. A. Griffin, S. D. Thompson, R. A. Colbert, D. N. Glass, M. G. Barnes, and A. A. Grom. Immature cell populations and an erythropoiesis gene-expression signature in systemic juvenile idiopathic arthritis: implications for pathogenesis. *Arthritis Res. Ther.*, 12(3):R123, 2010.

R. A. Hodos, B. A. Kidd, K. Shameer, B. P. Readhead, and J. T. Dudley. In silico methods for drug repurposing and pharmacology. *Wiley Interdiscip. Rev. Syst. Biol. Med.*, 8(3): 186–210, 2016.

A. P. Hollander, K. P. Corke, A. J. Freemont, and C. E. Lewis. Expression of hypoxia-inducible factor 1alpha by macrophages in the rheumatoid synovium: implications for targeting of therapeutic genes to the inflamed joint. *Arthritis Rheum.*, 44(7):1540–1544, 2001. ISSN 0004-3591. doi: 10.1002/1529-0131(200107)44:7<1540::AID-ART277>3.0.CO;2-7.

J. J. M. Hoozemans, E. S. van Haastert, P. Eikelenboom, R. A. I. de Vos, J. M. Rozemuller, and W. Scheper. Activation of the unfolded protein response in Parkinson's disease. *Biochem. Biophys. Res. Commun.*, 354(3):707–711, 2007. ISSN 0006-291X. doi: 10.1016/j.bbrc.2007.01.043.

A. Howard. Synthesis of deoxyribonucleic acid in normal and irradiated ceils and its relation to chromosome breakage. *Heredity Suppl*, 6:261–273, 1953.

G. Hu and P. Agarwal. Human disease-drug network based on genomic expression profiles. *PLoS One*, 4(8):e6536, 2009.

X.-T. Huang, Z.-M. Qian, X. He, Q. Gong, K.-C. Wu, L.-R. Jiang, L.-N. Lu, Z.-J. Zhu, H.-Y. Zhang, W.-H. Yung, and Y. Ke. Reducing iron in the brain: a novel pharmacologic mechanism of huperzine a in the treatment of Alzheimer's disease. *Neurobiol. Aging*, 35(5):1045–1054, 2014. ISSN 0197-4580, 1558-1497. doi: 10.1016/j.neurobiolaging.2013.11.004.

W. Huber, V. J. Carey, R. Gentleman, S. Anders, M. Carlson, B. S. Carvalho, H. C. Bravo, S. Davis, L. Gatto, T. Girke, R. Gottardo, F. Hahne, K. D. Hansen, R. A. Irizarry, M. Lawrence, M. I. Love, J. MacDonald, V. Obenchain, A. K. Ole's, H. Pag'es, A. Reyes, P. Shannon, G. K. Smyth, D. Tenenbaum, L. Waldron, and M. Morgan. Orchestrating high-throughput genomic analysis with Bioconductor. *Nature Methods*, 12(2):115–121, 2015. URL http://www.nature.com/nmeth/journal/v12/n2/full/nmeth.3252.html.

Y. H. Huh, G. Lee, K.-B. Lee, J.-T. Koh, J.-S. Chun, and J.-H. Ryu. HIF-2α-induced chemokines stimulate motility of fibroblast-like synoviocytes and chondrocytes into the cartilage-pannus interface in experimental rheumatoid arthritis mouse models. *Arthritis Res. Ther.*, 17:302, 2015.

L. Hunter and K. B. Cohen. Biomedical language processing: what's beyond PubMed? *Mol. Cell*, 21(5):589–594, 2006.

A. J. Hutt. The development of single-isomer molecules: why and how. *CNS Spectr.*, 7(4 Suppl 1):14–22, 2002.

S. Hwang. Comparison and evaluation of pathway-level aggregation methods of gene expression data. *BMC Genomics*, 13 Suppl 7(7):S26, 2012.

B. Imtiaz, M. Tuppurainen, T. Rikkonen, M. Kivipelto, H. Soininen, H. Kröger, and A.-M. Tolppanen. Postmenopausal hormone therapy and Alzheimer disease: A prospective cohort study. *Neurology*, 88(11):1062–1068, 2017. ISSN 0028-3878, 1526-632X. doi: 10.1212/WNL.0000000000003696.

I. H. Ismail, J.-P. Gagné, M.-C. Caron, D. McDonald, Z. Xu, J.-Y. Masson, G. G. Poirier, and M. J. Hendzel. CBX4-mediated SUMO modification regulates BMI1 recruitment at sites of DNA damage. *Nucleic Acids Res.*, 40(12):5497–5510, 2012. ISSN 0305-1048, 1362-4962. doi: 10.1093/nar/gks222.

E. Jadamba and M. Shin. A systematic framework for drug repositioning from integrated omics and drug phenotype profiles using Pathway-Drug network. *Biomed Res. Int.*, 2016:7147039, 2016. ISSN 2314-6133. doi: 10.1155/2016/7147039.

K. J. Janczura, C.-H. Volmar, G. C. Sartor, S. J. Rao, N. R. Ricciardi, G. Lambert, S. P. Brothers, and C. Wahlestedt. Inhibition of HDAC3 reverses Alzheimer's disease-related pathologies in vitro and in the 3xTg-AD mouse model. *Proc. Natl. Acad. Sci. U. S. A.*, 115(47):E11148–E11157, 2018. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1805436115.

S. Janelidze, D. Lindqvist, V. Francardo, S. Hall, H. Zetterberg, K. Blennow, C. H. Adler, T. G. Beach, G. E. Serrano, D. van Westen, E. Londos, M. A. Cenci, and O. Hansson. Increased CSF biomarkers of angiogenesis in Parkinson disease. *Neurology*, 85(21):1834–1842, 2015. ISSN 0028-3878, 1526-632X. doi: 10.1212/WNL.0000000000002151.

J. Jankovic. Parkinson's disease: clinical features and diagnosis. *J. Neurol. Neurosurg. Psychiatry*, 79(4):368–376, 2008. ISSN 0022-3050, 1468-330X. doi: 10.1136/jnnp.2007.131045.

J. N. Jarvis, T. Pousak, M. Krenz, M. Iobidze, and H. Taylor. Complement activation and immune complexes in juvenile rheumatoid arthritis. *J. Rheumatol.*, 20(1):114–117, 1993. ISSN 0315-162X.

E. Jednacz and L. Rutkowska-Sak. Atherosclerosis in juvenile idiopathic arthritis. *Mediators Inflamm.*, 2012:714732, 2012. ISSN 0962-9351, 1466-1861. doi: 10.1155/2012/714732.

W. Jian, X. Wei, L. Chen, Z. Wang, Y. Sun, S. Zhu, H. Lou, S. Yan, X. Li, J. Zhou, and B. Zhang. Inhibition of HDAC6 increases acetylation of peroxiredoxin1/2 and ameliorates 6-OHDA induced dopaminergic injury. *Neurosci. Lett.*, 658:114–120, 2017. ISSN 0304-3940, 1872-7972. doi: 10.1016/j.neulet.2017.08.029.

W. Jin, L.-F. Qu, P. Min, S. Chen, H. Li, H. Lu, and Y.-T. Hou. Identification of genes responsive to apoptosis in HL-60 cells. *Acta Pharmacol. Sin.*, 25(3):319–326, 2004. ISSN 1671-4083.

R. B. Joachim, G. M. Altschuler, J. N. Hutchinson, H. R. Wong, W. A. Hide, and L. Kobzik. The relative resistance of children to sepsis mortality: from pathways to drug candidates. *Mol. Syst. Biol.*, 14(5):e7998, 2018.

A. L. Johnstone, G. W. Reierson, R. P. Smith, J. L. Goldberg, V. P. Lemmon, and J. L. Bixby. A chemical genetic approach identifies piperazine antipsychotics as promoters of CNS neurite growth on inhibitory substrates. *Mol. Cell. Neurosci.*, 50(2):125–135, 2012.

K. Kandasamy, S. S. Mohan, R. Raju, S. Keerthikumar, G. S. S. Kumar, A. K. Venugopal, D. Telikicherla, J. D. Navarro, S. Mathivanan, C. Pecquet, S. K. Gollapudi, S. G. Tattikota, S. Mohan, H. Padhukasahasram, Y. Subbannayya, R. Goel, H. K. C. Jacob, J. Zhong, R. Sekhar, V. Nanjappa, L. Balakrishnan, R. Subbaiah, Y. L. Ramachandra, B. A. Rahiman, T. S. K. Prasad, J.-X. Lin, J. C. D. Houtman, S. Desiderio, J.-C. Renauld, S. N. Constantinescu, O. Ohara, T. Hirano, M. Kubo, S. Singh, P. Khatri, S. Draghici, G. D. Bader, C. Sander, W. J. Leonard, and A. Pandey. NetPath: a public resource of curated signal transduction pathways. *Genome Biol.*, 11(1):R3, 2010.

M. Kanehisa and S. Goto. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, 28(1):27–30, 2000.

M. Kanehisa, S. Goto, Y. Sato, M. Kawashima, M. Furumichi, and M. Tanabe. Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res.*, 42(Database issue):D199–205, 2014.

Y. Kang, S. Tiziani, G. Park, M. Kaul, and G. Paternostro. Cellular protection using flt3 and PI3Kα inhibitors demonstrates multiple mechanisms of oxidative glutamate toxicity. *Nat. Commun.*, 5:3672, 2014. ISSN 2041-1723. doi: 10.1038/ncomms4672.

A. Kauffmann, R. Gentleman, and W. Huber. arrayqualitymetrics–a bioconductor package for quality assessment of microarray data. *Bioinformatics*, 25(3):415–6, 2009.

T. M. Kauppinen, S. W. Suh, Y. Higashi, A. E. Berman, C. Escartin, S. J. Won, C. Wang, S.-H. Cho, L. Gan, and R. A. Swanson. Poly(ADP-ribose)polymerase-1 modulates microglial responses to amyloid β. *J. Neuroinflammation*, 8:152, 2011. ISSN 1742-2094. doi: 10.1186/1742-2094-8-152.

A. B. Keenan, S. L. Jenkins, K. M. Jagodnik, S. Koplev, E. He, D. Torre, Z. Wang, A. B. Dohlman, M. C. Silverstein, A. Lachmann, M. V. Kuleshov, A. Ma'ayan, V. Stathias, R. Terryn, D. Cooper, M. Forlin, A. Koleti, D. Vidovic, C. Chung, S. C. Schurer, J. Vasiliauskas, M. Pilarczyk, B. Shamsaei, M. Fazel, Y. Ren, W. Niu, N. A. Clark, S. White, N. Mahi, L. Zhang, M. Kouril, J. F. Reichard, S. Sivaganesan, M. Medvedovic, J. Meller, R. J. Koch, M. R. Birtwistle, R. Iyengar, E. A. Sobie, E. U. Azeloglu, J. Kaye, J. Osterloh, K. Haston, J. Kalra, S. Finkbiener, J. Li, P. Milani, M. Adam, R. Escalante-Chong, K. Sachs, A. Lenail, D. Ramamoorthy, E. Fraenkel, G. Daigle, U. Hussain, A. Coye, J. Rothstein, D. Sareen, L. Ornelas, M. Banuelos, B. Mandefro,

R. Ho, C. N. Svendsen, R. G. Lim, J. Stocksdale, M. S. Casale, T. G. Thompson, J. Wu, L. M. Thompson, V. Dardov, V. Venkatraman, A. Matlock, J. E. van Eyk, J. D. Jaffe, M. Papanastasiou, A. Subramanian, T. R. Golub, S. D. Erickson, M. Fallahi-Sichani, M. Hafner, N. S. Gray, J.-R. Lin, C. E. Mills, J. L. Muhlich, M. Niepel, C. E. Shamu, E. H. Williams, D. Wrobel, P. K. Sorger, L. M. Heiser, J. W. Gray, J. E. Korkola, G. B. Mills, M. LaBarge, H. S. Feiler, M. A. Dane, E. Bucher, M. Nederlof, D. Sudar, S. Gross, D. F. Kilburn, R. Smith, K. Devlin, R. Margolis, L. Derr, A. Lee, and A. Pillai. The library of integrated Network-Based cellular signatures NIH program: System-Level cataloging of human cells response to perturbations. *Cell systems*, 6(1):13–24, 2018.

M. J. Keiser, V. Setola, J. J. Irwin, C. Laggner, A. I. Abbas, S. J. Hufeisen, N. H. Jensen, M. B. Kuijer, R. C. Matos, T. B. Tran, R. Whaley, R. A. Glennon, J. Hert, K. L. H. Thomas, D. D. Edwards, B. K. Shoichet, and B. L. Roth. Predicting new molecular targets for known drugs. *Nature*, 462(7270):175–181, 2009.

T. Kelder, M. P. van Iersel, K. Hanspers, M. Kutmon, B. R. Conklin, C. T. Evelo, and A. R. Pico. WikiPathways: building research communities on biological pathways. *Nucleic Acids Res.*, 40(Database issue):D1301–7, 2012.

N. Kemppainen, M. Laine, M. P. Laakso, V. Kaasinen, K. Någren, T. Vahlberg, T. Kurki, and J. O. Rinne. Hippocampal dopamine D2 receptors correlate with memory functions in Alzheimer's disease. *Eur. J. Neurosci.*, 18(1):149–154, 2003. ISSN 0953-816X. doi: 10.1046/j.1460-9568.2003.02716.x.

R. Khare, J. Li, and Z. Lu. Toward creating a gold standard of drug indications from FDA drug labels. In *2013 IEEE International Conference on Healthcare Informatics*, pages 30–35, 2013.

R. Killick, E. M. Ribe, R. Al-Shawi, B. Malik, C. Hooper, C. Fernandes, R. Dobson, P. M. Nolan, A. Lourdusamy, S. Furney, K. Lin, G. Breen, R. Wroe, A. W. M. To, K. Leroy, M. Causevic, A. Usardi, M. Robinson, W. Noble, R. Williamson, K. Lunnon, S. Kellie, C. H. Reynolds, C. Bazenet, A. Hodges, J.-P. Brion, J. Stephenson, J. P. Simons, and S. Lovestone. Clusterin regulates β-amyloid toxicity via dickkopf-1-driven induction of the wnt-PCP-JNK pathway. *Mol. Psychiatry*, 19(1):88–98, 2014. ISSN 1359-4184, 1476-5578. doi: 10.1038/mp.2012.163.

S. Kim, P. A. Thiessen, E. E. Bolton, J. Chen, G. Fu, A. Gindulyte, L. Han, J. He, S. He, B. A. Shoemaker, J. Wang, B. Yu, J. Zhang, and S. H. Bryant. PubChem substance and compound databases. *Nucleic Acids Res.*, 44(D1):D1202–13, 2016. ISSN 0305-1048, 1362-4962. doi: 10.1093/nar/gkv951.

S. C. Kim, R. J. Glynn, E. Giovannucci, S. Hernández-Díaz, J. Liu, S. Feldman, E. W. Karlson, S. Schneeweiss, and D. H. Solomon. Risk of high-grade cervical dysplasia and cervical cancer in women with systemic inflammatory diseases: a population-based cohort study. *Ann. Rheum. Dis.*, 74(7):1360–1367, 2015. ISSN 0003-4967, 1468-2060. doi: 10.1136/annrheumdis-2013-204993.

S. L. Kinnings, N. Liu, N. Buchmeier, P. J. Tonge, L. Xie, and P. E. Bourne. Drug discovery using chemical systems biology: repositioning the safe medicine comtan to treat multi-drug and extensively drug resistant tuberculosis. *PLoS Comput. Biol.*, 5(7): e1000423, 2009.

M. Kinoshita, C. M. Field, M. L. Coughlin, A. F. Straight, and T. J. Mitchison. Self- and actin-templated assembly of mammalian septins. *Dev. Cell*, 3(6):791–802, 2002.

M. Kissa and G. Tsatsaronis. A benchmark dataset for computational drug repositioning. *BMDJ*, 1(2):10–12, 2015.

T. Kitamura, Y. Kabuyama, A. Kamataki, M. K. Homma, H. Kobayashi, S. Aota, S.-I. Kikuchi, and Y. Homma. Enhancement of lymphocyte migration and cytokine production by ephrinb1 system in rheumatoid arthritis. *Am. J. Physiol. Cell Physiol.*, 294(1):C189–96, 2008. ISSN 0363-6143. doi: 10.1152/ajpcell.00314.2007.

I. Kjos, K. Vestre, N. A. Guadagno, M. Borg Distefano, and C. Progida. Rab and arf proteins at the crossroad between membrane transport and cytoskeleton dynamics. *Biochim. Biophys. Acta Mol. Cell Res.*, 1865(10):1397–1409, 2018.

N. Knowlton, K. Jiang, M. B. Frank, A. Aggarwal, C. Wallace, R. McKee, B. Chaser, C. Tung, L. Smith, Y. Chen, J. Osban, K. O'Neil, M. Centola, J. L. McGhee, and J. N. Jarvis. The meaning of clinical remission in polyarticular juvenile idiopathic arthritis: gene expression profiling in peripheral blood mononuclear cells identifies distinct disease states. *Arthritis Rheum.*, 60(3):892–900, 2009.

M. Kobayashi, K. Nishikawa, T. Suzuki, and M. Yamamoto. The homeobox protein six3 interacts with the groucho corepressor and acts as a transcriptional repressor in eye and forebrain formation. *Dev. Biol.*, 232(2):315–326, 2001. ISSN 0012-1606. doi: 10.1006/dbio.2001.0185.

M. A. Koerper, D. A. Stempel, and P. R. Dallman. Anemia in patients with juvenile rheumatoid arthritis. *J. Pediatr.*, 92(6):930–933, 1978. ISSN 0022-3476. doi: 10.1016/s0022-3476(78)80363-1.

V. C. Kok, J.-T. Horng, J.-L. Huang, K.-W. Yeh, J.-J. Gau, C.-W. Chang, and L.-Z. Zhuang. Population-based cohort study on the risk of malignancy in east asian children with juvenile idiopathic arthritis. *BMC Cancer*, 14:634, 2014. ISSN 1471-2407. doi: 10.1186/1471-2407-14-634.

K. Koler, S. L. Morgan, D. R. Jones, D. Wang, and W. A. Hide. KATdb: a graph theoretic approach to unification of drug names. Manuscript in Preparation, 2020.

A. Koleti, R. Terryn, V. Stathias, C. Chung, D. J. Cooper, J. P. Turner, D. Vidovic, M. Forlin, T. T. Kelley, A. D'Urso, B. K. Allen, D. Torre, K. M. Jagodnik, L. Wang, S. L. Jenkins, C. Mader, W. Niu, M. Fazel, N. Mahi, M. Pilarczyk, N. Clark, B. Shamsaei, J. Meller, J. Vasiliauskas, J. Reichard, M. Medvedovic, A. Ma'ayan, A. Pillai, and S. C. Schürer. Data portal for the library of integrated network-based cellular signatures (LINCS) program: integrated access to diverse large-scale cellular perturbation response data. *Nucleic Acids Res.*, 46(D1):D558–D566, 2018.

T. Kondo, K. Imamura, M. Funayama, K. Tsukita, M. Miyake, A. Ohta, K. Woltjen, M. Nakagawa, T. Asada, T. Arai, S. Kawakatsu, Y. Izumi, R. Kaji, N. Iwata, and H. Inoue. iPSC-Based compound screening and in vitro trials identify a synergistic anti-amyloid β combination for Alzheimer's disease. *Cell Rep.*, 21(8):2304–2312, 2017. ISSN 2211-1247. doi: 10.1016/j.celrep.2017.10.109.

G. E. Kuehl, J. W. Lampe, J. D. Potter, and J. Bigler. Glucuronidation of nonsteroidal anti-inflammatory drugs: identifying the enzymes responsible in human liver microsomes. *Drug Metab. Dispos.*, 33(7):1027–1035, 2005.

S. D. Kunkel, M. Suneja, S. M. Ebert, K. S. Bongers, D. K. Fox, S. E. Malmberg, F. Alipour, R. K. Shields, and C. M. Adams. mRNA expression signatures of human skeletal muscle atrophy identify a natural compound that increases muscle mass. *Cell Metab.*, 13(6): 627–638, 2011.

M. A. Kurian, P. Gissen, M. Smith, S. Heales, Jr, and P. T. Clayton. The monoamine neurotransmitter disorders: an expanding range of neurological syndromes. *Lancet Neurol.*, 10(8):721–733, 2011. ISSN 1474-4422, 1474-4465. doi: 10.1016/S1474-4422(11) 70141-7.

N. Kustrimovic, C. Comi, L. Magistrelli, E. Rasini, M. Legnaro, R. Bombelli, I. Aleksic, F. Blandini, B. Minafra, G. Riboldazzi, A. Sturchio, M. Mauri, G. Bono, F. Marino, and M. Cosentino. Parkinson's disease patients have a complex phenotypic and functional th1 bias: cross-sectional studies of CD4+ Th1/Th2/T17 and treg in drug-naïve and drug-treated patients. *J. Neuroinflammation*, 15(1):205, 2018. ISSN 1742-2094. doi: 10.1186/s12974-018-1248-8.

N. Kutukculer, S. Caglayan, and F. Aydogdu. Study of pro-inflammatory (TNF-alpha, IL-1alpha, IL-6) and t-cell-derived (IL-2, IL-4) cytokines in plasma and synovial fluid of patients with juvenile chronic arthritis: correlations with clinical and laboratory parameters. *Clin. Rheumatol.*, 17(4):288–292, 1998.

S. S. Kwak, K. J. Washicosky, E. Brand, D. von Maydell, J. Aronson, S. Kim, D. E. Capen, M. Cetinbas, R. Sadreyev, S. Ning, E. Bylykbashi, W. Xia, S. L. Wagner, S. H. Choi, R. E. Tanzi, and D. Y. Kim. Amyloid-β42/40 ratio drives tau pathology in 3D human neural cell culture models of Alzheimer's disease. *Nat. Commun.*, 11(1):1377, 2020. ISSN 2041-1723. doi: 10.1038/s41467-020-15120-3.

O. V. Lagutin, C. C. Zhu, D. Kobayashi, J. Topczewski, K. Shimamura, L. Puelles, H. R. C. Russell, P. J. McKinnon, L. Solnica-Krezel, and G. Oliver. Six3 repression of wnt signaling in the anterior neuroectoderm is essential for vertebrate forebrain development. *Genes Dev.*, 17(3):368–379, 2003. ISSN 0890-9369. doi: 10.1101/gad.1059403.

S. Lal, T. L. Sourkes, K. Missala, and G. Belendiuk. Effects of aporphine and emetine alkaloids on central dopaminergic mechanisms in rats. *Eur. J. Pharmacol.*, 20(1):71–79, 1972. ISSN 0014-2999. doi: 10.1016/0014-2999(72)90217-8.

J. Lamb, E. D. Crawford, D. Peck, J. W. Modell, I. C. Blat, M. J. Wrobel, J. Lerner, J.-P. Brunet, A. Subramanian, K. N. Ross, M. Reich, H. Hieronymus, G. Wei, S. A. Armstrong, S. J. Haggarty, P. A. Clemons, R. Wei, S. A. Carr, E. S. Lander, and T. R. Golub. The connectivity map: using gene-expression signatures to connect small molecules, genes, and disease. *Science*, 313(5795):1929–1935, 2006.

H. Larbalestier, M. Keatinge, L. Trollope, E. White, S. Gowda, W. Wei, K. Koler, S. Semenova, N. Rimmer, S. Sweeney, J. Mazzolini, D. Sieger, W. A. Hide, R. Macdonald, J. McDearmid, P. Panula, and O. Bandmann. Tyrosine hydroxylase depletion and inflammatory dysregulation in a zebrafish gch1$^{-/-}$ Parkinson's disease model. Manuscript in Preparation, 2020.

M. Latapy, C. Magnien, and N. D. Vecchio. Basic notions for the analysis of large two-mode networks. *Soc. Networks*, 30(1):31–48, 2008.

C. Lauritzen and J. W. W. Studd. *Current Management of the Menopause*. CRC Press, 2005.

V. Law, C. Knox, Y. Djoumbou, T. Jewison, A. C. Guo, Y. Liu, A. Maciejewski, D. Arndt, M. Wilson, V. Neveu, A. Tang, G. Gabriel, C. Ly, S. Adamjee, Z. T. Dame, B. Han, Y. Zhou, and D. S. Wishart. DrugBank 4.0: shedding new light on drug metabolism. *Nucleic Acids Res.*, 42(Database issue):D1091–7, 2014.

H.-J. Lee, E.-J. Bae, and S.-J. Lee. Extracellular α–synuclein-a novel and crucial factor in lewy body diseases. *Nat. Rev. Neurol.*, 10(2):92–98, 2014a. ISSN 1759-4758, 1759-4766. doi: 10.1038/nrneurol.2013.275.

Y.-C. Lee, C.-H. Lin, R.-M. Wu, J.-W. Lin, C.-H. Chang, and M.-S. Lai. Antihypertensive agents and risk of Parkinson's disease: a nationwide cohort study. *PLoS One*, 9(6): e98961, 2014b. ISSN 1932-6203. doi: 10.1371/journal.pone.0098961.

L. Lepore, M. Pennesi, S. Saletta, S. Perticarari, G. Presani, and M. Prodan. Study of IL-2, IL-6, TNF alpha, IFN gamma and beta in the serum and synovial fluid of patients with juvenile chronic arthritis. *Clin. Exp. Rheumatol.*, 12(5):561–565, 1994.

J. Lever, M. Krzywinski, and N. Altman. Points of significance: Model selection and overfitting. *Nat. Methods*, 2016.

G. Levkowitz, H. Waterman, S. A. Ettenberg, M. Katz, A. Y. Tsygankov, I. Alroy, S. Lavi, K. Iwai, Y. Reiss, A. Ciechanover, S. Lipkowitz, and Y. Yarden. Ubiquitin ligase activity and tyrosine phosphorylation underlie suppression of growth factor signaling by c-Cbl/Sli-1. *Mol. Cell*, 4(6):1029–1040, 1999. ISSN 1097-2765. doi: 10.1016/s1097-2765(00)80231-2.

J. Li and Z. Lu. A new method for computational drug repositioning using drug pairwise similarity. In *2012 IEEE International Conference on Bioinformatics and Biomedicine*, volume 2012, pages 1–4. IEEE, 2012a.

J. Li and Z. Lu. A network approach for computational drug repositioning. In *2012 IEEE Second International Conference on Healthcare Informatics, Imaging and Systems Biology*, pages 83–83. IEEE, 2012b.

J. Li, S. Zheng, B. Chen, A. J. Butte, S. J. Swamidass, and Z. Lu. A survey of current trends in computational drug repositioning. *Brief. Bioinform.*, 17(1):2–12, 2016.

S. W. Li, T.-S. Lin, S. Minteer, and W. J. Burke. 3,4-dihydroxyphenylacetaldehyde and hydrogen peroxide generate a hydroxyl radical: possible role in Parkinson's disease pathogenesis. *Molecular Brain Research*, 93(1):1–7, 2001. ISSN 0169-328X. doi: 10.1016/S0169-328X(01)00120-6.

W. Li and D. Schuurmans. Modular community detection in networks. *Twenty-Second International Joint Conference on Artificial Intelligence*, 2011.

Y. H. Li, C. Y. Yu, X. X. Li, P. Zhang, J. Tang, Q. Yang, T. Fu, X. Zhang, X. Cui, G. Tu, Y. Zhang, S. Li, F. Yang, Q. Sun, C. Qin, X. Zeng, Z. Chen, Y. Z. Chen, and F. Zhu. Therapeutic target database update 2018: enriched resource for facilitating bench-to-clinic research of targeted therapeutics. *Nucleic Acids Res.*, 46(D1):D1121–D1127, 2018.

S. F. Lichtenthaler, R. Wang, H. Grimm, S. N. Uljon, C. L. Masters, and K. Beyreuther. Mechanism of the cleavage specificity of Alzheimer's disease gamma-secretase identified by phenylalanine-scanning mutagenesis of the transmembrane domain of the amyloid precursor protein. *Proc. Natl. Acad. Sci. U. S. A.*, 96(6):3053–3058, 1999. ISSN 0027-8424. doi: 10.1073/pnas.96.6.3053.

A. N. Lieberman and M. Goldstein. Bromocriptine in Parkinson disease. *Pharmacol. Rev.*, 37(2):217–227, 1985. ISSN 0031-6997.

J. Liu, Y. Zhou, Y. Wang, H. Fong, T. M. Murray, and J. Zhang. Identification of proteins involved in microglial endocytosis of alpha-synuclein. *J. Proteome Res.*, 6(9):3614–3627, 2007. ISSN 1535-3893. doi: 10.1021/pr0701512.

X. Liu, T. Murata, G. S. of Information Science, and ngineering, Tokyo Institute of Technology, 2-12-1 Ookayama, Meguro, Tokyo 152-8552, Japan. An efficient algorithm for optimizing bipartite modularity in bipartite networks. *J. Adv. Comput. Intell. Intell. Inform.*, 14(4):408–415, 2010.

Z. Liu, H. Fang, K. Reagan, X. Xu, D. L. Mendrick, W. Slikker Jr, and W. Tong. In silico drug repositioning: what we need to know. *Drug Discov. Today*, 18(3-4):110–115, 2013.

O. Lordan and M. Albareda-Sambola. Exact calculation of network robustness. *Reliab. Eng. Syst. Saf.*, 183:276–280, 2019.

S. Love, R. Barber, and G. K. Wilcock. Increased poly(ADP-ribosyl)ation of nuclear proteins in Alzheimer's disease. *Brain*, 122 (Pt 2):247–253, 1999. ISSN 0006-8950. doi: 10.1093/brain/122.2.247.

L. Lu and M. Zhang. Edge betweenness centrality. In W. Dubitzky, O. Wolkenhauer, K.-H. Cho, and H. Yokota, editors, *Encyclopedia of Systems Biology*, pages 647–648. Springer New York, New York, NY, 2013.

C. Ma, J. Chen, and P. Li. Geldanamycin induces apoptosis and inhibits inflammation in fibroblast-like synoviocytes isolated from rheumatoid arthritis patients. *J. Cell. Biochem.*, 120(9):16254–16263, 2019. ISSN 0730-2312, 1097-4644. doi: 10.1002/jcb.28906.

N. Maeno, S. Takei, H. Imanaka, I. Takasaki, I. Kitajima, I. Maruyama, K. Matsuo, and K. Miyata. Increased circulating vascular endothelial growth factor is correlated with disease activity in polyarticular juvenile rheumatoid arthritis. *J. Rheumatol.*, 26(10): 2244–2248, 1999.

Y. Mano, T. Usui, and H. Kamimura. Inhibitory potential of nonsteroidal anti-inflammatory drugs on UDP-glucuronosyltransferase 2B7 in human liver microsomes. *Eur. J. Clin. Pharmacol.*, 63(2):211–216, 2007.

C.-J. Mao, C.-K. Zhong, Y. Yang, Y.-P. Yang, F. Wang, J. Chen, J.-R. Zhang, H.-J. Zhang, H. Jin, L.-L. Xu, J.-Y. Huang, and C.-F. Liu. Serum sodium and chloride are inversely associated with dyskinesia in Parkinson's disease patients. *Brain Behav.*, 7(12):e00867, 2017. ISSN 2162-3279. doi: 10.1002/brb3.867.

Q. Marlier, F. Jibassia, S. Verteneuil, J. Linden, P. Kaldis, L. Meijer, L. Nguyen, R. Vandenbosch, and B. Malgrange. Genetic and pharmacological inhibition of cdk1 provides neuroprotection towards ischemic neuronal death. *Cell Death Discov*, 4:43, 2018. ISSN 2058-7716. doi: 10.1038/s41420-018-0044-7.

D. Masini, A. Bonito-Oliva, M. Bertho, and G. Fisone. Inhibition of mTORC1 signaling reverts cognitive and affective deficits in a mouse model of Parkinson's disease. *Front. Neurol.*, 9:208, 2018. ISSN 1664-2295. doi: 10.3389/fneur.2018.00208.

K. Matsuo, Y. Xiang, H. Nakamura, K. Masuko, K. Yudoh, K. Noyori, K. Nishioka, T. Saito, and T. Kato. Identification of novel citrullinated autoantigens of synovium in rheumatoid arthritis using a proteomic approach. *Arthritis Res. Ther.*, 8(6):R175, 2006.

L. Matthews, G. Gopinath, M. Gillespie, M. Caudy, D. Croft, B. de Bono, P. Garapati, J. Hemish, H. Hermjakob, B. Jassal, A. Kanapin, S. Lewis, S. Mahajan, B. May, E. Schmidt, I. Vastrik, G. Wu, E. Birney, L. Stein, and P. D'Eustachio. Reactome knowledgebase of human biological pathways and processes. *Nucleic Acids Res.*, 37 (Database issue):D619–22, 2009.

R. Mattsson, A. Mattsson, R. Holmdahl, A. Whyte, and G. A. Rook. Maintained pregnancy levels of oestrogen afford complete protection from post-partum exacerbation of collagen-induced arthritis. *Clin. Exp. Immunol.*, 85(1):41–47, 1991. ISSN 0009-9104. doi: 10.1111/j.1365-2249.1991.tb05679.x.

M. H. McAdoo, A. M. Dannenberg, Jr, C. J. Hayes, S. P. James, and J. H. Sanner. Inhibition of cathepsin d-type proteinase of macrophages by pepstatin, a specific pepsin inhibitor, and other substances. *Infect. Immun.*, 7(4):655–665, 1973. ISSN 0019-9567.

M. N. McCall, B. M. Bolstad, and R. A. Irizarry. Frozen robust multiarray analysis (fRMA). *Biostatistics*, 11(2):242–253, 2010.

M. N. McCall, P. N. Murakami, M. Lukk, W. Huber, and R. A. Irizarry. Assessing affymetrix GeneChip microarray quality. *BMC Bioinformatics*, 12:137, 2011a. ISSN 1471-2105. doi: 10.1186/1471-2105-12-137.

M. N. McCall, P. N. Murakami, M. Lukk, W. Huber, and R. A. Irizarry. Assessing affymetrix genechip microarray quality. *BMC Bioinformatics*, 12(1):137, 2011b.

M. N. McCall, H. A. Jaffee, S. J. Zelisko, N. Sinha, G. Hooiveld, R. A. Irizarry, and M. J. Zilliox. The gene expression barcode 3.0: improved data processing and mining tools. *Nucleic Acids Res.*, 42(Database issue):D938–43, 2014.

J. McConathy and M. J. Owens. Stereochemistry in drug action. *Prim. Care Companion J. Clin. Psychiatry*, 5(2):70–73, 2003.

A. D. McNaught, A. Wilkinson, and S. J. Chalk. *The IUPAC Compendium of Chemical Terminology (the "Gold Book").* IUPAC, 2nd edition edition, 1997.

N. E. Mencacci, I. U. Isaias, M. M. Reich, C. Ganos, V. Plagnol, J. M. Polke, J. Bras, J. Hersheson, M. Stamelou, A. M. Pittman, A. J. Noyce, K. Y. Mok, T. Opladen, E. Kunstmann, S. Hodecker, A. Münchau, J. Volkmann, S. Samnick, K. Sidle, T. Nanji, M. G. Sweeney, H. Houlden, A. Batla, A. L. Zecchinelli, G. Pezzoli, G. Marotta, A. Lees, P. Alegria, P. Krack, F. Cormier-Dequaire, S. Lesage, A. Brice, P. Heutink, T. Gasser, S. J. Lubbe, H. R. Morris, P. Taba, S. Koks, E. Majounie, J. Raphael Gibbs, A. Singleton, J. Hardy, S. Klebe, K. P. Bhatia, N. W. Wood, and International Parkinson's Disease Genomics Consortium and UCL-exomes consortium. Parkinson's disease in GTP cyclohydrolase 1 mutation carriers. *Brain*, 137(Pt 9):2480–2492, 2014. ISSN 0006-8950, 1460-2156. doi: 10.1093/brain/awu179.

M. P. Menden, D. Wang, M. J. Mason, B. Szalai, K. C. Bulusu, Y. Guan, T. Yu, J. Kang, M. Jeon, R. Wolfinger, T. Nguyen, M. Zaslavskiy, AstraZeneca-Sanger Drug Combination DREAM Consortium, I. S. Jang, Z. Ghazoui, M. E. Ahsen, R. Vogel, E. C. Neto, T. Norman, E. K. Y. Tang, M. J. Garnett, G. Y. D. Veroli, S. Fawell, G. Stolovitzky, J. Guinney, J. R. Dry, and J. Saez-Rodriguez. Community assessment to advance computational prediction of cancer drug combinations in a pharmacogenomic screen. *Nat. Commun.*, 10(1):2674, 2019. ISSN 2041-1723. doi: 10.1038/s41467-019-09799-2.

X.-Y. Meng, H.-X. Zhang, M. Mezei, and M. Cui. Molecular docking: a powerful approach for structure-based drug discovery. *Curr. Comput. Aided Drug Des.*, 7(2):146–157, 2011.

A. J. Mishizen-Eberz, R. P. Guttmann, B. I. Giasson, G. A. Day, 3rd, R. Hodara, H. Ischiropoulos, V. M.-Y. Lee, J. Q. Trojanowski, and D. R. Lynch. Distinct cleavage patterns of normal and pathologic forms of alpha-synuclein by calpain I in vitro. *J. Neurochem.*, 86(4):836–847, 2003. ISSN 0022-3042. doi: 10.1046/j.1471-4159.2003.01878.x.

A. Mo, U. M. Marigorta, D. Arafat, L. H. K. Chan, L. Ponder, S. R. Jang, J. Prince, S. Kugathasan, S. Prahalad, and G. Gibson. Disease-specific regulation of gene expression in a comparative analysis of juvenile idiopathic arthritis and inflammatory bowel disease. *Genome Med.*, 10(1):48, 2018.

R. L. Momparler. Optimization of cytarabine (ARA-C) therapy for acute myeloid leukemia. *Exp. Hematol. Oncol.*, 2:20, 2013. ISSN 2162-3619. doi: 10.1186/2162-3619-2-20.

S. L. Morgan, P. Naderi, K. Koler, Y. Pita-Juarez, I. Vlachos, and W. A. Hide. Are all pathways related to Alzheimer's disease? Manuscript in Preparation, 2020.

Y. Morishima, Y. Gotoh, J. Zieg, T. Barrett, H. Takano, R. Flavell, R. J. Davis, Y. Shirasaki, and M. E. Greenberg. Beta-amyloid induces neuronal apoptosis via a mechanism that involves the c-jun n-terminal kinase pathway and the induction of fas ligand. *J. Neurosci.*, 21(19):7551–7560, 2001. ISSN 0270-6474, 1529-2401.

F. Mosteller and R. A. Fisher. Questions and answers. *Am. Stat.*, 2(5):30–31, 1948. ISSN 0003-1305. doi: 10.2307/2681650.

A. Mouatt-Prigent, J. O. Karlsson, Y. Agid, and E. C. Hirsch. Increased m-calpain expression in the mesencephalon of patients with Parkinson's disease but not in other neurodegenerative disorders involving the mesencephalon: a role in nerve cell death? *Neuroscience*, 73(4):979–987, 1996. ISSN 0306-4522. doi: 10.1016/0306-4522(96) 00100-5.

A. Mullard. Drug repurposing programmes get lift off. *Nat. Rev. Drug Discov.*, 11(7): 505–506, 2012.

A. Mullard. Parsing clinical success rates. *Nat. Rev. Drug Discov.*, 15(7):447, 2016.

M. J. Murray, T. Tang, C. Ryder, D. Mabin, and J. C. Nicholson. Childhood leukaemia masquerading as juvenile idiopathic arthritis. *BMJ*, 329(7472):959–961, 2004. ISSN 0959-8138, 1756-1833. doi: 10.1136/bmj.329.7472.959.

L. K. Myers, G. C. Higgins, T. H. Finkel, A. M. Reed, J. W. Thompson, R. C. Walton, J. Hendrickson, N. C. Kerr, R. K. Pandya-Lipman, B. V. Shlopov, P. Stastny, A. E. Postlethwaite, and A. H. Kang. Juvenile arthritis and autoimmunity to type II collagen. *Arthritis Rheum.*, 44(8):1775–1781, 2001. ISSN 0004-3591. doi: 10.1002/1529-0131(200108)44:8<1775::AID-ART313>3.0.CO;2-V.

A. Naba, K. R. Clauser, S. Hoersch, H. Liu, S. A. Carr, and R. O. Hynes. The matrisome: in silico definition and in vivo characterization by proteomics of normal and tumor extracellular matrices. *Mol. Cell. Proteomics*, 11(4):M111.014647, 2012.

J. C. Nacher and J.-M. Schwartz. A global view of drug-therapy interactions. *BMC Pharmacol.*, 8:5, 2008.

S. Nagai, K. Takenaka, D. Nachagari, C. Rose, K. Domoney, D. Sun, A. Sparreboom, and J. D. Schuetz. Deoxycytidine kinase modulates the impact of the ABC transporter ABCG2 on clofarabine cytotoxicity. *Cancer Res.*, 71(5):1781–1791, 2011. ISSN 0008-5472, 1538-7445. doi: 10.1158/0008-5472.CAN-10-1919.

S. Nakamura, Y. Kawamoto, S. Nakano, I. Akiguchi, and J. Kimura. p35nck5a and cyclin-dependent kinase 5 colocalize in lewy bodies of brains with Parkinson's disease. *Acta Neuropathol.*, 94(2):153–157, 1997. ISSN 0001-6322. doi: 10.1007/s004010050687.

J. E. Nash and J. M. Brotchie. Characterisation of striatal NMDA receptors involved in the generation of parkinsonian symptoms: intrastriatal microinjection studies in the 6-OHDA-lesioned rat. *Mov. Disord.*, 17(3):455–466, 2002. ISSN 0885-3185. doi: 10.1002/mds.10107.

D. Naughton, M. Whelan, E. C. Smith, R. Williams, D. R. Blake, and M. Grootveld. An investigation of the abnormal metabolic status of synovial fluid from patients with rheumatoid arthritis by high field proton nuclear magnetic resonance spectroscopy. *FEBS Lett.*, 317(1-2):135–138, 1993. ISSN 0014-5793. doi: 10.1016/0014-5793(93)81508-w.

T.-M. Nguyen, A. Shafi, T. Nguyen, and S. Draghici. Identifying significantly impacted pathways: a comprehensive review and assessment. *Genome Biol.*, 20(1):203, 2019. ISSN 1465-6906. doi: 10.1186/s13059-019-1790-4.

NIH U.S. National Library of Medicine. DailyMed. https://dailymed.nlm.nih.gov/dailymed/, 2016. Accessed: 2017-3-16.

D. Nishimura. BioCarta. *Biotech Software & Internet Report*, 2(3):117–120, 2001.

T. Nomura and T. M. Kutchan. Three new o-methyltransferases are sufficient for all o-methylation reactions of ipecac alkaloid biosynthesis in root culture of psychotria ipecacuanha. *J. Biol. Chem.*, 285(10):7722–7738, 2010. ISSN 0021-9258, 1083-351X. doi: 10.1074/jbc.M109.086157.

M. J. Nunes, M. Moutinho, M. J. Gama, C. M. P. Rodrigues, and E. Rodrigues. Histone deacetylase inhibition decreases cholesterol levels in neuronal cells by modulating key genes in cholesterol synthesis, uptake and efflux. *PLoS One*, 8(1):e53394, 2013. ISSN 1932-6203. doi: 10.1371/journal.pone.0053394.

E. J. Oberle, J. G. Harris, and J. W. Verbsky. Polyarticular juvenile idiopathic arthritis - epidemiology and management approaches. *Clin. Epidemiol.*, 6:379–393, 2014.

K. Oen, P. N. Malleson, D. A. Cabral, A. M. Rosenberg, R. E. Petty, and M. Cheang. Disease course and outcome of juvenile rheumatoid arthritis in a multicenter cohort. *J. Rheumatol.*, 29(9):1989–1999, 2002.

W. H. Oertel. Recent advances in treating Parkinson's disease. *F1000Res.*, 6:260, 2017. ISSN 2046-1402. doi: 10.12688/f1000research.10100.1.

Y. Okada, D. Wu, G. Trynka, T. Raj, C. Terao, K. Ikari, Y. Kochi, K. Ohmura, A. Suzuki, S. Yoshida, R. R. Graham, A. Manoharan, W. Ortmann, T. Bhangale, J. C. Denny, R. J. Carroll, A. E. Eyler, J. D. Greenberg, J. M. Kremer, D. A. Pappas, L. Jiang, J. Yin, L. Ye, D.-F. Su, J. Yang, G. Xie, E. Keystone, H.-J. Westra, T. Esko, A. Metspalu, X. Zhou, N. Gupta, D. Mirel, E. A. Stahl, D. Diogo, J. Cui, K. Liao, M. H. Guo, K. Myouzen, T. Kawaguchi, M. J. H. Coenen, P. L. C. M. van Riel, M. A. F. J. van de Laar, H.-J. Guchelaar, T. W. J. Huizinga, P. D. é, X. Mariette, S. L. Bridges Jr, A. Zhernakova, R. E. M. Toes, P. P. Tak, C. Miceli-Richard, S.-Y. Bang, H.-S. Lee, J. Martin, M. A. Gonzalez-Gay, L. Rodriguez-Rodriguez, S. Rantapää-Dahlqvist, L. Arlestig, H. K. Choi, Y. Kamatani, P. Galan, M. Lathrop, RACI consortium, GARNET consortium, S. Eyre, J. Bowes, A. Barton, N. de Vries, L. W. Moreland, L. A. Criswell, E. W. Karlson, A. Taniguchi, R. Yamada, M. Kubo, J. S. Liu, S.-C. Bae, J. Worthington, L. Padyukov, L. Klareskog, P. K. Gregersen, S. Raychaudhuri, B. E. Stranger, P. L. De Jager, L. Franke, P. M. Visscher, M. A. Brown, H. Yamanaka, T. Mimori, A. Takahashi, H. Xu, T. W. Behrens, K. A. Siminovitch, S. Momohara, F. Matsuda, K. Yamamoto, and R. M. Plenge. Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature*, 506 (7488):376–381, 2014.

M. J. Ombrello, E. F. Remmers, I. Tachmazidou, A. Grom, D. Foell, J.-P. Haas, A. Martini, M. Gattorno, S. Özen, S. Prahalad, A. S. Zeft, J. F. Bohnsack, E. D. Mellins, N. T. Ilowite, R. Russo, C. Len, M. O. E. Hilario, S. Oliveira, R. S. M. Yeung, A. Rosenberg, L. R. Wedderburn, J. Anton, T. Schwarz, A. Hinks, Y. Bilginer, J. Park, J. Cobb, C. L. Satorius, B. Han, E. Baskin, S. Signa, R. H. Duerr, J. P. Achkar, M. I. Kamboh, K. M. Kaufman, L. C. Kottyan, D. Pinto, S. W. Scherer, M. E. Alarcón-Riquelme, E. Docampo, X. Estivill, A. Gül, British Society of Pediatric and Adolescent Rheumatology (BSPAR) Study Group, Childhood Arthritis Prospective Study (CAPS) Group, Randomized Placebo Phase Study of Rilonacept in sJIA (RAPPORT) Investigators, Sparks-Childhood Arthritis Response to Medication Study (CHARMS) Group, Biologically Based Outcome Predictors in JIA (BBOP) Group, P. I. W. de Bakker, S. Raychaudhuri, C. D. Langefeld, S. Thompson, E. Zeggini, W. Thomson, D. L. Kastner, P. Woo, and International Childhood Arthritis Genetics (INCHARGE) Consortium. HLA-DRB1*11 and variants of the MHC class II locus are strong risk factors for systemic juvenile idiopathic arthritis. *Proc. Natl. Acad. Sci. U. S. A.*, 112(52):15970–15975, 2015.

T. I. Oprea and J. P. Overington. Computational and practical aspects of drug repositioning. *Assay Drug Dev. Technol.*, 13(6):299–306, 2015.

T. F. Outeiro, E. Kontopoulos, S. M. Altmann, I. Kufareva, K. E. Strathearn, A. M. Amore, C. B. Volk, M. M. Maxwell, J.-C. Rochet, P. J. McLean, A. B. Young, R. Abagyan, M. B.

Feany, B. T. Hyman, and A. G. Kazantsev. Sirtuin 2 inhibitors rescue alpha-synuclein-mediated toxicity in models of Parkinson's disease. *Science*, 317(5837):516–519, 2007. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.1143780.

H. Paik, A.-Y. Chung, H.-C. Park, R. W. Park, K. Suk, J. Kim, H. Kim, K. Lee, and A. J. Butte. Repurpose terbutaline sulfate for amyotrophic lateral sclerosis using electronic medical records. *Sci. Rep.*, 5:8580, 2015.

E. Palomer, J. Buechler, and P. C. Salinas. Wnt signaling deregulation in the aging and Alzheimer's brain. *Front. Cell. Neurosci.*, 13:227, 2019. ISSN 1662-5102. doi: 10.3389/fncel.2019.00227.

Y.-J. Pan, W.-H. Wang, T.-Y. Huang, W.-H. Weng, C.-K. Fang, Y.-C. Chen, and J.-J. Hwang. Quetiapine ameliorates collagen-induced arthritis in mice via the suppression of the AKT and ERK signaling pathways. *Inflamm. Res.*, 67(10):847–861, 2018. ISSN 1023-3830, 1420-908X. doi: 10.1007/s00011-018-1176-1.

S. M. Papa, R. C. Boldry, T. M. Engber, A. M. Kask, and T. N. Chase. Reversal of levodopa-induced motor fluctuations in experimental parkinsonism by NMDA receptor blockade. *Brain Res.*, 701(1-2):13–18, 1995. ISSN 0006-8993. doi: 10.1016/0006-8993(95) 00924-3.

M. D. Paranjpe, A. Taubes, and M. Sirota. Insights into computational drug repurposing for neurodegenerative disease. *Trends Pharmacol. Sci.*, 40(8):565–576, 2019. ISSN 0165-6147, 1873-3735. doi: 10.1016/j.tips.2019.06.003.

K. Park. A review of computational drug repurposing. *Transl Clin Pharmacol*, 27(2): 59–63, 2019. ISSN 2383-5427, 2289-0882. doi: 10.12793/tcp.2019.27.2.59.

R. Patro, G. Duggal, M. I. Love, R. A. Irizarry, and C. Kingsford. Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods*, 14(4):417–419, 2017. ISSN 1548-7091, 1548-7105. doi: 10.1038/nmeth.4197.

G. A. Pavlopoulos, P. I. Kontou, A. Pavlopoulou, C. Bouyioukos, E. Markou, and P. G. Bagos. Bipartite graphs in systems biology and medicine: a survey of methods and applications. *Gigascience*, 7(4):1–31, 2018.

F. Peng, Y. Zhao, X. Huang, C. Chen, L. Sun, L. Zhuang, and L. Xue. Loss of polo ameliorates APP-induced Alzheimer's disease-like symptoms in drosophila. *Sci. Rep.*, 5:16816, 2015. ISSN 2045-2322. doi: 10.1038/srep16816.

M. F. Perutz, M. G. Rossmann, A. F. Cullis, H. Muirhead, G. Will, and A. C. North. Structure of haemoglobin: a three-dimensional fourier synthesis at 5.5-a. resolution, obtained by x-ray analysis. *Nature*, 185(4711):416–422, 1960. ISSN 0028-0836. doi: 10.1038/185416a0.

L. Petrucelli, D. Dickson, K. Kehoe, J. Taylor, H. Snyder, A. Grover, M. De Lucia, E. McGowan, J. Lewis, G. Prihar, J. Kim, W. H. Dillmann, S. E. Browne, A. Hall, R. Voellmy, Y. Tsuboi, T. M. Dawson, B. Wolozin, J. Hardy, and M. Hutton. CHIP and hsp70 regulate tau ubiquitination, degradation and aggregation. *Hum. Mol. Genet.*, 13 (7):703–714, 2004. ISSN 0964-6906. doi: 10.1093/hmg/ddh083.

R. E. Petty, T. R. Southwood, P. Manners, J. Baum, D. N. Glass, J. Goldenberg, X. He, J. Maldonado-Cocco, J. Orozco-Alcala, A.-M. Prieur, M. E. Suarez-Almazor, P. Woo, and International League of Associations for Rheumatology. International league of associations for rheumatology classification of juvenile idiopathic arthritis: second revision, edmonton, 2001. *J. Rheumatol.*, 31(2):390–392, 2004.

Pharma Intelligence. Pharmaprojects. https://pharmaintelligence.informa.com/products-and-services/data-and-analysis/pharmaprojects, 2017. Accessed: 2017-3-17.

E. C. Phillips, C. L. Croft, K. Kurbatskaya, M. J. O'Neill, M. L. Hutton, D. P. Hanger, C. J. Garwood, and W. Noble. Astrocytes and neuroinflammation in Alzheimer's disease. *Biochem. Soc. Trans.*, 42(5):1321–1325, 2014. ISSN 0300-5127, 1470-8752. doi: 10.1042/BST20140155.

K. Phillips, N. Kedersha, L. Shen, P. J. Blackshear, and P. Anderson. Arthritis suppressor genes TIA-1 and TTP dampen the expression of tumor necrosis factor alpha, cyclooxygenase 2, and inflammatory arthritis. *Proc. Natl. Acad. Sci. U. S. A.*, 101(7):2011–2016, 2004.

B. Phipson and G. K. Smyth. Permutation p-values should never be zero: calculating exact p-values when permutations are randomly drawn. *Stat. Appl. Genet. Mol. Biol.*, 9: Article39, 2010.

B. R. Pinho, S. D. Reis, P. Guedes-Dias, A. Leitão-Rocha, C. Quintas, P. Valentão, P. B. Andrade, M. M. Santos, and J. M. A. Oliveira. Pharmacological modulation of HDAC1 and HDAC6 in vivo in a zebrafish model: Therapeutic implications for Parkinson's disease. *Pharmacol. Res.*, 103:328–339, 2016. ISSN 1043-6618, 1096-1186. doi: 10.1016/j.phrs.2015.11.024.

Y. Pita-Juárez, G. Altschuler, S. Kariotis, W. Wei, K. Koler, C. Green, R. Tanzi, and W. Hide. The pathway coexpression network: Revealing pathway relationships. *PLoS Comput. Biol.*, 14(3):e1006042, 2018.

K. Poesen, D. Lambrechts, P. van Damme, J. Dhondt, F. Bender, N. Frank, E. Bogaert, B. Claes, L. Heylen, A. Verheyen, K. Raes, M. Tjwa, U. Eriksson, M. Shibuya, R. Nuydens, L. van den Bosch, T. Meert, R. D'Hooge, M. Sendtner, W. Robberecht, and P. Carmeliet. Novel role for vascular endothelial growth factor (VEGF) receptor-1 and its ligand VEGF-B in motor neuron degeneration. *J. Neurosci.*, 28(42):10451–10459, 2008. ISSN 0270-6474, 1529-2401. doi: 10.1523/JNEUROSCI.1092-08.2008.

W. Poewe, A. Antonini, J. C. Zijlmans, P. R. Burkhard, and F. Vingerhoets. Levodopa in the treatment of Parkinson's disease: an old drug still going strong. *Clin. Interv. Aging*, 5:229–238, 2010. ISSN 1176-9092, 1178-1998. doi: 10.2147/cia.s6456.

V. Prasad and S. Mailankody. Research and development spending to bring a single cancer drug to market and revenues after approval. *JAMA Intern. Med.*, 177(11):1569–1575, 2017. ISSN 2168-6106, 2168-6114. doi: 10.1001/jamainternmed.2017.3601.

A. B. Pupyshev, M. A. Tikhonova, A. A. Akopyan, M. V. Tenditnik, N. I. Dubrovina, and T. A. Korolenko. Therapeutic activation of autophagy by combined treatment with rapamycin and trehalose in a mouse MPTP-induced model of Parkinson's disease. *Pharmacol. Biochem. Behav.*, 177:1–11, 2019. ISSN 0091-3057, 1873-5177. doi: 10.1016/j.pbb.2018.12.005.

S. Pushpakom, F. Iorio, P. A. Eyers, K. J. Escott, S. Hopper, A. Wells, A. Doig, T. Guilliams, J. Latimer, C. McNamee, A. Norris, P. Sanseau, D. Cavalla, and M. Pirmohamed. Drug repurposing: progress, challenges and recommendations. *Nat. Rev. Drug Discov.*, 18(1): 41–58, 2019. ISSN 1474-1776, 1474-1784. doi: 10.1038/nrd.2018.168.

Z. M. Qian and Y. Ke. Huperzine a: Is it an effective Disease-Modifying drug for Alzheimer's disease? *Front. Aging Neurosci.*, 6:216, 2014. ISSN 1663-4365. doi: 10.3389/fnagi.2014.00216.

L. Qin, Q. Xu, Z. Li, L. Chen, Y. Li, N. Yang, Z. Liu, J. Guo, L. Shen, E. G. Allen, C. Chen, C. Ma, H. Wu, X. Zhu, P. Jin, and B. Tang. Ethnicity-specific and overlapping alterations of brain hydroxymethylome in Alzheimer's disease. *Hum. Mol. Genet.*, 29(1):149–158, 2020. ISSN 0964-6906, 1460-2083. doi: 10.1093/hmg/ddz273.

C. Qiu, G. Hu, M. Kivipelto, T. Laatikainen, R. Antikainen, L. Fratiglioni, P. Jousilahti, and J. Tuomilehto. Association of blood pressure and hypertension with the risk of Parkinson disease: the national FINRISK study. *Hypertension*, 57(6):1094–1100, 2011. ISSN 0194-911X, 1524-4563. doi: 10.1161/HYPERTENSIONAHA.111.171249.

C. M. Quiñonez-Flores, S. A. González-Chávez, and C. Pacheco-Tena. Hypoxia and its implications in rheumatoid arthritis. *J. Biomed. Sci.*, 23(1):62, 2016.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2019. URL https://www.R-project.org/.

M. S. Rafii, S. Walsh, J. T. Little, K. Behan, B. Reynolds, C. Ward, S. Jin, R. Thomas, P. S. Aisen, and Alzheimer's Disease Cooperative Study. A phase II trial of huperzine a in mild to moderate Alzheimer disease. *Neurology*, 76(16):1389–1394, 2011. ISSN 0028-3878, 1526-632X. doi: 10.1212/WNL.0b013e318216eb7b.

D. S. C. Raj. Role of interleukin-6 in the anemia of chronic disease. *Semin. Arthritis Rheum.*, 38(5):382–388, 2009. ISSN 0049-0172, 1532-866X. doi: 10.1016/j.semarthrit.2008.01.006.

M. M. Rajsombath, A. Y. Nam, M. Ericsson, and S. Nuber. Female sex and Brain-Selective estrogen benefit α-Synuclein tetramerization and the PD-like motor syndrome in 3K transgenic mice. *J. Neurosci.*, 39(38):7628–7640, 2019. ISSN 0270-6474, 1529-2401. doi: 10.1523/JNEUROSCI.0313-19.2019.

K. Ramanathan, A. Glaser, H. Lythgoe, J. Ong, M. W. Beresford, A. Midgley, and H. L. Wright. Neutrophil activation signature in juvenile idiopathic arthritis indicates the presence of low-density granulocytes. *Rheumatology*, 57(3):488–498, 2018.

T. M. Randis, K. D. Puri, H. Zhou, and T. G. Diacovo. Role of PI3Kdelta and PI3Kgamma in inflammatory arthritis and tissue localization of neutrophils. *Eur. J. Immunol.*, 38(5): 1215–1224, 2008. ISSN 0014-2980. doi: 10.1002/eji.200838266.

A. Ravelli, A. Consolaro, G. Horneff, R. M. Laxer, D. J. Lovell, N. M. Wulffraat, J. D. Akikusa, S. M. Al-Mayouf, J. Antón, T. Avcin, R. A. Berard, M. W. Beresford, R. Burgos-Vargas, R. Cimaz, F. De Benedetti, E. Demirkaya, D. Foell, Y. Itoh, P. Lahdenne, E. M. Morgan, P. Quartier, N. Ruperto, R. Russo, C. Saad-Magalhães, S. Sawhney, C. Scott, S. Shenoi, J. F. Swart, Y. Uziel, S. J. Vastert, and J. S. Smolen. Treating juvenile idiopathic arthritis to target: recommendations of an international task force. *Ann. Rheum. Dis.*, 77(6):819–828, 2018.

N. Rawal, O. Corti, P. Sacchetti, H. Ardilla-Osorio, B. Sehat, A. Brice, and E. Arenas. Parkin protects dopaminergic neurons from excessive wnt/beta-catenin signaling. *Biochem. Biophys. Res. Commun.*, 388(3):473–478, 2009. ISSN 0006-291X, 1090-2104. doi: 10.1016/j.bbrc.2009.07.014.

resTORbio, Inc. resTORbio announces initiation of phase 1b/2a trial of RTB101 in Parkinson's disease. https://ir.restorbio.com/news-releases/news-release-details/restorbio-announces-initiation-phase-1b2a-trial-rtb101, 2019. Accessed: 2020-3-11.

D. K. Rhee, S. C. Hockman, S.-K. Choi, Y.-E. Kim, C. Park, V. C. Manganiello, and K. K. Kim. SFPQ, a multifunctional nuclear protein, regulates the transcription of PDE3A. *Biosci. Rep.*, 37(4), 2017.

J. Riise, N. Plath, B. Pakkenberg, and A. Parachikova. Aberrant wnt signaling pathway in medial temporal lobe structures of Alzheimer's disease. *J. Neural Transm.*, 122(9): 1303–1318, 2015. ISSN 0300-9564, 1435-1463. doi: 10.1007/s00702-015-1375-7.

S. Ringold, P. F. Weiss, T. Beukelman, E. M. DeWitt, N. T. Ilowite, Y. Kimura, R. M. Laxer, D. J. Lovell, P. A. Nigrovic, A. B. Robinson, R. K. Vehe, and American Collge of Rheumatology. 2013 update of the 2011 american college of rheumatology recommendations for the treatment of juvenile idiopathic arthritis: recommendations for the medical therapy of children with systemic juvenile idiopathic arthritis and tuberculosis screening among children receiving biologic medications. *Arthritis Rheum.*, 65(10): 2499–2512, 2013. ISSN 0004-3591, 1529-0131. doi: 10.1002/art.38092.

S. Ringold, P. F. Weiss, R. A. Colbert, E. M. DeWitt, T. Lee, K. Onel, S. Prahalad, R. Schneider, S. Shenoi, R. K. Vehe, Y. Kimura, and Juvenile Idiopathic Arthritis Research Committee of the Childhood Arthritis and Rheumatology Research Alliance. Childhood arthritis and rheumatology research alliance consensus treatment plans for new-onset polyarticular juvenile idiopathic arthritis. *Arthritis Care Res.*, 66(7):1063–1072, 2014.

M. E. Ritchie, B. Phipson, D. Wu, Y. Hu, C. W. Law, W. Shi, and G. K. Smyth. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, 43(7):e47, 2015. doi: 10.1093/nar/gkv007.

J. Rius, M. Guma, C. Schachtrup, K. Akassoglou, A. S. Zinkernagel, V. Nizet, R. S. Johnson, G. G. Haddad, and M. Karin. NF-kappaB links innate immunity to the hypoxic response through transcriptional regulation of HIF-1alpha. *Nature*, 453(7196):807–811, 2008.

G. Rizzi and K. R. Tan. Dopamine and acetylcholine, a circuit point of view in Parkinson's disease. *Front. Neural Circuits*, 11:110, 2017. ISSN 1662-5110. doi: 10.3389/fncir.2017.00110.

M. D. Robinson, D. J. McCarthy, and G. K. Smyth. edger: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1): 139–140, 2010. doi: 10.1093/bioinformatics/btp616.

W. A. Rocca, B. R. Grossardt, and L. T. Shuster. Oophorectomy, menopause, estrogen treatment, and cognitive aging: clinical evidence for a window of opportunity. *Brain Res.*, 1379:188–198, 2011. ISSN 0006-8993, 1872-6240. doi: 10.1016/j.brainres.2010.10.031.

W. Rojo-Contreras, E. M. Olivas-Flores, J. I. Gamez-Nava, H. Montoya-Fuentes, B. Trujillo-Hernandez, X. Trujillo, A. E. Suarez-Rincon, L. M. Baltazar-Rodriguez, J. Sanchez-Hernandez, M. Ramirez-Flores, J. Vazquez-Salcedo, J. Rojo-Contreras, J. Morales-Romero, and L. Gonzalez-Lopez. Cervical human papillomavirus infection in mexican women with systemic lupus erythematosus or rheumatoid arthritis. *Lupus*, 21(4):365–372, 2012. ISSN 0961-2033, 1477-0962. doi: 10.1177/0961203311425517.

S. G. V. Rosa and W. C. Santos. Clinical trials on drug repositioning for COVID-19 treatment. *Rev. Panam. Salud Publica*, 44:e40, 2020. ISSN 1020-4989, 1680-5348. doi: 10.26633/RPSP.2020.40.

V. Rozani, T. Gurevich, N. Giladi, B. El-Ad, J. Tsamir, B. Hemo, and C. Peretz. Higher serum cholesterol and decreased Parkinson's disease risk: A statin-free cohort study. *Mov. Disord.*, 33(8):1298–1305, 2018. ISSN 0885-3185, 1531-8257. doi: 10.1002/mds.27413.

J. Rustenhoven, A. M. Smith, L. C. Smyth, D. Jansson, E. L. Scotter, M. E. V. Swanson, M. Aalderink, N. Coppieters, P. Narayan, R. Handley, C. Overall, T. I. H. Park, P. Schweder, P. Heppner, M. A. Curtis, R. L. M. Faull, and M. Dragunow. PU.1 regulates Alzheimer's disease-associated genes in primary human microglia. *Mol. Neurodegener.*, 13(1):44, 2018. ISSN 1750-1326. doi: 10.1186/s13024-018-0277-1.

J.-H. Ryu, C.-S. Chae, J.-S. Kwak, H. Oh, Y. Shin, Y. H. Huh, C.-G. Lee, Y.-W. Park, C.-H. Chun, Y.-M. Kim, S.-H. Im, and J.-S. Chun. Hypoxia-inducible factor-2α is an essential catabolic regulator of inflammatory rheumatoid arthritis. *PLoS Biol.*, 12(6):e1001881, 2014.

S. Sacre, M. Medghalchi, B. Gregory, F. Brennan, and R. Williams. Fluoxetine and citalopram exhibit potent antiinflammatory activity in human and murine models of rheumatoid arthritis and inhibit toll-like receptors. *Arthritis Rheum.*, 62(3):683–693, 2010. ISSN 0004-3591, 1529-0131. doi: 10.1002/art.27304.

T. Saito and M. Rehmsmeier. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS One*, 10(3): e0118432, 2015.

P. Sanseau, P. Agarwal, M. R. Barnes, T. Pastinen, J. B. Richards, L. R. Cardon, and V. Mooser. Use of genome-wide association studies for drug repositioning. *Nat. Biotechnol.*, 30(4):317–320, 2012.

A. M. Saunders, W. J. Strittmatter, D. Schmechel, P. H. St. George-Hyslop, M. A. Pericak-Vance, S. H. Joo, B. L. Rosi, J. F. Gusella, D. R. Crapper-Mac Lachlan, M. J. Alberts, C. Hulette, B. Crain, D. Goldgaber, and A. D. Roses. Association of apolipoprotein E allele ε4 with late-onset familial and sporadic Alzheimer's disease. *Neurology*, 43(8): 1467–1472, 1993. ISSN 0028-3878.

B. Saunders, S. Lyon, M. Day, B. Riley, E. Chenette, S. Subramaniam, and I. Vadivelu. The molecule pages database. *Nucleic Acids Res.*, 36(Database issue):D700–6, 2008.

M. J. Savage, Y.-G. Lin, J. R. Ciallella, D. G. Flood, and R. W. Scott. Activation of c-jun n-terminal kinase and p38 in an Alzheimer's disease model is associated with amyloid deposition. *J. Neurosci.*, 22(9):3376–3385, 2002. ISSN 0270-6474, 1529-2401. doi: 20026352.

C. F. Schaefer, K. Anthony, S. Krupa, J. Buchoff, M. Day, T. Hannay, and K. H. Buetow. PID: the pathway interaction database. *Nucleic Acids Res.*, 37(Database issue):D674–9, 2009.

J. Schafer, R. Opgen-Rhein, V. Zuber, M. Ahdesmaki, A. P. D. Silva, and K. Strimmer. *corpcor: Efficient Estimation of Covariance and (Partial) Correlation*, 2017. URL https://CRAN.R-project.org/package=corpcor. R package version 1.6.9.

M. Schirmer, E. Mur, K. P. Pfeiffer, J. Thaler, and G. Konwalinka. The safety profile of low-dose cladribine in refractory rheumatoid arthritis. a pilot trial. *Scand. J. Rheumatol.*, 26(5):376–379, 1997. ISSN 0300-9742. doi: 10.3109/03009749709065702.

K. Schmidt and B. J. Nichols. Functional interdependence between septin and actin cytoskeleton. *BMC Cell Biol.*, 5(1):43, 2004.

L. M. Schriml, E. Mitraka, J. Munro, B. Tauber, M. Schor, L. Nickle, V. Felix, L. Jeng, C. Bearer, R. Lichenstein, K. Bisordi, N. Campion, B. Hyman, D. Kurland, C. P. Oates, S. Kibbey, P. Sreekumar, C. Le, M. Giglio, and C. Greene. Human disease ontology 2018 update: classification, content and workflow expansion. *Nucleic Acids Res.*, 47 (D1):D955–D962, 2019.

D. J. Selkoe and J. Hardy. The amyloid hypothesis of Alzheimer's disease at 25 years. *EMBO Mol. Med.*, 8(6):595–608, 2016. ISSN 1757-4676, 1757-4684. doi: 10.15252/emmm.201606210.

Y. Shafranovich. Common format and MIME type for Comma-Separated values (CSV) files. Technical Report RFC 4180, Internet Engineering Steering Group, 2005.

K. Shameer, B. S. Glicksberg, R. Hodos, K. W. Johnson, M. A. Badgeley, B. Readhead, M. S. Tomlinson, T. O'Connor, R. Miotto, B. A. Kidd, R. Chen, A. Ma'ayan, and J. T. Dudley. Systematic analyses of drugs and disease indications in RepurposeDB reveal pharmacological, biological and epidemiological factors influencing drug repositioning. *Brief. Bioinform.*, 2017.

A.-W. Shao, H. Sun, Y. Geng, Q. Peng, P. Wang, J. Chen, T. Xiong, R. Cao, and J. Tang. Bclaf1 is an important NF-κB signaling transducer and C/EBPβ regulator in DNA damage-induced senescence. *Cell Death Differ.*, 23(5):865–875, 2016.

A. F. Shaughnessy. Old drugs, new tricks. *BMJ*, 342:d741, 2011.

H.-Y. Shen, J.-C. He, Y. Wang, Q.-Y. Huang, and J.-F. Chen. Geldanamycin induces heat shock protein 70 and protects against MPTP-induced dopaminergic neurotoxicity in mice. *J. Biol. Chem.*, 280(48):39962–39969, 2005. ISSN 0021-9258. doi: 10.1074/jbc.M505524200.

J. W. Shim and J. R. Madsen. VEGF signaling in neurological disorders. *Int. J. Mol. Sci.*, 19(1), 2018. ISSN 1422-0067. doi: 10.3390/ijms19010275.

I. Shoulson, J. Penney, M. McDermott, S. Schwid, E. Kayson, T. Chase, S. Fahn, J. T. Greenamyre, A. Lang, A. Siderowf, N. Pearson, M. Harrison, E. Rost, A. Colcher, M. Lloyd, M. Matthews, R. Pahwa, D. McGuire, M. F. Lew, S. Schuman, K. Marek, S. Broshjeit, S. Factor, D. Brown, A. Feigin, J. Mazurkiewicz, B. Ford, D. Jennings, S. Dilllon, C. Comella, L. Blasucci, K. Janko, L. Shulman, W. Wiener, D. Bateman-Rodriguez, A. Carrion, O. Suchowersky, A. L. Lafontaine, C. Pantella, E. Siemers, J. Belden, R. Davies, M. Lannon, D. Grimes, P. Gray, W. Martin, L. Kennedy, C. Adler, S. Newman, J. Hammerstad, C. Stone, P. Lewitt, K. Bardram, K. Mistura, J. Miyasaki, L. Johnston, J. H. Cha, M. Tennis, M. Panniset, J. Hall, J. Tetrud, J. Friedlander, R. Hauser, L. Gauger, R. Rodnitzky, A. Deleo, J. Dobson, L. Seeberger, C. Dingmann, D. Tarsy, P. Ryan, L. Elmer, D. Ruzicka, M. Stacy, M. Brewer, B. Locke, D. Baker, C. Casaceli, D. Day, M. Florack, K. Hodgeman, N. Laroia, R. Nobel, C. Orme, L. Rexo, K. Rothenburgh, K. Sulimowicz, A. Watts, E. Wratni, P. Tariot, C. Cox, C. Leventhal, V. Alderfer, A. M. Craun, J. Frey, L. McCree, J. McDermott, J. Cooper, T. Holdich, B. Read, and Parkinson Study Group. A randomized, controlled trial of remacemide for motor fluctuations in Parkinson's disease. *Neurology*, 56(4):455–462, 2001. ISSN 0028-3878. doi: 10.1212/wnl.56.4.455.

S. A. Shumaker, C. Legault, L. Kuller, S. R. Rapp, L. Thal, D. S. Lane, H. Fillit, M. L. Stefanick, S. L. Hendrix, C. E. Lewis, K. Masaki, L. H. Coker, and Women's Health Initiative Memory Study. Conjugated equine estrogens and incidence of probable dementia and mild cognitive impairment in postmenopausal women: Women's health initiative memory study. *JAMA*, 291(24):2947–2958, 2004. ISSN 0098-7484, 1538-3598. doi: 10.1001/jama.291.24.2947.

N. Simola, M. Morelli, and A. R. Carta. The 6-hydroxydopamine model of Parkinson's disease. *Neurotox. Res.*, 11(3-4):151–167, 2007. ISSN 1029-8428. doi: 10.1007/bf03033565.

J. Simón-Sánchez, C. Schulte, J. M. Bras, M. Sharma, J. R. Gibbs, D. Berg, C. Paisan-Ruiz, P. Lichtner, S. W. Scholz, D. G. Hernandez, R. Krüger, M. Federoff, C. Klein, A. Goate, J. Perlmutter, M. Bonin, M. A. Nalls, T. Illig, C. Gieger, H. Houlden, M. Steffens, M. S. Okun, B. A. Racette, M. R. Cookson, K. D. Foote, H. H. Fernandez, B. J. Traynor, S. Schreiber, S. Arepalli, R. Zonozi, K. Gwinn, M. van der Brug, G. Lopez, S. J. Chanock, A. Schatzkin, Y. Park, A. Hollenbeck, J. Gao, X. Huang, N. W. Wood, D. Lorenz, G. Deuschl, H. Chen, O. Riess, J. A. Hardy, A. B. Singleton, and T. Gasser. Genome-wide association study reveals genetic risk underlying Parkinson's disease. *Nat. Genet.*, 41(12):1308–1312, 2009. ISSN 1061-4036, 1546-1718. doi: 10.1038/ng.487.

S. Singh, S. Kumar, and M. Dikshit. Involvement of the mitochondrial apoptotic pathway and nitric oxide synthase in dopaminergic neuronal death induced by 6-hydroxydopamine and lipopolysaccharide. *Redox Rep.*, 15(3):115–122, 2010. ISSN 1351-0002, 1743-2928. doi: 10.1179/174329210X12650506623447.

M. Sirota, J. T. Dudley, J. Kim, A. P. Chiang, A. A. Morgan, A. Sweet-Cordero, J. Sage, and A. J. Butte. Discovery and preclinical validation of drug indications using compendia of public gene expression data. *Sci. Transl. Med.*, 3(96):96ra77–96ra77, 2011.

S. H. Sleigh and C. L. Barton. Repurposing strategies for therapeutics. *Pharmaceut. Med.*, 24(3):151–159, 2012.

A. M. Smith, H. M. Gibbons, R. L. Oldfield, P. M. Bergin, E. W. Mee, R. L. M. Faull, and M. Dragunow. The transcription factor PU.1 is critical for viability and function of human brain microglia. *Glia*, 61(6):929–942, 2013. ISSN 0894-1491, 1098-1136. doi: 10.1002/glia.22486.

M. J. Smith, J. B. Kwok, C. A. McLean, J. J. Kril, G. A. Broe, G. A. Nicholson, R. Cappai, M. Hallupp, R. G. Cotton, C. L. Masters, P. R. Schofield, and W. S. Brooks. Variable phenotype of Alzheimer's disease with spastic paraparesis. *Ann. Neurol.*, 49(1):125–129, 2001. ISSN 0364-5134. doi: 10.1002/1531-8249(200101)49:1<125::aid-ana21>3.0.co;2-1.

P. D. Smith, S. J. Crocker, V. Jackson-Lewis, K. L. Jordan-Sciutto, S. Hayley, M. P. Mount, M. J. O'Hare, S. Callaghan, R. S. Slack, S. Przedborski, H. Anisman, and D. S. Park. Cyclin-dependent kinase 5 is a mediator of dopaminergic neuron loss in a mouse model of Parkinson's disease. *Proc. Natl. Acad. Sci. U. S. A.*, 100(23):13650–13655, 2003. ISSN 0027-8424. doi: 10.1073/pnas.2232515100.

R. B. Smith. Repositioned drugs: integrating intellectual property and regulatory strategies. *Drug Discov. Today Ther. Strateg.*, 8(3):131–137, 2011. ISSN 1740-6773. doi: 10.1016/j.ddstr.2011.06.008.

W. L. Smith, D. L. DeWitt, and R. M. Garavito. Cyclooxygenases: structural, cellular, and molecular biology. *Annu. Rev. Biochem.*, 69:145–182, 2000.

S. G. Snowden, A. A. Ebshiana, A. Hye, Y. An, O. Pletnikova, R. O'Brien, J. Troncoso, C. Legido-Quigley, and M. Thambisetty. Association between fatty acid metabolism in the brain and Alzheimer disease neuropathology and cognitive performance: A nontargeted metabolomic study. *PLoS Med.*, 14(3):e1002266, 2017. ISSN 1549-1277, 1549-1676. doi: 10.1371/journal.pmed.1002266.

E. Södersten, M. Feyder, M. Lerdrup, A.-L. Gomes, H. Kryh, G. Spigolon, J. Caboche, G. Fisone, and K. Hansen. Dopamine signaling leads to loss of polycomb repression and aberrant gene activation in experimental parkinsonism. *PLoS Genet.*, 10(9):e1004574, 2014. ISSN 1553-7390, 1553-7404. doi: 10.1371/journal.pgen.1004574.

B. Song, K. Davis, X. S. Liu, H.-G. Lee, M. Smith, and X. Liu. Inhibition of polo-like kinase 1 reduces beta-amyloid-induced neuronal cell death in Alzheimer's disease. *Aging*, 3(9):846–851, 2011. ISSN 1945-4589. doi: 10.18632/aging.100382.

M. Steyvers and J. B. Tenenbaum. The large-scale structure of semantic networks: statistical analyses and a model of semantic growth. *Cogn. Sci.*, 29(1):41–78, 2005. ISSN 0364-0213. doi: 10.1207/s15516709cog2901_3.

J. B. Strosznajder, H. Jeśko, and R. P. Strosznajder. Effect of amyloid beta peptide on poly(ADP-ribose) polymerase activity in adult and aged rat hippocampus. *Acta Biochim. Pol.*, 47(3):847–854, 2000. ISSN 0001-527X.

J. B. Strosznajder, G. A. Czapski, A. Adamczyk, and R. P. Strosznajder. Poly(ADP-ribose) polymerase-1 in amyloid beta toxicity and Alzheimer's disease. *Mol. Neurobiol.*, 46(1): 78–84, 2012. ISSN 0893-7648, 1559-1182. doi: 10.1007/s12035-012-8258-9.

A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.*, 102(43):15545–15550, 2005.

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

P. Sun, J. Guo, R. Winnenburg, and J. Baumbach. Drug repurposing by integrated literature mining and drug-gene-disease triangulation. *Drug Discov. Today*, 2016.

H. Suo, P. Wang, J. Tong, L. Cai, J. Liu, D. Huang, L. Huang, Z. Wang, Y. Huang, J. Xu, Y. Ma, M. Yu, J. Fei, and F. Huang. NRSF is an essential mediator for the neuroprotection of trichostatin a in the MPTP mouse model of Parkinson's disease. *Neuropharmacology*, 99:67–78, 2015. ISSN 0028-3908, 1873-7064. doi: 10.1016/j.neuropharm.2015.07.015.

S. J. Swamidass. Mining small-molecule screens to repurpose drugs. *Brief. Bioinform.*, 12 (4):327–335, 2011.

J. Świdrowska, P. Smolewski, J. Stańczyk, and E. Smolewska. Serum angiogenesis markers and their correlation with Ultrasound-Detected synovitis in juvenile idiopathic arthritis. *J Immunol Res*, 2015:741457, 2015.

M. Takarabe, M. Kotera, Y. Nishimura, S. Goto, and Y. Yamanishi. Drug target prediction using adverse event report systems: a pharmacogenomic approach. *Bioinformatics*, 28 (18):i611–i618, 2012.

C. Tan, X. Liu, and J. Chen. Microarray analysis of the molecular mechanism involved in Parkinson's disease. *Parkinsons Dis.*, 2018:1590465, 2018. ISSN 2090-8083, 2042-0080. doi: 10.1155/2018/1590465.

R. E. Tanzi and L. Bertram. Twenty years of the Alzheimer's disease amyloid hypothesis: a genetic perspective. *Cell*, 120(4):545–555, 2005. ISSN 0092-8674. doi: 10.1016/j.cell.2005.02.008.

A. Tasneem, L. Aberle, H. Ananth, S. Chakraborty, K. Chiswell, B. J. McCourt, and R. Pietrobon. The database for aggregate analysis of ClinicalTrials.gov (AACT) and subsequent regrouping by clinical specialty. *PLoS One*, 7(3):e33677, 2012.

H. Teo, S. Ghosh, H. Luesch, A. Ghosh, E. T. Wong, N. Malik, A. Orth, P. de Jesus, A. S. Perry, J. D. Oliver, N. L. Tran, L. J. Speiser, M. Wong, E. Saez, P. Schultz, S. K. Chanda, I. M. Verma, and V. Tergaonkar. Telomere-independent rap1 is an IKK adaptor and regulates NF-kappaB-dependent gene expression. *Nat. Cell Biol.*, 12(8):758–767, 2010.

B. Thomas and M. F. Beal. Parkinson's disease. *Hum. Mol. Genet.*, 16 Spec No. 2: R183–94, 2007. ISSN 0964-6906. doi: 10.1093/hmg/ddm159.

M. G. Thomas, C. Welch, L. Stone, P. Allan, R. A. Barker, and R. B. White. PAX6 expression may be protective against dopaminergic cell loss in Parkinson's disease. *CNS Neurol. Disord. Drug Targets*, 15(1):73–79, 2016. ISSN 1871-5273, 1996-3181. doi: 10.2174/1871527314666150821101757.

S. D. Thompson, M. C. Marion, M. Sudman, M. Ryan, M. Tsoras, T. D. Howard, M. G. Barnes, P. S. Ramos, W. Thomson, A. Hinks, J.-P. Haas, S. Prahalad, J. F. Bohnsack, C. A. Wise, M. Punaro, C. D. Rosé, N. M. Pajewski, M. Spigarelli, M. Keddache, M. Wagner, C. D. Langefeld, and D. N. Glass. Genome-wide association analysis of juvenile idiopathic arthritis identifies a new susceptibility locus at chromosomal region 3q13. *Arthritis Rheum.*, 64(8):2781–2791, 2012.

B. Tolusso, M. de Santis, S. Bosello, E. Gremese, S. Gobessi, I. Cuoghi, M. C. Totaro, G. Bigotti, C. Rumi, D. G. Efremov, and G. Ferraccioli. Synovial B cells of rheumatoid arthritis express ZAP-70 which increases the survival and correlates with the inflammatory and autoimmune phenotype. *Clin. Immunol.*, 131(1):98–108, 2009.

D. Tripathy and P. Grammas. Acetaminophen inhibits neuronal inflammation and protects neurons from oxidative stress. *J. Neuroinflammation*, 6:10, 2009. ISSN 1742-2094. doi: 10.1186/1742-2094-6-10.

O.-B. Tysnes and A. Storstein. Epidemiology of Parkinson's disease. *J. Neural Transm.*, 124(8):901–905, 2017. ISSN 0300-9564. doi: 10.1007/s00702-017-1686-y.

O. Ursu, J. Holmes, J. Knockel, C. G. Bologa, J. J. Yang, S. L. Mathias, S. J. Nelson, and T. I. Oprea. DrugCentral: online drug compendium. *Nucleic Acids Res.*, 45(D1): D932–D939, 2017.

V. S. van Laar, P. A. Otero, T. G. Hastings, and S. B. Berman. Potential role of Mic60/Mitofilin in Parkinson's disease. *Front. Neurosci.*, 12:898, 2018. ISSN 1662-4548, 1662-453X. doi: 10.3389/fnins.2018.00898.

G. van Rossum and F. L. Drake Jr. *Python tutorial*. Centrum voor Wiskunde en Informatica Amsterdam, The Netherlands, 1995.

Q. Vanhaelen, P. Mamoshina, A. M. Aliper, A. Artemov, K. Lezhnina, I. Ozerov, I. Labat, and A. Zhavoronkov. Design of efficient computational workflows for in silico drug repurposing. *Drug Discov. Today*, 22(2):210–222, 2017.

Y. Vasilopoulos, V. Gkretsi, M. Armaka, V. Aidinis, and G. Kollias. Actin cytoskeleton dynamics linked to synovial fibroblast activation as a novel pathogenic principle in TNF-driven arthritis. *Ann. Rheum. Dis.*, 66 Suppl 3:iii23–8, 2007.

L. Verhagen Metman, P. J. Blanchet, P. van den Munckhof, P. Del Dotto, R. Natté, and T. N. Chase. A trial of dextromethorphan in parkinsonian patients with motor response complications. *Mov. Disord.*, 13(3):414–417, 1998a. ISSN 0885-3185. doi: 10.1002/mds.870130307.

L. Verhagen Metman, P. Del Dotto, R. Natté, P. van den Munckhof, and T. N. Chase. Dextromethorphan improves levodopa-induced dyskinesias in Parkinson's disease. *Neurology*, 51(1):203–206, 1998b. ISSN 0028-3878. doi: 10.1212/wnl.51.1.203.

A. Verkhratsky, M. Matteoli, V. Parpura, J.-P. Mothet, and R. Zorec. Astrocytes as secretory cells of the central nervous system: idiosyncrasies of vesicular secretion. *EMBO J.*, 35 (3):239–257, 2016. ISSN 0261-4189, 1460-2075. doi: 10.15252/embj.201592705.

V. L. Villemagne, S. Burnham, P. Bourgeat, B. Brown, K. A. Ellis, O. Salvado, C. Szoeke, S. L. Macaulay, R. Martins, P. Maruff, D. Ames, C. C. Rowe, C. L. Masters, and Australian Imaging Biomarkers and Lifestyle (AIBL) Research Group. Amyloid β deposition, neurodegeneration, and cognitive decline in sporadic Alzheimer's disease:

a prospective cohort study. *Lancet Neurol.*, 12(4):357–367, 2013. ISSN 1474-4422, 1474-4465. doi: 10.1016/S1474-4422(13)70044-9.

R. von Coelln, S. Kügler, M. Bähr, M. Weller, J. Dichgans, and J. B. Schulz. Rescue from death but not from functional impairment: caspase inhibition protects dopaminergic cells against 6-hydroxydopamine-induced apoptosis but not against the loss of their terminals. *J. Neurochem.*, 77(1):263–273, 2001. ISSN 0022-3042. doi: 10.1046/j.1471-4159.2001.t01-1-00236.x.

D. von Maydell, K. Koler, M. Jorfi, E. Brand, J. Aronson, K. J. Washicosky, S. S. Kwak, M. Cetinbas, R. Sadreyev, J. Park, S. L. Wagner, W. Hide, R. E. Tanzi, and D. Y. Kim. Identifying shared enriched pathways driven by pathogenic Aβ accumulation in 3D cellular models and human Alzheimer's brain. Manuscript in Preparation, 2020.

K. Wada, H. Arai, M. Takanashi, J. Fukae, H. Oizumi, T. Yasuda, Y. Mizuno, and H. Mochizuki. Expression levels of vascular endothelial growth factor and its receptors in Parkinson's disease. *Neuroreport*, 17(7):705–709, 2006. ISSN 0959-4965. doi: 10.1097/01.wnr.0000215769.71657.65.

L. Wadi, M. Meyer, J. Weiser, L. D. Stein, and J. Reimand. Impact of outdated gene annotations on pathway enrichment analysis. *Nat. Methods*, 13(9):705–706, 2016.

M. G. Waisberg, A. C. M. Ribeiro, W. M. Candido, P. B. Medeiros, C. N. Matsuzaki, M. C. Beldi, M. Tacla, H. H. Caiaffa-Filho, E. Bonfa, and C. A. Silva. Human papillomavirus and chlamydia trachomatis infections in rheumatoid arthritis under anti-TNF therapy: an observational study. *Rheumatol. Int.*, 35(3):459–463, 2015. ISSN 0172-8172, 1437-160X. doi: 10.1007/s00296-014-3157-1.

S. R. Walmsley, C. Print, N. Farahi, C. Peyssonnaux, R. S. Johnson, T. Cramer, A. Sobolewski, A. M. Condliffe, A. S. Cowburn, N. Johnson, and E. R. Chilvers. Hypoxia-induced neutrophil survival is mediated by HIF-1alpha-dependent NF-kappaB activity. *J. Exp. Med.*, 201(1):105–115, 2005.

M. J. Walport. Complement. second of two parts. *N. Engl. J. Med.*, 344(15):1140–1144, 2001. ISSN 0028-4793. doi: 10.1056/NEJM200104123441506.

M. R. Walton, H. Gibbons, G. A. MacGibbon, E. Sirimanne, J. Saura, P. D. Gluckman, and M. Dragunow. PU.1 expression in microglia. *J. Neuroimmunol.*, 104(2):109–115, 2000. ISSN 0165-5728. doi: 10.1016/s0165-5728(99)00262-3.

C.-Y. Wang, W. Zheng, T. Wang, J.-W. Xie, S.-L. Wang, B.-L. Zhao, W.-P. Teng, and Z.-Y. Wang. Huperzine a activates Wnt/β-catenin signaling and enhances the nonamyloidogenic pathway in an Alzheimer transgenic mouse model. *Neuropsychopharmacology*, 36(5):1073–1089, 2011. ISSN 0893-133X, 1740-634X. doi: 10.1038/npp.2010.245.

R. Wang and X. C. Tang. Neuroprotective effects of huperzine a. a natural cholinesterase inhibitor for the treatment of Alzheimer's disease. *Neurosignals*, 14(1-2):71–82, 2005. ISSN 1424-862X. doi: 10.1159/000085387.

W. Wang, L. T. T. Nguyen, C. Burlak, F. Chegini, F. Guo, T. Chataway, S. Ju, O. S. Fisher, D. W. Miller, D. Datta, F. Wu, C.-X. Wu, A. Landeru, J. A. Wells, M. R. Cookson, M. B. Boxer, C. J. Thomas, W. P. Gai, D. Ringe, G. A. Petsko, and Q. Q. Hoang. Caspase-1 causes truncation and aggregation of the Parkinson's disease-associated protein α-synuclein. *Proc. Natl. Acad. Sci. U. S. A.*, 113(34):9587–9592, 2016a. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1610099113.

Y.-C. Wang, S.-L. Chen, N.-Y. Deng, and Y. Wang. Network predicting drug's anatomical therapeutic chemical code. *Bioinformatics*, 29(10):1317–1324, 2013.

Z. Wang, N. R. Clark, and A. Ma'ayan. Drug-induced adverse events prediction with the LINCS L1000 data. *Bioinformatics*, 32(15):2338–2345, 2016b.

Z. Wang, A. Lachmann, A. B. Keenan, and A. Ma'ayan. L1000FWD: fireworks visualization of drug-induced transcriptomic signatures. *Bioinformatics*, 34(12):2150–2152, 2018. ISSN 1367-4803, 1367-4811. doi: 10.1093/bioinformatics/bty060.

G. L. Warren, T. D. Do, B. P. Kelley, A. Nicholls, and S. D. Warren. Essential considerations for using protein-ligand structures in drug discovery. *Drug Discov. Today*, 17(23-24): 1270–1281, 2012.

S. Wax, M. Piecyk, B. Maritim, and P. Anderson. Geldanamycin inhibits the production of inflammatory cytokines in activated macrophages by reducing the stability and translation of cytokine transcripts. *Arthritis Rheum.*, 48(2):541–550, 2003. ISSN 0004-3591. doi: 10.1002/art.10780.

W. Wei, D. D. Norton, X. Wang, and J. W. Kusiak. Abeta 17-42 in Alzheimer's disease activates JNK and caspase-8 leading to neuronal apoptosis. *Brain*, 125(Pt 9):2036–2043, 2002. ISSN 0006-8950. doi: 10.1093/brain/awf205.

M. J. Weiss, S. Zhou, L. Feng, D. A. Gell, J. P. Mackay, Y. Shi, and A. J. Gow. Role of alpha-hemoglobin-stabilizing protein in normal erythropoiesis and beta-thalassemia. *Ann. N. Y. Acad. Sci.*, 1054:103–117, 2005.

B. L. Welch. The generalisation of student's problems when several different population variances are involved. *Biometrika*, 34(1-2):28–35, 1947. ISSN 0006-3444. doi: 10.1093/biomet/34.1-2.28.

M. Whirl-Carrillo, E. M. McDonagh, J. M. Hebert, L. Gong, K. Sangkuhl, C. F. Thorn, R. B. Altman, and T. E. Klein. Pharmacogenomics knowledge for personalized medicine. *Clin. Pharmacol. Ther.*, 92(4):414–417, 2012.

WHOCC. WHOCC - ATC structure and principles. https://www.whocc.no/atc/structure_and_principles/, 2018. Accessed: 2019-9-13.

H. Wickham. *stringr: Simple, Consistent Wrappers for Common String Operations*, 2019. URL https://CRAN.R-project.org/package=stringr. R package version 1.4.0.

H. Wickham and L. Henry. *tidyr: Tidy Messy Data*, 2019. URL https://CRAN.R-project.org/package=tidyr. R package version 1.0.0.

H. Wickham, R. François, L. Henry, and K. Müller. *dplyr: A Grammar of Data Manipulation*, 2019. URL https://CRAN.R-project.org/package=dplyr. R package version 0.8.3.

Wikipedia contributors. Wikipedia. https://en.wikipedia.org/w/index.php?title=Wikipedia&oldid=838173151, 2018. Accessed: 2018-4-25.

D. B. Williams. Beyond lectins: the calnexin/calreticulin chaperone system of the endoplasmic reticulum. *J. Cell Sci.*, 119(Pt 4):615–623, 2006. ISSN 0021-9533. doi: 10.1242/jcs.02856.

J. A. Williams, R. Hyland, B. C. Jones, D. A. Smith, S. Hurst, T. C. Goosen, V. Peterkin, J. R. Koup, and S. E. Ball. Drug-drug interactions for UDP-glucuronosyltransferase substrates: a pharmacokinetic explanation for typically observed low exposure (AUCi/AUC) ratios. *Drug Metab. Dispos.*, 32(11):1201–1208, 2004.

K. M. Wilton and E. L. Matteson. Malignancy incidence, management, and prevention in patients with rheumatoid arthritis. *Rheumatol Ther*, 4(2):333–347, 2017. ISSN 2198-6576. doi: 10.1007/s40744-017-0064-4.

D. S. Wishart, Y. D. Feunang, A. C. Guo, E. J. Lo, A. Marcu, J. R. Grant, T. Sajed, D. Johnson, C. Li, Z. Sayeeda, N. Assempour, I. Iynkkaran, Y. Liu, A. Maciejewski, N. Gale, A. Wilson, L. Chin, R. Cummings, D. Le, A. Pon, C. Knox, and M. Wilson. DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res.*, 46(D1):D1074–D1082, 2018.

G. Wohlgemuth, P. K. Haldiya, E. Willighagen, T. Kind, and O. Fiehn. The chemical translation service–a web-based tool to improve standardization of metabolomic reports. *Bioinformatics*, 26(20):2647–2648, 2010.

L. Wong, K. Jiang, Y. Chen, T. Hennon, L. Holmes, C. A. Wallace, and J. N. Jarvis. Limits of peripheral blood mononuclear cells for gene Expression-Based biomarkers in juvenile idiopathic arthritis. *Sci. Rep.*, 6:29477, 2016.

World Health Organization, WHO International Working Group for Drug Statistics Methodology, WHO Collaborating Centre for Drug Statistics Methodology, WHO Collaborating Centre for Drug Utilization Research, and Clinical Pharmacological Services. *Introduction to Drug Utilization Research*. WHO, 2003.

C.-C. Wu, S.-Y. Wu, C.-Y. Liao, C.-M. Teng, Y.-C. Wu, and S.-C. Kuo. The roles and mechanisms of PAR4 and P2Y12/phosphatidylinositol 3-kinase pathway in maintaining thrombin-induced platelet aggregation. *Br. J. Pharmacol.*, 161(3):643–658, 2010a. ISSN 0007-1188, 1476-5381. doi: 10.1111/j.1476-5381.2010.00921.x.

G. Wu, X. Feng, and L. Stein. A human functional protein interaction network and its application to cancer data analysis. *Genome Biol.*, 11(5):R53, 2010b.

Z. Wu, Y. Wang, and L. Chen. Network-based drug repositioning. *Mol. Biosyst.*, 9(6): 1268–1281, 2013.

R. Xu and Q. Wang. A genomics-based systems approach towards drug repositioning for rheumatoid arthritis. *BMC Genomics*, 17 Suppl 7(7):518, 2016.

F. Yang, J. Xu, and J. Zeng. Drug-Target interaction prediction by integrating chemical, genomic, functional and pharmacological data. In *Biocomputing 2014*, pages 148–159. WORLD SCIENTIFIC, 2013a.

G. Yang, Y. Wang, J. Tian, and J.-P. Liu. Huperzine a for Alzheimer's disease: a systematic review and meta-analysis of randomized clinical trials. *PLoS One*, 8(9):e74916, 2013b. ISSN 1932-6203. doi: 10.1371/journal.pone.0074916.

L. Yang and P. Agarwal. Systematic drug repositioning based on clinical side-effects. *PLoS One*, 6(12):e28025, 2011.

Z. Yang, R. Algesheimer, and C. J. Tessone. A comparative analysis of community detection algorithms on artificial networks. *Sci. Rep.*, 6:30750, 2016.

L. Yao, Y. Zhang, Y. Li, P. Sanseau, and P. Agarwal. Electronic health records: Implications for drug discovery. *Drug Discov. Today*, 16(13-14):594–599, 2011.

T. Yasuhara, T. Shingo, K. Kobayashi, A. Takeuchi, A. Yano, K. Muraoka, T. Matsui, Y. Miyoshi, H. Hamada, and I. Date. Neuroprotective effects of vascular endothelial growth factor (VEGF) upon dopaminergic neurons in a rat model of Parkinson's disease. *Eur. J. Neurosci.*, 19(6):1494–1504, 2004. ISSN 0953-816X. doi: 10.1111/j.1460-9568.2004.03254.x.

T. Yasuhara, T. Shingo, K. Muraoka, M. Kameda, T. Agari, Y. Wen Ji, H. Hayase, H. Hamada, C. V. Borlongan, and I. Date. Neurorescue effects of VEGF on a rat model of Parkinson's disease. *Brain Res.*, 1053(1-2):10–18, 2005. ISSN 0006-8993. doi: 10.1016/j.brainres.2005.05.027.

H. Ye, Q. Liu, and J. Wei. Construction of drug network based on side effects and its application for drug repositioning. *PLoS One*, 9(2):e87864, 2014.

K. G. Yiannopoulou and S. G. Papageorgiou. Current and future treatments for Alzheimer's disease. *Ther. Adv. Neurol. Disord.*, 6(1):19–33, 2013. ISSN 1756-2856. doi: 10.1177/1756285612461679.

S. O. Yoon, D. J. Park, J. C. Ryu, H. G. Ozer, C. Tep, Y. J. Shin, T. H. Lim, L. Pastorino, A. J. Kunwar, J. C. Walton, A. H. Nagahara, K. P. Lu, R. J. Nelson, M. H. Tuszynski, and K. Huang. JNK3 perpetuates metabolic stress induced by Aβ peptides. *Neuron*, 75 (5):824–837, 2012. ISSN 0896-6273, 1097-4199. doi: 10.1016/j.neuron.2012.06.024.

S. Yoshida, M. Ono, T. Shono, H. Izumi, T. Ishibashi, H. Suzuki, and M. Kuwano. Involvement of interleukin-8, vascular endothelial growth factor, and basic fibroblast growth factor in tumor necrosis factor alpha-dependent angiogenesis. *Mol. Cell. Biol.*, 17(7):4015–4023, 1997.

G. Yu. *enrichplot: Visualization of Functional Enrichment Result*, 2019. URL https://github.com/GuangchuangYu/enrichplot. R package version 1.5.2.

G. Yu, L.-G. Wang, Y. Han, and Q.-Y. He. clusterprofiler: an r package for comparing biological themes among gene clusters. *OMICS: A Journal of Integrative Biology*, 16 (5):284–287, 2012. doi: 10.1089/omi.2011.0118.

G. Yu, L.-G. Wang, G.-R. Yan, and Q.-Y. He. DOSE: an r/bioconductor package for disease ontology semantic and enrichment analysis. *Bioinformatics*, 31(4):608–609, 2015. doi: 10.1093/bioinformatics/btu684. URL http://bioinformatics.oxfordjournals.org/content/31/4/608.

R.-L. Yu, C.-H. Tan, Y.-C. Lu, and R.-M. Wu. Aldehyde dehydrogenase 2 is associated with cognitive functions in patients with Parkinson's disease. *Sci. Rep.*, 6:30424, 2016. ISSN 2045-2322. doi: 10.1038/srep30424.

S. V. Zaichick, K. M. McGrath, and G. Caraveo. The role of ca2+ signaling in Parkinson's disease. *Dis. Model. Mech.*, 10(5):519–535, 2017. ISSN 1754-8403, 1754-8411. doi: 10.1242/dmm.028738.

M. A. Zampol and M. H. Barros. Melatonin improves survival and respiratory activity of yeast cells challenged by alpha-synuclein and menadione. *Yeast*, 35(3):281–290, 2018. ISSN 0749-503X, 1097-0061. doi: 10.1002/yea.3296.

A. Zangara. The psychopharmacology of huperzine a: an alkaloid with cognitive enhancing and neuroprotective properties of interest in the treatment of Alzheimer's disease. *Pharmacol. Biochem. Behav.*, 75(3):675–686, 2003. ISSN 0091-3057. doi: 10.1016/s0091-3057(03)00111-4.

U. M. Zanger, M. Turpeinen, K. Klein, and M. Schwab. Functional pharmacogenetics/genomics of human cytochromes P450 involved in drug biotransformation. *Anal. Bioanal. Chem.*, 392(6):1093–1108, 2008.

D. Zhang, V. Anantharam, A. Kanthasamy, and A. G. Kanthasamy. Neuroprotective effect of protein kinase C delta inhibitor rottlerin in cell culture and animal models of Parkinson's disease. *J. Pharmacol. Exp. Ther.*, 322(3):913–922, 2007. ISSN 0022-3565. doi: 10.1124/jpet.107.124669.

H. G. Zhang, Y. Wang, J. F. Xie, X. Liang, D. Liu, P. Yang, H. C. Hsu, R. B. Ray, and J. D. Mountz. Regulation of tumor necrosis factor alpha-mediated apoptosis of rheumatoid arthritis synovial fibroblasts by the protein kinase akt. *Arthritis Rheum.*, 44 (7):1555–1567, 2001. ISSN 0004-3591. doi: 10.1002/1529-0131(200107)44:7<1555::AID-ART279>3.0.CO;2-M.

M. Zhang, G. Schmitt-Ulms, C. Sato, Z. Xi, Y. Zhang, Y. Zhou, P. St George-Hyslop, and E. Rogaeva. Drug repositioning for Alzheimer's disease based on systematic 'omics' data mining. *PLoS One*, 11(12):e0168812, 2016. ISSN 1932-6203. doi: 10.1371/journal.pone.0168812.

P. Zhang, P. Agarwal, and Z. Obradovic. Computational drug repositioning by ranking and integrating multiple data sources. In *Machine Learning and Knowledge Discovery in Databases*, pages 579–594. Springer, Berlin, Heidelberg, 2013.

B. Zhou, R. Wang, P. Wu, and D.-X. Kong. Drug repurposing based on drug-drug interaction. *Chem. Biol. Drug Des.*, 85(2):137–144, 2015.

X. Zhu, A. K. Raina, C. A. Rottkamp, G. Aliev, G. Perry, H. Boux, and M. A. Smith. Activation and redistribution of c-jun n-terminal kinase/stress activated protein kinase in degenerating neurons in Alzheimer's disease. *J. Neurochem.*, 76(2):435–441, 2001. ISSN 0022-3042. doi: 10.1046/j.1471-4159.2001.00046.x.

R. M. Zinkernagel and P. C. Doherty. MHC-restricted cytotoxic T cells: studies on the biological role of polymorphic major transplantation antigens determining t-cell restriction-specificity, function, and responsiveness. *Adv. Immunol.*, 27:51–177, 1979.

Appendix A

Supplementary material to Chapter 3 Materials and Methods

Table (A.1) Summary of drug synonyms extracted from source databases. Summary of synonym contribution from individual databases with specified unique identifiers and synonym types extracted for the development of the drug synonym database. ATC — Anatomical Therapeutic Chemical; BRD — Broad ID; CasRN — CAS Registry Number; ChEBI — Chemical Entities of Biological Interest; ChEMBL — Chemicals database by European Molecular Biology Laboratory; CID — Compound ID; CMap — Connectivity Map; CTD — Comparative Toxicogenomics Database; DB — database; DPD — Drugs Product Database; EINECS — European Inventory of Existing Commercial Chemical Substances; EMA — European Medicines Agency; FDA — the US Food and Drug Administration; InChI — the IUPAC International Chemical Identifier; IUPAC — International Union of Pure and Applied Chemistry; KEGG — Kyoto Encyclopedia of Genes and Genomes; L1000CDS^2 — LINCS L1000 characteristic direction signature search engine; LINCS — Library of Integrated Network-based Cellular Signatures; MESH — Medical Subject Headings; NDFRT — National Drug File - Reference Terminology; PharmGKB — Pharmacogenomics Knowledge Base; SMILES — Simplified molecular-input line-entry system; TTD — Therapeutic Target Database; UMLS — Unified Medical Language System; UNII — Unique Ingredient Identifier; URL — Uniform Resource Locator; WHO — World Health Organisation.

Database	Unique Identifier	Extracted name types	Used name types (after curation)
ATC	ATC	RxNorm, Synonym, WHO Name	RxNorm, Synonym, WHO Name
BindingDB	BindingDB	ChEBI, ChEMBL, PubChem CID, PubChem SID	ChEBI, ChEMBL, PubChem CID
ChEMBL	ChEMBL	ATC, ChEMBL Name, WHO Name	ChEMBL Name, WHO Name
CMap	CMap Name	Catalog Name, CMap Instance Id	CMap Instance Id, synonym
CMaptoATC	CMap Instance	ATC, CMap Name	CMap Name
CTD	CTD Name	CasRN, CTD Name, DrugBank, Synonym	CasRN, CTD Name, DrugBank, MESH, PubChem CID, synonym
DNI	Drug name	CasRN	CasRN
DrugBank	DrugBank	ATC, BindingDB, CasRN, ChEBI, ChEMBL, ChemSpider, DrugBank Name, Drugs Product Database (DPD), GenAtlas, GenBank, GenBank Gene Database, GenBank Protein Database, Guide to Pharmacology, HUGO Gene Nomenclature Committee (HGNC), IUPHAR, KEGG Compound, KEGG Drug, PDB, PharmGKB, PubChem Compound, PubChem Substance, Therapeutic Targets Database, UniProt Accession, UniProtKB, Wikipedia	BindingDB, CasRN, ChEBI, ChEMBL, ChemSpider, DrugBank Name, Drugs Product Database (DPD), KEGG, PharmGKB, PubChem CID, Wikipedia

continues on the next page

Table A.1 continued

Database	**Unique Identifier**	**Extracted name types**	**Used name types (after curation)**
DrugCentral	DrugCentral	CasRN, InChI, InChIKey, SMILES, Synonym	CasRN, InChI, InChIKey, SMILES, synonym
EMA	EMA	Active substance, ATC, EMA Name, Synonym	Active substance, EMA, EMA Name, synonym
KEGG	KEGG	ATC code, CAS, ChEBI, ChEMBL, Chemical structure group, DrugBank, KEGG Drug, LigandBox, NIKKAJI, PDB-CCD, PubChem, Same as, Synonym, Therapeutic category	CasRN, ChEBI, C
LINCS	LINCS	ChEBI, ChEMBL, DrugCentral, InChI, InChI Key, LINCS, MESH, PubChem CID, SMILES, Synonym, Target, IUPAC InChI, LINCS, LINCS Center Id	ChEBI, ChEMBL, DrugCentral, InChI, InChI Key, LINCS, MESH, PubChem CID, SMILES, synonym, LINCS Name, LINCS Center Id
PharmGKB	PharmGKB	ATC, BindingDB, ChEBI, Chemical Abstracts Service, ChemSpider, ClinicalTrials.gov, DrugBank, Drugs Product Database (DPD), FDA Drug Label at DailyMed, GenBank, HET, HMDB, InChI, IUPHAR Ligand, KEGG Compound, KEGG Drug, MedDRA, MeSH, National Drug Code Directory, NDFRT, PDB, PharmGKB Generic Name, PharmGKB Name, PharmGKB Trade Name, PubChem Compound, PubChem Substance, RxNorm, SMILES, Therapeutic Targets Database, UMLS, UniProtKB, URL	BindingDB, CasRN, ChEBI, ChemSpider, DrugBank, Drugs Product Database (DPD), FDA Drug Label at DailyMed, InChI, KEGG, MESH, National Drug Code Directory, NDFRT, PharmGKB, PharmGKB Name, PubChem CID, RxNorm, SMILES, synonym, UMLS, URL
RepoDB	DrugBank	DrugBank, RepoDB Name	DrugBank, RepoDB Name
RepurposeDB	RepurposeDB	ATC, CasRN, ChEBI, ChEMBL, DrugBank, InChI, KEGG, MESH, PubChem, RepurposeDB, SMILES, Synonym, Target entrez, Target entrez SEA, Target entrez Union, Target names, Target names SEA, Target names Union	CasRN, ChEBI, ChEMBL, DrugBank, InChI, KEGG, MESH, PubChem CID, RepurposeDB, SMILES, synonym
TTD	TTD	CAS Number, ChEBI, DrugName, Formular, PubChem CID, PubChem SID, SuperDrug ATC, SuperDrug CAS, TTD	CasRN, ChEBI, PubChem CID, synonym, TTD

continues on the next page

Table A.1 continued

Database	Unique Identifier	Extracted name types	Used name types (after curation)
Wikipedia	Wikipedia	CASnumber, CASNo, ChEBI, ChEMBL, ChemSpiderID, DrugBank, InChI, KEGG, StdInChI, StdInChIKey, UNII, Wikipedia, ATCprefix, CASnumber, EINECS, IUPACname, IUPACName, MeSHName, OtherNames, PubChem, SMILES, synonyms, tradename	CasRN, ChEBI, ChEMBL, ChemSpider, DrugBank, InChI, InChI Key, KEGG, UNII, Wikipedia, EINECS, IUPAC, KEGG, MESH, PubChem CID, SMILES, synonym

Table (A.2) Test case details for KATdb versus manual translation. The search parameters (search input name type, search goal name type) used in the KATdb visual interface (Supplementary Fig. B.3) are listed for each test case. Results from each test case are summarised in Chapter 4 Table 4.2. All translations were performed ignoring the letter case. * — unique perturbagen IDs from drug signatures below significance threshold p-value < 0.05 in L1000CDS^2 signature database. ATC — Anatomical Therapeutic Chemical; BRD ID — Broad ID; CMap — Connectivity Map; CTD — Comparative Toxicogenomics Database; EMA — European Medicines Agency; L1000CDS^2 — LINCS L1000 characteristic direction signature search engine; LINCS — Library of Integrated Network-based Cellular Signatures; PharmGKB — Pharmacogenomics Knowledge Base.

Test case	**Method**	**Input name source**	**Input name type**	**Goal name type**	**Search input name type**	**Search goal name type**
A	manual	L1000CDS2	BRD ID	name	n/a	n/a
B	KATdb	L1000CDS2	BRD ID	name	any	Active substance, ChEMBL Name, CMap Name, CTD Name, DrugBank Name, EMA Name, L1000CDS2, LINCS Name, PharmGKB Name, RepoDB Name, synonym, WHO Name, Wikipedia
C	KATdb	L1000CDS2	BRD ID	ATC	any	ATC
D	KATdb	L1000CDS2	BRD ID*	ATC	any	ATC
E	manual	RepoDB	RepoDB name	LINCS name	n/a	n/a
F	manual	RepoDB	RepoDB name	BRD ID	n/a	n/a
G	KATdb	RepoDB	RepoDB name	BRD ID	any	BRD, LINCS center ID, LINCS ID
H	KATdb	RepoDB	RepoDB DB	BRD ID	any	BRD, LINCS center ID, LINCS ID
I	KATdb	EMA	EMA name	BRD ID	any	BRD, LINCS center ID, LINCS ID
J	KATdb	EMA	EMA INN	BRD ID	any	BRD, LINCS center ID, LINCS ID
K	KATdb	RepoDB, EMA	RepoDB + EMA name	BRD ID	any	BRD, LINCS center ID, LINCS ID

Table (A.3) Summary of Gene Expression Omnibus (GEO) DataSet search results for juvenile idiopathic arthritis (JIA). Reasons for not being included are listed in the last column. 18 studies with less than 20 samples were not considered due to low sample numbers. Six studies were excluded because of no control samples. Two studies were excluded because they did not include any JIA samples. Three studies were excluded because they only included treated JIA samples. Five were analysed on not commonly used arrays and thus excluded due to their platform. One was excluded because of the study design, it investigated monozygotic twins. One study was a duplicate of another study and one was removed because it represented a super series, from which the individual series were already included in the search results. The remaining 16 studies were then curated by Lester Kobzik. Ten studies were selected based on having untreated systemic or polyarticular JIA with control samples in either PBMC or whole blood (Table 3.2). In addition, GSE79970 was removed, because the data was identified to be unsuitable to study JIA by the authors (Wong et al., 2016).

Accession	Control Samples	Disease Samples	Included	Reason
GSE20307	56	20	yes	y
GSE112057	12	26	yes	y
GSE21521	29	18	yes	y
GSE7753	30	17	yes	y
GSE80060	22	33	yes	y
GSE8650	21	16	yes	y
GSE15645	13	14	yes	not sJIA
GSE26554	23	3, 38	yes	mixed JIA
GSE17590	21	22	no	treated, platform
GSE54629	46	48	no	treated
GSE80325	5	7	no	superseries
GSE24060	6	6	no	study design
GSE66896	3	0	no	sample no, platform
E-MEXP-987	0	17	no	sample no
GSE103170	3	3	no	sample no
GSE103501	5	7	no	sample no
GSE122552	5	4	no	sample no
GSE38849	0	11	no	sample no
GSE57183	7	7	no	sample no
GSE58667	4	11	no	sample no
E-MTAB-3201	5	5	no	sample no
GSE15083	0	21	no	sample no
GSE23687	0	11	no	sample no
GSE71595	4	3	no	sample no
GSE83415	0	11	no	sample no
GSE92293	2	3	no	sample no
GSE41744	0	12	no	sample no
GSE66895	3	0	no	sample no
GSE66898	3	0	no	sample no
GSE37107	8	6	no	sample no
GSE8361	8	51	no	platform
GSE17755	8	51	no	platform
GSE29536	19	67	no	platform
GSE71010	43+40	35+38	no	platform
GSE13849	59	61	no	not sJIA, treated
GSE41831	15	14	no	not sjIA, treated
GSE79970	16	85	no	not sJIA, platform
GSE61281	12	20	no	not sJIA
GSE67596	15	14	no	not sJIA
GSE11083	15	14	no	not sJIA
GSE55319	19	26	no	not sJIA
GSE93777	43	202	no	not JIA, treated
GSE43553	43	17	no	not JIA
GSE103500	0	0	no	no controls

continues on the next page

Table A.3 continued

Accession	Control Samples	Disease Samples	Included	Reason
GSE11907	0	47	no	no controls
GSE11908	0	46	no	no controls
GSE89252	0	68	no	no controls
GSE26112	0	17	no	no controls
GSE94354	0	0	no	no controls
GSE13501	59	21	no	duplicate

Appendix B

Supplementary material to Chapter 4 KATdb, the Drug Synonym Database

Table (B.1) KATdb synonym name types, aspirin example. At least one synonym result from searching "aspirin" for each name type in KATdb. * — the values were not found in KATdb, but with a manual online search. ** — related term, not fully correct. There are more synonyms, than listed here in KATdb for the aspirin connected component. ATC — Anatomical Therapeutic Chemical; BRD — Broad ID; CasRN — CAS Registry Number; ChEBI — Chemical Entities of Biological Interest; ChEMBL — Chemicals database by European Molecular Biology Laboratory; CID — Compound ID; CMap — Connectivity Map; CTD — Comparative Toxicogenomics Database; DB — database; DPD — Drugs Product Database; ECHA — European Chemicals Agency; EINECS — European Inventory of Existing Commercial Chemical Substances; EMA — European Medicines Agency; FDA — the US Food and Drug Administration; GSRS — Global Substance Registration System; InChI — the IUPAC International Chemical Identifier; IUPAC — International Union of Pure and Applied Chemistry; KEGG — Kyoto Encyclopedia of Genes and Genomes; L1000CDS^2 — LINCS L1000 characteristic direction signature search engine; LINCS — Library of Integrated Network-based Cellular Signatures; MESH — Medical Subject Headings; NDFRT — National Drug File - Reference Terminology; PharmGKB — Pharmacogenomics Knowledge Base; SMILES — Simplified molecular-input line-entry system; TTD — Therapeutic Target Database; UMLS — Unified Medical Language System; UNII — Unique Ingredient Identifier; URL — Uniform Resource Locator; VHA — Veterans Health Administration; WHO — World Health Organisation; WHOCC — WHO Collaborating Centre.

Name type	Value	Governing body
ATC	A01AD05	WHOCC
RxNorm	C0004057	National Library of Medicine
WHO Name	acetylsalicylic acid	WHO
synonym	aspirin	various
BindingDB	BDBM22360	BindingDB
ChEBI	CHEBI15365	ChEBI

continues on the next page

Table B.1 continued

Name type	Value	Governing body
ChEMBL	CHEMBL25	ChEMBL
PubChem CID	CID2244	PubChem
ChEMBL Name	ASPIRIN, Aspirin, aspirin, acetylsalicylic acid, Acetylsalicylic acid, Acetyl salicyclic acid	ChEMBL
CMap Name	acetylsalicylic acid	CMap
CMap Instance ID	25, 506, 513, 602, 603, 626, 650, 727, 725, 757, 750, 767, 765	CMap
CTD Name	Aspirin	CTD
MESH	D001241	National Library of Medicine
CasRN	50-78-2	Chemical Abstracts Service
DrugBank	DB00945	DrugBank
DrugBank Name	Acetylsalicylic acid	DrugBank
Drugs Product Database (DPD)	150	Health Canada
KEGG	D00109	KEGG
PharmGKB	PA448497	PharmGKB
ChemSpider	2157	Royal Society of Chemistry
DrugCentral	74	DrugCentral
InChI	1S/C9H8O4/c1-6(10)13-8-5-3-2-4-7(8)9(11)12/h2-5H,1H3,(H,11,12)	IUPAC
InChI Key	BSYNRYMUTXBXSQ-UHFFFAOYSA-N	IUPAC
SMILES	CC(=O)OC1=CC=CC=C1C(O)=O, CC(=O)Oc1ccccc1C(=O)O	various
EMA Name	**Clopidogrel Qualimed	EMA
Active substance	*Clopidogrel/Acetylsalicylic acid Zentiva (previously DuoCover)	various
EMA ID	**EMEA/H/C/001135	EMA
BRD	BRD-K11433652	LINCS
L1000CDS2	*-666	LINCS
LINCS ID	LSM-5288	LINCS
LINCS Center ID	184	LINCS
LINCS Name	Aspirin	LINCS
FDA Drug Label at DailyMed	*82cc404b-fcf5-4e48-ab1a-09d8c47f9e04	FDA
National Drug Code Directory	*49348-980-23	FDA
NDFRT	N0000145918(ASPIRIN)	U.S. Department of Veterans Affairs, VHA
PharmGKB Name	aspirin	PharmGKB
UMLS	C0004057(Aspirin)	National Library of Medicine
URL	//en.wikipedia.org/wiki/Aspirin	various
RepoDB Name	Acetylsalicylic acid	FDA

continues on the next page

Table B.1 continued

Name type	**Value**	**Governing body**
TTD	D0GY5Z	TTD
Wikipedia	Aspirin	Wikipedia contributors
RepurposeDB	acetylsalicylic acid	RepurposeDB
IUPAC	2-acetoxybenzoic acid	IUPAC
UNII	R16CO5Y76E	GSRS of the FDA
EINECS	O-acetylsalicylic acid	ECHA

Table (B.2) Summary of individual source database contributions to the KATdb drug graph. NB: The node % add up to >100%, because some nodes are shared between databases. ATC — Anatomical Therapeutic Chemical; ChEMBL — Chemicals database by European Molecular Biology Laboratory; CMap — Connectivity Map; CTD — Comparative Toxicogenomics Database; DB — database; DNI — Drugs of New Indications (Liu et al., 2013); EMA — European Medicines Agency; KEGG — Kyoto Encyclopedia of Genes and Genomes; L1000CDS^2 — LINCS L1000 characteristic direction signature search engine; LINCS — Library of Integrated Network-based Cellular Signatures; PharmGKB — Pharmacogenomics Knowledge Base; TTD — Therapeutic Target Database.

Database name	**Nodes**	**Edges**
total	3305952 (100%)	2851870 (100%)
total with RepurposeDB	3309159 (100.1%)	2863598 (100.4%)
ATC	16886 (0.51%)	11418 (0.4%)
BindingDB	1912493 (58%)	1167919 (41%)
ChEMBL	27597 (0.83%)	24388 (0.86%)
CMap	8793 (0.27%)	7484 (0.26%)
CMAPtoATC	8420 (0.25%)	7135 (0.25%)
CTD	513435 (16%)	342479 (12%)
CTD-katkoler	159 (0.0048%)	82 (0.0029%)
DrugBank	87842 (2.7%)	74503 (2.6%)
DrugCentral	23892 (0.72%)	19910 (0.7%)
EMA	5248 (0.16%)	3873 (0.14%)
katkoler	174823 (5.3%)	111797 (3.9%)
KEGG	51548 (1.6%)	40991 (1.4%)
L1000CDS2	9575 (0.29%)	5183 (0.18%)
LINCS	434403 (13%)	793034 (28%)
Liu2013	426 (0.013%)	213 (0.0075%)
PharmGKB	34635 (1%)	31325 (1.1%)
RepoDB	1974 (0.06%)	987 (0.035%)
RepurposeDB	13036 (n/a)	11728 (n/a)
TTD	96879 (2.9%)	61841 (2.2%)
Wikipedia	163510 (4.9%)	145414 (5.1%)
Wikipedia-katkoler	2482 (0.075%)	1894 (0.066%)

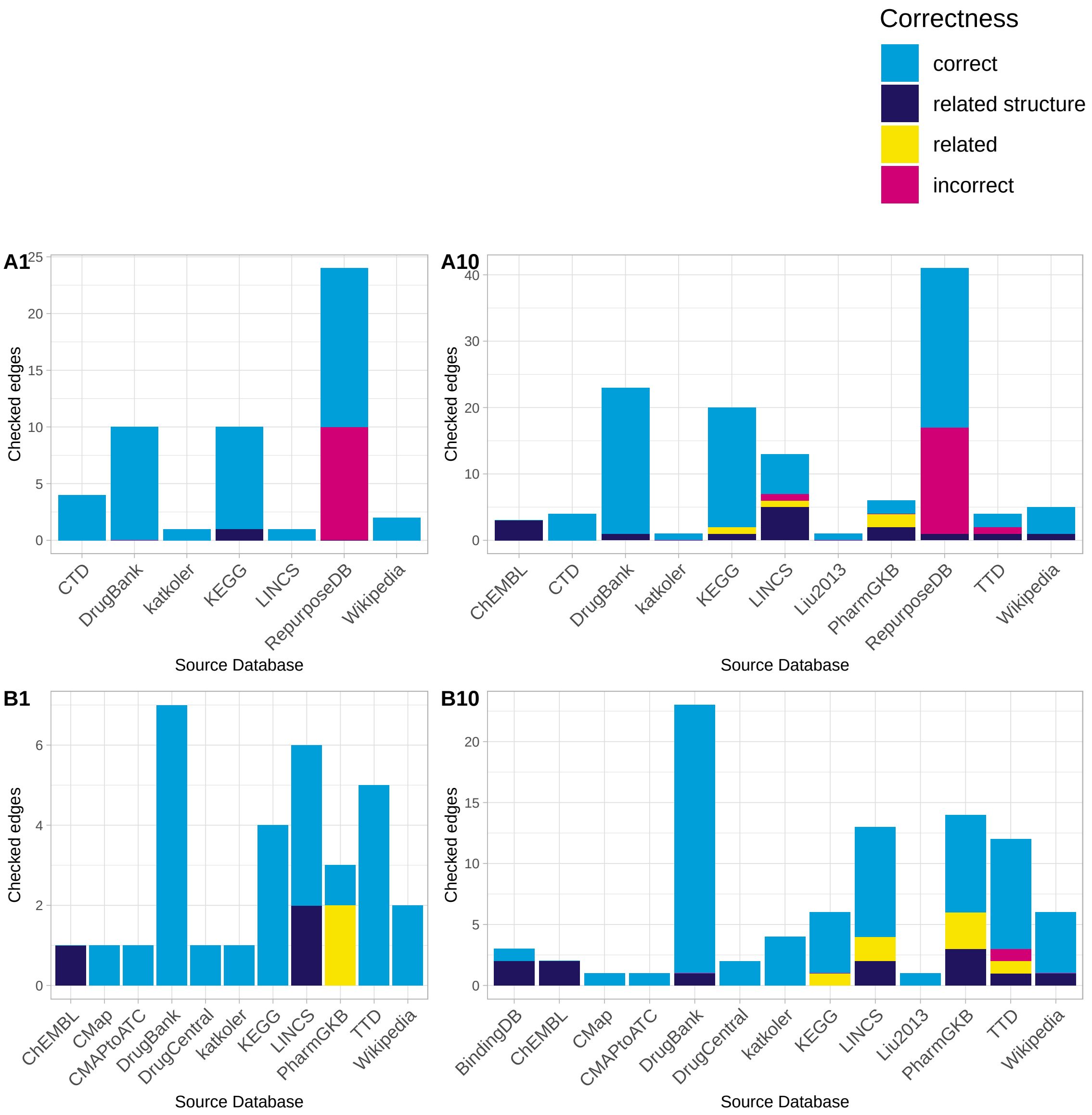
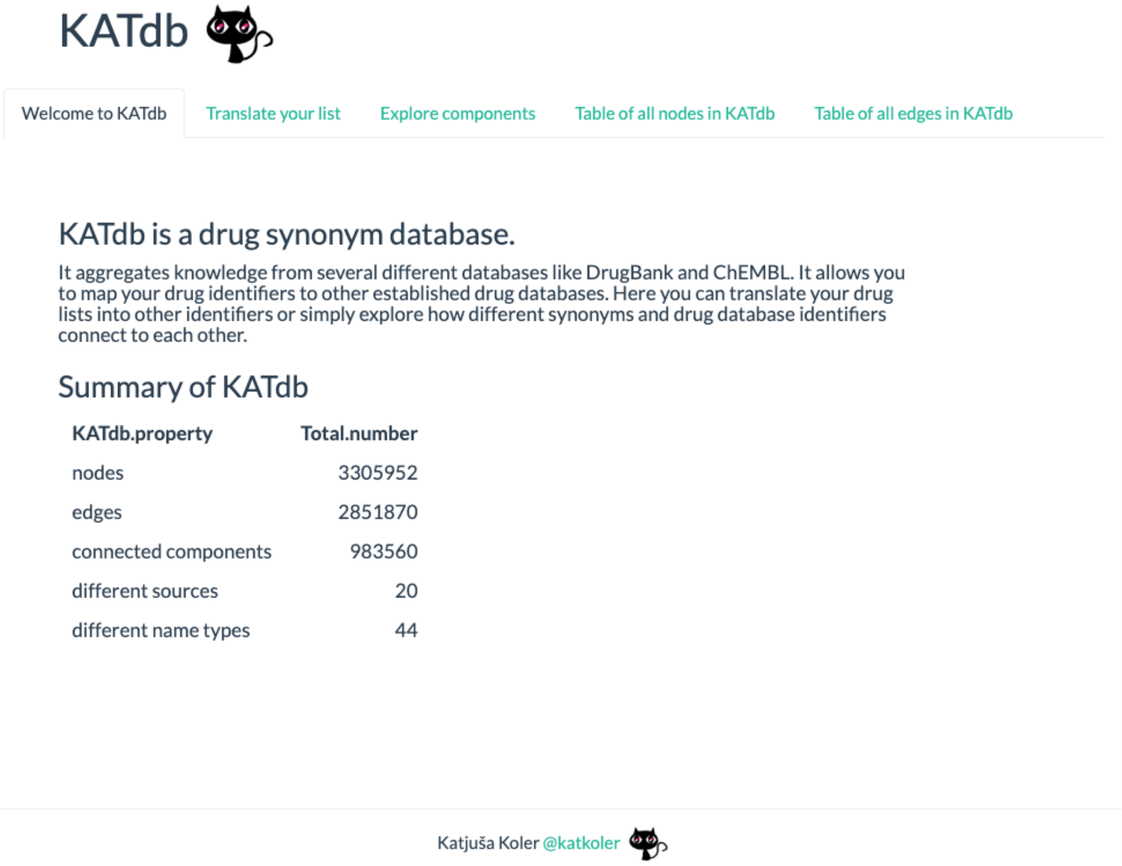


Fig. (B.1) KATdb synonym relationship correctness per source database. Correctness estimation using four different iterations. From the initial synonym database, the largest (A1) and ten largest (A10) components were assessed, and then reassessed (B1, B10, respectively) after removing RepurposeDB as a source database and reconstructing KATdb. We assigned one of 4 different levels of correctness to each manually checked synonym relationship. The correctness levels were summarised per source database. The ten largest components (A10, B10), representing the manually curated top 2% of edges with highest edge betweenness, are summarised. The labels on this plot match the experiment name on Fig.4.6 (page 75). NB: the *y*-axes are different on each plot. ATC — Anatomical Therapeutic Chemical; CMap — Connectivity Map; CTD — Comparative Toxicogenomics Database; DB — database; KEGG — Kyoto Encyclopedia of Genes and Genomes; LINCS — Library of Integrated Network-based Cellular Signatures; Liu2013 — Liu et al. (2013); PharmGKB — Pharmacogenomics Knowledge Base; TTD — Therapeutic Target Database.



KATdb

KATdb is a drug synonym database.

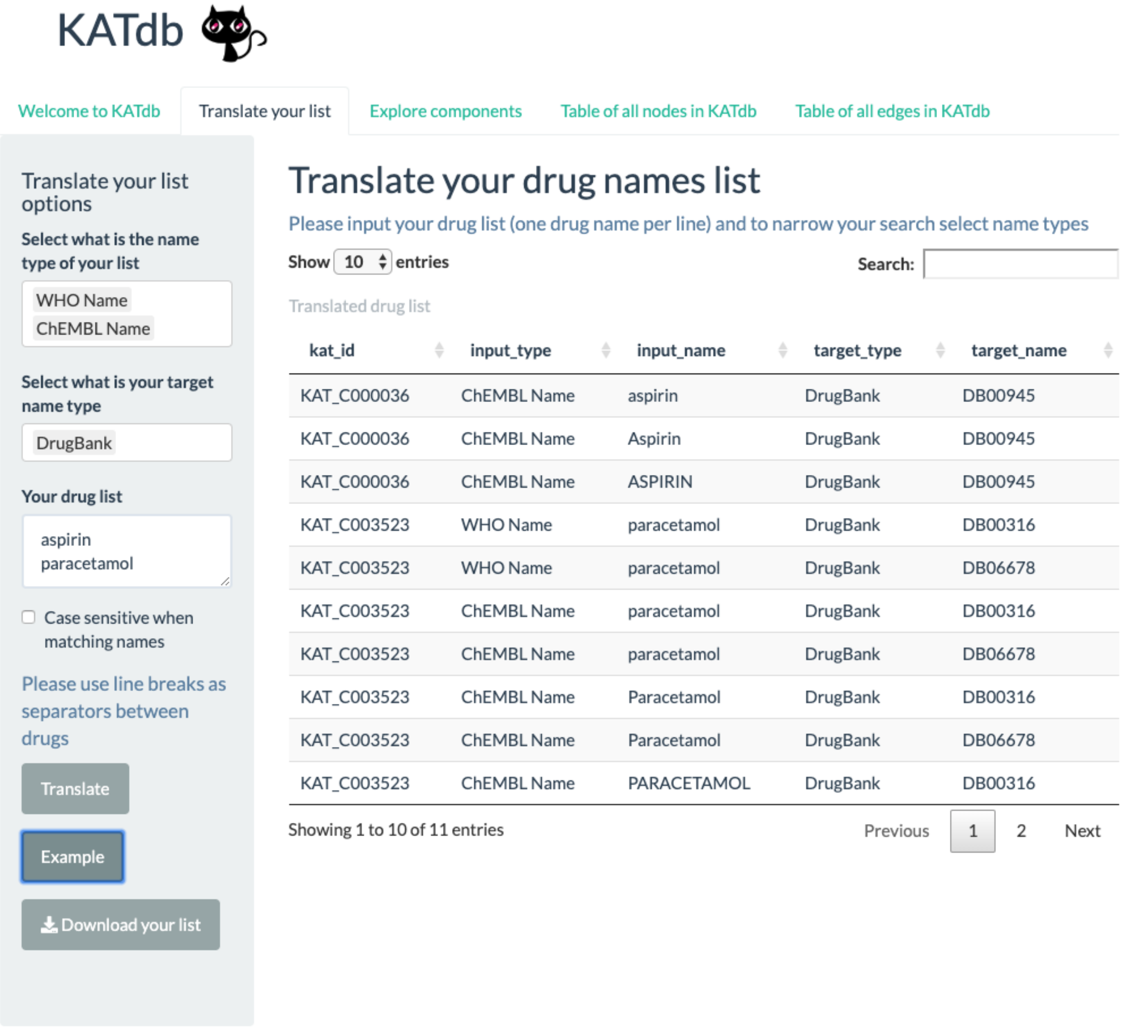
It aggregates knowledge from several different databases like DrugBank and ChEMBL. It allows you to map your drug identifiers to other established drug databases. Here you can translate your drug lists into other identifiers or simply explore how different synonyms and drug database identifiers connect to each other.

Summary of KATdb

KATdb.property	Total.number
nodes	3305952
edges	2851870
connected components	983560
different sources	20
different name types	44

Katjuša Koler @katkoler

Fig. (B.2) KATdb visual interface part 1 — landing page. The landing page introduces the main aim of the database and provides the database summary statistics.



Translate your list options

Select what is the name type of your list: WHO Name, ChEMBL Name

Select what is your target name type: DrugBank

Your drug list:
aspirin
paracetamol

□ Case sensitive when matching names

Please use line breaks as separators between drugs

Translate your drug names list

Please input your drug list (one drug name per line) and to narrow your search select name types

Show 10 entries

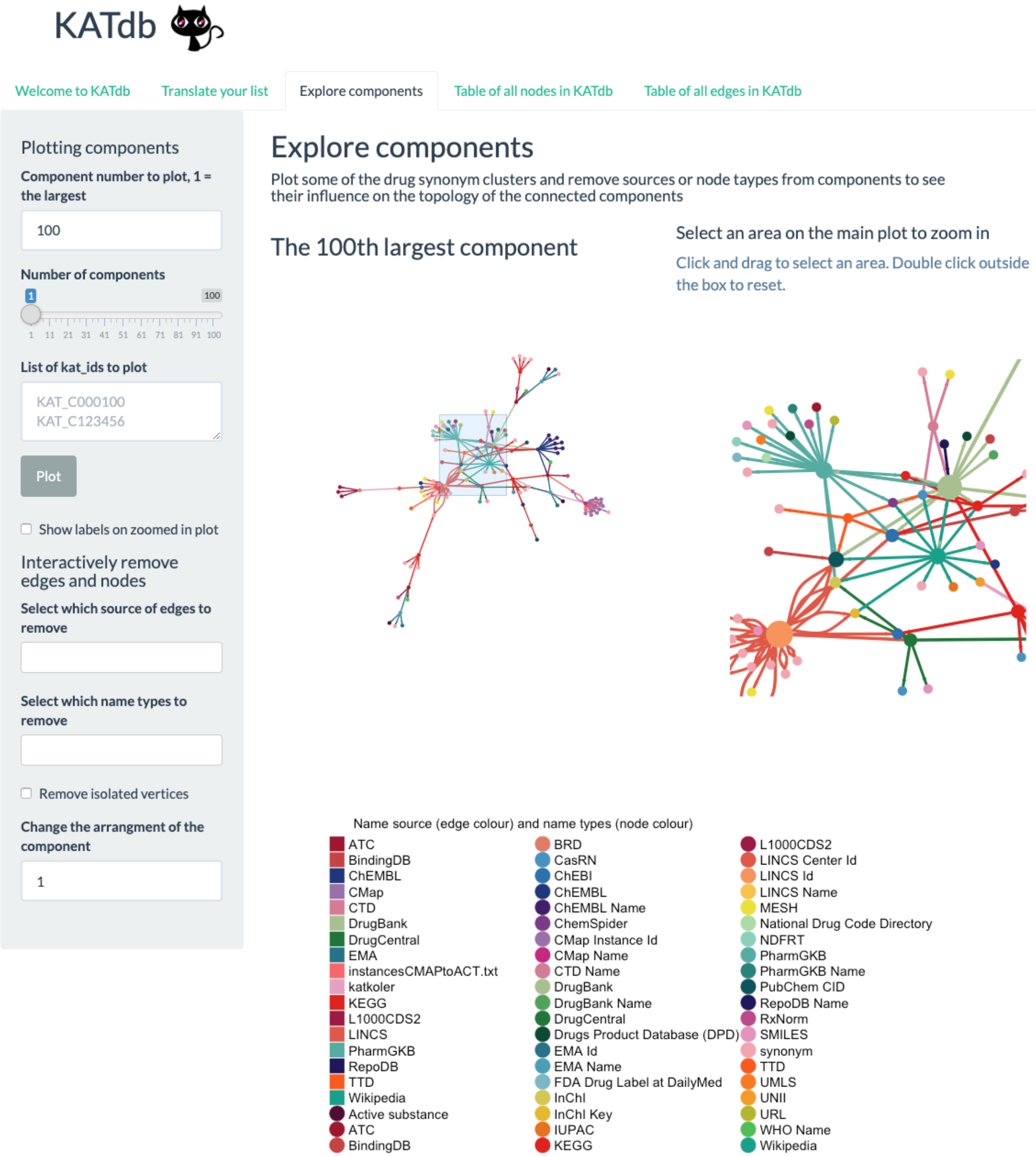
Search:

Translated drug list

kat_id	input_type	input_name	target_type	target_name
KAT_C000036	ChEMBL Name	aspirin	DrugBank	DB00945
KAT_C000036	ChEMBL Name	Aspirin	DrugBank	DB00945
KAT_C000036	ChEMBL Name	ASPIRIN	DrugBank	DB00945
KAT_C003523	WHO Name	paracetamol	DrugBank	DB00316
KAT_C003523	WHO Name	paracetamol	DrugBank	DB06678
KAT_C003523	ChEMBL Name	paracetamol	DrugBank	DB00316
KAT_C003523	ChEMBL Name	paracetamol	DrugBank	DB06678
KAT_C003523	ChEMBL Name	Paracetamol	DrugBank	DB00316
KAT_C003523	ChEMBL Name	Paracetamol	DrugBank	DB06678
KAT_C003523	ChEMBL Name	PARACETAMOL	DrugBank	DB00316

Showing 1 to 10 of 11 entries

Fig. (B.3) KATdb visual interface part 2 — translate your list. The main feature of KATdb can be accessed on the "translate your list" tab, where a user drug list can be translated into synonyms from specified or unspecified name types.



Plotting components

Component number to plot, 1 = the largest: 100

Number of components: 1 (1 11 21 31 41 51 61 71 81 91 100; max 100)

List of kat_ids to plot: KAT_C000100 KAT_C123456

□ Show labels on zoomed in plot

Interactively remove edges and nodes

Select which source of edges to remove:

Select which name types to remove:

□ Remove isolated vertices

Change the arrangement of the component: 1

Explore components

Plot some of the drug synonym clusters and remove sources or node taypes from components to see their influence on the topology of the connected components

The 100th largest component

Select an area on the main plot to zoom in

Click and drag to select an area. Double click outside the box to reset.

Name source (edge colour) and name types (node colour)

■ ATC
■ BindingDB
■ ChEMBL
■ CMap
■ CTD
■ DrugBank
■ DrugCentral
■ EMA
■ instancesCMAPtoACT.txt
■ katkoler
■ KEGG
■ L1000CDS2
■ LINCS
■ PharmGKB
■ RepoDB
■ TTD
■ Wikipedia
● Active substance
● ATC
● BindingDB
● BRD
● CasRN
● ChEBI
● ChEMBL
● ChEMBL Name
● ChemSpider
● CMap Instance Id
● CMap Name
● CTD Name
● DrugBank
● DrugBank Name
● DrugCentral
● Drugs Product Database (DPD)
● EMA Id
● EMA Name
● FDA Drug Label at DailyMed
● InChI
● InChI Key
● IUPAC
● KEGG
● L1000CDS2
● LINCS Center Id
● LINCS Id
● LINCS Name
● MESH
● National Drug Code Directory
● NDFRT
● PharmGKB
● PharmGKB Name
● PubChem CID
● RepoDB Name
● RxNorm
● SMILES
● synonym
● TTD
● UMLS
● UNII
● URL
● WHO Name
● Wikipedia

Fig. (B.4) KATdb visual interface part 3 — plotting components. Plotting components tab allows exploration of connections between synonyms. It can be used to estimate the "correctness" of the relationships (edges), highlighting any edges with high edge betweenness score (edge width).

Appendix C

Supplementary material to Chapter 5 The Drug Repositioning Pipeline

Table (C.1) Alzheimer's disease (AD) curated list used in Pita-Juárez et al. (2018) case study. Domain expert-curated (Professor Rudolph Tanzi, Harvard Medical School) list of genes associated with Alzheimer's disease identified via genome wide association studies (GWAS). This gene set was used in the AD case study in Pita-Juárez et al. (2018). Taken from Pita-Juárez et al. (2018) Supplementary Table S4. DOI: https://doi.org/10.1371/journal.pcbi.1006042.s007.

Association	Genes
Early Onset Linked	APP, PSEN1, PSEN2
Frontotemporal Lobar Degeneration Associated Genes	MAPT, GRN, CHMP2B, VCP, C9orf72, FUS, TARDBP, CTNND2, PTN, HAVCR1, NYAP2, RNASEL
Late Onset Genome-Wide Associated Highly Suggestive	ZNF3, NDUFS3, MTCH2, IGHV1-67, TP53INP1, ACE, ATXN1, HLA-DRA, HLA-DRB4, HLA-DQ-A1, HLA-DQB, HLA-DQB1, HLA-DQA1, DPYSL2, AX747894, RIN3, LGMN, GOLGA5, HS3ST1, SQSTM1, TREML2, NDUFAF6, ECHDC3, AP2A2, ADAMST20, IGH, SPPL2A, TRIP4, SCIMP
Late Onset Genome-Wide Significant	APOE, CD33, BIN1, PTK2B, CLU, ABCA7, CR1, PICALM, MS4A6A, MS4A4E, CD2AP, SORL1, SLC24A4, DSG2, INPP5D, MEF2C, NME8, ZCWPW1, CELF1, FERMT2, CASS4, ADAM10, TREM2, HLA-DRB5, HLA-DRB1

Table (C.2) Pathway membership to KEGG and Reactome pathway groups. Pathways used for annotating PCxN and PDxN. The PDxN and PCxN column indicate whether the pathway is present in PDxN or PCxN (q-value < 0.05). The pathway group entry applies to all member pathways in the following lines until a new pathway group is listed. Only pathways in either PDxN or PCxN are listed. The letter case has been lost in data processing, so some acronyms and capital letters are now in lowercase. KEGG — Kyoto Encyclopedia of Genes and Genomes; PCxN — Pathway Coexpression Network; PDxN — Pathway Drug Coexpression Network.

Pathway group	Member pathway	PDxN	PCxN
Amino acid metabolism (KEGG)	Alanine aspartate and glutamate metabolism	y	y
	Arginine and proline metabolism	y	y
	Cysteine and methionine metabolism	y	y
	Glycine serine and threonine metabolism	y	y
	Histidine metabolism	y	y
	Lysine degradation	y	y
	Phenylalanine metabolism	y	y
	Tryptophan metabolism	y	y
	Tyrosine metabolism	n	y
	Valine leucine and isoleucine biosynthesis	n	y
	Valine leucine and isoleucine degradation	y	y
Apoptosis (Reactome)	Apoptosis	y	y
	Apoptosis induced dna fragmentation	y	y
	Apoptotic cleavage of cellular proteins	y	y
	Apoptotic execution phase	y	y
	Intrinsic pathway for apoptosis	y	y
	Regulation of apoptosis	y	y
Cancer: overview (KEGG)	Pathways in cancer	y	y
Cancer: specific types (KEGG)	Acute myeloid leukemia	y	y
	Basal cell carcinoma	y	y
	Bladder cancer	y	y
	Chronic myeloid leukemia	y	y
	Colorectal cancer	y	y
	Endometrial cancer	y	y
	Glioma	y	y
	Melanoma	y	y
	Pancreatic cancer	y	y
	Prostate cancer	y	y
	Renal cell carcinoma	y	y
	Small cell lung cancer	y	y
	Thyroid cancer	y	y
Carbohydrate metabolism (KEGG)	Amino sugar and nucleotide sugar metabolism	y	y
	Ascorbate and aldarate metabolism	y	y
	Butanoate metabolism	y	y
	Citrate cycle tca cycle	y	y
	Fructose and mannose metabolism	y	y
	Galactose metabolism	y	y
	Glycolysis gluconeogenesis	y	y
	Glyoxylate and dicarboxylate metabolism	y	y
	Inositol phosphate metabolism	y	y
	Pentose and glucuronate interconversions	y	y
	Pentose phosphate pathway	y	y
	Propanoate metabolism	y	y
	Pyruvate metabolism	y	y
	Starch and sucrose metabolism	y	y

continues on the next page

Table C.2 continued

Pathway group	Member pathway	PDxN	PCxN
Cardiovascular disease (KEGG)	Arrhythmogenic right ventricular cardiomyopathy arvc	y	y
	Hypertrophic cardiomyopathy hcm	y	y
	Viral myocarditis	y	y
Cell cell communication (Reactome)	Adherens junctions interactions	y	y
	Cell junction organization	y	y
	Tight junction interactions	y	y
Cell cycle (Reactome)	Activation of atr in response to replication stress	y	y
	Autodegradation of the e3 ubiquitin ligase cop1	y	y
	Cell cycle	y	y
	Cell cycle checkpoints	y	y
	Chromosome maintenance	y	y
	Dna strand elongation	y	y
	E2f mediated regulation of dna replication	y	y
	Extension of telomeres	y	y
	G0 and early g1	y	y
	G1 phase	y	y
	Lagging strand synthesis	y	y
	Loss of nlp from mitotic centrosomes	y	y
	Mitotic prometaphase	y	y
	Orc1 removal from chromatin	y	y
	Packaging of telomere ends	y	y
	Processive synthesis on the lagging strand	y	y
	Recruitment of mitotic centrosome proteins and complexes	y	y
	Recruitment of numa to mitotic centrosomes	y	y
	Regulation of mitotic cell cycle	y	y
	S phase	y	y
	Telomere maintenance	y	y
	Unwinding of dna	y	y
Cell growth and death (KEGG)	Apoptosis	y	y
	Cell cycle	y	y
	Oocyte meiosis	y	y
	P53 signaling pathway	y	y
Cell motility (KEGG)	Regulation of actin cytoskeleton	y	y
Cellular community - eukaryotes (KEGG)	Adherens junction	y	y
	Focal adhesion	y	y
	Gap junction	y	y
	Tight junction	y	y
Circadian clock (Reactome)	Circadian clock	y	y
Circulatory system (KEGG)	Cardiac muscle contraction	y	y
	Vascular smooth muscle contraction	y	y
Development and regeneration (KEGG)	Axon guidance	y	y
Developmental biology (Reactome)	Axon guidance	y	y
	Crmps in sema3a signaling	y	y
	Dcc mediated attractive signaling	y	y
	Developmental biology	y	y

continues on the next page

Table C.2 continued

Pathway group	Member pathway	PDxN	PCxN
	Dscam interactions	y	y
	Interaction between l1 and ankyrins	y	y
	L1cam interactions	y	y
	Myogenesis	y	y
	Ncam1 interactions	y	y
	Other semaphorin interactions	y	y
	Recycling pathway of l1	y	y
	Regulation of gene expression in beta cells	y	y
	Sema3a pak dependent axon repulsion	y	y
	Sema4d in semaphorin signaling	y	y
	Semaphorin interactions	y	y
	Signal transduction by l1	y	y
	Transcriptional regulation of white adipocyte differentiation	y	y
Digestion absorption (Reactome)	Digestion of dietary carbohydrate	y	n
Disease (Reactome)	Activated point mutants of fgfr2	n	y
	Binding and entry of hiv virion	y	y
	Early phase of hiv life cycle	y	y
	Hiv infection	y	y
	Hiv life cycle	y	y
	Host interactions of hiv factors	y	y
	Influenza life cycle	y	y
	Influenza viral rna transcription and replication	y	y
	Integration of provirus	y	y
	Interactions of vpr with host cellular proteins	y	y
	Late phase of hiv life cycle	y	y
	Membrane binding and targetting of gag proteins	y	y
	Nef mediated downregulation of mhc class i complex cell surface expression	y	y
	Signaling by activated point mutants of fgfr1	n	y
	Signaling by egfr in cancer	y	y
	Signaling by fgfr in disease	y	y
	Transport of ribonucleoproteins into the host nucleus	y	y
	Viral messenger rna synthesis	y	y
DNA repair (Reactome)	Base excision repair	y	y
	Dna repair	y	y
	Fanconi anemia pathway	y	y
	Nucleotide excision repair	y	y
DNA replication (Reactome)	Dna replication	y	y
	Synthesis of dna	y	y
Endocrine and metabolic disease (KEGG)	Maturity onset diabetes of the young	y	y
	Type i diabetes mellitus	y	y
	Type ii diabetes mellitus	y	y
Endocrine system (KEGG)	Adipocytokine signaling pathway	y	y
	Gnrh signaling pathway	y	y
	Insulin signaling pathway	y	y
	Melanogenesis	y	y
	Ppar signaling pathway	y	y
Energy metabolism (KEGG)	Nitrogen metabolism	y	y
	Oxidative phosphorylation	y	y
	Sulfur metabolism	y	y

continues on the next page

Table C.2 continued

Pathway group	Member pathway	PDxN	PCxN
Excretory system (KEGG)	Proximal tubule bicarbonate reclamation	y	y
Extracellular matrix organisation (Reactome)	Collagen formation	y	y
	Degradation of the extracellular matrix	y	y
	Extracellular matrix organization	y	y
	Integrin cell surface interactions	y	y
Folding, sorting and degradation (KEGG)	Proteasome	y	y
	Protein export	y	y
	Rna degradation	y	y
	Snare interactions in vesicular transport	n	y
	Ubiquitin mediated proteolysis	y	y
Glycan biosynthesis and metabolism (KEGG)	Glycosaminoglycan biosynthesis keratan sulfate	y	y
	Glycosaminoglycan degradation	y	y
	Glycosphingolipid biosynthesis ganglio series	y	y
	Glycosphingolipid biosynthesis lacto and neolacto series	y	y
	Other glycan degradation	y	y
Hemostasis (Reactome)	Basigin interactions	y	y
	Cell surface interactions at the vascular wall	y	y
	Cgmp effects	y	y
	Factors involved in megakaryocyte development and platelet production	y	y
	Hemostasis	y	y
	Kinesins	y	y
	Nitric oxide stimulates guanylate cyclase	y	y
	P130cas linkage to mapk signaling for integrins	y	y
	Pecam1 interactions	y	y
	Platelet adhesion to exposed collagen	y	y
	Platelet calcium homeostasis	n	y
	Platelet homeostasis	y	y
	Platelet sensitization by ldl	y	y
	Prostacyclin signalling through prostacyclin receptor	y	y
	Signal amplification	y	y
	Thromboxane signalling through tp receptor	y	y
	Tie2 signaling	y	y
Immune disease (KEGG)	Allograft rejection	y	y
	Asthma	y	y
	Autoimmune thyroid disease	y	y
	Primary immunodeficiency	y	y
	Systemic lupus erythematosus	y	y
Immune system (KEGG)	Antigen processing and presentation	y	y
	B cell receptor signaling pathway	y	y
	Chemokine signaling pathway	y	y
	Complement and coagulation cascades	y	y
	Fc epsilon ri signaling pathway	y	y
	Hematopoietic cell lineage	y	y
	Intestinal immune network for iga production	y	y
	Leukocyte transendothelial migration	y	y
	Natural killer cell mediated cytotoxicity	y	y
	T cell receptor signaling pathway	y	y
Immune system (Reactome)	Activated tak1 mediates p38 mapk activation	y	y

continues on the next page

Table C.2 continued

Pathway group	Member pathway	PDxN	PCxN
	Adaptive immune system	y	y
	Advanced glycosylation endproduct receptor signaling	y	y
	Beta defensins	y	y
	Cd28 dependent vav1 pathway	y	y
	Complement cascade	y	y
	Costimulation by the cd28 family	y	y
	Creation of c4 and c2 activators	y	y
	Ctla4 inhibitory signaling	y	y
	Cytokine signaling in immune system	y	y
	Defensins	y	y
	Downstream tcr signaling	y	y
	Erks are inactivated	y	y
	Generation of second messenger molecules	y	y
	Growth hormone receptor signaling	y	y
	Ikk complex recruitment mediated by rip1	y	y
	Immune system	y	y
	Inflammasomes	y	y
	Initial triggering of complement	y	y
	Innate immune system	y	y
	Interferon gamma signaling	y	y
	Interferon signaling	y	y
	Irak1 recruits ikk complex	y	y
	Mhc class ii antigen presentation	y	y
	Phosphorylation of cd3 and tcr zeta chains	y	y
	Pi3k cascade	y	y
	Rap1 signalling	n	y
	Regulation of complement cascade	y	y
	Regulation of ifna signaling	y	y
	Regulation of ifng signaling	y	y
	Regulation of signaling by cbl	y	y
	Tak1 activates nfkb by phosphorylation and activation of ikks complex	y	y
	Tcr signaling	y	y
	The nlrp3 inflammasome	y	y
	Traf6 mediated irf7 activation	y	y
	Trafficking and processing of endosomal tlr	y	y
Infectious disease: bacterial (KEGG)	Epithelial cell signaling in helicobacter pylori infection	y	y
	Pathogenic escherichia coli infection	y	y
	Vibrio cholerae infection	y	y
Lipid metabolism (KEGG)	Arachidonic acid metabolism	y	y
	Biosynthesis of unsaturated fatty acids	y	y
	Ether lipid metabolism	y	y
	Glycerolipid metabolism	y	y
	Glycerophospholipid metabolism	y	y
	Linoleic acid metabolism	y	y
	Primary bile acid biosynthesis	y	y
	Sphingolipid metabolism	y	y
	Steroid biosynthesis	y	y
	Steroid hormone biosynthesis	y	y
Meiosis (Reactome)	Meiosis	y	y
	Meiotic recombination	y	y
	Meiotic synapsis	y	y
Membrane transport (KEGG)	Abc transporters	n	y

continues on the next page

Table C.2 continued

Pathway group	Member pathway	PDxN	PCxN
Metabolism (Reactome)	A tetrasaccharide linker sequence is required for gag synthesis	y	y
	Acyl chain remodelling of pc	y	y
	Acyl chain remodelling of pe	n	y
	Acyl chain remodelling of pg	n	y
	Acyl chain remodelling of pi	y	y
	Acyl chain remodelling of ps	n	y
	Androgen biosynthesis	n	y
	Bile acid and bile salt metabolism	y	y
	Biological oxidations	y	y
	Cholesterol biosynthesis	y	y
	Chondroitin sulfate biosynthesis	y	y
	Cytosolic sulfonation of small molecules	y	y
	Endogenous sterols	y	y
	Ethanol oxidation	y	y
	Formation of atp by chemiosmotic coupling	y	y
	Glucagon signaling in metabolic regulation	y	y
	Gluconeogenesis	y	y
	Glucose metabolism	y	y
	Glucuronidation	y	y
	Glutathione conjugation	y	y
	Glycerophospholipid biosynthesis	y	y
	Glycolysis	y	y
	Glycosaminoglycan metabolism	y	y
	Glycosphingolipid metabolism	y	y
	Hyaluronan metabolism	y	y
	Hyaluronan uptake and degradation	y	y
	Integration of energy metabolism	y	y
	Keratan sulfate biosynthesis	y	y
	Keratan sulfate degradation	y	y
	Metabolism of amino acids and derivatives	y	y
	Metabolism of carbohydrates	y	y
	Metabolism of nucleotides	y	y
	Metabolism of polyamines	y	y
	Metabolism of porphyrins	y	y
	Metabolism of vitamins and cofactors	y	y
	Peroxisomal lipid metabolism	y	y
	Phospholipid metabolism	y	y
	Pi metabolism	y	y
	Ppara activates gene expression	y	y
	Purine catabolism	y	n
	Purine ribonucleoside monophosphate biosynthesis	y	y
	Purine salvage	y	y
	Pyrimidine catabolism	y	y
	Pyruvate metabolism	y	y
	Recycling of bile acids and salts	y	n
	Regulation of glucokinase by glucokinase regulatory protein	y	y
	Regulation of insulin secretion	y	y
	Respiratory electron transport	y	y
	Reversible hydration of carbon dioxide	y	y
	Sphingolipid de novo biosynthesis	y	y
	Sphingolipid metabolism	y	y
	Sulfur amino acid metabolism	y	y
	Synthesis of bile acids and bile salts	y	y
	Synthesis of pa	n	y
	Synthesis of pc	n	y
	Synthesis of pe	y	y
	Synthesis of pips at the early endosome membrane	y	y

continues on the next page

Table C.2 continued

Pathway group	Member pathway	PDxN	PCxN
	Synthesis of pips at the golgi membrane	y	y
	Synthesis of pips at the late endosome membrane	y	y
	Synthesis of pips at the plasma membrane	y	y
	Triglyceride biosynthesis	y	y
	Tryptophan catabolism	y	n
	Xenobiotics	y	y
Metabolism of cofactors and vitamins (KEGG)	Folate biosynthesis	y	y
	Nicotinate and nicotinamide metabolism	y	n
	One carbon pool by folate	y	y
	Pantothenate and coa biosynthesis	y	y
	Porphyrin and chlorophyll metabolism	y	y
	Retinol metabolism	y	y
Metabolism of other amino acids (KEGG)	Glutathione metabolism	y	y
Metabolism of proteins (Reactome)	Cytosolic trna aminoacylation	y	y
	Glycoprotein hormones	y	y
	Metabolism of proteins	y	y
	Mitochondrial trna aminoacylation	y	y
	Peptide chain elongation	y	y
	Peptide hormone biosynthesis	y	y
	Protein folding	y	y
	Translation	y	y
	Transport to the golgi and subsequent modification	y	y
	Trna aminoacylation	y	y
Metabolism of RNA (Reactome)	Deadenylation of mrna	y	y
	Metabolism of rna	y	y
	Mrna capping	y	y
	Mrna splicing	y	y
	Transport of mature mrna derived from an intronless transcript	y	y
	Transport of mature transcript to cytoplasm	y	y
Metabolism of terpenoids and polyketides (KEGG)	Limonene and pinene degradation	n	y
	Terpenoid backbone biosynthesis	y	y
Muscle contraction (Reactome)	Muscle contraction	y	y
	Smooth muscle contraction	y	y
	Striated muscle contraction	y	y
Nervous system (KEGG)	Neurotrophin signaling pathway	y	y
Neurodegenerative disease (KEGG)	Amyotrophic lateral sclerosis als	y	y
	Prion diseases	y	y
Neuronal system (Reactome)	Acetylcholine binding and downstream events	y	y
	Acetylcholine neurotransmitter release cycle	y	y
	Activation of kainate receptors upon glutamate binding	y	y
	Adenylate cyclase inhibitory pathway	y	y
	Dopamine neurotransmitter release cycle	y	y
	Gaba b receptor activation	y	y

continues on the next page

Table C.2 continued

Pathway group	Member pathway	PDxN	PCxN
	Gaba receptor activation	y	y
	Glutamate neurotransmitter release cycle	y	y
	Highly calcium permeable postsynaptic nicotinic acetylcholine receptors	y	y
	Ionotropic activity of kainate receptors	y	y
	Neuronal system	y	y
	Neurotransmitter release cycle	y	y
	Norepinephrine neurotransmitter release cycle	y	y
	Post nmda receptor activation events	y	y
	Potassium channels	y	y
	Presynaptic nicotinic acetylcholine receptors	y	y
	Tandem pore domain potassium channels	y	y
	Trafficking of ampa receptors	y	y
	Transmission across chemical synapses	y	y
	Voltage gated potassium channels	y	y
Nucleotide metabolism (KEGG)	Purine metabolism	y	y
	Pyrimidine metabolism	y	y
Protein families: genetic information processing (KEGG)	Proteasome	y	y
	Ribosome	y	y
	Spliceosome	y	y
Protein localisation (Reactome)	Mitochondrial protein import	y	y
Replication and repair (KEGG)	Base excision repair	y	y
	Dna replication	y	y
	Homologous recombination	y	y
	Mismatch repair	y	y
	Nucleotide excision repair	y	y
Sensory system (KEGG)	Olfactory transduction	y	y
	Taste transduction	y	y
Signal transduction (KEGG)	Calcium signaling pathway	y	y
	Erbb signaling pathway	y	y
	Hedgehog signaling pathway	y	y
	Mapk signaling pathway	y	y
	Mtor signaling pathway	y	y
	Notch signaling pathway	y	y
	Phosphatidylinositol signaling system	y	y
	Vegf signaling pathway	y	y
	Wnt signaling pathway	y	y
Signaling molecules and interaction (KEGG)	Cell adhesion molecules cams	y	y
Signalling (Reactome)	Activated notch1 transmits signal to the nucleus	n	y
	Adenylate cyclase activating pathway	y	y
	Akt phosphorylates targets in the cytosol	y	y
	Chemokine receptors bind chemokines	y	y
	Dag and ip3 signaling	y	y
	Downstream signal transduction	y	y
	Effects of pip2 hydrolysis	y	y
	Egfr downregulation	y	y
	Fgfr1 ligand binding and activation	y	y

continues on the next page

Table C.2 continued

Pathway group	Member pathway	PDxN	PCxN
	Fgfr2c ligand binding and activation	n	y
	Fgfr4 ligand binding and activation	n	y
	Gab1 signalosome	y	y
	Gpcr ligand binding	y	y
	Grb2 events in erbb2 signaling	y	y
	Insulin receptor recycling	n	y
	Insulin receptor signalling cascade	y	y
	Integrin alphaiib beta3 signaling	y	y
	Notch1 intracellular domain regulates transcription	y	y
	Nrage signals death through jnk	y	y
	Nrif signals cell death from the nucleus	n	y
	Nuclear signaling by erbb4	y	y
	Olfactory signaling pathway	y	y
	Opioid signalling	y	y
	Opsins	y	n
	P2y receptors	y	y
	P38mapk events	y	y
	P75ntr recruits signalling complexes	y	y
	Pi3k events in erbb2 signaling	y	y
	Pi3k events in erbb4 signaling	y	y
	Pip3 activates akt signaling	y	y
	Plc beta mediated events	y	y
	Prolonged erk activation events	y	y
	Prostanoid ligand receptors	y	y
	Regulated proteolysis of p75ntr	y	y
	Regulation of kit signaling	y	y
	Retrograde neurotrophin signalling	y	y
	Serotonin receptors	y	y
	Shc1 events in egfr signaling	y	y
	Shc1 events in erbb4 signaling	y	y
	Signal attenuation	y	y
	Signaling by bmp	y	y
	Signaling by erbb2	y	y
	Signaling by erbb4	y	y
	Signaling by fgfr	y	y
	Signaling by gpcr	y	y
	Signaling by hippo	y	y
	Signaling by insulin receptor	y	y
	Signaling by notch	y	y
	Signaling by notch1	y	y
	Signaling by notch2	y	y
	Signaling by notch3	y	y
	Signaling by notch4	y	y
	Signaling by pdgf	y	y
	Signaling by rho gtpases	y	y
	Signaling by wnt	y	y
	Signalling to erks	y	y
	Signalling to p38 via rit and rin	y	y
	Signalling to ras	y	y
	Spry regulation of fgf signaling	y	y
Transcription (KEGG)	Basal transcription factors	y	y
	Rna polymerase	y	y
	Spliceosome	y	y
Transcription (Reactome)	Formation of rna pol ii elongation complex	y	y
	Generic transcription pathway	y	y
	Nuclear receptor transcription pathway	y	y
Translation (KEGG)	Ribosome	y	y

continues on the next page

Table C.2 continued

Pathway group	Member pathway	PDxN	PCxN
Transport and catabolism (KEGG)	Endocytosis	y	y
	Lysosome	y	y
	Peroxisome	y	y
Transport of small molecules (Reactome)	Amino acid transport across the plasma membrane	y	y
	Ion channel transport	n	y
	Iron uptake and transport	y	y
	Metal ion slc transporters	y	y
	Passive transport by aquaporins	y	y
	Transferrin endocytosis and recycling	y	y
	Zinc transporters	y	y
Vesicle mediated transport (Reactome)	Gap junction assembly	y	y
	Gap junction degradation	y	y
	Gap junction trafficking	y	y
	Golgi associated vesicle biogenesis	y	y
	Lysosome vesicle biogenesis	y	y
	Membrane trafficking	y	y
Xenobiotics biodegradation and metabolism (KEGG)	Drug metabolism cytochrome p450	y	y
	Drug metabolism other enzymes	y	y
	Metabolism of xenobiotics by cytochrome p450	y	y

Table (C.3) Anatomical Therapeutic Chemical (ATC) level 1 and level 2 classification present in PDxN. Relationships between ATC level 1 and level 2 labels are presented. Only labels present in PDxN (q-value < 0.05) are listed. ATC — Anatomical Therapeutic Chemical; Lvl — level; PDxN — Pathway Drug Coexpression Network.

Lvl 1 code	Lvl 1 label	Lvl 2 code	Lvl 2 label
A	Alimentary tract and metabolism drugs	A01	Stomatological preparations
		A02	Drugs for acid related disorders
		A03	Drugs for functional gastrointestinal disorders
		A04	Antiemetics and antinauseants
		A05	Bile and liver therapy drugs
		A06	Drugs for constipation
		A07	Antidiarrheals, intestinal antiinflammatory/antiinfective agents
		A08	Antiobesity preparations, excl. diet products
		A10	Drugs used in diabetes
		A11	Vitamins
		A14	Anabolic agents for systemic use
		A16	Other alimentary tract and metabolism products
B	Blood and blood forming organ drugs	B01	Antithrombotic agents
		B02	Antihemorrhagics
		B03	Antianemic preparations
		B05	Blood substitutes and perfusion solutions
C	Cardiovascular system drugs	C01	Cardiac therapy drugs
		C02	Antihypertensives
		C03	Diuretics
		C04	Peripheral vasodilators
		C05	Vasoprotectives
		C07	Beta-adrenergic blocking agents
		C08	Calcium channel blockers
		C09	Agents acting on the renin-angiotensin system
		C10	Lipid modifying agents
D	Dermatologicals	D01	Antifungals for dermatological use
		D04	Antipruritics, incl. antihistamines, anesthetics, etc.
		D05	Antipsoriatics
		D06	Antibiotics and chemotherapeutics for dermatological use
		D07	Corticosteroids, dermatological preparations
		D08	Antiseptics and disinfectants
		D09	Medicated dressings
		D10	Anti-acne preparations
		D11	Other dermatological preparations
G	Genito urinary system and sex hormones	G01	Gynecological antiinfectives and antiseptics
		G02	Other gynecologicals
		G03	Sex hormones and modulators of the genital system
		G04	Urologicals
H	Systemic hormonal preparations, excl. sex hormones and insulins	H01	Pituitary and hypothalamic hormones and analogues

continues on the next page

Table C.3 continued

Lvl 1 code	Lvl 1 label	Lvl 2 code	Lvl 2 label
		H02	Corticosteroids for systemic use
		H03	Thyroid therapy drugs
		H05	Calcium homeostasis
J	Antiinfectives for systemic use	J01	Antibacterials for systemic use
		J02	Antimycotics for systemic use
		J04	Antimycobacterials
		J05	Antivirals for systemic use
L	Antineoplastic and immunomodulating agents	L01	Antineoplastic agents
		L02	Endocrine therapy antineoplastic and immunomodulating agents
		L03	Immunostimulants
		L04	Immunosuppressants
M	Musculo-skeletal system drugs	M01	Antiinflammatory and antirheumatic products
		M02	Topical products for joint and muscular pain
		M03	Muscle relaxants
		M04	Antigout preparations
		M05	Drugs for treatment of bone diseases
		M09	Other drugs for disorders of the musculo-skeletal system
N	Nervous system drugs	N01	Anesthetics
		N02	Analgesics
		N03	Antiepileptics
		N04	Anti-parkinson drugs
		N05	Psycholeptics
		N06	Psychoanaleptics
		N07	Other nervous system drugs
P	Antiparasitic products, insecticides and repellents	P01	Antiprotozoals
		P02	Anthelmintics
		P03	Ectoparasiticides, incl. scabicides, insecticides and repellents
R	Respiratory system drugs	R01	Nasal preparations
		R02	Throat preparations
		R03	Drugs for obstructive airway diseases
		R05	Cough and cold preparations
		R06	Antihistamines for systemic use
		R07	Other respiratory system products
S	Sensory organ drugs	S01	Ophthalmologicals
		S02	Otologicals
		S03	Ophthalmological and otological preparations
V	Various drug classes	V03	All other therapeutic products
		V04	Diagnostic agents
		V08	Contrast media
		V09	Diagnostic radiopharmaceuticals
		V10	Therapeutic radiopharmaceuticals

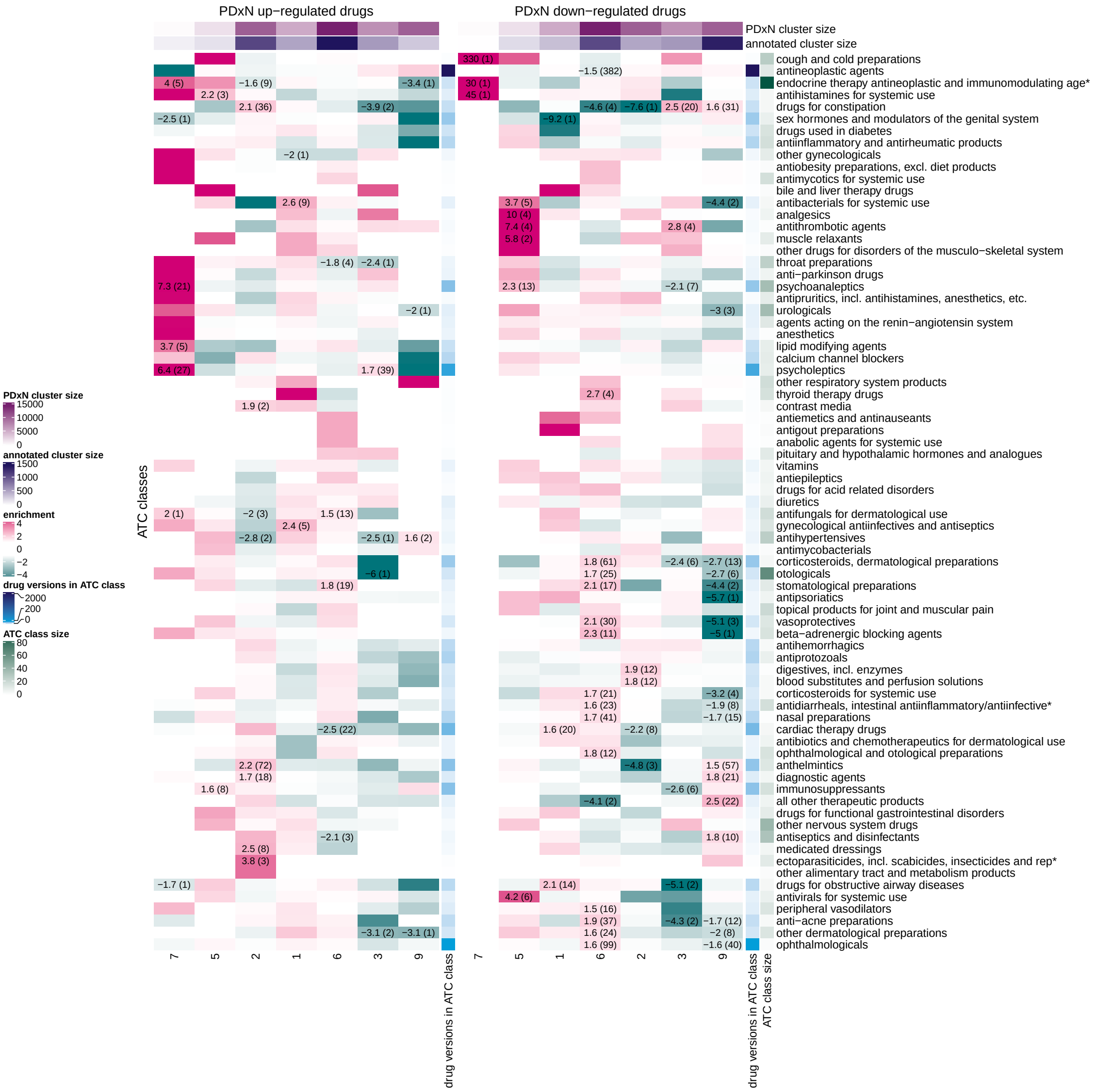


Fig. (C.1) ATC level 2 drug annotation enrichment of PDxN bipartite clusters. The heatmap summarises enrichment (pink) or depletion (teal) scores for drug nodes consisting of either up- or down-regulated genes. Absolute enrichment values of 0 or above 1.5 and below significance threshold p-value < 0.05 are displayed. The first number displayed is the enrichment or depletion score, while the number in () is the observed number. When the observed number is 0, the depletion score cannot be calculated, thus we displayed number-of-observed : number-of-expected terms in (). Row names are ATC level 2 drug annotation terms. Only clusters with at least one pathway group annotation are shown. ATC level 1 clustering can be seen in Chapter 5 Fig. 5.11. * — ATC term was shortened, full terms are listed in Supplementary Table C.3; ATC — Anatomical Therapeutic Chemical; PDxN — Pathway Drug Coexpression Network.

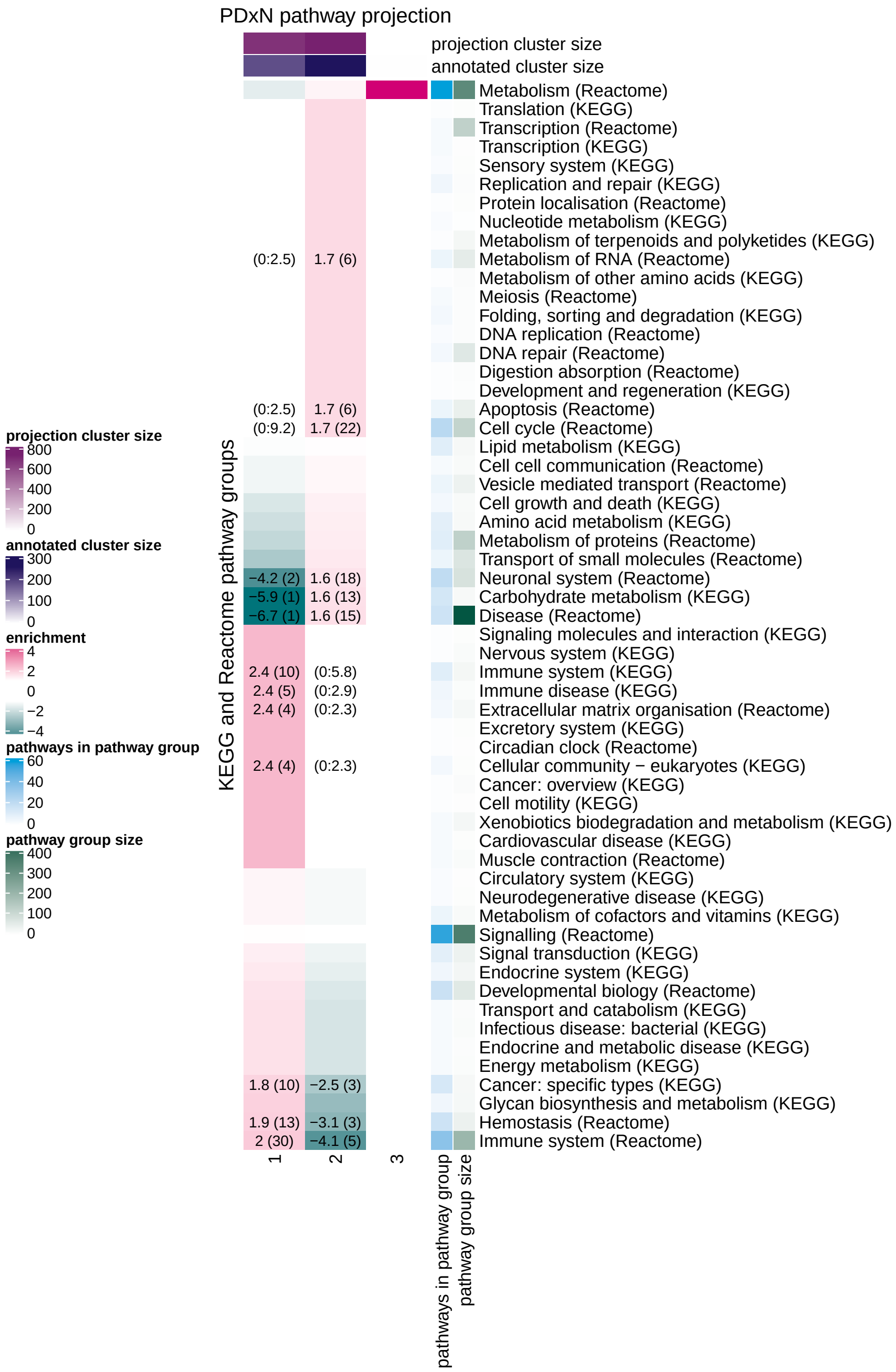


Fig. (C.2) KEGG and Reactome pathway annotation enrichment of PDxN pathway projection. The heatmap summarises enrichment (pink) or depletion (teal) scores for drug nodes consisting of either up- or down-regulated genes. Absolute enrichment values of 0 or above 1.5 and below significance threshold p-value < 0.05 are displayed. The first number displayed is the enrichment or depletion score, while the number in () is the observed number. When the observed number is 0, the depletion score cannot be calculated, thus we displayed number-of-observed : number-of-expected terms in (). Cluster numbers are not corresponding to PDxN bipartite clusters. Row names are KEGG and Reactome pathway group names. Only clusters with at least one pathway group annotation are shown. KEGG — Kyoto Encyclopedia of Genes and Genomes; PDxN — Pathway Drug Coexpression Network.

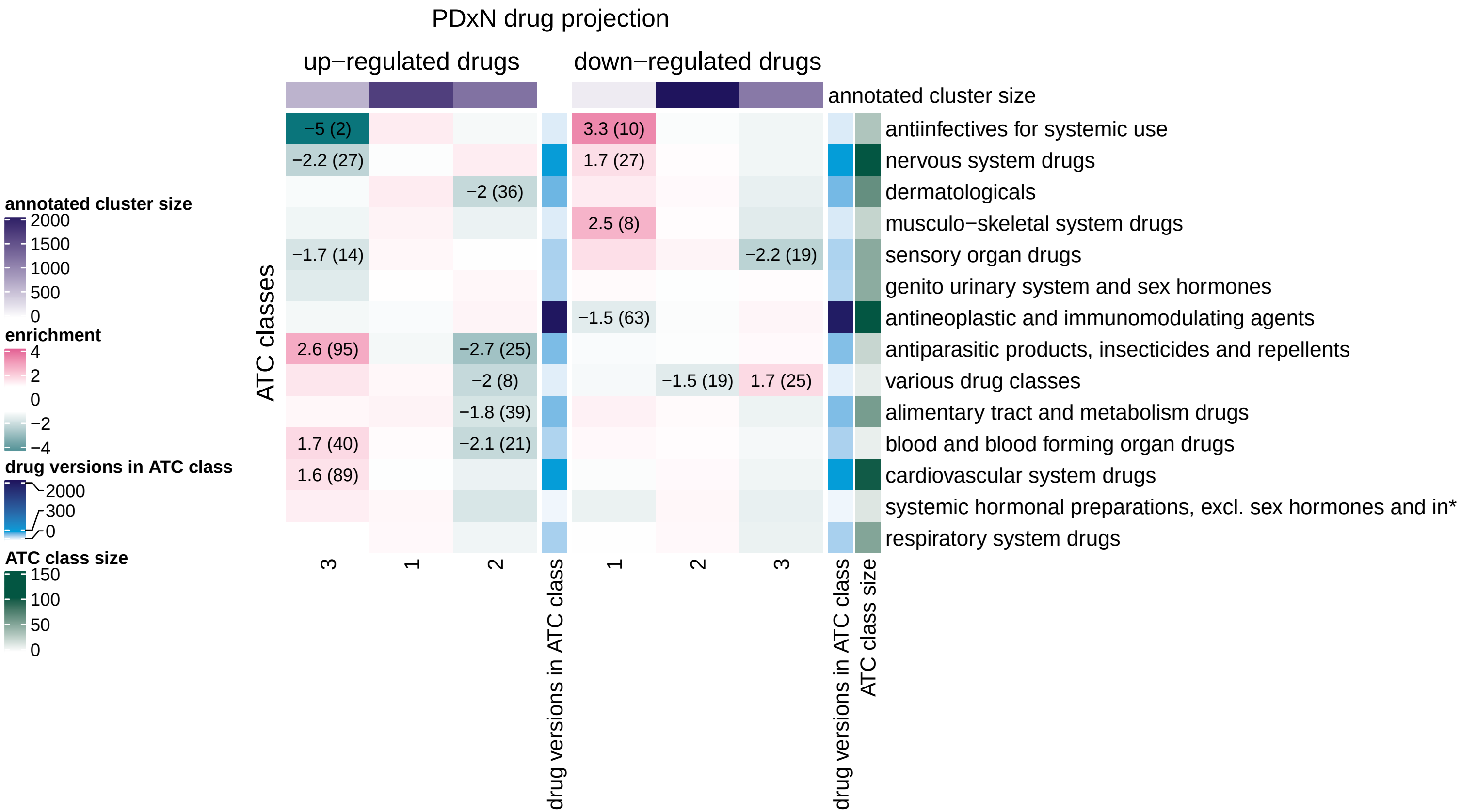


Fig. (C.3) ATC level 1 drug annotation enrichment of PDxN drug projections. The heatmap summarises enrichment (pink) or depletion (teal) scores. Absolute enrichment values of 0 or >1.5 (p-value < 0.05) are displayed. The first number displayed is the enrichment or depletion score, while the number in () is the observed number. When the observed number is 0, the depletion score cannot be calculated, thus we displayed number-of-observed : number-of-expected terms in (). Row names are ATC level 1 drug annotation terms. * — ATC term was shortened, full terms are listed in Supplementary Table C.3; ATC — Anatomical Therapeutic Chemical; PDxN — Pathway Drug Coexpression Network.

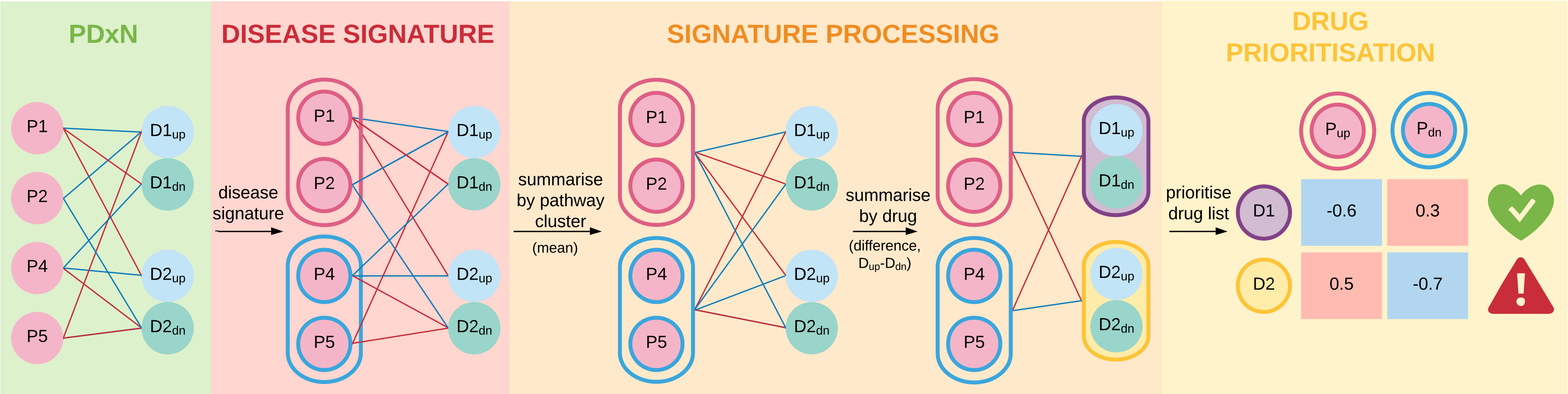


Fig. (C.4) Simplified example of edge summarisation through the pipeline. The pathway gene sets nodes (P, pink) in PDxN are grouped into two clusters, a cluster of up- and a cluster of down-regulated disease pathways. The correlation edges between each cluster pathway and drug node are first averaged across the pathway cluster. Followed by a summarisation step combining correlation edges for up- and down-regulated parts for each drug signature (D_{up}, blue and D_{down}, teal, respectively). The summary score for each pathway cluster↔drug pair is prioritised based on best predicted directionality outcomes in Fig. 5.15, where we predict that drugs with the most negative score for the up-regulated cluster and the most positive score for the down-regulated pathway cluster are best.

Appendix D

Supplementary material to Chapter 6 Evaluation of the System: Application to juvenile idiopathic arthritis (JIA)

Table (D.1) GSE7753 sJIA disease signature genes. The top 20 most up- (rank: 1 to 20) and down- (rank: -1 to -20) regulated genes (q-value < 0.05). Drugs were prioritised with LINCS clue.io for up- and down-regulated gene sets at: the top 20, 50, 100, 150 genes by decreasing log fold change (LogFC) for up- and decreasing for down-regulated genes. LINCS — Library of Integrated Network-based Cellular Signatures; sJIA — systemic juvenile idiopathic arthritis.

Rank	Gene	Description	LogFC	q-value
1	SLC25A37	solute carrier family 25 member 37	14.10	1.02e-06
2	HBA2	hemoglobin subunit alpha 2	13.70	1.95e-04
3	SNCA	synuclein alpha	9.91	1.83e-04
4	HBG2	hemoglobin subunit gamma 2	8.13	1.75e-04
5	GYPA	glycophorin A (MNS blood group)	7.32	9.21e-05
6	GYPB	glycophorin B (MNS blood group)	7.28	1.37e-04
7	TNS1	tensin 1	6.64	1.94e-04
8	ANK1	ankyrin 1	6.05	1.40e-04
9	FAM20A	FAM20A golgi associated secretory pathway pseudokinase	6.02	2.85e-11
10	MYL4	myosin light chain 4	5.77	5.37e-05
11	HBB	hemoglobin subunit beta	5.63	7.40e-04
12	SLC6A8	solute carrier family 6 member 8	5.55	6.36e-05
13	SESN3	sestrin 3	5.46	1.54e-03
14	SOX6	SRY-box 6	5.43	4.38e-04
15	CEACAM1	carcinoembryonic antigen related cell adhesion molecule 1	5.38	2.87e-05
16	MS4A4A	membrane spanning 4-domains A4A	5.38	3.81e-07
17	CR1	complement C3b/C4b receptor 1 (Knops blood group)	5.16	3.81e-07
18	SLC4A1	solute carrier family 4 member 1 (Diego blood group)	4.16	6.84e-05
19	RHD	Rh blood group D antigen	3.93	1.52e-03
20	ALAS2	5’-aminolevulinate synthase 2	3.78	1.63e-06

continues on the next page

Table D.1 continued

Rank	Gene	Description	LogFC	q-value
-1	BNC2	basonuclin 2	-3.64	2.07e-04
-2	ADAMTS5	ADAM metallopeptidase with thrombospondin type 1 motif 5	-3.48	9.75e-05
-3	BMPR1A	bone morphogenetic protein receptor type 1A	-3.16	3.15e-05
-4	RORA	RAR related orphan receptor A	-3.00	9.37e-05
-5	CAMTA1	calmodulin binding transcription activator 1	-2.44	2.11e-04
-6	SLC4A4	solute carrier family 4 member 4	-2.40	1.51e-04
-7	PRSS23	serine protease 23	-2.36	2.57e-04
-8	NKTR	natural killer cell triggering receptor	-2.28	3.94e-02
-9	DST	dystonin	-2.22	9.22e-04
-10	GPM6B	glycoprotein M6B	-2.16	7.49e-04
-11	KLRD1	killer cell lectin like receptor D1	-2.11	1.20e-03
-12	CLEC4C	C-type lectin domain family 4 member C	-2.07	2.27e-04
-13	PTGDR	prostaglandin D2 receptor	-1.96	1.25e-06
-14	PHLDB2	pleckstrin homology like domain family B member 2	-1.96	1.05e-05
-15	NEFL	neurofilament light	-1.95	2.24e-02
-16	TTC3	tetratricopeptide repeat domain 3	-1.95	1.99e-02
-17	AKT3	AKT serine/threonine kinase 3	-1.92	5.28e-04
-18	TPD52	tumor protein D52	-1.91	3.28e-02
-19	SYTL2	synaptotagmin like 2	-1.89	4.53e-05
-20	CCDC65	coiled-coil domain containing 65	-1.87	1.72e-04

Table (D.2) GSE112057 sJIA disease signature genes. The top 20 most up- (rank: 1 to 20) and down- (rank: -1 to -20) regulated genes (q-value < 0.05). Drugs were prioritised with LINCS clue.io for up- and down-regulated gene sets at: the top 20, 50, 100, 150 genes by decreasing log fold change (LogFC) for up- and decreasing for down-regulated genes. LINCS — Library of Integrated Network-based Cellular Signatures; sJIA — systemic juvenile idiopathic arthritis.

Rank	Gene	Description	LogFC	q-value
1	DAAM2	dishevelled associated activator of morphogenesis 2	1.250	0.03650
2	CD177	CD177 molecule	1.150	0.03850
3	C4BPA	complement component 4 binding protein alpha	1.080	0.02510
4	RAP1GAP	RAP1 GTPase activating protein	0.863	0.02310
5	ANKRD22	ankyrin repeat domain 22	0.833	0.01900
6	AOC1	amine oxidase copper containing 1	0.767	0.04490
7	SPATC1	spermatogenesis and centriole associated 1	0.670	0.03000
8	GPR84	G protein-coupled receptor 84	0.669	0.04870
9	KREMEN1	kringle containing transmembrane protein 1	0.620	0.01760
10	CD274	CD274 molecule	0.603	0.01090
11	HP	haptoglobin	0.573	0.03270
12	SLC1A3	solute carrier family 1 member 3	0.570	0.04960
13	SLC26A8	solute carrier family 26 member 8	0.563	0.01960
14	LRRN1	leucine rich repeat neuronal 1	0.555	0.01680
15	NSUN7	NOP2/Sun RNA methyltransferase family member 7	0.553	0.02230
16	ETV7	ETS variant 7	0.533	0.02290
17	KCNH7	potassium voltage-gated channel subfamily H member 7	0.527	0.03330
18	ST6GALNAC3	ST6 N-acetylgalactosaminide alpha-2,6-sialyltransferase 3	0.526	0.01750
19	FCGR1B	Fc fragment of IgG receptor Ib	0.510	0.02030
20	TRPM6	transient receptor potential cation channel subfamily M member 6	0.485	0.01180
-1	LGR6	leucine rich repeat containing G protein-coupled receptor 6	-0.886	0.00123

continues on the next page

Table D.1 continued

Rank	**Gene**	**Description**	**LogFC**	***q*-value**
-2	KRT72	keratin 72	-0.831	0.02210
-3	IGFBP3	insulin like growth factor binding protein 3	-0.810	0.01160
-4	LTK	leukocyte receptor tyrosine kinase	-0.777	0.00123
-5	SLC4A10	solute carrier family 4 member 10	-0.731	0.01090
-6	NSG1	neuronal vesicle trafficking associated 1	-0.712	0.01180
-7	RORC	RAR related orphan receptor C	-0.709	0.00229
-8	B3GAT1	beta-1,3-glucuronyltransferase 1	-0.706	0.00527
-9	NR4A1	nuclear receptor subfamily 4 group A member 1	-0.693	0.01740
-10	LGALS9B	galectin 9B	-0.634	0.04110
-11	DLG5	discs large MAGUK scaffold protein 5	-0.625	0.00123
-12	KRT73	keratin 73	-0.616	0.02080
-13	NEO1	neogenin 1	-0.606	0.00147
-14	DCANP1	dendritic cell associated nuclear protein	-0.604	0.01180
-15	BOK	BCL2 family apoptosis regulator BOK	-0.598	0.01480
-16	CLDND2	claudin domain containing 2	-0.597	0.00431
-17	FXYD7	FXYD domain containing ion transport regulator 7	-0.594	0.01950
-18	FEZ1	fasciculation and elongation protein zeta 1	-0.589	0.00248
-19	B3GALT2	beta-1,3-galactosyltransferase 2	-0.587	0.00660
-20	NMUR1	neuromedin U receptor 1	-0.584	0.00227

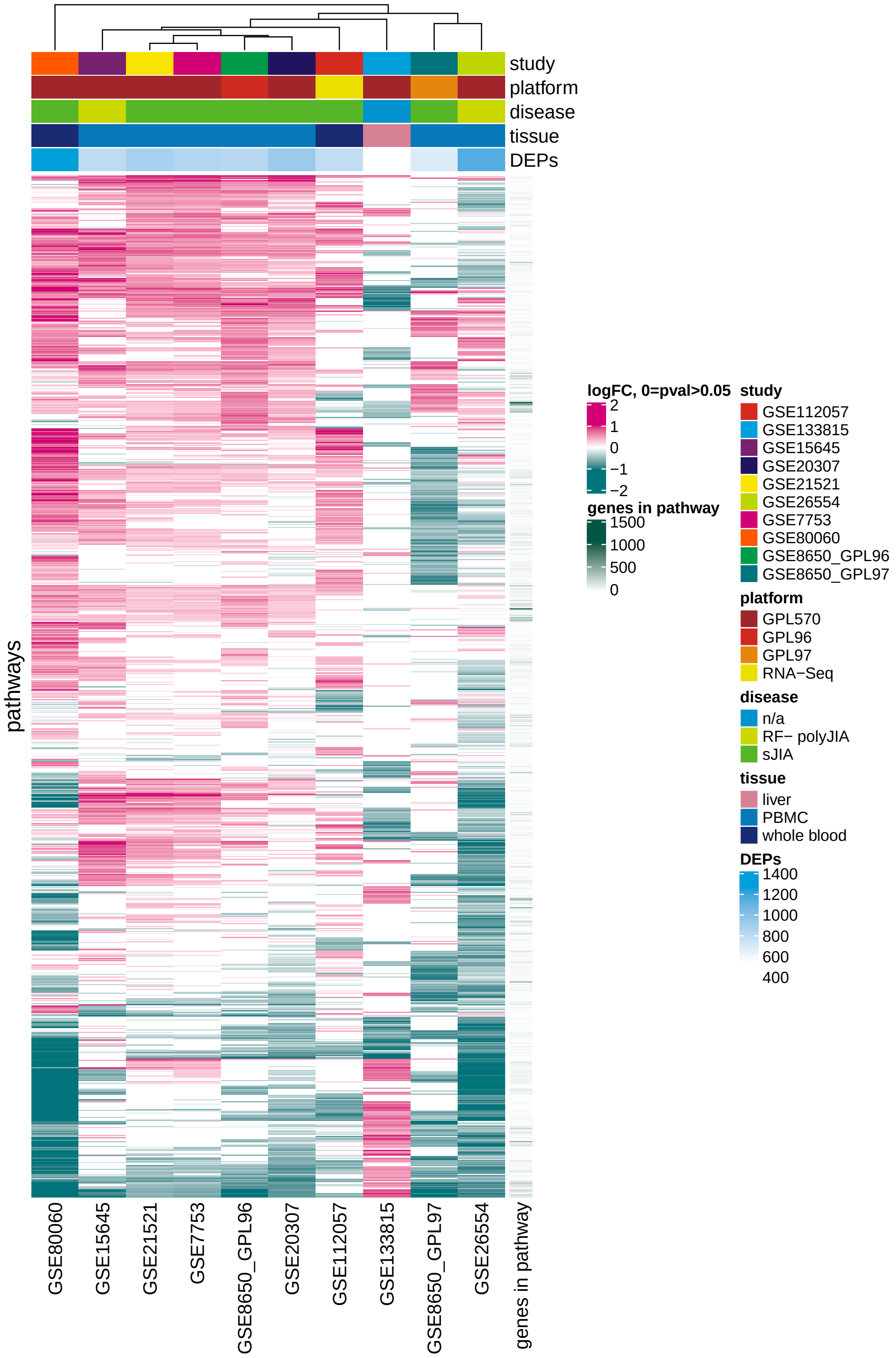


Fig. (D.1) JIA differential pathway expression profiles. Log$_2$ fold changes for differentially expressed pathways in 10 JIA and one non-JIA study. Columns represent studies, rows represent pathways from the PDxN pathway set. logFC values not meeting significance threshold p-value < 0.05 were assigned logFC = 0 (white). Note that logFC values > 2 or < -2 are coloured with colours for 2 (pink) and -2 (teal), respectively. DEPs — differentially expressed pathways, JIA — juvenile idiopathic arthritis; logFC — log$_2$ fold change; n/a — not applicable; PBMC — peripheral blood mononuclear cells; polyJIA — polyarticular JIA; pval — adjusted p-value, q-value; sJIA — systemic JIA.

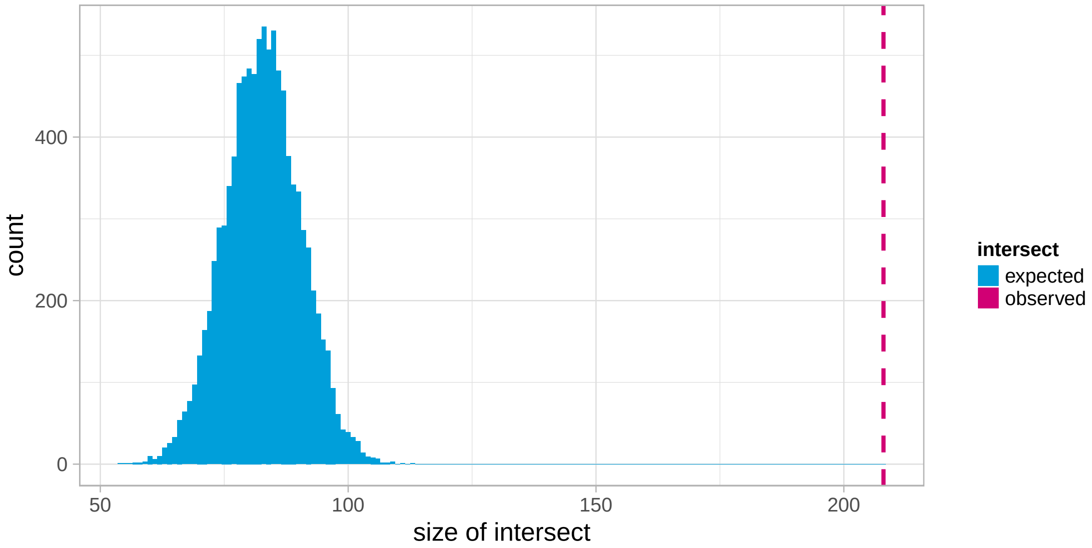


Fig. (D.2) Distribution of expected and observed size of overlap between sJIA differentially expressed pathways. We investigated the size of overlap from 5 PBMC sJIA studies. The expected range of intersect size from a random permutation test (number of permutations = 10,000) was 54-113 pathways (expected mean = 82.8), with the observed overlap size of 208 (*p*-value = 1e-04). Distribution of expected size of intersect by random permutation in blue. Pink line indicates the observed intersect size. JIA — juvenile idiopathic arthritis; PBMC — peripheral blood mononuclear cells; sJIA — systemic JIA.

Table (D.3) Number of drug signatures prioritised by sJIA disease pathway signatures. Cluster direction and size refer to direction and number of pathways in the disease signature used for drug prioritisation. sJIA — systemic juvenile idiopathic arthritis.

Cluster direction	Cluster size	GSE7753	GSE20307	GSE21521	GSE8650 (GPL96)	GSE8650 (GPL97)	sJIA overlap
up	5	3584	3467	3584	11150	12663	13263
up	10	4634	4523	5672	12252	14031	14112
up	15	7970	4667	12113	12326	14124	14113
up	20	12761	7151	14035	12477	14170	14542
down	5	3615	4380	2740	2811	2860	3966
down	10	5617	7057	5227	3380	3662	5170
down	15	7321	8150	6889	7114	6879	6613
down	20	8487	8645	9810	8888	7936	8516

Table (D.4) The top 10 drugs prioritised for GSE7753 pathway clusters. Results for top drugs prioritised based on top 5, 10, 15 and 20 up- and down-regulated pathways. * — drug name was mapped with KATdb.

Pathway cluster	Rank	Drug Name	Drug ID	Batch	Cell Type	Dose	Pert time
up 5	1	SERTRALINE HYDROCHLORIDE	BRD-K82036761	CPC015	MCF7	10.00	24
up 5	2	serdemetan	BRD-K60219430	CPC006	HT29	10.00	6
up 5	3	gemcitabine	BRD-K15108141	CPC006	VCAP	0.08	6
up 5	4	2-chloro-2-deoxyadenosine	BRD-K93034159	CPC020	MCF7	10.00	6
up 5	5	Clofarabine	BRD-A82371568	CPC016	MCF7	10.00	6
up 5	6	cytarabine*	BRD-K33106058	CPC011	A549	10.00	6
up 5	7	cytarabine*	BRD-K33106058	CPC011	PC3	10.00	6
up 5	8	wortmannin	BRD-A75409952	CPC005	HT29	10.00	24
up 5	9	2-chloro-2-deoxyadenosine	BRD-K93034159	CPC010	PC3	10.00	6
up 5	10	MLS002729057*	BRD-K78385490	CPC010	PC3	10.00	24
up 10	1	cytarabine*	BRD-K33106058	CPC011	A549	10.00	6
up 10	2	cytarabine*	BRD-K33106058	CPC011	PC3	10.00	6
up 10	3	4-Demethoxydaunorubicin hydrochloride (65)	BRD-A71390734	CPC006	A549	0.08	6
up 10	4	2-chloro-2-deoxyadenosine	BRD-K93034159	CPC010	A375	10.00	6
up 10	5	BML-259	BRD-K71799778	CPC006	MCF7	80.00	6
up 10	6	TW 37	BRD-K28360340	CPC006	HT29	10.00	6
up 10	7	PF 750	BRD-K83213911	CPC006	VCAP	80.00	6
up 10	8	BML-259	BRD-K71799778	CPC006	HCC515	80.00	6
up 10	9	4-Demethoxydaunorubicin hydrochloride (65)	BRD-A71390734	CPC006	VCAP	0.08	6
up 10	10	PIK-90	BRD-K99818283	CPC006	MCF7	10.00	24
up 15	1	SUGA1_008424*	BRD-K33164466	CPC013	HEPG2	10.00	6
up 15	2	Cytarabine	BRD-K33106058	CPC011	PC3	10.00	6
up 15	3	gemcitabine	BRD-K15108141	CPC006	SW620	0.08	6
up 15	4	cobaltous chloride*	BRD-K90864987	CPC020	HA1E	10.00	6
up 15	5	ALW-II-38-3	BRD-K68191783	CPC013	HEPG2	10.00	6
up 15	6	2541665-P1	BRD-K79382620	CPC006	SW620	11.10	6
up 15	7	FPA1_000240*	BRD-K37340241	CPC013	HEPG2	10.00	6
up 15	8	Lylamine hydrochloride	BRD-K62289640	CPC017	A549	10.00	6
up 15	9	10162	BRD-A67438293	CPC012	HA1E	10.00	6
up 15	10	wortmannin	BRD-A75409952	CPC007	HT29	10.00	24

continues on the next page

Table D.4 continued

Pathway cluster	Rank	Drug Name	Drug ID	Batch	Cell Type	Dose	Pert time
up 20	1	SUGA1_008424*	BRD-K33164466	CPC013	HEPG2	10.00	6
up 20	2	gemcitabine	BRD-K15108141	CPC006	SW620	0.08	6
up 20	3	S1018	BRD-K85402309	CPC014	HA1E	10.00	6
up 20	4	ALW-II-38-3	BRD-K68191783	CPC013	HEPG2	10.00	6
up 20	5	NVP-BEZ235	BRD-K12184916	CPC006	A673	0.63	6
up 20	6	2541665-P1	BRD-K79382620	CPC006	SW620	11.10	6
up 20	7	cobaltous chloride*	BRD-K90864987	CPC020	HA1E	10.00	6
up 20	8	FPA1_000240*	BRD-K37340241	CPC013	HEPG2	10.00	6
up 20	9	wortmannin	BRD-A75409952	CPC007	HT29	10.00	24
up 20	10	7910663	BRD-K03176945	CPC013	HEPG2	10.00	6
down 5	1	CP466722	BRD-K15592317	LJP005	A375	10.00	24
down 5	2	YM-201636	BRD-K48488978	LJP005	A549	0.37	24
down 5	3	ALW-II-38-3	BRD-K68191783	CPC013	HEPG2	10.00	6
down 5	4	withaferin-a	BRD-K88378636	LJP006	PC3	10.00	24
down 5	5	QL-XII-47	BRD-K99252563	LJP006	SKBR3	10.00	3
down 5	6	AZD-7762	BRD-K46056750	LJP006	HT29	3.33	24
down 5	7	gemcitabine	BRD-K15108141	CPC006	HA1E	0.08	6
down 5	8	BRL 54443	BRD-K17868609	CPC002	HA1E	10.00	6
down 5	9	GDC-0980	BRD-A18328003	LJP005	MDAMB231	1.11	24
down 5	10	Pepstatin A	BRD-K13571841	CPC005	A549	10.00	6
down 10	1	DIETHYLSTILBESTROI	BRD-K45330754	CPC004	VCAP	10.00	24
down 10	2	GDC-0980	BRD-A18328003	LJP005	MDAMB231	1.11	24
down 10	3	H5902	BRD-K15402119	CPC012	VCAP	10.00	24
down 10	4	MK-2206	BRD-K68065987	LJP006	BT20	0.12	3
down 10	5	SPECTRUM_000090*	BRD-A80151636	CPC015	MCF7	10.00	6
down 10	6	NCGC00182371-01	BRD-K44366801	CPC008	PC3	10.00	6
down 10	7	SPB02303	BRD-K99532291	CPC012	PC3	10.00	6
down 10	8	geldanamycin	BRD-A19500257	CPD001	MCF7	10.00	6
down 10	9	N-Benzylnaltrindole hydrochloride	BRD-A06276885	CPC016	HT29	10.00	6
down 10	10	amlodipine base	BRD-A64297288	CPC011	VCAP	10.00	6
down 15	1	H5902	BRD-K15402119	CPC012	VCAP	10.00	24
down 15	2	DIETHYLSTILBESTROI	BRD-K45330754	CPC004	VCAP	10.00	24
down 15	3	GDC-0980	BRD-A18328003	LJP005	MDAMB231	1.11	24
down 15	4	MK-2206	BRD-K68065987	LJP006	BT20	0.12	3
down 15	5	SPECTRUM_000090*	BRD-A80151636	CPC015	MCF7	10.00	6
down 15	6	amlodipine base	BRD-A64297288	CPC011	VCAP	10.00	6
down 15	7	SPB02303	BRD-K99532291	CPC012	PC3	10.00	6
down 15	8	geldanamycin	BRD-A19500257	CPD001	MCF7	10.00	6
down 15	9	NCGC00182371-01	BRD-K44366801	CPC008	PC3	10.00	6
down 15	10	NP-004527	BRD-K97951054	CPC012	VCAP	10.00	24
down 20	1	GDC-0980	BRD-A18328003	LJP005	MDAMB231	1.11	24
down 20	2	H5902	BRD-K15402119	CPC012	VCAP	10.00	24
down 20	3	DIETHYLSTILBESTROI	BRD-K45330754	CPC004	VCAP	10.00	24
down 20	4	MK-2206	BRD-K68065987	LJP006	BT20	0.12	3
down 20	5	SPECTRUM_000090*	BRD-A80151636	CPC015	MCF7	10.00	6
down 20	6	amlodipine base	BRD-A64297288	CPC011	VCAP	10.00	6
down 20	7	SPB02303	BRD-K99532291	CPC012	PC3	10.00	6
down 20	8	geldanamycin	BRD-A19500257	CPD001	MCF7	10.00	6
down 20	9	Doconexent*	BRD-K39965020	CPC018	A375	10.00	6
down 20	10	Compound 58	BRD-K80672993	CPC009	VCAP	10.00	24

Table (D.5) List of approved drugs for JIA. EMA and FDA approved list of drugs for treatment of JIA. Each drug can be present in multiple disease signatures under different experimental conditions (various combinations of drug id, concentration, cell line, perturbation time, and batch). One drug name can be mapped to many BRD IDs. Drug synonyms can be mapped to one BRD ID. Mapping from drug name to BRD ID was done with the KATdb app. BRD ID — Broad ID; EMA — European Medicines Agency; FDA — the US Food and Drug Administration; JIA — juvenile idiopathic arthritis

Drug name	BRD ID	Number of signatures
nordimet	BRD-K59456551	4
methotrexate	BRD-K59456551	4
jylamvo	BRD-K59456551	4
nordimet	BRD-A55424491	1
methotrexate	BRD-A55424491	1
jylamvo	BRD-A55424491	1
matever	BRD-K49404994	n/a
levetiracetam	BRD-K49404994	n/a
keppra	BRD-K49404994	n/a
stiripentol	BRD-A72441487	n/a
diacomit	BRD-A72441487	n/a
tocilizumab	n/a	n/a
simponi	n/a	n/a
roactemra	n/a	n/a
orencia	n/a	n/a
matever	n/a	n/a
lifmior	n/a	n/a
levetiracetam teva	n/a	n/a
levetiracetam sun	n/a	n/a
levetiracetam ratiopharm	n/a	n/a
levetiracetam hospira	n/a	n/a
levetiracetam actavis group	n/a	n/a
levetiracetam actavis	n/a	n/a
levetiracetam accord	n/a	n/a
kromeya	n/a	n/a
imraldi	n/a	n/a
ilaris	n/a	n/a
idacio	n/a	n/a
hyrimoz	n/a	n/a
humira	n/a	n/a
hefiya	n/a	n/a
halimatoz	n/a	n/a
golimumab	n/a	n/a
golimumab	n/a	n/a
etanercept	n/a	n/a
erelzi	n/a	n/a
enbrel	n/a	n/a
enbrel	n/a	n/a
canakinumab	n/a	n/a
benepali	n/a	n/a
amgevita	n/a	n/a
adalimumab	n/a	n/a
abatacept	n/a	n/a

Table (D.6) List of approved drugs for JIA and rheumatoid arthritis. EMA and FDA approved list of drugs for treatment of JIA or rheumatoid arthritis. Each drug can be present in multiple disease signatures under different experimental conditions (various combinations of drug id, concentration, cell line, perturbation time, and batch). One drug name can be mapped to many BRD IDs. Drug synonyms can be mapped to one BRD ID. Mapping from drug name to BRD ID was done with the KATdb app. BRD ID — Broad ID; EMA — European Medicines Agency; FDA — the US Food and Drug Administration; JIA — juvenile idiopathic arthritis.

Drug name	BRD ID	Number of signatures
auranofin	BRD-A79465854	31
cyclosporine	BRD-A38030642	30
cyclosporine	BRD-A69815203	14
cyclosporine	BRD-K80970344	9
cyclosporine	BRD-K13533483	9
dexamethasone	BRD-A35108200	7
betamethasone	BRD-A35108200	7
dexamethasone	BRD-A69951442	6
azathioprine	BRD-K32821942	5
prednisolone	BRD-A27887842	5
betamethasone	BRD-A02180903	5
nordimet	BRD-K59456551	4
methotrexate	BRD-K59456551	4
jylamvo	BRD-K59456551	4
dexamethasone	BRD-K38775274	4
meloxicam	BRD-A84174393	4
etodolac	BRD-A16998493	4
dexamethasone	BRD-A10188456	4
triamcinolone	BRD-K77554836	3
betamethasone	BRD-K39188321	3
diclofenac	BRD-K08252256	3
piroxicam	BRD-A57382968	3
triamcinolone	BRD-A37780065	3
leflunomide	BRD-K78692225	2
arava	BRD-K78692225	2
cortisone acetate	BRD-K43736954	2
methylprednisolone	BRD-K35240538	2
diflunisal	BRD-K22031190	2
celecoxib	BRD-K02637541	2
hydrocortisone	BRD-A75172220	2
hydrocortisone	BRD-A23290232	2
hydrocortisone	BRD-K93568044	1
cimzia	BRD-K88358234	1
certolizumab pegol	BRD-K88358234	1
cortisone acetate	BRD-K86161929	1
hydrocortisone	BRD-K73978287	1
nabumetone	BRD-K65146499	1
indomethacin	BRD-K57222227	1
hydrocortisone	BRD-K53342282	1
oxaprozin	BRD-K25394294	1
ketoprofen	BRD-A97739905	1
naproxen	BRD-A87719232	1
etodolac	BRD-A74667430	1
tiaprofenic acid	BRD-A72988804	1
nordimet	BRD-A55424491	1
methotrexate	BRD-A55424491	1
jylamvo	BRD-A55424491	1
fenoprofen	BRD-M61246020	n/a
prednisolone	BRD-K98039984	n/a
salicylic acid	BRD-K93632104	n/a

continues on the next page

Table D.6 continued

Drug name	BRD ID	Number of signatures
prednisone	BRD-K85883481	n/a
prednisone	BRD-K82624463	n/a
tolmetin	BRD-K82562631	n/a
prednisolone	BRD-K70504303	n/a
naproxen	BRD-K59197931	n/a
penicillamine	BRD-K58676198	n/a
olumiant	BRD-K53581288	n/a
baricitinib	BRD-K53581288	n/a
meclofenamic acid	BRD-K50398167	n/a
matever	BRD-K49404994	n/a
levetiracetam	BRD-K49404994	n/a
keppra	BRD-K49404994	n/a
salsalate	BRD-K48892307	n/a
auranofin	BRD-K45995181	n/a
xeljanz	BRD-K31283835	n/a
tofacitinib	BRD-K31283835	n/a
triamcinolone	BRD-K23714869	n/a
dexibuprofen	BRD-K14965640	n/a
meclofenamic acid	BRD-K13296708	n/a
hydrocortisone	BRD-K11612998	n/a
acetylsalicylic acid	BRD-K11433652	n/a
sulfasalazine	BRD-K10670311	n/a
acetylsalicylic acid	BRD-K07753030	n/a
betamethasone	BRD-K00835182	n/a
hydroxychloroquine	BRD-A99117172	n/a
chloroquine	BRD-A91699651	n/a
flurbiprofen	BRD-A86044036	n/a
fenoprofen	BRD-A81129465	n/a
stiripentol	BRD-A72441487	n/a
diacomit	BRD-A72441487	n/a
dexamethasone	BRD-A69951442-001-01-3	n/a
prednisone	BRD-A62525898	n/a
cortisone acetate	BRD-A54487287	n/a
loxoprofen	BRD-A43082555	n/a
teriflunomide	BRD-A42699921	n/a
ibuprofen	BRD-A17655518	n/a
sulindac	BRD-A13946108	n/a
sulindac	BRD-A03427350	n/a
zessly	n/a	n/a
truxima	n/a	n/a
tocilizumab	n/a	n/a
sodium aurothiomalate	n/a	n/a
simponi	n/a	n/a
sarilumab	n/a	n/a
sarilumab	n/a	n/a
roactemra	n/a	n/a
riximyo	n/a	n/a
rixathon	n/a	n/a
rituximab	n/a	n/a
remsima	n/a	n/a
remicade	n/a	n/a
orencia	n/a	n/a
matever	n/a	n/a
magnesium salicylate	n/a	n/a
mabthera	n/a	n/a
lithium	n/a	n/a
lithium	n/a	n/a
lifmior	n/a	n/a

continues on the next page

Table D.6 continued

Drug name	BRD ID	Number of signatures
levetiracetam teva	n/a	n/a
levetiracetam sun	n/a	n/a
levetiracetam ratiopharm	n/a	n/a
levetiracetam hospira	n/a	n/a
levetiracetam actavis group	n/a	n/a
levetiracetam actavis	n/a	n/a
levetiracetam accord	n/a	n/a
leflunomide zentiva (previously leflunomide winthrop)	n/a	n/a
leflunomide ratiopharm	n/a	n/a
leflunomide medac	n/a	n/a
kromeya	n/a	n/a
kineret	n/a	n/a
kevzara	n/a	n/a
infliximab	n/a	n/a
inflectra	n/a	n/a
imraldi	n/a	n/a
ilaris	n/a	n/a
idacio	n/a	n/a
hyrimoz	n/a	n/a
humira	n/a	n/a
hulio	n/a	n/a
hefiya	n/a	n/a
halimatoz	n/a	n/a
golimumab	n/a	n/a
golimumab	n/a	n/a
flixabi	n/a	n/a
etanercept	n/a	n/a
erelzi	n/a	n/a
enbrel	n/a	n/a
enbrel	n/a	n/a
certolizumab pegol	n/a	n/a
canakinumab	n/a	n/a
benepali	n/a	n/a
anakinra	n/a	n/a
amgevita	n/a	n/a
adalimumab	n/a	n/a
abatacept	n/a	n/a

Table (D.7) List of drugs in ATC class M01: Anti-inflammatory and antirheumatic products. Each drug can be present in multiple disease signatures under different experimental conditions (various combinations of drug id, concentration, cell line, perturbation time, and batch). One drug name can be mapped to many BRD IDs. Drug synonyms can be mapped to one BRD ID. ATC code can be mapped to BRD ID without also being mapped to the drug name. Only ATC codes that were mapped to a BRD ID are listed. Mapping from drug name to BRD ID was done with the KATdb app. ATC — Anatomical Therapeutic Chemical; BRD ID — Broad ID.

ATC code	Drug name	BRD ID	Number of signatures
M01CB03	auranofin	BRD-A79465854	31
M01AH03	valdecoxib	BRD-K12994359	21
M01AG03	n/a	BRD-K44067360	4
M01AC06	meloxicam	BRD-A84174393	4
M01AB15	ketorolac	BRD-A40639672	4
M01AB08	etodolac	BRD-A16998493	4

continues on the next page

Table D.7 continued

ATC code	Drug name	BRD ID	Number of signatures
M01AB05	diclofenac	BRD-K08252256	3
M01AC01	piroxicam	BRD-A57382968	3
M01AE17	n/a	BRD-K43764301	2
M01AH01	celecoxib	BRD-K02637541	2
M01AX01	nabumetone	BRD-K65146499	1
M01AB01	indometacin	BRD-K57222227	1
M01AE12	oxaprozin	BRD-K25394294	1
M01AH02	rofecoxib	BRD-K21733600	1
M01AB10	n/a	BRD-K16077845	1
M01AE05	n/a	BRD-K12513978	1
M01AE03	ketoprofen	BRD-A97739905	1
M01AE56	naproxen and misoprostol	BRD-A87719232	1
M01AE52	naproxen and esomeprazole	BRD-A87719232	1
M01AE02	naproxen	BRD-A87719232	1
M01AB08	etodolac	BRD-A74667430	1
M01AE11	n/a	BRD-A72988804	1
M01AE04	fenoprofen	BRD-M61246020	n/a
M01AX02	niflumic acid	BRD-K98763141	n/a
M01AG01	mefenamic acid	BRD-K92778217	n/a
M01AB03	tolmetin	BRD-K82562631	n/a
M01AX17	nimesulide	BRD-K76775527	n/a
M01AX07	n/a	BRD-K76133116	n/a
M01AX21	n/a	BRD-K69122748	n/a
M01AB16	n/a	BRD-K68538666	n/a
M01AB11	n/a	BRD-K67563174	n/a
M01AE56	naproxen and misoprostol	BRD-K59197931	n/a
M01AE52	naproxen and esomeprazole	BRD-K59197931	n/a
M01AE02	naproxen	BRD-K59197931	n/a
M01AH05	etoricoxib	BRD-K54770957	n/a
M01AG04 M02AA18	rimonabant	BRD-K50398167	n/a
M01AG04	meclofenamic acid	BRD-K50398167	n/a
M01AG02	n/a	BRD-K50133271	n/a
M01CB03	auranofin	BRD-K45995181	n/a
M01AB17	n/a	BRD-K36660044	n/a
M01AX07	n/a	BRD-K28542495	n/a
M01AE14	dexibuprofen	BRD-K14965640	n/a
M01AE	Propionic acid derivatives	BRD-K14965640	n/a
M01AH03	valdecoxib	BRD-K13800121	n/a
M01BA03	acetylsalicylic acid and corticosteroids	BRD-K11433652	n/a
M01AA01	phenylbutazone	BRD-K10843433	n/a
M01AE09	flurbiprofen	BRD-A86044036	n/a
M01AX04	n/a	BRD-A70182876	n/a
M01AE10	n/a	BRD-A44090213	n/a
M01AE07	suprofen	BRD-A34006693	n/a
M01AA03	oxyphenbutazone	BRD-A33749298	n/a
M01AC02	tenoxicam	BRD-A22844106	n/a
M01AE01	ibuprofen	BRD-A17655518	n/a
M01AB02	sulindac	BRD-A13946108	n/a
M01AB02	sulindac	BRD-A03427350	n/a

Table (D.8) List of drugs in ATC class L04A: Immunosuppressants. Each drug can be present in multiple disease signatures under different experimental conditions (various combinations of drug id, concentration, cell line, perturbation time, and batch). One drug name can be mapped to many BRD IDs. Drug synonyms can be mapped to one BRD ID. ATC code can be mapped to BRD ID without also being mapped to the drug name. Mapping from drug name to BRD ID was done with the KATdb app. ATC — Anatomical Therapeutic Chemical; BRD ID — Broad ID.

ATC code	Drug name	BRD ID	Number of signatures
L04AD01	ciclosporin	BRD-A38030642	30
L04AA40	n/a	BRD-K93034159	15
L04AA18	everolimus	BRD-K13514097	15
L04AD01	ciclosporin	BRD-A69815203	14
L04AD01	ciclosporin	BRD-K80970344	9
L04AD01	ciclosporin	BRD-K13533483	9
L04AA06	mycophenolic acid	BRD-K92428153	6
L04AA10	sirolimus	BRD-K84937637	6
L04AX02	thalidomide	BRD-A93255169	6
L04AX01	azathioprine	BRD-K32821942	5
L04AX03	methotrexate	BRD-K59456551	4
L04AD02	tacrolimus	BRD-K69608737	3
L04AA13	leflunomide	BRD-K78692225	2
L04AD02	tacrolimus	BRD-K44094599	2
L04AX04	lenalidomide	BRD-A17883755	2
L04AB05	n/a	BRD-K88358234	1
L04AX03	methotrexate	BRD-A55424491	1
L04AX05	n/a	BRD-K96862998	n/a
L04AA29	n/a	BRD-K31283835	n/a
L04AX04	lenalidomide	BRD-K05926469	n/a

Table (D.9) Number of drug signatures in PDxN (q-value < 0.05) per cell type with cell annotations. PDxN — Pathway Drug Coexpression Network.

Cell ID	Number of signatures	Tissue extraction site	Tissue type
A375	2048	skin	malignant melanoma
A549	1902	lung	non-small cell lung carcinoma
A673	96	bone/soft tissue around bone	Ewing sarcoma
AGS	82	stomach	gastric adenocarcinoma
ASC	315	adipose	adipose-derived mesenchymal stem cell
BT20	593	breast	invasive ductal carcinoma
CL34	31	colon	colon adenocarcinoma
CORL23	9	lung/pleural effusion	large cell lung carcinoma
COV644	18	ovary	ovarian carcinoma
DV90	25	lung/pleural effusion	lung adenocarcinoma
EFO27	55	omentum	ovarian mucinous adenocarcinoma
H1299	31	lymph node	large cell lung carcinoma
HA1E	2109	embryonic kidney	kidney epithelial immortalized
HCC15	47	lung	squamous cell lung carcinoma
HCC515	1774	lung	non-small cell lung adenocarcinoma
HCT116	74	colon	colon carcinoma
HEC108	44	uterus	Endometrial adenocarcinoma
HEPG2	1348	liver	hepatocellular carcinoma cell line, shown to be hepatoblastoma
HME1	394	breast	breast mammary immortalized
HS578T	603	breast	invasive ductal carcinoma
HT115	58	colon	colon carcinoma
HT29	1634	colon/large intestine	colorectal adenocarcinoma/rectosigmoid adenocarcinoma

continues on the next page

Table D.9 continued

Cell ID	Number of signatures	Tissue extraction site	Tissue type
JHUEM2	17	uterus	endometrial adenocarcinoma
LNCAP	250	prostate/left supraclavicular lymph node	prostate carcinoma
LOVO	81	colon/left supraclavicular lymph node	colon adenocarcinoma
MCF10A	556	breast	breast adenocarcinoma, immortalised line
MCF7	3177	breast	breast adenocarcinoma, invasive ductal carcinoma
MDAMB231	491	breast/pleural effusion	triple negative breast adenocarcinoma
MDST8	101	colon	colon carcinoma
NCIH1694	27	lung/ascites	small cell lung carcinoma
NCIH1836	29	lung	small cell lung carcinoma
NCIH2073	78	lung	lung adenocarcinoma
NCIH508	14	cecum/abdominal wall	cecum adenocarcinoma
NCIH596	73	lung	adenosquamous lung carcinoma
NEU	42	neuron	neuron cells primary terminally differentiated in-plate from NPC
NOMO1	19	blood	adult acute monocytic leukaemia
NPC	280	neuron	primary human iPS-derived neural progenitor cell line
OV7	34	ovary	ovarian carcinoma
PC3	2605	prostate	prostate adenocarcinoma
PHH	100	liver	primary human hepatocyte cells co-cultured with 3T3J2 mouse fibroblasts
PL21	102	blood	acute myeloid leukaemia
RKO	37	colon	colon carcinoma
RMGI	38	ovary/ascites	ovarian clear cell adenocarcinoma
RMUGS	73	ovary	ovarian mucinous cystadenocarcinoma
SKB	295	n/a	skeletal myoblast cells
SKBR3	583	breast	breast adenocarcinoma
SKLU1	60	lung	lung adenocarcinoma
SKM1	89	blood	adult acute myeloid leukaemia
SKMEL1	24	skin/thoracic lymph duct	melanoma
SKMEL28	58	skin	cutaneous melanoma
SNGM	50	uterus/obturator lymph node	endometrial adenocarcinoma
SNU1040	1	colon	colon adenocarcinoma
SNUC4	35	colon	colon adenocarcinoma
SNUC5	8	cecum	cecum adenocarcinoma
SW480	30	colon	colon adenocarcinoma
SW620	168	colon/lymph node	colon adenocarcinoma
SW948	136	colon	colon adenocarcinoma
T3M10	13	lung	large cell lung carcinoma
THP1	112	blood	acute monocytic leukaemia
TYKNU	17	ovary	high grade ovarian serous adenocarcinoma
U937	107	blood	adult acute monocytic leukaemia
VCAP	2131	prostate/vertebra	metastatic prostate cancer
WSUDLCL2	79	blood	diffuse large B-cell lymphoma

Appendix E

Supplementary material to Chapter 7 Case Studies: Neurodegenerative Diseases

Table (E.1) Number of drug signatures prioritised by neurodegenerative disease pathway signatures. Cluster direction and size refer to direction and number of pathways in the disease signature used for drug prioritisation. Lyso dysfun — lysosomal dysfunction; Mito dysfun — mitochondrial dysfunction.

Cluster direction	Cluster size	A5	H10	I47F	I45F	Mayo	Lyso dysfun	Mito dysfun	GCH1
up	5	8325	1620	4209	209	5660	6609	10139	1186
up	10	8678	3881	4330	822	11042	6644	10277	1958
up	15	9216	4080	4727	2794	11746	7638	12870	7980
up	20	11612	5699	12419	3175	15546	8987	12981	8950
down	5	4065	1554	2027	2271	710	3768	2872	4740
down	10	5012	6702	2233	5198	2281	4530	4437	4778
down	15	6938	7204	4493	7977	2937	5712	7133	5260
down	20	7364	10831	9815	8122	3174	10087	7348	6092

Table (E.2) The top 10 drugs prioritised for A5 3D cell model pathway clusters. Results for top drugs prioritised based on top 10 and 20 up- and down-regulated pathways. * — drug name was mapped with KATdb.

Pathway cluster	Rank	Drug Name	Drug ID	Batch	Cell Type	Dose	Pert time
up 10	1	BI 2536	BRD-K64890080	CPC006	SKM1	10.0	6
up 10	2	CARBAMAZEPINE	BRD-K71799949	CPC004	VCAP	10.0	6
up 10	3	TG101348	BRD-K12502280	CPC006	SKM1	11.1	6
up 10	4	torin-1	BRD-K40175214	CPC014	HEPG2	10.0	6
up 10	5	TG101348	BRD-K12502280	CPC006	U937	11.1	6
up 10	6	KU 0060648 trihydrochloride	BRD-K09499853	CPC006	WSUDLCL2	10.0	6
up 10	7	ISOLIQUIRITIGENIN	BRD-K33583600	CPC006	THP1	10.0	6
up 10	8	CAM-9-027-3	BRD-K45399554	CPC006	LOVO	10.0	6
up 10	9	S1042	BRD-M64432851	CPC014	VCAP	10.0	6
up 10	10	L-sulforophane	BRD-A58955223	CPC006	THP1	10.0	6
up 20	1	S-8599	BRD-K49810818	CPC013	A549	10.0	24
up 20	2	2-chloro-2-deoxyadenosine	BRD-K93034159	CPC010	MCF7	10.0	24
up 20	3	VU0410183-2	BRD-K74710236	CPC008	PC3	10.0	24
up 20	4	evista	BRD-K63828191	CPC020	MCF7	10.0	24
up 20	5	Hinokitiol	BRD-K37691127	CPC003	PC3	10.0	24
up 20	6	QL-X-138*	BRD-U33728988	CPC014	MCF7	10.0	24
up 20	7	AS-601245	BRD-A60245366	CPC014	MCF7	10.0	6
up 20	8	GR-103	BRD-K89085489	CPC014	PC3	10.0	24
up 20	9	BRD-K04695623*	BRD-K04695623	CPC019	PC3	10.0	24
up 20	10	Nutlin-3	BRD-A12230535	CPC006	HT29	44.4	24
down 10	1	H5902	BRD-K15402119	CPC012	VCAP	10.0	24
down 10	2	DIETHYLSTILBESTROL	BRD-K45330754	CPC004	VCAP	10.0	24
down 10	3	SPECTRUM_000090*	BRD-A80151636	CPC015	MCF7	10.0	6
down 10	4	MLS002607805*	BRD-A42737819	CPC009	VCAP	10.0	24
down 10	5	SCHEMBL2560033*	BRD-K17739445	CPC009	VCAP	10.0	24
down 10	6	SPB02303	BRD-K99532291	CPC012	PC3	10.0	6
down 10	7	geldanamycin	BRD-A19500257	CPD001	MCF7	10.0	6
down 10	8	OXIBENDAZOLE	BRD-K52075715	CPC004	VCAP	10.0	24
down 10	9	Triazolothiadiazine, 28*	BRD-K96704648	CPC009	A549	10.0	6
down 10	10	amlodipine base	BRD-A64297288	CPC011	VCAP	10.0	6
down 20	1	H5902	BRD-K15402119	CPC012	VCAP	10.0	24
down 20	2	DIETHYLSTILBESTROL	BRD-K45330754	CPC004	VCAP	10.0	24
down 20	3	SPECTRUM_000090*	BRD-A80151636	CPC015	MCF7	10.0	6
down 20	4	3,5-dichloro-2-hydroxy-N-(2-methoxy-5-phenylphenyl)benzenesulfonamide	BRD-K43620258	CPC006	THP1	80.0	6
down 20	5	SPB02303	BRD-K99532291	CPC012	PC3	10.0	6
down 20	6	geldanamycin	BRD-A19500257	CPD001	MCF7	10.0	6
down 20	7	Daunorubicin hydrochloride	BRD-A68009927	CPC015	A375	10.0	6
down 20	8	amlodipine base	BRD-A64297288	CPC011	VCAP	10.0	6
down 20	9	Methapyrilene hydrochloride	BRD-K47323024	CPD001	MCF7	10.0	6
down 20	10	SCHEMBL2560033*	BRD-K17739445	CPC009	VCAP	10.0	24

Table (E.3) The top 10 drugs prioritised for AD Mayo dataset pathway clusters. Results for top drugs prioritised based on top 10 and 20 up- and down-regulated pathways. * — drug name was mapped with KATdb.

Pathway cluster	Rank	Drug Name	Drug ID	Batch	Cell Type	Dose	Pert time
up 10	1	MLS002264465*	BRD-K42499654	CPC009	MCF7	10.00	6
up 10	2	L-6307	BRD-K23192422	CPC014	SKB	10.00	24
up 10	3	FCCP	BRD-K14821540	CPC017	HEPG2	10.00	6
up 10	4	LUPANINE PERCHLORATE	BRD-A92826379	CPC004	VCAP	10.00	24
up 10	5	nadolol	BRD-A87606379	CPC020	A549	10.00	6
up 10	6	Proscillaridin A	BRD-A34806832	CPC015	ASC	10.00	24
up 10	7	NP-004102	BRD-A14178283	CPC013	SKB	10.00	24
up 10	8	penfluridol	BRD-K15409150	CPC006	H1299	30.00	6
up 10	9	NCGC00183401-01	BRD-K95080525	CPC007	HA1E	10.00	6
up 10	10	BAS 02002358*	BRD-A83255679	CPC006	HA1E	20.00	6
up 20	1	PENFLURIDOL	BRD-K15409150	CPC015	A549	10.00	24
up 20	2	JNK-9L	BRD-K19220233	CPC014	A549	10.00	6
up 20	3	GBR 13069 dihydrochloride	BRD-K11634954	CPC001	HA1E	10.00	24
up 20	4	JAK3 Inhibitor II	BRD-K52850071	CPC017	A549	10.00	24
up 20	5	PROMAZINE HYDROCHLORIDE	BRD-K06980535	CPC006	HA1E	10.00	24
up 20	6	Mibefradil dihydrochloride	BRD-K09549677	CPC001	VCAP	10.00	24
up 20	7	BAS 02002358*	BRD-A83255679	CPC006	HA1E	20.00	6
up 20	8	dibenzyline	BRD-A67799922	CPC020	A549	10.00	6
up 20	9	MLS002264403*	BRD-A75301702	CPC009	VCAP	10.00	6
up 20	10	CHEMBL2135524*	BRD-K98834634	CPC019	VCAP	10.00	6
down 10	1	GSK-461364	BRD-K92428232	LJP008	HT29	0.12	24
down 10	2	Hoechst 33342 (cell permeable) (BisBenzimide)	BRD-K08554278	CPC003	HA1E	10.00	6
down 10	3	NCGC00241726-01	BRD-K05979026	CPC008	A549	10.00	6
down 10	4	NCGC00182609-01	BRD-K95858622	CPC008	A549	10.00	6
down 10	5	"3-cyclohexyl-6-4-[3-(trifluoromethyl)phenyl]piperazin-1-ylpyrimidine-2,4(1H,3H)-dione"	BRD-K77547509	CPC008	HEPG2	10.00	6
down 10	6	MLS003530001*	BRD-K05549170	CPC008	A549	10.00	6
down 10	7	528116.cdx	BRD-K49371609	CPC006	VCAP	0.09	6
down 10	8	"4-[(1-methyl-2-oxo-1,2-dihydroquinolin-4-yl)oxy]-N-(4-methylpyridin-2-yl)butanamide"	BRD-K84203638	CPC007	A549	10.00	6
down 10	9	AG14361	BRD-K00615600	CPC006	HT29	25.00	6
down 10	10	cercosporin	BRD-A78360835	CPC005	HT29	10.00	24
down 20	1	GSK-461364	BRD-K92428232	LJP008	HT29	0.12	24
down 20	2	MGCD-265	BRD-K56277358	LJP009	A549	10.00	24
down 20	3	mitoxantrone	BRD-K21680192	LJP006	HCC515	1.11	24
down 20	4	YM-155	BRD-K76703230	CPC006	PC3	0.31	24
down 20	5	NCGC00241077-01	BRD-A31946439	CPC008	A375	10.00	6
down 20	6	SCHEMBL15444220*	BRD-K83336168	CPC013	MCF7	10.00	24
down 20	7	MLS003329219*	BRD-K26304855	CPC009	HCC515	10.00	6
down 20	8	gefitinib	BRD-K64052750	LJP006	MCF10A	1.11	3
down 20	9	auranofin	BRD-A79465854	CPC006	A375	10.00	6
down 20	10	prucalopride*	BRD-A36630025	CPC006	SKM1	0.35	6

Table (E.4) List of approved drugs for AD. EMA and FDA approved list of drugs for treatment of AD. Each drug can be present in multiple disease signatures under different experimental conditions (various combinations of drug id, concentration, cell line, perturbation time, and batch). One drug name can be mapped to many BRD IDs. Drug synonyms can be mapped to one BRD ID. Mapping from drug name to BRD ID was done with the KATdb app. AD — Alzheimer's disease; BRD ID — Broad ID; EMA — European Medicines Agency; FDA — the US Food and Drug Administration.

Drug name	BRD ID	Number of signatures
donepezil	BRD-A49160188	3
nemdatine	BRD-A79803969	2
memantine hydrochloride	BRD-A79803969	2
memantine	BRD-A79803969	2
ebixa	BRD-A79803969	2
axura	BRD-A79803969	2
nemdatine	BRD-K91938660	n/a
memantine hydrochloride	BRD-K91938660	n/a
memantine	BRD-K91938660	n/a
ebixa	BRD-K91938660	n/a
axura	BRD-K91938660	n/a
galantamine	BRD-K49481516	n/a
rivastigmine	BRD-K10706131	n/a
prometax	BRD-K10706131	n/a
nimvastid	BRD-K10706131	n/a
exelon	BRD-K10706131	n/a
vizamyl	n/a	n/a
rivastigmine sandoz	n/a	n/a
rivastigmine hexal	n/a	n/a
rivastigmine actavis	n/a	n/a
rivastigmine 1 a pharma	n/a	n/a
prometax	n/a	n/a
nimvastid	n/a	n/a
neuraceq	n/a	n/a
nemdatine	n/a	n/a
memantine ratiopharm	n/a	n/a
memantine mylan	n/a	n/a
memantine merz	n/a	n/a
memantine lek	n/a	n/a
memantine hydrochloride	n/a	n/a
memantine accord	n/a	n/a
memantine	n/a	n/a
marixino (previously maruxa)	n/a	n/a
ioflupane (123l)	n/a	n/a
flutemetamol (18f)	n/a	n/a
florbetapir (18f)	n/a	n/a
florbetaben (18f)	n/a	n/a
ebixa	n/a	n/a
datscan	n/a	n/a
amyvid	n/a	n/a

Table (E.5) The top 10 drugs prioritised for sporadic Parkinson's disease lysosomal dysfunction pathway clusters. Results for top drugs prioritised based on top 10 and 20 up- and down-regulated pathways. * — drug name was mapped with KATdb.

Pathway cluster	Rank	Drug Name	Drug ID	Batch	Cell Type	Dose	Pert time
up 10	1	HYDROQUINIDINE	BRD-A06390036	CPC015	A375	10.00	6
up 10	2	CGP-60474	BRD-K79090631	LJP006	LNCAP	1.11	3
up 10	3	EI-346	BRD-U08759356	CPC014	SKB	10.00	24
up 10	4	HY-11001	BRD-K64800655	CPC014	SKB	10.00	24
up 10	5	5284616	BRD-K89626439	CPC012	MCF7	10.00	6
up 10	6	MENADIONE	BRD-K78126613	CPC006	PC3	10.00	24
up 10	7	BRD-K60067222*	BRD-K60067222	CPC019	PC3	10.00	6
up 10	8	U 18666A	BRD-A81795050	CPC016	SKB	10.00	24
up 10	9	1-benzylimidazole	BRD-K32795028	CPC010	VCAP	10.00	6
up 10	10	Metergoline	BRD-A30435184	CPC005	MCF7	10.00	6
up 20	1	cephaeline*	BRD-K80348542	CPC002	HA1E	10.00	24
up 20	2	Homoharringtonine	BRD-K76674262	CPC001	HA1E	10.00	24
up 20	3	EMETINE	BRD-A25687296	CPC004	HA1E	10.00	24
up 20	4	HYDROQUINIDINE	BRD-A06390036	CPC015	A375	10.00	6
up 20	5	CGP-60474	BRD-K79090631	LJP006	LNCAP	1.11	3
up 20	6	HY-11001	BRD-K64800655	CPC014	SKB	10.00	24
up 20	7	5284616	BRD-K89626439	CPC012	MCF7	10.00	6
up 20	8	MENADIONE	BRD-K78126613	CPC006	PC3	10.00	24
up 20	9	BRD-K60067222*	BRD-K60067222	CPC019	PC3	10.00	6
up 20	10	U 18666A	BRD-A81795050	CPC016	SKB	10.00	24
down 10	1	DIETHYLSTILBESTROL	BRD-K45330754	CPC004	VCAP	10.00	24
down 10	2	H5902	BRD-K15402119	CPC012	VCAP	10.00	24
down 10	3	SPECTRUM_000090*	BRD-A80151636	CPC015	MCF7	10.00	6
down 10	4	geldanamycin	BRD-A19500257	CPD001	MCF7	10.00	6
down 10	5	SPB02303	BRD-K99532291	CPC012	PC3	10.00	6
down 10	6	Doconexent*	BRD-K39965020	CPC018	A375	10.00	6
down 10	7	amlodipine base	BRD-A64297288	CPC011	VCAP	10.00	6
down 10	8	Triazolothiadiazine, 28*	BRD-K96704648	CPC009	A549	10.00	6
down 10	9	Akt inhibitor X	BRD-K70792160	CPC006	HCC515	24.00	6
down 10	10	NP-004527	BRD-K97951054	CPC012	VCAP	10.00	24
down 20	1	H5902	BRD-K15402119	CPC012	VCAP	10.00	24
down 20	2	DIETHYLSTILBESTROL	BRD-K45330754	CPC004	VCAP	10.00	24
down 20	3	SPECTRUM_000090*	BRD-A80151636	CPC015	MCF7	10.00	6
down 20	4	trichostatin A	BRD-A19037878	CPC015	MCF7	10.00	6
down 20	5	amlodipine base	BRD-A64297288	CPC011	VCAP	10.00	6
down 20	6	SPB02303	BRD-K99532291	CPC012	PC3	10.00	6
down 20	7	ST019366	BRD-A38275906	CPC013	SKB	10.00	24
down 20	8	vorinostat	BRD-K81418486	CPC006	RMGI	11.10	6
down 20	9	geldanamycin	BRD-A19500257	CPD001	MCF7	10.00	6
down 20	10	8-[2-oxo-2-(1-pyrrolidinyl)ethyl]thioquinoline	BRD-K36772364	CPC010	HCC515	10.00	6

Table (E.6) List of approved drugs for PD. EMA and FDA approved list of drugs for treatment of PD. Each drug can be present in multiple disease signatures under different experimental conditions (various combinations of drug id, concentration, cell line, perturbation time, and batch). One drug name can be mapped to many BRD IDs. Drug synonyms can be mapped to one BRD ID. Mapping from drug name to BRD ID was done with the KATdb app. BRD ID — Broad ID; EMA — European Medicines Agency; FDA — the US Food and Drug Administration; PD — Parkinson's disease.

Drug name	BRD ID	Number of signatures
carbidopa	BRD-A69512159	4
metixene	BRD-A33711280	2
selegiline	BRD-K86434416	1
carbidopa	BRD-K78712176	1
ropinirole	BRD-K15933101	1
procyclidine	BRD-A31800922	1
biperiden	BRD-A00546892	1
entacapone	BRD-K92977333	n/a
comtess	BRD-K92977333	n/a
comtan	BRD-K92977333	n/a
xadago	BRD-K92613113	n/a
safinamide	BRD-K92613113	n/a
rotigotine	BRD-K91111634	n/a
neupro	BRD-K91111634	n/a
leganto	BRD-K91111634	n/a
entacapone	BRD-K83636919	n/a
comtess	BRD-K83636919	n/a
comtan	BRD-K83636919	n/a
apomorphine	BRD-K76022557	n/a
amantadine	BRD-K70330367	n/a
benzatropine	BRD-K68804560	n/a
pergolide	BRD-K60770992	n/a
rasagiline	BRD-K58114536	n/a
azilect	BRD-K58114536	n/a
dopamine	BRD-K43887077	n/a
stalevo	BRD-K34730807	n/a
levodopa / carbidopa / entacapone	BRD-K34730807	n/a
levodopa	BRD-K34730807	n/a
corbilta (previously levodopa/carbidopa/entacapone sandoz)	BRD-K34730807	n/a
bromocriptine	BRD-K14496212	n/a
tolcapone	BRD-K10852020	n/a
tasmar	BRD-K10852020	n/a
rivastigmine	BRD-K10706131	n/a
prometax	BRD-K10706131	n/a
nimvastid	BRD-K10706131	n/a
exelon	BRD-K10706131	n/a
pramipexole	BRD-K06388322	n/a
oprymea	BRD-K06388322	n/a
mirapexin	BRD-K06388322	n/a
norepinephrine	BRD-A91555231	n/a
orphenadrine	BRD-A53576514	n/a
trihexyphenidyl	BRD-A48180038	n/a
selegiline	BRD-A28545468	n/a
benzatropine	BRD-A04322457	n/a
stalevo	n/a	n/a

continues on the next page

Table E.6 continued

Drug name	BRD ID	Number of signatures
stalevo	n/a	n/a
sifrol	n/a	n/a
rivastigmine sandoz	n/a	n/a
rivastigmine hexal	n/a	n/a
rivastigmine actavis	n/a	n/a
rivastigmine 1 a pharma	n/a	n/a
rasagiline ratiopharm	n/a	n/a
rasagiline mylan	n/a	n/a
prometax	n/a	n/a
pramipexole teva	n/a	n/a
pramipexole accord	n/a	n/a
oprymea	n/a	n/a
opicapone	n/a	n/a
ongentys	n/a	n/a
numient	n/a	n/a
nimvastid	n/a	n/a
neupro	n/a	n/a
mirapexin	n/a	n/a
levodopa/carbidopa/entacapone orion	n/a	n/a
levodopa / carbidopa	n/a	n/a
ioflupane (123I)	n/a	n/a
entacapone teva	n/a	n/a
entacapone orion	n/a	n/a
droxidopa	n/a	n/a
datscan	n/a	n/a
cycrimine	n/a	n/a
comtess	n/a	n/a
benzatropine	n/a	n/a

Table (E.7) List of neuroprotective drugs for PD. Expert-curated (Professor Oliver Bandmann, University of Sheffield) list of predicted neuroprotective drugs. Each drug can be present in multiple disease signatures under different experimental conditions (various combinations of drug id, concentration, cell line, perturbation time, and batch). One drug name can be mapped to many BRD IDs. Drug synonyms can be mapped to one BRD ID. Mapping from drug name to BRD ID was done with the KATdb app. BRD ID — Broad ID; PD — Parkinson's disease.

Drug name	BRD ID	Number of signatures
nilotinib	BRD-K81528515	25
atorvastatin	BRD-U88459701	7
azathioprine	BRD-K32821942	5
nifedipine	BRD-K96354014	4
atorvastatin	BRD-K69726342	1
ambroxol	BRD-K56558538	1
atorvastatin	BRD-A82307304	1
nifedipine	BRD-A30977374	n/a
exenatide	n/a	n/a

Data Acknowledgements

I would like to thank my collaborators Prof. Oliver Bandmann from the University of Sheffield and Asst. Prof. Doo Yeon Kim from Harvard Medical School for trusting me with their data.

In addition to thanking funding bodies that support and require public sharing of data, I wish to make my final thanks to every patient and volunteer who participated in any of the studies used in this thesis.

The results published here are in part based on data obtained from the AMP-AD Knowledge Portal (doi:10.7303/syn2580853). Study data were provided by the following sources: The Mayo Clinic Alzheimer's Disease Genetic Studies, led by Dr. Nilufer Taner and Dr. Steven G. Younkin, Mayo Clinic, Jacksonville, FL using samples from the Mayo Clinic Study of Aging, the Mayo Clinic Alzheimer's Disease Research Center, and the Mayo Clinic Brain Bank. Data collection was supported through funding by NIA grants P50 AG016574, R01 AG032990, U01 AG046139, R01 AG018023, U01 AG006576, U01 AG006786, R01 AG025711, R01 AG017216, R01 AG003949, NINDS grant R01 NS080820, CurePSP Foundation, and support from Mayo Foundation. Study data includes samples collected through the Sun Health Research Institute Brain and Body Donation Program of Sun City, Arizona. The Brain and Body Donation Program is supported by the National Institute of Neurological Disorders and Stroke (U24 NS072026 National Brain and Tissue Resource for Parkinson's Disease and Related Disorders), the National Institute on Aging (P30 AG19610 Arizona Alzheimer's Disease Core Center), the Arizona Department of Health Services (contract 211002, Arizona Alzheimer's Research Center), the Arizona Biomedical Research Commission (contracts 4001, 0011, 05-901 and 1001 to the Arizona Parkinson's Disease Consortium) and the Michael J. Fox Foundation for Parkinson's Research.